Reviewer #1 (Remarks to the Author):

The paper describes a computational method for imputing transcriptome profiles from a small number of marker genes. The main argument presented in the paper is that the proposed method Tradict would allow reducing the need to profile genome-wide expresso profiles and instead focus on a smaller number of markers.  
  
My main concern with this work is the lack of a clear demonstration of the broad utility of the method. The idea to predict gene expression profiles from informative marker sets is not new and most previous approaches are not particularly widely used. Hence, a compelling demonstration of the practical advance enabled by Tradict in the context of an application to a major biological problem will be critical. Additionally, I note several technical concerns that would need to be addressed.  
  
- Validation and demonstration of the method  
The presented experiments are exclusively synthetic benchmarks based on sub sampling of genome-wide transcriptome profiles. These results partially validate the method, however the practical utility remains unclear. A stronger validation and clear demonstration could be a direct comparison of targeted RNA-seq coupled with Tradict versus conventional deep RNA-seq of the same cell populations. If this is not possible, the authors should demonstrate that Tradict applied to existing datasets can improve the biological interoperation compared to conventional analysis strategies.  
  
- Extended comparison to alternative methods  
The imputation of gene expression profiles from marker genes is not a new per se. There are several existing methods, from which the authors select a subset for comparison. I would request to include additional comparisons, such as k nearest neighbour (kNN) and factor analysis methods (see e.g. De Souto et al. for a comparison). An important question is whether single-gene imputation techniques (such as kNN) are less accurate than Tradict predictions evaluates for single genes. Also, the method is conceptually related to factor analysis, and hence this class of models would be a naturalcomparison partner. Additionally, the authors claim that the improved noise model is a major advance that enables improved accuracy. This could be tested experimentally, for example by comparing the performance of alternative methods for genes in different abundance bins. The results from the reduced Tradict model (Supplementary Fig. S5) do currently not address this in full.  
  
- Robustness and impact on the annotation used  
The method builds on reference gene sets derived from GO ontologies or Kegg pathways to define and train expression signatures of transcriptional programmes. It will be essential to demonstrate that the method is robust w.r.t. to the specific choice of annotation. What is the impact of false positive gene sets in the set or missing markers ? The impact of size of the program, its variance in the training population (not considered), mean expression levels and false positive assignments should be more clearly explored and assessed. I am also missing insights as to which proportion of the transcriptome can be accurate imputed for a given reference annotation. Are genes that participate in several programs easier to predict? A clearer assessment of the prediction accuracy at a single-gene level will be needed.  
  
- Algorithmic details  
The authors emphasise that a rigorous and statistically sound model is proposed. I appreciate the statistical basis but the there some aspects that are not solved satisfactory. First, the scalability of the method should be explored and stated explicitly. How much compute time to the coding and decoding steps take ? Second, the method appears to be based on MAP inference, which means each predicted gene expression value are a best guess estimate. What seems to be missing are predictive uncertainties are an alternative diagnostics that enables the user to identify genes that can or cannot be reconstructed for given dataset.  
  
- Limitations when imputation dataset with single-gene perturbations.  
A related point is the the paper does not discuss the limitations of imputation approaches in general. There numerous use case of tarnscriptome profiling where expression changes at the single-gene level are of primary interest and unlikely to be picked up by transciprtional programmes. This includes for example the genetic analysis of gene expression (e.g. eQTL), which frequently affects individual genes. A demonstration that the model can cope with these settings of sub-program variation will be required. I believe the CRISPR dataset may contain some instances where only a single or a small number of genes are differentially regulated in the knock down condition. An interesting question is an assessment of the imputation accuracy as a function of the size of the affected gene set. A second limitation is that the method will not pickup more subtle variation such as splicing or isoform specific differences.  
  
Minor comments:  
  
- The method nicely leverages large references datasets. Why not use the data to infer the modules directly fro these data ? This has previously been considered (e.g. Fehrmann et al.) and would seem a better use of the data.  
  
- The single-cell motivation is weak. This needs to be substantially expanded or removed. I don't think the presented experiment demonstrates anything.  
  
- The methods description should be extended. I am missing some details on how the iterative selection of marker genes is performed. Are there issues with local optima due to the greedy nature of the selection ?  
  
References:  
Fehrmann, Rudolf SN, et al. "Gene expression analysis identifies global gene dosage sensitivity in cancer." Nature genetics 47.2 (2015): 115-125.  
  
De Souto, Marcilio CP, Pablo A. Jaskowiak, and Ivan G. Costa. "Impact of missing data imputation methods on gene expression clustering and classification." BMC bioinformatics 16.1 (2015): 1.  
  
Donner, Yoni, et al. "Imputing gene expression from selectively reduced probe sets." Nature methods 9.11 (2012): 1120-1125.  
  
  
Reviewer #3 (Remarks to the Author):  
  
In this manuscript the authors introduce Tradict, a software to evaluate gene expression levels from a limited set of marker genes. Obtaining reliable estimates of gene expressions from a small sample of targeted genes could have the important consequence of reducing the cost of RNA-seq while still obtaining transcriptome-wide information. I appreciate the effort the authors put into developing Tradict and I found it technically and statistically sound. However I have some major criticisms of this manuscript, as detailed below.  
  
\* Tradict is claimed to "reconstruct" an entire eukaryotic transcriptome. I think this is a grossly exaggerated claim.   
1) First of all the "reconstruct" term suggests more than determining gene expression levels, but also unknown transcripts and transcriptome structures, which is not the case in this manuscript. We still don't have a complete picture of all the existing transcripts produced from an eukaryotic genome, but the authors assume that all transcripts and all their structures are already known. They do not attempt to "reconstruct" them.  
2) Tradict accurately determines expression levels of some well defined transcriptional programs, but this level of accuracy is not as high for gene expression levels. The authors should acknowledge that gene level expression could still be more accurate when performing transcriptome-wide RNA sequencing.   
3) Isoform expression levels are not determined by Tradict. In some cases, differential alternative splicing usage is what differentiate a certain condition from another, but this is not captured by Tradict. While I don't think this should necessarily be the goal of Tradict, a \*complete\* reconstruction of the transcriptome will also include determining isoform-level expressions.  
4) The authors exclude non-coding transcripts from their analysis, which is another point against their complete transcriptome reconstruction argument. We do include lncRNAs  
  
\* Also, I didn't find compelling arguments for the practical usage of Tradict:  
5) The authors train Tradict on a large number of RNA-seq samples that come from just two species. However many species have not been sequenced so extensively. It would seem that before the cost of RNA-seq would be reduced by only doing targeted sequencing one must wait a long time until a large number of samples are sequenced for their species of interest. The applicability of Tradict is therefore reduced and limited to just a few species, at least for the near future. The authors should point this in their manuscript.   
6) Related to the point above, what would be a good number of samples for which one would have to wait until Tradict could produce reliable results? In other words, how does the accuracy vary when the number of training samples is reduced?  
7) What is the diversity of the biological processes captured by the training samples that is needed to produce reliable results?  
8) What is the running time, and what computer resources are needed in order to run Tradict?  
  
  
Reviewer #4 (Remarks to the Author):  
  
This manuscript describes an approach for representing gene expression  
programs in such a way that the measurement of the expression level of only  
100 marker genes is predictive of the expression level of pathways. This is  
done by grouping genes by GO terms, and then selecting marker genes (via a  
greedy selection procedure). The approach is applied to mouse and a. thaliana  
expression, using a large number of samples drawn from a public database. The  
reported reconstruction accuracy is very good at the pathway level.  
  
The computational methodology seems sound, and the study represents progress  
toward the important goal of making use of large expression data sets.  
  
Comments.  
  
1. Line 44: "Pareto optimality...": it is not clear (a) what this sentence  
means or (b) how it implies that the dimension should be low.  
  
Dataset:   
  
2. Line 103-113 and supplemental note 1: It is important to quantify better  
the diversity and sampling bias in the set of expression experiments that were  
used. Though some qualitative arguments are given that the set represents a  
"biologically realistic" set, more results / discussion are needed. If the SRA  
experiments, for example, are highly biased towards particular experimental  
conditions (e.g. environmental stress), then it would (a) be less surprising  
that you could represent the full expression matrix using only a few marker  
genes, and (b) would limit the generalizability of the results.  
  
Pathways used:   
  
3. Line 133 and supplementary material: The selection of pathways is crucial  
to the results presented. A robustness analysis is needed: (a) how well does  
the method perform on higher-level or lower-level groupings, and (b) how well  
does the method perform on \*random\* pathways. This latter test will show  
whether the method is adding something beyond the observation that genes with  
similar GO terms tend to be co-expressed.  
  
Measuring success:   
  
4. Only a single measure is used for reconstruction accuracy: the PCC of  
mean-subtracted expression values. Other metrics for how accurately the  
reconstruction is done must be presented (there are many discussed in the  
literature on RNA-seq quantification methods).   
  
5. The motivation for the "intra-submission" accuracy is not sufficient.  
First, the name "intra-submission accuracy" suggests that training and testing  
are done on the same submission, when this is not the case. Second, why  
doesn't subtracting the mean of a submission leave mostly noise (esp. e.g. in  
the case of a submission containing biological replicates)? Third, the goal of  
the transformation is not clear: the phase "the total biological signal  
contained in the test set" well defined, or why one would want to eliminate  
the "total biological signal". Fourth, why subtracting the mean of a  
submission achieves the desired goal (whatever it is) should be explicitly  
stated.  
  
6. For cross validation, does one need to remove experiments that are  
conducted on the same tissue under the same conditions but in different  
submissions? Even if these experiments can stay in, this issue should be  
discussed.  
  
7. Were the \*measured\* genes contained in the expression computed for  
pathways? How does this affect the PCC? What correlation for pathways would  
be achieved if the measured gene alone was used as a proxy for the entire  
pathway (without Tradict)?  
  
8. The PCC of 0.94 and 0.93 are extremely high. What would the average PCC be  
between biological or technical replicates --- without Tradict --- after the  
fully-measured expression values are collapsed into pathways? Can repeatedly  
even performing the same complete RNA-seq experiment get PCCs of 0.94 at the  
pathway level?  
  
Coverage of genes:  
  
9. line 440: "have only 'Biological Process' as a GO annotation, and therefore  
we do not need to capture these genes": I don't think that you can say you  
don't "need" to capture these genes. In fact, these genes with unknown  
function are the ones that most people conducting experiments would like to  
learn about.  
  
10. How was the range [50,2000] chosen for the GO term size cutoffs? How do  
the results change when this range is adjusted?  
  
11. Only 54% (mouse) and 63% (A. thaliana) of the genes appear among the genes  
that are being reconstructed. More should be said about the genes where the  
method could not reconstruct the expression value (b/c they were excluded).   
The lack of demonstrated ability to reconstruct 46% - 37% of the genes may   
limit the method's usefulness.  
  
Minor comments:  
  
12. "its" is frequently misspelled "it's"  
  
13. line 283: "just the right balance" is too strong.  
  
14. Ref. 15 is wrong.  
  
15. Supp note 4: How was the figure ~10,000 reads are need for measuring the  
markers computed? That amounts to 100 reads / gene, which seems low.  
  
16. Tradict - mathematical details, line 26: "[REF]" should be a reference.  
  
  
  
  
Reviewer #5 (Remarks to the Author):  
  
Biswas et al present a method for encoding transcriptome-wide expression profiles into a compressed representation using only 100 marker genes, which can then be decoded to recover transcriptional program expression and transcriptome-wide expression. The authors show that the Tradict method can predict test set transcriptional programs in A thaliana and M musculus with average correlation higher than 0.9. In addition Tradict is used to predict transcriptional programs and specific genes in a hormone perturbation experiment, again using only 100 marker genes. A key consequence is that the methodology could be used to greatly reduce expenses and increase sample sizes by requiring only small sets of transcripts to be assayed in order to obtain transcriptome-wide expression.  
  
Major comments:  
  
The results on the recovery of transcriptome-wide expression do not support the claims in the Abstract, Introduction and Conclusion, that the prediction is accurate or could be used to replace transcriptome-wide screening. The Pearson correlations of predicted gene expression to actual gene expression are 0.64 and 0.59 for A thaliana and M musculus respectively (Fig 3), which means that Tradict explains less than half of the total variability in gene expression across the thousands of transcriptomes surveyed. Furthermore, the relevant measure here is not correlation but residual mean squared error. From the figures this appears to be quite large.  
  
Furthermore, the impact and utility of the proposed method is questionable, as achieving a high Pearson correlation (explaining a large amount of total variability) for gene expression or the first PC of a transcriptional programs might not help for a given experiment, in which the expression of a single gene or program could be relevant. Targeting only 100 genes and then using Tradict to extrapolate to the entire transcriptome might get close to the general trend of gene expression patterns, but it could just as well mis-estimate expression of critical genes or programs whose actual expression, if it were assayed would correlate with the phenotype of interest, e.g. response to treatment, patient status, etc. The critical measure for assessing whether Tradict can replace current transcriptome-wide assays would be a bound on the worst performance for a single gene or program, as we cannot know in advance which genes or programs will be relevant for a given experiment. The  
manuscript make the claim that "the 100 markers Tradict learns are likely to be predictive independent of most contexts and applications", but such a claim is not demonstrated through data.  
  
Finally, I want to point out that there is something very strange about figures 3a and 3c. Note for example the wide horizontal blue bands. This means that the great majority of tr. programs are not expressed for a very large number of transcriptomes. Also note the blocks of transcriptomes that are almost identical when looking across programs. This seems to imply that tr. prgrams are basically the same. How can this be? First possibilities that come to mind are a bug in how the programs are defined, experiments from different experimental protocols incorrectly being compared, or lack of proper normalization.