nature
structural &
molecular biology

Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells

Liying Yan[1,2,5], Mingyu Yang[1,5], Hongshan Guo[1], Lu Yang[1], Jun Wu[1], Rong Li[1,2], Ping Liu[1], Ying Lian[1], Xiaoying Zheng[1], Jie Yan[1], Jin Huang[1], Ming Li[1], Xinglong Wu[1], Lu Wen[1], Kaiqin Lao[3], Ruiqiang Li[1,4], Jie Qiao[1,2] & Fuchou Tang[1]

Measuring gene expression in individual cells is crucial for understanding the gene regulatory network controlling human embryonic development. Here we apply single-cell RNA sequencing (RNA-Seq) analysis to 124 individual cells from human preimplantation embryos and human embryonic stem cells (hESCs) at different passages. The number of maternally expressed genes detected in our data set is 22,687, including 8,701 long noncoding RNAs (lncRNAs), which represents a significant increase from 9,735 maternal genes detected previously by cDNA microarray. We discovered 2,733 novel lncRNAs, many of which are expressed in specific developmental stages. To address the long-standing question whether gene expression signatures of human epiblast (EPI) and *in vitro* hESCs are the same, we found that EPI cells and primary hESC outgrowth have dramatically different transcriptomes, with 1,498 genes showing differential expression between them. This work provides a comprehensive framework of the transcriptome landscapes of human early embryos and hESCs.

The identity and behavior of a cell is determined by its gene expression network, which is regulated by genetic and epigenetic mechanisms. Hence, deciphering the temporal and spatial patterns of gene expression in human embryos is a crucial step toward understanding early developmental processes[1–4]. Pioneering work in transcriptional profiling has been carried out by either multiplex quantitative PCR (qPCR) or microarray technology using multiple human embryos, single embryos or single blastomeres at various preimplantation stages[5–15]. These studies have identified genes involved in important developmental events, such as maturation of oocytes[6], maternal-zygotic transition[10], embryonic genome activation[13] and segregation of inner cell mass and trophectoderm[7,9,11]. For example, the gene expression of human embryos at the 8-cell stage has been compared with that of human ESCs and fibroblasts to uncover unique cell-cycle control patterns of 8-cell blastomeres[16,17]. It has also been found that during the first 3 d of embryonic development, the expression of 1,896 genes changes substantially, with most being downregulated[10]. Moreover, Xie *et al.* have shown that 40.2% of orthologous gene triplets show different expression patterns among mouse, bovine and human[15]. However, these previous studies recovered only small sets of relevant markers limited to known protein-coding genes[9–11,13,14]. To fully understand embryonic development, it would be highly informative to generate comprehensive whole transcriptomes that include both known and novel protein-coding genes and lncRNAs[18]. Furthermore, to make it possible to characterize the genetic basis of cell fate determination and coordination among individual embryonic cells, data generation and analyses of gene expression should be done at the single-cell level.

The same strategy and technology should also be applied to the genetic characterization of hESCs. HESCs, derived from the ICM of human blastocyst-stage embryos, have greatly facilitated study of the mechanisms underlying pluripotency—the potential of a cell to differentiate into all cell types in the adult body[19]. However, the extent to which hESCs serve as an accurate model of early embryos remains unclear[20,21]. Because of technical and analytic difficulties, the differences in gene expression between the ICM *in vivo* and hESCs *in vitro* are still largely unknown. In addition, the dynamics of gene expression during the process of hESC derivation have not been documented.

Recently, several single-cell RNA-Seq techniques have been developed, making it feasible to analyze the transcriptome of human embryos at single-cell and single-base resolution[22–25]. Here, using the single-cell RNA-Seq technique[23], we sequenced the transcriptome of human oocytes and early embryos at seven developmental stages—from metaphase II mature oocytes to late blastocysts—and of the primary outgrowth during hESC derivation (hESCs of passage 0) and hESCs of passage 10 (refs. 22,26) (**Table 1**). In total, we analyzed 124 single cells. The sequencing reads mapped to the human genome but not to any known transcripts for each of these cells were then assembled into transcripts and compared with known protein-coding genes or noncoding RNAs in the RefSeq, Ensembl and Noncode v3.0 lncRNA databases. This process yielded 253 novel protein-coding

[1]Biodynamic Optical Imaging Center and Center for Reproductive Medicine, College of Life Sciences, Third Hospital, Peking University, Beijing, China. [2]Key Laboratory of Assisted Reproduction, Ministry of Education, Beijing, China. [3]Genetic Systems, Applied Biosystems, Life Technologies, Foster City, California, USA. [4]Peking-Tsinghua Center for Life Sciences, College of Life Sciences, Peking University, Beijing, China. [5]These authors contributed equally to this work. Correspondence should be addressed to R.L. (lirq@pku.edu.cn), J.Q. (jie.qiao@263.net) or F.T. (tangfuchou@pku.edu.cn).

Received 18 July; accepted 5 August; published online 11 August 2013; doi:10.1038/nsmb.2660

Table 1 Numbers of embryos and cells analyzed by single-cell RNA-Seq analysis

Stage	No. of embryos	No. of single cells
Oocyte	3	3
Zygote	3	3
2-cell	3	6
4-cell	3	12
8-cell	3	20
Morula	2	16
Late blastocyst	3	30
hESC passage 0	NA	8
hESC passage 10	NA	26
Total	20	124

NA, not applicable.

transcripts and 2,733 novel lncRNAs that had not been previously identified in human cells of any type.

Unsupervised hierarchical clustering showed that global gene expression profiles of individual cells at the same stage, from oocyte through blastocyst, are very similar, whereas the profiles at different stages of development differ significantly. Single-cell analysis further revealed insights that could not have been gained through ensemble analysis, such as the separation of the EPI and primitive endoderm (PE) lineages of cells within a blastocyst, and alternative splicing patterns within individual cells.

RESULTS

We used donated oocytes and embryos to perform RNA-Seq analysis of 90 individual cells from 20 oocytes and embryos (**Table 1**). The 20 samples were chosen from 226 cultured oocytes and embryos on the basis of stringent morphological criteria, as described in detail in the Online Methods. The embryos were at seven crucial stages of preimplantation development: metaphase II oocyte, zygote, 2-cell, 4-cell, 8-cell, morula and late blastocyst at hatching stage. Notably, we recovered all the individual blastomeres of three 2-cell, three 4-cell and two 8-cell embryos for analysis (**Fig. 1a**,**b**). We also obtained 8 individual cells from the primary hESC outgrowth (hESCs of passage 0) and 26 individual cells from hESCs of passage 10 (**Fig. 1a**,**b**). Using the Illumina HiSeq2000 sequencer, we generated 438 Gb of sequencing data from the 124 single cells, with, on average, 35.3 million reads per cell with read length of 100 bp.

Transcriptional profiles across different cell types

We first analyzed how many known genes were expressed in each of the 90 embryonic cells. On average, we detected expression of 11,006 (49%) out of 22,092 RefSeq genes (**Supplementary Table 1**) and of 18,022 (48%) out of 36,983 RefSeq transcripts. Thus, approximately half of the known human genes and transcripts were expressed in the sampled oocytes and embryo blastomeres (**Supplementary Table 2**).

To determine whether these gene expression profiles correlated with developmental stages, we analyzed RNA-Seq data of the oocytes and embryonic cells by unsupervised hierarchical clustering. Cells that clustered together were at the same developmental stages in all cases, with the exception that two blastomeres of a morula-stage embryo clustered with those of the blastocysts (**Fig. 2a**). Moreover, the developmental order was also accurately captured from mature oocytes to late blastocysts, as neighboring stages cluster together in the analysis as expected, similar to what has been previously reported[15]. The greatest changes in gene expression were seen between the 4- and 8-cell stages, which may be explained by a major maternal-zygotic transition[27–29] (**Fig. 2a** and **Supplementary Fig. 1**). Similar within-stage and different between-stage expression patterns were also supported by principal-component analysis (PCA) (**Fig. 2b**). Another dramatic difference occurs between EPI cells of blastocysts and all other cells from embryos after the 8-cell stage (as discussed later). As many as 3,906 genes showed differential expression between EPI and the remaining cells. We carried out Gene Ontology (GO) analysis for these differentially expressed genes and found that they are enriched for GO terms related to transcriptional regulation and germ cell development, indicating that EPI cells have lower expression of genes related to gamete generation, germ cell development and reproduction compared to the other cell lineages in blastocysts (**Supplementary Table 3**).

Interestingly, one of the embryos at the morula stage clustered with an 8-cell-stage embryo, indicating that the transcriptomes of these two stages of embryos are still relatively similar. This was also supported by the relatively short distance between these two groups of embryos in the PCA analysis. The majority of expressed genes showed stage-specific expression patterns and could be clustered into eight distinct groups (**Fig. 2c**).

We then compared the transcriptional profiles of single cells within each of the seven developmental stages. Whereas expression patterns are generally rather homogeneous among cells of the same stage, blastomeres of the 8-cell embryos were more variable in the expression of downregulated genes (FC[8-cell/4-cell] < 0.5, P < 0.01) than in that of upregulated genes (FC[8-cell/4-cell] > 2, P < 0.01), suggesting that the degradation of maternal transcripts may be heterogeneous among 8-cell-stage blastomeres. To our knowledge, this is the first time that the transcriptomes of all of the blastomeres within the same 8-cell-stage embryo have been analyzed simultaneously.

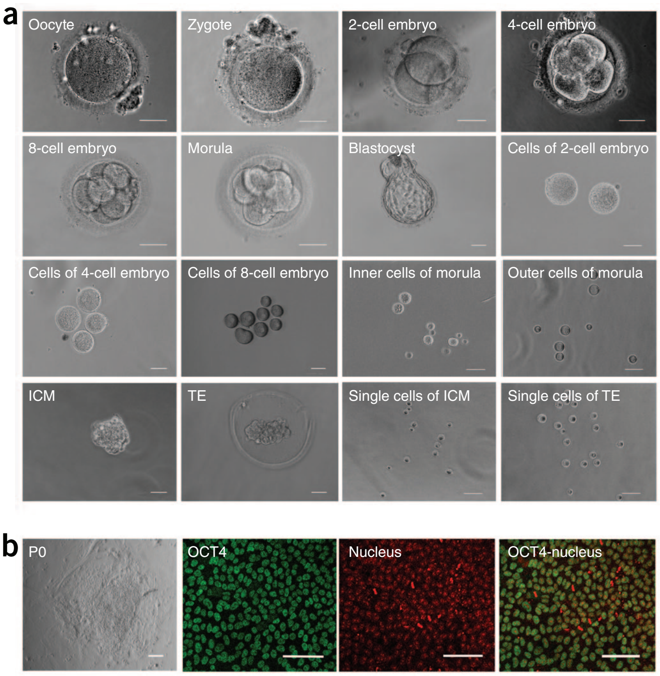

Figure 1 Morphology and marker gene expression of human early embryos and hESCs. (**a**) Microscopy imaging of mature human oocytes and preimplantation embryos at zygote (2PN), 2-cell, 4-cell, 8-cell, morula and late blastocyst stages and corresponding isolated single blastomeres from these embryos. Note the full recovery of all of the individual blastomeres of 2-cell, 4-cell and 8-cell embryos. Scale bar, 50 μm. (**b**) Microscopy imaging of colonies of hESCs at passage 0 (P0). Immunostaining of *OCT4* (green) and corresponding nuclear staining (PI staining; red), alone and merged, of hESCs at passage 10 are also shown. Scale bar, 100 μm.

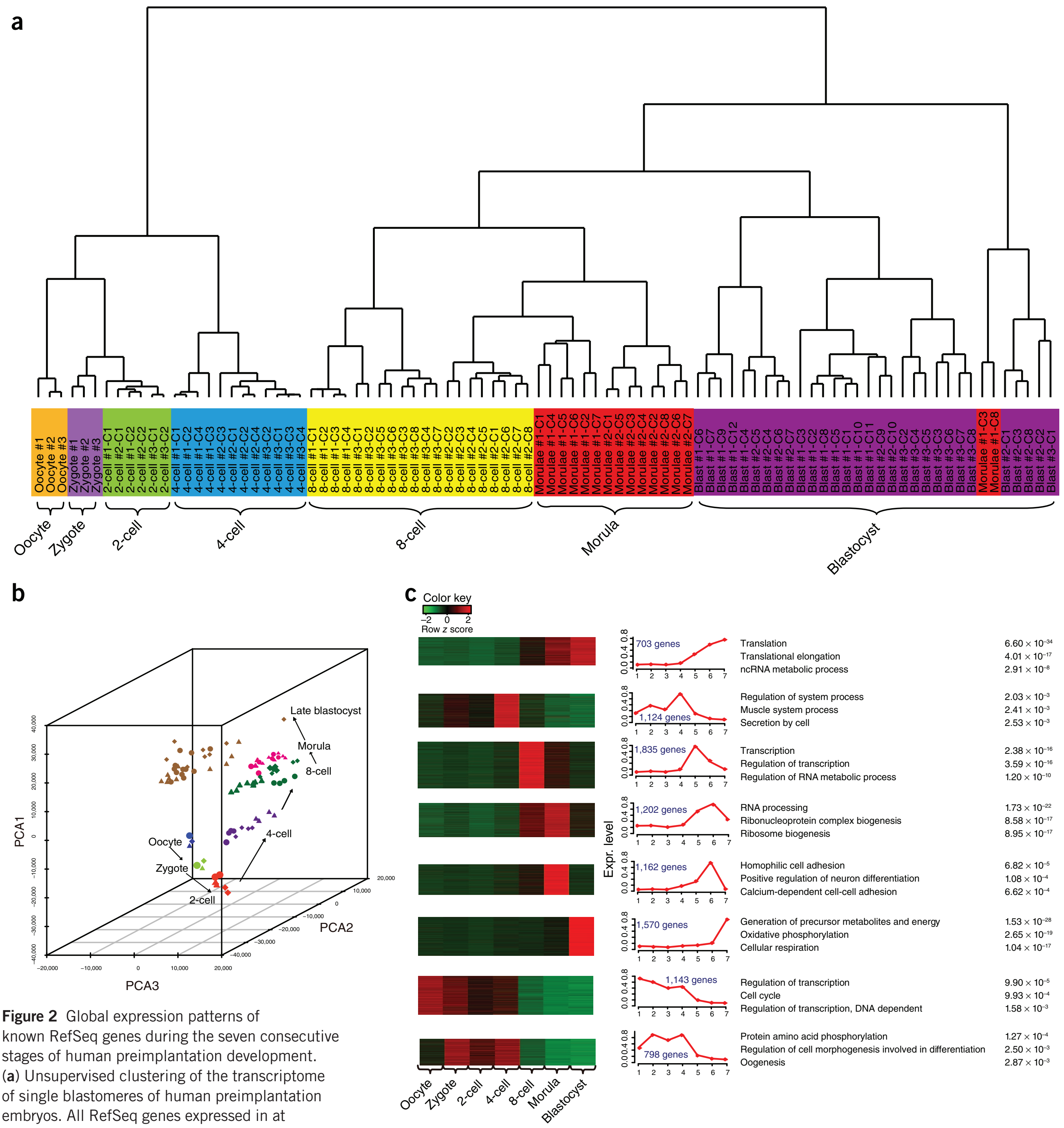

Figure 2 Global expression patterns of known RefSeq genes during the seven consecutive stages of human preimplantation development. (**a**) Unsupervised clustering of the transcriptome of single blastomeres of human preimplantation embryos. All RefSeq genes expressed in at least one of the samples with RPKM ≥ 0.1 were used for the analysis. (**b**) Principal-component analysis (PCA) of the transcriptome of single blastomeres of human preimplantation embryos. Blastomeres from the same embryo are shown as symbols of the same shape. The arrows indicate the developmental direction between consecutive stages of the embryos. PCA1, PCA2 and PCA3 represent the top three dimensions of the genes showing differential expression among these preimplantation blastomeres, which accounts for 43.8%, 16.1% and 9.8% of the expressed RefSeq genes, respectively. (**c**) Clusters of genes showing representative expression patterns during human preimplantation development. First, we selected all of the genes that were differentially expressed between any two consecutive stages (fold change > 2 or < 0.5, $P < 0.01$). Then, we used the SOTA function in the clValid package to classify these genes into 25 categories. Finally, we chose the eight categories with the most significant variations, as shown here. The top GO terms and corresponding enrichment *P* values are shown on the right side. Statistical analyses are described in the Online Methods.

In human embryos the major maternal-zygotic transition happens at the 8-cell stage[27]. We observed significant upregulation of 2,495 genes (zygotic gene activation (ZGA); FC[8-cell/4-cell] > 2, $P < 0.001$) and significant downregulation of 2,675 genes (FC[8-cell/4-cell] < 0.5, $P < 0.001$) between the 4-cell and 8-cell stages (**Supplementary Table 1** and **Supplementary Fig. 1a,b**). For the upregulated (that is, zygotically

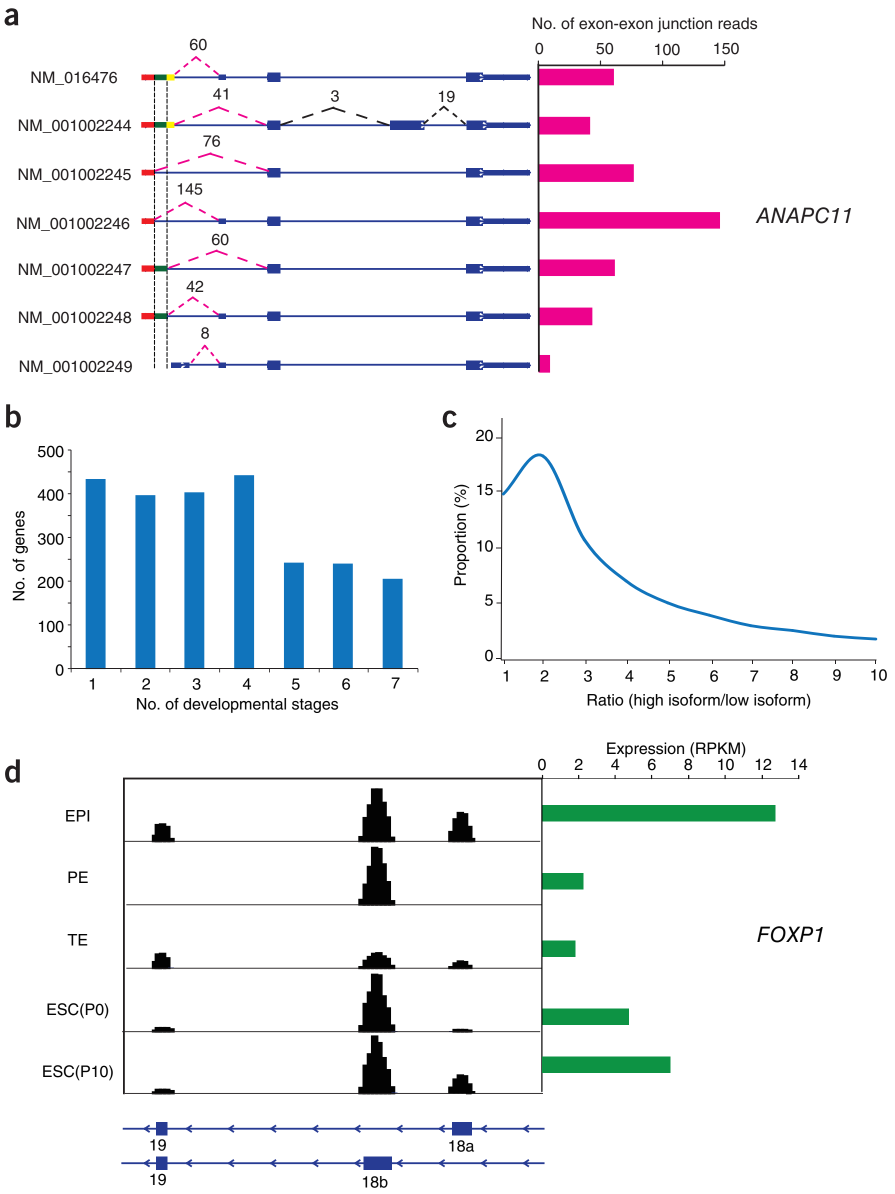

Figure 3 Dynamic patterns of alternative splicing during the seven consecutive stages of human preimplantation development and derivation of hESCs. (**a**) Exon-exon junction plots of all of the seven transcript variants of *ANAPC11* in an individual hESC. Note that these transcript variants will be translated into two different proteins, as they contain two different open reading frames (ORFs). (**b**) The developmental stage–specific expression pattern of the genes with two or more different transcript isoforms detected simultaneously within an individual cell. The *x* axis represents the number of developmental stages during which a gene expresses multiple transcript isoforms within an individual cell (at one developmental stage, if more than 50% of the cells expressed two or more transcript isoforms of a gene within individual cells, the gene is considered to express multiple isoforms at this stage). (**c**) The percentage of genes that express two different isoforms at different ratios. The *x* axis represents the ratio of high-expression isoform to low-expression isoform of a gene within an individual cell. (**d**) Expression dynamics of different transcript isoforms of *FOXP1* in three lineages of late blastocysts and hESCs. The corresponding expression values (RPKM) for the sum of the transcript isoforms are shown at right. TE, trophectoderm; PE, primitive endoderm; EPI, epiblast; ESC, embryonic stem cell; P0, P10, passage 0 or 10.

expressed) genes, there was a clear enrichment for genes whose products are involved in RNA metabolism and translation, such as RNA processing ($P = 5.5 \times 10^{-51}$), RNA splicing ($P = 3.5 \times 10^{-34}$), ribonucleoprotein complex biogenesis ($P = 2.4 \times 10^{-27}$) and ribosome biogenesis ($P = 5.6 \times 10^{-22}$), indicating that the zygotic-specific transcription and translation machinery is establishing. Genes related to chromosome organization ($P = 7.8 \times 10^{-8}$), cell division ($P = 1.2 \times 10^{-8}$) and DNA packaging ($P = 8.8 \times 10^{-7}$) were also strongly enriched, implying that the epigenetic and cell-cycle regulation are also shifting after the zygotic genes are activated. Interestingly, from the 2-cell to the 4-cell stage, there are already 983 genes upregulated, probably reflecting a minor zygotic gene activation wave, as previously reported (**Supplementary Fig. 1a**)[13].

Dynamic patterns of alternative splicing

Alternative splicing is an important mechanism that increases the complexity of the transcriptome and proteome. However, the global pattern of alternative splicing within single human cells remains unclear. We evaluated whether any known genes were expressed with two or more transcript isoforms within the same embryonic cell or hESC. On average, individual cells expressed 4,822 genes with at least two known isoforms. Based on reads that were uniquely mapped to transcript isoform–specific exons or exon-exon junctions, 981 (20%) of these 4,822 genes had more than two transcript isoforms that could be detected simultaneously within the same individual embryonic cell, with a maximum of seven isoforms detected (**Fig. 3a**, **Supplementary Table 4** and **Supplementary Note**).

Notably, the level of alternative splicing within individual cells varied between developmental stages. Among the 2,372 genes with two or more transcript isoforms in the same cell, 1,484 were in oocytes, 1,699 were in zygotes, 1,885 were in 2-cell embryos, 1,045 were in 4-cell embryos, 914 were in 8-cell embryos, 877 were in morulae and 424 were in late blastocysts. Only 206 genes were expressed with multiple transcript isoforms within the same individual cell at all of the seven developmental stages (**Fig. 3b**). These data suggest that the pattern of alternative splicing changed at different stages during preimplantation development. Finally, we analyzed the expression ratio of different transcript isoforms within individual cells. We found that for 66% of the cases, the expression of the major isoform is more than twofold higher than that of the minor isoform (FC[high isoform/low isoform] ≥ 2) (**Fig. 3c**); and only for 34% of the cases are the two isoforms expressed at comparable levels (FC[high isoform/low isoform] < 2). This indicates that for genes with multiple isoforms expressed simultaneously within an individual cell, one of the isoforms usually dominates that expression.

Recently, a forkhead box (FOX) transcriptional factor, *FOXP1*, has been shown to express an ESC-specific transcript isoform with an alternative exon 18 (exon 18b), and this isoform encodes another FOXP1

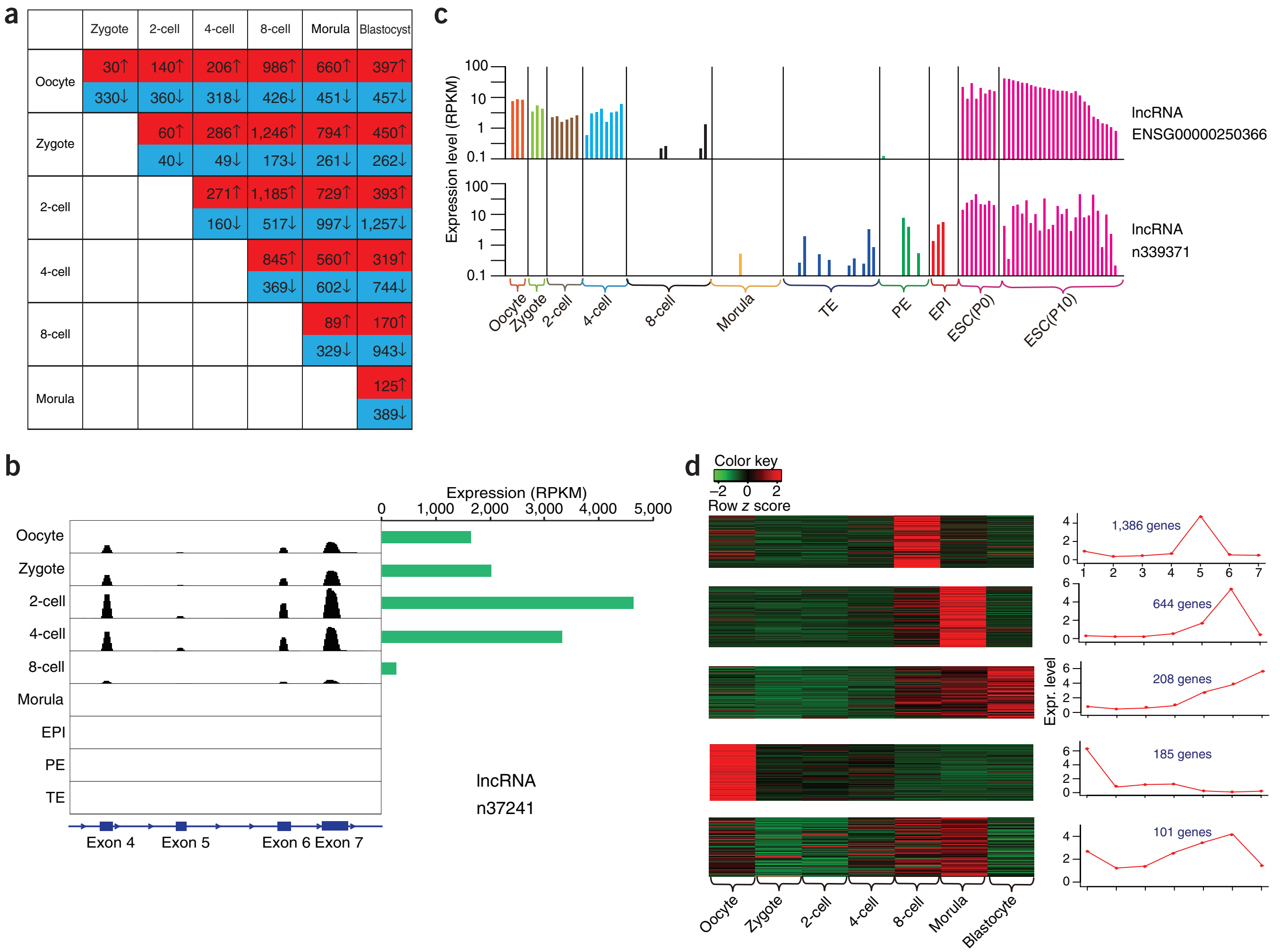

Figure 4 Expression patterns of known long noncoding RNA (lncRNA) genes during human preimplantation development and derivation of hESCs. (**a**) Number of known lncRNA genes showing up- or downregulation during preimplantation development (fold change > 2 or < 0.5, $P < 0.01$). (**b**) Coverage plots of RNA-Seq reads of a known lncRNA gene during preimplantation development. (**c**) Expression patterns of two lncRNA genes within individual cells during preimplantation development and derivation of hESCs. Each vertical bar represents the expression of the lncRNA in an individual cell. RPKM values are shown on the *y* axis. TE, trophectoderm; PE, primitive endoderm; EPI, epiblast; ESC, embryonic stem cell; P0, P10, passage 0 or 10. (**d**) Clusters of lncRNAs showing representative expression patterns during human preimplantation development. First, we selected all of the lncRNAs that were differentially expressed between any two consecutive stages (fold change > 2 or < 0.5, $P < 0.01$). Then, we used the SOTA function in the clValid package to classify these genes into nine categories. Finally, we chose the five categories with the most significant variations, as shown here. Statistical analyses are described in the Online Methods.

protein product targeting to different DNA sequences and is functionally crucial for maintenance of pluripotency[30]. We found that in the late blastocyst, the EPI cells express higher level of *FOXP1* than trophectoderm cells, with more than two-thirds being the ESC cell–specific isoform (with exon 18b), indicating that *FOXP1* is involved in the creation of pluripotency in EPI cells *in vivo*. In undifferentiated hESC cells *in vitro*, reads corresponding to exon 18b are 25 times more abundant than those of exon 18a (**Fig. 3d**), similar to what has been previously reported[30].

Dynamic expression of long noncoding RNAs

Recently, tens of thousands of long noncoding RNA (lncRNA) genes with diverse functional roles were discovered in the mouse and human genomes[18,31–34]. We investigated whether lncRNAs were expressed during human early embryonic development and detected a total of 18,383 (64%) out of 28,640 known human lncRNAs in the 90 single embryonic cells analyzed (**Fig. 4a**, **Supplementary Fig. 2a**,**b** and **Supplementary Table 5**). Individual cells expressed, on average, 3,337 (11%) of 28,640 lncRNAs, which accounts for 5.5% of the transcriptome of a human embryonic cell (**Supplementary Table 5**). A stage-specific expression pattern implies that lncRNA has a regulatory role during development. For example, the expression level of a maternally inherited lncRNA is variable between developmental stages and drastically reduced after the 4-cell stage. The reads clearly mapped specifically to the exon regions, confirming their identification as spliced mature RNAs (**Fig. 4b**). The heterogeneous expression patterns of two highly expressed lncRNAs within individual hESCs are also shown in **Figure 4c**.

Next, we analyzed our data from all 124 cells to compare the relative abundance of known lncRNAs and known protein-coding transcripts. When these data were pooled in an ensemble analysis, the averaged copy number of transcripts for a lncRNA gene was approximately one-tenth (10.1%) that of a protein-coding gene, which is consistent with a previous estimate[35]. However, within single cells, the average copy number of a lncRNA gene jumped to ~40.5% of the average copy number of a protein-coding gene. This implies that lncRNAs are relatively abundant within individual cells and are potentially important regulators of cellular phenotype.

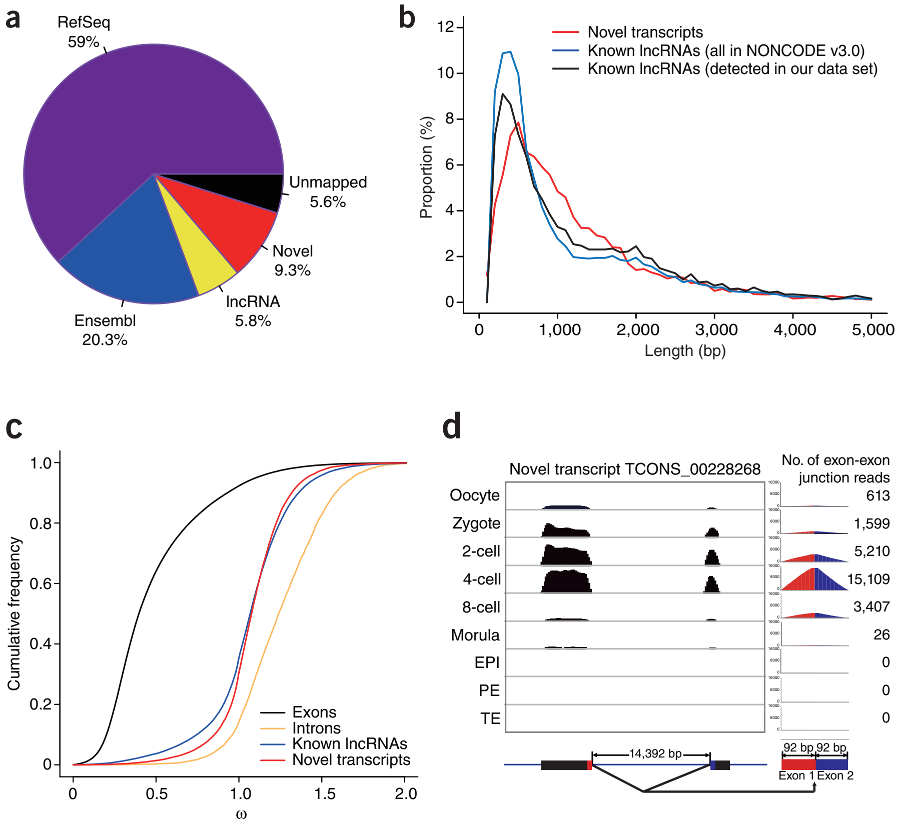

Figure 5 Expression patterns of novel lncRNAs during human preimplantation development. (**a**) Pie chart of the percentage of reads aligned to different classes of genes. (**b**) Length distribution of the 7,167 potential novel lncRNA transcripts. The known lncRNAs in NONCODE v3.0 database are used as a control (both the whole data set and the ones detected in our single-cell RNA-Seq data). (**c**) Conservation level (ω metric) of the 7,167 potential novel lncRNA transcripts[38]. This reflects the total contraction of the branch length of the evolutionary tree connecting the 46 mammalian species. (**d**) Coverage plots of RNA-Seq reads of a novel lncRNA during preimplantation development. The corresponding exon-exon junction reads are plotted at right, and the corresponding numbers of exon-exon junction reads are listed at far right.

To further evaluate whether the heterogeneity of lncRNA expression is likely to be functional or merely the consequence of leaky transcription, we looked specifically at the maternal-zygotic transition between the 4-cell and 8-cell stages, when the blastomeres within an embryo are still relatively homogenous. If the expression of a lncRNA gene were merely due to leaky transcription, it is very unlikely that the lncRNA would be expressed in all sampled embryos of the same stage. Among the 8,123 lncRNAs detected in at least one blastomere (reads per kilobase per million (RPKM) value ≥0.1) of the 12 blastomeres from the three 4-cell embryos, 4,200 (51.7%) were detected in all three embryos. Similarly, among the 10,594 lncRNAs detected in at least one blastomere of the three 8-cell embryos, 3,405 (32.1%) were detected in all three embryos. Many lncRNAs are variably expressed among cells of the same embryo but consistently present in different embryos of the same stage, which suggests that the lncRNA expression in human embryos is potentially functional. We carried out hierarchical clustering analysis and found that, like protein-coding genes, they show very distinct developmental stage–specific expression patterns, indicating their potential involvement in early embryonic development (**Fig. 4d**).

Novel protein-coding transcripts and lncRNAs

We searched across all 124 cells for transcripts that had not been reported in the RefSeq, Ensembl or Noncode v3.0 lncRNA databases. We used the 27.8 Gb of data that were directly or indirectly mappable to the human reference genome but had no overlap with sequences in these databases and carried out *de novo* transcript assembly using the reference-guided assembly software Cufflinks[36]. We removed all candidate transcripts within 10 kb upstream or downstream of any known genes, as these sequences could have derived from extended exons of the known genes. This analysis yielded 7,420 potential novel transcripts constituting 3,866 potential transcription units (the transcripts that from a genomic locus with the distance between transcripts less than 10 kb apart are called a transcription unit) that were at least 10 kb apart from each other (**Fig. 5a** and **Supplementary Table 6**). Using the CPC (Coding Potential Calculator)[37] software, we analyzed the coding potential of these novel transcripts and identified 253 possible protein-coding genes, while determining that the remaining 7,167 are most likely novel lncRNAs. The average length of these novel transcripts was 1,384 nucleotides, similar to that of known lncRNAs (**Fig. 5b**). Their conservation level (calculated with the metric ω) was also similar to that of known lncRNAs[38] (**Fig. 5c**). We further selected novel transcripts that (i) were >1 kb, (ii) comprised at least two exons and (iii) had detected exon-exon junction reads, generating a highly reliable set of 2,733 novel lncRNA transcripts for subsequent analyses (that is, ensuring that these lncRNA candidates were relatively long and had multiple exons). Moreover, 1,348 (49.3%) of the 2,733 novel lncRNA transcripts have intact 3′ ends, as determined by the presence of the canonical polyadenylation site (AAUAAA) (ref. 39) at around 100 bp from the 3′ ends, which in comparison is found in 22,327 (63.6%) of the 35,105 RefSeq transcripts.

Among the 2,733 novel lncRNAs, 1,197 (44%) were expressed variably across the seven embryonic stages listed in **Supplementary Table 6** (**Fig. 5d** and **Supplementary Fig. 2c**–**e**). Among the 2,181 novel transcripts detected in at least one blastomere of the 12 blastomeres from the three 4-cell embryos, 1,088 (49.9%) were detected in all three embryos. Similarly, among the 2,282 novel transcripts detected in at least one blastomere of the three 8-cell embryos, 974 (42.7%) were detected in all of the three 8-cell embryos. Although preliminary, these findings suggest that transcripts expressed in all the embryos at the 4-cell or 8-cell stages may be of functional significance—a conclusion that would be unachievable from ensemble data alone.

It has been shown by cDNA microarray analysis that 9,735 genes are maternally expressed in human MII oocytes[40]. Now, by using RNA-Seq combined with *de novo* RNA assembly, we have found that 13,986 RefSeq genes, 7,214 known lncRNAs and 1,487 novel lncRNAs were maternally expressed in human mature oocytes. That means that we detected 12,952 more maternally expressed genes, including 8,701 lncRNAs. To our knowledge, this is the first time that maternally expressed lncRNAs have been systematically analyzed.

Tracing pluripotency during the derivation of hESCs

HESCs are derived from epiblast cells in the inner cell mass of blastocysts[19], but differences between hESCs and epiblast cells have not been

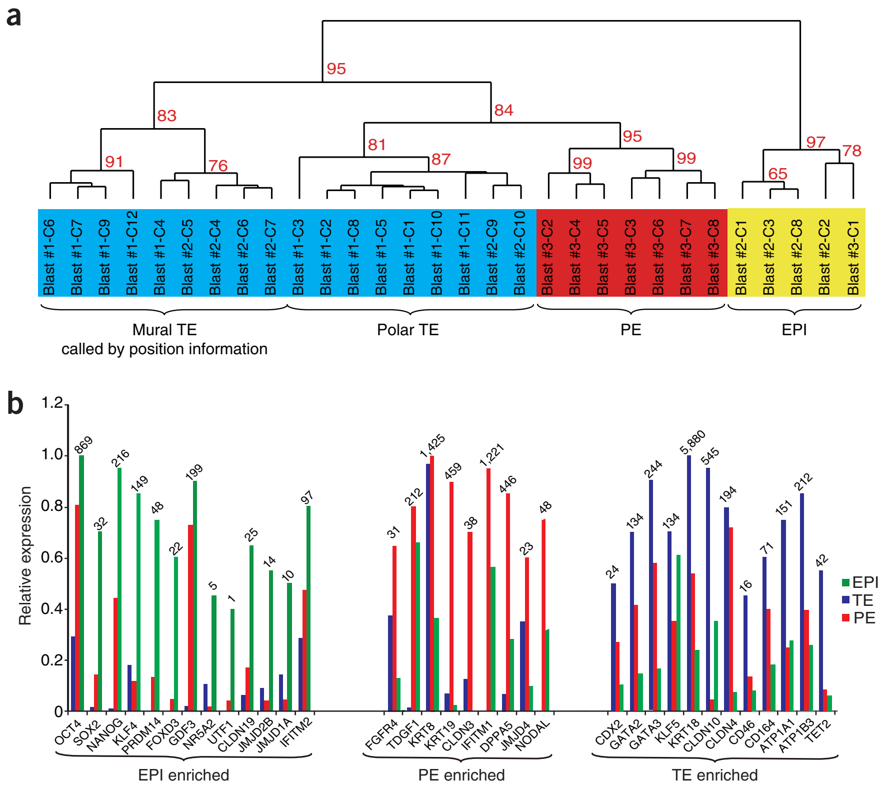

Figure 6 The EPI, PE and TE lineage segregation in the blastocysts. (**a**) Unsupervised clustering of the expression profiles of known RefSeq genes in individual blastomeres from late blastocysts. All of the RefSeq genes expressed in at least one of the samples with RPKM ≥ 0.1 were used for the analysis. To verify the accuracy of the clustering analysis, we used an R package named pvclust to do the bootstrap analysis to check the significance and robustness of the clustering[45]. We set nboot to 1,000 and clustered them independently for every time, and then, from these 1,000 independent clusterings, counted the percentage of times every branch was recovered. TE, trophectoderm; PE, primitive endoderm; EPI, epiblast. (**b**) Relative expression pattern of marker genes in the three lineages of late blastocysts. RPKM values of genes are labeled at the top of the highest bar for every gene. TE, trophectoderm; PE, primitive endoderm; EPI, epiblast.

comprehensively studied[20,21]. We analyzed our data on the 30 cells from three late blastocysts by unsupervised hierarchical clustering. The blastocyst-cell data separated clearly into three groups according to the specific transcript level of genes that have been previously shown to be characteristically expressed in trophectoderm (TE), epiblast (EPI) and primitive endoderm (PE): high expression of *CDX2* in TE; high expression of *SOX2*, *NANOG* and *KLF4* in EPI; and low expression of *SOX2*, *NANOG* and *KLF4*, but high expression of *FGFR4* and *CLDN3*, in PE.

In the pluripotent EPI lineage, 885 genes were uniquely expressed as compared with the TE lineage (**Fig. 6a**,**b**, **Supplementary Fig. 3a** and **Supplementary Table 7**). The PE lineage was relatively similar to the EPI lineage, as only 164 genes were uniquely expressed in PE compared with EPI (**Supplementary Fig. 3b** and **Supplementary Table 7**). The TE lineage uniquely expressed 493 genes compared with EPI and PE (**Supplementary Fig. 3c** and **Supplementary Table 7**). Several TE cells did cluster with PE cells; these were likely polar TE cells in physical proximity to EPI cells, similarly to PE cells, which are proximal or even mixed with EPI cells (**Fig. 6a**,**b**, **Supplementary Fig. 3d** and **Supplementary Fig. 4**). However, we cannot exclude the possibility that some of the PE cells in the 'ICM portion' were misclassified as TE cells. Although the blastocysts sampled at the same early hatching stage are morphologically similar, we cannot formally exclude the possibility that they may be at subtly different developmental stages, resulting in the variable allocation blastomeres to EPI and PE lineages.

To compare EPI cells and primary hESC outgrowth (hESCs of passage 0), we examined the expression of all known genes and of previously published pluripotency marker genes. We found that the majority of genes were expressed at comparable level between EPI cells and hESCs. (**Supplementary Table 8**). However, we also identified significant differences. From EPI to primary hESC outgrowth (called the "post-ICM intermediate" in ref. 21), 975 genes were upregulated (FC[hESC/EPI] > 2, $P < 0.01$) (**Supplementary Fig. 5a** and **Supplementary Table 8**). Notably, certain genes associated with pluripotency, such as those encoding the transcription factor genes *SALL2*, *TCF3*, *ZIC2* and *ZIC3* and the Id family genes (*ID1*, *ID2* and *ID3*), were strongly upregulated (to at least fivefold of the original values), perhaps contributing to the pluripotency of hESCs *in vitro*. In addition, the genes encoding the WNT signaling pathway receptor FZD7 and the FGF signaling cytokine FGF2 were also upregulated from EPI cells to primary hESC outgrowth.

The transition from EPI cells to primary hESC outgrowth also involved downregulation of 523 genes (FC[hESC/EPI] < 0.5, $P < 0.01$) (**Supplementary Table 1**). The downregulated genes included pluripotency-related genes such as *NANOG*, *GDF3*, *KLF4*, *KLF5*, *KLF17*, *ZFP57*, *ESRRB*, *PECAM1*, *STELLA* (also known as *DPPA3*), *DPPA2*, *DPPA5*, *IGF1* and *ABCG2* (**Supplementary Fig. 5a** and **Supplementary Table 8**). We carried out similar comparisons between passage 0 and passage 10 hESCs and found that only 851 genes showed differential expression between them (FC[passage 10/passage 0] > 2 or < 0.5, $P < 0.01$). Moreover, we did unsupervised hierarchical clustering analysis for EPI and hESCs. The passage 0 hESCs in general clustered together with passage 10 hESCs, and the EPI cells relatively separated from them (**Supplementary Fig. 5b**). Together, these analyses indicate that the main changes of gene expression signature during hESC derivation happen between EPI and primary outgrowth of hESCs, compatible with previous reports about further development of the ICM during the hESC derivation process[21].

We also carried out comparisons between EPI cells and primary hESC outgrowth with regard to known lncRNAs and novel transcripts. 138 known lncRNAs and 37 novel lncRNAs were expressed in EPI but not in hESCs, whereas 2,286 known lncRNAs and 194 novel lncRNAs were expressed in hESCs but not in EPI, again highlighting gene-expression differences between EPI and primary hESC outgrowth derived from them.

It has been proposed that hESCs are more similar to mouse epiblast stem cells (mEpiSCs), which derive from post-implantation epiblast cells, than to mouse embryonic stem cells[41–43] (mESCs), derived from blastocyst-stage preimplantation epiblast cells. We chose a set of pluripotency-related marker genes that can distinguish between mESCs and mEpiSCs[26,44] and compared expression of these genes, as previously reported[26], with our present RNA-Seq data on undifferentiated hESCs (**Supplementary Fig. 5c**). We found that mouse mESCs clearly express higher levels of mESC-enriched genes, such as *Pecam1*, *Fbxo15*, *Stella*, *Fgf4*, *Stra8* and *Cdh1* (encoding E-cadherin). The expression level of *Rex1* and *Klf4* is not higher in mESCs than in hESCs. In contrast, hESCs express higher levels of mEpiSC-enriched genes such as *FGF2*, *FGF5*, *FGFR1* and *FGFR4*. The expression level of *EOMES*, *LEFTY1* and *LEFTY2* is not higher in hESCs than in mESCs, and hESCs and mESCs express similar levels of *NODAL*, *TDGF1*, *BMP4* and *BMPR1*. Our analysis suggests that hESCs are more similar to mEpiSCs than to mESCs because hESCs express the majority of the mEpiSC-specific marker genes at high levels and the majority of the mESC-specific marker genes at low levels. Nevertheless, hESCs have distinct features that set them apart from mEpiSCs. In particular, they express significantly higher levels of *POU5F1* (also known as *OCT4*) and *NANOG* than do mESCs, by factors of 4.5 ($P = 2.4 \times 10^{-10}$) and 5.3 ($P = 8.0 \times 10^{-5}$) times, respectively.

Comparing single-cell RNA-Seq and Smart-Seq data

Recently, an alternative single-cell RNA-Seq approach, Smart-Seq, was developed[24] and eight single cells of human embryonic stem cells (hESCs) were analyzed. Smart-Seq detected on average 7,869 RefSeq genes expressed in an individual hESC cell (with RPKM ≥ 0.1), compared with on average 11,784 RefSeq genes in an individual hESC cell by our approach (with RPKM ≥ 0.1) (**Supplementary Fig. 6a**). Therefore, our approach revealed 3,915 (49.8%) more genes than Smart-Seq within an individual cell of the same cell type, indicating that it is more sensitive. For the 2,575 genes detected in at least one of our eight single cells at RPKM ≥ 1, but undetectable in Smart-Seq data set (RPKM < 0.1 in every of their eight single cells), there were 1,110 with RPKM ≥ 10 in at least one of our single cells. Because Smart-Seq also suggested better coverage of the full-length cDNAs, we compared the coverage of cDNAs in hESCs using both techniques and found that our approach gives better 5′-end coverage than Smart-Seq (**Supplementary Fig. 6b**). This demonstrates the relatively even coverage for cDNAs using our single-cell RNA-Seq approach, at least when analyzing single cells of hESCs[24]. Moreover, our approach showed better technical reproducibility than Smart-Seq: the average correlation coefficient among individual hESCs by our method is 0.956, as compared to 0.925 by Smart-Seq (**Supplementary Fig. 6a**).

DISCUSSION

This study uses single-cell transcriptome sequencing to profile gene expression during human preimplantation embryonic development and the derivation of hESCs from blastocysts. These results provide a comprehensive framework of the transcriptome landscape of 90 single cells of 20 morphologically normal human oocytes and early embryos and 34 single cells of hESCs. Our major findings include the following.

The number of maternally expressed genes detected increased from 9,735 (microarray) to 22,687 (RNA-Seq), including 8,701 long noncoding RNAs (lncRNAs). To our knowledge, this is the first time that maternally expressed lncRNAs have been systematically analyzed in human early embryos.

Nearly 1,000 genes were shown to express unambiguously multiple transcript isoforms in the same human blastomere, which demonstrates the high complexity of alternative splicing within individual human cells at a whole-genome scale. Given that our single-cell RNA-Seq technology preferentially captures the 3′ end of the RNA sequences and so will favor alternative splicing detection toward the 3′ end but disfavor alternative splicing detection toward the 5′ end of cDNAs because of technical limitations[22], the actual alternative splicing complexity within individual cells is potentially higher than reported here. As far as we know, the global pattern of alternative splicing had not been analyzed in any types of human cells before this work.

Through *de novo* assembly, we identified 2,733 potential novel lncRNAs; these transcripts are at least 10 kb away from any known genes in RefSeq, Ensembl or known lncRNA databases and from each other. A large proportion of these novel lncRNAs show developmental stage–specific expression patterns, indicating that they are involved in preimplantation development. In the period from oocyte to late blastocyst, we identified expression of 32,934 RefSeq transcripts, 18,383 known lncRNA transcripts and 2,733 potential novel transcripts, of which 58%, 21% and 44%, respectively, were variably present throughout preimplantation development (FC > 2 or < 0.5, $P < 0.01$).

We found a large set of lineage-specific genes that can discriminate the epiblast (EPI), primitive endoderm (PE) and trophectoderm (TE) lineage cells in hatching blastocysts. Because the formation of the PE from the ICM is heterogeneous among different embryos, it is still possible that the PE cells that we analyzed were at a relatively early stage of the segregation or even at a PE 'precursor' stage.

The transcriptomes of human EPI and the primary hESC outgrowth derived from them showed dramatic global changes, with 1,498 genes differentially expressed between them, probably as a result of the adaptation of hESCs to the culture condition where cytokines of bFGF and IGF2 are present. To our knowledge, this is the first time that the transcriptome of hESC cells has been analyzed as early as passage 0, which permitted the tracing of the earliest gene expression changes from EPI to hESCs.

In 8-cell embryos, the variation between individual blastomeres in the expression of upregulated genes (FC[8-cell/4-cell] > 2, $P < 0.01$) was much less than that of downregulated genes (FC[8-cell/4-cell] < 0.5, $P < 0.01$), suggesting that the degradation of maternal transcripts is heterogeneous among 8-cell-stage blastomeres. This may contribute to subsequent differentiation of 8-cell-stage blastomeres into different lineages.

Our expression data of known RefSeq genes, known lncRNAs and novel lncRNAs in all of these 124 individual embryonic cells were in **Supplementary Tables 1**, **5** and **6**, and the raw data have been deposited in GEO as GSE36552. One can navigate through these data to look for developmental stage–specific or cell type–specific features of gene expression for human preimplantation development, derivation of hESCs processes, either within individual cells or the average of the embryos at same stages. Because the embryos we analyzed from the 8-cell stage and later were developed from cryopreserved embryos, whereas the embryos from the 4-cell and earlier stage were developed from freshly isolated oocytes, it is possible that some of the gene expression changes observed between 4-cell and 8-cell stage were instead due to the cryopreservation treatment. Nonetheless, our results pave the way for dissecting the molecular regulation of early human embryonic development and provide insight into pluripotency and the molecular identity of hESCs.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The raw sequencing data, mapped data and data for visualization of all of the single-cell RNA-Seq analyses for 124 human early embryonic cells and hESCs have been deposited to the GEO database under accession number GSE36552. The hESC line used in this work is freely available to other laboratories upon request.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS
F.T. was supported by grants from the National Basic Research Program of China (2012CB966704 and 2011CB966303) and National Natural Science Foundation of China (31271543). J.Q. was supported by grants from the National Basic Research Program of China (2011CB944500) and the National Natural Science Funds for Distinguished Young Scholar (30825038). L.Y. was supported by a grant from the National Science Foundation of China (81000275). We would like to thank S. Gao, Y. Zhang, P. Xu, S. Lin, X. Ren, Q. Zhang, Y. Jiang, M. Fan, J. Li, X. Zhuang, W. Song and Y. Chen for their great help.

AUTHOR CONTRIBUTIONS

J.Q. and F.T. conceived and designed the project, and R.L. was in charge of the bioinformatic analysis. L. Yan, H.G., L. Yang, X.W. and L.W. conducted the majority of the experiments. M.Y. and J.W. did all of the data analysis. R.L., P.L., Y.L., X.Z., J.Y., J.H. and M.L. contributed to oocyte collection, sperm treatment and embryo culture *in vitro*. L. Yan, M.Y., K.L., R.L., J.Q. and F.T. prepared the manuscript. All authors contributed to the revision of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Martinez Arias, A. & Brickman, J.M. Gene expression heterogeneities in embryonic stem cell populations: origin and function. *Curr. Opin. Cell Biol.* **23**, 650–656 (2011).
2. Hardy, K. *et al.* Future developments in assisted reproduction in humans. *Reproduction* **123**, 171–183 (2002).
3. Niakan, K.K., Han, J., Pedersen, R.A., Simon, C. & Pera, R.A.R. Human pre-implantation embryo development. *Development* **139**, 829–841 (2012).
4. Reijo Pera, R.A. Non-invasive imaging of human embryos to predict developmental competence. *Placenta* **32** (Suppl. 3), S264–S267 (2011).
5. Aghajanova, L. *et al.* Comparative transcriptome analysis of human trophectoderm and embryonic stem cell-derived trophoblasts reveal key participants in early implantation. *Biol. Reprod.* **86**, 1–21 (2012).
6. Assou, S. *et al.* Dynamic changes in gene expression during human early embryo development: from fundamental aspects to clinical applications. *Hum. Reprod. Update* **17**, 272–290 (2011).
7. Assou, S. *et al.* Transcriptome analysis during human trophectoderm specification suggests new roles of metabolic and epigenetic genes. *PLoS ONE* **7**, e39306 (2012).
8. Assou, S. *et al.* A gene expression signature shared by human mature oocytes and embryonic stem cells. *BMC Genomics* **10**, 10 (2009).
9. Bai, Q. *et al.* Dissecting the first transcriptional divergence during human embryonic development. *Stem Cell Rev. Rep.* **8**, 150–162 (2012).
10. Dobson, A.T. *et al.* The unique transcriptome through day 3 of human preimplantation development. *Hum. Mol. Genet.* **13**, 1461–1470 (2004).
11. Galán, A. *et al.* Functional genomics of 5- to 8-cell stage human embryos by blastomere single-cell cDNA analysis. *PLoS ONE* **5**, e13615 (2010).
12. Haouzi, D. *et al.* Transcriptome analysis reveals dialogues between human trophectoderm and endometrial cells during the implantation period. *Hum. Reprod.* **26**, 1440–1449 (2011).
13. Vassena, R. *et al.* Waves of early transcriptional activation and pluripotency program initiation during human preimplantation development. *Development* **138**, 3699–3709 (2011).
14. Wong, C.C. *et al.* Non-invasive imaging of human embryos before embryonic genome activation predicts development to the blastocyst stage. *Nat. Biotechnol.* **28**, 1115–1121 (2010).
15. Xie, D. *et al.* Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species. *Genome Res.* **20**, 804–815 (2010).
16. Kiessling, A.A. *et al.* Evidence that human blastomere cleavage is under unique cell cycle control. *J. Assist. Reprod. Genet.* **26**, 187–195 (2009).
17. Kiessling, A.A. *et al.* Genome-wide microarray evidence that 8-cell human blastomeres over-express cell cycle drivers and under-express checkpoints. *J. Assist. Reprod. Genet.* **27**, 265–276 (2010).
18. Guttman, M. & Rinn, J.L. Modular regulatory principles of large non-coding RNAs. *Nature* **482**, 339–346 (2012).
19. Thomson, J.A. *et al.* Embryonic stem cell lines derived from human blastocysts. *Science* **282**, 1145–1147 (1998).
20. Reijo Pera, R.A. *et al.* Gene expression profiles of human inner cell mass cells and embryonic stem cells. *Differentiation* **78**, 18–23 (2009).
21. O'Leary, T. *et al.* Tracking the progression of the human inner cell mass during embryonic stem cell derivation. *Nat. Biotechnol.* **30**, 278–282 (2012).
22. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
23. Tang, F., Lao, K. & Surani, M.A. Development and applications of single-cell transcriptome analysis. *Nat. Methods* **8**, S6–S11 (2011).
24. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
25. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Reports* **2**, 666–673 (2012).
26. Tang, F. *et al.* Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell* **6**, 468–478 (2010).
27. Braude, P., Bolton, V. & Moore, S. Human gene expression first occurs between the four- and eight-cell stages of preimplantation development. *Nature* **332**, 459–461 (1988).
28. Cockburn, K. & Rossant, J. Making the blastocyst: lessons from the mouse. *J. Clin. Invest.* **120**, 995–1003 (2010).
29. Rossant, J. & Tam, P.P.L. Blastocyst lineage formation, early embryonic asymmetries and axis patterning in the mouse. *Development* **136**, 701–713 (2009).
30. Gabut, M. *et al.* An alternative splicing switch regulates embryonic stem cell pluripotency and reprogramming. *Cell* **147**, 132–146 (2011).
31. Khalil, A.M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. USA* **106**, 11667–11672 (2009).
32. Ørom, U.A. *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46–58 (2010).
33. Wang, K.C. & Chang, H.Y. Molecular mechanisms of long noncoding RNAs. *Mol. Cell* **43**, 904–914 (2011).
34. Guttman, M. *et al.* lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**, 295–300 (2011).
35. Derrien, T., Guigó, R. & Johnson, R. The long non-coding RNAs (IncRNAs): a new (p) layer in the "dark matter". *Frontiers Genet.* **2**, 107 (2012).
36. 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).
37. Kong, L. *et al.* CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* **35**, W345-9 (2007).
38. Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54–i62 (2009).
39. Guttman, M. *et al.* *Ab initio* reconstruction of transcriptomes of pluripotent and lineage committed cells reveals gene structures of thousands of lincRNAs. *Nat. Biotechnol.* **28**, 503–510 (2010).
40. Wells, D. & Patrizio, P. Gene expression profiling of human oocytes at different maturational stages and after in vitro maturation. *Am. J. Obstet. Gynecol.* **198**, 455.e1–455.e11 (2008).
41. Roode, M. *et al.* Human hypoblast formation is not dependent on FGF signalling. *Dev. Biol.* **361**, 358–363 (2012).
42. Brons, I.G.M. *et al.* Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature* **448**, 191–195 (2007).
43. Tesar, P.J. *et al.* New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448**, 196–199 (2007).
44. Boyer, L.A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–956 (2005).
45. Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**, 1540–1542 (2006).

ONLINE METHODS

Informed consent. This study was approved by the Reproductive Study Ethics Committee of Peking University Third Hospital (Research License 2011S2003 and 2011S2018). All embryos were obtained with written informed consent signed by the donor couples. The informed consent confirmed that the couple donors were voluntarily donating oocytes and embryos (including sperm) for research on human early embryonic development mechanisms with no financial payment (Research License 2011S2003). The couples donating eggs and embryos before the 8-cell stage were informed that the donation posed a potential risk to their fertility success for that cycle. The embryos used for the derivation of hESCs were obtained for this purpose with the signed informed consent from the donor couples (Research License 2011S2018).

Patients. Embryos were obtained from women undergoing *in vitro* fertilization (IVF) at the Center for Reproductive Medicine of Peking University Third Hospital using standard clinical protocols as described[46]. The women had an average age of 30 years (25–35 years), tubal-factor infertility and partners with normal semen parameters.

Preparation of single-cell cDNAs. The step-by-step single-cell RNA-Seq method has been described previously[22,47]. Briefly, we use a mouth pipette to pick an individual cell manually and transfer it into lysate buffer. We perform the reverse transcription reaction directly on the whole-cell lysate. We use terminal deoxynucleotidyl transferase to add a poly(A) tail to the 3′ end of the first-strand cDNA, and then perform 20 + 10 cycles of PCR to amplify the single-cell cDNA.

RNA-Seq library preparation, sequencing and alignment. After the generation of cDNA from a single cell, 200 ng of cDNA was sheared into 150- to 350-bp fragments by Covaris S2, and a TrueSeq DNA library preparation kit (Illumina) was used to prepare a sequencing library following the manufacturer's suggested procedures. The fragmented cDNA was end-repaired, dA-tailed, adaptor ligated and then subjected to 10–12 cycles of PCR amplification.

We obtained 20–60 million 100-bp reads for each individual cell, and a total of 438 Gb of data was obtained for all of the samples together. The correlation coefficient between the two blastomeres within the same 2-cell-stage embryo is 0.994, which verifies that the method is accurate (**Supplementary Fig. 1c,d**).

Read mapping. The hg19 RefSeq (RNA sequences, GRCh37) was downloaded from the UCSC Genome Browser (http://genome.ucsc.edu). We used the Burrows-Wheeler Aligner[48] (BWA, Version 0.5.9-r16) to align the filtered reads to the hg19 RefSeq, using the options "aln -o 1 -e 60 -i 15 -q 10 -t 8". Reads that failed to be mapped were progressively mapped to the ensemble gene set (GRCh37.65, http://www.ensembl.org/info/data/ftp/), the lncRNA database[49] (NONCODE v3.0, http://www.noncode.org/NONCODERv3) and the reference genome (UCSC hg19, http://genome.ucsc.edu), respectively.

New transcript prediction. First, we used the *de novo* transcriptome reconstruction software Trinity[50] (version r2011-08-20) and the eukaryotic genome annotation tool PASA[51] (version r2011-05-20) to perform the *de novo* assembly of reads that could not be mapped to the RefSeq or Ensembl transcript databases. Primary assembly results were aligned to the reference genome using blat[52],with the highest sequence identity and >90% coverage of the contig.

Second, we combined the reads that could be mapped to the genome with assembled candidate transcripts to predict new genes using Cufflinks[36] (v1.3.0). The predicted new genes were also required to be at least 10 kb away from either end of known genes in the nonredundant gene set (the union of RefSeq, Ensembl and lncRNA) and other predicted new genes.

Collection and culture of human oocytes and embryos. Oocytes were donated for this study with the informed consent of couples who had more than 20 oocytes derived from the same IVF cycle. Embryos that were produced by routine fertilization were cultured individually in G1.3 medium (Vitrolife, Sweden) in 20-μl droplets covered with mineral oil (Sigma, 6% CO_2) until day 3. Pronuclear and 2-cell- and 4-cell-stage embryos were collected at the appropriate time according to embryonic development. Day 3 cleavage-stage embryos were donated for this study with the informed consent of couples who already had a healthy baby from the same cycle and wished to donate the remaining cryopreserved embryos. The embryos were thawed rapidly by taking straws from the liquid nitrogen storage tank, exposing them to air for 40 s, and immersing them in a water bath at 30 °C for 1 min. Embryos were sequentially placed into thawing medium drops for 5 min with decreasing PROH concentrations (1.0 mol/l, then 0.5 mol/l and finally 0 mol/l), each with 0.2 mol/l sucrose at room temperature to remove the cryoprotectant. Thawed embryos were transferred to PBS with 20% HSA for 10 min at 37 °C and then to G2 culture medium (Vitrolife) to evaluate blastomere survival. The 8-cell, morula and late-blastocyst-stage embryos were collected and treated to obtain single blastomeres at 2 h, 24 h or 72 h after thawing, respectively.

Embryos were obtained from women with average age of 30 years old and with only tubal factor infertility whose partners had normal semen parameters undergoing *in vitro* fertilization (IVF). Among these patients, there are usually more than three good-quality embryos in one oocyte retrieval cycle. Two or three good embryos were selected to transfer on day 3 after oocyte collection, and the remaining morphologically good-quality embryos were cryopreserved. After these patients had delivered a healthy baby either from the initial IVF cycle or from a later cycle with thawed frozen embryos, some of them wished to donate their remaining frozen embryos for scientific research and signed written informed consent. The embryos of 8-cell, morula and blastocyst stages analyzed in this study are from these donated frozen embryos.

All of the oocytes, zygotes and 2-cell- and 4-cell-stage embryos are from oocyte donation. The zygotes and the 2-cell- and 4-cell-stage embryos were derived by routine *in vitro* fertilization of these donated oocytes by donated sperm from the same couple. Those female patients, having with tubular factors and being less than 30 years old, had high probability of successful IVF, and when they had more than 20 oocytes, some were willing to donate a few of their oocytes to our scientific research. The couples donating eggs and embryos before 8-cell-stage were informed that the donation posed a potential risk to their fertility success for that cycle.

The oocytes and zygotes analyzed are selected according to their morphology and timing of development. The oocytes were picked randomly by the embryologists at the clinical IVF lab of the Center for Reproductive Medicine, Peking University Third Hospital, after patients signed informed consent and then were transferred to the scientific lab. The cumulus cells around the oocytes were removed by hyaluronidase treatment. Only the mature MII oocytes were used in our study. The two-pronucleate (2PN) zygotes with good morphology were selected through pronuclear assessment at 19 h after routine *in vitro* fertilization.

All of the human embryos used in this study had a relatively high quality and good morphology. The embryonic assessment was performed according to a previous study[53], and the embryonic stages were defined as follows.

Oocytes: oocytes up to 4 h after IVF oocyte retrieval.

Zygotes: embryos with two pronuclei and two polar bodies, collected 19 h after routine fertilization.

2-cell-stage embryos: embryos with two blastomeres of equal size and without any fragments, collected 27 h after routine fertilization.

4-cell-stage embryos, 4A1: embryos with four blastomeres of similar size and with negligible fragmentation, collected 48 h after routine fertilization.

8-cell-stage embryos, 8A1: day 3 embryos with eight blastomeres of similar size and no cytoplasmic fragments.

Morulae: day 4 embryos developed from day 3 8A1 embryos, in the process of becoming compacted, with more than 15 cells and the cell borders becoming fuzzy.

Early hatching blastocysts, 5AA: day 6 embryos with a thin zona pellucida, smooth trophectoderm, clearly visible blastocyst cavity and well-developed inner cell mass.

Isolation and separation of individual blastomeres. The zona pellucida was removed using an acidic solution (pH 2.5, 1 ml of PBS supplemented with 1 μl of 36% HCl). The embryos were exposed to an acidic PBS solution briefly for 5–10 s and were washed thoroughly in PBS with 1% HSA. The denuded embryos were then treated with 0.02% EDTA for 5 min and placed into Accutase medium (Chemicon, SCR005 for morulae or blastocysts) or 0.005% trypsin (GIBCO, for

doi:10.1038/nsmb.2660

2-cell-, 4-cell- or 8-cell-stage embryos) for 30–60 min. Single blastomeres were isolated by gentle, repeated pipetting. When all of the blastomeres were separated, they were removed from the manipulation drops, washed 3–5 times in prewarmed PBS with 1% HSA medium, and placed into lysis buffer immediately for the preparation of the single-cell cDNA library.

The morulae were placed in droplets for separate single-cell treatment. The potential outer layer of cells in morulae detached from the inner cells after being incubated in Accutase for 30 min. We called the cells first isolated 'outer' cells and cells isolated later 'inner' cells.

The trophectoderm (TE) of late blastocysts was isolated under a stereoscope by mechanical dissection of the mural trophectoderm section of the embryo. The ICM portion was also mechanically isolated and then further dissociated into a single-cell suspension by Accutase treatment. The epiblast (EPI) and primitive endoderm (PE) were discriminated retrospectively using a set of marker genes. The result was also verified by unsupervised hierarchical clustering of these blastocyst cells. It is well known that in the blastocyst, TE cells specifically activate the Na^+/K^+-ATPase pump to drive vectorial transport for blastocoel formation and the exchange of ions, amino acids, energy substrates and other metabolites[54]. Similarly, we found a clear upregulation of the alpha 1 (*ATP1A1*) and beta3 (*ATP1B3*) subunits of the Na^+/K^+ transporting ATPase in TE cells compared with EPI and PE cells (**Fig. 6b**).

The data from individual ICM cells was queried for a set of genes to discriminate EPI and PE cells. We first physically separated ICM and mural TE by cutting using syringe needle. Then the ICM portion (including EPI, PE and some remaining TE cells) was dissociated into single-cell suspension and randomly picked. The mural TE cells physically isolated by syringe needle showed high expression of *CDX2*, *KRT18*, *GATA2*, *GATA3* and *CLDN10*, as previously reported[7,11]. We therefore used these markers to identify left TE cells in the 'ICM section'. After that, the cells not showing high expression of these maker genes were considered as potential ICM cells. This is validated by the fact that they showed clear expression of *SOX2*, *NANOG* and *KLF4*. Further, they contained an EPI population with high expression of *SOX2*, *NANOG*, *KLF4*, *GDF3*, *FOXD3*, *ESRRB* and *PRDM14* (partially shown in **Supplementary Fig. 3d**)[11,55], and a PE population with low expression of *SOX2*, *NANOG*, *KLF4*, *GDF3*, *FOXD3*, *ESRRB* and *PRDM14* but high expression of *FGFR4* and *CLDN3*. The identification of these lineages was consistent with their separation in the unsupervised hierarchical clustering analysis. For the unsupervised hierarchical clustering analysis shown in **Figure 6a**, when we added the two morula-stage blastomeres clustered together with EPI cells in **Figure 2a** and redid the analysis, the pattern did not change. That is, the two morula-stage blastomeres clustered with the EPI cells (**Supplementary Fig. 7a**). To see if the EPI cells separate from other lineages of cells in the blastocysts, we carried out PCA analysis of blastomeres in morulae and blastocysts and found that the five EPI cells clearly separated from other cells of blastocysts (**Supplementary Fig. 7b**).

Derivation and culture of human embryonic stem cells. hESC cells were derived from the ICM of blastocysts that were discarded in a routine IVF embryo transfer program with the written informed consent of the couples.

ICM isolation was performed mechanically with a 29-G syringe needle under a stereoscope. The isolated ICM was cultured on mouse embryonic fibroblast (MEF) feeder cells that had been mitotically inactivated with 10 μg/ml mitomycin C (Sigma, M4287).

hESCs were cultured on mitotically inactivated MEF feeder layers derived from E13.5 embryos of the ICR strain. The basic culture medium for hESC cell maintenance consisted of knockout-Dulbecco's modified Eagle's medium (KO-DMEM; Invitrogen, 10829018) supplemented with 15% knockout serum replacement (Ko-SR, Invitrogen, 10828028), 5% FBS (FBS; HyClone, SH30070-03), 1 mM glutamine (Chemicon,TMS-002-C), 0.1 mM β-mercaptoethanol (Chemicon, ES-007-E), 1% nonessential amino acids (NEAAs; Chemicon, TMS-001-C) and 4 ng/ml human basic fibroblast growth factor (bFGF; Invitrogen, 13256029). For the initial culture of ICM cells and the first five passages, 1,000 U/ml of LIF (Chemicon, LIF1010) and 10 ng/ml of bFGF were added. After an initial growth period of 10 d, the cell colonies derived from the ICM were removed from the dish by mechanical slicing using glass capillaries drawn on a flame and with a sealed tip. The cell clusters were transferred to new plates containing fresh feeder cells. After the first splitting, the new colonies were again split and transferred to new dishes every 5 d. Half of the culture medium was changed every day. After five mechanical passages, the hESC colonies were treated with collagenase type IV for passaging.

Small, undifferentiated cell clumps from passage 0 were treated with Accutase for 30–60 min at 37 °C for dissociation into single-cell suspensions that were used to establish single-cell cDNA libraries.

Western blot analyses. Western blots were performed as previously described[56]. The antibodies used for western blotting were anti-OCT4 polyclonal antibody (Abcam, ab19857), anti-SOX2 polyclonal antibody (Abcam, ab59776) and anti-NANOG polyclonal antibody (Abcam, ab21624), with anti–α-tubulin monoclonal antibody (Sigma, T6074) used as a control.

Immunostaining of human embryonic stem cells. Immunostaining was performed as previously described with some modifications[56]. The primary antibodies used were rabbit anti-human OCT4 (Abcam, 19857) or mouse anti–TRA1-60 (Abcam, 841619) antibody diluted 1:100 in blocking solution.

Karyotype analysis of human embryonic stem cells. The culture medium was changed approximately 4 h before karyotyping. hESCs were incubated with 0.2 μg/ml colcemid for 2 h at 37 °C (5% CO_2) and then with 0.25% trypsin-EDTA for 2 min at 37 °C. After pipetting, the single-cell suspension was washed twice with PBS. The pellet was resuspended in prewarmed 75 mM KCl for 15 min at 37 °C. The cells were then fixed with methanol:glacial acetic acid (1:3) three times and dropped onto glass slides. The slides containing cells were stained in Giemsa solution for 3 min for standard G-banding analysis (**Supplementary Fig. 7**).

Teratoma formation. To examine their pluripotency *in vivo*, hESCs grown on MEF feeders with an undifferentiated morphology were collected by 1 mg/ml collagenase type IV (Gibco, Invitrogen Corporation) treatment at 37 °C for 20 min. A 200-μl volume of the hESC suspension (hESCs from two 60-mm dishes for each transplanted mouse) was injected subcutaneously into the inguinal groove of 6-week-old SCID mice. Eight weeks after the injection, teratomas were dissected, fixed in PBS containing 10% formaldehyde, embedded in paraffin and processed with hematoxylin and eosin staining (**Supplementary Fig. 7**).

Real-time PCR. The quality of single-cell cDNA was analyzed by real-time PCR[47]. The PCR was performed as follows using an AB7500 with 96-well plates: first, 95 °C for 10 min to activate the Taq polymerase, then 40 cycles of 95 °C for 15 s and 60 °C for 1 min.

Preparation of single-cell cDNA and RNA-Seq libraries. As described in our previous study[47], we used qPCR analysis of a set of house-keeping genes to check the quality of the amplified single-cell cDNA. After the sequencing library was prepared, we used an Agilent 2100 bioanalyzer to analyze the quality of the libraries.

Deep sequencing and quality control (QC). The libraries were sequenced on the Illumina HiSeq 2000 platform using the 100-bp single-end sequencing strategy.

In total, we generated 438 Gb (raw data) for 124 single-cell cDNA samples. The original image data generated by the sequencing machine were converted into sequence data via base calling (Illumina pipeline CASAVA v1.8.0) and then subjected to standard QC criteria to remove all of the reads that fit any of the following parameters:

1. The reads that aligned to adaptors or primers with no more than two mismatches.

2. The reads with more than 10% unknown bases (N bases).

3. The reads with more than 50% of low-quality bases (quality value ≤ 5) in one read.

We subsequently only used samples with Q20 > 75% and a rate of clean data (percentage of obtained final data with respect to raw data) of >45% for further analysis. We also mapped the clean data to the rRNA database (ftp://ftp.sanger.ac.uk/pub/databases/); only samples with <5% rRNA contamination were left for final analysis (**Supplementary Table 9**). In total, 124 individual cells within the RNA-Seq data set met all of these criteria for the final analysis. Finally, 371.9 Gb (84.9%) of filtered reads were left for further analysis after QC (**Supplementary Table 10**). From these 371.9 Gb of filtered reads, 352.2 Gb (80.4%) of data

doi:10.1038/nsmb.2660

were mapped to the RefSeq, Ensembl, lncRNA and hg19 reference databases (**Supplementary Table 10**).

We also analyzed the distribution of mapped reads in the genome and found that 91% of the reads mapped to exon regions (exons for known RefSeq genes, known Ensembl genes and known lncRNA genes), with the remaining reads mapping to introns and intergenic regions (**Supplementary Table 11**).

Expression analysis of RefSeq genes and lncRNAs. Gene expression was calculated using the RPKM method (Reads Per Kilobase transcriptome per million reads)[57,58]. We used all of the genes with an RPKM ≥ 0.1 as the expressed genes in the following analysis. We sequenced on average 35 million 100-bp reads for each individual cell.

We used BWA alignment to calculate gene expression level of the novel transcripts we assembled. In detail, after we got the assembly results (.gtf file) by Cufflinks, we first converted the file from .gtf format to .fa format and used the .fa file as a reference in the later alignment step. Next, we used Burrows-Wheeler Aligner (BWA, Version 0.5.9-r16) (ref. 48) to map the sequencing reads onto the .fa file. Specifically we aligned the filtered reads that can be mapped to the human reference genome, but with no overlap to any known RefSeq genes, known Ensembl genes or known lncRNAs to the *.fa file, with the options "aln -o 1 -e 60 -i 15 -q 10 -t 8". We obtained the read counts of each novel transcript from the BWA's result. Then we converted read counts for each novel transcript into RPKM by a perl script we developed.

Analysis of expression patterns of RefSeq genes and lncRNAs. Genes with similar expression patterns are likely to have functional correlations, so we performed a cluster analysis of the gene expression patterns using Cluster 3.0 (ref. 59) and JavaTreeview[60] software. Expression differences in RefSeq genes and lncRNAs were clustered by the Hierarchical Complete Linkage Clustering method using an uncentered correlation similarity matrix.

Note that we added two additional steps before clustering to identify differently expressed lncRNAs.

First, lncRNAs with only one read covered were removed because a lack of reliability.

Then, we pretreated the data set using standardization tools in Cluster 3.0. (i) Log Transform Data: replace all data values x by $\log_2 (x)$. (ii) Center Arrays [mean]: Subtract the column-wise mean from the values in each column of data so that the mean value of each column is 0. (iii) Normalize Arrays: Multiply all values in each column of data by a scale factor S so that the sum of the squares of the values in each column is 1.0 (a separate S is computed for each column).

Screening of differentially expressed genes. To analyze differences in gene expression among the seven consecutive stages of pre-implantation embryos, the P value (two-tailed) was calculated accordingly (two-sample t-test with homoscedasticity), whereas for other cases, the P value (one-tailed) was calculated using a two-sample t-test with heteroscedasticity. Because the analysis of differentially expressed genes generates large multiplicity problems in which thousands of hypotheses (i.e., Is gene x differentially expressed between the two groups?) are tested simultaneously, corrections for false-positive (type I errors) and false-negative (type II) errors were performed using Benjamini and Yekutieli's false discovery rate (FDR) method in some cases, as indicated in the main text[61]. We use '$P < 0.01$ and FC (fold change) > 2' as the threshold to judge the significance of gene expression differences, except for the cases specially stated in the main text. The heat maps were drawn by using the R packages as follows: function 'heatmap.2' of 'gplots' package and function 'sota' of 'clValid' package, with default Euclidean distance and hierarchical clustering method.

Analysis of isoform-specific exon-exon junction reads. We generated a gene set of multiple known transcript isoforms from the RefSeq database that contains all of the genes with at least two known transcript isoforms. Then, from all of the known exon-exon junctions of these genes, we isolated all of the exon-exon junctions unique to each transcript isoform. We then analyzed all of the reads in the transcriptome data of individual cells that mapped to these isoform-specific exon-exon junctions with at least 8 bp corresponding to the opposite side of the junction.

Gene Ontology (GO) analysis. GO enrichment was performed using DAVID[62] (http://david.abcc.ncifcrf.gov/). A hypergeometric test with the Benjamini and Hochberg false discovery rate (FDR) was performed using the default parameters to adjust the P value[61].

46. Huang, J. *et al.* Characteristics of embryo development in Robertsonian translocations' preimplantation genetic diagnosis cycles. *Prenat. Diagn.* **29**, 1167–1170 (2009).
47. Tang, F. *et al.* RNA-seq analysis to capture the transcriptome landscape of a single cell. *Nat. Protoc.* **5**, 516–535 (2010).
48. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
49. Bu, D. *et al.* NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res.* D210-5 (2012).
50. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
51. Haas, B.J. et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
52. Kent, W.J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656–664 (2002).
53. Baczkowski, T., Kurzawa, R. & Glabowski, W. Methods of embryo scoring in in vitro fertilization. *Reprod. Biol.* **4**, 5–22 (2004).
54. Madan, P., Rose, K. & Watson, A.J. Na/K-ATPase β1 subunit expression is required for blastocyst formation and normal assembly of trophectoderm tight junction-associated proteins. *J. Biol. Chem.* **282**, 12127–12134 (2007).
55. Guo, G. *et al.* Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev. Cell* **18**, 675–685 (2010).
56. Bao, S. *et al.* Epigenetic reversion of post-implantation epiblast to pluripotent embryonic stem cells. *Nature* **461**, 1292–1295 (2009).
57. Audic, S.p. & Claverie, J.-M. The significance of digital gene expression profiles. *Genome Res.* **7**, 986–995 (1997).
58. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
59. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
60. Saldanha, A.J. Java Treeview—extensible visualization of microarray data. *Bioinformatics* **20**, 3246–3248 (2004).
61. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B* **57**, 289–300 (1995).
62. Huang, D.W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2008).

doi:10.1038/nsmb.2660