SEMIPs: Structural Equation Modeling of In silico Perturbations

**Jianying Li1,2,3** †**, Pierre R. Bushel3,4** †**, Lin Lin5, Kevin Day6, Tianyuan Wang1,2, Francesco J. DeMayo7, San-Pin Wu7 \*, and Jian-Liang Li1 \***

1 Integrative Bioinformatics, Epigenetics and Stem Cell Biology Laboratory, Division of Intramural Research, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA

2 Kelly Government Solutions, Research Triangle Park, NC 27709, USA

3 Massive Genome Informatics Group, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA

4 Biostatistics and Computational Biology Branch, Division of Intramural Research, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA

5 Department of Family Health Care Nursing, University of California at San Francisco, San Francisco, CA 94143, USA

6 Duke University, Durham NC 27713

7 Reproductive and Developmental Biology Laboratory, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA

†These authors have contributed equally to this work and share first authorship

**\* Correspondence:**Jian-Liang Li ([jianliang.li@nih.gov](mailto:jianliang.li@nih.gov))

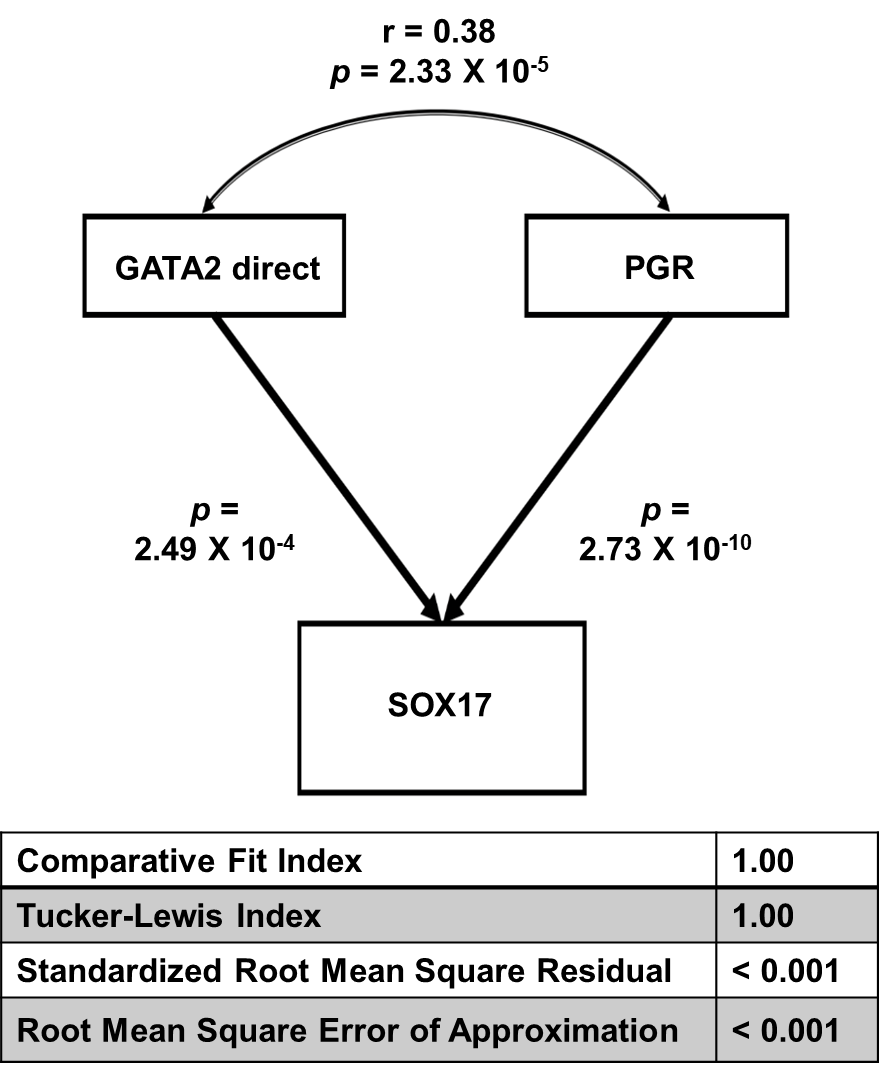
San-Pin Wu ([steve.wu@nih.gov](mailto:steve.wu@nih.gov))

Keywords: Structural Equation Modeling, gene expression, *in silico* perturbation, molecular interaction, R, Shiny

Diagram

Description automatically generated

**Supplemental Figure 1**. A two-class (elimination with or without replacement) bootstrap resampling simulation. From the initial GATA2 significant gene list in the yellow rectangle, the same number of genes as that of the subset of genes (represented by the white oval shape inside the yellow rectangle) are eliminated either without replacement (left side) or with replacement other than those in the subset” (right side). The resulting shrunken GATA2 gene list or reduced GATA2 then restored by the same number of irrelevant genes are tested in the SEM model. The simulation can be repeated for a large “number of bootstraps” to generate a non-parametric distribution for statistics inference.



**Supplemental Figure 2**. Model fit statistics for joint regulation of the SOX17 gene expression levels by GATA2 and PGR activities in the GEO accession: GSE58144 dataset using SEM. “GATA2 direct” depicts GATA2 activities that were derived from the GATA2 direct downstream targets.

**Supplemental Methods**

**Gene list preparation**

The microarray gene expression data was analyzed using The Partek Genomics Suite 7.17 software (Partek Inc., St. Louis, MO). The Robust Multichip Analysis (RMA) algorithm with quantile for normalization and log2 transformation was applied to generate gene expression values of all samples. The one-way analysis of variance (ANOVA) model was used to compare expression profiles from different groups. Differentially expressed genes (DEGs) were identified using the filters of ANOVA unadjusted p value < 0.01 and absolute fold change >1.3.

The published GATA2 occupancy information GEO accession: GSE40659 (Rubel et al. 2016) was first lifted from mm9 to mm10 genome assembly and then annotated by HOMER (Heinz et al. 2010) for the nearby genes. The obtained GATA2 ChIP-seq targets were mapped to the GATA2 signature from microarray data to identify the putative GATA2 direct downstream targets (GATA2 direct signature - Supplemental Table 1). The criteria used to selected GATA2 ChIP-seq targets was GATA2 binding at immediate promoter regions (+/-2kb of TSS).

**The main steps to follow the use case example**

**Step 1.** **To get the T score**: Users can launch the App and import the 634 genes list (Supplemental Table 1) and HumanArray4Shiny comes with the App. By clicking the green “Go” button, the corresponding T score will then be calculated and can be download (shown in Supplemental Figure 3). We also provided this calculated T score in Supplemental Table 2.

**Step 2.** **To construct the dataset**: Users need to open the \_sampleDAT.txt under the “app\_installation\_dir/dataSEM/”, i.e. /Users/li11/myGit/SEMIPs/dataSEM, append the new T score column from step 1 and name the header accordingly. We use “GATA2 Direct” in this use case. Please save the new file as “app\_installation\_dir/dataSEM/sampleDAT.txt”.

**Step 3.**  **To run the SEM model**: Users need to re-launch the app. Under the SEM tab, from the drop-down list select “GATA2 Direct”, “PGR\_act\_FC13\_P01”, and “SOX17\_lev” as show in Supplemental Figure U. Then the structural equation model will be fitted accordingly. User can download the 3-node SEM image as well as the model fitting details as shown in Supplemental Figure 4.

Graphical user interface, table

Description automatically generated

**Supplemental Figure 3**. An illustration for using the App to calculate T-score for Supplemental Table 1.

Graphical user interface, text, application

Description automatically generated

**Supplemental Figure 4**. An illustration for using the App to fit the structural equation model for Supplemental Table 2 (GATA2 direct gene list). The fitting statistics can be downloaded by clicking the “Download Zip” button.

**References**

Heinz, S., et al. (2010). "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities." Mol Cell **38**(4): 576-589.

Rubel, C. A., et al. (2016). "A Gata2-Dependent Transcription Network Regulates Uterine Progesterone Responsiveness and Endometrial Function." Cell Rep **17**(5): 1414-1425.