

Supplementary materials

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

Koichiro Nishino^{1,2*}, Ken Takasawa^{1#a}, Kohji Okamura³ Yoshikazu Arai¹, Asato Sekiya¹, Hidenori Akutsu⁴ and Akihiro Umezawa⁴

¹Laboratory of Veterinary Biochemistry and Molecular Biology, Graduate School of Medicine and Veterinary Medicine/Faculty of Agriculture, University of Miyazaki, Miyazaki, Japan.

²Center for Animal Disease Control, University of Miyazaki, Miyazaki, Japan.

³Department of Systems BioMedicine, National Research Institute for Child Health and Development, Tokyo, Japan.

⁴Department of Reproductive Biology, Center for Regenerative Medicine, National Research Institute for Child Health and Development, Tokyo, Japan.

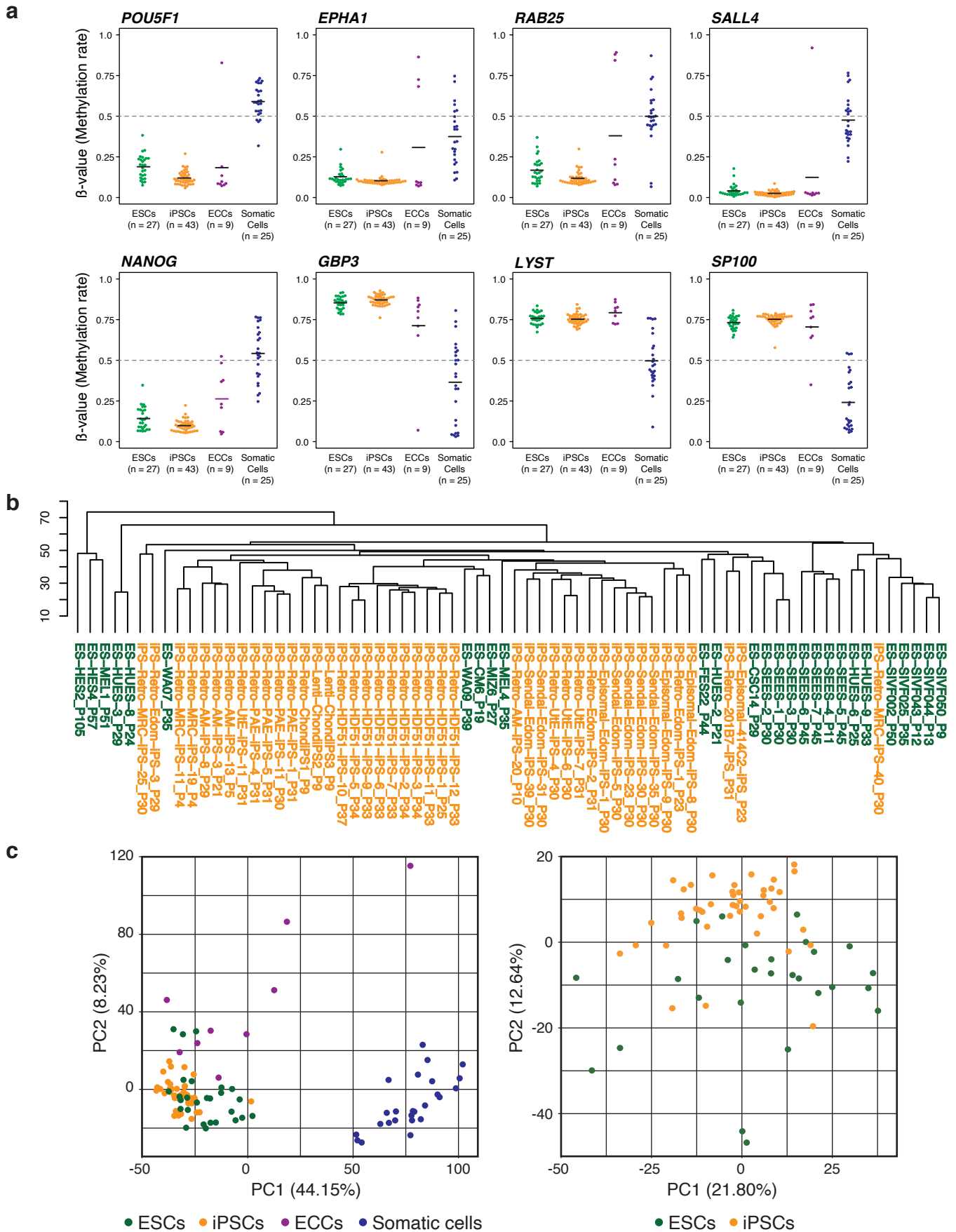
^{#a} Current address: Division of Molecular Modification and Cancer Biology, National Cancer Center Research Institute, Tokyo, Japan and Cancer Translational Research Team, RIKEN Center for Advanced Intelligence Project, Tokyo, Japan

* Corresponding author

E-mail: aknishino@cc.miyazaki-u.ac.jp (KN)

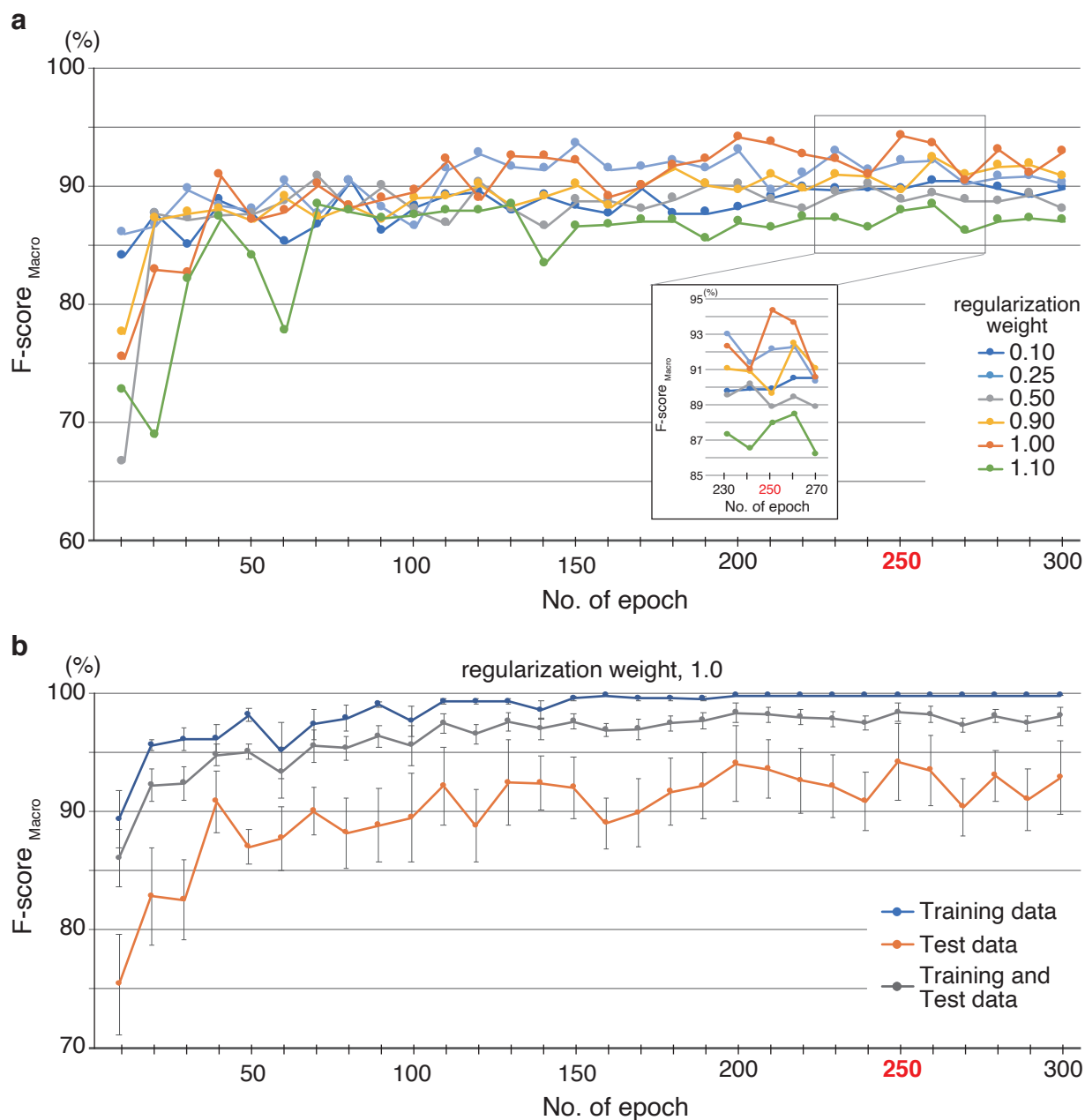
Tel: +81-985-58-7263

Fax: +81-985-58-7263



Supplemental Fig. 1 DNA methylation rate at the promoter regions of pluripotency-associated genes and principal component analysis

(a) DNA methylation rate at the promoter regions of *POU5F1*, *EPHA1*, *RAB25*, *SALL4*, *NANOG*, *GBP3*, *LYST* and *SP100*. A plot indicates a cell line. **(b)** Unsupervised HCA based on DNA methylation. Green - ESCs, orange - iPSCs. **(c)** PCA based on DNA methylation. Green - ESCs, orange - iPSCs, purple - ECCs and blue - somatic cells.



Supplemental Fig. 2 Learning curve obtained from the four learning series

(a) Cross validation of the regularization weight value and epochs in the average of F-score_{Macro} rate in test data set. The highest average of the F-score_{Macro} rate was recorded at the 250th epoch with regularization weight value of “1.00”. **(b)** The average of the F-score_{Macro} rate of the training data set and test data set at the 250th epoch with regularization weight value of “1.00”. The values are shown as the means $\pm$ S.E.

Learning Model: BCD-250

Class \ Predicted	ESC	iPSC	ECC	Somatic cells	Precision	Recall
ESCs (n = 7)	6 (a1)	1 (b1)	0 (c1)	0 (d1)	75.00% (e1)	85.71% (f1)
iPSCs (n = 11)	2 (a2)	9 (b2)	0 (c2)	0 (d2)	90.00% (e2)	81.82% (f2)
ECCs (n = 2)	0 (a3)	0 (b3)	1 (c3)	1 (d3)	100.00% (e3)	50.00% (f3)
Somatic cells (n = 6)	0 (a4)	0 (b4)	0 (c4)	6 (d4)	85.71% (e4)	100.00% (f4)

Precision _{Macro}	Recall _{Macro}	F-score _{Macro}	Accuracy
87.68%	79.38%	83.32%	84.62%

Learning Model: CDA-250

Class \ Predicted	ESC	iPSC	ECC	Somatic cells	Precision	Recall
ESCs (n = 7)	6 (a1)	1 (b1)	0 (c1)	0 (d1)	100.00% (e1)	85.71% (f1)
iPSCs (n = 11)	0 (a2)	11 (b2)	0 (c2)	0 (d2)	91.67% (e2)	100.00% (f2)
ECCs (n = 2)	0 (a3)	0 (b3)	2 (c3)	0 (d3)	100.00% (e3)	100.00% (f3)
Somatic cells (n = 6)	0 (a4)	0 (b4)	0 (c4)	6 (d4)	100.00% (e4)	100.00% (f4)

Precision _{Macro}	Recall _{Macro}	F-score _{Macro}	Accuracy
97.92%	96.43%	97.17%	96.15%

Learning Model: DAB-250

Class \ Predicted	ESC	iPSC	ECC	Somatic cells	Precision	Recall
ESCs (n = 7)	7 (a1)	0 (b1)	0 (c1)	0 (d1)	100.00% (e1)	100.00% (f1)
iPSCs (n = 11)	0 (a2)	11 (b2)	0 (c2)	0 (d2)	100.00% (e2)	100.00% (f2)
ECCs (n = 2)	0 (a3)	0 (b3)	2 (c3)	0 (d3)	100.00% (e3)	100.00% (f3)
Somatic cells (n = 6)	0 (a4)	0 (b4)	0 (c4)	6 (d4)	100.00% (e4)	100.00% (f4)

Precision _{Macro}	Recall _{Macro}	F-score _{Macro}	Accuracy
100.00%	100.00%	100.00%	100.00%

Learning Model: ABC-250

Class \ Predicted	ESC	iPSC	ECC	Somatic cells	Precision	Recall
ESCs (n = 6)	6 (a1)	0 (b1)	0 (c1)	0 (d1)	85.71% (e1)	100.00% (f1)
iPSCs (n = 10)	1 (a2)	9 (b2)	0 (c2)	0 (d2)	100.00% (e2)	90.00% (f2)
ECCs (n = 3)	0 (a3)	0 (b3)	3 (c3)	0 (d3)	100.00% (e3)	100.00% (f3)
Somatic cells (n = 7)	0 (a4)	0 (b4)	0 (c4)	7 (d4)	100.00% (e4)	100.00% (f4)

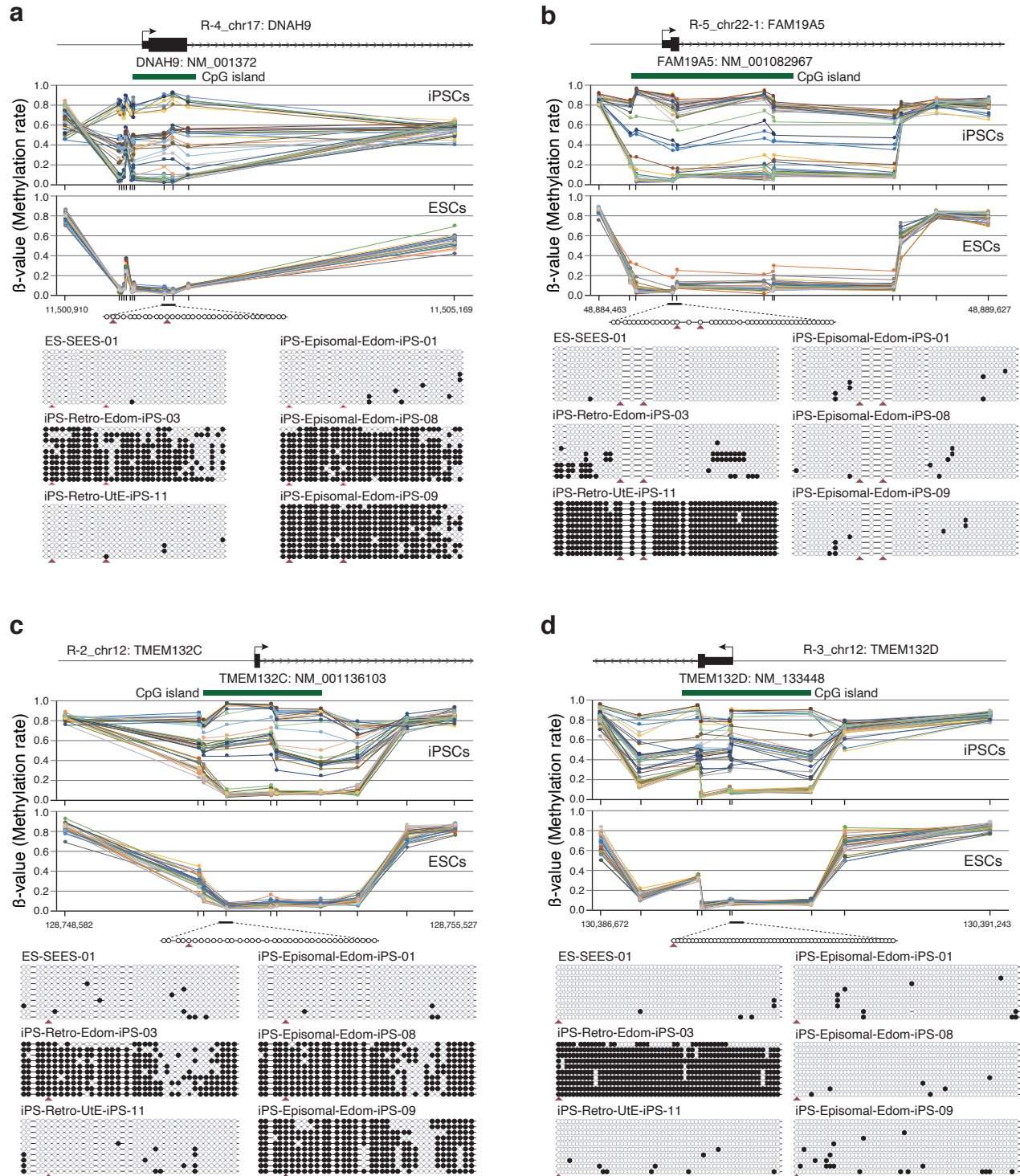
Precision _{Macro}	Recall _{Macro}	F-score _{Macro}	Accuracy
96.43%	97.50%	96.96%	96.15%

Average

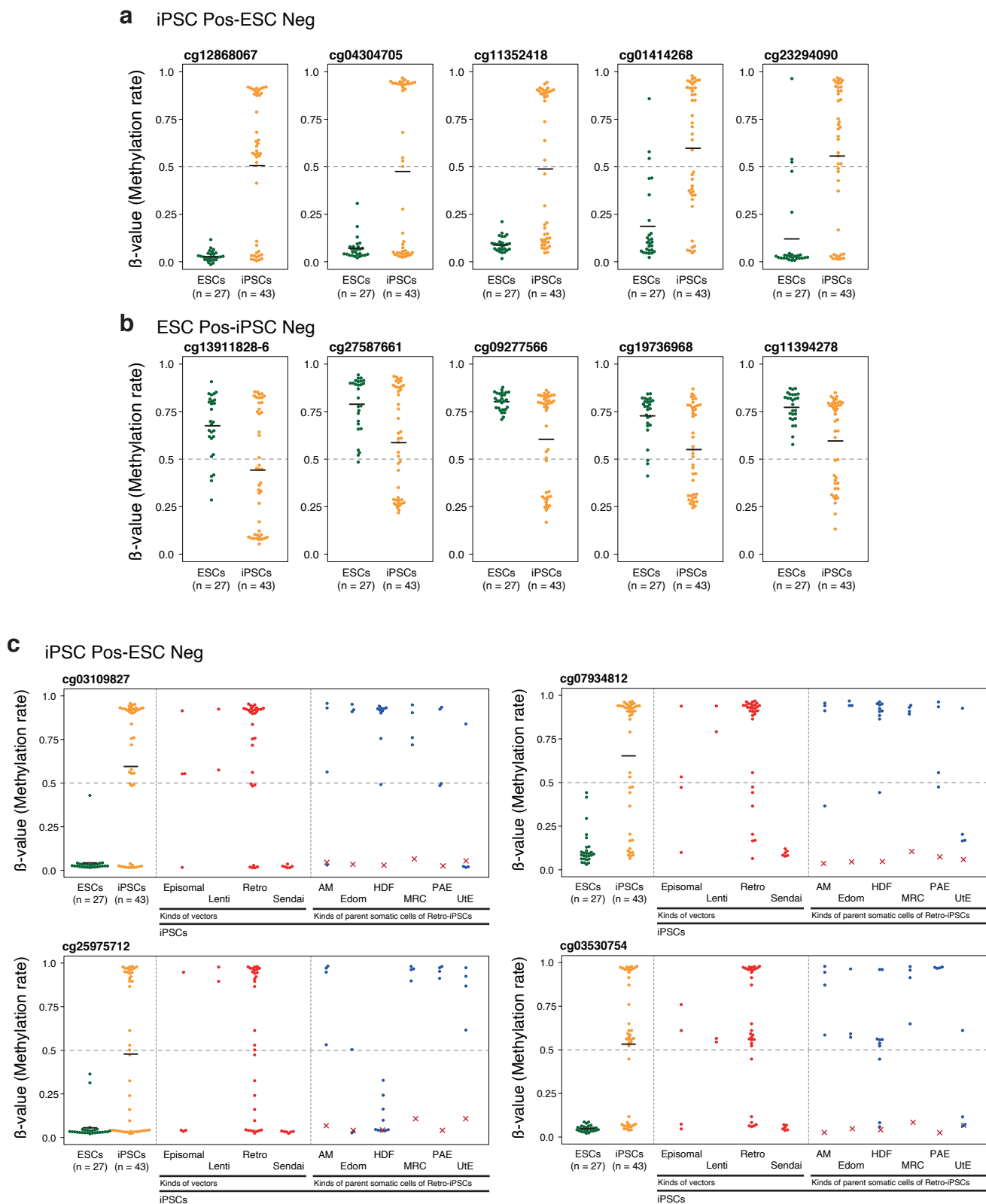
Precision _{Macro}	Recall _{Macro}	F-score _{Macro}	Accuracy
95.51%	93.33%	94.36%	94.23%

$\text{Precision (e1)} = a1 / (a1 + a2 + a3 + a4)$ $\text{Recall (f1)} = a1 / (a1 + b1 + c1 + d1)$
 $\text{Precision (e2)} = b2 / (b1 + b2 + b3 + b4)$ $\text{Recall (f2)} = b2 / (a2 + b2 + c2 + d2)$
 $\text{Precision (e3)} = c3 / (c1 + c2 + c3 + c4)$ $\text{Recall (f3)} = c3 / (a3 + b3 + c3 + d3)$
 $\text{Precision (e4)} = d4 / (d1 + d2 + d3 + d4)$ $\text{Recall (f4)} = d4 / (a4 + b4 + c4 + d4)$
 $\text{Precision Macro} = (e1 + e2 + e3 + e4) / 4$ $\text{Recall Macro} = (f1 + f2 + f3 + f4) / 4$
 $\text{F-score Macro} = 2 \cdot \text{Precision Macro} \cdot \text{Recall Macro} / (\text{Precision Macro} + \text{Recall Macro})$
 $\text{Accuracy} = (a1 + b2 + c3 + d4) / 26$

Supplemental Fig. 3 The prediction accuracy, F-score_{Macro}, Precision_{Macro}, and Recall_{Macro} rates of the test data set in each four models at the 250-th epoch with regularization weight “1.00” .



Supplemental Fig. 4 DNA methylation rate of the genes that have the iPSC Pos-ESC Neg sites
DNA methylation rate of *DNAH9* (a), *FAM19A5* (b), *TMEM132C* (c), and *TMEM132D* (d) genes loci and sodium bisulfite sequencing analysis. (Top) Upper and lower graphs show DNA methylation rates in iPS cells 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.



Supplemental Fig. 5 Analysis of the high weight CpG sites

(a) DNA methylation rate of the top 6-10th high weight CpG sites in the iPSC Pos-ESC Neg sites. (b) DNA methylation rate of the top 6-10th high weight CpG sites in the ESC Pos-iPSC Neg sites. (c) 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.