scientific reports

Check for updates

OPEN

Identification of genes associated with altered gene expression and m6A profiles during hypoxia using tensor decomposition based unsupervised feature extraction

Sanjiban Sekhar Roy[1] & Y.-H. Taguchi[2]✉

Although hypoxia is a critical factor that can drive the progression of various diseases, the mechanism underlying hypoxia itself remains unclear. Recently, m6A has been proposed as an important factor driving hypoxia. Despite successful analyses, potential genes were not selected with statistical significance but were selected based solely on fold changes. Because the number of genes is large while the number of samples is small, it was impossible to select genes using conventional feature selection methods with statistical significance. In this study, we applied the recently proposed principal component analysis (PCA), tensor decomposition (TD), and kernel tensor decomposition (KTD)-based unsupervised feature extraction (FE) to a hypoxia data set. We found that PCA, TD, and KTD-based unsupervised FE could successfully identify a limited number of genes associated with altered gene expression and m6A profiles, as well as the enrichment of hypoxia-related biological terms, with improved statistical significance.

Hypoxia[1], also known as tissue hypoxia, is a serious symptom with various causes. For example, hypoxia could result in death, such as in the case of COVID-19, a serious pandemic[2]. Hypoxia also plays a critical role in cancer[3]. Both brain hypoxia[4] and lung cell hypoxia[5] can be fatal. Despite the significance of hypoxia, the critical factors of hypoxia are not yet fully understood[6]. Recently, m6A was reported to be a newly discovered regulator of hypoxia[7]. Wang et al.[8] found that many genes are simultaneously associated with altered m6A and gene expression profiles in hypoxia. Although the investigations were successful, there was one methodological issue with their study; they selected genes associated with altered m6A and gene expression in hypoxia without determining statistical significance. They selected genes based on fold change (FC). Usually, only using FC to select altered expression or any other measurements might be erroneous because a sufficiently large FC might be observed simply by chance when a large number of candidates are considered. In their analysis, all human genes (as many as a few tens of thousands) and whole genome m6A were considered. In this case, if the FC was not validated statistically, a sufficiently large FC might have been observed simply by chance. The genes associated with altered m6A and gene expression based on statistical significance could not be identified because of the small number of samples; there were only four time points (including the control) measured without any replicates. If we consider the large number of genes as well as m6A peaks in the genome, it is unlikely that four samples are enough to achieve statistical significance; small samples result in larger P-values, whereas a large number of genes and m6A peaks result in relatively larger P-values. In this study, we applied principal component analysis (PCA) and tensor decomposition (TD)-based unsupervised feature extraction (FE) to select genes associated with altered m6A as well as gene expression in hypoxia to determine statistical significance. Enrichment analyses of selected genes are reasonable and consistent with previous findings[8] and can now be supported with statistical significance. Thus, not only were the critical roles of m6A in hypoxia validated but also the usefulness of PCA-and TD-based unsupervised FE in the case where there are very few samples with a large number of variables.

There are a limited number of genomic studies using TD[9,10]. Fang proposed tightly integrated genomic and epigenomic data mining using TD[11] (445 samples for TCGA-OV and 480 samples for TCGA-HNSC), Hore et al applied TD to multi-tissue gene expression experiments[12] (845 related individuals), Ramdhani et al applied TD to

[1]School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India. [2]Department of Physics, Chuo University, Tokyo 112-8551, Japan. ✉email: tag@granular.com

stimulated monocyte and macrophage gene expression profiles[13] (432 samples), Wang et al. applied TD to multi-tissue multi-individual gene expression[14] (544 individuals), Li et al. applied TD to clinical gene-sample-time microarray expression[15] (53 genes and 27 samples), Hu et al. applied TD to gene expression of tumor samples[16] (more than 11,000 tumor samples), Diaz et al. applied TD to genomic data[17] (503 patients), and Bradley et al. applied TD to DNA copy-number alterations[18] (a few hundred samples). All methods other than that used by Li et al. included as few as 53 genes and required as many as several hundred samples, whereas our methods generally require only a few samples (in this study as few as eight samples). To our knowledge, ours is the only method applicable to a data set that includes a few samples with as many as 10^4 genes.

Results

Figure 1 shows the flow chart of analyses performed in this study.

PCA based unsupervised FE applied to gene expression profiles. Because gene expression profiles are measured solely over four time points, it is formatted as a matrix $x_{it} \in \mathbb{R}^{N\times 4}$, and PCA-based unsupervised FE was applied to x_{it}. Figure 2 shows the PC loading attributed to time points. Because the second PC loading is mostly correlated with time, we decided to employ the second PC score, u_{2i}, in order to attribute P-values to genes i. Using Eq. (7) with $\ell = 2$, P_is are attributed to gene i. Then 52 is (genes) with associated corrected P-values less than 0.01, were selected. Table 1 shows the enrichment terms in "KEGG 2019 Human" categories in Enrichr for 52 selected gene symbols. Not all terms are related to hypoxia, whereas a term such as "Oxidative phosphorylation" is known to be related to hypoxia[19]. "Cardiac muscle contraction" is also known to be related to hypoxia[20]. Retrograde endocannabinoid signaling is known to be related to hypoxia[21]. Although three representative neurodegenerative diseases, "Parkinson's disease," "Alzheimer's disease," and "Huntington's disease", are listed, hypoxia is known to be related to neurodegenerative diseases[22]. Glycolysis is also related to hypoxia[23]. Most importantly, HIF-1, a hypoxia-inducible factor, is listed. There are additional identified enrichments that can support the success of PCA-based unsupervised FE. Although they are not always top-ranked, the 52 identified genes are also known to be up/downregulated in independent hypoxia experiments (Table 2). In the "GO Biological Process 2018" category in Enrichr, various glucose/glucogenesis related terms are enriched (Table 3). These results suggested that the analyses performed by PCA-based unsupervised FE were successful.

TD-based unsupervised FE applied to m6A profiles. Although we successfully applied PCA-based unsupervised FE to some gene expression profiles of hypoxia, the identification of the relationship between altered gene expression and hypoxia was not the primary purpose of this study. Instead, its purpose was to identify the relationship between m6A profiles and hypoxia. To identify the relationship between hypoxia and m6A profiles, HOSVD was applied to x_{ktj}, as described in "Materials and methods". The left panel in Fig. 3 shows the singular value vectors attributed to time points; the second singular value vector is most significantly correlated with time. Remarkably, u_{2t} is almost identical to v_{2t}, which is the 2nd PC loading attributed to the time points when PCA was applied to x_{it} (right panel in Fig. 3). Considering that gene expression and m6A profiles are distinct from each other, this coincidence between u_{2t}, which is attributed to m6A profiles, and v_{2t}, which is attributed to gene expression profiles, suggests that our analysis correctly detects the regulatory relationship between m6A and the gene expression profile. In addition to the fact that u_{2t} is most significantly correlated with time points, u_{2j}s have opposite signs between $j = 1$ and $j = 2$ (not shown here), which means that $\ell_3 = 2$ is associated with the distinction between the input and m6A. We then determined which $G(\ell_1, 2, 2)$ has the largest absolute value to determine which $u_{\ell_1 k}$ is used to select genomic regions, k (Table 4). Because it is obvious that $G(2, 2, 2)$ has the largest absolute value, we decided to employ u_{2k} to select genomic regions. P_ks are attributed to k using Eq. (9) with $\ell_1 = 2$. Then 106 ks (genomic regions 25,000 nucleotides in length, see "Materials and methods") associated with corrected P-values less than 0.01, were selected. These 106 genomic regions included 196 unique gene symbols that were uploaded to Enrichr to evaluate enrichment.

In contrast to the 52 genes identified by PCA-based unsupervised FE applied to gene expression, no KEGG pathway terms or GO BP terms were enriched in these 196 gene symbols. Nevertheless, there are some hypoxia experiments in which genes with altered expression are enriched in 196 gene symbols (Table 5). Therefore, even though 196 genes are less biologically significant than the 52 genes identified in the gene expression analysis, they still have some potential to be related to hypoxia.

Integrated analysis of gene expression and m6A profiles using KTD-based unsupervised FE. Since TD-based unsupervised FE applied to m6A profiles was not fully successful, we needed to employ more advanced methodology: Kernel TD-based (KTD) unsupervised FE. HOSVD was applied to $x_{tjt'j'}$, as described in "Materials and methods". Figure 4 shows the results that are consistent with the results obtained by non-integrated analysis (Figs. 2, 3). The second singular value vectors, u_{2t} and $u_{2t'}$, are most consistent with time points; it is coincident with the second PC loading attributed to gene expression, and the second singular value vectors attributed to m6A are coincident with time points. u_{2t} and $u_{2t'}$ are also identical; it is coincident with the second PC loading attributed to gene expression, and the second singular value vectors attributed to m6A are identical. In addition, u_{2j} has the opposite sign between $j = 1$ (control) and $j = 2$ (m6A). Then $u_{\ell_1 i}$ was computed using Eq. (16), with $\ell_1 = 2$, and $u_{\ell_2 \ell_3 k}$ was computed using Eq. (17), with $\ell_2 = \ell_3 = 2$. P_i and P_k are attributed to i and k, respectively, with Eqs. (18) and (19), respectively. The 53 is (genes) and 128 ks (genome regions) associated with adjusted P-values less than 0.01 were selected. Two hundred gene symbols were retrieved from 128 genomic regions, as previously described.

The 53 and 200 gene symbols were uploaded to Enrichr. Table 6 shows the results of the "KEGG 2019 Human" category in Enrichr. When compared with Table 1, the enrichment for gene expression profiles is similar. Four

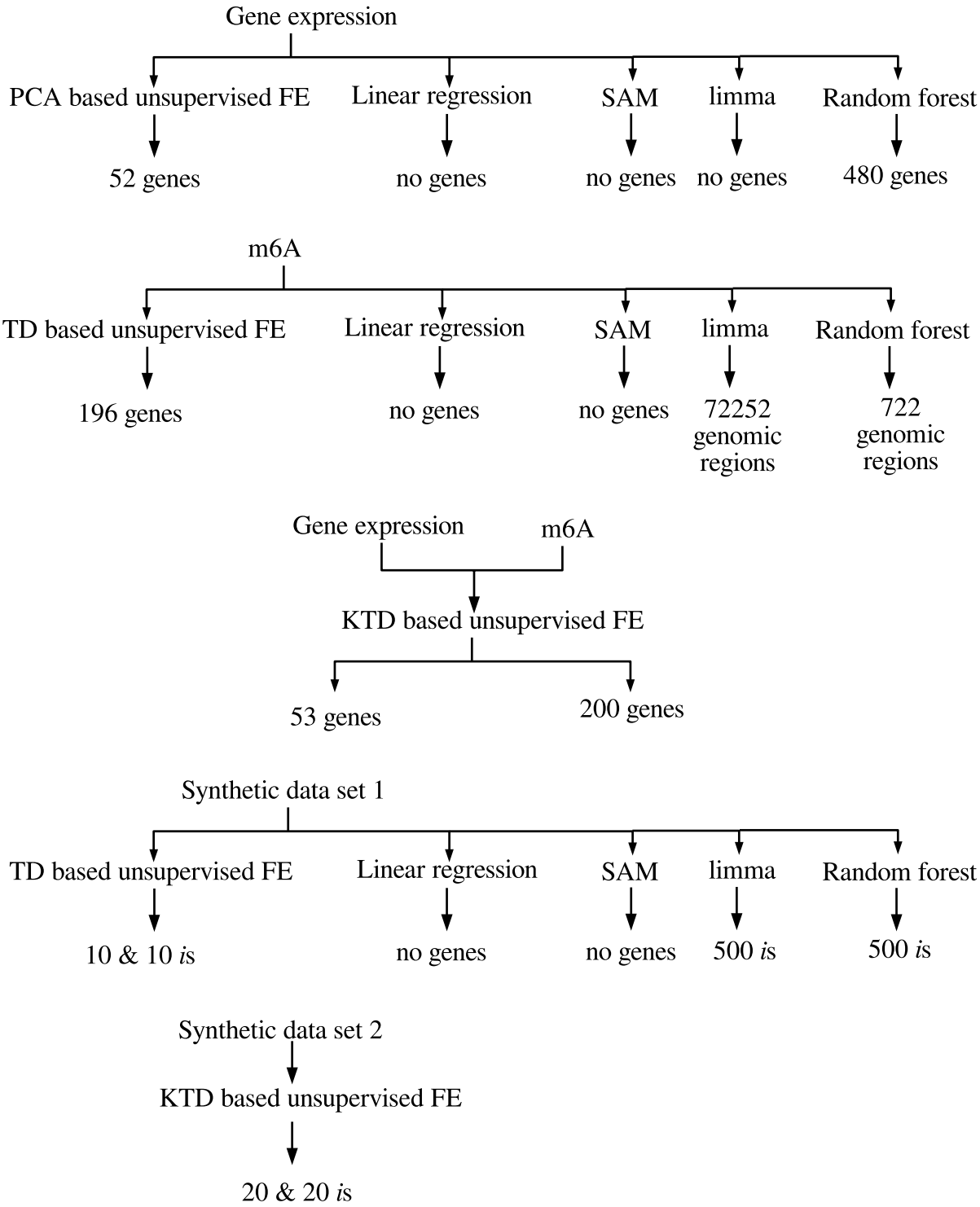

Figure 1. Flow chart of analyses performed in this study.

enrichment terms were identified, whereas no terms were identified when TD-based unsupervised FE was applied to the m6A profile.

Table 7 shows the results of the "Disease Perturbations from GEO down/up" category in Enrichr. Compared with Tables 2 and 5, although enrichment in the "Disease Perturbations from GEO down" for m6A is missing, it still has enrichment for both gene expression and m6A profiles. Table 8 shows the enrichment in the "GO Biological Process 2018" category of Enrichr. Compared with Table 3, enrichment for gene expression does not change, and enrichment for m6A is identified, whereas it was not identified when TD-based unsupervised FE was applied to the m6A profile. Thus, KTD-based unsupervised FE improved the enrichment of m6A profiles without affecting the enrichment for gene expression.

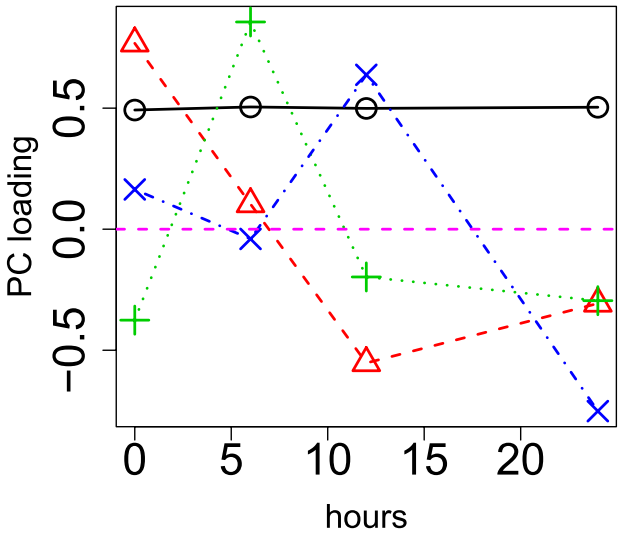

Figure 2. PC loading, $v_{\ell t}$, computed by PCA applied to time points by applying PCA to gene expression profiles. Open black circles: 1st, (0.61) open red triangles: 2nd (− 0.76), green crosses: 3rd (− 0.24), blue crosses: 4th PC loading (− 0.60). The numbers in parentheses are the Pearson's correlation coefficients. Hours (horizontal axis) represent the duration after the treatments. The horizontal magenta broken line indicates baseline (zero).

Term	Overlap	*P*-value	Adjusted *P*-value
Ribosome	19/153	1.20×10^{-27}	3.68×10^{-25}
Thermogenesis	13/231	1.98×10^{-14}	3.05×10^{-12}
Oxidative phosphorylation	11/133	3.54×10^{-14}	3.63×10^{-12}
Parkinson disease	11/142	7.35×10^{-14}	5.66×10^{-12}
Glycolysis/gluconeogenesis	8/68	7.80×10^{-12}	4.80×10^{-10}
HIF-1 signaling pathway	6/100	2.27×10^{-7}	1.16×10^{-5}
Alzheimer disease	7/171	2.86×10^{-7}	1.26×10^{-5}
Huntington disease	6/193	1.05×10^{-5}	4.05×10^{-4}
Retrograde endocannabinoid signaling	5/148	4.07×10^{-5}	1.39×10^{-3}
Cardiac muscle contraction	4/78	5.03×10^{-5}	1.55×10^{-3}
Non-alcoholic fatty liver disease (NAFLD)	4/149	6.07×10^{-4}	1.70×10^{-2}

Table 1. "KEGG 2019 Human" category of Enrichr for 52 genes selected by applying PCA-based unsupervised FE to gene expression profiles. Eleven terms with adjusted *P*-values less than 0.05 are listed.

Term	Overlap	*P*-value	Adjusted *P*-value
Disease perturbations from GEO down			
Hypoxia C0242184 human GSE4630 sample 250	13/349	3.83×10^{-12}	2.19×10^{-11}
Hypoxia C0242184 mouse GSE3195 sample 70	10/328	1.05×10^{-8}	3.77×10^{-8}
Hypoxia C0242184 human GSE4483 sample 440	9/275	3.38×10^{-8}	1.11×10^{-7}
Disease perturbations from GEO up			
Hypoxia C0242184 human GSE4483 sample 440	27/325	5.48×10^{-35}	9.19×10^{-33}
Hypoxia C0242184 human GSE4630 sample 250	10/251	8.07×10^{-10}	2.38×10^{-9}
Hypoxia C0242184 mouse GSE3195 sample 70	10/272	1.76×10^{-9}	5.01×10^{-9}

Table 2. "Disease Perturbations from GEO down/up" category of Enrichr for 52 genes selected by applying PCA-based unsupervised FE to gene expression profiles. Six hypoxia experiments with adjusted *P*-values less than 0.05 are listed.

Term	Overlap	*P*-value	Adjusted *P*-value
Canonical glycolysis (GO:0061621)	7/25	2.45×10^{-13}	6.57×10^{-11}
Glycolytic process through glucose-6-phosphate (GO:0061620)	7/25	2.45×10^{-13}	6.93×10^{-11}
Glucose catabolic process to pyruvate (GO:0061718)	7/25	2.45×10^{-13}	6.24×10^{-11}
Gluconeogenesis (GO:0006094)	6/41	9.62×10^{-10}	2.04×10^{-7}
Glycolytic process (GO:0006096)	5/23	3.17×10^{-9}	6.22×10^{-7}
Glucose metabolic process (GO:0006006)	6/64	1.53×10^{-8}	2.79×10^{-6}

Table 3. "GO Biological Process 2018" category of Enrichr for 52 genes selected by applying PCA-based unsupervised FE to gene expression profiles. Six glucose-related terms with adjusted *P*-values less than 0.05 are listed.

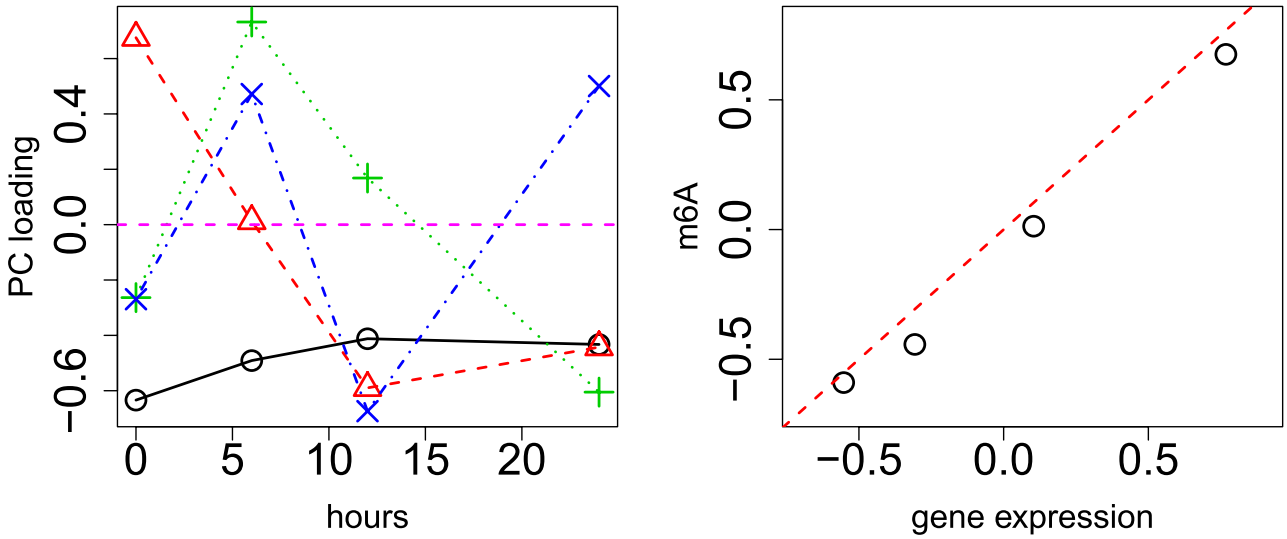

Figure 3. Left: Singular value vector, $u_{\ell_2 t}$, computed by HOSVD applied to time points with applying HOSVD to m6A profiles. Open black circles: 1st (0.78). Open red triangles: 2nd (− 0.80), green crosses: 3rd (− 0.47), blue crosses: 4th PC loading (0.36). The numbers in parentheses are the Pearson's correlation coefficients. Hours (horizontal axis) represent the duration after the treatments. Horizontal magenta broken line indicates baseline (zero). Right: scatter plot between v_{2t}, which is the 2nd PC loading attributed to time points when PCA was applied to x_{it}, (horizontal axis) and u_{2t} (vertical axis). The red broken line indicates $v_{2t} = u_{2t}$.

ℓ_1	$G(\ell_1, 2, 2)$
1	-6.28×10^4
2	5.89×10^5
3	-5.47×10^4
4	9.27×10^4
5	-2.24×10^5
6	7.59×10^4
7	-6.23×10^4
8	3.40×10^4
9	0.0
10	-1.15×10^{-14}

Table 4. $G(\ell_1, 2, 2)$ computed by applying HOSVD to x_{ktj}.

Additional improvement from PCA- and TD-based unsupervised FE to integrated analysis using KTD-based unsupervised FE identified significant associations between gene expression and m6A (Table 9). When PCA-and TD-based unsupervised FE were separately applied to gene expression and m6A profiles, 52 and 196 genes were identified, respectively. The number of common genes between them was seven. However, integrated analysis of gene expression and m6A profiles with TKD-based unsupervised FE identified 53 genes for gene expression

Term	Overlap	*P*-value	Adjusted *P*-value
Disease perturbations from GEO down			
Hypoxia C0242184 mouse GSE3195 sample 70	12/328	$9,99 \times 10^{-5}$	$9,86 \times 10^{-4}$
Hypoxia C0242184 human GSE4483 sample 440	10/275	3.98×10^{-4}	2.79×10^{-3}
Hypoxia C0242184 human GSE4630 sample 250	10/349	2.41×10^{-3}	1.22×10^{-2}
Disease perturbations from GEO up			
Hypoxia C0242184 human GSE4483 sample 440	22/325	1.17×10^{-12}	4.92×10^{-10}
Hypoxia C0242184 human GSE4630 sample 250	15/251	2.82×10^{-8}	1.18×10^{-6}
Hypoxia C0242184 mouse GSE3195 sample 70	14/272	5.15×10^{-7}	9.39×10^{-6}

Table 5. "Disease Perturbations from GEO down/up" category of Enrichr for 196 genes selected by applying TD-based unsupervised FE to m6A profiles. Six hypoxia experiments with adjusted *P*-values less than 0.05 are listed.

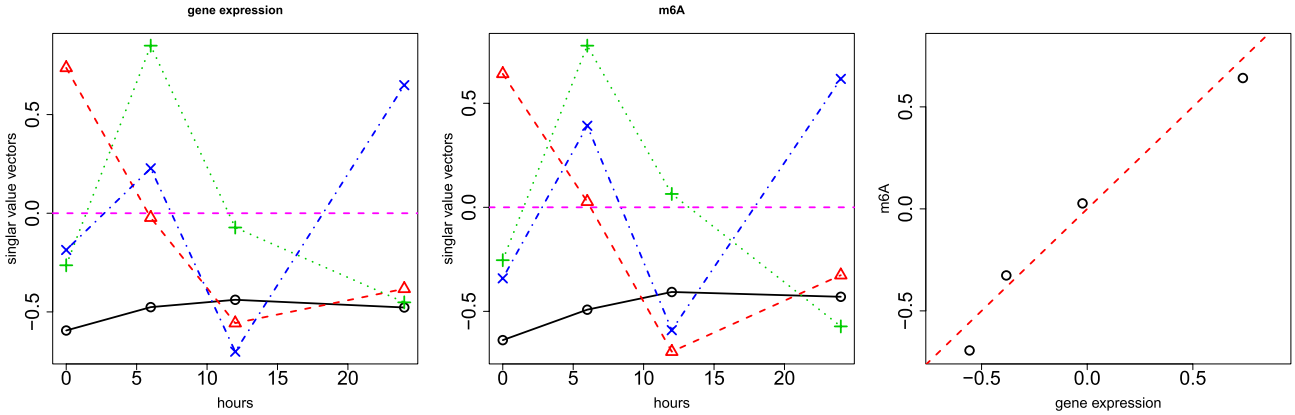

Figure 4. Left: Singular value vector, $u_{\ell_1 t}$, computed by HOSVD applied to $x_{tjt'j'}$. Open black circles: 1st, (0.61) open red triangles: 2nd (− 0.77), green crosses: 3rd (− 0.41), blue crosses: 4th PC loading (0.48). The numbers in parentheses are the Pearson's correlation coefficients. Hours (horizontal axis) represent the duration after the treatments. Horizontal magenta broken line indicates baseline (zero). Middle: singular value vector, $u_{\ell_3 t'}$, computed by HOSVD applied to $x_{tjt'j'}$. Open black circles: 1st (0.78). open red triangles: 2nd (− 0.70), green crosses: 3rd (− 0.47), blue crosses: 4th PC loading (0.52). The numbers in parentheses are the Pearson's correlation coefficients. Hours (horizontal axis) represent the duration after the treatments. Horizontal magenta broken line indicates baseline (zero). Right: scatter plot between u_{2t} (horizontal axis) and $u_{2t'}$ (vertical axis). The red broken line indicates $u_{2t} = u_{2t'}$.

and 200 genes for m6A profiles. The number of common genes increased to 12. Although we cannot estimate their significance very accurately, if we can tentatively assume that there are 20,000 human genes in total, both coincidences are significant, and coincidence in KTD-based unsupervised FE is more significant.

In conclusion, integrated analysis of gene expression and m6A profiles using KTD-based unsupervised FE substantially increased the results compared with applying PCA- and TD-based unsupervised FE separately to gene expression and m6A profiles, respectively. KTD-based unsupervised FE could identify the relationship of gene expression and m6A with hypoxia simultaneously for the first time in a statistically significant manner.

Comparisons with other conventional methods. Although we have shown that integrated analysis of gene expression and m6A profiles simultaneously identified the relationship between hypoxia and gene expression as well as hypoxia and m6A profiles, if other simpler conventional methods can achieve similar performances, it is useless to employ complicated methods such as KTD-based unsupervised FE. To confirm that other conventional methods cannot identify similar relationships, we applied a few conventional feature selection methods. As can be seen in the following text, no conventional feature selections were found to be useful.

Linear regression analysis. Linear regressions were applied to the gene expression profiles, x_{it}, and m6A profiles, x_{ktj}. No genes or genomic regions was associated with adjusted *P*-values less than 0.05, respectively; thus, no genes or genomic regions were associated with adjusted *P*-values less than 0.01 either.

SAM. Although we tried to apply SAM to gene expression profiles, x_{it}, and m6A profiles, x_{ktj}, we found that SAM requires at least two replicates for each class. In this study, there are no replicated classes in four classes in x_{it} or eight classes in x_{ktj}; therefore, we could not apply SAM to these data sets.

Term	Overlap	*P*-value	Adjusted *P*-value
Gene expression			
Ribosome	20/153	2.13×10^{-29}	6.55×10^{-27}
Thermogenesis	13/231	2.59×10^{-14}	3.99×10^{-12}
Oxidative phosphorylation	11/133	4.44×10^{-14}	4.56×10^{-12}
Parkinson disease	11/142	9.22×10^{-14}	7.10×10^{-12}
Glycolysis/gluconeogenesis	8/68	9.16×10^{-12}	5.64×10^{-10}
HIF-1 signaling pathway	6/100	2.55×10^{-7}	1.31×10^{-5}
Alzheimer disease	7/171	3.27×10^{-7}	1.44×10^{-5}
Huntington disease	6/193	1.18×10^{-5}	4.53×10^{-4}
Retrograde endocannabinoid signaling	5/148	4.47×10^{-5}	1.53×10^{-3}
Cardiac muscle contraction	4/78	5.42×10^{-5}	1.67×10^{-3}
Non-alcoholic fatty liver disease (NAFLD)	4/149	6.52×10^{-4}	1.83×10^{-2}
m6A			
Glycolysis/gluconeogenesis	7/68	5.21×10^{-6}	1.60×10^{-3}
Central carbon metabolism in cancer	6/65	4.69×10^{-5}	7.22×10^{-3}
HIF-1 signaling pathway	6/100	5.07×10^{-4}	5.21×10^{-2}
Glucagon signaling pathway	6/103	5.93×10^{-4}	4.57×10^{-2}

Table 6. "KEGG 2019 Human" category of Enrichr for 53 genes (based on gene expression) and 200 genes (based on m6A) selected by applying KTD-based unsupervised FE to integration of gene expression and m6A profile. Fifteen terms with adjusted *P*-values less than 0.05 are listed.

Term	Overlap	*P*-value	Adjusted *P*-value
Gene expression			
Disease perturbations from GEO down			
Hypoxia C0242184 human GSE4630 sample 250	15/349	1.08×10^{-14}	7.85×10^{-14}
Hypoxia C0242184 mouse GSE3195 sample 70	12/328	4.53×10^{-11}	2.03×10^{-10}
Hypoxia C0242184 human GSE4483 sample 440	10/275	2.38×10^{-9}	8.67×10^{-9}
Disease perturbations from GEO up			
Hypoxia C0242184 human GSE4483 sample 440	27/325	1.10×10^{-34}	1.15×10^{-32}
Hypoxia C0242184 mouse GSE3195 sample 70	10/272	2.14×10^{-9}	5.79×10^{-9}
Hypoxia C0242184 human GSE4630 sample 250	9/251	1.83×10^{-8}	4.51×10^{-8}
m6A			
Disease perturbations from GEO up			
Hypoxia C0242184 human GSE4483 sample 440	30/325	1.91×10^{-20}	1.60×10^{-17}
Hypoxia C0242184 mouse GSE3195 sample 70	21/272	4.63×10^{-13}	1.94×10^{-10}
Hypoxia C0242184 human GSE4630 sample 250	17/251	6.56×10^{-10}	4.24×10^{-8}

Table 7. "Disease Perturbations from GEO down/up" category of Enrichr for 53 genes (based on gene expression) and 200 genes (based on m6A) selected by applying KTD-based unsupervised FE to integration of gene expression and m6A profiles. Nine hypoxia experiments with adjusted *P*-values less than 0.05 are listed.

Limma. Limma was applied to the gene expression profiles, x_{it}, and m6A profiles, x_{ktj}. No genes but 72252 genomic regions were associated with adjusted *P*-values of less than 0.01, respectively. Therefore, limma was not useful.

Random forest. Random forest was applied to gene expression profiles, x_{it}, and m6A profiles, x_{ktj}. Four hundred and eighty genes and 722 genomic regions had non-zero importance, respectively. Thus, random forest successfully selected a reasonable number of genes and genomic regions. Nevertheless, no hypoxia-related biological terms were enriched in the 480 genes or gene symbols included in the 722 genomic regions. Thus, random forest was not a useful method.

Term	Overlap	*P*-value	Adjusted *P*-value
Gene expression			
Glycolytic process through glucose-6-phosphate (GO:0061620)	7/25	2.82×10^{-13}	7.56×10^{-11}
Glucose catabolic process to pyruvate (GO:0061718)	7/25	2.82×10^{-13}	6.84×10^{-11}
Canonical glycolysis (GO:0061621)	7/25	2.82×10^{-13}	7.18×10^{-11}
Gluconeogenesis (GO:0006094)	6/41	1.08×10^{-9}	2.40×10^{-7}
Glycolytic process (GO:0006096)	5/23	3.49×10^{-9}	7.13×10^{-7}
Glucose metabolic process (GO:0006006)	6/64	1.72×10^{-8}	3.14×10^{-6}
m6A			
Glycolytic process through glucose-6-phosphate (GO:0061620)	5/25	4.29×10^{-6}	2.19×10^{-2}
Glucose catabolic process to pyruvate (GO:0061718)	5/25	4.29×10^{-6}	7.30×10^{-3}
Canonical glycolysis (GO:0061621)	5/25	4.29×10^{-6}	1.10×10^{-2}

Table 8. "GO Biological Process 2018" category of Enrichr for 53 genes (based on gene expression) and 200 genes (based on m6A) selected by applying KTD-based unsupervised FE to integration of gene expression and m6A profiles. Nine glucose-related terms with adjusted *P*-values less than 0.05 are listed.

		PCA-and TD-based unsupervised FE		KTD-based unsupervised FE	
m6A	Gene expression	Not selected	Selected	Not selected	Selected
	Not selected	19773	45	19783	41
	Selected	189	7	188	12
	Odds ratio	16.26		30.77	
	P-value	7.15×10^{-7}		1.32×10^{-13}	

Table 9. Confusion matrices of selected genes between gene expression and m6A profiles. PCA-and TD-based unsupervised FE were separately applied to gene expression and m6A profiles, or KTD-based unsupervised FE was applied to the integration of gene expression and m6A profiles

Discussion

One might wonder why linear regression, SAM, limma, and random forest failed to select genes associated with altered gene expression, genomic regions associated with m6A profiles in hypoxia, or genes biologically related to hypoxia. This is because it is a very difficult problem. There are more than 17140 genes as well as 123817 genomic regions, whereas the number of samples measured was four and eight, respectively, which were too small to obtain sufficiently significant *P*-values. These numbers of genes and genomic regions were too large to obtain significant *P*-values; although random forest is free from *P*-values, too small sample numbers often prevent random forest from obtaining results that are not obtainable by chance.

To demonstrate how the KTD-based unsupervised FE outperforms the other four methods, we applied them to two synthetic data sets with $N = 1000$ and $N_1 = 10$. When linear regression was applied to the 1st synthetic data set, there were no *i*s associated with adjusted *P*-values less than 0.01. When limma and random forest were applied to the 1st synthetic data set, there were as many as 500 *i*s associated with adjusted *P*-values less than 0.01. Thus, neither linear regression nor limma was useful.

The problem is when time points are regarded as a quantitative property (i.e., in linear regression); in this case, eight samples were too small to give significant *P*-values because there are 1000 features on which *P*-values must be corrected by considering multiple comparison criteria. However, if they are classified into eight classes with one replicate, too many *i*s were regarded as distinct between eight classes because those values that are constant between any pairs of eight classes are unlikely to be fulfilled when using a null hypothesis.

However, when TD-based unsupervised FE was applied to the first synthetic data set, the situation was very different. Figures 5 and 6 show the two combinations of the $u_{\ell_2 j}$, $u_{\ell_3 k}$, and $G(\ell_1\ell_2\ell_3)$. Figure 5 ($\ell_2 = 2, \ell_3 = 1$) corresponds to $x_{ijk}, i \le N_1$ since $x_{ij1} = x_{ij2}$ is coincident with $u_{\ell_3 1} = u_{\ell_3 2}$ (see Fig. 7). However, Fig. 6 ($\ell_2 = \ell_3 = 2$) corresponds to $x_{ijk}, N_1 < i \le 2N_1$ because $x_{ij1} = -x_{ij2}$, which coincides with $u_{\ell_3 1} = -u_{\ell_3 2}$ (see Fig. 7). Thus, in contrast to other supervised methods, TD-based unsupervised FE can detect the distinction between $x_{ij1} = x_{ij2}$ and $x_{ij1} = -x_{ij2}$. To determine whether TD-based unsupervised FE can correctly identify x_{ijk}s, we need to attribute *P*-values to *i*s. Because $|G(\ell_1, 2, 1)|$ and $|G(\ell_1, 2, 2)|$ have the largest values when $\ell_1 = 3$ and $\ell_1 = 2$, respectively, we decided to attribute *P*-values to *i*s using Eq. (13) by assigning $\ell_1 = 3$ and $\ell_1 = 2$, respectively. Computed *P*-values were corrected, and *i*s associated with adjusted *P*- values less than 0.01 were selected. Table 10 shows the performance of TD-based unsupervised FE applied to the first synthetic data set. It perfectly selects *i* coincident with Eq. (1). As a result, the reason why only PCA and TD-based unsupervised FE could select a reasonable number of genes, while other methods failed, is because PCA and TD-based

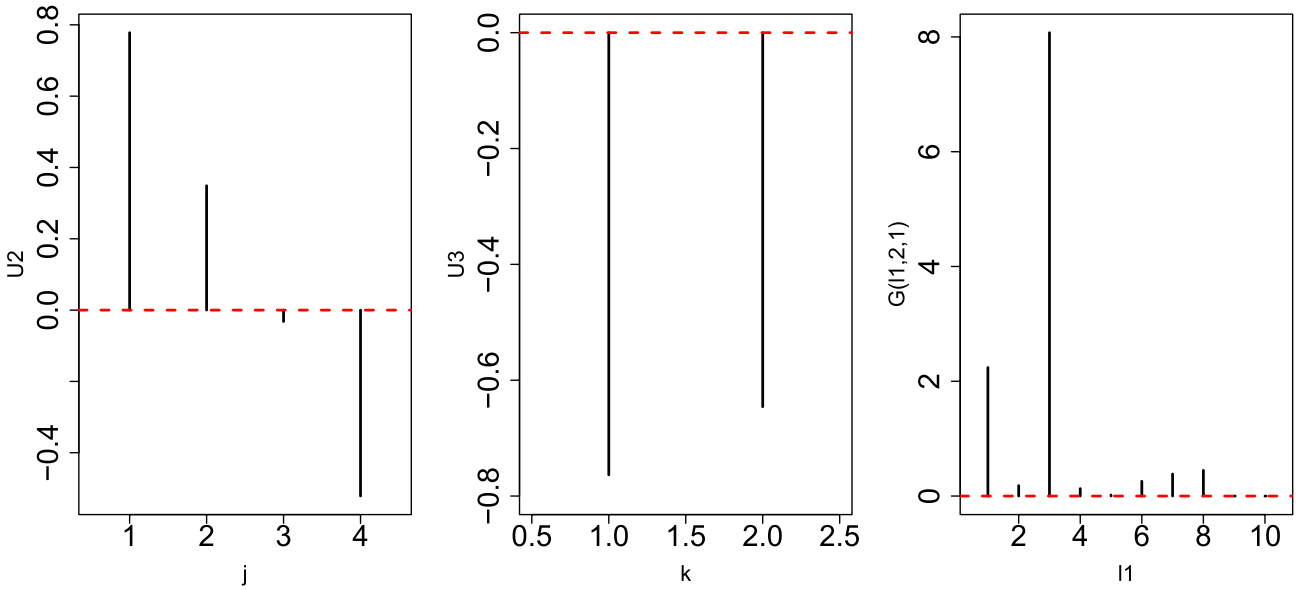

Figure 5. Left: Singular value vector, u_{2j}, computed by HOSVD applied to the 1st synthetic data, x_{ijk}. Middle: Singular value vector, u_{1k}, computed by HOSVD applied to 1st synthetic data, x_{ijk}. Right: $|G(\ell_1, 2, 1)|$. In all three sub-panels, the horizontal red broken lines indicate the baseline (zero).

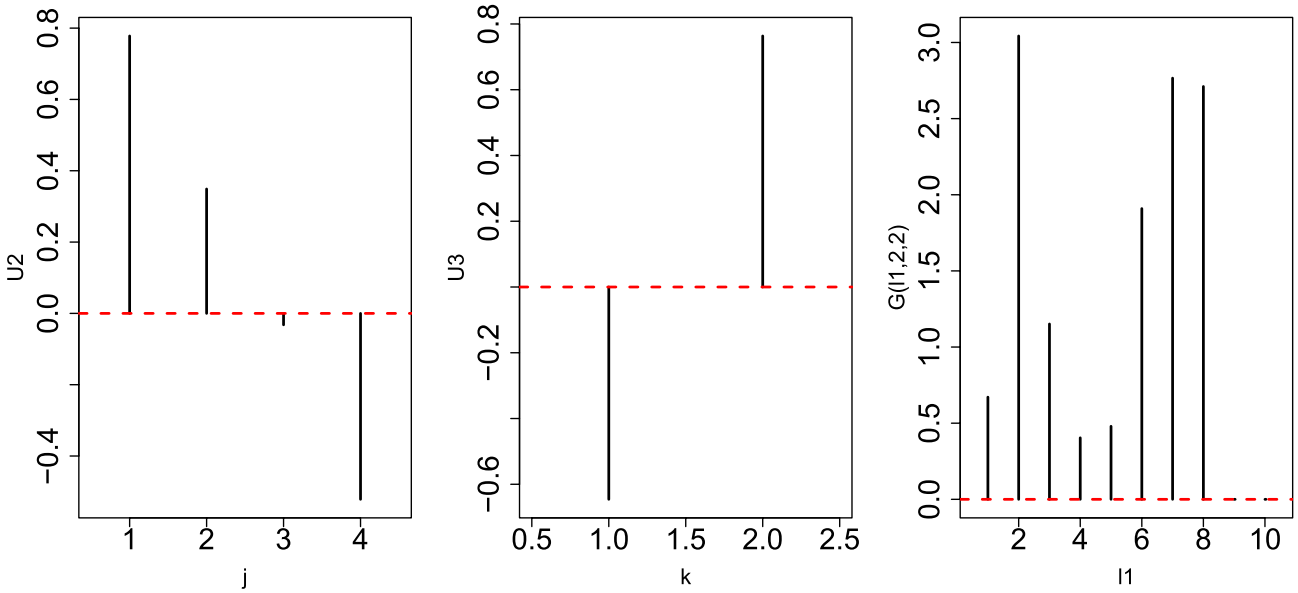

Figure 6. Left: Singular value vector, u_{2j}, computed by HOSVD applied to the 1st synthetic data, x_{ijk}. Middle: singular value vector, u_{2k}, computed by HOSVD applied to 1st synthetic data, x_{ijk}. Right: $|G(\ell_1, 2, 2)|$. In all three sub-panels, the horizontal red broken lines indicate the baseline (zero).

unsupervised FE are suitable for situations where there are a very small number of samples with a large number of features (observations).

Next, we attempted to demonstrate how KTD-based unsupervised FE integrates two data sets in order to identify common features between the two. Figures 8 and 9 show singular value vectors obtained when KTD-based unsupervised FE was applied to $x_{jk_1k_2j'k'_1k'_2}$.

Figure 8 shows u_{1j}, u_{1k_1}, $u_{1j'}$, and $u_{1k'_1}$, which is coincident with x_{ijk}, $i \leq N_1$, and $2N_1 < i \leq 3N_1$ and x'_{ijk}, $i \leq N_1$, and $4N_1 < i \leq 5N_1$, since both singular value vectors and x_{ijk} and x'_{ijk} for these is increase as j increases and do not depend on k (see Fig. 10). To determine whether we can select these x_{ijk} and x'_{ijk} in these is, we reproduced singular value vectors attributed to is using Eq. (22) with $\ell_1 = \ell_2 = 1$ or Eq. (23) with $\ell_4 = \ell_5 = 1$. Then P-values were attributed to is using Eq. (24) or Eq. (25). Computed P-values were corrected, and is associated with adjusted P-values less than 0.01 were selected. Table 11 shows a perfect performance.

Figure 9 shows u_{1j}, u_{2k_1}, $u_{1j'}$, and $u_{2k'_1}$, which is coincident with x_{ijk}, $N_1 < i \leq 2N_1$, and $3N_1 < i \leq 4N_1$, and x'_{ijk}, $N_1 < i \leq 2N_1$, and $5N_1 < i \leq 6N_1$, since u_{1j} and $u_{1j'}$ increase as j increases, and u_{2k_1} and $u_{2k'_1}$ have opposite

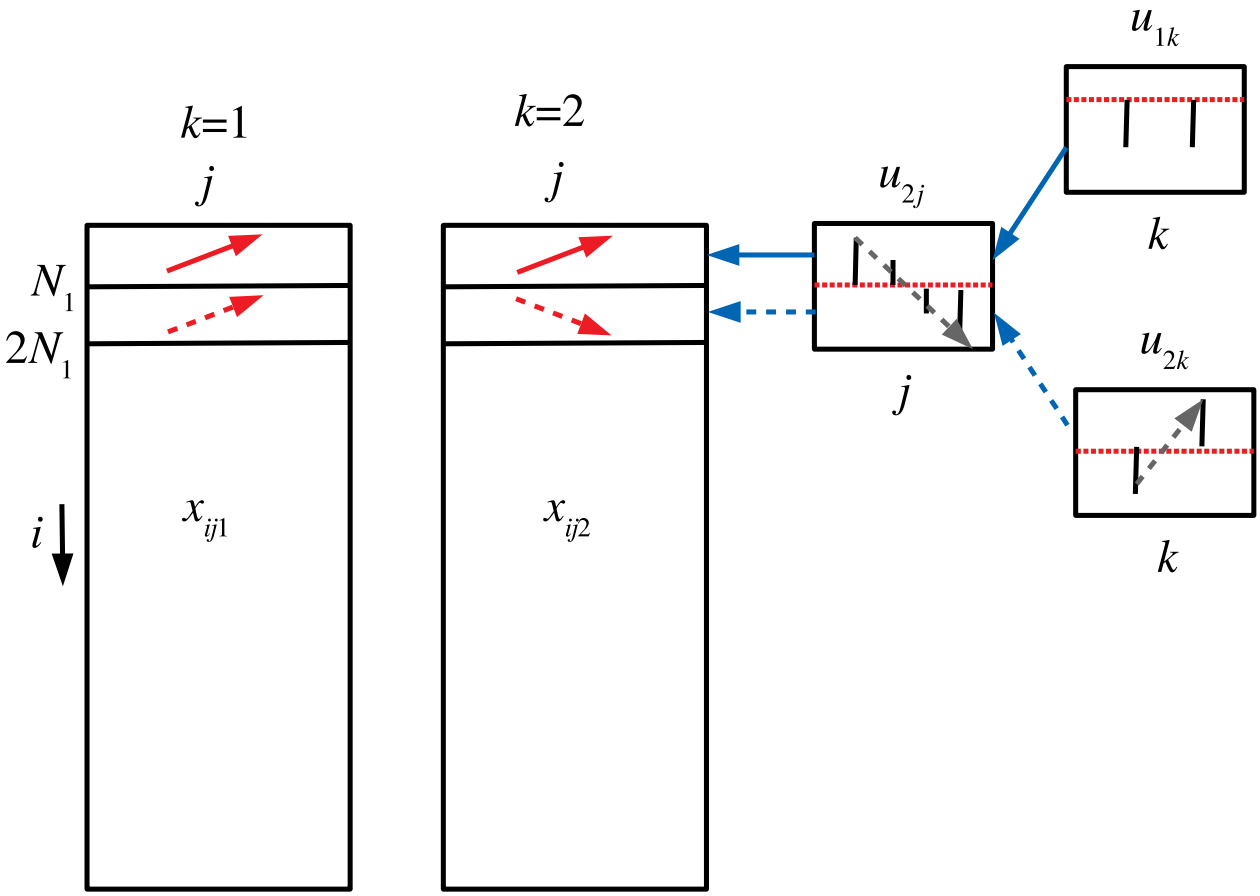

Figure 7. Schematic figure that explains the correspondence between $x_{ijk}, i \leq N_1$ (red solid arrows) and u_{2j}, u_{1k} and that between $x_{ijk}, N_1 < i \leq 2N_1$ (red broken arrows) and u_{2j} and u_{2k}, respectively.

TD based unsupervised FE	$\ell_1 = 3$			$\ell_1 = 2$	
Adjusted *P*-values	> 0.01	≤ 0.01		> 0.01	≤ 0.01
$i \notin [1, N_1]$	990	0	$i \notin [N_1 + 1, 2N_1]$	990	0
$i \in [1, N_1]$	0	10	$i \in [N_1 + 1, 2N_1]$	0	10

Table 10. Confusion matrices of selected is between true and those selected by TD-based unsupervised FE applied to synthetic data set 1.

signs between $k_1 = k'_1 = 1$ and $k_2 = k'_2 = 2$, while x_{ij1} and x'_{ijk1} for these *i*s increase as *j* increases, and x_{ij2} and x'_{ij2} for these *i*s decrease as *j* increases (see Fig. 10). To determine whether we can select these x_{ijk} and x'_{ijk} in these *i*s, we reproduced singular value vectors attributed to *i*s using Eq. (22) with $\ell_1 = 1, \ell_2 = 2$ or Eq. (23) with $\ell_4 = 1, \ell_5 = 2$. Then, *P*-values were attributed to *i*s using Eq. (24) or Eq. (25). Computed *P*-values were corrected, and *i*s associated with adjusted *P*-values less than 0.01 were selected. Table 11 shows a perfect performance.

Table 12 shows the confusion matrix between *i*s selected for x_{ijk} and those selected for x'_{ijk}. This corresponds to Table 9, where genes were selected based on gene expression and m6A profiles. This might be the reason why KTD-based unsupervised FE could identify a significantly overlapping set of genes between gene expression and m6A profiles.

Although TD- and KTD-based unsupervised FE can outperform conventional supervised methods when applied to a small number of samples with a large number of features, TD- and KTD-based unsupervised FE have yet another advantage: *j* dependence is not monotonic (see the open red triangles in Figs. 2, 3, and 4). Such a non-linear dependence on *j* cannot be assumed by supervised methods in advance. Wrongly assumed *j* dependence results in decreased feature selection performance. This is another reason why PCA-, TD-, and KTD-based unsupervised FE can outperform other conventional supervised feature selection methods.

In order to see if our findings are robust, we tried to find alternative data sets in which gene expression and m6A were simultaneously measured for hypoxia, but we could not find any such data sets. Thus, we employed GSE120860, in which only m6A was measured. TD-based unsupervised FE applied to these data sets gave us 54 genes associated with adjusted *P*-values less than 0.01 ($\ell_2 = \ell_3 = 1$ and $\ell_4 = 2$ were selected, and u_{4i} was used

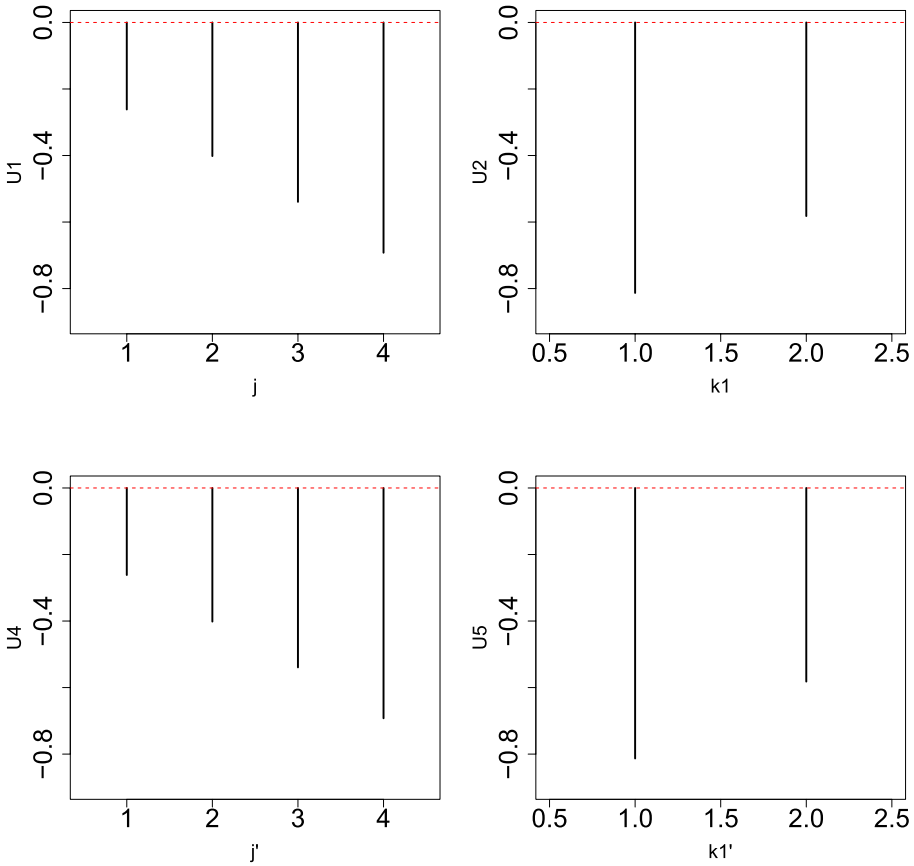

Figure 8. Singular value vectors obtained when KTD-based unsupervised FE was applied to the 2nd synthetic data set. Upper left: u_{1j}, upper right: u_{1k_1}, lower left: $u_{1j'}$, lower right: $u_{1k_1'}$. The horizontal red broken line indicates baseline (zero).

to attribute *P*-values to gene *i* with Eq. (11) because $G(4, 1, 1, 2)$ has the largest absolute value given $\ell_2 = \ell_3 = 1$ and $\ell_4 = 2$). Uploading these genes to Enrichr did not identify any terms associated with both hypoxia and significant *P*-values. As shown in Table 5, when genes were selected by m6A only, there were fewer significant terms. Therefore, we may need to have alternative data sets associated with both gene expression and m6A simultaneously in order to validate the robustness of our results.

One might wonder why we did not compare the proposed methods with conventional unsupervised methods using PCA and TD but only with supervised methods. The reasons for this are as follows. Although there are many papers whose titles include "Feature selection using principal component analysis", feature selections in these papers mean selecting limited numbers of latent vectors generated by PCA or TD. Thus, they are not applicable to the present study, which needed to select not generated features but original ones (i.e., genomic regions). Although there are a few studies that aim to select original features, and not generated latent vectors, they did not attribute *P*-values to the features, which would have allowed us to evaluate the significance of the feature selections. For example, Song et al.[24] selected a limited number of original features associated with relatively larger absolute values of eigenvectors, and no *P*-values were attributed to the individual original features. The purpose of the present study is not simply to select features but to evaluate the significance of selected features; as denoted above, Song et al's study could not help us to evaluate the significance of the feature selections. This is why we did not compare our method with other unsupervised methods using PCA or TD but compared ours with the supervised methods that could give us *P*-values, by which we could evaluate the significance of the feature selections.

In this study, we applied PCA-, TD-, and KTD-based unsupervised FE to gene expression and m6A profiles in hypoxia. Although these methods identified a limited number of genes significantly related to hypoxia, other conventional methods failed. To understand why PCA-, TD-, and KTD-based unsupervised FE could outperform other conventional methods, we applied these methods to synthetic data sets with small numbers of samples and large numbers of features. As a result, we successfully reproduced the superior performance of TD-and KTD-based unsupervised FE over other conventional methods. Thus, the superiority of PCA-, TD-, and KTD-based unsupervised FE is possibly due to having a small number of samples with a large number of features. In conclusion, despite the limitations of previous studies, we validated a set of genes associated with altered gene expression and m6A profiles in hypoxia in a statistically significant manner.

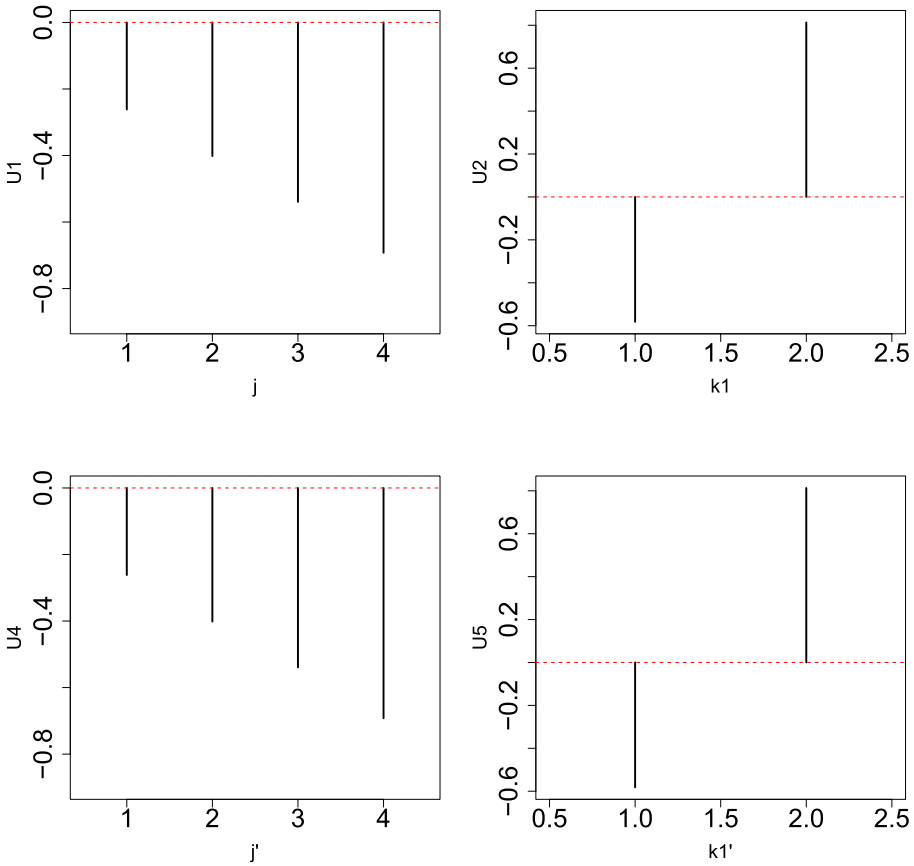

Figure 9. Singular value vectors obtained when KTD-based unsupervised FE was applied to the 2nd synthetic data set. Upper left: u_{1j}, upper right: u_{2k_1}, lower left: $u_{1j'}$, lower right: $u_{2k'_1}$. The horizontal red broken line indicates baseline (zero).

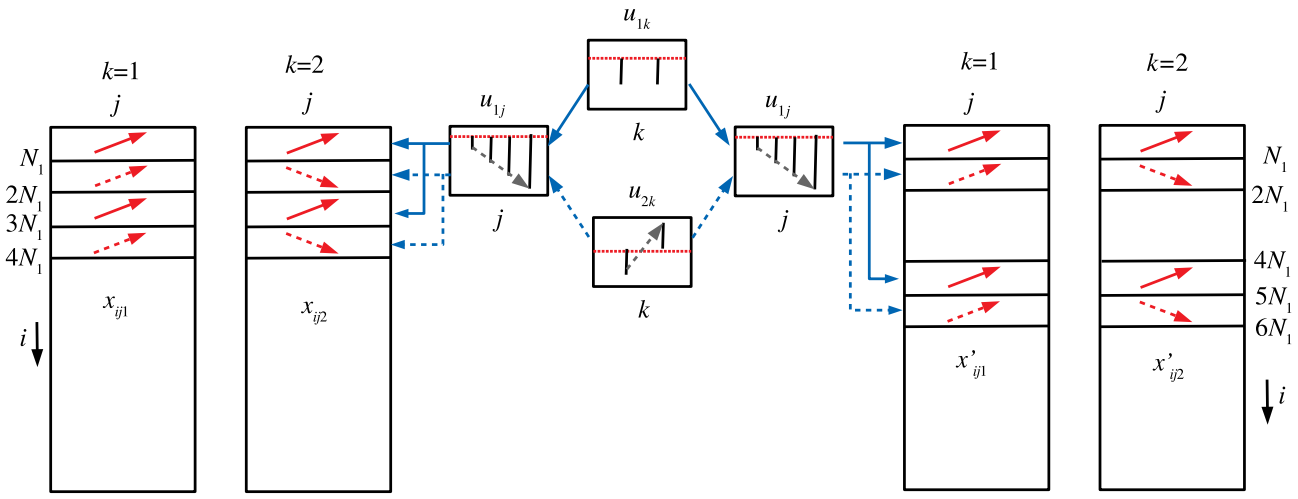

Figure 10. Schematic figure that explains the correspondence between $x_{ijk}, i \leq N_1, 2N_1 < i \leq 3N_1$, $x'_{ijk}, i \leq N_1$, and $4N_1 < i \leq 5N_1$ (red solid arrows) and u_{1j} and u_{1k} and that between $x_{ijk}, N_1 < i \leq 2N_1, 3N_1 < i \leq 4N_1, x'_{ijk}, N_1 < i \leq 2N_1$, and $5N_1 < i \leq 6N_1$ (red broken arrows) and u_{1j} and u_{2k}, respectively.

Materials and methods

m6A and gene expression profiles. m6A and gene expression profiles were downloaded from Gene Expression Omnibus (GEO) using GEO ID GSE141941. For m6A, eight files included in GSE141941_RAW.tar available as part of the Supplementary Information were employed. m6A profiles were summed up within 25,000-nucleotide intervals sequentially divided over the whole genome. As a result, 123,817 genomic regions of 25,000 nucleotides in

KTD based unsupervised FE	$\ell_1 = \ell_2 = 1$			$\ell_4 = \ell_5 = 1$	
Adjusted *P*-values	> 0.01	≤ 0.01		> 0.01	≤ 0.01
$i \notin [1, N_1] \cup [2N_1 + 1, 3N_1]$	980	0	$i \notin [1, N_1] \cup [4N_1 + 1, 5N_1]$	980	0
$i \in [1, N_1] \cup [2N_1 + 1, 3N_1]$	0	20	$i \in [1, N_1] \cup [4N_1 + 1, 5N_1]$	0	20
	$\ell_1 = 1, \ell_2 = 2$			$\ell_4 = 1, \ell_5 = 2$	
Adjusted *P*-values	> 0.01	≤ 0.01		> 0.01	≤ 0.01
$i \notin [N_1 + 1, 2N_1] \cup [3N_1 + 1, 4N_1]$	980	0	$i \notin [N_1 + 1, 2N_1] \cup [5N_1 + 1, 6N_1]$	980	0
$i \in [N_1 + 1, 2N_1] \cup [3N_1 + 1, 4N_1]$	0	20	$i \in [N_1 + 1, 2N_1] \cup [5N_1 + 1, 6N_1]$	0	20

Table 11. Confusion matrices of selected is between true and those selected by KTD-based unsupervised FE applied to synthetic data set 2.

		x'_{ijk}	
x_{ijk}	Adjusted *P*-values	> 0.01	≤ 0.01
	> 0.01	970	10
	≤ 0.01	10	10

Table 12. Confusion matrix of selected *i*s between x_{ijk} and x'_{ijk} by KTD-based unsupervised FE applied to synthetic data set 2.

length were obtained. For gene expression, four profiles included in GSE141941_normoxiaVShypoxia6h.12h.24h_RNA-seq.PROCESSED.DATA.xlsx, which is also available as a part of the Supplementary Information, were employed. Eight files for m6A were composed of four time points (including the control), time input, or treated files. Four profiles for gene expression were composed of four times points as well.

As an alternative data set, we employed GSE120860, in which only the m6A profile was measured. We downloaded 16 bed files provided in the Supplementary Information section in GEO, which correspond to four healthy controls and four patients, of which the tumor and paratumor were measured.

Synthetic data set. In order to demonstrate how well KTD-based unsupervised FE can work when there are a small number of samples associated with a large number of features (observations) and where other conventional supervised methods fail, we prepared a synthetic data set. It has N variables attributed to eight samples whose number is the same as that of the m6A profiles. In the first data set, we aimed to demonstrate the performance when KTD-based unsupervised FE is applied to a single data set. It is composed of $x_{ijk} \in \mathbb{R}^{N\times 4\times 2}$,

$$x_{ijk} = \begin{cases} j + \epsilon_{ijk} & i \leq N_1, \\ j + \epsilon_{ijk} & N_1 < i \leq 2N_1,\ k = 1 \\ -j + \epsilon_{ijk} & N_1 < i \leq 2N_1,\ k = 2, \end{cases} \tag{1}$$

where j corresponds to the first to fourth time points, and $k = 1$ and $k = 2$ correspond to two distinct experimental conditions. ϵ_{ijk} obeys $\mathcal{N}(0, \frac{1}{2})$, where $\mathcal{N}(\mu, \sigma)$ is a Gaussian distribution with a mean of μ and a standard deviation of σ. For $i \leq N_1$, the two conditions have the same dependence on time points, whereas for $N_1 < j \leq 2N_2$, the two conditions have opposite time point dependence.

In the second dataset, we aimed to demonstrate how KTD-based unsupervised FE can identify features that share the same time point dependence on two measurements represented as two tensors, x_{ijk} and $x'_{ijk} \in \mathbb{R}^{N\times 4\times 2}$, which obey Eq. (1) for $i \leq 2N_1$, and for $i > 2N_1$,

$$x_{(i+2N_1)jk} = x_{ijk}, i \leq 2N_1 \tag{2}$$

$$x'_{(i+4N_1)jk} = x_{ijk}, i \leq 2N_1. \tag{3}$$

Thus, for $i \leq 2N_1$, x_{ijk}, $x_{(i+2N_1)jk}$, x'_{ijk}, and $x'_{(i+4N_1)jk}$ share the same time-point dependence.

PCA-based unsupervised FE applied to gene expression. Although the details of PCA-based unsupervised FE have been described in a recently published book[25], we briefly outline this method. Suppose we have gene expression profiles as a matrix, $x_{it} \in \mathbb{R}^{N\times 4}$, which represents the gene expression of the ith gene at the tth time point. The ℓth PC score attributed to gene i, $u_{\ell i} \in \mathbb{R}^N$, can be obtained as the ith component of the ℓth eigenvector, $\boldsymbol{u}_\ell$, of the gram matrix $XX^T \in \mathbb{R}^{N\times N}$, where X is an $N \times 4$ matrix composed of x_{it}, and X^T is a transposed matrix of X as

$$XX^T \boldsymbol{u}_\ell = \lambda_\ell \boldsymbol{u}_\ell. \tag{4}$$

In contrast, the PC loading attributed to the tth time point can be computed as the tth component of the ℓth PC loading vector, $\boldsymbol{v}_\ell \in \mathbb{R}^4$, which can be computed as

$$\boldsymbol{v}_\ell = X^T \boldsymbol{u}_\ell. \tag{5}$$

This is also the ℓth eigenvector of $X^T X \in \mathbb{R}^{4\times 4}$ because

$$X^T X \boldsymbol{v}_\ell = X^T X X^T \boldsymbol{u}_\ell = X^T \lambda_\ell \boldsymbol{u}_\ell = \lambda_\ell X^T \boldsymbol{u}_\ell = \lambda_\ell \boldsymbol{v}_\ell. \tag{6}$$

In order to select genes, we first need to determine which $\boldsymbol{v}_\ell$ is associated with time dependence. After identifying the time-dependent $\boldsymbol{v}_\ell$, we attribute P-values to the ith gene using $u_{\ell i}$ by assuming that $u_{\ell i}$ follows a Gaussian distribution (null hypothesis)

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell i}}{\sigma_\ell}\right)^2 \right], \tag{7}$$

where $P_\chi[> x]$ is the cumulative χ^2 distribution in which the argument is larger than x, and σ_ℓ is the standard deviation. The computed P-values were corrected by the BH criterion[25], and genes associated with adjusted P-values less than 0.01 were selected.

TD-based unsupervised FE applied to the m6A profile. Although the details of TD-based unsupervised FE are described in a recently published book[25], we briefly outline this method. For GSE141941, suppose that a tensor $x_{ktj} \in \mathbb{R}^{K\times 4\times 2}$ represents the m6A of the kth genomic region at the tth time point of input (control, $j = 1$) or m6A ($j = 2$) sample. Individual genomic regions are 25,000-nucleotide sequence length regions sequentially defined over the whole genome without overlaps and adjusted with each other. Higher-order singular value decomposition[25] (HOSVD) was applied to x_{ktj}, and TD was obtained as

$$x_{ktj} = \sum_{\ell=1}^{K}\sum_{\ell_2=1}^{4}\sum_{\ell_3=1}^{2} G(\ell_1\ell_2\ell_3) u_{\ell_1 k} u_{\ell_2 t} u_{\ell_3 j}, \tag{8}$$

where $G \in \mathbb{R}^{K\times 4\times 2}$ is the core tensor, $u_{\ell_1 k} \in \mathbb{R}^{K\times K}$, $u_{\ell_2 t} \in \mathbb{R}^{4\times 4}$, and $u_{\ell_3 j} \in \mathbb{R}^{2\times 2}$ are singular value matrices and orthogonal matrices.

To select genomic regions that are associated with time dependence and to distinguish between input and m6A treatment, we need to specify which $u_{\ell_2 t}$ and $u_{\ell_3 j}$ are associated with time dependence and the distinction between control (input) and m6a, respectively. Once $u_{\ell_2 t}$ and $u_{\ell_3 j}$ are fixed, we attempt to find $G(\ell_1\ell_2\ell_3)$ with the largest absolute value, given ℓ_2 and ℓ_3. Finally, we attribute the P-value to the kth genomic region by assuming that $u_{\ell_1 k}$ obeys a Gaussian distribution (null hypothesis) as

$$P_k = P_{\chi^2}\left[> \left(\frac{u_{\ell_1 k}}{\sigma_{\ell_1}}\right)^2 \right], \tag{9}$$

where σ_{ℓ_1} is the standard deviation. The computed P-values were corrected by the BH criterion, and genes associated with adjusted P-values less than 0.01 were selected.

For GSE126860, m6A profiles were formatted as $x_{ijkm} \in \mathbb{R}^{N\times 4\times 2\times 2}$, which represents the m6A profiles of the ith gene at the jth subject of the kth group ($k = 1$: annotated as 0204 in GEO, $k = 2$: annotated as patient in GEO) of the mth tissue ($m = 1$:tumor, $m = 2$:paratumor). HOSVD was applied, and we obtained

$$x_{ijkm} = \sum_{\ell=1}^{N}\sum_{\ell_2=1}^{4}\sum_{\ell_3=1}^{2}\sum_{\ell_4=1}^{2} G(\ell_1\ell_2\ell_3\ell_4) u_{\ell_1 i} u_{\ell_2 j} u_{\ell_3 k} u_{\ell_3 m}, \tag{10}$$

where $G \in \mathbb{R}^{N\times 4\times 2\times 2}$ is the core tensor, and $u_{\ell_1 i} \in \mathbb{R}^{N\times N}$, $u_{\ell_2 j} \in \mathbb{R}^{4\times 4}$, and $u_{\ell_3 k}, u_{\ell_4 m}, \in \mathbb{R}^{2\times 2}$ are singular value matrices and orthogonal matrices.

In order to select genes that are associated with the distinction between tumor and paratumor, but independent of subjects as well as groups, we need to specify which $u_{\ell_4 m}$ are associated with the distinction between the tumor and paratumor, but $u_{\ell_2 j}$ and $u_{\ell_3 k}$ take constant values. Once $u_{\ell_2 j}$, $u_{\ell_3 k}$, and $u_{\ell_4 m}$ are fixed, we then attempt to find that $G(\ell_1\ell_2\ell_3\ell_4)$ with the largest absolute value, given ℓ_2, ℓ_3, and ℓ_4. Finally, we attribute the P-value to the ith gene by assuming that $u_{\ell_1 i}$ obeys a Gaussian distribution (null hypothesis) as

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell_1 i}}{\sigma_{\ell_1}}\right)^2 \right], \tag{11}$$

where σ_{ℓ_1} is the standard deviation. The computed P-values were corrected by the BH criterion, and genes associated with adjusted P-values less than 0.01 were selected.

When HOSVD was applied to the 1st synthetic data set, we obtained

$$x_{ijk} = \sum_{\ell_1=1}^{N} \sum_{\ell_2=1}^{4} \sum_{\ell_3=1}^{2} G(\ell_1 \ell_2 \ell_3) u_{\ell_1 i} u_{\ell_2 j} u_{\ell_3 k}, \tag{12}$$

where $G \in \mathbb{R}^{N \times 4 \times 2}$ is the core tensor and $u_{\ell_1 i} \in \mathbb{R}^{N \times N}$, $u_{\ell_2 j} \in \mathbb{R}^{4 \times 4}$, and $u_{\ell_3 k} \in \mathbb{R}^{2 \times 2}$ are singular value matrices and orthogonal matrices.

After identifying $u_{\ell_2 j}$ and $u_{\ell_3 k}$ associated with properties of interest, we attempt to find $G(\ell_1 \ell_2 \ell_3)$ with the largest absolute value, given ℓ_2, ℓ_3. Using the selected ℓ_1, *P*-values are attributed to the *i*th by assuming that $u_{\ell_1 i}$ obeys a Gaussian distribution,

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell_1 i}}{\sigma_{\ell_1}} \right) \right]. \tag{13}$$

The computed *P*-values were corrected by the BH criterion, and genes associated with adjusted *P*-values less than 0.01 were selected.

Integrated analysis of gene expression and m6A profiles. To integrate gene expression and m6A profiles, we employed a recently proposed KTD-based unsupervised FE[26]. We define a tensor $x_{tjt'j'} \in \mathbb{R}^{4 \times 2 \times 4 \times 2}$ as

$$x_{tjt'j'} = \sum_{t''} \left(\sum_{i} x_{it} x_{it''} \right) \left(\sum_{k} x_{kt''j} x_{kt'j'} \right). \tag{14}$$

HOSVD was applied to $x_{tjt'j'}$, and we obtained

$$x_{tjt'j'} = \sum_{\ell_1=1}^{4} \sum_{\ell_2=1}^{2} \sum_{\ell_3=1}^{4} \sum_{\ell_4=1}^{2} G(\ell_1 \ell_2 \ell_3 \ell_4) u_{\ell_1 t} u_{\ell_2 j} u_{\ell_3 t'} u_{\ell_4 j'}, \tag{15}$$

where $G \in \mathbb{R}^{4 \times 2 \times 4 \times 2}$ is the core tensor, and $u_{\ell_1 t} \in \mathbb{R}^{4 \times 4}$, $u_{\ell_2 t'} \in \mathbb{R}^{2 \times 2}$, $u_{\ell_3 j} \in \mathbb{R}^{4 \times 4}$, and $u_{\ell_4 j'} \in \mathbb{R}^{2 \times 2}$ are singular value matrices and orthogonal matrices. Here, it should be noted that $u_{\ell_2 j}, u_{\ell_3 t'}$, and $u_{\ell_4 j'}$ are attributed to m6A profiles, and only $u_{\ell_1 t}$ is attributed to the gene expression profiles.

In order to identify genes whose expression profiles depend on time and genomic regions where m6A profiles depend on time associated with the distinction between m6A and control, we need to find which $u_{\ell_1 t}$ and $u_{\ell_3 t'}$ depend on time and which $u_{\ell_2 j}$ and $u_{\ell_4 j'}$ are distinct between control and m6A (since $x_{tjt'j'}$ does not change even if j is replaced with j', $u_{\ell_2 j} = u_{\ell_4 j'}$). Once ℓ_1, ℓ_2, ℓ_3, and ℓ_4 are identified, we can compute the singular value vectors attributed to gene expression samples, $u_{\ell_1 i}$, and m6A profiles, $u_{\ell_2 \ell_3 k}$, can be computed as

$$u_{\ell_1 i} = \sum_{t=1}^{4} u_{\ell_1 t} x_{it} \tag{16}$$

$$u_{\ell_2 \ell_3 k} = \sum_{j=1}^{2} \sum_{t=1}^{4} u_{\ell_2 j} u_{\ell_3 t} x_{ktj}. \tag{17}$$

The *P*-values are attributed to *i*s and *k*s as

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell_1 i}}{\sigma_{\ell_1}} \right)^2 \right], \tag{18}$$

$$P_k = P_{\chi^2}\left[> \left(\frac{u_{\ell_2 \ell_3 k}}{\sigma_{\ell_2 \ell_3}} \right)^2 \right], \tag{19}$$

where σ_{ℓ_1} and $\sigma_{\ell_2 \ell_3}$ are standard deviations. The computed *P*-values were corrected by the BH criterion, and genes, *i*, and genomic regions, *k*, associated with adjusted P-values less than 0.01 were selected.

Integrated analysis of the 2nd synthetic data set. To integrate x_{ijk} and x'_{ijk} in the second synthetic data set, we define a tensor, $x_{jk_1k_2j'k'_1k'_2} \in \mathbb{R}^{4 \times 2 \times 2 \times 4 \times 2 \times 2}$, as

$$x_{jk_1k_2j'k'_1k'_2} = \sum_{j''} \left(\sum_{i} x_{ijk_1} x_{ij''k_2} \right) \left(\sum_{i} x'_{ij''k'_1} x_{ij'k'_2} \right). \tag{20}$$

After applying HOSVD to $x_{jk_1k_2j'k'_1k'_2}$, we obtained

$$x_{jk_1k_2j'k_1'k_2'} = \sum_{\ell_1=1}^{4}\sum_{\ell_2=1}^{2}\sum_{\ell_3=1}^{2}\sum_{\ell_4=1}^{4}\sum_{\ell_5=1}^{2}\sum_{\ell_5=1}^{2} G(\ell_1\ell_2\ell_3\ell_4\ell_5\ell_6) \times u_{\ell_1 j}u_{\ell_2 k_1}u_{\ell_3 k_2}u_{\ell_4 j'}u_{\ell_5 k_1'}u_{\ell_6 k_2'}. \tag{21}$$

Singular value vectors attributed to *i*s can be reproduced as

$$u_{\ell_1\ell_2 i} = \sum_{j=1}^{4}\sum_{k_1=1}^{2} x_{ijk_1}u_{\ell_1 j}u_{\ell_2 k_1} \tag{22}$$

$$u_{\ell_4\ell_5 i} = \sum_{j=1}^{4}\sum_{k_1=1}^{2} x'_{ijk_1}u_{\ell_4 j}u_{\ell_5 k_1}. \tag{23}$$

After identifying $u_{\ell_1 j}, u_{\ell_2 k_1}, u_{\ell_4 j},$ and $u_{\ell_5 k_1}$ is considered, *P*-values are attributed to *i*, as

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell_1\ell_2 i}}{\sigma_{\ell_1\ell_2}}\right)^2\right], \tag{24}$$

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell_4\ell_5 i}}{\sigma_{\ell_4\ell_5}}\right)^2\right]. \tag{25}$$

The computed *P*-values were corrected by the BH criterion, and *i*s associated with adjusted P-values less than 0.01 were selected.

Retrieval of gene symbols included in selected genomic regions. After selecting genomic regions, we needed to retrieve the gene symbols included in the selected genomic regions. This could be done using the biomaRt package implemented in R by specifying the hg19 human genome to which short reads were mapped.

Ensembl gene ID to gene symbol. Since gene expression profiles are defined using Ensembl gene IDs, we needed to convert these IDs to gene symbols. This was done by uploading gene symbols selected by TD-based unsupervised FE to DAVID[27]. Uploaded Ensembl gene IDs were converted to gene symbols using the gene ID conversion tool implemented in DAVID by specifying the official gene symbol as the target of conversion.

Enrichment analysis. Identified gene symbols were uploaded to Enrichr[28], which is an enrichment server, to evaluate various enrichments within sets of identified gene symbols.

Various conventional feature selections. *Linear regression-based feature selection.* To select genes or genomic regions using linear regression analysis, the `ls` function in the base package in R was used. *P*-values computed by `ls` were corrected by the BH criterion, and genes or genomic regions associated with adjusted *P*-values less than 0.01 or 0.05 were selected.

When linear regression was applied to gene expression, x_{it},

$$x_{it} = a_i + b_i T(t) \tag{26}$$

was assumed, where $T(1) = 0, T(2) = 6, T(3) = 12,$ and $T(4) = 24$.

When linear regression was applied to m6A, x_{ktj},

$$x_{ktj} = a_k + b_k T(t)j \tag{27}$$

was assumed.

When linear regression was applied to the 1st synthetic data,

$$x_{ijk} = a_i + b_i jk \tag{28}$$

was assumed.

SAM. When SAM[29] was applied to gene expression, x_{it}, or m6A, x_{ktj}, are assumed to be classified into four classes based on *t* (for gene expression) or eight classes based on the combination of *t* and *j* (for m6A), respectively. The `sam` function was implemented in the siggenes package in R.

Limma. When limma[30] was applied to gene expression, x_{it}, or m6A, x_{ktj}, respectively, they were assumed to be classified in the same way as in SAM. Limma was applied to logarithmically converted x_{it} or x_{ktj}. The `limma` function was implemented in the limma package in R.

When limma was applied to the first synthetic data set, x_{ijk} was classified into eight classes based on the pairs of *j* and *k*. Because x_{ijk} takes both positive and negative values, x_{ijk}s themselves were regarded as logarithmically converted valRes.

Random forest. When random forest[31] was applied to gene expression, x_{it}, or m6A, x_{ktj}, respectively, they are assumed to be classified in the same way as in SAM. When it was applied to the first synthetic data set, x_{ijk} was classified into eight classes, as in the case of limma. The `randomForest` function was implemented in the randomForest package. Features included in OOB were selected by selecting features with non-zero importance given by the `importance` function implemented in the `randomForest` package in R.

Received: 9 January 2021; Accepted: 5 April 2021
Published online: 26 April 2021

References

1. Roach, R. C. *et al.* (eds) *Hypoxia* (Springer, 1999).
2. Dhont, S., Derom, E., Braeckel, E. V., Depuydt, P. & Lambrecht, B. N. The pathophysiology of 'happy' hypoxemia in COVID-19. *Respir. Res.* **21**, 198. https://doi.org/10.1186/s12931-020-01462-5 (2020).
3. Muz, B., de la Puente, P., Azab, F. & Azab, A. K. The role of hypoxia in cancer progression, angiogenesis, metastasis, and resistance to therapy. *Hypoxia* **2015**(3), 83–92. https://doi.org/10.2147/hp.s93413 (2015).
4. Hossmann, K.-A. The hypoxic brain. In *Advances in Experimental Medicine and Biology* 155–169. https://doi.org/10.1007/978-1-4615-4711-2_14 (Springer, New York, 1999).
5. Schumacker, P. T. Lung cell hypoxia: Role of mitochondrial reactive oxygen species signaling in triggering responses. *Proc. Am. Thorac. Soc.* **8**, 477–484. https://doi.org/10.1513/pats.201103-032mw (2011).
6. Sarkar, M., Niranjan, N. & Banyal, P. Mechanisms of hypoxemia. *Lung India* **34**, 47. https://doi.org/10.4103/0970-2113.197116 (2017).
7. Fry, N. J., Law, B. A., Ilkayeva, O. R., Holley, C. L. & Mansfield, K. D. N6-methyladenosine is required for the hypoxic stabilization of specific mRNAs. *RNA* **23**, 1444–1455. https://doi.org/10.1261/rna.061044.117 (2017).
8. Wang, Y.-J. *et al.* Reprogramming of m6a epitranscriptome is crucial for shaping of transcriptome and proteome in response to hypoxia. *RNA Biol.* **18**(1), 131–143. https://doi.org/10.1080/15476286.2020.1804697 *(2020).*
9. Luo, Y., Wang, F. & Szolovits, P. Tensor factorization toward precision medicine. *Brief. Bioinform.* **18**, 511–514. https://doi.org/10.1093/bib/bbw026 (2016).
10. Yahyanejad, F., Albert, R. & DasGupta, B. A survey of some tensor analysis techniques for biological systems. *Quant. Biol.* **7**, 266–277. https://doi.org/10.1007/s40484-019-0186-5 (2019).
11. Fang, J. Tightly integrated genomic and epigenomic data mining using tensor decomposition. *Bioinformatics* **35**, 112–118. https://doi.org/10.1093/bioinformatics/bty513 (2018).
12. Hore, V. *et al.* Tensor decomposition for multiple-tissue gene expression experiments. *Nat. Genet.* **48**, 1094–1100. https://doi.org/10.1038/ng.3624 (2016).
13. Ramdhani, S. *et al.* Tensor decomposition of stimulated monocyte and macrophage gene expression profiles identifies neurodegenerative disease-specific trans-eqtls. *PLoS Genet.* **16**, 1–23. https://doi.org/10.1371/journal.pgen.1008549 (2020).
14. Wang, M., Fischer, J. & Song, Y. S. Three-way clustering of multi-tissue multi-individual gene expression data using semi-nonnegative tensor decomposition. *Ann. Appl. Stat.* **13**, 1103–1127. https://doi.org/10.1214/18-AOAS1228 (2019).
15. Li, Y. & Ngom, A. Classification of clinical gene-sample-time microarray expression data via tensor decomposition methods. In *Computational Intelligence Methods for Bioinformatics and Biostatistics* (eds Rizzo, R. & Lisboa, P. J. G.) 275–286 (Springer, 2011).
16. Hu, Y., Liu, J.-X., Gao, Y.-L., Li, S.-J. & Wang, J. Differentially expressed genes extracted by the tensor robust principal component analysis (TRPCA) method. *Complexity* **1–13**, 2019. https://doi.org/10.1155/2019/6136245 (2019).
17. Diaz, D., Bollig-Fischer, A. & Kotov, A. Tensor decomposition for sub-typing of complex diseases based on clinical and genomic data. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 647–651. https://doi.org/10.1109/BIBM47256.2019.8983014 (2019).
18. Bradley, M. W., Aiello, K. A., Ponnapalli, S. P., Hanson, H. A. & Alter, O. GSVD- and tensor GSVD-uncovered patterns of DNA copy-number alterations predict adenocarcinomas survival in general and in response to platinum. *APL Bioeng.* **3**, 036104. https://doi.org/10.1063/1.5099268 (2019).
19. Solaini, G., Baracca, A., Lenaz, G. & Sgarbi, G. Hypoxia and mitochondrial oxidative metabolism. *Biochim. Biophys. Acta (BBA) Bioenergy* **1797**, 1171–1177. https://doi.org/10.1016/j.bbabio.2010.02.011 (2010) **(16th European Bioenergetics Conference 2010)**.
20. Chan, C. K. & Vanhoutte, P. M. Hypoxia, vascular smooth muscles and endothelium. *Acta Pharm. Sin. B* **3**, 1–7. https://doi.org/10.1016/j.apsb.2012.12.007 (2013).
21. Sugimoto, N., Ishibashi, H., Nakamura, H., Yachie, A. & Ohno-Shosaku, T. Hypoxia-induced inhibition of the endocannabinoid system in glioblastoma cells. *Oncol. Rep.* **38**(6), 3702–3708. *Rep.*https://doi.org/10.3892/or.2017.6048 *(2017).*
22. Jha, N. K. *et al.* Hypoxia-induced signaling activation in neurodegenerative diseases: Targets for new therapeutic strategies. *J. Alzheimer's Dis.* **62**, 15–38. https://doi.org/10.3233/JAD-170589 (2018).
23. Semenza, G. L., Roth, P. H., Fang, H. M. & Wang, G. L. Transcriptional regulation of genes encoding glycolytic enzymes by hypoxia-inducible factor 1. *J. Biol. Chem.* **269**, 23757–23763 (1994).
24. Song, F., Guo, Z. & Mei, D. Feature selection using principal component analysis. In *2010 International Conference on System Science, Engineering Design and Manufacturing Informatization*. https://doi.org/10.1109/icsem.2010.14 (IEEE, 2010).
25. Taguchi, Y.-H. *Unsupervised Feature Extraction Applied to Bioinformatics* (Springer International Publishing, 2020).
26. Taguchi, Y. H. & Turki, T. Application of tensor decomposition to gene expression of infection of mouse hepatitis virus can identify critical human genes and efffective drugs for SARS-CoV-2 infection. *IEEE J. Sel. Top. Signal Process.* **15**(3), 746–758. https://doi.org/10.1109/JSTSP.2021.3061251 *(2021).*
27. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57. https://doi.org/10.1038/nprot.2008.211 (2008).
28. Kuleshov, M. V. *et al.* Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97. https://doi.org/10.1093/nar/gkw377 (2016).
29. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci.* **98**, 5116–5121. https://doi.org/10.1073/pnas.091062498 (2001).
30. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47. https://doi.org/10.1093/nar/gkv007 (2015).
31. Liaw, A. & Wiener, M. Classification and regression by randomforest. *R News* **2**, 18–22 (2002).

Acknowledgements

This study was supported by KAKENHI 20K12067, 20H04848, and 19H05270.

Author contributions

All authors planned the study. Y.H.T. performed the analyses. All authors validated and discussed the results. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-87779-7.

Correspondence and requests for materials should be addressed to Y.-H.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021