Response letter

We are grateful to the reviewers for their careful reading and thoughtful comments. We have addressed each concern and hope that the revised manuscript is now acceptable. Each concern is discussed in detail below. Please note that the **Response to Major Comments Summarized by Editor** consists of some general comments raised by referees. We address these issues upfront.

Response to Major Comments Summarized by Editor (Kafadar):

1. *Identifiability: Referee 1 raises a serious issue (#1) echoed by the AE: Are the estimates of the loadings really unique? As the AE writes, you need to either show that the solution is unique or explain why constraints to ensure uniqueness are unnecessary.*

   Response: We have added a new section to address identifiability. In particular, we have stated on p.7 the conditions under which the loading vectors are unique up to sign flips and permutations of indices. These two conditions apply to both the general and constrained order-3 tensor decomposition. The first one is useful for checking uniqueness of a given tensor, while the second one gives a general condition for uniqueness almost everywhere. The new section now reads:

   > "...Before presenting the algorithm, we first state some conditions for the model identifiability. The first complication is the indeterminacy due to sign flips and permutation:
   >
   > – Sign flips: changing the factors from $(\boldsymbol{G}_r, \boldsymbol{I}_r, \boldsymbol{T}_r)$ to $(-\boldsymbol{G}_r, -\boldsymbol{I}_r, \boldsymbol{T}_r)$ does not affect the likelihood.
   >
   > – Permutation: applying permutation to the index set $[R]$ does not affect the likelihood.
   >
   > To deal with the above indeterminacy, we adopt the following convention. The sign of $\boldsymbol{I}_r$ is chosen such that $\max_{j\in[n_I]} I_{r,j} = \max_{j\in[n_I]} |I_{r,j}|$ for all $r \in [R]$. Because of the nonnegativity constraints on $\boldsymbol{T}_r$, this convention fixes the sign of $\boldsymbol{I}_r$ (and thus $\boldsymbol{G}_r$). Furthermore, component indices are arranged such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_R$. In the degenerate case where not all eigenvalues are unique, we break ties by first choosing the module r with larger $\max_{j\in[n_I]} I_{r,j}$.
   >
   > The second complication comes from the possible non-uniqueness of tensor decomposition even after accounting for sign and permutation indeterminacy. Fortunately, we are able to utilize sufficient conditions for the uniqueness of tensor decomposition. These conditions were initially developed for unconstrained tensor decomposition, but they also apply to our semi-nonnegative tensor decomposition.
   >
   > – (Kruskal, 1977) A rank-R semi-nonnegative tensor decomposition is unique if $k_G + k_T + k_I \geq 2R + 2$, where k_G is the Kruskal-rank of the gene factor matrix $\boldsymbol{G} = [\boldsymbol{G}_1, \ldots, \boldsymbol{G}_R]$, i.e., the maximum value k such that any k columns are linearly independent. The definitions for k_T and k_I are similar, except that the tissue factor matrix $\boldsymbol{T} = [\boldsymbol{T}_1, \ldots, \boldsymbol{T}_R]$ is nonnegative in our case.
   >
   > – (De Lathauwer, 2006) Suppose $n_G > n_I > n_T$ (as in the GTEx data). If $R \leq n_T$ and $R(R-1) \leq n_G(n_G-1)n_I(n_I-1)/2$, then the rank-$R$ semi-nonnegative tensor decomposition is unique for almost all such tensors except on a set of Lebesgue measure zero."

2. *Comparison with other approaches: Your simulation compares your* MultiCluster *method to* HOSVD *and* SDA*, but your Figure 4 shows the ROC curves for only different choices of R for your* MultiCluster*. Where do the ROC curves for* HOSVD *and* SDA *fall if they were included in Figure 4? More importantly, Referee 3 wonders how much benefit is provided by your tensor approach, over a simple test of the significance of the gene expression. Please see paragraphs 2-3 in the report from Referee 3 and comments in my report.*

Response: Neither of the previously-developed tensor software packages (*SDA* Hore et al. (2016) and *HOSVD* Omberg et al. (2007)) allows association tests on a gene-specific basis, so we were not able to compare their ROCs with our method. Per the reviewers' suggestions, we have added the standard significance tests in the ROC comparison (see the new Figure 5). Furthermore, we have also added a new Section **4.1 A simple example** to highlight the settings when *MultiCluster* captures the patterns that would be difficult to find using simpler methods such as a meta-analysis or 2D approaches.

The revised Section **4.3 Power to detect differentially-expressed genes** now reads:

> "...To compare to single-tissue tests, we performed standard linear regressions in each tissue separately and declared a gene age-related if its p-value was less than the nominal level in at least one of the 10 tissues. We also performed a fixed-effect meta-analyses by aggregating the age effects across single-tissue tests using z-score method. Neither *SDA* (Hore et al., 2016) nor *HOSVD* (Omberg et al., 2007) allow association tests on single-gene bases, so we did not consider them here.
>
> Figure 5 shows the receiver operating characteristic (ROC) curves for each method. We found that the testing procedure based on tensor projection had higher detection power than single-tissue analyses, demonstrating the advantage gained when tensor-based methods incorporate information from similar tissues. Notably, the power appears stable when the decomposition rank R increases from 3 (the number of latent tissue groups) to 10 (the number of total tissues). We note that the power of a meta-analysis relies on genes being age-related in several tissues with effects primarily in the same direction (Figure 5a). Violations of these assumptions are frequent in applications and result in substantial losses in power (Figure 5b). In contrast, our tensor approach teases apart tissue-specific expression patterns by using eigen-tissues to synthesize information from sufficiently similar tissues. Subsequent examination of the entries of eigen-tissues allows one to determine in which tissues DE patterns are present, something that requires additional steps in meta-analyses."

The new Section **4.1 A simple example** explains what the simple, naive methods miss that *MultiCluster* captures.

> "As a basic illustration, we generated an expression tensor consisting of ....
>
> This example represents a challenging scenario in which traditional methods may fail. For example, if we average the expression over individuals and apply matrix PCA to the resulting data, then neither the mode-specific grouping nor the three-way interaction can be recovered. In fact, matrix PCA (Figure 3a) reveals little information on the gene/tissue clustering. This is because the matricization destroys the three-way structure encoded in the higher-order tensor data.
>
> The standard (fixed-effect) meta-analysis also suffers from low power for detecting

DE genes in this example. To see this, we tested the age effects in each tissue separately and combined the test statistics into a pooled estimate using z-score method (Kelley and Kelley, 2012). This approach detected few DE genes in group A and also exhibited limited power in groups B and C (Figure 3b). The meta-analysis' poor performance is due to the tissue-specificity of DE genes: genes in Gene Group A have opposite age effects in two of the tissue groups, so the signals partially cancel out; moreover, genes in Gene Groups B and C have age effects in only subsets of tissues, potentially diluting observed DE patterns.

In contrast to matrix PCA, the factors from our tensor decomposition ably capture the true clustering patterns (Figure 3c). Furthermore, tensor projection significantly improves detection power across all three gene groups (Figure 3b). As the tissue loadings are used as the weights in the tensor projection (Section 3.3.3), testing based on eigen-tissues allows us to test for age effects in a group-specific fashion. Consider Gene Group A as an example. Genes in this group have opposite age effects in Tissue Groups A and Group B. Since the first eigen-tissue has nearly-zero loadings in Tissue Group A, it only contains information about differential expression in Tissue Group B without including unwanted noise from Tissue Group A. This toy example demonstrates the ability of *MultiCluster* to improve detection power buy automatically identifying similar tissues and borrowing information among them."

3. *Robust Imputation: You allude briefly to "imputation" on p7 (caption of Figure 2), and again briefly on p23 ("robust imputation"). Can you elaborate on this imputation? How many values were "imputed"? Referee 3 notes serious consequences to the reliability of the conclusions depending on where the imputation was done (e.g., all or most genes on a single tissue, or randomly throughout the* $18,841 \times 53 \times 544$ *array?)*

Response: There are two aspects involved: one is how our algorithm handles missing data, and another is the robustness of our biological findings. We address the two aspects separately.

Regarding the algorithm, we have chosen to exclude the missing entries from the cost function. To clarify, we added the following paragraph on p.9:

> "...when some entries Y_{ijk} are missing, tensor decomposition is not well-defined. In such a case, one could instead use the cost function $\sum_{[i,j,k]\in\Omega}(Y_{ijk}-\sum_r \lambda_r G_{r,i} I_{r,j} T_{r,k})^2$, where $\Omega \subset [n_G] \times [n_I] \times [n_T]$ is the index set for non-missing entries. To implement this, we iteratively approximate missing data with fitted values based on current parameter estimates and proceed with the algorithm until convergence. This procedure has been commonly used in matrix factorization (Lee et al., 2010; Lee and Huang, 2014), and we adopt it for tensor factorization."

While this approach alleviates the impact of missing data in each iteration, we still need a complete tensor to initialize the algorithm. This requires a data set-specific imputation. In GTEx, the number of samples (individuals) per tissue ranges from 103 to 544. We excluded 5 tissues with low sample sizes and 4 sex-specific tissues (testis, prostate, pituitary, ovary), which resulted in the 44 somatic tissues analyzed in Section 5.1. Note that if a tissue is missing from an individual, it will be missing data at *every* gene in that tissue (i.e., not missing at random). We therefore took a similar approach as in Wang et al. (2016) to impute the unobserved expression for each gene separately. Briefly, we implemented a k-nearest neighbors imputation scheme which fills in missing entries with the average read

counts from the corresponding tissue in the ten individuals most similar in terms of age, ancestry, and gender. Unlike methods such as matrix completion, this approach preserves the pre-imputation signal in the data and does not appear to introduce erroneous clusterings due to the non-random sample collection procedure. We vetted this procedure by comparing hierarchical tissue trees (Supplemental Figure S5) before and after imputation, and we also validated top genes by looking at the original, non-imputed data.

We have added the following sentence in the revised Section 5:

> "Prior to analysis, we performed a standard data processing procedure described in depth in the Supplementary Material. Briefly, these steps included correction for sequencing depth, removal of lowly expressed genes, log transformation of the data, correction for nuisance variation arising due to technical effects, removal of sex-specific tissues, and imputation of missing data."

4. *Measures of uncertainty: You mention "further extending" statistical inference with "measures of uncertainty such as confidence intervals for [loadings]" using, for example, the bootstrap (pp 23-24). Some might regard such "extensions" as essential. Exactly how would you use the bootstrap to do this? Point estimates alone are rarely sufficient so this item is important.*

   Response: It is beyond the scope of this paper to fully solve the problem of producing confidence regions for loadings, but we provide a brief discussion here for the reviewer's interest.

   Measures of uncertainty, such as confidence intervals for tissue-, gene-, or individual-loadings, could be further extended. One possible approach would be performing parametric bootstrap (Efron and Tibshirani, 1994) to assess the uncertainty in the estimation. For example, one can simulate tensors from the fitted low-rank model (3.1) based on the estimates, and then assess the empirical distribution of the loadings. This approach has been applied in matrix factorization (Milan and Whittaker, 1995) and can be extended to tensor factorization. While being simple, the parametric bootstrap assesses uncertainty only in the estimation procedure but not the modeling. A more comprehensive assessment would involve resampling the RNA-seq data as in *kallisto* (Afgan et al., 2016) which requires a complicated bioinformatics pipeline. We leave it for future study.

5. *Figures and exposition: Reports from me and Referee 3 provide editorial corrections and stylistic changes to the text and figures.*

   Response: We have removed the colors in Figures 4-6 to save the cost for printing.

Other comments by Editor (Kafadar):

1. *p3, l.5: "this reduces to detecting three-way blocks in the expression tensor": I think I understand Fig 5: interest lies in finding those genes that are "significantly" expressed in all (or some) individuals and/or in all (or some) tissues. They may not be ordered so nicely as to appear in a nice "block". They don't really need to be "re-ordered": we just want to know which genes, tissues, and individuals have similarly expressed genes. But visually your approach is clearer if they are re-ordered so we can see that block. Is that correct? BTW, another re-ordering may be necessary to visualize a second or third or fourth block, yes?*

   Response: Exactly, yes. (No alterations to the text.)

2. *p3, l.-10: How much higher is the computational cost, especially a robust one that requires a constrained tensor decomposition? On p14, you describe one scenario where the* MultiCluster *time is 1.7 hours, compared to* HOSVD*'s 1.6 hr and* SDA*'s 20.1 hours, so maybe you want to say here "the computational cost is comparable to some methods and lower than others (see Section 4.4)."*

   Response: The revised text now reads:

   > "...*MultiCluster* is computationally competitive with *HOSVD* while being more computationally efficient than *SDA*."

3. *p3, l.-2: "automatically learns the interplay across different modes of the data": this sentence is not clear to me.*

   Response: The revised text now reads:

   > "This approach...learns the clustering patterns across different modes of the data in an unsupervised manner analogous to PCA and SVD."

4. *p4, para 2, l.1: "The remainder of this paper is organized as follows:" You can delete this sentence; the paragraph itself describes the organization.*

   Response This sentence has been removed.

5. *p7, Fig 2 legend, l.-2: First appearance of "GO". In text, first appearance is page 9, where you do explain the abbreviation, but may need it here also.*

   Response: We have added the explanation for "GO" (Gene Ontology) in the Figure 2 legend.

6. *p10, para 1, l.2: "to those of a number of" to "those from two other tensor methods".*

   Response: We have removed this phrase.

7. *p14, Fig 4: Do you really need color? Won't different line types (solid, dot, dash) suffice to distinguish them? Color is very expensive to print, so we appreciate your attention to this cost-saving measure. Also, won't the reader want to see the ROC curves for HOSVD and SDA also?*

   Response: We have removed the color in Figure 4 (numbered as Figure 4 in the current manuscript). To save the cost, we also removed unnecessary colored panels in Figures 4-6.

   For the comparison with other methods, please see our **Response to Major Comments Summarized by Editor, p.2. #2 Comparison with other approaches.**

8. *p16, Fig 5: Color is not mentioned in the legend. It seems important for only the bar graphs in panels a and d. I'm not sure what the color adds in panel a: all bars are red except for the lone black bar. Panel d has only 5 bars, and if the point of color is to indicate the p-value, then you can do so equally (if not more) effectively by indicating the p-value inside the bar. I don't think 9e-14 is all that much different from 3e-14 anyway, and, again, the legend does not even mention color. Same for Figure 6 on p.17.*

   Response: We have removed the color in Figure 5 (numbered as Figure 6 in the current manuscript).

9. *pp22-23: Referee 3 (paragraphs 2-3) asks how much benefit is provided by your tensor approach, over a simple test of the significance of gene expressions. For example, does your tensor approach "borrow information" across dimensions, or does it have other advantages over simpler approaches? You mention on p.22-23, "MultiCluster identifies coherent clusters across each mode of the data in a single step."' Perhaps you add 1-2 sentences here. In any event, it will be important to state the advantages or show what the simple, naive method misses that MultiCluster captures.*

   Response: To address this concern, we have added a new Section **4.1 A simple example** and revised Section **4.3 Power to detect differentially-expressed genes**. Our response is detailed in **Response to Major Comments Summarized by Editor, p.2. #2 Comparison with other approaches.**

   We have also elaborated this sentence on p.23 into:

   > "...*MultiCluster* constructs clusters across each mode of the data and associates the resulting variation with biological contexts via eigen-genes, -tissues, and -individuals. Each of these resulting components can then serve as the basis for testing, removing the need for many marginal tests."

10. *p23, l.-13: "adopting a robust imputation scheme": Can you explain what this 'robust imputation scheme' is, and how many of the $n_G x n_I x n_T$ values you imputed? Were they missing at random, or were whole tissues missing for some individuals? The former would not concern me but the latter certainly would. [n_G = 18,481 genes, n_T = 53 tissues, n_I = 544 individuals (357 females, 187 males)]*

    Response: Please see our **Response to Major Comments Summarized by Editor, p.3. #3 Robust Imputation.**

11. *p23, l.-2: "Measures of uncertainty, such as confidence intervals for tissue-, gene-, or individual-loadings, would be useful." The paper is already 27 pages long (35% longer than our usual submission), but I think this item is important: measures of uncertainty would be, in my view, not just useful but imperative. I think that your idea of doing some sort of a bootstrap makes sense, but I'm not altogether sure how it would work because you need to retain the 3-D structure. Can you be more specific here, and even calculate them? Given MultiCluster's computational cost (1.7 hours), a small bootstrap might be over 50 hours, but it might be worth it, especially because, if Referee 1 is right that the estimates may not be uniquely identifiable, you'll find out quickly by looking at the bootstrap estimates of the loadings.*

    Response: Please see our **Response to Major Comments Summarized by Editor, p.3. #4 Measures of uncertainty.**

Response to Referee 1

1. *A major concern is the uniqueness property of the proposed method. Is the constraint that all loading vectors are norm-1 sufficient for obtaining unique loading vectors? If the Sub-algorithm (Modified two-mode HOSVD) is used, is the unique estimation of loading vectors guaranteed?*

   Response: Please see our **Response to Major Comments Summarized by Editor, p.1. #1 Identifiability.**

2. *In clustering problem, an important and difficult task is the determination of the optimal cluster number (number of modules R in this manuscript). In the simulation study, for comparing the performance, the best rank R which minimizes the relative error is selected. However, in real data analysis, this determination is not available because $\mathcal{Y}_{true}$ is unknown. With the other performance measures, reasonable comparisons are necessary for the validity of the proposed method. Also, in the GTEx data example, the authors just used $R = 10$ and interpreted those modules biologically. If possible, the assessment of goodness-of-fit (or explained variance) for this tensor model with $R = 10$ would be useful.*

   Response: To address this comments, we have added the following paragraph in Section 3.2. The paragraph on p.9 now reads:

   > "Before concluding this section, we briefly comment on two implementation details. First, the algorithm assumes that R is given. In practice, the rank R is often unknown and must be determined from the data $\mathcal{Y}$. There are many heuristics developed for choosing R in the matrix case, and similar ideas can be adopted here. For example, one can plot the sum of squared residuals (3.1) as a function of R and identify the elbow point in the curve..."

   Furthermore, we also added an assessment of goodness-of-fit for the real data (see Section **5.2 Brain transcriptome Data**). The revised text now reads:

   > "...To assess the goodness-of-fit, we plotted the sum of squared residuals (see equation (3.2)) as a function of rank R (Supplemental Figure S1). Visual inspection suggested $R = 6$ in our case...."

3. *The proposed method is similar to the non-negative tensor factorization (NTF) method which is frequently used in computer vision. Is there any special reason that non-negativity constraint is required only on Eigen-Tissue? If all three loading vectors (tissue, gene, individual) are constrained to be non-negative, the proposed method is same to NTF?*

   Response: Because the GTEx tensor has both positive and negative entries, it is impossible to impose all loading vectors (tissue, gene, individuals) to be non-negative. We require non-negative tissue loadings mainly for interpretability; in this regards, our method enjoys similar merits as NTF. The tissue loadings indicate the "activity" or "strength" of the expression module for each tissue. By examining the entries of the tissue vector, one can infer which tissues drive the signal of differential expression.

   Having said that, the algorithm for NTF is totally different from our semi-nonnegative decomposition. NTF typically uses a multiplicative update rule. In our case, we carefully designed a fast algorithm that is specifically for semi-nonnegativity constraints (Section **3.2 Estimation via optimization**). It compares favorably to competing tensor approaches (Section **4 Numerical comparison**).

The application is also new. We focused on multi-tissue multi-individual expression studies and designed a framework that involves not only tensor decomposition but also the subsequent analysis including characterizing the biological modules, testing for covariate effects, identifying DE-genes, etc (Figure 2). The nonnegative tensor decomposition suits our needs and permits the investigation of transcriptome variation across individuals and tissues simultaneously.

Response to Referee 2

1. *The simulations are quite small in scope and seem designed to show the benefits of the proposed method. The authors' key application is a gene expression dataset in which the gene dimension is large (∼18,000) and there is a strong biological prior that gene loadings are sparse. The authors should include simulations of sparse gene components in a dataset of the same size as the GTEx dataset they analyze.*

   Response: In Section 4.4, we did describe a numerical experiment with 18,000 genes × 500 individuals × 40 tissues. This is of the similar size the GTEx data. Note that it took ≈ 20.1 hours for the competing method, *SDA*, to decompose this tensor once. So a small-scale simulation with 50 trials would take ≈ 40 days for SDA to complete the task, which is computationally onerous.

   We agree that the sparse gene expression is a realistic scenario in many RNA-seq studies. We have added simulations with sparse tensors to the revised manuscript. The revised part in Section **4.2 Accuracy of three-way clustering** now reads:

   > "...Block means $\{\mu_{lmn}\}$ were generated according to the following two block models (as well as sparse versions)..."
   >
   > "For both the additive- and multiplicative-mean models, we also considered a sparse setting in which expression matrices $\mathcal{Y}_{\text{true}}(i,\cdot,\cdot)$ were zeroed out for 90% of genes $i = 1,\ldots,500$..."
   >
   > "The simulation models we consider here span a range of scenarios. ...The sparse setting represents a realistic scenario in RNA-seq studies in which a high number of genes are lowly expressed across individuals and tissues...."
   >
   > "...In particular, the recovery error of *MultiCluster* grows noticeably more slowly than that of *SDA* in the non-sparsity settings (Figure 4a and Figure 4b)... We also found that, even in the sparse settings, *MultiCluster* compares favorably with the other two methods (Figure 4c and Figure 4d). Note that these three methods adopt different regularization schemes: tissue nonnegativity for *MultiCluster*, gene sparsity for *SDA*, and orthogonality for *HOSVD*. Our results suggest the flexibility of *MultiCluster* to handle a range of models."

2. *Simulations can tell you only so much. Can the authors apply one of the other methods to the real dataset and compare?*

   Response: In Section **5.2 Brain transcriptome data**, we have added a new Section **5.2.1 Comparison with other tensor methods**. In particular, we compared *MultiCluster*'s performance with *HOSVD* and *SDA* for this brain tensor; the results are summarized in the revised text and **Supplementary Material**. The new section now reads:

   > "...We also applied *HOSVD* and *SDA* to the brain tensor; the results are summarized in the Supplementary Material (Supplemental Figure S2 and Supplemental Figure S3). Both *MultiCluster* and *HOSVD* successfully clustered the 13 tissues into functionally similar groups, while *SDA* failed in tissue clustering. Furthermore, *MultiCluster* enjoyed better interpretability as it yielded sparse tissue factors. In particular, we found that most expression modules are spatially restricted to specific brain regions, such as the two cerebellum tissues (component 2), three cortex tissues (component 4), and three basal ganglia tissues (component 5). "

3. *Does the algorithm really use two-mode unfolding (as referenced in the AISTATS paper)? It looks to me from p2 of the AISTATS paper and page 8 of this paper, in the section on "Sub-algorithm Modified Two-Mode HOSVD" that one-mode unfolding is being used.*

   Response: We did use the two-mode unfolding as described in the AISTATS paper. The confusion may come from the fact that, for an order-3 tensor, the two-mode unfolding $\mathcal{Y}_{(12)(3)} \in \mathbb{R}^{d_1 d_2 \times d_3}$ (i.e., unfolding along the first two modes) can also be represented as the transpose matrix of the one-mode unfolding $\mathcal{Y}_{(3)(12)} \in \mathbb{R}^{d_3 \times d_1 d_2}$ (i.e., unfolding along the third mode). We were attempting to avoid unnecessary complication in the original manuscript so we implicitly referred to the latter representation.

   Because of the confusion, we realized that we needed to rewrite the algorithm to make it clearer. The detailed algorithm is now described in the **Supplementary Material**.

4. *Can the authors add more comments on their method of non-negativity on the tissue loadings on page 6? This seems to be done by setting un-constrained estimates to zero. Does this give the same result as constrained optimization estimates?*

   Response: We have added the rationale for such a practice. The constrained optimization (3.3) is separable into each of its factors $\boldsymbol{T}_r, \boldsymbol{G}_r, \boldsymbol{I}_r$, so we can optimize it in an iterative block-wise manner. In particular, the solution of $\boldsymbol{T}_r$ (for fixed $\boldsymbol{G}_r$ and $\boldsymbol{I}_r$) has a closed form, which is precisely the thresholded vector we provided.

   We have added related comments at p.8 of the revision:

   > "...As the optimization (3.3) is separable into each of its factors, we can optimize this in an iterative block-wise manner:
   >
   > **Property 1.** *Let $(\widehat{\lambda}_r, \widehat{\boldsymbol{G}}_r, \widehat{\boldsymbol{I}}_r, \widehat{\boldsymbol{T}}_r)$ be the optimizer of (3.3). Then the following properties hold (assuming the denominators are non-zero):*
   >
   > $$\begin{aligned} \widehat{\boldsymbol{G}}_r &= \mathcal{Y}(\cdot,\ \widehat{\boldsymbol{I}}_r,\ \widehat{\boldsymbol{T}}_r)/\|\mathcal{Y}(\cdot,\ \widehat{\boldsymbol{I}}_r,\ \widehat{\boldsymbol{T}}_r)\|_2, \\ \widehat{\boldsymbol{I}}_r &= \mathcal{Y}(\widehat{\boldsymbol{G}}_r,\ \cdot,\ \widehat{\boldsymbol{T}}_r)/\|\mathcal{Y}(\widehat{\boldsymbol{G}}_r,\ \cdot,\ \widehat{\boldsymbol{T}}_r)\|_2, \\ \widehat{\boldsymbol{T}}_r &= \mathcal{Y}(\widehat{\boldsymbol{G}}_r,\ \widehat{\boldsymbol{I}}_r,\ \cdot)_+/\|\mathcal{Y}(\widehat{\boldsymbol{G}}_r,\ \widehat{\boldsymbol{I}}_r,\ \cdot)_+\|_2, \\ \widehat{\lambda}_r &= \mathcal{Y}(\widehat{\boldsymbol{G}}_r,\ \widehat{\boldsymbol{I}}_r,\ \widehat{\boldsymbol{T}}_r), \end{aligned} \tag{1}$$
   >
   > *where $a_+ := \max(a, 0)$ and we allow this operator to be applied to vectors in an element-wise manner.*
   >
   > A proof is provided in Supplementary Material. The above result suggests an alternating optimization scheme. The tensor factors $\widehat{\boldsymbol{G}}_r$, $\widehat{\boldsymbol{I}}_r$ and $\widehat{\boldsymbol{T}}_r$ are initialized using outputs from unconstrained tensor decomposition (Wang and Song, 2017). Each factor is then updated alternatively while keeping the other two factors fixed. The update step requires solving a (either constrained or unconstrained) least-square problem and the optimal solution is given explicitly by the right-hand side of equality (1). In particular, the entrywise nonnegativity of the tissue loading vectors $\widehat{\mathbf{T}}_r$ is imposed by setting negative values of $\widehat{\mathbf{T}}_r$ to 0. As each coordinate update reduces the objective function, which is bounded below by 0, convergence of this scheme is assured....The full algorithm is provided in the Supplementary Material."

5. *Can you say a bit more in the main paper about the imputation method you have used. The GTEx dataset has much missing data, so the performance of this step seems quite crucial. It would be useful to know what the limits of this imputation step are. When will it breakdown?*

   Response: Please see our **Response to Major Comments Summarized by Editor, p.3. #3 Robust Imputation.**

Minor comments:

1. *In Section 4 it could be made clearer how Y_{ijk} is related to μ_{lmn}.*

   Response: We have added the following sentence to clarify:

   > "Let $\mathcal{Y}_{\text{true}}$ denote the noiseless tensor with three-way block means generated from each of the above schemes, i.e., $\mathcal{Y}_{\text{true}}(i,j,k) = \mu_{lmn}$ when i is in block l, j in block m, and k in block n....The observed expression data were then simulated as $\mathcal{Y} = \mathcal{Y}_{\text{true}} + \mathcal{E}$, where $\mathcal{E} \in \mathbb{R}^{500\times 50\times 10}$ is a random Gaussian tensor with i.i.d. $N(0,\sigma^2)$ entries."

2. *The idea of using tensor components to help reduce multiple testing burden is not novel to the authors' paper. The authors should cite the relevant literature.*

   Response: We have added the citation and revised the text as:

   > "One benefit of *MultiCluster*...is the substantially reduced number of comparisons which must be considered (Hore et al., 2016)."

3. *Will the authors be making software available? If so, can they provide a URL?*

   Response: Yes, we have added the following sentence in the Section **Supplementary Material**.

   > "Our software, including the GTEx data used in the analysis, will be publicly available at https://github.com/songlab-cal/MultiCluster."

Response to Referee 3

Major comments:

1. *While the paper contains some interesting results, it seemed to me that many of these results could likely have been obtained by simpler and more direct methods.*

   For example, the paper highlights several genes that show interesting associations with covariates (eg age, sex) across several tissues. However, it seems that there are much simpler ways to go about identifying such genes. The most obvious way is to do a test (gene expression vs age say) in each tissue, and then combine tests across tissues with some kind of meta-analysis.

   For example, most simply by summing the z scores across tissues (effectively a fixed effects meta-analysis). Is it really better to first do a somewhat-complex unsupervised dimension reduction before incorporating the covariate? If so, why? The comparisons with single-tissue analysis in Fig 4 are not really convincing because a meta-analysis is the obvious comparison here. (Furthermore, many of the weaknesses of traditional meta-analyses seem not to be solved by the tensor approach. For example, if you find a gene associated with age through either a tensor or meta-analysis approach, how do you know in which tissues it is associated? The authors here return to the single-tissue results, which is understandable but also not ideal...)

   Response: Please see our detailed **Response to Major Comments Summarized by Editor, p2. #2 Comparison with other approaches.**

   Briefly, we have added the meta-analysis in the ROC comparison (see the new Figure 5). We have also added a new Section **4.1 A simple example** to highlight the settings when MultiCluster captures the patterns that would be difficult to find using meta-analysis. Unlike standard meta-analysis, the tensor projection does not require single-tissue results to identify the driving tissue. To clarify this, we have added the following sentence in Section 3.3.3:

   > "...By examining the entries of the tissue vector $\widehat{\mathbf{T}}$, we can infer which tissues drive the signal of differential expression...."

2. *The paper also highlights several interesting modules identified in the tensor factorization. However, in most cases the modules appear to be defined primarily by variation across tissue and across genes, with relatively little variation across individuals. This is evident in Figures 5 and 6 for example. Would you not obtain equally good results by just averaging across individuals at each gene in each tissue and doing an appropriate matrix factorization of the 2-d matrix? It seems likely that other factors - such as the decision whether or not to allow negative loadings on genes, and whether to analyze the gene expression on the log scale or unlogged scale - may well be more influential than incorporating individual as a third mode.*

   More generally, since matrix factorization methods are so plentiful and widely used, it would be really helpful to show illustrative examples (even hypothetical, or simulated) that are designed to showcase the benefits of tensor factorization? that is, settings where the tensor factorization can illuminate structure that would be difficult to find by using 2-d approaches.

   To summarize, the authors need to provide a more thorough demonstration that, for at least some tasks, the tensor-based analysis gives you insights that simpler methods do not. A concrete starting point would be to compare with simple fixed effects meta-analysis across tissues to Fig 4, and with a 2-d analysis that averages across individuals in Figs 5 and 6. And a bigger-picture explanation of when tensor methods are likely to be preferred over simpler 2-d counterparts would be very helpful.

Response: Please see our detailed **Response to Major Comments Summarized by Editor, p2. #2 Comparison with other approaches.**

There seem to be three questions here involving the variability across individuals: 1. visualization; 2. the specific data set; 3. comparison with 2D methods. We address each of them in turn.

We realized that Figure 5c (numbered as Figure 6c in the revised manuscript) is not the perfect device to visualize the variation – both the mean and variation of the loadings are low, partly due to normalization. To make it more clear, we have added Figure 6e (boxplot of individual loading vs. age). The trend of the loadings with age is clearly depicted. In fact, using a linear model, age explains explains 24.4% ($p < 2 \times 10^{-16}$) variation in the individual loadings (Section 5.1.2).

In this "specific dataset", we focused on two tissue collections, one consisting of 44 somatic tissues and another consisting of 13 brain tissues. For the former collection, we did make the following comment in the original paper,

> "...most eigen-individuals have limited descriptive power compared to eigen-genes and eigen-tissues (Supplemental Table S1). This was expected because variation in gene expression is usually lower among individuals than among tissues (Melé et al., 2015). Consequently, we turned our attention to smaller tensors of similar tissues to fully showcase *MultiCluster*'s three-way clustering capabilities...."

For the 13 brain tissues, we found that "many expression modules in the brain also exhibit considerable individual-specificity" (Section 5.2.3), supported by Figure 7 and Section 2. In particular, we are able to characterize the source of individual variability due to age and sex (Section 5.2.3).

Regarding "comparison with 2D methods", we have added a new Section **4.1 A simple example** to highlight the settings when *MultiCluster* captures the patterns that would be difficult to find using 2D approaches. As suggested by reviewers, we have focused on two aspects of the comparison: (1) comparing *MultiCluster* with simple fixed effects meta-analysis across tissues (Figure 3c and Figure 5), and (2) comparing *MultiCluster* with a 2D analysis that averages across individuals (Figure 3a andFigure 3b). Higher-level explanations have also been added in the new Section 4.1 and revised Section 4.3. Please see the **Response to Major Comments Summarized by Editor, p2. #2 Comparison with other approaches** for details.

3. *There needs to be more discussion and assessment of the practical issues of applying these methods to the GTEx data. How were the data processed? Is it log-transformed or raw counts? Does the tensor decomposition model make sense on both the original and log scale? (This also comes up indirectly in the simulations, where you consider both the additive mean model and multiplicative mean model. Are your methods equally suited to both? Which one makes sense where?) Is it corrected for any covariates, surrogate variables, sequence depth etc? Only later in the discussion do we learn that the data were imputed to make a complete tensor. This seems potentially problematic because, when a tissue is missing from an individual, it will be missing data at *every* gene in that tissue, which is impossible to impute reliably. What about the sex-specific tissues - are they removed?*

Response: In the original manuscript, we did describe the detailed preprocessing steps in the **Supplementary Material**. To make the reference more explicit, we have added the

following sentences in the beginning of Section 5.

> "...Prior to analysis, we performed a standard data processing procedure described in depth in the Supplementary Material. Briefly, these steps included correction for sequencing depth, removal of lowly expressed genes, log transformation of the data, correction for nuisance variation arising due to technical effects, removal of sex-specific tissues, and imputation of missing data...."

For further discussions on imputation, please see our **Response to Major Comments Summarized by Editor, p.2. #3 Robust imputation.**

4. *The 3d figures are difficult to assess. How do you deal with over-plotting in these figures? Are they transparent colors? It is not crucial in a schematic like Figure 1, but for Figure 3, where you are trying to compare an estimate with the truth, I think it would be helpful to unfold the 3-d cube to show a 2-d representation so that accuracy is easier to judge. Also, in Fig 3, are the results all on the same color scale? Why are the input data so white?*

   Response: The 3D figures are purely a display device to visualize the effect of our tensor method – the numerical accuracy was listed at the side. We realized that such color panels may be distracting (and costly to print, pointed out by the editor), so we have removed the 3D figures here.

5. *throughout: generally I suggest the term "ancestry" should be used instead of the more controversial term "race".*

   Response: "Race" has been replaced with "ancestry" in both the main text and the supplement.

6. *Is "semi-nonnegative" a widely used term? I did not understand it here.*

   Response: A nonnegative decomposition would be one in which all singular vectors are nonnegative. Since we only require this for the tissue mode, we called it semi-nonnegative. We have added the following sentence in the last paragraph of Section 3:

   > "...Note that no sign constraint is imposed on individual and gene loadings, so our method is flexible enough to handle mixed-sign data tensors. We refer to such constraints as "semi-nonnegative" tensor decomposition."

7. *"modes" of the data may not be familiar to readers unfamiliar with tensor jargon.*

   Response: A clarification was added in Section 3, p.4:

   > "...We use $\mathcal{Y} = [\![Y_{i_1 i_2 \ldots i_k}]\!] \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_k}$ to denote a $(d_1, d_2, \ldots, d_k)$-dimensional real-valued tensor, where k corresponds to the number of modes of $\mathcal{Y}$ and is called the order..."

   and in Introduction:

   > "...can be organized into a three-way array, or order-3 tensor, with gene, tissue, and individual modes..."

   We also added the explanation in the first appearance of "mode" in the Introduction. The revised text now reads:

   > "...a statistical method that integrates multiple modes (defined in Section 3)...."

8. *p3: " may not be powerful enough.." this reads a bit oddly. If you show evidence that is not powerful enough then say so. (Maybe power is not the right way to talk about a prior?)*

   Response: The wording has been altered to:

   > "...may not be suitable to detect tissue- and individual-wise correlation."

9. *ep4: "elderly" seems an inaccurate description of these subjects (these are post-mortem samples so the studied individuals are deceased; and from the age ranges given, many of them died quite young it seems).*

   Response: "Elderly" has been replaced by the following statement:

   > "The GTEx data set contains categorical clinical variables such as... age (1st and 3rd age quantiles of 47 and 62, respectively)...."

10. *p11, the definition of RelErr does not make sense to me. It is a function of $\hat{\mathcal{Y}}_{est}$ so you can't minimize over rank($\hat{\mathcal{Y}}_{est}$) ≤ 10. Also I don't understand the need for minimizing over π you need to do that kind of thing only if you try to examine accuracy of individual loadings that are not identifiable; the reconstruction itself is identifiable. Also where $\hat{Y}_{est}$ denotes the rank R approximation ($R = 1, \ldots, 10$) is unclear - it should have superscript R if it depends on R.*

    Response: For a given tensor $\mathcal{Y}_{\text{true}}$ and its noisy observation $\mathcal{Y}$, RelErr is a function of R. We have changed the notation $\hat{Y}_{\text{est},R}$ to make the dependence on R more explicit. The revised text now reads:

    > "We assessed the recovery accuracy of each algorithm using the relative error, defined as
    >
    > $$\text{RelErr} = \min_{R \leq 10} \frac{\left\| \widehat{\mathcal{Y}}_{\text{est},R} - \mathcal{Y}_{\text{true}} \right\|_F^2}{\|\mathcal{Y}_{\text{true}}\|_F^2},$$
    >
    > where $\widehat{\mathcal{Y}}_{\text{est},R}$ denotes the rank-R approximation obtained from tensor decomposition."

11. *p11: why do you look at only the reconstruction of Y, and not accuracy of individual estimated loadings?*

    Response: The simulated tensors are generated from more complicated models than assumed in (3.1). The intent of our simulations is to assess the robustness of our tensor method, and for a wide range of simulated models (additive, sparse, etc), there are no ground-truth loading vectors to compare against.

12. *p12: simulation studies are only as convincing as the way the data are simulated. How about simulating by adding random noise to real data and comparing methods in their ability to reconstruct the original data? (Obviously the real data themselves will not exactly fit the tensor model, but the additional error associated with that would be the same for every method?)*

    Response: Our tensor model assumes that the real data is a low-rank tensor plus random noise. The reconstruction error would depend on how close the real data is to the low-rank assumption. We interpret the referee's comment as a concern on goodness-of-fit. To address this concern, we have added the Section **5.2.1. Comparison with other tensor methods** to the real data:

"To assess the goodness-of-fit, we plotted the sum of squared residuals (see equation (3.2)) as a function of rank R (Supplemental Figure S1). Visual inspection suggested $R = 6$ in our case. We also applied *HOSVD* and *SDA* to the brain tensor; the results are summarized in the Supplementary Material (Supplemental Figure S2 and Supplemental Figure S3). Both *MultiCluster* and *HOSVD* successfully clustered the 13 tissues into functionally similar groups, while *SDA* failed in tissue clustering. Furthermore, *MultiCluster* enjoyed better interpretability as it yielded sparse tissue factors...."

13. *p13: "Additionally, the number of ...": this is not true: in both cases there is one test per gene. The minimum p value across tissues produces a single test statistic for each gene, and you are plotting the ROC curve for that test statistic. Thus the ROC curves are already directly comparable.*

    Response: We have removed this sentence.

14. *p13: typo "was was"*

    Response: We have corrected it to "what was".

15. *p14: is the high abundance of mitochondrial genes not simply due to the large number of mitochondria in every cell?*

    Response: The sentence now reads

    "The top genes in the corresponding eigen-gene (Supplemental Table S1) are mainly mitochondrial genes (15/20 top genes), comporting with their high transcription rates and the large number of mitochondria within most cells (Melé et al., 2015)..."

16. *p16: what does it mean "a gene clustering that is biologically coherent with brain × aging"? Are you saying your results are consistent with previous studies on aging in brain? If so, citations would be helpful.*

    Response: We have added a citation, and the revised text now reads:

    "...a gene clustering that is biologically coherent with aging signals in the brain (Yang et al., 2015)..."

17. *p16: "specific to tissues with roles in the immune system" seems inaccurate, in that Fig 6 shows appreciable loadings on many non-brain tissues. "Primarily driven by two blood tissues" also seems a stretch - they do have the highest loadings.*

    Response: The revised text now reads:

    "...captures an expression module heavily loaded on tissues with roles in the immune system...."

    "Primarily driven by..." has been altered into:

    "..the eigen-tissue is led by two blood tissues, the spleen, and the liver."

18. *p20: "the X-Y gene pair PCDH11X/Y" appears to be a pair of genes. So it cannot be "the top sex-related gene".*

    We have changed "the top sex-related gene" into "the sex-related signal".

19. *p20: emboldened → bold*

    Response: We have replaced "emboldened" by "bold".

20. *Fig 5d: the p value color scale is odd: they are all around* 10^{-14}*, and this is inconsistent with the text. Similar comment on Fig 6d.*

    Response: We have updated the Figure 5d (now numbered as Figure 6d in the revised manuscript). We have labeled the p-value for each GO in the figure.

21. *provide a reference for "BH correction". (Or, perhaps better, simply give the uncorrected p values.)*

    Response: A citation was added.

References

Afgan, E., Baker, D., Van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., *et al.*, 2016. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic acids research*, **44**(W1):W3–W10.

De Lathauwer, L., 2006. A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. *SIAM journal on Matrix Analysis and Applications*, **28**(3):642–666.

Efron, B. and Tibshirani, R. J., 1994. *An introduction to the bootstrap*. CRC press.

Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M. I., Small, K., and Marchini, J., 2016. Tensor decomposition for multiple-tissue gene expression experiments. *Nature Genetics*, **48**(9):1094–1100.

Kelley, G. A. and Kelley, K. S., 2012. Statistical models for meta-analysis: A brief tutorial. *World journal of methodology*, **2**(4):27.

Kruskal, J. B., 1977. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, **18**(2):95–138.

Lee, S. and Huang, J. Z., 2014. A biclustering algorithm for binary matrices based on penalized bernoulli likelihood. *Statistics and Computing*, **24**(3):429–441.

Lee, S., Huang, J. Z., and Hu, J., 2010. Sparse logistic principal components analysis for binary data. *The annals of applied statistics*, **4**(3):1579.

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., *et al.*, 2015. The human transcriptome across tissues and individuals. *Science*, **348**(6235):660–665.

Milan, L. and Whittaker, J., 1995. Application of the parametric bootstrap to models that incorporate a singular value decomposition. *Applied Statistics*, :31–49.

Omberg, L., Golub, G. H., and Alter, O., 2007. A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. *Proceedings of the National Academy of Sciences*, **104**(47):18371–18376.

Wang, J., Gamazon, E. R., Pierce, B. L., Stranger, B. E., Im, H. K., Gibbons, R. D., Cox, N. J., Nicolae, D. L., and Chen, L. S., 2016. Imputing gene expression in uncollected tissues within and beyond gtex. *The American Journal of Human Genetics*, **98**(4):697–708.

Wang, M. and Song, Y. S., 2017. Tensor Decompositions via Two-Mode Higher-Order SVD (HOSVD). In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 614–622.

Yang, J., Huang, T., Petralia, F., Long, Q., Zhang, B., Argmann, C., Zhao, Y., Mobbs, C. V., Schadt, E. E., Zhu, J., *et al.*, 2015. Synchronized age-related gene expression changes across multiple tissues in human and the link to complex diseases. *Scientific reports*, **5**:15145.