Tensor decomposition-Based Unsupervised Feature Extraction Applied to Single-Cell Gene Expression Analysis

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

[1] *Department of Physics, Chuo University, Tokyo 112-8551, Japan*
[2] *King Abdulaziz University, Department of Computer Science, Jeddah 21589, Saudi Arabia*

Correspondence*:
Corresponding Author
tag@granular.com

ABSTRACT

Although single cell RNA sequencing (scRNA-seq) technology is newly invented and promising one, because of lack of enough information that labels individual cells, it is hard to interpret the obtained gene expression of each cell. Because of this insufficient information available, unsupervised clustering, e.g., tSNE and UMAP, is usually employed to obtain low dimensional embedding that can help to understand cell-cell relationship. One possible drawback of this strategy is that the outcome is highly dependent upon genes selected for the usage of clustering. In order to fulfill this requirements, there are many methods that performed unsupervised gene selection. In this study, a tensor decomposition (TD) based unsupervised feature extraction (FE) was applied to the integration of two scRNA-seq expression profiles that measure human and mouse midbrain development. TD based unsupervised FE could not only select coincident genes between human and mouse, but also biologically reliable genes. Coincidence between two species as well as biological reliability of selected genes is increased compared with principal component analysis (PCA) based FE applied to the same data set in the previous study. Since PCA based unsupervised FE outperformed other three popular unsupervised gene selection methods, highly variable genes, bimodal genes and dpFeature, TD based unsupervised FE can do so as well. In addition to this, ten transcription factors (TFs) that might regulate selected genes and might contribute to midbrain development are identified. These ten TFs, BHLHE40, EGR1, GABPA, IRF3, PPARG, REST, RFX5, STAT3, TCF7L2, and ZBTB33, were previously reported to be related to brain functions and diseases. TD based unsupervised FE is promising method to integrate two scRNA-seq profiles effectively.

Keywords: tensor decomposition, enrichment analysis, sngle cell RNA-sequencing, midbrain development, inter-species analysis

1 INTRODUCTION

Single cell RNA sequencing (scRNA-seq) (17) is a newly invented technology that enables us to measure amount of RNA in single cell basis. In spite of its promising potential, it is not easy to interpret the measurements. The primary reason of this difficulty is the lack of sufficient information that characterizes

individual cells. In contrast to the huge number of cells measured, which is often as many as several thousands, the number of labeling is limited, e.g., measurement of conditions as well as the amount of expression of key genes measured by fluorescence-activated cell sorting, whose number is typically as little as tens. This prevents us from selecting genes that characterize the individual cell properties.

In order to deal with samples without suitable number of labelling, unsupervised method is frequently used, since it does not make use of labeling information directly. K-means clustering as well as hierarchical clustering are the popular methodology that are often applied to gene expression analysis. The popular clustering methods specifically applied to scRNA-seq is tSNE (52) or UMAP (9), which is known to be useful to get low dimensional embedding of a set of cells. In spite of that, the obtained clusters are highly dependent upon genes used for clustering. Thus, the next issue is, without labeling (i.e., pre-knowledge), to select genes that might be biologically meaningful.

The various unsupervised gene selection methods were invented, e.g., highly variable genes, bimodal genes and dpFeature and principal component analysis (PCA) based unsupervised feature extraction (FE) (50, 34, 49, 42, 25, 31, 27, 40, 47, 4, 5, 44, 45, 13, 12, 11, 51, 39, 43, 22, 23, 24, 26, 41, 48). Chen et al. (2) recently compared genes selected by these methods and concluded that the genes selected are very diverse and have their own (unique) biological features. In this sense, it is required to invent more advanced unsupervised gene selection methods that can select more biologically reliable genes.

In this paper, we propose the application of tensor decomposition (TD) based unsupervised FE (37, 46, 35, 36, 30, 32, 29, 33, 28), which is an advanced method of PCA based unsupervised FE, to scRNA-seq analysis (for more details about PCA based unsupervised FE and TD based unsupervised FE, see the recently published book (38)), especially focusing on the integration of two scRNA-seq profiles, The advantages of TD based unsupervised FE when compared with PCA based unsupervised FE are that the former can integrated more than two gene expression prior to the analysis while the latter can only integrate the results obtained by applying the method to individual data set. In the following, following the previous study (34) where PCA based unsupervised FE was employed, we try to integrate human and mouse midbrain development gene expression profiles to obtain key genes that contribute to this process, by applying TD based unsupervised FE. It turned out that TD based unsupervised FE can identify biologically more reliable and more common genes between human and mouse than PCA based unsupervised FE that outperformed other methods compared with.

2 METHODS AND MATERIALS

2.1 scRNA-seq data

ScRNA-seq data used in this study were downloaded from gene expression omnibus (GEO) under the GEO ID GSE76381; the files named "GSE76381_EmbryoMoleculeCounts.cef.txt.gz" (for human) and "SE76381_MouseEmbryoMoleculeCounts.cef.txt.gz" (for mouse) were downloaded. These two gene expression profiles are generated from scRNA-seq data set that represents human embryo ventral midbrain cells between 6 and 11 weeks of gestation (287 cells for six weeks, 131 cells for seven weeks, 331 cells for eight weeks, 322 cells for nine weeks, 509 cells for ten weeks, and 397 cells for eleven weeks, in total 1977 cells), mouse ventral midbrain cells at six developmental stages between E11.5 to E18.5 (349 cells for E11.5, 350 cells for E12.5, 345 cells for E13.5, 308 cells for E14.5, 356 cells for E15.5, 142 cells for E18.5 and 57 cells for unknown, in total 1907 cells).

2.2 TD based unsupervised FE

TD based unsupervised FE is recently proposed methods successfully applied to various biological problems. TD based unsupervised FE can be used for integration of multiple measurements applied to the common set of genes. Suppose $x_{ij} \in \mathbb{R}^{N \times M}$ and $x_{ik} \in \mathbb{R}^{N \times K}$ are the ith expression of jth and kth cells under the two distinct conditions (in the present study, they are human and mouse), respectively. Then 3-mode tensor, $x_{ijk} \in \mathbb{R}^{N \times M \times K}$, where $N(= 13889)$ is total number of common genes between human and mouse, which share gene symbols, between human and mouse, $M(= 1977)$ is the number of human cells and $K(= 1907)$ is total number of mouse cells, is defined as

$$x_{ijk} = x_{ij} \cdot x_{ik}. \tag{1}$$

It is Type I Case II tensor (32). Since it is too large to be decomposed, we further transform it into type II tensor as

$$x_{jk} = \sum_{i=1}^{N} x_{ijk}, \tag{2}$$

where $x_{jk} \in \mathbb{R}^{M \times K}$ is now not a tensor but a matrix. In this case, TD is equivalent to singular value decomposition (SVD). After applying SVD to x_{jk}, we get SVD,

$$x_{jk} = \sum_{\ell=1}^{\min(M,K)} \lambda_\ell u_{\ell j} v_{\ell k}. \tag{3}$$

where $u_{\ell j} \in \mathbb{R}^{M \times M}$ and $v_{\ell k} \in \mathbb{R}^{K \times K}$ are singular value vectors attributed to cells of human scRNA-seq and those of mouse scRNA-seq, respectively.

Singular value vectors attributed to genes of human and mouse scRNA-seq, $u_{\ell i} \in \mathbb{R}^{N \times M}$ and $v_{\ell i} \in \mathbb{R}^{N \times K}$, are defined as

$$u_{\ell i} = \sum_{j=1}^{M} u_{\ell j} x_{ij} \tag{4}$$

$$v_{\ell i} = \sum_{k=1}^{K} v_{\ell k} x_{ik} \tag{5}$$

respectively.

In order to find genes associated with biological functions, we need to select $u_{\ell j}$ and $v_{\ell k}$ which are coincident with biological meaning. In this study, we employ time points of measurements as biological meanings. In other words, we seek genes associated with time development. Since we would like to find any kind of time dependence, we simply deal with time points as un-ordered labelling. Thus, we apply categorical regression

$$u_{\ell j} = a_\ell + \sum_{t=1}^{T} a_{\ell t} \delta_{jt} \tag{6}$$

(T=6; t=1 to T correspond to 6, 7, 8, 9, 10 and 11 weeks, see Methods and Materials) or

$$v_{\ell k} = b_\ell + \sum_{t=1}^{T} b_{\ell t}\delta_{kt} \tag{7}$$

(T=7; t=1 to T correspond to E11.5, E12.5, E13.5, E14.5, F15.5, E18.5 and unknown, see Methods and Materials) where $\delta_{jt}(\delta_{kt}) = 1$ when the jth(kth) cell is taken from the tth time point otherwise $\delta_{jt}(\delta_{kt}) = 0$. $a_\ell, a_{\ell t}, b_\ell$ and $b_{\ell t}$ are the regression coefficients.

P-values are attributed to ℓth singular value vectors using the above categorical regression (`lm` function in R (15) is used to compute P-values). P-values attributed to singular value vectors are corrected by BH criterion (1). Singular value vectors associated with corrected P-values less than 0.01 are selected for the download analysis. Hereafter, the set of selected singular value vectors of human and mouse are denoted as $\Omega_\ell^{\text{human}}$ and $\Omega_\ell^{\text{mouse}}$, respectively.

P-values are attributed to genes with assuming χ^2 distribution for the gene singular value vectors, $u_{\ell i}$ and $v_{\ell i}$, corresponding to the cell singular value vectors selected by categorical regression as

$$P_i^{\text{human}} = P_{\chi^2}\left[> \sum_{\ell \in \Omega_\ell^{\text{human}}} \left(\frac{u_{\ell i} - \langle u_{\ell i}\rangle}{\sigma_\ell^{\text{human}}}\right)^2\right] \tag{8}$$

for human genes and

$$P_i^{\text{mouse}} = P_{\chi^2}\left[> \sum_{\ell \in \Omega_\ell^{\text{mouse}}} \left(\frac{v_{\ell i} - \langle v_{\ell i}\rangle}{\sigma_\ell^{\text{mouse}}}\right)^2\right] \tag{9}$$

for mouse genes, respectively. Here

$$\langle u_{\ell i}\rangle = \frac{1}{N}\sum_{i=1}^{N} u_{\ell i} \tag{10}$$

and

$$\langle v_{\ell i}\rangle = \frac{1}{N}\sum_{i=1}^{N} v_{\ell i}. \tag{11}$$

$\sigma_\ell^{\text{human}}$ and $\sigma_\ell^{\text{mouse}}$ are the standard deviations of ℓth gene singular value vectors for human and mouse respectively, $\Omega_\ell^{\text{human}}$ and $\Omega_\ell^{\text{mouse}}$ are sets of ℓs selected by categorical regression for human (eq. (6)) and mouse (eq.(7)), respectively, and $P_{\chi^2}[> x]$ is the cumulative probability of χ^2 distribution when the argument takes values larger than x. P_i^{human} and P_i^{mouse} are corrected by BH criterion and genes associated with corrected P-values less than 0.01 are selected.

2.3 Enrichment analyses

Various enrichment analysis methods are performed with separately uploading selected human and mouse gene symbols to Enrichr (6).

3 RESULTS

As a result, following the procedure described in Methods and Materials, we identified 55 and 44 singular value vectors attributed to cells, $u_{\ell j}$s and $v_{\ell k}$s, for human and mouse, respectively. One possible validation of selected $u_{\ell j}$s and $v_{\ell k}$s are coincidence. Although cells measured are not related between human and mouse at all, if SVD works well, corresponding singular value vectors (i.e., $u_{\ell j}$ and $v_{\ell k}$ sharing the same ℓs) attributed to cells should share something biological, e.g., time dependence. This suggests that it is more likely that corresponding singular value vectors attributed to cells, $u_{\ell j}$ and $v_{\ell k}$, are simultaneously associated with significant P-vales computed by categorical regression. As expected, they are highly significantly correlated. Table 1 shows confusion matrix of the coincidence of selected singular value vectors between human and mouse. Odds ratio is as large as 227 and P-value is as small as 1.44×10^{-44}.

Figure 1 shows the coincidence of selected singular value vectors between human and mouse. Singular value vectors with smaller ℓs, i.e., with more contributions, are more likely selected and coincident between human and mouse. This can be the side evidence that guarantees that TD based unsupervised FE successfully integrated human and mouse scRNA-seq data.

Next, we selected genes with following the procedures described in Methods and Materials. The first validation of selected genes is the coincidence between human and mouse. In my previous study (34), more number of common genes were selected by PCA based unsupervised FE than other methods compared, i.e., highly variables genes, bimodal genes and dpFeature. Table 2 shows the confusion matrix that describes the coincidence of selected genes between human and mouse. Odds ratio is as large as 133 and P-value is 0 (i.e., less than numerical accuracy), which is significantly better than coincidence of selected genes between human and mouse (53 common genes between 116 genes selected for human and 118 genes selected mouse), previously achieved by PCA based unsupervised FE (34), which outperformed other methods, i.e., highly variable genes, bimodal genes and dpFeature.

On the other hands, most of genes selected by PCA based unsupervised FE in the previous study (34) are included in the genes selected by TD based unsupervised FE in the present study. One hundred and two genes are selected by TD based unsupervised FE among 116 human genes selected by PCA based unsupervised FE in the previous study (34) while 91 genes are selected by TD based unsupervised FE among 118 mouse genes by PCA based unsupervised FE. Thus, TD based unsupervised FE is quite consistent with PCA based unsupervised FE.

Biological significance tested by enrichment analysis is further enhanced. Most remarkable advances achieved by TD based unsupervised FE is "Allen Brain Atlas", to which only downregulated genes were enriched in the previous study (34). As can be seen in Table 3, now many enrichment associated with upregulated genes. In addition to this, most of top ranked five terms are related to paraventricular, which is adjusted to midbrain. This suggests that TD based unsupervised FE successfully identified genes related to midbrain.

In addition to this, "Jensen TISSUES" (Table 4) for Embryonic_brain is highly enhanced (i.e., more significant (smaller) P-values $\sim 10^{-100}$, which were as large as 10^{-10} to 10^{-20} in the previous study (34)). On the other hand, "ARCHS4 tissues" also strongly supports the biological reliability of selected genes (Table 5). The term "MIDBRAIN" is enriched highly and it is top ranked for both human and mouse.

There are some brain related enrichment found in other categories, although they are not strong enough compared with the above three. Brain related terms in "GTEx Tissue Sample Gene Expression Profiles up" (Table 6) are also enhanced for mouse brain (top three terms are brain), although no brain terms are

enriched within top five raked terms for human (This discrepancy cannot be understood at the moment). Contrary, brain related terms in "MGI Mammalian Phenotype 2017" (Table 7) are enhanced for human brain (fourth and fifth ranked), although no brain terms are enriched within top five raked terms for mouse (This discrepancy also cannot be understood at the moment). The above observations suggest that TD based unsupervised FE could identify genes related to mouse and human embryonic midbrain.

We also uploaded selected 456 human genes and 505 mouse genes to STRING server (21) that evaluates protein-protein interaction (PPI) enrichment. 7488 PPI are reported among 456 human genes while expected number of PPI is as small as 3524 (P-value is less than 1×10^{-6}). 6788 PPI are reported among 505 mouse genes while expected number of PPI is as small as 3290 (P-value is again less than 1×10^{-6}). Thus, TD based unsupervised FE can successfully identify significantly interacting protein coding genes.

Finally, we checked if transcription factor (TF) that targets selected genes are common between human and mouse (Table 8). These TFs are associated with adjusted P-values less than 0.01 in "ENCODE and ChEA Consensus TFs from ChIP-X" of Enrichr. They are highly overlapped between human and mouse (There are ten common TFs between 16 TFs found in human and 24 TFs found in mouse). Although selected TFs are very distinct from those in the previous study (34), they are highly interrelated with each other (see below). These TFs are up-loaded to regnetworkweb server (8) and TF networks shown in Figure 2 are identified. Clearly, even in partial, these TFs interact highly with each other.

We also checked if commonly selected ten TFs (bold in Table 8) are related to brains. Lack of BHLHE40 was found to result in brain malfunction (3). The function of EGR1 was found in embryonic rat brain (54). GABPA is essential for human cognition (16). IRF3 is related to brain disease (18). PPAR, which PPARG belongs to, is believed to be therapeutic target of neurodegenerative diseases (53). REST is a master regulator of neurogenesis (10). RFX5 is known to be expressive in fetal brain (20). STAT3 promotes brain metastasis (14). TCF7L2 regulated brain gene expression (19). ZBTB33 affects the mouse behaviour through regulating brain gene expression (7). Thus all commonly selected ten TFs are related to brains.

4 DISCUSSIONS AND FUTURE WORK

In this study, we applied TD based unsupervised FE to the integration of scRNA-seq data sets taken from two species: human and mouse. In the sense of identification of biologically more reliable set of genes, TD based unsupervised FE can outperform PCA based unsupervised FE that previously (34) could outperform three more popular methods: highly variable genes, bimodal genes, and dpFeature. Thus, it is supposed that TD based unsupervised FE can do so, too.

For the purpose of integration of two scRNA-seq data set, TD based unsupervised FE has many advantages than other four methods, i.e., PCA based unsupervised FE, highly variable genes, bimodal genes and dpFeature. At first, TD based unsupervised FE can integrate two scRNA-seq data set, not after the selection of genes, but before the selection of genes. This enables us to identify more coincident genes set between two scRNA-seq, in this study human and mouse. As a result, we could identify more coincident results between human and mouse.

The criterion of genes selection is quite robust; they should be dependent upon time points when they are measured. We do not have to specify how they are actually correlated with time. It is another advantages of TD based unsupervised FE.

With applying enrichment analysis to genes selected, we can find many valuable insights about biological process. As a result, we can identify ten key TFs that might regulate embryonic midbrain developments. All of ten selected TFs turn out to be related to brains.

TD based unsupervised FE turn out to be quite effective to integrate two scRNA-seq data set. This method should be applied to various scRNA-seq data sets considering broader scope of investigations.

In future work, we plan to (1) utilize the proposed TD based unsupervised FE under the transfer learning setting; (2) extend the proposed approach to handle the data integration from multiple related tasks; and (3) investigate the performance of the proposed approach when coupled with machine and deep learning algorithms.

CONFLICT OF INTEREST STATEMENT

No conflict of interest is declared.

AUTHOR CONTRIBUTIONS

YHT planed the research, performed analyses, and wrote a paper. TT discussed the results and wrote a paper.

FUNDING

It was supported by KAKENHI, 17K00417 and Okawa foundation, grant number 17-10.

ACKNOWLEDGMENTS

Nothing acknowldegded.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the GEO.

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76381.

REFERENCES

1 .Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 289–300

2 .Chen, B., Herring, C. A., and Lau, K. S. (2018). pyNVR: investigating factors affecting feature selection from scRNA-seq data for lineage reconstruction. *Bioinformatics* doi:10.1093/bioinformatics/bty950

3 .Hamilton, K. A., Wang, Y., Raefsky, S. M., Berkowitz, S., Spangler, R., Suire, C. N., et al. (2018). Mice lacking the transcriptional regulator Bhlhe40 have enhanced neuronal excitability and impaired synaptic plasticity in the hippocampus. *PLOS ONE* 13, 1–22. doi:10.1371/journal.pone.0196223

4 .Ishida, S., Umeyama, H., Iwadate, M., and Taguchi, Y.-h. (2014). Bioinformatic Screening of Autoimmune Disease Genes and Protein Structure Prediction with FAMS for Drug Discovery. *Protein Pept. Lett.* 21, 828–39

5 .Kinoshita, R., Iwadate, M., Umeyama, H., and Taguchi, Y.-h. (2014). Genes associated with genotype-specific DNA methylation in squamous cell carcinoma as candidate drug targets. *BMC Syst Biol* 8 Suppl 1, S4

6 .Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* 44, W90–W97. doi:10.1093/nar/gkw377

7 .Kulikov, A. V., Korostina, V. S., Kulikova, E. A., Fursenko, D. V., Akulov, A. E., Moshkin, M. P., et al. (2016). Knockout zbtb33 gene results in an increased locomotion, exploration and pre-pulse inhibition in mice. *Behavioural Brain Research SreeTestContent1* 297, 76 – 83. doi:https://doi.org/10.1016/j.bbr.2015.10.003

8 .Liu, Z.-P., Wu, C., Miao, H., and Wu, H. (2015). RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database* 2015. doi:10.1093/database/bav095

9 .McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction ArXiv:1802.03426

10 .Mozzi, A., Guerini, F. R., Forni, D., Costa, A. S., Nemni, R., Baglio, F., et al. (2017). REST, a master regulator of neurogenesis, evolved under strong positive selection in humans and in non human primates. *Scientific Reports* 7. doi:10.1038/s41598-017-10245-w

11 .Murakami, Y., Kubo, S., Tamori, A., Itami, S., Kawamura, E., Iwaisako, K., et al. (2015). Comprehensive analysis of transcriptome and metabolome analysis in Intrahepatic Cholangiocarcinoma and Hepatocellular Carcinoma. *Sci Rep* 5, 16294

12 .Murakami, Y., Tanahashi, T., Okada, R., Toyoda, H., Kumada, T., Enomoto, M., et al. (2014). Comparison of Hepatocellular Carcinoma miRNA Expression Profiling as Evaluated by Next Generation Sequencing and Microarray. *PLoS ONE* 9, e106314

13 .Murakami, Y., Toyoda, H., Tanahashi, T., Tanaka, J., Kumada, T., Yoshioka, Y., et al. (2012). Comprehensive miRNA expression analysis in peripheral blood can diagnose liver disease. *PLoS ONE* 7, e48366

14 .Priego, N., Zhu, L., Monteiro, C., Mulders, M., Wasilewski, D., Bindeman, W., et al. (2018). STAT3 labels a subpopulation of reactive astrocytes required for brain metastasis. *Nature Medicine* 24, 1024–1035. doi:10.1038/s41591-018-0044-4

15 .R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria

16 .Reiff, R. E., Ali, B. R., Baron, B., Yu, T. W., Ben-Salem, S., Coulter, M. E., et al. (2014). METTL23, a transcriptional partner of GABPA, is essential for human cognition. *Human Molecular Genetics* 23, 3456–3466. doi:10.1093/hmg/ddu054

17 .Sasagawa, Y., Hayashi, T., and Nikaido, I. (2019). *Strategies for Converting RNA to Amplifiable cDNA for Single-Cell RNA Sequencing Methods* (Singapore: Springer Singapore). 1–17. doi:10.1007/978-981-13-6037-4_1

18 .Schultz, K. L. W., Troisi, E. M., Baxter, V. K., Glowinski, R., and Griffin, D. E. (2019). Interferon regulatory factors 3 and 7 have distinct roles in the pathogenesis of alphavirus encephalomyelitis. *Journal of General Virology* 100, 46–62

19 .Shao, W., Wang, D., Chiang, Y.-T., Ip, W., Zhu, L., Xu, F., et al. (2013). The Wnt signaling pathway effector TCF7L2 controls gut and brain proglucagon gene expression and glucose homeostasis. *Diabetes* 62, 789–800. doi:10.2337/db12-0365

20 .Sugiaman-Trapman, D., Vitezic, M., Jouhilahti, E.-M., Mathelier, A., Lauter, G., Misra, S., et al. (2018). Characterization of the human RFX transcription factor family by regulatory and target gene analysis. *BMC Genomics* 19, 181. doi:10.1186/s12864-018-4564-6

21 .Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2014). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research* 43, D447–D452. doi:10.1093/nar/gku1003

22 .Taguchi, Y.-h. (2014). Integrative analysis of gene expression and promoter methylation during reprogramming of a non-small-cell lung cancer cell line using principal component analysis-based unsupervised feature extraction. In *Intelligent Computing in Bioinformatics*, eds. D.-S. Huang, K. Han, and M. Gromiha (Heidelberg: Springer International Publishing), vol. 8590 of *LNCS*. 445–455

23 .Taguchi, Y.-h. (2015). Identification of aberrant gene expression associated with aberrant promoter methylation in primordial germ cells between E13 and E16 rat F3 generation vinclozolin lineage. *BMC Bioinformatics* 16 Suppl 18, S16

24 .Taguchi, Y.-h. (2016). Identification of More Feasible MicroRNA-mRNA Interactions within Multiple Cancers Using Principal Component Analysis Based Unsupervised Feature Extraction. *Int J Mol Sci* 17, E696

25 .Taguchi, Y. H. (2016). microRNA-mRNA interaction identification in Wilms tumor using principal component analysis based unsupervised feature extraction. In *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*. 71–78. doi:10.1109/BIBE.2016.14

26 .Taguchi, Y.-h. (2016). Principal component analysis based unsupervised feature extraction applied to budding yeast temporally periodic gene expression. *BioData Min* 9, 22

27 .Taguchi, Y. H. (2016). Principal component analysis based unsupervised feature extraction applied to publicly available gene expression profiles provides new insights into the mechanisms of action of histone deacetylase inhibitors. *Neuroepigenetics* 8, 1 – 18. doi:http://dx.doi.org/10.1016/j.nepig.2016.10.001

28 .Taguchi, Y.-H. (2017). Identification of candidate drugs for heart failure using tensor decomposition-based unsupervised feature extraction applied to integrated analysis of gene expression between heart failure and DrugMatrix datasets. In *Intelligent Computing Theories and Application* (Springer International Publishing). 517–528

29 .Taguchi, Y.-H. (2017). Identification of candidate drugs using tensor-decomposition-based unsupervised feature extraction in integrated analysis of gene expression between diseases and DrugMatrix datasets. *Sci Rep* 7, 13733

30 .Taguchi, Y.-H. (2017). One-class differential expression analysis using tensor decomposition-based unsupervised feature extraction applied to integrated analysis of multiple omics data from 26 lung adenocarcinoma cell lines. In *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*. 131–138. doi:10.1109/BIBE.2017.00-66

31 .Taguchi, Y.-h. (2017). Principal Components Analysis Based Unsupervised Feature Extraction Applied to Gene Expression Analysis of Blood from Dengue Haemorrhagic Fever Patients. *Sci Rep* 7, 44016

32 .Taguchi, Y.-H. (2017). Tensor decomposition-based unsupervised feature extraction applied to matrix products for multi-view data processing. *PLoS ONE* 12, e0183933

33 .Taguchi, Y.-H. (2017). Tensor decomposition-based unsupervised feature extraction identifies candidate genes that induce post-traumatic stress disorder-mediated heart diseases. *BMC Med. Genomics* 10, 67

34 .Taguchi, Y.-h. (2018). Principal component analysis-based unsupervised feature extraction applied to single-cell gene expression analysis. In *Intelligent Computing Theories and Application*, eds. D.-S. Huang, K.-H. Jo, and X.-L. Zhang (Cham: Springer International Publishing), 816–826

35 .Taguchi, Y.-H. (2018). Tensor decomposition-based unsupervised feature extraction can identify the universal nature of sequence-nonspecific off-target regulation of mRNA mediated by microRNA transfection. *Cells* 7, 54. doi:10.3390/cells7060054

36 .Taguchi, Y.-H. (2018). Tensor decomposition/principal component analysis based unsupervised feature extraction applied to brain gene expression and methylation profiles of social insects with multiple castes. *BMC Bioinformatics* 19, 99

37 .Taguchi, Y.-h. (2019). Drug candidate identification based on gene expression of treated cells using tensor decomposition-based unsupervised feature extraction for large-scale data. *BMC Bioinformatics* 19, 388. doi:10.1186/s12859-018-2395-8

38 .Taguchi, Y.-h. (2019). *Unsupervised Feature Extraction Applied to Bioinformatics* (Switzerland: Springer International)

39 .Taguchi, Y.-h., Iwadate, M., and Umeyama, H. (2015). Heuristic principal component analysis-based unsupervised feature extraction and its application to gene expression analysis of amyotrophic lateral sclerosis data sets. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on*. 1–10. doi:10.1109/CIBCB.2015.7300274

40 .Taguchi, Y. H., Iwadate, M., and Umeyama, H. (2015). Principal component analysis-based unsupervised feature extraction applied to in silico drug discovery for posttraumatic stress disorder-mediated heart disease. *BMC Bioinformatics* 16, 139

41 .Taguchi, Y.-h., Iwadate, M., and Umeyama, H. (2016). SFRP1 is a possible candidate for epigenetic therapy in non-small cell lung cancer. *BMC Medical Genomics* 9, 28. doi:10.1186/s12920-016-0196-3

42 .Taguchi, Y. H., Iwadate, M., Umeyama, H., and Murakami, Y. (2017). Principal component analysis based unsupervised feature extraction applied to bioinformatics analysis. In *Computational Methods with Applications in Bioinformatics Analysis*, eds. J. J. P. Tsai and K.-L. Ng (Singapore: World Scientific), chap. 8. 153–182. doi:10.1142/9789813207981_0008

43 .Taguchi, Y.-h., Iwadate, M., Umeyama, H., Murakami, Y., and Okamoto, A. (2015). Heuristic principal component analysis-aased unsupervised feature extraction and its application to bioinformatics. In *Big Data Analytics in Bioinformatics and Healthcare*, eds. B. Wang, R. Li, and W. Perrizo. 138–162

44 .Taguchi, Y.-h. and Murakami, Y. (2013). Principal component analysis based feature extraction approach to identify circulating microRNA biomarkers. *PLoS ONE* 8, e66714

45 .Taguchi, Y.-h. and Murakami, Y. (2014). Universal disease biomarker: can a fixed set of blood microRNAs diagnose multiple diseases? *BMC Res Notes* 7, 581

46 .Taguchi, Y. H. and Ng, K.-L. (2018). Tensor decomposition–based unsupervised feature extraction for integrated analysis of TCGA data on micrRNA expression and promoter methylation of genes in ovarian cancer. In *2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE)*. 195–200. doi:10.1109/BIBE.2018.00045

47 .Taguchi, Y.-h. and Okamoto, A. (2012). Principal component analysis for bacterial proteomic analysis. In *Pattern Recognition in Bioinformatics*, eds. T. Shibuya, H. Kashima, J. Sese, and S. Ahmad (Heidelberg: Springer International Publishing), vol. 7632 of *LNCS*. 141–152

48 .Taguchi, Y. H. and Wang, H. (2017). Genetic association between amyotrophic lateral sclerosis and cancer. *Genes* 8, 243. doi:10.3390/genes8100243

49 .Taguchi, Y.-h. and Wang, H. (2018). Exploring microrna biomarker for amyotrophic lateral sclerosis. *International Journal of Molecular Sciences* 19. doi:10.3390/ijms19051318

50 .Taguchi, Y.-h. and Wang, H. (2018). Exploring microrna biomarkers for parkinson disease from mrna expression profiles. *Cells* 7. doi:10.3390/cells7120245

51 .Umeyama, H., Iwadate, M., and Taguchi, Y.-h. (2014). TINAGL1 and B3GALNT1 are potential therapy target genes to suppress metastasis in non-small cell lung cancer. *BMC Genomics* 15 Suppl 9, S2

52 .van der Maaten, L. and Hinton, G. (2008). *Journal of Machine Learning Research* 9, 2579–2605

53 .Warden, A., Truitt, J., Merriman, M., Ponomareva, O., Jameson, K., Ferguson, L. B., et al. (2016). Localization of PPAR isotypes in the adult mouse and human brain. *Scientific Reports* 6. doi:10.1038/srep27618

54 .Wells, T., Rough, K., and Carter, D. (2011). Transcription mapping of embryonic rat brain reveals EGR-1 induction in SOX2+ neural progenitor cells. *Frontiers in Molecular Neuroscience* 4, 6. doi:10.3389/fnmol.2011.00006

FIGURE CAPTIONS

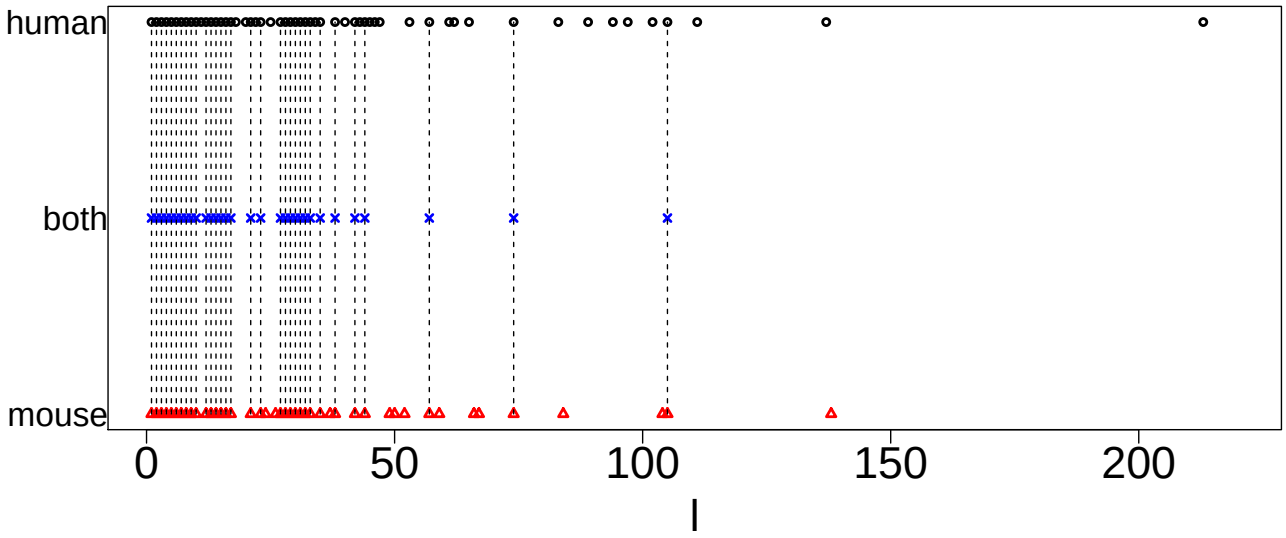
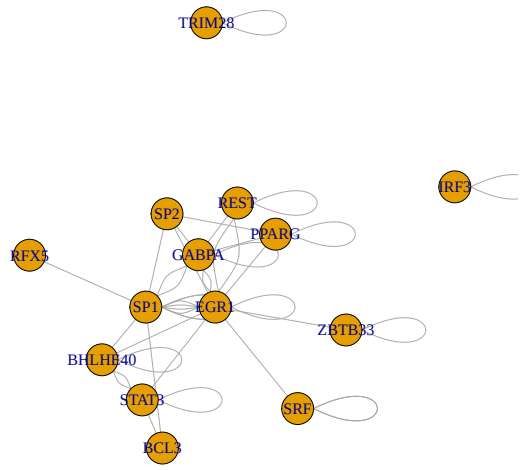


Figure 1. Coincidence between singular value vectors shown in Table 1. Horizontal axis: singular value vector numbering ℓ. Black open circles: ℓs selected for human, blue crosses: ℓs selected for both human and mouse, red open triangles: ℓs selected for mouse. Vertical black broken lines connect ℓs selected for both human and mouse.

(A)



(B)

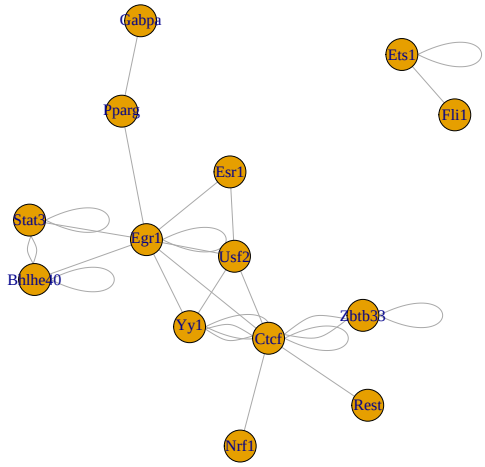


Figure 2. TF network identified by regnetworkweb for TFs in Table 8. (A) human, (B) mouse.

Table 1. Confusion matrix of coincidence between selected 55 singular value vectors selected among all 1977 singular value vectors, $u_{\ell j}$, attributed to human cells and 44 singular value vectors selected among all 1907 singular value vectors, $v_{\ell k}$, attributed to mouse cells. For human cells, only top 1907 singular value vectors among all 1977 singular value vectors are considered, since total number of singular value vectors attributed to mouse cells is 1907. selected: corrected P-values, computed with regression analysis eqs. (6) and (7), are less than 0.01, not selected: otherwise. Odds ratio is as many as 227 and P-values computed by Fisher's exact test is 1.44×10^{-44}.

		human	
		not selected	selected
mouse	not selected	1833	12
	selected	23	32

Table 2. Confusion matrix of coincidence between selected 456 genes for human and selected 505 genes for mouse among all 13384 common genes. Selected: corrected P-values, computed with χ^2 distribution eqs. (8) and (9), are less than 0.01, not selected: otherwise. Odds ratio is as many as 133 and P-values computed by Fisher's exact test is 0 (i.e. less than numerical accuracy).

		human	
		not selected	selected
mouse	not selected	13233	151
	selected	200	305

Table 3. Five top ranked terms from "Allen Brain Atlas up" by Enrichr for selected 456 human genes and 505 mouse genes.

Term	Overlap	P-value	Adjusted P-value
human			
Paraventricular hypothalamic nucleus, magnocellular division, medial magnocellular part	31/301	2.68×10^{-12}	2.91×10^{-9}
Paraventricular hypothalamic nucleus, magnocellular division	31/301	2.68×10^{-12}	2.91×10^{-9}
Paraventricular hypothalamic nucleus, magnocellular division, posterior magnocellular part	28/301	3.39×10^{-10}	1.47×10^{-7}
Paraventricular hypothalamic nucleus	29/301	7.02×10^{-11}	5.08×10^{-8}
paraventricular nucleus, dorsal part	27/301	1.57×10^{-9}	4.88×10^{-7}
mouse			
Paraventricular hypothalamic nucleus, magnocellular division, medial magnocellular part	31/301	4.03×10^{-11}	2.19×10^{-8}
Paraventricular hypothalamic nucleus, magnocellular division	31/301	4.03×10^{-11}	2.19×10^{-8}
Paraventricular hypothalamic nucleus, magnocellular division, posterior magnocellular part	31/301	4.03×10^{-11}	2.19×10^{-8}
lower dorsal lateral hypothalamic area	29/301	8.40×10^{-10}	3.65×10^{-7}
Paraventricular hypothalamic nucleus, magnocellular division, posterior magnocellular part, lateral zone	31/301	4.03×10^{-11}	2.19×10^{-8}

Table 4. Enrichment of Embryonic brain by "JENSEN TISSUES" in Enrichr

Term	Overlap	P-value	adjusted P-value
human			
Embryonic_brain	330/4936	3.36×10^{-104}	4.30×10^{-102}
mouse			
Embryonic_brain	366/4936	3.59×10^{-115}	4.59×10^{-113}

Table 5. Enrichment of Embryonic brain by "ARCHS4 Tissues" in Enrichr

Term	Overlap	P-value	adjusted P-value
		human	
MIDBRAIN	248/2316	1.02×10^{-129}	1.11×10^{-127}
		mouse	
MIDBRAIN	248/2316	1.44×10^{-99}	1.56×10^{-97}

Table 6. Five top ranked terms from "GTEx Tissue Sample Gene Expression Profiles up" by Enrichr for selected 456 human genes and 505 mouse genes. Brain related terms are asterisked.

Term	Overlap	P-value	Adjusted P-value
human			
GTEX-QCQG-1426-SM-48U22_ovary_female_50-59_years	105/1165	3.56×10^{-35}	1.04×10^{-31}
GTEX-RWS6-1026-SM-47JXD_ovary_female_60-69_years	116/1574	7.96×10^{-31}	7.74×10^{-28}
GTEX-TMMY-1726-SM-4DXTD_ovary_female_40-49_years	117/1582	2.97×10^{-31}	4.33×10^{-28}
GTEX-RU72-0008-SM-46MV8_skin_female_50-59_years	94/1103	1.99×10^{-29}	1.45×10^{-26}
GTEX-R55E-0008-SM-48FCG_skin_male_20-29_years	111/1599	3.67×10^{-27}	1.78×10^{-24}
mouse			
*GTEX-WVLH-0011-R4A-SM-3MJFS_brain_male_50-59_years	139/1957	1.93×10^{-30}	5.63×10^{-27}
*GTEX-X261-0011-R8A-SM-4E3I5_brain_male_50-59_years	135/1878	5.24×10^{-30}	7.65×10^{-27}
*GTEX-T5JC-0011-R4A-SM-32PLT_brain_male_20-29_years	129/1948	3.51×10^{-25}	3.42×10^{-22}
GTEX-R55E-0008-SM-48FCG_skin_male_20-29_years	109/1599	4.93×10^{-22}	2.40×10^{-19}
GTEX-TMMY-1726-SM-4DXTD_ovary_female_40-49_years	107/1582	2.37×10^{-21}	7.69×10^{-19}

Table 7. Five top ranked terms from "MGI Mammalian Phenotype 2017" by Enrichr for selected 456 human genes and 505 mouse genes. Brain related terms are asterisked.

Term	Overlap	P-value	Adjusted P-value
human			
MP:0002169_no_abnormal_phenotype_detected	82/1674	2.52×10^{-11}	5.53×10^{-8}
MP:0001262_decreased_body_weight	63/1189	3.40×10^{-10}	3.72×10^{-7}
MP:0001265_decreased_body_size	46/774	3.20×10^{-9}	$2.33 \times s10^{-6}$
*MP:0009937_abnormal_neuron_differentiation	15/106	1.81×10^{-8}	9.90×10^{-6}
*MP:0000788_abnormal_cerebral_cortex_morphology	17/145	3.64×10^{-8}	1.60×10^{-5}
mouse			
MP:0002169_no_abnormal_phenotype_detected	89/1674	$1.36 \times s10^{-11}$	3.09×10^{-8}
MP:0011091_prenatal_lethality,_complete_penetrance	27/272	1.68×10^{-9}	1.91×10^{-6}
MP:0001262_decreased_body_weight	65/1189	3.93×10^{-9}	2.97×10^{-6}
MP:0011100_preweaning_lethality,_complete_penetrance	42/674	8.55×10^{-8}	3.88×10^{-5}
MP:0001265_decreased_body_size	46/774	8.22×10^{-8}	3.88×10^{-5}

Table 8. TFs enriched in "ENCODE and ChEA Consensus TFs from ChIP-X" by Enrichr for human and mouse. Bold TFs are common.

human	BCL3, **BHLHE40**, **EGR1**, **GABPA**, **IRF3**, **PPARG**, **REST**, **RFX5**, SP1, SP2, SRF, **STAT3**, **TCF7L2**, TRIM28, TRIM28, **ZBTB33**,
mouse	**BHLHE40**, CTCF, E2F4, E2F6, **EGR1**, ESR1, ETS1, FLI1, **GABPA**, **IRF3**, NFIC, NRF1, **PPARG**, RCOR1, **REST**, **RFX5**, SPI1, **STAT3**, **TCF7L2**, USF1, USF2, YY1, **ZBTB33**, ZNF384,