SEMIPs: Structural Equation Modeling of In silico Perturbations

Supplemental Information

Jianying Li[1,2,3,8], Pierre R. Bushel[3,4,8], Lin Lin[5], Kevin Day[6], Tianyuan Wang[1,2], Francesco J. DeMayo[7], San-Pin Wu[7 *], and Jian-Liang Li[1 *]

[1] Integrative Bioinformatics, Epigenetics and Stem Cell Biology Branch, 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

[8]These authors contributed equally

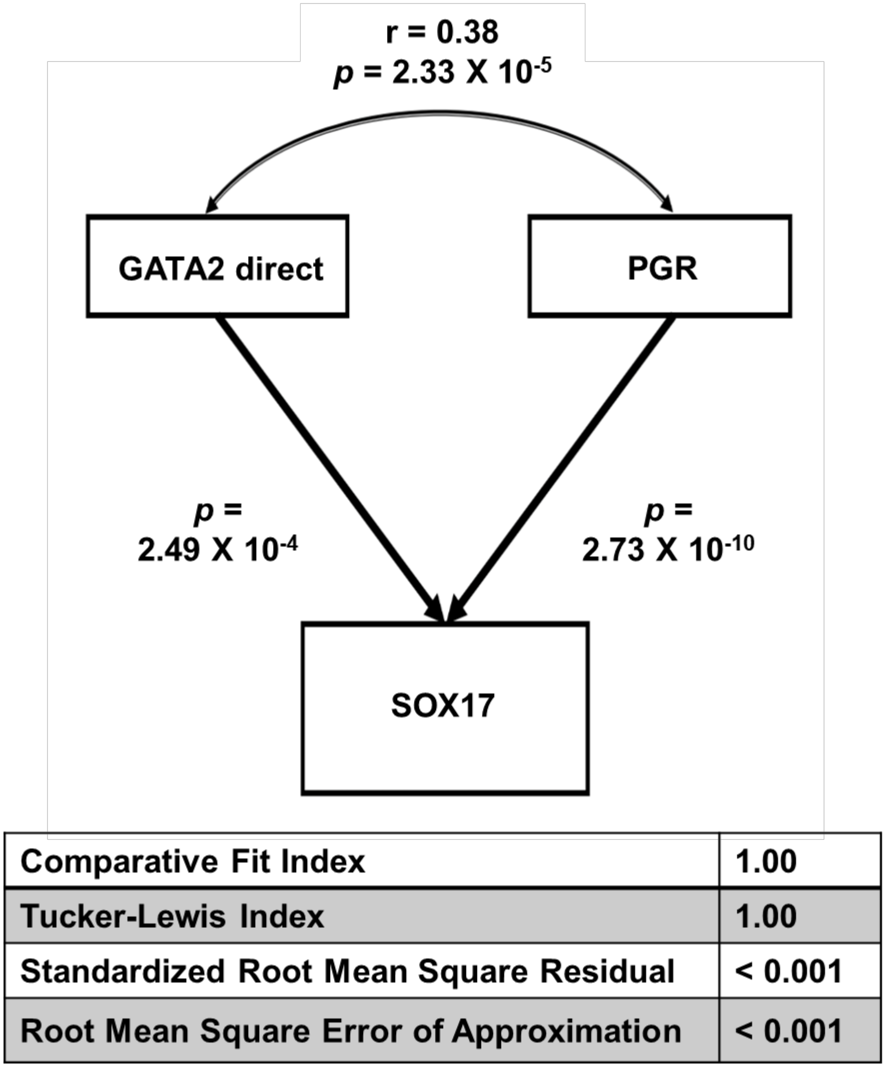
* Correspondence
Jian-Liang Li (jianliang.li@nih.gov)
San-Pin Wu (steve.wu@nih.gov)

Overview

This SEMIPs RShiny App allows users to compute a two-sided t-statistic (T score) from gene expression data to infer the activities of genes of interest in a quantitative manner. This app also provides a 3-node model fitting function using structural equation modeling to test the joint regulation of a target gene by two upstream regulators *in silico*. In addition, for hypothesis generation purposes, a two-way bootstrap method, elimination with replacement or elimination without replacement, is included in the app to examine the impact of removing genes that belong to the same signaling cascade from the downstream targets of the gene of interest. As an example, here we applied SEMIPs to evaluate latent gene interactions that mediate the progesterone signaling in the uterus for female fertility.

A user case application

Previously we demonstrated that the mouse gene signatures of GATA2 and PGR allow inference of the interaction between GATA2 and PGR for regulation of SOX17 expression in the human endometrial tissues (Rubel et al. 2016). The full GATA2 gene signature consists of both direct and indirect downstream genes of GATA2 in the uterus (Rubel et al. 2016). Since GATA2 is known as a transcription factor that occupies cis-acting elements and confers genomic actions, we hypothesize that expression levels of GATA2's direct downstream targets reflect its activities *in silico.* Here, a GATA2 direct downstream target is defined as a GATA2 regulated gene with GATA2 genome occupancy within 2-kilobase vicinity of the said gene's transcription start site in the uterus (Gene Expression Omnibus (GEO) accession: GSE40659, (Rubel et al. 2016)). This stringent criterion led to the identification of 634 genes (Supplemental Table 1), which is termed "GATA2 direct signature". The GATA2 activity, as represented by the GATA2 direct signature in a T-score, was quantified by the SEMIPs app from gene expression data of the endometrium tissue for each individual human subject (GEO accession: GSE58144, (Koot et al. 2016)). T scores for the uterine GATA2 in all 115 patients were calculated by the app with the GATA2 direct signature and the data matrix of GEO accession: GSE58144 (Supplemental Table 2). Similarly, T scores for the uterine PGR (termed PGR signature) were obtained using the GEO accession: GSE39920 dataset (Rubel et al. 2016) on the same data matrix via the application's T score calculation function. To test whether the GATA2 direct signature fits the model of the 3-node PGR-GATA2-SOX17 genetic network, the application was fed with T scores of GATA2 direct signature and PGR signature as exogenous variables and the SOX17 expression levels as the endogenous variable under the "SEM" function. The output data shows that, with GATA2 direct signature in place of the full gene signature, the model significantly fits the GEO accession: GSE58144 dataset with all proposed paths (Supplemental Figure 1) and this model is considered not rejected by the human data. This finding suggests that the expression levels of GATA2 direct downstream targets, a subset of the full GATA2 regulated genes, can mathematically serve as surrogate reporters of the GATA2 activities in the human endometrium tissues, which supports our hypothesis. Results of this analysis not only reduce the number of reporter genes for GATA2 activities to 634, but also implicate possibilities of a further reduction with additional filtering criteria on the gene list. A small and manageable panel of markers for GATA2 activities could serve as a future diagnostic tool for pregnancy failure (Díaz-Gimeno et al. 2011).

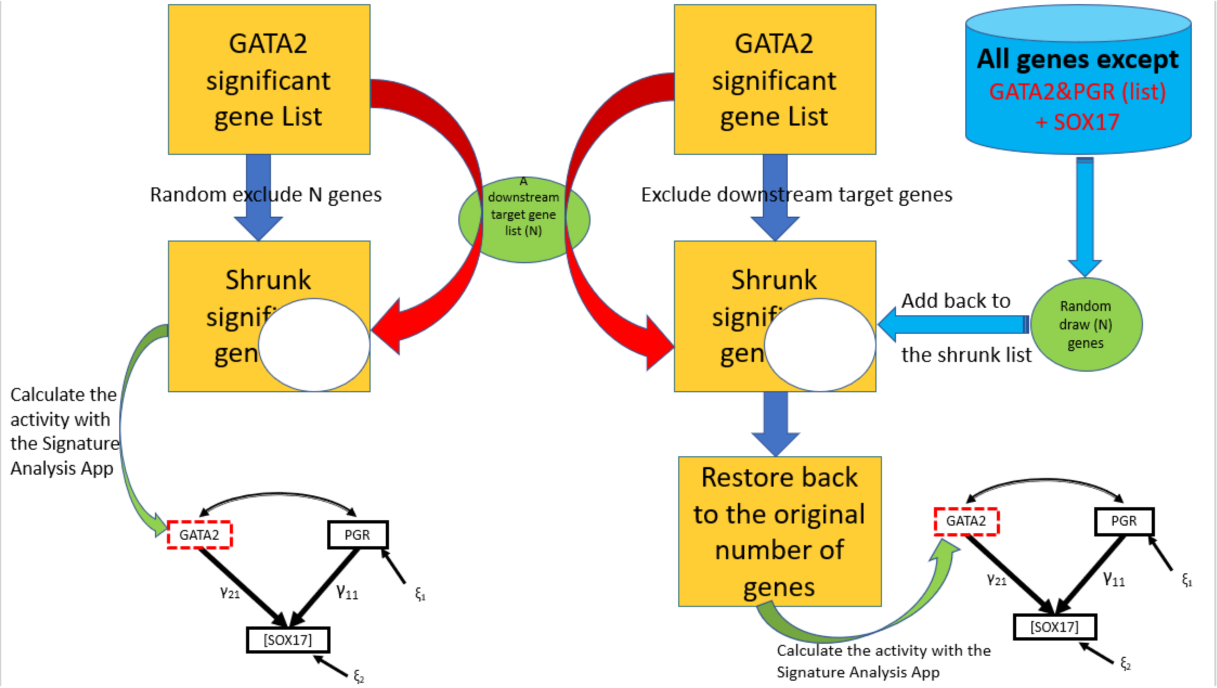


Comparative Fit Index	1.00
Tucker-Lewis Index	1.00
Standardized Root Mean Square Residual	< 0.001
Root Mean Square Error of Approximation	< 0.001

Supplemental Figure 1. 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.

Another feature of this app is a framework for hypothesis generation beyond simple model fitting. Under the Bootstrap tab of this app, we implemented a two-class (elimination with or without replacement) bootstrap resampling simulation for statistical inference (Supplemental Figure 2).

The overall concept is illustrated in Figure 1 in the main paper by the left most four rectangles. The idea is that the model fitting results would be altered if a subset of genes that has a significant role in the genetic network is removed from the gene signature. Results of this function may aid in prioritizing subsequent wet bench biological tests.


GATA2 significant gene List
Random exclude N genes
Shrunk signif gen
A downstream target gene list (N)
Calculate the activity with the Signature Analysis App
GATA2
PGR
γ_{21}
γ_{11}
ξ_1
[SOX17]
ξ_2
GATA2 significant gene List
All genes except GATA2&PGR (list) + SOX17
Exclude downstream target genes
Shrunk signif gen
Add back to the shrunk list
Random draw (N) genes
Restore back to the original number of genes
Calculate the activity with the Signature Analysis App
GATA2
PGR
γ_{21}
γ_{11}
ξ_1
[SOX17]
ξ_2

Supplemental Figure 2. 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.

Discussion

The SEMIPs R Shiny app offers an *in silico* perturbation testing system with multiple useful features. This user-friendly app allows quick assessments on genetic interactions and subsequent hypothesis generation without the requirement of extensive knowledge on computation languages and statistical analyses. Due to its simplicity in design, this app is limited to a 3-node model fitting capability. Models of higher complexity can be tested on the MplusAutomation, another R package that focuses on automating the SEM modeling currently done via a commercial Mplus software (Hallquist and Wiley 2018).

Currently, the two-class bootstrap analysis can only be conducted separately. Integration of these into the SEMIPs methodology for formulation into a single test will be investigated for future design, development and implementation. As noted in the manuscript and mentioned previously, the SEMIPs app has been adopted by researchers in the field with a few papers published recently (Liu et al. 2019, Wetendorf et al. 2020). We hope that it can serve a wider research community to address additional scientific questions.

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.

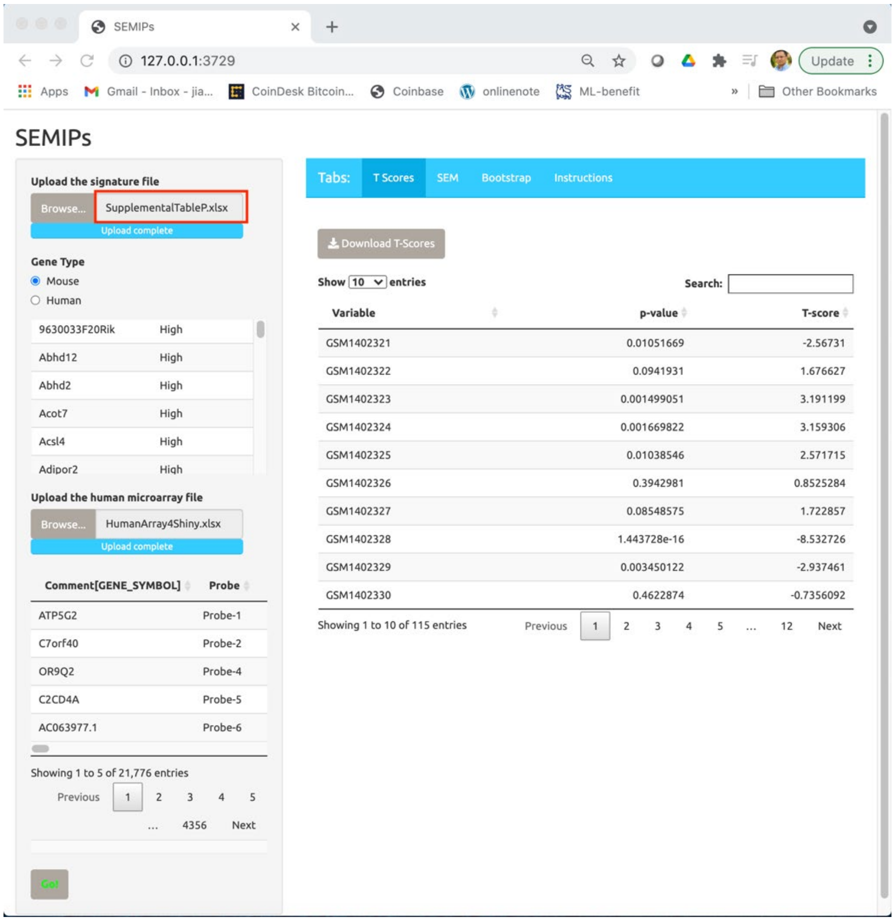
The published GATA2 occupancy information GEO accession: GSE40659 (Rubel et al. 2016) was first lifted from mm9 to mm10 genome assembly and then annotated by HOMER (Heinz et al. 2010) for the nearby genes. The obtained GATA2 ChIP-seq targets were mapped to the GATA2 signature from microarray data to 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.



SEMIPs

Upload the signature file

SupplementalTableP.xlsx

Upload complete

Gene Type

● Mouse

○ Human

9630033F20Rik	High
Abhd12	High
Abhd2	High
Acot7	High
Acsl4	High
Adipor2	High

Upload the human microarray file

HumanArray4Shiny.xlsx

Upload complete

Comment[GENE_SYMBOL]	Probe
ATP5G2	Probe-1
C7orf40	Probe-2
OR9Q2	Probe-4
C2CD4A	Probe-5
AC063977.1	Probe-6

Showing 1 to 5 of 21,776 entries

Previous 1 2 3 4 5 ... 4356 Next

Show 10 entries

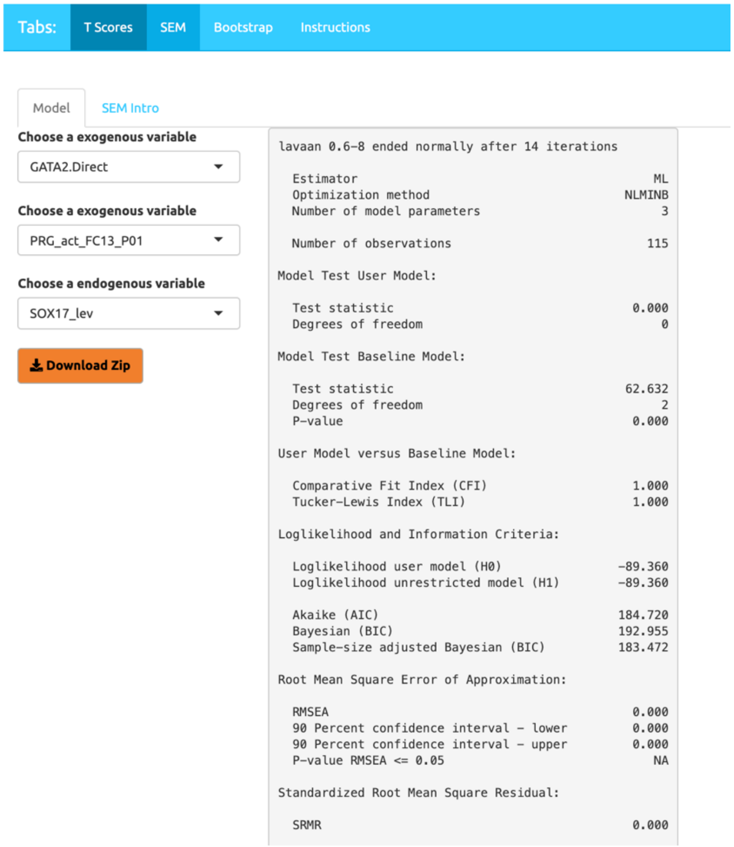
Search:

Variable	p-value	T-score
GSM1402321	0.01051669	-2.56731
GSM1402322	0.0941931	1.676627
GSM1402323	0.001499051	3.191199
GSM1402324	0.001669822	3.159306
GSM1402325	0.01038546	2.571715
GSM1402326	0.3942981	0.8525284
GSM1402327	0.08548575	1.722857
GSM1402328	1.443728e-16	-8.532726
GSM1402329	0.003450122	-2.937461
GSM1402330	0.4622874	-0.7356092

Showing 1 to 10 of 115 entries

Previous 1 2 3 4 5 ... 12 Next

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



Choose a exogenous variable: GATA2.Direct

Choose a exogenous variable: PRG_act_FC13_P01

Choose a endogenous variable: SOX17_lev

```
lavaan 0.6-8 ended normally after 14 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                         3

  Number of observations                           115

Model Test User Model:

  Test statistic                                 0.000
  Degrees of freedom                                 0

Model Test Baseline Model:

  Test statistic                                62.632
  Degrees of freedom                                 2
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    1.000
  Tucker-Lewis Index (TLI)                       1.000

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)                -89.360
  Loglikelihood unrestricted model (H1)        -89.360

  Akaike (AIC)                                 184.720
  Bayesian (BIC)                               192.955
  Sample-size adjusted Bayesian (BIC)          183.472

Root Mean Square Error of Approximation:

  RMSEA                                          0.000
  90 Percent confidence interval - lower         0.000
  90 Percent confidence interval - upper         0.000
  P-value RMSEA <= 0.05                             NA

Standardized Root Mean Square Residual:

  SRMR                                           0.000
```

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

Díaz-Gimeno, P., et al. (2011). "A genomic diagnostic tool for human endometrial receptivity based on the transcriptomic signature." Fertil Steril **95**(1): 50-60, 60.e51-15.

Hallquist, M. N. and J. F. Wiley (2018). "MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus." Structural Equation Modeling: A Multidisciplinary Journal **25**(4): 621-638.

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.

Koot, Y. E. M., et al. (2016). "An endometrial gene expression signature accurately predicts recurrent implantation failure after IVF." Scientific reports **6**: 19411-19411.

Liu, J., et al. (2019). "JNK(1/2) represses Lkb(1)-deficiency-induced lung squamous cell carcinoma progression." Nat Commun **10**(1): 2148.

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

Wetendorf, M., et al. (2020). "Constitutive expression of progesterone receptor isoforms promotes the development of hormone-dependent ovarian neoplasms." Sci Signal **13**(652).