MDPI

Article

Tensor-decomposition-based unsupervised feature extraction applied to prostate cancer multiomics data

Y-h. Taguchi [1,*], Turki Turki [2,]

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

Version December 2, 2020 submitted to Genes

Abstract: The large p small n problem is a challenge without a de facto standard method available to it. In this study, we propose a tensor-decomposition (TD)-based unsupervised feature extraction (FE) formalism applied to multiomics datasets, in which the number of features is more than 100,000 whereas the number of samples is as small as about 100, hence constituting a typical large p small n problem. The proposed TD-based unsupervised FE outperformed other conventional supervised feature selection methods, random forest, categorical regression (also known as analysis of variance, or ANOVA), penalized linear discriminant analysis, and two unsupervised methods, multiple non-negative matrix factorization and principal component analysis (PCA) based unsupervised FE when applied to synthetic datasets and four methods other than PCA based unsupervised FE when applied to multiomics datasets. The genes selected by TD-based unsupervised FE were enriched in genes known to be related to tissues and transcription factors measured. TD-based unsupervised FE was demonstrated to be not only the superior feature selection method but also the method that can select biologically reliable genes. To our knowledge, this is the first study in which TD-based unsupervised FE has been successfully applied to the integration of this variety of multiomics measurements.

Keywords: prostate cancer; gene expression; genomic regions; protien-coding genes; tensor decomposition; unsupervised learning

1. Introduction

The term "big data" often indicates a high number of instances as well as features [1,2]. Typical big data comprise a few million images (photos) each composed of several million pixels. The number of features is as high as the number of instances, or often higher. Most recently developed machine-learning methods including deep learning (DL) aim to address this manner of problem [1,3–8]. Nonetheless, although less popular than these typical big data problems, there is another branch of big data problem known as the "large p small n" problem, or "high-dimensional data analysis". For these cases, the number of instances is typically smaller than the number of features [9–11]. In this case, there can be many unique problems that do not exist for the typical big data problems mentioned above. For example, in discrimination tasks, because there are more features than instances, if we do not reduce the number of features, the accuracy will always be 100% for no other reason than overfitting. Alternatively, the so-called "curse of dimensionality" problem describes a shortage of instances covering sufficiently large spaces. Because of these difficulties, compared with the popular big data problems, the "large p small n" problem has relatively rarely been investigated comprehensively, although there are many studies to tackle "large p small n" problem [12–15].

This might not seem significant, as the "large p small n" problem is relatively rare and not as important. Nevertheless, dependent on the domains, the "large p small n" problem will inevitably

become typical. For example, in genomic science, the typical number of features is as many as or more than that of genes (i.e., about 10^4), whereas the number of samples is the number of patients, typically at most a few hundred. The length of the DNA sequence of the human genome is approximately 3×10^9 [16], whereas the human population is about 10^{10}. Nevertheless, only a very small fraction of these populations can be considered in an individual study. Thus, in typical genomic studies, the ratio of the number of features to the number of samples can be in the region of 10^2 or more.

Nonetheless, Taguchi has proposed a very different method to the typical machine-learning methods that are applicable to large p small n problems: tensor-decomposition (TD)-based unsupervised feature extraction (FE) [17]. m-mode tensor is associated with more than two suffix whereas matrix is associated with two suffix, row and column. In this method, a smaller number of representative features, referred to as singular value vectors, are generated with linear combinations of the original large number of features, without considering labeling. These singular value vectors are attributed to both samples and original features. We have investigated the singular value vectors attributed to samples. Finally, the selected singular value vectors attributed to the original features are used to evaluate the importance of the original features; because the singular value vectors are the linear combinations of original features, their coefficients of linear combinations can be used to evaluate the importance of individual original features. The original features with larger absolute values of coefficients are selected. As a result, we can avoid many difficulties in the "large p small n" problem; for example, we can perform standard discriminant analysis when there is a higher number of features than instances.

In this paper, we propose a TD-based unsupervised FE formalism, applied to synthetic data that imitate typical genomic datasets. The proposed method is first compared with four conventional methods applicable to the "large p small n" problem: random forest (RF), penalized linear discriminant analysis (LDA) [18], categorical regression, and multiple non-negative matrix factorization (MNMF); the latter has been used especially often for analysis of multiomics datasets [19]. These five methods are then also applied to a real dataset consisting of multiomics measurements of prostate cancers, a typical "large p small n" dataset. The results for the synthetic and real datasets demonstrate the superiority of our proposed method when compared against the four conventional methods in feature selections.

In this study, we need to deal with categorical data set. Since the labeling is neither binary nor numeric, no standard regression-based approaches are applicable. Thus, the methods applicable are restricted to either supervised methods that can deal with categorical labeling, e.g., tissues or unsupervised methods that do not require labeling and do not consider sample annotations at all. The above six methods were selected as representative methods that can satisfy the above requirements.

To our knowledge, this is the first study in which TD-based unsupervised FE has been successfully applied to the integration of this variety of multiomics measurements. The performance of TD-based unsupervised FE is highly data-type-dependent, since we cannot intentionally generate singular value vectors of interest. When singular value vectors generated by TD is not of interest, we cannot do anything since TD does not have tunable parameters. In this sense, it is challenging to estimate how many types of multiomics datasets (here, three transcription factor bindings as well as four histone modifications and one chromosomal state measurement, making eight types in total) can be successfully integrated with this method.

The types of multiomics data [20] that are possibly dealt with the proposed technology are as follows: gene (or mRNA) expression profiles, DNA methylation profiles (e.g., promoter methylation), various histone modification, chromatin structure, protein binding to genome (e.g., transcription factor biding to DNA), and RNA modification and so on.

A recent review [20] listed handful multi-omics data sets analysis tools (Table 1). As can be seen, the number of samples are from 10^2 to 10^3 and numbers of omics data used are at most four. TD based unsupervised FE (see below) can deal with mode extreme case; the number of samples per omics and tissue is from three to thirty while number of omics is as many as eight. We do not know any methods

Table 1. List of number of sample as well as types of omics data analyzed in the review [20]

		Omics data	
Model name	number of samples	numbers	types
PARADIGM [21]	230 patient samples and 10 adjacent normal tissues	two omics data	copy number and mRNA expression
iCluster [22]	37 primary breast cancers and four breast cancer cell lines	two omics data	copy number and mRNA expression
	91 lung adenocarcinomas	two omics data	copy number and mRNA expression
iClusterPlus [23]	729 human cancer cell lines	three omics data	chromosomal copy number, gene expression, and mutation
	189 tumors	four omics data	exome sequence, DNA copy number, promoter methylation, and mRNA expression
LRAcluster [24]	3,319 samples	four omics data	mutation, CNV, DNA methylation, and gene expression
PSDF [25]	106 breast cancer samples	two omics data	Copy number and gene expression
BCC [26]	348 tumor samples	four omics data	RNA expression, methylation, miRNA expression, Reverse phase protein array
SNF [27]	215 GBM data samples	three omics data	DNA methylation, miRNA expression, and gene expression
PFA [28]	415 cell lines	two omics	gene expression and copy number
PINSPlus [29]	12,158 samples	—	—
NEMO [30]	173 samples DNA methylation data from 194 samples, and miRNA expression data from 188 samples	three omics data	—
DIABLO [31]	150 breast cancer samples	three moics	mRNA, miRNA, and protein expression
moCluster [32]	83 samples of colorectal cancer	three omics data	DNA methylation, gene expression, and protein expression
MCIA [33]	266 samples	two omics data	proteomics and transcriptomics
JIVE [34]	348 breast cancer samples	three omics data	gene expression, DNA methylation, and miRNA data
MFA [35]	43 glioma samples	two omics data	CGH-array and transcriptome
rMKL-LPP [36]	glioblastoma multiforme (GBM) with 213 samples, breast invasive carcinoma (BIC) with 105 samples, kidney renal clear cell carcinoma (KRCCC) with 122 samples, lung squamous cell carcinoma (LSCC) with 106 samples and colon adenocarcinoma (COAD) with 92 samples	three omics data	gene expression, DNA methylation and miRNA expression
iNMF [37]	592 samples	three omics data	gene expression, DNA methylation, miRNA expression

other than ours that can deal with this many kinds of omics data set as well as this small number of samples.

The way by which multiomics data is formatted as a tensor as follows. Suppose that $\mathcal{X} \in \mathbb{R}^{N \times M \times K \times S}$ composed of x_{ijks} that represents jth kind of omics data (e.g., gene expression, promoter methylation, histone modification etc) attributed to ith gene of sth biological replicates of kth organism. In this case, for kth patients, M kind of omics data was measured using microarray or next generation sequencing technology measuring variables attributed to N genes at once. Thus, the multiomics measurements are naturally formatted as a tensor without any special reformat.

The primary difference between PCA based unsupervised FE and TD based unsupervised FE is as follows. When condition of experiments is restricted to one, e.g., "patients vs healthy control", PCA is very useful. However, if it is a combination of multiple ones, e.g., "patients vs healthy controls" and

"multiple tissues", since tensor is more reasonable format than matrix, TD based unsupervised FE is more suitable method than PCA based unsupervised FE.

2. Materials and Methods

2.1. Synthetic data

In the synthetic dataset, we assumed that there are 81 samples composed of three tissue subclasses (each class has three replicates) and three disease subclasses (each class also has three replicates). That is, the 81 samples are composed of all possible pairs of one of the nine tissue samples (= three tissues $\times$ three replicates) and one of the nine disease samples (= three diseases $\times$ three replicates). Typically, the three tissues are supposed to be, for example, heart, brain, and skin, while the three diseases are supposed to be a few exclusive diseases, and healthy controls; this means that we do not have any biological knowledge of which combinations are more likely to be associated with one another. The number of genes, which are features, is as many as 10^5; although so-called protein-coding genes are at most a few tens of thousands, a higher number of non-coding genes have come to be considered. Unlike the synthetic data, in the real dataset to which the methods are applied in the later part of this paper, the number of features corresponds to the number of genomic regions with fixed DNA sequence length. Thus, it can be even higher than that of the synthetic data. Among these 10^5 features, only the first 100 features are supposed to be associated with distinction between diseases and tissues. The purpose of the analysis is to identify these 100 features correctly based upon the available datasets in a fully data-driven approach. To emulate this situation, we introduced tensor $x_{ijk} \in \mathbb{R}^{10^4 \times 9 \times 9}$, where i, j, k stand for features, tissues, and diseases, as

$$x_{ijk} = \begin{cases} \left\lceil \frac{j}{3} \right\rceil \left(\left\lceil \frac{k}{3} \right\rceil + 3 \right) + \epsilon_{ijk}, & i \leq 100 \\ \epsilon_{ijk}, & i > 100 \end{cases} \tag{1}$$

where $\epsilon_{ijk} \sim \mathcal{N}(0, 3)$, $\mathcal{N}(\mu, \sigma)$ is a Gaussian distribution with mean μ and standard deviation σ, and $\lceil x \rceil$ is a ceiling function that gives the smallest integer that is not less than x. The x_{ijk} values are classified into nine subclasses dependent upon pairs $\left(\left\lceil \frac{j}{3} \right\rceil, \left\lceil \frac{k}{3} \right\rceil \right), 1 \leq \left\lceil \frac{j}{3} \right\rceil, \left\lceil \frac{k}{3} \right\rceil \leq 3$ because $1 \leq j, k \leq 9$.

The synthetic data set intentionally model the simplest cases of integrated analysis of multiple omics data set. j and k are assumed to be two independent omics measurements while replicates are supposed to be biological replicates in real experiments.

All performances reported in this study is averaged value over one hundred independent trials.

2.2. Real dataset

The real dataset consists of multiomics measurements of prostate cancer retrieved from the Gene Expression Omnibus (GEO) [38], using GEO ID GSE130408. The processed file named GSE130408_RAW.tar was downloaded and individual files with names starting with "GSM" were extracted from it. The individual files are composed of eight omics measurements: three transcription factor bindings, AR, FOXA1, and HOXB13; four histone modifications, H3K27AC, H3K27me3, H3K4me3, and K4me2; and ATAC-seq. Two additional subclasses, tumor and normal tissue, were attributed to each. Because the number of biological replicates varies from subclass to subclass, we randomly selected the following number of ChIP-seq subclasses: two AR, four FOXA1, four HOXB13, ten H3K27AC, one H3K27me3, one H3K4me3, one K4me2, and one ATAC-seq. In total, 24 multiomics measurement groups were constructed. Each multiomics measurement group is composed of six samples, comprising two tissue subclasses (tumor and normal prostate), each of which has three biological replicates (Table 2).

Table 2. Sumary of real dataset

Omics	number of multiomics	tissues	biological replicates
AR	2	2	3
FOXA1	4	2	3
HOXB13	4	2	3
H3K27AC	10	2	3
H3K27me3	1	2	3
H3K4me3	1	2	3
K4me2	1	2	3
ATAC	1	2	3
total	24	—	—

The values in each file were averaged over 25,000 DNA sequence intervals, giving a total number of 123,817, as 3×10^9, the total DNA length of the human genome, divided by 25,000 is equal to 120,000. The averaged values were treated as a dataset to be analyzed further.

2.3. Categorical regression

Categorical regression, also known as analysis of variance or ANOVA, is expressed as

$$x_{ijk} = \sum_{J=1}^{3} a_J \delta_{Jj} + \sum_{K=1}^{3} b_K \delta_{Kk}, \tag{2}$$

where δ_{Jj} and δ_{Kk} take 1 only when j or k belong to the Jth tissue or Kth disease subclass, respectively. a_J and b_K are regression coefficients (for synthetic data).

$$x_{ijkm} = \sum_{J=1}^{8} a_J \delta_{Jj} + \sum_{K=1}^{2} b_K \delta_{Kk}, \tag{3}$$

where δ_{Jj} and δ_{Kk} take 1 only when j or k belong to the Jth multiomics measurement group or Kth tissue ($K = 1$: tumor, $K = 2$: normal prostate) subclass, respectively. a_J and b_K are regression coefficients (for real data).

The computation was performed by the `lm` function in the base package in R [39]. The computed P-values were adjusted by the BH criterion [17] and the ith features associated with adjusted P-values less than 0.01 were selected.

2.4. RF

RF was performed using the `randomForest` function implemented in the `randomForest` package [40] in R. Synthetic and real data were regarded as $3 \times 3 = 9$ and $8 \times 2 = 16$ subclasses, respectively, each of which corresponds to one of the (J, K) pairs defined in Eqs. (2) or (3). Features included in the out-of-bag error were selected that had non-zero importance given by the `importance` function implemented in the `randomForest` package.

2.5. PenalizeLDA

PenalizedLDA was performed using the `PenalizedLDA.cv` function implemented in the `PenalizedLDA` package [41] in R. Synthetic and real data were regarded as $3 \times 3 = 9$ and $8 \times 2 = 16$ subclasses, respectively, each of which corresponds to one of the (J, K) pairs defined in Eqs. (2) or (3). λ was taken to be $0.01, 0.02, 0.03, 0.04$ for synthetic data. PenalizedLDA could not be performed for the real dataset because of the zero in-subclass standard deviations of some features (for more details, see [18]).

2.6. TD-based unsupervised FE

Although the process was fully described in the recently published book [17], here we propose a variant, which is outlined briefly. First, a dataset must be formatted as a tensor. For synthetic data, $x_{ijk} \in \mathbb{R}^{10^5 \times 9 \times 9}$. For real data, $x_{ijkm} \in \mathbb{R}^{N \times 24 \times 2 \times 3}$ represents the averaged value of the ith interval of the jth omics measurement group of the kth tissue ($k = 1$: tumor, $k = 2$: normal prostate) of the mth biological replicates ($1 \leq m \leq 3$). Here, N = 123,817 is the total number of genomic regions of 25,000 DNA sequence length.

Higher-order singular value decomposition (HOSVD) [17] was then applied to the tensor, obtaining

$$x_{ijk} = \sum_{\ell_1=1}^{9} \sum_{\ell_2=1}^{9} \sum_{\ell_3=1}^{10^5} G(\ell_1 \ell_2 \ell_3) u_{\ell_1 j} u_{\ell_2 k} u_{\ell_3 i} \tag{4}$$

for synthetic data. $G \in \mathbb{R}^{9 \times 9 \times 10^5}$ is a core tensor and $u_{\ell_1 j} \in \mathbb{R}^{9 \times 9}, u_{\ell_2 k} \in \mathbb{R}^{9 \times 9}, u_{\ell_3 i} \in \mathbb{R}^{10^5 \times 10^5}$ are singular value matrices that are orthogonal. Conversely, for the real dataset,

$$x_{ijkm} = \sum_{\ell_1=1}^{24} \sum_{\ell_2=1}^{2} \sum_{\ell_3=1}^{3} \sum_{\ell_4=1}^{N} G(\ell_1 \ell_2 \ell_3 \ell_4) u_{\ell_1 j} u_{\ell_2 k} u_{\ell_3 m} u_{\ell_4 i}, \tag{5}$$

where $G \in \mathbb{R}^{24 \times 2 \times 3 \times N}$ is a core tensor and $u_{\ell_1 j} \in \mathbb{R}^{24 \times 24}, u_{\ell_2 k} \in \mathbb{R}^{2 \times 2}, u_{\ell_3 m} \in \mathbb{R}^{3 \times 3}, u_{\ell_4 i} \in \mathbb{R}^{N \times N}$ are singular value matrices that are orthogonal.

The first step is to check which singular value vectors attributed to sample subclasses ($u_{\ell_1 j}$ attributed to tissue subclasses and $u_{\ell k}$ to disease subclasses for the synthetic dataset, and $u_{\ell_1 j}$ attributed to multiomics measurement groups and $u_{\ell_2 k}$ to tissues for the real dataset) represent distinction between samples of interest. After identifying biologically interesting singular value vectors attributed to samples ($u_{\ell_1 j}$ and $u_{\ell_2 k}$ for the synthetic data, and $u_{\ell_1 j}$, $u_{\ell_2 k}$, and $u_{\ell_3 m}$ for the real data), we next attempt to find singular value vectors attributed to features ($u_{\ell_3 i}$ for the synthetic data and $u_{\ell_4 i}$ for the real data) that share core tensor G with larger absolute values with those identified singular value vectors attributed to samples. For example, suppose that ℓ'_1, ℓ'_2 are selected for synthetic data and $\ell'_1, \ell'_2, \ell'_3$ are selected for real data. In this case we seek ℓ_3 that has $G(\ell'_1 \ell'_2 \ell_3)$ with larger absolute values for the synthetic data and ℓ_4 that has $G(\ell'_1 \ell'_2 \ell'_3 \ell_4)$ with larger absolute values for the synthetic data. Next, using the selected $u_{\ell_3 i}$ and $u_{\ell_4 i}$, P-values are attributed to the ith feature as [17]

$$P_i = P_{\chi^2} \left[> \left(\frac{u_{\ell_3 i}}{\sigma_{\ell_3}} \right)^2 \right] \tag{6}$$

for synthetic data and

$$P_i = P_{\chi^2} \left[> \left(\frac{u_{\ell_4 i}}{\sigma_{\ell_4}} \right)^2 \right] \tag{7}$$

for real data. $P_{\chi^2}[> x]$ is the cumulative probability of χ^2 distribution that the argument is larger than x. σ_{ℓ_3} and σ_{ℓ_4} are the standard deviations. The computed P-values are corrected by the BH criterion and features associated with adjusted P-values less than 0.01 are selected.

2.7. MNMF

Suppose that we have multiple matrices sharing the same number of rows, i.e., $X_k \in \mathbb{R}^{N \times M_k}$. When there are no non-positive components in these matrices, we can apply MNMF as

$$\min_{Q, H_k} \sum_k ||X_k - QH_k||_F, \tag{8}$$

where $Q \in \mathbb{R}^{N \times n}$ and $H_k \in \mathbb{R}^{n \times M_k}$, the column and row vectors of which are n latent variable vectors attributed to is and samples, respectively, and there are no non-positive components in Q and H_ks. $||\cdot||_F$ is the Frobenius norm. n is an integer smaller than any N or M_k. MNMF can be easily performed by applying non-negative matrix factorization (NMF) to the contraction matrix of X_ks, $X \in \mathbb{R}^{N \times \sum_k M_k}$, which is an unfolded matrix of tensors in this study (see Fig. 1 and below), because

$$\sum_k ||X_k - QH_k||_F = ||X - QH||_F, \tag{9}$$

where $H \in \mathbb{R}^{n \times \sum_k M_k}$ is a contraction matrix of H_ks. NMF was performed by the `nmf` function in the NMF package of R.

When MNMF is applied to the synthetic dataset, it is applied to the matrix $x_{i(j)} \in \mathbb{R}^{10^5 \times 81}$ generated by unfolding tensor $x_{ijk} \in \mathbb{R}^{10^5 \times 9 \times 9}$. When MNMF is applied to the real dataset, it is applied to the matrix $x_{i(jkm)} \in \mathbb{R}^{123817 \times 144}$ generated by unfolding tensor $x_{ijkm} \in \mathbb{R}^{123817 \times 24 \times 2 \times 3}$. Because NMF cannot be applied to matrices with non-positive values, $\min(x_{ijk})$ are extracted from x_{ijk} so that NMF can be applied to matrices without non-positive values when applied to synthetic data. When NMF is applied to the real dataset, 1×10^{-10} is added to x_{ijkm} when $x_{ijkm} = 0$ (no negative values are included in x_{ijkm}).

2.8. PCA based unsupervised FE

Although the process was fully described in the recently published book [17], we briefly explain the method. PCA was applied to matrices to whch MNMF was applied in the previous subsection so that PC scores are attributed to i whereas PC loading applied to either (j) (for synthetic data) or (jkm) (for real data). After identifying PC loading of interest, using corresponding PC score, P-values are attributed to is as [17]

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell i}}{\sigma_\ell}\right)^2\right] \tag{10}$$

where $u_{\ell i}$ is the ℓth PC score that corresponds to the PC loading of interest. P-values are corrected by BH criterion and is associated with adjusted P-values less than 0.01 are selected.

3. Results

Figure 1 shows the flowchart of analyses performed in this study.

3.1. Synthetic data

We applied RF, PenalizedLDA, categorical regression, TD-based unsupervised FE, and PCA based unsupervised FE to the synthetic dataset. Please note that all performances reported in the below are averaged over one hundred independent trials as denoted in Materials and Methods. When RF was applied to the synthetic dataset, there were up to 8456 features in average with non-zero importance, although the 100 features with subclass dependence were correctly selected. Thus, RF has the ability to select all the correct features. Nevertheless, because it selects too many false positives, it is ineffective. Reflecting this poor ability, as small as 16 samples among 81 samples were correctly classified in average. Although one might wonder if we select top 100 features with larger importance values provided by random forest, only as small as 73 features are correctly selected among top ranked 100 features by random forest. In contrast, when PenalizedLDA was applied to the synthetic dataset, the 100 features with subclass dependence were selected without any false positives being selected. Thus, PenalizedLDA can outperform RF. The only disadvantage is that we have to find the optimal λ, which is the weight coefficient of the L1 norm. Through massive trials and errors, we found that when λ is taken to be 0.01, it achieves the ideal performance mentioned above for all of one hundred independent trials. Reflecting this, PenalizedLDA can classify 81 samples with the accuracy of as high as 84 % in average. In contrast, if we select λ as 0.02, it selects as small as only nine features in average,

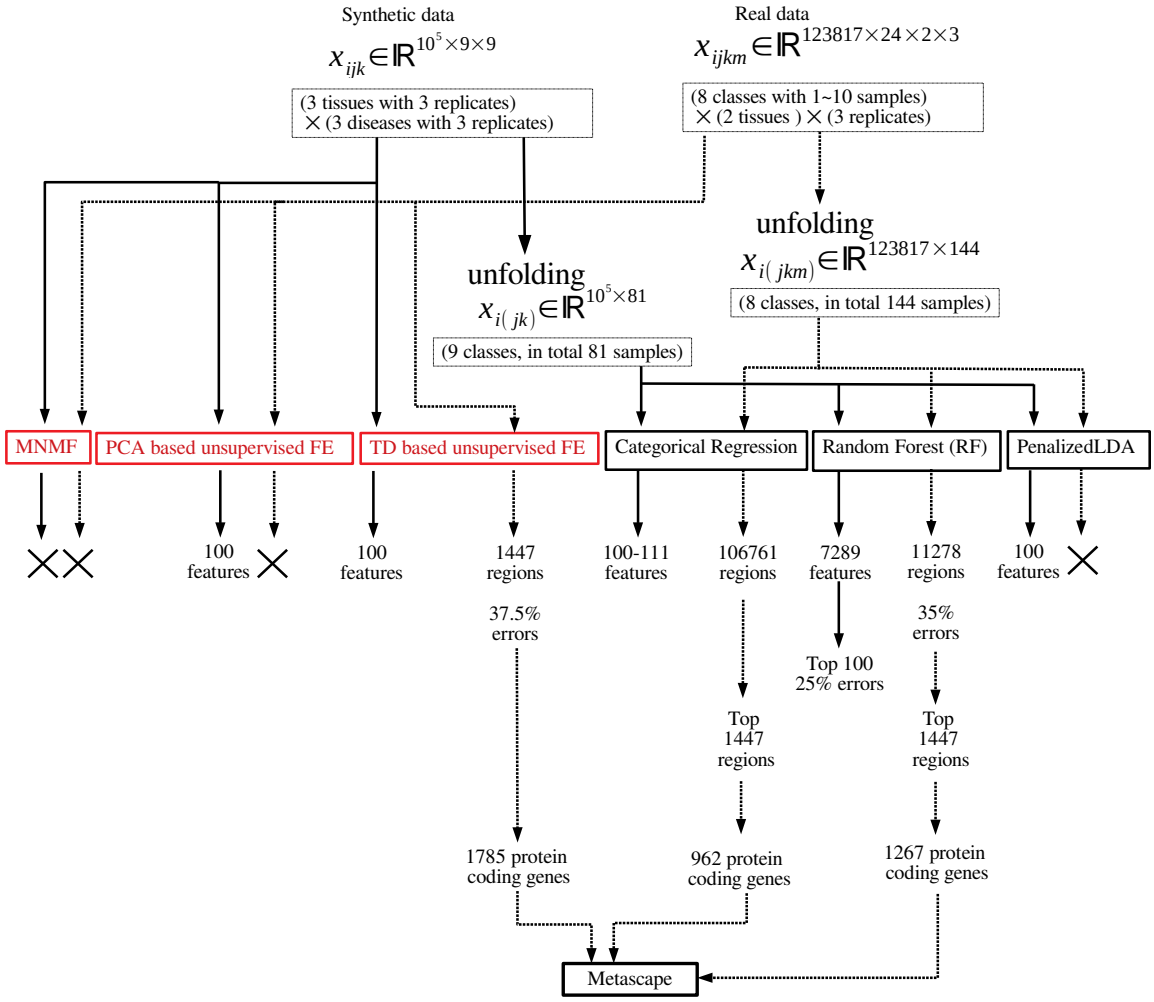

Figure 1. Flowchart of analyses performed in this study. Those in red and black are unsupervised and supervised feature selection, respectively.

Table 3. Confusion matrices when categorical regression is applied to synthetic data. P is the threshold P-value. Features are selected when adjusted P-values less than the threshold P-values are associated with them. Averaged over 100 independent trials.

P	0.1		0.05		0.01		0.001	
	$i > 100$	$i \leq 100$	$i > 100$	$i \leq 100$	$i > 100$	$i \leq 100$	$i > 100$	$i \leq 100$
not selected	99889	0	99895	0	99899	0	99900	0
selected	11	100	5	100	1	100	0	100

although no features with subclass dependence are correctly selected for $\lambda = 0.03$ and 0.04. If we do not have prior knowledge that the correct number is 100, it is impossible to tune the optimal parameter. Thus, in this sense, PenalizedLDA cannot be regarded as a perfect method, although it can achieve perfect performance if we can successfully tune the optimal parameter. Categorical regression is far superior to PenalizedLDA, as it achieves good performances regardless of the threshold P-value (Table 3). It exhibits a maximum of just 11 false positives even if the threshold P-values vary from 0.01 to 0.1. This is in significant contrast to PenalizedLDA that requires sophisticated parameter tuning.

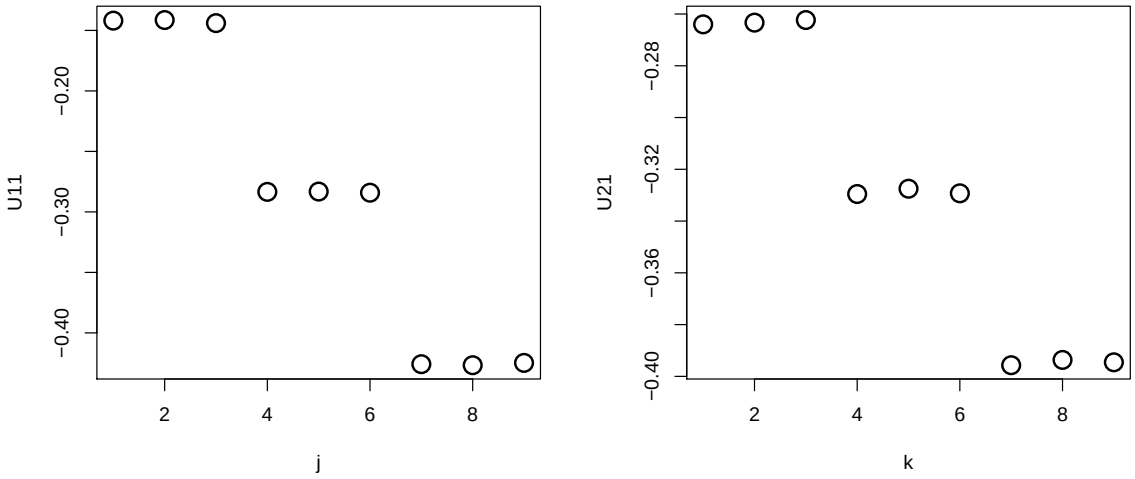

Figure 2. u_{1j} and u_{1k} obtained when HOSVD is applied to the synthetic dataset. Averaged over 100 independent trials.

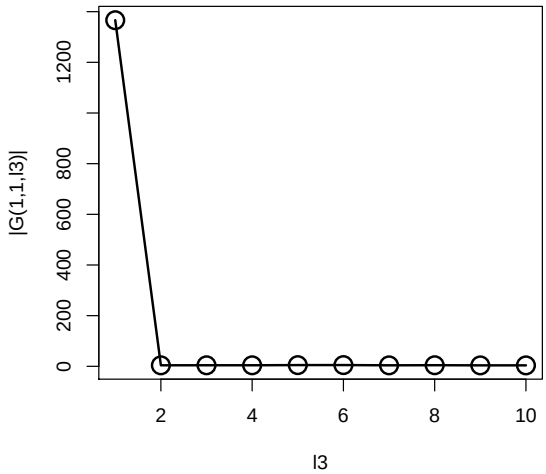

Figure 3. $G(1, 1, \ell_3)$ obtained when HOSVD is applied to the synthetic dataset. Averaged over 100 independent trials.

Nevertheless, TD-based unsupervised FE generated good performance results when compared against categorical regression. Initially, after applying HOSVD to the synthetic dataset, we noticed that u_{1j} and u_{2k} were associated with distinction between the three subclasses (Fig. 2). Thus, we attempted to find which ℓ_3 is associated with $G(1, 1, \ell_3)$ of the largest absolute value (Fig. 3). From this, it was obvious that $G(1, 1, 1)$ had the largest absolute values. Equation (6) with $\ell_3 = 1$ was used to attribute P_i to the ith feature. P_i was corrected by the BH criterion and features associated with adjusted P-values less than the threshold values were selected (Table 4). As a result, TD-based unsupervised FE achieved perfect performances regardless of the threshold P-values.

Table 4. Confusion matrices when TD-based unsupervised FE is applied to the synthetic data. P is the threshold P-value. Features are selected when adjusted P-values less than the threshold P-values are associated with them. $\left\lceil \frac{j}{3} \right\rceil \left(\left\lceil \frac{k}{3} \right\rceil + 3 \right)$. Averaged over 100 independent trials.

P	0.1		0.05		0.01		0.001	
	$i > 100$	$i \leq 100$	$i > 100$	$i \leq 100$	$i > 100$	$i \leq 100$	$i > 100$	$i \leq 100$
not selected	99900	0	99900	0	99900	0	99900	0
selected	0	100	0	100	0	100	0	100

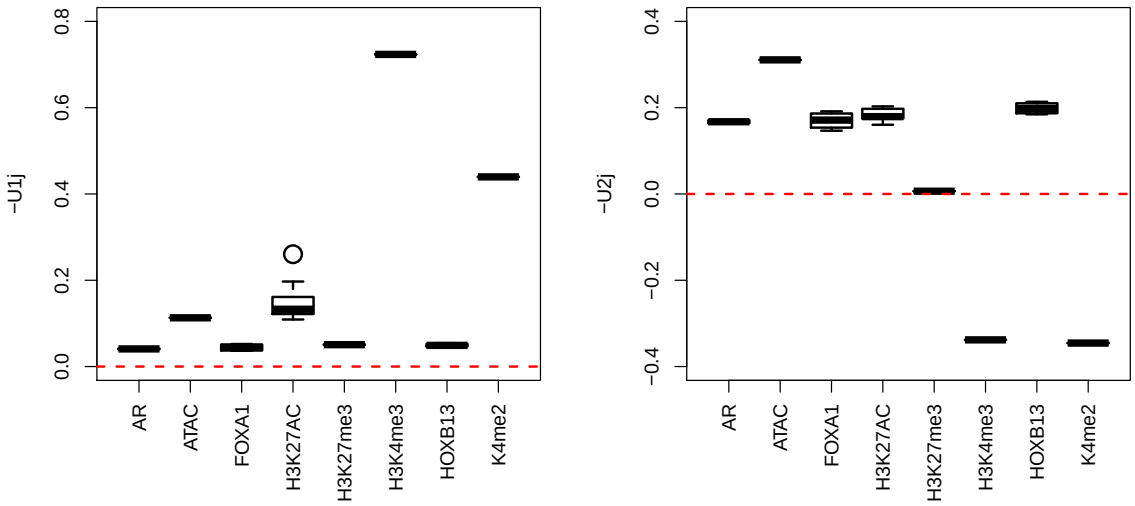

Figure 4. Boxplot of u_{1j} and u_{2j} when HOSVD is applied to the real dataset.

3.2. Real data

Next, we apply TD-based unsupervised FE to a real dataset, as it was shown to be the best of the four methods tested on synthetic data in the previous subsection. After applying HOSVD to x_{ijkm}, we realized that u_{1j} and u_{2j} have significant biological dependence upon multiomics measurement groups (Fig. 4). To explain why u_{1j} and u_{2j} are biologically reliable, we need to introduce genomic science briefly as follows:

A genome, which is a sequence of DNA, stores the information to be translated into a protein. Because all cells store the same genome, the distinct functions and structures of individual cells are controlled by whichever parts of the genome are translated into proteins through transcription to RNA. This process is controlled by the various proteins that bind to genomes. In these measurements, the amounts of three such proteins, AR, FOXA1, and HOXB13, are measured. Another factor that controls this process is the small molecules that bind to histone. Histone is a core protein around which DNA wraps itself, and the tightness of this wrapping is the factor that controls the process from DNA to RNA. H3K27AC, H3K27me3, H3K4me3, and K4me2 are the measures of which small molecules bind to which parts of histone. Finally, ATAC-seq measures how "open" the DNA is. As DNA becomes more open, it can produce more RNA. Thus, the number of proteins that bind to DNA, the number of small molecules that bind to histone, and ATAC-seq are the measures of how many genomes are transcribed to RNA. The difficulty in multiomics analysis is to identify the combinations of proteins binding to DNA and small molecules binding to histone that can enhance the openness of DNA, which is assumed to be observed by ATAC-seq. As we do not know what combination is likely to be observed, the supervised approach is not easy to employ.

We now explain why u_{1j} and u_{2j} are biologically reliable. H3K4me3 and H3K27ac are known to be active/open chromatin marks, whereas H3K27me3 is known to be an inactive mark [42]. u_{1j}

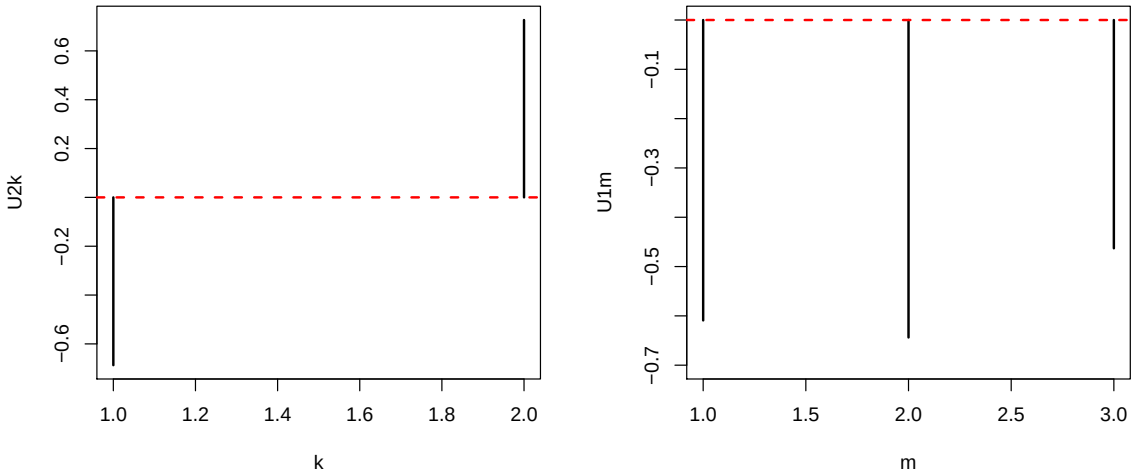

Figure 5. u_{2k} and u_{1m} when HOSVD is applied to the real dataset.

has higher values for H3K4me3 and H3K27ac but smaller values for H3K27me3, with larger values for ATAC-seq, which measures the amount of openness of genome. Thus, the dependences of u_{1j} upon H3K4me3, H3K27ac, H3K27me3, and ATAC-seq are reliable. H3K4me2 [43] is also known to activate gene transcription in tissue-specific ways. The overall dependence of u_{1j} upon multiomics measurement groups is reasonable. Nonetheless, three proteins, AR [44], FOXA1 [45], and HOXB13 [46], are also known to mediate prostate cancer progression, and their enrichment corresponded with u_{2j}. We decided that u_{1j} and u_{2j} successfully captured the biological aspects of multiomics measurements.

In contrast, u_{2k} represents the distinction between a tumor ($k = 1$) and normal prostate ($k = 2$), while u_{1m} represents the independence of biological replicates (Fig. 5). As we are seeking multiomics measurements that are distinct between a tumor and normal prostate as well as common between biological replicates, u_{2j} and u_{1m} represent the properties that we seek. Although we can select either $\ell_1 = 1$ or $\ell_1 = 2$, as both are biologically reliable, we select $\ell_1 = 1$ because it is the first (i.e., mathematically primary) factor. To find the $u_{\ell_4 i}$ used for attributing P-values to genes i, we seek $G(1, 2, 1, \ell_4)$ with the largest absolute value (Fig. 6).

It is obvious that $G(1, 2, 1, 8)$ has the largest absolute value. Hence, u_{8i} is employed to assign P-values to the ith genomic regions using (7) with $\ell_4 = 8$. The P_i values are corrected using the BH criterion, and 1447 genomic regions associated with adjusted P_i less than 0.01 are selected from the total 123,817 genomic regions. This indicates that TD-based unsupervised FE selects approximately 2% of genomics regions of the whole human genome.

Various biological evaluations of selected regions are possible, one of which is the number of protein-coding regions included in these 1447 selected regions. As the genomic regions that include protein-coding genes comprise less than a few percent of the human genome, and the total number of human protein-coding genes is up to 2×10^4, the expected number of protein-coding genes included in these 1447 selected genomic regions is at most a few hundred. Nevertheless, as many as 1785 protein-coding genes can be counted in these regions, which is much higher than expected. This indicates that TD-based unsupervised FE can select genomic regions that include protein-coding genes, correctly considering the altered multiomics variables between normal and tumor tissues, although non-coding RNAs (ncRNAs) have also a key role in regulating the behavior of cells and their over- and underexpression strongly correlated with cancer.

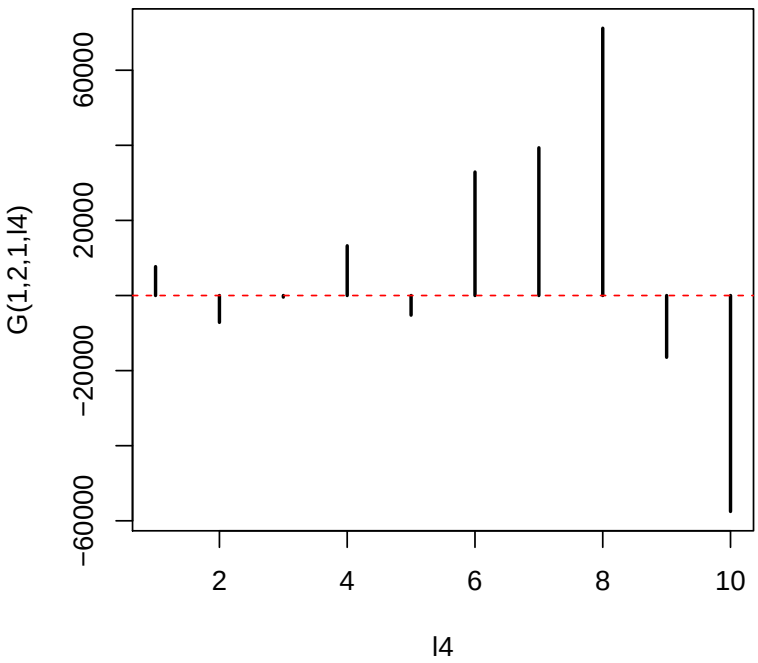

Figure 6. $G(1, 2, 1, \ell_4)$ when HOSVD is applied to the real dataset.

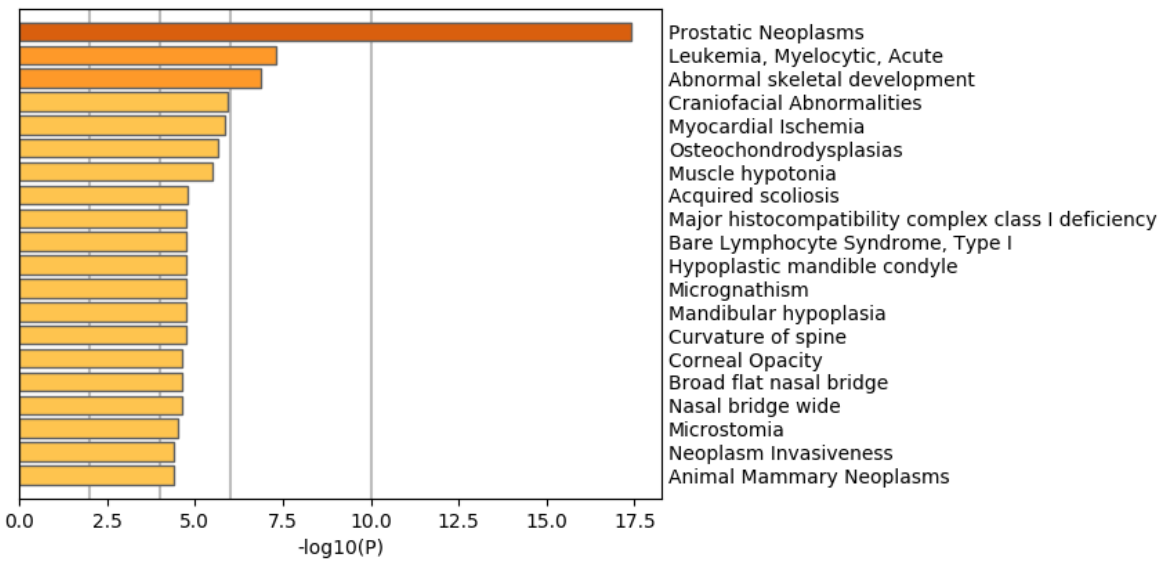

Figure 7. Summary of enrichment of the DisGeNet category of Metascape when 1785 protein-coding genes included in 1447 genomic regions selected by TD-based unsupervised FE are considered.

Another evaluation of these selected 1447 genomic regions is the biological properties of the 1785 protein-coding genes included in these 1447 genomic regions. If biologically reliable protein-coding genes are selected, they should be enriched with the genes related to prostate cancer. To confirm this point, we uploaded these genes to Metascape [47], which evaluates the enrichment of various biological terms with reducing redundancy. Figure 7 shows the results of the DisGeNet [48] category of Metascape. It is obvious that prostatic neoplasms are not only top ranked but also outstandingly

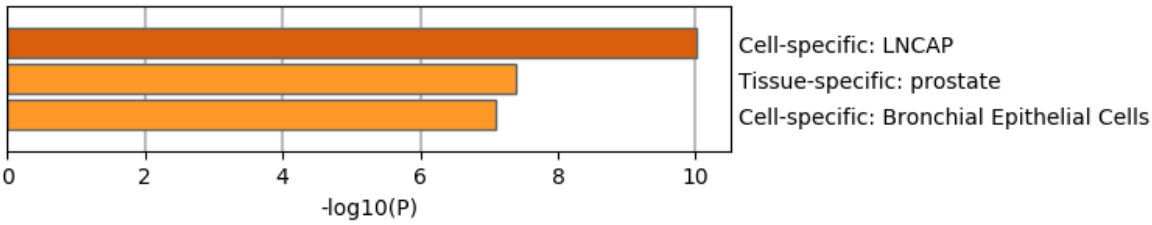

Figure 8. Summary of enrichment of the PaGenBase category of Metascape when 1785 protein-coding genes included in 1447 genomic regions selected by TD-based unsupervised FE are considered.

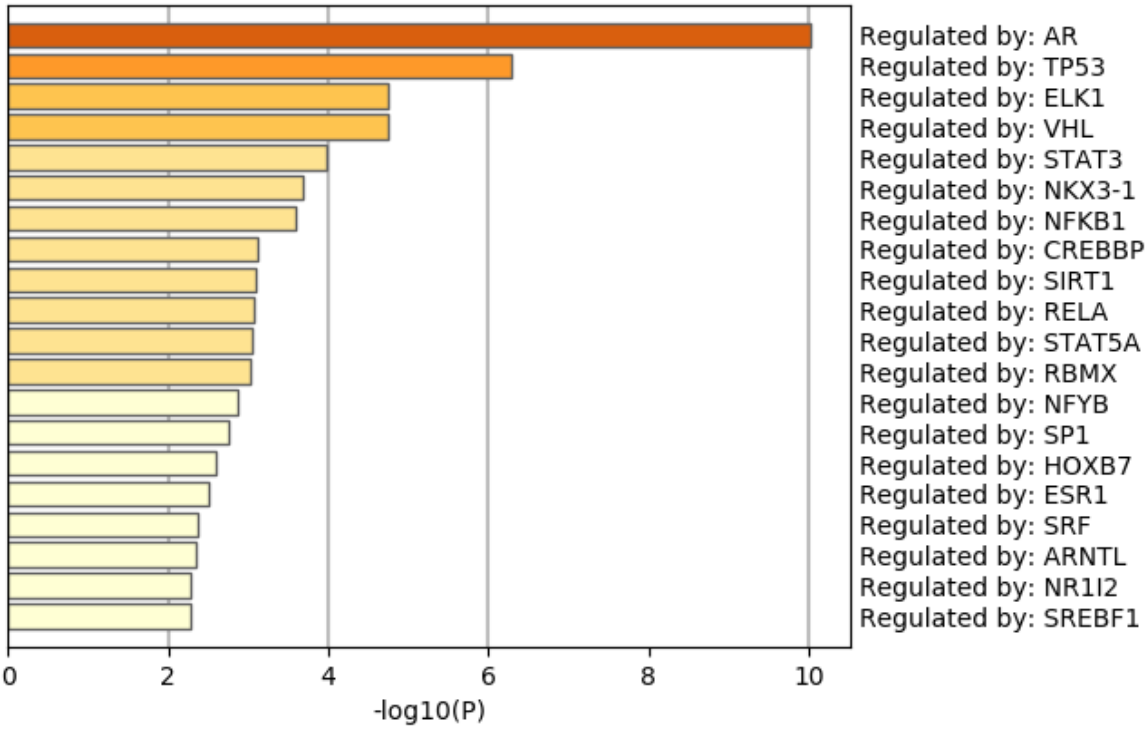

Figure 9. Summary of enrichment of the TRRUST category of Metascape when 1785 protein-coding genes included in 1447 genomic regions selected by TD-based unsupervised FE are considered.

enriched. This suggests that TD-based unsupervised FE can specifically select genes related to prostate cancer. Figure 8 shows the results of the PaGenBase [49] category of Metascape. It is obvious that LNCaP cells are not only top ranked but also outstandingly enriched. LNCAP [50] is a model cell line of a human prostatic carcinoma. This also suggests that TD-based unsupervised FE can specifically select genes related to prostate cancer. Figure 9 shows the results of the TRRUST [51] category of Metascape. It is obvious that the androgen receptor (AR) is not only top ranked but also outstandingly enriched. As mentioned above, AR is a protein that binds to the genome and mediates prostate cancer progression. This also suggests that TD-based unsupervised FE can specifically select genes related to prostate cancer. Although there are more convincing results, the supplementary materials include the full report provided by Metascape.

Discussion

Synthetic data

To determine whether TD-based unsupervised FE outperformed the three supervised feature selections because of its unsupervised nature, we also applied an alternative unsupervised feature

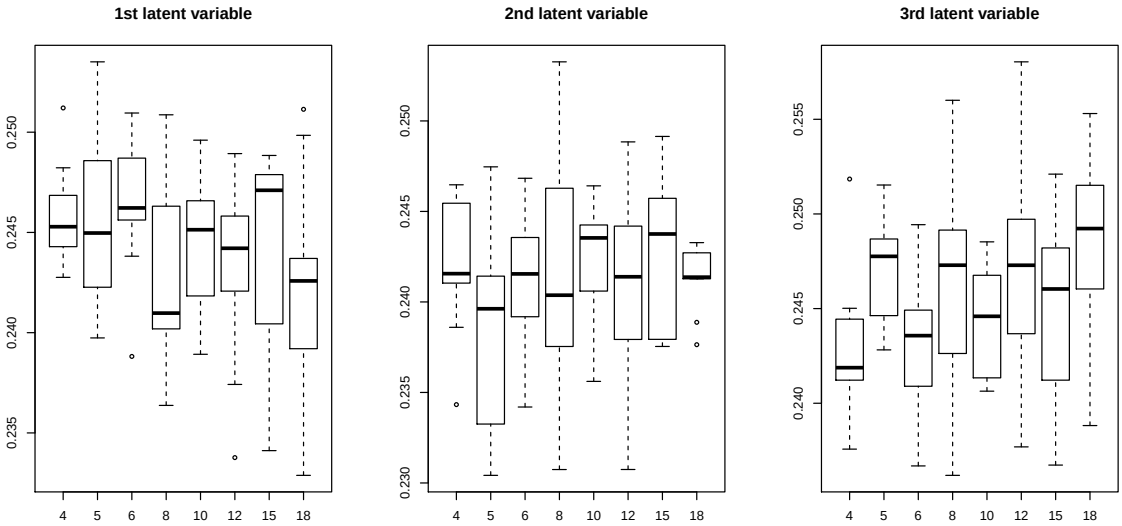

Figure 10. Boxplot of the first to third latent variable vectors attributed to the 81 samples. Horizontal axis: $\left\lceil \frac{j}{3} \right\rceil \left(\left\lceil \frac{k}{3} \right\rceil + 3 \right), 1 \leq j,k \leq 3$. Averaged over 100 independent trials.

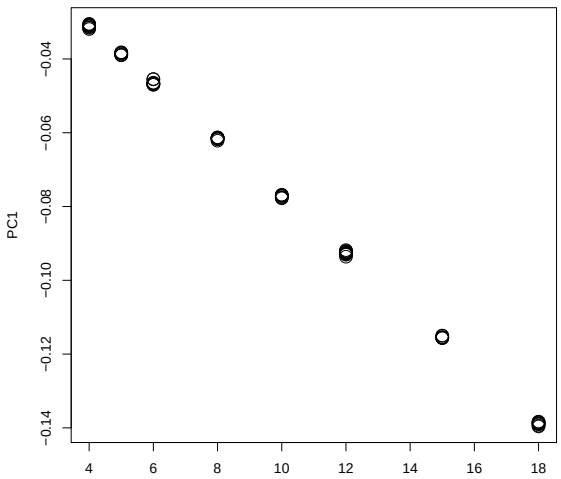

Figure 11. The 1st PC loading computed by PCA applied to synthetic data. Horizontal axis: $\left\lceil \frac{j}{3} \right\rceil \left(\left\lceil \frac{k}{3} \right\rceil + 3 \right), 1 \leq j,k \leq 3$. Averaged over 100 independent trials.

selection, MNMF, to the synthetic data, specifying $n = 3$ as the number of vectors each of which is composed of latent variables. Figures 10 shows the latent variable vectors attributed to the samples and *i*s, respectively, computed by MNMF. Up to the third latent vector, none is coincident with 9 classes defined in eq. (1). Thus, it is obvious that MNMF is inferior to TD based unsupervised FE. Thus the reason why TD based unsupervised FE could outperform other supervised methods is not simply because it is a unsupervised method.

In order to see if TD based unsupervised FE is a only unsupervised method that can outperform other supervised method, we also tried PCA based unsupervised FE [17] from which TD based unsupervised FE developed. It is obvious that it is well coincident with : $\left\lceil \frac{j}{3} \right\rceil \left(\left\lceil \frac{k}{3} \right\rceil + 3 \right), 1 \leq j,k \leq 3$. Then we attributed P-values to i using eq. (10) with $\ell = 1$. Table 5 shows the confusion matrix when PCA based unsupervised FE was applied to synthetic data. It is identical to Table 4. Thus Thus,

Table 5. Confusion matrices when PCA-based unsupervised FE is applied to the synthetic data. P is the threshold P-value. Features are selected when adjusted P-values less than the threshold P-values are associated with them. $\left\lceil \frac{j}{3} \right\rceil \left(\left\lceil \frac{k}{3} \right\rceil + 3 \right)$. Averaged over 100 independent trials.

P	0.1		0.05		0.01		0.001	
	$i > 100$	$i \leq 100$	$i > 100$	$i \leq 100$	$i > 100$	$i \leq 100$	$i > 100$	$i \leq 100$
not selected	99900	0	99900	0	99900	0	99900	0
selected	0	100	0	100	0	100	0	100

Table 6. CPU times of the five methods required for the synthetic and real datasets. The CPU time for PenalizedLDA is for testing four λ values with between one and eight discriminant functions. The CPU time for an individual run is 18 s. Cpu times for synthetic data were averaged over 100 independent trials.

	methods	PCA based unsupervised FE	TD based unsupervised FE	categorical regression	random forest	penalized LDA	MNMF
cpu time [sec]	synthetic data	1.55	3.5	51.5	118.2	283 (18)	290
	real data	17	20	87	321	—	223

reported results show that TD and PCA-based unsupervised FE outperforms the other six methods tested on the synthetic dataset.

TD-based unsupervised FE also has the advantage of its CPU time (Table 6). PCA and TD-based unsupervised FE are the fastest of the six methods tested on the synthetic data. It can finish within a few seconds, whereas the others take longer, ranging from a few tens of seconds for categorical regression to a few minutes for RF. In this sense, PCA and TD-based unsupervised FE are the fastest, most accurate, and most robust of the six methods tested on the synthetic dataset.

Real data

We now compare the performances of TD-based unsupervised FE with those of the other five methods: RF, PenalizedLDA, categorical regression, MNMF and PCA based unsupervised FE. First, RF was applied to the real dataset, which is classified into 16 categorical classes composed of any pairs of eight mulitiomics measurement groups and two tissues (tumor and normal prostate). RF results in as many as 11,278 genomic regions having non-zero importance. This number is almost eight times larger than the number of regions selected by TD-based unsupervised FE (1447). One of the evaluations of RF performance is its accuracy of discrimination, as RF is a supervised method. RF should have selected the minimum features that can successfully discriminate between the 16 classes. The error rate given by RF was as small as 0.35 (Table 7). This indicates that 65% of the 144 samples were correctly classified into the 16 classes (see Table 2 for details of how the 144 samples were classified into the 16 subclasses). To validate the discrimination performances of the 1447 genomic regions selected by TD-based unsupervised FE, we carried out the following. u_{1j}, u_{2k}, u_{3k} were recomputed by HOSVD using only the 1447 genomic regions selected by TD-based unsupervised FE. The 144 samples were then discriminated using linear discriminant analysis, employing the product $u_{1j}u_{2k}u_{3k}$ as input vectors (Table 8 with leave-one-out cross-validation). The error rate was as small as 37.5%, similar to that of RF (0.35). Considering the fact that TD-based unsupervised FE employed only one-eighth of the number of genomic regions selected by RF, we concluded that TD-based unsupervised FE can outperform RF.

In spite of the comparisons presented in the above, one might wonder that RF can achieve comparative performances if top ranked restricted number of genomic regions are intentionally selected. In order to this, we selected 1447 top ranked regions (they are as many as those selected by TD based unsupervised FE) and identified as many as 1267 gene symbols included in these genomic regions. These 1267 genes are uploaded to Metascape to evaluate them biologically. Figure 12 shows the results of the DisGeNet category of Metascape. In contrast to Fig. 7 where prostatic neoplasms was

Table 7. Confusion matrix for RF applied to the real dataset. N: normal prostate, T: tumor. The error rate is 35%.

	AR		ATAC		FOXA1		H3K27AC		H3K27me3		H3K4me3		HOXB13		K4me2		class
	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	error
AR_N	5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1/6
AR_T	0	0	0	0	1	3	1	0	0	0	0	0	0	1	0	0	1
ATAC_N	0	0	0	2	0	0	0	1	0	0	0	0	0	0	0	0	1
ATAC_T	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	2/3
FOXA1_N	1	0	0	0	9	0	0	0	0	0	0	0	2	0	0	0	1/4
FOXA1_T	0	0	0	0	4	5	0	0	0	0	0	0	0	3	0	0	7/12
H3K27AC_N	0	0	0	0	0	0	26	4	0	0	0	0	0	0	0	0	2/15
H3K27AC_T	0	0	0	0	0	0	6	24	0	0	0	0	0	0	0	0	1/5
H3K27me3_N	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	2/3
H3K27me3_T	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	1/3
H3K4me3_N	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	1
H3K4me3_T	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	1
HOXB13_N	0	0	0	0	2	0	0	0	0	0	0	0	9	1	0	0	1/4
HOXB13_T	0	0	0	0	0	0	0	0	0	0	0	0	1	11	0	0	1/12
K4me2_N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	1
K4me2_T	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	0	1

Table 8. Confusion matrix for LDA using $u_{1j}u_{2k}u_{3m}$ given by TD applied to 1447 genomic regions selected by TD-based unsupervised FE applied to the real dataset. N: normal prostate, T: tumor. The error rate is 37.5%.

	AR		ATAC		FOXA1		H3K27AC		H3K27me3		H3K4me3		HOXB13		K4me2		class
	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	error
AR_N	0	0	0	0	5	0	0	0	0	0	0	0	1	0	0	0	1
AR_T	0	0	0	0	0	5	0	0	0	0	0	0	0	1	0	0	1
ATAC_N	0	0	0	0	0	0	2	0	0	0	1	0	0	0	0	0	1
ATAC_T	0	0	0	0	0	0	0	2	0	0	0	1	0	0	0	0	1
FOXA1_N	0	0	0	0	9	0	1	0	0	0	0	0	2	0	0	0	1/4
FOXA1_T	0	0	0	0	0	9	0	1	0	0	0	0	0	2	0	0	1/4
H3K27AC_N	0	0	2	0	0	0	25	0	0	0	1	0	2	0	0	0	1/6
H3K27AC_T	0	0	0	2	0	0	0	25	0	0	0	1	0	2	0	0	1/6
H3K27me3_N	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0
H3K27me3_T	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0
H3K4me3_N	0	0	1	0	0	0	0	0	0	0	2	0	0	0	0	0	1/3
H3K4me3_T	0	0	0	1	0	0	0	0	0	0	0	2	0	0	0	0	1/3
HOXB13_N	0	0	0	0	4	0	2	0	0	0	0	0	6	0	0	0	1/2
HOXB13_T	0	0	0	0	0	4	0	2	0	0	0	0	0	6	0	0	1/2
K4me2_N	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	1
K4me2_T	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	1

top ranked, it was even not ranked at all in Figure 12. Thus, it is obvious that genes selected by RF is inferior to those by TD based unsupervised FE. Figure 13 shows the results of the PaGenBase category of Metascape. In cotrast to Fig. 8 where prostate related cell lines are top ranked, no prostate related cell lines are top ranked in Fig. 13. Thus, again, it is obvious that genes selected by RF is inferior to those by TD based unsupervised FE. Figure 14 shows the results of the TRRUST category of Metascape. In contrast to Fig. 9 where AR is top ranked, AR is not ranked at all. Thus, it is obvious that genes selected by RF is inferior to those by TD based unsupervised FE as well.

Next, we attempted to apply PenalizedLDA to the real dataset, but found that we could not. To perform LDA, all in-subclass variance must be non-zero. This condition was not fulfilled for the real data. Hence, PenalizedLDA cannot be employed to process real datasets.

Third, we applied categorical regression to the real dataset by regarding it as 16 categorical subclasses. The number of features associated with adjusted P-values less than 0.01 was as high as 106,761 among all 123,817 features. This indicates that categorical regression selected almost all features and is thus clearly ineffective. The reason categorical regression failed, despite being relatively successful (Table 3) when applied to the synthetic dataset, is as follows. In the synthetic dataset, there was only one dependence upon subclasses. Selecting features with subclass dependence automatically gives us targeted features. In the real dataset, this cannot stand as it is. There are many variations of subclass dependence. In TD-based unsupervised FE, by assessing the dependence of singular value vectors upon subclasses (see Fig. 4), we can specify the type of dependence upon subclasses that

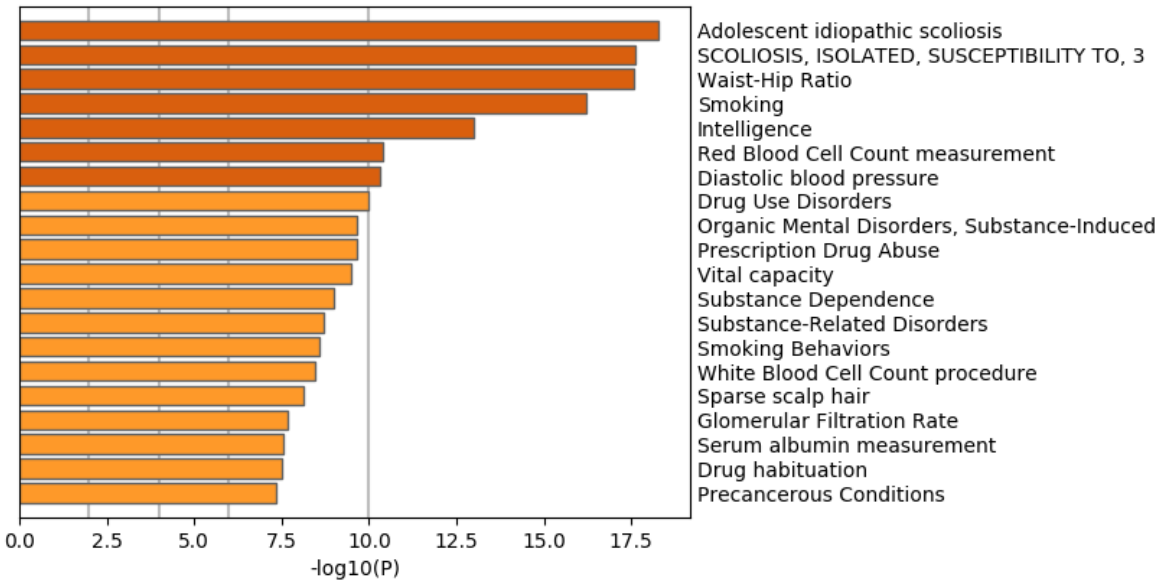

Figure 12. Summary of enrichment of the DisGeNet category of Metascape when 11267 protein-coding genes included in 1447 genomic regions selected by RF are considered.

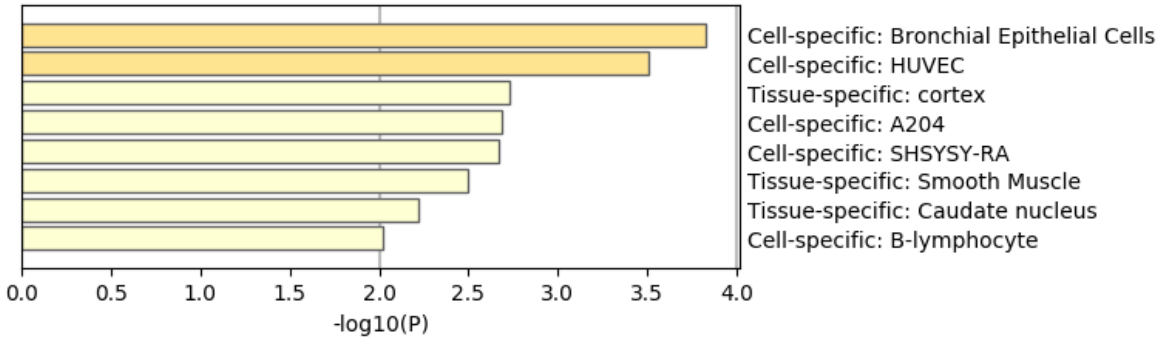

Figure 13. Summary of enrichment of the PaGenBase category of Metascape when 1267 protein-coding genes included in 1447 genomic regions selected by RF are considered.

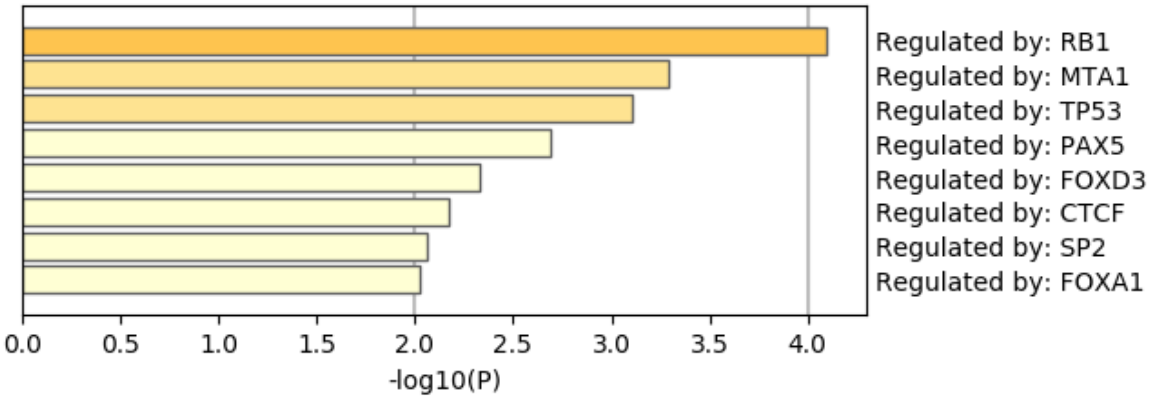

Figure 14. Summary of enrichment of the TRRUST category of Metascape when 1267 protein-coding genes included in 1447 genomic regions selected by RF are considered.

should be considered. This allows us to identify a restricted number of biologically reliable genomic regions. Categorical regression cannot distinguish between various subclass dependences and can only identify genomic regions associated with any type of subclass dependence. This results in the

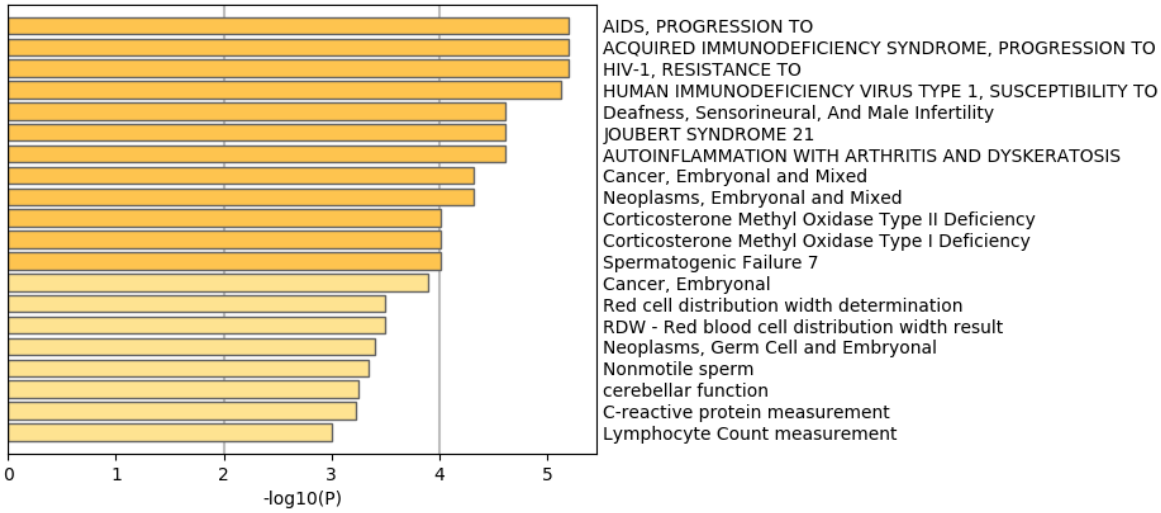

Figure 15. Summary of enrichment of the DisGeNet category of Metascape when 962 protein-coding genes included in 1447 genomic regions selected by categorical regression are considered.

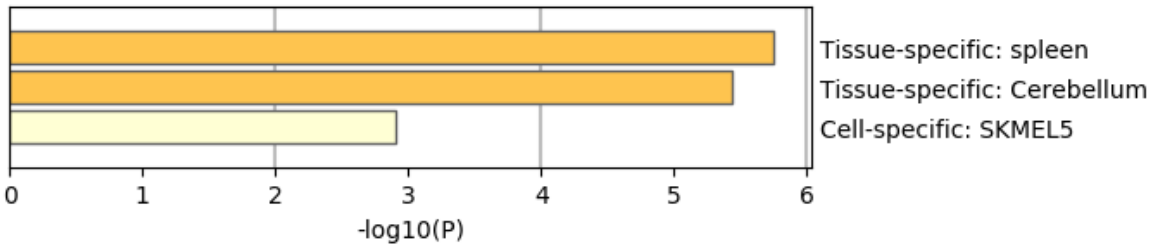

Figure 16. Summary of enrichment of the PaGenBase category of Metascape when 962 protein-coding genes included in 1447 genomic regions selected by categorical regression are considered.

identification of almost all genomic regions, as most are likely to be associated with a type of subclass dependence.

In spite of the comparisons presented in the above, one might wonder that categorical regression can achieve comparative performances if top ranked restricted number of genomic regions are intentionally selected. In order to this, we selected 1447 top ranked regions (they are as many as those selected by TD based unsupervised FE) and identified as many as 962 gene symbols included in these genomic regions. These 1267 genes are uploaded to Metascape to evaluate them biologically. Figure 15 shows the results of the DisGeNet category of Metascape. In contrast to Fig. 7 where prostatic neoplasms was top ranked, it was even not ranked at all in Figure 15. Thus, it is obvious that genes selected by RF is inferior to those by TD based unsupervised FE. Figure 16 shows the results of the PaGenBase category of Metascape. In cotrast to Fig. 8 where prostate related cell lines are top ranked, no prostate related cell lines are top ranked in Fig. 16. Thus, again, it is obvious that genes selected by RF is inferior to those by TD based unsupervised FE. In contrast to Fig. 9 where AR is top ranked, we could find no significant enrichment in TRRUST category of Metascape. Thus, it is obvious that genes selected by RF is inferior to those by TD based unsupervised FE as well.

Finally, MNMF ($n = 3$), one of the state-of-the-art methods that are often applied to multiomics datasets [19], and PCA were applied to x_{ijkm} as described in Materials and Methods. Figs. 17 and 18 show the latent variable vectors and PC loading attributed to 144 samples, respectively. Although the dependence upon eight multiomics measurements are similar to those of TD-based unsupervised FE, because the correspondence with eight types of measurement is clearly inferior to that of TD-based unsupervised FE (within class variances much larger than those of the singular value vectors obtained by TD-based unsupervised FE, Fig. 4), we employ TD-based unsupervised FE. An additional reason

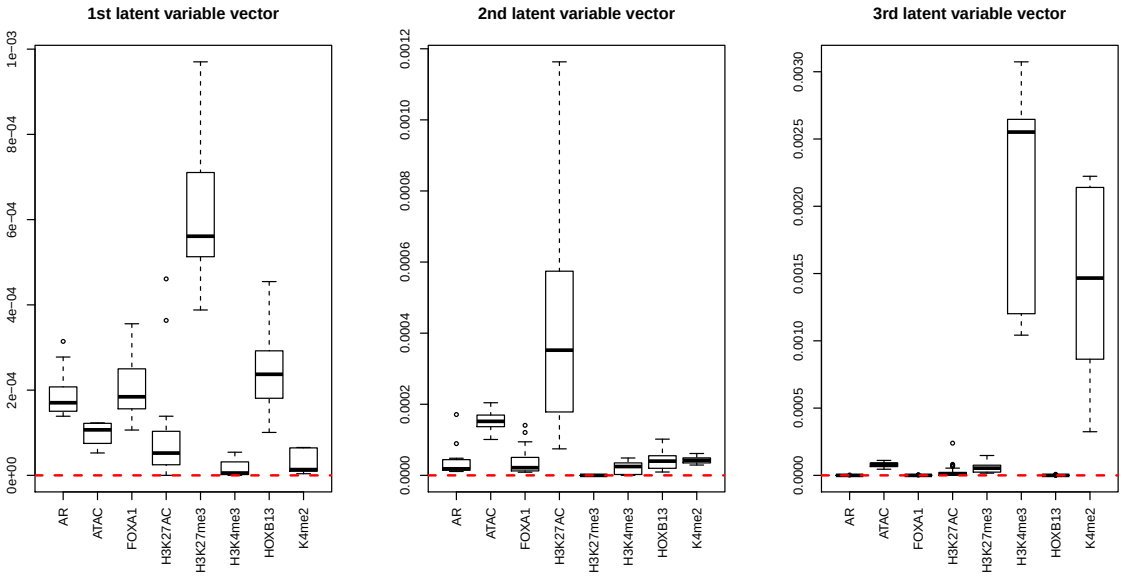

Figure 17. Boxplots of the first to third latent variable vectors attributed to 144 samples classified into eight subclasses, each of which corresponds to a multiomics measurement. The horizontal red broken lines are base lines (zero).

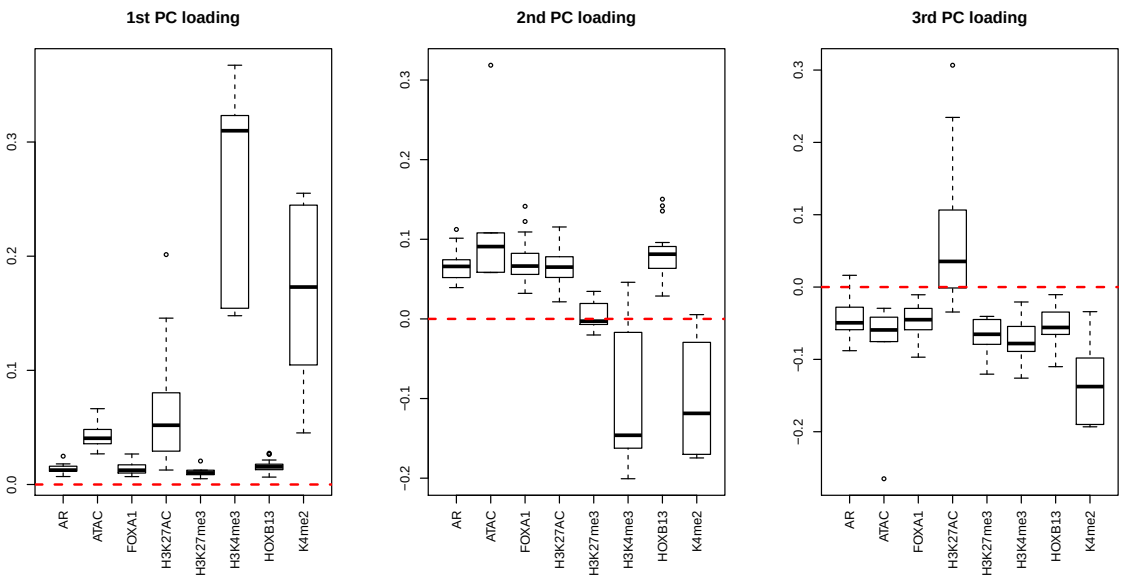

Figure 18. Boxplots of the first to third PC loading attributed to 144 samples classified into eight subclasses, each of which corresponds to a multiomics measurement. The horizontal red broken lines are base lines (zero).

f

not to employ MNMF instead of TD-based unsupervised FE is that, although singular value vectors derived by HOSVD can be assumed to obey a Gaussian distribution, latent variables computed by MNMF are unlikely to because they do not take negative values. Thus, there is no way to attribute *P*-values to genes, as there is no suitable null hypothesis by which we can compute the *P*-values.

As a result, none of the five methods (RF, PenalizedLDA, categorical regression, MNMF, or PCA) can compete with TD-based unsupervised FE when applied to a real dataset.

In addition, the CPU time required of PCA and TD-based unsupervised FE are also the shortest compared to the other four methods when applied to the real dataset (Table 6). From this perspective, TD-based unsupervised FE outperforms the other four methods.

Discussions not to specific to either synthetic or real data

Unlike "genomic regions + samples" structure used in traditional unsupervised learning that requires a matrix representation, Our TD employs "genomic regions + samples + tissues + biological replicates" structure that requires a four-axes tensor representation. Reported results via enrichment analysis show the superiority of our TD, attributed to taking into account the relationships among genomic regions, samples, tissues, and biological replicates at once. One might wonder why TD based unsupervised FE is more coincident with subclasses than other supervised or unsupervised methods. It is simple because of the nature of tensor. As can be seen in eqs. (4) and (5), using TD, dependence of x_{ijk} or x_{ijkm} upon i, j, k or i, j, k, m is decomposed. Thus distinction between i, j, k, m or i, j, k is specifically identified. This means that we can identify distinction of x_{ijkm} or x_{ijk} between distinct i independent of other index, j, k or j, k, m. Such a detection is impossible since the distinction between $x_{ijk'}$ and $x_{ij'k}$ or that between $x_{ijk'm}$ and $x_{ij'km}$ where more than one index are altered simultaneously. This weaken the ability for methods other than TD based unsupervised FE to identify subclass dependence specifically. This is possibly because TD based unsupervised FE can outperform other methods on the identification of subclasses.

Conclusions

In this study, a TD-based unsupervised FE formalism is successfully applied for the first time to a variety of multiomics measurements for both synthetic and real datasets that represented a typical large p small n problem. The proposed method outperformed the conventional supervised feature selection methods of RF, categorical regression, PenalizedLDA, and MNMF, and was demonstrated to be superior not only in feature selection but also in selection of biologically reliable genes.

Acknowledgment

This study was supported by KAKENHI, 20K12067, 20H04848, and 19H05270. This project was also funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, under grant no. KEP-8-611-38. The authors, therefore, acknowledge DSR with thanks for providing technical and financial support.

Supplementary data

Output from Metascape.

References

1. Richter, A.N.; Khoshgoftaar, T.M. Efficient learning from big data for cancer risk modeling: a case study with melanoma. *Computers in biology and medicine* **2019**, *110*, 29–39.
2. Awan, M.G.; Eslami, T.; Saeed, F. GPU-DAEMON: GPU algorithm design, data management & optimization template for array based big omics data. *Computers in biology and medicine* **2018**, *101*, 163–173.
3. Nashaat, M.; Ghosh, A.; Miller, J.; Quader, S.; Marston, C.; Puget, J.F. Hybridization of active learning and data programming for labeling large industrial datasets. 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018, pp. 46–55.
4. Shah, R.; Zhang, S.; Lin, Y.; Wu, P. xSVM: Scalable Distributed Kernel Support Vector Machine Training. 2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019, pp. 155–164.
5. Bekkerman, R.; Bilenko, M.; Langford, J. *Scaling up machine learning: Parallel and distributed approaches*; Cambridge University Press, 2011.
6. Chatterjee, A.; Gupta, U.; Chinnakotla, M.K.; Srikanth, R.; Galley, M.; Agrawal, P. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior* **2019**, *93*, 309–317.
7. Ngiam, K.Y.; Khor, W. Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology* **2019**, *20*, e262–e273.
8. Santosh, T.; Ramesh, D.; Reddy, D. LSTM based prediction of malaria abundances using big data. *Computers in Biology and Medicine* **2020**, p. 103859.

9. Ge, J.; Li, X.; Jiang, H.; Liu, H.; Zhang, T.; Wang, M.; Zhao, T. Picasso: A Sparse Learning Library for High Dimensional Data Analysis in R and Python. *J. Mach. Learn. Res.* **2019**, *20*, 44–1.
10. Wen, F.; Chu, L.; Ying, R.; Liu, P. Fast and Positive Definite Estimation of Large Covariance Matrix for High-Dimensional Data Analysis. *IEEE Transactions on Big Data* **2019**.
11. Yang, S.; Wen, J.; Zhan, X.; Kifer, D. ET-lasso: a new efficient tuning of lasso-type regularization for high-dimensional data. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 607–616.
12. MEI, B.; WANG, Z. An efficient method to handle the 'large p, small n' problem for genomewide association studies using Haseman–Elston regression. *Journal of Genetics* **2016**, *95*, 847–852. doi:10.1007/s12041-016-0705-3.
13. Johnstone, I.M.; Titterington, D.M. Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2009**, *367*, 4237–4253, [https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2009.0159]. doi:10.1098/rsta.2009.0159.
14. Zhang, M.; Zhang, D.; Wells, M.T. Variable selection for large p small n regression models with incomplete data: Mapping QTL with epistases. *BMC Bioinformatics* **2008**, *9*. doi:10.1186/1471-2105-9-251.
15. Huynh, P.H.; Nguyen, V.H.; Do, T.N. Improvements in the Large p, Small n Classification Issue. *SN Computer Science* **2020**, *1*. doi:10.1007/s42979-020-00210-2.
16. Hood, L.; Rowen, L. The human genome project: big science transforms biology and medicine. *Genome Medicine* **2013**, *5*, 79. doi:10.1186/gm483.
17. Taguchi, Y.H. *Unsupervised Feature Extraction Applied to Bioinformatics*; Springer International Publishing, 2020. doi:10.1007/978-3-030-22456-1.
18. Witten, D.M.; Tibshirani, R. Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2011**, *73*, 753–772.
19. Baldwin, E.; Han, J.; Luo, W.; Zhou, J.; An, L.; Liu, J.; Zhang, H.H.; Li, H. On fusion methods for knowledge discovery from multi-omics datasets. *Computational and Structural Biotechnology Journal* **2020**, *18*, 509–517. doi:10.1016/j.csbj.2020.02.011.
20. Subramanian, I.; Verma, S.; Kumar, S.; Jere, A.; Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and Biology Insights* **2020**, *14*, 1177932219899051, [https://doi.org/10.1177/1177932219899051]. PMID: 32076369, doi:10.1177/1177932219899051.
21. Vaske, C.J.; Benz, S.C.; Sanborn, J.Z.; Earl, D.; Szeto, C.; Zhu, J.; Haussler, D.; Stuart, J.M. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **2010**, *26*, i237–i245, [https://academic.oup.com/bioinformatics/article-pdf/26/12/i237/16893608/btq182.pdf]. doi:10.1093/bioinformatics/btq182.
22. Shen, R.; Olshen, A.B.; Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **2009**, *25*, 2906–2912, [https://academic.oup.com/bioinformatics/article-pdf/25/22/2906/16891356/btp543.pdf]. doi:10.1093/bioinformatics/btp543.
23. Mo, Q.; Wang, S.; Seshan, V.E.; Olshen, A.B.; Schultz, N.; Sander, C.; Powers, R.S.; Ladanyi, M.; Shen, R. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences* **2013**, *110*, 4245–4250, [https://www.pnas.org/content/110/11/4245.full.pdf]. doi:10.1073/pnas.1208949110.
24. Wu, D.; Wang, D.; Zhang, M.Q.; Gu, J. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC Genomics* **2015**, *16*. doi:10.1186/s12864-015-2223-8.
25. Chin, S.F.; Teschendorff, A.E.; Marioni, J.C.; Wang, Y.; Barbosa-Morais, N.L.; Thorne, N.P.; Costa, J.L.; Pinder, S.E.; van de Wiel, M.A.; Green, A.R.; Ellis, I.O.; Porter, P.L.; Tavar?, S.; Brenton, J.D.; Ylstra, B.; Caldas, C. High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol* **2007**, *8*, R215.
26. Lock, E.F.; Dunson, D.B. Bayesian consensus clustering. *Bioinformatics* **2013**, *29*, 2610–2616.
27. Wang, B.; Mezlini, A.M.; Demir, F.; Fiume, M.; Tu, Z.; Brudno, M.; Haibe-Kains, B.; Goldenberg, A. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* **2014**, *11*, 333–337.

28. Shi, Q.; Zhang, C.; Peng, M.; Yu, X.; Zeng, T.; Liu, J.; Chen, L. Pattern fusion analysis by adaptive alignment of multiple heterogeneous omics data. *Bioinformatics* **2017**, *33*, 2706–2714, [https://academic.oup.com/bioinformatics/article-pdf/33/17/2706/25163881/btx176.pdf]. doi:10.1093/bioinformatics/btx176.
29. Nguyen, H.; Shrestha, S.; Draghici, S.; Nguyen, T. PINSPlus: a tool for tumor subtype discovery in integrated genomic data. *Bioinformatics* **2018**, *35*, 2843–2846, [https://academic.oup.com/bioinformatics/article-pdf/35/16/2843/29154723/bty1049.pdf]. doi:10.1093/bioinformatics/bty1049.
30. Rappoport, N.; Shamir, R. NEMO: cancer subtyping by integration of partial multi-omic data. *Bioinformatics* **2019**, *35*, 3348–3356, [https://academic.oup.com/bioinformatics/article-pdf/35/18/3348/30024606/btz058.pdf]. doi:10.1093/bioinformatics/btz058.
31. Rohart, F.; Gautier, B.; Singh, A.; Lê Cao, K.A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Computational Biology* **2017**, *13*, 1–19. doi:10.1371/journal.pcbi.1005752.
32. Meng, C.; Helm, D.; Frejno, M.; Kuster, B. moCluster: Identifying Joint Patterns Across Multiple Omics Data Sets. *Journal of Proteome Research* **2015**, *15*, 755–765. doi:10.1021/acs.jproteome.5b00824.
33. Meng, C.; Kuster, B.; Culhane, A.C.; Gholami, A. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* **2014**, *15*, 162. doi:10.1186/1471-2105-15-162.
34. O'Connell, M.J.; Lock, E.F. R.JIVE for exploration of multi-source molecular data. *Bioinformatics* **2016**, *32*, 2877–2879, [https://academic.oup.com/bioinformatics/article-pdf/32/18/2877/25416720/btw324.pdf]. doi:10.1093/bioinformatics/btw324.
35. de Tayrac, M.; L?, S.; Aubry, M.; Mosser, J.; Husson, F. Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. *BMC Genomics* **2009**, *10*, 32.
36. Speicher, N.K.; Pfeifer, N. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics* **2015**, *31*, i268–i275, [https://academic.oup.com/bioinformatics/article-pdf/31/12/i268/572569/btv244.pdf]. doi:10.1093/bioinformatics/btv244.
37. Integrated genomic analyses of ovarian carcinoma. *Nature* **2011**, *474*, 609–615. doi:10.1038/nature10166.
38. Edgar, R.; Domrachev, M.; Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **2002**, *30*, 207–210, [https://academic.oup.com/nar/article-pdf/30/1/207/9901036/300207.pdf]. doi:10.1093/nar/30.1.207.
39. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.
40. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22.
41. Witten, D. *penalizedLDA: Penalized Classification using Fisher's Linear Discriminant*, 2015. R package version 1.1.
42. Igolkina, A.A.; Zinkevich, A.; Karandasheva, K.O.; Popov, A.A.; Selifanova, M.V.; Nikolaeva, D.; Tkachev, V.; Penzar, D.; Nikitin, D.M.; Buzdin, A. H3K4me3, H3K9ac, H3K27ac, H3K27me3 and H3K9me3 Histone Tags Suggest Distinct Regulatory Evolution of Open and Condensed Chromatin Landmarks. *Cells* **2019**, *8*. doi:10.3390/cells8091034.
43. Pekowska, A.; Benoukraf, T.; Ferrier, P.; Spicuglia, S. A unique H3K4me2 profile marks tissue-specific gene regulation. *Genome Research* **2010**, *20*, 1493–1502, [http://genome.cshlp.org/content/20/11/1493.full.pdf+html]. doi:10.1101/gr.109389.110.
44. Fujita, K.; Nonomura, N. Role of Androgen Receptor in Prostate Cancer: A Review. *The World Journal of Men's Health* **2019**, *37*, 288. doi:10.5534/wjmh.180040.
45. Gerhardt, J.; Montani, M.; Wild, P.; Beer, M.; Huber, F.; Hermanns, T.; Müntener, M.; Kristiansen, G. FOXA1 Promotes Tumor Progression in Prostate Cancer and Represents a Novel Hallmark of Castration-Resistant Prostate Cancer. *The American Journal of Pathology* **2012**, *180*, 848–861. doi:10.1016/j.ajpath.2011.10.021.
46. Navarro, H.I.; Goldstein, A.S. HoxB13 mediates AR-V7 activity in prostate cancer. *Proceedings of the National Academy of Sciences* **2018**, *115*, 6528–6529, [https://www.pnas.org/content/115/26/6528.full.pdf]. doi:10.1073/pnas.1808196115.

47. Zhou, Y.; Zhou, B.; Pache, L.; Chang, M.; Khodabakhshi, A.H.; Tanaseichuk, O.; Benner, C.; Chanda, S.K. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications* **2019**, *10*. doi:10.1038/s41467-019-09234-6.
48. Piñero, J.; Ramírez-Anguita, J.M.; Saüch-Pitarch, J.; Ronzano, F.; Centeno, E.; Sanz, F.; Furlong, L.I. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* **2019**, *48*, D845–D855, [https://academic.oup.com/nar/article-pdf/48/D1/D845/31697865/gkz1021.pdf]. doi:10.1093/nar/gkz1021.
49. Pan, J.B.; Hu, S.C.; Shi, D.; Cai, M.C.; Li, Y.B.; Zou, Q.; Ji, Z.L. PaGenBase: A Pattern Gene Database for the Global and Dynamic Understanding of Gene Function. *PLOS ONE* **2013**, *8*, 1–9. doi:10.1371/journal.pone.0080747.
50. Horoszewicz, J.S.; Leong, S.S.; Kawinski, E.; Karr, J.P.; Rosenthal, H.; Chu, T.M.; Mirand, E.A.; Murphy, G.P. LNCaP Model of Human Prostatic Carcinoma. *Cancer Research* **1983**, *43*, 1809–1818, [https://cancerres.aacrjournals.org/content/43/4/1809.full.pdf].
51. Han, H.; Shim, H.; Shin, D.; Shim, J.E.; Ko, Y.; Shin, J.; Kim, H.; Cho, A.; Kim, E.; Lee, T.; Kim, H.; Kim, K.; Yang, S.; Bae, D.; Yun, A.; Kim, S.; Kim, C.Y.; Cho, H.J.; Kang, B.; Shin, S.; Lee, I. TRRUST: a reference database of human transcriptional regulatory interactions. *Scientific Reports* **2015**, *5*. doi:10.1038/srep11432.

© 2020 by the authors. Submitted to *Genes* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).