MDPI

Article

Integrated Analysis of Tissue-specific Gene Expression in Diabetes by Tensor Decomposition Can Identify Possible Associated Diseases.

Y-H. Taguchi [1,*] and Turki Turki [2]

[1] Department of Physics, Chuo University, Tokyo 112-8551, Japan
[2] Department of Computer Science, King Abdulaziz University, Jeddah 21589, Saudi Arabia; tturki@kau.edu.sa
* Correspondence:tag@granular.com; Tel.: +81-3-3817-1791

Abstract: In the field of gene expression analysis, methods of integrating multiple gene expression profiles are still being developed and the existing methods have scope for improvement. The previously proposed tensor decomposition-based unsupervised feature extraction method was improved by introducing standard deviation optimization. The improved method was applied to perform an integrated analysis of three tissue-specific gene expression profiles (namely, adipose, muscle, and liver) for diabetes mellitus, and the results showed that it can detect diseases that are associated with diabetes (e.g., neurodegenerative diseases) but that cannot be predicted by individual tissue expression analyses using state-of-the-art methods. Although the selected genes differed from those identified by the individual tissue analyses, the selected genes are known to be expressed in all three tissues. Thus, compared with individual tissue analyses, an integrated analysis can provide more in-depth data and identify additional factors, namely, the association with other diseases.

Keywords: gene expression; tensor decomposition; diabetes mellitus; neurodegenerative diseases

Citation: Taguchi, Y.-H.; Turki, T. Title. *Genes* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Submitted to *Genes* for possible open access publication under the terms and conditions of the Creative Commons Attri- bution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Gene expression analysis is an important step for investigating diseases and identifying genes that can be used as therapeutic targets or biomarkers or genes that are causes of disease. Although the development of high throughput sequencing technology (HST) has led to continuous increases in the amount of gene expression profile data, methods of integrating multiple gene expression profiles are still being developed. Tensor decomposition (TD) is a promising candidate method for integrating multiple gene expression profiles. Using this method, gene expression profiles from multiple tissues of individuals can be stored as a tensor $x_{ijk} \in \mathbb{R}^{N\times M\times K}$, which represents the gene expression of the ith gene in the jth individual of the kth tissue. TD provides a method of decomposing a tensor into a series expansion of the product of singular value vectors, each of which represents a gene assigned to a specific individual or tissue. For example, by applying the higher-order singular value decomposition (HOSVD) method to x_{ijk}, we can obtain the following:

$$x_{ijk} = \sum_{\ell_1=1}^{N}\sum_{\ell_2=1}^{M}\sum_{\ell_3=1}^{K} G(\ell_1\ell_2\ell_3)u_{\ell_1 i}u_{\ell_2 j}u_{\ell 3k} \tag{1}$$

where $G \in \mathbb{R}^{N\times M\times K}$ is a core tensor, $u_{\ell_1 i} \in \mathbb{R}^{N\times N}$, $u_{\ell_2 j} \in \mathbb{R}^{M\times M}$, $u_{\ell 3k} \in \mathbb{R}^{M\times M}$ are singular value matrices and orthogonal matrices. We previously proposed a TD-based unsupervised feature extraction (FE) method [1] and applied it to a wide range of genomic sciences. Recently, this method was improved by the introduction of standard deviation (SD) optimization and applied to gene expression [2], DNA methylation [3], and histone modification analyses [4]. Nevertheless, because the updated method was only previously applied to gene expression measured by HST, whether it is also applicable to gene expression profiles

retrieved by microarray technology remains to be clarified. In this paper, an integrated analysis was performed by applying the recently proposed TD-based unsupervised FE method with SD optimization to microarray-measured gene expression data for diabetes mellitus from multiple tissues. We found that applying the TD-based unsupervised FE with SD optimization to gene expression profiles from individual tissues can identify diseases associated with diabetes that cannot be identified by the other state-of-the-art methods.

There are multiple benefits in using TD to identify DEGs. First, since it is not a supervised method, it can select DEGs that are biologically more plausible than those selected using supervised methods. This can be explained using the following example wherein the aim is to identify DEGs that are distinct between two classes, e.g., patients and healthy controls. Supervised methods attempt to identify DEGs associated with a smaller divergence within individual classes, whereas TD allows to select DEGs with within-class divergence to some extent (since TD tries to identify the representative state of distinction between two classes). If the representative state is associated with within-class divergence that has biological origins, e.g., age and sex, this divergence should not be penalized. However, supervised methods often do so whereas the unsupervised method allows biological within-class divergence. Second, TD can select more stable DEGs; i.e., those independent of specific sets of samples considered in the analysis. This is because TD attempts to identify DEGs coincident with those of the representative state, which should be robust. Since sub-sampling does not change the representative state drastically, the gene set selected by TD is not altered drastically either. Third, TD can deal with multiple conditions. For example, if gene expression is measured in various tissues of several people, it is natural to format them as gene $\times$ person $\times$ tissue, which results in a tensor form. We have listed only a few important advantages here. Readers interested in acquiring information on other advantages of TD can refer to our recent book [1].

2. Materials and Methods

2.1. Gene expression

Gene expression profiles (GSE13268, GSE13269, and GSE13270 [5]) were retrieved from the Gene Expression Omnibus (GEO), and they were obtained from a study of the progression of diabetes biomarker diseases in the rat liver, gastrocnemius muscle, and adipose tissue. Each of these profiles is composed of gene expression profiles from five individuals seen in two strains, Goto-Kakizaki and WistarKyoto, and they include data for three tissues (adipose, muscle, and liver) obtained at five time points after treatment. Three files named GSE13268_series_matrix.txt.gz, GSE13269_series_matrix.txt.gz, and GSE13270_series_matrix.txt.gz were downloaded from the Supplementary Files in GEO.

Gene expression profiles were formatted as a tensor, with $x_{ijkmst} \in \mathbb{R}^{31099\times5\times5\times2\times2\times3}$, representing the expression of the ith probe in the tth tissue ($t = 1$: adipose, $t = 2$:muscle, $t = 3$:liver) at the jth time point for the kth replicate and mth treatment at sth strain. These values are normalized as follows:

$$\sum_i x_{ijkmst} = 0 \tag{2}$$

$$\sum_i x^2_{ijkmst} = 31099 \tag{3}$$

2.2. The brief description of the HOSVD algorithm

For demonstration, consider the tensor $x_{ijk} \in \mathbb{R}^{N\times M\times K}$. HOSVD is an algorithm that applies SVD to the unfolded tensor, either $x_{i(jk)} \in \mathbb{R}^{N\times MK}$, $x_{j(ik)} \in \mathbb{R}^{M\times NK}$, or $x_{k(ij)} \in \mathbb{R}^{K\times NM}$,

$$x_{i(jk)} = \sum_{\ell_1=1}^{\min(N,MK)} \lambda_{\ell_1} u_{\ell_1 i} v_{\ell_1 jk} \tag{4}$$

$$x_{j(ik)} = \sum_{\ell_2=1}^{\min(M,NK)} \lambda_{\ell_2} u_{\ell_2 j} v_{\ell_2 ik} \tag{5}$$

$$x_{k(ij)} = \sum_{\ell_3=1}^{\min(K,NM)} \lambda_{\ell_3} u_{\ell_3 k} v_{\ell_3 ij} \tag{6}$$

$$\tag{7}$$

where $u_{\ell_1 i} \in \mathbb{R}^{\min(N,MK)\times N}$, $v_{\ell_1 jk} \in \mathbb{R}^{\min(N,MK)\times MK}$, $u_{\ell_2 j} \in \mathbb{R}^{\min(M,NK)\times M}$, $v_{\ell_2 ik} \in \mathbb{R}^{\min(M,NK)\times NK}$, $u_{\ell_3 k} \in \mathbb{R}^{\min(K,MM)\times K}$, $v_{\ell_3 ij} \in \mathbb{R}^{\min(K,NM)\times NM}$. A core tensor $G(\ell_1\ell_2\ell_3)$ can be calculated as

$$G(\ell_1\ell_2\ell_3) = \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{K} x_{ijk} u_{\ell_1 i} u_{\ell_2 j} u_{\ell_3 k}. \tag{8}$$

Subsequently, we obtain the following

$$x_{ijk} = \sum_{\ell_1=}^{N}\sum_{\ell_2=1}^{M}\sum_{\ell_3=1}^{K} G(\ell_1\ell_2\ell_3) u_{\ell_1 i} u_{\ell_2 j} u_{\ell_3 k} \tag{9}$$

where $u_{\ell_1 i} \in \mathbb{R}^{N\times N}$, $u_{\ell_2 j} \in \mathbb{R}^{M\times M}$, and $u_{\ell_3 k} \in \mathbb{R}^{K\times K}$. When $u_{\ell_1 i}$, $u_{\ell_2 j}$, and $u_{\ell_3 k}$, are smaller than those computed in eqs. (4), (5), and (6), missing values are filled with zero. Then, $u_{\ell_1 i}$, $u_{\ell_2 j}$, and $u_{\ell_3 k}$ are the resultant orthogonal matrices obtained.

2.3. TD-based unsupervised FE with SD optimization

HOSVD was applied to x_{ijkmst}, and we obtained the following:

$$x_{ijkmst} = \sum_{\ell_1=1}^{5}\sum_{\ell_2=1}^{5}\sum_{\ell_3=1}^{2}\sum_{\ell_4=1}^{2}\sum_{\ell_5=1}^{3}\sum_{\ell_6=1}^{31099} G(\ell_1\ell_2\ell_3\ell_4\ell_5\ell_6) u_{\ell_1 j} u_{\ell_2 k} u_{\ell_3 m} u_{\ell_4 s} u_{\ell_5 t} u_{\ell_6 i} \tag{10}$$

where $G(\ell_1\ell_2\ell_3\ell_4\ell_5\ell_6) \in \mathbb{R}^{5\times5\times2\times2\times3\times31099}$ is a core tensor and $u_{\ell_1 j} \in \mathbb{R}^{5\times5}$, $u_{\ell_2 k} \in \mathbb{R}^{5\times5}$, $u_{\ell_3 m} \in \mathbb{R}^{2\times2}$, $u_{\ell_4 s} \in \mathbb{R}^{2\times2}$, $u_{\ell_5 t} \in \mathbb{R}^{3\times3}$, and $u_{\ell_6 i} \in \mathbb{R}^{31099\times31099}$ are singular value matrices and orthogonal matrices.

After identifying $u_{\ell_1 j}$, $u_{\ell_2 k}$, $u_{\ell_3 m}$, $u_{\ell_4 s}$, and $u_{\ell_5 t}$ of interest, we attempted to identify $G(\ell_1\ell_2\ell_3\ell_4\ell_5\ell_6)$ with the largest absolute value and fixed ℓ_1 to ℓ_5. Then, we attributed P-values to i by assuming that $u_{\ell_6 i}$ obeys a Gaussian distribution as follows:

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell_6 i}}{\sigma_{\ell_6}}\right)^2\right] \tag{11}$$

where $P_{\chi^2}[> x]$ is the cumulative χ^2 distribution under the assumption that it is larger than x and σ_{ℓ_6} is the SD. P_i is corrected with the BH criterion [1].

The SD is optimized so that $u_{\ell_6 i}$ obeys the Gaussian distribution (null hypothesis) as much as possible [2], and the optimization process is as follows. Initially, we compute the histogram h_p, which includes the number of i values that satisfy the following:

$$\frac{p}{N_p} \leq 1 - \text{adjusted}P_i \leq \frac{p+1}{N_p}, 0 \leq p \leq N_p - 1 \tag{12}$$

Then, the standard deviation of h_p for $\text{adjusted}P_i > p_0$ for the threshold value p_0,

$$\Delta h_p = \frac{1}{N'_p} \sum_{\text{adjusted}\frac{p}{N_P} > p_0} \left(h_p - \langle h_p \rangle\right)^2 \tag{13}$$

$$\langle h_p \rangle = \frac{1}{N'_p} \sum_{\text{adjusted}\frac{p}{N_P} > p_0} h_p \tag{14}$$

$$N'_P = \sum_{\text{adjusted}\frac{p}{N_P} > p_0} 1 \tag{15}$$

is minimized. Therefore, for $1 - \text{adjusted}P_i < 1 - p_0$, h_p takes a constant value (i.e., it obeys the null hypothesis) as much as possible, whereas for $1 - \text{adjusted}P_i > 1 - p_0$, h_p presents a peak value.

The P_i values are finally calculated using the optimized SD, and i values associated with adjusted P-values less than 0.01 are selected. Probes, is, are converted to gene symbols using the gene ID conversion tool in DAVID [6].

The genes selected are presumed to be coincident with the profiles associated with selected $u_{\ell_1 j}$, $u_{\ell_2 k}$, $u_{\ell_3 m}$, $u_{\ell_4 s}$, and $u_{\ell_5 t}$; this corresponds to the identification of differentially expressed genes.

For details, check sample R source code in supplementary materials.

3. Results

The results indicated that u_{2j} represents the dependence on time, u_{2m} represents the distinction between the controls and treatment, u_{1k} represents the independence of the replicate, u_{1s} represents the independence of strains, and u_{1t} represents the independence of tissues (Fig. 1).

Table 1. The core tensor defined in eq. (10) when HOSVD is applied to data sets, $G(2,1,2,1,1,\ell_6)$.

ℓ_6	$G(2,1,2,1,1,\ell_6)$	ℓ_6	$G(2,1,2,1,1,\ell_6)$
1	3.505367	6	6.227675
2	8.092320	7	6.890919
3	-15.707937	8	-3.287872
4	9.134575	9	-1.529140
5	6.723113	10	-9.158979

Table 1 presents $G(2,1,2,1,1,\ell_6)$, and it shows that $G(2,1,2,1,1,3)$ has the highest absolute value. Then, u_{3i} (i.e., $\ell_6 = 3$) is used to select probes. The P_i values are determined using eq. (11), with $\ell_6 = 3$. SD optimization was performed, and $\sigma_3 = 0.001514603$.

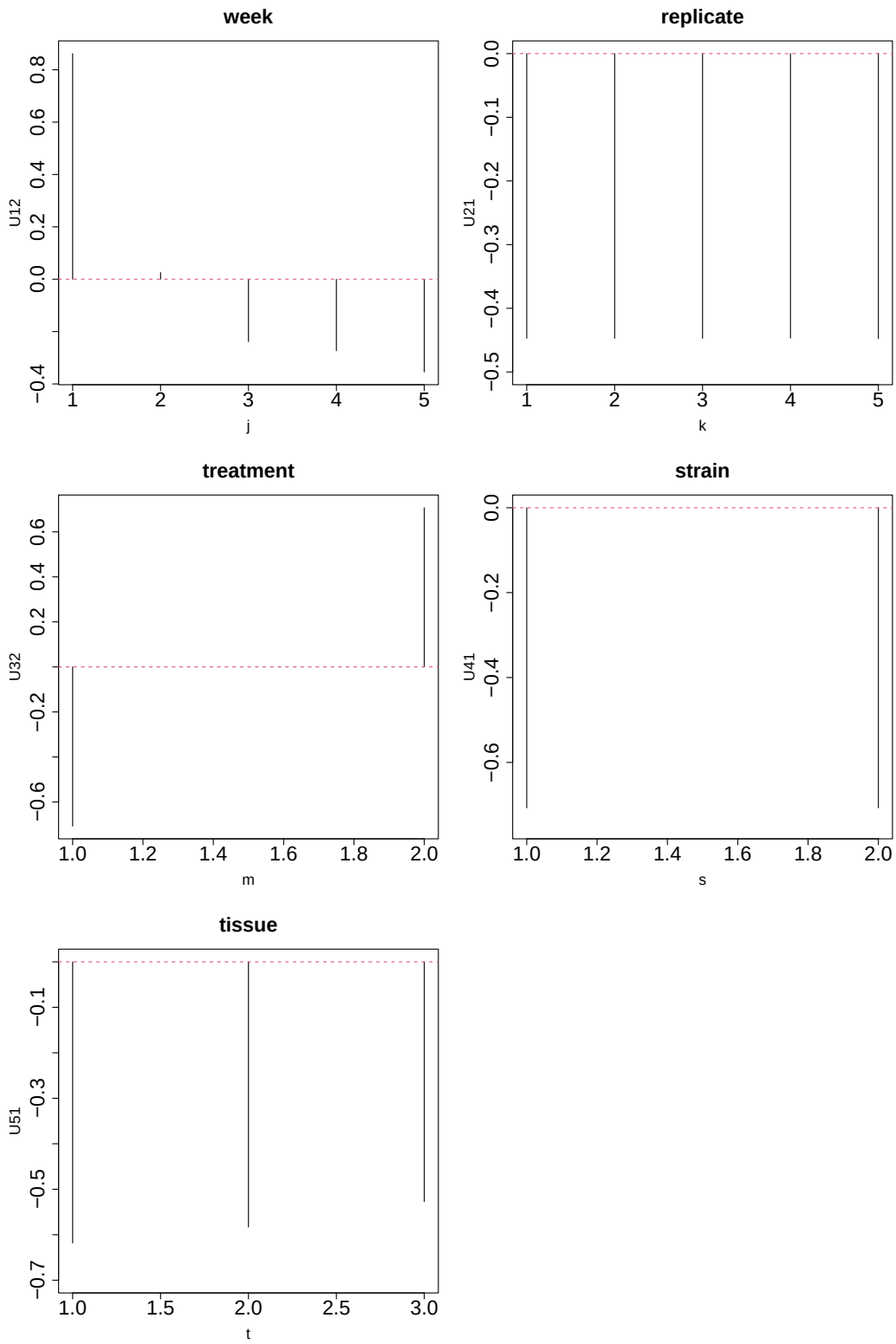

Figure 1. Singular value vectors defined in eq. (10) when HOSVD is applied to data sets. u_{2j} (week), u_{1k} (replicate), u_{2m} (treatment), u_{1s} (strain), and u_{1t} (tissue).

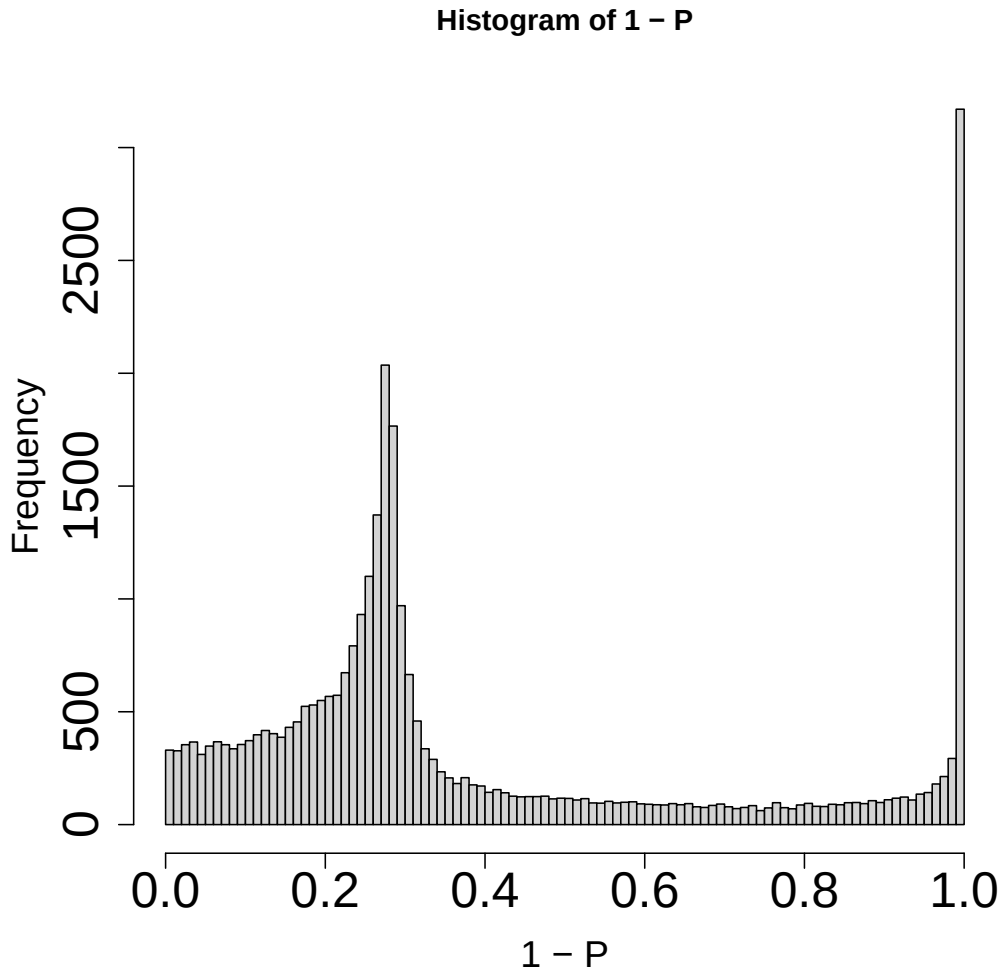

Figure 2. Histogram of 1-P_i.

A histogram of 1-P_i after SD optimization is presented in Figure 2 and the histogram is more similar to that of the null hypothesis, i.e., the combination of a flat region and a sharp peak. Then, the P-values corresponding to the ith probe are determined with eq. (11), and the P_i values are corrected by BH criterion. Finally, 2,452 selected probes are associated with adjusted P-values less than 0.01, and they are further converted to 2,281 gene symbols.

Table 2. Top 10 "KEGG 2021 Human" category terms in Enrichr.

Term	Overlap	P-value	Adjusted P-value
Diabetic cardiomyopathy	83/203	1.89×10^{-31}	5.80×10^{-29}
Prion disease	93/273	7.40×10^{-28}	1.13×10^{-25}
Parkinson disease	86/249	2.50×10^{-26}	2.55×10^{-24}
Oxidative phosphorylation	60/133	7.92×10^{-26}	6.06×10^{-24}
Nonalcoholic fatty liver disease	65/155	1.19×10^{-25}	7.30×10^{-24}
Thermogenesis	76/232	8.28×10^{-22}	4.22×10^{-20}
Complement and coagulation cascades	42/85	2.26×10^{-20}	9.85×10^{-19}
PPAR signaling pathway	39/74	2.58×10^{-20}	9.85×10^{-19}
Alzheimer disease	94/369	3.99×10^{-18}	1.36×10^{-16}
Huntington disease	83/306	6.48×10^{-18}	1.98×10^{-16}

To validate the selected genes, 2,281 gene symbols are uploaded to Enrichr [7] (For the full list of selected probes, genes and enrichment analyses, check the supplementary materials). Table 2 shows the results of the "KEGG 2021 Human" category in Enrichr. Since none of the terms are related to diabetes except for the top term, i.e., "diabetic cardiomyopathy", the process initially appears to be a failure. Nevertheless, a number of the identified diseases are deeply related to diabetes mellitus. For example, many neurodegenerative diseases are listed, and diabetes mellitus is widely known to be a risk factor for neurodegenerative diseases [8–12]. Moreover, diabetes mellitus is known to be associated with thermogenesis [13], oxidative phosphorylation [14], and the PPAR signaling pathway [15]. Thus, the proposed method is successful in contrast to the first impression and can identify many diseases associated with diabetes mellitus.

Table 3. Top 10 terms in the "ARCHS4 Tissues" category in Enrichr.

Term	Overlap	P-value	Adjusted P-value
LIVER (BULK TISSUE)	481/2316	3.49×10^{-63}	3.77×10^{-61}
VENTRICLE	449/2316	1.67×10^{-49}	9.04×10^{-48}
SKELETAL MUSCLE (BULK TISSUE)	428/2316	2.34×10^{-41}	8.42×10^{-40}
ADIPOSE (BULK TISSUE)	410/2316	6.46×10^{-35}	1.75×10^{-33}
MYOBLAST	409/2316	1.42×10^{-34}	3.08×10^{-33}
SUBCUTANEOUS ADIPOSE TISSUE	401/2316	6.92×10^{-32}	1.25×10^{-30}
ATRIUM	366/2316	2.38×10^{-21}	3.67×10^{-20}
HEART (BULK TISSUE)	363/2316	1.53×10^{-20}	2.07×10^{-19}
HEPATOCYTE	362/2316	2.82×10^{-20}	3.39×10^{-19}
OMENTUM	350/2316	3.25×10^{-17}	3.51×10^{-16}

Table 3 shows the top 10 terms in the category "ARCHS4 tissues" in Enrichr. Remarkably, gene expression is measured for three of the top four tissues. Similar results are found for the "Mouse Gene Atlas" category in Enrichr (Table 4). In conclusion, the proposed method is successful.

Table 4. Top 10 terms in the "Mouse Gene Atlas" category in Enrichr.

Term	Overlap	P-value	Adjusted P-value
mammary gland non-lactating	116/201	7.92×10^{-64}	7.61×10^{-62}
skeletal muscle	229/710	5.23×10^{-63}	2.51×10^{-61}
liver	243/928	3.58×10^{-48}	1.14×10^{-46}
adipose brown	148/456	5.78×10^{-41}	1.39×10^{-39}
heart	154/568	2.53×10^{-32}	4.86×10^{-31}
kidney	80/554	3.98×10^{-4}	5.90×10^{-3}
osteoblast day 21	44/264	4.30×10^{-4}	5.90×10^{-3}
bladder	33/195	1.63×10^{-3}	1.96×10^{-2}
adipose white	33/199	2.29×10^{-3}	2.44×10^{-2}
MEF	45/300	3.33×10^{-3}	3.20×10^{-2}

4. Discussion

Although the proposed method successfully integrated gene expression data measured in three tissues and identified diseases associated with diabetes mellitus, the identified genes also included genes expressed in all three tissues. If other methods that do not require an integrated analysis can perform similarly, then complicated methods, such as the proposed method, will not be required. To determine whether methods without integration can achieve similar performance, we tested three methods: t test, SAM [16] and limma [17]. Since the t test and SAM methods cannot simultaneously consider the distinction between the control and treatment as well as the dependent on time, we attempted to identify genes that presented expression differences between the control and treatment (no consideration

of time dependence). For more details on how to perform these three methods, check the sample R source code in supplementary materials.

Table 5. Number of probes selected by other methods.

Tissue	*t* test	sam	limma
Adipose	556	773	116
Muscle	100	119	2
liver	947	1090	211
ComBat	4009	180	0

Table 5 shows the number of probes selected by the other methods. These methods select fewer probes than the proposed method (2,542 probes), and the number selected in muscle is relatively low. According to the limma method, only two probes could be selected for muscle; thus, the method was not successful. The integrated analysis likely helped identify more probes, which resulted in more significant enrichment.

To further validate the genes selected by other methods, we converted probe IDs to gene symbols and uploaded them to Enrichr. Table 6 presents the results for the other methods on the "Mouse Gene Atlas" category in Enrichr. For muscle, neither SAM nor *t* test could select muscle as top ranked tissues whereas limma could identify only two probes as muscle-specific genes (see Table 5). Thus, the other methods are not better than the proposed method that could identify muscle specificity correctly (Table 4). Figure 3 shows the Venn diagrams between selected genes. Since the proposed method selects different genes from those specifically selected in individual tissues, an integrated analysis is a valuable method.

Table 6. Top three terms by other methods in the "Mouse Gene Atlas" category in Enrichr

Term	Overlap	P-value	Adjusted P-value
t test			
Adipose			
adipose brown	38/456	4.17×10^{-12}	3.92×10^{-10}
mammary gland lact	12/104	3.87×10^{-6}	1.82×10^{-4}
macrophage peri LPS thio 0 hrs	18/353	1.14×10^{-3}	3.59×10^{-2}
Muscle			
adipose brown	29/456	4.34×10^{-26}	2.26×10^{-24}
heart	21/568	3.88×10^{-14}	1.01×10^{-12}
mammary gland lact	4/104	1.15×10^{-3}	1.99×10^{-2}
Liver			
liver	90/928	2.53×10^{-16}	2.38×10^{-14}
adipose brown	40/456	9.53×10^{-7}	4.48×10^{-5}
kidney	40/554	9.11×10^{-5}	2.86×10^{-3}
ComBat			
bone marrow	107/413	1.04×10^{-10}	9.98×10^{-9}
osteoblast day 21	75/264	8.31×10^{-10}	3.99×10^{-8}
embryonic stem line V26 2 p16	149/728	9.44×10^{-7}	3.02×10^{-5}
sam			
Adipose			
adipose brown	51/456	2.61×10^{-16}	2.48×10^{-14}
mammary gland lact	12/104	5.67×10^{-5}	2.69×10^{-3}
macrophage peri LPS thio 0 hrs	23/353	3.54×10^{-4}	1.12×10^{-2}
Muscle			
adipose brown	33/456	2.16×10^{-29}	1.21×10^{-27}
heart	23/568	7.01×10^{-15}	1.96×10^{-13}
mammary gland lact	4/104	1.91×10^{-3}	3.47×10^{-2}
Liver			
liver	93/928	3.10×10^{-14}	2.91×10^{-12}
adipose brown	43/456	1.63×10^{-6}	7.66×10^{-5}
kidney	43/554	1.76×10^{-4}	5.51×10^{-3}
ComBat			
Cell cycle	11/124	2.95×10^{-9}	4.71×10^{-7}
Oocyte meiosis	9/129	6.65×10^{-7}	5.32×10^{-5}
Progesterone-mediated oocyte maturation	8/100	1.00×10^{-6}	5.34×10^{-5}
limma			
Adipose			
adipose brown	14/456	4.19×10^{-8}	2.85×10^{-6}
adipose white	4/199	1.61×10^{-2}	5.46×10^{-1}
intestine small	6/466	2.59×10^{-2}	5.87×10^{-1}
Liver			
liver	33/928	2.39×10^{-11}	1.60×10^{-9}
adipose brown	7/456	1.31×10^{-1}	1.00×10^{0}
heart	8/568	1.57×10^{-1}	1.00×10^{0}

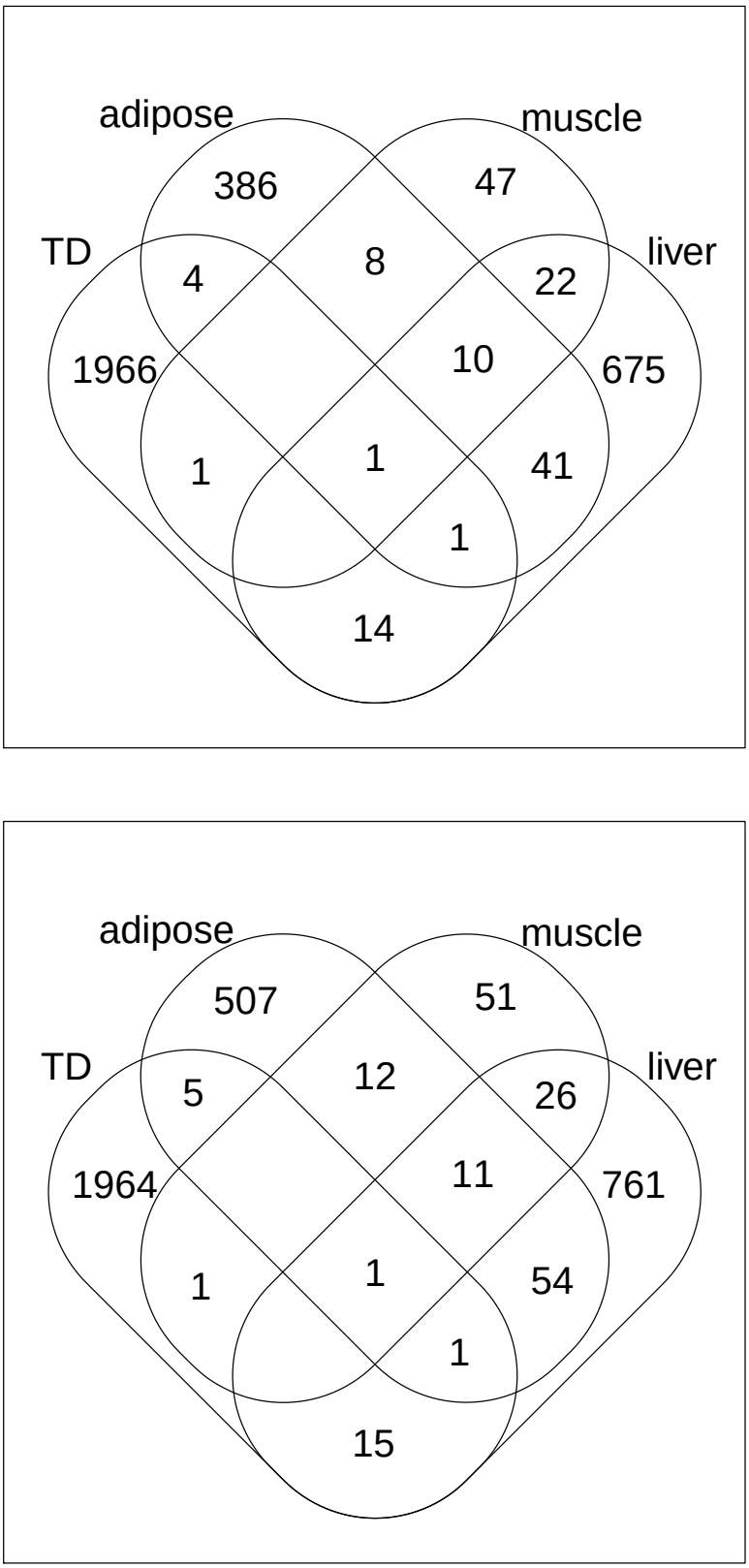

Figure 3. Venn diagrams between genes selected by various methods. Upper: t test, lower: SAM.

Finally, based on the genes associated with probes shown in Table 5, we found that the "KEGG 2021 Human" category in Enrichr does not include neurodegenerative diseases (see

the supplementary materials). Thus, the association between neurodegenerative diseases and diabetes mellitus can be found only when an integrated analysis, such as the proposed method, is employed. In this sense, an integrated analysis is more than a simple union of individual analysis and can identify factors that cannot be identified by individual analyses, such as potentially associated diseases. Thus, an integrated analysis of gene expression profiles in individual tissues provide more in-depth information than individual analyses, at least for certain cases. Thus, integrated analyses of gene expression profiles in individual tissues should be encouraged.

It may be plausible for other integrated methods to perform similarly. If this is true, the advanced methods that we have proposed here are not required. To rule out this possibility, we apply ComBat [18] to remove the batch effect between the three tissue typed, since we selected genes whose expression is independent of tissues as can be seen in Fig. 1; Table 5 shows the results. It is seldom reported to be successful. Limma failed to select any DEGs, and the number of genes selected by the t test and SAM is markedly different from each other in contrast to the identification of tissue-specific DEGs, whose numbers are more coincident across the three methods (Table 5).

Biological validation is also worse; Table 6 shows the result of the "Mouse Gene Atlas". None of tissues used in the experiments are listed whereas the proposed method is (Table 4). In addition to this, based on the genes associated with probes shown in Table 5, we found that the "KEGG 2021 Human" category in Enrichr does not include neurodegenerative diseases (see the supplementary materials) that were detected using the proposed method (Table 2). In conclusion, integrated analysis using ComBat is inferior to the proposed method.

One might wonder why an integrated analysis of three tissues from patients with diabetes mellitus can identify associations with neurodegenerative diseases. The PCA and TD-based unsupervised FE methods are frequently able to detect disease associations. We previously identified an association between cancer and amyotrophic lateral sclerosis [19] without investigating cancer gene expression and an association between heart diseases and posttraumatic stress disorder [20] without investigating brain gene expression. Therefore, we were not surprised that the integrated analysis using the proposed method was able to identify disease associations. To our knowledge, few studies have attempted to predict the association between diseases using gene expression, although many studies have focused on the associations between genes and disease [21–23] and between drugs and disease association [24–26]. Our proposed strategy would be useful for such studies.

5. Conclusions

In this study, we applied the proposed TD-based unsupervised FE with SD optimization method to perform an integrated analysis of gene expression measured in three distinct tissues using microarray architecture; moreover, the proposed method has not been applied to such data in previous studies. The results show that the proposed method can identify more genes than individual analyses. The selected genes are known to be expressed in all three tissues, and they are also enriched in many neurodegenerative diseases that have a known association with diabetes mellitus but cannot be identified by individual analysis. In this sense, integrated analyses might have the ability to identify additional factors relative to individual analyses.

Author Contributions: Y.-h.T. planned the research and performed the analyses. Y.-h.T. and T.T. evaluated the results, discussions, and outcomes and wrote and reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by KAKENHI (Grant Numbers 20H04848, and 20K12067) to Y.-h.T.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: All of the data used in this study are available in GEO ID GSE13268, GSE13269, and GSE13270.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Taguchi, Y.H. *Unsupervised Feature Extraction Applied to Bioinformatics*; Springer International Publishing, 2020. doi:10.1007/978-3-030-22456-1.
2. Taguchi, Y.h.; Turki, T. Tensor decomposition- and principal component analysis-based unsupervised feature extraction to select more reasonable differentially expressed genes: Optimization of standard deviation versus state-of-art methods. *bioRxiv* **2022**, [https://www.biorxiv.org/content/early/2022/02/22/2022.02.18.481115.full.pdf]. doi:10.1101/2022.02.18.481115.
3. Taguchi, Y.H.; Turki, T. Principal component analysis- and tensor decomposition-based unsupervised feature extraction to select more reasonable differentially methylated cytosines: Optimization of standard deviation versus state-of-the-art methods. *bioRxiv* **2022**, [https://www.biorxiv.org/content/early/2022/04/05/2022.04.02.486807.full.pdf]. doi:10.1101/2022.04.02.486807.
4. Roy, S.S.; Taguchi, Y.h. Tensor decomposition and principal component analysis-based unsupervised feature extraction outperforms state-of-the-art methods when applied to histone modification profiles. *bioRxiv* **2022**, [https://www.biorxiv.org/content/early/2022/ doi:10.1101/2022.04.29.490081.
5. Xue, B.; Nie, J.; Wang, X.; DuBois, D.C.; Jusko, W.J.; Almon, R.R. Effects of High Fat Feeding on Adipose Tissue Gene Expression in Diabetic Goto-Kakizaki Rats. *Gene Regulation and Systems Biology* **2015**, *9*, GRSB.S25172, [https://doi.org/10.4137/GRSB.S25172]. PMID: 26309393, doi:10.4137/GRSB.S25172.
6. Huang, D.W.; Sherman, B.T.; Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **2008**, *4*, 44–57. doi:10.1038/nprot.2008.211.
7. Xie, Z.; Bailey, A.; Kuleshov, M.V.; Clarke, D.J.B.; Evangelista, J.E.; Jenkins, S.L.; Lachmann, A.; Wojciechowicz, M.L.; Kropiwnicki, E.; Jagodnik, K.M.; et al. Gene Set Knowledge Discovery with Enrichr. *Current Protocols* **2021**, *1*, e90, [https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/cpz1.90]. doi:https://doi.org/10.1002/cpz1.90.
8. Umegaki, H., Neurodegeneration in Diabetes Mellitus. In *Neurodegenerative Diseases*; Ahmad, S.I., Ed.; Springer US: New York, NY, 2012; pp. 258–265. doi:10.1007/978-1-4614-0653-2_19.
9. Ristow, M. Neurodegenerative disorders associated with diabetes mellitus. *Journal of Molecular Medicine* **2004**, *82*. doi:10.1007/s00109-004-0552-1.
10. Nasrolahi, A.; Mahmoudi, J.; Noori-Zadeh, A.; Haghani, K.; Bakhtiyari, S.; Darabi, S. Shared Pathological Mechanisms Between Diabetes Mellitus and Neurodegenerative Diseases. *Current Pharmacology Reports* **2019**, *5*, 219–231. doi:10.1007/s40495-019-00191-8.
11. Madhusudhanan, J.; Suresh, G.; Devanathan, V. Neurodegeneration in type 2 diabetes: Alzheimer's as a case study. *Brain and Behavior* **2020**, *10*, e01577, [https://onlinelibrary.wiley.com/doi/pdf/10.1002/brb3.1577]. doi:https://doi.org/10.1002/brb3.1577.
12. León, K.I.L.D.; Bertadillo-Jilote, A.D.; García-Gutiérrez, D.G.; Meraz-Ríos, M.A. Alzheimer's Disease and Type 2 Diabetes Mellitus: Molecular Mechanisms and Similarities. In *Neurodegenerative Diseases*; Tunalı, N.E., Ed.; IntechOpen: Rijeka, 2020; chapter 4. doi:10.5772/intechopen.92581.
13. Sun, H.; Wang, Y. A new branch connecting thermogenesis and diabetes. *Nature Metabolism* **2019**, *1*, 845–846. doi:10.1038/s42255-019-0112-1.
14. Lewis, M.T.; Kasper, J.D.; Bazil, J.N.; Frisbee, J.C.; Wiseman, R.W. Quantification of Mitochondrial Oxidative Phosphorylation in Metabolic Disease: Application to Type 2 Diabetes. *International Journal of Molecular Sciences* **2019**, *20*. doi:10.3390/ijms20215271.
15. Holm, L.J.; Mnsted, M.O.; Haupt-Jorgensen, M.; Buschard, K. PPARs and the Development of Type 1 Diabetes. *PPAR Res* **2020**, *2020*, 6198628.
16. Tusher, V.G.; Tibshirani, R.; Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* **2001**, *98*, 5116–5121, [https://www.pnas.org/content/98/9/5116.full.pdf]. doi:10.1073/pnas.091062498.
17. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **2015**, *43*, e47–e47, [https://academic.oup.com/nar/article-pdf/43/7/e47/7207289/gkv007.pdf]. doi:10.1093/nar/gkv007.
18. Johnson, W.E.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **2006**, *8*, 118–127, [https://academic.oup.com/biostatistics/article-pdf/8/1/118/25435561/kxj037.pdf]. doi:10.1093/biostatistics/kxj037.
19. Taguchi, Y.H.; Wang, H. Genetic Association between Amyotrophic Lateral Sclerosis and Cancer. *Genes* **2017**, *8*. doi:10.3390/genes8100243.
20. Taguchi, Y.H.; Iwadate, M.; Umeyama, H. Principal component analysis-based unsupervised feature extraction applied to in silico drug discovery for posttraumatic stress disorder-mediated heart disease. *BMC Bioinformatics* **2015**, *16*. doi:10.1186/s12859-015-0574-4.
21. Babbi, G.; Martelli, P.L.; Profiti, G.; Bovo, S.; Savojardo, C.; Casadio, R. eDGAR: a database of Disease-Gene Associations with annotated Relationships among genes. *BMC Genomics* **2017**, *18*. doi:10.1186/s12864-017-3911-3.

22. Luo, P.; Xiao, Q.; Wei, P.J.; Liao, B.; Wu, F.X. Identifying Disease-Gene Associations With Graph-Regularized Manifold Learning. *Frontiers in Genetics* **2019**, *10*. doi:10.3389/fgene.2019.00270.
23. Opap, K.; Mulder, N. Recent advances in predicting gene?disease associations [version 1; peer review: 2 approved]. *F1000Research* **2017**, *6*. doi:10.12688/f1000research.10788.1.
24. Huang, F.; Qiu, Y.; Li, Q.; Liu, S.; Ni, F. Predicting Drug-Disease Associations via Multi-Task Learning Based on Collective Matrix Factorization. *Frontiers in Bioengineering and Biotechnology* **2020**, *8*. doi:10.3389/fbioe.2020.00218.
25. Jiang, H.; Huang, Y. An effective drug-disease associations prediction model based on graphic representation learning over multi-biomolecular network. *BMC Bioinformatics* **2022**, *23*. doi:10.1186/s12859-021-04553-2.
26. Yu, Z.; Huang, F.; Zhao, X.; Xiao, W.; Zhang, W. Predicting drug–disease associations through layer attention graph convolutional network. *Briefings in Bioinformatics* **2020**, *22*, [https://academic.oup.com/bib/article-pdf/22/4/bbaa243/39135298/bbaa243.pdf]. bbaa243, doi:10.1093/bib/bbaa243.