Contents lists available at ScienceDirect

Regenerative Therapy

journal homepage: http://www.elsevier.com/locate/reth

Original Article

Variation of DNA methylation on the *IRX1/2* genes is responsible for the neural differentiation propensity in human induced pluripotent stem cells

Asato Sekiya [a], Ken Takasawa [a, 1], Yoshikazu Arai [a], Shin-ichi Horike [b], Hidenori Akutsu [c], Akihiro Umezawa [c], Koichiro Nishino [a, *]

[a] *Laboratory of Veterinary Biochemistry and Molecular Biology, Graduate School of Medicine and Veterinary Medicine/Faculty of Agriculture, University of Miyazaki, 1-1 Gakuen-Kibanadai-Nishi, Miyazaki 889-2192, Japan*

[b] *Division of Integrated Omics Research, Research Center for Experimental Modeling of Human Disease, Kanazawa University, Kanazawa 920-8640, Japan*

[c] *Center for Regenerative Medicine, National Center for Child Health and Development Research Institute, 2-10-1 Okura, Setagaya-ku, Tokyo 157-8535, Japan*

ARTICLE INFO

Article history:
Received 29 September 2022
Received in revised form
5 November 2022
Accepted 17 November 2022

Keywords:
Human iPSCs
Neural stem cells
DNA methylation
Differentiation propensity
Machine learning

ABSTRACT

Introduction: Human induced pluripotent stem cells (hiPSCs) are useful tools for reproducing neural development *in vitro*. However, each hiPSC line has a different ability to differentiate into specific lineages, known as differentiation propensity, resulting in reduced reproducibility and increased time and funding requirements for research. To overcome this issue, we searched for predictive signatures of neural differentiation propensity of hiPSCs focusing on DNA methylation, which is the main modulator of cellular properties.

Methods: We obtained 32 hiPSC lines and their comprehensive DNA methylation data using the Infinium MethylationEPIC BeadChip. To assess the neural differentiation efficiency of these hiPSCs, we measured the percentage of neural stem cells on day 7 of induction. Using the DNA methylation data of undifferentiated hiPSCs and their measured differentiation efficiency into neural stem cells as the set of data, and HSIC Lasso, a machine learning-based nonlinear feature selection method, we attempted to identify neural differentiation-associated differentially methylated sites.

Results: Epigenome-wide unsupervised clustering cannot distinguish hiPSCs with varying differentiation efficiencies. In contrast, HSIC Lasso identified 62 CpG sites that could explain the neural differentiation efficiency of hiPSCs. Features selected by HSIC Lasso were particularly enriched within 3 Mbp of chromosome 5, harboring *IRX1*, *IRX2*, and *C5orf38* genes. Within this region, DNA methylation rates were correlated with neural differentiation efficiency and were negatively correlated with gene expression of the *IRX1/2* genes, particularly in female hiPSCs. In addition, forced expression of the *IRX1/2* impaired the neural differentiation ability of hiPSCs in both sexes.

Conclusion: We for the first time showed that the DNA methylation state of the *IRX1/2* genes of hiPSCs is a predictive biomarker of their potential for neural differentiation. The predictive markers for neural differentiation efficiency identified in this study may be useful for the selection of suitable undifferentiated hiPSCs prior to differentiation induction.

Abbreviations: PSC, pluripotent stem cell; NSCs, neural stem cells; PAX6, Paired Box 6; Lasso, least absolute shrinkage and selection operator; dpc, days post coitum; IRX, Iroquois Homeobox; C5orf38, Chromosome 5 open reading frame 38; EGFP, enhanced green fluorescent protein.

[*] Corresponding author.
E-mail address: aknishino@cc.miyazaki-u.ac.jp (K. Nishino).

Peer review under responsibility of the Japanese Society for Regenerative Medicine.

[1] Present address: Cancer Translational Research Team, RIKEN Center for Advanced Intelligence Project, Tokyo, 103-0027, Japan

https://doi.org/10.1016/j.reth.2022.11.007

1. Introduction

The ability of human induced pluripotent stem cells (hiPSCs) to self-renew and differentiate into a variety of tissues *in vitro* [1] makes them a very useful tool for regenerative medicine and drug screening [2,3]. However, the differentiation propensity of human PSCs differs with cell lines [4–6], leading to reduced reproducibility and increased time and cost burdens for research.

Recently, there has been growing evidence of DNA methylation variations among iPSC lines, such as residual patterns in the origin cells [7], random aberrations accompanying epigenome-wide rewriting [8,9], and fluctuations associated with continuous PSC culturing [10,11]. Since DNA methylation is a key component of epigenetic mechanisms that regulate a wide variety of nuclear events [12], fluctuations in DNA methylation in individual iPSC lines are likely to be the cause of quality variation, notably variation in differentiation propensity.

Most of the reported biomarkers that predict the ability of human PSCs to differentiate into specific tissues are based on their transcriptome [13–18]. The transcriptome reflects the profile of the cell at the time of observation. The epigenetic signature, on the other hand, carries the memory of environments and events, and regulates current and future gene expression patterns. Epigenetic marks such as DNA methylation gives us insight into the potential as well as the existing properties of the cell, suggests its usefulness as a predictive biomarker.

Machine learning has been actively used for biomarker discovery [19–21]. Supervised machine learning trains models are based on a set of input data and the resulting response (output), thereby eliminating the need for prior hypotheses and promising novel discoveries beyond human cognition. Yamada et al. developed the Hilbert-Schmidt Independence Criterion Lasso (HSIC Lasso) [21,22], which is a supervised machine learning-based feature selection method. HSIC Lasso has two advantages: it can consider input–output nonlinear relationships and it is effective for high-dimensional data. Therefore, HSIC Lasso is superior to existing methods such as classical Lasso in machine learning-based feature selection from omics data [24].

The objective of this study was to identify signatures in undifferentiated hiPSCs that could predict neural differentiation efficiency. We used HSIC Lasso and comprehensive DNA methylation data from 32 undifferentiated hiPSCs and identified loci harboring *IRX1*, *IRX2*, and *C5orf38* genes. Our study provides a clue for understanding the differentiation propensity of human iPSCs.

2. Materials and methods

2.1. Preparations of mouse embryonic fibroblasts (MEFs) and MEF feeder cells

MEFs were isolated from 13.5 dpc fetuses of pregnant CD1 (ICR) mice (Charles River Japan, Inc., Yokohama, Japan) and cultured in Dulbecco's modified Eagle's high-glucose medium (DMEM; Sigma-Aldrich, St Louis, MO, USA) containing 10% fetal bovine serum (Thermo Fisher Scientific, Inc., Waltham, MA, USA), 1% penicillin and streptomycin (Thermo Fisher Scientific), and 0.1% 2-mercaptoethanol (Thermo Fisher Scientific). MEFs were irradiated with 30 Gy of gamma radiation to generate MEF feeder cells. All procedures were performed in accordance with the Guidelines for Animal Care and Use of Laboratory Animals of the University of Miyazaki, and the experimental protocols were approved by the Animal Experiment Committee of the University of Miyazaki (no. 2012-017, 2017-009).

2.2. Human cell culture

Human endometrial (UtE1104) and menstrual blood (Edom22) cell lines were independently established [25]. Human fetal lung fibroblast cells (MRC-5 and IMR-90) [26,27] were obtained from the Japanese Collection of Research Bioresources (JCRB) Cell Bank, Japan. Human dermal fibroblasts (DFM1, DFM2, DFMF1, and DFMF2) were purchased from ZenBio, Inc (Research Triangle Park, NC, USA). These human somatic cells were maintained in Dulbecco's modified Eagle's medium-low glucose (DMEM-LG; Sigma-Aldrich, St. Louis, MO, USA) supplemented with 10% fetal bovine serum, 1% GlutaMAX™ supplement (Thermo Fisher Scientific), 1% penicillin and streptomycin (Thermo Fisher Scientific), and 0.1% 2-mercaptoethanol (Thermo Fisher Scientific). Human Retro-iPSCs were generated using the retroviral vector pMXs, which contain cDNAs for human *OCT3/4*, *SOX2*, *c-MYC*, and *KLF4* [11,28]. Episomal-iPSCs were obtained from the JCRB Cell Bank or established using episomal vectors pCXLE-hOCT3/4-shp53, pCXLE-hSK, and pCXLE-hUL [29]. RNA-iPSCs were established using the StemRNA 3rd Gen Reprogramming Kit (REPROCELL, Inc., Kanagawa, Japan) according to the manufacturer's recommendations. Human iPSCs were maintained on irradiated MEF feeder cells in KnockOut™ Dulbecco's modified Eagle medium (KO-DMEM; Thermo Fisher Scientific) containing 20% knockout serum replacement (Thermo Fisher Scientific), 1% GlutaMAX™ (Thermo Fisher Scientific), 1% nonessential amino acids (Thermo Fisher Scientific), 1% penicillin and streptomycin (Thermo Fisher Scientific), 0.1% 2-mercaptoethanol (Thermo Fisher Scientific), and 10 ng/mL recombinant human basic fibroblast growth factor (bFGF; FUJIFILM Wako Pure Chemical Corp., Ltd., Osaka, Japan). All the human cell lines used in this study are summarized in Supplemental Table 1. Ethical approval for the use of human cell lines, in this study, was obtained from the Institutional Review Board of the National Institute for Child Health and Development and University of Miyazaki (no. 2016-1). All procedures performed in this study that involved the handling of human cells were in accordance with the ethical standards of the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

2.3. Genome-wide DNA methylation analysis

DNA methylation profiles were obtained for each sample using the Infinium MethylationEPIC BeadChip (Illumina Inc., San Diego, CA, USA). Genomic DNA was extracted from cells using DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany). From each sample, 1 μg of genomic DNA was subjected to bisulfite conversion using EZ DNA Methylation kit (Zymo Research, Orange, CA, USA) according to the manufacturer's recommendations. Following bisulfite conversion, genomic DNA was hybridized with the MethylationEPIC BeadChip, and each BeadChip was scanned on an iScan (Illumina Inc.) according to the manufacturer's instructions. GenomeStudio (Illumina Inc.) was used for background subtraction and data normalization. Methylated and unmethylated signals were used to compute the β-value, a quantitative score of the DNA methylation rate that ranges from “0.00” for a completely unmethylated state to “1.00” for a completely methylated state. Detailed information on the cell lines and GEO accession numbers used in this study are provided in Supplemental Table 1. Probes harboring common SNPs (minor allele frequency [MAF] > 1%) within 10 bases of the 3′ end, based on the 1000 Genomes Project and the Japanese Multi Omics Reference Panel [30,31], were eliminated from further analysis. Probes with a detection p-value ≥ 0.05 (computed from the background based on negative controls) or blank β-values in more than 10% of samples were also eliminated from further analysis. The blank or non-significant β-values were replaced with the median β-value of the common probe. A total of 806,442 CpG sites were analyzed in 32 samples (Supplemental Fig. 1). Unsupervised hierarchical clustering analysis (HCA) with Euclidean distance, Ward's method, and uniform manifold approximation and projection (UMAP)-based two-dimensional embedding of principal components holding approximately 90% of the variance obtained using principal component analysis (PCA) were used for analysis [32]. DNA methylation pluripotency markers of hiPSCs and fibroblasts were assessed based on previous reports [28,33]. The Epi-Pluri-Score was calculated as the difference in the DNA methylation rate on *ANKRD46* (β-value in cg23737055) and *VRTN* (β-value in cg22247240) [34]. Pearson's rank correlation coefficient was used for correlation analysis between DNA methylation rates and differentiation efficiency and between DNA methylation rates for individual probes.

2.4. Genome-wide gene expression analysis

Gene expression array data were obtained from each sample using SurePrint G3 human GE microarrays 8 × 60 K (Agilent Technologies, Santa Clara, CA). Total RNA was extracted from cells using TRIzol (Thermo Fisher Scientific) and purified using RNeasy columns (Qiagen). The total RNA quality was checked for integrity using a high-sensitivity RNA ScreenTape (Agilent) on an Agilent TapeStation 2200 instrument by following the manufacturer's protocol. From each sample, 50 ng of total RNA was amplified and labeled using an Agilent Low-Input QuickAmp labeling kit, according to the manufacturer's instructions. Subsequently, Cy3-labeled cRNA was fragmented and hybridized onto a SurePrint G3 human GE microarrays 8 × 60 K slide. The slides were then washed and scanned using an Agilent microarray scanner system. The features of the scanned image files were extracted using Agilent feature extraction. Background correction and quantile normalization between arrays were performed using the limma package in R. The expression levels shown by different probes for the same transcript were summarized by median polishing. A total of 48,588 transcripts were analyzed in 28 samples. Unsupervised HCA with Euclidean distance, Ward's method, and UMAP-based two-dimensional embedding of principal components holding approximately 90% of the variance obtained using PCA were used for analysis.

2.5. Neural stem cell differentiation

Neural stem cell induction was conducted using the dual-SMAD inhibition protocol [35] with STEMdiff™ neural induction medium (NIM; StemCell Technologies, Inc., Vancouver, Canada). Human iPSCs were dissociated using the Gentle cell dissociation reagent (StemCell Technologies), and cell suspension were passed through a 40 μm cell strainer. Cells were resuspended in NIM supplemented with 10 μM SB431542 (FUJIFILM Wako Pure Chemical Corp.), 100 nM LD193189 (FUJIFILM Wako Pure Chemical Corp.), and 10 μM Rock inhibitor (Y27632) (FUJIFILM Wako Pure Chemical Corp.) and plated at a density of 2 × 10^5 cells/cm^2 on Matrigel-coated plates (BD Biosciences, Bedford, MA, USA). NIM supplemented with 10 μM SB431542 and 100 nM LD193189 was replaced every 24 h after plating. On day 7 of neural induction, the differentiated cells were collected and used for flow cytometric analysis.

2.6. Immunohistochemistry

Human iPSCs and induced neural stem cells were fixed at room temperature in 4% paraformaldehyde for 10 min. The cells were then permeabilized in 0.1% Triton X-100/PBS for 10 min and blocked with 1% BSA/PBS for 1 h. The cells were then incubated with the diluted primary antibody in 1% BSA/PBS for 12 h at 4 °C and subsequently incubated with the diluted secondary antibody in 1% BSA/PBS for 1 h at room temperature. The antibodies used in this study are summarized in Supplemental Table 2.

2.7. Flow cytometry analysis

Differentiated cells were dissociated by treatment with StemPro Accutase Cell Dissociation Reagent (Thermo Fisher Scientific) for 8–10 min. Collected cells were fixed, permeabilized, blocked, and incubated with antibody as mentioned in section 2.6. The antibodies used in this study are summarized in Supplemental Table 2. The analysis was performed using SONY SA3800 spectral cell analyzer (Sony Biotechnology, San Jose, CA, USA) (see also Supplemental Fig. 2A).

2.8. Hilbert-Schmidt independence criterion (HSIC) Lasso based feature selection

For supervised feature selection to find CpG sites related to neural differentiation efficiency of hiPSCs, we used HSIC Lasso [22,23]. HSIC Lasso is a kernel-based minimum redundancy maximum relevance (mRMR) algorithm that uses HSIC to measure the dependency between variables [36,37], which is useful for high-dimensional and small-sample (n << p) datasets and enables the assessment of the nonlinear dependency between the output variable and a feature. For input data, CpG sites showing a standard deviation of DNA methylation greater than 0.1 between the 32 hiPSC lines were extracted. The neural differentiation efficiency of each hiPSC line was used for the output data. HSIC Lasso regression was performed for selecting neural differentiation efficiency-related CpG sites using Python 2/3 package pyHSICLasso (http://github.com/riken-aip/pyHSICLasso). To determine where the features (primary selected neural differentiation efficiency-related CpG sites) and their neighbors selected by HSIC Lasso were enriched in the genome, we performed a one-sided Fisher's exact test for each region defined by the sliding-window method (1 Mbp in each window, sliding at 100 kbp steps). Based on the computed *p*-value, False Discovery Rate (FDR) was estimated using the q value package in R and defined as q-value [38]. Genes located within the window (q-value < 0.01) were analyzed for Gene Ontology (GO) enrichment using the Database for Annotation, Visualization and Integrated Discovery (DAVID; https://david.ncifcrf.gov/) [39].

2.9. Combined bisulfite restriction analysis (COBRA) and bisulfite sequencing

Sodium bisulfite treatment of genomic DNA was performed using the EZ DNA Methylation-Gold Kit (Zymo Research, Irvine, CA, USA). PCR amplification was performed using BIOTAQ HS DNA Polymerase (Bioline Ltd., London, UK) with specific primers. The primer sequences used in this study are summarized in Supplemental Table 3. All PCR experiments were performed under the following thermocycling conditions: 95 °C for 10 min; 35 cycles of 95 °C for 30 s, 60 °C for 30 s, and 72 °C for 1 min; and a final extension at 72 °C for 10 min. For COBRA [40], the PCR product was treated with HpyCH4IV (New England Biolabs Inc., Ipswich, MA, USA) or Taq$^{\alpha}$I (New England Biolabs Inc.). The concentrations of the treated PCR products were measured using MultiNA (SHIMADZU, Kyoto, Japan). To determine the methylation states of individual CpG sites in the *IRX1/2* genes, the PCR product was gel-extracted, subcloned into the T-Vector pMD20 (Takara), and sequenced. Methylation sites were visualized and quality control was performed using the QUMA web-based tool (http://quma.cdb.riken.jp/) [41].

2.10. Quantitative gene expression analysis

Total RNA was extracted from tissues and cells using ISOGEN II (FUJIFILM Wako Pure Chemical Corp.), following the manufacturer's instructions. For reverse transcription-polymerase chain reaction (RT-PCR), first-strand cDNA was synthesized using total RNA (1 μg) with random hexamers and ReverTra Ace reverse transcriptase (TOYOBO Co., Ltd., Osaka, Japan). Quantitative real-time PCR (qPCR) was performed using SYBR® Green PCR master mix (Applied Biosystems, Woburn, MA, USA). Data were normalized to *GAPDH* expression levels. The primer sequences used in this study are summarized in Supplemental Table 3. Gene expression levels are presented as fold-change in expression, which was calculated using the Pfaffl method [42].

2.11. Generation of hiPSC lines with forced IRX1/2 gene expression

We generated an all-in-one PiggyBac transposon vector system consisting of three units flanked by PiggyBac inverted terminal repeats [43]. First, the gene of interest and EGFP reporter gene were placed under the control of the doxycycline-inducible bi-directional TRE3G promoter; CAG promoter drove the expression of Tet-On 3G gene, and PGK promoter drove that of puromycin resistance gene. The human *IRX1*, *IRX2* and *C5orf38* genes were amplified by PCR using Prime STAR HS DNA polymerase (TaKaRa) and cloned into the vector. The constructed PiggyBac vectors were co-transfected with pCMV-hyPBase vector. The plasmid vector was transfected at a 5 μg: 1 μg transposon: transposase ratio in two hiPSC lines. Approximately 1×10^6 cells were transfected using the Neon® transfection system (Thermo Fisher Scientific) with one pulse at 1200 V for 30 ms. Three days after electroporation, the transfected hiPSCs were selected using puromycin (0.5–1 μg/mL) for 4 days and then were cultured with 1 μg/mL doxycycline. The selected cells were confirmed to be more than 99% EGFP-positive by FCM. For neural stem cell differentiation, doxycycline was added to the medium three days before the start of induction until the end.

2.12. Statistical analysis

Differences between two independent samples were evaluated using two-tailed Student's t-test. All error bars represent standard error of the mean. Linear regression and Pearson's product–moment correlation coefficients were used to analyze the correlations between the two variables.

2.13. Accession numbers

NCBI GEO: Infinium MethylationEPIC BeadChip data obtained in this study have been submitted under the accession number GSE214021. SurePrint G3 human GE microarrays 8 × 60 K data obtained in this study have been submitted under the accession number GSE214020.

3. Results

3.1. Assessment of neural differentiation potential of hiPSCs

We first obtained DNA methylation profiles of 32 hiPSC lines and seven original somatic cell lines using the Illumina Infinium MethylationEPIC array. For the preprocessing step of the DNA methylation array data, probes containing SNPs with minor allele frequency (MAF) > 1% within 10 bp of the methylation site were deleted. Of the remaining data, those with low quality of detection were treated as described in section 2.3. Consequently, 806,442 sites were used for the downstream analysis. Using DNA methylation data, we assessed the pluripotency of hiPSCs using previously identified epigenetic markers for pluripotent stem cells [11,28,33]. DNA methylation pattern on the promoters of 17 genes including *OCT4* in all hiPSCs used in this study showed to be pluripotency (Supplemental Fig. 3A). The pluripotency of hiPSC lines was also validated by the Epi-pluri-Score based on DNA methylation at two CpG sites (Supplemental Fig. 3B) [34]. Furthermore, we obtained gene expression profiles using the Agilent microarray and confirmed hiPSC pluripotency by high expression levels of pluripotency genes and low expression levels of fibroblast genes (Supplemental Figs. 3C and 3D).

Next, we assessed the neural stem cell differentiation efficiency of each hiPSC line by measuring the number of PAX6-positive cells that differentiated using the dual SMAD inhibition protocol (Fig. 1A and B, and Supplemental Fig. 2). Among the 32 hiPSC lines, differentiation efficiencies were uniformly distributed, ranging from nearly 0% to 95% (Fig. 1C). The neural differentiation propensity was observed among the hiPSC lines in this study. At first, we analyzed the relationship between the neural differentiation ability and reprogramming methods. The neural differentiation ability of each hiPSC line was not related to reprogramming methods. In addition, we also analyzed about gender, origin cell types or passage number. However, these did not related to the differentiation efficiency (Supplemental Figs. 2B–2E).

There was no difference in DNA methylation rates on epigenetic pluripotency markers between hiPSCs with low and high abilities to differentiate into neural stem cells (Supplemental Fig. 3A). In addition, there was no correlation between the Epi-Pluri-Score and neural differentiation efficiency (Supplemental Fig. 3B). The expression levels of pluripotency marker genes also did not correlate with neural differentiation efficiency (data not shown). These results are consistent with previous reports that state the expression levels of pluripotency markers, such as *OCT4* and *NANOG*, in undifferentiated hiPSCs were unrelated to their capacity for subsequent neural differentiation [13,18].

We next performed HCA based on DNA methylation rates of 806,442 CpG sites. The hiPSC lines were divided into three groups (Clusters 1, 2, and 3; Fig. 1C). We did not find relevance between neural differentiation ability and these three groups (Fig. 1D). Two-dimensional visualization of the whole genome DNA methylation profiles by dimensionality reduction of 806,441 features using UMAP also did not show a distribution of hiPSC lines dependent on differentiation efficiency (Fig. 1E). The same analyses were performed on transcriptome data. HCA using 48,588 transcripts classified the cells into three cell groups, but no association between the group and neural differentiation efficiency was found (Supplemental Figs. 4A and 4B). UAMP using gene expression profiles also did not show a distribution of hiPSC lines dependent on differentiation efficiency (Supplemental Fig. 4C).

3.2. Screening of neural differentiation efficiency-related differentially methylated sites by HSIC Lasso

Recent high-throughput platforms for biological assays provide a large number of features (p); thus, in most cases, the number of observations (n) is much less than p (p>>n). Consequently, conventional statistical hypothesis testing and machine learning cannot adequately assess the similarities and differences between samples, making it difficult to discover important features. To address this problem, we used HSIC Lasso, which is a supervised machine learning-based feature selection method.

Firstly, 60,728 CpG sites exhibiting standard deviations of DNA methylation rates greater than 0.1 among 32 hiPSC lines were extracted to obtain a dataset of DNA methylation sites that could more robustly affect the phenotype (Fig. 2A and Supplemental Fig. 5A). HCA and UMAP using 60,728 high standard deviation

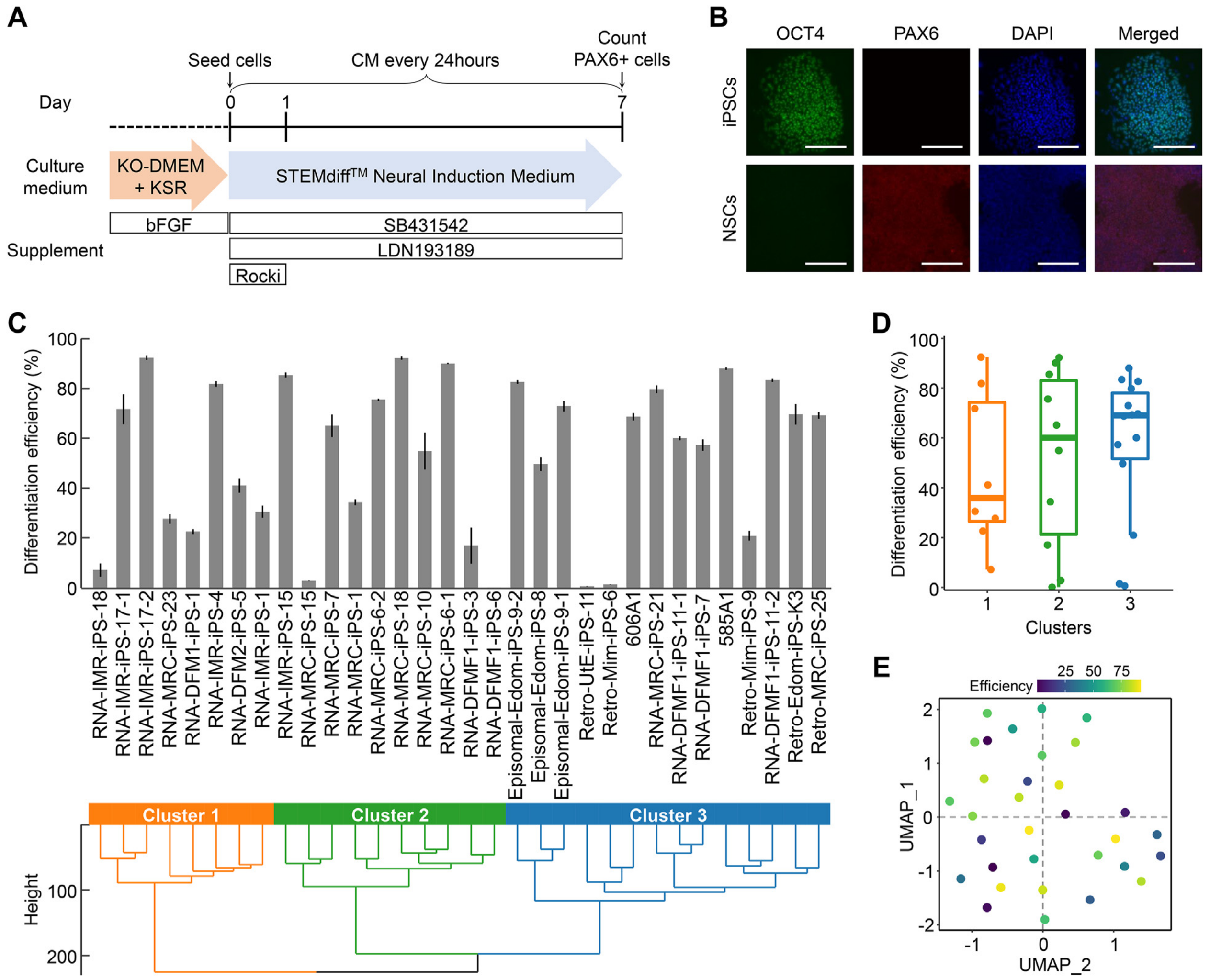


Fig. 1. Neural induction and comprehensive DNA methylation analysis (A) Schematic of the dual-SMAD inhibition protocol used for neural differentiation. (B) Representative images of hiPSCs and induced neural stem cells (NSCs) after immunostaining for the OCT-4 and PAX6, and DAPI staining for DNA. Scale bars represent 200 μm. (C) Neural differentiation efficiency in 32 hiPSC lines and unsupervised hierarchical clustering analysis (HCA) using 806,442 CpG sites. The differentiation efficiency was calculated by the positive rate of PAX6 cells after neural stem cell induction. (D) Comparisons of neural differentiation efficiency between the three groups separated based on HCA (C). (E) 2D representation of Uniform Manifold Approximation and Projection (UMAP)-based dimensionality reduction of the top 18 principal components obtained from PCA analysis using 806,442 CpG sites. The distance between hiPSC lines, indicated by the dots, shows the degree of similarity between the samples as represented by DNA methylation. Each hiPSC sample indicated as a dot is colored by differentiation efficiency. There was no association between sample location and neural differentiation efficiency.

sites was not able to distinguish hiPSCs with different neural differentiation efficiencies (Supplemental Figs. 5B–5D). These results did not explain the differentiation efficiency of hiPSCs by unsupervised methods. Therefore, we conducted subsequent analysis using supervised machine learning. The DNA methylation rates of the 60,728 sites as input features and the differentiation efficiency as output values were used for HSIC Lasso feature selection. As a result, 62 CpG sites were extracted (Fig. 2A and Supplemental Table 4). UMAP with DNA methylation rates of 62 CpG sites showed that 16 of the 17 hiPSC lines with neural differentiation efficiencies >60% were located in the third and fourth quadrants, whereas all other lines were located in the first and second quadrants (Fig. 2B), suggesting that distribution of hiPSC lines depended on differentiation efficiency. In addition, there was a strong correlation between the coordinates on each axis of UMAP and neural differentiation efficiency (UMAP_1, Pearson's $r = -0.84$, $p < 0.01$; UMAP_2, $r = -0.62$, $p < 0.01$), indicating that the 62 CpG sites could explain the neural differentiation propensity of hiPSCs.

Because HSIC Lasso is designed specifically to select non-redundant features, features that exhibit similar DNA methylation patterns tend to be eliminated except for one [22]. Although the 62 selected features were representative CpG sites describing the neural differentiation ability of hiPSCs, other sites exhibiting similar DNA methylation behavior among the hiPSC lines (hereinafter called "neighboring features") were not considered. The neighboring features were only mechanically excluded in a non-redundant feature selection step, but those might include important features. Therefore, 100 neighboring features of each 62 selected features were extracted for further analysis. Consequently, a total of 5393 CpG sites, excluding duplicates, were analyzed as neural differentiation efficiency-related differentially methylated sites (ND-DMSs) (Fig. 2A and Supplemental Table 4). We examined whether ND-DMSs were enriched at specific coordinates in the genome. To reduce the variability of the calculated results owing to differences in bin ranges, a sliding window approach (1 Mb in each window, sliding by a step of 100 kb) was used to evaluate the enrichment of ND-DMSs within each window. Here, the significance of the proportion of probes located within the window among all ND-DMSs relative to that among all probes designed in the methylation EPIC was calculated using Fisher's exact test.

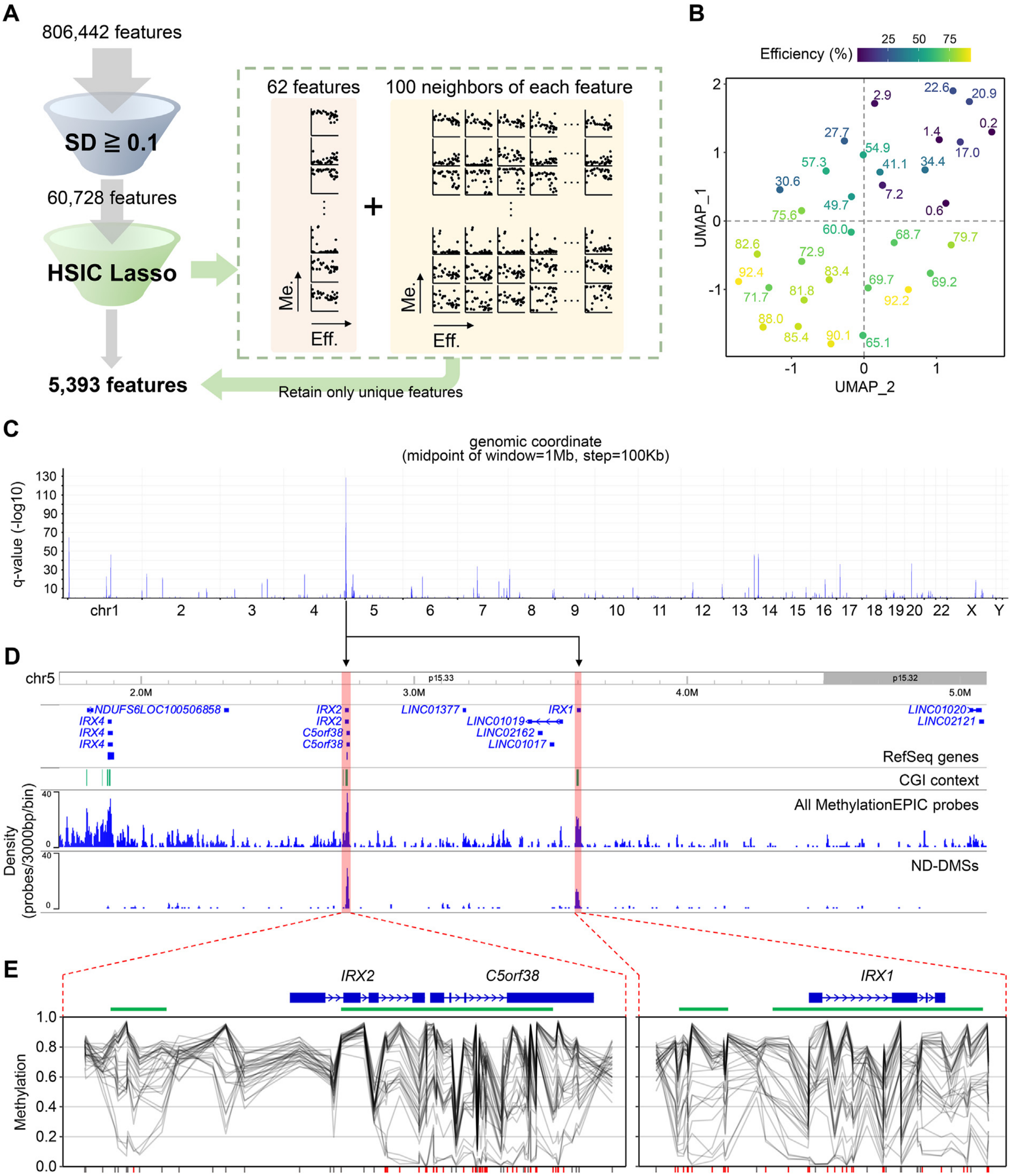
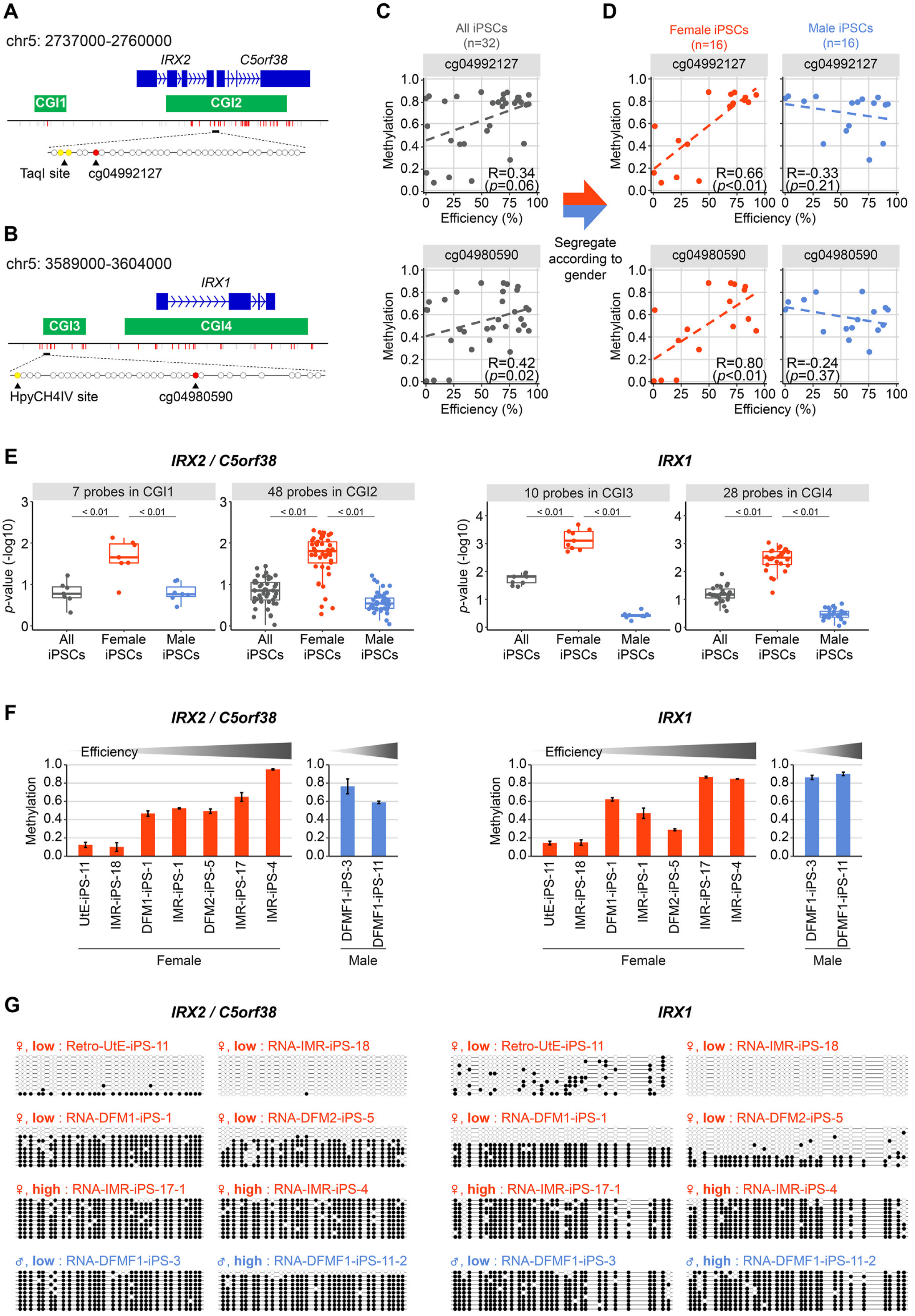


Fig. 2. HSIC-Lasso based feature selection and analysis of DNA methylation sites related to neural differentiation efficiency. (A) Schematic of the feature selection in this study. After primary filtering based on standard deviation among hiPSC lines, 60,728 CpG sites were remained. Supervised feature selection using HSIC Lasso selected 62 non-redundant feature sets and 100 neighboring features for each of those features. Finally, after eliminating duplicate CpG sites, 5393 of neural differentiation-associated differentially methylated sites (ND-DMSs) were extracted. (B) 2D representation of Uniform Manifold Approximation and Projection (UMAP)-based dimensionality reduction of the top 17 principal components obtained from PCA analysis using 62 selected features. Each hiPSC sample indicated as a dot is colored and labeled by differentiation efficiency. (C) ND-DMS enrichment within each genomic window defined by the sliding window method. Enrichment is represented by the negative logarithm of the adjusted *p*-value (q-value), which assesses the significance of ND-DMS in the probes designed for each window. (D) A screenshot from WashU Epigenome Browser showing region with strongest enrichment peak of ND-DMSs (chr5:1,700,000–5,100,000). From the top panel, the location of transcripts registered in NCBI Reference Sequence Database, the location of CpG islands, the density plot of the number of ND-DMSs per 3000 bp, and the density plot of all probes within methylationEPIC per 3000 bp are shown. Most of the ND-DMSs in this peak are located around the *IRX2*, *C5orf38* and *IRX1* genes, which are highlighted in light red. (E) DNA methylation rate of the *IRX2*, *C5orf38*, and *IRX1* genes loci. Blue boxes and lines represent exon and intron structures of the genes and green boxes represent CpG islands. Line in the panels indicates DNA methylation levels in a hiPSC line. Ticks on the horizontal axis indicate the position of the methylationEPIC probes, with ND-DMSs colored red and the others in gray.


A
chr5: 2737000-2760000
IRX2
C5orf38
CGI1
CGI2
TaqI site
cg04992127
B
chr5: 3589000-3604000
IRX1
CGI3
CGI4
HpyCH4IV site
cg04980590
C
All iPSCs (n=32)
cg04992127
R=0.34 (p=0.06)
cg04980590
R=0.42 (p=0.02)
Methylation
Efficiency (%)
Segregate according to gender
D
Female iPSCs (n=16)
Male iPSCs (n=16)
cg04992127
R=0.66 (p<0.01)
R=-0.33 (p=0.21)
cg04980590
R=0.80 (p<0.01)
R=-0.24 (p=0.37)
E
IRX2 / C5orf38
7 probes in CGI1
48 probes in CGI2
IRX1
10 probes in CGI3
28 probes in CGI4
p-value (-log10)
< 0.01
All iPSCs
Female iPSCs
Male iPSCs
F
IRX2 / C5orf38
IRX1
Efficiency
Methylation
UtE-iPS-11
IMR-iPS-18
DFM1-iPS-1
IMR-iPS-1
DFM2-iPS-5
IMR-iPS-17
IMR-iPS-4
DFMF1-iPS-3
DFMF1-iPS-11
Female
Male
G
IRX2 / C5orf38
IRX1
♀, low : Retro-UtE-iPS-11
♀, low : RNA-IMR-iPS-18
♀, low : RNA-DFM1-iPS-1
♀, low : RNA-DFM2-iPS-5
♀, high : RNA-IMR-iPS-17-1
♀, high : RNA-IMR-iPS-4
♂, low : RNA-DFMF1-iPS-3
♂, high : RNA-DFMF1-iPS-11-2

Windows with q-values (corrected *p*-values) less than 0.01 were extracted, and a series of contiguous windows were defined as a single peak, resulting in 141 peaks including 249 genes (Supplemental Table 5). Gene ontology analysis suggested that these 249 genes consisted of transcription factors associated with neurogenesis and genes associated with cell–cell adhesion and signaling (Supplemental Fig. 6).

The highest enrichment peaks were found within the p15.33 of chromosome 5 (Fig. 2C). We also examined Pearson's correlation coefficients of DNA methylation between high standard deviation CpG sites on chromosome 5 in 32 hiPSC lines and detected a strong correlation of DNA methylation between CpG sites across a large region of approximately 3 Mbp (chromosome 5:1,891,789–4,851,084), where ND-DMSs were enriched (Supplemental Fig. 7). The selected features in this region were particularly frequent in two loci; one included *IRX2* and *C5orf38* genes and the other included *IRX1* gene (Fig. 2D). Visualization of the DNA methylation levels of each hiPSC line in these regions showed that DNA methylation fluctuations were particularly pronounced within the CpG islands around the three genes (Fig. 2E). For further analyses, we focused on two loci: one was the *IRX2* locus (chr5: 2,737,000–2,760,000) including *IRX2* and *C5orf38* genes and the other was the *IRX1* locus (chr5: 3,589,000–3,604,000) including *IRX1* gene.

3.3. Differentiation efficiency-related DNA methylation variations on the IRX1/2 genes

Pearson's correlation test for each ND-DMS within the *IRX1/2* loci showed a weak correlation between the DNA methylation rate and neural differentiation efficiency. The CpG site (methylationEPIC ProbeID; cg04992127), which was located on the promoter of *IRX2* and *C5orf38* genes, showed Pearson's $r = 0.34$ and $p = 0.06$ (Fig. 3A and C). The CpG site (cg04980590), which was located 5 kb upstream of *IRX1* gene showed Pearson's $r = 0.42$ and $p = 0.02$ (Fig. 3B and C). Interestingly, when hiPSCs were separated by sex for correlation analysis, female hiPSCs showed a strong positive correlation (cg04992127, $r = 0.65$, $p < 0.01$; cg04980590, $r = 0.80$, $p < 0.01$) between DNA methylation rate and differentiation efficiency (Fig. 3D). On the other hand, male hiPSCs did not show significant correlation (cg04992127, $r = -0.33$, $p = 0.22$; cg0498 0590, $r = -0.24$, $p = 0.37$). Pearson's correlation test for each probe designed on CpG islands within the *IRX1/2* loci showed a tendency to correlate significantly only in female hiPSCs when cells were gender-segregated (Fig. 3E).

We validated the DNA methylation rate at the *IRX1/2* loci using COBRA and sodium bisulfite sequencing analysis. DNA methylation rates were confirmed to correlate with differentiation efficiency in female hiPSCs, similar to the methylationEPIC assay (Fig. 3F). Bisulfite sequencing of the region which contains CpG sites analyzed by COBRA and methylationEPIC also showed that the regions were almost completely DNA methylated in highly differentiated female hiPSCs, whereas, hypo-methylated alleles were detected in poorly differentiated female hiPSCs (Fig. 3G). In contrast, male hiPSCs were hypermethylated with both high and low neural differentiation abilities (Fig. 3F and G).

3.4. Forced expression of IRX1, IRX2 and C5orf38 genes inhibits neural stem cell differentiation

Next, we investigate whether the DNA methylation state of *IRX1*, *IRX2* and *C5orf38* genes affects their expression. In Pearson's correlation tests between the DNA methylation array and gene expression array, DNA methylation rates were negatively correlated with gene expression levels for *IRX2* ($r = -0.93$, $p < 0.01$), *C5orf38* ($r = -0.57$, $p < 0.01$), and *IRX1* ($r = -0.46$, $p = 0.01$) genes (Fig. 4A–C). Similarly, in a correlation analysis of DNA methylation, quantified by COBRA, and gene expression, calculated by RT-qPCR for nine hiPSC lines, gene expression levels were inversely correlated with DNA methylation rates for three genes (Supplemental Fig. 8). These results suggest that the expression of the three genes induced by DNA hypomethylation reduces differentiation efficiency.

To test this hypothesis, we performed experiments in which the three genes were forced to be expressed in hiPSCs. The three genes (*IRX2*, *C5orf38* or *IRX1*) and EGFP under the control of the tetracycline operator were introduced into hiPSCs (Fig. 4D). After neomycin selection and culture with doxycycline (1 μg/mL), we confirmed that more than 99% of the cells were EGFP-positive by flow cytometry (Fig. 4E and Supplemental Fig. 9). Non-transduced hiPSCs did not show reduced differentiation efficiency. On the other hand, doxycycline-inducible forced expression of the target genes *IRX2, C5orf38* or *IRX1* reduced the differentiation efficiency by 16% and 48%, 63% and 83%, and 55% and 62% in female and male hiPSCs, respectively (Fig. 4F). Interestingly, overexpression of the three genes suppressed neural differentiation ability not only in female hiPSCs but also in male hiPSCs. Some male hiPSC lines have poor differentiation abilities, even when expression of the three genes was suppressed by DNA hypermethylation. In male hiPSC lines, a downstream cascade of the three genes may be involved in the inhibition of neural differentiation. These findings indicate that the induction of *IRX1*, *IRX2* and *C5orf38* genes by aberrant DNA methylation in the corresponding region is one of the factors influencing differentiation propensity in hiPSCs.

4. Discussion

In this study, we identified biomarkers that predict the efficiency of hiPSC differentiation into neural stem cells prior to induction by comparing the DNA methylation profiles of undifferentiated hiPSC lines. We found that DNA methylation fluctuations were dominant in a broad region of chromosome 5, where *IRX1*,

Fig. 3. Analysis of DNA methylation within *IRX2*, *C5orf38* and *IRX1* loci. (A and B) Overview of *IRX2*, *C5orf38* (A) and *IRX1* (B) loci. Exon and intron structures of the genes are shown as blue boxes and lines. CpG islands (CGI) are shown as green boxes and are numbered in correspondence to (E). Vertical lines indicate the position of the designed probes in methylationEPIC, with ND-DMSs colored red and the others in gray. Circles represent position of CpG sites analyzed by bisulfite sequencing analysis shown in (G). Yellow circles indicate CpG sites analyzed by COBRA shown in (F). Red circles indicate CpG sites shown in (C) and (D). (C) Scatter plot of DNA methylation levels in the CpG sites (cg04992127 and cg04980590) associated with *IRX2*, *C5orf38* and *IRX1* genes and neural differentiation efficiency for hiPSC lines ($n = 32$). Dashed line indicates linear regression line. The results of the Pearson's correlation test are shown in the panels. (D) Scatter plot of DNA methylation levels in the CpG sites (cg04992127 and cg04980590) and neural differentiation efficiency for gender-segregated hiPSC lines ($n = 16$ for females and $n = 16$ for males). Dashed line indicates linear regression line. The results of the Pearson's correlation test are shown in the panels. (E) Comparisons of the significance of Pearson's correlation between DNA methylation and neural differentiation efficiency for each ND-DMS for different groupings by gender. (F) DNA methylation levels measured by COBRA for 7 female hiPSCs and 2 male hiPSCs. (left) Average DNA methylation rates of the two very close neighboring CpG sites that are cleaved by TaqI are shown in (A). (right) DNA methylation rates of the CpG sites that are cleaved by HpyCH4IV are shown in (B). The data shown represent the mean ± standard error ($n = 3$). (G) Bisulfite sequencing analysis of the genomic regions shown in (A) and (B). Open and closed circles indicate unmethylated and methylated sites, respectively. Each hiPSC line used in this analysis is listed above the plot in the following order; gender (♀: female, ♂: male), differentiation efficiency, and line name.

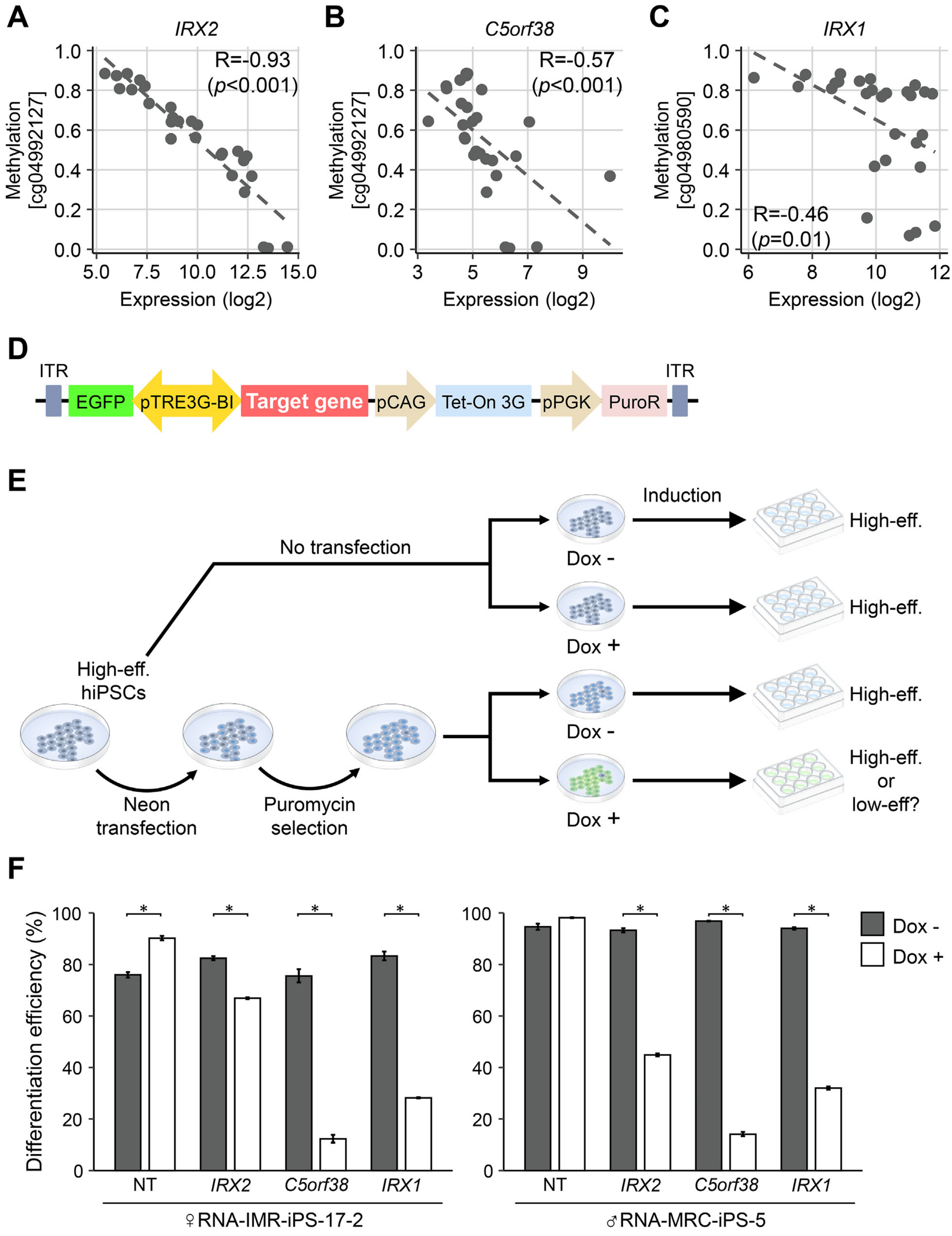


Fig. 4. Relationship between DNA methylation and gene expression and forced gene expression experiments (A–C) Scatter plot of gene expression and DNA methylation levels in 28 hiPSC lines. Dashed line indicates linear regression line. The results of the Pearson's correlation test are shown in the plots. (D) Schematic of the PiggyBac transposon vector. PiggyBac inverted terminal repeats (ITR) flank three units: a bi-directional TRE3G promoter unit for doxycycline (Dox)-inducible expression of EGFP and target gene (*IRX2*, *C5orf38* or *IRX1*), a Tet-On 3G gene unit to regulate response to Dox, and a puromycin resistance unit for selection. (E) Experimental scheme for forced expression studies. The constructed PiggyBac vectors as shown (D) were introduced into hiPSCs by the Neon transfection system. The transfected hiPSCs were then purified by puromycin. The cells were inducted into neural stem cells with or without Dox. (F) Comparisons of neural differentiation efficiency under forced expression of *IRX2*, *C5orf38* or *IRX1* genes. The data shown represent the mean ± standard error (n = 3). $^*p < 0.001$.

IRX2 and *C5orf38* were located, and their DNA methylation rates were strongly correlated with the neural differentiation efficiency in female hiPSCs. Human iPS cell-derived neural stem cells have tremendous potential as a tool for *in vitro* modeling of nervous system development and disease, and for the treatment of neurodegenerative diseases and nervous system injuries. In addition, the selection of a suitable hiPSCs is important for reproducibility and safety. Our findings suggest that DNA methylation levels in the *IRX1/2* loci can predict the ability of undifferentiated hiPSCs to differentiate into neural stem cells. Furthermore, the reproducibility of the evaluation of DNA methylation levels in this region by a simple quantitative method, such as COBRA, also supports its usefulness as a biomarker. In the present study, we used the monolayer dual-SMAD inhibition protocol to assess the ability of each hiPSC line because it is the basic and by far the most popular method for obtaining neural stem cells from PSCs [44]. However, many methods of inducing neural differentiation are known. The other induction methods that can improve low differentiation efficiency may correct abnormal hypomethylation of *IRX1/2* gene region.

The Iroquois homeobox (*IRX*) gene family contains homeobox domains and plays multiple roles during patterning processes in vertebrate embryos [45,46]. *IRX1* is known as a tumor suppressor in humans rather than a regulator of neurogenesis [47–49]. The function of *C5orf38* remains unclear. However, forced expression of *IRX1* and *C5orf38* in undifferentiated hiPSCs strongly inhibited neural differentiation in this study. This is the first report to show that these genes affect the differentiation of hiPSCs.

IRX2 was shown to be upregulated in the later stages of neural differentiation of human embryonic stem cells (hESCs) but not in early stages of commitment [50]. In addition, knockdown experiments using hESCs have shown that suppression of *IRX2* impairs differentiation into neural stem cells [51]. In the present study, we showed that forced expression of *IRX2* in undifferentiated hiPSCs impaired their ability to differentiate into neural stem cells. This evidence suggests that the *IRX2* gene is required for neural stem cell maturation, but its expression in the early stages suppresses neural differentiation.

Forced expression of *IRX1*, *IRX2* and *C5orf38* genes in female hiPSCs that showed high differentiation led to decreased neural differentiation efficiency. These results indicated that DNA methylation in the *IRX1/2* loci is a major regulator, and aberrant expression of *IRX1*, *IRX2* and *C5orf38* due to DNA methylation variation in this region was found to be responsible for neural differentiation propensity in female hiPSCs. On the other hand, some male hiPSCs showed low differentiation, even though *IRX1*, *IRX2* and *C5orf38* genes were repressed by DNA methylation. We did not find a prominent biomarker, such as the *IRX1/2* loci, in the dataset of the 16 male hiPSC lines using HSIC Lasso. However, forced expression of *IRX1*, *IRX2* and *C5orf38* genes in male hiPSCs, in which the three genes were repressed by DNA methylation and that showed high differentiation, led to decreased neural differentiation efficiency. These facts suggested that there may be random perturbations downstream of the networks regulated by the three genes. Another possibility is that there may be a fundamental distinction in the regulation of neurogenesis between males and females.

We identified other ND-DMSs besides the *IRX1/2* loci by HSIC Laaso. These CpG sites were enriched in genes involved in neurogenesis and cell communication, which may constitute a network that underlies neural differentiation propensity. In future studies, we will explore in detail the epigenetic and transcriptional networks associated with the inhibition of neural differentiation to determine the cause of neural differentiation propensity for both sex. It also remains to be explored whether the ND-DMSs can affect the subsequent induction of differentiation into neurons and glial cells.

In conclusion, we identified the *IRX1/2* loci as a DNA methylation biomarker that predicts the differentiation of human iPSCs from an undifferentiated state to neural stem cells, providing an epigenetic basis for understanding neural differentiation propensity.

Author contributions

AS and KN conceived and designed the experiments; AS, KT, YA, SH and KN performed the experiments; AS and KN analyzed the data; AS, KT, YA, AH, AU and KN contributed reagents/materials/analysis tool; and AS and KN wrote the manuscript.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Acknowledgments

We would like thank Ms. Haruka Tomaru, Ms. Yukie Toyonaga, Ms. Yuka Todaka, Ms. Junko Sasaki, Ms. Wakana Goto, and Ms. Kae Inoue for performing cell culture. This research was supported in part by AMED under Grant number JP18bm0704003 and JP21bk0104092 to KN; JSPS KAKENHI (B) under Grant number 21H02381 to KN.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.reth.2022.11.007.

References

[1] Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 2007;131:861–72. https://doi.org/10.1016/j.cell.2007.11.019.

[2] Kitagawa T, Nagoshi N, Kamata Y, Kawai M, Ago K, Kajikawa K, et al. Modulation by DREADD reveals the therapeutic effect of human iPSC-derived neuronal activity on functional recovery after spinal cord injury. Stem Cell Rep 2022;17:127–42. https://doi.org/10.1016/j.stemcr.2021.12.005.

[3] Linker SB, Mendes APD, Marchetto MC. IGF-1 treatment causes unique transcriptional response in neurons from individuals with idiopathic autism. Mol Autism 2020;11:55. https://doi.org/10.1186/s13229-020-00359-w.

[4] Bock C, Kiskinis E, Verstappen G, Gu H, Boulting G, Smith ZD, et al. Reference maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. Cell 2011;144:439–52. https://doi.org/10.1016/j.cell.2010.12.032.

[5] Nasu A, Ikeya M, Yamamoto T, Watanabe A, Jin Y, Matsumoto Y, et al. Genetically matched human iPS cells reveal that propensity for cartilage and bone differentiation differs with clones, not cell type of origin. PLoS One 2013;8:e53771. https://doi.org/10.1371/journal.pone.0053771.

[6] Osafune K, Caron L, Borowiak M, Martinez RJ, Fitz-Gerald CS, Sato Y, et al. Marked differences in differentiation propensity among human embryonic stem cell lines. Nat Biotechnol 2008;26:313–5. https://doi.org/10.1038/nbt1383.

[7] Kim K, Zhao R, Doi A, Ng K, Unternaehrer J, Cahan P, et al. Donor cell type can influence the epigenome and differentiation potential of human induced pluripotent stem cells. Nat Biotechnol 2011;29:1117–9. https://doi.org/10.1038/nbt.2052.

[8] Panopoulos AD, Smith EN, Arias AD, Shepard PJ, Hishida Y, Modesto V, et al. Aberrant DNA methylation in human iPSCs associates with MYC-binding motifs in a clone-specific manner independent of genetics. Cell Stem Cell 2017;20:505–17. https://doi.org/10.1016/j.stem.2017.03.010. e6.

[9] Yagi M, Kabata M, Ukai T, Ohta S, Tanaka A, Shimada Y, et al. De novo DNA methylation at imprinted loci during reprogramming into naive and primed pluripotency. Stem Cell Rep 2019;12:1113–28. https://doi.org/10.1016/j.stemcr.2019.04.008.

[10] Wutz A. Epigenetic alterations in human pluripotent stem cells: a tale of two cultures. Cell Stem Cell 2012;11:9–15. https://doi.org/10.1016/j.stem.2012.06.012.

[11] Nishino K, Toyoda M, Yamazaki-Inoue M, Fukawatase Y, Chikazawa E, Sakaguchi H, et al. DNA methylation dynamics in human induced pluripotent

stem cells over time. PLoS Genet 2011;7:e1002085. https://doi.org/10.1371/journal.pgen.1002085.

[12] Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. Nat Rev Mol Cell Biol 2019;20: 590–607. https://doi.org/10.1038/s41580-019-0159-6.

[13] Kim H, Lee G, Ganat Y, Papapetrou EP, Lipchina I, Socci ND, et al. miR-371-3 expression predicts neural differentiation propensity in human pluripotent stem cells. Cell Stem Cell 2011;8:695–706. https://doi.org/10.1016/j.stem.2011.04.002.

[14] Kuroda T, Yasuda S, Tachi S, Matsuyama S, Kusakawa S, Tano K, et al. SALL3 expression balance underlies lineage biases in human induced pluripotent stem cell differentiation. Nat Commun 2019;10:2175. https://doi.org/10.1038/s41467-019-09511-4.

[15] Mo C-F, Wu F-C, Tai K-Y, Chang W-C, Chang K-W, Kuo H-C, et al. Loss of non-coding RNA expression from the DLK1-DIO3 imprinted locus correlates with reduced neural differentiation potential in human embryonic stem cell lines. Stem Cell Res Ther 2015;6:1. https://doi.org/10.1186/scrt535.

[16] Nishizawa M, Chonabayashi K, Nomura M, Tanaka A, Nakamura M, Inagaki A, et al. Epigenetic variation between human induced pluripotent stem cell lines is an indicator of differentiation capacity. Cell Stem Cell 2016;19:341–54. https://doi.org/10.1016/j.stem.2016.06.019.

[17] Ohashi F, Miyagawa S, Yasuda S, Miura T, Kuroda T, Itoh M, et al. CXCL4/PF4 is a predictive biomarker of cardiac differentiation potential of human induced pluripotent stem cells. Sci Rep 2019;9:4638. https://doi.org/10.1038/s41598-019-40915-w.

[18] Zhu L, Gomez-Duran A, Saretzki G, Jin S, Tilgner K, Melguizo-Sanchis D, et al. The mitochondrial protein CHCHD2 primes the differentiation potential of human induced pluripotent stem cells to neuroectodermal lineages. JCB (J Cell Biol) 2016;215:187–202. https://doi.org/10.1083/jcb.201601061.

[19] Chang C-H, Lin C-H, Lane H-Y. Machine learning and novel biomarkers for the diagnosis of alzheimer's disease. Int J Math Stat 2021;22:2761. https://doi.org/10.3390/ijms22052761.

[20] Glaab E, Rauschenberger A, Banzi R, Gerardi C, Garcia P, Demotes J. Biomarker discovery studies for patient stratification using machine learning analysis of omics data: a scoping review. BMJ Open 2021;11:e053674. https://doi.org/10.1136/bmjopen-2021-053674.

[21] Marcos-Zambrano LJ, Karaduzovic-Hadziabdic K, Loncar Turukalo T, Przymus P, Trajkovik V, Aasmets O, et al. Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. Front Microbiol 2021;12:634511. https://doi.org/10.3389/fmicb.2021.634511.

[22] Climente-González H, Azencott C-A, Kaski S, Yamada M. Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data. Bioinformatics 2019;35:i427–35. https://doi.org/10.1093/bioinformatics/btz333.

[23] Yamada M, Jitkrittum W, Sigal L, Xing EP, Sugiyama M. High-dimensional feature selection by feature-wise kernelized Lasso. Neural Comput 2014;26: 185–207. https://doi.org/10.1162/NECO_a_00537.

[24] Takahashi Y, Ueki M, Yamada M, Tamiya G, Motoike IN, Saigusa D, et al. Improved metabolomic data-based prediction of depressive symptoms using nonlinear machine learning with feature selection. Transl Psychiatry 2020;10: 157. https://doi.org/10.1038/s41398-020-0831-9.

[25] Cui C-H, Uyama T, Miyado K, Terai M, Kyo S, Kiyono T, et al. Menstrual blood-derived cells confer human dystrophin expression in the murine model of duchenne muscular dystrophy via cell fusion and myogenic Transdifferentiation. Mol Biol Cell 2007;18:9. https://doi.org/10.1091/mbc.e06-09-0872.

[26] Jacobs JP, Jones CM, Baille JP. Characteristics of a human diploid cell designated MRC-5. Nature 1970;227:168–70. https://doi.org/10.1038/227168a0.

[27] Nichols WW, Murphy DG, Cristofalo VJ, Toji LH, Greene AE, Dwight SA. Characterization of a new human diploid cell strain, IMR-90. Science 1977;196:60–3. https://doi.org/10.1126/science.841339.

[28] Nishino K, Toyoda M, Yamazaki-Inoue M, Makino H, Fukawatase Y, Chikazawa E, et al. Defining hypo-methylated regions of stem cell-specific promoters in human iPS cells derived from extra-embryonic amnions and lung fibroblasts. PLoS One 2010;5:e13017. https://doi.org/10.1371/journal.pone.0013017.

[29] Nishino K, Arai Y, Takasawa K, Toyoda M, Yamazaki-Inoue M, Sugawara T, et al. Epigenetic-scale comparison of human iPSCs generated by retrovirus, Sendai virus or episomal vectors. Regen Ther 2018;9:71–8. https://doi.org/10.1016/j.reth.2018.08.002.

[30] Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature 2015;526:68–74. https://doi.org/10.1038/nature15393.

[31] Tadaka S, Katsuoka F, Ueki M, Kojima K, Makino S, Saito S, et al. 3.5KJPNv2: an allele frequency panel of 3552 Japanese individuals including the X chromosome. Hum Gen Variation 2019;6:28. https://doi.org/10.1038/s41439-019-0059-5.

[32] Dorrity MW, Saunders LM, Queitsch C, Fields S, Trapnell C. Dimensionality reduction by UMAP to visualize physical and genetic interactions. Nat Commun 2020;11:1537. https://doi.org/10.1038/s41467-020-15351-4.

[33] Takasawa K, Arai Y, Yamazaki-Inoue M, Toyoda M, Akutsu H, Umezawa A, et al. DNA hypermethylation enhanced telomerase reverse transcriptase expression in human-induced pluripotent stem cells. Hum Cell 2018;31: 78–86. https://doi.org/10.1007/s13577-017-0190-x.

[34] Lenz M, Goetzke R, Schenk A, Schubert C, Veeck J, Hemeda H, et al. Epigenetic biomarker to support classification into pluripotent and non-pluripotent cells. Sci Rep 2015;5:8973. https://doi.org/10.1038/srep08973.

[35] Chambers SM, Fasano CA, Papapetrou EP, Tomishima M, Sadelain M, Studer L. Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. Nat Biotechnol 2009;27:275–80. https://doi.org/10.1038/nbt.1529.

[36] Peng Hanchuan, Long Fuhui, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27:1226–38. https://doi.org/10.1109/TPAMI.2005.159.

[37] Gretton A, Bousquet O, Smola A, Schölkopf B. Algorithm Learn Theory. In: Jain S, Simon HU, Tomita E, editors. Measuring statistical dependence with Hilbert-Schmidt norms. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005. p. 63–77. https://doi.org/10.1007/11564089_7.

[38] Dabney A, Storey JD. qvalue: Q-value estimation for false discovery rate control. 2021. R package version 2.24.0, http://github.com/jdstorey/qvalue.

[39] Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res 2022;50:W216–21. https://doi.org/10.1093/nar/gkac194.

[40] Xiong Z, Laird PW. COBRA: a sensitive and quantitative DNA methylation assay. Nucleic Acids Res 1997;25:2532–4. https://doi.org/10.1093/nar/25.12.2532.

[41] Kumaki Y, Oda M, Okano M. QUMA: quantification tool for methylation analysis. Nucleic Acids Res 2008;36:W170–5. https://doi.org/10.1093/nar/gkn294.

[42] Pfaffl MW. A new mathematical model for relative quantification in real-time RT–PCR. Nucleic Acids Res 2001;29:e45. https://doi.org/10.1093/nar/29.9.e45.

[43] Yusa K, Zhou L, Li MA, Bradley A, Craig NL. A hyperactive *piggyBac* transposase for mammalian applications. Proc Natl Acad Sci USA 2011;108:1531–6. https://doi.org/10.1073/pnas.1008322108.

[44] Galiakberova AA, Dashinimaev EB. Neural stem cells and methods for their generation from induced pluripotent stem cells in vitro. Front Cell Dev Biol 2020;8:815. https://doi.org/10.3389/fcell.2020.00815.

[45] Bosse A, Zülch A, Becker M-B, Torres M, Gómez-Skarmeta JL, Modolell J, et al. Identification of the vertebrate Iroquois homeobox gene family with overlapping expression during early development of the nervous system. Mech Dev 1997;69:169–81. https://doi.org/10.1016/S0925-4773(97)00165-2.

[46] Bürglin TR. Analysis of TALE superclass homeobox genes (MEIS, PBC, KNOX, Iroquois, TGIF) reveals a novel domain conserved between plants and animals. Nucleic Acids Res 1997;25:4173–80. https://doi.org/10.1093/nar/25.21.4173.

[47] Bennett KL, Karpenko M, Lin M, Claus R, Arab K, Dyckhoff G, et al. Frequently methylated tumor suppressor genes in Head and neck squamous cell carcinoma. Cancer Res 2008;68:4494–9. https://doi.org/10.1158/0008-5472.CAN-07-6509.

[48] Guo X, Liu W, Pan Y, Ni P, Ji J, Guo L, et al. Homeobox gene IRX1 is a tumor suppressor gene in gastric carcinoma. Oncogene 2010;29:3908–20. https://doi.org/10.1038/onc.2010.143.

[49] Jiang J, Liu W, Guo X, Zhang R, Zhi Q, Ji J, et al. IRX1 influences peritoneal spreading and metastasis via inhibiting BDKRB2-dependent neovascularization on gastric cancer. Oncogene 2011;30:4498–508. https://doi.org/10.1038/onc.2011.154.

[50] Kreimer A, Ashuach T, Inoue F, Khodaverdian A, Deng C, Yosef N, et al. Massively parallel reporter perturbation assays uncover temporal regulatory architecture during neural differentiation. Nat Commun 2022;13:1504. https://doi.org/10.1038/s41467-022-28659-0.

[51] Zhang Y, Schulz VP, Reed BD, Wang Z, Pan X, Mariani J, et al. Functional genomic screen of human stem cell differentiation reveals pathways involved in neurodevelopment and neurodegeneration. Proc Natl Acad Sci USA 2013;110:12361–6. https://doi.org/10.1073/pnas.1309725110.