Identification of aberrant gene expression associated with aberrant promoter methylation in primordial germ cells between E13 and E16 rat F3 generation vinclozolin lineage

Y-H. TAGUCHI[1,a)]

Abstract:
Background Transgenerational epigenetics (TGE) are currently considered important in disease, but the mechanisms involved are not yet fully understood. TGE abnormalities expected to cause disease are likely to be initiated during development and to be mediated by aberrant gene expression associated with aberrant promoter methylation that is heritable between generations. However, because methylation is removed and then re-established during development, it is not easy to identify promoter methylation abnormalities by comparing normal lineages with those expected to exhibit TGE abnormalities.
Methods This study applied the recently proposed principal component analysis (PCA)-based unsupervised feature extraction to previously reported and publically available gene expression/promoter methylation profiles of rat primordial germ cells, between E13 and E16 of the F3 generation vinclozolin lineage that are expected to exhibit TGE abnormalities, to identify multiple genes that exhibited aberrant gene expression/promoter methylation during development.
Results The biological feasibility of the identified genes were tested via enrichment analyses of various biological concepts including pathway analysis, gene ontology terms and protein–protein interactions.

1. Introduction

Transgenerational epigenetics (TGE) [1] describes the transfer of phenotypes between generations without the modification of genome sequences. Because the plant germline arises from somatic cells, TGE is often observed in plants. However, TGE was also reported in the offspring of mammals, when pregnant females are exposed to endocrine disruptions. Many factors are affected by TGE including male infertility [2], anxious behavior [3], mate preference [4], various diseases [5], reprogramming of primordial germ cells [6], and stress responses [7].

In contrast to reports studying the relationship of TGE to various abnormalities, few studies have investigated how TGE occurs. The main difficulty of studying TGE mechanisms is that epigenetic markers such as promoter methylation are not only heritable, but also vary over time during development in the generation associated with TGE. For example, for promoter methylation to affect development, it must be switched on/off during various stages of development [1]. Thus, TGE that affects development is expected to follow a similar time course. Therefore, abnormalities caused by TGE must be related to the aberrant timing of promoter methylation/demethylation when compared with normal organisms. Detecting small irregularities of promoter methylation timing based on comparisons with normal organisms is not easy. For example, Skinner et al [6] recently tried to identify aberrant gene expression associated with aberrant promoter methylation between E13 and E16 germ line in F3 generation vinclozolin lineages, where vinclozolin functions as an endocrine disruptor. Endocrine disruption is thought to cause various diseases especially in reproductive organs, because it is often misrecognized as a hormone effect on the development of reproductive organs. Thus, usage of endocrine disruptors is usually forbidden for public health. Furthermore, vinclozolin was recently observed to cause TGE abnormalities. However, Skinner et al failed to identify strict pairs of aberrant gene expression and promoter methylation for specific genes. They concluded "A comparison between the germ cell DMR (differential DNA methylated regions) and the differentially expressed genes indicated no significant overlap". Thus, our understanding of the mechanisms by which TGE occurs remains poor.

In the present study we applied the recently proposed principal component analysis (PCA)-based unsupervised feature extraction (FE) [8–17] to the data set obtained by Skinner et al [6] and successfully identified a significant overlap between DMR and differentially expressed genes. Various methods for enrichment analyses supported the biological feasibility of the 48 identified RefSeq mRNAs.

1.1 Previous usage of PCA-based unsupervised FE

Here, we briefly review previous studies [8–17] that used PCA-based unsupervised FE. In Refs. [8–11], we applied PCA-based unsupervised FE to microRNA expression for biomarker identi-

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

fication between patients (of various diseases including various cancers, chronic obstructive pulmonary disease, and Alzheimer's disease, etc) and healthy controls; microRNA extracted in an unsupervised manner was combined with linear discriminant analysis. We found a combination of 10–20 microRNAs generally achieved about 80% accuracy. It was also confirmed that the identified set of microRNAs were stable. Thus, this method is robust for the selection of samples. In Ref. [12], we applied PCA-based unsupervised FE to the proteome in a bacterial culture and identified critical proteins in an unsupervised manner. In Ref. [13], we applied PCA-based unsupervised FE to mRNA and miRNA expression of stressed mouse heart. After identifying potential disease causing genes, we performed *in silico* drug discovery of the identified genes. In Ref. [14], we performed integrated analysis of promoter methylation profiles of three distinct autoimmune diseases using PCA-based unsupervised FE and identified many genes commonly associated with aberrant promoter methylation. In Ref. [15], we applied PCA-based unsupervised FE to genotyping/DNA methylation profiles of cancer and identified genotype specific DNA methylation profiles that occurred in cancer genetics. In Refs [16, 17], PCA-based unsupervised FE of mRNA expression and promoter methylation profiles of normal/treated cancer cell lines was investigated. Based upon the integrated analysis of mRNA expression and promoter methylation profiles, we identified potential disease causing genes.

In summary, PCA-based unsupervised FE has mainly been used to compare between patients (or cancer cell lines) and healthy controls excluding one exception [12]. Because it is likely that healthy controls and patients (or control and treated cancer cell lines) exhibit distinct expressions, it is not surprising that PCA-based unsupervised FE detected significant differences, even if most of the biomarker/disease causing genes were identified only by PCA-based unsupervised FE, but not by other methodologies. In this study, we applied PCA-based unsupervised FE to a different factor, the difference between two time points (E13 and E16). These time points represent different developmental stages and thus some differences are expected; however, the time points are separated by only 3 days, and therefore the differences should be much smaller than between healthy controls and patients (or control and treated cancer cell lines). Of note, although Skinner et al [6] reported no aberrant gene expression associated with aberrant promoter methylation between E13 and E16 germ lines in F3 generation vinclozolin lineages, the study was still published. Thus, from a methodological point of view, the purpose of this study was to investigate whether PCA-based unsupervised FE could identify slight differences; thus it is a new challenge for this methodology.

2. Methods

2.1 Gene expression and promoter methylation profiles

Gene expression/promoter methylation profiles were retrieved from the gene expression omnibus (GEO) using GEO ID GSE59511. This super series consists of two subseries, GSE43559 and GSE59510, each of which includes gene expression (using Affymetrix Rat Gene 1.0 ST Array) and promoter methylation (using NimbleGen Rat CpG Island Plus RefSeq Promoter 720k array) information, respectively. Gene expression profiles were directly loaded from GEO to R [18] by `getGEO` function while six files whose names ended with `ratio_peaks_mapToFeatures_All_Peaks.txt.gz` were downloaded and loaded into R using `read.csv` for promoter methylation. Table 1 shows a list of the samples analyzed. GSE43559 (gene expression) consists of eight samples classified into four categories, E13 control, E13 treated, E16 control, and E16 treated. GSE59510 (promoter methylation) consists of six samples classified into two categories, E13 and E16 (all from F3 generation primordial germ lines). Using the ratio between treated and control groups, eight gene expression profiles were converted to alternative eight profiles as follows:

Table 1 Gene expression and promoter methylation profiles.

GEO ID	Description
GSE43559 (gene expression)	
GSM1065332	PGC_E13_F3-Control_biological rep1
GSM1065333	PGC_E13_F3-Control_biological rep2
GSM1065334	PGC_E13_F3-Vinclozolin_biological rep1
GSM1065335	PGC_E13_F3-Vinclozolin_biological rep2
GSM1065336	PGC_E16_F3-Control_biological rep1
GSM1065337	PGC_E16_F3-Control_biological rep2
GSM1065338	PGC_E16_F3-Vinclozolin_biological rep1
GSM1065339	PGC_E16_F3-Vinclozolin_biological rep2
GSE59510 (promoter methylation)	
GSM1438556	E16-Vip2/Cip2
GSM1438557	E13-Vip2/Cip1
GSM1438558	E13-Vip1/Cip1
GSM1438559	E16-Vip1/Cip1
GSM1438560	E16-Vip2/Cip1
GSM1438561	E13-Vip2/Cip2

$$\left(\begin{array}{c} \frac{\text{E13 Control rep1}}{\text{E13 treated rep1}} \\ \\ \frac{\text{E13 Control rep2}}{\text{E13 treated rep2}} \\ \\ \frac{\text{E13 Control rep2}}{\text{E13 treated rep1}} \\ \\ \frac{\text{E13 Control rep1}}{\text{E13 treated rep2}} \\ \\ \frac{\text{E16 Control rep1}}{\text{E16 treated rep1}} \\ \\ \frac{\text{E16 Control rep2}}{\text{E16 treated rep2}} \\ \\ \frac{\text{E16 Control rep2}}{\text{E16 treated rep1}} \\ \\ \frac{\text{E16 Control rep1}}{\text{E16 treated rep2}} \end{array}\right).$$

These were further normalized to have a mean of zero and a variance of one within each sample. Because six samples in GSE59510 were already transformed to a ratio between treated/control samples, these were not normalized. In total, 14 (8+6) samples that exhibited a ratio between control/treated samples were pooled and prepared for further analyses. The only difference between control and treated samples was whether oil or vinclozolin was injected to F1 pregnant rats between E8 and E14. Any other treatments were identical between E13 and E16.

2.2 Principal component analysis–based unsupervised feature extraction

Although this method was described in detail in a recently published review article [19], this methodology is briefly introduced here. Example: x_{ij} is the gene expression/promoter methylation of the ith gene ($i = 1, \ldots, N$) in the jth sample ($j = 1, \ldots, M$). For simplicity, it is assumed that the mean of x_{ij} over i within each j is zero. Then, in contrast to the ordinary usage of PCA where samples are embedded into the low dimensional space, genes are embedded into the low dimensional space by applying PCA. Thus, principal component (PC) scores of the ℓth component, $x_{i\ell}$, ($\ell = 1, \ldots, M$) are attributed to each gene while each sample has contributed $c_{\ell j}$ to the ℓth component. By this definition, $x_{i\ell}$ is expressed as

$$x_{i\ell} = \sum_j c_{\ell j} x_{ij}$$

PCA-based unsupervised FE attempts to extract features (in this specific application, genes) with larger absolute PC scores along the specified ℓth PC.

In the specific application described in the present study, $N'_{\text{expression}}$ probes using gene expression and $N'_{\text{methylation}}$ probes using promoter methylation were selected, respectively. For the computation of P-values of coincident analysis with binomial distribution, $N'_{\text{expression}} = N'_{\text{methylation}} = N'$ for simplicity.

Although there are several ways to determine which PC is employed for FE, the most straightforward and intuitive strategy is to identify PCs that are mostly coincident with categories by employing categorical regression:

$$c_{\ell j} = a_\ell + \sum_k a_{k\ell} \delta_{kj}$$

where a_ℓ and $a_{k\ell}$ are numerical (regression) coefficients. Then, the ℓth PC associated with the (most) significant regression is employed as the PC for FE. Because this study only contained two categories (E13 and E16), we used the t test instead of categorical regression to measure the significance of coincidence between $c_{\ell j}$ and categories.

2.3 Protein–protein interaction enrichment analysis

The obtained RefSeq mRNA IDs were converted to gene names ("official gene symbol") via a gene ID conversion tool implemented in DAVID [20], and the obtained gene names were uploaded to STRING [21] server. Then, "protein–protein interactions" was selected among the pull-down menu of "enrichment", where the expected number of PPIs for the set of genes uploaded and the P-value attributed to identified PPIs are available.

2.4 Gene ID identification for literature searches

Literature searches were performed using gene symbols that were converted from RefSeq mRNAs using DAVID as explained above.

3. Results and Discussion

3.1 Gene selection using PCA-based unsupervised FE

Fig. 1 illustrates the strategy to identify aberrant gene expression associated with aberrant promoter methylation between controls and vinclozolin treated samples during development from E13 to E16. Gene expression and promoter methylation of vinclozolin treated F3 samples were normalized relative to controls. Then, by separately applying PCA-based unsupervised FE to each sample group, the top $N'(\ll N)$ genes were independently selected. The number of commonly selected genes N'' was counted. If N'' was much larger than expected, the selection of aberrant gene expression associated with aberrant promoter methylation was determined to be successful.

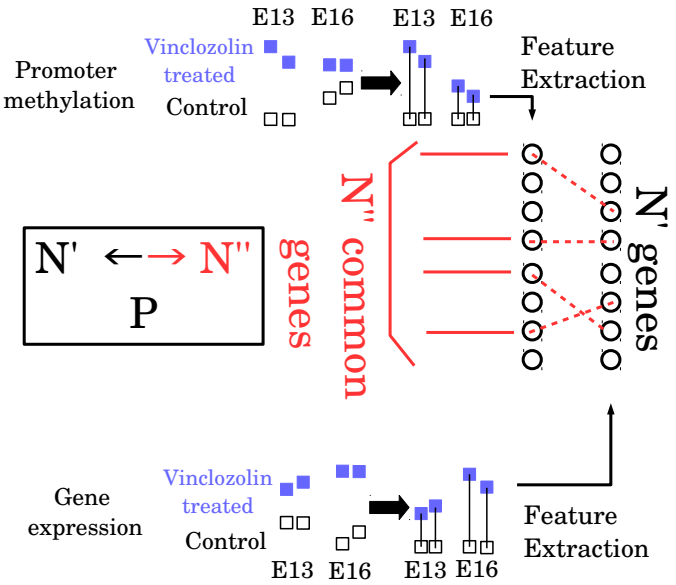

Fig. 1 Schematics that illustrate the procedure of PCA-based unsupervised FE applied to data set analyzed in the present study

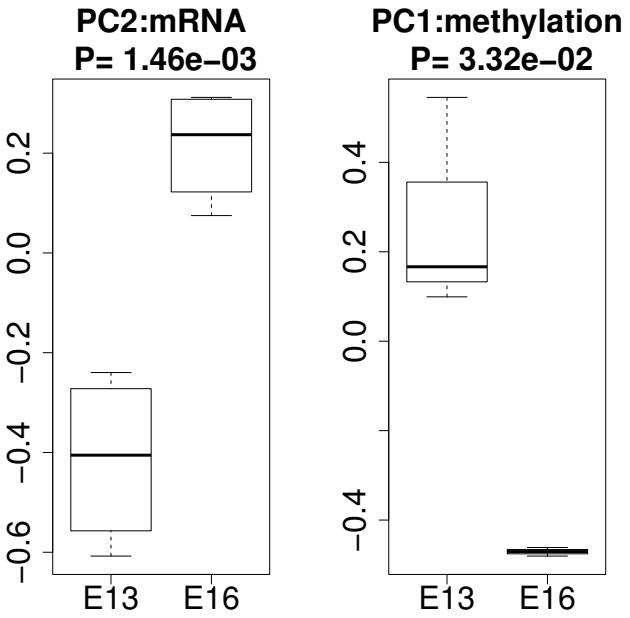

Fig. 2 Boxplots of PCs used for FE in this study, PC2 for mRNA and PC1 for methylation. P-values are computed by t test.

At first, the PCs used for FE shown in Fig. 1 were specified and

a boxplot (PC2 for mRNA and PC1 for methylation) is shown in Fig. 2. These two PCs exhibited a significant distinction between the two categories, E13 and E16. Using the specified PCs, PCA-based unsupervised FE was performed. Then, the most significant N' genes were extracted for gene expression and promoter methylation, respectively. P-values to determine whether the coincidence and the number of commonly selected genes among N' genes occurred accidentally was computed by binomial distribution. How the P-values varied dependent upon N' was determined. Fig. 3 shows the dependence of P-values upon N' when N = 13324, the number of genes commonly included in gene expression and promoter methylation profiles. P-values were smaller for larger N'. However, the minimum N' with P-values less than 0.05 were selected (i.e., N' = 1000) to validate the performance of methodology by enrichment analysis performed in the later part of this study, since smaller number of genes have less abilities to be enhanced. Among the 1000 genes selected in either gene expression or promoter methylation, 48 RefSeq mRNAs were commonly selected (a list of gene names are shown in Table 2). The P-value for N' = 1000 was 0.04 (see Fig. 3). Thus, we successfully selected genes that were significantly associated with simultaneous aberrant gene expression/promoter methylation.

Table 2 48 genes selected by PCA based unsupervised FE when N'=1000

Refseq	gene symbol	Refseq	gene symbol
NM_021866	CCR2	NM_013025	CCL3
NM_030856	lrrn3	NM_001013952	RGD1566251
NM_001099492	Vom2r19	NM_001001053	Olr545
NM_013149	ahr	NM_001024805	HBE2
NM_001000650	Olr624	NM_001000566	Olr542
NM_017061	lox	NM_001000384	Olr408
NM_001109617	Pramel1	NM_022218	cmklr1
NM_012523	Cd53	NM_013158	DBH
NM_001033998	ITGAL	NM_001109374	Lrrtm1
NM_001013177	Sult1c2	NM_021853	KCNT1
NM_053843	Fcgr2b	NM_175586	Taar7b
NM_001109118	Elovl2	NM_001047891	RGD1310507
NM_001106056	TRIM52	NM_138537	LOC171573
NM_001007729	PF4	NM_001000896	Olr1726
NM_001000551	Olr218	NM_001080938	Tas2r124
NM_001000523	Olr1381	NM_001001017	Olr1143
NM_023968	NPY2R	NM_020071	fgb
NM_001000080	Olr1583	NM_017105	BMP3
NM_053994	pdhA2	NM_012893	Actg2
NM_001111321	Vom2r80	NM_001000619	Olr727
NM_001107036	MPO	NM_001012112	Ankrd9
NM_020104	MYL1	NM_012909	AQP2
NM_001000600	Olr796	NM_001108651	HEBP1
NM_022696	HAND2	NM_001014222	Dmrtc1c

To biologically validate these 48 RefSeq mRNAs, we uploaded them to three enrichment analyses servers, DAVID [20], TargetMine [22] and g:Profiler [23]. We observed some biological terms were enriched among the selected genes (Table 3) in spite of the selection of minimum number of significant genes. Almost 50% of the genes selected belonged to G-protein coupled receptors (GPCR) or cell surface receptor pathways, which was expected because an endocrine disruptor such as vinclozolin targets cell surface receptors. We also estimated PPI enrichment (see methods). Because it is rare for proteins to function in the absence of collaboration with other proteins, enriched PPIs among the selected genes (proteins) can provide supporting evidence for

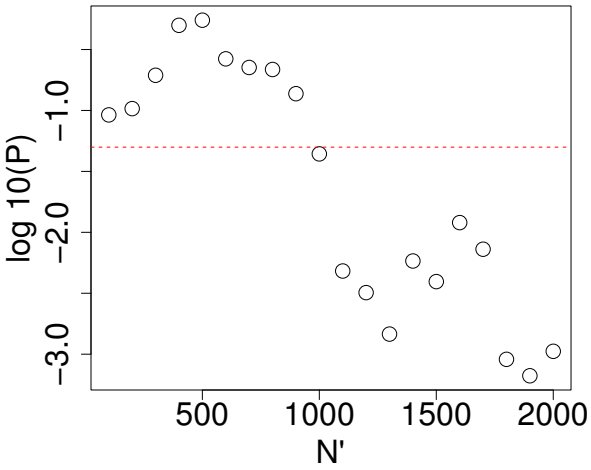

Fig. 3 Dependence of logarithmic P-values that represent the significance of commonly selected genes between gene expression and promoter methylation upon N' when PCA-based unsupervised FE was employed. Horizontal broken red line represents P = 0.05.

the biological significance of selected genes. There were seven PPIs although the expected number of PPIs was three. This resulted in P = 0.05; thus there was significant PPI enrichment among the genes selected by PCA-based unsupervised FE.

Table 3 Enrichment analysis of 48 RefSeq mRNAs commonly selected in the top most 1000 genes by applying PCA-based unsupervised FE to gene expression and promoter methylation. # = the number of genes included.

Biological terms	#	description	P-values
DAVID			
GO BP			
GO:0007186	19	G-protein coupled receptor protein signaling pathway	5.35E-03
GO:0007166	21	Cell surface receptor linked signal transduction	4.19E-03
g:proflier			
GO BP			
GO:0003008	17	System process	4.37E-02
GO:0007166	22	Cell surface receptor signaling pathway	8.91E-03
GO MF			
GO:0060089	17	Molecular transducer activity	4.49E-02
GO:0004871	17	Signal transducer activity	1.82E-02
GO:0004872	17	Receptor activity	1.13E-02
GO:0038023	17	Signaling receptor activity	3.98E-03
GO:0004888	16	Transmembrane signaling receptor activity	1.08E-02
GO:0004930	14	G-protein coupled receptor activity	4.43E-02

P-values shown in Fig. 3 remained significant even when N' increased from 1000 to 2000. Thus, we tried to obtain more genes by setting N' = 2000, because the greater number of genes uploaded would have a tendency to enhance enrichment. There were 179 mRNAs commonly selected between gene expression and promoter methylation (gene names are not shown here). Uploading these genes to three enrichment analyses servers resulted in greater enrichment for these 179 genes as expected (Tables 4,

Table 4 Enrichment analysis of 179 genes commonly selected in the top most 2000 genes by applying PCA-based unsupervised FE to gene expression and promoter methylation. # = the number of genes included.

Biological terms	#	description	*P*-values
DAVID			
KEGG			
rno04740	50	Olfactory transduction	1.63E-15
GO BP			
GO:0007186	79	G-protein coupled receptor protein signaling pathway	2.04E-20
GO:0007166	85	Cell surface receptor linked signal transduction	2.39E-18
GO:0050911	59	Detection of chemical stimulus involved in sensory perception of smell	1.99E-18
GO:0050907	59	Detection of chemical stimulus involved in sensory perception	2.22E-18
GO:0009593	59	Detection of chemical stimulus	3.09E-18
GO:0007608	59	Sensory perception of smell	3.38E-18
GO:0050906	59	Detection of stimulus involved in sensory perception	3.26E-18
GO:0007606	60	Sensory perception of chemical stimulus	2.89E-18
GO:0051606	60	Detection of stimulus	2.88E-18
GO:0007600	61	Sensory perception	3.31E-16
GO:0050890	62	Cognition	2.44E-15
GO:0050877	62	Neurological system process	1.94E-12
GO CC			
GO:0016021	101	Integral to membrane	3.57E-12
GO:0031224	101	Intrinsic to membrane	1.65E-11
GO:0031983	7	Vesicle lumen	1.49E-03
GO:0060205	6	Cytoplasmic membrane-bounded vesicle lumen	7.41E-03
GO:0031091	6	Platelet alpha granule	1.59E-02
GO:0031093	5	Platelet alpha granule lumen	3.82E-02
GO MF			
GO:0004984	60	Olfactory receptor activity	1.59E-19

Table 5 Enrichment analysis of 179 genes commonly selected in the top most 2000 genes by applying PCA-based unsupervised FE to gene expression and promoter methylation. # = the number of genes included.

Biological terms	#	description	*P*-values
g:profiler			
GO BP			
GO:0007606	54	Sensory perception of chemical stimulus	9.14E-21
GO:0007186	65	G-protein coupled receptor signaling pathway	7.61E-20
GO:0050911	50	Detection of chemical stimulus involved in sensory perception of smell	1.44E-19
GO:0007600	58	Sensory perception	2.89E-19
GO:0050907	50	Detection of chemical stimulus involved in sensory perception	5.26E-19
GO:0007608	50	Sensory perception of smell	5.65E-19
GO:0009593	50	Detection of chemical stimulus	1.72E-18
GO:0050906	50	Detection of stimulus involved in sensory perception	3.39E-18
GO:0007166	84	Cell surface receptor signaling pathway	4.19E-18
GO:0003008	69	System process	8.92E-18
GO:0051606	51	Detection of stimulus	1.26E-17
GO:0050877	59	Neurological system process	3.82E-16
GO:0051716	106	Cellular response to stimulus	6.09E-13
GO:0042221	84	Response to chemical	9.54E-13
GO:0050896	116	Response to stimulus	4.65E-12
GO:0007154	98	Cell communication	4.91E-12
GO:0007165	92	Signal transduction	2.84E-11
GO:0044700	95	Single organism signaling	6.05E-11
GO:0023052	95	Signaling	6.70E-11
GO:0065007	131	Biological regulation	3.40E-10
GO:0050789	128	Regulation of biological process	3.48E-10
GO:0050794	120	Regulation of cellular process	1.92E-07
GO:0044707	94	Single-multicellular organism process	9.54E-07
GO:0032501	94	Multicellular organismal process	8.75E-06
GO:0044763	129	Single-organism cellular process	1.17E-05
GO:0044699	135	Single-organism process	1.86E-04
GO:0046010	3	Positive regulation of circadian sleep/wake cycle, non-REM sleep	2.21E-02
GO CC			
GO:0016021	88	Integral component of membrane	1.13E-12
GO:0031224	88	Intrinsic component of membrane	3.85E-12
GO:0071944	79	Cell periphery	1.19E-08
GO:0044425	92	Membrane part	1.43E-08
GO:0005886	77	Plasma membrane	3.24E-08
GO:0016020	97	Membrane	1.09E-02
GO MF			
GO:0038023	70	Signaling receptor activity	5.11E-023
GO:0004930	64	G-protein coupled receptor activity	5.42E-023
GO:0004888	68	Transmembrane signaling receptor activity	1.3E-022
GO:0004871	72	Signal transducer activity	1E-021
GO:0004872	70	Receptor activity	4.63E-021
GO:0060089	72	Molecular transducer activity	5.95E-020
GO:0004984	50	Olfactory receptor activity	1.39E-019
KEGG			
KEGG:04740	42	Olfactory transduction	6.46E-014
KEGG:05144	5	Malaria	1.96E-02

5, and 6).

GPCR and cell surface receptors were enhanced and olfactory transduction related biological terms were vastly enriched. Careful investigation of the selected genes indicated that many olfactory receptor proteins were newly identified when N' was increased from 1000 to 2000. Olfactory receptor proteins were also recognized by Skinner et al [6]. Thus, the identification of many olfactory receptor proteins suggested the correctness and superiority of our methodology, because Skinner et al [6] did not identify reciprocal relationships between gene expression and promoter methylation, probably owing to a lack of suitable statistical methods, although they noted their importance.

PPI enrichment significance was also enhanced when N' increased from 1000 to 2000. There were 360 PPIs among 179 genes while the expected number of PPIs was 191. This resulted in $P = 0$ (within the numerical accuracy adopted); thus the significance of PPI enrichment was enhanced. The increase of PPIs was mostly due to the newly identified olfactory receptor proteins.

These data suggest the biological suitability of our methodology.

4. Conclusions

This study re-analyzed the gene expression/promoter methylation profiles of primordial germ cells between E13 and E16 rat F3 generation vinclozolin lineage [6]. In contrast to analyses performed previously [6], we successfully identified various genes associated with aberrant promoter methylation/gene expression using treated and control samples. Identified genes were related to previously reported diseases in F3 generation vinclozolin lineage. The success of the study methodology suggests the possibility that abnormalities in F3 generation vinclozolin lineage are mediated by heritable aberrant promoter methylation during development between generations.

References

[1] Heard, E. and Martienssen, R. A.: Transgenerational epigenetic inheritance: myths and mechanisms, *Cell*, Vol. 157, No. 1, pp. 95–109

Table 6 Enrichment analysis of 179 genes commonly selected in the top most 2000 genes by applying PCA-based unsupervised FE to gene expression and promoter methylation. # = the number of genes included.

Biological terms	#	description	*P*-values
TargetMine			
GO BP			
GO:0007600	43	Sensory perception	5.81E-12
GO:0007606	40	Sensory perception of chemical stimulus	8.20E-12
GO:0050907	38	Detection of chemical stimulus involved in sensory perception	2.64E-11
GO:0051606	39	Detection of stimulus	2.64E-11
GO:0009593	38	Detection of chemical stimulus	3.56E-11
GO:0050906	38	Detection of stimulus involved in sensory perception	3.60E-11
GO:0003008	48	System process	3.93E-11
GO:0050877	43	Neurological system process	5.27E-11
GO:0007186	41	G-protein coupled receptor signaling pathway	3.63E-09
GO:0007166	46	Cell surface receptor signaling pathway	2.01E-06
GO:0042221	59	Response to chemical	3.63E-06
GO:0044707	60	Single-multicellular organism process	3.94E-05
GO:0032501	61	Multicellular organismal process	6.44E-05
GO:0050911	24	Detection of chemical stimulus involved in sensory perception of smell	9.90E-05
GO:0007608	24	Sensory perception of smell	1.24E-04
GO:0051716	59	Cellular response to stimulus	1.04E-03
GO:0007165	49	Signal transduction	1.30E-03
GO:0050896	68	Response to stimulus	2.64E-03
GO:0065007	75	Biological regulation	4.11E-03
GO:0007154	52	Cell communication	4.97E-03
GO:0050789	71	Regulation of biological process	1.07E-02
GO:0023052	49	Signaling	2.56E-02
GO:0044700	49	Single organism signaling	2.56E-02
GO:0044699	84	Single-organism process	4.23E-02
GO CC			
GO:0016021	46	Integral component of membrane	8.14E-07
GO:0031224	46	Intrinsic component of membrane	9.96E-07
GO:0044425	51	Membrane part	3.43E-04
GO:0016020	56	Membrane	2.37E-02
GO MF			
GO:0004871	45	Signal transducer activity	5.80E-10
GO:0004888	43	Transmembrane signaling receptor activity	5.80E-10
GO:0038023	44	Signaling receptor activity	5.80E-10
GO:0004872	44	Receptor activity	7.10E-10
GO:0060089	45	Molecular transducer activity	7.10E-10
GO:0004984	24	Olfactory receptor activity	8.31E-05
KEGG			
rno04740	50	Olfactory transduction	1.05E-13

(2014).

[2] Guerrero-Bosagna, C., Savenkova, M., Haque, M. M., Nilsson, E. and Skinner, M. K.: Environmentally induced epigenetic transgenerational inheritance of altered Sertoli cell transcriptome and epigenome: molecular etiology of male infertility, *PLoS ONE*, Vol. 8, No. 3, p. e59922 (2013).

[3] Skinner, M. K., Anway, M. D., Savenkova, M. I., Gore, A. C. and Crews, D.: Transgenerational epigenetic programming of the brain transcriptome and anxiety behavior, *PLoS ONE*, Vol. 3, No. 11, p. e3745 (2008).

[4] Skinner, M. K., Savenkova, M. I., Zhang, B., Gore, A. C. and Crews, D.: Gene bionetworks involved in the epigenetic transgenerational inheritance of altered mate preference: environmental epigenetics and evolutionary biology, *BMC Genomics*, Vol. 15, p. 377 (2014).

[5] Anway, M. D., Leathers, C. and Skinner, M. K.: Endocrine disruptor vinclozolin induced epigenetic transgenerational adult-onset disease, *Endocrinology*, Vol. 147, No. 12, pp. 5515–5523 (2006).

[6] Skinner, M. K., Guerrero-Bosagna, C., Haque, M., Nilsson, E., Bhandari, R. and McCarrey, J. R.: Environmentally induced transgenerational epigenetic reprogramming of primordial germ cells and the subsequent germ line, *PLoS ONE*, Vol. 8, No. 7, p. e66318 (2013).

[7] Crews, D., Gillette, R., Scarpino, S. V., Manikkam, M., Savenkova, M. I. and Skinner, M. K.: Epigenetic transgenerational inheritance of altered stress responses, *Proc. Natl. Acad. Sci. U.S.A.*, Vol. 109, No. 23, pp. 9143–9148 (2012).

[8] Murakami, Y., Toyoda, H., Tanahashi, T., Tanaka, J., Kumada, T., Yoshioka, Y., Kosaka, N., Ochiya, T. and Taguchi, Y. H.: Comprehensive miRNA expression analysis in peripheral blood can diagnose liver disease, *PLoS ONE*, Vol. 7, No. 10, p. e48366 (2012).

[9] Murakami, Y., Tanahashi, T., Okada, R., Toyoda, H., Kumada, T., Enomoto, M., Tamori, A., Kawada, N., Taguchi, Y. H. and Azuma, T.: Comparison of Hepatocellular Carcinoma miRNA Expression Profiling as Evaluated by Next Generation Sequencing and Microarray, *PLoS ONE*, Vol. 9, No. 9, p. e106314 (2014).

[10] Taguchi, Y. H. and Murakami, Y.: Principal component analysis based feature extraction approach to identify circulating microRNA biomarkers, *PLoS ONE*, Vol. 8, No. 6, p. e66714 (2013).

[11] Taguchi, Y. H. and Murakami, Y.: Universal disease biomarker: can a fixed set of blood microRNAs diagnose multiple diseases?, *BMC Res Notes*, Vol. 7, p. 581 (2014).

[12] Taguchi, Y.-h. and Okamoto, A.: Principal Component Analysis for Bacterial Proteomic Analysis, *Pattern Recognition in Bioinformatics* (Shibuya, T., Kashima, H., Sese, J. and Ahmad, S., eds.), LNCS, Vol. 7632, Springer International Publishing, Heidelberg, pp. 141–152 (2012).

[13] Taguchi, Y. H., Iwadate, M. and Umeyama, H.: Principal component analysis-based unsupervised feature extraction applied to in silico drug discovery for posttraumatic stress disorder-mediated heart disease, *BMC Bioinformatics*, Vol. 16, No. 1, p. 139 (2015).

[14] Ishida, S., Umeyama, H., Iwadate, M. and Taguchi, Y. H.: Bioinformatic Screening of Autoimmune Disease Genes and Protein Structure Prediction with FAMS for Drug Discovery, *Protein Pept. Lett.*, Vol. 21, No. 8, pp. 828–39 (2014).

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

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

[17] 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, *Intelligent Computing in Bioinformatics* (Huang, D.-S., Han, K. and Gromiha, M., eds.), LNCS, Vol. 8590, Springer International Publishing, Heidelberg, pp. 445–455 (2014).

[18] R Core Team: *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2014).

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

[20] Huang, d. a. W., Sherman, B. T. and Lempicki, R. A.: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat Protoc*, Vol. 4, No. 1, pp. 44–57 (2009).

[21] Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J. and von Mering, C.: STRING v10: protein-protein interaction networks, integrated over the tree of life, *Nucleic Acids Res.*, Vol. 43, No. Database issue, pp. D447–452 (2015).

[22] Chen, Y. A., Tripathi, L. P. and Mizuguchi, K.: TargetMine, an integrated data warehouse for candidate gene prioritisation and target discovery, *PLoS ONE*, Vol. 6, No. 3, p. e17844 (2011).

[23] Reimand, J., Arak, T. and Vilo, J.: g:Profiler–a web server for functional interpretation of gene lists (2011 update), *Nucleic Acids Res.*, Vol. 39, No. Web Server issue, pp. W307–315 (2011).