Contents lists available at ScienceDirect

Genomics Data

Data in brief

Generating and evaluating a ranked candidate gene list for potential vertebrate heart field regulators

G. Musso [a], C. Mosimann [b,c,d,e], D. Panáková [f], A. Burger [e], Y. Zhou [b,c,d], L.I. Zon [b,c,d], C.A. MacRae [a,*]

[a] Cardiovascular Division, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA
[b] Howard Hughes Medical Institute, Boston, MA 02115, USA
[c] Stem Cell Program, Boston Children's Hospital, MA 02115, USA
[d] Division of Hematology/Oncology, Boston Children's Hospital, Harvard Stem Cell Institute, Harvard Medical School, Boston, MA 02115, USA
[e] Institute of Molecular Life Sciences (IMLS), University of Zürich, 8057 Zürich, Switzerland
[f] Max-Delbrück Center for Molecular Medicine (MDC), 13125 Berlin, Buch, Germany

ARTICLE INFO

ABSTRACT

The vertebrate heart develops from two distinct lineages of cardiomyocytes that arise from the first and second heart fields (FHF and SHF, respectively). The FHF forms the primitive heart tube, while adding cells from the SHF allows elongation at both poles of the tube. Initially seen as an exclusive characteristic of higher vertebrates, recent work has demonstrated the presence of a distinct FHF and SHF in lower vertebrates, including zebrafish. We found that key transcription factors that regulate septation and chamber formation in higher vertebrates, including Tbx5 and Pitx2, influence relative FHF and SHF contributions to the zebrafish heart tube. To identify molecular modulators of heart field migration, we used microarray-based expression profiling following inhibition of tbx5a and pitx2ab in embryonic zebrafish (Mosimann & Panakova, et al, 2015; GSE70750). Here, we describe in more detail the procedure used to process, prioritize, and analyze the expression data for functional enrichment.

Specifications	
Organism/cell line/tissue	Zebrafish
Sex	Mix
Sequencer or array type	Affymetrix array (Zebrafish Gene 1.0 ST Array)
Data format	Raw (CEL Files)
Experimental factors	Comparison of whole-embryo transcript expression from embryos injected with morpholinos targeting pitx2ab or tbx5a, versus control embryos
Experimental features	Extracted RNA from at least 40 embryos per experimental condition (in triplicate) was used to generate cDNA libraries that were hybridized to an Affymetrix array. Data were subsequently background subtracted, normalized, and analyzed for enrichment of functional gene sets using Gene Set Enrichment Analysis (GSEA).
Consent	NA
Sample source location	All samples were generated using embryos from fish colonies maintained at the Children's Hospital zebrafish facility, Boston, MA, USA

1. Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70750.

* Corresponding author.

2. Experimental design

Value of the data.

– The raw data files linked here provide unique expression profiles from two conditions highly relevant to vertebrate developmental biology.
– The scripts included in the supplement not only provide the means to replicate the referenced study, but also are immediately applicable to any dataset available in the Gene Expression Omnibus [1] (GEO).
– The resulting ranked list is a resource of potential markers of the cardiomyocyte lineages.

3. Data, materials and methods

3.1. Generation of microarray expression data

To generate the expression data accompanying our initial study [2], adult wild-type (AB) zebrafish were kept in breeding cages overnight with dividers separating males and females, then allowed to breed the following morning. Resulting embryos were collected over a short

time window (approximately 30 minutes), and injected into the yolk with a morpholino targeting either *tbx5a* (5-CCTGTACGATGTCTACCG TGAGGC-3; as previously designed [3]) or *pitx2ab* (5′-TGGGAGTCCATT TAGTAGGTTATAT-3′) before the 4-cell developmental stage using a standard microinjection platform. At least 40 embryos were injected per morpholino treatment, with the experiment being performed in

triplicate (at least 120 embryos total per group). At 56 hours post fertilization embryos were visually inspected for efficiency of morpholino-based knockdown by assessing their resulting developmental phenotypes; *tbx5a* morphants lack pectoral fins and develop cardiac edema, while *pitx2ab* morphants display left-right asymmetry defects (*i. e.* heart looping randomized). Occasional wildtype-looking and thus
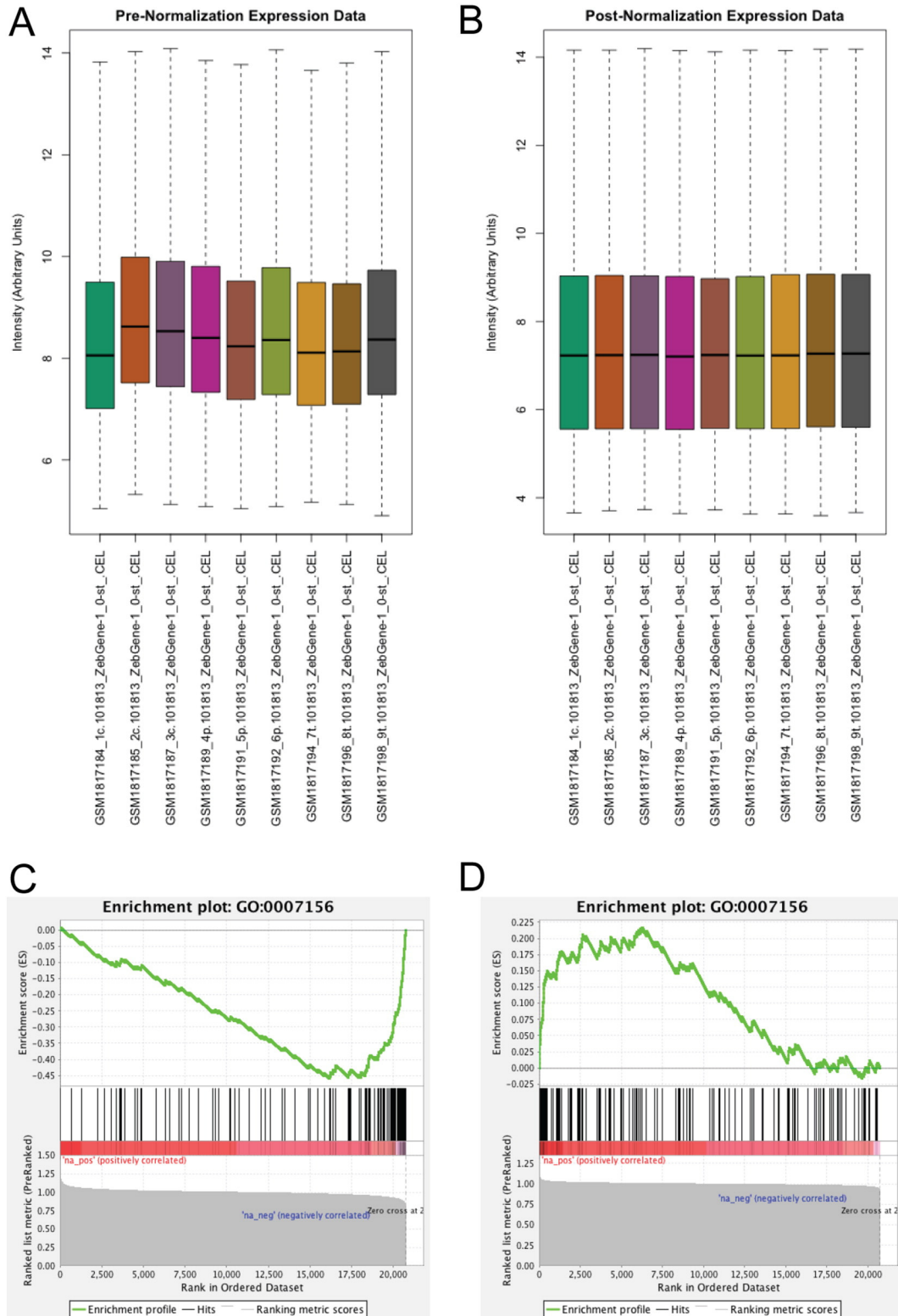
Fig. 1. Normalization and GSEA enrichment analysis results. Boxplots (A & B) show the expression intensity values for each CEL file both before (A) and after (B) RMA normalization. Plots outputted by GSEA (C & D) show enrichment for genes annotated as being involved in homophilic cell adhesion following *pitx2* (C) and *tbx5* (D) knockdown.

possibly inefficiently injected embryos in clutches of morpholino-injected cohorts were manually sorted out. Embryos were homogenized in Trizol LS (Life Technologies) using a tissue ruptor (Qiagen), and mRNA extracted using a standard phenol/chloroform extraction protocol as per manufacturer's instructions with an extra chloroform cleanup step thereafter. Resulting total RNA was checked for quality using a Bioanalyzer nano RNA chip, and 100 ng total RNA went into a standard Affymetrix pipeline for probe synthesis, labeling, hybridization and scanning of the hybridized array at the Boston Children's Hospital Molecular Genetics Core.

3.2. Normalization and processing of microarray data

Raw CEL files were processed using the Oligo package [4] as part of the Bioconductor suite (www.bioconductor.org) in the R statistical framework (www.r-project.org), using a customized script (Supplementary Script 1). Specifically, background subtraction and normalization were performed using the Robust Multiarray Average (RMA) method implemented in the Oligo package. Boxplots of intensity values were compared for all chips before and after normalization to visualize the corresponding effects on mean and quartile values (Fig. 1A–B). Following normalization, probeset IDs were matched to corresponding transcript IDs. Specifically, the zebrafish 1.0 ST array NetAffx annotation file was downloaded in CSV format from the Affymetrix website (www.affymetrix.com), and transcript/gene IDs corresponding to given probe IDs were extracted (*Supplementary Table 1*). In the interest of representing the data using a single gene annotation framework, all transcript and gene IDs were mapped to corresponding Ensemble gene IDs. For Ensemble transcript IDs, corresponding Ensemble gene IDs were obtained using the BioMart community portal [5]. The Synergizer web application [6] was used to convert gene IDs from other annotation frameworks to Ensemble gene IDs. The resulting table (*Supplementary Table 2*) was then merged with the table of probeset expression data. In instances where a gene ID matched multiple probesets, probeset values were averaged (see Supplementary Script 1).

3.3. Examination for highly expressed subgroups using GSEA pre-ranked analysis

Transcript data were averaged for each gene over the three experimental replicates. Genes were then ranked by fold-change in experimental conditions: *pitx2ab* morpholino injected divided by control, and *tbx5a* morpholino injected divided by control. These pre-ranked datasets were used as the input for gene set enrichment analysis [7] (GSEA). GSEA pre-ranked analysis also requires a list of gene sets that will be examined for enrichment at the top or bottom of the ranked expression datasets. Gene Ontology [8] (GO) annotations were used for this purpose. GO SLIM biological process, molecular function, and cellular compartment annotations for zebrafish were downloaded from the

BioMart community portal [5] and converted to GSEA-compatible GMT format using a Perl script (Supplementary Script 2). GSEA pre-ranked analysis (1000 permutations, minimum term size of 5, maximum term size of 500) was implemented using the stand-alone GSEA tool [2]. GSEA analysis found genes annotated with the term 'homophilic cell adhesion' (GO:0007156) as being significantly (q < 0.001) down regulated in expression following inhibition of pitx2, and up regulated following inhibition of tbx5 (Fig. 1C–D). This reinforced the notion that *pitx2* and *tbx5* control alternate aspects of cardiac cell adhesion and migration, and provided a novel group of genes for experimental follow-up that, notably, would not have been identified through traditional univariate statistical analysis. Ongoing experimental work is aimed at experimentally determining how downstream targets of *pitx2* and *tbx5* mitigate homophlic (like-to-like) cell adhesion, and the consequent physiological effects on cellular function.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2015.09.015.

References

[1] T. Barrett, R. Edgar, Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Methods Enzymol. 411 (2006) 352–369.
[2] C. Mosimann, D. Panakova, A.A. Werdich, G. Musso, A. Burger, K.L. Lawson, L.A. Carr, K.R. Nevis, M.K. Sabeh, Y. Zhou, A.J. Davidson, A. DiBiase, C.E. Burns, C.G. Burns, C.A. MacRae, et al., Chamber identity programs drive early functional partitioning of the heart. Nat. Commun. 6 (2015) 8146.
[3] D.G. Ahn, M.J. Kourakis, L.A. Rohde, L.M. Silver, R.K. Ho, T-box gene tbx5 is essential for formation of the pectoral limb bud. Nature 417 (2002) 754–758.
[4] B.S. Carvalho, R.A. Irizarry, A framework for oligonucleotide microarray preprocessing. Bioinformatics 26 (2010) 2363–2367.
[5] D. Smedley, S. Haider, S. Durinck, L. Pandini, P. Provero, J. Allen, O. Arnaiz, M.H. Awedh, R. Baldock, G. Barbiera, P. Bardou, T. Beck, A. Blake, M. Bonierbale, A.J. Brookes, et al., The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. 43 (2015) W589–598.
[6] G.F. Berriz, F.P. Roth, The Synergizer service for translating gene, protein and other biological identifiers. Bioinformatics 24 (2008) 2272–2273.
[7] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, J.P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 15545–15550.
[8] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, et al., Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat. Genet. 25 (2000) 25–29.