Genomics Data 2 (2014) 73–77

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Identification of the new gene *Zrsr1* to associate with the pluripotency state in induced pluripotent stem cells (iPSCs) using high throughput sequencing technology

Shuai Gao [a,b], Gang Chang [a,b], Jianhui Tian [a], Shaorong Gao [b], Tao Cai [b,*]

[a] Ministry of Agriculture Key Laboratory of Animal Genetics, Breeding and Reproduction; National Engineering Laboratory for Animal Breeding; College of Animal Sciences and Technology, China Agricultural University, Beijing 100193, China
[b] National Institute of Biological Sciences, NIBS, Beijing 102206, China

ARTICLE INFO

Article history:
Received 28 March 2014
Received in revised form 17 April 2014
Accepted 17 April 2014
Available online 30 April 2014

ABSTRACT

Finding the markers to predict the quality of induced pluripotent stem cells (iPSCs) will accelerate its practical application. The fully pluripotent iPSCs has been determined as viable all-iPSC mice can be generated through tetraploid (4N) complementation. The activation of the imprinted Dlk1-Dio3 gene cluster was reported to correlate with the pluripotency of iPSCs. However, recent studies demonstrated that the loss of imprinting at the Dlk1-Dio3 locus does not strictly correlate with the reduced pluripotency of iPSCs. In our study (ref [1]), iPSC lines with the same genetic background and proviral integration sites were established, and the pluripotency state of each iPSC line was well characterized using tetraploid (4N) complementation assay. The gene expression and global epigenetic modifications of "4N-ON" and the corresponding "4N-OFF" iPSC lines were compared through deep sequencing analysis of mRNA expression, small RNA profiling, histone modifications (H3K4me3, H3K27me3 and H3K4me2) and DNA methylation. Very few differences were detected in the iPSC lines that were investigated. However, an imprinted gene, *Zrsr1* was disrupted in the "4N-OFF" iPSC lines. Here we provide more detail about the dataset and the R script with additional data for others to repeat the finding.

Specifications	
Organism/cell line/tissue	*Mus musculus/induced pluripotent stem cells (iPSCs), embryonic stem cells (ESCs), mice embryonic fibroblast (MEF), adipose progenitor cells (APC), hematopoietic progenitor cells (HPC)/*
Sex	
Sequencer or array type	*Hiseq2000*
Data format	*Raw: Fastq, analyzed: bed, fpkm*
Experimental factors	*"4N-ON" vs. "4N-OFF" using tetraploid (4N) complementation assay; iPSC lines vs. somatic cell lines*
Experimental features	*The epigenomics study in iPSCs and somatic cells using NGS technology*
Consent	
Sample source location	NIBS, Beijing, China

Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36294.

* Corresponding author at: 7 Science Park Road, Zhongguancun Life Science Park, Beijing 102206, China.
E-mail address: caitao@nibs.ac.cn (T. Cai).

Experimental design, materials and methods

The cell lines in the study

The lineage of cell lines in the study was illustrated in the Fig. 1:

There are total 13 cell lines sequenced: R1-ESC, 1^0-MEF-iPSC-37, 2^0-APC-iPSC-32, 2^0-HPC-iPSC-16, 3^0-APC-iPSC-3, 1^0-MEF-iPSC-42, 2^0-APC-iPSC-4, 2^0-HPC-iPSC-18, 3^0-APC-iPSC-9, 1^0-MEF, 2^0-APC, 2^0-HPC, 3^0-APC.

The first generation MEF (1^0-MEF) were derived from 13.5 dpc embryos collected from female 129S2/Sv mice that were mated with Rosa26-M2rtTA transgenic mice. Then the viral supernatants containing the TetO-FUW-Oct4, Sox2, Klf4, and c-Myc plasmids and the packaging plasmids ps-PAX-2 and pMD2G were harvested, and the MEFs were infected with supernatants containing viruses at a density of ~5×10^5 cells per 6 cm dish. The infection medium was replaced with ES medium supplemented with 1 μg/ml doxycycline (Dox) 12 h after infection. The ES-like colonies appeared at approximately 12 days, and 4 days after the withdrawal of Dox, smooth colonies were isolated and passaged 3 days later for the derivation of 1^0-MEF-iPSC lines. 1^0-MEF-iPSC-37 is proved true pluripotence by the capability of generation of all-iPSC mice. Subsequently, adipocyte progenitor cells (2^0-APC) and hematopoietic progenitor cells (2^0-HPC) were retrieved from the adipose

http://dx.doi.org/10.1016/j.gdata.2014.04.008

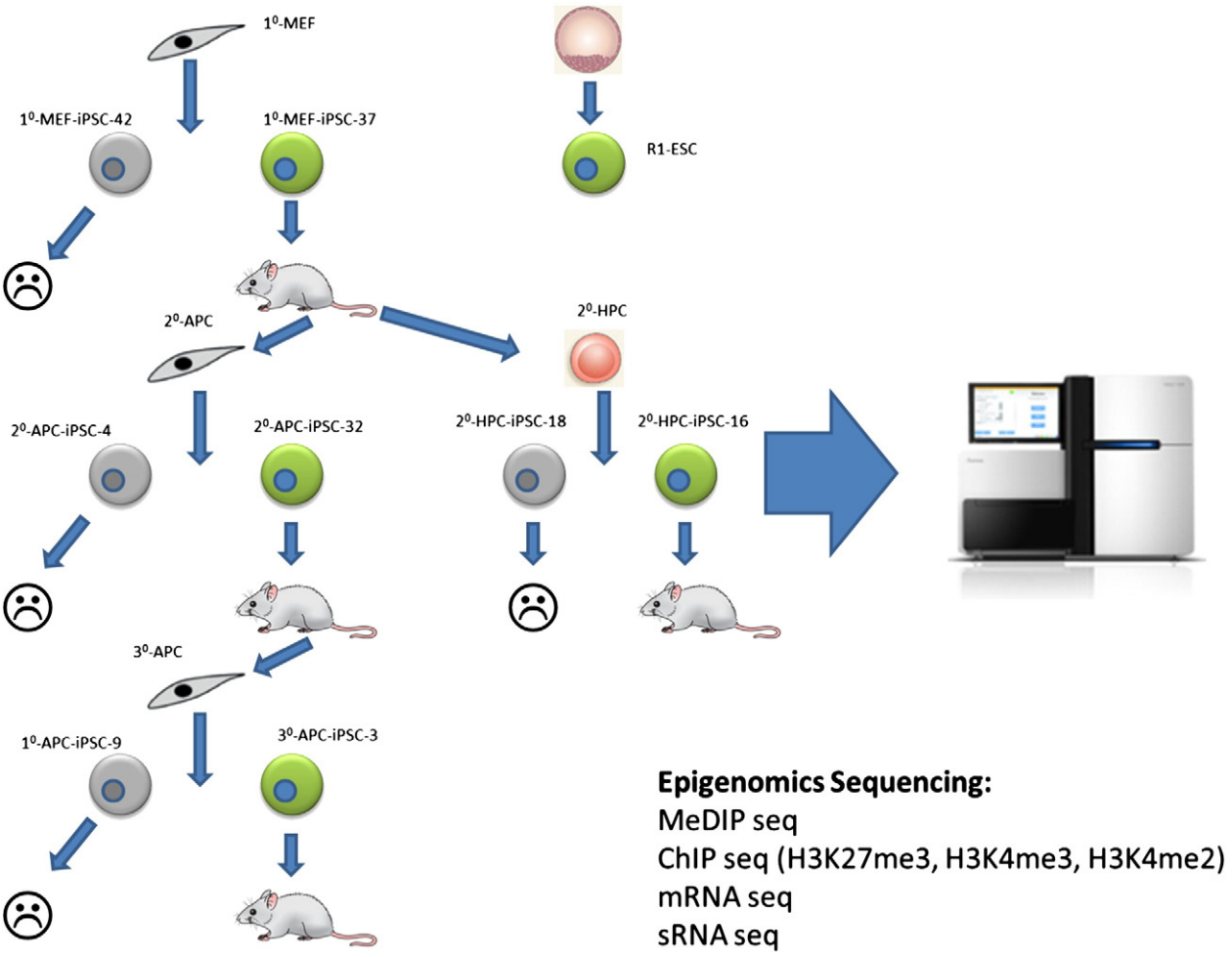

Fig. 1. The lineage of the cell lines. For each of cell lines, the DNA methylation, Histone modification (H3K27me3, H3K4me3, H3K4me2), RNA expression and small RNA expression were sequenced using HiSeq2000 platform.

tissue and the bone marrow of the 1^0-all-iPS mice produced by the 1^0-MEF-iPSC-37. After the addition of Dox to the induction medium, 2^0-iPS colonies emerged and viable all-iPSC mice could be generated from both 2^0-APC-iPSC-32 and 2^0-HPC-iPSC-16 cell lines through tetraploid complementation. Subsequently, a third generation of viable all-iPSC mice could be produced from the 3^0-APC-iPSC-3 cell lines that were established from 3^0-APC retrieved from the 2^0-all-iPS mice generated by the 2^0-APC-iPSC-32. On the contrary, 1^0-MEF-iPSC-42, 2^0-APC-iPSC-4, 2^0-HPC-iPSC-18, 3^0-APC-iPSC-9 cannot generate the all-iPSC mice through tetraploid complementation. Finally, R1-ESC is a typical normal fertilization-derived ES cell lines.

Sample preparing and sequencing

Samples prepared for RNA-Seq

Total RNA was isolated from cell pellets using TRIzol reagent (Life Technologies) according to the manufacturer's instructions. The RNA integrity was confirmed by using a Bioanalyzer 2100 (Agilent Technologies) with a minimum RNA integrity number (RIN) of 8. The mRNA was enriched by using oligo(dT) magnetic beads and sheared to create short fragments of approximately 200 bp. cDNA was synthesized by using random hexamer primers and purified by using PCR product extraction kit (Qiagen). Finally, the sequencing primers linked to the cDNA fragments (approximately 200 bp) were isolated by gel electrophoresis and enriched by PCR amplification to construct the library.

Samples prepared for MeDIP-Seq

Genomic DNA was extracted from the cell pellets by using the DNeasy Mini Kit (Qiagen). The DNA quality was analyzed by using a Qubit 2.0 Fluorometer (Life Technologies). The gDNA was sonicated to 100–500 bp and repaired to contain a 3′-dA overhang; adapters were then ligated to the end of the DNA fragments according to the Paired-End DNA Sample Prep Kit (Illumina). For immunoprecipitation, the DNA was first denatured, then immunoprecipitated with the 5mC antibody using the Magnetic Methylated DNA Immunoprecipitation kit (Diagenod). Q-PCR was performed to validate the efficiency of the enriched products. Next, the immunoprecipitated DNA was amplified for approximately 12–15 cycles, and fragments of the proper size (usually 200–300 bp) were gel-purified using the Gel Extraction Kit (Qiagen) and quantified by using a 2100 Bioanalyzer (Agilent Technologies).

Samples prepared for Small RNA-Seq

sRNAs of approximately 18–30 nt were first separated from the 5–10 μg of total RNA by size fractionation with a 15% TBE urea polyacrylamide gel (TBU). Next, the 5′ RNA adapter (GUUCAGAGUUCUACAGUCCGACGAUC-3′) was ligated to the RNA pool with T4 RNA ligase. The ligated RNA was size-fractionated on a 15% agarose gel, and the 40–65 nt fraction was excised. The 3′ RNA adapter (5′-pUCGUAUGCCGUCUUCUGCUUGidT-3′; p, phosphate; idT, inverted deoxythymidine) was subsequently ligated to the precipitated RNA by using T4 RNA ligase. The ligated RNA was size-fractionated on a 10% agarose gel, and the 70–90 nt fraction was excised. The samples were reverse transcribed, and amplified for about 15 cycles to generate the sRNA libraries.

Samples prepared for ChIP-Seq

ChIP experiments were performed as previously described (http://www.abcam.com). Briefly, ~1.5 × 10^8 cells were resuspended in lysis buffer and digested with micrococcal nuclease (Takara) for approximately 5 min at 37 °C. Then, the lysate was immunoprecipitated with the following antibodies: anti-H3K4me2 (Millipore 07–030), anti-H3K4me3 (Abcam ab8580), and anti-H3K27me3 (Millipore 07–449). A fraction of input 'whole-cell extract' was retained as a sequencing control. The DNA isolated from the ChIP was quantified using a Qubit 2.0 Fluorometer (Life Technologies), and Q-PCR was performed to validate the enrichment efficiency. Then, the enriched DNA was sonicated to 100–500 bp fragments. The DNA ends were repaired to create 3′-dA overhangs, and the adapters were then ligated to the ends of the DNA fragments. DNA fragments of approximately 100–300 bp were recovered and amplified to construct the sequencing library. Thirty-nine ChIP libraries and thirteen input controls were used for sequencing.

High throughput sequencing

The samples were sequenced in BGI by using HiSeq2000. All the samples took the Illumina 50 nt single end sequencing protocol except for the MeDIP samples (50 nt pair-end protocol). The fastq format raw data were generated for the following analysis.

The data analysis

The raw data were analyzed by using tophat/cufflink, MACS/CCAT, etc. (described in ref. 1). The quality of each dataset was estimated. The summary table for the mapping ratio, reads distribution,

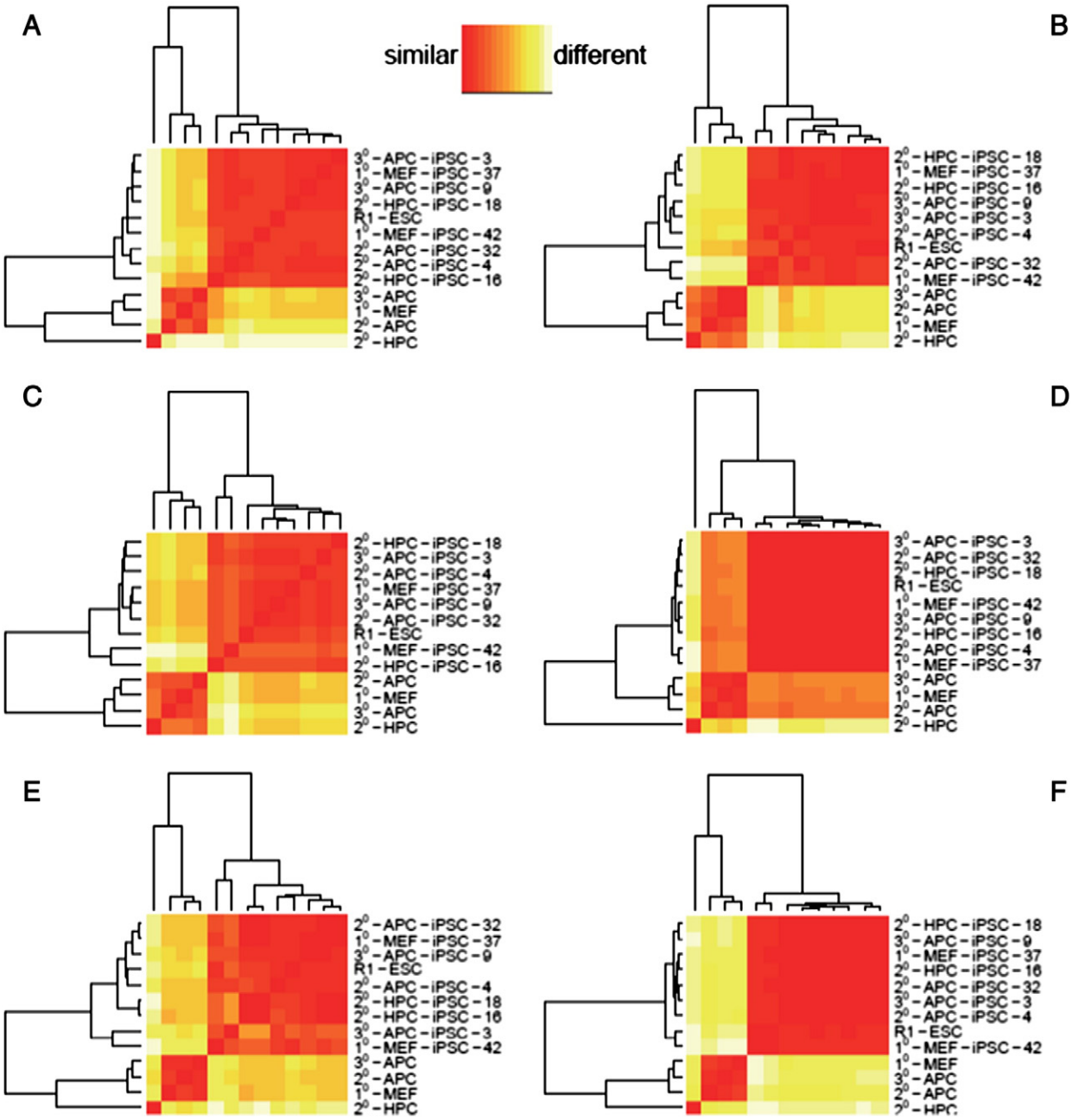

Fig. 2. The hierarchical cluster analysis for the similarity between iPSCs and ESC. The cluster analysis used "gplots" library in R language, the input was based on the Pearson correlation matrix for genes' expression (A) and Spearman correlation matrix for reads count in the miRNAs(B) and in the enriched regions of MeDIP (C), H3K27me3(D), H3K4me3(E), H3K4me2(F). The results are consistent with the expectation of great similarity between iPSCs and ESC.

etc. can be found in the appendix "Table 1.xls". The results are in concordant with the expectation. The global cluster analysis (Fig. 2) and the reads distribution plot for the marker genes (Fig. 3) also showed that the datasets are of good quality. The R script for the key finding of DNA methylation difference of *Zrsr1* (Fig. 4) was attached in the appendix ("compare.R" with the related data and description).

Conflict of interest

The authors declare no conflict interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2014.04.008.

Reference

[1] G. Chang, S. Gao, X. Hou, Z. Xu, et al., High-throughput sequencing reveals the disruption of methylation of imprinted gene in induced pluripotent stem cells. Cell Res. 24 (3) (2014 Mar) 293–306 (PMID: 24381111).

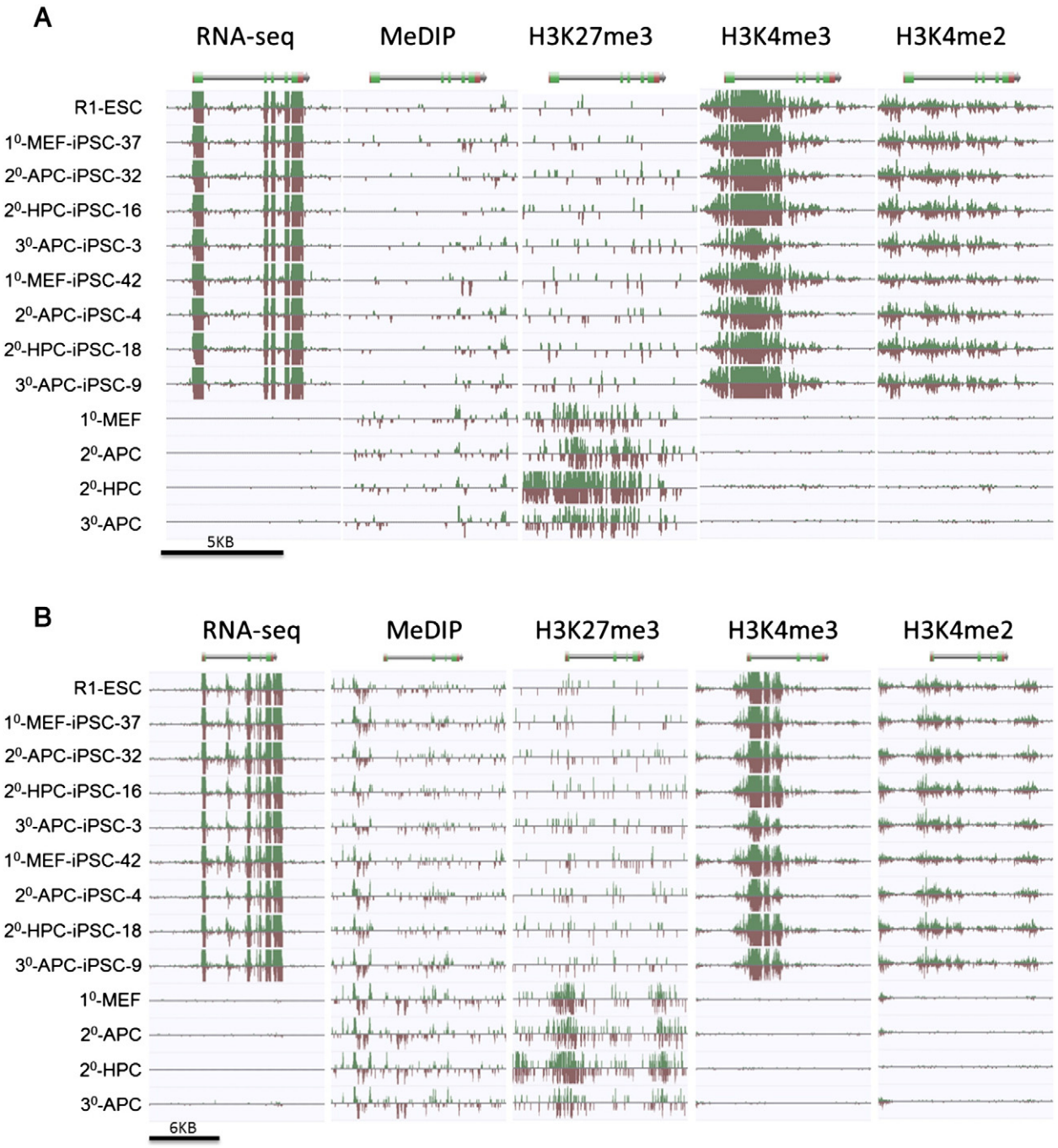

Fig. 3. The reads distribution for the iPSC marker genes (genome browser display snapshot). The *Oct4* (A) and *Nanog* (B) are marker genes for iPSCs. Both of them are highly transcribed in iPSCs but not in the somatic cells. The epigenetic modification and gene expression in the genes are in concordant with the knowledge.

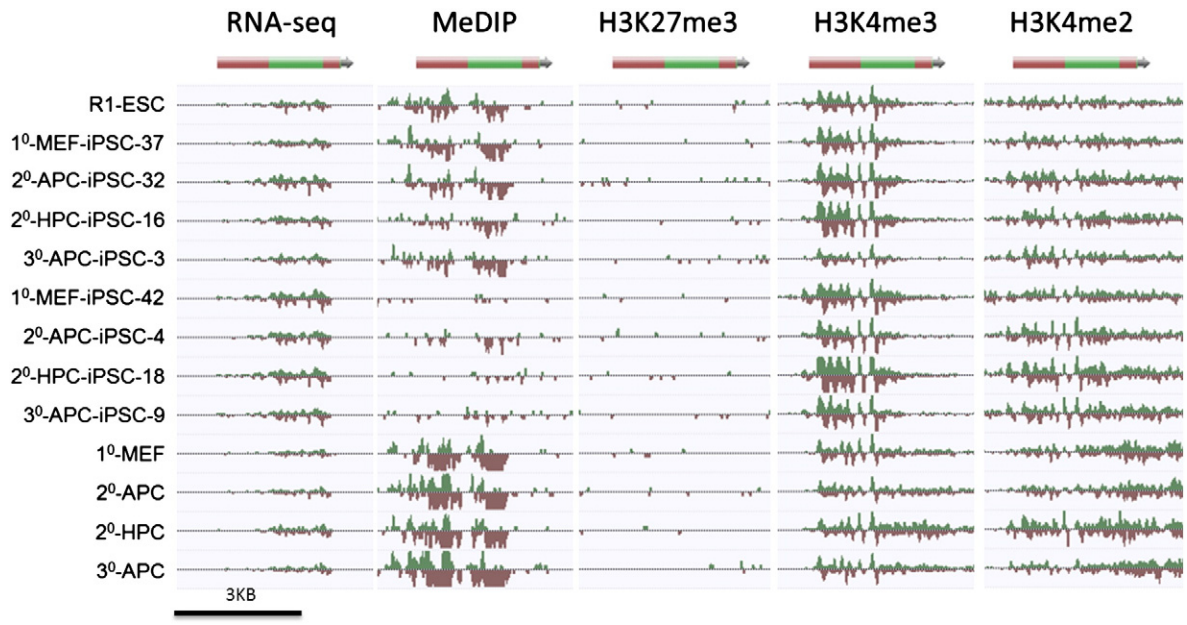

Fig. 4. The reads distribution for *Zrsr1*. The loss of MeDIP reads in "4N-OFF" iPSC lines (1^0-MEF-iPSC-42, 2^0-APC-iPSC-4, 2^0-HPC-iPSC-18, 3^0-APC-iPSC-9) can be detected.