Major comments:  
Line 73-75: More details about the t-score should be added here, i.e. the assumptions and observed data for the t-test.

The main motivation using a “t-score” was to achieve the cross-species projection from a model animal (mice or rats) experiment to another species or human when a perturbation was not directly applicable (Wu, S.P. et al. 2015). With a model animal (mice or rats) experiment, normally the animals are randomly assigned into two groups, where one group will receive “placebo” and/or no treatment and another group will receive the perturbation treatment. Experimental measurement will be properly collected from both groups (i.e. gene expression profile from a Microarray experiment) according to some thresholds followed by a statistical analysis (t-test). Significantly changed genes/probes (signatures) will be obtained from this analysis with directionality (up/down regulation). Such a group of genes/probes are deemed collectively as the “signature profiling” of biological responses to a particular perturbation. This finding and information will be projected into another animal system (i.e. human) of interest bearing the assumption that the experimental animal of interest would respond similarly if the perturbation were applied.

In a separate experiment of interest that is done with species of interest (i.e. human), the homologous genes of those previously identified genes/probes from the experimental species will be selected, and the experimental measurement from this study (GEO accession: GSE58144, (Koot et al. 2016)) will be used. The directionality information will be used to group these genes into two separate groups. A normal t-statistics will be calculated from these two groups of measurement to represent the responses in a new species and new experiment set up of interest. Since this calculation was originated from a standard t-test statistics and the term “T-score” was coined firstly by Wu, S.P. et al (Wu, S.P. et al. 2015) and have been widely used in other research projects.

Line 96: Regarding the two bootstrapping methods, do they have different assumptions, computational costs and/or test power? Any recommendations for when either method is preferred?  
  
Yes, the two-class (elimination with or without replacement) bootstrap simulations do carry different assumptions. The primary application of SEM in our research and fundamental advantage of SEM is to allow researchers to derive the relationship between variables of interest and when these variables are not directly measurable. In our research, we tested the relationship via a three-node SEM model among three genomic regulators in a complex genomic system. Each of these factors can be a regulator that regulates a group of downstream genes or a readout of impact from some upstream regulator. And the T-scores will be calculated based on the direction of these upstream/downstream signatures(genes), then used in the SEM modeling.

When we have a group of signatures (genes) obtained from an experiment (i.e. a KEGG pathway analysis in our paper), we are interested in whether a regulator (upstream/downstream) is associated with a factor (i.e. GATA2 in our example) in our SEM model. We chose to eliminate these group of signatures (genes) from the GATA2-related signatures. To provide an unbiased assessment of such analysis, we implemented “an elimination without replacement” bootstrap analysis by randomly eliminated the same number of signatures from this originate GATA2-related signatures, the re-calculate the T-score and re-evaluate the SEM model, we do this 1000 round of simulation to provide an empirical distribution of all important statistics.

As a counterpart of this analysis, we also implemented “an elimination with replacement” bootstrap analysis. Such “an elimination with replacement” is similar to “an elimination without replacement” except that after eliminating the signatures from this originate GATA2-related signatures, we will replace the same number of “irrelevant” signatures back to the “shrunken” list. Then, we will re-calculate the T-score and re-evaluate the SEM model, we do this 1000 round of simulation to provide an empirical distribution of all important statistics.

The elimination without replacement simulation was used to test whether a regulator has any impact on our factor (i.e. GATA2) in term of function association; and the elimination with replacement simulation was used to rule out the possibility that the number of downstream signatures of a factor (i.e. GATA2) has any impact on its function. Both empirical distributions serve as the null hypothesis for any statistical testing.

Minor comments:  
Line 70: projects -> projected  
Line 73: Such an information -> Such information  
  
Line 262: Figure 1 is a little confusing. The green shape is not a rectangle but was referred to as one. Varying both shape and color without appropriate annotation is confusing/distracting. The red boxes, dashed vs solid, do they have different meaning?  
  
Supplementary Figure 1  
- Why is SOX17 in brackets?

When the “brackets” are used, it refers that “gene expression value” is used instead of “activity” value.

- The top two thicker blue arrows seem to indicate the same processing step, but the text annotations are different, which is confusing and distracting.

- Figure legend: “The resulting shrunken GATA2 gene list or reduced GATA2 [gene list] then restored by the same number of irrelevant genes are tested in the SEM model.”  
  
Source code: Coding style in the source code could use some standardization.