Neuroepigenetics 8 (2016) 1–18

Contents lists available at ScienceDirect

Neuroepigenetics

journal homepage: WWW.ELSEVIER.COM/LOCATE/NEPIG

Research paper

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

Y.-H. Taguchi

Department of Physics, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan

ARTICLE INFO

Article history:
Received 4 June 2016
Received in revised form 20 July 2016
Accepted 4 October 2016

Keywords:
Principal component analysis
Feature extraction
Contextual fear conditioning
Histone deacetylase inhibitor

ABSTRACT

Publicly available gene expression profiles of the hippocampus measured during the successful administration of the histone deacetylase inhibitor, CI-994, to assist the extinction of mice contextual fear conditioning were re-analyzed using the recently proposed principal component analysis based unsupervised feature extraction. We identified 30 genes associated with differential gene expression in the hippocampus of mice treated with the HDAC inhibitor compared to controls; most of these genes code for postsynaptic density proteins. These 30 genes significantly overlapped with those detected by treatment with another HDAC inhibitor, FTY720, during similar contextual fear conditioning. However, because the 30 genes did not strongly overlap with genes associated with histone acetylation during contextual fear conditioning, altered histone modification in response to HDAC inhibitor treatment might not be the primary mechanism of effective extinction of contextual fear conditioning. Based on the results of our analyses we propose that HDAC inhibitors affect the temporal expression of the above genes via direct as well as indirect mechanisms that involve calcium signaling.

1. Introduction

Post-traumatic stress disorder (PTSD) is a psychologically important disorder that affects human behavior, but also causes other diseases including heart failure (Taguchi et al., 2015a). The extinction of fear conditioning is critically important for the treatment of PTSD (VanElzakker et al., 2014). The use of rodent models of fear memory conditioning is useful to investigate the mechanisms involved in memory formation (Baldi and Bucherelli, 2015). Fear conditioning as well as its memory retention can be detected in animals by careful observation of animal behavior (freezing reaction). The use of omics data analyses (Hong et al., 2013), e.g., gene expression analyses, has helped our understanding of the molecular biological background of memory formation. Primary important brain regions for fear memory formation include the amygdala and hippocampus (Zelikowsky et al., 2014). Gene expression and epigenetic markers have been observed for formation, extinction, and retrieval of fear memory in the short and long term, in drug treatment and in gene knock out studies (Mamiya et al., 2009; Lin et al., 2011; Havekes et al., 2008; Yamada et al., 2009).

Recently, the epigenetic effect on fear memory formation has attracted researchers' interests. Epigenetic effects are considered to be related to long term memory (Gupta et al., 2010), because it can be maintained for longer periods while displaying plasticity. Thus, epigenetics might have a role in long term memory including fear memory formation.

The extinction of fear memory by retrieval of the fear memory itself is an interesting subject, because the extinction of fear memory is not thought to erase the old memory but rather to form new memory that overwrites the old fear memory. Generally, the extinction of a short term fear memory by retrieval is easy but it is more difficult to extinguish a remote long term memory (Inda et al., 2011). Recently, Gräff et al. (2014) showed that the successful extinction of a remote contextual fear memory was aided by treatment with an inhibitor of histone deacetylation, CI-994 (HDACi). Histone deactylation affects brain function; specifically it has been implicated in fear memory extinction (Volmar and Wahlestedt, 2015; Whittle and Singewald, 2014). Gräff et al. suggested that the altered histone acetylation of key genes in the hippocampus by CI-994, was essential for the effective extinction of contextual fear memory.

In this study, we re-analyzed the gene expression profiles measured by Gräff et al. (2014) and found that 30 genes in the hippocampus were significantly and differently expressed between

E-mail address: tag@granular.com.

http://dx.doi.org/10.1016/j.nepig.2016.10.001

controls and CI-994 treated samples. Moreover, these 30 genes significantly overlapped with the altered expression of hippocampal genes induced by treatment with another HDAC inhibitor, FTY720 (Hait et al., 2014). In addition, these genes did not overlap with genes associated with a previous study (Park et al., 2013) investigating histone acetylation during contextual fear conditioning in the hippocampus. This suggests that the extinction of a contextual fear memory cannot be fully mediated by the alteration of histone acetylation in the hippocampus. We propose that the altered gene expression in the hippocampus by HDACi treatment is mediated not through histone acetylation but rather by inhibition of the direct regulation of target genes by HDAC, specifically HDAC4, as well as regulation through transcription factors including EGR1. This indicates the difficulty and complexity of interpreting altered gene expression induced by HDACi treatment.

2. Methods

2.1. Hippocampus gene expression

We used mouse gene expression profiles from two remote contextual fear memory extinction experiments (Gräff et al., 2014; Hait et al., 2014) mediated by two distinct HDAC inhibitors. GSE53794 (Gräff et al., 2014) is a gene expression profile measured by next generation sequencing (NGS) technology. It is composed of six files with three HDACi CI-994 treated samples and three control samples. Downloaded SRA format files were converted to fastq files by fastq-dump in SRA Toolkit 2.4.5 (Bethesda, 2011). Then, the obtained fastq files were mapped to the mm10 genome via TopHat (Kim et al., 2013). The obtained sam files were processed via SAMtools (Li et al., 2009) and htseq-count (Anders et al., 2015). Finally, we obtained the gene expression profiles of the RefSeq genes. GSE57015 (Hait et al., 2014) was measured by microarray and was composed of eight files with four HDACi FTY720 treated samples and four control samples. The provided CEL files were standardized via mas5 function implemented in the AFFY (Gautier et al., 2004) package (Bioconductor Gentleman et al., 2004) and gene expression profiles were extracted. GSE3963 (Keeley et al., 2006) was also measured by microarray and was composed of 26 samples, containing 8 fear conditioning (FC), 8 conditional stimulation (CS) and 10 normal (N) samples. Gene expression profiles stored in a series matrix were normalized to have a mean of zero and a variance of one, and were used because no raw data was provided.

2.2. Hippocampus histone acetylation

Histone acetylation profiles were obtained from GSE30325 (Park et al., 2013) where genome wide H4K5ac was measured after a contextual fear conditioning mice experiment with Chip-Seq technology. Two wig files (GSM751963_CON_H4K5_IC_norm.wig.gz and GSM751966_FC_H4K5_IC_norm.wig.gz) that corresponded to control and H4K5ac treated profiles were downloaded. wig files were processed with the findOverlaps function in rtracklayer (Lawrence et al., 2009) packages (Bioconductor) towards the down/upstream of 1000 bps from transcription start sites of refseq genes (mm9). Refseq gene ids were converted to gene symbols by the biomaRt (Durinck et al., 2009) package (Bioconductor).

2.3. Principal component analysis based unsupervised feature extraction

Although our proposed method, PCA based unsupervised FE, was extensively and successfully applied to various biological problems (Taguchi, 2014, 2015; Taguchi et al., 2015a, 2015b, 2015c; Umeyama et al., 2014; Murakami et al., 2012, 2014, 2015; Taguchi and Murakami, 2013, 2014; Kinoshita et al., 2014; Ishida et al., 2014; Taguchi and Okamoto, 2012), we briefly review the methodology here. Basically, it is composed of two parts, gene embedding parts and gene selection parts (Fig. 1).

In brief, PCA based unsupervised FE, in contrast to the ordinary usage of PCA, uses features (genes) embedded into the low dimensional space rather than samples. After specifying PCs that exhibit biological significance, features as outliers along the specified PC are extracted as important features. The philosophy behind this methodology is that if a set of features have common dependence upon samples, no matter what they are, they are more likely to construct PCs because PCs represents the majority of behaviors. PCs that exhibit clear sample dependence likely represent biological significance, e.g., the distinction between control and treated samples. Although there is no evidence to support this hypothesis, it is such a simple methodology that it is not computationally challenging at all, thus is worthwhile trying. Gene expression profiles are normalized so as to have a mean of zero and unit variance before applying PCA.

2.4. Gene embedding by PCA

Suppose that we have mRNA expression x_{ij} of ith mRNA of jth sample. It is also supposed that $\frac{1}{N}\sum_i x_{ij} = 0$ and $\frac{1}{N}\sum_i x_{ij}^2 = 1$ where N is the number of genes. X is the matrix whose element is x_{ij}. In contrast to the usual usage of PCA, where samples are embedded, genes (mRNAs) are embedded in the PCA based upon unsupervised FE. Then kth principal component (PC) score u_{ki} attributed to ith gene can be computed as the element of eigen vector $\boldsymbol{u}_k$ of the gram matrix $G \equiv XX^T$, $XX^T\boldsymbol{u}_k = \lambda_k \boldsymbol{u}_k$ where λ_k is eigen value ordered such that $\lambda_{k+1} < \lambda_k$. The kth PC loading v_{kj} attributed to jth sample can be computed as the element of $\boldsymbol{v}_k = X^T\boldsymbol{u}_k$, which is eigen vector of the matrix X^TX, since $X^TX\boldsymbol{v}_k = X^TXX^T\boldsymbol{u}_k = X^T\lambda_k\boldsymbol{u}_k = \lambda_k\boldsymbol{v}_k$.

2.5. Sample embedding PCA

Gene embedding has the tight relationship with the ordinary sample embedding. In sample embedding, it is also supposed that $\frac{1}{M}\sum_j x_{ij} = 0$ instead of that $\frac{1}{N}\sum_i x_{ij} = 0$ and $\frac{1}{N}\sum_i x_{ij}^2 = 1$ where M is the number of samples. Then instead of gram matrix $G \equiv XX^T$, covariance matrix $S \equiv X^TX$ was diagonalized. Eigen vector $\boldsymbol{u}_k$ of the covariance matrix S is PC scores attributed to each sample, while $\boldsymbol{v}_k = X\boldsymbol{u}_k$ is PC loadings attributed to each genes.

Thus the principal difference between sample embedding and gene embedding is either $\frac{1}{N}\sum_i x_{ij} = 0$ or $\frac{1}{M}\sum_j x_{ij} = 0$, but this difference can generally matter so much, since PCA is the diagonalazation of the product of X, not X itself. Thus, the effect of row-wise or column-wise mean extraction is unpredictable. These two generally give us distinct outcomes.

2.6. Feature extraction

In PCA based unsupervised FE, gene embedding was performed. Then after identifying a set Ω_k of PCs whose PC loading are coincident with the distinction between treated and control samples, outlier genes were identified by assuming Gaussian distribution of PC scores using χ squared distribution, $P_i = P\left[>\sum_{k\in\Omega_k}\left(\frac{u_{ki}}{\sigma_k}\right)^2\right]$, where $P[>x]$ is cumulative probability of χ squared distribution when the argument is larger than x and σ_k is standard deviation of kth PC scores. Then, if BH criterion (Benjamini and Hochberg, 1995) adjusted $P_i < 0.01$, ith gene is identified as outlier.

P-values were attributed to PC scores associated with each gene by assuming χ squared distribution (degree of freedom is one for GSE53794 and two for GSE57015 and GSE3963, based on the number of PCs used for extraction). P-values were adjusted by Benjamini and