CORRECTION NOTICE

Nat. Methods; doi:10.1038/nmeth.1315

mRNA-Seq whole-transcriptome analysis of a single cell

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

In the version of this supplementary file originally posted online, Supplementary Figure 5a in was a duplicate of Supplementary Figure 5b. The error has been corrected in this file as of 19 April 2009.

nature | methods

mRNA-Seq whole-transcriptome analysis of a single cell

Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu,

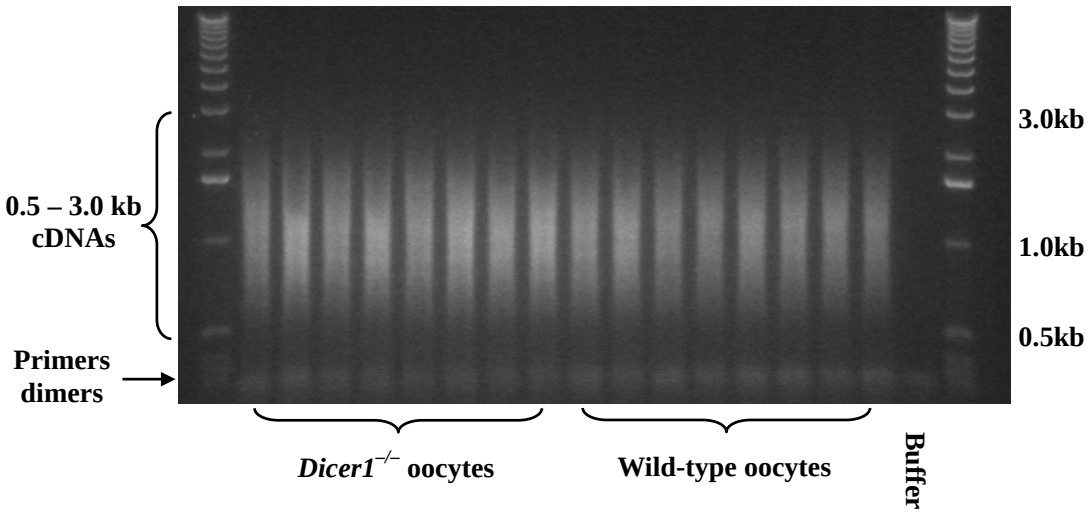
Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao & M Azim Surani

Supplementary figures and text:

Supplementary Figure 1	The cDNA products amplified from single wild-type and *Dicer1*$^{-/-}$ oocytes.
Supplementary Figure 2	The pie charts of the number of the mRNA-Seq reads for single cells analyzed.
Supplementary Figure 3	The estimated instrument error rate of SOLiD system.
Supplementary Figure 4	Workflow of matching analysis for 50 bases reads.
Supplementary Figure 5	The reproducibility of SOLiD library preparation.
Supplementary Figure 6	The correlation plots of the fold changes that are determined by mRNA-Seq reads and TaqMan real-time PCR.
Supplementary Figure 7	Compared upregulated genes listed in Hannon's paper in *Dicer1*$^{-/-}$ oocytes with our mRNA-Seq.
Supplementary Figure 8	Alignment score for mRNA-Seq reads.
Supplementary Figure 9	Base coverage of the single cell mRNA-Seq assay in a single mature oocyte.
Supplementary Figure 10	21,436 known transcripts are compared between two different wild-type oocytes using three normalization methods
Supplementary Figure 11	Visualization (UCSC genome browser) of chromosome 9 using the wiggle output file (showing details for three genes: *Omt2a*, *Omt2b*, and *Ooep*).
Supplementary Figure 12	Seven patterns that can be used for matching two sequences of size seven with at most two mismatches.

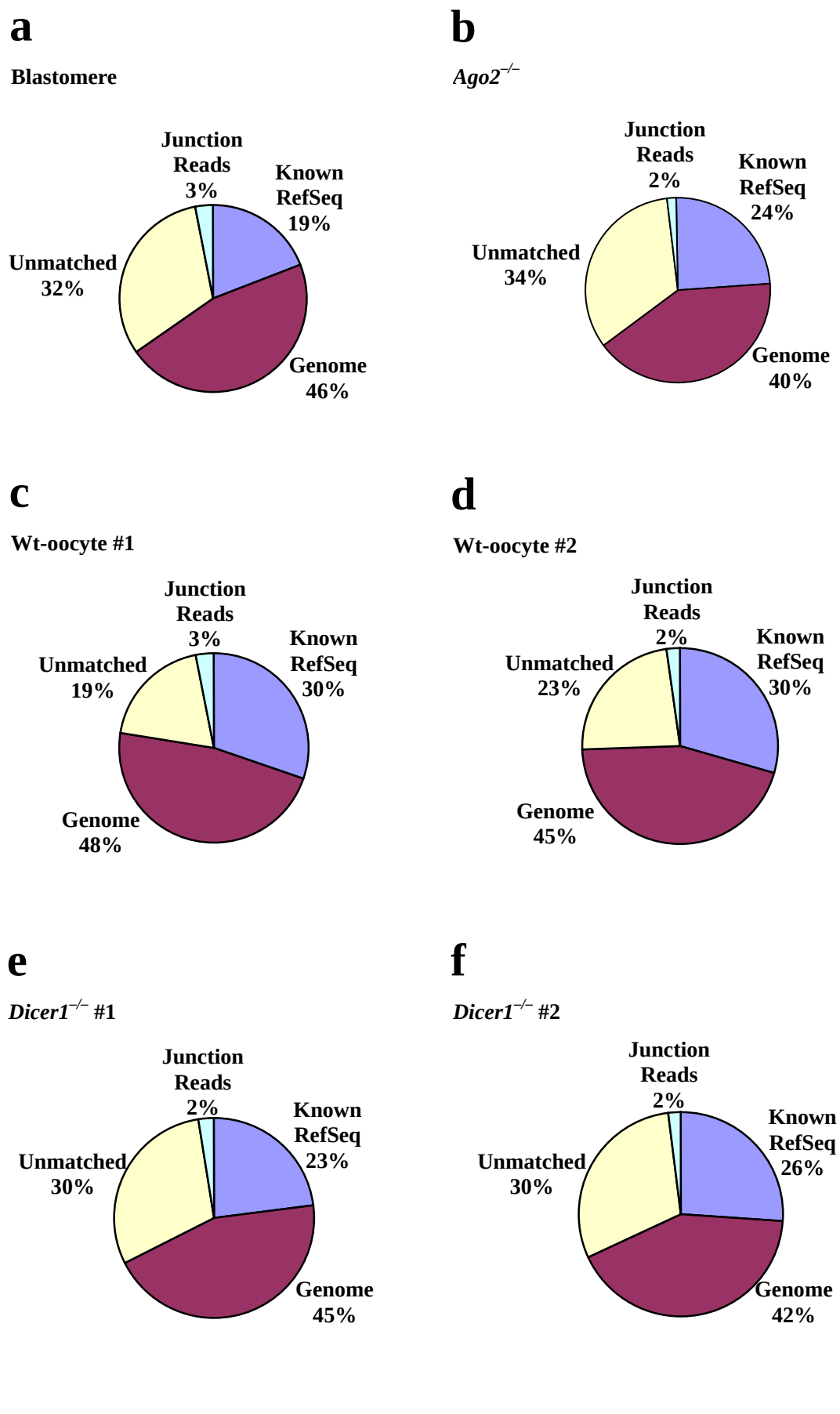
Note: Supplementary Tables 1–9 are available on the Nature Methods website.

Supplementary Figure 1. The cDNA products amplified from single wild-type and *Dicer1*$^{-/-}$ oocytes.



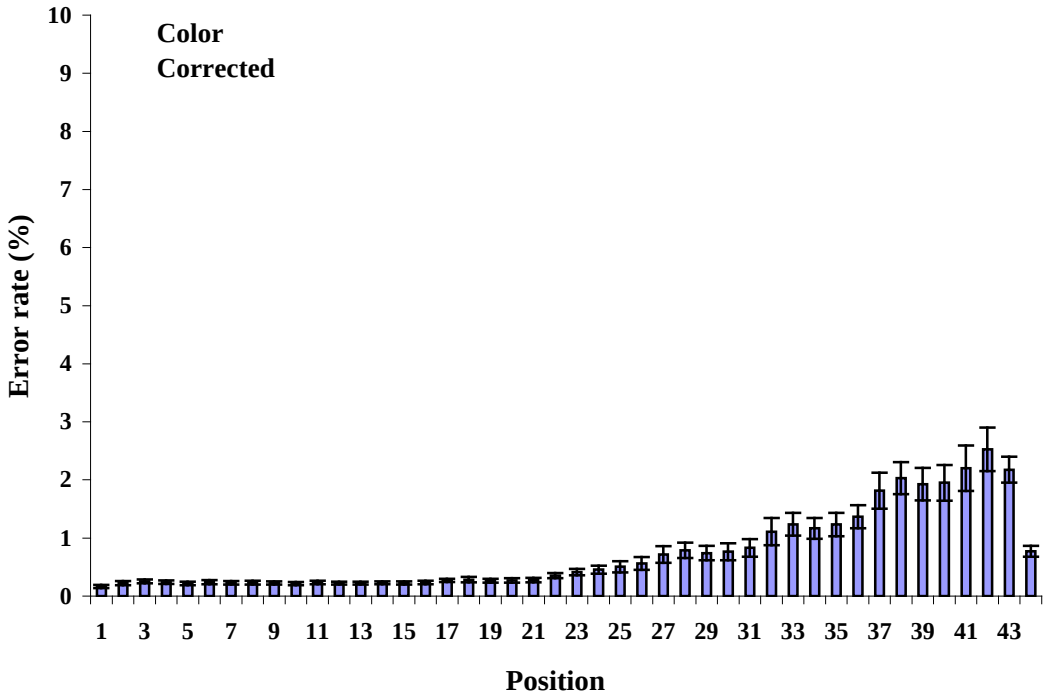
The buffer lane is the negative control that omits the single cell but just picks buffer carryover for the reverse transcription reaction.

Supplementary Figure 2. The pie charts of the number of the mRNA-Seq reads for single cells analyzed.



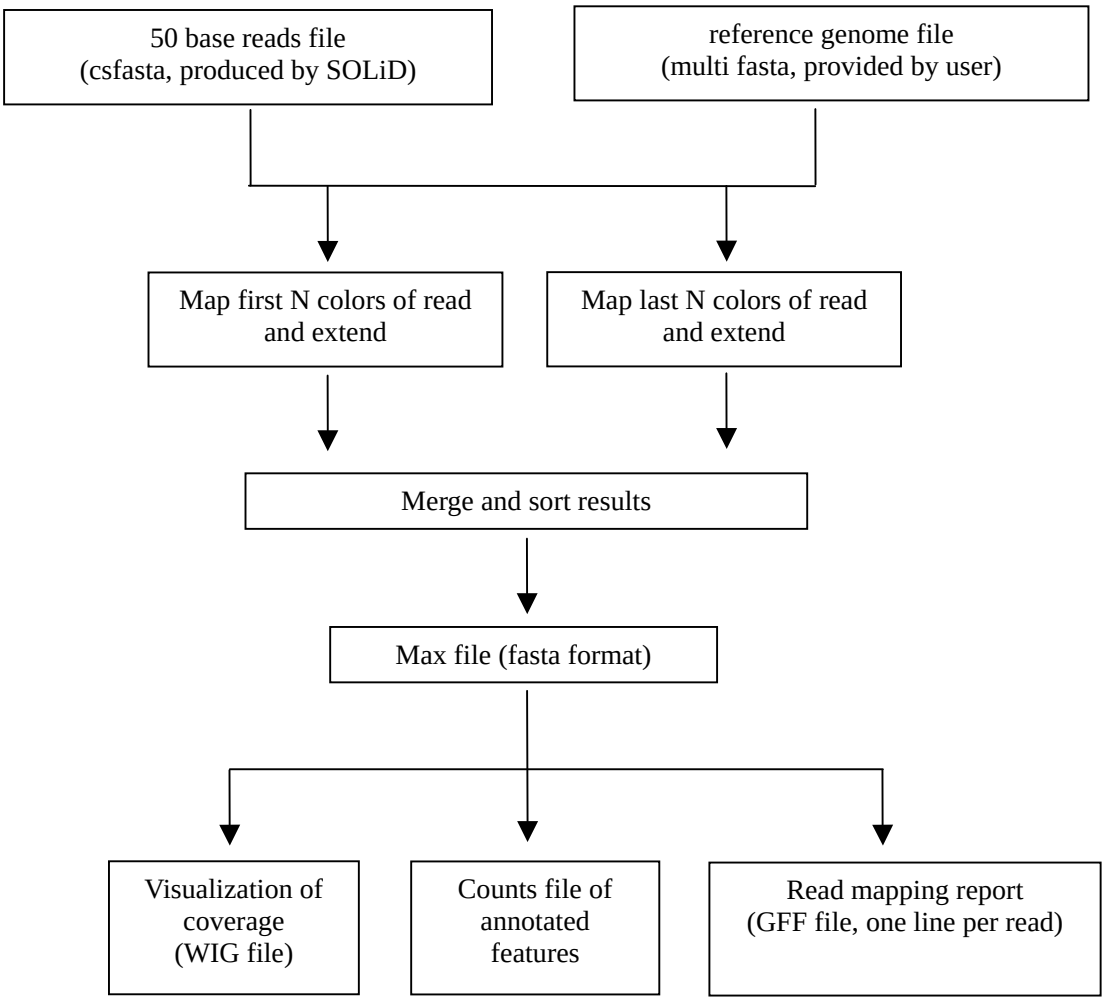
Reads are mapped to the mouse genome (mm9, NCBI Build 37) as described in Alignment and Algorithm section. We use UCSC annotation database (mm9) to determine if matching locations of individual reads correspond to exon regions, or exon-exon junctions of known transcripts. The number of these reads as a fraction of the total number of reads produced by each run is represented in these pie charts. We generated 350 millions reads in total (**Table 1**). We obtained about 66 - 81% of reads that mapped uniquely to Refseq, known junctions, and genome. There are about 2 - 3 % reads that mapped uniquely to known exon junctions.

Supplementary Figure 3. The estimated instrument error rate of SOLiD system.

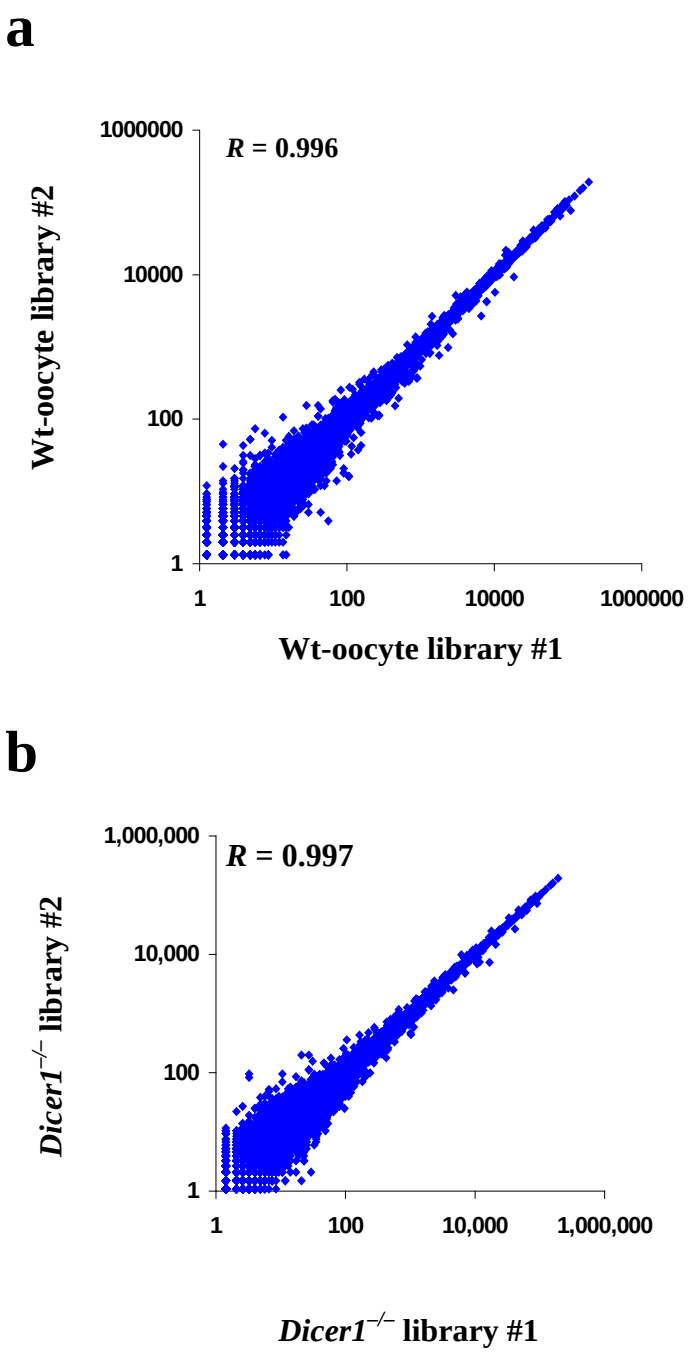


Reads that aligned contiguously to the genome (as described in Alignment and Algorithm section), on full length and in a unique place are used to estimate instrument error rate as a function of position within the read (see Error detection section). For a given position, the number of times we see a difference between the read call and corresponding matching location, calculated as a fraction of all reads considered, is represented on the y-axis. Data produced from all runs are aggregated and represented as average error rates, and 95% confidence intervals of these estimates are inferred. The averaged error rate is about 0.5% for the first 30 positions. For 50-mer reads, we only plotted the first 44 bases since the extension step reaches full length of the read if reads have fewer errors on the last positions (and so the error rates of the last positions for the subset of reads we used are under estimated).

Supplementary Figure 4. Workflow of matching analysis for 50 bases reads.

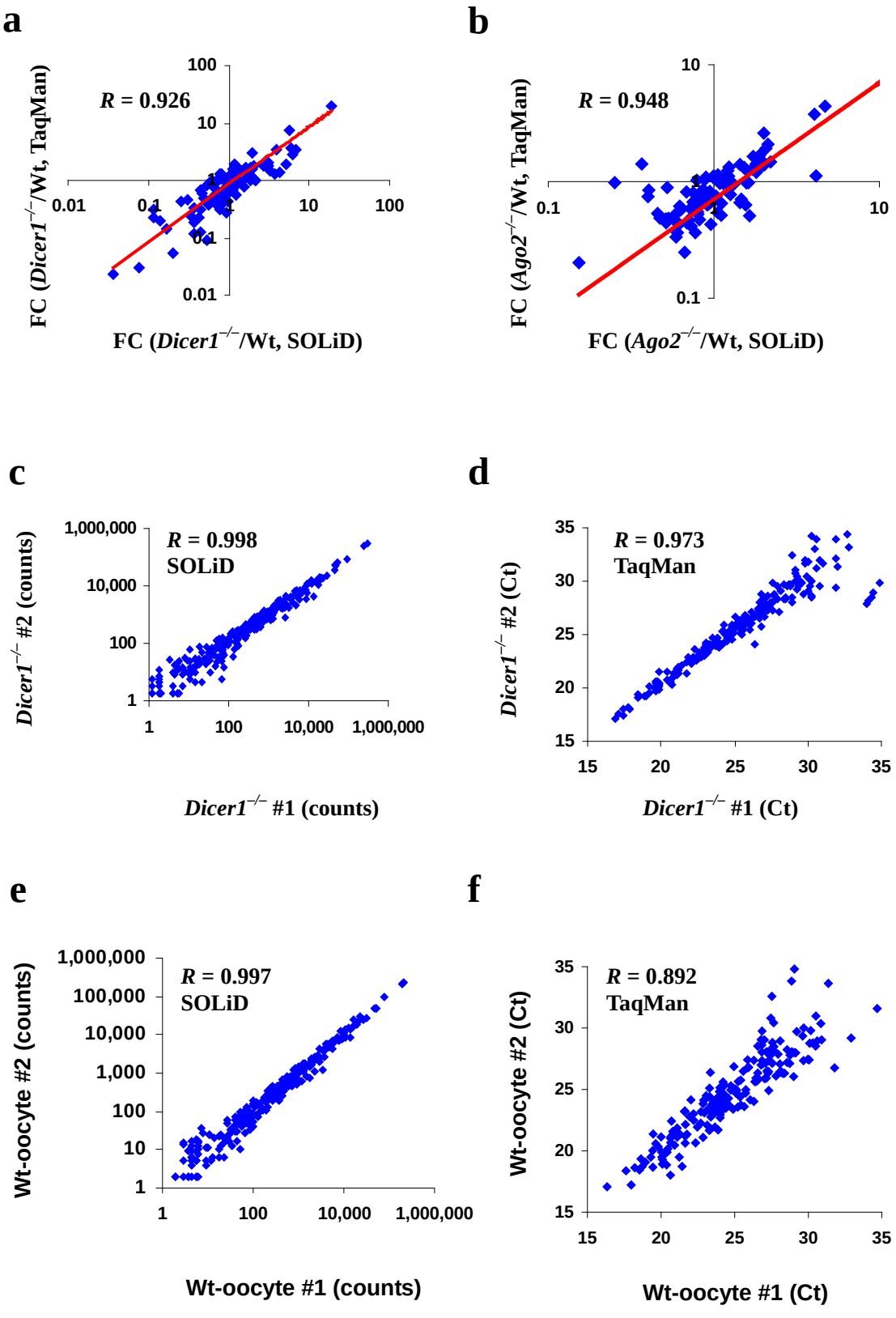


Supplementary Figure 5: The reproducibility of SOLiD library preparation.



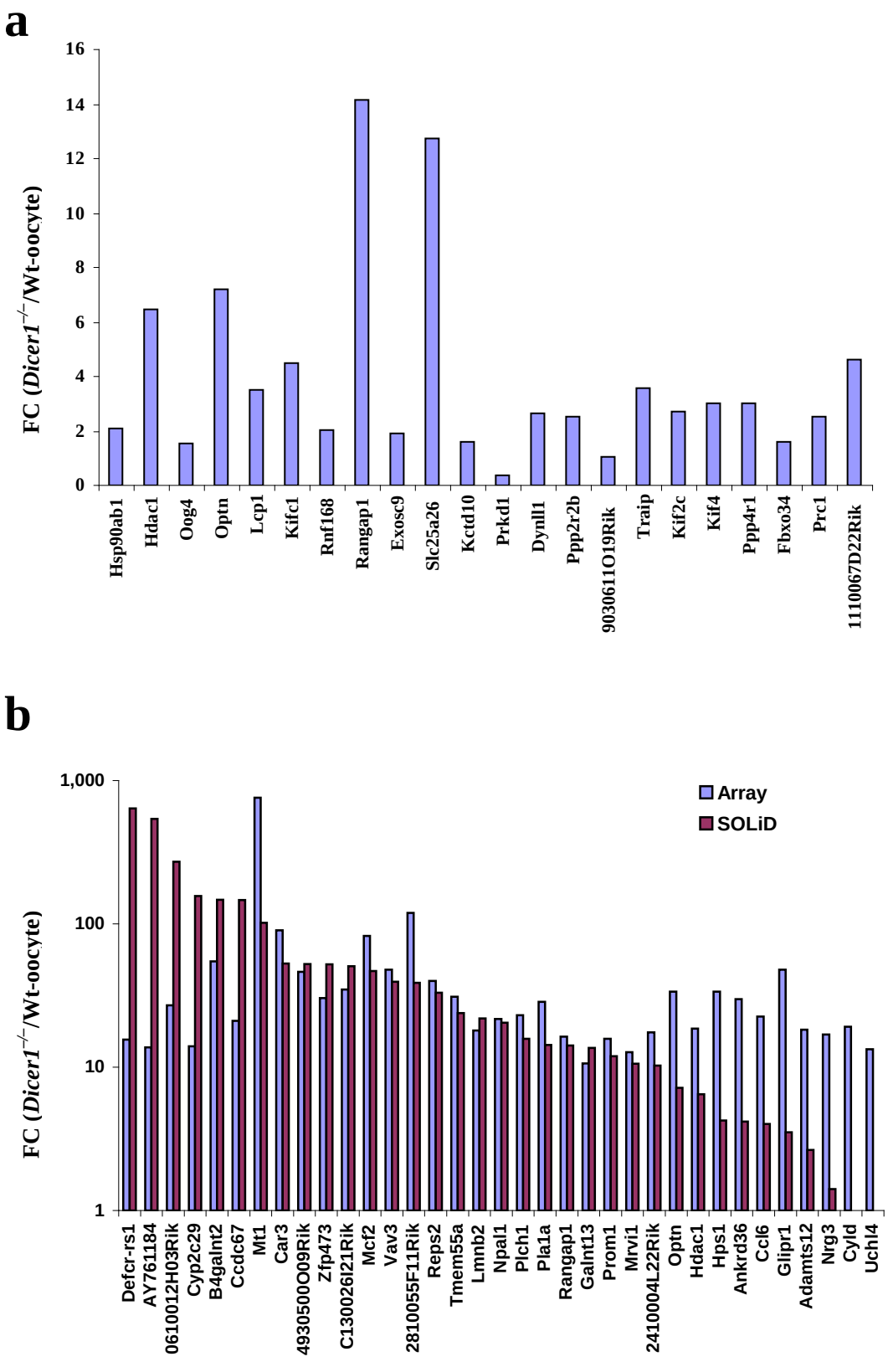
Two independent libraries were prepared from the same cDNA samples of (a) a single Wildtype oocyte and (b) a single *Dicer1*$^{-/-}$ oocyte. This confirms the accuracy of the sampling of the single cell cDNAs and the reproducibility of the library construction.

Supplementary Figure 6: The correlation plots of the fold changes that are determined by mRNA-Seq reads and TaqMan real-time PCR.



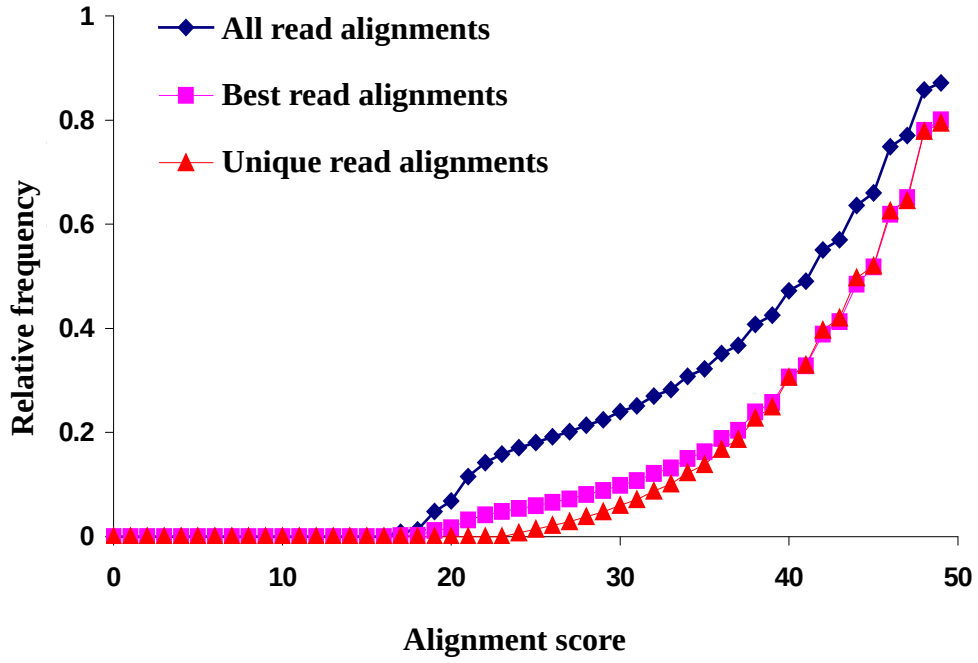
(a) *Dicer1*$^{-/-}$/Wt-oocyte and (b) *Ago2*$^{-/-}$/Wt-oocyte. Here, the top 100 most abundant genes based on the Ct values of Wt-oocyte were plotted. All of the concordance Pearson coefficients are > 0.99 for the positive and negative hits of wild-type, *Dicer1*$^{-/-}$, and *Ago2*$^{-/-}$ oocytes (**Supplementary Table 1** online). The Ct values were normalized based on the *Hprt1* gene that has been verified to have similar expression level for 20 of single wild-type, *Dicer1*$^{-/-}$, and *Ago2*$^{-/-}$ mouse mature oocytes. The Pearson correlation significantly decreases for these genes with Ct value > 33. The main reason was due to the sampling errors in the TaqMan assays where 200 ng cDNAs were diluted by 800-fold in each reaction well for duplication of 384 assays, while mRNA-Seq used all 200 ng cDNAs to make the libraries that result in better detection sensitivity. The number of reads that match each gene annotated in the UCSC database mm9 is used to estimate the fold change between samples. For TaqMan measurements the "delta Ct" method was used to estimate (log2) fold change between samples. The mRNA-Seq reads and Ct values are also highly correlated for *Dicer1*$^{-/-}$ #1 vs *Dicer1*$^{-/-}$ #2 (c, d) and Wt-oocyte #1 vs Wt-oocyte #2 mature oocytes (e, f).

Supplementary Figure 7. Compared upregulated genes listed in Hannon's paper in *Dicer1*$^{-/-}$ oocytes with our mRNA-Seq.



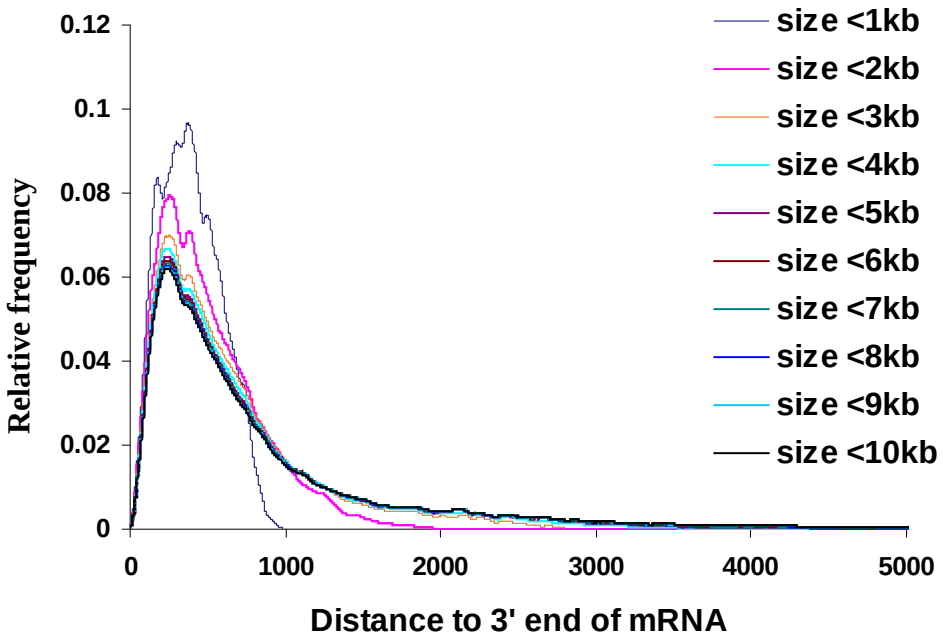
(a) Fold change by our mRNA-Seq for the 22 upregulated genes controlled by endogenous siRNA in *Dicer1*$^{-/-}$ oocytes determined by Affymetrix mouse microarray[26], 20 of them (91%) were confirmed by our mRNA-Seq assay. The 8 genes at the left side were also showed upregulated by real-time PCR in Hannon's paper[26]. (b) Fold change by Hannon's Affymetrix microarray and by our mRNA-Seq assay[23]. 33 of 36 genes (92%) were confirmed by our mRNA-Seq assay.

Supplementary Figure 8. Alignment score for mRNA-Seq reads.



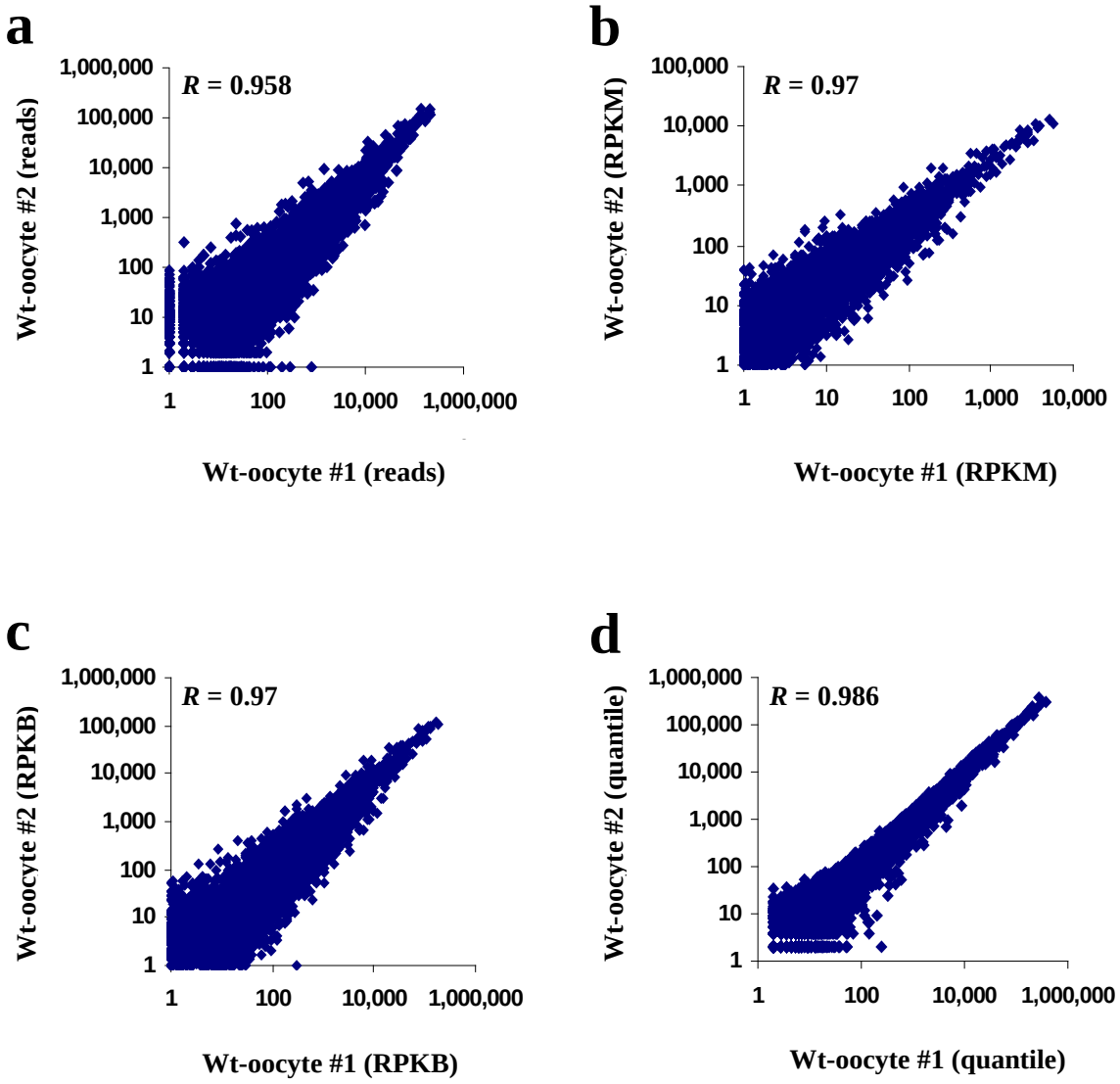
For each read alignment we associate the score obtained by adding one for a color match and subtracting one for a mismatch. This scoring function is used in the extension step (as described in Alignment and Algorithm section) and the (contiguous) alignment producing the highest score is reported. Cumulative distribution of aligned reads from an *Ago2*$^{-/-}$ oocyte run is shown. Low scoring alignments are expected to be produced by reads aligning over splice junctions or low quality reads, while high scoring reads are expected to represent exonic regions.

Supplementary Figure 9. Base coverage of the single cell mRNA-Seq assay in a single mature oocyte.



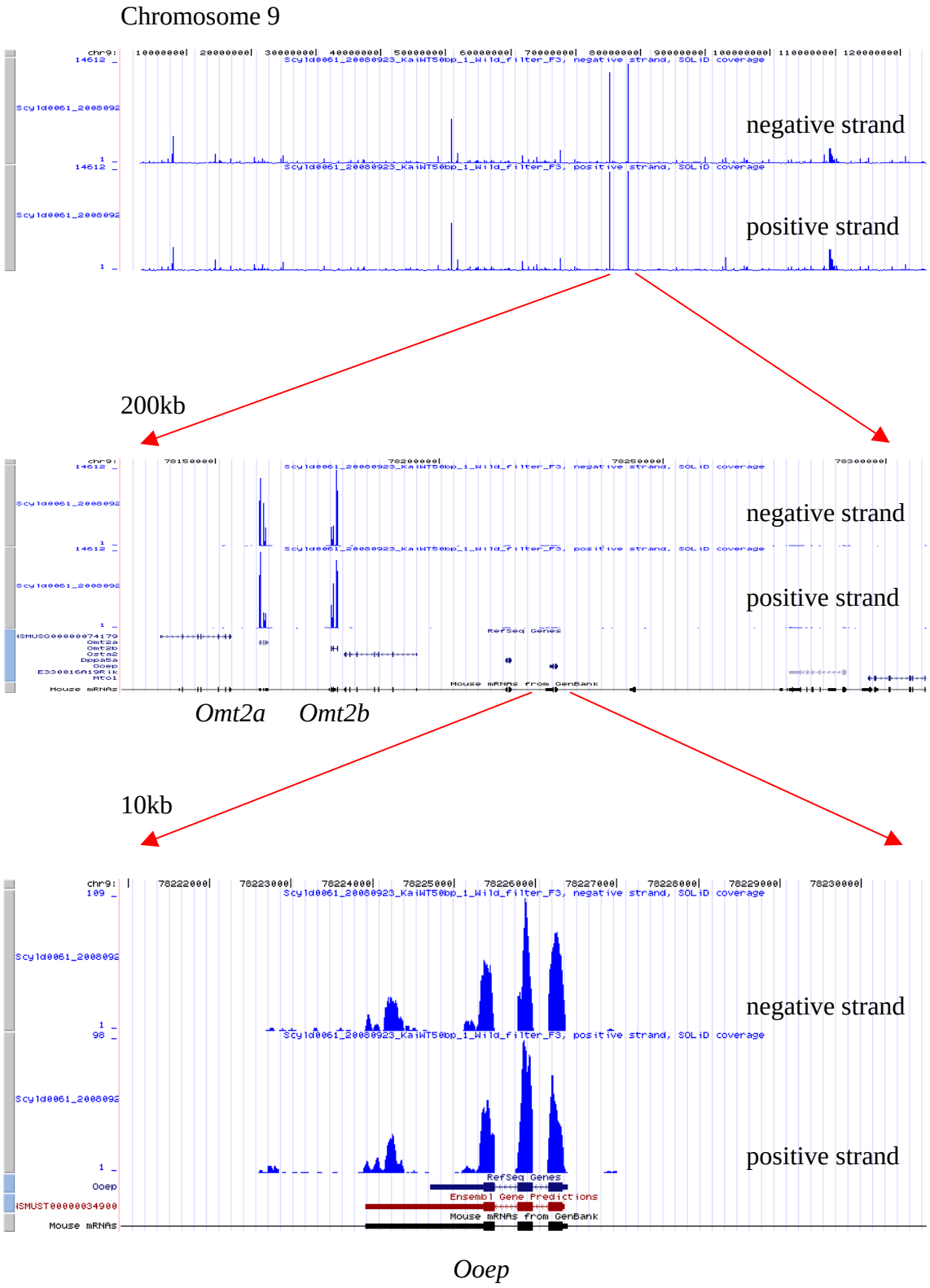
To obtain the coverage length distributions of our cDNAs, we binned all 21,436 transcripts based on their sizes, bin n containing all transcripts of size less than n kb. Base coverage is generated for each bin of transcripts and scaled to the total number of aligned reads. The obtained distribution is represented as a function of the distance to the 3' end of the transcripts. The reads distribution for regions 3 kb away from the 3'end is very limited that agrees with our gel results (in **Supplementary Fig. 1**).

Supplementary Figure 10. 21,436 known transcripts are compared between two different wild-type oocytes using three normalization methods.

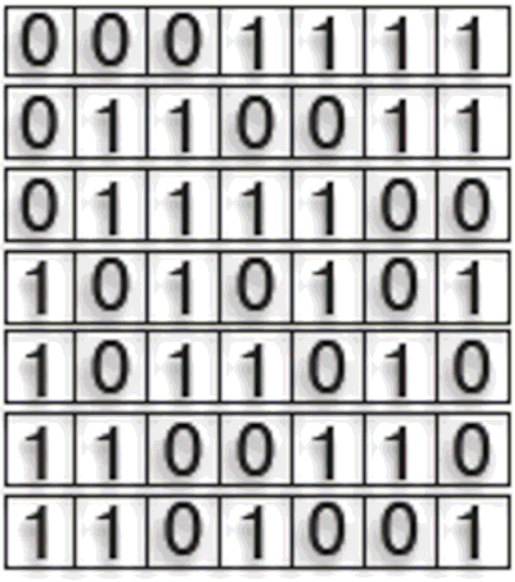


(a) number of reads (without normalization), (b) number of reads per kilobase per million reads[6] (RPKM), (c) number of reads per kilobase (RPKB), and (d) quantile normalized counts. The three normalization methods generate relatively similar Pearson correlation coefficients (0.97, 0.97 and 0.986 respectively) while the quantile normalization generates better Pearson correlation for our mRNA-Seq reads.

Supplementary Figure 11. Visualization (UCSC genome browser) of chromosome 9 using the wiggle output file (showing details for three genes: *Omt2a*, *Omt2b*, and *Ooep*).



Supplementary Figure 12. Seven patterns that can be used for matching two sequences of size seven with at most two mismatches.



These patterns are designed so that for any pair of integers (i,j), i,j $\leq$ 7, it corresponds to a pattern that has zeros on positions i and j.