

Three-way clustering of multi-tissue multi-individual gene expression data using constrained tensor decomposition

Miaoyan Wang¹, Jonathan Fischer², and Yun S. Song^{1,2,3}

December 5, 2017

¹*Computer Science Division, University of California, Berkeley, CA 94720, USA*

²*Department of Statistics, University of California, Berkeley, CA 94720, USA*

³*Chan Zuckerberg Biohub, San Francisco, CA 94158, USA*

Abstract

The advent of next generation sequencing methods has led to an increasing availability of large, multi-tissue datasets which contain gene expression measurements across different tissues and individuals. In this setting, variation in expression levels arises due to contributions specific to genes, tissues, individuals, and interactions thereof. Classical clustering methods are ill-suited to explore these three-way interactions, and struggle to fully extract the insights into transcriptome complexity and regulation contained in the data. Thus, to exploit the multi-mode structure of the data, new methods are required. To this end, we propose a new method, called *MultiCluster*, based on constrained tensor decomposition which permits the investigation of transcriptome variation across individuals and tissues simultaneously. Through simulation and application to the GTEx RNA-seq data, we show that our tensor decomposition identifies three-way clusters with higher accuracy, while being 11x faster, than the competing Bayesian method. For several age-, race-, or gender-related genes, the tensor projection approach achieves increased significance over single-tissue analysis by two orders of magnitude. Our analysis finds gene modules consistent with existing knowledge while further detecting novel candidate genes exhibiting either tissue-, individual-, or tissue-by-individual specificity. These identified genes and gene modules offer bases for future study, and the uncovered multi-way specificities provide a finer, more nuanced snapshot of transcriptome variation than previously possible.

Introduction

Owing to advances in high-throughput sequencing technology, multi-tissue expression studies have provided unprecedented opportunities to investigate transcriptome variation across tissues and individuals (Lonsdale et al. 2013; Melé et al. 2015; Walker et al. 2004; Nica et al. 2011; Hawrylycz et al. 2012). A typical multi-tissue experiment collects gene expression profiles (e.g. via RNA-seq or microarrays) from different individuals across a number of different tissues, and variation in expression levels often results from complex interactions among the different modes of the data (Melé et al. 2015). For example, a group of genes may perform coordinated biological functions in certain contexts (e.g. specific tissues or individuals), but may behave differently in other settings through tissue- and/or individual-dependent gene regulation mechanisms. An improved understanding of these interactions will yield insight into fundamental biological questions with clinical applications such as changes in cellular function (Dönertaş et al. 2017), personalized transcriptomics (Montgomery and Dermitzakis 2011), and disease susceptibility (Melé et al. 2015).

Clustering has proven useful to reveal latent structure in high-dimensional expression data (Kluger et al. 2003; Kiselev et al. 2017; Wiwie et al. 2015). Traditional clustering methods (such as K-means, PCA, and t-SNE (Maaten and Hinton 2008)) assume that gene expression patterns persist across one of the different contexts (either the tissue or individual mode), or assume that samples within one mode are i.i.d. or homogeneous. Direct application of these algorithms to multi-tissue expression data requires concatenating all available samples from different tissues into a single matrix, precluding potential insights into tissue \times individual specificity (Bahcall 2015). Alternatively, inferring gene modules separately for each tissue ignores commonalities among tissues and may hinder the discovery of differentially expressed genes that characterize tissues or tissue groups. Likewise, individuals vary by their biological attributes (such as ethnicity, gender and age), and ignoring such heterogeneity impedes the accurate estimation of gene- and/or tissue-wise correlation (McCall et al. 2016). The development of a statistical method that integrates multiple modes simultaneously is therefore essential for elucidating the complex biological interactions present in multi-tissue multi-individual gene expression data.

Several methods have been proposed in multi-tissue multi-individual expression studies, but they are often unable to fully exploit the three-mode structure of the data. Pierson et al. (2015) propose a hierarchical transfer learning algorithm to learn gene networks in which they first construct a global tissue hierarchy based on mean expression values and subsequently infer gene networks for each tissue conditioned on the tissue hierarchy. Dey et al. (2017) instead use topic models to cluster samples (i.e. tissues or individuals), and based on tissue clusters they identify genes that are distinctively expressed in each cluster. Both algorithms take a two-step procedure to uncover expression patterns in tissues and genes. Other methods aim to take a one-shot approach by identifying subsets of correlated genes that are exclusive to, for example, female individuals. Gao et al. (2016) adopt the biclustering framework and propose decomposing the expression matrix into biclusters consisting of subsets of samples and features with latent structure unique to the overlap of particular subsets. However, in the case of multi-tissue measurements across individuals, concatenating the data sample-wise to create a single expression matrix will not explore the three-way interactions among genes, tissues, and individuals. A more recent work (Hore et al. 2016) develops sparse decomposition of arrays (*SDA*) for multi-tissue expression experiments. Because their focus is not on clustering tissues or individuals, the proposed i.i.d. prior on individual/tissue loadings may not be powerful enough to detect tissue- and individual-wise correlation.

In this paper, motivated by recent success in tensor decomposition (Wang and Song 2017; Kuleshov et al. 2015; Sidiropoulos et al. 2017), we address the aforementioned challenges by using a tensor-based approach to jointly cluster genes, tissues, and individuals. We utilize triplets of

sorted loading vectors in a constrained tensor decomposition to represent three-way clusters specific to subsets of genes, tissues, and individuals. This approach handles heterogeneity in each mode and automatically learns the interplay among different modes of the data. We propose a simple but principled statistical procedure to characterize the identified three-way clusters on the basis of available metadata. Our method uncovers several different types of gene expression modules, including (i) global, shared expression modules; (ii) expression modules specific to certain subsets of tissues; (iii) modules with differentially expressed (DE) genes across individual-level covariates (e.g., age, sex or race); and (iv) expression modules that are specific to both tissues and individuals.

Results

Overview of our *MultiCluster* method

Below, we briefly describe a new clustering method, *MultiCluster*, for identifying three-way clustering patterns in multi-tissue multi-individual gene expression data. Additional details can be found in [Materials and Methods](#).

As illustrated in Figure 1a, multi-tissue multi-individual gene expression measurements can be organized into a 3-way array, or order-3 tensor, with gene, tissue, and individual modes. Our goal is to identify subsets of genes that are similarly expressed in subsets of tissues and individuals; mathematically, this reduces to detecting 3-way blocks in the expression tensor (Figure 1b). These local blocks may correspond to, for example, gene expression modules that are active in some but not all tissues and individuals. We utilize the flexible tensor decomposition framework to directly identify gene modules in a tissue \times individual specific fashion, which traditional clustering methods would fail to capture.

Figure 2a–d provide a schematic illustration of the *MultiCluster* method. Briefly, *MultiCluster* takes as input the multi-tissue multi-individual gene expression data, $\mathcal{Y} = \llbracket Y_{ijk} \rrbracket \in \mathbb{R}^{n_G \times n_I \times n_T}$, where Y_{ijk} represents the expression value (possibly after a suitable transformation) of gene i measured in individual j and tissue k (Figure 2b). Our algorithm decomposes \mathcal{Y} into a sum of rank-1 components,

$$\mathcal{Y} \approx \sum_{r=1}^R \lambda_r \mathbf{G}_r \otimes \mathbf{I}_r \otimes \mathbf{T}_r, \quad (1)$$

where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_R \geq 0$ are singular values in descending order, and $\mathbf{G}_r, \mathbf{I}_r$ and \mathbf{T}_r are norm-1 loading vectors indicating the relative contribution of each gene, individual, and tissue to the r -th component (Figure 2c). We interpret the rank-1 tensor $\mathbf{G}_r \otimes \mathbf{I}_r \otimes \mathbf{T}_r$ as the basic unit of expression pattern (called an expression module), in which the (i, j, k) -th entry of $\mathbf{G}_r \otimes \mathbf{I}_r \otimes \mathbf{T}_r$ is the multiplicative product of the corresponding entries in the three modes, i.e., $(\mathbf{G}_r \otimes \mathbf{I}_r \otimes \mathbf{T}_r)_{(i,j,k)} = G_{r,i} I_{r,j} T_{r,k}$. Each module may represent a particular interaction of biological processes. Genes with large G_r -values are affected more greatly by the tissues and individuals in the module r , whereas these effects are also greater for individuals with larger I_r -values and tissues with larger T_r -value (Figure 2d). To facilitate interpretation, we impose entrywise non-negativity on the tissue loading vectors \mathbf{T}_r by zeroing negative values of $T_{r,k}$. By examining the entries in the tissue vectors \mathbf{T}_r , one can learn the tissue activity pattern of the associated module. Note that no sign constraint is imposed on individual and gene loadings, therefore the method handles mixed-sign data tensors. We adopted our earlier algorithm ([Wang and Song 2017](#)) based on successive rank-1 approximations to solve the tensor decomposition (1).

Extending earlier work ([Alter et al. 2000](#); [Omberg et al. 2007](#)), we refer to the loading vectors $\mathbf{G}_r, \mathbf{I}_r, \mathbf{T}_r$ as “eigen-genes”, “eigen-individuals”, and “eigen-tissues”, respectively. To characterize

the biological significance of the inferred modules, we utilize metadata to identify the sources of variation in each loading (Figure 2d). For each eigen-gene, we identify gene ontology (GO) terms enriched among the top-ranked genes, and annotate each eigen-tissue using its positively-loaded tissues. An expression module is declared “tissue-specific” if the eigen-tissue is dominated by only a few tissues, and “age-, sex-, or race-related” if the eigen-individual loadings are correlated with age, sex, or race, respectively. We also test whether these individual-level covariates can account for the variation in eigen-individuals and employ tensor projection onto the eigen-tissues to pinpoint the genes which drive the signal in covariate-associated expression modules.

Summary of simulation results

To test the performance of *MultiCluster*, we carried out a simulation study and compared the results with a number of alternative methods. The simulations served two purposes: 1) to evaluate the ability of *MultiCluster* to recover multi-way clusters under various cluster models, and 2) to assess the power of our tensor-based procedure to detect DE genes and compare with the power in standard single-tissue analysis.

Accuracy of three-way clustering. We applied *MultiCluster* to identify 3-way blocks in noisy expression tensors simulated from three different cluster models: additive-, multiplicative-, and combinatorial-mean models, as detailed in [Materials and Methods](#). We compared the recovery accuracy with two recently developed tensor methods: (i) sparse decomposition of arrays (*SDA*) ([Hore et al. 2016](#)), and (ii) tensor higher-order singular value decomposition (*HOSVD*) ([Omberg et al. 2007](#)). Both *MultiCluster* and *SDA* are built upon the Canonical Polyadic (CP) decomposition which decompose a tensor into a sum of rank-1 matrices, whereas *HOSVD* decomposes a tensor into a core tensor multiplied by an orthogonal matrix in each mode.

As seen in the example tensors depicted in Figure 3, *MultiCluster* is able to recover the block structure well in all three scenarios, demonstrating its robustness to model misspecification. We used the relative estimation error (up to permutation) across 50 simulations to assess the recovery accuracy of each method ([Materials and Methods](#)). We found that *MultiCluster* consistently outperforms the other two methods (Figure 3a-c). For the additive and multiplicative models, the two non-Bayesian methods (*MultiCluster* and *HOSVD*) tend to recover the block structure better than *SDA* (Figure 3a-b). A possible explanation is that *SDA* is designed to cluster genes rather than tissues and individuals, so the i.i.d. prior imposed on tissues/individuals may not be powerful enough to detect local blocks, especially the blocks are small. It may also be due the algorithmic stability of *MultiCluster* relative to *SDA*; the latter usually requires multiple restarts in order to reduce spurious components ([Hore et al. 2016](#)). For the more complicated combinatorial model, however, the two CP decompositions (*MultiCluster* and *SDA*) yield lower errors than the *HOSVD* (Figure 3c). Note that neither *MultiCluster* nor *SDA* forces orthogonality in the loading vectors; instead, they adopt some sparsity and regularity (tissue nonnegativity for *MultiCluster* and gene sparsity for *SDA*). The way we incorporated tissue non-negativity essentially introduces zeros into the rank-1 tensor output, thereby fitting the 3-way blocks more flexibly in a local fashion.

Power to detect differentially-expressed genes. To study how our tensor projection procedure affects the ability to detect covariate-associated genes, we simulated age-related genes. Expression tensors were generated according to the additive model with 10 tissues (3 tissue groups), and 100 genes (out of the 500 in total) were randomly chosen to be age-related. Each age-related gene was active in at least one tissue cluster, and for every tissue in the active cluster, the age effect size was independently drawn from a $\text{Unif}[0, 0.05]$ (up-regulated) or $\text{Unif}[-0.05, 0]$ (down-regulated)

distribution. We decomposed each simulated tensor into $R = 3, 5, 10$ components and applied our tensor-projection procedure ([Materials and Methods](#)) to test for age-related genes. We declared a gene age-related if its p -value is less than the nominal significance level in at least one of the R eigen-tissues. To see whether we have improved power over single-tissue tests, we performed standard linear regressions in each tissue separately and declared a gene age-related if its p -value is less than the nominal level in at least one of the 10 tissues.

Figure 4 shows the receiver operating characteristic (ROC) curves that compare the true positive rate vs. false positive rate for each method. We found that the testing procedures based on tensor projection have higher detection power than the single-tissue analysis, demonstrating the advantage of tensor-based methods in incorporating information across similar tissues. In particular, the power seems to be stable when the decomposition rank R is increased from 3 (the number of latent tissue groups) to 10 (the number of total tissues). It is worth noting that the number of association tests conducted in the tensor projection based approaches are fractions of those conducted when tissues are considered one at a time. Since the burden imposed by multiple testing corrections is a common concern in genome-scale testing, our tensor approach is especially attractive because one may perform fewer tests while achieving higher power.

Run time. To compare computational performance, we simulated a large order-3 tensor of 18,000 genes \times 500 individuals \times 40 individuals. The size of this tensor is to mimic the size of GTEx RNA-seq dataset. We recorded the run times for each method while decomposing this tensor into 10 components. It took 6,047 seconds (≈ 1.7 hrs) for *MultiCluster* and 73,989 seconds (≈ 20.1 hrs) for *SDA* to finish the run. The run time of *HOSVD* (5,849 seconds) is roughly the same as that of *MultiCluster*.

Identifying expression modules across tissues and individuals in GTEx data

We applied our method to the GTEx V6 gene expression data, which consists of 8,555 RNA-seq samples collected from 544 human individuals across 53 tissues. Each RNA-seq sample measures the read counts of 56,238 annotated transcripts, of which roughly 25,000 are putative protein-coding regions, 10,000 are long non-coding RNAs (lncRNA), and 10,000 are pseudogenes. For convenience, we henceforth refer to all of these transcripts as “genes.” After performing our data processing procedure and removing low-expressed genes ([Materials and Methods](#)), we organized the expression measurements into a gene \times individual \times tissue tensor $\mathcal{Y} \in \mathbb{R}^{n_G \times n_I \times n_T}$, where $n_G = 18,481$, $n_I = 544$ and $n_T = 44$. This tensor, which we named the GTEx tensor, is used in the global tissue analysis.

To investigate the dominant features in the human transcriptome, we performed a global, all-tissue clustering analysis to identify gene \times tissue \times individual expression modules. We applied *MultiCluster* to the GTEx tensor, excluding chromosome Y-linked genes and sex-specific tissues. Supplemental Table S1 summarizes the top expression modules along with their characterization.

Component I – shared, global expression. Examining the eigen-tissue loadings informs us in which tissues the corresponding module is active. As expected, the first eigen-tissue and eigen-individual are essentially flat (Supplemental Figure S1a and Supplemental Figure S1c), so this expression module captures baseline global expression common to all samples. The top genes in the corresponding eigen-gene (Supplemental Figure S1b) are mainly mitochondrial genes (15/20 top genes), comporting with their pervasive transcription across tissues as a result of the mitochondrion’s critical role in cellular energy production (Kelly et al. 2012; Melé et al. 2015). In addition, we detected several non-mitochondrial genes, most of which are related to essential protein synthesis

functions and eukaryotic cell activities (Supplemental Figure S1d). For example, *ACTB* encodes highly conserved proteins (Valente et al. 2009) and is known to be involved in various types of cell motility (Fishilevich et al. 2016). Two other nuclear genes, *EEF1A1* and *EEF2*, encode eukaryotic translation elongation factors, and their isoforms are widely expressed in the brain, placenta, liver, kidney, pancreas, heart, and skeletal muscle (Fishilevich et al. 2016).

Component II – brain tissues. The second eigen-tissue clearly separates brain tissues from non-brain tissues, with the pituitary gland being the only non-brain tissue in the cluster (Figure 5a). We note that while not explicitly labeled as a brain tissue, the pituitary gland protrudes from the base of the brain. The sharp decline in tissue loadings (Figure 5a) highlights the distinctive expression pattern in the brain. We found that, in the eigen-individual (Figure 5c), age explains more variation (24.4%, $p < 2 \times 10^{-16}$) than sex (0.3%, $p = 0.12$) or race (4.3%, $p = 2.3 \times 10^{-8}$) (Figure 5e; Materials and Methods). The eigen-gene (Figure 5b) produces a gene clustering that is biologically coherent with the brain \times aging signal. Indeed, we observed an enrichment of genes from the glutamate receptor signaling pathway (e.g. *CACNG7*, *CACNG3*, *GRIN11*; $p = 1.2 \times 10^{-20}$), chemical synaptic transmission (e.g. *SEZ6*, *GRM5*, *GRIA2*; $p = 1.8 \times 10^{-16}$), excitatory postsynaptic potential (e.g. *RIMS2*, *CHRNA2*, *RIMS1*; $p = 2.4 \times 10^{-16}$), and memory (e.g. *SLC24A2*, *JPH3*, *SYT4*; $p = 1.2 \times 10^{-11}$) (Figure 5d). Among the 899 genes in this cluster, we identified 675 age-related genes using tensor-projection (with significance threshold $\alpha = 10^{-3}/899 \approx 10^{-7}$ via Bonferroni correction; see Materials and Methods), 556 of which exhibit decreased expression with age (Supplementary Data). The association of brain disease and neurological disorders with age is well-documented, and our findings support that aging affects brain tissues in a manner not shared by other tissues. We present further evidence of multi-way clustering in the brain in Fine structures in subtensors of similar tissues.

Component III – tissues involved in immune response. The third component captures an expression module specific to tissues with roles in the immune system. The eigen-tissue is primarily driven by two blood tissues (whole blood and EBV-transformed lymphocytes), the spleen, and the liver (Figure 6a). These tissues serve as mediators of direct immune response (whole blood and lymphocytes), the production and storage of antibodies (spleen), and filtering of antigens (spleen and liver). Our eigen-tissue also reveals the subtle involvement of esophagus in this module (Figure 6a). In fact, mast cells (part of immune system) have been found at the basement membrane and in the lamina propria of normal esophageal mucosa (Collins 2014), reflecting the mixture of cell types in esophagus tissues.

Correspondingly, the eigen-gene loads heavily on immunity-related genes (e.g. *IGHM*, *FCRL5*, *IGJ*, *MS4A1*) (Figure 6b). The eigen-individual does not correlate with any covariate as strikingly as the brain does with age, but we do find a significant correlation with race (explaining 4.5% variation among individuals, $p = 5.8 \times 10^{-7}$; Figure 6e). The top genes in the eigen-gene are functionally related to the B cell receptor signaling pathway (e.g. *BLK*, *CD79A*, *IGHG3*; $p = 3.0 \times 10^{-15}$), humoral immune response mediated by circulating immunoglobulin (e.g. *IGHM*, *IGHD*, *IGHA1*; $p = 7.5 \times 10^{-13}$), the phagocytosis recognition (e.g. *IGHA2*, *IGHG1*, *IGHG2*; $p = 5.3 \times 10^{-10}$), and plasma membrane invagination (e.g. *AURKB*, *IGHV3-23*, *IGHG4*; $p = 2.1 \times 10^{-9}$) (Figure 6d).

Component IV – tissues with structural similarities. The fourth gene expression module is enriched for collagen catabolic process (e.g. *COL8A1*, *COL15A1*, *COL3A1*; $p = 4.5 \times 10^{-12}$), collagen fibril organization (e.g. *DPT*, *LUM*, *COL14A1*; $p = 1.1 \times 10^{-11}$), and multicellular organismal catabolic process (e.g. *MMP2*, *COL6A2*, *OL1A1*; $p = 1.5 \times 10^{-11}$) (Figure 7d). These GO terms

represent the synthesis of extracellular matrix proteins that give structure support and elasticity to tissues. For example, the top-ranked gene, *MYOC*, encodes the protein *myocilin*, which is believed to play an important role in cytoskeletal structure (Fishilevich et al. 2016). The second-ranked gene, *PLN*, is associated with the protein *phospholamban*, which regulates the activity of cardiac and smooth muscle (Fishilevich et al. 2016). Coupled with this gene cluster, the tissue cluster is heavily loaded on three artery subtypes (tibial, aorta, coronary). Other highly-ranked tissues include adipose, esophagus, heart, and colon (Figure 7a). These organs are rich in connective fibers with stretch-recoil properties, endowed with the molecular functions we identified earlier. Furthermore, we found that the genes driving this module are expressed at slightly lower rates in women and African-Americans (Figure 7e). The top sex-related gene is XG (X-linked blood group), and the top race-related gene is TUSC5 (Tumor suppressor candidate 5).

Other expression modules identified in the global analysis. Each of the remaining expression modules is active in only a subset of tissues, indicating the presence of tissue specificity (Supplemental Table S1). These detected modules are specific to skin (exposed and non-exposed), cell lines (EBV-transformed lymphocytes and transformed fibroblasts), liver, muscle (skeletal and cardiac), and cerebellar regions (Supplemental Table S1). Of note is the strong signal of underexpression in the cerebellum in women relative to men. We also found that tissues derived from the same embryologic origin tend to be clustered together; for example, the 4th eigen-tissue groups together blood vessels, heart, smooth muscles and connective tissues (mesoderm), whereas the 7th eigen-tissue groups together liver, pancreas and kidney (endoderm). As seen in Supplemental Figure S2–Supplemental Figure S7, the skin-specific module is enriched with keratin-related genes, the two cell lines are enriched with genes responsible for cell division (e.g. chromosome segregation, meiosis, sister chromatid segregation), the liver-specific module is enriched with inflammation and triglyceride related genes, the muscle-specific module is enriched with myofibril related genes, and the cerebellum-specific gene module is enriched for excretion and dorsal spinal cord development. All of these ontologies are consistent with the function of the associated tissues. Conversely, most eigen-individuals have limited descriptive power compared to the eigen-gene and eigen-tissue (Supplemental Table S1). This was expected because variation in gene expression is usually much lower among individuals than among tissues (Melé et al. 2015). In a subsequent section (Fine structures in subtensors of similar tissues), we detail how targeted analyses of subtensors comprised of similar tissues help mitigate this effect by revealing patterns of variation which are overwhelmed by the large heterogeneity in tissues at the global level.

Relationship between different expression modules. The aforementioned modules are naturally ranked according to the amount of variability explained by the associated tensor components (Materials and Methods). We chose to focus on the top 10 components based on their interpretability. Unlike matrix SVD, the components obtained from tensor decomposition are usually not guaranteed to be orthogonal (Supplemental Figure S8). Examination of pair-wise correlations revealed relatively low levels of overlap between most eigenvectors. We did find modest agreement (Pearson’s $r = 0.44$) between gene cluster 4 (tissues with smooth and cardiac muscle) and gene cluster 9 (skeletal and cardiac muscle) (Supplemental Figure S8). Given that these clusters both correspond to muscle-driven eigen-tissues, the similarity is sensible. Likewise, tissue cluster 8 represents a mixture of blood and muscle tissues and overlaps most with the tissue cluster 3 (blood cluster, $r = 0.59$) and tissue cluster 9 (muscle cluster, $r = 0.43$). Nevertheless, we found that the various components in our GTEx analysis tend to capture non-redundant information (Supplemental Table S1), as later components are able to uncover fine-scale structure nested in earlier components. For example, the two cerebellum tissues are separated from other brain regions in the 10th com-

ponent (Supplemental Figure S7), although they all belong to the larger brain cluster in the 2nd component (Figure 5a). In fact, the eigen-genes in these two clusters prioritize different sets of genes; the brain cluster prioritizes general neural genes such as *OPALIN* and *GFAP*, whereas the cerebellum cluster prioritizes more specific genes such as *ZP2* and *SLC22A31*. Recent study shows that the expression of *ZP2* in brain is both spatially-restricted and time-regulated (i.e. transiently enriched during a narrow time window) (Kang et al. 2011), and our analysis demonstrates that *ZP2* is almost exclusively expressed in the cerebellum but not in other tissues, including non-cerebellar regions of the brain.

Fine structures in subtensors of similar tissues

Although our global analysis successfully uncovers distinctive expression patterns in the GTEx data, it may miss finer-scale structure within similar tissues or within similar individuals because of the high degree of inter-tissue heterogeneity. We suspected that restricting the analysis to relatively homogeneous tissues would increase the discriminative power of eigen-individuals and provide information about which genes separate similar tissues. In order to reveal the crucial individual \times tissue specificity, we considered six tissue groups, each consisting of morphologically and functionally similar tissues (Materials and Methods). For each tissue group, we built a subtensor and applied *MultiCluster* to identify gene \times individual \times tissue modules in the subtensor.

Spatially-restricted and sex/age-related expression in the brain. Table 1 shows the expression modules detected in the brain subtensor. 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). The corresponding gene clusters capture distinctly-expressed genes that are over- or under-expressed in each region. Genes over-expressed in the cerebellum region are strongly enriched for dorsal spinal cord regulation ($p = 9.8 \times 10^{-7}$) whereas the under-expressed genes are most strongly enriched for forebrain development ($p = 3.4 \times 10^{-8}$). An opposite enrichment pattern is observed for basal ganglia region. This is consistent with the spatial location of the cerebellum (located in the hindbrain) and basal ganglia (which is situated at the base of the forebrain). In addition, we noticed an abundance of over-expressed *HOX* genes in the spinal cord (cervical C-1) compared to other brain regions (Figure 8a). The *HOX* gene family (*HOXA-HOXD*) is a group of related genes that control the body plan and orientation of an embryo. The non-uniform expression of *HOX* genes across brain regions may suggest the particularly important role of the spinal cord during early embryogenesis.

In addition to tissue-specificity, most expression modules also exhibit considerable individual-specificity. We identified two sex-related and six age-related expression modules from the top tensor components (bolded in Table 1). The second gene module is found to be both cerebellum-specific and sex-related. By ranking genes based on their p -values for sex effect in the direction of eigen-tissue (Materials and Methods), we found that the top sex-related gene in this module is X-Y gene pair *PCDH11X/Y*. In fact, the combined expression of *PCDH11X/Y* is significantly lower in the cerebellum (paired t -test p -value $< 2 \times 10^{-16}$) and in females ($p = 8.0 \times 10^{-11}$), and the expression level also decreases with age ($p = 3 \times 10^{-3}$). Notably, *PCDH11X* was the first reported gender-linked susceptibility gene for late-onset Alzheimer’s disease (Carrasquillo et al. 2009), and it may also be implicated in developmental dyslexia (Veerappa et al. 2013). However, its homologous gene on the Y-chromosome, *PCDH11Y*, is believed to have different expression regulation. Studies show that *PCDH11X* and *PCDH11Y* are differentially regulated by retinoic acid. This acid stimulates the activity of *PCDH11Y* but suppresses *PCDH11X* (Priddle and Crow 2013), perhaps explaining the sex-specificity we observed for this gene pair in most brain tissues.

Significant age effects are widely present in the identified expression modules (Table 1). In particular, age explains over 15% individual-level variation in module 4 (cortex), module 7 (hippocampus), and module 8 (all brain tissues). Notably, the hippocampus is associated with memory, in particular long-term memory, and is vulnerable to Alzheimer’s disease (Lam et al. 2017). In the module 4, *GPR26* is found to be the top age-related gene. Using linear regression models (Materials and Methods), we confirmed the significant decrease of *GPR26* expression with age in all three cortex tissues (cortex, $p = 1.9 \times 10^{-18}$; frontal cortex, $p = 8.8 \times 10^{-12}$, anterior cingulate cortex, $p = 1.9 \times 10^{-7}$) but not in the substantia nigra ($p = 0.17$) or cerebellum ($p = 0.64$). It is worth noting that both the substantia nigra and cerebellum have zero loadings in the eigen-tissue (Table 1), so our tensor-based approach automatically detects the tissue-specificity of this aging pattern. In line with our findings, a recent study shows that *GPR26* plays an important role in the degradation of intranuclear inclusions in several age-related neurodegenerative diseases (Mori et al. 2016).

Another age-related module is component 8 (Table 1; age explains 22.1% individual-level variation), in which every tissue has a non-zero loading in the eigen-tissue. We found that the 2nd top gene in this expression module is *SERPINA3*, a gene implicated in Alzheimer’s disease (Ciriyam et al. 2016). This gene has a moderate age effect in each single tissue (p -values ranging from 1.1×10^{-6} to 0.05 with all effects in the same direction). Using tensor projection, we obtain $p = 1.0 \times 10^{-8}$ in the direction of eigen-tissue, improving the single-tissue analysis by two orders of magnitude. This increased power is one example of an advantage our tensor-based approach via the sharing information across similar tissues. See Supplementary Data for the complete list of p -values for covariate-associated genes obtained using our approach.

Tissue-specific and race/sex-related expression in cardiac and skeletal muscles. The muscle sub-tensor consists of gene expression profiles sampled from two heart regions of cardiac muscle (atrial appendage, left ventricle) and skeletal muscle. As seen from Supplemental Table S2, the top five eigen-tissues reveal the hierarchy-based similarity among the three tissues. Eigen-tissues 2 and 3 represent the muscle and heart clades, respectively, whereas eigen-tissues 4 and 5 further partition the heart clade into each of its constituent components. The corresponding gene clusters capture the differentially expressed genes which drive the tissue partition. In particular, the 3rd gene cluster comprises 511 genes that are similarly expressed in the heart tissues but have distinctive expression patterns in the heart relative to skeletal muscle. By projecting the expression tensor through the 3rd eigen-tissue (Materials and Methods), we identified 122 race-related genes and 95 sex-related genes in this gene cluster (Supplementary Data). Comparatively, the corresponding single-tissue analyses uncover only 91 race-related (86 sex-related) in the left ventricle and 107 race-related (91 sex-related) genes in the atrial appendage, again displaying the increased detection power of our tensor method for genes with moderate but concordant effects across tissues. One such covariate-associated gene, *TCF21*, is a tumor suppressor gene involved in dilated cardiomyopathy which exhibits consistently decreased expression in African-Americans ($p = 4.7 \times 10^{-12}$ in the ventricle; $p = 8.1 \times 10^{-17}$ in the atrial appendage), in females ($p = 1.8 \times 10^{-6}$ in the ventricle; $p = 2.3 \times 10^{-12}$ in the atrial appendage), and in the elderly ($p = 9.4 \times 10^{-4}$ in the ventricle; $p = 2.8 \times 10^{-9}$ in the atrial appendage). Using our tensor-based joint analysis to test for individual effects, we improve the statistical significance to $p = 2.9 \times 10^{-20}$ for race effect, 3.3×10^{-13} for gender effect, and 1.3×10^{-9} for age effect, respectively.

Gender-driven distinction between breast and two adipose tissues. In the adipose sub-tensor, we detected several modules representing highly expressed genes in breast tissue and females. For example, the eigen-tissue in module 2 (Supplemental Table S3) loads on breast tissue only and 40% of the individual-level variation is attributable to sex. Such a pattern highlights the gender-driven

distinction between breast tissue and the other two adipose (subcutaneous and visceral) tissues. More importantly, the top five genes in this module (*SCGB2A2*, *KRT17*, *VTCN1*, *PIP*, *MUCL1*) are breast cancer biomarkers (Barh 2014; Lacroix 2006; Merkin et al. 2017; Naderi and Vanneste 2014). Each of these five genes has a significant sex effect in the breast tissue (p ranging from 9.8×10^{-8} to 1.1×10^{-18}), but not in the other two adipose tissues (p ranging from 0.003 to 0.62). In addition to the risk genes themselves, genes known to be co-expressed with them also tend to be included in this module. Indeed, the secretoglobins *SCGB1D2* and *SCGB2A1*, known to be reliably co-expressed with *SCGB2A2* (Lacroix 2006), were also highly ranked (13rd and 51st, respectively) in the gene cluster.

Tissue-specific and age-related expression in three artery types. In the artery subtensor, the most distinct tissue is the tibial artery, with the 2nd eigen-tissue clearly separating it from the other two arteries (coronary and aorta). The corresponding eigen-gene peaks in the *HOXA* and *HOXC* regions (Figure 8b), indicating the overexpression of *HOX* genes in the tibial artery relative to the coronary artery and aorta. Of note is the famous lncRNA *HOTAIR*, the first RNA gene found to regulate distantly located genes throughout the genome. *HOTAIR* gene is located inside the *HOXC* locus and plays a key role in the initiation and progression of different types of cancer. In addition, this tibial-specific expression module exhibits significant age-relatedness; in particular, 14.9% of the individual-level variation is attributable to age. A further investigation reveals that this aging signal is mostly driven by the group of genes at the negative end of the eigen-gene (Supplemental Table S4). Among the 517 genes in the cluster, we detected 207 age-related genes (with significance threshold $\alpha = 10^{-3}/517 \approx 1.9 \times 10^{-6}$ via Bonferroni correction), 206 of which are over-expressed with age (Supplementary Data). The top age-related gene is *ARHGEF28*, encoding a member of the Rho guanine nucleotide exchange factor family. The encoded protein may be involved in amyotrophic lateral sclerosis (ALS), a neurodegenerative disorder that affects the movement of arms, legs, and body (Droppelmann et al. 2013).

Expression patterns in reproductive tissues. The subtensor analyses of gender-specific reproductive tissues (ovary, uterus, and vagina for female; prostate and testis for male) also reveal interesting gene expression patterns. We observed a clear uterus \times age specificity in the female subtensor (Supplemental Table S5) and a prostate \times age/race specificity in the male subtensor (Supplemental Table S6).

In the female tensor, component 4 is an age-related expression module which also distinguishes the uterus from the other two tissues (ovary, vagina). The top genes in this gene cluster are *CHRD12* (also known as *BNF1*, Breast Tumor Novel Factor 1), *DPP6* (Dipeptidyl Peptidase VI), *TEX15* (Testis Expression 15) and *ZCCHC12* (Zinc Finger CCHC-Type Containing 12). These genes tend to be related to reproductive functions such as DNA double-strand break repair (*TEX15*), or be involved in X-linked disease (*ZCCHC12*). *DPP6* expression changes are associated with age and is preferably expressed in the uterus compared to the other two tissues (paired t -test $p < 2 \times 10^{-16}$). In particular, *DPP6* expression decreases significantly with age in the uterus only (p -value for age = 1.3×10^{-16} compared to $p = 0.05$ in ovary and $p = 0.51$ in vagina). A recent study (Chettier et al. 2014) reveals that *DPP6* harbors a copy number variant locus, rs758316, that is significantly associated with endometriosis. While the function of *DPP6* in uterus remains unclear, its unique aging pattern makes it a good candidate for further investigation.

In the male subtensor, we found that prostate and testis are characterized by two distinct clusters of genes (components 2 and 3 in Supplemental Table S6). Genes over-expressed in prostate are mostly related to prostate glandular acinus development (e.g. *HOXB13*, *FOXA1*) and fluid transport (e.g. *SLC14A1*, *UPK3A*). Among the top five genes, *KLK3*, *ACPP*, and *MSMB* encode

the three predominant proteins secreted by a normal human prostate gland. Their protein level in serum is commonly used for monitoring prostate disorders such as prostatitis (*KLK3* and *MSMB*) or prostate cancer (*ACPP*, *MSMB*, *HOXB13*). Genes over-expressed in testis, on the other hand, are mostly enriched for sperm motility (e.g. *TNP1*, *AKAP4*, *SMCP*; $p = 1.4 \times 10^{-17}$), meiosis I (e.g. *DMRTC2*, *SYCP1*, *BRDT*; $p = 7.9 \times 10^{-17}$), and male meiosis (e.g. *BRDT*, *DDX4*, *TEX15*; $p = 5.1 \times 10^{-7}$). In addition, we found that the prostate-specific modules tend to be race- or age-related (components 2, 4, 7, 8 and 9 in Table S6). The top race-related gene in module 2 is *SPINK2*, with a higher average expression in black than white Americans ($p = 9.0 \times 10^{-4}$). *SPINK2* encodes a serine protease inhibitor located in the spermatozoa, and recent evidence shows that *SPINK2* deficiency leads to fertility changes by causing sperm defects in individuals with one defective copy and azoospermia in those with two defective copies (Kherraf et al. 2017). Given the high degree of race-relatedness among gene expression patterns in the prostate, it is relevant to note the large discrepancy between prostate cancer rates in black and white Americans (153.9 vs 86.8 per 100,000, respectively, according to U.S. Cancer Statistics Working Group (2017)). Although the GTEx cohort does not include individuals with cancer, the strong dependence of prostate cancer incidence rates on race suggests that some of the genes identified as race-related may be involved in the development of prostate cancer and merit further study.

Common expression features in subtensors

Analysis of subtensors allows us to focus on one tissue group at a time, revealing a finer-scale characterization of transcriptional variation in different parts of the body. In addition to tissue comparisons within each subtensor, it is interesting to examine how expression modules identified in different tissue groups compare, and several intriguing features emerge from this meta-analysis.

We found that genes belonging to the same family tend to be clustered closely together. For example, the genes *ZIC1* and *ZIC4* always co-occur in gene clusters (Supplemental Figure S9, component 3 in Supplemental Table S3, component 6 in Supplemental Table S1). *KRT13* is often paired with *KRT4* (component 10 in Supplemental Table S2 and Supplemental Table S4), and *HBA1/HBA2* have similar tissue loadings (Supplemental Figure S10). As we reduce the dimension of expression data from thousands of genes to a handful of eigen-genes, these co-expressed genes provide a validation of our gene groups. Many gender-related expression modules also prioritize *XIST* (e.g. component 5 in Supplemental Table S4 and components 5, 6, 8 in Supplemental Table S3). In fact, *XIST* is one of the most famous lncRNA genes essential for X-inactivation process and female survival (da Rocha and Heard 2017). The wide presence of *XIST* in our analysis reinforces its crucial role in gender-differentiated expression in tissues.

Several subtensors yield similar eigen-genes, suggesting the presence of common expression patterns shared by seemingly unrelated tissues. In particular, we identified three eigen-genes, one in each of the female and male subtensors and another in the artery subtensor (Supplemental Figure S11) that exhibit clearly similar gene loadings. The three eigen-genes are mainly loaded in four genomic loci encoding immunoglobulin: the *IGK*, *IGJ*, *IGH*, and *IGL* regions on chromosomes 2, 4, 14, and 22, respectively. Among other top genes, we identified *FCRL5*, *CH3L1*, and *CHIT1*, all of which are related to immune response. The repeated appearance of this expression module in different tissues and individuals highlights the similar roles of distinct tissues in disparate bodily systems. Despite its presence in each of these tissue groups, this module exhibits differential expression between tissues within each tissue group. Namely, these immune genes are more expressed in the vagina, prostate, and aorta compared to the other reproductive tissues (ovary/uterus, testis) (Supplemental Figure S11) and artery types (tibial/coronary). Interestingly, this module exhibits a strong race effect in the prostate, with higher average expression in black than white men (ex-

plaining 14% variation in the corresponding eigen-individual, $p = 3.1 \times 10^{-9}$; see component 7 in Supplemental Table S6). Such non-uniform expression reflects the complex relationship between related tissues and individuals which our method is well suited to uncover.

Discussion

We have presented a new multi-way clustering method, *MultiCluster*, and demonstrated its utility in identifying three-way gene expression patterns in multi-tissue multi-individual experiments. To the best of our knowledge, our model is the first tensor decomposition that uses nonnegativity constraints and the sharing of information across modes to detect multi-modal specificities in this context. We are able to highlight gene expression modules that are common to all tissues/individuals or exclusive to particular tissue \times individual combinations. Our method uncovers three-way specificities with clear statistical and biological significance in both simulations and the GTEx dataset. We have provided evidence that the distinctions among human tissue gene expression profiles are usually driven by a small set of functionally coherent genes and that many age-, race-, gender-related genes exhibit tissue-specificity even within functionally similar tissues.

A major benefit of *MultiCluster* (and tensor-based methods in general) over existing tissue comparison method (Consortium et al. 2015) is the substantially reduced number of comparisons which must be considered. If one wanted to analyze every possible tissue pairing in a set of n tissues, roughly n^2 analyses would have to be performed and the results would need to be synthesized via a meta-analysis. The analysis can be even more prohibitive if one wanted to examine the 2^n possible tissue-specificity and sharing configurations (Consortium et al. 2015). In contrast, *MultiCluster* identifies coherent clusters across each mode of the data in a single step and associates the resulting variation with biological contexts. Though prior knowledge of tissue function can greatly reduce the number of pairwise comparisons, doing so constrains potential insights to the set of hypothesized tissue modules. For instance, components III and IV of our global tensor decomposition consist of diverse tissues which may not have been grouped together a priori. An additional benefit of *MultiCluster* is the ranking of tissue modules by amounts of variation.

We also implemented a tensor projection procedure to test whether differentially-expressed genes correlate with biological attributes (age, sex or race) and found that we achieve improved power relative to single-tissue tests. This approach can be naturally extended to (trans-)eQTL analyses by testing the projected expression of each gene against genetic variants across the genome. Alternatively, one can test each individual loading vector against genetic variants to identify eQTLs (Hore et al. 2016). Existing multi-tissue eQTL analyses usually proceed by identifying eQTLs in each tissue separately before combining single-tissue results via meta analysis (Battle et al. 2017). However, the large numbers of genes, tissues, and genetic variants potentially incur a substantial penalty for multiple testing and there is also the risk of under-powered tests due to limited sample sizes. An avenue worthy of pursuit is to apply *MultiCluster* for eQTL discovery in large multi-tissue expression studies.

Several consortium efforts — including the Cancer Genome Atlas (Weinstein et al. 2013), the Allen Human Brain Atlas (Hawrylycz et al. 2012), and the Encyclopedia of DNA Elements (Consortium et al. 2004) — have been recently completed or continue to collect data, and our method provides a powerful computational tool to analyze the genome-scale transcriptional profiles that they produce. Although we have presented *MultiCluster* in the context of multi-tissue multi-individual gene expression data, the general framework applies to more general multi-way datasets. One possible extension is integrative analysis of omics data (Berger et al. 2013), in which multiple types of omics measurements (such as gene expression, DNA methylation, microRNA) are collected

in the same set of individuals (Weinstein et al. 2013). In such cases, tensor decomposition may be applied to a stack of data matrices or correlation matrices, depending on the specific goals of the project. Other applications include multi-tissue gene expression studies under different experimental conditions in which one may be interested in identifying 4-way expression modules arising from the interactions among individuals, genes, tissues, and conditions. The tensor framework can also be applied to time-course multi-tissue gene expression (Almon et al. 2003). In this instance one may treat time as the 4th mode and extend the tensor projection approach to identify the time trajectories of 3-way expression modules.

One assumption made by our algorithm is that expression matrices for different tissues are of the same dimension. In the present work, we obtain a complete tensor prior to decomposition by adopting a robust imputation scheme. A possible approach which avoids the need for imputation is to make use of the connection between tensor decomposition and joint matrix factorization (Hore et al. 2016; Xiao et al. 2014; Kuleshov et al. 2015). For example, one could model the n_G -by- n_{I_t} expression matrix \mathbf{M}_t , where t indexes the tissue, as $\mathbf{M}_t \approx \mathbf{A}\mathbf{\Lambda}_t\mathbf{B}_t$ with some identifiability conditions. This model is a relaxation of tensor decomposition because it allows different tissues to have different column (individual) spaces \mathbf{B}_t while sharing the same row (gene) space \mathbf{A} . The diagonal matrix $\mathbf{\Lambda}_t$ captures the tissue-sharing and specificity as before. Another potential approach is to implement tensor imputation and decomposition simultaneously via a low-rank approximation, an idea which has roots in the matrix literature (Troyanskaya et al. 2001; Candès and Recht 2009). We did not pursue this direction beyond initial investigation because of the computational cost and bias introduced by the non-random sample collection in the GTEx data.

Materials and Methods

Data Processing

Here we describe our data processing steps.

Normalization and quality control. To prepare for comparisons across samples, normalization was performed using the size factors produced by the *estimateSizeFactors* function of DESeq2 (Love et al. 2014). After normalizing, we applied quality control measures at both the tissue and gene levels to refine our results and restrict our analyses to informative features. Specifically, we required at least 15 samples to include a given tissue and an average of at least 500 normalized reads in one or more tissues to retain a gene.

Correction for nuisance variation. There were several technical covariates whose effects we wished to remove in order to focus on the correlation between gene expression and biological and phenotypic characteristics. The choice of these factors was driven by a preliminary step in which we looked for signs of significant correlations between any one technical covariate and expression levels. After curating the list of technical covariates in this manner, we were left with the sample collection cohort (postmortem, organ donor, surgical), ischemic time (IT, in minutes), whether the patient died while on a ventilator, and the date of RNA sequencing. Evidence of effects due to some of these factors has been discussed previously elsewhere (McCall et al. 2016).

To correct for the variation due to these factors while preserving the impact of phenotypes, we ran multiple linear regression for every tissue-gene pair per the following linear model:

$$\log(Z_{ijk} + 1) = \beta_1^{ik} + \beta_2^{ik} 1_{\text{female}}^j + \beta_3^{ik} 1_{\text{African-American}}^j + \beta_4^{ik} \text{Age}^j + \beta_5^{ik} 1_{\text{organ donor}}^j + \beta_6^{ik} 1_{\text{surgical}}^j \\ + \beta_7^{ik} 1_{\text{IT} \leq 300}^{jk} + \beta_8^{ik} 1_{\text{IT} \in (300, 900)}^{jk} + \beta_9^{ik} 1_{\text{ventilator}}^j + \beta_{10}^{ik} 1_{\text{sequencing} \leq 7/01/12}^{jk} + \varepsilon^{ijk},$$

where $\epsilon^{ijk} \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma_{ik}^2)$. Here Z_{ijk} is the normalized read count in gene i , individual j , and tissue k . The superscripts on coefficients and covariates indicate to which attribute(s) (gene, individual, tissue) they correspond.

After fitting this set of models, we removed the estimated effects due to the aforementioned technical covariates. To obtain the log-transformed corrected expression value Y_{ijk} , we computed

$$Y_{ijk} = \log(Z_{ijk} + 1) - \hat{\beta}_5^{ik} 1_{\text{organ donor}}^j - \hat{\beta}_6^{ik} 1_{\text{surgical}}^j - \hat{\beta}_7^{ik} 1_{\text{IT} \leq 300}^{jk} - \hat{\beta}_8^{ik} 1_{\text{IT} \in (300, 900)}^{jk} - \hat{\beta}_9^{ik} 1_{\text{ventilator}}^j - \hat{\beta}_{10}^{ik} 1_{\text{sequencing} \leq 7/01/12}^{jk},$$

for all $i = 1, \dots, n_G$, $j = 1, \dots, n_I$, and $k = 1, \dots, n_T$, where n_G , n_I , and n_T denote the number of genes, individuals, and tissues, respectively.

Imputation of unobserved entries. Applying our tensor decomposition method necessitates a complete set of observations in which we have the RNA-seq gene read counts for all individuals in all considered tissues. To obtain the requisite data structure from the initial incomplete set of observations, we implemented a k -nearest neighbors imputation scheme which fills missing entries with the averaged read counts from the corresponding tissue in the ten individuals most similar in terms of age, race, and gender. This method preserves the pre-imputation signal in the data and does not appear to introduce erroneous clusterings due to the non-random sample collection procedure as validated by comparing hierarchical tissue trees (Supplemental Figure S12) before and after imputation. For tissue hierarchy, we took the mean across individuals to produce a gene-by-tissue matrix and computed the distance matrix based on the tissue-tissue Spearman correlation matrix. The hierarchy tree was constructed using UPGMA (Unweighted Pair Group Method with Arithmetic Mean) algorithm (Sokal 1958).

Handling sex-specificity at the tissue and gene levels. As the GTEx cohort comprises both male and female samples, it does not make sense to compare some tissues and genes when using the full set of individuals. To remedy these concerns, we held out sex-specific tissues (e.g. testis, uterus, etc) and only considered them in smaller analyses in the appropriate gender. Further, we also removed all Y chromosome genes save those in the pseudoautosomal region, whose reads we combine with their X chromosome paralogs.

Tissue groups considered in subtensor analysis. The following six tissue groups are considered in the subtensor analysis: (i) 13 brain tissues; (ii) three artery tissues (tibial, aorta, coronary); (iii) two adipose tissues (subcutaneous, visceral) and breast - mammary tissue; (iv) three muscle tissues (heart - atrial appendage, heart - left ventricle, and muscle - skeletal); (v) three female-specific tissues (ovary, uterus, vagina); and (vi) two male-specific tissues (testis, prostate).

MultiCluster via semi-nonnegative tensor decomposition

In a multi-tissue gene expression experiment, the data take the form of an order-3 tensor, $\mathcal{Y} = [\![Y_{ijk}]\!] \in \mathbb{R}^{n_G \times n_I \times n_T}$, where Y_{ijk} denotes the (corrected or imputed) expression value of gene i measured in individual j and tissue k . We propose to model the expression tensor \mathcal{Y} as a perturbed rank- R tensor,

$$\mathcal{Y} = \sum_{r=1}^R \lambda_r \mathbf{G}_r \otimes \mathbf{I}_r \otimes \mathbf{T}_r + \mathcal{E},$$

where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_R \geq 0$ are the singular values in descending order, and \mathbf{G}_r , \mathbf{I}_r , and \mathbf{T}_r are norm-1 singular vectors in \mathbb{R}^{n_G} , \mathbb{R}^{n_I} , and \mathbb{R}^{n_T} , respectively, and $\mathcal{E} = \llbracket E_{ijk} \rrbracket$ is noise tensor with each entry $E_{ijk} \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. We refer to the vector \mathbf{G}_r (respectively, \mathbf{T}_r and \mathbf{I}_r) as the r -th eigen-gene (respectively, eigen-tissue and eigen-individual). We take a successive rank-1 approximation to \mathcal{Y} (or its residual) by solving the following optimization:

$$\begin{aligned} & \underset{\lambda_r, \mathbf{G}_r, \mathbf{I}_r, \mathbf{T}_r}{\text{minimize}} \left\| \mathcal{Y} - \lambda_r \mathbf{G}_r \otimes \mathbf{I}_r \otimes \mathbf{T}_r \right\|_F, \\ & \text{subject to } \|\mathbf{G}_r\|_2 = \|\mathbf{I}_r\|_2 = \|\mathbf{T}_r\|_2 = 1. \end{aligned}$$

where $\|\cdot\|_F$ is defined entrywise as $\|\mathcal{Y}\|_F = \sqrt{\sum_{i=1}^{n_G} \sum_{j=1}^{n_I} \sum_{k=1}^{n_T} Y_{ijk}^2}$. In each iteration, we impose the entrywise non-negativity on the tissue loading vectors $\mathbf{T}_r \geq 0$ by thresholding negative values of \mathbf{T}_r to 0. Then we take the residual tensor $\mathcal{Y} \leftarrow \mathcal{Y} - \lambda_r \mathbf{G}_r \otimes \mathbf{I}_r \otimes \mathbf{T}_r$ as the new input and repeat the algorithm to find the next component. The full algorithm is provided in Supplemental Material.

The non-negativity constraint on each \mathbf{T}_r eases interpretation of the interaction at the tissue level; a sparse vector \mathbf{T}_r implies that the module r is “active” in only a few tissues, whereas a dense vector \mathbf{T}_r implies that the module r is common to several tissues. Without the nonnegativity constraint, it is possible (in fact likely) that each \mathbf{T}_r actually contains two expression modules, one which corresponds to the positively-loaded tissues and one to the negatively-loaded tissues. Consequently, gene and individual loadings become less informative because it is difficult to determine with which module they associate.

Characterizing the expression modules

We applied the *MultiCluster* algorithm to the GTEx tensor with R set to 10. Note that *MultiCluster* finds the tensor components via successive rank-1 approximations, so one can always apply the algorithm to the residual tensor if more components are wanted. Here we chose to focus on the top $R = 10$ components based on biological interpretability. For each $r = 1, \dots, R$, we used the procedure discussed below to characterize the biological significance of the loading vectors, $\mathbf{G}_r, \mathbf{I}_r, \mathbf{T}_r$. For ease of presentation, in what follows we drop the subscript r and simply write \mathbf{G} , \mathbf{I} , and \mathbf{T} .

GO enrichment based on gene loadings. Let $\mathcal{G} = \{1, \dots, n_G\}$ denote all genes in the analysis, and $\mathbf{G} = (G_1, \dots, G_{n_G})^T$ be the eigen-gene of interest. We performed GO enrichment analyses in both the top gene cluster $\mathcal{G}_{\text{top}} = \{i \in \mathcal{G} : G_i \geq c_{\text{top}}\}$, and bottom gene cluster $\mathcal{G}_{\text{bottom}} = \{i \in \mathcal{G} : G_i \leq c_{\text{bottom}}\}$. Here c_{top} and c_{bottom} are cut-off values which determine the cluster sizes. For each declared gene cluster, we assessed the overrepresentation of GO terms using the hypergeometric test as implemented in the *enrichGO* function of ClusterProfiler (Yu et al. 2012). The GO annotations were loaded from *org.Hs.eg.db* (Carlson 2017) and only biological processes (BP) terms with at least 10 genes were considered. To account for multiple testing, we report enrichment p -values after applying the Benjamini-Hochberg (B-H) correction.

The cut-off values c_{top} and c_{bottom} control the significance level of the genes in the declared gene clusters. We employ a permutation-based procedure to determine the cut-off values (and thus gene cluster sizes) with a significance level $\alpha = 0.005$. Specifically, we generated ten null tensors by randomly and independently permuting genes for every individual-tissue pair (j, k) , i.e.,

$$\mathcal{Y}^{\text{null}}(\mathcal{G}, \text{individual } j, \text{tissue } k) \stackrel{\text{def}}{=} \mathcal{Y}(\mathcal{P}\mathcal{G}, \text{individual } j, \text{tissue } k),$$

where $\mathcal{P}\mathcal{G}$ represents a random permutation of the set \mathcal{G} at the pair (j, k) . We then decomposed each of the null tensors $\mathcal{Y}^{\text{null}}$ and used their eigen-genes \mathbf{G}^{null} to represent the null distribution of \mathbf{G} -values. The cut-off value c_{top} (respectively, c_{bottom}) was determined using the top 0.5%-quantile (respectively, bottom 0.5%-quantile) of the empirical distribution of $\{\mathbf{G}_i^{\text{null}}\}$.

Effects of biological attributes on individual loadings. To identify the sources of variation in the individual loadings, we considered the following linear model for an eigen-individual $\mathbf{I} = (I_1, \dots, I_{n_I})^T$,

$$I_j = \beta_1 + \beta_2 1_{\text{female}}^j + \beta_3 1_{\text{African-American}}^j + \beta_4 \text{Age}^j + \varepsilon_j, \text{ where } \varepsilon_j \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2) \text{ for all } j = 1, \dots, n_I.$$

Upon fitting the model, we calculated the proportion of variance explained by each covariate (gender, race, or age) using ANOVA.

Tensor projection for detecting DE genes

Here we describe our tensor projection procedure for detecting covariate-associated genes. Let $\mathcal{Y} \in \mathbb{R}^{n_G \times n_I \times n_T}$ denote the expression tensor and $\{\mathbf{T}_r \in \mathbb{R}^{n_T}\}$ be the set of eigen-tissues from the decomposition. As we have seen, \mathbf{T}_r captures the degree of similarity across tissues in the expression module r . Let $\mathcal{Y}(\cdot, \cdot, \mathbf{T}_r) \in \mathbb{R}^{n_G \times n_I}$ denote the projection of \mathcal{Y} through the eigen-tissue $\mathbf{T}_r = (T_{r,1}, \dots, T_{r,n_T})^T$; that is, $\mathcal{Y}(\cdot, \cdot, \mathbf{T}_r) = \sum_{k=1}^{n_T} T_{r,k} \mathcal{Y}(\cdot, \cdot, k)$. For each gene to be tested, we proposed the following linear model,

$$\mathcal{Y}(\text{test gene}, \cdot, \mathbf{T}_r) = \beta_1 \mathbf{1}_{n_I} + \beta_2 \mathbf{1}_{\text{female}} + \beta_3 \mathbf{1}_{\text{African-American}} + \beta_4 \mathbf{Age} + \varepsilon, \quad (2)$$

where $\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n_I \times n_I})$, $\mathbf{1}_{n_I}$ denotes a vector of length n_I with every element equal to 1, and $\mathbf{I}_{n_I \times n_I}$ denotes an n_I -by- n_I identity matrix. The age-effect was assessed by testing $\mathcal{H}_0: \beta_4 = 0$ against $\mathcal{H}_a: \beta_4 \neq 0$. Similar hypothesis testing can be performed for gender and race effects.

In the single-tissue analysis, we considered each gene-tissue pair one at a time and performed the following regression analysis,

$$\mathcal{Y}(\text{test gene}, \cdot, \text{test tissue}) = \beta_1 \mathbf{1}_{n_I} + \beta_2 \mathbf{1}_{\text{female}} + \beta_3 \mathbf{1}_{\text{African-American}} + \beta_4 \mathbf{Age} + \varepsilon, \quad (3)$$

where $\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{n_I \times n_I})$. Comparing model (3) with model (2), we see that the tensor projection makes use of the similarity across tissues and thus may boost power for detecting covariate-associated genes.

Simulation models for assessing three-way clustering

We simulated a set of expression tensors $\mathcal{Y} = [\![Y_{ijk}]\!] \in \mathbb{R}^{n_G \times n_I \times n_T}$, where $n_G = 500$ (genes), $n_I = 500$ (individuals), and $n_T = 10$ (tissues). In each tensor, we created $K_G = 5$ gene clusters, $K_I = 4$ individual clusters, and $K_T = 3$ tissue clusters by randomly assigning each gene (respectively, individual and tissue) to a gene cluster (respectively, individual cluster and tissue cluster) with uniform probability. We generated mean expressions according to $\mathbb{E}(Y_{ijk}) = \mu_{lmn}$, where l, m, n denote the corresponding gene/individual/tissue cluster that Y_{ijk} belongs to. We employed three models to simulate the 3-way block means $\{\mu_{lmn}\}$.

1. Additive-mean model: $\mu_{lmn} = \mu_l^g + \mu_m^i + \mu_n^t$, where μ_l^g , μ_m^i , and μ_n^t represent the marginal mean for gene cluster l , tissue cluster m and individual cluster n , respectively. The marginal means, μ_l^g , μ_m^i , and μ_n^t , are i.i.d. drawn from $N(1, 1)$.
2. Multiplicative-mean model: $\mu_{lmn} = \mu_l^g \mu_m^i \mu_n^t$, where the notation remains the same.

3. Combinatorial-mean model: $\mu_{lmn} \stackrel{\text{i.i.d.}}{\sim} N(1, 1)$; that is, each three-way block has its own mean, independently of each other.

Let $\mathcal{Y}_{\text{true}}$ denote the noiseless tensor with 3-way block means generated from each of the above schemes. The observed expression data were then simulated from $\mathcal{Y} = \mathcal{Y}_{\text{true}} + \mathcal{E}$, where $\mathcal{E} \in \mathbb{R}^{n_G \times n_I \times n_T}$ is a random Gaussian tensor with each entry i.i.d. drawn from $N(0, \sigma^2)$.

We simulated 50 expression tensors under each model, and then assessed the accuracy of recovery using the relative error, defined by

$$\text{RelErr}(\hat{\mathcal{Y}}_{\text{est}}, \mathcal{Y}_{\text{true}}) = \min_{\substack{\pi \in \{\text{all permutation of 3-way blocks}\}, \\ \text{rank}(\hat{\mathcal{Y}}_{\text{est}}) \leq 10}} \frac{\|\hat{\mathcal{Y}}_{\text{est}} - \mathcal{Y}_{\text{true}}^{\pi}\|_F}{\|\mathcal{Y}_{\text{true}}\|_F},$$

where $\hat{\mathcal{Y}}_{\text{est}}$ denotes the rank- R approximation ($R = 1, \dots, 10$) obtained from tensor decomposition.

Simulation models for detecting DE genes

In order to simulate tensors with both block structure and age signals, we modified the earlier additive model into

$$Y_{ijk} = \mu_l^g + \mu_{[i:k]} \text{Age}(j) + \mu_n^t + \varepsilon_{ijk}, \quad \text{where } \varepsilon_{ijk} \stackrel{\text{i.i.d.}}{\sim} N(0, 2^2), \quad (4)$$

where Y_{ijk} denotes the expression level of gene i , individual j and tissue k ; μ_l^g and μ_n^t denote the same parameters as before (the marginal means for gene cluster l and for tissue cluster n); and

$$\mu_{[i:k]} \sim \begin{cases} 0, & \text{if gene } i \text{ is not an age-related gene in the tissue cluster } n = n(k), \\ \text{Unif}[0, 0.05], & \text{if gene } i \text{ is an age-upregulated gene in the tissue cluster } n = n(k), \\ \text{Unif}[-0.05, 0], & \text{if gene } i \text{ is an age-downregulated gene in the tissue cluster } n = n(k). \end{cases} \quad (5)$$

We simulated 50 tensors $\mathcal{Y} \in \mathbb{R}^{n_G \times n_I \times n_T}$ with $n_G = 500$ (genes), $n_I = 50$ (individuals), and $n_T = 10$ (tissues). In each tensor, we planted five gene clusters, three tissue clusters, and further assigned 100 genes to be age-related (50 up-regulated and 50 down-regulated) in at least one of the three tissue clusters. The ages of individuals were generated using i.i.d. $\text{Unif}[40, 70]$, and the effect sizes were simulated using (5). The final expression data was generated based on model (4).

Runtime

The run times were evaluated using a single processor on an Macbook (Mac OS High Sierra 10.13) with Intel Core i5 2.9GHz CPU and 8GB RAM. We chose the multi-threading option in *SDA* with 4 threads. Both *MultiCluster* and *HOSVD* were implemented in Matlab R2016a 64-bits.

Data and software availability

Supplementary data containing inferred gene modules and the software *MultiCluster* are available.

Acknowledgments

We thank Junhyong Kim for helpful discussions and comments on our work. This research is supported in part by a Math+X Research Grant from the Simons Foundation, a Packard Fellowship for Science and Engineering, and a National Institutes of Health grant R01-GM094402. YSS is a Chan Zuckerberg Biohub investigator.

References

- Almon, R. R., Chen, J., Snyder, G., DuBois, D. C., Jusko, W. J., and Hoffman, E. P., 2003. In vivo multi-tissue corticosteroid microarray time series available online at public expression profile resource (PEPR). *Pharmacogenomics*, **4**(6):791–799.
- Alter, O., Brown, P. O., and Botstein, D., 2000. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, **97**(18):10101–10106.
- Bahcall, O. G., 2015. Human genetics: GTEx pilot quantifies eQTL variation across tissues and individuals. *Nature Reviews Genetics*, **16**(7):375.
- Barh, D., 2014. *Omics Approaches in Breast Cancer: Towards Next-Generation Diagnosis, Prognosis and Therapy*. Springer.
- Battle, A., Brown, C. D., Engelhardt, B. E., Montgomery, S. B., Consortium, G., et al., 2017. Genetic effects on gene expression across human tissues. *Nature*, **550**(7675):204–213.
- Berger, B., Peng, J., and Singh, M., 2013. Computational solutions for omics data. *Nature Reviews Genetics*, **14**(5):333.
- Candès, E. J. and Recht, B., 2009. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, **9**(6):717.
- Carlson, M., 2017. org.hs.eg.db: Genome wide annotation for human. *R package version 3.4.1.*, .
- Carrasquillo, M. M., Zou, F., Pankratz, V. S., Wilcox, S. L., Ma, L., Walker, L. P., Younkin, S. G., Younkin, C. S., Younkin, L. H., Bisceglia, G. D., et al., 2009. Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer’s disease. *Nature Genetics*, **41**(2):192–198.
- Chettier, R., Ward, K., and Albertsen, H. M., 2014. Endometriosis is associated with rare copy number variants. *PLoS ONE*, **9**(8):e103968.
- Ciryam, P., Kundra, R., Freer, R., Morimoto, R. I., Dobson, C. M., and Vendruscolo, M., 2016. A transcriptional signature of Alzheimer’s disease is associated with a metastable subproteome at risk for aggregation. *Proceedings of the National Academy of Sciences*, **113**(17):4753–4758.
- Collins, M. H., 2014. Histopathologic features of eosinophilic esophagitis and eosinophilic gastrointestinal diseases. *Gastroenterology Clinics of North America*, **43**(2):257–268.
- Consortium, E. P. et al., 2004. The ENCODE (ENCyclopedia of DNA elements) project. *Science*, **306**(5696):636–640.
- Consortium, G. et al., 2015. The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, **348**(6235):648–660.

- da Rocha, S. T. and Heard, E., 2017. Novel players in X inactivation: insights into xist-mediated gene silencing and chromosome conformation. *Nature Structural & Molecular Biology*, **24**(3):197–204.
- Dey, K. K., Hsiao, C. J., and Stephens, M., 2017. Visualizing the structure of RNA-seq expression data using grade of membership models. *PLoS Genetics*, **13**(3):e1006599.
- Dönertaş, H. M., İzgi, H., Kamacıoğlu, A., He, Z., Khaitovich, P., and Somel, M., 2017. Gene expression reversal toward pre-adult levels in the aging human brain and age-related loss of cellular identity. *Scientific Reports*, **7**.
- Droppelmann, C. A., Wang, J., Campos-Melo, D., Keller, B., Volkening, K., Hegele, R. A., and Strong, M. J., 2013. Detection of a novel frameshift mutation and regions with homozygosis within ARHGEF28 gene in familial amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, **14**(5-6):444–451.
- Fishilevich, S., Zimmerman, S., Kohn, A., Iny Stein, T., Olender, T., Kolker, E., Safran, M., and Lancet, D., 2016. Genic insights from integrated human proteomics in GeneCards. *Database*, .
- Gao, C., McDowell, I. C., Zhao, S., Brown, C. D., and Engelhardt, B. E., 2016. Context specific and differential gene co-expression networks via bayesian biclustering. *PLoS Computational Biology*, **12**(7):e1004791.
- Hawrylycz, M. J., Lein, S., Guillozet-Bongaarts, A. L., Shen, E. H., Ng, L., Miller, J. A., Van De Lagemaat, L. N., Smith, K. A., Ebbert, A., Riley, Z. L., *et al.*, 2012. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, **489**(7416):391.
- 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.
- Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A. M., Pletikos, M., Meyer, K. A., Sedmak, G., *et al.*, 2011. Spatiotemporal transcriptome of the human brain. *Nature*, **478**(7370):483.
- Kelly, R. D., Mahmud, A., McKenzie, M., Trounce, I. A., and St John, J. C., 2012. Mitochondrial DNA copy number is regulated in a tissue specific manner by DNA methylation of the nuclear-encoded DNA polymerase gamma A. *Nucleic Acids Research*, **40**(20):10124–10138.
- Kherraf, Z.-E., Christou-Kent, M., Karaouzene, T., Amiri-Yekta, A., Martinez, G., Vargas, A. S., Lambert, E., Borel, C., Dorphin, B., Aknin-Seifer, I., *et al.*, 2017. SPINK2 deficiency causes infertility by inducing sperm defects in heterozygotes and azoospermia in homozygotes. *EMBO Molecular Medicine*, :e201607461.
- Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R., *et al.*, 2017. SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods*, **14**(4):483–486.
- Kluger, Y., Basri, R., Chang, J. T., and Gerstein, M., 2003. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Research*, **13**(5):703–716.
- Kuleshov, V., Chaganty, A., and Liang, P., 2015. Tensor factorization via matrix factorization. In *Artificial Intelligence and Statistics*, pages 507–516.

- Lacroix, M., 2006. Significance, detection and markers of disseminated breast cancer cells. *Endocrine-Related Cancer*, **13**(4):1033–1067.
- Lam, A. D., Deck, G., Goldman, A., Eskandar, E. N., Noebels, J., and Cole, A. J., 2017. Silent hippocampal seizures and spikes identified by foramen ovale electrodes in alzheimer’s disease. *Nature Medicine*, **23**(6):678–680.
- Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., *et al.*, 2013. The genotype-tissue expression (GTEx) project. *Nature Genetics*, **45**(6):580–585.
- Love, M., Huber, W., and Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**(550).
- Maaten, L. v. d. and Hinton, G., 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, **9**(Nov):2579–2605.
- McCall, M., Illei, P., and Halushka, M., 2016. Complex sources of variation in tissue expression data: Analysis of the GTEx lung transcriptome. *Am J Hum Genet.*, **99**(3):624–635.
- 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.
- Merkin, R. D., Vanner, E. A., Romeiser, J. L., Shroyer, A. L. W., Escobar-Hoyos, L. F., Li, J., Powers, R. S., Burke, S., and Shroyer, K. R., 2017. Keratin 17 is overexpressed and predicts poor survival in estrogen receptor–negative/human epidermal growth factor receptor-2–negative breast cancer. *Human Pathology*, **62**:23–32.
- Montgomery, S. B. and Dermitzakis, E. T., 2011. From expression QTLs to personalized transcriptomics. *Nature Reviews Genetics*, **12**(4):277.
- Mori, F., Tanji, K., Miki, Y., Toyoshima, Y., Yoshida, M., Kakita, A., Takahashi, H., Utsumi, J., Sasaki, H., and Wakabayashi, K., *et al.*, 2016. G protein-coupled receptor 26 immunoreactivity in intranuclear inclusions associated with polyglutamine and intranuclear inclusion body diseases. *Neuropathology*, **36**(1):50–55.
- Naderi, A. and Vanneste, M., 2014. Prolactin-induced protein is required for cell cycle progression in breast cancer. *Neoplasia*, **16**(4):329–342.
- Nica, A. C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K., *et al.*, 2011. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genetics*, **7**(2):e1002003.
- 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.
- Pierson, E., Koller, D., Battle, A., Mostafavi, S., Consortium, G., *et al.*, 2015. Sharing and specificity of co-expression networks across 35 human tissues. *PLoS Computational Biology*, **11**(5):e1004220.

- Priddle, T. H. and Crow, T. J., 2013. The protocadherin 11X/Y (PCDH11X/Y) gene pair as determinant of cerebral asymmetry in modern homo sapiens. *Annals of the New York Academy of Sciences*, **1288**(1):36–47.
- Sidiropoulos, N. D., De Lathauwer, L., Fu, X., Huang, K., Papalexakis, E. E., and Faloutsos, C., 2017. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, **65**(13):3551–3582.
- Sokal, R. R., 1958. A statistical method for evaluating systematic relationship. *University of Kansas science bulletin*, **28**:1409–1438.
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B., 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**(6):520–525.
- U.S. Cancer Statistics Working Group, 2017. United states cancer statistics: 1999-2014 incidence and mortality web-based report. *Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute*, . Available at: www.cdc.gov/uscs.
- Valente, V., Teixeira, S. A., Neder, L., Okamoto, O. K., Oba-Shinjo, S. M., Marie, S. K., Scrideli, C. A., Paço-Larson, M. L., and Carlotti, C. G., 2009. Selection of suitable housekeeping genes for expression analysis in glioblastoma using quantitative RT-PCR. *BMC Molecular Biology*, **10**(1):17.
- Veerappa, A. M., Saldanha, M., Padakannaya, P., and Ramachandra, N. B., 2013. Genome-wide copy number scan identifies disruption of PCDH11X in developmental dyslexia. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, **162**(8):889–897.
- Walker, J. R., Su, A. I., Self, D. W., Hogenesch, J. B., Lapp, H., Maier, R., Hoyer, D., and Bilbe, G., 2004. Applications of a rat multiple tissue gene expression data set. *Genome Research*, **14**(4):742–749.
- 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.
- Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., *et al.*, 2013. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, **45**(10):1113–1120.
- Wiwie, C., Baumbach, J., and Röttger, R., 2015. Comparing the performance of biomedical clustering methods. *Nature Methods*, **12**(11):1033.
- Xiao, X., Moreno-Moral, A., Rotival, M., Bottolo, L., and Petretto, E., 2014. Multi-tissue analysis of co-expression networks by higher-order generalized singular value decomposition identifies functionally coherent transcriptional modules. *PLoS Genetics*, **10**(1):e1004006.
- Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y., 2012. clusterprofiler: an R package for comparing biological themes among gene clusters. *OMICS*, **16**(5):284–287.

Figures and Tables

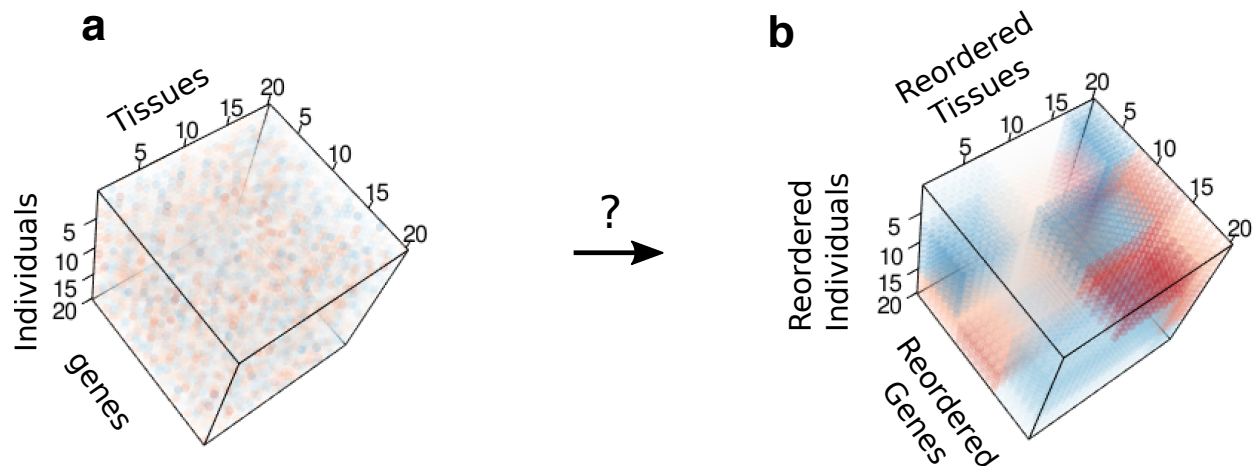


Figure 1: **Three-way clustering problem.** (a) Input tensor of gene expression. (b) Shuffled, de-noised output tensor containing local blocks. Both (a) and (b) are color images of a data tensor $\mathcal{Y} = \llbracket Y_{ijk} \rrbracket$, with each entry colored according to the value of Y_{ijk} .

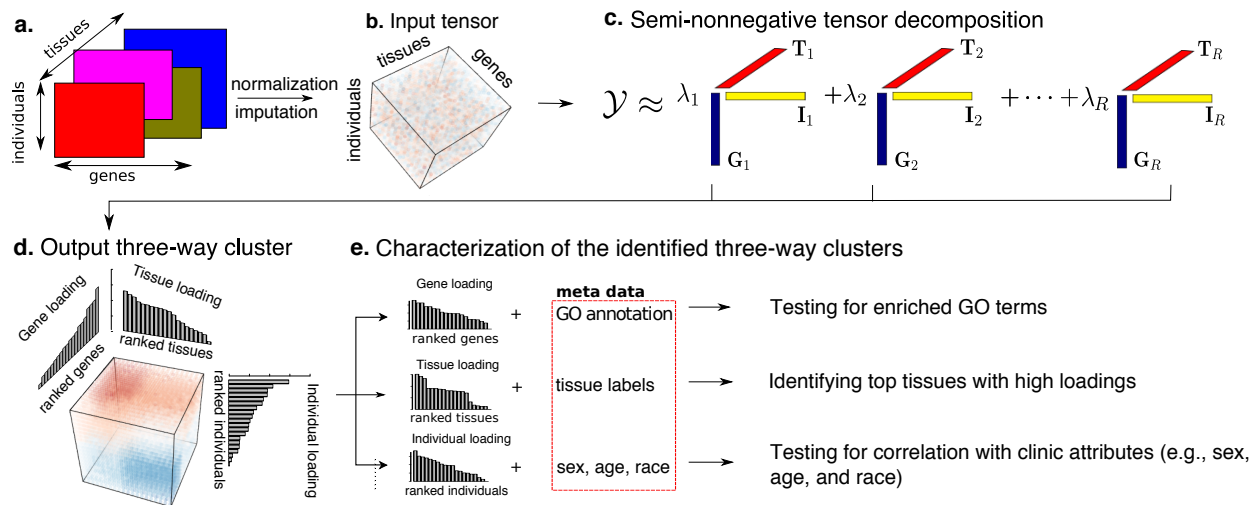


Figure 2: Schematic diagram of *MultiCluster* algorithm. (a) Multi-tissue gene expression data. (b) Input expression tensor after normalization and imputation. (c) The algorithm decomposes the expression tensor into a set of rank-1 tensors, $\mathbf{G}_r \otimes \mathbf{I}_r \otimes \mathbf{T}_r$, where \mathbf{G}_r , \mathbf{I}_r , and \mathbf{T}_r are, respectively, gene, individual, and tissue singular vectors. (d) Each three-way cluster is represented by the three sorted singular vectors. (e) We utilize metadata, such as GO annotation, tissue labels, and individual-level covariates, to identify the sources of variation in each loading vector.

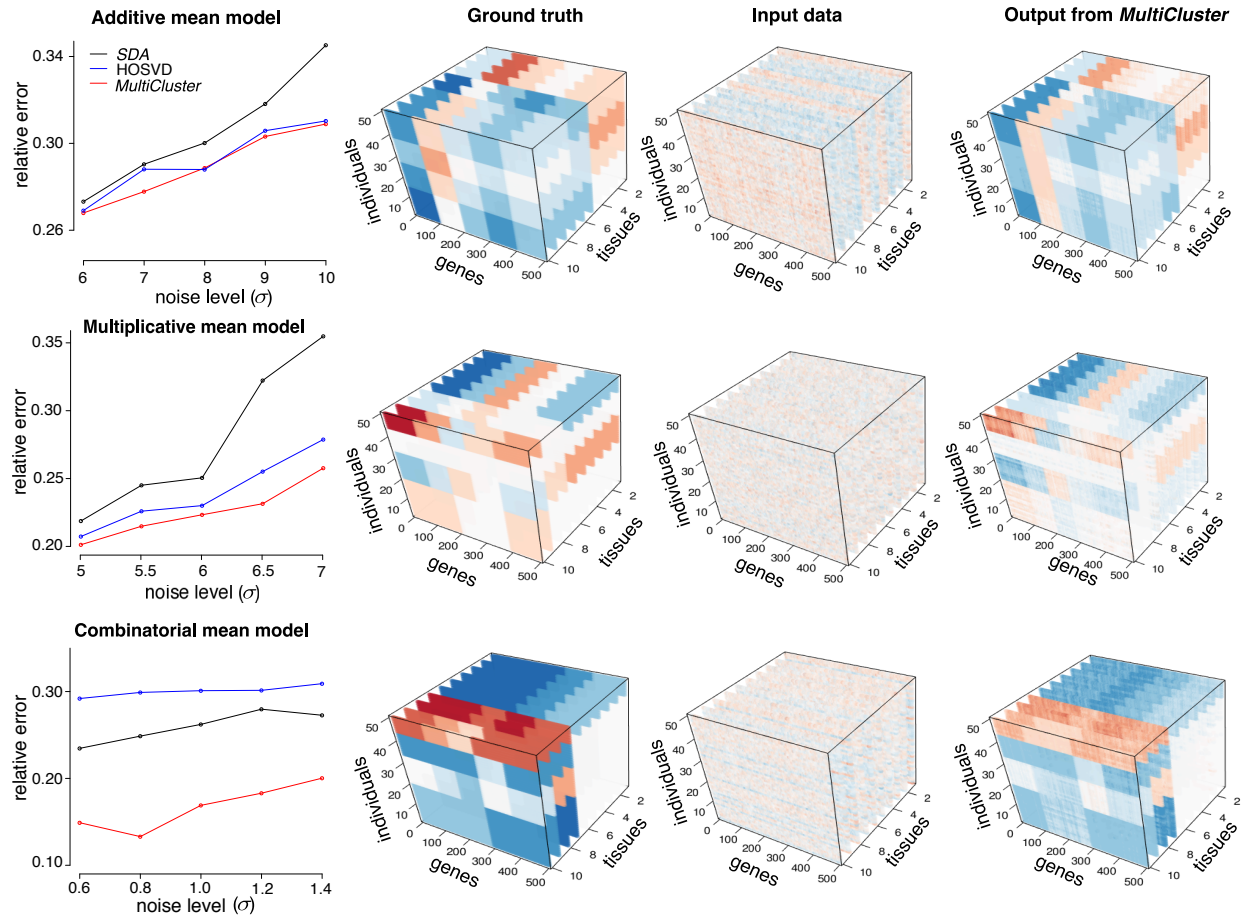


Figure 3: **Three-way clustering performance in simulations.** The first column shows the recovery accuracy of different tensor-based methods. The columns 2–4 are color images of example tensors in simulations under different block-mean models. The example tensor $\mathcal{Y} = [Y_{ijk}]$ consists of 500 (genes) \times 50 (individuals) \times 10 (tissues), with each entry colored according to the value of Y_{ijk} .

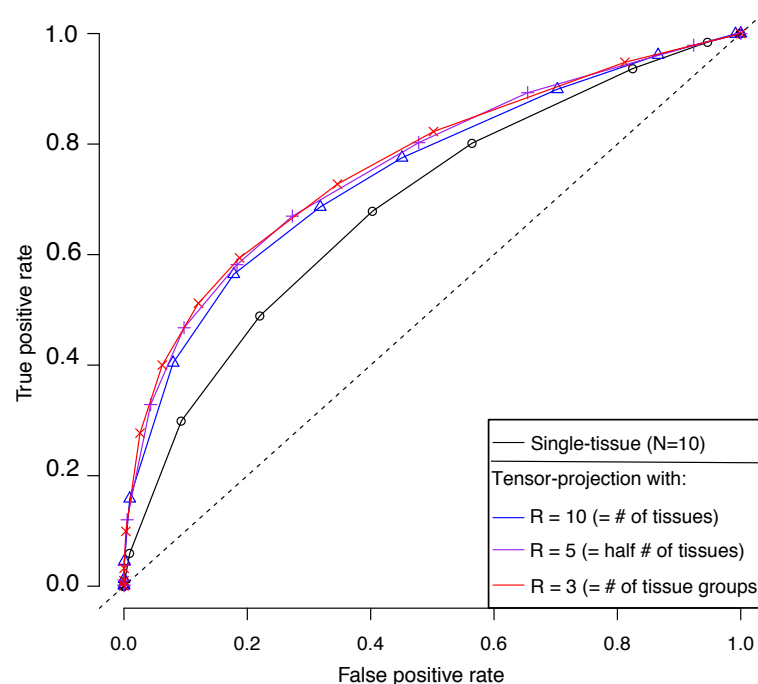


Figure 4: **ROC curves for tensor projection and single-tissue analysis.** We performed tensor projection and single-tissue analyses to detect age-related genes in simulated expression tensors. For tensor projection, we decomposed each tensor into R components with $R = 3, 5$ and 10 , and declared an age-related gene if its p -value was less than the nominal significance level in at least one of the R eigen-tissues. For single-tissue analyses, we tested age-related genes in each tissue separately and declared an age-related gene if its p -value was less than the significance level in at least of the 10 tissues. The ROC curves were obtained under various nominal significance levels using 50 simulations.

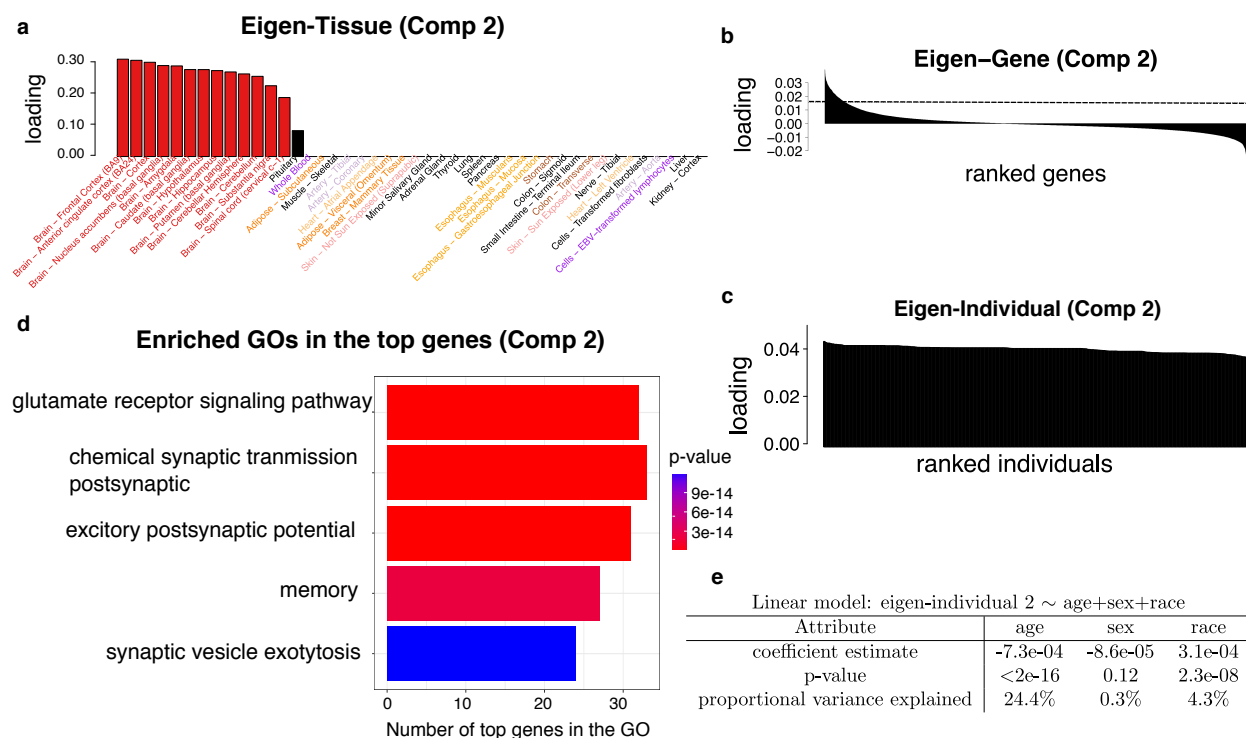


Figure 5: Expression module II – brain tissues. Panels (a)–(c) represent the triplets of sorted singular vectors, whereas panels (d)–(e) represent the biological contexts of the identified three-way cluster. (a) Barplot of the sorted tissue loading vector, where each tissue is colored based on their functional similarity. (b) Barplot of the sorted gene loading vector, where the dotted line represents the threshold for the top genes. (c) Barplot of the sorted individual loading vector. (d) Enriched GO annotations among the top 899 genes identified from the gene loading vector. The enrichment p -values are obtained from hypergeometric tests with B-H correction. The GO size on the axis represents the number of top genes that belong to the GO annotation. (e) Linear regression analysis of individual loadings against individual-level covariates (age, sex and race). The proportional variance explained is computed from ANOVA analysis.

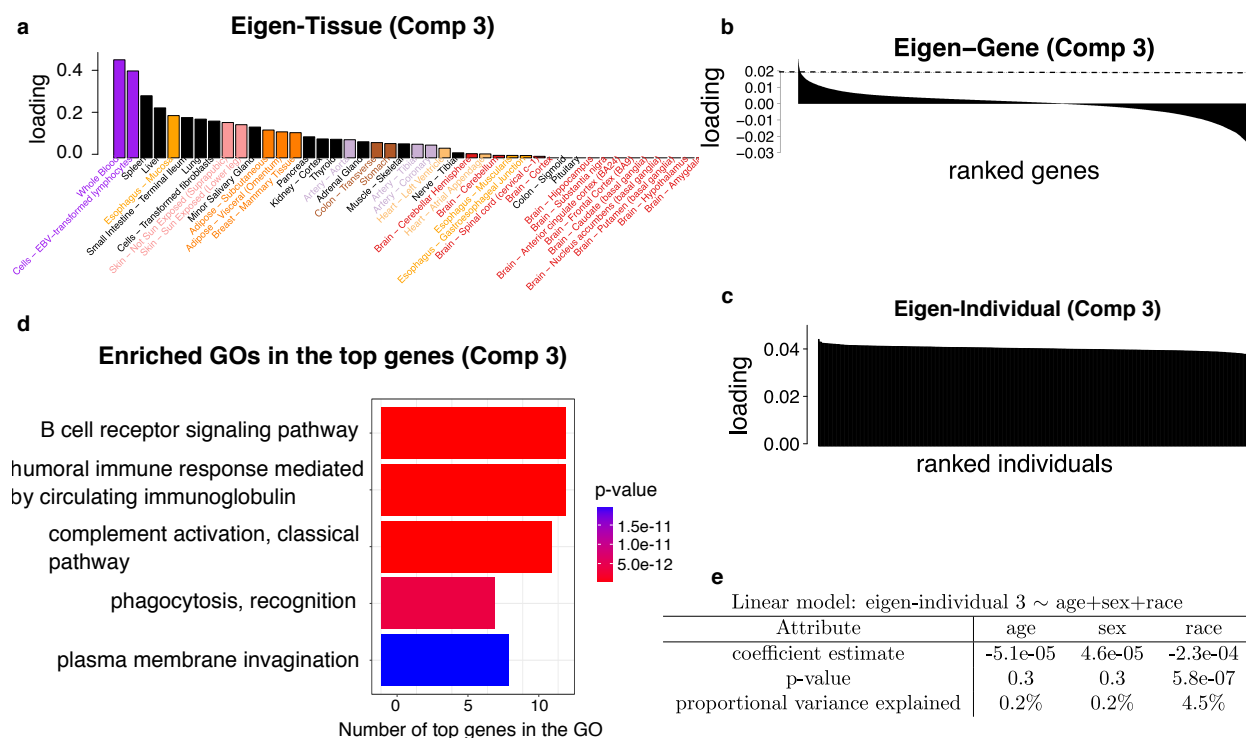


Figure 6: **Expression module III – tissues involved in immune response.** Panels (a)–(c) represent the triplets of sorted singular vectors, whereas panels (d)–(e) represent the biological contexts of the identified three-way cluster. (a) Barplot of the sorted tissue loading vector, where each tissue is colored based on their functional similarity. (b) Barplot of the sorted gene loading vector, where the dotted line represents the threshold for the top genes. (c) Barplot of the sorted individual loading vector. (d) Enriched GO annotations among the top 89 genes identified from the gene loading vector. The enrichment p -values are obtained from hypergeometric tests with B-H correction. The GO size on the axis represents the number of top genes that belong to the GO annotation. (e) Linear regression analysis of individual loadings against individual-level covariates (age, sex and race). The proportional variance explained is computed from ANOVA analysis.

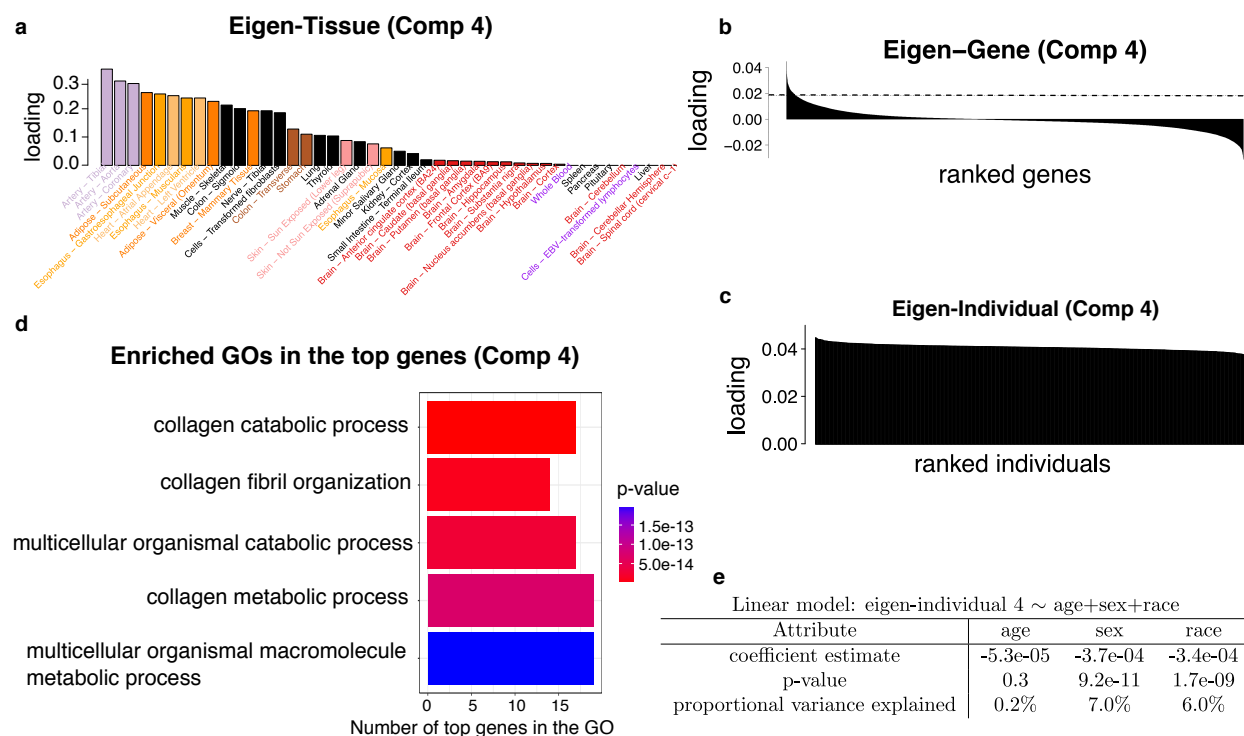


Figure 7: Expression module VI – tissues with structural similarities. Panels (a)–(c) represent the triplets of sorted singular vectors, whereas panels (d)–(e) represent the biological contexts of the identified three-way cluster. (a) Barplot of the sorted tissue loading vector, where each tissue is colored based on their functional similarity. (b) Barplot of the sorted gene loading vector, where the dotted line represents the threshold for top genes. (c) Barplot of the sorted individual loading vector. (d) Enriched GO annotations among the top 352 genes identified from the gene loading vector. The enrichment *p*-values are obtained from hypergeometric tests with B-H correction. The GO size on the axis represents the number of top genes that belong to the GO annotation. (e) Linear regression analysis of individual loadings against individual-level covariates (age, sex and race). The proportional variance is computed from ANOVA analysis.

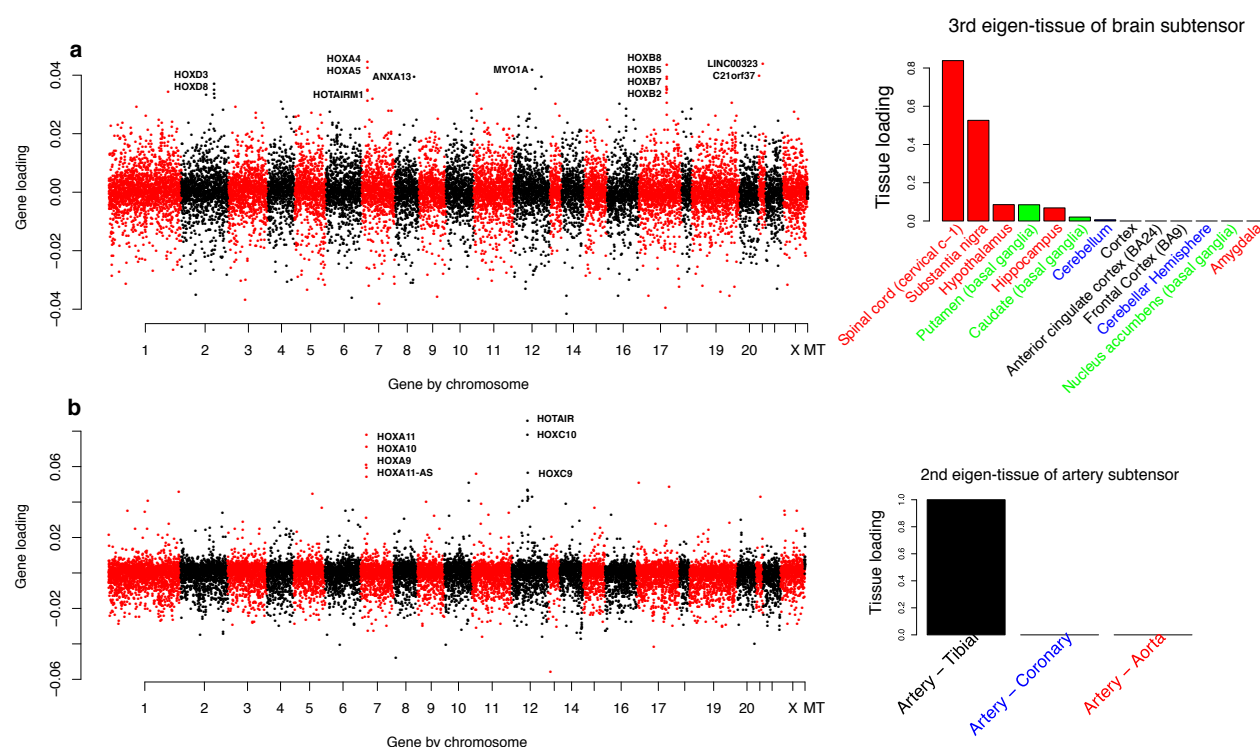


Figure 8: ***HOX* gene expressions and associated tissue loadings in different subtensors.** (a) Over-expression of *HOX* genes in spinal cord (cervical C-1) compared to other brain tissues. This expression pattern was identified from the 3rd tensor component of the brain subtensor. (b) Over-expression of *HOX* genes in tibial tissues compared to other artery tissues. This expression pattern was identified from the 2nd tensor component of the artery subtensor. In each panel, the left figure plots gene loadings against gene positions on the chromosomes. Genes with extreme loadings (e.g., *HOXD* genes, *HOXB* genes, *HOXA* genes, etc) are labeled on the plot. The right figure shows the barplot of tissue loadings in the corresponding eigen-tissue.

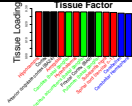
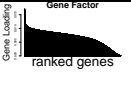

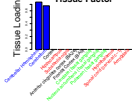
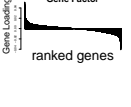
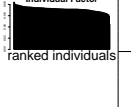
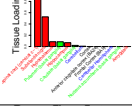
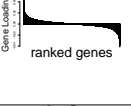
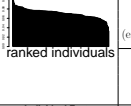
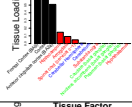
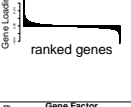

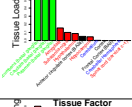
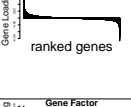

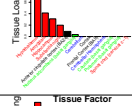

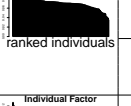
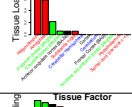
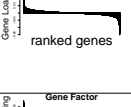

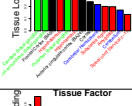


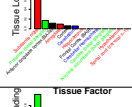
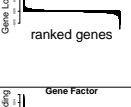
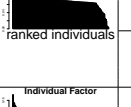
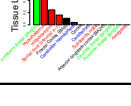
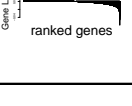
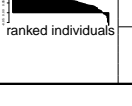
Module	Component			Eigen-gene	Eigen-tissue	Eigen-individual		
					leading tissue	% variance explained		
						age	sex	race
1				Top 5 Genes: MT-RNR2, MT-CO1, MT-ND4, MT-CO2, MT-CO3 Enriched GOs: neuronal synaptic plasticity, memory	All	1.5	7.8	2.2
2				Over-expressed cluster: 122 genes: C16orf11, ZP2, SLC22A31, CDH15, BARHL2 (dorsal spinal cord development, spinal cord development, excretion) Under-expressed cluster: 168 genes: FOXG1, SST, DDN, GDA, NRGN (forebrain generation of neurons, forebrain neuron differentiation)	Cerebellum	0.0	8.0	0.2
3				Over-expressed cluster: 216 genes: HOXA4, LINC00323, HOXB8, HOXA5, MYO1A (embryonic skeletal system morphogenesis, axon ensheathment in central nervous system) Under-expressed cluster: 288 genes: FOXG1, RPRML, DLX6-AS1, FEZF2, GPR6 (regulation of calcium ion-dependent exocytosis, synaptic vesicle exocytosis)	Spinal cord	10.5	0.7	5.2
4				Over-expressed cluster: 375 genes: LINC01007, LINC00507, TBR1, TMEM155, EMX1 (fear response, behavior defense response) Under-expressed cluster: 103 genes: OTX2, IRX1, SLITRK6, IRX5, CRNDE (proximo/distal pattern formation, forebrain regionalization)	Cortex	16.7	0.6	1.4
5				Over-expressed cluster: 331 genes: SYNDIGL1, CTA-920C8.8, SIX3, GPR88, RP11-481J2.2 (forebrain generation of neurons, associative learning) Under-expressed cluster: 204 genes: NEUROD1, SLC17A7, NEUROD2, SLC6A7, SLC17A6 (dorsal spinal cord development, forelimb morphogenesis)	Basal ganglia	1.3	0.8	1.7
6				Over-expressed cluster: 444 genes: NTS, SLC17A6, AVP, PMCH, RP11-566J3.4 (neuropeptide signaling pathway, diencéphalon development) Under-expressed cluster: 151 genes: TESPA1, RP11-481J2.2, O6orf141, RP11-766F14.2, LRRC38 (embryonic skeletal system morphogenesis, visual learning)	Hypothalamus	14.8	2.2	0.1
7				Over-expressed cluster: 157 genes: P2RX2, SLC17A7, EMX1, NEUROD6, NEUROD2 (central nervous system myelination, axon ensheathment in central nervous system) Under-expressed cluster: 325 genes: ISL1, SV2C, HCHT, RP11-466F24.7, C22orf42 (neuropeptide signaling pathway, pituitary gland development)	Hippocampus	22.0	0.1	3.6
8				Over-expressed cluster: 120 genes: LINC00996, SV2C, SYT2, SLC32A1, RAB3B (response to ammonium ion, synaptic transmission - cholinergic) Under-expressed cluster: 361 genes: IL1RL1, SERPINA3, SOCS3, S100A12, S100A8 (neutrophil migration, neutrophil chemotaxis)	All brain regions	22.1	0.6	1.9
9				Over-expressed cluster: 207 genes: SLC6A3, LINC00261, TH, FOXA2, PITX3 (cellular response to zinc ion, dopaminergic neuron differentiation) Under-expressed cluster: 296 genes: HOXB8, HOXD8, HOXB5, HOXA5, TRH (embryonic skeletal system morphogenesis, ciliun movement)	Substantia nigra	0.4	1.4	3.0
10				Over-expressed cluster: 254 genes: DGKK, CARTPT, PNMA5, LINC00202-2, FAT2 (neuropeptide signaling pathway, amine transport) Under-expressed cluster: 252 genes: NEUROG2, ALDH1A1, TESPA1, RP11-454G4.2, MOXD1 (oligodendrocyte development, oligodendrocyte differentiation)	Nucleus accumbens (basal ganglia)	7.2	2.0	1.4

Table 1: **Top 10 expression modules in the brain subtensor.** The brain subtensor contains two cerebellar tissues (cerebellar hemisphere, cerebellum; coded in blue), three basal ganglia tissues (caudate, nucleus accumbens, and putamen; coded in green), three cortex tissues (cortex, anterior cingulate cortex–BA24, frontal cortex–BA9; coded in black) and other four tissues (spinal cord–cervical c-1, substantia nigra, hypothalamus, hippocampus; coded in red). The top 10 expression modules (ranked by their singular values) were identified from the *MultiCluster* method. For each component, we plotted the barplots for the sorted tissue loadings, gene loadings, and individual loadings, respectively. In each eigen-gene, we listed top/bottom 5 genes as well as the enriched GO annotations in the identified clusters. In each eigen-tissue, we reported the leading tissue with the largest tissue loading. In each eigen-individual, we reported the proportional variance of the individual loadings that was explained by age, sex, or race, respectively. Number in bold indicates $p < 10^{-3}$.