Cell Reports, Volume *23*

Supplemental Information

Reprogramming of Chromatin Accessibility in Somatic Cell Nuclear Transfer Is DNA Replication Independent

Mohamed Nadhir Djekidel, Azusa Inoue, Shogo Matoba, Tsukasa Suzuki, Chunxia Zhang, Falong Lu, Lan Jiang, and Yi Zhang

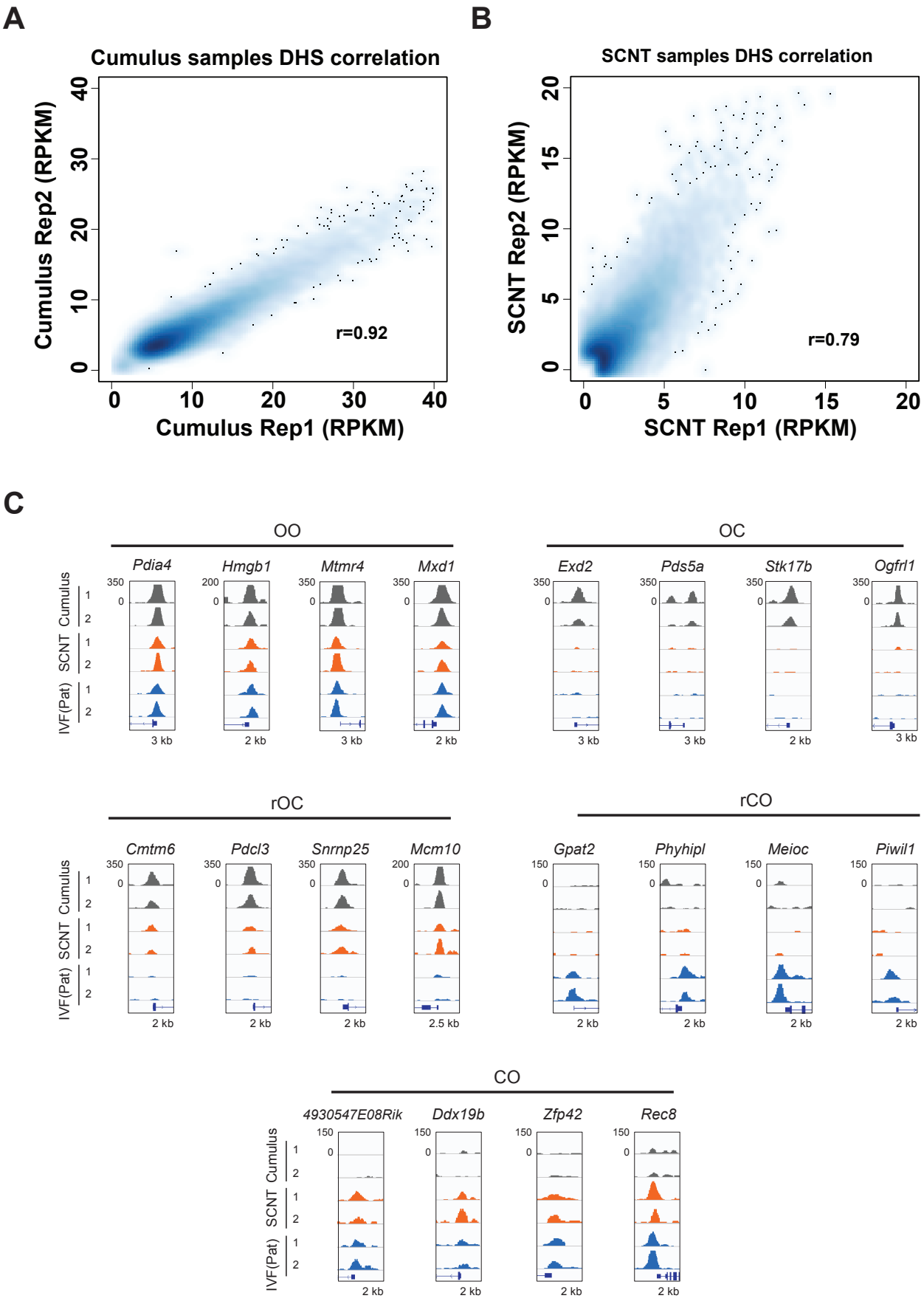
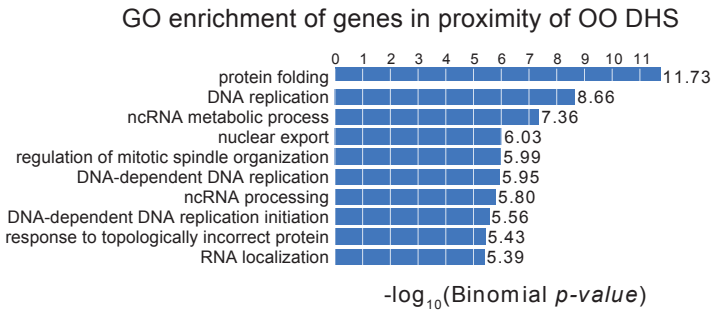
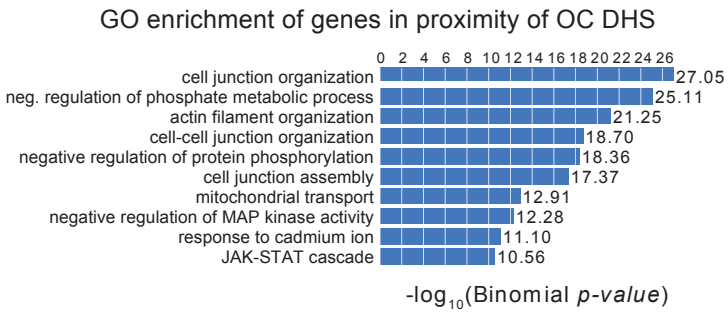


Figure S1. liDNase-Seq data reproducibility and DHS examples, Related to Figure 1. (A) Scatterplot showing the correlation of liDNase-seq signal of two biological replicates of cumulus cells. **(B)** Scatterplot showing the correlation of liDNase-seq signal of two biological replicates of 1-cell stage SCNT embryos. **(C)** Genome browser view of representative liDNase-seq loci for the different DHS categories.

A



B



C

de novo motifs enriched in rOC DHS

Motif	Best match TF	*p−value*	% target	% background
	AP−1(bZIP)	1e−19	16.98	5.59
	*Elf1*(ETS)	1e−13	20.23	9.23
	CCAAT−box (NFY)	1e−21	17.56	5.51

Figure S2. Functional and motif analysis of different DHS categories, Related to Figure 2. (A) Bar plot showing the top 10 enriched GO terms of OO-associated genes using GREAT enrichment. **(B)** Bar plot showing the top 10 enriched GO terms of CO-associated genes using GREAT enrichment. **(C)** TFs with binding motif enriched in the rOC DHSs.

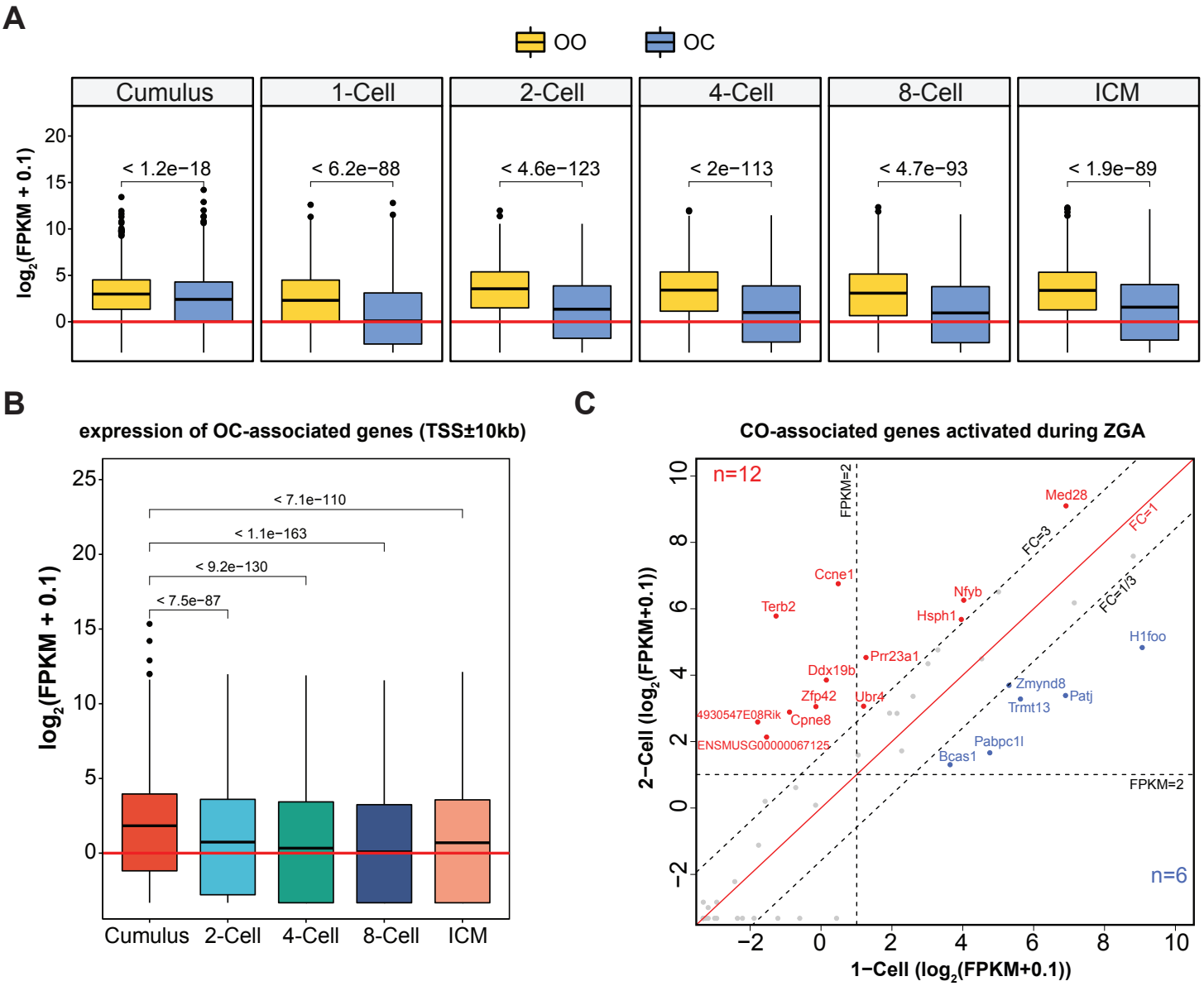


Figure S3. Reprogrammed DHS are associated with loss of somatic cell transcription program and activation of ZGA genes, Related to Figure 3. (**A**) Averaged gene expression levels of the OO and OC-associated genes (TSS±1 kb) in cumulus and preimplantation embryos. The *p-values* were calculated using the Wilcoxon rank sum test. (**B**) Averaged gene expression levels of the OC-associated genes (TSS±10 kb) in cumulus and preimplantation embryos. The *p-values* were calculated using the paired t-test**.** (**C**) Dot plot showing activation of CO-associated ZGA genes [red, FC>3, mean FPKM(2-Cell) >2] as well as the 2-Cell down regulated genes [blue, FC<1/3, mean FPKM(1-Cell) >2].

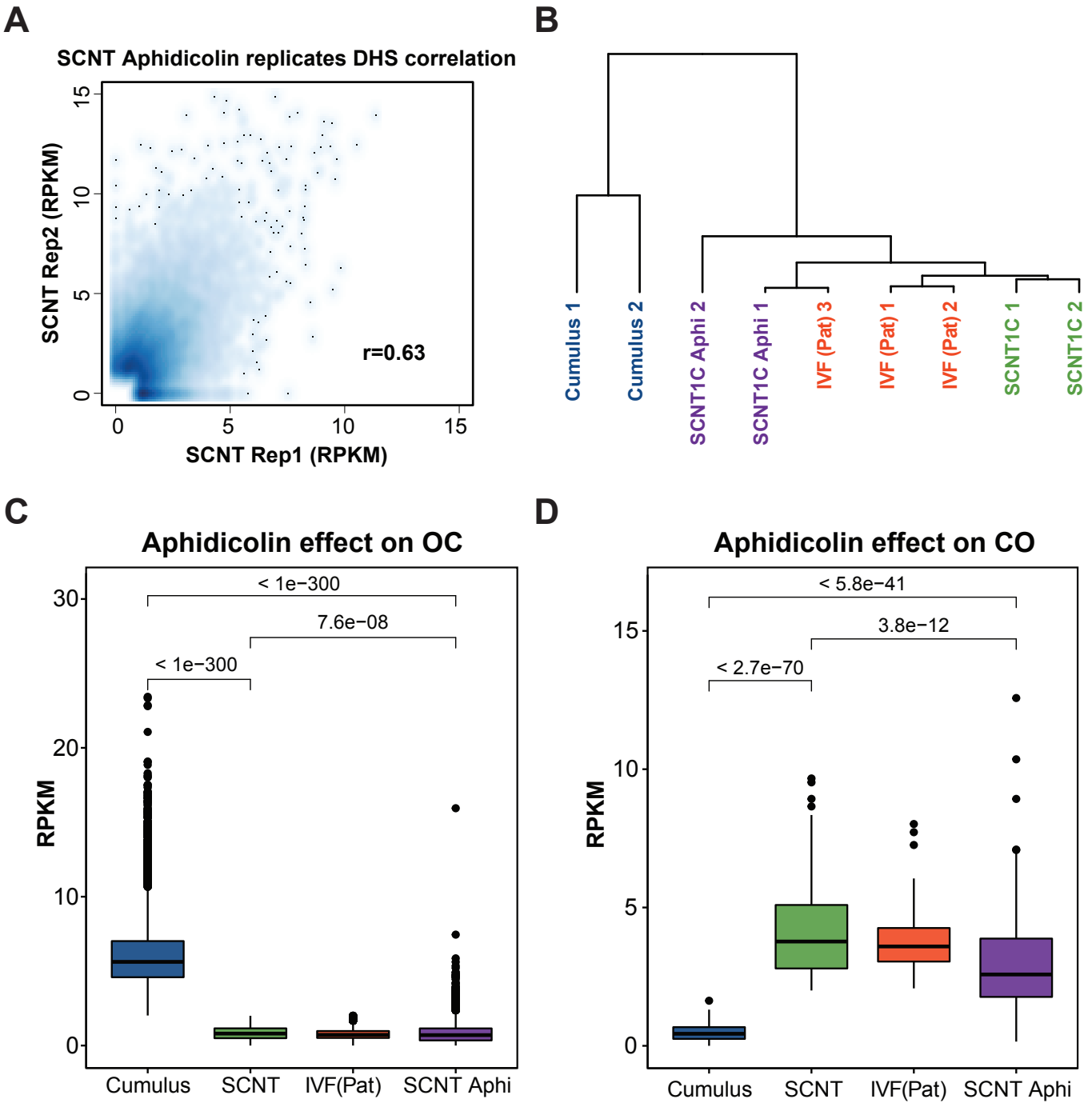


Figure S4. DHS reprogramming in SCNT is DNA replication independent, Related to Figure 4. (A) Scatterplot showing the correlation of liDNase-seq signal between two biological replicates of aphidicolin-treated SCNT samples. (**B**) Dendrogram showing the clustering of DHS profiles. (**C, D**) Box plot showing the normalized read per kilobase million (RPKM) intensities of the OC (C) and CO (D) with aphidicolin treatment. The *p-values* were calculated using paired t-test. In (C) the *p-values* were equal to 0 (due to the machine limitation), thus we showed them as <1e-300.

	Sample	Treatment	# Total reads	# Reads after trimming	# Uniquely Mapped reads	Ratio Mapped	# reads after PCR duplicates removal	Ratio Retained
liDNase-Seq	**Cumulus Rep1**	**None**	66,351,339	66,190,050	11,788,242	17.81%	6,774,772	10.24%
	Cumulus Rep2	**None**	65,038,144	64,899,520	17,283,041	26.63%	10,063,816	15.51%
	SCNT Rep1	**None**	63,929,455	63,824,336	24,893,889	39%	9,352,487	14.65%
	SCNT Rep2	**None**	61,753,526	61,613,759	8,302,330	13.47%	3,501,401	5.68%
	SCNT Rep1	**Aphidicolin**	57,559,043	57,531,340	13,937,720	24.23%	4,623,070	8.04%
	SCNT Rep2	**Aphidicolin**	51,452,039	51,416,901	8,615,459	16.76%	1,879,840	3.66%

Table S4. Summary of the datasets generated in this study. Related to Figure 1 and Figure 4.

Supplemental Experimental Procedures

Data analysis of liDNase-seq

The single-end 100-bp reads generated from liDNase-seq were first trimmed of low quality and adapter sequences using trim_galore (v0.4.4) (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Then, reads were mapped to the mm9 reference genome using Bowtie1 (Langmead et al., 2009) (v0.12.9) with parameters '-v 3 -m 1 -S -t'. Unmapped and Multi-mapped reads were filtered using samtools (Li et al., 2009) (v1.3.1-38-g9df6321). PCR duplicates where removed using the MarkDuplicates command from Picard tools (http://picard.sourceforge.net/)(v2.6.0). For each replicate, an initial list of DHS was generated using hotspot (John et al., 2011) (v4.1.0). The most reliable peaks between replicates were then selected using the IDR method (Li et al., 2011) , available at https://github.com/nboley/idr (v2.0.4), with a cutoff of IDR < 0.05. Overlapping DHS detected in Cumulus, 1-Cell SNCT, paternal pronuclei and the SCNT+Aphidicolin samples were merged using the findOverlapsOfPeaks function of the R/Bioconductor package ChIPpeakAnno (Zhu et al., 2010). The number of reads mapping to each peak were calculated using the summarizeOverlaps function from the R/Bioconductor GenomicAlignments package (Lawrence et al., 2013) then normalized to RPKM values. For each sample, the Pearson correlation between replicates was calculated using the peaks that had an RPKM>1 in at least one of the replicates. The PCA analysis was performed using all the merged peaks.

ATAC-seq analysis

The preimplantation embryos ATAC-seq data were from a previous publication (Wu et al., 2016). Raw reads were first trimmed using Trimmomatic (Bolger et al., 2014)(v.0.36). 4-bp sliding window was used to scan the reads and trim when the window mean quality dropped below 15. The trimmed pair-end reads were then mapped to the mm9 genome using Bowtie2 (v2.2.9) with parameter -X 2000 --mm -1. Unmapped reads, reads with mapping quality less than 10 were removed using samtools (v1.3.1-38-g9df6321). Reads mapping to the mitochondrial genome were discarded. PCR duplicates were removed using the MarkDuplicates command from Picard tools (v2.6.0). The bam files of the replicates were then merged together using samtools.

RNA-seq analysis

All of the RNA-seq data used in this study were download from public sources. Cumulus cell RNA-seq data were from a previous study (Matoba et al., 2014) under the GEO accession numbers GSM1426195 and GSM1426196. The mouse oocyte data were from the public data of a previous study (Xue et al., 2013) available under GEO accession numbers GSM1080195 and GSM1080196. The raw data of preimplantation embryos were from a previous study (Wu et al., 2016) under the GEO accession number GSE66582. The downloaded raw reads were trimmed using Trimmomatic(v.0.36) and then mapped to the mm9 genome. Reads with a length of at least 50-bp were mapped to the genome using STAR (Dobin et al., 2013) (v.2.5.2b) with the following parameters: --outSAMtype BAM SortedByCoordinate –outSAMunmapped Within –outFilterType BySJout --outSAMattributes NH HI AS NM MD --outFilterMultimapNmax 20 --outFilterMismatchNmax 999 --quantMode TranscriptomeSAM GeneCounts. Gene expression was then estimated using RSEM (Li and Dewey, 2011) given the Ensemble NCBIM37.67 transcriptome assembly.

H3K9me3 ChIP-seq data analysis

The cumulus and 2-cell embryo SCNT H3K9me3 ChIP-seq data were from a previous publication (Liu et al., 2016) under GEO accession numbers GSM1811770, GSM1811771, GSM1811775, and GSM1811776. Reads were trimmed using Trimmomatic (Bolger et al., 2014) and then mapped to the mm9 genome using Bowtie1 (Langmead et al., 2009) (v0.12.9) with parameters '--best --strata --sam -m 1'. Multi-mapped and unmapped and low-quality reads were removed using samtools (Li et al., 2009) (v1.3.1-38-g9df6321) and PCR duplicates were removed using the MarkDuplicates command from Picard tools (v2.6.0).

Genomic annotation of DHS

The Ensemble NCBIM37.67 genomic annotation was used to associate DHS peaks with the different genomic regions. Briefly, the makeTxDbFromGFF function from the GenomicFeatures (Lawrence et al., 2013) R/Bioconductor package was used to load the GTF file, then we called the annotatePeak function from the R/Bioconductor ChIPpeakAnno (Zhu et al., 2010) package for genomic annotation. Promoters were considered to be ±1Kb from TSS and all the regions that did not fall within exons, introns, or UTRs were classified as distal intergenic regions.

Functional enrichment analysis

The GREAT utility (McLean et al., 2010) was used to perform functional enrichment analysis of DHS. All the functional analysis were performed under the "Basal plus extension" mode in which the gene regulatory domain was defined as ±10kb around the TSS with a distal domain up to 1Mb. Only the Gene Ontology terms or pathways with an FDR < 0.05 were considered.

Motif analysis

All the *de-novo* motifs in this study were detected using HOMER (Heinz et al., 2010) (v4.9) with parameters mm9 and -size 500. *De-novo* motifs that had a p-value < 0.0001 and at least a two-fold enrichment between foreground and background were considered for downstream analysis.

Heatmap generations

The DHS peaks were extended by ±5kb from their peak center to form 10kb windows, each window was then split into 100 bin using the binner function from the R/Bioconductor package genomation (Akalin et al., 2015). The number in each bin were counted using the summarizeOverlaps function from the GenomicAlignments R/Bioconductor package (Lawrence et al., 2013). A matrix was then generated and displayed using the EnrichedHeatmap R/Bioconductor package (Zuguang, 2017).

PCA analysis

The DHS peaks detected in Cumulus, SCNT, SCNT+ Aphidicolin treatment and IVF(Pan) samples were merged together, their RPKM values were calculated in each biological replicate then the 'prcomp' function available in R (http://www.r-project.org/) was used with parameters 'center = FALSE, scale. = FALSE'.

Genomic tracks generation

Smoothed bigwig tracks for each sample were generated using the bamCoverage utility from the deeptools suite (Ramirez et al., 2014) (v2.5.1) with parameters '--binSize 25 --extendReads 200 --normalizeUsingRPKM --outFileFormat bigwig --scaleFactor 1'. All genomic track visualization was performed using IGV (Thorvaldsdottir et al., 2013).

Statistical analyses

Statistical analyses were implemented with R (http://www.r-project.org/). Pearson's r coefficient was calculated using the 'cor' function with default parameters.

Pipeline automation

All of the RNA-seq, ChIP-seq, and ATAC-seq mapping pipelines were automated using Bpipe (Sadedin et al., 2012).

Code availability

Additional custom codes used for bioinformatics analysis are available upon request.

Supplemental References

Akalin, A., Franke, V., Vlahovicek, K., Mason, C.E., and Schubeler, D. (2015). Genomation: a toolkit to summarize, annotate and visualize genomic intervals. Bioinformatics *31*, 1127-1129.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114-2120.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell *38*, 576-589.

John, S., Sabo, P.J., Thurman, R.E., Sung, M.H., Biddie, S.C., Johnson, T.A., Hager, G.L., and Stamatoyannopoulos, J.A. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat Genet *43*, 264-268.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. PLoS Comput Biol *9*, e1003118.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Li, Q., Brown, J.B., Huang, H., and Bickel, P.J. (2011). Measuring reproducibility of high-throughput experiments. Ann Appl Stat *5*, 1752-1779.

Liu, W., Liu, X., Wang, C., Gao, Y., Gao, R., Kou, X., Zhao, Y., Li, J., Wu, Y., Xiu, W.*, et al.* (2016). Identification of key factors conquering developmental arrest of somatic cell cloned embryos by combining embryo biopsy and single-cell sequencing. Cell Discov *2*, 16010.

Matoba, S., Liu, Y., Lu, F., Iwabuchi, K., Shen, L., Inoue, A., and Zhang, Y. (2014). Embryonic Development following Somatic Cell Nuclear Transfer Impeded by Persisting Histone Methylation. Cell *159*, 884-895.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol *28*, 495-501.

Ramirez, F., Dundar, F., Diehl, S., Gruning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res *42*, W187-191.

Sadedin, S.P., Pope, B., and Oshlack, A. (2012). Bpipe: a tool for running and managing bioinformatics pipelines. Bioinformatics *28*, 1525-1526.

Thorvaldsdottir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform *14*, 178-192.

Wu, J., Huang, B., Chen, H., Yin, Q., Liu, Y., Xiang, Y., Zhang, B., Liu, B., Wang, Q., Xia, W.*, et al.* (2016). The landscape of accessible chromatin in mammalian preimplantation embryos. Nature *534*, 652-657.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.Y., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y.E.*, et al.* (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature *500*, 593-597.

Zhu, L.J., Gazin, C., Lawson, N.D., Pages, H., Lin, S.M., Lapointe, D.S., and Green, M.R. (2010). ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics *11*, 237.

Zuguang, G. (2017). EnrichedHeatmap: Making Enriched Heatmaps. R package version 192.