

# SCIENTIFIC DATA

OPEN

## Data Descriptor: Single-cell transcriptome of early embryos and cultured embryonic stem cells of cynomolgus monkeys

Received: 16 December 2016

Accepted: 4 April 2017

Published: 20 June 2017

Tomonori Nakamura<sup>1,2</sup>, Yukihiro Yabuta<sup>1,2</sup>, Ikuhiro Okamoto<sup>1,2</sup>, Kotaro Sasaki<sup>1,2</sup>, Chizuru Iwatani<sup>3</sup>, Hideaki Tsuchiya<sup>3</sup> & Mitinori Saitou<sup>1,2,4,5</sup>

In mammals, the development of pluripotency and specification of primordial germ cells (PGCs) have been studied predominantly using mice as a model organism. However, divergences among mammalian species for such processes have begun to be recognized. Between humans and mice, pre-implantation development appears relatively similar, but the manner and morphology of post-implantation development are significantly different. Nevertheless, the embryogenesis just after implantation in primates, including the specification of PGCs, has been unexplored due to the difficulties in analyzing the embryos at relevant developmental stages. Here, we present a comprehensive single-cell transcriptome dataset of pre- and early post-implantation embryo cells, PGCs and embryonic stem cells (ESCs) of cynomolgus monkeys as a model of higher primates. The identities of each transcriptome were also validated rigorously by other way such as immunofluorescent analysis. The information reported here will serve as a foundation for our understanding of a wide range of processes in the developmental biology of primates, including humans.

<b>Design Type(s)</b>	cell type comparison design • organism development design
<b>Measurement Type(s)</b>	transcription profiling assay
<b>Technology Type(s)</b>	RNA sequencing
<b>Factor Type(s)</b>	cell type
<b>Sample Characteristic(s)</b>	Macaca fascicularis • blastocyst • embryo • embryonic stem cell

<sup>1</sup>Department of Anatomy and Cell Biology, Graduate School of Medicine, Kyoto University, Yoshida-Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan. <sup>2</sup>JST, ERATO, Yoshida-Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan. <sup>3</sup>Research Center for Animal Life Science, Shiga University of Medical Science, Seta-Tsukinowa-cho, Otsu, Shiga 520-2192, Japan. <sup>4</sup>Center for iPS Cell Research and Application, Kyoto University, 53 Kawahara-cho, Shogoin, Sakyo-ku, Kyoto 606-8507, Japan. <sup>5</sup>Institute for Integrated Cell-Material Sciences, Kyoto University, Yoshida-Ushinomiya-cho, Sakyo-ku, Kyoto 606-8501, Japan. Correspondence and requests for materials should be addressed to T.N. (email: t-nakamu@anat2.med.kyoto-u.ac.jp) or to M.S. (email: saitou@anat2.med.kyoto-u.ac.jp).

## Background & Summary

For more than half a century, mice have been exploited as a representative model organism for mammalian development and physiology. The results acquired from these researches have greatly contributed to our understanding of such processes as well as diseased states. Nonetheless, it has been widely recognized that there are significant differences in development and physiology between mice and humans.

The epiblast (EPI) is made up of cells specified during the peri-implantation period of early embryogenesis and differentiates into three primary germ layers and the germ cell lineages; the EPI therefore bears the pluripotency. Even within such a transient period, the EPI cells show dynamic changes in pluripotency from a so-called naive to a primed state<sup>1</sup>, and both states in mice have been captured *in vitro*: the naive state is replicated in ESCs/induced pluripotent stem cells (iPSCs), which have essentially the same status as the pre-implantation EPI *in vivo*<sup>2</sup>, while the primed state is replicated in epiblast stem cells (EpiSCs), which are derived from the post-implantation EPI and are homologs of the EPI of the gastrula<sup>3–5</sup>. On the other hand, although human ESCs/iPSCs have been established, their characteristics—including morphology, culture requirements, and the molecular networks underlying the pluripotent state—have been considered to more closely resemble those of mouse EpiSCs. The underlying mechanisms of these differences have remained unresolved, in part due to the lack of *in vivo* analyses of human/primate early post-implantation development.

The PGCs emerge from the posterior EPI at the onset of gastrulation in mice<sup>6</sup>. The specification process of mouse PGCs has been extensively studied and was reconstructed *in vitro*<sup>7</sup>. Remarkably, the resultant cells (PGC-like cells: PGCLCs) have the ability to produce functional gametes, and the dynamics of PGCLC specification is highly similar to that *in vivo*. In humans, PGCLCs have also been induced from human ESCs/iPSCs<sup>8,9</sup>, but it has still remained unclear whether their properties recapitulate those *in vivo*, again due to the lack of *in vivo* information.

Here, we present a comprehensive transcriptome dataset at the single-cell level from pre- and post-implantation embryo cells, PGCs and ESCs of cynomolgus monkeys, one of the primates amenable to experiment and most closely related to humans. We employed the single-cell mRNA 3' end sequencing (SC3-seq), which was designed to enrich the reads of the 3' end of transcripts and enables highly quantitative and effective analysis<sup>10</sup>. We successfully amplified a total of 1,241 single-cell cDNAs and generated 474 transcriptomes (Table 1 (available online only)). The qualities of the transcriptomes and the representations of gene expression profile were validated by qPCR. The sample annotations were defined comprehensively by comparing the expression of key genes in transcriptome data with that obtained through the histological analysis such as immunofluorescent analysis and/or *in situ* hybridization<sup>11,12</sup>. Thus, the dataset in this Data Descriptor defined the first comprehensive molecular dynamics of primate early development, including early post-implantation embryogenesis, and will provide a foundation for future studies of primate development.

## Methods

The description of the method is extended from the related research manuscripts<sup>11,12</sup>.

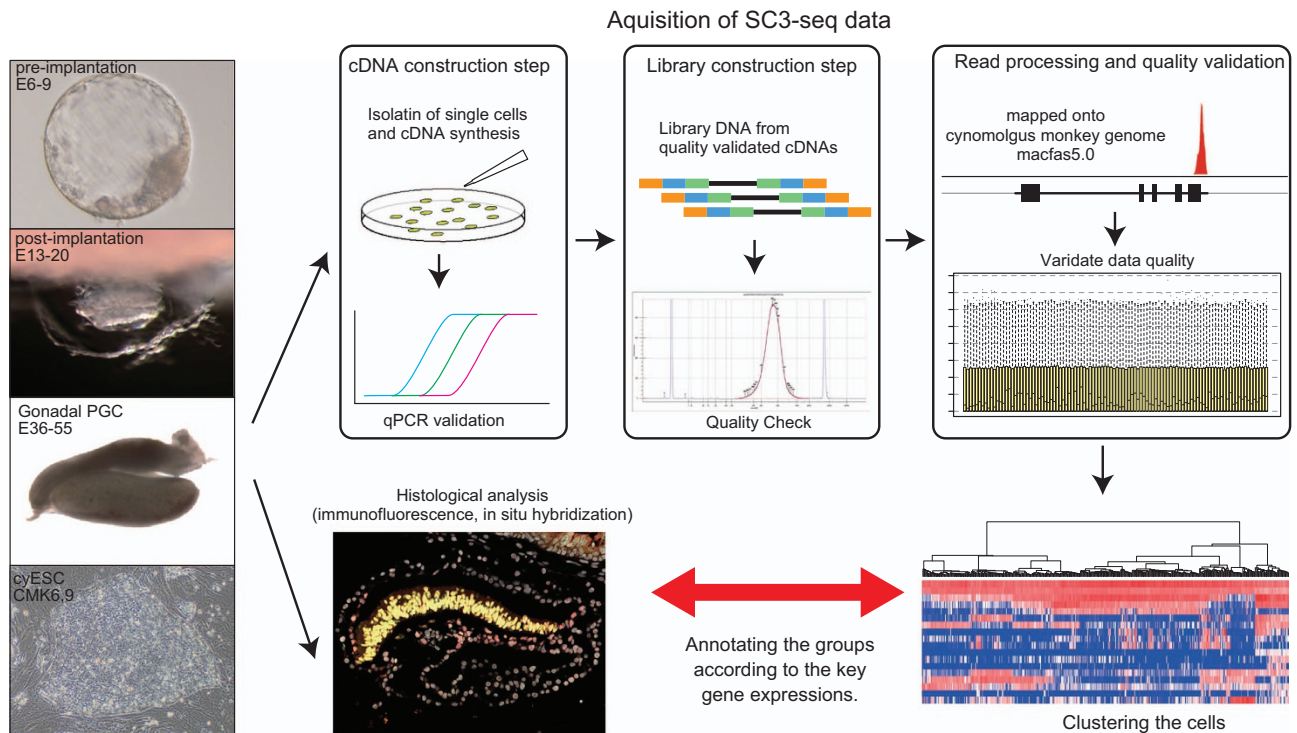
### Experimental study design

The overall experimental design is illustrated in Fig. 1. For the generation of SC3-seq data, single cells were prepared from pre- and post-implantation embryos, genital ridges and ESCs. They were picked manually and the SC3-seq cDNAs were amplified. After the quality validation and the selection of cDNAs, the DNA libraries for massive parallel sequencers were constructed (see the results of prior qPCR analyses<sup>11,12</sup>). The sequence reads were acquired by SOLiD5500xl and were mapped onto the cynomolgus monkey genome, *Macaca fascicularis* 5.0 (MacFas5.0). Then, the reads count was converted into reads per million (RPM). Finally, we verified the global distributions of gene expression and obtained quality-validated transcriptomes. Generally, single-cell transcriptome analysis drops off the positional information. Therefore, the expression patterns of key genes were also examined by immunofluorescent analysis or *in situ* hybridization of their expression, and each transcriptome data point was annotated accurately by comparison with its histologically examined counterpart<sup>11,12</sup>.

### Animals

The experimental procedures were approved by the Animal Care and Use Committee of Shiga University of Medical Science, and the methods were carried out in accordance with the approved guidelines. The procedures in cynomolgus monkeys for housing, oocyte collection, intra-cytoplasmic sperm injection (ICSI), pre-implantation embryo culture, and transfer of pre-implantation embryos into foster mothers were performed as described previously with some modifications<sup>9,11–14</sup>. Briefly, monkeys were housed individually in appropriate cages, and the light cycle consisted of 12 h of artificial light from 8 AM to 8 PM. Temperature and humidity in the animal rooms were maintained at  $25 \pm 2$  °C and  $50 \pm 5\%$ , respectively. Each animal was fed 20 g kg<sup>-1</sup> of body weight of commercial pellet monkey chow (CMK-1; CLEA Japan Inc., Tokyo, Japan) in the morning, supplemented with 20–50 g of sweet potato in the afternoon. Water was available *ad libitum*.

For super-ovulation, ovarian stimulation with follicle-stimulating hormone (Gonapure; ASKA) was performed by embedding an implantable and programmable micro-fusion device (iPRECIO; Primetech



**Figure 1. Overview of the study design and work flow.** The overall experimental design and work flow. The isolated cells dissociated from each samples were picked up for the cDNA synthesis. Then the quality and key gene expression of the generated cDNAs were examined by qPCR and the selected samples were used for the library construction. The libraries were sequenced and the resultant reads were mapped onto the cynomolgus monkey genome. The global distribution of gene expression in each cell was validated. Finally, the annotations of cell clusters were defined by comparing the gene expression patterns of the transcriptome and histological analysis such as immunofluorescent analysis.

Corporation) subcutaneously. The day when the ICSI was performed was designated as embryonic day 0 (E0). For embryo transfer, 4 to 5 two-cell to blastocyst-stage embryos were selected and transferred into an appropriate recipient female. For the detection of pregnancy of post-implantation embryos, implanted embryos were monitored by ultrasound scanning. For the collection of early post-implantation embryos (until around E20), the implanted uterus was surgically removed and bisected for the isolation of embryos. For the germ cell collection in gonads, the embryos were delivered by Caesarean section and bisected for the isolation of gonads. All the embryos used in this Data Descriptor are listed in Table 2 (available online only).

### Cell culture

The cynomolgus ESCs [CMK6 (male) and CMK9 (female)] were gifts from H. Suemori<sup>15</sup>. For cultivation on feeders, they were cultured with conventional human ESC medium [DMEM/F12 (D6421; Sigma-Aldrich) supplemented with 20% (vol/vol) of KSR (10828-028; Thermo Fisher Scientific), 1 mM of sodium pyruvate (11360-070; Thermo Fisher Scientific), 2 mM of GlutaMax (35050-061; Thermo Fisher Scientific), 0.1 mM of non-essential amino acids (11143-050; Thermo Fisher Scientific), 0.1 mM of 2-mercaptoethanol (M3148; Sigma-Aldrich), 1,000 U ml<sup>-1</sup> of ESGRO mouse LIF (ESG1107; Millipore), and 4 ng ml<sup>-1</sup> of recombinant human bFGF (060-04543; Wako Pure Chemical Industries)] on mouse embryonic feeders (MEFs)<sup>16</sup>. For feeder-free cultivation, cynomolgus ESCs were cultured under the same condition as human iPSCs (StemFit; Ajinomoto) on recombinant LAMININ511 (iMatrix511, Nippi), as described previously<sup>17</sup>. All of the cell lines were tested for mycoplasma contamination by MycoAlert (LT07-118; Lonza Japan), according to the manufacturer's instructions.

### Single-cell preparation

For pre-implantation embryos, the zona pellucida was removed by acid tyrode solution treatment (T1788; Sigma-Aldrich). Then the whole embryo was incubated with 0.25% trypsin/PBS (T4799; Sigma-Aldrich) for around 10 min at 37 °C, then dissociated into single cells by repeated pipetting, and dispersed in 0.1 mg ml<sup>-1</sup> of PVA/PBS (P8136; Sigma-Aldrich) for preparation of single-cell cDNAs.

For early post-implantation embryos, the implantation site was dissected out from the uterus and the embryonic fragment containing the EPI, amnion, hypoblast, and yolk-sac endoderm was isolated manually. For isolation of PGCs, in several instances relatively posterior parts of the embryonic fragments were dissected. The embryo information and dissected positions are summarized in Table 1 (available online only). For the embryonic gonads, the genital ridges were dissected out from embryos manually. Then each fragment was incubated with 0.25% trypsin/PBS for around 10 min at 37 °C, dissociated into single cells by repeated pipetting, and dispersed in 0.1 mg ml<sup>-1</sup> of PVA/PBS.

For cynomolgus ESCs, the cells were first detached as clumps with CTK solution [0.25% trypsin (15090-046; Thermo Fisher Scientific), 0.1 mg ml<sup>-1</sup> of collagenase IV (17104-019; Thermo Fisher Scientific), and 1 mM of CaCl<sub>2</sub> (06729-55; Nacalai Tesque)], incubated in TrypLE Select (12563029; Thermo Fisher Scientific) for around 10 min at 37 °C, and dispersed into single cells in 1% (vol/vol) KSR/PBS containing 10 μM of the ROCK inhibitor Y-27632 (257-00511; Wako Pure Chemical Industries)<sup>16</sup>. Cells under the feeder-free condition were directly incubated in TrypLE Select for around 5 min at 37 °C, and dispersed into single cells in 1% (vol/vol) KSR/PBS containing 10 μM of the ROCK inhibitor Y-27632.

### Acquisition of SC3-seq data

The SC3-seq process consists of two steps: the synthesis and amplification of cDNAs from isolated single cells, and the construction of a DNA library for sequencing by the SOLiD5500xl sequencer. The first step, the synthesis and amplification of the single-cell cDNA, was performed essentially as described previously<sup>10</sup>. Briefly, cells were picked up manually into the lysis buffer containing tagged dT primer [V1dT(24)], and lysed by heating. Then the Reverse transcription buffer was added and the 800 ~ 1,500 nt cDNA strand was synthesized with short reaction time (5 min). The reaction mixture contained an excess amount of tagged dT primers, which would have interrupted proper amplification. Therefore, the remaining dT primers were digested by exonuclease. Next, in order to add the poly A tail at the end of the synthesized cDNA strand, a terminal deoxynucleotidyl transferase mixture was added into the reaction. Then the second cDNA strand with another tag was synthesized by using a V3-tagged dT primer [V3(dT)24]. Finally, the synthesized double-strand cDNAs were amplified by PCR using V1 and V3 sequences, and the cDNAs with V3 and V1 tags at the 5' and 3' sides of the mRNA were obtained.

Before the construction of the DNA library, the quality of the amplified cDNAs was evaluated by examining the Ct values of the qPCR of several endogenous genes and by examining the distribution of the lengths of cDNA fragments using a LabChip GX (CLS760672; Perkin Elmer) or Bioanalyzer 2,100 (5,067-4,626; Agilent Technologies) system. qPCR was performed using Power SYBR Green PCR Master Mix (4367659; Thermo Fisher Scientific) with a CFX384 real-time qPCR system (Bio-Rad, Hercules, CA) according to the manufacturer's instructions. According to the pilot experiments, the samples whose Ct values of *GAPDH* were more than 21 tended to have poor quality (data not shown). Therefore, we considered the samples whose Ct values of *GAPDH* and *PPIA* were less than 19 and 20 as those with good quality. Then we chose appropriate samples from good quality cDNAs for library construction based on the combinations of the lineage-specific gene expression (Table 3 (available online only)). For the cells from pre-implantation embryos, NANOG, GATA4 and GATA6 were used for the marker of EPI, hypoblast and hypoblast/Trophectoderm. For the cells from post-implantation embryos, EPI cells were defined as POU5F1(+)/ NANOG(+)/ SOX2(+)/ PRDM14(+)/ T(-)/ GATA4(-), and gastrulating cells were POU5F1(+)/ NANOG(low)/ PRDM14(low)/ some of T, GATA4, GATA6(+). The extraembryonic cells such as visceral endoderm, yolk sac endoderm and extraembryonic mesenchyme were classified as POU5F1(low) and other lineage-specific genes (+). The early PGCs were identified as PRDM1(+)/ TFAP2C(+)/ SOX17(+)/SOX2(-). The late PGCs from embryonic gonads were identified as POU5F1(+)/ NANOG(+)/ TFAP2C(+)/ SOX2(-). Most of the primer sets were designed using Primer-Blast (NCBI) within a distance of 500 base pairs (bp) from the transcription termination sites (TTSs). The primer sets and oligo DNA sequences used in this Data Descriptor are given in Table 3 (available online only).

SC3-seq libraries of quality-checked cDNAs were constructed as described previously<sup>10</sup>. Briefly, the cDNAs were digested into 150–250 bp fragment by sonication (M210, Covaris). Then the damaged cDNA fragments were end-polished by T4 DNA polymerase (M203; NEB) and T4 polynucleotide kinase (M201; NEB). The cDNAs are now expected to be broken into three types of fragments; the ones with the V3-tag (5' side of the original mRNA), the ones with the V1-tag (3' side of the original mRNA), and the ones without any tag (internal part of the original mRNA). This fragment mixture was subjected to a one-cycle DNA polymerization step to allocate the internal adaptor sequence (the essential tag for the SOLiD sequencer) at the end of V1 tag. Then the P1-adaptor (another essential tag for the SOLiD sequencer) was added by ligation at the internal adaptor-free end. Finally, the index sequence and P2 tag (tag for the SOLiD sequencer) were added by PCR, and the library DNAs were obtained.

The quality and quantity of the constructed libraries were evaluated by using a LabChip GX or Bioanalyzer 2,100 system, a Qubit dsDNA HS assay kit (Q32851; Thermo Fisher Scientific), and a SOLiD Library TaqMan Quantitation kit (4449639; Thermo Fisher Scientific). The clonal amplification of the libraries on beads by emulsion PCR was performed using a SOLiD EZ Bead System (4449639; Thermo Fisher Scientific) at the E120 scale according to the manufacturer's instruction. The resulting bead

libraries were loaded onto flowchips and sequenced for 50 and 5 bp barcode plus Exact Call Chemistry (ECC) on a SOLiD 5500XL system (4449639; Thermo Fisher Scientific).

Note that the SC3-seq method has now been modified to be applicable to the illumina sequencers. The detailed protocol has been recently published<sup>18</sup>.

### Reference sequences

The genome sequence for MacFas5.0 (*mfa\_ref\_Macaca\_fascicularis\_5.0\_X.fa* (X: chr1 - chr20, chrX)) and transcript definition file (GFF3) for MacFas5.0 (*ref\_Macaca\_fascicularis\_5.0\_top\_level.gff3*) were obtained from the NCBI ftp site. ERCC spike-in sequences were obtained from the Thermo Fisher Scientific website ([https://tools.thermofisher.com/content/sfs/manuals/cms\\_095047.txt](https://tools.thermofisher.com/content/sfs/manuals/cms_095047.txt)).

### Modification of transcript definition file

The SC3-seq protocol enriches ~300 bp from the 3'-end of transcription termination sites (TTSs). Therefore, an inaccurate definition of a TTS, especially one that defines TTS as too far short of the appropriate site, could result in an absence of expression values, even if there are reads accumulated further downstream. To ensure that all possible transcript signals are covered, we extended all 3'-end of TTSs by up to 10 kb in the transcript definition file (GFF3 file, all 'gene', 'transcript' and 'exon' type category, Supplementary File 1) by applying the following rules according to the previous report<sup>10</sup>:

1. Identify genes sharing identical TTSs. Remove all these genes but one at each TTS position.
2. Extend TTS by 10 kb if no gene is found within the targeted area.
3. If the extended TTS hits a transcription start site (TSS) of a downstream gene (same strand), stop the extension at 1 bp upstream of the TSS.
4. If the extended TTS hits a TTS of a downstream gene (opposite strand), stop the extension at the mid of two TTS.
5. Remove all pseudogenes (marked as 'pseudo=True'), tRNAs or ncRNAs.

Since the gene names in these MacFas5.0 annotations were not fully annotated like their human counterparts, we searched for genomic coordinates of the MacFas5.0 genes on the human genome, hg19, using the LiftOver tool of UCSC, and matched the MacFas5.0 gene name with the human one if the names were different<sup>12</sup>.

### Read processing, mapping and conversion to expression values

All reads were pre-processed by cutadapt v1.11 with '-c -e 0.1 -q 20 -n 2 -O 1 -m 30' options and '-a' and '-g' options for sequence CTCGAGGGCGCGCCGGATCCATATACGCCTTGGCCGTACAG-CAG, and -a option for sequence A(20) to remove adaptor and poly-A sequences, and low quality bases<sup>19</sup>. Untrimmed and trimmed reads of 30 bp or longer were mapped onto the MacFas5.0 genome and ERCC spike-in RNA sequences with tophat v2.0.11/bowtie1.0.1 with '-bowtie1 -C -no-coverage search' options<sup>20</sup>. Mapped reads on the genome and ERCC in the bam file were separated and reads on the genome were processed with cufflinks v2.2.0 with '-compatible-hit-norm --no-length-correction--library-type fr-secondstrand' options and Maffas5.0 reference gene annotation with extended TTSs<sup>21</sup>.

### Data Records

The raw csfasta and QV.qual files were deposited in the Gene Expression Omnibus (GEO) database under acquisition numbers GSE67259, GSE76267 and GSE74767 (Data Citation 1, Data Citation 2 and Data Citation 3). The deposited data also contain the abundance of processed expression data, including the Entrez gene IDs and gene names. The sample information is summarized in Table 1 (available online only).

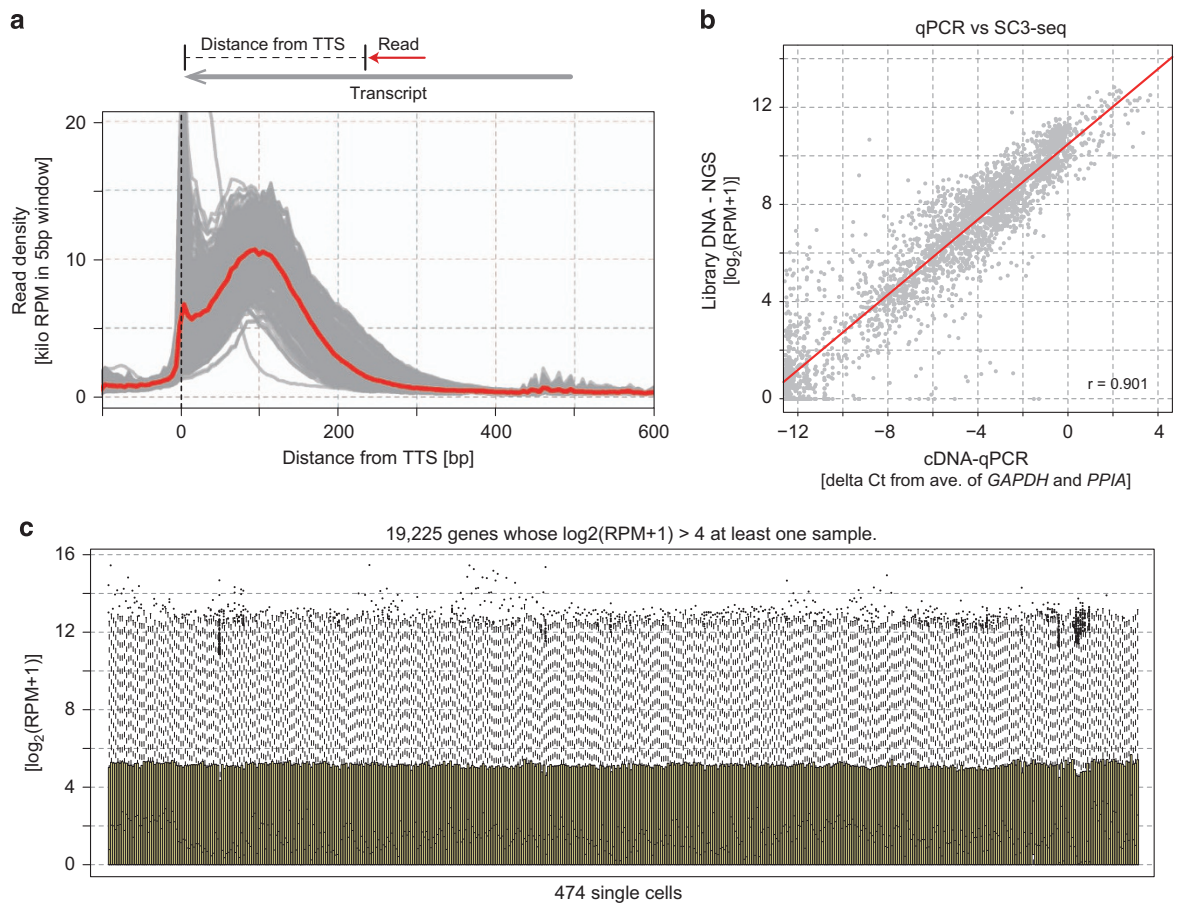
### Technical Validation

#### Verification of the qualities of SC3-seq data

A total of 474 cDNAs from each single cells were sequenced to a depth of between 0.64–4.73 mega mapped reads (Table 1 (available online only)). According to the previous study<sup>10</sup>, 0.5 mega reads were sufficient for an accurate analysis by SC3-seq. The positions of mapped reads were significantly enriched at the very 3' ends of mapped genes (Fig. 2a). The comparison of data between cDNAs and subsequent library DNAs indicated that the expression profiles were highly conserved during the experiments (Fig. 2b). Finally, in order to validate the quality of the SC3-seq libraries, we evaluated the profile of expression levels of all the expressed genes in at least one cell (Fig. 2c). Although the single-cell transcriptome is affected by both technical and biological variation, the expression values of the 75th percentile gene showed uniform distribution, indicating that all samples were successfully generated with appropriate qualities.

#### Clustering and annotating of the cells

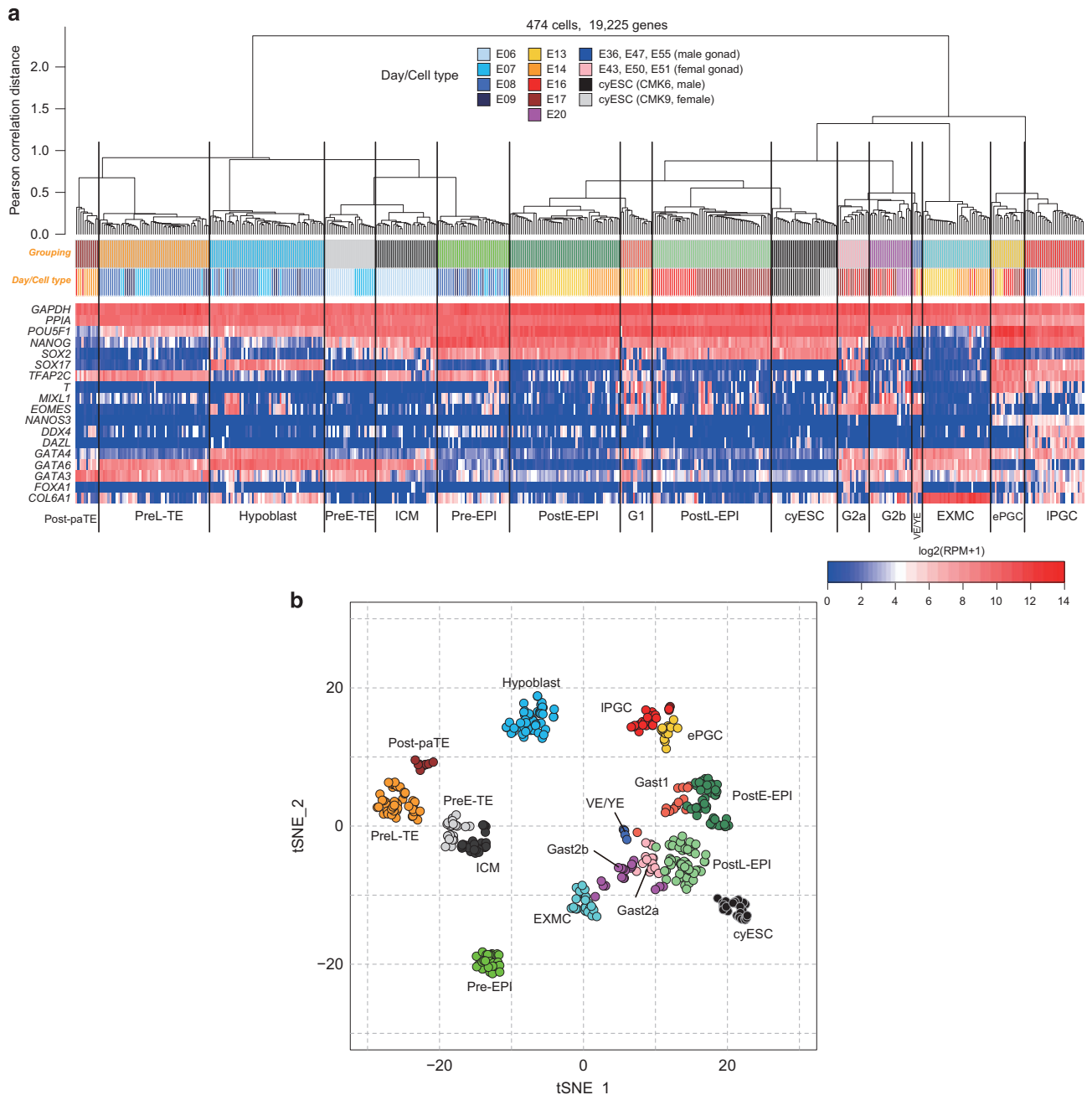
The cells were grouped by unsupervised hierarchical clustering (UHC) using the hclust function of R software with Pearson correlation distance and Ward's method (ward.D2) (Fig. 3a), and t-SNE analysis using the Rtsne function in the Rtsne package with default parameters<sup>22</sup> (Fig. 3b). Their annotations were defined by comparing the expression of key genes in transcriptome data with that obtained through the histological analysis<sup>11,12</sup> and are summarized in Table 1 (available online only).



**Figure 2. Verification of the qualities of SC3-seq data.** (a) The SC3-seq track [read density (kilo RPM) plotted against the read position from the annotated TTSs] of all cells (gray) and the average of all cells (red) as indicated above. (b) Scatter plot comparison of the gene expressions acquired by qPCR [delta Ct] and by NGS analysis [ $\log_2(\text{RPM}+1)$ ]. The delta Ct values of qPCR analysis were calculated from mean values of *GAPDH* and *PPIA*. The regression line is shown in red, and the correlation coefficient ( $r$ ) is indicated at the bottom right of the plot. (c) Distribution of gene expression levels of all expressed genes. All the expressed genes (19,225 genes) were defined as the genes whose  $\log_2(\text{RPM}+1)$  values were more than 4 in at least one sample among all 474 cells. The bars at the top, middle and bottom of the box indicate the 75th, 50th, and 25th percentile expression levels, and the top bar of the whisker indicates the expression levels encompassing the expression of 2 s.d. from the median of the genes, respectively.

In UHC analysis, all cells were classified into two large clusters: 6 clusters mainly from pre-implantation embryos and 10 clusters from post-implantation embryos, genital ridges and ESCs (Fig. 3a). In the left side of the dendrogram in Fig. 3a, the green, blue and orange groups of the pre-implantation cluster consisted of cells from E7 to E9, and each of them expressed the key genes for the pre-implantation epiblast (Pre-EPI) (*NANOG* and *SOX2*), hypoblast (*SOX17*, *GATA4* and *GATA6*) and late trophoctoderm (PreL-TE) (*TFAP2C*, *GATA3* and *GATA6*) according to the histological analysis<sup>12</sup>. On the other hand, the light gray and dark gray groups in the pre-implantation cluster consisted of the cells from E6 and E7 only, and while the light gray group expressed a high level of trophoctoderm-related genes, the dark gray group did not. Consistent with this, in the t-SNE analysis, the cells in the light gray cluster were plotted close to the PreL-TE, while those in the dark gray cluster were located on the side of Pre-EPI and Hypoblast (Fig. 3b). Therefore we annotated them as pre-implantation early trophoctoderm (PreE-TE) and inner cell mass (ICM) cells. The brown group, which was derived from post-implantation embryos but located close to PreL-TE in the pre-implantation cluster (Fig. 3a,b), had a similar gene expression pattern to PreL-TE (Fig. 3a), suggesting that the members of the brown group are the derivative of the PreL-TE. Therefore we annotated these cells as post-implantation parietal trophoctoderm (Post-paTE) cells.

In the other large cluster in UHC dendrogram (Fig. 3a), the PGCs from genital ridges (late PGCs; IPGCs) were grouped into one cluster (red), and the cells in yellow clusters from the early post-implantation embryo (E13–20) were annotated as early PGCs (ePGCs) because they were clustered



**Figure 3. Clustering and annotation of the cells.** (a) Unsupervised hierarchical clustering (UHC) with all the expressed genes and a heat map of the levels of selected marker genes. The colored bars under the dendrogram indicate grouping (top) and embryonic days/cell type (bottom), respectively. Post-paTE, post-implantation parietal trophoblast; PreL-TE, pre-implantation late TE; PreE-TE, pre-implantation early TE; ICM, inner cell mass; Pre-EPI, pre-implantation epiblast; PostE-EPI, post-implantation early epiblast; G1, gastrulating cells group 1; PostL-EPI, post-implantation late epiblast; G2a/G2b, gastrulating cells group 1/2; VE/YE, visceral endoderm/yolk sac endoderm; EXMC, extraembryonic mesenchyme; ePGC, early PGC; IPGC, late PGC. (b) Plot of two-dimensional t-SNE analysis with all the expressed genes. The color codes is as indicated.

with IPGCs both in the UHC dendrogram and t-SNE plot, and also shared characteristic gene expression patterns with IPGCs (Fig. 3a,b). Next, the other clusters of post-implantation cells were classified into two groups according to their *POU5F1* expression (Fig. 3a); *POU5F1* is expressed in embryonic cells during the early post-implantation embryo stage<sup>12</sup>. The light and dark blue groups were annotated as extraembryonic mesenchyme (EXMC) and visceral endoderm/yolk sac endoderm (VE/YE) cells because they showed high and uniform expression of *COL6A1* and *FOXA1* (Fig. 3a), whose expression

patterns were confirmed by histological analysis<sup>12</sup>. The remaining clusters were classified as EPI [post-implantation early epiblast, PostE-EPI (E13, 14); post-implantation late epiblast, PostL-EPI (E16,17)] or gastrulating cells (Gast1, 2a, 2b) due to the expression of the pluripotency-associated genes and differentiation-related genes. We could not provide a detailed explanation of the cell types of gastrulating cells due to the highly variable expression of the differentiation-related genes. Consistent with this, the cells in Gast2a and Gast2b were not separated clearly in the t-SNE analysis (Fig. 3b). Both the male and female ESCs (CMK6 and CMK9) were clustered close to PostL-EPI (Fig. 3a,b).

## References

- Nichols, J. & Smith, A. Naive and primed pluripotent states. *Cell Stem Cell* **4**, 487–492 (2009).
- Boroviak, T., Loos, R., Bertone, P., Smith, A. & Nichols, J. The ability of inner-cell-mass cells to self-renew as embryonic stem cells is acquired following epiblast specification. *Nat. Cell Biol.* **16**, 516–528 (2014).
- Brons, I. G. M. *et al.* Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature* **448**, 191–195 (2007).
- Kojima, Y. *et al.* The Transcriptional and Functional Properties of Mouse Epiblast Stem Cells Resemble the Anterior Primitive Streak. *Cell Stem Cell* **14**, 107–120 (2014).
- Tesar, P. J. *et al.* New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448**, 196–199 (2007).
- Saitou, M., Kagiwada, S. & Kurimoto, K. Epigenetic reprogramming in mouse pre-implantation development and primordial germ cells. *Development* **139**, 15–31 (2012).
- Hayashi, K., Ohta, H., Kurimoto, K., Aramaki, S. & Saitou, M. Reconstitution of the mouse germ cell specification pathway in culture by pluripotent stem cells. *Cell* **146**, 519–532 (2011).
- Irie, N. *et al.* SOX17 is a critical specifier of human primordial germ cell fate. *Cell* **160**, 253–268 (2015).
- Sasaki, K. *et al.* Robust In Vitro Induction of Human Germ Cell Fate from Pluripotent Stem Cells. *Cell Stem Cell* **17**, 178–194 (2015).
- Nakamura, T. *et al.* SC3-seq: a method for highly parallel and quantitative measurement of single-cell gene expression. *Nucleic Acids Res.* **43**, e60 (2015).
- Sasaki, K. *et al.* The Germ Cell Fate of Cynomolgus Monkeys Is Specified in the Nascent Amnion. *Dev. Cell* **39**, 1–17 (2016).
- Nakamura, T. *et al.* A developmental coordinate of pluripotency among mice, monkeys and humans. *Nature* **537**, 57–62 (2016).
- Seita, Y. *et al.* Generation of transgenic cynomolgus monkeys that express green fluorescent protein throughout the whole body. *Sci. Rep.* **6**, 24868 (2016).
- Yamasaki, J. *et al.* Vitrification and transfer of cynomolgus monkey (*Macaca fascicularis*) embryos fertilized by intracytoplasmic sperm injection. *Theriogenology* **76**, 33–38 (2011).
- Suemori, H. *et al.* Establishment of embryonic stem cell lines from cynomolgus monkey blastocysts produced by IVF or ICSI. *Dev. Dyn.* **222**, 273–279 (2001).
- Watanabe, K. *et al.* A ROCK inhibitor permits survival of dissociated human embryonic stem cells. *Nat. Biotechnol.* **25**, 681–686 (2007).
- Nakagawa, M. *et al.* A novel efficient feeder-free culture system for the derivation of human induced pluripotent stem cells. *Sci. Rep.* **4**, 3594 (2014).
- Ishikura, Y. *et al.* In Vitro Derivation and Propagation of Spermatogonial Stem Cell Activity from Mouse Pluripotent Stem Cells. *Cell reports* **17**, 2789–2804 (2016).
- Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).
- Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
- Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
- Van Der Maaten, L. J. P. & Hinton, G. E. Visualizing high-dimensional data using t-sne. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

## Data Citations

- Nakamura, T. *et al.* *Gene Expression Omnibus* GSE74767 (2016).
- Sasaki, K. *et al.* *Gene Expression Omnibus* GSE76267 (2016).
- Sasaki, K. *et al.* *Gene Expression Omnibus* GSE67259 (2015).

## Acknowledgements

This work was supported in part by a Grant-in-Aid from MEXT and by JST-ERATO. We thank Y. Nagai, R. Kabata, N. Konishi, Y. Sakaguchi, M. Kasawaki, T. Sato, M. Kabata, T. Yamamoto, J. Matsushita, C. Iwatani, H. Tsuchiya, Y. Seita, S. Nakamura, and M. Matsutani for their technical assistance. We are grateful to H. Suemori for the provision of CMK6/9, to M. Ema for encouragement, and to the animal care staff at the Research Center for Animal Life Science, Shiga University of Medical Science for their assistance. This work was supported by JST ERATO Grant Numbers JPMJER1104, Japan.

## Author Contributions

T.N. and M.S. conceived the project, designed the experiments and wrote the manuscript. T.N. conducted all the experiments and analyzed the data, I.O. and K.S. assisted with the SC3-seq experiments, and Y.Y. assisted in the process of analyzing SC3-seq data.

## Additional Information

Tables 1, 2 and 3 are only available in the online version of this paper.

**Competing interests:** The authors declare no competing financial interests.

**How to cite this article:** Nakamura, T. *et al.* Single-cell transcriptome of early embryos and cultured embryonic stem cells of cynomolgus monkeys. *Sci. Data* **4**:170067 doi: 10.1038/sdata.2017.67 (2017).



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



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

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files made available in this article.

© The Author(s) 2017