SUMMARY OF RESPONSES TO REVIEWERS

We feel that all the reviewers appreciate and are comfortable with the core statistical and methodological contributions of the paper. In our opinion, the major reason the paper was rejected from Nature Biotechnology was due to the lack of a complete practical demonstration of the method, including prospectively designed experiments that actually use targeted RNA-sequencing with Tradict.

The major requests of the reviewers can be summarized as follows:

1. A complete and practical demonstration/validation of Tradict, including prospectively designed experiments that use targeted RNA-Sequencing + Tradict.

2. A greater exploration of Tradict's limitations and methodology:
    - Investigation into what drives Tradict's prediction errors.
    - Sample size requirements.
    - Robustness to choice of annotation for transcriptional programs.
    - Availability of confidence measures over predictions
    - Timing analysis

Item 1 is beyond the scope of what we hoped to achieve in this work. We nonetheless believe our work is an important theoretical advance, and immediately accessible for practical use to labs that perform screens, diagnostics, or transcriptomic analyses. Targeted RNA-Sequencing approaches, such as RASL-Seq, for screening or diagnostic purposes are lightweight, cheap, well developed, and increasingly commonplace, making their adoption easy and likely.

We have therefore focused our efforts toward addressing each sub-item under item 2. To summarize, we have:

- Performed an extensive error analysis, in which we studied the association between Tradict's prediction errors and several properties of the gene or transcriptional program to be predicted (Supplemental Analysis 3.I). We were able to build a model of errors that could account for most of Tradict's mis-predictions.
- Performed a power analysis in which we investigated Tradict's prediction performance as a function of increasing training set sizes. We found Tradict's performance begins to saturate at 1000 samples (Supplemental Analysis 3.II). The major model organisms all have far more than 1000 publicly available transcriptomes. Furthermore, investigators are easily able to incorporate their own samples.
- Performed a transcriptional program annotation robustness analysis, in which we examined how gene-to-transcriptional program mis-annotation rates influence Tradict's prediction accuracies (Supplemental Analysis 3.III). We found Tradict is robust up to a 20% mis-annotation rate, which is a comfortable cushion for what one might encounter in practice, especially for well maintained annotations for model organisms such as GO and KEGG.
- Added functionality to Tradict to calculate conditional posterior distributions over the expression of programs and genes. From this the user can derive point estimates as well as credible intervals. We validated these intervals with cross-validation, and showed that on samples not used for training an X% credible interval on average captures X% of test-set observations for any X between 0 and 100 (Supplemental Analysis IV). This is highly reassuring that Tradict's probabilistic/modeling assumptions are quite reasonable and accurate.
- Performed a timing analysis and showed that Tradict's training times and test times are resonable. The only limitation is the size of the training database, which may not be easily stored on the disk or in memory for a standard laptop, and may require up to a 32GB RAM desktop.

Please find below a point-by-point response to all reviewer comments. Our responses are in BLUE.

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.

An experimental demonstration of the method is beyond what we hope to achieve in this work. We argue though that targeted RNA-Sequencing is very well developed and as accurate as RNA-Seq read outs for the same set of genes targeted. Tradict interfaces easily with such technology.

Though the idea to predict gene expression profiles from informative marker sets is not new, we argue that no previous work has achieved the training size, data type, or accuracy that we have. Additionally, no other method has considered formulating and predicting the expression of transcriptional programs, which give a more nuanced view of the state of the organism than a single phenotype, but are more interpretable than gene-expression alone.

All previous methods that we know of were developed on microarray, a noisy technology (compared to RNA-Seq) that is increasingly outdated by sequencing based readouts. We argue that part of the reason these methods have not been adopted is because 1) their prediction accuracies were modest (perhaps a consequence of training on microarray data, and/or training only on a few hundred samples), and 2) they focus only on gene expression and not higher level targets, such as transcriptional programs, which are more interpretable and lend mechanistic insight.

Thus, Tradict is the first of its kind to leverage publicly available RNA-Seq datasets in very large numbers (1-2 orders of magnitude more than previous methods). It is also substantially more accurate than previous methods for predicting gene expression and the first to formulate and predict the expression of transcriptional programs, which we argue is a compelling advance that enables simultaneous screening and mechanistic investigation at large scales.

We have made these assertions more clear throughout the text.

- 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.

We agree that these experiments are an ideal demonstration of the method. However, as mentioned before, targeted RNA-Sequencing is very well established, and even commercialized (see Illumina's targeted RNA-Sequencing kit, and TempO-Seq from BioSpyder technologies, http://biospyder.com). Tradict's software is intuitive and easy to use. Consequently, we feel that use of targeted RNA-Sequencing coupled with Tradict is straightforward, up to of course, the usual optimization a lab has to do at the bench whenever a new method is adopted.

- 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.

With respect to the kNN baseline, the LWA baseline we discuss (now both in the main text and in Supplemental Analysis 5) is conceptually similar, but better empirically with respect to the results of Donner *et al.* (2012), where they perform a comparison to a nearest neighbors approach and conclude that locally weighted averaging (LWA) is more powerful.

We don't think factor analysis is an appropriate baseline with respect to program/gene expression prediction from a subset of the transcriptome. In factor analysis, the assumption is that there is some latent, unobserved (and unmeasured/unmeasurable) "factor" or set of "factors" from which the variables under study (which are more than the number of assumed factors) are generated (as linear combinations of the factors). Though we could run a (regularized) factor analysis on our data, the learned factors will be some linear combination of all genes or programs. This is not the type of object we'd like to measure with targeted sequencing, though we can think of some ways of how to do it. Any measurement target that is a linear combination of many genes will likely be difficult and expensive to sample given that many genes would somehow have to be probed. In this work we don't consider the existence of latent factors, but rather assume that a subset of genes themselves encode enough information to form tight predictive distributions over transcriptional programs (especially) and the remaining genes in the transcriptome.

- 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.

Point by point:

- "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 ?" In Supplemental Analysis 3.III show that gene expression prediction is completely robust to the choice of annotation. Gene expression prediction accuracy is completely unchanged even in the face of 100% mis-annotated gene sets. This is due to the statistical decoupling between gene and transcriptional program expression prediction. Transcriptional program expression prediction remains robust up until a 20% mis-annotation rate, which we argue is a comfortable cushion for well maintained annotations such as GO and KEGG. Supplemental Analysis 2 and 6 examine situations where many (~40-60%) of the markers are missing or measured noisily. But even here Tradict remains robust and is able to, at the very least, return the most salient expression pattern changes between samples.

- “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.” These are now deeply explored in Supplemental Analysis 3.I. Program size correlates weakly with prediction performance. However, training expression variance and mean are significant correlates of prediction error.
- “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” This is also considered in Supplemental Analysis 3.I. Again due to the statistical decoupling of program annotation and gene expression prediction, gene expression prediction is completely robust with respect to program annotation or size.

- 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.

We have now added a supplemental analysis (Supplemental Analysis 4) that performs a timing and memory analysis. We have also provided a summary of these results in the main text.

We have improved the method to return posterior distributions as predictions over gene and program expression instead of MAP point-estimates. This allows the user to derive their own point estimator (posterior mode or mean) and a credible interval. In Supplemental Analysis 4 we show via a quantile-quantile analysis that in a cross-validation setting Tradict’s X% credible interval statistically includes X% of actual test-set expression values for all X between 0 and 100. This is strong evidence that Tradict’s error model is highly appropriate for gene and transcriptional program expression.

- 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 transcipritonal 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.

We acknowledge that it may be difficult to detect single gene changes or sub-program variation. For this reason, and others offered by the other reviewers, we have toned down our claims for Tradict’s accuracy with respect to single gene expression prediction. However, we argue there may be few instances of total independence of a few genes from the rest of the transcriptome. Extensive connectivity and feedback - a consequence of the energy constraints on a cell - should ensure that most transcriptional changes, even if initiated locally, should ultimately feedback onto the expression of the selected markers.

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.

We were aware of this work and several others that try to define expression modules in an unsupervised manner. Our dataset would be perfect for this task, but we have purposefully avoided defining our own modules without being guided by some knowledge-based approach. Though co-expression suggests co-regulation or involvement in the same function, this is not always the case. Furthermore, one cannot ascribe function in the form of a text description (e.g. “response to salt stress”) when defining modules purely in such an unsupervised manner. For this reason we have restricted ourselves to using GO annotations, which are either manually curated based on experimental results, or defined using strict homology criterion. Both of these approaches, we argue are stronger criteria for defining pure and labeled transcriptional programs than some other statistical definition.

- The single-cell motivation is weak. This needs to be substantially expanded or removed. I don't think the presented experiment demonstrates anything.

This has been removed from the text.

- 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 ?

A detailed description of the method (beyond what is presented in the main text) can be found in the Supplemental Information, Section “Materials and Methods”, subsection “Tradict algorithm”). We have made a more clear reference to this in the main text.

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.

We agree with this critique and have rephrased our text to not say “reconstruct”.

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.

We have toned down our claims throughout the text and made more clear Tradict’s predictive performance for genes. We argue that it is impressive that gene expression can be predicted from 100 markers with a pearson correlation of ~0.6-0.7, and this performance is superior to previously published methods; however, this must be improved if the individual gene expression predictions are to be taken as seriously as one would from whole transcriptome 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.

As previously mentioned, we have replaced our usage of “reconstruction.” We have made it more explicit that we are not reconstructing isoform level expression levels.

4) The authors exclude non-coding transcripts from their analysis, which is another point against their complete transcriptome reconstruction argument.

We do not include any non poly-adenylated transcripts in our analysis because a great fraction of RNA-Seq samples on the public domain were produced from polyA enriched samples. However, this means that we DO include long non-coding RNAs (lncRNAs), which are poly-adenylated.

* 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?

Response to 5&6) Toward this end, we performed a power analysis (Supplemental Analysis 3.II) in which we determined in a cross-validation setting how many samples Tradict would need to achieve (statistically) its best performance. We found this number to be about 1000 samples. This level of sampling exists for nearly all model and popularly studied organsims including *M. musculus*, *H. sapiens*, *A. thaliana*, *D. rerio*, *C. elegans*, *D. melanogaster,* and *S. cerevisiae*.

7) What is the diversity of the biological processes captured by the training samples that is needed to produce reliable results?

This is a difficult question to test directly since we have unstandardized annotation over the experiments submitted to the SRA and have not prospectively designed the training sets ourselves. Further more, it is hard to quantitatively define "diversity." However, the power analysis in Supplemental Analysis 3.II let's us know that ~1000 samples deposited to the SRA is sufficient for good training. Additionally, from Figure 1 in the main text, we see that a large proportion of expression variation is driven by tissue and developmental stage of the organism. The first 1000 samples deposited to the SRA, for both organisms, contain a sampling of all the major tissues and a range of developmental contexts, ranging from embryonic tissue to fully differentiated, mature tissue. Thus we would conclude that sufficient sampling of tissues and developmental contexts is necessary for a good training set. This will certainly perturb the "housekeeping" biological processes (which are the majority of the processes), related to growth and development.

Beyond this, one would be interested in response regulons related to biotic/abiotic stresses. For these we would suggest that the training set include examples generally related to the application of interest. For example, investigators working in innate immune signaling should have training transcriptomes with immune perturbations, generally, though of course, these perturbations don't need to be exactly the ones being investigated. Many response processes are integrated into housekeeping properties (e.g. plant defense responses are dependent on circadian rhythm, which is an essential regulator of proper growth and development).

Finally, though we cannot answer the reviewer's question as precisely as we would like, we have added a feature to Tradict that allows a user to see if some test samples they have generated are "well-captured" by the training set in hand. This method works by computing the PCA of the training transcriptome collection with respect to the selected markers only. Then for a query sample of marker measurements, the method checks to see if the query is linearly spanned by a set of training samples in low-dimensional (3 dimensions) PC space. If this is the case (i.e. the query samples are "well-contained" by the training collection), then we would conclude the training set is appropriate for the user's application.

8) What is the running time, and what computer resources are needed in order to run Tradict?

As mentioned for a previous reviewer comment, we have now added a supplemental analysis (Supplemental Analysis 4) that performs a timing and memory analysis. We have also provided a summary of and reference to these results in the main text.

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.

We have clarified this in the main text.

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.

Conceptually, we agree. We provide two arguments as to why what we have presented in the manuscript already address this concern.
- First, the major (perhaps only publicly accessible) databases for deposition of RNA-Sequencing data are the SRA and GEO, the latter of which mandates upload of raw sequencing data to the former. Thus, these databases contain as much variety as is achievable through repurposing already generated data.
- Second, with regard to the point of generalizability, which is the most relevant criteria to judge Tradict by, our cross-validation experiments (Figure 3 in the main text) are aimed to test exactly this. By training on a subset of experiments and accurately predicting expression in held-out experiments we demonstrate directly Tradict's generalizability.

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.

These concerns are addressed in Supplemental Analysis 3.I and Supplemental Analysis 3.III, In the first, we have performed an error analysis. Higher-level groupings, which will contain more genes, should be easier to predict more accurately (Fig S6a-b, right), though accuracy is already high for nearly all programs. However, these categories will be more general and not as informative about transcriptional status.

In Supplemental Analysis 3.III we perform a program annotation robustness analysis, part of which includes a test of 100% randomly assigned genes to programs. As illustrated there and proven mathematically in the "Tradict -mathematical details" section of the Supplemental Information, gene expression prediction accuracy is completely invariant and therefore robust to with respect program annotation. We have made it more clear in the main text that statistically gene and program expression prediction are decoupled, even if conceptually (e.g. as presented in Fig 1) it is useful to think that they are not).

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).

The PCC is the ratio of the covariance between the prediction and the target divided by the square root of the product of the variances of the prediction and target. This a common measure of accuracy in machine learning and statistics. The mean squared error (MSE) is also popular and measures the total squared deviation between the prediction and the target (NOTE: Reviewer #5 suggests this criteria). This is equivalently the residual variance or unexplained variance. Because our target's (programs and genes) can be on different scales (different mean and variances in expression), in order to compare them we must normalize the MSE criterion. To do this, we consider the normalized unexplained variance, which is given by the unexplained variance divided by the variance of the target. This is statistically equivalent to one minus the coefficient of determination.

In Supplemental Analysis 3.I we perform side-by-side comparisons of PCC and normalized unexplained variance and find that while their relationship isn't completely linear, they are highly correlated in rank. Thus we conclude that the PCC results presented are a sufficient exposition of the methods performance.

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.

We have clarified what we mean by intra-submission accuracy. Our updated text makes it clear that training and testing are NOT occurring within the same submission.

As our updated text now better reflects, Tradict's ultimate goal should be to train on publicly available data. Then when a lab want's to use Tradict for their next experiment, which will likely include some specific biological context, we want Tradict to be quantitative within the signal contained in this single experiment. During cross validation, our test set contains multiple submissions worth of data, and in aggregate, this contains far more biological signal (e.g. different tissues) than one might encounter in practice. By subtracting the submission specific mean, we are regressing out inter-submission effects and are considering Tradict's quantitative performance exactly as one would want to do in practice, one experiment at a time.

The intra-submission performance is the most stringent criteria for assessing Tradict's predictive performance.

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.

We argue that this does not need to be the case. First, we qualitatively argue this situation is not common simply because its unlikely the same or another group would perform such similar experiments under two different submissions (which usually happen after a lab submits a manuscript for publication). However, even in such cases (e.g. one lab replicates a previous one's findings as a control result), we think we should include these during training as it more closely mimics reality and gives more realistic estimates of the performance we might expect from Tradict.

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)?

The measured genes are selected from the entire transcriptome and are chosen greedily to decompose the program expression matrix. However, in each greedy iteration the program expression matrix is reweighted such that a marker is selected to preferentially decompose the current program with the largest unexplained variance. This process is detailed in the Tradict algorithm subsection of Materials and Methods in the Supplemental Information.
With this in mind, this request is not directly addressable. Nevertheless, one could still ask if whether the "best" (e.g. mean, medoid) gene in each transcriptional program could serve as a proxy for the entire pathway (without Tradict). However, one would need at minimum as many genes as there are transcriptional programs (which is >100). Additionally, the markers Tradict picks tend to be explanatory of multiple pathways at once. Thus we do not consider this an immediately useful baseline.

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?

Note that we are looking at PCC values for each gene across multiple samples of varying condition. We are not looking at the PCC of two samples across all genes. In the former, we are trying to detect expression **changes** to **each** gene across different samples. The latter is just a measure of similarity between two samples (e.g. useful for comparing technical replicates to make sure protocols are robust). We think the reviewer may have accidentally confused these two scenarios.

The baselines requested here do not make sense. In the case of true technical replicates, differences in gene expression are due only to the measurement protocol and not due to any biological signal that affects the expression of genes in a manner that ultimately feeds back into changes in our selected marker genes. Thus, across many technical replicates we would expect our marker expression levels to be unchanged on average, and consequently our predictions for the rest of the transcriptome as well. Because any changes in gene expression across technical replicates should be pure measurement noise, we would expect the PCC between predicted and actual expression to be 0, not high. In fact, this should be true of complete RNA-Seq - a given transcript should have 0 correlation across many technical replicates.

In the case of biological replicates, there should be some biological signal (due to differences between individuals) driving changes in gene expression and in theory this should be captured by our markers if our claim that they are representative of the transcriptome is true. However, because these individuals haven't been perturbed in anyway, this signal should be low, and our predicted expression across many biological replicates should be correlated with actual expression, but it will be low.

We could not find a sufficient number of biological replicates (>5) within any submission for which we could test this empirically.

We argue that the PCC values of 0.94 and 0.93 are, in fact, expected. This is because each transcriptional program is defined to be a linear combination (in log-latent space) of the expression values of its constituent genes. As a consequence of central limit theorem, taking a weighted sum (if the weights are not to unevenly distributed) over all of these genes will “average-out” orthogonal noise that’s associated with the prediction of each gene individually. Thus predicting this weighted sum -- the expression of the transcriptional program -- is a substantially easier task than predicting the expression of a single gene.

As example evidence, consider this example where we have 1 marker that is lowly correlated (PCC = 0.45) to each of 50 genes. These 50 genes participate in a transcriptional program, where the weights in the linear combination are 1/50 (i.e. the transcriptional program is the mean expression of all the genes. In practice, these weights are more uneven, but this example illustrates the point). In this case, the correlation between the transcriptional program and the marker will be very high (PCC = 0.95). The following MATLAB code illustrates this:

```
x = randn(10000,1); % Marker gene
X = repmat(x,1,50) + 2*randn(10000,50); % 50 genes each with low correlation to marker
corr(x,X) % Average PCC of 0.45

Y = sum(X,2); % Transcriptional program. Linear combination of constituent genes.
corr(x,Y) % PCC ~ 0.95
```

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.

We think we have inadvertently made this a point of confusion. Our reference to “we do not need to capture these genes” was only with respect to defining transcriptional programs. Our point was that, when choosing an appropriate level of the GO hierarchy at which to define programs, we observed that 1/3 of the genes do not have an informative GO annotation (just ‘biological process’). Thus, we were arguing that these should not be considered when using GO to define the transcriptional programs. We were NOT saying that we do not care to predict the expression of these genes. Tradict predicts the expression of ALL genes with detectable expression in the training set (21,000+ genes), regardless of their GO annotation. We have clarified this in the text.

10. How was the range [50,2000] chosen for the GO term size cutoffs? How do the results change when this range is adjusted?

Our rationale for settling on [50,2000] was a data driven decision that we’ve discussed in detail in Part 2 - “Defining transcriptional programs” of the “Tradict algorithm” section of Materials and Methods in the Supplemental Information. Our goal was to define programs that were not too specific but also not too general. At the same time we wanted to maximize coverage of the transcriptome in defining transcriptional programs. It was optimizing these criteria that made us conclude, fairly clearly, that [50, 2000] is the right range.

Though we could adjust this range, we feel it would violate the criteria we’ve laid out above. Nevertheless, it is important to note that while our predictions for program expression would change (since we are changing the programs themselves), gene expression prediction transcriptome wide would not change. This is due to decoupling between gene and program expression prediction, as discussed for previous reviewers comments and in Supplemental Analysis 3.III.

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.

Again, this is a misunderstanding that we have clarified in Part 2 - “Defining transcriptional programs” of the “Tradict algorithm” section of Materials and Methods in the Supplemental Information. See our response to comment #9, two above.

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.

These minor comments have been addressed

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.

As mentioned before for other reviwers' comments, we have toned down the language in our main text to claim less about reconstructing gene-expression and focusing more on program expression. As we show in Supplemental Analysis 3, the PCC and normalized unexplained variance (equivalent to the residual mean squared error divided by the variance of the target being predicted) are highly correlated in rank.

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.

As mentioned before for another reviewer's comments, we acknowledge that it may be difficult for Tradict to detect subprogram variation. However, we argue such situations may be less common. Genes rarely operate in such isolation of others and there is considerable feedback. Disruption of truly critical genes, in general, should have disruptive effects on a detectable portion of the transcriptome. We have made it more clear that our goal is NOT to replace existing transcriptome-wide assays.

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.

At the level of transcriptional programs, we believe we have demonstrated this through our cross validation experiments. All of the test-sets cycle through the entire collection of transcriptomes that we have and program prediction accuracy remains high on all of them. Thus, we feel that at least for program expression our 100 markers are predictive independent of most contexts and applications. Again, we've made it explicit that these claims are specific for program expression prediction.

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.

We think the patterns observed make sense. First we note that the scale shows program z-score across all samples. Thus blue does not mean ‘not expressed.’ It simply means less expressed than yellow. The heatmaps contain samples that span many different biological contexts including different tissues. In our dataset are subsets of samples (e.g. developing/germinating seeds) with the expression of just a handful of programs (e.g. during seed development only a small subset of the transcriptome is expressed). This is what creates the blue bands, in which just a few hundred transcriptomes have “niche” expression of just a handful of transcriptional programs.

There is no question that many transcriptional programs are correlated with each other. This is especially the case when what’s plotted includes the entire spectrum of samples and conditions in the SRA (e.g. the plant programs, “defense response to fungus” and “defense response to bacterium” will look very correlated when looking at their expression patterns across tissues. however their separation will be more clear when examining their expression in plants treated with bacterium vs fungus).

Figure 3c is meant to give a broad pictorial representation of Tradict’s reconstruction performance. We have made this more explicit in the main text.