



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

의학박사 학위논문

An epigenomic roadmap to  
induced pluripotency

– DNA methylation as a reprogramming modulator –

유도만능 줄기세포로의 역분화 과정 중 후성  
유전체의 변화에 관한 연구

–DNA 메틸화의 역분화 조절인자로서의 역할–

2015년 2월

서울대학교 대학원  
의과학과 의과학 전공  
이 동 성

유도만능 줄기세포로의 역분화 과정 중  
후성 유전체의 변화에 관한 연구  
-DNA 메틸화의 역분화 조절인자로서의 역할-

지도교수 서 정 선

이 논문을 의학박사 학위논문으로 제출함  
2014년 10월

서울대학교 대학원  
의과학과 의과학전공  
이 동 성

이동성의 의학박사 학위논문을 인준함  
2015년 1월

위원장	金孝洙	(인)
부위원장	李 弼 瑞	(인)
위원	金 鍾 侑	(인)
위원	吳 一 煥	(인)
위원	崔 棻 林	(인)

# An epigenomic roadmap to induced pluripotency

– DNA methylation as a reprogramming modulator –

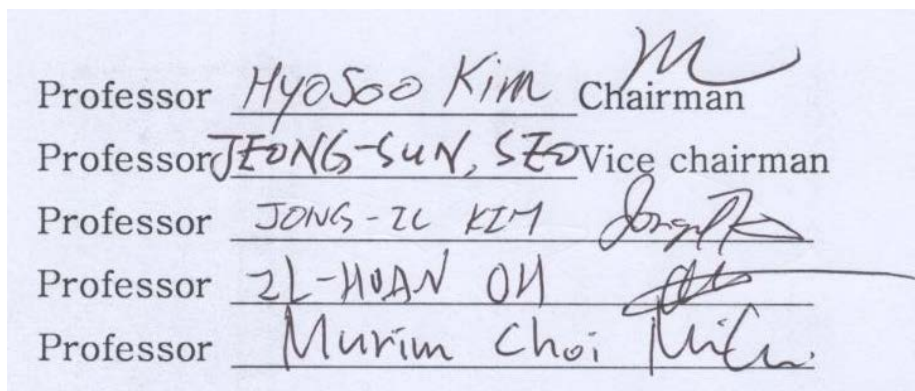
by


Dong-Sung Lee

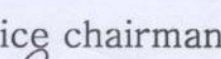
A thesis submitted to the Department of  
Biomedical Science in partial fulfillment of the  
requirement of the Degree of Doctor of Philosophy  
in Medical Science at Seoul National University  
College of Medicine

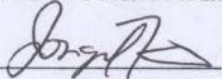
December 2014

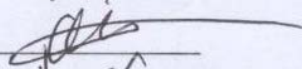
Approved by Thesis Committee:

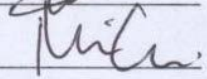


Professor Hyosoo Kim Chairman 

Professor JEONG-SUN, SEO Vice chairman 

Professor JONG-IL KIM 

Professor IL-HWAN OH 

Professor Murim Choi 

# ABSTRACT

## An epigenomic roadmap to induced pluripotency

- DNA methylation as a reprogramming modulator-

Dong-Sung Lee

Major in Biomedical Science

Department of Biomedical Science

Seoul National University Graduate School

**Introduction:** During cellular reprogramming to induced pluripotent stem cells (iPSCs), somatic cells rebuild their epigenetic architecture to acquire a steady self-renewing state. The biological significance and mechanisms of this epigenetic remodeling have remained unclear.

**Methods:** Here we characterize the epigenomic roadmap to pluripotency at base resolution by performing whole genome bisulfite sequencing of samples from secondary reprogramming system. We investigated the changes in differentially methylated regions (DMRs) and integrated this with analysis of histone modifications.

**Results:** We observed that methylation gain in DMRs occurred gradually during reprogramming. In contrast, methylation loss in DMRs was achieved only at the transition to the ESC-like state. Supporting a prominent role for DNA methylation in reprogramming, DMRs were enriched for transcription factor binding sites (TFBSs) and histone mark H3K4me3. Cells exhibited focal DNA demethylation at the binding sites of activated reprogramming factors during high transgene expression leading to a pluripotent ‘F-class’ state. ESC-like pluripotent cells were distinguished by extension of demethylation to the wider neighborhood of these sites. Our data indicated contrasting modes of control for genes with CpG rich promoters, which demonstrated stable low DNA methylation and strong engagement of histone marks H3K4me3 and H3K27me3, and genes with CpG poor promoters whose repression was driven by DNA methylation. Such DNA methylation driven control is key to the expression of several ESC-pluripotency predictor genes, including *Dppa4*, *Dppa5a* and *Esrrb*.

**Conclusions:** These results reveal the crucial role that DNA methylation plays in the epigenetic switch that drives somatic cells to pluripotency.

\* This work is published in Nature Communications (1).

---

**Keywords:** induced pluripotent stem cell; embryonic stem cell; epigenomics; DNA methylation; histone modification; transcription factor binding site

**Student number:** 2010-21914

# CONTENTS

<b>Abstract .....</b>	<b>i</b>
<b>Contents .....</b>	<b>ii</b>
<b>List of tables .....</b>	<b>iii</b>
<b>List of figures .....</b>	<b>iv</b>
<b>List of Abbreviations .....</b>	<b>v</b>
<b>Introduction .....</b>	<b>1</b>
<b>Material and Methods .....</b>	<b>4</b>
<b>Results.....</b>	<b>20</b>
Dynamic changes in DNA methylation during reprogramming .....	20
TFBSs and histone modification are enriched in the DMRs .....	22
Dynamic changes of TFBS methylation during reprogramming...	23
Demethylation leads to precise control of gene expression.....	25
<b>Discussion .....</b>	<b>50</b>
<b>References.....</b>	<b>52</b>
<b>Abstract in Korean .....</b>	<b>59</b>

## LIST OF TABLES

Table 1 MethylC-Seq data summary .....	16
Table 2 Gene separation strategy based on expression .....	17
Table 3 H3K4me3 and H3K27me3 occupancy in DMRs depend on methylation level .....	28
Table 4 Enrichment of Transcription factor binding sites in each DMR group .....	29
Table 5 Correlation coefficient between gene expression and epigenomic changes .....	30
Table 6 . Correlation coefficient between differentially expressed genes and epigenomic changes .....	31
Table 7 H3K4me3 and H3K27me3 occupancy in promoters depend on methylation level .....	32



# LIST OF FIGURES

Figure 1 Experimental and computational analysis overview of the study .....	3
Figure 2 Scheme for identifying differentially methylated regions (DMRs) .....	18
Figure 3 Scheme for identifying histone mark clusters.....	19
Figure 4 Heatmap of Differentially Methylated Regions (DMRs) ....	33
Figure 5 Base-level visualization of Differentially Methylated Regions (DMRs) .....	34
Figure 6 General features of DNA methylation on whole genome and DMRs .....	35
Figure 7 DMR accumulation during reprogramming .....	36
Figure 8 Average methylation change levels.....	37
Figure 9 Features affecting DNA methylation change during reprogramming .....	38
Figure 10 Percentage of DMRs containing H3K4me3 or H3K27me3 based on methylation level.....	39
Figure 11 Relationship between DNA methylation level and histone modification .....	40
Figure 12 Boxplots of CpG methylation change within each histone	

mark cluster .....	41
Figure 13 Histone modification and DNA methylation change at transcription factor binding sites .....	42
Figure 14 Histone modification and DNA methylation change around transcription factor binding sites .....	43
Figure 15 Relationship between histone modification and RNA expression.....	44
Figure 16 Relationship between DNA methylation and histone modification .....	45
Figure 17 Epigenetic features of gene classes.....	46
Figure 18 Epigenetic features of gene classes.....	47
Figure 19 Relationship between DNA methylation, histone modification, RNA expression, and CpG density .....	48
Figure 20 DNA methylation level in promoter of 2oMEF and engagement of ESC specific histone marks .....	49
Figure 21 A model summarizing DNA methylation and histone modification driven control of gene expression.....	51

## **LIST OF ABBREVIATIONS**

iPSC: induced pluripotent stem cell

ESC: embryonic stem cell

MEF: mouse embryonic fibroblast

Dox: doxycycline

WGBS: whole-genome bisulfite sequencing

ChIP-Seq: chromatin immunoprecipitation sequencing

NGS: next generation sequencing

DMR: differentially methylated region

TF: transcription factor

TFBS: transcription factor binding site

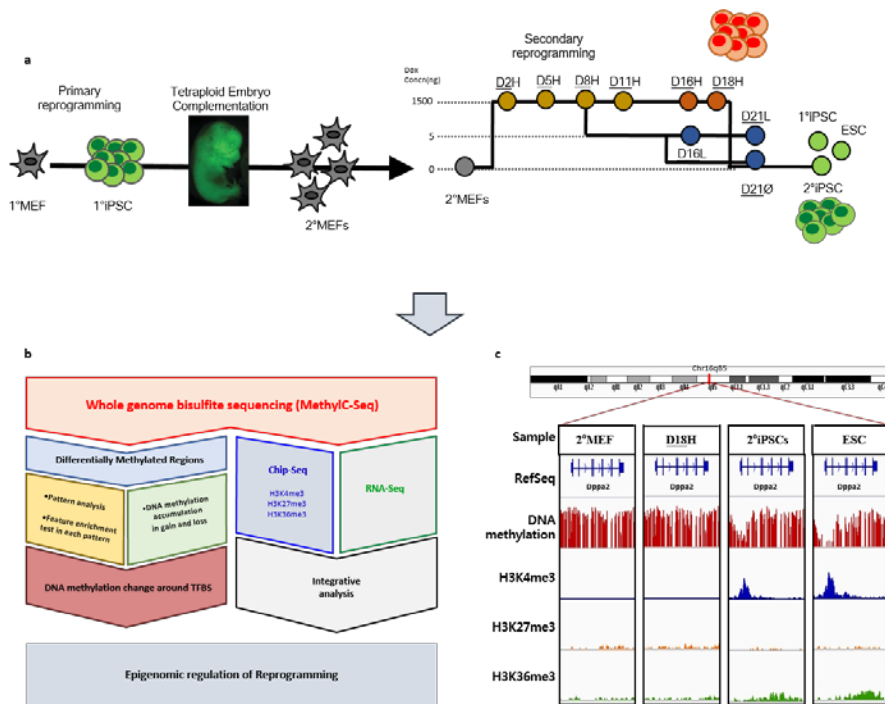
# INTRODUCTION

Somatic cells can be reprogrammed into induced pluripotent stem cells (iPSCs) by the expression of defined transcription factors (2-6). During the reprogramming process, the global epigenetic landscape has to be reset to establish the epigenetic marks of the pluripotent state through DNA methylation and chromatin remodeling processes (3, 7-10). Through the development of a secondary reprogramming system (11), iPSC generation was initially described as a multistep process characterized by transcriptional, DNA methylation and chromatin changes (12-15). Genome wide analysis of specific chromatin modification dynamics at early stages of reprogramming indicated that this process might be constrained by repressive epigenetic modifications, such as H3K9me3 and DNA methylation (16-19).

More recently, it has been proposed that DNA methylation during iPSC generation functions in the silencing of genes involved in differentiation, while also facilitating chromatin remodeling (19-21). DNA demethylation appears to play an important role in reactivating pluripotency genes, which are hypermethylated and silenced in somatic cells, particularly in the late stages of the reprogramming process (14). However, overall understanding of the global dynamics of epigenetic modification at different stages during reprogramming remains poor.

In this work, we have utilized a murine secondary reprogramming system to sample cellular trajectories during reprogramming and performed whole-genome bisulfite sequencing, ChIP-seq (H3K4me3, H3K27me3, and H3K36me3), and RNA-Seq to

characterize the epigenomic roadmap to pluripotency at base resolution (**Figs. 1a-c**) (22, 23). Our observations provide a deeper understanding of the reprogramming process and reveal the crucial role that DNA methylation plays in the epigenetic switch that drives somatic cells to pluripotency.



**Figure 1 | Experimental and computational analysis overview of the study.**

a) Establishment of secondary system and sample collection. b) MethylC-Seq was performed on samples from secondary system. Differentially methylated regions were identified. RNA-Seq and ChIP-Seq data were integrated with MethylC-Seq data based on transcripts. c) Base-level visualization of DNA methylation and histone distribution around Dppa2.

# MATERIALS AND METHODS

## 1. Cell culture and Secondary Reprogramming

ROSA26-rtTA-IRES-GFP mouse ESC, iPSCs and mouse embryonic fibroblasts (MEFs) were cultured as previously described (24). ESCs and iPSCs were cultured in 5% CO<sub>2</sub> at 37°C on irradiated MEFs in DMEM containing 15% FCS, leukemia-inhibiting factor, penicillin/streptomycin, l-glutamine, nonessential amino acids, sodium pyruvate, and 2-mercaptoethanol. 1B 1° iPS cells were aggregated with tetraploid host embryos as described (11) and MEFs established from E13.5 embryos. High doxycycline cell samples were collected at days 0, 2, 5, 8, 11, 16 and 18 (D2H, D5H, D8H, D11H, D16H and D18H). A subculture of the reprogramming cells was established from day 19 and cultured in the absence of dox, to develop a factor independent 2° iPS cell line by day 30 (2°iPSC). Low dox samples were maintained from day 8 to day 14 cells in 5ng dox. At day 14 the culture was diverged in two with some of the cells being cultured until day 21 in the absence of dox (D21Ø) and the remainder were cultured in 5 ng/mL of dox and collected at day 16 (D16L) and (D21L). Rosa26rtTA ESCs, and 1B 1° iPSCs were collected as controls.

## **2. MethylC-Seq Library Generation**

For all 13 samples (2MEF, D2H, D5H, D8H, D11H, D16H, D18H, D16L, D21L, D21Ø, 1°iPSC, 2°iPSC and rtTA ESC), five micrograms of genomic DNA was mixed with 25 ng unmethylated cl857 Sam7 Lambda DNA (Promega, Madison, WI, USA). The DNA was fragmented by sonication to 300–500 bp with a Covaris S2 system (Covaris) followed by end repair with the End-It DNA End-Repair Kit (Epicenter). Paired-end universal library adaptors provided by Illumina (Illumina) were ligated to the sonicated DNA as per manufacturer's instructions for genomic DNA library construction. Ligated products were purified with AMPure XP beads (Beckman, Brea, CA). Adaptor-ligated DNA was bisulfite treated using the EpiTect Bisulfite Kit (QIAGEN) following the manufacturer's instructions and then PCR amplified using PfuTurboC<sub>x</sub> Hotstart DNA polymerase (Agilent, Santa Clara, CA) with the following PCR conditions (2 min at 95°C, 4 cycles of 15 s at 98°C, 30 s at 60°C, 4 min at 72°C then 10 min at 72°C). The reaction products were purified using the MinElute gel purification kit (QIAGEN). The sodium bisulfite non-conversion rate was calculated as the percentage of cytosines sequenced at cytosine reference positions in the lambda genome.

## **3. ChIP Library Generation**

ChIP was carried out as described in (25). 40-150 million cells were fixed with 1% formaldehyde for 10min at room temperature, scraped and stored as pellets



(-80°C). Samples were lysed at 20 million cells/mL Farnham lysis buffer for 10min and subsequently at 10 million cells/mL nuclear lysis buffer. The released chromatin was sheared to 100-500 bp (250 bp average) on ice using a SonicsVibraCell Sonicator equipped with a 3 mm probe. For each sample, 50 µL of solubilized chromatin was used as input DNA to normalize sequencing results and the remaining chromatin was immunoprecipitated with 10 µg of H3K4me3 (ab8580) (26), 10 µg H3K27me3 (Millipore 07-449) (17) or 10 µg H3K36me3 (ab9050) (17) antibodies, separately. Antibody-chromatin complexes were pulled down with 100 µL magnetic Protein G Dynal beads (Invitrogen) and washed six times. The chromatin was then eluted, reverse cross-linked at 65°C overnight and subjected to RNaseA / proteinase K treatment. ChIP and input DNA was purified using a Qiagen Purification Column and quantified using a Quant-it dsDNA High Sensitivity Assay (Invitrogen). For ChIP sequencing, ChIP-seq libraries were prepared according to the protocols described in the Illumina ChIP-seq library preparation kit. Briefly, 50 ng of immunopurified DNA or 100 ng of genomic DNA from an input sample was end-repaired, followed by the 3' addition of a single adenosine nucleotide and ligation to universal library adapters. Ligated material was separated on a 2.0% agarose gel, followed by the excision of a 250–350-bp fragment and column purification (QIAGEN). DNA libraries were prepared by PCR amplification (18 cycles).

#### **4. High-Throughput Sequencing**

MethylC-Seq DNA and ChIP DNA libraries were sequenced using the Illumina HiSeq 2000 as per manufacturer's instructions. Sequencing of libraries was performed up to 2x 101cycles. Image analysis and base calling were performed with the standard Illumina pipeline version RTA 2.8.0

#### **5. Processing and alignment of MethylC-Seq data**

MethylC-Seq sequencing data was processed using the Illumina analysis pipeline and FastQ format reads were aligned to the NCBI37/mm9 mouse reference using the Bismark/Bowtie alignment algorithm (19, 27-30). Paired-read MethylC-Seq sequences produced by the Illumina pipeline in FastQ format were trimmed with trim threshold 1500, we removed the last 2 bases from sequences that were not trimmed, and removed 3 bases from sequences that were trimmed. The Bismark package version 0.7.7 was used as the aligner using the following parameters: -e 90 -n 2 -l 32 -X 550. As up to six independent libraries from each biological replicate were sequenced, we first removed duplicate reads. Subsequently, the reads from all libraries of a particular sample were combined. Unique read alignments were then subjected to post-processing. The number of calls for each base at every reference sequence position and on each strand was calculated. All results of aligning a read to both the Watson and Crick converted genome sequences were combined. The CpG methylation levels were calculated using bisulfite conversion rates by  $(\text{Number of not converted Cs} / \text{read depth})$  for each

position (**Table 1**).

## **6. RNA-Seq Library Generation and Sequencing**

Total RNA was subjected to two rounds of on column DNaseI treatment to remove contaminating DNA using the RNase-Free DNase set (Qiagen PN 79254) as per manufacturer's protocol. The total RNA was then analyzed using Agilent RNA 6000 Nano Kit (PN 5067-1511) on the Agilent Bioanalyzer 2100 (PN G2939AA) to quantify yield, qualify integrity and confirm removal of DNA contamination.

Following DNaseI treatment, 5ug total RNA from each sample was depleted of Ribosomal RNA using the Ribo-Zero<sup>TM</sup>rRNA Removal Kit (Epicenter PN RZH110424) as per manufacturer's instructions. The Ribosomal depleted RNA were then run on an Agilent RNA 6000 Pico Kit (PN 5067-1513) on the Agilent Bioanalyzer 2100 to confirm Ribosomal RNA depletion. Sequencing libraries were generated from the Ribosomal depleted RNA using the SOLiD<sup>TM</sup> Transcriptome Multiplexing Kit (PN 4427046) from Applied Biosystems following the manufacturer's publication. Final libraries were quantified and qualified using Agilent High Sensitivity DNA Kit (PN 5067-4626) on the Agilent Bioanalyzer 2100.

Sequencing libraries were subsequently pooled in equimolar ratios (four libraries per pool) and clonally amplified onto SOLiD Nanobeads. Clonal amplification was completed via emulsion PCR using the SOLiD EZ Bead System (PN

4448419, 4448418 and 4448420) coupled with SOLiD EZ Bead N200 amplification reagents (PN 4467267, 4457185, 4467281, 4467283, 4467282). Following emulsion PCR clonally amplified Nanobeads were enriched using the SOLiD EZ Bead Enricher Kits (PN 4467276, 4444140, 4453073) before being deposited into SOLiD™ 6-Lane FlowChip (PN 4461826) using the SOLiD Flowchip Deposition Kit v2 (PN 4468081) as per the manufacturers recommendations.

In total two flowchips were sequenced yielding a total of 8 lanes of data; with sequencing reads generated using the SOLiD 5500xl platform generating paired 75bp forward and 35bp reverse reads. To allow de-convolution of the pooled libraries a single 5bp index read was generated. A total of 1,204,676,394 fragments (2,409,352,788 reads) were generated post de-convolution, ranging from 35,714,748 to 147,282,580 fragments per library.

## **7. Processing and alignment of RNA-Seq data**

Sequence mapping was performed using Applied Biosystems LifeScope v2.5 whole transcriptome (paired-end) analysis pipeline against the NCBI37 (mm9) genome and exon-junction libraries constructed from the Ensembl v64 gene model. Briefly, this pipeline first removes potential contaminant reads by aligning to a filter set containing rRNA, tRNA, adaptor sequences and retrotransposon sequences. Following filtering, LifeScope then aligns all reads

to the genome and F3 reads to the junction library. F5 reads are additionally aligned at a higher sensitivity to exonic sequences within insert size distance from the paired (F3) read alignment. Read alignments are merged and disambiguated, and a single BAM (Binary Alignment/Mapped) file output per library.

BAM files were then additionally filtered to remove reads with a mapping quality (MAPQ)  $< 9$ , and all mitochondrial reads. Alignments were then assembled using Cufflinks (v2.0.2) using the  $-G$  parameter to quantify gene and isoform FPKM expression values against the reference gene model (Ensembl v67).

## **8. Identification of methylated cytosines**

At each reference cytosine the binomial distribution was used to identify whether at least a subset of the genomes within the sample were methylated, using a 0.01 FDR corrected P-value. We identified methyl cytosines while keeping the number of false positive methylcytosine calls below 1% of the total number of methyl cytosines we identified. The probability  $p$  in the binomial distribution  $B(n,p)$  was estimated from the number of cytosine bases sequenced in reference cytosine positions in the unmethylated Lambda genome (referred to as the error rate: non-conversion plus sequencing error frequency). We interrogated the sequenced bases at each reference cytosine position one at a time, where read depth refers to the number of reads covering that position. For each position, the

number of trials (n) in the binomial distribution was the read depth. For each possible value of n we calculated the number of cytosines sequenced (k) at which the probability of sequencing k cytosines out of n trials with an error rate of p was less than the value M, where  $M * (\text{number of unmethylated cytosines}) < 0.01 * (\text{number of methylated cytosines})$  and if the error rate of p was over 0.01, we assumed the cytosine was not methylated. In this way, we established the minimum threshold number of cytosines sequenced at each reference cytosine position at which the position could be called as methylated, so that out of all methyl cytosines identified no more than 1% would be due to the error rate.

## 9. Calculation of DNA methylation level

If the error rate is less than 0.01 we calculated adjusted DNA methylation level for cytosine as follow:

$$\text{Adjusted cytosine methylation level} = \frac{\{a - (\frac{b}{cr})\}}{a} \quad (1)$$

(a=total Cs, b=number of converted Cs, cr=bisulfite conversion rate)

## 10. Identification of differentially methylated regions (DMRs)

DMRs were identified using a sliding window approach (**Fig. 2**). A window size of 30 CpGs less than 6kb with coverage more than 5X in 15 CpGs/window in all samples were considered, progressing 1 CpG per iteration. Total of 20,214,978 windows were assessed. Windows showing Maximum difference and fold

enrichment of 30% and 4-fold with Benjamini-Hochberg corrected FDR from anova-test P values of less than 1% were identified as differentially methylated windows. 188,529 differentially methylated windows were then joined if regions were overlapped or progressing region and the succeeding regions were covering more than 60% of the region.

DMRs were then defined as Hyper-DMRs and Hypo-DMRs if average methylation level difference of each DMR in each sample was higher or lower by more than 20% relative to 2MEF.

## **11. Mapping and enrichment analysis of ChIP-Seq reads**

Paired-end ChIP-Seq data was processed using the Illumina analysis pipeline and mapping was conducted using Bowtie version 0.12.8 with the following parameters: `--pairtries 100 -y -k 1 -n 3 -l 50 -I 0 -X 1000`. Enrichment analysis was conducted using MACS (31) with parameters of `--nomodel -S -w -n -space 30`.

## **12. ChIP-Seq data analysis**

Enriched peaks from ChIP-Seq data were joined into clusters where at least one sample has a peak for each modification (H3K4me3, H3K27me3, and H3K36me3) (**Fig. 3**). The total peak width of each sample within cluster was calculated as histone mark score within clusters.

### **13. TFBS epigenomic change analysis**

Transcription factor binding sites (ESC-TFBSs) of mouse ESCs were obtained from different studies (32-34). CpG methylation level of each transcription factor binding site in each sample was calculated. The average CpG methylation change of each transcription factor binding site was then calculated in each sample relative to 2MEF. For calculating CpG methylation change around ESC-TFBSs, the same procedure was applied for 200 bp 400 bins around each ESC-TFBS. The same procedure using enrichment score for 30 bp window was applied for calculating average histone modification change.

### **14. Genome annotation**

Genomic regions and CpG islands were defined based on NCBI37/mm9 coordinates downloaded from the UCSC website (<http://genome.ucsc.edu/>). Promoters were arbitrarily defined as 5 kb upstream and 1 kb downstream of transcriptional start site for each Ensembl release-67 transcript. Gene bodies are defined as from transcription start to end site for each transcript. Histone modification clusters and DMRs were annotated if they overlap with their promoters.



## **15. Fold enrichment test**

Fold enrichment was calculated as follows: (Observed number of X in examining region/total length of examining region (bp))/(Total number of X in reference region/reference region length(bp)), X=genomic feature).

## **16. Gene expression pattern separation**

We selected genes of expression patterns as described in Table 2.

## **17. Data integration and normalization**

DNA methylation levels of promoters were calculated from 5kb upstream and 1kb downstream of transcription start site. H3K4me3 and H3K27me3 marks were considered if their cluster of peaks were overlapped with promoters. Overlapped H3K36me3 peaks were calculated for whole gene. For calculating normalized histone modification scores, maximum peak width was considered as 1 and relative widths were calculated for each sample in each gene.

## **18. Accession codes**

Methylome sequencing data is available under the European Nucleotide Archive accessions #ERP004116 (<http://www.ebi.ac.uk/ena/data/view/PRJEB4795>).

Long RNA seq and Chip-seq sequencing data are available under the NCBI Sequence Read Archive (SRA) accessions #SRP046744

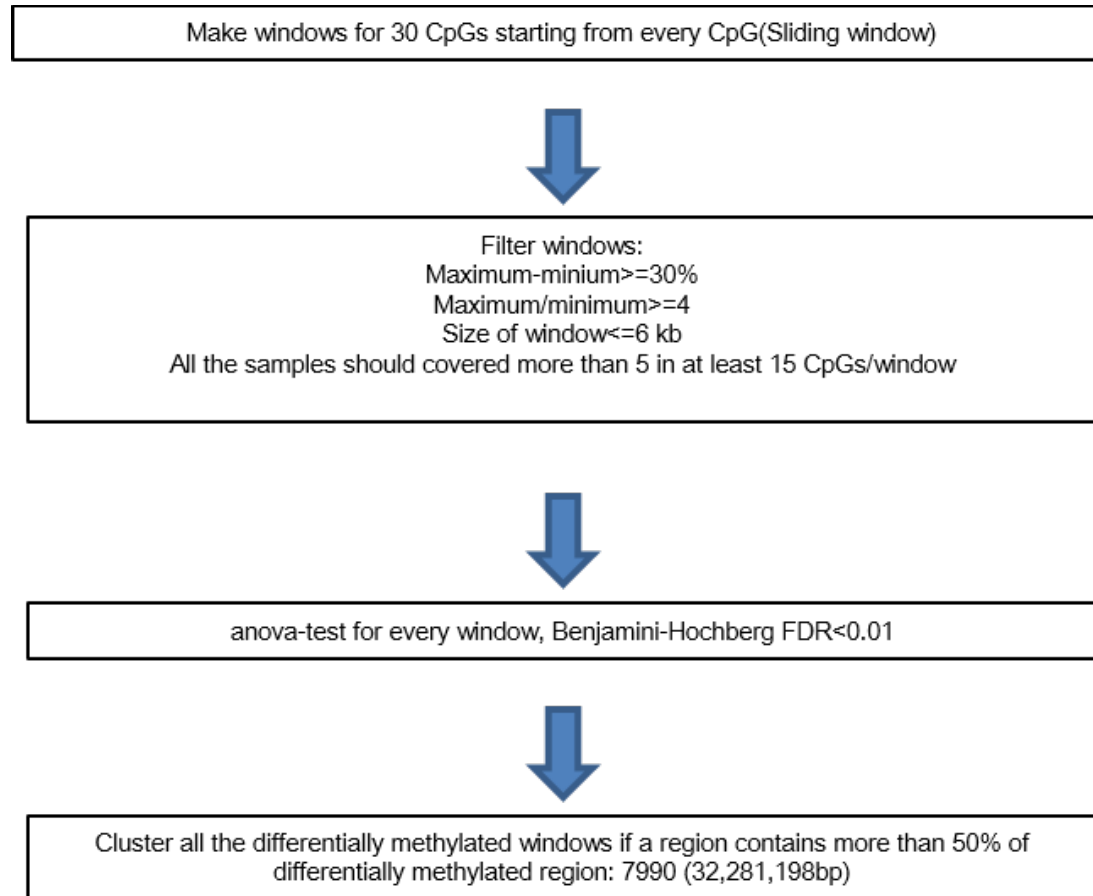
(<http://www.ncbi.nlm.nih.gov/sra>). Analyzed data sets can be obtained from Stemformatics ([www.stemformatics.org](http://www.stemformatics.org)) (35).

**Table 1.** MethylC-Seq data summary

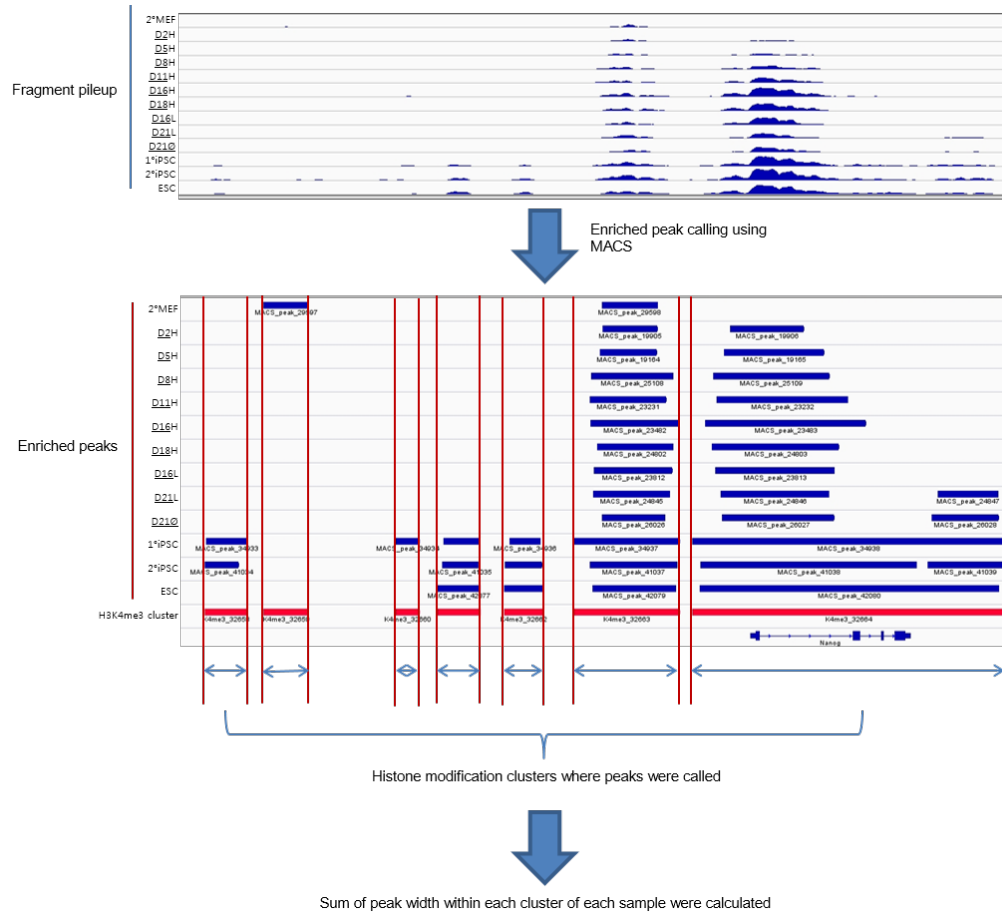
<b>Sample ID</b>	<b>Total methylated C in CpG context</b>	<b>Total methylated C in CHG context</b>	<b>Total methylated C in CHH context</b>	<b>Total C to T conversions in CpG context</b>	<b>Total C to T conversions in CHG context</b>	<b>Total C to T conversions in CHH context</b>
<u>2°MEF</u>	439825482	4726039	37437558	201039072	3009142302	9458047813
<u>D2H</u>	480932349	4635253	42464217	311503846	3825228622	12651058898
<u>D5H</u>	444287068	4309734	40434576	275768157	3395441370	11119623360
<u>D8H</u>	545699906	5170538	48456632	246548329	3845896311	12014237257
<u>D11H</u>	476600931	4994705	45849467	300699423	3787302481	12421114608
<u>D16H</u>	471032795	4899535	45009484	286204989	3835321604	12653401981
<u>D18H</u>	473155217	5994978	48782843	304714369	3843619363	12549976502
<u>D16L</u>	552223868	6124752	44208260	366468101	4528759821	14804418392
<u>D21L</u>	634523160	24789039	73090494	341893392	4824511488	15616979331
<u>D21Ø</u>	693833566	32013168	93729438	287264670	4861045022	15960060823
1°iPSC	1033694646	58775778	163737386	478008184	7287844374	24115979815
2°iPSC	420611840	24162827	64044181	175740365	2843232561	9159533662
ESC	414609494	20867163	66918909	198997550	2925328531	9365529172

**Table 2.** Gene separation strategy based on expression

			Sample_context(X_FPKM<=1.5, O_FPKM>=5)					Number of genes
			2°MEF	<u>D16H</u>	<u>D18H</u>	2°iPSC	ESC	
ESC like	Activation	in F-class and ESC-like cells (1a)	X	O	O	O	O	87
		Only in ESC-like cells (1b,1c)	X	X	X	O	O	93
	Repression	in F-class and ESC-like cells (2a, 2b)	O	X	X	X	X	221
		Only in ESC-like cells	O	O	O	X	X	14
Fail	Activation		X	X	X	X	O	47
	Repression		O	O	O	O	X	9
F-class specific	Activation (3a)		X	O	O	X	X	41
	Repression (3b)		O	X	X	O	O	35



**Figure 2 | Scheme for identifying differentially methylated**



**Figure 3 | Scheme for identifying histone mark clusters**

# RESULTS

## Dynamic changes in DNA methylation during reprogramming

The Project Grandiose secondary reprogramming samples present a unique opportunity to profile cellular state changes at various time points during reprogramming (11, 22, 23). These consisted of secondary fibroblasts (2MEF), six intermediate time points at high doxycycline (dox) concentrations (D2H, D5H, D8H, D11H, D16H and D18H), three alternative intermediate time points collected for samples treated with reduced dox concentrations (D16L, D21L, D21Ø), the secondary iPSCs (2°iPSCs), the primary iPSCs (1°iPSCs) used to generate the chimeric mouse, and a mouse Rosa rtTA embryonic stem cell line (ESC) for standard comparison (**Fig. 1a-c**). As described by Tonge et al., these samples showed reprogramming to two distinct pluripotent states: ESC-like cells and the “F-class” consisting of stages D16H and D18H (22).

In this manuscript, we describe base-resolution bisulfite sequencing of the 13 Project Grandiose samples and investigation of global DNA methylation changes during reprogramming (**Table 1**). The sample methylomes were scanned using a sliding window of 30 CpGs, identifying 7,890 differentially methylated (DMRs) covering 22 Mb, representing 0.81% of the mouse genome (**Figs. 4-6**). Unsupervised hierarchical clustering performed on the DNA methylation state of DMRs (**Fig. 4**) distinguished the intermediate states (D2H-D18H, and D16L-D21L) from the ESC-like pluripotent states (D21Ø, 1°iPSCs,

2iPSCs and ESCs). DMRs were categorized into 3 groups based on the changing pattern of DNA methylation (**Fig. 4**). The DMR-1 group exhibited increased methylation levels after (DMR-1a) or during (DMR-1b) high level reprogramming factor expression and included genes related to development and cell differentiation, such as the Hox family, *Col25a1*, and *Meox2*. The DMR-2 group represented differential methylation changes between two pluripotent states: either gradual demethylation to F-class and methylation in the ESC-like state (DMR-2a) or gradual methylation to F-class and acquired demethylation in the ESC-like state (DMR-2b). A final group (DMR-3) was identified as exhibiting low methylation levels in the ESC-like state (1iPSCs, 2iPSCs and ESCs), with stable methylation persisting in the F-class state and intermediate reprogramming samples, which included multiple pluripotency genes such as *Dppa2*, *Dppa4*, *Dppa5a*, *Esrrb*, *Tcl1*, and *Eras* (**Fig. 4**).

We annotated the DMRs in each sample as Hyper- or Hypo-DMRs where they differed from a corresponding 2MEF baseline by over 20% (**Fig. 7**). We observed a widespread gradual increase in methylation to generate Hyper-DMRs during reprogramming, whereas limited demethylation was observed as cells reprogrammed to the F-class state (D16H and D18H). The steady increase in Hyper-DMRs during both high-dox and low-dox reprogramming challenges the notion that most changes in DNA methylation occur at a late stage when cells acquire stable pluripotency (14). A similar trend was observed for the average methylation level of DMRs, as methylation occurred gradually while demethylation did not change significantly during transgene expression (**Fig.**



**7,8a,b**). Almost all Hypo-DMRs found in iPSCs were also observed in ESCs (98.94%), but this was not the case for Hyper-DMRs (61.88%), suggesting that demethylation during reprogramming occurred more conservatively.

### **TFBSs and histone modification are enriched in the DMRs**

To assay the distribution of histone marks, we performed ChIP-Seq for H3K4me3, H3K27me3 and H3K36me3 (see Methods). We determined the distribution and enrichment of these histone marks within DMRs, as well as other genomic features including ESC-TFBSs from published data (32-34, 36). Notably, we found that 98% of DMRs contained H3K4me3 clusters and 68% contained ESC-TFBSs (**Fig. 9a**). When we assessed enrichment of each feature relative to the whole genome, H3K4me3 marks, ESC-TFBSs, CpG islands, CpG shores, and enhancers showed more than 10-fold enrichment, followed by promoters, and H3K27me3 clusters. (**Fig. 9b**).

Our finding that histone marks were highly enriched within DMRs led us to explore the relationship between DNA methylation levels and H3K4me3/H3K27me3 marks within DMRs (**Fig.s 10,11, Table 3**). DMRs exhibiting low level methylation (less than 30%) were frequently associated (96.9%) with H3K4me3 and H3K27me3. In contrast, the absence of both histone marks was most frequently associated (79.7%) with DMRs with high levels of methylation ( $\geq 70\%$ ), supporting the inverse relationship between DNA methylation and these two histone modifications. Furthermore, CpGs inside H3K4me3 and H3K27me3 marks exhibit more methylation change, in

comparison to CpGs inside H3K36me3 mark (**Fig. 12**).

To investigate the involvement of ESC-TFBSs in reprogramming, we performed separate enrichment analysis for each DMR group (**Table 4**). Polycomb repressive complex (PRC) binding sites, including SUZ12, EZH2, and RING1B, were enriched in DMR-1 and DMR-2b. On the other hand, sequence specific pluripotency-associated ESC-TFBSs such as Nanog, Oct4 and Klf4 (but not CTCF and TET1) binding sites were enriched in DMR-3, the group of DMRs that are demethylated only in the ESC-like state. These results demonstrate the dynamic changes in DNA methylation at TFBSs, and the connection between the pattern of changes and TFBS enrichment.

### **Dynamic changes of TFBS methylation during reprogramming**

Interrogating methylation changes at ESC-TFBSs resulted in the detection of methylation depletion during high-dox treatment, which was not apparent by examining DMRs (**Fig. 13; Methods**). This was most obvious at the binding sites for activated or over-expressed transcription factors during early time points, such as OCT4, SOX2, KLF4, and NANOG. These TFBSs also accumulated H3K4me3 modifications proceeding after the methylation depletion. H3K27me3 marks diminished at binding sites of expressed transcription factors early in reprogramming. In contrast, ESC-TFBSs for genes that were not activated during high-dox reprogramming but are known to play critical roles in ESC-like pluripotent state, such as ESRRB and TCF2L1(15, 37, 38), showed no change in DNA methylation and were demethylated only in

the ESC-like state. The PRC (SUZ12 and EZH2) binding sites underwent a gain of DNA methylation during reprogramming but showed baseline levels of methylation in ESC.

We assessed DNA methylation changes occurring within  $\pm 40$ kb of ESC-TFBSs (**Fig. 14**). At the binding sites of core ESC-pluripotency transcription factors, (OCT4, SOX2, KLF4, and NANOG), we observed rapid focal demethylation during high-dox treatment (D2H-D18H) if the factors were expressed. On the other hand, ESC-like cells (1<sup>o</sup>iPSC, 2<sup>o</sup>iPSC, ESC) exhibited extensive demethylation, up to 20 kb distal from the binding sites. A similar but more delayed process was also observed for H3K4me3 modifications. The broad neighborhoods around PRC binding sites were hyper-methylated in all samples examined. Interestingly, although methylation accumulated broadly around PRC (SUZ12, EZH2, RING1B) binding sites (**Fig. 14**), these underwent focal renormalization at the ESC-like pluripotent state. These sites also demonstrate bivalent marks of H3K4me3 and H3K27me3 in ESC-like state (33). The patterns of change to DNA methylation and histone marks were distinct for the three types of transcription factor shown (**Figs. 13-14**). Our results show an interesting contrast between the focal demethylation induced early in reprogramming and broader demethylated regions at ESC-like pluripotent state, perhaps representing a key distinguishing feature of the pluripotent state where broader demethylation is required for completion of the reprogramming to ESC-like state.

We attempted to show that the dynamics of methylation change at transcription factor binding sites could act as a predictor of importance to the reprogramming process. We proposed criteria for DNA-binding transcription factors of  $>1.2X$  enrichment and  $>10\%$  overlap in DMR-3, implying over-representation in DMRs that underwent demethylation at transition to the ESC-like state, but little change early in reprogramming. We tested a set of 118 transcription factors with computationally predicted binding sites against these criteria (39, 40). We found only three transcription factors (SOX2, MYC, and OCT4) that fulfilled our criteria, all of which are known to be important in reprogramming to iPSCs. This suggests a high specificity for the prediction criteria, although sensitivity is low as other factors known to be involved in reprogramming were not identified. Transcription factors whose binding sites show significant change in methylation late in a transition can be called important to that transition with high confidence. We believe that methylome-based tests of this nature could have useful application in prediction of transcription factors involved in other cellular transitions.

### **Demethylation leads to precise control of gene expression**

We integrated corresponding RNA expression data (23) with our DNA methylation and histone modification datasets (**Tables 5-7; Methods**). Activation of genes was associated with H3K4me3 occupancy in promoter regions, and repression was associated with either H3K27me3 occupancy or no histone mark (**Fig. 15**). Moreover, as we observed in DMRs, engagement of

both H3K4me3 and H3K27me3 marks in promoters was dependent on DNA methylation levels with a strong inverse relationship (**Fig. 16**).

We selected 477 genes segregating into 7 clusters on the basis of expression and epigenetic change over the course of reprogramming (**Fig. 17-18, Table 2; Methods**). These groups represent: activated early in reprogramming (Expr-1a), activated late in reprogramming with either low- (Expr-1b) or full- (Expr-1c) DNA methylation in 2°MEF, and repressed during reprogramming with either low- (Expr-2a) or full- (Expr-2b) DNA methylation in ESC. Genes in Expr-3a were turned on while those in Expr-3b were turned off in high-dox, therefore they were differentially expressed between D16H/D18H (F-class cells) and ESC-like cells. Expression changes of genes in Expr-1a and Expr-2a/b are likely responsible for pluripotency, as they were differentially expressed between 2°MEF and pluripotent cells (22). Finally, the presence of genes in Expr-1b/c explains why F-class cells are distinct from ESC-like state cells.

The expression dynamics through reprogramming of these genes was clear upon visualization of the categories and representative genes from each class (**Fig. 15-19**). Genes repressed by H3K27me3 with low methylated promoters in 2°MEF tended to be activated early in reprogramming and had CpG-rich promoters (Expr-1a/b). These loci were enriched in genes involved in cell-adhesion, such as *Epcam* and *Cdh1* (**Fig. 17**, Expr-1a). In contrast, quiescence of Expr-1c genes was initially safeguarded by DNA methylation of CpG-poor

promoters, and H3K4me3 was only acquired after late demethylation. The same two modes of control were observed for the genes repressed by reprogramming. However, as in the analysis of DMRs, DNA methylation in promoter regions happened early in reprogramming (Expr-2b) whereas demethylation was detected exclusively in the ESC-like state, revealing that a gain of methylation is kinetically favored over demethylation. This is also true for histone marks in relation to changes in gene expression, where histone modifications, specifically the modulation of H3K27me3, occurred early during reprogramming (Expr-2a) within low methylated promoters. Interestingly, the dynamic process of histone modification alterations during reprogramming was strongly influenced by the starting methylation state of gene promoters (**Fig. 20**). Genes with low-methylated promoters at 2°MEF showed a significantly higher rate of transition to the ESC-like state for both ESC-specific histone marks compared to those with fully-methylated promoters. This suggests that DNA methylation presents a major barrier during somatic cell reprogramming to ESC-like cells and that the methylation status of a given region determines its control by histone modifications.

**Table 3.** H3K4me3 and H3K27me3 occupancy in DMRs depend on methylation level

		<b>2°MEF</b>	<b>D2H</b>	<b>D5H</b>	<b>D8H</b>	<b>D11H</b>	<b>D16H</b>	<b>D18H</b>	<b>D16L</b>	<b>D21L</b>	<b>D21Ø</b>	<b>1°iPSC</b>	<b>2°iPSC</b>	<b>ESC</b>
DMRs methylation % <=0.3	K4me3 only	2055 (45.9%)	2143 (45.0%)	2170 (49.9%)	2214 (55.3%)	2025 (48.2%)	1534 (43.7%)	1524 (45.7%)	1264 (45.0%)	540 (43.4%)	852 (48.6%)	1322 (48.9%)	1569 (51.6%)	3544 (62.1%)
	Both K4/K27me3	415 (9.3%)	475 (10.0%)	343 (7.9%)	232 (5.8%)	339 (8.1%)	371 (10.6%)	362 (10.9%)	350 (12.5%)	17 (1.4%)	3 (0.2%)	0 (0.0%)	0 (0.0%)	3 (0.1%)
	K27me3 only	1910 (42.7%)	1846 (38.8%)	1572 (36.2%)	1409 (35.2%)	1650 (39.3%)	1494 (42.5%)	1281 (38.4%)	1110 (39.5%)	659 (53.0%)	890 (50.8%)	1378 (50.9%)	1472 (48.4%)	2137 (37.5%)
	no K4me3 or K27me3	93 (2.1%)	295 (6.2%)	260 (6.0%)	152 (3.8%)	187 (4.5%)	115 (3.3%)	165 (5.0%)	83 (3.0%)	28 (2.3%)	8 (0.5%)	5 (0.2%)	2 (0.1%)	20 (0.4%)
	Total	4473	4759	4345	4007	4201	3514	3332	2807	1244	1753	2705	3043	5704
DMRs methylation % >=0.7	K4me3 only	90 (8.3%)	67 (7.1%)	78 (7.8%)	107 (10.2%)	74 (7.8%)	87 (8.4%)	95 (9.1%)	113 (10.3%)	447 (45.9%)	470 (40.2%)	94 (17.8%)	197 (37.6%)	29 (8.7%)
	Both K4/K27me3	3 (0.3%)	2 (0.2%)	0 (0.0%)	3 (0.3%)	2 (0.2%)	4 (0.4%)	5 (0.5%)	9 (0.8%)	8 (0.8%)	16 (1.4%)	18 (3.4%)	23 (4.4%)	7 (2.1%)
	K27me3 only	28 (2.6%)	13 (1.4%)	12 (1.2%)	10 (1.0%)	11 (1.2%)	17 (1.6%)	14 (1.3%)	21 (1.9%)	12 (1.2%)	50 (4.3%)	64 (12.1%)	49 (9.4%)	32 (9.6%)
	no K4me3 or K27me3	965 (88.9%)	868 (91.4%)	904 (90.9%)	931 (88.6%)	856 (90.8%)	925 (89.5%)	932 (89.1%)	953 (87.0%)	507 (52.1%)	634 (54.2%)	353 (66.7%)	255 (48.7%)	266 (79.6%)
	Total	1086	950	994	1051	943	1033	1046	1096	974	1170	529	524	334

**Table 4** Enrichment of Transcription factor binding sites in each DMR group

DMR Groups	DMR Number	sequence-specific transcription factors														Transcription regulators				
		TET1	CTCF	Oct4	SOX2	NANOG	ESRRB	ZFX	KLF4	cMYC	nMYC	E2F1	TCF2L1	SMAD1	STAT3	p300	EZH2	SUZ12	RING1B	
<b>DMR-1</b>	<b>1a</b>	1819	NE	NE	-	-	-	-	-	-	-	-	-	-	-	-	-	NE	NE	NE
	<b>1b</b>	1453	NE	NE	-	-	-	-	-	-	-	-	-	-	-	-	-	++	++	+
<b>DMR-2</b>	<b>2a</b>	553	NE	NE	-	NE	NE	NE	NE	++	NE	NE	NE	NE	NE	NE	-	-	-	-
	<b>2b</b>	1291	+	NE	NE	NE	NE	+	++	NE	NE	+	NE	+	-	NE	+	+++	+++	+++
<b>DMR-3</b>		2774	NE	NE	+++	+++	+++	++	++	++	+++	+++	+++	+	+++	+++	+++	-	-	-
TFBS Enrichment																				
of	Total DMRs	7890	13.25	3.84	17.26	15.99	17.26	18.65	13.25	18.65	13.25	17.26	18.65	13.25	17.26	13.25	17.26	18.65	13.25	17.26
vs Whole genome																				
<b>Fold enrichment vs Total DMRs: - &lt; 0.75X ≤ NE (Not enriched) ≤ 1.25X ≤ + &lt; 1.5X ≤ ++ &lt; 1.75X ≤ +++</b>																				



**Table 5.** Correlation coefficient between gene expression and epigenomic changes

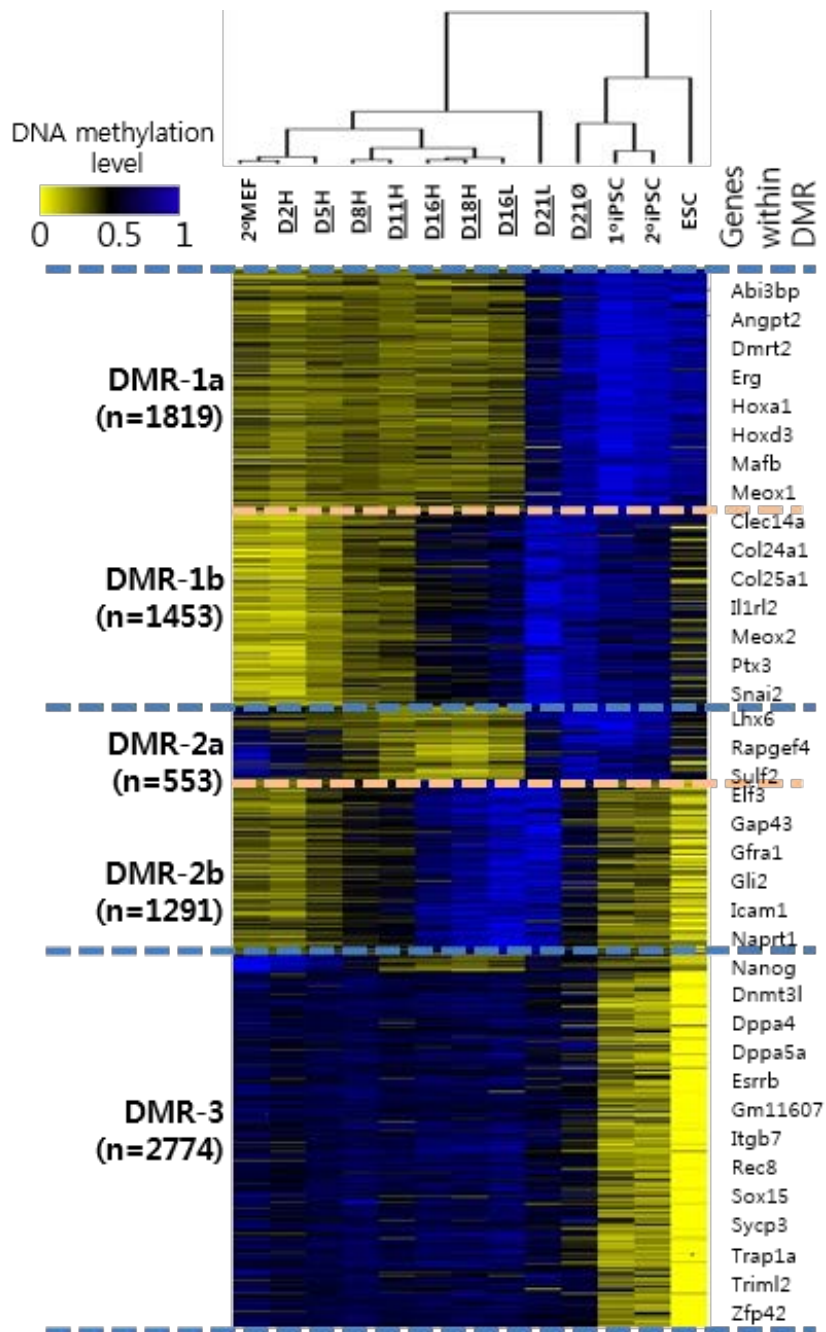
<b>Relationship with Total Gene(n=37412)</b>	<b>H3K4me3</b>	<b>DMR</b>	<b>H3K27me3</b>	<b>H3K36me3</b>	<b>Promoter CpG methylation</b>	<b>Gene body CpG methylation</b>
<b>Containing number</b>	20816	4320	13497	14198	37412	37412
<b>Average correlation</b>	0.25	-0.25	-0.14	0.23	-0.12	-0.06
<b>Strong correlation (R&gt;=0.5)</b>	6226	206	511	3682	1525	2104
<b>Strong anti-correlation (R&lt;=-0.5)</b>	414	1298	1691	226	5081	3838

**Table 6.** Correlation coefficient between differentially expressed genes and epigenomic changes

<b>Relationship with DEG(n=547)</b>	<b>H3K4me3</b>	<b>DMR</b>	<b>H3K27me3</b>	<b>H3K36me3</b>	<b>CpG methylation</b>	<b>Gene body CpG methylation</b>
<b>Containing number</b>	438	180	315	283	547	547
<b>Average correlation</b>	0.57	-0.40	-0.29	0.54	-0.13	-0.07
<b>Strong correlation (R&gt;=0.5)</b>	300	7	14	165	48	39
<b>Strong anti-correlation (R&lt;=-0.5)</b>	5	83	102	2	119	94

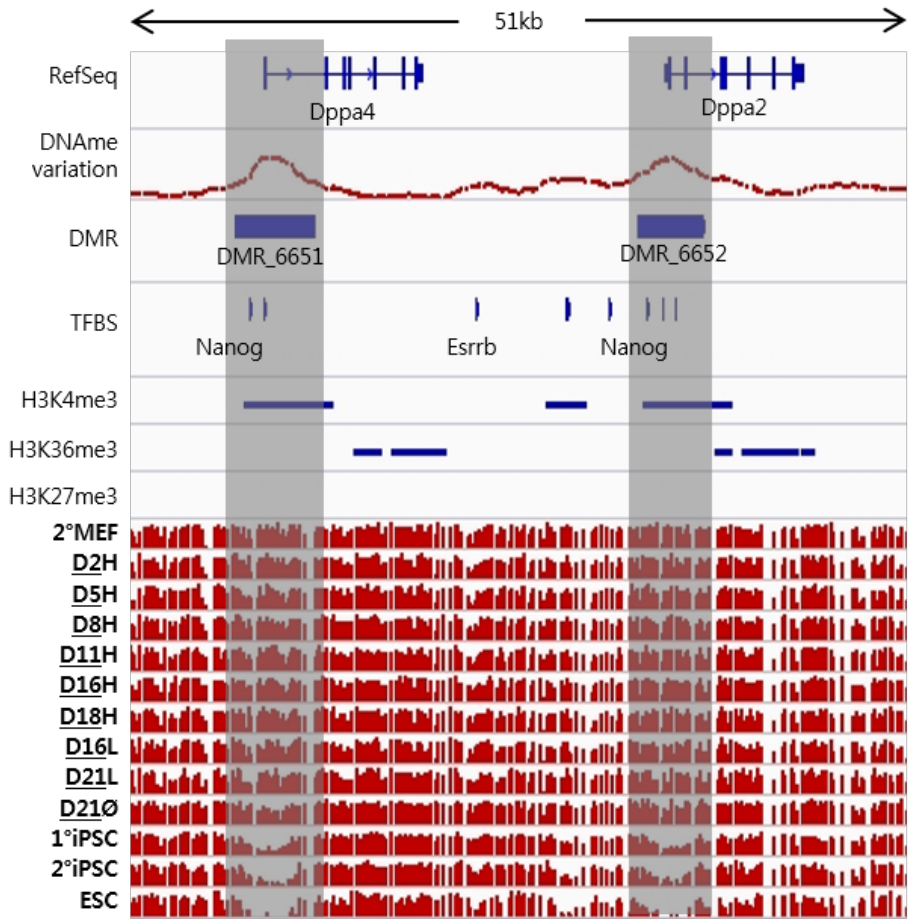
**Table 7.** H3K4me3 and H3K27me3 occupancy in promoters depend on methylation level

		<b>2°MEF</b>	<b>D2H</b>	<b>D5H</b>	<b>D8H</b>	<b>D11H</b>	<b>D16H</b>	<b>D18H</b>	<b>D16L</b>	<b>D21L</b>	<b>D21Ø</b>	<b>1°iPSC</b>	<b>2°iPSC</b>	<b>ESC</b>
Gene Promoters methylation % <=0.3	K4me3 only	6026 (62.6%)	6736 (66.2%)	7144 (72.4%)	7143 (74.7%)	7128 (69.6%)	6471 (63.7%)	6699 (65.4%)	6444 (64.9%)	5929 (67.3%)	5756 (69.6%)	5535 (62.0%)	5517 (60.6%)	6809 (64.6%)
	Both K4/K27me3	3351 (34.8%)	3047 (29.9%)	2434 (24.7%)	2154 (22.5%)	2679 (26.2%)	3120 (30.7%)	2897 (28.3%)	2798 (28.2%)	2551 (28.9%)	2374 (28.7%)	3312 (37.1%)	3498 (38.4%)	3614 (34.3%)
	K27me3 only	105 (1.1%)	190 (1.9%)	103 (1.0%)	69 (0.7%)	77 (0.8%)	99 (1.0%)	99 (1.0%)	83 (0.8%)	5 (0.1%)	6 (0.1%)	5 (0.1%)	7 (0.1%)	6 (0.1%)
	no K4me3 or K27me3	143 (1.5%)	202 (2.0%)	190 (1.9%)	190 (2.0%)	352 (3.4%)	464 (4.6%)	554 (5.4%)	610 (6.1%)	330 (3.7%)	138 (1.7%)	71 (0.8%)	76 (0.8%)	111 (1.1%)
	Total	9625	10175	9871	9556	10236	10154	10249	9935	8815	8274	8923	9098	10540
	<hr/>		<hr/>											
Gene Promoters methylation % >=0.7	K4me3 only	1231 (9%)	754 (7%)	801 (7%)	1156 (10%)	821 (9%)	877 (10%)	884 (11%)	961 (11%)	1307 (11%)	1628 (10%)	1642 (9%)	2005 (11%)	1209 (8%)
	Both K4/K27me3	109 (1%)	39 (0%)	48 (0%)	49 (0%)	41 (0%)	49 (1%)	33 (0%)	52 (1%)	119 (1%)	308 (2%)	362 (2%)	434 (2%)	211 (1%)
	K27me3 only	1428 (11%)	452 (4%)	486 (4%)	334 (3%)	203 (2%)	236 (3%)	127 (2%)	262 (3%)	440 (4%)	1587 (9%)	1059 (6%)	1075 (6%)	555 (4%)
	no K4me3 or K27me3	10553 (79%)	9247 (88%)	9798 (88%)	10413 (87%)	7749 (88%)	7323 (86%)	7207 (87%)	7308 (85%)	9749 (84%)	13609 (79%)	15314 (83%)	14209 (80%)	12327 (86%)
	Total	13321	10492	11133	11952	8814	8485	8251	8583	11615	17132	18377	17723	14302
	<hr/>		<hr/>											

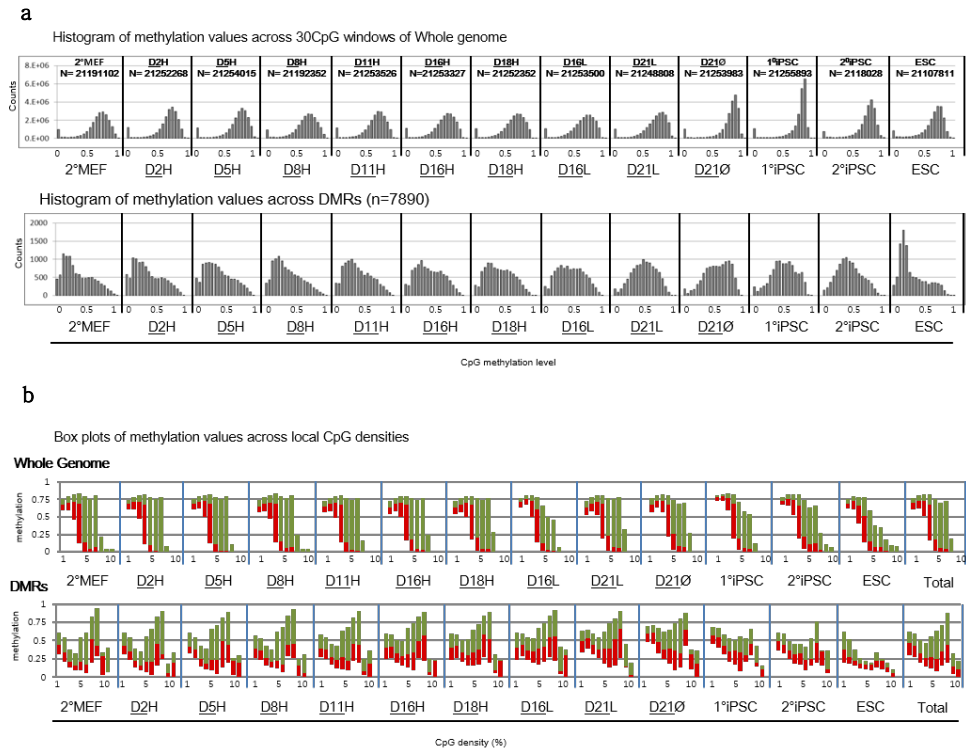


**Figure 4 | Differentially Methylated Regions (DMRs)**

Hierarchical clustering based on the DNA methylation level of DMRs in each sample. DMRs were clustered into 6 groups based on pairwise correlations.

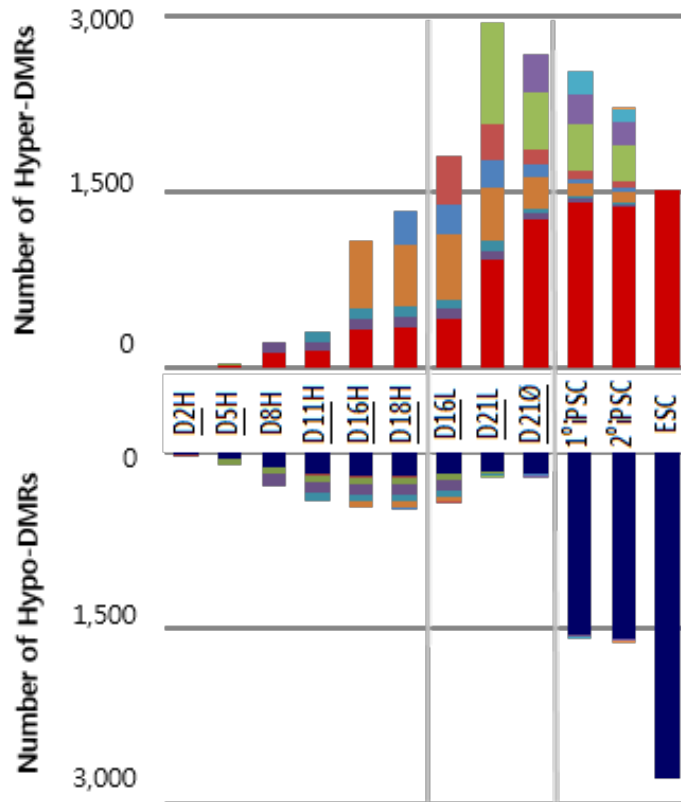


**Figure 5 | Base-level visualization of Differentially Methylated Regions (DMRs)**



**Figure 6 | General features of DNA methylation on whole genome and DMRs**

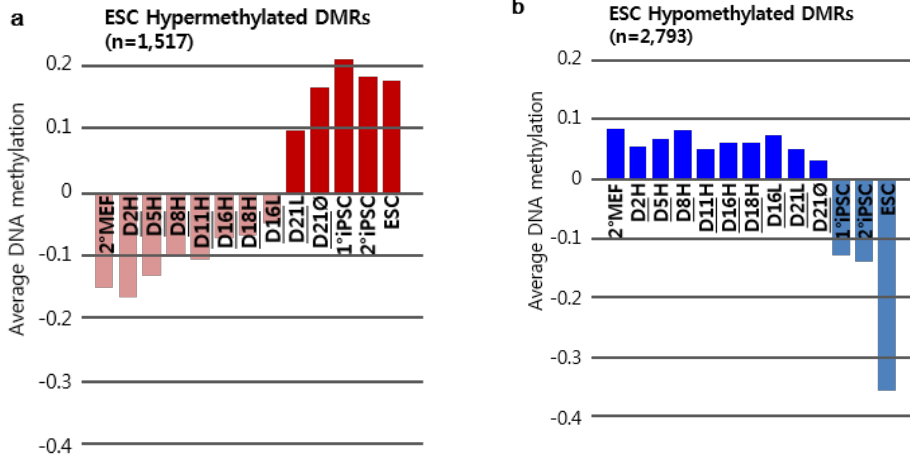
a) Histograms of methylation values across 30 CpG windows of whole genome and across DMRs for each sample. n is number of windows for each stage. b) Boxplots of methylation values across local CpG densities. Edges of green and red boxes indicate the 75<sup>th</sup> and 25<sup>th</sup> percentile, respectively. Boundary lines of red and green boxes indicate median.



**Figure 7 | DMR accumulation during reprogramming**

Dark red and dark blue bars represent ESC specific Hyper- and Hypo-DMRs.

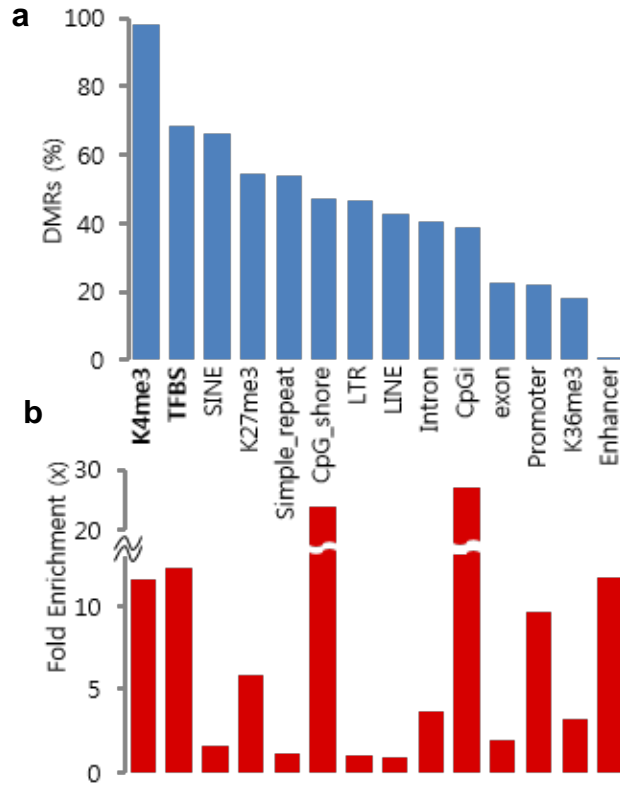
Other colors indicate Hyper- and Hypo-DMRs in the order of left to right.



**Figure 8 | General features of DNA methylation change**

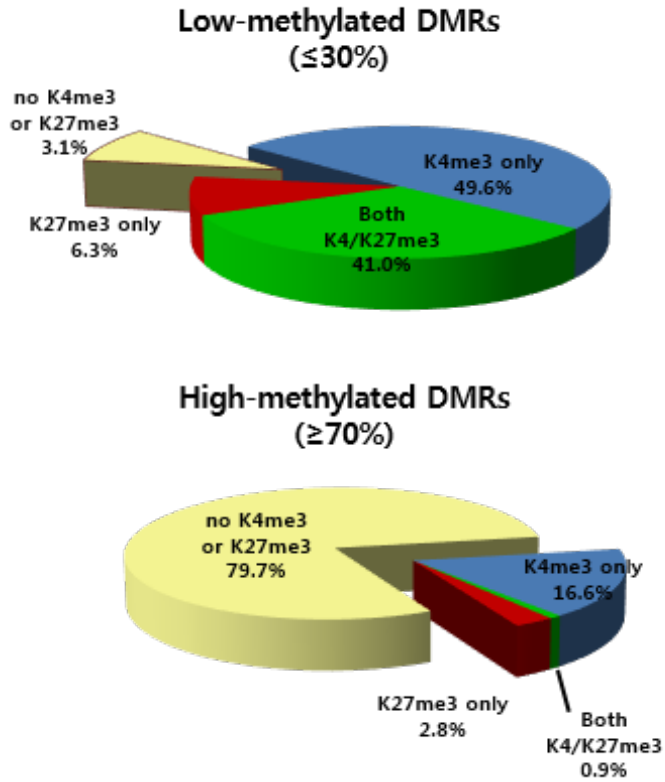
a) Average methylation levels for ESC Hyper-DMRs. b) Average methylation levels for ESC Hypo-DMRs.



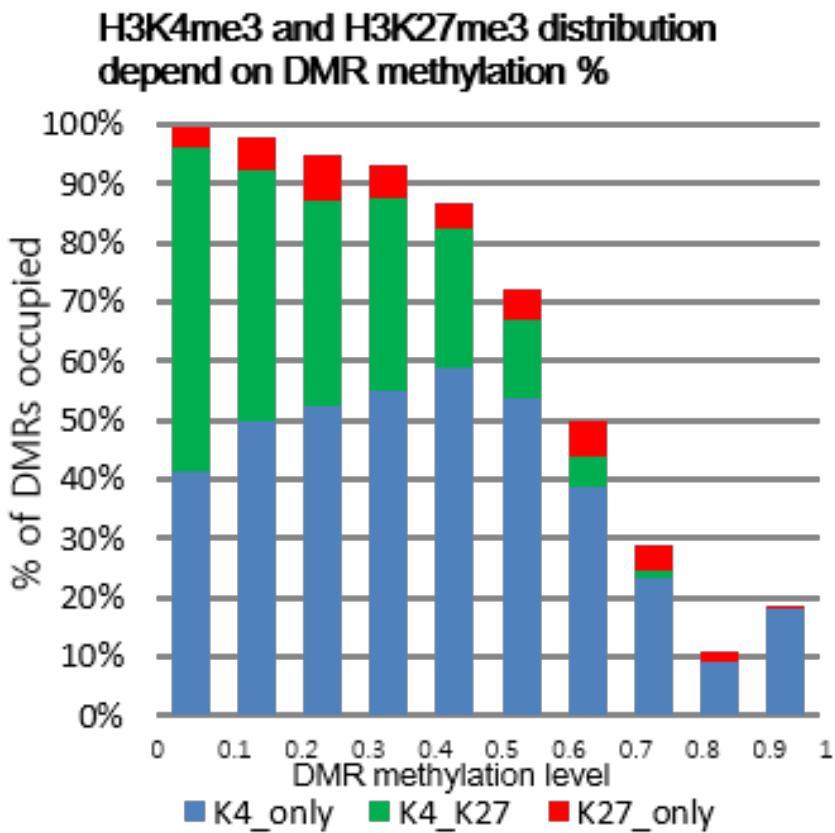


**Figure 9 | Features affecting DNA methylation change during reprogramming**

a) Proportion of DMRs containing various genomic features. b) Fold enrichment of examined genomic features within DMRs.

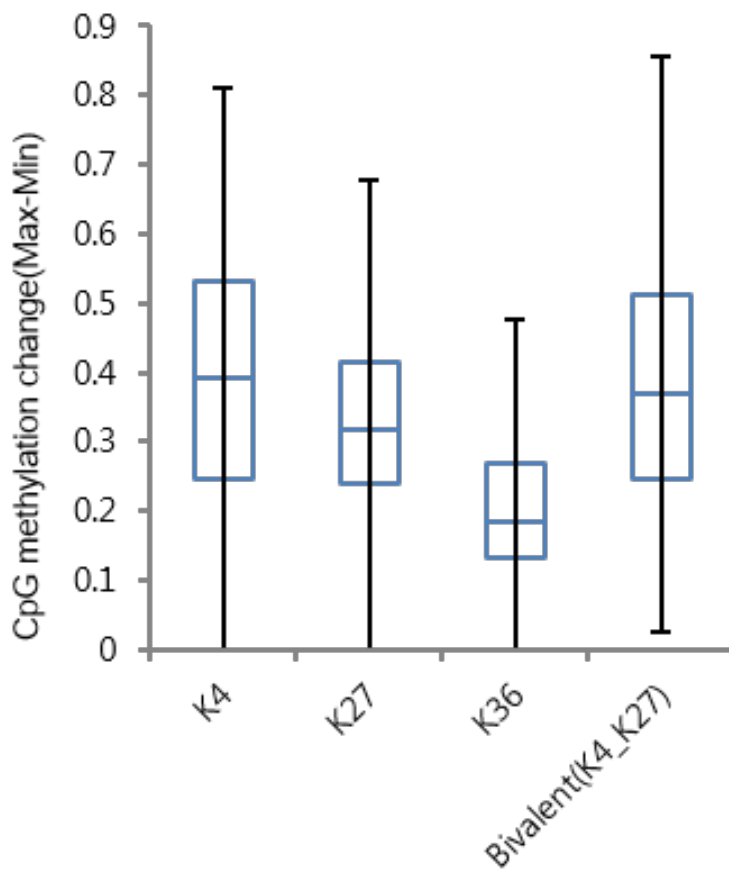


**Figure 10 | Percentage of DMRs containing H3K4me3 or H3K27me3 based on methylation level**

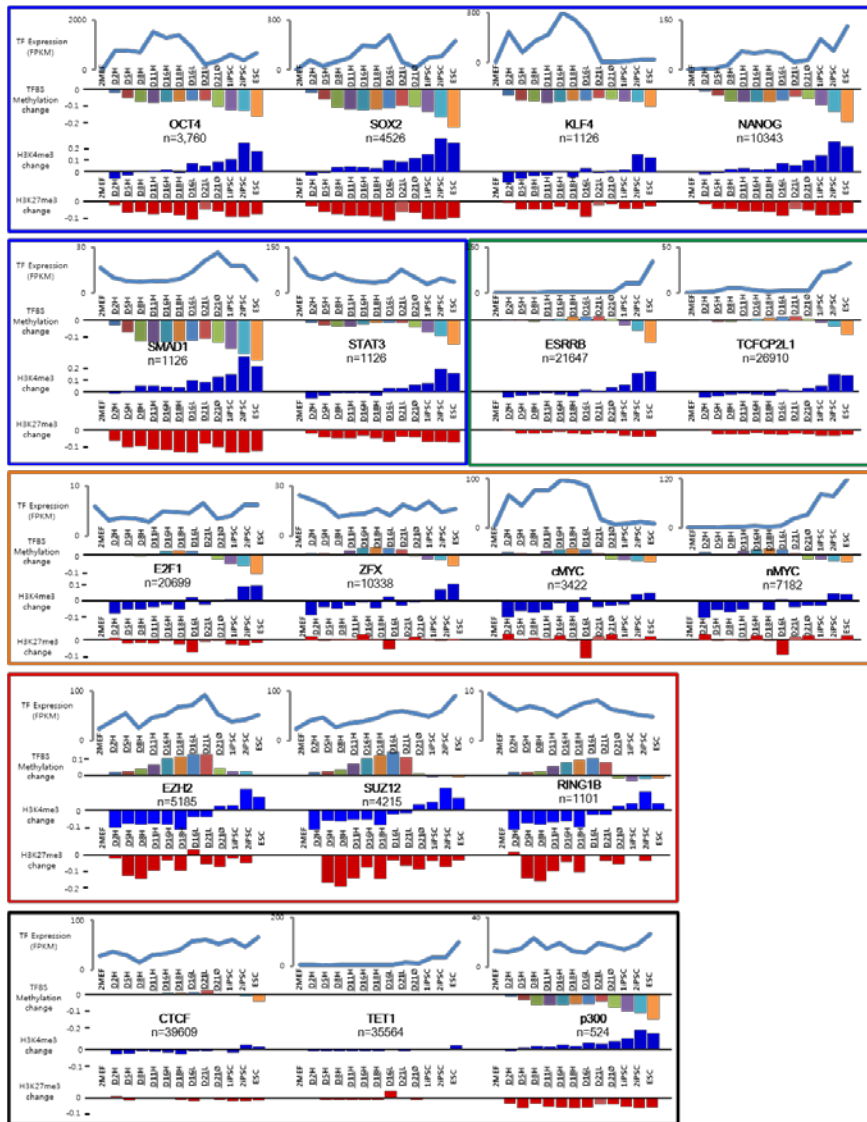


**Figure 11 | Relationship between DNA methylation level and histone modification**

H3K4me3 and H3K27me3 occupancy for each DMR methylation level in all samples.

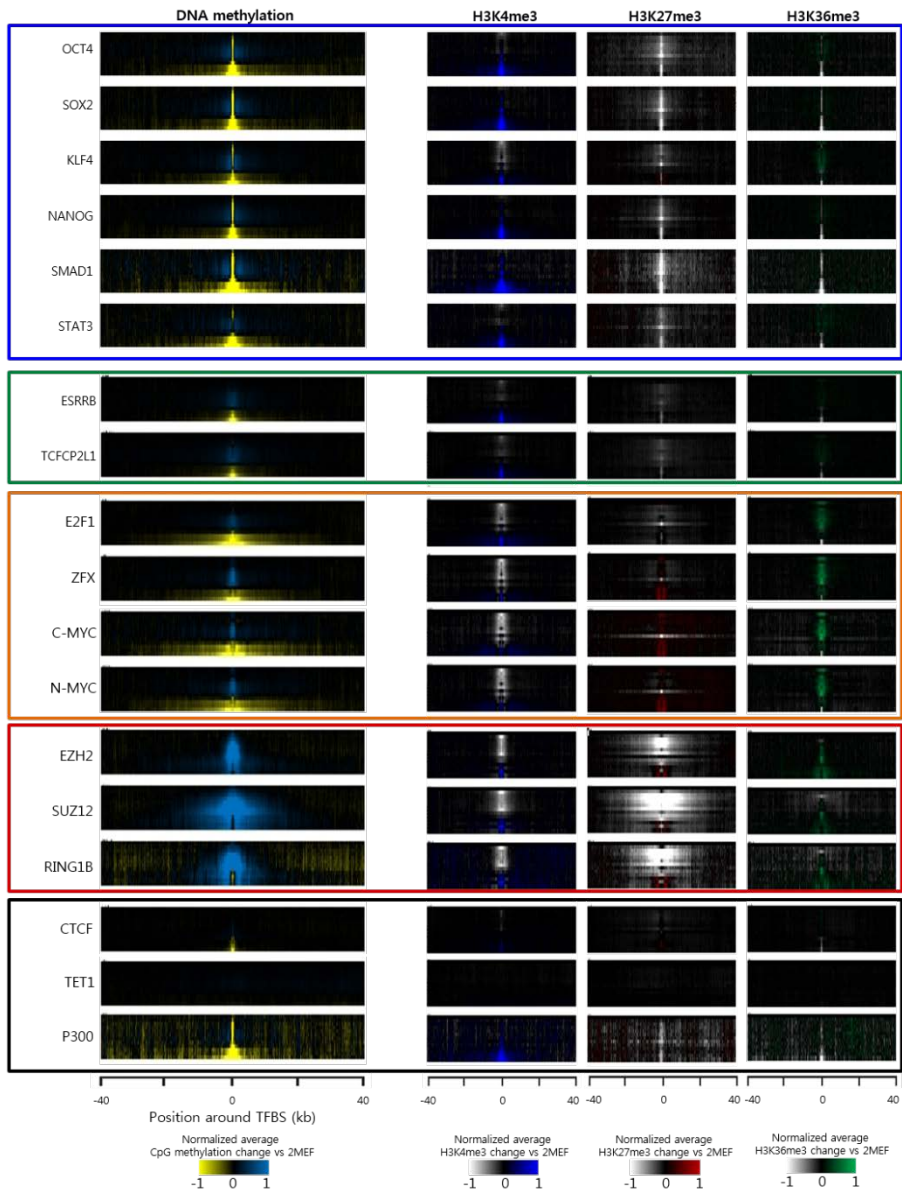


**Figure 12 | Boxplots of CpG methylation change within each histone mark cluster**



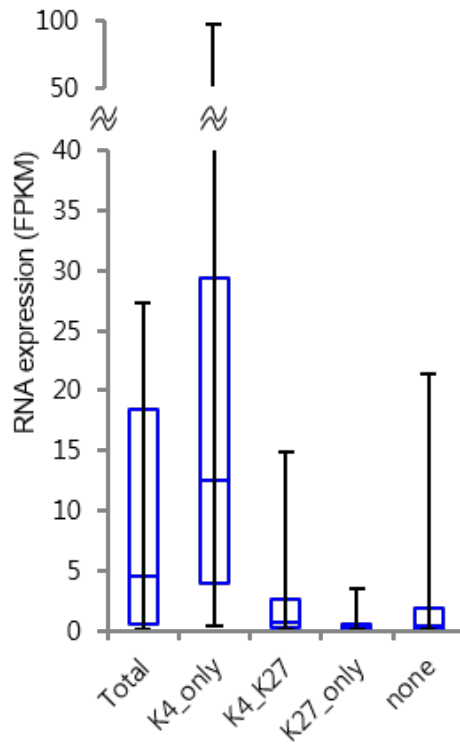
**Figure 13 | Histone modification and DNA methylation change at transcription factor binding sites**

RNA expression level (FPKM) of transcription factors (line plots), average DNA methylation change (upper bar plots), average H3K4me3 change (blue bar plots) and average H3K27me3 change (red bar plots) at binding sites of each transcription factor. Transcriptionally active genes during high-dox treatment (blue box), transcriptionally silent genes during high-dox treatment (green box), and polycomb repressive complexes (PRCs) (red box) are shown.



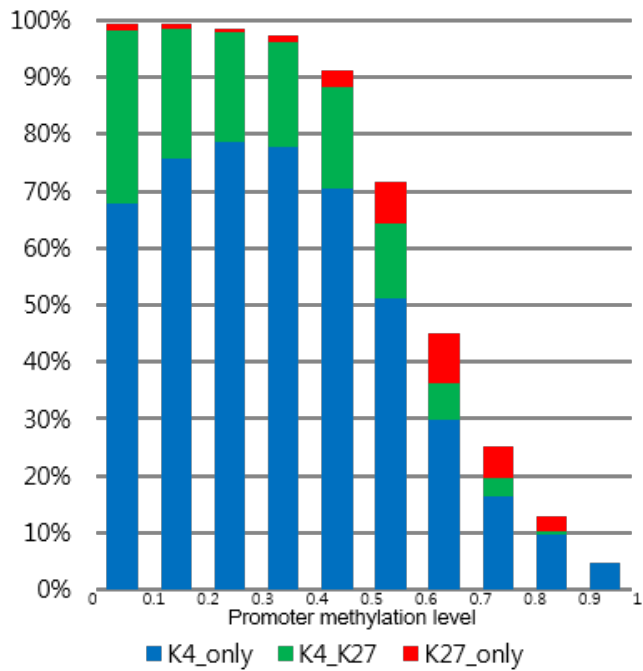
**Figure 14 | Histone modification and DNA methylation change around transcription factor binding sites**

Average DNA methylation change (left), average H3K4me3 change (middle) and average H3K27me3 change (right) in the 80 kb neighborhood of transcription factor binding sites. Grouped (coloured boxes) as in Figure 6.



**Figure 15 | Relationship between histone modification and RNA expression.**

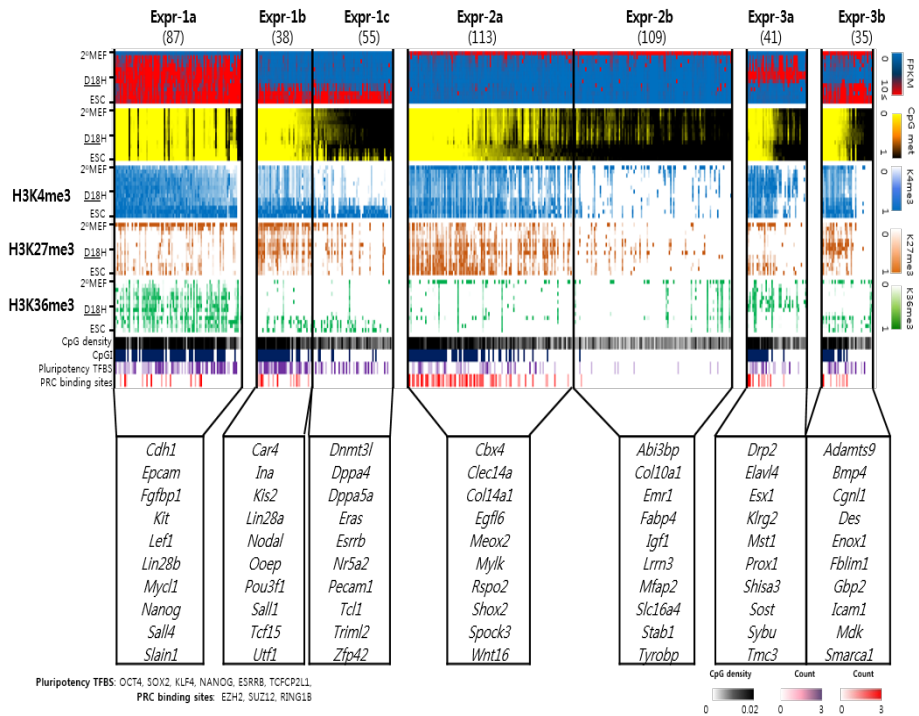
Boxplots of expression levels for genes with different histone mark occupancy in promoter regions.

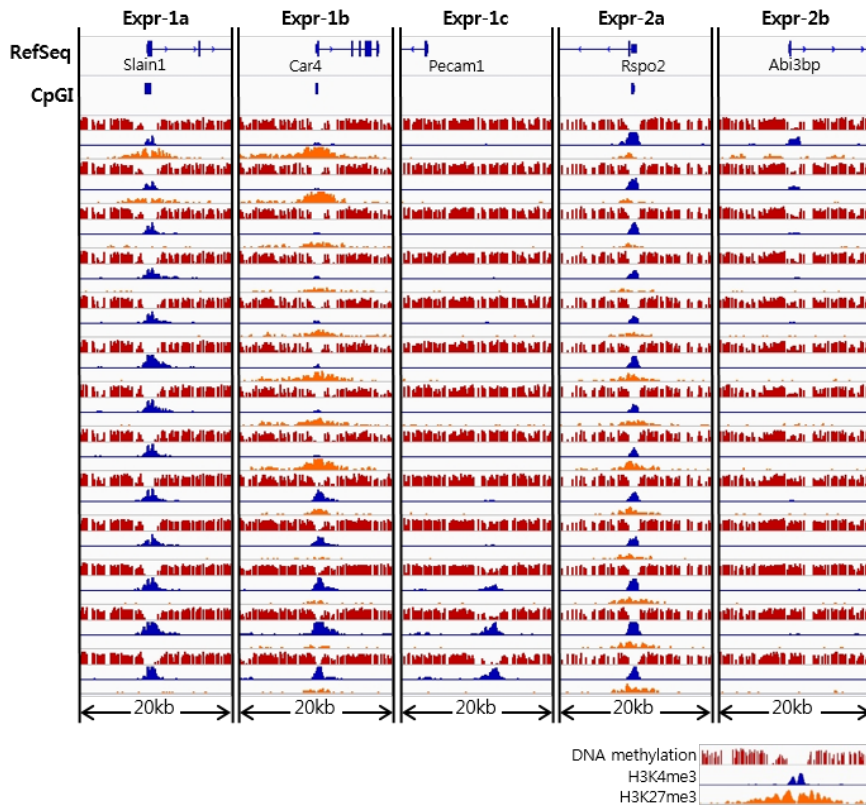


**Figure 16 | Relationship between DNA methylation and histone modification**

Occupancy of H3K4me3 and H3K27me3 marks in promoters, for each methylation level of promoters.

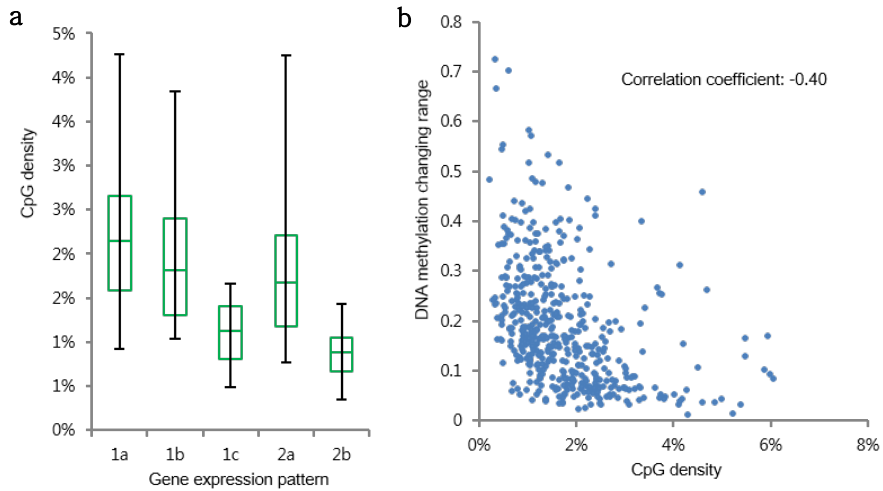






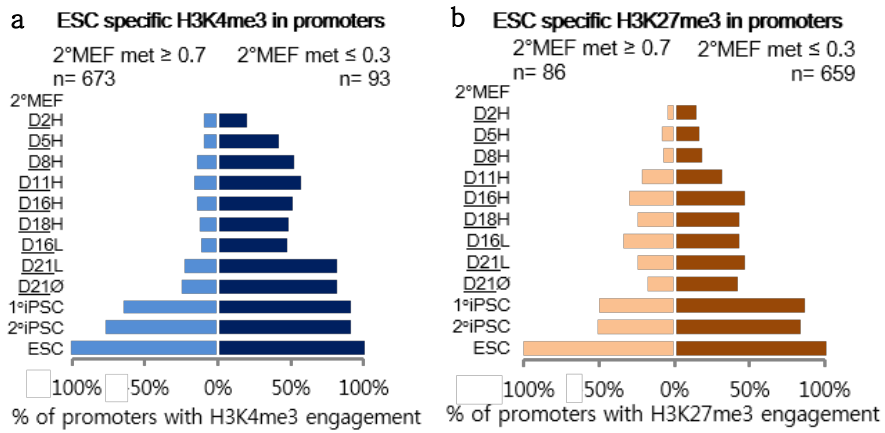
**Figure 18 | Epigenetic features of gene classes**

Base-level visualization of DNA methylation and histone modifications in the promoter regions of representative genes for each class across all samples



**Figure 19 | Relationship between DNA methylation, histone modification, RNA expression, and CpG density**

a) Boxplots of CpG density in promoters of genes in each expression group as described in Fig. 3a. b) Relationship between promoter CpG density and range of change in DNA methylation levels across samples in all genes.



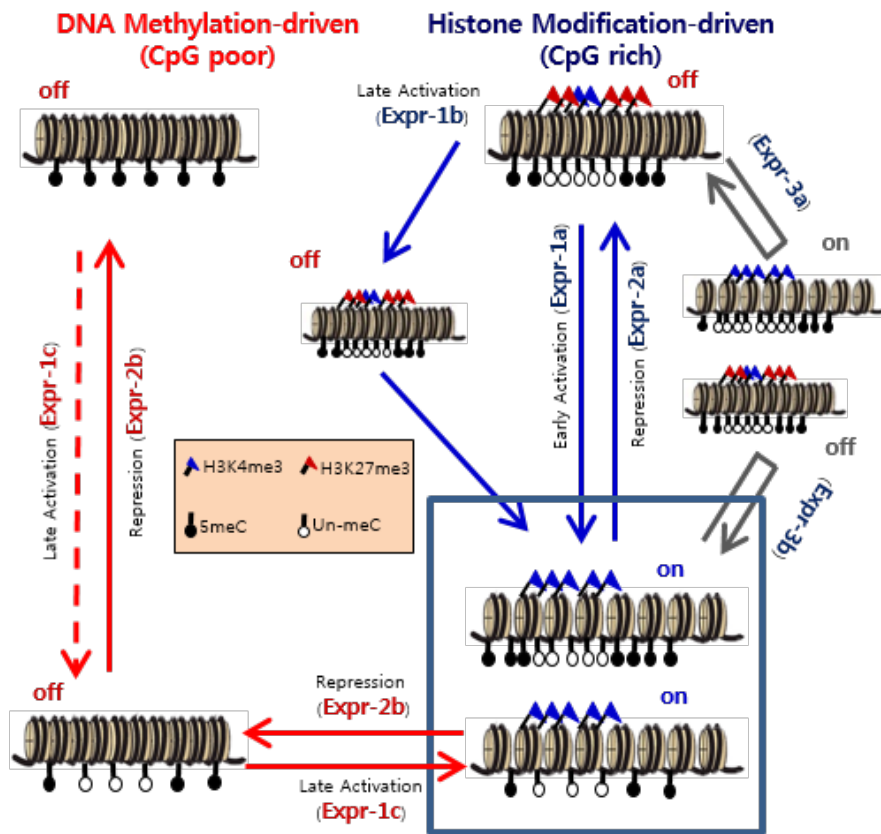
**Figure 20 | DNA methylation level in promoter of 2°MEF and engagement of ESC specific histone marks**

a) Percentage of ESC specific H3K4me3 mark for promoters with high and low initial methylation. b) Percentage of ESC specific H3K27me3 mark for promoters with high and low initial methylation.

## DISCUSSION

We propose a model that describes the key mechanism of epigenetic control of gene expression during reprogramming (**Fig. 21**). In genes with CpG-poor promoters, control is driven by DNA methylation. Such genes may be activated by demethylation followed by H3K4me3 engagement, producing expression profiles characteristic of class Expr-1c/2b. In genes with CpG-rich promoters, low methylation levels allow histone modification driven control. This model is supported by data showing the role of initial methylation status as a modulator of the dynamic changes to histone modification, and the sequential modification of DNA methylation followed by histone marks in TFBSs. The model also accounts for characteristic gene expression classes (detailed in **Figs. 17-18**). We predict that this mechanism may not only apply to iPSC reprogramming but also to lineage specification of cells. Therefore, our insights into how DNA methylation controls the epigenetic landscape in reprogramming to pluripotency could be crucial to a better understanding of the mechanisms underlying general cell fate change, and could have ramifications for stem cell based therapies.

## Epigenomic control of gene expression



**Figure 21 | A model summarizing DNA methylation and histone modification driven control of gene expression**

Dashed arrow represents the strict control of demethylation. Gene classes affected by changes are shown in brackets accompanying arrows.

## REFERENCES

1. Lee DS, Shin JY, Tonge PD, Puri CM, Lee S, Park HS, et al. An epigenomic roadmap to induced pluripotency reveals DNA methylation as a reprogramming modulator. *Nature Communications*. 2014.
2. 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.
3. Maherali N, Sridharan R, Xie W, Utikal J, Eminli S, Arnold K, et al. Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. *Cell stem cell*. 2007;1(1):55-70.
4. 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(5):861-72.
5. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, et al. Induced pluripotent stem cell lines derived from human somatic cells. *Science*. 2007;318(5858):1917-20.
6. Park IH, Zhao R, West JA, Yabuuchi A, Huo H, Ince TA, et al.

Reprogramming of human somatic cells to pluripotency with defined factors. *Nature*. 2008;451(7175):141-6.

7. Kang L, Wang J, Zhang Y, Kou Z, Gao S. iPS cells can support full-term development of tetraploid blastocyst-complemented embryos. *Cell stem cell*. 2009;5(2):135-8.

8. Onder TT, Kara N, Cherry A, Sinha AU, Zhu N, Bernt KM, et al. Chromatin-modifying enzymes as modulators of reprogramming. *Nature*. 2012;483(7391):598-602.

9. Singhal N, Graumann J, Wu G, Arauzo-Bravo MJ, Han DW, Greber B, et al. Chromatin-Remodeling Components of the BAF Complex Facilitate Reprogramming. *Cell*. 2010;141(6):943-55.

10. Zhao XY, Li W, Lv Z, Liu L, Tong M, Hai T, et al. iPS cells produce viable mice through tetraploid complementation. *Nature*. 2009;461(7260):86-90.

11. Woltjen K, Michael IP, Mohseni P, Desai R, Mileikovsky M, Hamalainen R, et al. piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature*. 2009;458(7239):766-70.

12. Samavarchi-Tehrani P, Golipour A, David L, Sung HK, Beyer TA,



Datti A, et al. Functional genomics reveals a BMP-driven mesenchymal-to-epithelial transition in the initiation of somatic cell reprogramming. *Cell stem cell*. 2010;7(1):64-77.

13. Mikkelsen TS, Hanna J, Zhang X, Ku M, Wernig M, Schorderet P, et al. Dissecting direct reprogramming through integrative genomic analysis. *Nature*. 2008;454(7200):49-55.

14. Polo JM, Anderssen E, Walsh RM, Schwarz BA, Nefzger CM, Lim SM, et al. A molecular roadmap of reprogramming somatic cells into iPS cells. *Cell*. 2012;151(7):1617-32.

15. Buganim Y, Faddah DA, Cheng AW, Itskovich E, Markoulaki S, Ganz K, et al. Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase. *Cell*. 2012;150(6):1209-22.

16. Chen J, Guo L, Zhang L, Wu H, Yang J, Liu H, et al. Vitamin C modulates TET1 function during somatic cell reprogramming. *Nature genetics*. 2013;45(12):1504-9.

17. Wang T, Chen K, Zeng X, Yang J, Wu Y, Shi X, et al. The histone demethylases Jhdm1a/1b enhance somatic cell reprogramming in a vitamin-C-dependent manner. *Cell stem cell*. 2011;9(6):575-87.

18. Plath K, Lowry WE. Progress in understanding reprogramming to the induced pluripotent state. *Nature reviews Genetics*. 2011;12(4):253-65.
19. Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, Hon G, et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*. 2011;471(7336):68-73.
20. Papp B, Plath K. Epigenetics of Reprogramming to Induced Pluripotency. *Cell*. 2013;152(6):1324-43.
21. Surani MA, Hayashi K, Hajkova P. Genetic and epigenetic regulators of pluripotency. *Cell*. 2007;128(4):747-62.
22. Tonge PD, Corso AJ, Monetti C, Hussein SMI, Puri MC, Michael IP, et al. Divergent reprogramming routes lead to alternative stem cell states *Nature*. 2014.
23. Hussein S, Puri CM, Tonge PD, Benevento M, Corso AJ, Clancy JL, et al. Genome-wide characterization of the routes to pluripotency. *Nature*. 2014.
24. Nagy A, Gertsenstein M, Vintersten K, Behringer R. *Manipulating the Mouse Embryo: A Laboratory Manual*. Cold Spring Harbor Press. 2003.
25. O'Geen H, Echipare L, Farnham PJ. Using ChIP-seq technology to

generate high-resolution profiles of histone modifications. *Methods in molecular biology*. 2011;791:265-86.

26. Gaspar-Maia A, Alajem A, Polesso F, Sridharan R, Mason MJ, Heidersbach A, et al. Chd1 regulates open chromatin and pluripotency of embryonic stem cells. *Nature*. 2009;460(7257):863-8.

27. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*. 2009;462(7271):315-22.

28. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*. 2009;10(3):R25.

29. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*. 2011;27(11):1571-2.

30. Krueger F, Kreck B, Franke A, Andrews SR. DNA methylome analysis using short bisulfite sequencing data. *Nature methods*. 2012;9(2):145-51.

31. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). *Genome biology*.

2008;9(9):R137.

32. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*. 2008;133(6):1106-17.

33. Ku M, Koche RP, Rheinbay E, Mendenhall EM, Endoh M, Mikkelsen TS, et al. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS genetics*. 2008;4(10):e1000242.

34. Wu H, D'Alessio AC, Ito S, Xia K, Wang Z, Cui K, et al. Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature*. 2011;473(7347):389-93.

35. Wells CA, Mosbergen R, Korn O, Choi J, Seidenman N, Matigian NA, et al. Stemformatics: visualisation and sharing of stem cell gene expression. *Stem cell research*. 2013;10(3):387-95.

36. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic acids research*. 2007;35(Database issue):D88-92.

37. Feng B, Jiang J, Kraus P, Ng JH, Heng JC, Chan YS, et al. Reprogramming of fibroblasts into induced pluripotent stem cells with orphan

nuclear receptor Esrrb. *Nature cell biology*. 2009;11(2):197-203.

38. Fishedick G, Klein DC, Wu G, Esch D, Hoing S, Han DW, et al. Zfp296 is a novel, pluripotent-specific reprogramming factor. *PloS one*. 2012;7(4):e34645.

39. Stormo GD. DNA binding sites: representation and discovery. *Bioinformatics*. 2000;16(1):16-23.

40. Ho Sui SJ, Mortimer JR, Arenillas DJ, Brumm J, Walsh CJ, Kennedy BP, et al. oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic acids research*. 2005;33(10):3154-64.

## 국문초록

# 유도만능 줄기세포로의 역분화 과정 중 후성 유전체의 변화에 관한 연구 -DNA 메틸화의 역분화 조절인자로서의 역할-

서울대학교 대학원 의과학과 의과학 전공

이 동 성

**서론:** 체세포에서의 유도만능 줄기세포로의 역분화 과정 중 체세포는 후성유전체의 구조를 변화시킴으로써 안정적인 자기 재생 상태를 얻는다. 하지만 이러한 후성유전체의 역분화 과정과 그 의미는 아직 분명하게 밝혀지지 않았다.

**방법:** 2 차 역분화 시스템(secondary reprogramming system)을 이용하여 이 과정 중의 샘플들을 대상으로 전장 유전체의 중아황산염 처리 염기서열 분석을 시행하였다. 이를 통하여 샘플간에 DNA 메틸화에 변화를 보이는 지역들(differentially methylated regions, DMRs)을 찾고 변화의 경향성을 보았으며 이를 히스톤 변이(histone modification) 분석과 통합하였다.

**결과:** DNA 메틸화에 변화를 보이는 지역에 있어서 메틸화의 증가는 역분화 과정 전반에 걸쳐 점차적으로 일어나는데 반해 메틸화의 감소는 배아줄기세포와 비슷한 상태로 변한 샘플들에서만 관찰 되었다. 이러한 변화를 보이는 지역들은 전사인자들이 결합하는 자리들과 히스톤 변화 중 H3K4me3 가 생기는 자리들이 풍부하게 발견 되었다. 이 중에서도 특히 역분화 과정 중 활성화된 인자들이 결합하는 자리들은 좁은 지역에 있어서의 탈메틸화가 일어나는 반면 배아줄기세포와 비슷한 상태로 변한 샘플들에서는 결합자리 주변으로 넓은 지역에서의 탈메틸화가 관찰되었다. 끝으로 우린 유전자의 발현을 조절하는 두가지의 형태를 밝혔다. 이는 크게 유전자의 프로모터 지역에 있어서의 CpG 밀도에 따라 나눌 수 있었는데, CpG 가 풍부한 프로모터를 가진 유전자의 경우 DNA 메틸화가 낮게 유지되며 H3K4me3 와 H3K27me3 의 변화가 빠르게 일어남으로써 발현의 조절이 되는 반면 CpG 의 밀도가 낮은 프로모터를 가진 유전자는 발현이 억제될 때에 DNA 메틸화가 높고 탈메틸화에 제한이 되는 모습을 보였다. 특히 후자의 유전자 조절 형태는 *Dppa4*, *Dppa5a*, *Esrrb* 등의 유전자들에게 적용되고 있었으므로 배아줄기세포와 비슷한 전분화능을 얻는 데에 결정적으로 작용하는 것으로 보인다.

**결론:** 결론적으로 우리의 연구는 체세포가 전분화능을 얻는 데에 있어서 DNA 메틸화의 후생유전학적 스위치로서 역할의 중요성을 밝힐 수 있었다.

\* 본 내용은 Nature Communications 에 출판 완료된 내용임 (1)

-----  
**주요어 :** 배아줄기세포, 전분화능, 역분화줄기세포, DNA 메틸화, 중아황산염 처리 염기서열 분석, 후생유전체, DNA 메틸화에 변화를 보이는 지역, 히스톤 변화, 전사인자 결합 자리

**학 번 : 2010-21914**