Principal component analysis based unsupervised feature extraction applied to single cell gene expression analysis

Y-h. Taguchi[1][0000-0003-0867-8986]

[1] Department of Physics, Chuo University, Tokyo 112-8551, Japan
tag@granular.com

Abstract. Because of missing sample labeling, single cell (sc) RNA-seq naturally needs the unsupervised feature selection in order to identify critical genes under the experimental conditions considered. In this paper, in order to identify biologically relevant genes, we applied principal component analysis (PCA) based unsupervised feature extraction (FE) to gene expression profiles during mice and human embryonic brain development retrieved by scRNA-seq. When we consider biological relevance of selected genes evaluated by various enrichment analyses of biological term, PCA based unsupervised FE outperforms conventional unsupervised approaches that selects highly variable genes as well as bimodal genes and recently proposed dpFeature.

Keywords: principal component analysis, feature selection, embryonic brain development.

1 Introduction

Single cell analysis is a newly developed high through put technology that enables us to identify gene expression profiles of individual genes. There is a critical difference between single cell analysis and conventional tissue specific analysis; tissue samples are naturally labeled distinctively (e.g., patients vs healthy controls) while single cell samples are not always. Inevitably, we need to have unsupervised methodology, e.g., highly variable genes [1, 2] and bimodal genes [3] or recently proposed dpFeature [4]. The aim of employment of highly variable genes is to select genes that can discriminate underlying cluster structure that can be depicted by unsupervised clustering, e.g., tSNE [5]. On the other hand, bimodal genes are selected because unimodal genes are unlikely expressive distinctly between multiple classes, e.g., healthy controls and patients Although the combination of tSNE and highly variable genes or bimodal genes approach is often employed and empirically successful, biological validation of selected genes was rarely addressed. Generally, very few studies tried to evaluate multiple gene selection procedures for single cell RNA-seq. The purpose of this paper is to compare the multiple gene selection procedures and to identify the best one.
On the other hand, principal component analysis (PCA) based unsupervised feature extraction (FE), which was proposed by the author [6–28], was known to work well when tissue specific gene expression profiles are investigated. In this paper, PCA based unsupervised FE was applied to single cell gene expression analysis and its

effectiveness was evaluated from the biological point of views, based upon the comparisons with conventional approaches, highly variable genes approach as well as bimodal genes approach.

2 Materials and Methods

2.1 Gene expression

Gene expression profiles 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. Gene expression profiles were standardized such that they have zero mean and unit standard deviation within each sample. That is, when x_{ij} represents the expression of ith gene in jth sample, $\sum_i x_{ij} = 0$ and $\sum_i x_{ij}^2 = N$ where N is the number of genes. These two gene expression profiles are generated from single cell RNA-seq data set that represents human embryo ventral midbrain cells between 6 and 11 weeks of gestation, mouse ventral midbrain cells at six developmental stages between E11.5 to E18.5, Th+ neurons at P19-P27, and putative dopaminergic neurons at P28-P56 FACS-sorted from Slc6a3-Cre/tdTomato mice.

2.2 PCA based unsupervised FE

Suppose that matrix X has element x_{ij} that is gene expression of ith gene of jth sample. kth PC score attributed to ith gene u_{ki} can be computed as ith element of kth eigen vector of Gram matrix XX^T

$$XX^T\boldsymbol{u_k} = \lambda_k \boldsymbol{u_k}$$

kth PC loading attributed to jth sample, v_{kj}, can be obtained by $\boldsymbol{v}_{\mathrm{k}} = X^T\boldsymbol{u}_{\mathrm{k}}$ since

$$X^T X\boldsymbol{v}_{\mathrm{k}} = X^T XX^T\boldsymbol{u}_{\mathrm{k}} = X^T\lambda_{\mathrm{k}}\boldsymbol{u}_{\mathrm{k}} = \lambda_{\mathrm{k}}\boldsymbol{v}_{\mathrm{k}}$$

At first, we have to identify which PC loading attributed to samples is coincident with distinction between considered class labels attributed to samples. Since single cell RNA-seq (scRNA-seq) lacks sample labeling, the first K PCs are employed. Then, assuming multiple Gaussian distribution to PC scores, P-values are attributed to gene i using χ^2 distribution,

$$P_{\mathrm{i}} = P_{\chi^2}\left[> \sum_{k=1}^{K}\left(\frac{u_{ki}}{\sigma_k}\right)^2\right]$$

where σ_{k} is standard deviation and $P_{\chi^2}[> x]$ is the cumulative probability of χ^2distribution that the argument is larger than x. The summation is taken over the selected first K PC scores. Obtained P-values are adjusted with Benjamini and Hochberg (BH) criterion [29] and genes associated with adjusted P-values less than 0.01 were selected.

2.3 Enrichment analysis

In order to perform enrichment analyses, selected genes were uploaded to Enrichr [30] that include various enrichment analysis.

2.4 Highly variable genes

The procedure is basically following Ref. [1], although a brief description is given in the below. Suppose that x_{ij} is gene expression of ith gene of jth sample. The mean expression of ith gene is defines as

$$\mu_{\mathrm{i}} = \frac{\sum_{\mathrm{i}} \mathrm{x}_{\mathrm{ij}}}{M}$$

where M is the number of samples. The standard deviation σ_{i} is defines as

$$\sigma_{\mathrm{i}}^2 = \sum_{\mathrm{i}} \frac{\left(\mathrm{x}_{\mathrm{ij}} - \mu_{\mathrm{i}}\right)^2}{M}$$

Then the regression relation between σ_{i} and μ_{i} is assumed as

$$\log_{10}\left(\frac{\sigma_{\mathrm{i}}}{\mu_{\mathrm{i}}}\right) = \frac{1}{2}\log_{10}\left(\frac{\beta}{\mu_{\mathrm{i}}} + \alpha\right) + \epsilon_i$$

where α and β are the regression coefficients. Then P-value P_{i}' is attributed to ϵ_i assuming χ^2 distribution as

$$\mathrm{P}_{\mathrm{i}}' = P_{\chi^2}\left[> \left(\frac{\epsilon_{\mathrm{i}}}{\sigma'}\right)^2\right]$$

where σ' is standard deviation. Then genes i associated with adjusted P-value by BH criterion less than 0.01 were selected as highly variable genes.

2.5 Bimodal genes

P-values attributed to genes that reject null hypothesis (unimodal genes) are computed by dip.test function in R with default setting. Obtained P-values are adjusted by BH criterion and genes associated with adjusted P-values less than 0.01 are selected.

3 Results

Applying PCA based unsupervised FE to human and mouse embryonic brain developmental gene expression profiles, we selected 116 genes for human ($K = 2$, i.e, the first two PC scores were used for gene selection) and 118 genes for mouse ($K = 3$, i.e., the first three PC scores were used for gene selection), respectively. Interestingly, 53 genes, i.e., half of genes, are selected commonly both in human and mouse samples. Since such a large overlap between two genes sets with highly restricted number of genes cannot occur by chance, it is very likely that selected genes play critical roles in embryonic midbrain development.

In order to confirm this proposition by biologically validating the selected genes, we applied various enrichment analyses to the selected genes. Table 1 shows an Enrichment analysis by Enrichr, “MGI Mammalian Phenotype 2017”, of 118 genes se-

lected for mouse. Among top ranked five terms, top four are brain related terms. All of them reveal the abnormal morphology; it is reasonable since fetal gene expression is often distinct from adult one that lacks expression of fetal specific genes. Table 2 and Table 3 show another Enrichment analysis by Enrichr, "Allen Brain Atlas down", of 116 genes selected for human and 118 genes selected for mouse, respectively. All of them reveal downregulation of genes in brain regions. It is again reasonable since fatal genes expressive in embryo is unlikely expressive in adult tissue. The lack of fetal brain specific gene expression in the selected genes can also be seen in Table 4 and Table 5.

Table 1 Enrichment analysis by Enrichr, "MGI Mammalian Phenotype 2017", of 118 genes selected for mouse. Top 5 ranked terms.

Term	Overlap	P-value	Adjusted P-value
MP:0000788_abnormal_cerebral_cortex_morphology	7/145	2.45×10^{-5}	4.55×10^{-5}
MP:0003651_abnormal_axon_extension	5/48	9.18×10^{-6}	4.55×10^{-3}
MP:0000812_abnormal_dentate_gyrus_morphology	5/58	2.34×10^{-5}	4.55×10^{-3}
MP:0000807_abnormal_hippocampus_morphology	5/86	1.56×10^{-4}	2.04×10^{-2}
MP:0000819_abnormal_olfactory_bulb_morphology	4/48	1.83×10^{-4}	2.04×10^{-2}

Table 2 Enrichment analysis by Enrichr, "Allen Brain Atlas down", of 116 genes selected for human. Top 5 ranked terms.

Term	Overlap	P-value	Adjusted P-value
periventricular stratum of cerebellar vermis	18/300	1.33×10^{-13}	3.72×10^{-11}
Simple lobule	18/300	1.33×10^{-13}	3.72×10^{-11}
Simple lobule, molecular layer	18/300	1.33×10^{-13}	3.72×10^{-11}
Simple lobule, granular layer	18/300	1.33×10^{-13}	3.72×10^{-11}
white matter of cerebellar vermis	18/300	1.33×10^{-13}	3.72×10^{-11}

Table 3 Enrichment analysis by Enrichr, "Allen Brain Atlas down", of 118 genes selected for mouse. Top 5 ranked terms.

Term	Overlap	P-value	Adjusted P-value
Pyramus (VIII), granular layer	18/300	1.81×10^{-13}	4.66×10^{-11}
Pyramus (VIII)	18/300	1.81×10^{-13}	4.66×10^{-11}
Pyramus (VIII), molecular layer	18/300	1.81×10^{-13}	4.66×10^{-11}
Paraflocculus, molecular layer	18/300	1.81×10^{-13}	4.66×10^{-11}
Cerebellar cortex	18/300	1.81×10^{-13}	4.66×10^{-11}

Table 4 Enrichment analysis by Enrichr, "GTEx Tissue Sample Gene Expression Profiles down", of 116 genes selected for human. Top 5 ranked terms.

Term	Overlap	P-value	Adjusted P-value
GTEX-Q2AG-0011-R10A-SM-2HMLA_brain_female_40-49_years	51/1467	1.47×10^{-27}	3.29×10^{-24}
GTEX-TSE9-3026-SM-3DB76_brain_female_60-69_years	49/1384	1.06×10^{-26}	1.19×10^{-23}
GTEX-S7SE-0011-R10A-SM-2XCDF_brain_male_50-59_years	44/1278	3.20×10^{-23}	1.43×10^{-20}
GTEX-QMR6-1426-SM-32PLA_brain_male_50-59_years	41/1066	2.57×10^{-23}	1.43×10^{-20}
GTEX-RNOR-2326-SM-2TF4I_brain_female_50-59_years	47/1484	2.02×10^{-23}	1.43×10^{-20}

Table 5 Enrichment analysis by Enrichr, "GTEx Tissue Sample Gene Expression Profiles down", of 118 genes selected for mouse. Top 5 ranked terms

Term	Overlap	P-value	Adjusted P-value
GTEX-U8XE-0126-SM-4E3I3_testis_male_30-39_years	15/376	6.13×10^{-9}	3.45×10^{-6}
GTEX-X4XX-0011-R10B-SM-46MWO_brain_male_60-69_years	23/938	5.25×10^{-9}	3.45×10^{-6}
GTEX-U4B1-1526-SM-4DXSL_testis_male_40-49_years	13/282	1.23×10^{-8}	3.71×10^{-6}
GTEX-Q2AG-0011-R10A-SM-2HMLA_brain_female_40-49_years	29/1467	5.11×10^{-9}	3.45×10^{-6}
GTEX-RNOR-2326-SM-2TF4I_brain_female_50-59_years	29/1484	6.62×10^{-9}	3.45×10^{-6}

Although these above results are highly significant, since they are negative results, one may wonder if these are not strong enough to support the confidence of the selected genes. In order to show positive results, we check that if these genes are expressive in embryonic brain. Then, we found that these genes are enriched in "Jensen TISSUES" by Enrichr; Table 6 shows that selected genes are enriched in embryonic brain for human and mouse, respectively. Thus, confidence of selected genes is supported by not only negative selections but also a positive selection.

We further seek the regulatory elements that can regulate selected genes, since if the selected genes are truly co-expressive, there are likely common regulatory elements. In order to perform this, we check "ENCODE and ChEA Consensus TFs from ChIP-X" by Enrichr for both human and mouse, respectively. 42 TFs for human and 23 TFs for mouse are associated with adjusted *P*-values less than 0.01 (Table 7). Thus, they are likely co-regulated by these TFs. Moreover, most of TFs for mouse are

also identified in human (bold faces in Table 7). Therefore, it is very likely that we successfully identified common (species non-specific or conserved) TFs that regulate genes expressive during embryonic brain development.

We also checked if they construct cooperatively function or not. These TFs are uploaded to regnetwork server [31] and TF networks shown in Figure 1 are identified. Clearly, even in partial, these TFs interact with each other.

We also checked if the identified TFs are related to fetal brain development. TAF7 was reported to play critical role in embryonic development [32]. KAT2A, ATF2 and TAF1 are also suggested to be included into brain development [33]. BRCA1 was also known to play critical roles in brain development [34].CEBPD and CREB were reported to be related to brain disease [35, 36], respectively. E2F1 was reported to be related to postnatal brain development [37]. The function of EGR1 was found in embryonic rat brain [38]. PML was also reported in brain development [39]. SIN3A was reported to be included in brain development [40]. TCF3 was included into zebrafish brain development [41]. YY1 was also reported in brain development [42]. In conclusion, most of genes selected commonly (bold faces in Table 7) for human and mouse are related to brain development. Thus, the selection of genes is possibly reasonable.

Table 6 Enrichment of selected genes in Embryonic brain by "Jensen TISSUES" by Enrichr.

Term	Overlap	P-value	Adjusted P-value
Human			
Embryonic_brain	71/4936	2.52×10^{-16}	4.07×10^{-15}
Mouse			
Embryonic_brain	75/4936	8.90×10^{-20}	1.06×10^{-18}

Table 7 TF enrichment enriched in "ENCODE and ChEA Consensus TFs from ChIP-X" by Enrichr for human and mouse. Bold TFs are common.

human	**ATF2**, **BCL3**, BCLAF1, BHLHE40, **BRCA1**, **CEBPB**, **CEBPD**, **CHD1**, **CREB1**, CTCF, **E2F1**, E2F4, **EGR1**, ELF1, ETS1, FLI1, GABPA, **KAT2A**, KLF4, **MAX**, **MYC**, NANOG, **NELFE**, NFYA, NFYB, NR2C2, **PBX3**, **PML**, **RELA**, SALL4, **SIN3A**, SIX5, SOX2, SP1, SPI1, **TAF1**, **TAF7**, **TCF3**, USF2, **YY1**, ZBTB33, **ZMIZ1**
mouse	**ATF2**, **BCL3**, **BRCA1**, **CEBPB**, **CEBPD**, **CHD1**, **CREB1**, **E2F1**, **EGR1**, **KAT2A**, KLF, **MAX**, **MYC**, **NELFE**, **PBX3**, **PML**, **RELA**, **SIN3A**, **TAF1**, **TAF7**, **TCF3**, **YY1**, **ZMIZ1**

One might wonder why we did not use popular methods like highly variable genes [1] but used unpopular PCA based unsupervised FE. Thus, we apply highly variable genes approach to the present data set and decided which is more coincident with enrichment analysis. After applying highly variable genes strategy, we got 168 genes for human and 171 genes for mouse. The numbers of selected genes are similar to those of genes selected by PCA based unsupervised FE. In addition to this, there are 44 commonly selected genes between human and mouse. Thus, apparently, highly variable genes seem to work as well.

This impression cannot be hold after detailed investigation. At first, there are very few overlaps between genes selected by PCA based unsupervised FE and highly variable genes (only four genes are commonly selected by PCA based unsupervised FE and highly variable genes for both human and mouse). It does not sound very promising, since genes selected by PCA based unsupervised FE are likely biologically reliable as demonstrated in the above, although there is still a possibility that highly variable genes are coincident with enrichment analysis without significant overlap with genes selected by PCA based unsupervised FE.

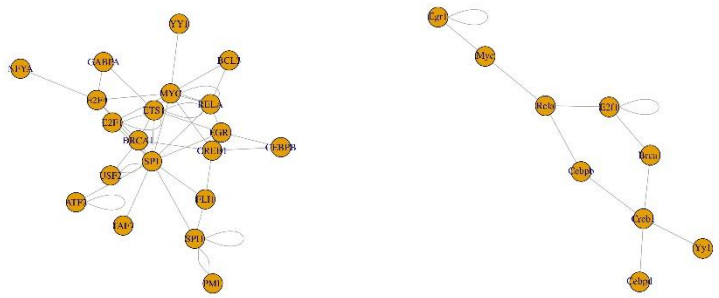

Figure 1 TF network identified by regnetworkweb for TFs in Table 7. Left:human, right: mouse.

In order to confirm this point, we uploaded genes to Enrichr. The result is substantially distinct from that given by PCA based unsupervised FE. At first, top ranked five enriched terms in “MGI Mammalian Phenotype 2017” for mouse did not include anything related to brain. It is inferior to the results in Table 1. On the other hand, no biological terms are significantly enriched in “Allen Brain Atlas down” for human, although substantially large number of terms are enriched in mouse. This is not a good sign, since highly variable genes were substantially common between mouse and human; the inconsistency between human and mouse suggests the selection of highly variable genes might be abiological. This impression was confirmed when “GTEx Tissue Sample Gene Expression Profiles down” was considered. Top five ranked terms did not include anything related to brain but those related to skin and blood. This suggests that highly variable genes are not likely more biologically reliable than genes selected by PCA based unsupervised FE. Finally, we consider “Jensen TISSUES” as has been done in the above (Table 6). Then we found no brain related terms significantly enriched in highly variable genes. We also checked if there are TFs that can regulate highly variable genes. Nevertheless, “ENCODE and ChEA Consensus TFs from ChIP-X” can give us only one TF whose target genes are significantly enriched in highly variable genes identified for human or mouse. One TF identified is definitely less than that in Table 7. All of these suggest that highly variable genes are unlikely biologically more reliable than genes selected by PCA based unsupervised FE.

We also checked bimodal genes as an alternative strategy (see Supplementary document). Genes associated with adjusted P-values less than 0.01 are 11344 and 10849 for human and mouse, respectively. Since these are too many, it is obvious that bimodal genes approach has no ability to select reasonable (restricted) number of genes. Nevertheless, in order to evaluate bimodal genes further, 200 top ranked (i.e., associated with smaller P-values) are intentionally selected. At first, number of commonly selected genes between 200 genes selected for human and mouse is as small as 21, which should be compared with 53 and 44 commonly selected genes between human and mouse by PCA based unsupervised FE and by highly variable genes. Furthermore, no genes are commonly selected between bimodal genes and PCA based unsupervised FE. Enrichment analyses by Enricher for bimodal genes are also inferior to that of PCA based unsupervised FE. "MGI Mammalian Phenotype 2017", of 200 bimodal genes selected for mouse include no terms associated with adjusted P-values less than 0.01 (Table S1). Five top ranked terms by "Allen Brain Atlas down" (Tables S2 and S3) are less significant than PCA based unsupervised FE (Tables 2 and 3), since P-vales are generally larger for bimodal genes that genes selected by PCA based unsupervised FE. Five top ranked terms by "GTEx Tissue Sample Gene Expression Profiles down" include no brain related one (Tables S4 and S5). All of these suggest that bimodal genes are unlikely biologically more reliable than genes selected by PCA based unsupervised FE.

Nevertheless, bimodal genes are slightly better than highly variable genes. "Jensen TISSUES" by Enrichr includes Embryonic_brain as a significant term (Table S6). "GTEx Tissue Sample Gene Expression Profiles down" identifies 40 TFs associated with adjusted P-vales less than 0.01 for human and mouse, among which as many as 30 TFs are commonly selected (Table S7). These TFs are also highly connected due to regnetworkweb (Figure S1).

Interestingly, among 30 commonly selected TFs between human and mouse in Table S7, eleven TFs are also commonly selected between human and mouse in Table 7. When considering that no genes are commonly selected between top ranked 200 bimodal genes and genes selected by PCA based unsupervised FE, the high number of commonly selected TFs between PCA based unsupervised FE and bimodal genes suggests the robustness of TFs selected. Thus, as an overall evaluation PCA based unsupervised FE is better than other two.

Although we have also compared with the newly proposed approach, dpFeature, because of lack of space, it was included in Supplementary document. dpFeature could not select biologically more reliable genes than PCA based unsupervised FE, either.

One may wonder if this study is not innovative at all, since PCA is not a new technology. Nevertheless, highly variable genes and bimodal genes are also not new concepts at all. Thus, as for the application to scRNA-seq, it is not important if the method is new or not, but important if the method is valuable or not. In this sense, even if methods themselves applied are not new, their application to new technology, scRNA-seq, can be innovative, if they have never been applied to scRNA-seq yet and more successful than newly proposed approach, e.g., dpFeature.

4 Conclusion

We have applied PCA based unsupervised FE to gene expression profiles retrieved by scRNA-seq analysis. Since scRNA-seq primarily lacks the labeling of sample (each cell), unsupervised approach inevitable. Genes selected by PCA based unsupervised FE for human and mice embryonic brain developments are not only associated with numerous significant biological terms enrichment but also highly coincident between mouse and human. On the other hand, frequently employed highly variable genes approach as well as bimodal genes approach or recently proposed dpFeature could not identify genes that are associated with as many significant biological terms enrichment as PCA based unsupervised FE achieved. Thus, PCA based unsupervised FE is more favorable than highly variable genes approach, bimodal genes approach or dpFeature from the biological point of views.

5 Supplementary materials

Supplementary materials are available at https://github.com/tagtag/SC.

Full list of genes, enrichment analyses for genes selected by PCA based unsupervised FE, bimodal genes and dpFeature, supplementary document that includes supplementary tables and figure for bimodal gene analyses and dpFeature, R codes that identify genes selected by PCA based unsupervised FE, highly variable genes, bimodal genes and dpFeature.

References

1. Chen, H.-I.H., Jin, Y., Huang, Y., Chen, Y.: Detection of high variability in gene expression from single-cell RNA-seq profiling. BMC Genomics. 17, 508 (2016).
2. Costa-Silva, J., Domingues, D., Lopes, F.M.: RNA-Seq differential expression analysis: An extended review and a software tool, (2017).
3. DeTomaso, D., Yosef, N.: FastProject: A tool for low-dimensional analysis of single-cell RNA-Seq data. BMC Bioinformatics. 17, (2016).
4. Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., Trapnell, C.: Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods. 14, 979–982 (2017).
5. Maaten, L. Van Der, Hinton, G.: Visualizing Data using t-SNE. J. Mach. Learn. Res. 1. 620, 267–84 (2008).
6. Ishida, S., Umeyama, H., Iwadate, M., Taguchi, Y.H.: Bioinformatic Screening of Autoimmune Disease Genes and Protein Structure Prediction with FAMS for Drug Discovery. Protein Pept. Lett. 21, 828–839 (2014).
7. Taguchi, Y.-H.: microRNA-mRNA interaction identification in Wilms tumor using principalcomponent analysis based unsupervised feature extraction. In:

2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE). pp. 71–78 (2016).
8. Murakami, Y., Kubo, S., Tamori, A., Itami, S., Kawamura, E., Iwaisako, K., Ikeda, K., Kawada, N., Ochiya, T., Taguchi, Y.-H.: Comprehensive analysis of transcriptome and metabolome analysis in Intrahepatic Cholangiocarcinoma and Hepatocellular Carcinoma. Sci. Rep. 5, 16294 (2015).
9. Taguchi, Y.-H.: Identification of More Feasible MicroRNA-mRNA Interactions within Multiple Cancers Using Principal Component Analysis Based Unsupervised Feature Extraction. Int. J. Mol. Sci. 17, 696 (2016).
10. Murakami, Y., Toyoda, H., Tanahashi, T., Tanaka, J., Kumada, T., Yoshioka, Y., Kosaka, N., Ochiya, T., Taguchi, Y. h.: Comprehensive miRNA Expression Analysis in Peripheral Blood Can Diagnose Liver Disease. PLoS One. 7, e48366 (2012).
11. Taguchi, Y.-H.: 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 (2017).
12. Murakami, Y., Kubo, S., Tamori, A., Itami, S., Kawamura, E., Iwaisako, K., Ikeda, K., Kawada, N., Ochiya, T., Taguchi, Y.H.: Comprehensive analysis of transcriptome and metabolome analysis in Intrahepatic Cholangiocarcinoma and Hepatocellular Carcinoma. Sci Rep. 5, 16294 (2015).
13. Tamori, A., Murakami, Y., Kubo, S., Itami, S., Uchida-Kobayashi, S., Morikawa, H., Enomoto, M., Takemura, S., Tanahashi, T., Taguchi, Y.-H., Kawada, N.: MicroRNA expression in hepatocellular carcinoma after the eradication of chronic hepatitis virus C infection using interferon therapy. Hepatol. Res. 46, (2016).
14. Taguchi, Y.-H., Iwadate, M., Umeyama, H., Murakami, Y.: Principal component analysis based unsupervised feature extraction applied to bioinformatics analysis. Comput. Methods with Appl. Bioinforma. Anal. 153–182 (2017).
15. Taguchi, Y.H.: Principal Components Analysis Based Unsupervised Feature Extraction Applied to Gene Expression Analysis of Blood from Dengue Haemorrhagic Fever Patients. Sci. Rep. 7, 44016 (2017).
16. Taguchi, Y.-H., Iwadate, M., Umeyama, H., Murakami, Y., Okamoto, A.: Heuristic principal component analysis-based unsupervised feature extraction and its application to bioinformatics. (2014).
17. Taguchi, Y.-H.: 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 (2016).
18. Taguchi, Y.-H., Murakami, Y.: Universal disease biomarker: can a fixed set of blood microRNAs diagnose multiple diseases? BMC Res. Notes. 7, 581 (2014).

19. Taguchi, Y.-H.: Principal component analysis based unsupervised feature extraction applied to budding yeast temporally periodic gene expression. BioData Min. 9, 22 (2016).
20. Umeyama, H., Iwadate, M., Taguchi, Y.-H.: TINAGL1 and B3GALNT1 are potential therapy target genes to suppress metastasis in non-small cell lung cancer. BMC Genomics. 15, S2 (2014).
21. Taguchi, Y.H., Murakami, Y.: Principal component analysis based feature extraction approach to identify circulating microRNA biomarkers. PLoS One. 8, e66714 (2013).
22. Taguchi, Y.-H., Wang, H.: Genetic Association between Amyotrophic Lateral Sclerosis and Cancer. Genes (Basel). 8, 243 (2017).
23. Taguchi, Y.-H., Iwadate, M., Umeyama, H.: SFRP1 is a possible candidate for epigenetic therapy in non-small cell lung cancer. BMC Med. Genomics. 9, (2016).
24. Taguchi, Y. -h., Iwadate, M., Umeyama, H.: Principal component analysis-based unsupervised feature extraction applied to in silico drug discovery for posttraumatic stress disorder-mediated heart disease. BMC Bioinformatics. 16, 139 (2015).
25. Taguchi, Y.-H., Iwadate, M., Umeyama, H.: 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. pp. 1–10 (2015).
26. Y-H. Taguchi, Hideaki Umeyama, Mitsuo Iwadate, Yoshiki Murakami, Akira Okamoto: Heuristic Principal Component Analysis-Based Unsupervised Feature Extraction and Its Application to Bioinformatics. In: Baoying Wang, Ruowang Li, and William Perrizo (eds.) Big Data Analytics in Bioinformatics and Healthcare. pp. 138–162. IGI global (2015).
27. Murakami, Y., Tanahashi, T., Okada, R., Toyoda, H., Kumada, T., Enomoto, M., Tamori, A., Kawada, N., Taguchi, Y.H., Azuma, T.: Comparison of hepatocellular carcinoma miRNA expression profiling as evaluated by next generation sequencing and microarray. PLoS One. 9, (2014).
28. Taguchi, Y.-H.: 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: ICIC2014. pp. 445–455 (2014).
29. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B57, 289–300 (1995).
30. Kuleshov, M. V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., McDermott, M.G., Monteiro, C.D., Gundersen, G.W., Ma'ayan, A.: Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 44, W90–W97 (2016).

31. Liu, Z.-P., Wu, C., Miao, H., Wu, H.: RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database. 2015, bav095 (2015).
32. Gegonne, A., Tai, X., Zhang, J., Wu, G., Zhu, J., Yoshimoto, A., Hanson, J., Cultraro, C., Chen, Q., Guinter, T., Yang, Z., Hathcock, K., Singer, A., Rodriguez-canales, J., Tessarollo, L., Mackem, S., Meerzaman, D., Buetow, K., Singer, D.S.: The General Transcription Factor TAF7 Is Essential for Embryonic Development but Not Essential for the Survival or Differentiation of Mature T Cells. Mol. Cell. Biol. 32, 1984–1997 (2012).
33. Tapias, A., Wang, Z.Q.: Lysine Acetylation and Deacetylation in Brain Development and Neuropathies. Genomics, Proteomics Bioinforma. 15, 19–36 (2017).
34. Pao, G.M., Zhu, Q., Perez-Garcia, C.G., Chou, S.-J., Suh, H., Gage, F.H., O'Leary, D.D.M., Verma, I.M.: Role of BRCA1 in brain development. Proc. Natl. Acad. Sci. 111, E1240–E1248 (2014).
35. Sun, Y., Jia, L., Williams, M.T., Zamzow, M., Ran, H., Quinn, B., Aronow, B.J., Vorhees, C. V., Witte, D.P., Grabowski, G.A.: Temporal gene expression profiling reveals CEBPD as a candidate regulator of brain disease in prosaposin deficient mice. BMC Neurosci. 9, 1–20 (2008).
36. Mantamadiotis, T., Lemberger, T., Bleckmann, S.C., Kern, H., Kretz, O., Villalba, A.M., Tronche, F., Kellendonk, C., Gau, D., Kapfhammer, J., Otto, C., Schmid, W., Schütz, G.: Disruption of CREB function in brain leads to neurodegeneration. Nat. Genet. 31, 47–54 (2002).
37. Suzuki, D.E., Ariza, C.B., Porcionatto, M.A., Okamoto, O.K.: Upregulation of E2F1 in cerebellar neuroprogenitor cells and cell cycle arrest during postnatal brain development. Vitr. Cell. Dev. Biol. - Anim. 47, 492–499 (2011).
38. Wells, T., Rough, K., Carter, D.A.: Transcription Mapping of Embryonic Rat Brain Reveals EGR-1 Induction in SOX2+ Neural Progenitor Cells. Front. Mol. Neurosci. 4, 1–12 (2011).
39. Korb, E., Finkbeiner, S.: PML in the Brain: From Development to Degeneration. Front. Oncol. 3, 1–5 (2013).
40. Witteveen, J. S., Willemsen, M. H., Dombroski, T. C. D., van Bakel, N. H. M., Nillesen, W. M., van Hulten, J. A., Jansen, E. J. R., Verkaik, D., Veenstra-Knol, H. E., van Ravenswaaij-Arts, C. M. A., Wassink-Ruiter, J. S. K., Vincent, M., David, A., Le Cai, S.M.: Haploinsufficiency of MeCP2-interacting transcriptional co-repressor SIN3A causes mild intellectual disability by affecting the development of cortical integrity. Nat. Genet. 48, 877–87 (2016).
41. Dorsky, R.I.: Two tcf3 genes cooperate to pattern the zebrafish brain. Development. 130, 1937–1947 (2003).
42. Beagan, J.A., Duong, M.T., Titus, K.R., Zhou, L., Cao, Z., Ma, J., Lachanski, C. V., Gillis, D.R., Phillips-Cremins, J.E.: YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. Genome Res. 27, 1139–1152 (2017).