



# Identification of an epigenetic signature in human induced pluripotent stem cells using a linear machine learning model

Koichiro Nishino<sup>1,2</sup> · Ken Takasawa<sup>1</sup> · Kohji Okamura<sup>3</sup> · Yoshikazu Arai<sup>1</sup> · Asato Sekiya<sup>1</sup> · Hidenori Akutsu<sup>4</sup> · Akihiro Umezawa<sup>4</sup>

Received: 24 June 2020 / Accepted: 2 October 2020 / Published online: 12 October 2020  
© The Author(s) 2020

## Abstract

The use of human induced pluripotent stem cells (iPSCs), used as an alternative to human embryonic stem cells (ESCs), is a potential solution to challenges, such as immune rejection, and does not involve the ethical issues concerning the use of ESCs in regenerative medicine, thereby enabling developments in biological research. However, comparative analyses from previous studies have not indicated any specific feature that distinguishes iPSCs from ESCs. Therefore, in this study, we established a linear classification-based learning model to distinguish among ESCs, iPSCs, embryonal carcinoma cells (ECCs), and somatic cells on the basis of their DNA methylation profiles. The highest accuracy achieved by the learned models in identifying the cell type was 94.23%. In addition, the epigenetic signature of iPSCs, which is distinct from that of ESCs, was identified by component analysis of the learned models. The iPSC-specific regions with methylation fluctuations were abundant on chromosomes 7, 8, 12, and 22. The method developed in this study can be utilized with comprehensive data and widely applied to many aspects of molecular biology research.

**Keywords** Machine learning · Human iPSCs · Human ESCs · DNA methylation · Epigenetic signature of hiPSCs

## Introduction

The application of human induced pluripotent stem cells (iPSCs) in medicine requires prior assessment of the cells with respect to quality, including identity, equivalence, and safety. For evaluation of the iPSCs, comprehensive

molecular analysis of characteristics, such as DNA methylation, rather than tests based on a few marker genes, is considered to be more useful. DNA methylation is an epigenetic modification with important roles in normal development and differentiation [1–6]. DNA methylation profiles vary depending on tissue types and cell lineage [5, 7]; therefore, the DNA methylation profile of a cell can be useful for the identification and validation of its cell type. Epigenetic reprogramming, which involves conversion of the DNA methylation profile from somatic to pluripotent cell type, is an essential for the transformation of somatic cells into iPSCs; the cells that acquire the DNA methylation profile of embryonic stem cells (ESCs) become iPSCs [8, 9].

Human iPSCs lower the rate of immune rejection and help in resolving ethical issues associated with the use of ESCs in regenerative medicine [10]. Since the successful development of iPSCs [11–13], comparative analyses between iPSCs and ESCs have been performed by many researchers. Choi et al. [14] reported that there are no molecular or functional differences between genetically matched human ESCs and iPSCs. On the other hand, several studies have identified differentially methylated DNA regions between human iPSCs and ESCs [8, 15–17].

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s13577-020-00446-3>) contains supplementary material, which is available to authorized users.

✉ Koichiro Nishino  
aknishino@cc.miyazaki-u.ac.jp

- <sup>1</sup> Laboratory of Veterinary Biochemistry and Molecular Biology, Graduate School of Medicine and Veterinary Medicine/Faculty of Agriculture, University of Miyazaki, Miyazaki, Japan
- <sup>2</sup> Center for Animal Disease Control, University of Miyazaki, Miyazaki, Japan
- <sup>3</sup> Department of Systems BioMedicine, National Research Institute for Child Health and Development, Tokyo, Japan
- <sup>4</sup> Department of Reproductive Biology, Center for Regenerative Medicine, National Research Institute for Child Health and Development, Tokyo, Japan

However, these studies only analyzed single point of passage of human iPSCs. In a previous study, we comparatively analyzed several points of passages of 22 human iPSC lines and the results indicated the presence of aberrant hypermethylated sites in iPSCs; however, aberrant hypermethylation in iPSCs occurs stochastically throughout the genome and there is no iPSC-specific aberrant methylated site common to all iPSCs [9]. Despite the lack of DNA methylation hotspots in iPSCs, previous studies have suggested that there are fundamental differences between ESCs and iPSCs, raising questions regarding the extent of similarity between ESC-type epigenome and the reconstructed whole genome of iPSCs. For comparative analysis of cell types with no clear differences, machine learning technology may be useful.

Machine learning is a data analysis technique that attempts to train computers to learn through experience with datasets, in manner similar to natural learning in human. Supervised machine learning can be used to build models for evidence-based prediction, even when there is uncertainty. A supervised learning algorithm trains a machine learning model on a set of input data and the resultant responses (outputs), so that it can reasonably predict the response to new data. In supervised machine learning, classification or regression methods are used to construct predictive models. Classification models are trained to classify the data into categories. Regression models are used to estimate one variable based on the data.

If a model capable of discriminating between ESCs and iPSCs can be constructed using supervised machine learning, the difference between the two cell types could be elucidated. Such a model could help identify the factors underlying the differences between ESCs and iPSCs, as well as enable visualization of these differences, which cannot be distinguished by the naked human eyes.

In this study, we used classification method-based machine learning to create a model that can discriminate between iPSCs and ESCs on the basis of DNA methylation profiles. Further, we attempted to determine the difference between iPSCs and ESCs by analyzing the components of the learning model. Our machine learning-based analysis method and the identified epigenetic indices are useful for evaluating the therapeutic application of human iPSCs. We propose a new method for molecular analysis of the cells that combines comprehensive DNA methylation data and machine learning.

## Materials and methods

### 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 (FBS) (Thermo Fisher Scientific, Inc., Waltham, MA, USA, Cat. No. SH3091003), 55  $\mu$ M 2-mercaptoethanol (Thermo Fisher Scientific), 1% penicillin and streptomycin (Thermo Fisher Scientific). MEFs were irradiated with 30 Gy of gamma irradiation to generate MEF feeder cells. All procedures were performed in accordance with the guidelines for animal care and use of laboratory animals, University of Miyazaki, and the experimental protocols were approved by the Animal Experiment Committee of University of Miyazaki (no. 2012-017, 2017-009).

### Human cell culture

Human endometrium (UtE1104), amnion (AM936EP), placental artery endothelium (PAE551) and menstrual blood (Edom22) cell lines were independently established [18, 19]. Fetal lung fibroblast cells (MRC-5) [20] were obtained from JCRB Cell Bank, Japan. UtE1104, AM936EP, MRC-5, and Edom22 were maintained in POWEREDBY10 medium (Glyco Technica Ltd., Sapporo, Japan). PAE551 were cultured in EGM-2MV BulletKit medium (Lonza, Walkersville, MD, USA) containing 5% FBS (Thermo Fisher Scientific). Human Retro-iPSCs were generated using the retroviral vector pMXs, which contains the cDNAs for human *OCT3/4*, *SOX2*, *c-MYC*, and *KLF4* [8, 9, 19, 21, 22], according to previously described procedures [12] with slight modifications. Episomal-iPSCs were established using the episomal vectors, pCXLE-hOCT3/4-shp53, pCXLE-hSK, and pCXLE-hUL [23], according to previously described procedures [24]. Sendai-iPSCs were produced using the Sendai viral vector SeVdp-iPS, which contains the polycistronic cDNAs for mouse *Oct3/4*, *Sox2*, *c-Myc*, and *Klf4* [23], according to previously described procedures [25]. The SEES lines of human ESCs were generated in the Center for Regenerative Medicine, National Research Institute for Child Health and Development, Tokyo, Japan [26]. Genomic DNA of the HUES lines of human ESCs [27, 28], was kindly gifted by Drs. C. Cowan and T. Tenzan (Harvard Stem Cell Institute, Harvard University, Cambridge, MA, USA). 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), 55  $\mu$ M 2-mercaptoethanol (Thermo Fisher Scientific), 1% penicillin and streptomycin (Thermo Fisher Scientific), and 10 ng/ml recombinant human basic fibroblast growth factor (bFGF) (Wako Pure Chemical Industries, Ltd., Osaka, Japan). The human embryonal carcinoma cell lines NCR-G2, NCR-G3 and NCR-G4, which were established in the National Research Institute for Child Health and Development, Tokyo, Japan [29], were cultured in G031101 medium [21]. The human embryonal carcinoma cell lines NCC-IT-A3 [30], PA-1 [31], NEC8, and NEC14 [32] were obtained from JCRB Cell Bank, Japan. NCC-IT-A3, NEC8 and NEC14 were cultured in RPMI1640 medium (Sigma-Aldrich) supplemented with 10% FBS (Thermo Fisher Scientific), and PA-1 was cultured in MEM supplemented with nonessential amino acids and 10% FBS (Thermo Fisher Scientific). The human ECC lines 1777N Rpmet [33] and NTERA-2 [34] were obtained from DS Pharma Biological Co. LTD, Japan, and were cultured in DMEM supplemented with 10% FBS (Thermo Fisher Scientific). All human cell lines used in this study are summarized in Supplemental Table 1.

### DNA methylation analysis

DNA methylation profiles were obtained from each sample using the Illumina Infinium assay with the Infinium HumanMethylation450K BeadChip and Infinium MethylationEPIC BeadChip (Illumina Inc., San Diego, CA, USA). Genomic DNA was extracted from the cells using the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany). From each sample, 1  $\mu$ g of genomic DNA was subjected to bisulfite conversion using the EZ DNA Methylation kit (Zymo Research, Orange, CA, USA), according to the manufacturer's recommendations. Following bisulfite conversion, the genomic DNA was hybridized with the Infinium HumanMethylation450K BeadChip and 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 normalization of data. Methylated and unmethylated signals were used to compute the  $\beta$  value, a quantitative score of the DNA methylation rate that ranges from "0.00", for completely unmethylated state to "1.00", for completely methylated state. Additional DNA methylation data were obtained from the NCBI database. Detailed information of cell lines and accession numbers used in this study is mentioned in Supplemental Table 1. Common probes between 450K and EPIC were selected. The probes with sequences that overlapped

with variants showing minor allele frequency (MAF)  $\geq 5\%$  [35] and detection  $p$  value  $\geq 0.05$  (computed from the background based on negative controls) were eliminated from further analysis. A total of 385,683 CpG sites were analyzed in 104 samples including 27 ESC lines, 43 iPSC lines, 9 ECC lines, and 25 somatic cell lines. Unsupervised hierarchical clustering (HCA) with Euclidian distance and group average method and principal component analysis (PCA) were used for data analysis. A differentially methylated region (DMR) was characterized by a CpG site having a score that differed by  $\geq 0.3$  points with respect to the  $\beta$  values between two groups. For comparing the average number of DMRs between ESCs and iPSCs, 15 samples were randomly selected from 27 ESC lines and 47 iPSC lines, and the number of DMRs was counted. This step was repeated 100 times and the average number of DMRs was calculated. For comparing the average number of CpG sites within a certain range of standard deviation (SD) between ESCs and iPSCs, 15 samples were randomly selected from 27 ESC lines and 47 iPSC lines, and the number of CpG sites within a certain range of SD was counted. This step was repeated 100 times and the average number of CpG sites within a certain range of SD was calculated.

### Machine learning

Jubatus, a machine learning analytical platform, is an online open-source software (<https://jubat.us/en/>) developed by Preferred Infrastructure, Inc. (Tokyo, Japan) and NTT SIC (Tokyo, Japan). Multi-class classification (one-vs-others) of the cell types was performed using the classification module Jubaclassifier with Adaptive Regularization of Weight vectors (AROW) [36], which is a linear classification model supported by Jubatus. To perform 4-fold cross-validation, each cell line was divided into four groups, A–D (Supplemental Table 1), and the following four learning series were used: Series-1 comprising training dataset, BCD and test dataset, A; Series-2 comprising training dataset, CDA and test dataset, B; Series-3 comprising training dataset, DAB and test dataset, C; and Series-4 comprising training dataset, ABC and test dataset, D. The training datasets were used for learning in constructing learning models, and test datasets as unknowns were used for validation of the learned models. The construction of learning models was entailed by the random selection of one sample from the training dataset, followed by the input of the DNA methylation rates of 385,683 CpG sites and the cell type of the selected sample into Jubatus followed by learning, thereby updating the learning model. This process was repeated for all the samples in the training dataset, and learning once with all the samples in the training dataset was designated as 1 epoch. In total, 300 epochs were performed and the learned model was assessed every 10 epochs. The adaptive regularization parameter

was evaluated using variable regularization weight values of “0.10”, “0.25”, “0.50”, “0.90”, “1.00”, and “1.10”. The learned model was delineated using four classification models corresponding to the cell types (ESCs, iPSCs, ECCs and somatic cells). The source code is available on GitHub ([https://github.com/aknishino/20191212\\_Jub](https://github.com/aknishino/20191212_Jub)). For evaluating the learned models, Precision, Recall and F-score, Macro-average Precision ( $\text{Precision}_{\text{Macro}}$ ), Macro-average Recall ( $\text{Recall}_{\text{Macro}}$ ) and Macro-average F-score ( $\text{F-score}_{\text{Macro}}$ ) of the each learned model were calculated using the formulae shown in Table 1.

## Sodium bisulfite sequencing

Sodium bisulfite treatment of genomic DNA was carried out using the EZ DNA Methylation-Gold kit (Zymo Research). PCR amplification was performed using BIOTAQ™ HS DNA polymerase (Bioline Ltd, London, UK) with specific primers for *CSMD1*, *FZD10*, *DNAH9*, *FAM19A5*, *TMEM132C*, and *TMEM132D*. The primers used in this study are summarized in Supplemental Table 2. To determine the methylation states of individual CpG sites, the PCR product was gel-extracted and subcloned into Eco RV cut-pBluescriptII vector using NEBuilder HiFi DNA Assembly Master Mix (New England BioLabs, Ipswich, MA, USA), and then sequenced. Methylation sites were visualized and quality control was carried out using the web-based tool QUMA (<https://quma.cdb.riken.jp/>) [37].

## Accession numbers

NCBI GEO: Infinium HumanMethylation450K BeadChip and Infinium MethylationEPIC BeadChip data obtained in this study have been submitted under the accession number GSE141521. Additional DNA methylation data were obtained from the NCBI database. Accession numbers are given in Supplemental Table 1.

## Results

### Comparison of DNA methylation between ESCs and iPSCs

The DNA methylation profiles of 104 human samples, including 27 ESC lines, 43 iPSC lines, 9 ECC lines, and 25 somatic cell lines, were obtained using the Illumina Infinium HumanMethylation array. The methylation rates of 385,683 CpG sites were further analyzed (see details in “Materials and methods”). The promoter regions of the pluripotency-associated genes *POU5F1*, *NANOG*, *SALL4*, *RAB25*, and *EPHA1* showed low levels of methylation, whereas those of

the somatic cell-associated genes *GBP3*, *LYST*, and *SP100* were highly methylated in ESC and iPSC lines (Supplemental Fig. 1a). Unsupervised hierarchical cluster analysis (HCA) (Fig. 1a and Supplemental Fig. 1b) and principal component analysis (PCA) (Supplemental Fig. 1c) revealed that iPSCs were clearly distinct from somatic cells and ECCs, but not from ESCs. Comparison between the two types of cells showed that there was no differentially methylated region (DMR) between ESCs and iPSCs (Fig. 1b). These results indicate that there was no clear difference between ESCs and iPSCs.

### Construction of a machine learning model for the classification of cell types

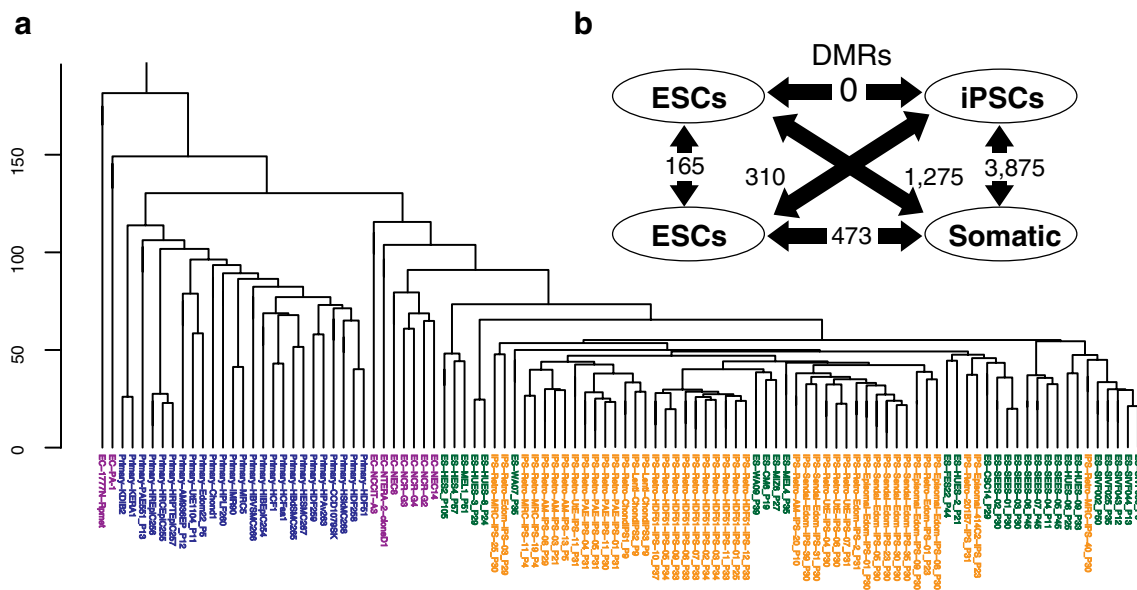
The DNA methylation data of 385,683 CpG sites and information on the cell type of the training samples were used for machine learning (Fig. 2a). In this study, machine learning involved 4-fold cross-validation method, wherein each cell line was divided into four groups to create four datasets (training dataset and test dataset) (Supplemental Table 1) and six different regularization weight values were validated. With each training dataset and regularization weight, 300 epochs were performed; thus, the total number of epochs performed was 7,200 (4 data sets  $\times$  6 regularization weight values  $\times$  300 epochs). After every 10 epochs, learning results were saved and thus, 720 learning results were obtained as learning models from each dataset. Each of the 720 learning models was used to discriminate the training dataset and the unknowns (test data set) (Fig. 2b), and comparative analyses of the average of the  $\text{F-score}_{\text{Macro}}$  rate were performed (Supplemental Fig. 2b). The learning models from the 250th epoch with AROW regularization weight value of “1.00” had the highest average of the  $\text{F-score}_{\text{Macro}}$  rate from the four models for both training dataset and test dataset, and were therefore selected as the optimal learning models. The highest average of the  $\text{F-score}_{\text{Macro}}$  rate of the test data set, which was achieved by the optimal learning model, was 94.36% (Supplemental Fig. 2a, b). The accuracy,  $\text{Precision}_{\text{Macro}}$ ,  $\text{Recall}_{\text{Macro}}$ , and  $\text{F-score}_{\text{Macro}}$  rates of the test data set in the mixed four models were 94.23%, 95.17%, 93.63% and 94.39%, respectively (Table 1). The accuracy,  $\text{Precision}_{\text{Macro}}$ , and  $\text{Recall}_{\text{Macro}}$  rates of the test data set in each four models were shown in Supplemental Fig. 3. The learning model distinguished ESCs from iPSCs with an accuracy of  $\geq 81.82\%$  (Supplemental Fig. 3). These results indicated that the learning model generated in the 250th epoch, with a regularization weight value of 1.00, is able to distinguish iPSCs from ESCs with a high efficiency.



**Table 1** Prediction accuracy, precision, recall and F-score in test samples

Class predicted	ESC	iPSC	ECC	Somatic cells	Precision	Recall
ESCs ( <i>n</i> = 27)	25 (a1)	2 (b1)	0 (c1)	0 (d1)	89.29% (e1)	92.59% (f1)
iPSCs ( <i>n</i> = 43)	3 (a2)	40 (b2)	0 (c2)	0 (d2)	95.24% (e2)	93.02% (f2)
ECCs ( <i>n</i> = 9)	0 (a3)	0 (b3)	8 (c3)	1 (d3)	100.00% (e3)	88.89% (f3)
Somatic cells ( <i>n</i> = 25)	0 (a4)	0 (b4)	0 (c4)	25 (d4)	96.15% (e4)	100.00% (f4)
Precision (e1) = a1/(a1 + a2 + a3 + a4)	Recall (f1) = a1/(a1 + b1 + c1 + d1)				Precision <sub>Macro</sub>	Recall <sub>Macro</sub>
Precision (e2) = b2/(b1 + b2 + b3 + b4)	Recall (f2) = b2/(a2 + b2 + c2 + d2)				95.17%	93.63%
Precision (e3) = c3/(c1 + c2 + c3 + c4)	Recall (f3) = c3/(a3 + b3 + c3 + d3)				Accuracy	F-score <sub>Macro</sub>
Precision (e4) = d4/(d1 + d2 + d3 + d4)	Recall (f4) = d4/(a4 + b4 + c4 + d4)				94.23%	94.39%
Precision <sub>Macro</sub> = (e1 + e2 + e3 + e4)/4					Recall <sub>Macro</sub> = (f1 + f2 + f3 + f4)/4	
F-score <sub>Macro</sub> = 2 × Precision <sub>Macro</sub> × Recall <sub>Macro</sub> / (Precision <sub>Macro</sub> + Recall <sub>Macro</sub> )					Accuracy = (a1 + b2 + c3 + d4)/(27 + 43 + 9 + 25)	

<sup>a</sup>This prediction was obtained from the learned model at the 250-th epoch with regularization weight “1.00”

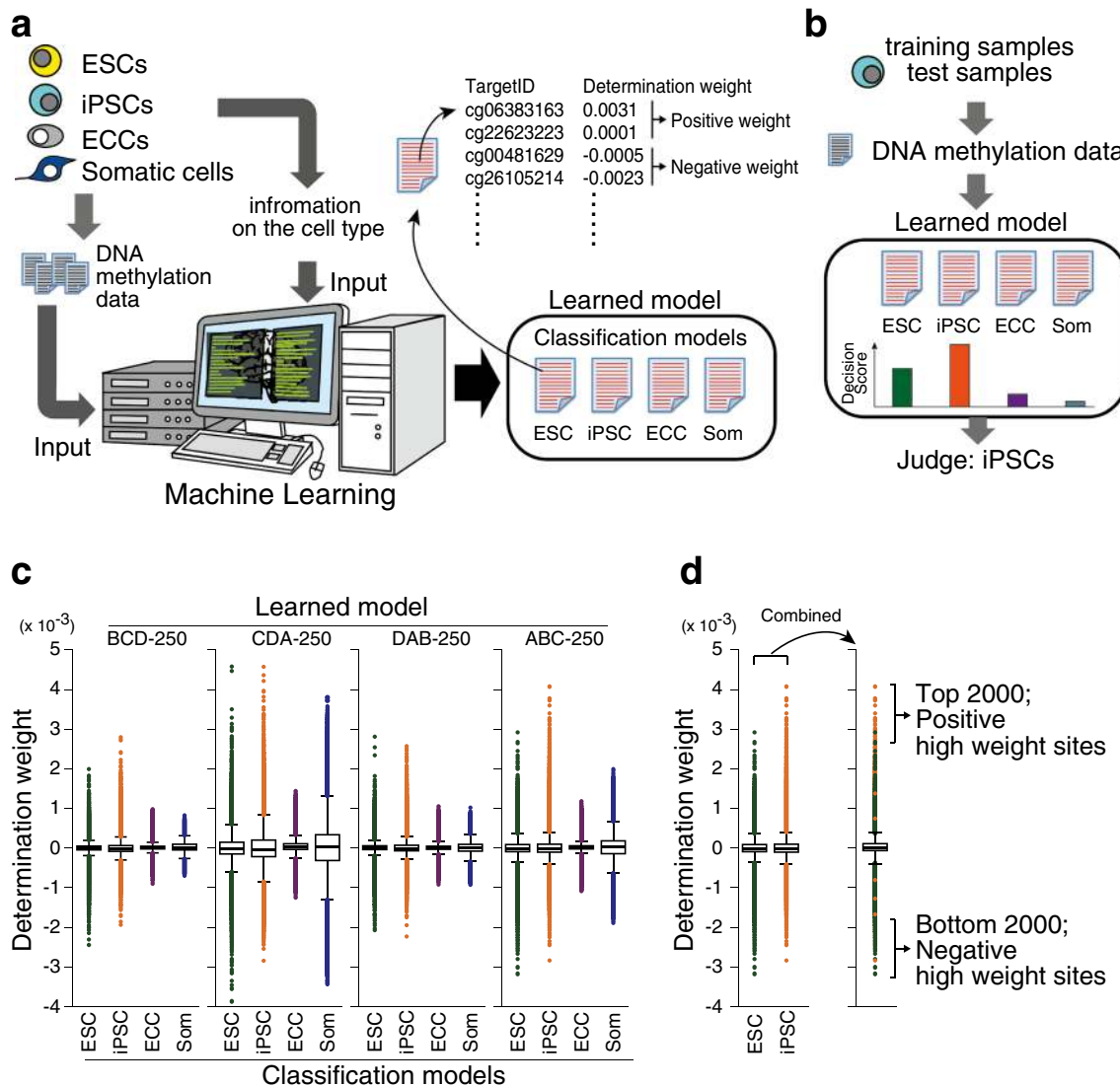


**Fig. 1** Comparison of DNA methylation between ESCs and iPSCs. **a** Unsupervised HCA based on DNA methylation. Green—ESCs, orange—iPSCs, purple—ECCs and blue—somatic cells. **b** The number of Differentially methylated regions (DMRs) between two types of cell lines

**Analysis of components of the learned models**

Analysis of components of the learned models in the 250th epoch with a regularization weight value of 1.00 was required for better understanding of its capacity to recognize ESCs and iPSCs, as well as distinguish between them. The learned model was delineated by four classification models, each corresponding to one of the following cell types: ESCs, iPSCs, ECCs, and somatic cells; each classification model comprised a list of determination weights for the 385,683 CpG sites (Fig. 2a). Since the determination weights in the various classification models for the same learned model can be compared directly (Fig. 2c), we selected highly weighted CpG sites for each learned model. The classification models for ESCs and

iPSCs in each learned model were combined and the top 2,000 highest-weight CpG sites with positive and negative values were selected from four learned models (Fig. 2d). Using comparative analysis, we found that the average number of highly weighted sites was 2.3 times higher in iPSCs than in ESCs. Interestingly, the average number of negative highly weighted sites was higher than that of the positive highly weighted sites in ESCs, whereas the average number of positive highly weighted sites was higher in iPSCs (Fig. 3a). These results suggest that the machine learned model detected more characteristic CpG sites in iPSCs than in ESCs. By extracting common highly weighted sites, 61 and 479 positive high weight sites and 93 and 181 negative high weight sites were identified in ESCs and iPSCs, respectively (Fig. 3b). By comparing

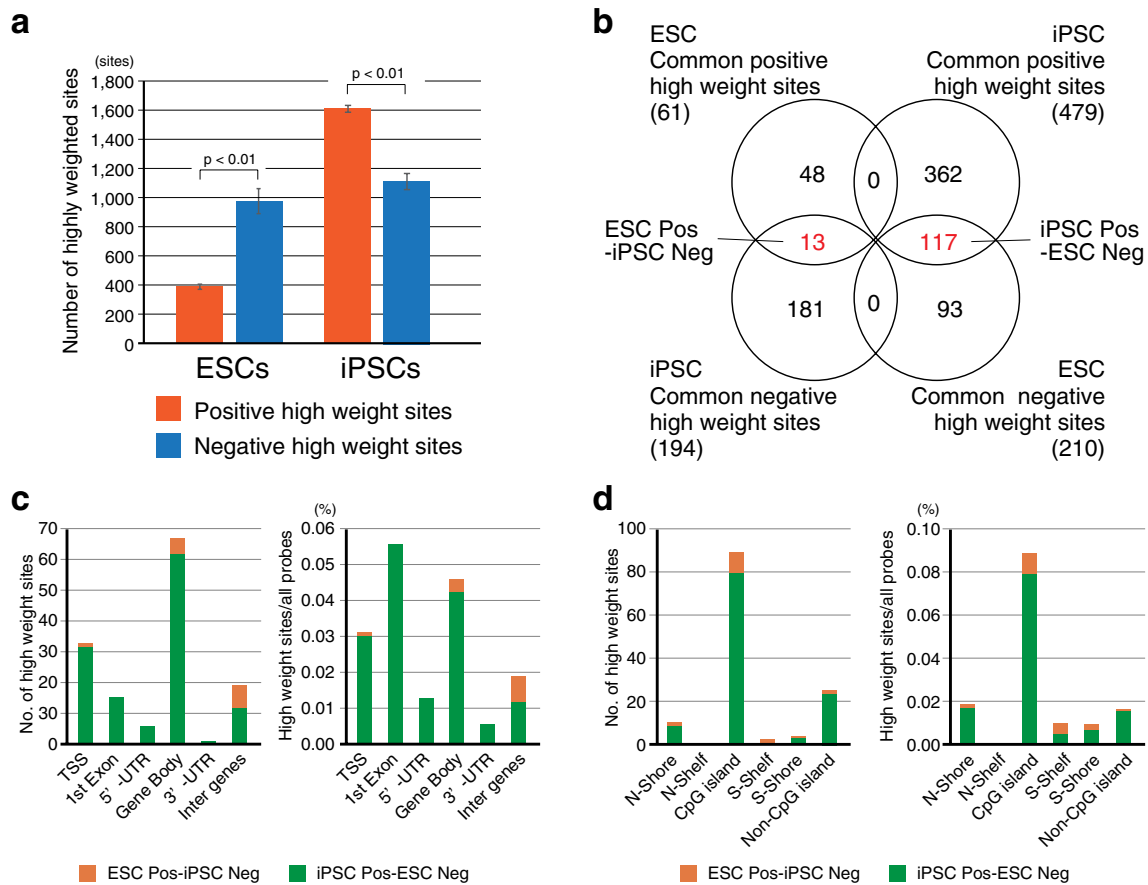


**Fig. 2** Scheme for machine learning. **a** Constructing a machine learned model. The DNA methylation data and cell type information in the training data set were used as input for the Jubatus classifier program. The four-class classification model was composed of four sets of determination weights corresponding to each cell type. Each classification model comprised a list of determination weights for each CpG sites. TargetID: ID assigned to each CpG site in Illumina HumanMethylation Array. **b** Prediction of the cell type of a sample. The cell type of a sample data was predicted by the learned model using the DNA methylation data. The four-class classification model calculated the integrated quantity obtained by multiplying the DNA

methylation rate and the determination weight, and the cell type corresponding to the classification model that produced the largest value was predicted as the cell type of the test data set. For example, if the decision score shows the highest value in iPSC in the learned model, as shown in the figure, the cell type is determined to be iPSC. **c** Boxplots of the determination weight. Dots indicate outliers in each boxplot. **d** Selection of high positive and negative weight sites. The determination files for ESCs and iPSCs in the learned models were combined. Top 2,000 and bottom 2,000 probes were selected as positive and negative high weight sites, respectively

the common highly weighted sites in ESCs with those in iPSCs, we found 13 sites common to ESC positive high weight sites and iPSC negative high weight sites (designated as ESC Pos-iPSC Neg), and 117 sites common to

iPSC positive high weight sites and ESC negative high weight sites (designated as iPSC Pos-ESC Neg) (Fig. 3b and Supplemental Table 3). The iPSC Pos-ESC Neg sites were found to be abundant around the transcription start



**Fig. 3** Analysis of determination weights of the learned models. **a** The average number of high weight CpG sites in iPSCs and ESCs. Data are represented as mean  $\pm$  SEM. **b** Venn-like diagram showing overlapping high weight CpG sites of iPSCs and ESCs. **c** The num-

ber (left graph) and proportion (right graph) of the overlapping high weight CpG sites associated with gene figures. **d** The number (left graph) and proportion (right graph) of the overlapping high weight CpG sites associated with CpG islands

site (TSS), first exon and gene body (Fig. 3c), and CpG island (Fig. 3d).

### Distribution of the iPSC Pos-ESC Neg high weight sites on chromosome

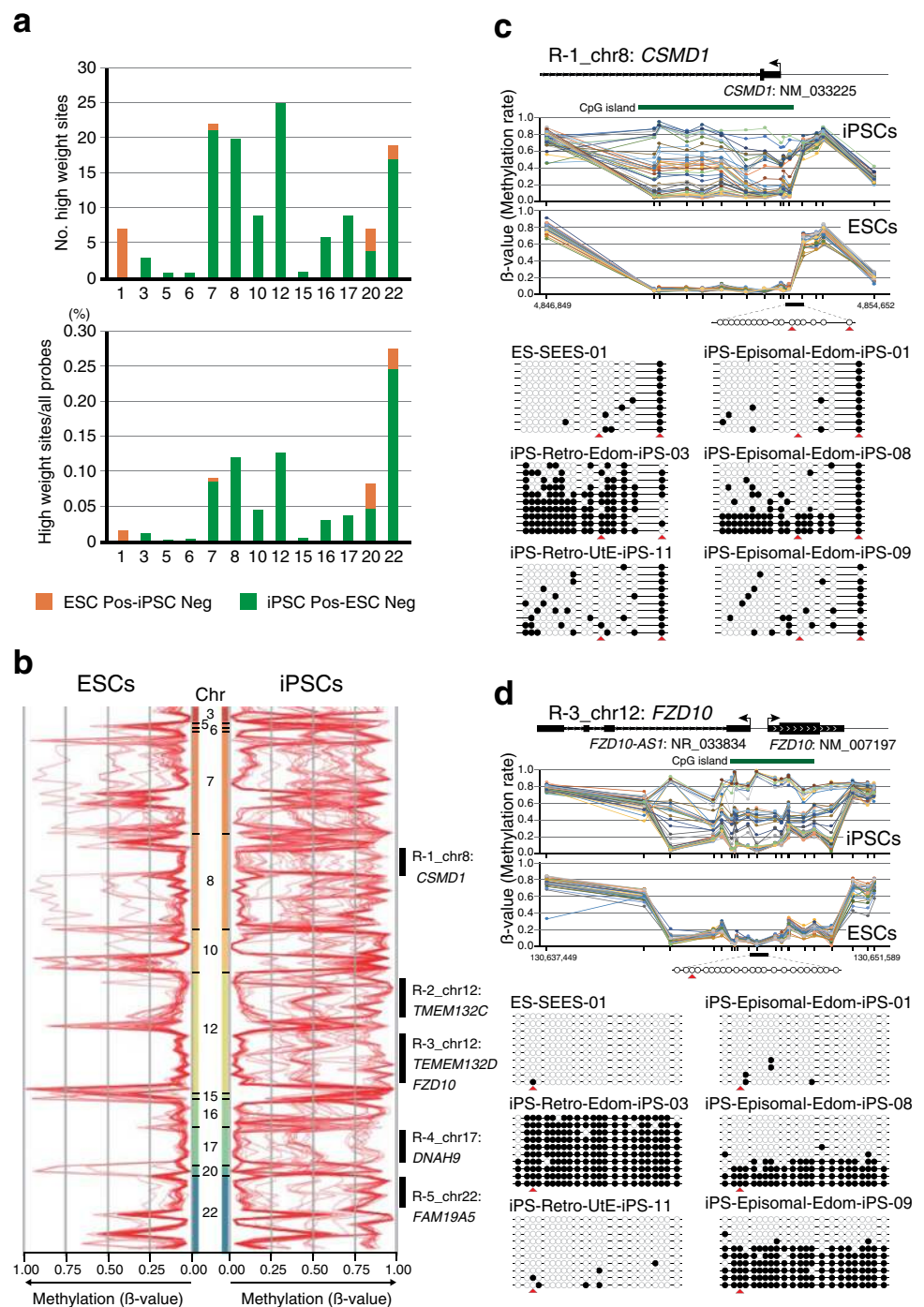
The iPSC Pos-ESC Neg sites were found to be abundant on chromosomes 7, 8, 12, and 22 (Fig. 4a). Next, we focused on the DNA methylation rate of high weighted sites. We compared the DNA methylation rates of iPSC Pos-ESC Neg sites in ESCs and iPSCs, and identified five regions in which the DNA methylation fluctuated only in iPSCs (Fig. 4b); these regions were found in the following genes: CUB And Sushi Multiple Domains 1 (*CSMD1*), Transmembrane Protein 132C (*TMEM132C*), Transmembrane Protein 132D (*TMEM132D*), Frizzled 10 (*FZD10*), dynein axonemal heavy chain 9 (*DNAH9*), and Tafa chemokine like family member 5 (*FAM19A5*). The fluctuating regions in these genes were located around the TSS (Fig. 4c, d and Supplemental Fig. 4). To confirm variable methylation at

those regions, sodium bisulfite sequencing analysis was performed. Consistent with the results of Infinium HumanMethylation assay, these regions showed variable methylation in iPSCs (Fig. 4c, d and Supplemental Fig. 4). However, these genes are rarely expressed in ESCs and iPSCs.

### Analysis of the high weight sites

The top ten highly weighted sites were selected and the DNA methylation rates in these sites were compared. The DNA methylation rates of the iPSC Pos-ESC Neg sites in an individual line of iPSCs were found to be widely distributed, whereas ESCs generally had a low methylation rate (Fig. 5a and Supplemental Fig. 5a). On the contrary, the ESC Pos-iPSC Neg sites had varied methylation rates in both ESCs and iPSCs (Fig. 5b and Supplemental Fig. 5b). Variations in the DNA methylation of the high weight sites in iPSCs were not due to the differences in the methods of iPSC generation or the types of the parental cells (Fig. 6a and Supplemental Fig. 5c). Interestingly, the ESCs showed a larger number of variably

**Fig. 4** Distribution of the iPSC Pos-ESC Neg high weight sites on chromosomes. **a** The number (upper graph) and proportion (lower graph) of the overlapping high weight CpG sites associated with chromosomes. **b** DNA methylation rate of the iPSC Pos-ESC Neg sites. Five regions (R1–R5) in which the DNA methylation fluctuations were seen only in iPSCs were identified. A red line indicates a cell line. **c, d** DNA methylation rate of *CSMD1* (**c**) and *FZD10* (**d**) genes loci and sodium bisulfite sequencing analysis. (Top) Upper and lower graphs shows DNA methylation rates in iPSCs and ESCs, respectively. A line indicates a cell line. (Bottom) Bisulfite sequencing results. Open and closed circles indicate unmethylated and methylated sites, respectively. Red arrowheads represent the position of CpG sites in the Infinium assay. See also Supplemental Fig. 4



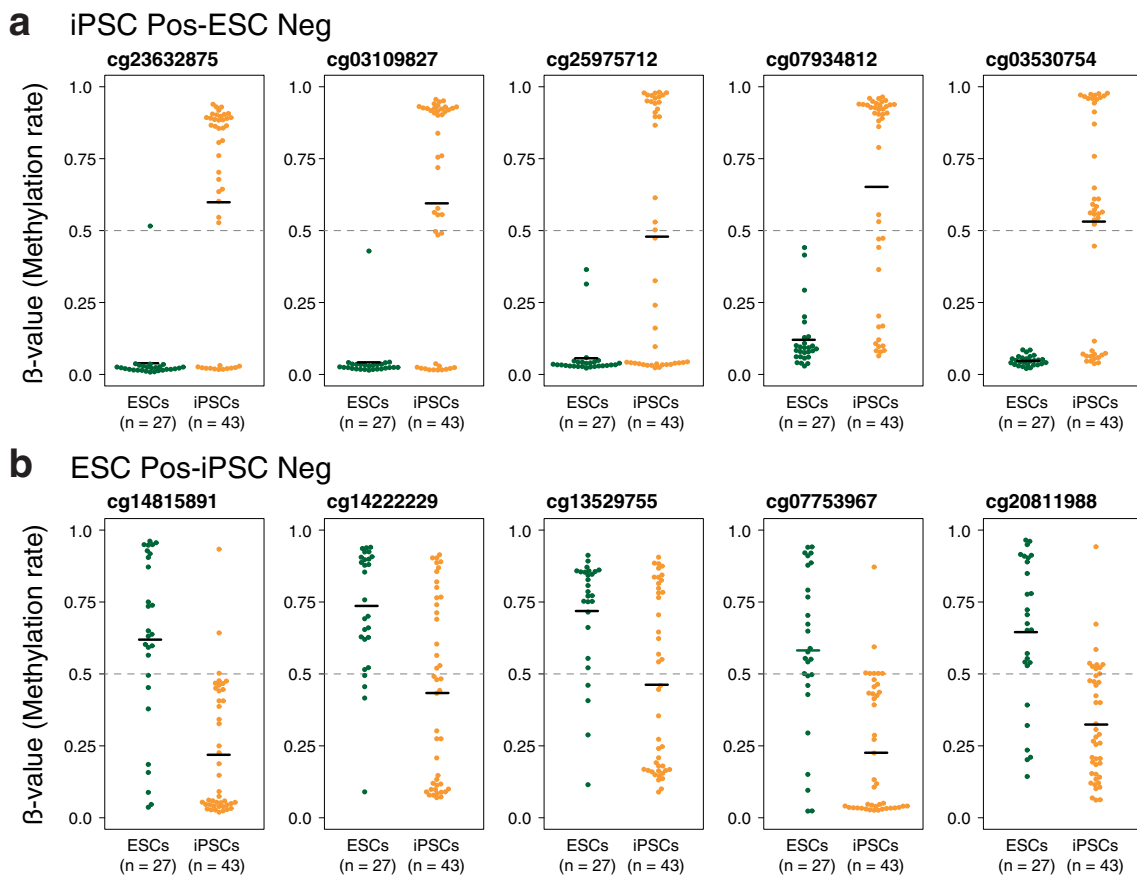
methylated regions (Fig. 6b) and CpG sites with high standard deviation compared to the iPSCs (Fig. 6c) in the analysis of all CpG sites, indicating that ESCs have more variability in DNA methylation rates than iPSCs. However, in the high weight sites, iPSCs had more CpG sites with high standard deviation compared to ESCs, indicating high variability in iPSCs with respect to methylation levels at the high weight sites (Fig. 6d). These results suggest that the machine learning method was able to determine CpG sites with DNA methylation diversity

specific to iPSCs, which can be considered as a characteristic for distinguishing iPSCs from ESCs.

## Discussion

In this study, we developed a new method to distinguish between iPSCs and ESCs on the basis of their DNA methylation profiles. We constructed a learning model based on





**Fig. 5** Analysis of the high weight CpG sites. **a** DNA methylation rate of the top five high weight CpG sites in the iPSC Pos-ESC Neg sites. See also Supplemental Fig. 5a. **b** DNA methylation rate of the

top five high weight CpG sites in the ESC Pos-iPSC Neg sites. See also Supplemental Fig. 5b

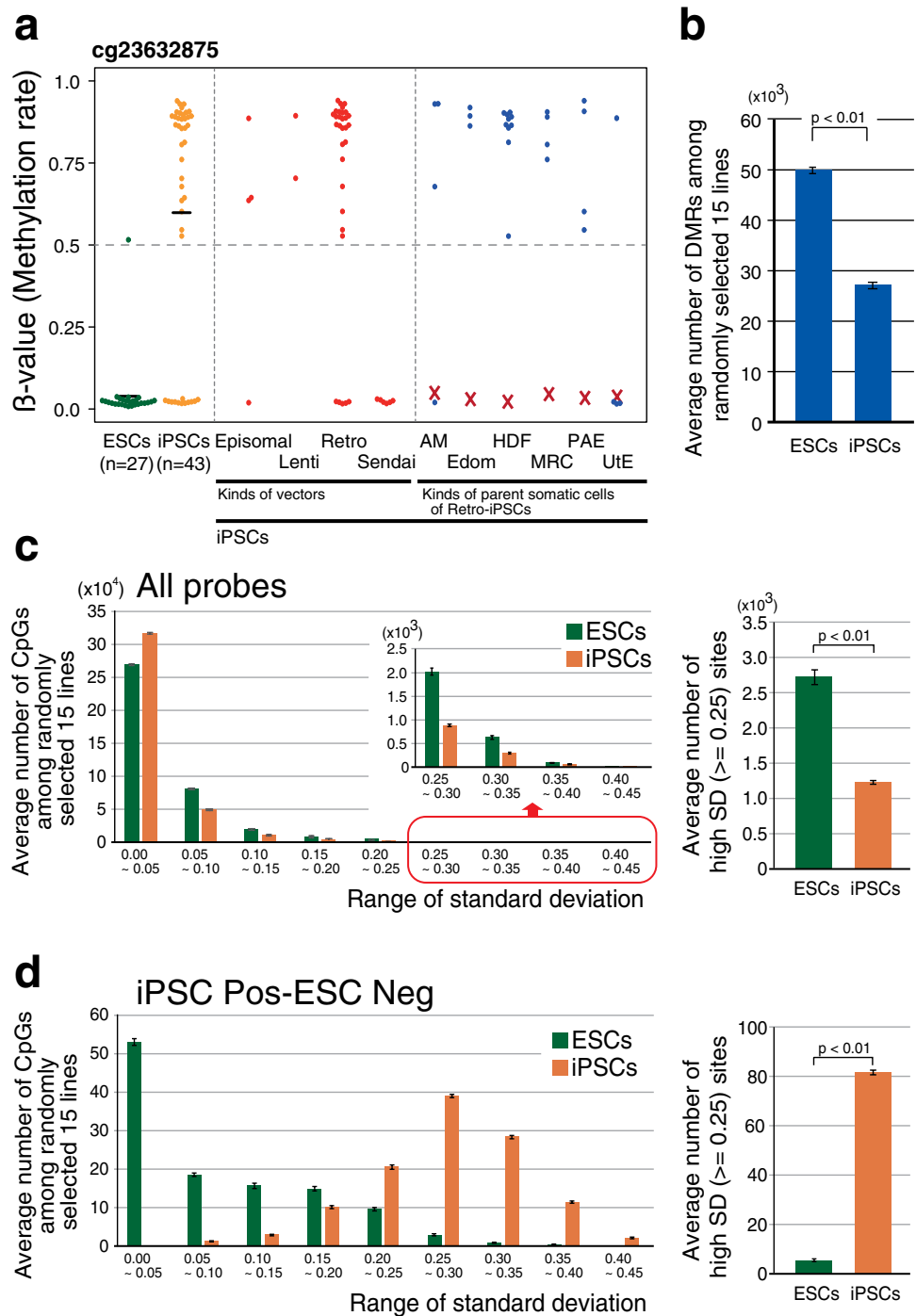
the linear model for multi-class classification using Jubatus, a machine learning platform. In recent years, deep learning methods have often been used for biological analysis; however, these methods usually require at least 10,000 samples. The availability of only 10–100 variants of human iPSC lines makes the linear model classification system ideal for the analysis of human iPSCs.

iPSCs are essentially an alternative to ESCs, with almost no difference between the two in terms of their properties. iPSC lines generated with non-genome integration methods, such as episomal vector or RNA transfection, are indistinguishable from ESCs in terms of morphology, differentiation ability, gene expression, DNA methylation, etc. [24, 38]. The results obtained in this study are in agreement with previous reports, as no epigenetic features that clearly distinguished iPSCs from ESCs were found. However, our analyses, using a collection of DNA methylation profiles from different types of cells, including 43 iPSC lines which contained Retro-, Lenti-, Sendai-, and Episomal-iPSCs, 27 ESC lines, 9 ECC lines, and 25 somatic cell lines, demonstrated that machine learning with AROW, a linear model for

classification, is effective for the discrimination of cell types, especially iPSCs and ESCs. The learned models achieved high-accuracy prediction rates in distinguishing iPSCs from ESCs. In other words, our learned models recognized the differences between iPSCs and ESCs and were able to discriminate between the cell types. Interestingly, the learned models recognized the iPSC lines as iPSCs, irrespective of the production methods used.

One of the advantages of a linear classification-based learning model is the ability to select and analyze components, such as determination weights corresponding to each CpG site. The analysis of the high weight components revealed that the learned models searched for genomic regions that are characteristic of iPSCs and used them to distinguish iPSCs from ESCs. This resulted in the identification of fluctuating iPSC-specific methylation regions, which are especially abundant on chromosomes 7, 8, 12, and 22. DNA methylation was more variable in each of the ESC lines than in iPSCs, indicating that the ESCs possessed more fluctuating methylation regions than the iPSCs. Despite the methylation variation in ESCs, the learned models selected

**Fig. 6** Variations in the DNA methylation rates of the high weight sites. **a** DNA methylation rate of the representative iPSC Pos-ESC Neg sites associated with the methods of iPSC production or type of the parental cells.  $X$  in the plot indicates the DNA methylation rate of the parental cells. See also Supplemental Fig. 5c. **b** Comparison of the average number of variably methylated regions in the same cell type of ESCs and iPSCs. Data are represented as mean  $\pm$  SEM. **c**, **d** Average number of CpG sites associated with the range of standard deviation (SD) of DNA methylation rates (left) and the average number of high SD CpG sites (right) in all probes (**c**) and the iPSC Pos-ESC Neg sites (**d**). Data are represented as mean  $\pm$  SEM



CpG sites with DNA methylation diversity specific to iPSCs as characteristics for distinguishing iPSCs from ESCs. Comparison of the DNA methylation rates of the iPSC Pos-ESC Neg sites led to the identification of fluctuating methylation regions in six genes, including *CSMD1*, *TMEM132C*, *TMEM132D*, *FZD10*, *DNAH9*, and *FAM19A5*. *CSMD1* is known to be a tumor-suppressor gene under the control of DNA methylation in liver cancer and head and neck squamous cell carcinoma [39–41]. *TMEM132C* has been reported

to show differential methylation and is downregulated by DNA hypermethylation in breast tumors [42]. *TMEM132D* [43] and *DNAH9* [44] are cancer-associated genes in small cell lung cancer, and *FDZ10* has a role in cancer reactivation [45]. The expression of *FAM19A5*, also known as *TAF5A*, is influenced by the activation of  $\beta$ -catenin [46] and c-Myc promotes the Wnt/ $\beta$ -catenin activity in breast cancers [47]. These genes are involved in carcinogenesis, and the fluctuating regions in these genes are located around the TSS;

this suggests that variations in DNA methylation in these genes influence the risk of iPSCs. However, it is seen that these variations in DNA methylation do not affect the gene expression profiles in either ESCs or iPSCs, and also do not exert any influence on pluripotency. Nevertheless, it is possible that the fluctuations in methylation may affect the differentiation properties of iPSCs. The possible effects of such methylation fluctuations on the differentiation properties of iPSCs need to be evaluated through further detailed investigations.

Comparison of iPSCs obtained through different production methods revealed that Sendai-iPSCs were the least diverse in terms of fluctuating methylation regions, and their DNA methylation pattern showed maximum similarity with that of the ESCs. The similarity observed between Sendai-iPSCs and ESCs is consistent with the result of a previously reported comprehensive DNA methylation analysis [23]. However, no significant differences were detected in pluripotency between the Sendai-iPSCs and the iPSCs derived from other production methods [23]. Aberrant DNA methylation at some imprinted gene loci in ESCs and iPSCs has been reported [9, 48, 49], and this abnormality was detected in 68 imprinted genes [23], indicating that aberrant DNA methylation occurs widely in human ESCs and iPSCs. In this study, we identified 130 high weight sites, including 13 ESC Pos-iPSC Neg and 117 iPSC Pos-ESC Neg CpG sites; however, there were no imprinted genes in the 130 high weight sites, suggesting that the abnormalities of imprinted genes are not specific to either iPSCs or ESCs.

In conclusion, we were able to distinguish human iPSCs from ESCs using machine learning methods, even when the cells lacked specific markers. The results of this study will have a significant effect on the use of these cell lines in various in vitro research studies for specific purposes. In addition, an epigenetic signature of iPSCs was identified by component analysis using our learned models. The learned models developed in this study contribute towards enhancing our understanding of the iPSCs at the gene level and hold potential for achieving remarkable advances in various fields of biology research, including computational biology, molecular biology, cell biology, and cancer biology. The approach of the machine learning method used in this study is useful for comprehensive data analysis and can be widely applied to iPSC research as well as many other fields of research in life sciences.

**Acknowledgements** We would like thank Ms. Haruka Tomaru, Ms. Yukie Toyonaga, and Ms. Kae Inoue for performing cell culture. This research was supported in part by AMED under Grant number JP18bm0704003 and 19bk0104092h0001 to K. N.

**Author contributions** Conception and experiment design: KN and AU; Experimentation: KN, KT, YA, and AS; Data analysis: KN, KT, and

AS; Contribution of reagents/materials/analysis tool: KN, KO, HA, and AU; Manuscript writing: KN.

## Compliance with ethical standards

**Conflict of interest** The authors declare no conflicts of interest.

**Ethical approval** 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.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Li E. Chromatin modification and epigenetic reprogramming in mammalian development. *Nat Rev Genet.* 2002;3(9):662–73.
- Nishino K, Hattori N, Tanaka S, Shiota K. DNA methylation-mediated control of Sry gene expression in mouse gonadal development. *J Biol Chem.* 2004;279(21):22306–13.
- Hattori N, Nishino K, Ko YG, et al. Epigenetic control of mouse Oct-4 gene expression in embryonic stem cells and trophoblast stem cells. *J Biol Chem.* 2004;279(17):17063–9.
- Ko YG, Nishino K, Hattori N, Arai Y, Tanaka S, Shiota K. Stage-by-stage change in DNA methylation status of Dnmt1 locus during mouse early development. *J Biol Chem.* 2005;280(10):9627–34.
- Reik W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature.* 2007;447(7143):425–32.
- Hattori N, Imao Y, Nishino K, et al. Epigenetic regulation of Nanog gene in embryonic stem and trophoblast stem cells. *Genes Cells.* 2007;12(3):387–96.
- Shiota K, Kogo Y, Ohgane J, et al. Epigenetic marks by DNA methylation specific to stem, germ and somatic cells in mice. *Genes Cells.* 2002;7(9):961–9.
- Nishino K, Toyoda M, Yamazaki-Inoue M, et al. Defining hypomethylated regions of stem cell-specific promoters in human iPSC cells derived from extra-embryonic amnions and lung fibroblasts. *PLoS ONE.* 2010;5(9):e13017.
- Nishino K, Toyoda M, Yamazaki-Inoue M, et al. DNA methylation dynamics in human induced pluripotent stem cells over time. *PLoS Genet.* 2011;7(5):e1002085.
- Park IH, Arora N, Huo H, et al. Disease-specific induced pluripotent stem cells. *Cell.* 2008;134(5):877–86.
- Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell.* 2006;126(4):663–76.

12. Takahashi K, Tanabe K, Ohnuki M, et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*. 2007;131(5):861–72.
13. Yu J, Vodyanik MA, Smuga-Otto K, et al. Induced pluripotent stem cell lines derived from human somatic cells. *Science*. 2007;318(5858):1917–20.
14. Choi J, Lee S, Mallard W, et al. A comparison of genetically matched cell lines reveals the equivalence of human iPSCs and ESCs. *Nat Biotechnol*. 2015;33(11):1173–81.
15. Doi A, Park IH, Wen B, et al. Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat Genet*. 2009;41(12):1350–3.
16. Lister R, Pelizzola M, Kida YS, et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*. 2011;471(7336):68–73.
17. Bock C, Kiskinis E, Verstaappen G, et al. Reference Maps of human ES and iPSC cell variation enable high-throughput characterization of pluripotent cell lines. *Cell*. 2011;144(3):439–52.
18. Cui CH, Uyama T, Miyada K, 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(5):1586–94.
19. Nagata S, Toyoda M, Yamaguchi S, et al. Efficient reprogramming of human and mouse primary extra-embryonic cells to pluripotent stem cells. *Genes Cells*. 2009;14(12):1395–404.
20. Jacobs JP, Jones CM, Baille JP. Characteristics of a human diploid cell designated MRC-5. *Nature*. 1970;227(5254):168–70.
21. Makino H, Toyoda M, Matsumoto K, et al. Mesenchymal to embryonic incomplete transition of human cells by chimeric OCT4/3 (POU5F1) with physiological co-activator EWS. *Exp Cell Res*. 2009;315(16):2727–40.
22. Toyoda M, Yamazaki-Inoue M, Itakura Y, et al. Lectin microarray analysis of pluripotent and multipotent stem cells. *Genes Cells*. 2011;16(1):1–11.
23. Nishino K, Arai Y, Takasawa K, et al. Epigenetic-scale comparison of human iPSCs generated by retrovirus, Sendai virus or episomal vectors. *Regen Ther*. 2018;9:71–8.
24. Okita K, Matsumura Y, Sato Y, et al. A more efficient method to generate integration-free human iPS cells. *Nat Methods*. 2011;8(5):409–12.
25. Nishimura K, Sano M, Ohtaka M, et al. Development of defective and persistent Sendai virus vector: a unique gene delivery/expression system ideal for cell reprogramming. *J Biol Chem*. 2011;286(6):4760–71.
26. Akutsu H, Machida M, Kanzaki S, et al. Xenogeneic-free defined conditions for derivation and expansion of human embryonic stem cells with mesenchymal stem cells. *Regener Ther*. 2015;1:18–29.
27. Cowan CA, Klimanskaya I, McMahon J, et al. Derivation of embryonic stem-cell lines from human blastocysts. *N Engl J Med*. 2004;350(13):1353–6.
28. Osafune K, Caron L, Borowiak M, et al. Marked differences in differentiation propensity among human embryonic stem cell lines. *Nat Biotechnol*. 2008;26(3):313–5.
29. Hata J, Fujimoto J, Ishii E, et al. Differentiation of human germ cell tumor cells in vivo and in vitro. *Acta Histochem Cytochem*. 1992;25:563–76.
30. Teshima S, Shimosato Y, Hirohashi S, et al. Four new human germ cell tumor cell lines. *Lab Invest*. 1988;59(3):328–36.
31. Giovanella BC, Stehlin JS, Williams LJ Jr. Heterotransplantation of human malignant tumors in "nude" thymusless mice. II. Malignant tumors induced by injection of cell cultures derived from human solid tumors. *J Natl Cancer Inst*. 1974;52(3):921–30.
32. Motoyama T, Watanabe H, Yamamoto T, Sekiguchi M. Human testicular germ cell tumors in vitro and in athymic nude mice. *Acta Pathol Jpn*. 1987;37(3):431–48.
33. Bronson DL, Andrews PW, Vessella RL, Fraley EE. In vitro differentiation of human embryonal carcinoma cells. In: Silver LM, Martin GR, Strickland S, editors. *Teratocarcinoma stem cells*, vol. 10. New York: Cold Spring Harbour Conferences on Cell Proliferation; 1983. p. 597–605.
34. Andrews PW, Damjanov I, Simon D, et al. Pluripotent embryonal carcinoma clones derived from the human teratocarcinoma cell line Tera-2 differentiation in vivo and in vitro. *Lab Invest*. 1984;50(2):147–62.
35. Okamura K, Kawai T, Hata K, Nakabayashi K. Lists of Human-Methylation450 BeadChip probes with nucleotide-variant information obtained from the Phase 3 data of the 1000 Genomes Project. *Genom Data*. 2016;7:67–9.
36. Crammer K, Kulesza A, Dredze M. Adaptive regularization of weight vectors. *Mach Learn*. 2013;91:155–87.
37. Kumaki Y, Oda M, Okano M. QUMA: quantification tool for methylation analysis. *Nucleic Acids Res*. 2008;36(Web Server issue):W170–W175.
38. Warren L, Manos PD, Ahfeldt T, et al. Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. *Cell Stem Cell*. 2010;7(5):618–30.
39. Toomes C, Jackson A, Maguire K, et al. The presence of multiple regions of homozygous deletion at the CSMD1 locus in oral squamous cell carcinoma question the role of CSMD1 in head and neck carcinogenesis. *Genes Chromosomes Cancer*. 2003;37(2):132–40.
40. Richter TM, Tong BD, Scholnick SB. Epigenetic inactivation and aberrant transcription of CSMD1 in squamous cell carcinoma cell lines. *Cancer Cell Int*. 2005;5:29.
41. Midorikawa Y, Yamamoto S, Tsuji S, et al. Allelic imbalances and homozygous deletion on 8p23.2 for stepwise progression of hepatocarcinogenesis. *Hepatology*. 2009;49(2):513–22.
42. de Almeida BP, Apolonio JD, Binnie A, Castelo-Branco P. Roadmap of DNA methylation in breast cancer identifies novel prognostic biomarkers. *BMC Cancer*. 2019;19(1):219.
43. Iwakawa R, Kohno T, Totoki Y, et al. Expression and clinical significance of genes frequently mutated in small cell lung cancers defined by whole exome/RNA sequencing. *Carcinogenesis*. 2015;36(6):616–21.
44. Kusakabe M, Kutomi T, Watanabe K, et al. Identification of G0S2 as a gene frequently methylated in squamous lung cancer by combination of in silico and experimental approaches. *Int J Cancer*. 2010;126(8):1895–902.
45. Scavo MP, Depalo N, Rizzi F, et al. FZD10 carried by exosomes sustains cancer cell proliferation. *Cells*. 2019;8(8):8.
46. Karner CM, Das A, Ma Z, et al. Canonical Wnt9b signaling balances progenitor cell expansion and differentiation during kidney development. *Development*. 2011;138(7):1247–57.
47. Cowling VH, Cole MD. Turning the tables: Myc activates Wnt in breast cancer. *Cell Cycle*. 2007;6(21):2625–7.
48. Rugg-Gunn PJ, Ferguson-Smith AC, Pedersen RA. Epigenetic status of human embryonic stem cells. *Nat Genet*. 2005;37(6):585–7.
49. Rugg-Gunn PJ, Ferguson-Smith AC, Pedersen RA. Status of genomic imprinting in human embryonic stem cells as revealed by a large cohort of independently derived and maintained lines. *Hum Mol Genet*. 2007;16(2):R243–R251.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.