mRNA-Seq whole-transcriptome analysis of a single cell

Fuchou Tang[1,3], Catalin Barbacioru[2,3], Yangzhou Wang[2], Ellen Nordman[2], Clarence Lee[2], Nanlan Xu[2], Xiaohui Wang[2], John Bodeau[2], Brian B Tuch[2], Asim Siddiqui[2], Kaiqin Lao[2] & M Azim Surani[1]

Next-generation sequencing technology is a powerful tool for transcriptome analysis. However, under certain conditions, only a small amount of material is available, which requires more sensitive techniques that can preferably be used at the single-cell level. Here we describe a single-cell digital gene expression profiling assay. Using our mRNA-Seq assay with only a single mouse blastomere, we detected the expression of 75% (5,270) more genes than microarray techniques and identified 1,753 previously unknown splice junctions called by at least 5 reads. Moreover, 8–19% of the genes with multiple known transcript isoforms expressed at least two isoforms in the same blastomere or oocyte, which unambiguously demonstrated the complexity of the transcript variants at whole-genome scale in individual cells. Finally, for *Dicer1*$^{-/-}$ and *Ago2*$^{-/-}$ (*Eif2c2*$^{-/-}$) oocytes, we found that 1,696 and 1,553 genes, respectively, were abnormally upregulated compared to wild-type controls, with 619 genes in common.

Next-generation sequencing technology is a powerful and cost-efficient tool for ultra-high-throughput transcriptome analysis[1–5]. By analyzing the transcriptome at spectacular and unprecedented depth and accuracy, thousands of new transcript variants and isoforms have been shown to be expressed in mammalian tissues or organs[6–12]. These advances greatly accelerate our understanding of the complexity of gene expression, regulation and networks for mammalian cells. These new techniques usually need microgram amounts of total RNA for analysis, which corresponds to hundreds of thousands of mammalian cells. However, under many important conditions, it is practically impossible to get such large amounts of material, for example, in early embryonic development studies. In fact, during mouse early development, when the founder population of germline, primordial germ cells have just emerged, there are only around 30 primordial germ cells in the embryo[13]. Even for *in vitro*–cultured stem cells, for which the number of cells would appear to be unlimited, there are serious limitations. For example, mouse embryonic stem cells, probably the most thoroughly analyzed type of stem cells, contain multiple subpopulations with strong differences in both gene expression and physiological function[14,15]. Therefore, a more sensitive mRNA-Seq assay, ideally an assay capable of working at single cell resolution, is needed to meaningfully study crucial developmental processes and stem cell biology.

Here we modified a widely used single-cell whole-transcriptome amplification method to generate cDNAs as long as 3 kilobases (kb) efficiently and without bias[16,17]. With Applied Biosystems' next-generation sequencing SOLiD system, we found that it is feasible to get digital gene expression profiles at single-cell resolution. Using our mRNA-Seq assay with only a single mouse blastomere, we detected expression of 5,270 more genes than microarrays using hundreds of blastomeres. Using only a single blastomere, we also identified 1,753 previously unknown splice junctions, which have never been detected by microarrays at single-cell resolution. We found that hundreds of genes expressed two or more transcript variants in the same cell. We also found that in *Dicer1*$^{-/-}$ and *Ago2*$^{-/-}$ mature oocytes, 1,696 and 1,553 genes, respectively, were abnormally upregulated, and 1,571 and 1,121 genes, respectively, were downregulated compared to wild-type controls, which illustrates the global importance of small RNAs (including microRNAs and endogenous small interfering RNAs) for oogenesis. This single-cell mRNA-Seq assay will greatly enhance our ability to analyze transcriptome complexity in individual cells during mammalian development, especially for early embryonic development and for stem cells, which are usually rare cell populations *in vivo*.

RESULTS

Characterization of single cell whole transcriptome analysis

First, to make the single-cell cDNA amplification method previously used for microarray analyses[16,17] suitable for mRNA-Seq, we modified the protocol. We increased the reverse transcription step from 5 min to 30 min to get full-length first-strand cDNAs. Correspondingly, we extended the extension time for PCR from 3 min to 6 min. We also modified the PCR primers by adding an amine at the 5′ end to prevent the ligation of the 5′ end fragments of the double-stranded cDNA to the SOLiD library adaptors, thereby eliminating end bias during sequencing. The size distribution of the amplified cDNAs was 0.5–3 kb, which

[1]Wellcome Trust–Cancer Research UK Gurdon Institute of Cancer and Developmental Biology, University of Cambridge, Cambridge, UK. [2]Molecular Cell Biology Division, Applied Biosystems, Foster City, California, USA. [3]These authors contributed equally to this work. Correspondence should be addressed to M.A.S. (as10021@mole.bio.cam.ac.uk) or K.L. (laokq@appliedbiosystems.com).

RECEIVED 2 DECEMBER 2008; ACCEPTED 2 MARCH 2009; PUBLISHED ONLINE 6 APRIL 2009; CORRECTED ONLINE 19 APRIL 2009 (DETAILS ONLINE); DOI:10.1038/NMETH.1315

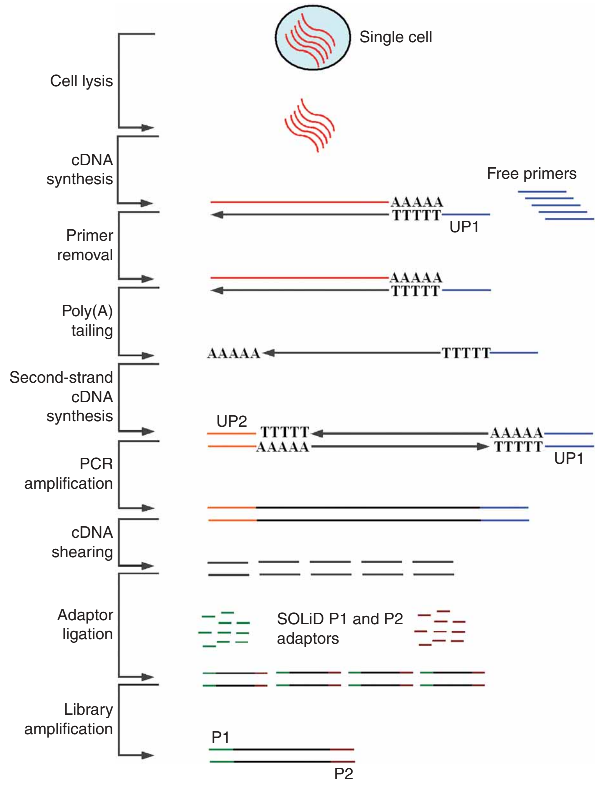

Figure 1 | Schematic of the single-cell whole-transcriptome analysis. A single cell is manually picked under a microscope and lysed. Then mRNAs are reverse-transcribed into cDNAs using a poly(T) primer with anchor sequence (UP1) and unused primers are digested. Poly(A) tails are added to the first-strand cDNAs at the 3′ end, and second-strand cDNAs are synthesized using poly(T) primers with another anchor sequence (UP2). Then cDNAs are evenly amplified by PCR using UP1 and UP2 primers, fragmented, and P1 and P2 adaptors are ligated to the ends. Finally, emulsion PCR is performed by mixing libraries with 1 μm diameter beads with P1 primers covalently attached to their surfaces.

means that the majority (64%) of expressed genes were converted into full-length cDNAs, based on the mouse RefSeq database (**Supplementary Fig. 1** online). Then we subjected these amplified single-cell cDNAs to SOLiD library preparation procedures. We sonicated cDNAs to fragments of 80–130 base pairs (bp). We generated fragment libraries by the SOLiD low-input fragment library construction workflow that includes end repair, blunt-end ligation, PCR and emulsion PCR (**Fig. 1**).

To test the sensitivity of this assay, we obtained the cDNA expression profile of a single blastomere from a four-cell stage embryo (MF1 strain). We obtained more than 100 million 35-base reads and 50-base reads from the single blastomere (**Table 1** and **Supplementary Fig. 2** online; in the latter analysis we mainly focused on the 50-base reads). Then we mapped these reads to the mouse genome, and the counts of the reads that aligned to the 189,620 known exons of the mm9 mouse genome were assigned to all known mouse RefSeq transcripts. We compared our mRNA-Seq data with those from Affymetrix microarray studies of about 80 pooled four-cell stage embryos (320 blastomeres) and found that 94.3% (6,650 genes) of the genes detected by Affymetrix microarrays had at least five reads in our single-blastomere mRNA-Seq data based on 15,776 RefSeq transcripts that have probes on the array[18]. The concordance between the sequences of the plus and minus cDNA strands was high (**Fig. 2a** and **Supplementary Table 1** online). As these complementary strands were annealed from sample preparation until the emulsion PCR step, the high concordance illustrated the accuracy of our sequencing technique (**Supplementary Fig. 3** online) and mapping algorithms (**Supplementary Fig. 4** online). This was also the case for wild-type, *Dicer1*$^{-/-}$ and *Ago2*$^{-/-}$ oocytes (**Fig. 2b–d**). mRNA-Seq analysis missed 5.7% of the transcripts (400 genes) detected by microarray analysis (**Fig. 3a**). Most of these genes were only marginally detected by microarray analysis (327/400 genes had fluorescence intensity on the chip lower than 100; **Fig. 3b**), which is similar to the result of the mRNA-Seq analysis using total RNAs from hundreds of thousands of cells[8,10]. We used real-time PCR to analyze expression of 11 genes detected by microarray analysis but not by our mRNA-Seq assay in blastomeres of the four-cell stage embryos. Nine of these genes had no expression (cycle threshold (Ct) = 40) and two of them had extremely low expression (Ct > 36) (**Supplementary Table 2** online). This suggests that the majority of these 400 genes detected by microarray but not detected by our mRNA-Seq were likely false positives by microarray

Table 1 | Single cell mRNA-Seq mapping summary

	Blastomere (50-base reads)	Blastomere (35-base reads)	Wild-type oocyte 1 (50-base reads)	Wild-type oocyte 2 (50-base reads)	*Dicer1*$^{-/-}$ oocyte 1 (50-base reads)	*Dicer1*$^{-/-}$ oocyte 2 (50-base reads)	*Ago2*$^{-/-}$ oocyte (50-base reads)
Reads processed	85,807,979	24,424,339	76,584,432	20,998,366	65,023,554	37,652,933	43,311,094
Known RefSeq transcripts[a]	20,677,262	6,534,175	33,038,025	9,334,408	23,459,955	11,121,519	19,179,784
Reads with 0 mismatches to 21,436[b] RefSeq transcripts	9,166,378	4,471,940	13,960,414	5,071,973	10,059,424	6,588,271	7,873,465
Reads with 1 mismatch to 21,436 RefSeq transcripts	5,733,123	1,741,229	9,898,861	2,382,288	6,917,296	2,658,557	6,300,824
Reads with 2 or more mismatches to 21,436 RefSeq transcripts	5,777,761	321,006	9,178,750	1,880,147	6,483,235	1,874,691	5,005,495
Reads matching mouse genome[c]	49,910,707	14,896,842	50,320,364	14,623,581	37,725,544	21,299,168	31,861,036
Unmatched reads	35,489,449	9,507,906	26,084,368	6,003,943	26,741,547	14,472,985	11,273,117
Reads matching filtered sequences (primers, rRNAs)	407,823	19,591	179,700	370,842	556,463	1,880,780	176,941
Splice junction reads	2,976,501	1,238,737	2,395,418	922,656	1,818,176	1,170,273	1,337,043

[a] Number of reads matched to the RefSeq sequences. [b] There is a total 21,436 known transcripts in UCSC RefSeq database. [c] Number of reads matching to the mouse genome (mm9, NCBI build 37).

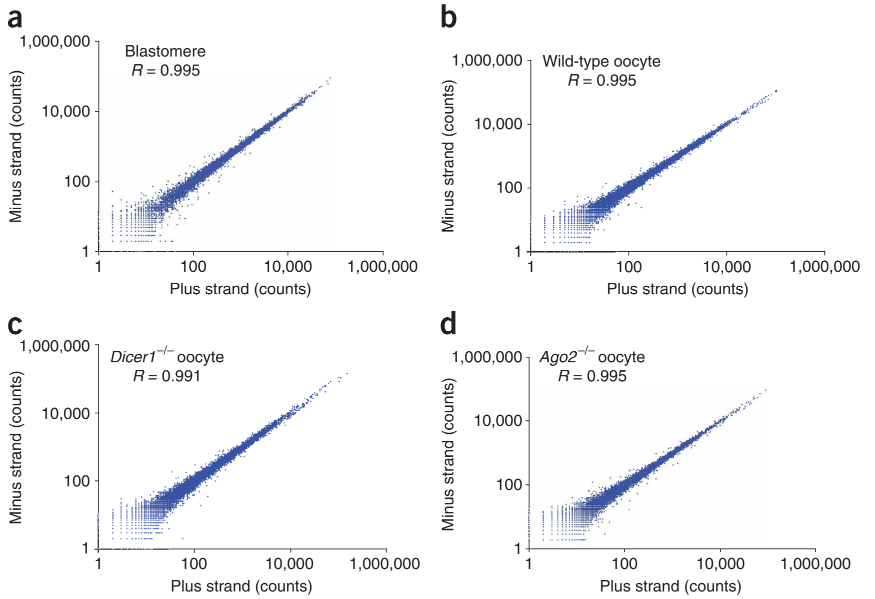

Figure 2 | mRNA-Seq of single blastomeres and oocytes. (**a**–**d**) The correlation plots of the reads of plus and minus strands for a single blastomere of a four-cell stage embryo (**a**), a single wild-type oocyte (**b**), a single *Dicer1*$^{-/-}$ oocyte (**c**) and a single *Ago2*$^{-/-}$ oocyte (**d**).

detection, which probably results from cross-hybridization on the microarrays. It is also possible that for some genes with low expression, their expression can be stochastically on or off in single cell, and these genes were probably not expressed in the individual cell analyzed by our mRNA-Seq assay[19,20].

Then we determine whether our mRNA-Seq assay identified more RefSeq transcripts compared to the microarray assay. For the four-cell stage embryo, our mRNA-Seq assay detected 4,243 more known transcripts, which is 60% more genes compared with the Affymetrix microarray data[18]. Notably, the mRNA-Seq assay detected another 1,027 transcripts that did not have probes on the arrays. Using our mRNA-Seq assay we found that in a single blastomere of a four-cell embryo, at least 61.4% of the known genes (11,920 out of 19,400 genes) were expressed and coexisted in the same cell at the same time point. To get another independent line of support for our assay, we compared our mRNA-Seq assay with the NIH mouse 60-mer array data for four-cell stage embryos[21] (**Fig. 3c**). We found very similar gene expression patterns. To confirm the accuracy of our mRNA-Seq data, we choose 380 genes known to be involved in early embryonic development and found that for the 266 genes detected by our mRNA-Seq assay, 188 (71%) of them were clearly detected by real-time PCR (**Fig. 3d** and **Supplementary Table 3** online).

Next we determined whether our mRNA-Seq assay could find new splice isoforms from the 50 base reads. We started from known gene exons and generated all possible combinations of exon-exon junctions as 84-bp sequences with 42 bases from each exon (~2 million splice junctions for all exons in Refseq). Then, we removed the known exon junctions. We used the remaining junctions as a reference and matched the reads already aligned to the genome (but not matching to known junctions) to this reference. For one blastomere, there were 6,701 and 1,753 new junctions with at least two reads or five reads, respectively, which illustrates the power of mRNA-Seq to find new splice isoforms *de novo* (**Supplementary Table 4** online). To confirm these splice junctions, we checked eight of them by real-time PCR and found that all were clearly detected (**Supplementary Table 5** online). We also found 9,012 and 2,070 new splice junctions with at least two reads or five reads, respectively, from a single mature oocyte. This also demonstrated the splice complexity within an individual cell. Meanwhile, we mapped 2–3 million reads to the known exon-exon junctions and asked whether there were any genes expressed with more than two transcript isoforms in the same cell. We found that about 335 genes (19% of all known genes with at least two known isoforms) expressed more than two transcript isoforms in a single blastomere (**Supplementary Table 6** online). This was also the case for wild-type, *Dicer1*$^{-/-}$ and *Ago2*$^{-/-}$ oocytes. To our knowledge, this was the first time that hundreds of genes have been shown to express multiple transcript isoforms in the same cell at the same time point, which unambiguously demonstrated the complexity of the transcript variants within individual cells at whole-genome scale.

Applying the analysis to *Dicer1*$^{-/-}$ and *Ago2*$^{-/-}$oocytes

Finally, we determined whether mRNA-Seq assay could be used to dissect the functional consequences when one of the critical genes for microRNA synthesis, *Dicer1*, or when a core component of the RNA-induced silencing complex, *Ago2*, was conditionally knocked out during oocyte development[22–24]. We obtained ~50 million

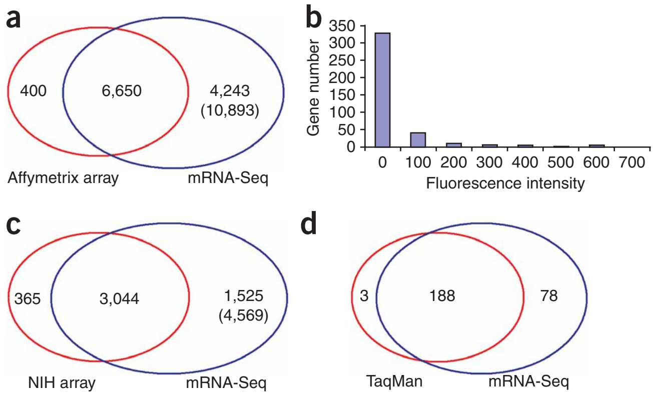

Figure 3 | Comparison of mRNA-Seq and microarray assays. (**a**) Analysis of RefSeq genes that have probes on the Affymetrix array and were detected by array and mRNA-Seq. (**b**) Fluorescence intensity distribution for 400 transcripts detected by microarray only. (**c**) Analysis of RefSeq genes that have probes on the NIH array and were detected by array and mRNA-Seq. (**d**) Of 380 selected genes known to be involved in early embryonic development, 266 genes detected by mRNA-Seq were also analyzed by TaqMan real-time PCR assays. For microarray platforms, we used manufacturer's recommendation on detection calls. For TaqMan measurements, detection was defined as Ct < 33.

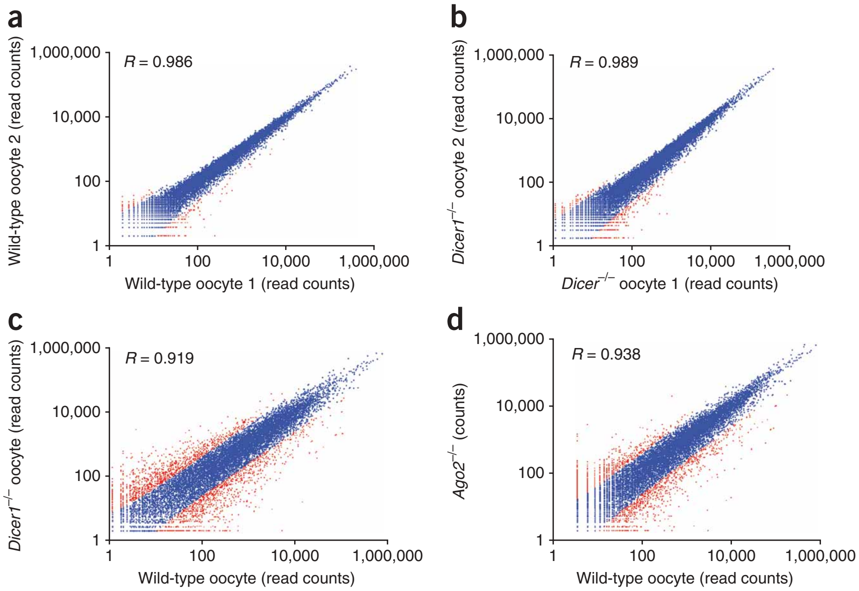

Figure 4 | Correlation plots of the quantile-normalized mRNA-Seq reads for oocytes. (**a**) One wild-type oocyte versus another wild-type oocyte. (**b**) One *Dicer1*$^{-/-}$ oocyte versus another *Dicer1*$^{-/-}$ oocyte. (**c**) Wild-type oocyte versus *Dicer1*$^{-/-}$ oocyte. (**d**) Wild-type oocyte versus *Ago2*$^{-/-}$ oocyte. All of the reads with more than fourfold changes were plotted in red.

reads for *Dicer1*$^{-/-}$, *Ago2*$^{-/-}$ and wild-type mature oocytes (of a mixed genetic background of 129/sv, C57BL/6J and MF1 strains) (**Fig. 4**). To determine the reproducibility of our assay, we compared the sequence data for two separately processed single wild-type mature oocytes and found that they showed very similar transcriptome profiles ($R = 0.986$; **Fig. 4a**). This was also the case for two independently processed *Dicer1*$^{-/-}$ single oocytes ($R = 0.989$; **Fig. 4b** and **Supplementary Fig. 5** online). The transcriptome difference between *Ago2*$^{-/-}$ and wild-type oocytes ($R = 0.938$; **Fig. 4d**) was clearly less than that between *Dicer1*$^{-/-}$ and wild-type oocytes ($R = 0.919$; **Fig. 4c**), which is consistent with the fact that the phenotype of *Ago2*$^{-/-}$ oocytes is similar to but milder than that of *Dicer1*$^{-/-}$ oocytes (M. Kaneda, F.T. and M.A.S.; unpublished data). We mapped 50-base reads to the mouse genome and counted reads aligning to known exons. In **Figure 5**, we present the coverage we obtained with this method for *Dicer1*, *Ccne1*, *Dppa5* (also known as *Esg1*) and *Klf2* (**Fig. 5**). The reads were found in exons with sharp boundaries at the exon-intron junction, confirming the single-exon resolution of the mRNA-Seq reads. There was a loss of reads that map to exon 23 whereas the reads from neighboring exons were intact in the single *Dicer1*$^{-/-}$ oocyte. In the *Dicer1*$^{-/-}$ oocytes, exon 23 of the *Dicer1* gene is deleted by *loxP*-directed Cre recombination[25]. We observed a loss of reads mapped to exon 23 of *Dicer1*, whereas the reads from neighboring exons remained intact (**Fig. 5a**) and we confirmed the loss of exon 23 in *Dicer1*$^{-/-}$ oocyte by exon-specific TaqMan PCR assays (**Supplementary Table 7** online). This demonstrated that our single cell mRNA-Seq assay is accurate and has low or even no background noise. We also found clear upregulation of *Ccne1*, *Dppa5* and *Klf2* at single-exon resolution in *Dicer1*$^{-/-}$ and *Ago2*$^{-/-}$ oocytes compared with wild-type controls, which we confirmed by real-time PCR (**Supplementary Fig. 6** online). Abnormal upregulation of these

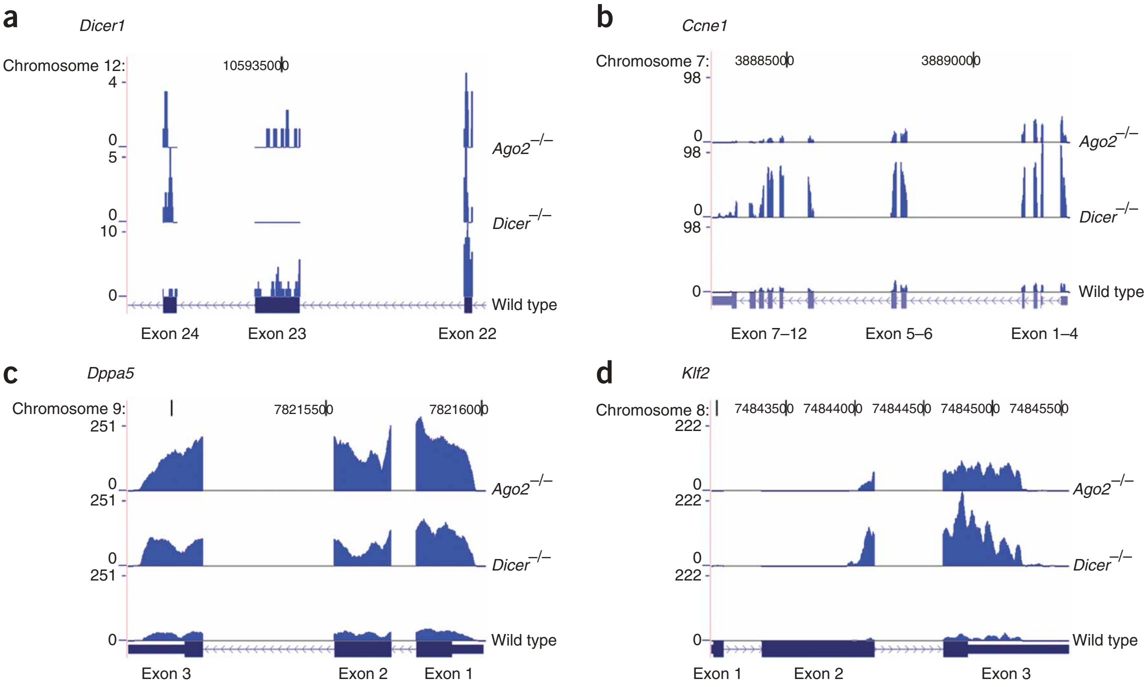

Figure 5 | Coverage plots. (**a**–**d**) Coverage plots of mRNA-Seq reads for *Dicer1* (**a**), *Ccne1* (**b**), *Dppa5* (**c**) and (**d**) *Klf2* on the UCSC genome browser in single wild-type, *Dicer1*$^{-/-}$ and *Ago2*$^{-/-}$ oocytes. The chromosome positions are shown at the top and genomic loci of the genes are shown at the bottom of each panel.

genes may contribute to the compromised developmental potential of *Dicer1*$^{-/-}$ and *Ago2*$^{-/-}$ oocytes.

Some expressed genes in mature oocytes are repressed by endogenous short interfering RNA through *Dicer1*[26]. For the 22 genes that were upregulated in *Dicer1*$^{-/-}$ oocytes as determined by microarray analysis, we found that 20 of them (91%) were upregulated by our single-cell mRNA-Seq assay. In addition, our mRNA-Seq method detected upregulation of all eight genes that had been previously shown to be upregulated in *Dicer1*$^{-/-}$ oocytes by real-time PCR[26] (**Supplementary Fig. 7** online).

Moreover, in *Dicer1*$^{-/-}$ and *Ago2*$^{-/-}$ oocytes, 1,696 and 1,553 transcripts, respectively, were upregulated compared with wild-type controls (fold change > 4, $P < 0.05$ and false discovery rate < 0.05) (**Supplementary Table 8** online). Among them, 619 transcripts were upregulated in both the *Dicer1*$^{-/-}$ and the *Ago2*$^{-/-}$, which is consistent with the fact that *Dicer1* and *Ago2* are both crucial for the function of microRNAs and endogenous siRNAs[22–24,26]. We also found 1,571 and 1,121 transcripts were downregulated in *Dicer1*$^{-/-}$ and *Ago2*$^{-/-}$ oocytes, respectively, with 589 of them downregulated in both (fold change < 0.25, $P < 0.05$ and false discovery rate < 0.05). These 619 commonly upregulated genes and 589 commonly downregulated genes offer the core candidates to dissect the function of microRNAs and endogenous small interfering RNAs for oogenesis. This also proves that Dicer1 and Ago2 globally control the transcriptome stability of the female germ cell.

DISCUSSION

One of the most widely used single-cell cDNA amplification methods[16,17] achieves quantitative accuracy by restricting the cDNA fragment to 0.85 kb from the 3′ end, which corresponds to about 7% of all full-length cDNAs. We showed that increasing the cDNA length up to 3 kb by extending the incubation time for reverse transcription and the extension time for PCR does not decrease the accuracy of counting the copy number of mRNAs. This improvement permitted us to capture full-length cDNAs for the majority (64%) of expressed genes. In single cells at the same cell cycle stage, hundreds of genes (8–19% of all known genes with at least two known isoforms) simultaneously expressed at least two transcript isoforms. We also identified 1,753 previously unknown splice junctions called by at least five reads from only a single blastomere of a four-cell-stage embryo.

Any fragmentation method followed by a size-selection process during cDNA library preparation will result in a loss of cDNA species shorter than the cutoff of the size selection. Our size selection during set up of the mRNA-Seq library was 80–130 bp, which is unlikely to introduce considerable bias for relatively short mRNAs. In fact, our data showed that there were 283 transcripts in RefSeq shorter than 500 bp; 143 of these short cDNA fragments (50.5%) had at least five reads, which is consistent with 61.4% of all known genes being detected by our mRNA-Seq assay.

The main technical novelty of this work is the combination of an improved unbiased amplification of cDNA from single cells with well over a 100 million reads, or a few gigabases of cDNAs on SOLiD system. This not only allowed us to discover many new transcripts that have been overlooked but also to quantitatively estimate their abundance in the cell by the frequency with which the sequence occurs in the mRNA-Seq reads. The assay can also be used to discover new transcripts and alternative splicing isoforms. This will be of great importance for studies on early embryonic development because there is a high probability that early embryos express unique, new transcripts. As mRNA-Seq provides a digital gene expression measurement, a wider dynamic range of gene expression should be covered with no background noise, which will make expression profiles more accurate. This is particularly important for early embryo studies because some of the key transcription factors are expressed at very low levels. These low-level transcripts would likely be missed by microarrays because of substantial noise. We calculated the widely recognized quantification measurement, reads per kilobase of exon model per million mapped reads (RPKM)[6], for the blastomere sequence reads. We obtained RPKM values of 0.2–12,000 that corresponded to 5–280,000 reads for the blastomere[6,27]. About 0.2 RPKM is equal to one copy of mRNA in a blastomere of the four-cell stage embryo[6,27], which showed that our single cell mRNA-Seq assay can cover five orders of dynamic range.

Our technique has considerable limitations. First, because the single-cell cDNA amplification method relies on poly(T) primer, which can only capture mRNA with a poly(A) tail, mRNAs without poly(A) tails, such as histone mRNAs, will not be detected by our mRNA-Seq assay[28]. Second, for most of the mRNAs longer than 3 kb, the 5′ end that is more than 3 kb away from the 3′ end of the mRNA will not be detected. Third, the assay uses double-stranded cDNAs but cannot discriminate between sense and antisense transcripts. All these limitations await future technical improvement.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Accession codes. Gene Expression Omnibus (GEO): GSE14605.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank C. Lee for excellent technical help. The work was supported by grants from the Wellcome Trust to M.A.S.

AUTHOR CONTRIBUTIONS

K.L. designed the project. C.B., B.B.T., A.S., X.W. and K.L. contributed to data analysis, F.T. and M.A.S. contributed to the cDNA sample preparation, E.N., N.X. and Y.W. constructed libraries, C.L. and J.B. contributed to the library sequencing, F.T., E.N. and K.L. contributed to experimental validation, F.T., K.L. and M.A.S. wrote manuscript.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Mardis, E.R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24**, 133–141 (2008).
2. Wold, B. & Myers, R.M. Sequence census methods for functional genomics. *Nat. Methods* **5**, 19–21 (2008).
3. Schuster, S.C. Next-generation sequencing transforms today's biology. *Nat. Methods* **5**, 16–18 (2008).
4. Cloonan, N. & Grimmond, S.M. Transcriptome content and dynamics at single-nucleotide resolution. *Genome Biol.* **9**, 234 (2008).
5. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).

6. 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).
7. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5**, 613–619 (2008).
8. Sultan, M. *et al.* A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**, 956–960 (2008).
9. Wang, E.T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
10. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517 (2008).
11. Pan, Q., Shai, O., Lee, L.J., Frey, B.J. & Blencowe, B.J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415 (2008).
12. Li, H. *et al.* Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. *Proc. Natl. Acad. Sci. USA* **105**, 20179–20184 (2008).
13. Saitou, M., Barton, S.C. & Surani, M.A. A molecular programme for the specification of germ cell fate in mice. *Nature* **418**, 293–300 (2002).
14. Chambers, I. *et al.* Nanog safeguards pluripotency and mediates germline development. *Nature* **450**, 1230–1234 (2007).
15. Toyooka, Y., Shimosato, D., Murakami, K., Takahashi, K. & Niwa, H. Identification and characterization of subpopulations in undifferentiated ES cell culture. *Development* **135**, 909–918 (2008).
16. Kurimoto, K. *et al.* An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Res.* **34**, e42 (2006).
17. Kurimoto, K., Yabuta, Y., Ohinata, Y. & Saitou, M. Global single-cell cDNA amplification to provide a template for representative high-density oligonucleotide microarray analysis. *Nat. Protoc.* **2**, 739–752 (2007).
18. Maekawa, M., Yamamoto, T., Kohno, M., Takeichi, M. & Nishida, E. Requirement for ERK MAP kinase in mouse preimplantation development. *Development* **134**, 2751–2759 (2007).
19. Blake, W.J., Kærn, M., Cantor, C.R. & Collins, J.J. Noise in eukaryotic gene expression. *Nature* **422**, 633–637 (2003).
20. Raser, J.M. & O'Shea, E.K. Noise in gene expression: origins, consequences, and control. *Science* **309**, 2010–2013 (2005).
21. Hamatani, T., Carter, M.G., Sharov, A.A. & Ko, M.S. Dynamics of global gene expression changes during mouse preimplantation development. *Dev. Cell* **6**, 117–131 (2004).
22. Tang, F. *et al.* Maternal microRNAs are essential for mouse zygotic development. *Genes Dev.* **21**, 644–648 (2007).
23. Murchison, E.P. *et al.* Critical roles for Dicer in the female germline. *Genes Dev.* **21**, 682–693 (2007).
24. O'Carroll, D. *et al.* A Slicer-independent role for Argonaute 2 in hematopoiesis and the microRNA pathway. *Genes Dev.* **21**, 1999–2004 (2007).
25. de Vries, W.N. *et al.* Expression of Cre recombinase in mouse oocytes: A means to study maternal effect genes. *Genesis* **26**, 110–112 (2000).
26. Tam, O.H. *et al.* Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**, 534–538 (2008).
27. Rambhatla, L., Patel, B., Dhanasekaran, N. & Latham, K.E. Analysis of G protein alpha subunit mRNA abundance in preimplantation mouse embryos using a rapid, quantitative RT-PCR approach. *Mol. Reprod. Dev.* **41**, 314–324 (1995).
28. Marzluff, W.F., Wagner, E.J. & Duronio, R.J. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. *Nat. Rev. Genet.* **9**, 843–854 (2008).

ONLINE METHODS

Isolation of oocytes and embryos. Four-cell stage embryos were recovered from MF1 strain females mated with MF1 strain male mice[29]. The zona pellucida was removed by treatment with acidic tyrode solution. The individual blastomeres were separated by gentle pipetting using a glass capillary in calcium-free medium. Mature oocytes were isolated from *Dicer* knockout [*Dicer*$^{-/Flox}$, *Zp3-Cre*] and *Ago2* knockout [*Ago2*$^{-/Flox}$, *Zp3-Cre*] female mice[22,24]. Cumulus cells were removed from oocytes by treatment with hyaluronidase. All the 'mature oocytes' mentioned in the text are ovulated mature oocytes.

Knockout mice. The knockout mice carrying *Dicer1* conditional allele (floxed) or *Ago2* floxed allele were described previously[22,24]. The *Dicer1*$^{Flox/Flox}$ mice were mated with *Zp3-Cre* transgenic mice, which express *Cre* recombinase under the control of zona pellucida glycoprotein 3 promoter[25]. Then the [*Dicer1*$^{Flox/+}$, *Zp3-Cre*] female mice were mated with *Dicer1*$^{Flox/Flox}$ male mice. From this mating, we obtained [*Dicer1*$^{-/Flox}$, *Zp3-Cre*] mice, and the deletion of the floxed allele in the oocyte generated oocytes that are the null mutant for *Dicer1*. The *Ago2*$^{Flox/Flox}$ mice were mated with *Zp3-Cre* transgenic mice. Then the [*Ago2*$^{Flox/+}$, *Zp3-Cre*] female mice were mated with *Ago2*$^{Flox/Flox}$ male mice. From this mating, we obtained [*Ago2*$^{-/Flox}$, *Zp3-Cre*] mice, and the deletion of the floxed allele in the oocyte generated oocytes that are the null mutant for *Ago2*.

Single-cell cDNA preparation. Single-cell cDNA amplification was based on a previous protocol[17], with modifications to make it more efficient and suitable for mRNA-Seq[16,17]. Briefly, a single oocyte or blastomere was picked and lysed. Then the mRNAs in the lysate were reverse-transcribed into cDNAs by poly(T) primer with anchor sequence (UP1) by incubating at 50 °C for 30 min and then reverse transcriptase was inactivated by incubation at 70 °C for 15 min. After this, the nonreactive primers were digested by exonuclease I. Then a poly(A) tail was added to the first-strand cDNAs at the 3′ end by terminal deoxynucleotidyl transferase. Next, the second-strand cDNAs were synthesized by poly(T) primer with another anchor sequence (UP2). Then these cDNAs were evenly amplified by PCR of 20 cycles of 95 °C for 30 s, 67 °C for 1 min, 72 °C for 6 min (plus 6 s more after each cycle). After purification, a portion of these cDNAs was further amplified by nine cycles of PCR using a poly(T) primer with an anchor sequence containing a 5′ end–blocked by amine at the C6 position (NH2-UP1 and NH2-UP2; see **Supplementary Table 9** online for sequences). After purification, these cDNAs are ready for subsequent mRNA-Seq assays.

mRNA-Seq library preparation and sequencing. After the generation of target cDNA from a single cell, typically 200–500 ng of cDNA was used with SOLiD system's low-input fragment library preparation. Using the Covaris S2 system (Covaris, Inc.), cDNA (0.5–3 kb) was sheared into 80–130 bp short fragments according to the protocol. The ends of the target DNA were repaired and subsequently ligated to SOLiD P1 and P2 adaptors (**Supplementary Table 9**). The resulting ligated population was resolved on a 6% polyacrylamide gel and the 150–200 bp fraction was excised. The fractionated, adaptor ligated DNA was subjected to 8–10 cycles of PCR amplification. The number of the cycles necessary was determined by the ability to visualize the amplified product on a standard FlashGel (Lonza) from an aliquot of the PCR sample. The amplified PCR products were purified to yield the SOLiD Fragment Library ready for emulsion PCR. Emulsion PCR reactions were performed by mixing 500 pg single cell libraries with 1.6 billion 1-μm-diameter beads with P1 primers covalently attached to their surfaces. The 50-base sequences were obtained using SOLiD sequencing.

TaqMan assays. For TaqMan real-time PCR, 1.0 μl of diluted cDNAs was used for each 10 μl real-time PCR (1× PCR Universal master mix, 250 nM TaqMan probe, 900 nM of each primer, that are commercially available as ready-to-use assays from Applied Biosystems). All reactions were duplicated. The PCR was done as follows in an AB7900 thermocycler (Applied Biosystems) with 384-well plates: 95 °C for 10 min; then 40 cycles of 95 °C for 15 s; and 60 °C for 1 min.

Alignment and algorithm. mRNA-Seq sequencing reads were analyzed using Applied Biosystems' whole transcriptome software tools (http://www.solidsoftwaretools.com/). Briefly, the reads generated by each sample are matched to the mouse genome (mm9, US National Center for Biotechnology Information (NCBI) build 37). Given the size of our 50-base reads relative to average exon length (150 bases) we anticipated that a substantial fraction of reads (up to one third) will cover a splice junction. Hence, these reads will not align contiguously to the genome and standard read mapping methods (for example, mapping and assembly with qualities (MAQ)) will fail. Making the assumption that at least half of each read sequence originates from a contiguous region of the genome, we circumvented this problem by dividing each read into two 25 base fragments from both ends of the reads (with 10 base overlap for 35 bp long reads) and then mapping each fragment to the genome independently using Applied Biosystems' color mapping tool. During this mapping phase we allowed up to two mismatches and removed reads that aligned to more than 10 locations. The mapping of each half was extended along the mapped genomic region using colors from the other half until a maximal score was reached (+1 for a match and −1 for a mismatch; **Supplementary Fig. 8** online). Where this extension yielded a full-length read, the results from the two halves were merged. Matching locations were subsequently used to generate counts for annotated features, exons, transcripts (**Supplementary Figs. 9 and 10** online) or genes (we used University of California Santa Cruz (UCSC) RefSeq Genes track for exons genomic locations of known transcripts) or coverage files (wiggle format; **Supplementary Fig. 11** online). Reads that were not aligned to their full length were subjected to the alternative splicing analysis, in which reads were aligned to a reference-containing exon-exon junctions, 45 bases on each side for known junctions and 42 bases on each side for new junctions, allowing up to four mismatches for the full length of the read (50 bases). The observed junction position within reads ranged uniformly between positions 8 and 42 (data not shown), which suggested that this approach may generate a misalignment error rate less than 1%. Approximately 90% of reads matching to known exon-exon junctions reference, mapped to a unique location.

Coverage design. The core engine of the whole-transcriptome software is in the Applied Biosystems color mapping tool. Covering designs have been described previously[30]. When applied to the short sequence mapping problem, they can help reduce the size of the hash table and the number of table look-ups while matching a read sequence. Basically, when a covering (s, t, m) is used for matching sequences of size s to the genome, instead of strictly considering m mismatches, initially $t > m$ mismatches are tolerated. Then, found hits are processed and those having at most m mismatches are retained. This process is repeated for each pattern in the cover, which ensures that all possible m mismatches are considered.

As an example, consider the cover (7, 3, 2) consisting of seven patterns (**Supplementary Fig. 12** online). This can be used to match two sequences of size seven with at most two mismatches. When this cover is used, a hit was reported for a particular pattern if the two sequences being matched are the same at the positions denoted by the ones even though they may have mismatches at the remaining positions. If at least one hit is detected over all patterns in the cover, exact number of mismatches is determined, and the hit is retained if there are no more than two mismatches between the two sequences. By initially tolerating three mismatches, this cover ensures that all cases with at most two mismatches are detected. Note that there are 21 cases that need to be checked if an enumeration-based method is used to achieve the same result. Hence, for this example the number of comparisons is reduced threefold owing to using this covering.

Error detection. Dual interrogation dramatically reduces sequencing errors (http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_057511.pdf). In the simplest case of an individual single-nucleotide polymorphisms, a true polymorphism will require a change at two adjacent positions in the sequence. Changes at a single position are identified as random errors and can be removed by the software in data analysis. Furthermore, only 3 possible consecutive mismatches that can be caused by a single base change out of all 16 possible pairs, the remaining 12 possible two-color changes being removed. In case of more complex variations such as multiple single-nucleotide polymorphisms or insertion-deletions, more complicated algorithms are used.

29. Nagy, A., Gertsenstein, M., Vintersten, K. & Behringer, R. Recovery and in vitro culture of preimplantation stage embryos. in *Manipulating the Mouse Embryo* 3rd edn. 194–200 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2003).

30. Gordon, D.M., Patashnik, O. & Kuperberg, G. New constructions for covering designs. *J. Comb. Designs* **3**, 269–284 (1995).

doi:10.1038/nmeth.1315