# Abstract

Gene expression is controlled by multiple regulators and their interactions. We hypothesize that cross-species conserve baseline mechanism exists and propose to derive a projection from the gene signature of a biological model system to gene expression profiles of a human study via a T-score. This approach helps us to understand the complex human biology system and its potential clinical implications. Structural equation modeling (SEM) reports any significant relationships among variables, which ultimately helps us with biological hypothesis testing, i.e. unveiling the concurrent regulatory effects of two or more upstream regulators on the expression level or activity of a downstream reporter gene. We developed an R Shiny application, termed “Structural Equation Modeling of In silico Perturbations (SEMIPs)” to carry out these tasks including the T-score as a surrogate gene activity in a given human specimen. SEMIPs can be used in either correlational studies between outcome variables of interest or subsequent model fitting on multiple variables. This application implements a 3-node SEM model that consists of two upstream regulators as exogenous variables and one downstream reporter as an endogenous variable, it will reveal any potential significance of interactions among these variables based on the statistical outputs. SEMIPs enables scientists to investigate genetic interactions among variables of interest *in silico*. In a case study using SEMIPs, we show that putative direct downstream genes of the GATA Binding Protein 2 (GATA2) transcription factor are sufficient to infer its activities *in silico* for the conserved progesterone receptor (PGR)-GATA2-SRY-box transcription factor 17 (SOX17) genetic network in the human uterine endometrium.

# Introduction

Genome-wide gene expression assays allow us to study the relationship between exogenous perturbations and biological responses so that important underline mechanism can be revealed. Gene expression data in public repositories provides a valuable resource for investigators to study such regulatory processes (Edgar, Domrachev et al. 2002) and the causal relationships among variables of interest. However, it is challenging to test the knowledge obtained from experimental model systems in humans due to undetermined clinical outcomes and ethical considerations. Assuming that homologous gene functions are conserved across different species, the degree of similarity between the gene of interest in a model animal system and that in the target species will largely be preserved. Such projection can be achieved by a T-score calculated based on a two-side t-distribution. The T-score calculation has been used in previous studies and successfully represent the activities of the regulator in the targeted human system (Creighton, Casa et al. 2008, Creighton, Li et al. 2009, Luo, Emanuele et al. 2009, Qin, Lee et al. 2014). This scoring system helped to establish correlations between the prognosis outcome and manifestation of activities of the factor of interest in corresponding tumors (Creighton, Casa et al. 2008, Creighton, Li et al. 2009, Luo, Emanuele et al. 2009, Qin, Wu et al. 2013, Qin, Lee et al. 2014). In addition, the T-score calculation has also been utilized to determine the association among activities of factors of interest or between the activities of an upstream regulator and levels of its downstream targets within a set of human specimens (Wu, Kao et al. 2015, Rubel, Wu et al. 2016). Results of these studies demonstrated applications of such a surrogate score of molecular activities in investigation of gene functions and inference of regulatory processes (Grace 2006).

To determine the relationships among multiple variables, we choose the structural equation modeling (SEM) system which itself is a statistical technique to test the strength of influence among variables and therefore reveal the structural relationships (Edgar, Domrachev et al. 2002, Grace 2006). One of the advantages of the SEM is to model a system when the underline variables are not directly measurable. We were motivated to develop a Structural Equation Modeling of In silico Perturbations (SEMIPs) R Shiny application (app) to facilitate casual inference of gene regulatory processes, especially on multi-factorial impacts on outcome variables concurrently. SEMIPs enables quantification of a projected activity metric (T-score) and allows users to fit desired SEM models on variables of interest. For hypothesis generation purpose, SEMIPs provides two different bootstrap random sampling procedures (elimination with or without replacement) to test the significance of a model (Creighton, Casa et al. 2008). Previously, the T-score and SEM were applied to gene expression data to evaluate gene interactions that regulate the progesterone signaling pathway in the mouse uterus and infer gene regulation processes in human uterine specimens (Rubel, Wu et al. 2016). SEMIPs streamlines these processes and allows scientists to perform the analyses through a user-friendly interface.