

# Supplemental Material

## Three-way clustering of multi-tissue multi-individual gene expression data using semi-nonnegative tensor decomposition

Miaoyan Wang<sup>1</sup>, Jonathan Fischer<sup>2</sup>, and Yun S. Song<sup>2,3,4</sup>

<sup>1</sup>*Department of Statistics, University of Wisconsin-Madison, WI 53706, USA*

<sup>2</sup>*Department of Statistics, University of California, Berkeley, CA 94720, USA*

<sup>3</sup>*Computer Science Division, University of California, Berkeley, CA 94720, USA*

<sup>4</sup>*Chan Zuckerberg Biohub, San Francisco, CA 94158, USA*

### 1 Supplemental notes

#### 1.1 Proofs

*Proof of Property 1.* Let  $\mathcal{Y} \in \mathbb{R}^{n_G \times n_I \times n_T}$  be the observed tensor. As described in the main manuscript, we aim to solve the following rank-1 optimization,

$$\underset{\lambda, \mathbf{G}, \mathbf{I}, \mathbf{T}}{\text{minimize}} \|\mathcal{Y} - \lambda \mathbf{G} \otimes \mathbf{I} \otimes \mathbf{T}\|_F^2, \quad (1)$$

$$\text{subject to } \|\mathbf{G}\|_2 = \|\mathbf{I}\|_2 = \|\mathbf{T}\|_2 = 1, \quad \text{and } \mathbf{T} \geq 0.$$

The objective function can be re-written as

$$\|\mathcal{Y} - \lambda \mathbf{G} \otimes \mathbf{I} \otimes \mathbf{T}\|_F^2 = \|\mathcal{Y}\|_F^2 - 2\lambda \mathcal{Y}(\mathbf{G}, \mathbf{I}, \mathbf{T}) + \lambda^2. \quad (2)$$

Note that partial derivative of (2) with respect to  $\lambda$  vanishes at the optimizer  $(\hat{\lambda}, \hat{\mathbf{G}}, \hat{\mathbf{I}}, \hat{\mathbf{T}})$ . Hence we obtain that

$$\hat{\lambda} = \mathcal{Y}(\hat{\mathbf{G}}, \hat{\mathbf{I}}, \hat{\mathbf{T}}). \quad (3)$$

Plugging (3) back to the objective function, we find that the optimization (1) is equivalent to

$$\underset{\mathbf{G}, \mathbf{I}, \mathbf{T}}{\text{maximize}} \mathcal{Y}(\mathbf{G}, \mathbf{I}, \mathbf{T})$$

$$\text{subject to } \|\mathbf{G}\|_2 = \|\mathbf{I}\|_2 = \|\mathbf{T}\|_2 = 1, \quad \text{and } \mathbf{T} \geq 0.$$

The above optimization is separable into each of its factors  $\mathbf{G}$ ,  $\mathbf{I}$ ,  $\mathbf{T}$ , so we can optimize one factor at a time while fixing other factors. Specifically,

$$\hat{\mathbf{G}} = \underset{\mathbf{G} \in \mathbb{R}^G: \|\mathbf{G}\|_2=1}{\arg \max} \mathcal{Y}(\mathbf{G}, \hat{\mathbf{I}}, \hat{\mathbf{T}}), \quad \hat{\mathbf{I}} = \underset{\mathbf{I} \in \mathbb{R}^I: \|\mathbf{I}\|_2=1}{\arg \max} \mathcal{Y}(\hat{\mathbf{G}}, \mathbf{I}, \hat{\mathbf{T}}), \quad \hat{\mathbf{T}} = \underset{\mathbf{T} \in \mathbb{R}^T: \|\mathbf{T}\|_2=1, \mathbf{T} \geq 0}{\arg \max} \mathcal{Y}(\hat{\mathbf{G}}, \hat{\mathbf{I}}, \mathbf{T}).$$

Now consider the first sub-optimization  $\widehat{\mathbf{G}} = \max_{\mathbf{G} \in \mathbb{R}^G: \|\mathbf{G}\|_2=1} \mathcal{Y}(\mathbf{G}, \widehat{\mathbf{I}}, \widehat{\mathbf{T}})$ . By definition,  $\mathcal{Y}(\mathbf{G}, \widehat{\mathbf{I}}, \widehat{\mathbf{T}}) = \langle \mathbf{G}, \mathcal{Y}(\cdot, \widehat{\mathbf{I}}, \widehat{\mathbf{T}}) \rangle$ , where  $\mathcal{Y}(\cdot, \widehat{\mathbf{I}}, \widehat{\mathbf{T}})$  is a length- $n_G$  vector. Hence, the solution for  $\widehat{\mathbf{G}}$  is

$$\widehat{\mathbf{G}} = \mathcal{Y}(\cdot, \widehat{\mathbf{I}}, \widehat{\mathbf{T}}) / \|\mathcal{Y}(\cdot, \widehat{\mathbf{I}}, \widehat{\mathbf{T}})\|_2. \quad (4)$$

Similar argument applies to the second sub-optimization,

$$\widehat{\mathbf{I}} = \mathcal{Y}(\cdot, \widehat{\mathbf{I}}, \widehat{\mathbf{T}}) / \|\mathcal{Y}(\cdot, \widehat{\mathbf{I}}, \widehat{\mathbf{T}})\|_2. \quad (5)$$

For the third sub-optimization, we note that  $\mathcal{Y}(\widehat{\mathbf{G}}, \widehat{\mathbf{I}}, \cdot)$  is a length- $n_T$  vector. Without loss of generality, assume  $\mathcal{Y}(\widehat{\mathbf{G}}, \widehat{\mathbf{I}}, \cdot)$  has at least one positive entry. Otherwise, we can flip the sign of  $\widehat{\mathbf{G}}$  (or  $\widehat{\mathbf{I}}$ ) and work with  $-\mathcal{Y}(\widehat{\mathbf{G}}, \widehat{\mathbf{I}}, \cdot)$ . By Lemma 1, the solution for  $\widehat{\mathbf{T}}$  is

$$\widehat{\mathbf{T}} = \mathcal{Y}(\widehat{\mathbf{G}}, \widehat{\mathbf{I}}, \cdot)_+ / \|\mathcal{Y}(\widehat{\mathbf{G}}, \widehat{\mathbf{I}}, \cdot)_+\|_2. \quad (6)$$

Combining (3),(4),(5) and (6) yields the desired conclusion.  $\square$

**Remark 1.** In practice, we can combine the first two sub-optimizations into a single step of matrix SVD. Specifically,

$$(\widehat{\mathbf{G}}, \widehat{\mathbf{I}}) = \arg \max_{(\mathbf{G}, \mathbf{I}): \|\mathbf{G}\|_2=\|\mathbf{I}\|_2=1} \mathcal{Y}(\mathbf{G}, \mathbf{I}, \widehat{\mathbf{T}}) = \arg \max_{(\mathbf{G}, \mathbf{I}): \|\mathbf{G}\|_2=\|\mathbf{I}\|_2=1} \mathbf{G}\mathbf{M}\mathbf{I}^T \quad (7)$$

where  $\mathbf{M} := \mathcal{Y}(\cdot, \cdot, \widehat{\mathbf{T}})$  is a  $n_G \times n_T$  matrix. So the solution to the above optimization is simply the leading singular-vector pairs of the matrix  $\mathbf{M}$ . This trick is in spirit similar to that used in the two-mode HOSVD algorithm (Wang and Song 2017).

**Lemma 1.** Suppose  $\mathbf{a} \in \mathbb{R}^d$  and  $\mathbf{a}$  has at least one positive entry. Let  $f(\mathbf{x}) = \langle \mathbf{x}, \mathbf{a} \rangle$ , where  $\mathbf{x} \in \mathbb{R}^d$ . Then

$$\arg \max_{\mathbf{x} \in \mathbb{R}^d: \mathbf{x} \geq 0, \|\mathbf{x}\|_2=1} f(\mathbf{x}) = \frac{\mathbf{a}_+}{\|\mathbf{a}_+\|_2}.$$

*Proof.* Let  $\mathbf{a} = (a_1, \dots, a_d)^T$ . Define  $a_+ := \max(a, 0)$  and  $a_- = \min(a, 0)$  for any  $a \in \mathbb{R}$ . We write  $\mathbf{a} = \mathbf{a}_+ + \mathbf{a}_-$ , where  $\mathbf{a}_+ = (a_{1,+}, \dots, a_{d,+})$  and  $\mathbf{a}_- = (a_{1,-}, \dots, a_{d,-})$ . Note that for any  $\mathbf{x} \geq 0$ ,

$$f(\mathbf{x}) = \langle \mathbf{x}, \mathbf{a}_+ \rangle + \langle \mathbf{x}, \mathbf{a}_- \rangle.$$

Let  $\mathcal{I} = \{i \in [d] : a_i < 0\}$ . We claim that the optimizer  $\mathbf{x}^* = (x_1^*, \dots, x_d^*)^T$  must satisfy  $x_i^* = 0$  for all  $i \in \mathcal{I}$ , i.e.,  $\sum_{i \notin \mathcal{I}} (x_i^*)^2 = \|\mathbf{x}^*\|_2^2 = 1$ . Otherwise, we can construct another vector  $\mathbf{y} = (y_1, \dots, y_d)^T$  by

$$y_i = \begin{cases} \frac{x_i^*}{\sqrt{\sum_{i \notin \mathcal{I}} (x_i^*)^2}}, & \text{if } i \notin \mathcal{I}, \\ 0, & \text{if } i \in \mathcal{I}. \end{cases}$$

Then it follows that  $\mathbf{y} \geq 0$ ,  $\|\mathbf{y}\|_2 = 1$ , and

$$f(\mathbf{y}) = \langle \mathbf{y}, \mathbf{a}_+ \rangle = \frac{\langle \mathbf{x}^*, \mathbf{a}_+ \rangle}{\sqrt{\sum_{i \notin \mathcal{I}} (x_i^*)^2}} > \langle \mathbf{x}^*, \mathbf{a}_+ \rangle \geq \langle \mathbf{x}^*, \mathbf{a}_+ \rangle + \langle \mathbf{x}^*, \mathbf{a}_- \rangle = f(\mathbf{x}^*),$$

which contradicts with the assumption. So the optimizer  $\mathbf{x}^*$  must satisfy

$$f(\mathbf{x}^*) = \sum_{i \notin \mathcal{I}} x_i^* a_i = \langle \mathbf{x}^*, \mathbf{a}_+ \rangle. \quad (8)$$

The optimizer for (8) is thus obtained at  $\mathbf{x}^* = \frac{\mathbf{a}_+}{\|\mathbf{a}_+\|_2}$ .  $\square$

## 1.2 Pseudocode

---

**Algorithm 1** Semi-nonnegative tensor decomposition

---

**Input:** Expression tensor  $\mathcal{Y} \in \mathbb{R}^{n_G \times n_I \times n_T}$ , number of components  $R$ .

**Output:** Singular values  $\{\hat{\lambda}_r\}$ , triplets of singular vectors  $\{(\hat{\mathbf{G}}_r, \hat{\mathbf{I}}_r, \hat{\mathbf{T}}_r)\}$ , where  $\hat{\mathbf{T}}_r \geq 0$  entrywise.

**for**  $r=1$  to  $R$  **do**

    Initialize  $(\hat{\mathbf{G}}_r, \hat{\mathbf{I}}_r, \hat{\mathbf{T}}_r)$  by running two-mode HOSVD algorithm (Wang and Song 2017) on  $\mathcal{Y}$ ;

**while** convergence is not reached **do**

        Update  $(\hat{\mathbf{G}}_r, \hat{\mathbf{I}}_r) = \arg \max_{(\mathbf{G}_r, \mathbf{I}_r): \|\mathbf{G}_r\|_2 = \|\mathbf{I}_r\|_2 = 1} \mathcal{Y}(\mathbf{G}_r, \mathbf{I}_r, \hat{\mathbf{T}}_r)$  by matrix SVD as in (7);

        Update  $\hat{\mathbf{T}}_r = \mathcal{Y}(\hat{\mathbf{G}}_r, \hat{\mathbf{I}}_r, \cdot)_+ / \|\mathcal{Y}(\hat{\mathbf{G}}_r, \hat{\mathbf{I}}_r, \cdot)_+\|_2$  as in (6);

        Update  $\hat{\lambda}_r = \mathcal{Y}(\hat{\mathbf{G}}_r, \hat{\mathbf{I}}_r, \hat{\mathbf{T}}_r)$  as in (3);

**end while**

    Update  $\mathcal{Y} \leftarrow \mathcal{Y} - \hat{\lambda}_r \hat{\mathbf{G}}_r \otimes \hat{\mathbf{I}}_r \otimes \hat{\mathbf{T}}_r$ ;

**end for**

---

## 1.3 Permutation-based procedure for selecting top genes

Let  $[n_G]$  denote all genes in the analysis, and let  $\hat{\mathbf{G}}_r = (\hat{G}_{r,1}, \dots, \hat{G}_{r,n_G})^T$  be the  $r$ -th eigen-gene estimate. Genes with extreme loadings contribute more to this expression module, and we are particularly interested in the overexpressed and underexpressed gene clusters  $\mathcal{G}_{\text{top}} = \{i \in [n_G]: \hat{G}_{r,i} \geq c_{\text{top}}\}$  and  $\mathcal{G}_{\text{bottom}} = \{i \in [n_G]: \hat{G}_{r,i} \leq c_{\text{bottom}}\}$ . Here  $c_{\text{top}}$  and  $c_{\text{bottom}}$  are thresholds which control the cluster sizes.

We propose a permutation-based procedure to determine the cut-off values (and thus gene cluster sizes) at significance level  $\alpha$ . Specifically, we generate a set of null tensors by randomly and independently permuting genes for every individual-tissue pair  $(j, k)$ , i.e.,

$$\mathcal{Y}^{\text{null}}([n_G], \text{individual } j, \text{tissue } k) \stackrel{\text{def}}{=} \mathcal{Y}(\text{Perm}([n_G]), \text{individual } j, \text{tissue } k),$$

where  $\text{Perm}([n_G])$  represents a random permutation of the set  $[n_G]$  for individual  $i$  and tissue  $k$ . The shuffled tensor therefore represents a null expression tensor without any genuine gene clusters across samples. We then decompose each of the null tensors  $\mathcal{Y}^{\text{null}}$  and use their eigen-genes  $\hat{\mathbf{G}}^{\text{null}} = (\hat{G}_{r,1}^{\text{null}}, \dots, \hat{G}_{r,n_G}^{\text{null}})^T$  to approximate the null distribution of  $\hat{G}_r$ -values. The cut-off value  $c_{\text{top}}$  (respectively,  $c_{\text{bottom}}$ ) is determined using the top  $\alpha$ -quantile (respectively, bottom  $\alpha$ -quantile) of the empirical distribution of  $\{\hat{G}_{r,i}^{\text{null}}\}$ .

## 1.4 Data Processing

Here we describe our data processing steps and additional results in the GTEx analysis.

**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:

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

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. We log-transformed the expression data because otherwise the raw counts are substantially skewed. Furthermore, our downstream tensor model assumes i.i.d. normal noise, and such assumption makes more sense on the log scale (otherwise, genes that are highly expressed will have high variance).

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

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

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 S5) 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.

## 1.5 Fine structures in subtensors of similar tissues

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). The analyses for the tissue group (i) has been detailed in the main manuscript, and now we describe the results for the tissue groups (ii)–(vi).

**Tissue-specific and race/sex-related expression in cardiac and skeletal muscles.** The muscle subtensor consists of gene expression profiles sampled from two heart regions of cardiac muscle (atrial appendage, left ventricle) and skeletal muscle. As seen from 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, 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,  $3.3 \times 10^{-13}$  for gender effect, and  $1.3 \times 10^{-9}$  for age effect, respectively.

**Gender-driven distinction between breast and two adipose tissues.** In the adipose subtensor, we detected several modules representing highly expressed genes in breast tissue and fe-

males. For example, the eigen-tissue in module 2 (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. For example, the increased expression of *SCGB2A2* is thought to be linked with low-grade, steroid receptor-positive tumors in postmenopausal patients, and as a biomarker, *SCGB2A2* currently has the highest diagnostic accuracy for the screening of metastatic breast cancer (Barh 2014; Lacroix 2006). *KRT17* is known to be up-regulated in triple-negative breast tumors and is a marker of poor prognosis for patients with advanced stage ER(-)/HER2(-) breast cancer (Merkin et al. 2017). Meanwhile, *PIP* encodes a protein which promotes the growth of breast cancer cells (Naderi and Vanneste 2014). Each of these five genes has a significant sex effect in the breast tissue ( $p$  ranging from  $9.8 \times 10^{-8}$  to  $1.1 \times 10^{-18}$ ), 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 (Supplemental Figure S4b), 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 (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, 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 *CHRD2*, *DPP6*, *TEX15* and *ZCCHC12*. 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 this gene is preferably expressed in the uterus relative 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 =  $1.3 \times 10^{-16}$  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 the prostate and testis are characterized by two distinct clusters of genes (components 2 and 3 in Supplemental Table S6). Genes overexpressed in the 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 overexpressed in testis, on the other hand, are mostly enriched for sperm motility ( $p = 1.4 \times 10^{-17}$ ), meiosis I ( $p = 7.9 \times 10^{-17}$ ), and male meiosis ( $p = 5.1 \times 10^{-7}$ ). 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).

**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 S6, 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 S7). As we reduce the dimension of expression data from thousands 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 S8) 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 S8) 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 (explaining 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.

## 1.6 Measures of uncertainty

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



## 2 Supplemental figures and tables

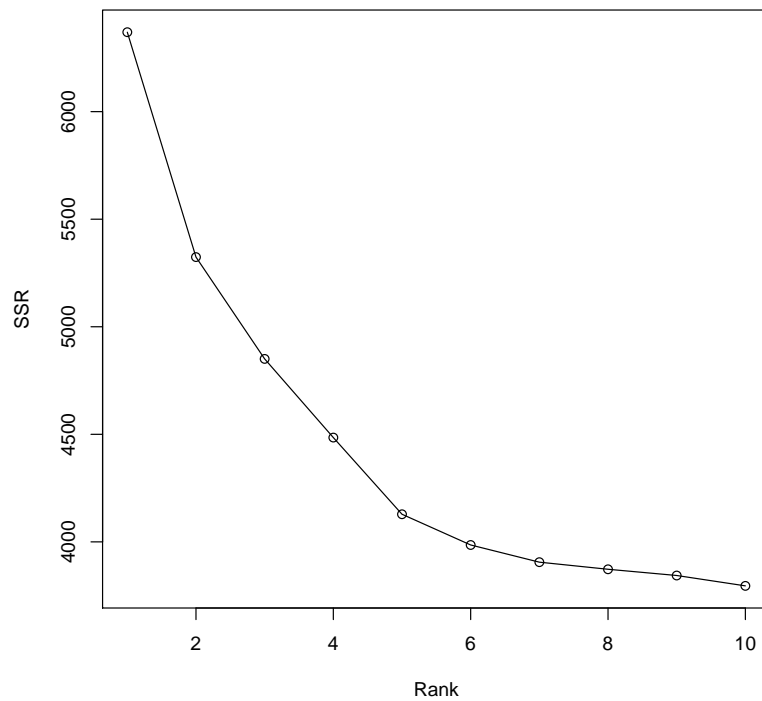


Figure S1: **Sum of squared residuals in the brain tensor.** The sum of squared residual (SSR) was calculated as  $\|\mathcal{Y} - \sum_{r=1}^R \hat{\lambda}_r \hat{\mathbf{G}}_r \otimes \hat{\mathbf{I}}_r \otimes \hat{\mathbf{T}}_r\|_F^2$ , which is a function of the rank  $R$ .

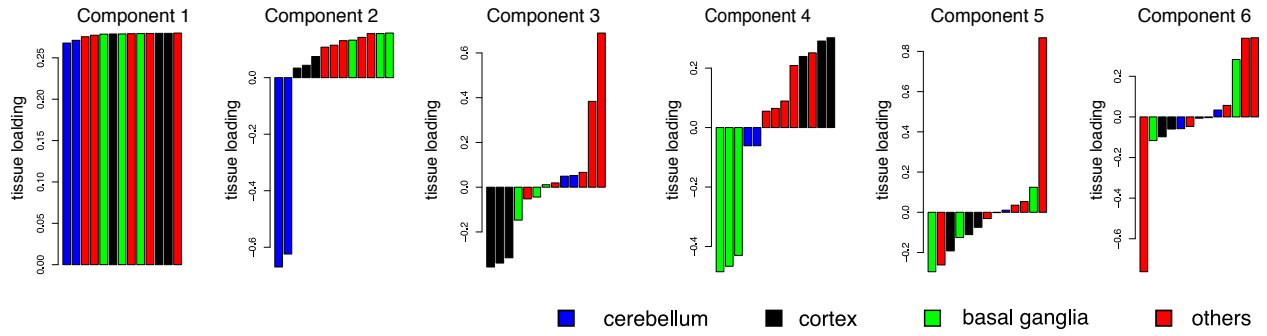


Figure S2: **Barplot for tissue factors in the brain tensor using *HOSVD*.** We applied *HOSVD* to the brain tensor and depicted the top six tissue components. Each panel is the barplot of the sorted tissue loading vector, where tissues are colored based on functional similarity.

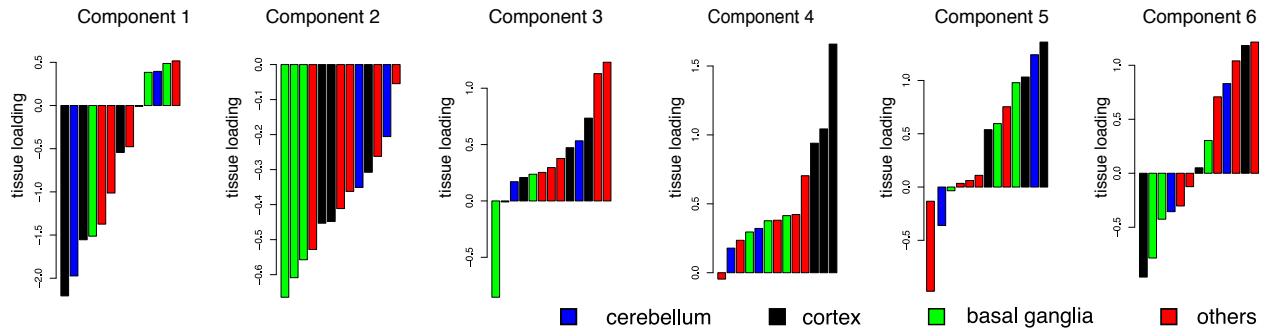


Figure S3: **Barplot for tissue factors in the brain tensor using *SDA*.** We applied *SDA* to the brain tensor and depicted the top six tissue components. Each panel is the barplot of the sorted tissue loading vector, where tissues are colored based on functional similarity.

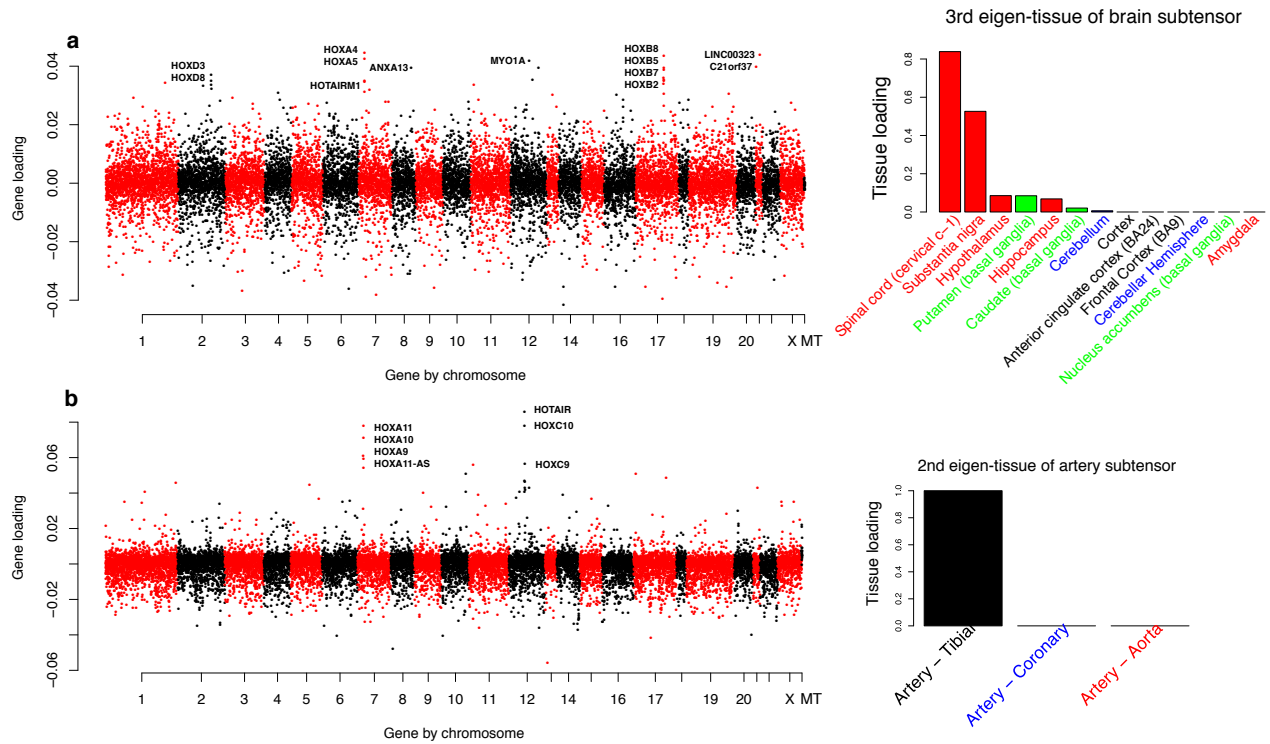


Figure S4: ***HOX*** gene expressions and associated tissue loadings in different subtensors. (a) Over-expression of *HOX* genes in spinal cord (cervical C-1) compared to other brain tissues. This expression pattern was identified from the 3rd tensor component of the brain subtensor. (b) Over-expression 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.

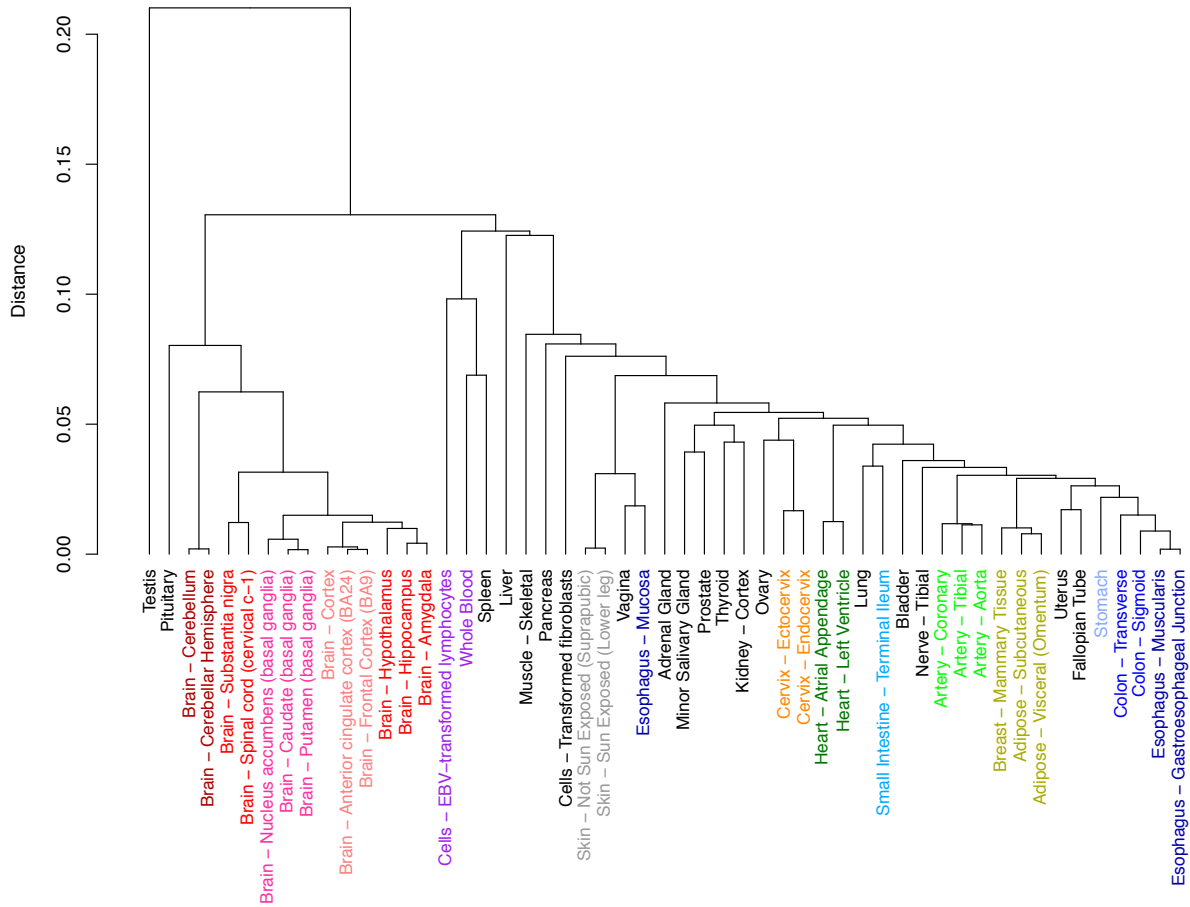


Figure S5: **Hierarchical clustering of GTEx tissues.** To build the hierarchical tree, we took computed the mean gene expressions across individuals in a given tissue to produce a gene-by-tissue matrix. We then constructed the distance matrix based on the tissue-tissue Spearman correlation matrix. The hierarchical tree was then assembled using UPGMA (Unweighted Pair Group Method with Arithmetic Mean) Sokal (1958) based on tissue-tissue correlation matrix. Tissues are colored based on functional similarity.

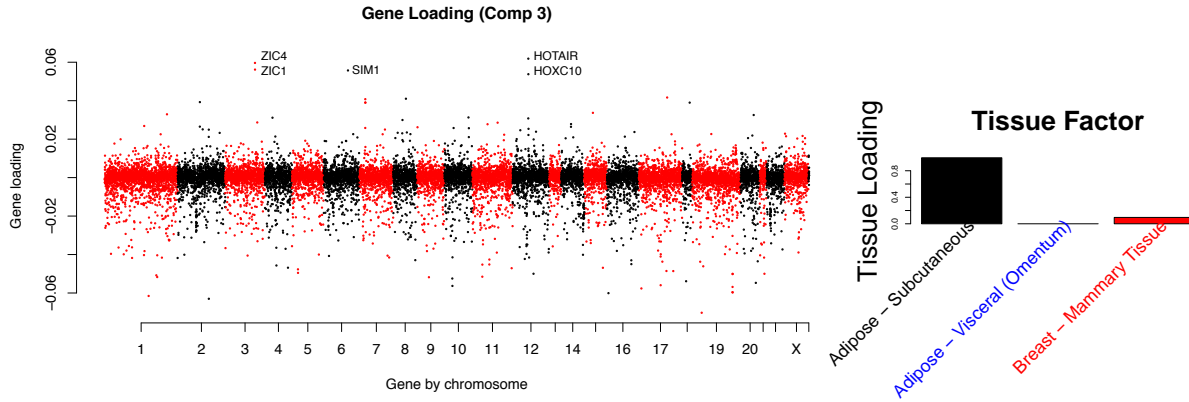


Figure S6: **Over-expression of *HOX* and *ZIC* genes in subcutaneous adipose compared to visceral adipose and breast.** This expression pattern was identified from the 3rd tensor component of the adipose subtensor. The left panel shows the gene loadings in the eigen-gene against the gene positions on the chromosomes. Genes with extreme loadings (e.g., *ZIC1*, *ZIC4*, *HOTAIR*, *HOXC10*) were labeled on the plot. The right figure shows the barplot of sorted tissue loadings in the corresponding eigen-tissue.

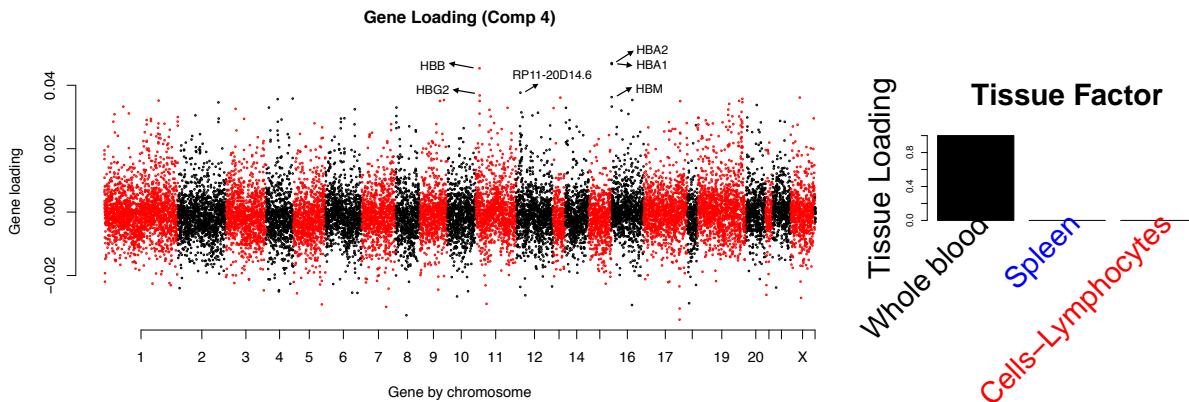


Figure S7: **Over-expression of *HB* genes in whole blood compared to spleen and lymphocytes.** This expression pattern was identified from the 4th tensor component of the blood subtensor. The left panel shows the gene loadings in the eigen-gene against the gene positions on the chromosomes. Genes with extreme loadings (e.g., *HBB*, *HBG2*, *HBA*, *HBM*) were labeled on the plot. The right figure shows the barplot of sorted tissue loadings in the corresponding eigen-tissue.

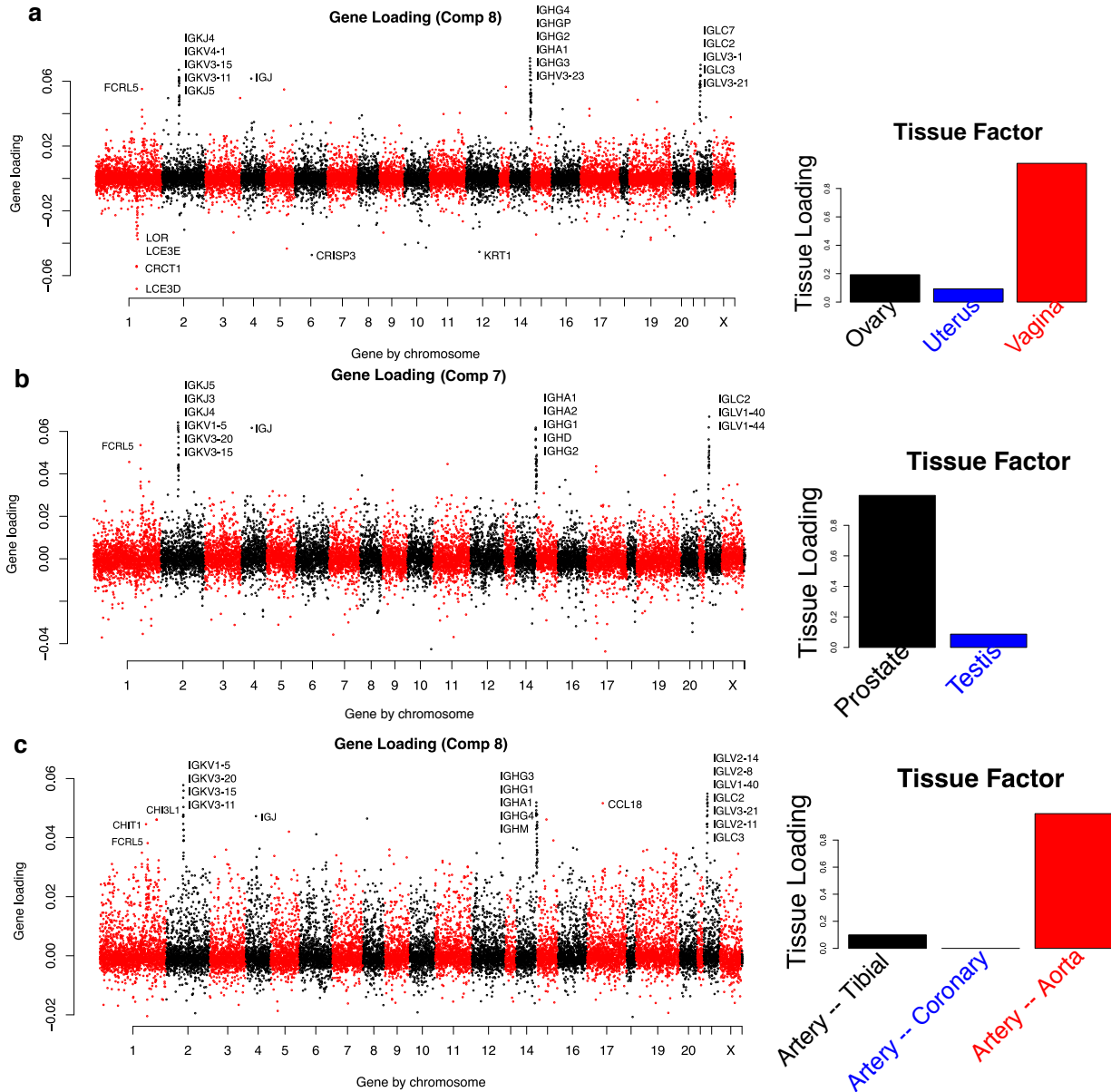


Figure S8: **Immune gene expressions and associated tissue loadings in different sub-tensors.** (a) Over-expression of immune genes in vagina compared to other female tissues. This expression pattern was identified from the 8th tensor component of the female subtensor. (b) Over-expression of immune genes in prostate compared to other male tissues. This expression pattern was identified from the 7th tensor component of the male subtensor. (c) Over-expression of immune genes in aorta compared to the other two artery tissues. This expression pattern was identified from the 8th 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., *IGK* genes, *IGH* genes, *IGL* 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 genes: MT-ND4, MT-RNR2, MT-CO1, MT-CO3, MT-CO2	All	<b>9.2</b>	<b>6.1</b>	<b>7.4</b>
2				Over-expressed cluster: 899 genes: OPALIN, GFAP, C1orf61, MOBP, CPLX2	Brain	<b>23.8</b>	0.3	<b>4.3</b>
				Under-expressed cluster: 43 genes: CCL21, CCL14, CPA3, HOXB4, TCF21				
3				Over-expressed cluster: 89 genes: IGHM, FCRL5, IGJ, MS4A1, BLK	Blood	0.2	0.2	<b>4.5</b>
				Under-expressed cluster: 473 genes: CPLX2, FRRS1L, GFAP, MT3, ATCAY				
4				Over-expressed cluster: 352 genes: MYOC, PLN, DPT, HAND2, HAND2-AS1	Artery	0.2	<b>7.0</b>	<b>6.0</b>
				Under-expressed cluster: 266 genes: SLCO1A2, SSTR3, C1orf61, CAMSAP3, OLIG1				
5				Over-expressed cluster: 655 genes: GGT6, ESRP1, GRHL2, RAB25, PRSS8	Skin	<b>2.7</b>	<b>8.4</b>	0.9
				Under-expressed cluster: 425 genes: CCL21, CCL14, FBP1, PLVAP, STAB1				
6				Over-expressed cluster: 119 genes: ZIC4, ZIC1, FGF5, NEUROG2, PAX3	Cell lines	0.5	<b>4.6</b>	0.0
				Under-expressed cluster: 454 genes: NR0B2, ALDOB, ALB, GSTA2, PAH				
7				Over-expressed cluster: 454 genes: NR0B2, ALDOB, ALB, GSTA2, PAH	Liver	1.7	8.2	<b>2.2</b>
				Under-expressed cluster: 135 genes: KLK5, NIPAL4, TGM5, COL17A1, KLK7				
8				Over-expressed cluster: 95 genes: ATP1A3, MPO, UNC45B, HBB, DEFA3	Blood/Muscle	0.0	1.5	0.1
				Under-expressed cluster: 496 genes: FAM180A, AOX1, GREM1, KRT7, HOXC8				
9				Over-expressed cluster: 348 genes: MB, NRAP, CSRP3, LMOD2, MYH7	Muscle	0.9	0.8	1.3
				Under-expressed cluster: 50 genes: TMEM130, FSD1, PTPRN, APBA2, RAB39B				
10				Over-expressed cluster: 102 genes: ZP2, SLC22A31, C16orf11, CDC42BPG, GFRA3	Cerebellum	0.6	<b>11.9</b>	0.2
				Under-expressed cluster: 181 genes: DDN, GDA, LHX2, NRGN, KCNJ4				

Table S1: **Top 10 expression modules in the GTex tensor.** The top 10 expression modules (ranked by their singular values) were identified from the *MultiCluster* method. For each component, we plot the barplots for the sorted tissue loadings, gene loadings, and individual loadings, respectively. In each eigen-gene, we list the top 5 genes as well as the top enriched GO annotations in the identified gene clusters. In each eigen-tissue, we report the leading tissue with the largest tissue loading. In each eigen-individual, we report the proportion of the variance of the individual loadings explained by age, sex, or race, respectively. A value in bold indicates  $p < 10^{-3}$ .

Module	Component			Eigen-gene	Eigen-tissue	Eigen-individual		
						variance explained (%)		
					driving tissue	age	sex	race
1				Top genes: MT-ND4, MT-CO1, MT-CO3, MT-CYB, MT-ATP6	All	1.9	0.1	0.1
2				Over-expressed cluster: 256 genes: MYH2, AMPD1, MYL1, MYLPF, MYBPC2	Muscle - skeletal	1.2	0.2	0.1
				Under-expressed cluster: 203 genes: MYL7, NKX2-5, MYBPC3, GATA4, TBX5				
3				Over-expressed cluster: 262 genes: MYL7, NPPA, BMP10, NKX2-5, NPPB	Heart	0.7	<b>6.4</b>	<b>7.9</b>
				Under-expressed cluster: 249 genes: MYH2, MYL1, MYBPC2, MYLPF, RP11-451G4.2				
4				Over-expressed cluster: 233 genes: DLK1, IRX4, MYL2, GUCA1C, RP1-46F2.2	Heart - Left ventricle	0.1	0.3	<b>1.7</b>
				Under-expressed cluster: 288 genes: BMP10, SHD, HAMP, SLN, NPPA				
5				Over-expressed cluster: 265 genes: BMP10, HAMP, SHD, SLN, NPPA	Heart - trial appendage	0.0	0.1	<b>9.7</b>
				Under-expressed cluster: 240 genes: DLK1, IRX4, MYL2, GUCA1C, P2RX3				
6				Over-expressed cluster: 176 genes: AC008592.8, C5orf27, SAA2, CHI3L1, SAA1	Muscle - skeletal	2.1	0.2	1.5
				Under-expressed cluster: 377 genes: SMC01, PAQR9, GADL1, PPP1R1C, KY				
7				Over-expressed cluster: 415 genes: SAA1, SAA2, ARHGAP36, MYH8, COL19A1	Muscle - skeletal	<b>3.6</b>	0.4	0.1
				Under-expressed cluster: 43 genes: MYH6, TNNI3, MYBPC3, MYL7, FABP3				
8				Over-expressed cluster: 256 genes: PTX3, SERPINE1, NPPB, LIPG, FOSL1	Heart - left ventricle	0.4	0.5	<b>2.8</b>
				Under-expressed cluster: 210 genes: KLHL38, LRRRC14B, AC007126.1, CTD-2083E4.7, SEC14L5				
9				Over-expressed cluster: 342 genes: MT1A, MYL3, CGA, ANKRD2, MT1M	Muscle - skeletal	1.6	0.0	0.0
				Under-expressed cluster: 118 genes: MAFG-AS1, RP11-28001.2, FAM186A, KCNK9, AKR1C3				
10				Over-expressed cluster: 18 genes: MTND4P12, MTRNR2L2, MTND5P11, PMP2, MTATP8P1	Muscle - skeletal	1.6	0.4	0.1
				Under-expressed cluster: 281 genes: KRT13, PRSS1, KRT4, GP2, PNLIP				

Table S2: **Top 10 expression modules in the muscle subtensor.** The muscle subtensor contains muscle-skeletal (in black), heart-atrial appendage (in blue), and heart-left ventricle (in red). The top 10 expression modules (ranked by their singular values) were identified from the *MultiCluster* method. For each component, we plot the barplots for the sorted tissue loadings, gene loadings, and individual loadings, respectively. In each eigen-gene, we list the top 5 genes as well as the top enriched GO annotations in the identified gene clusters. In each eigen-tissue, we report the leading tissue with the largest tissue loading. In each eigen-individual, we report the proportion of the variance of the individual loadings explained by age, sex, or race, respectively. A value in bold indicates  $p < 10^{-3}$ .



Module	Component			Eigen-gene	Eigen-tissue leading tissue	Eigen-individual		
						variance explained (%)		
						age	sex	race
1				Top genes: MT-ND4, MT-CO3, MT-ND2, MT-CYB, MT-ATP6	All	1.1	<b>4.6</b>	0.2
2				Over-expressed cluster: 608 genes: SCGB2A2, KRT17, VTCN1, PIP, MUCL1	Breast - Mammary Tissue	0.6	<b>39.7</b>	2.2
				Under-expressed cluster: 35 genes: ITLN1, WT1, PRG4, ADIPOQ, HAS1				
3				Over-expressed cluster: 68 genes: HOTAIR, ZIC4, ZIC1, SIM1, HOXC10	Adipose - Subcutaneous	0.6	0.1	0.0
				Under-expressed cluster: 584 genes: MUC16, LRP2, CGN, MSLN, KLK11				
4				Over-expressed cluster: 289 genes: ITLN1, WT1, MSLN, ALOX15, WT1-AS	Adipose - Visceral (Omentum)	0.1	2.2	1.0
				Under-expressed cluster: 241 genes: SIM1, IRX5, HOXC10, PAX3, MMP3				
5				Over-expressed cluster: 172 genes: DHRS2, WIF1, TAT, PIP, PROK1	Breast - Mammary Tissue	0.4	<b>71.2</b>	0.6
				Under-expressed cluster: 400 genes: XIST, DPYS, MSLN, IGHG3, ITLN1				
6				Over-expressed cluster: 406 genes: GABRA4, SCUBE3, PDZRN4, LGI1, PCSK2	All	0.1	<b>9.8</b>	<b>4.2</b>
				Under-expressed cluster: 139 genes: XIST, SAA2, MT1A, SAA1, PTX3				
7				Over-expressed cluster: 98 genes: LBP, NKX2-3, KIAA1239, SAA2, LEP	Adipose - Visceral (Omentum)	0.4	0.0	0.6
				Under-expressed cluster: 521 genes: LRP2, CGN, AARD, RORC, TMPRSS3				
8				Over-expressed cluster: 77 genes: BMP3, C12orf39, NDRG4, ASPG, ACE2	All	<b>4.1</b>	0.9	0.0
				Under-expressed cluster: 586 genes: OLR1, SPP1, CHI3L1, TREM2, IL1RN				
9				Over-expressed cluster: 281 genes: CSF3, PTX3, SERPINE1, PCSK1, CCL20	All	0.4	<b>5.1</b>	0.7
				Under-expressed cluster: 216 genes: XIST, HRSP, GLYAT, ABCC6P1, C14orf180				
10				Over-expressed cluster: 261 genes: LIPG, APLN, SLC2A5, CCL2, MXRA5	All	<b>10.7</b>	0.2	0.3
				Under-expressed cluster: 266 genes: ALOX15B, MMP3, CYP4B1, PNMT, RP11-434D9.1				

Table S3: **Top 10 expression modules in the adipose subtensor.** The artery subtensor contains adipose – subcutaneous (in black), adipose – visceral (in blue), and breast (in red). The top 10 expression modules (ranked by their singular values) were identified from the *MultiCluster* method. For each component, we plot the barplots for the sorted tissue loadings, gene loadings, and individual loadings, respectively. In each eigen-gene, we list the top 5 genes as well as the top enriched GO annotations in the identified gene clusters. In each eigen-tissue, we report the leading tissue with the largest tissue loading. In each eigen-individual, we report the proportion of the variance of the individual loadings explained by age, sex, or race, respectively. A value in bold indicates  $p < 10^{-3}$ .

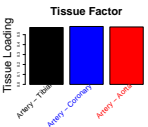
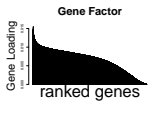

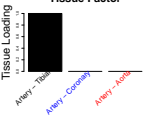
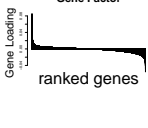
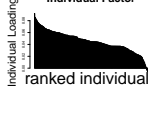
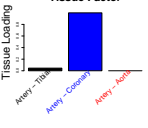
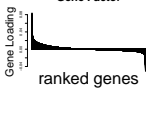
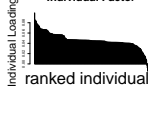
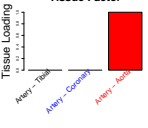
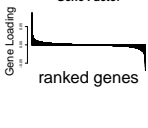
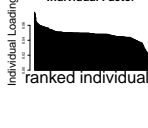
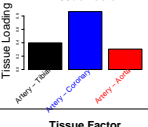

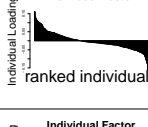
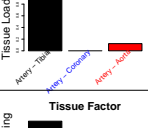

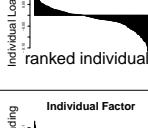
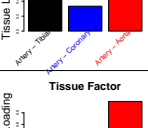


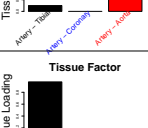
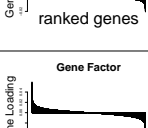
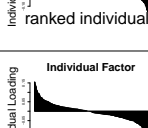
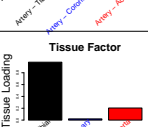
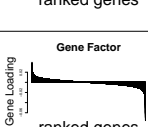
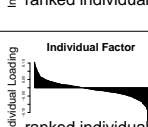
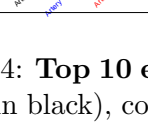
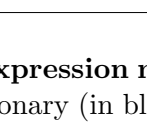
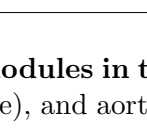
Module	Component			Eigen-gene	Eigen-tissue	Eigen-individual		
						variance explained (%)		
					leading tissue	age	sex	race
1				Top genes: FLNA, ACTB, FN1, MYH11, MT-ND4	All	1.4	0.0	<b>2.8</b>
2				Over-expressed cluster: 121 genes: HOTAIR, HOXC10, HOXA11, HOXA10, HOXA9	Tibial	14.9	0.1	0.2
				Under-expressed cluster: 517 genes: MTUS2, GATA4, CCL18, KCNK17, IL2RA				
3				Over-expressed cluster: 464 genes: GATA4, UNC45B, TBX5, THRSF, TCF21	Coronary	0.0	0.0	<b>6.2</b>
				Under-expressed cluster: 76 genes: CARTPT, HOXC10, PAK7, HOTAIR, HOXA10				
4				Over-expressed cluster: 243 genes: CARTPT, IL13RA2, GALNTL6, EDN2, MTUS2	Aorta	2.4	<b>3.2</b>	0.0
				Under-expressed cluster: 374 genes: FRMD1, OR51E2, CAV3, SRL, PAX3				
5				Over-expressed cluster: 560 genes: IGHG3, IGHG1, IGKV1-5, IGLV3-19, IGKJ4	Coronary	<b>8.7</b>	<b>12.9</b>	<b>6.0</b>
				Under-expressed cluster: 84 genes: XIST, ASPG, WNK2, COL6A6, ABCC8				
6				Over-expressed cluster: 83 genes: KCNK17, GATA4, MTUS2, COL4A3, LGR6	Tibial	4.4	0.9	0.3
				Under-expressed cluster: 572 genes: HOTAIR, EN1, HOXC10, HOXA11, HMCN2				
7				Over-expressed cluster: 255 genes: PKD1L2, GSTM1, DBP, SOST, PLD4	Tibial	<b>3.0</b>	0.1	<b>8.0</b>
				Under-expressed cluster: 213 genes: LBP, PTX3, CALCA, MT1A, ADAMTS4				
8				Over-expressed cluster: 704 genes: IGKV1-5, IGKV3-20, IGLV2-14, IGLV2-8, IGKJ4	Aorta	<b>9.3</b>	0.9	0.0
				Under-expressed cluster: 5 genes: MAPK4, ATP1A2, SP5, SLC8A2, C10orf71				
9				Over-expressed cluster: 190 genes: CHRDL2, LRRC15, THY1, ADAMTSS, FAM43B	Tibial	<b>5.1</b>	0.2	<b>3.8</b>
				Under-expressed cluster: 211 genes: RP1-193H18.3, ALOX15B, PRODH, PDK4, IGSF10				
10				Over-expressed cluster: 118 genes: XKR4, PMP2, MTRNR2L2, MTND4P12, GFRA3	Tibial	<b>8.0</b>	0.1	0.1
				Under-expressed cluster: 301 genes: KRT13, PRSS1, KRT4, CELA3A, PNLIP				

Table S4: **Top 10 expression modules in the artery subtensor.** The artery subtensor contains tibial (in black), coronary (in blue), and aorta (in red). The top 10 expression modules (ranked by their singular values) were identified from the *MultiCluster* method. For each component, we plot the barplots for the sorted tissue loadings, gene loadings, and individual loadings, respectively. In each eigen-gene, we list the top 5 genes as well as the top enriched GO annotations in the identified gene clusters. In each eigen-tissue, we report the leading tissue with the largest tissue loading. In each eigen-individual, we report the proportion of the variance of the individual loadings explained by age, sex, or race, respectively. A value in bold indicates  $p < 10^{-3}$ .

Module	Component			Eigen-gene	Eigen-tissue	Eigen-individual		
						leading tissue	variance explained (%)	
						age	sex	race
1				Top genes: EEF1A1, MT-ND4, MT-ND2, MT-RNR2, MT-CO3	All	0.1	-	0.9
2				Over-expressed cluster: 699 genes: MIR205HG, SERPINB13, TMPRSS11D, KRT15, KRT5	Vagina	1.5	-	1.6
				Under-expressed cluster: 18 genes: SSTR3, LEFTY2, TCF23, SIGLEC11, DLK1				
3				Over-expressed cluster: 339 genes: ARX, NR5A1, GATA4, C4BPB, SIGLEC11	Ovary	0.1	-	2.0
				Under-expressed cluster: 244 genes: HOXA13, MFAP5, HPGD, CHRDL2, HOXA11-AS				
4				Over-expressed cluster: 158 genes: CHRDL2, DPP6, TEX15, ZCCHC12, RSPO3	Uterus	<b>14.5</b>	-	2.7
				Under-expressed cluster: 495 genes: PROK1, RHBG, EDN3, SUSD4, MIR205HG				
5				Over-expressed cluster: 461 genes: PI16, HOXD13, GREM1, ISL1, EYA1	Vagina	3.7	-	0.9
				Under-expressed cluster: 121 genes: SPRR2F, KLHDC8A, SPRR2B, LEMD1-AS1, USH1G				
6				Over-expressed cluster: 276 genes: LRRC55, SEZ6L2, CCL21, THY1, RGS11	Uterus	<b>41.6</b>	-	2.7
				Under-expressed cluster: 243 genes: DCX, TMEM196, TUBA3E, SLC6A11, GRIA2				
7				Over-expressed cluster: 331 genes: RP1-193H18.3, MT1JP, PLA2G2A, ALOX15B, ADRA1A	All	<b>25.6</b>	-	0.1
				Under-expressed cluster: 302 genes: TMEM215, CASKIN1, SEZ6L2, ADAMTS18, EFN3				
8				Over-expressed cluster: 396 genes: IGHG4, IGHGP, IGLC7, IGHG2, IGHA1	Vagina	6.1	-	0.5
				Under-expressed cluster: 118 genes: LCE3D, CRCT1, LCE3E, CRISP3, KRT1				
9				Over-expressed cluster: 301 genes: KIAA1210, SLC24A2, RP11-548O1.3, HSD11B2, LCN8	All	2.5	-	0.9
				Under-expressed cluster: 208 genes: CHI3L1, HSPA6, TNFSF14, PLA2G2A, IL6				
10				Over-expressed cluster: 23 genes: NKX3-2, AP000350.5, CYP4F12, ASCL1, VSX1	Ovary	0.4	-	2.3
				Under-expressed cluster: 546 genes: IGFBP1, UNC13C, CASR, AREG, CHGB				

Table S5: **Top 10 expression modules in the female subtensor.** The female subtensor contains ovary (in black), uterus (in blue), and vagina (in red). The top 10 expression modules (ranked by their singular values) were identified from the *MultiCluster* method. For each component, we plot the barplots for the sorted tissue loadings, gene loadings, and individual loadings, respectively. In each eigen-gene, we list the top 5 genes as well as the top enriched GO annotations in the identified gene clusters. In each eigen-tissue, we report the leading tissue with the largest tissue loading. In each eigen-individual, we report the proportion of the variance of the individual loadings explained by age, sex, or race, respectively. A value in bold indicates  $p < 10^{-3}$ .

Module	Component			Eigen-gene	Eigen-tissue	Eigen-individual		
						leading tissue	variance explained (%)	
						age	sex	race
1				Top genes: MT-ND4, MT-RNR2, MT-CO3, MT-ND2, MT-CO2	All	0.6	-	4.1
2				Over-expressed cluster: 56 genes: KLK3, ACP, MIR205HG, HOXB13, MSMB	Prostate	1.7	-	<b>6.7</b>
				Under-expressed cluster: 651 genes: BOD1L2, C16orf82, PRM1, TNP1, PRM2				
3				Over-expressed cluster: 705 genes: BOD1L2, C16orf82, PRM1, LINC00202-2, TNP1	Testis	0.2	-	0.5
				Under-expressed cluster: 40 genes: KLK3, ACP, MIR205HG, HOXB13, KLK4				
4				Over-expressed cluster: 162 genes: BMP5, CHRDL2, TBX4, HSPB3, H19	Prostate	<b>25</b>	-	2.1
				Under-expressed cluster: 359 genes: CHRNA2, LA16c-83F12.6, CTD-2311B13.7, CD177, SLC6A19				
5				Over-expressed cluster: 399 genes: OLFM4, TMPRSS4, MUC4, PADI3, CLCA2	Prostate	1.1	-	0.0
				Under-expressed cluster: 148 genes: SLITRK3, MFAP5, NPY, OR51E2, SERTM1				
6				Over-expressed cluster: 134 genes: REG3G, DEFB119, PRND, FATE1, EPPIN	Testis	0.8	-	1.1
				Under-expressed cluster: 251 genes: KLK3, MSMB, SERPINB11, CHRMI1, GATA3-AS1				
7				Over-expressed cluster: 362 genes: IGLC2, IGKJ5, IGKJ3, IGLV1-40, IGHA1	Prostate	<b>9.2</b>	-	<b>14.1</b>
				Under-expressed cluster: 101 genes: CSF3, DMBT1, KRT16P2, PADI3, FOSL1				
8				Over-expressed cluster: 306 genes: MUC13, GHA2, C2CD4A, IGHA1, IGKJ3	Prostate	3.2	-	<b>11.0</b>
				Under-expressed cluster: 134 genes: PITX2, AQP2, CYP4F8, RP11-844P9.2, NDP				
9				Over-expressed cluster: 290 genes: TGM4, SLC39A2, CYP4F8, HOTAIR, SERPINB11	Prostate	0.3	-	<b>17.2</b>
				Under-expressed cluster: 221 genes: KCNQ2, LA16c-83F12.6, AC022596.6, AC131180.4, LINC00668				
10				Over-expressed cluster: 383 genes: SERTM1, TFAP2B, FOXD3, ANGPTL7, TMEM155	Prostate	1.5	-	2.3
				Under-expressed cluster: 121 genes: KCNQ2, BMP5, TBX4, VSTM2A, HAPLN1				

Table S6: **Top 10 expression modules in the male subtensor.** The female subtensor contains prostate (in black) and testis (in blue). The top 10 expression modules (ranked by their singular values) were identified from the *MultiCluster* method. For each component, we plot the barplots for the sorted tissue loadings, gene loadings, and individual loadings, respectively. In each eigen-gene, we list the top 5 genes as well as the top enriched GO annotations in the identified gene clusters. In each eigen-tissue, we report the leading tissue with the largest tissue loading. In each eigen-individual, we report the proportion of the variance of the individual loadings explained by age, sex, or race, respectively. A value in bold indicates  $p < 10^{-3}$ .

## References

- Afgan, E., Baker, D., Van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., *et al.*, 2016. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic acids research*, **44**(W1):W3–W10.
- Barh, D., 2014. *Omics Approaches in Breast Cancer: Towards Next-Generation Diagnosis, Prognosis and Therapy*. Springer.
- Chettier, R., Ward, K., and Albertsen, H. M., 2014. Endometriosis is associated with rare copy number variants. *PLoS ONE*, **9**(8):e103968.
- 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.
- 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 homozygosity within ARHGEF28 gene in familial amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, **14**(5-6):444–451.
- Efron, B. and Tibshirani, R. J., 1994. *An introduction to the bootstrap*. CRC press.
- 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.
- Lacroix, M., 2006. Significance, detection and markers of disseminated breast cancer cells. *Endocrine-Related Cancer*, **13**(4):1033–1067.
- 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).
- 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.
- 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.
- Milan, L. and Whittaker, J., 1995. Application of the parametric bootstrap to models that incorporate a singular value decomposition. *Applied Statistics*, :31–49.
- Naderi, A. and Vanneste, M., 2014. Prolactin-induced protein is required for cell cycle progression in breast cancer. *Neoplasia*, **16**(4):329–342.
- Sokal, R. R., 1958. A statistical method for evaluating systematic relationship. *University of Kansas science bulletin*, **28**:1409–1438.

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.