Biology of Reproduction, 2023, **109(5)**, 601–617
https://doi.org/10.1093/biolre/ioad107
Advance access publication date 5 September 2023
Research Article

Relationship of quantitative reverse transcription polymerase chain reaction (RT-PCR) to RNA Sequencing (RNAseq) transcriptome identifies mouse preimplantation embryo reference genes†

Allison K. Tscherner[1,2], Taylor McClatchie[1,2], Angus D. Macaulay[1] and Jay M. Baltz[1,2,3,*]

[1]Ottawa Hospital Research Institute, Ottawa, ON, Canada
[2]Department of Obstetrics and Gynecology, University of Ottawa Faculty of Medicine, Ottawa, ON, Canada
[3]Department of Cellular and Molecular Medicine, University of Ottawa Faculty of Medicine, Ottawa, ON, Canada

***Correspondence:** Ottawa Hospital Research Institute, mailbox 411, 501 Smyth Rd., Ottawa, ON K1H 8L6 Canada. Tel: (613) 737-8899 x79763; E-mail: jbaltz@ohri.ca

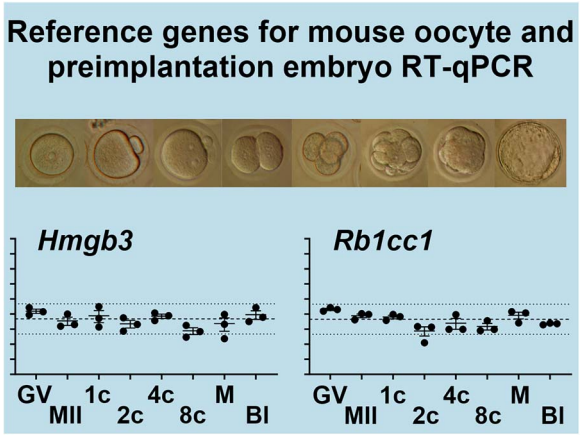
†**Grant Support:** This work was funded by Canadian Institutes of Health Research grant PJT152991.

Abstract

Numerous reference genes for use with quantitative reverse transcription polymerase chain reaction (RT-qPCR) have been used for oocytes, eggs, and preimplantation embryos. However, none are actually suitable because of their large variations in expression between developmental stages. To address this, we produced a standardized and merged RNA sequencing (RNAseq) data set by combining multiple publicly available RNAseq data sets that spanned mouse GV oocytes, MII eggs, and 1-cell, 2-cell, 4-cell, 8-cell, morula, and blastocyst stage embryos to identify transcripts with essentially constant expression across all stages. Their expression was then measured using RT-qPCR, with which they did not exhibit constant expression but instead revealed a fixed quantitative relationship between measurements by the two techniques. From this, the relative amounts of total messenger RNA at each stage from the GV oocyte through blastocyst stages were calculated. The quantitative relationship between measurements by RNAseq and RT-qPCR was then used to find genes predicted to have constant expression across stages in RT-qPCR. Candidates were assessed by RT-qPCR to confirm constant expression, identifying *Hmgb3* and *Rb1cc1* or the geometric mean of those plus either *Taf1d* or *Cd320* as suitable reference genes. This work not only identified transcripts with constant expression from mouse GV oocytes to blastocysts, but also determined a general quantitative relationship between expression measured by RNAseq and RT-qPCR across stages that revealed the relative levels of total mRNA at each stage. The standardized and merged RNA data set should also prove useful in determining transcript expression in mouse oocytes, eggs, and embryos.

Summary Sentence
The quantitative relationship between transcript expression levels determined by RNAseq and RT-qPCR for mouse oocytes through blastocysts was determined and used to find reference genes that have constant expression across all stages by RT-qPCR.

Graphical Abstract



Key words: egg, oocyte, PCR, preimplantation embryo, reference genes, RNAseq.

Received: May 16, 2023. **Revised:** August 4, 2023. **Accepted:** August 25, 2023

Introduction

The transcriptome changes profoundly during mammalian oocyte maturation and preimplantation embryo development. Fully grown oocytes are transcriptionally silent and remain so through meiotic maturation until fertilization of the mature egg [1, 2], which contains only maternal transcripts. After this, there is only limited transcription from either the paternal or maternal genomes until the global switch to expression of the embryonic genome known as zygotic genome activation (ZGA) [2, 3]. The stage at which the major ZGA occurs depends on the species, taking place at the 2-cell stage in the mouse [4]. After ZGA, development is driven by new transcript expression that varies greatly as the embryo progresses through the preimplantation stages [5, 6]. Maternal transcripts that were stored in the oocyte are degraded during maturation and early development with a round of transcript degradation in the oocyte during meiotic maturation [7, 8] and another round of degradation during ZGA [2, 3], when a large proportion of maternal transcripts are eliminated and embryonic transcripts from the maternal and paternal genes replace them. Finally, cell lineage differentiation proceeds during blastocyst development with fully expanded blastocysts having developed three distinct cell types—trophectoderm, epiblast, and primitive endoderm—each with their own gene expression patterns [9]. Thus, the expression profiles of individual gene transcripts and the total amount of mRNA vary substantially over the course of oocyte maturation and preimplantation development because of stage-dependent changes in expression and global maternal transcript degradation and replacement.

Technologies have been developed that allow gene expression to be quantitatively measured even for very small samples such as single preimplantation mammalian embryos or small pools of embryos. Global transcriptomes have been determined using gene array technologies [10] and, most recently, using RNA sequencing (RNAseq) [11]. Application of these technologies has confirmed that virtually every transcript expressed in oocytes, mature eggs, or preimplantation embryos undergoes substantial changes in their levels over the course of development [5, 6].

Commonly, quantitative reverse transcription polymerase chain reaction (RT-qPCR) has been used when determining the expression of only one or a few selected genes. RT-qPCR is often used to follow gene expression over multiple stages of oocyte maturation or preimplantation embryo development. The accepted standard for RT-qPCR is to determine the expression of each gene of interest relative to a reference gene or a combination of several reference genes whose expression remains essentially constant, to control for technical variation [12]. For somatic cells and tissues, it is often relatively straightforward to identify "housekeeping" genes that are expressed at essentially constant levels and which can be used as reference genes to reveal changes in the expression of specific genes of interest. However, this is problematic for preimplantation embryos where no such constantly expressed genes have been identified. Various investigators have tried to find such genes for mouse oocytes and preimplantation embryos, usually by identifying the most "stably expressed" genes determined by ranking a set of candidate transcripts by how much they vary within a set of samples [13–16]. The definition of gene expression stability is not standardized but has usually been determined using software such as NormFinder [17], GeNorm [18], and similar programs [19] which rank a set of transcripts whose expression has been measured across several cell or tissue types or under different experimental conditions to determine the transcripts that exhibit the greatest stability as defined by the program used.

Genes that have been employed as reference genes for RT-qPCR in mouse preimplantation embryos include housekeeping genes commonly used as reference genes in somatic cells and tissues, such as beta actin (*Actb*), glyceraldehyde 3-phosphate dehydrogenase (*Gapdh*), 18S ribosomal RNA (*Rn18s*), and mitochondrial 16S ribosomal RNA (*mt-Rnr2*) [13–16]. However, the levels of these transcripts vary by at least 10-fold and up to 500-fold during meiotic maturation and preimplantation development in mouse and are clearly not suitable for use as reference genes (these relative expression levels of these proposed reference genes are provided in one study [15] and can be estimated from figures in two others [13, 16]). More stable reference genes proposed specifically for mouse embryos include peptidylprolyl isomerase A (*Ppia*), histone H2A.Z (a.k.a. H2AFZ; *H2az1*), hypoxanthine phosphoribosyltransferase 1 (*Hprt1*), and ubiquitin C (*Ubc*). However, despite these being the genes reportedly exhibiting the least variation across oocytes and preimplantation embryos among the small sets of genes tested, they vary considerably among these stages, with variations of 30–150-fold for *Ppia*, 10–130-fold for *H2az1*, 7–24-fold for *Ppia*, and 6–30-fold for *Ubc* [13, 15, 16]. Therefore, none of these are actually suitable for normalizing gene expression across oocyte and preimplantation embryo stages. Similar issues have arisen with reference genes for preimplantation embryos in other species, e.g., bovine [20] or rabbit [21].

Any search for transcripts whose levels remain nearly constant from oocytes through the end of preimplantation embryo development would seem to be very unlikely to succeed if it relies on simply selecting candidate genes and measuring their expression in oocytes and embryos, since even housekeeping genes that have been established to have nearly constant expression in other tissues vary considerably between stages in oocytes and early embryos. Thus, there has been no clear strategy for choosing candidates. Fortunately, with the advent of methods such as RNAseq for determining the levels of global transcript expression for essentially all genes in the genome, it would appear to be possible to identify those transcripts that vary the least among all expressed genes. Conceptually, a viable strategy would be to obtain an RNAseq data set that spans all the stages from oocytes through blastocysts and identify those transcripts with the most constant expression, which could then be used with RT-qPCR as reference genes.

One major issue with such a strategy, however, is that RNAseq and RT-qPCR do not provide directly comparable measures of gene expression. It is generally underappreciated that RT-qPCR and RNAseq provide fundamentally different measurements. For RT-qPCR, the relative or absolute amount of a specific transcript is determined. For relative measurements, the cycle threshold (C_T) is reported and is often compared to that of a reference gene, whereas for absolute measurements, a calibration curve constructed using known amounts of the target sequence is used to calculate the mass or number of transcripts [22]. For RNAseq, however, the quantity of a specific transcript is instead expressed as the fraction of that transcript relative to the total transcripts sequenced, often normalized to transcript length [23]. Because of this, the expression patterns for a transcript determined

by RNAseq and RT-qPCR across oocyte and preimplantation embryo development would be theoretically identical only if the total pool of transcripts remained constant at each stage of development. However, as discussed above, the total amount of mRNA in oocytes and preimplantation embryos varies considerably because of maternal mRNA degradation and stage-specific embryonic gene expression. For example, the total amount of mRNA in mouse embryos falls to a minimum at the 2-cell stage when maternal transcripts have been degraded but embryonic genome expression is just being initiated [24]. For this reason, a transcript that is found to be constant through the 2-cell stage by RNAseq (i.e., it is present at a similar fraction of the entire mRNA pool as at other stages) will necessarily be lower than at other stages when measured by RT-qPCR since it is a fraction of a smaller total pool of transcripts. Thus, to identify transcripts that would exhibit constant expression by RT-qPCR requires knowledge of the relationship between measurements using RNAseq and those using RT-qPCR in oocytes and preimplantation embryos.

A second limitation on using RNAseq data to identify reference genes is that there is considerable variability between independent determinations of oocyte and preimplantation embryo transcriptomes. This could be addressed by obtaining a sufficiently large number of independently obtained transcriptomes covering all stages, although this would likely be prohibitively costly. Fortunately, a substantial number of RNAseq data sets have now been publicly deposited. While these are not directly comparable in their deposited forms, since they are commonly expressed in different units and were mapped to different builds of the mouse genome, the raw data are available which can be remapped and expressed in the same units.

Our strategy to identify a set of constantly expressed genes therefore relied on standardizing multiple available RNAseq data sets to the same genome build and normalizing to the same units for expression levels. These data sets could then be merged into one large set with multiple replicates at each stage and used to find transcripts with the least variation across stages as determined by RNAseq. A set of such genes could then be measured by RT-qPCR to reveal the relationship between expression measured by RNAseq and that measured by RT-qPCR. This relationship would then predict genes in the merged RNAseq data set that should yield the closest approximation to constant expression when measured by RT-qPCR.

Materials and methods

Mouse oocyte and preimplantation embryo RNAseq data sets

The Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo) was searched manually (in 2020) for recent RNAseq data sets containing mouse GV oocytes, MII eggs, or preimplantation embryos at the 1-cell, 2-cell, 4-cell, 8-cell, morula, or blastocyst stages. GEO data series are identified here by their data series accession numbers (GSE#), whereas individual samples within GEO series are identified by their sample accession numbers (GSM#). Where possible, preference was given to data sets spanning multiple developmental stages. For each sample, unprocessed reads were downloaded from the Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) and processed by the Ottawa Bioinformatics Core Facility (https://www.ohri.ca//bioinformatics). In cases where there were multiple SRA files from a single source, these were merged. The GEO series and samples used and their characteristics are listed in Table 1 (see Results). Reads were mapped to the GENCODE mouse vM25 annotation release of the mouse genome using salmon (https://combine-lab.github.io/salmon). Detailed descriptions of the processing of the data sets are provided in the Results section. We set a threshold of 5 million mapped reads below which individual samples would be rejected. PubMed identifiers (PMID#) were obtained for each data series from GEO and were used to access the associated published studies to confirm the stages and treatments of oocytes, eggs, or embryos. Principal component analysis to assess whether oocytes and embryos from different RNAseq data sets clustered together at each stage was carried out with DESeq2 (https://bioconductor.org/packages/release/bioc/html/DESeq2.html) using the default setting that plots the 500 most variable transcripts.

Oocyte and embryo collection

All animal protocols were approved by the University of Ottawa Animal Care Committee and comply with Canadian Council on Animal Care regulations. Mice were maintained on a 12-h light:dark cycle and had unrestricted access to water and Teklad Global 18% protein rodent diet 2018 (Envigo, Indianapolis, IN). Oocytes and embryos were obtained from 5- to 8-week-old female CD1 mice (Charles River Canada, St. Constant, QC, Canada) essentially as previously described [25]. Females were superovulated by intraperitoneal injection of 5 IU equine chorionic gonadotropin (eCG; Prospec, Sturgeon County, AB, Canada). For mature eggs and embryos, females were injected with 5 IU human chorionic gonadotropin (hCG, Prospec) 47 h post-eCG. For embryos, females were mated overnight with BDF1 males (Charles River) after hCG injection.

GV oocytes were obtained 44–46 h post-eCG from ovaries that were minced in Hepes-KSOM to release cumulus–oocyte complexes (COCs). Cumulus cells were removed mechanically by repeated pipetting. MII eggs were obtained 14–16 h post-hCG by flushing oviducts with Hepes-KSOM medium [26] using a blunt-end syringe. Eggs were exposed for 1–2 min to 300 μg/mL hyaluronidase in Hepes-KSOM to remove the expanded cumulus matrix. Embryos were obtained from mated females by flushing oviducts (1-cell, 2-cell, 4-cell, 8-cell) or oviducts with the uterus attached (morulae, blastocysts) at the following times post-hCG: 21–24 h (1-cell), 42–44 h (2-cell), 56 h (4-cell), 66–68 h (8-cell), 76 h (morulae), and 93–94 h (blastocysts). Pools of 34 oocytes or embryos were flash-frozen on dry ice in <5 μL Hepes-KSOM immediately after collection and were stored at -80°C until RNA isolation.

Quantitative reverse transcription polymerase chain reaction

Total RNA was isolated from COCs, oocytes, and embryos using the Arcturus PicoPure RNA isolation system, which is optimized for small samples (Applied Biosystems, Foster City, CA, catalog #KIT0204). Each sample had a spike-in of exogenous cRNA of *Xenopus laevis* elongation factor 1α gene (*eef1a1*) added that was used in RT-qPCR to control for variations introduced by subsequent handling and reverse transcription. The exogenous spike-in cRNA was produced from 1 μg of linearized *pTRI-Xef* TRIPLEscript plasmid by in vitro transcription using the mMessage mMachine T7 Transcription Kit that includes the plasmid (Thermo Fisher

Table 1. RNAseq data sets and characteristics.

GEO accession	GV	MII	1-cell	2-cell	4-cell	8-cell	Morula	Blastocyst	Superov.	In vivo/in vitro	Mouse strains	cDNA	Ref.
GSE118562	GSM3333253 GSM3333254	GSM3333257 GSM3333258	GSM3333259 GSM3333260						Yes	GV,1c in vivo; MII in vitro	C57Bl/6 (F)	polyA[a]	[27]
GSE126687	GSM3611277 GSM3611279 GSM3611280		GSM3611291 GSM3611293 GSM3611294	GSM3611305 GSM3611306 GSM3611307				GSM3611320 GSM3611321 GSM3611322	No	In vivo	C57Bl/6 (F)	polyA[a]	[28]
GSE138760		GSM4118733 GSM4118734	GSM4118735 GSM4118736	GSM4118737 GSM4118738	GSM4118739 GSM4118740	GSM4118741 GSM4118742		GSM4118743 GSM4118744	Yes	MII in vivo; embryos in vitro from 1c	C57Bl/6 (F) DBA/2 (M)	polyA[a]	[29]
GSE44183		GSM1080195 GSM1080196	GSM1080197 GSM1080198 GSM1080199	GSM1080200 GSM1080201 GSM1080202	GSM1080203 GSM1080204 GSM1080205	GSM1080206 GSM1080207 GSM1080208	GSM1080209 GSM1080210 GSM1080211		Yes	In vivo	C57Bl/6 (F)	polyA[b]	[30]
GSE68150	GSM1664630 GSM1664631 GSM1664632								No	In vivo	C57BlxCBA (F)	polyA[a]	[31]
GSE70605		GSM1811753 GSM1811754 GSM1811755	GSM1811765 GSM1811766 GSM1811767		GSM1811727 GSM1811728	GSM1811732 GSM1811733	GSM1811747 GSM1811748 GSM1811749	GSM1811738 GSM1811739 GSM1811740	Yes	MII, 1c, 2c in vivo; others in vitro from 2c	B6D2F1 (F)	polyA[a]	[32]
GSE71442					GSM1714227	GSM1714236	GSM1714280 GSM1714281		No	In vivo	C57Bl/6 (F)	polyA[a]	[33]
GSE73803	GSM1903780 GSM1903781 GSM1903782								Yes	In vivo	Mixed C57Bl/6129 (F)	Mixed[c]	[34]
GSE86470	GSM2303784 GSM2303785 GSM2303786	GSM2303787 GSM2303788 GSM2303789	GSM2303790 GSM2303791 GSM2303792						Yes	In vivo	CBA/CA-JxC57Bl6/J (F)	Mixed[c]	[35]
GSE98150		GSM2588668 GSM2588669		GSM2588670 GSM2588671	GSM2588674 GSM2588675 GSM2588676	GSM2588678 GSM2588679	GSM2588681 GSM2588682		Yes	In vivo	B6D2F1 or C57Bl/6 (F) B6D2F1 or DBA2 (M)	Mixed[c]	[36]

[a]Total RNA isolation, first-strand cDNA synthesized with oligo-dT. [b]PolyA RNA isolation, first-strand cDNA synthesized with oligo-dT. [c]Total RNA isolation, first-strand cDNA synthesized with mixed oligo-dT and internal primers.

Scientific, Waltham, MA, catalog #AM1344). The transcription reaction was incubated for 2 h, and unincorporated nucleotides and proteins were removed by lithium chloride precipitation. Messenger RNA was dried and resolubilized in nuclease-free water, to yield 26.4 μg. COC, oocyte, and embryo samples were spiked with 264 fg *pTRI-Xef* cRNA at the initial step of RNA extraction. On-column DNase digestion using the RNase-Free DNase Set (Qiagen, Toronto, ON, catalog #79254) was then performed.

The oocyte, egg, or embryo mRNA samples including the *pTRI-Xef* spike-in were reverse transcribed with the Superscript IV Kit (Thermo Fisher Scientific, catalog #18091050), using random hexamer primers. Complementary DNA from oocytes or embryos was diluted so that an equivalent of 0.36 oocytes/eggs/embryos was used as a template for subsequent RT-qPCR reactions. To confirm that quantification of gene expression was within the linear range for the RT-qPCR quantification, the amount of oocyte/embryo equivalents to use per reaction had been determined using serial dilutions of pooled oocyte/embryo cDNA and linear regression (not shown). For the control experiment designed to detect *Slc7a6* as an indicator of any cumulus cell contamination of oocyte or egg samples, an equivalent of 1.5 COCs/oocytes/embryos was used as the template.

Since transcripts were selected from RNAseq data sets that had been mapped to the genome, it was not possible to determine whether multiple transcript variants had contributed to mapped reads. Therefore, primers were designed to capture common regions of the greatest possible number of predicted or verified transcript variants for each gene. All transcripts for each gene of interest were downloaded using transcript tables from the *Mus musculus* NLM Gene database (https://www.ncbi.nlm.nih.gov/gene). Primers were designed using the "Primers common for a group of sequences" tool within NCBI Primer BLAST (https://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi?GROUP_TARGET=on).

Our strategy was to design a set of nested primers for each gene (Supplemental Table S1). The outer primer pairs were used solely for conventional RT-PCR to amplify cDNA that was quantified to be used as starting material for constructing calibration curves. The inner primer pairs were nested within the outer primers. These inner primers were used for amplifying from the outer primer products for calibration and for RT-qPCR to quantify transcript expression in oocytes, eggs, and preimplantation embryos. Primers for *pTRI-Xef* were designed from *X. laevis eef1a1*.

To produce the PCR products for constructing the calibration curves for RT-qPCR, conventional RT-PCR was carried out using the outer primers on samples of COC cDNA. In the cases of *Taf1d*, *Hmgb3*, and *Rb1cc1*, faint non-specific products were observed in COC cDNA. RT-PCR was repeated using a template of pooled oocyte/embryo cDNA, which resulted in single PCR products at the expected sizes. RT-PCR was carried out on a T100 Thermocycler (Bio-Rad, Mississauga, ON, Canada) with an initial denaturation at 95°C (3 min), 45 cycles of 95°C (30 s), 60°C (30 s), 72°C (1 min), then 5 min at 72°C. PCR products were visualized by electrophoresis on a 1.75% agarose gel with ethidium bromide and single bands at the predicted sizes were confirmed. The amplified DNA was then recovered using the QIAquick Gel Extraction kit (Qiagen, Toronto, ON, catalog #28704) and the identity of the PCR products was confirmed by DNA Sequencing (StemCore Laboratories, Ottawa, ON). All amplicons had sequences of common regions entirely shared between alternate transcripts except *Ubl5*, where it was not possible to include all variants. However, the *Ubl5* amplicons from oocytes and embryos were exclusively transcript variant 4 (NR_153857.1; not shown) which is apparently the only variant expressed in mouse oocytes, eggs, or embryos. The purified PCR products derived from the outer primer sets were quantified with the NanoDrop One Microvolume UV–Vis Spectrophotometer (Fisher Scientific Company, Ottawa, ON) and used to prepare standard curves for the mouse transcripts and for *pTRI-Xef* in subsequent RT-qPCR experiments.

RT-qPCR was performed on a 7500 FAST Real-Time system (Applied Biosystems, Foster City, CA). Reactions were carried out with PowerUp SYBRGreen Master Mix (Applied Biosystems, Foster City, CA, catalog #A25742) in MicroAmp Fast optical 96-well reaction plates covered with Adhesive Film (Applied Biosystems, catalog #4346907 and #4311971). All RT-qPCR was done using inner primer sequences including for *pTRI-Xef* (Supplemental Table S1A–C). Only inner primers were used for *Slc7a6* (Supplemental Table S1D) employing a previously validated primer pair [27].

Samples were loaded in duplicate wells and these technical replicates were averaged for each independent repeat. The RT-qPCR program used for all primer sets was 50°C (3 min), 95°C (10 min), and 35 cycles of 95°C (15 s), 60°C (15 s), and 72°C (1 min). Following amplification, a melt curve from 60 to 95°C was performed to confirm the presence of a single PCR product. Where specified, the PCR products amplified from oocytes and embryos by qPCR were collected from the wells, pooled for each transcript, and visualized by 1.75% agarose gel electrophoresis with ethidium bromide. The amplified products were then recovered, and the identity of the PCR products confirmed by DNA sequencing as above. Messenger RNA abundance in oocytes, eggs, and embryos was calculated relative to a standard curve produced from known cDNA concentrations (1:10 serial dilution from 1000 to 0.01 fg) that had been obtained from mouse COCs using the outer primers for each primer set as described above. The standard curve was run adjacent to oocyte, egg, and embryo samples for each independent repeat and was also used to calculate the amplification efficiency of each primer set (acceptable range within 90–110%). A total of $N = 3$ independent biological repeats were performed for each transcript at each stage.

For each transcript measured by qPCR, the cycle threshold (C_T) was recorded. Calibration curves were constructed using known amounts of starting templates made using the outer primer sets as described above. The calibrations were done in the same qPCR run as the samples and expressed as C_T versus amount of template. The calibrations for each transcript were used to convert C_T to mass of transcript in each sample. The *pTRI-Xef* spike-in was similarly quantified for each sample. Transcript abundance was then expressed as the ratio of each transcript to *pTRI-Xef* in the same sample and is reported here as a dimensionless number. Transcript abundance is calculated based on the same number of oocytes, eggs, or embryos per sample.

Measures of variation across oocyte and embryo stages

The coefficient of variation (CV) was used as a measure of the inter-stage variation of the expression of each transcript

across the eight stages assessed: GV oocytes, MII eggs, or preimplantation embryos at the 1-cell, 2-cell, 4-cell, 8-cell, morula, and blastocyst stages. For each transcript, the mean expression at each of the stages was first calculated as

$$\mu_{s,t} = \frac{1}{n}\sum_{i=1}^{n} E_{i,s,t}$$

where $\mu_{s,t}$ is the mean of the expression at stage s for transcript t, $E_{i,s,t}$ is the expression value for repeat i at stage s for transcript t (i.e., the expression by RNAseq or RT-qPCR), and n is the total number of repeats at stage s. The overall mean among all stages for each transcript, μ_t, was then calculated as

$$\mu_t = \frac{1}{N}\sum_{s=\mathrm{GV}}^{\mathrm{Bl}} \mu_{s,t}$$

and the standard deviation for that transcript was calculated as

$$\sigma_t = \left(\sum_{s=\mathrm{GV}}^{\mathrm{Bl}} \frac{(\mu_{s,t} - \mu_t)^2}{(N-1)}\right)^{\frac{1}{2}}$$

where $N = 8$ stages (i.e., GV oocyte to blastocyst, Bl). The inter-stage CV for transcript t was then

$$\mathrm{CV}_t = \frac{\sigma_t}{\mu_t}$$

Where specified, the intra-stage CVs ($\mathrm{CV}_{s,t}$) were similarly calculated using the mean ($\mu_{s,t}$) of all repeats within a given stage s for transcript t and the corresponding standard deviation.

The maximum and minimum expression levels were also used to characterize the variation of expression of a given transcript between stages. For this, the maximum and minimum expression levels among stages ($\mu_{t,\max}$ and $\mu_{t,\min}$) were recorded for each transcript. The ratio $\mathrm{MAX}_t/\mathrm{MIN}_t = \mu_{t,\max}/\mu_{t,\min}$ was then used to characterize the range of expression values for each transcript.

Euclidean vector distance was used as a measure of deviation from a specified expression pattern. For each transcript, the means at each stage ($\mu_{s,t}$) were expressed as a vector of the form

$$\mathrm{v}_t = \left[\mu_{\mathrm{GV},t}, \mu_{\mathrm{MII},t}, \mu_{1\mathrm{c},t}, \mu_{2\mathrm{c},t}, \mu_{4\mathrm{c},t}, \mu_{8\mathrm{c},t}, \mu_{\mathrm{M},t}, \mu_{\mathrm{Bl},t}\right]$$

Means for each transcript were normalized to expression at the GV stage for the same transcript (i.e., $\mu_{\mathrm{GV},t} = 1$). The idealized expression pattern was similarly expressed in vector form ($\mathrm{v}_{\mathrm{ideal}}$). The Euclidean vector distance (EVD_t) between the expression of each transcript and the idealized pattern was then calculated as the square root of the sum of the squares of the differences between the corresponding elements of the vectors:

$$\mathrm{EVD}_t = \sqrt{\sum_{\mathrm{GV}}^{\mathrm{Bl}} \left[\mu_{s,t} - \mathrm{v}_{s,\mathrm{ideal}}\right]^2}$$

where $\mathrm{v}_{s,\mathrm{ideal}}$ is the element in the vector $\mathrm{v}_{\mathrm{ideal}}$ describing the idealized value for stage s.

Data analysis

Data were graphed and analyzed with Prism 9.5 or 10.0 (GraphPad Software, San Diego, CA). Except where otherwise specified, individual values and the means ± SEM are graphed. Statistical significance of the differences among more than two means was tested using Welch ANOVA with Dunnett multiple comparisons test when standard deviations were significantly different or ordinary one-way ANOVA with Tukey multiple comparisons test if the standard deviations were not significantly different. To test for a significant difference between two means, *t*-tests were used. *P*-values <0.05 were considered significant. Calculations and data sorting were carried out using Excel for Microsoft Office 365. In Excel, means and standard deviations were obtained with the AVERAGE() and STDEV() functions, respectively, and geometric means were calculated using GEOMEAN().

Results

Mouse oocyte and preimplantation embryo RNAseq data sets

To assemble a set of transcriptome data covering mouse developmental stages from GV oocytes through blastocysts, we downloaded RNAseq data series from GEO. Initially, 11 separate series of data that contained 115 individual sets of transcriptome data for mouse oocyte or preimplantation embryo samples were identified, all from published studies [27–37]. These were mapped to 54 347 genes in the mouse genome (GENCODE vM25) and the number of mapped reads was calculated for each sample (Supplemental File S1). Of the oocyte and embryo samples that were mapped, 22 (19%) were found to have fewer than the 5 million mapped reads and thus were rejected, leaving 93 transcriptomes (Table 1) covering GV oocytes ($N = 14$), MII eggs ($N = 14$), 1-cell embryos ($N = 16$), 2-cell embryos ($N = 10$), 4-cell embryos ($N = 11$), 8-cell embryos ($N = 10$), morulae ($N = 10$), and blastocysts ($N = 8$). One data series (GSE111039 [37]) included only samples with <5 million reads each and this entire series was omitted, whereas two others (GSE70605 and GSE71442) had subsets of samples removed by this criterion (Supplemental File S1). The final transcriptome data then comprised 3–6 independent series for each stage from 10 separate data series in total (Table 1). The full standardized and merged set of transcript expression data mapped to 54 347 genes for all RNAseq data sets at each of the stages of oocytes, eggs, or embryos is provided as a supplementary file (Supplemental File S2).

Principal component analysis (Supplemental Figure S1) showed that the transcriptomes at the same stages from different series clustered together and exhibited the expected pattern [6]. The samples that fell below the 5-million-read threshold and were subsequently removed generally exhibited poorer clustering, further supporting their exclusion.

Identifying transcripts with approximately constant expression across oocyte and preimplantation embryo stages in the RNAseq data sets

The intra-stage $\mathrm{CV}_{s,t}$ within each stage of oocyte, egg, or embryo was first calculated for each transcript (54 347 genes)

to provide a measure of the variability between the different samples within stages, and the $CV_{s,t}$ versus the mean expression level within each stage ($\mu_{s,t}$) was plotted for every transcript. To allow visualization of the $CV_{s,t}$ across expression levels that varied by orders of magnitude between genes, we transformed the mean expression level for each transcript at each stage ($\mu_{s,t}$) to $\log_{10}(\mu_{s,t}+0.01)$ for the purpose of graphing only (Figure 1A). Variability between replicate samples in RNAseq is expected to be higher at lower expression levels [38], which was clearly evident here.

All further analysis was then performed using the mean expression levels at each stage (μ_t) for a given transcript with the intent of identifying transcripts in the RNAseq data set that varied minimally across stages from GV oocyte to blastocyst. To investigate the inter-stage variability, the CV (CV_t) of the means at each stage ($\mu_{s,t}$) was calculated and plotted for each transcript (Figure 1B) to show the variation across stages as a function of the mean expression among the eight stages (μ_t).

We chose a threshold of $\mu_{s,t} = 20$ TPM ($\log_{10}$(mean expression level + 0.01) ≈ 1.3) above which the distribution of intra-stage $CV_{s,t}$ values (Figure 1A) appeared low. We thus carried out the subsequent analyses below using the subset of transcripts that had a mean expression ($\mu_{s,t}$) of at least 20 TPM at one or more stages of oocyte or embryo. There were 9404 transcripts that met this threshold representing 17% of the total mapped transcripts (Figure 1C). To then identify transcripts likely to have the most constant expression across stages, those whose CV_t among stages was ≤0.33 (290 transcripts, 3% of mapped genes) were selected for further analysis (Figure 1C).

It is, however, possible for a transcript to have a low CV_t but not have nearly constant expression across all stages, for example, if the expression at only one stage was much higher or lower than the others that were themselves nearly equal. To select against this, we calculated the ratio of the maximum to minimum expression across the stages (MAX_t/MIN_t) for each transcript. All transcripts with $MAX_t/MIN_t \leq 2.00$ were selected (39 transcripts) along with an additional four with $MIN_t \geq 150$ TPM and MAX_t/MIN_t between 2.00 and 2.13 to provide a greater representation at higher expression levels (43 transcripts in total, Supplemental File S3). These were plotted to assess their expression across stages (Supplemental Figure S2). From these, eight that had relatively high expression levels and low intra-stage variation were selected for RT-qPCR analysis: *Anp32b*, *Dppa3*, *Exosc8*, *Hspa9*, *Mcm7*, *Ociad1*, *Snrpg*, and *Ubl5* (Figure 2A–H). Three transcripts, *Actb*, *Gapdh*, and *Ppia*, that have been previously used as reference genes in preimplantation embryos are also plotted for comparison (Figure 2I–K). For six of the chosen transcripts (Figure 2A, B, D, E, G, H), the mean expression was not significantly different between any stages (Welch ANOVA with Dunnett multiple comparisons test), whereas the remaining two were each significantly different between two stages: *Exosc8* (Figure 2C) for 1-cell versus morula ($P = 0.04$) and *Ociad1* (Figure 2F) for 1-cell versus 2-cell ($P = 0.04$). In contrast, the variation between the means of the example reference genes (Figure 2I–K) was highly significantly different ($P < 0.0001$ by Welch ANOVA).

RT-qPCR of transcripts with approximately constant expression in RNAseq data sets

Before performing RT-qPCR for the eight candidate genes, we first confirmed that the oocyte samples were not detectably contaminated with cumulus cells by measuring expression of *Slc7a6*, which is highly expressed in cumulus cells but not in oocytes or eggs [39]. *Slc7a6* was quantified by RT-qPCR in intact COCs, GV oocytes, MII eggs, and 2-cell embryos using a higher concentration of template (1.5 COCs/oocytes/embryos per reaction) and was confirmed to be present in COCs but not in GV oocytes, MII eggs, or 2-cell embryos (Supplemental Figure S3) indicating negligible cumulus cell contamination in oocytes or eggs.

Three independent sets of GV oocytes, MII eggs, 1-cell, 2-cell, 4-cell, 8-cell, morula, and blastocyst stage embryos were collected with pTriXEF RNA added as an internal control. RT-qPCR was then carried out and the results were normalized to pTriXEF in each of the samples (Figure 3). The eight genes each showed single melt peaks by RT-qPCR. The qPCR products were then collected and the presence of a single product at the expected size confirmed on agarose gels (Supplemental Figure S4A). From the RT-qPCR data, it is evident that none of these transcripts exhibited constant expression across stages, but instead decreased to a minimum around the 2–8-cell stages and then increased again (Figure 3A–H), in contrast to their nearly constant expression when measured by RNAseq (Figure 2A–H). The variation between means was significantly different for all eight transcripts (Figure 3A–H; $P < 0.0001$ except *Snrpg*, $P = 0.0003$). This pattern is consistent with total transcripts in mouse embryos generally reaching a minimum at the 2-cell stage and then increasing again as the embryonic genome becomes expressed, as discussed above.

The expression patterns measured by RT-qPCR for the eight transcripts (Figure 3A–H), although similar, were not identical, counter to what would be expected if their RNAseq expression patterns were absolutely constant across stages. We hypothesized that at least a part of these discrepancies could be because of the apparent differences in the RNAseq expression patterns between transcripts, for example, where *Hspa9* trended slightly upward across stages, whereas *Ociad1* has a slight downward trend (Figure 2A–H). To facilitate direct comparisons, we expressed the RT-qPCR data for each transcript relative to its mean at the GV stage (set to 1.0; Supplemental File S4) which revealed differences in expression patterns between the transcripts (Figure 3I). We then normalized the RT-qPCR data to account for non-constant expression in the RNAseq data by dividing the mean expression at each stage determined by RT-qPCR by the mean expression determined by RNAseq (Supplemental File S4), which made the curves more similar and decreased the standard errors of the means at all stages except MII (Figure 3J). This provided a relatively consistent quantitative relationship between RNAseq and RT-qPCR data which indicated that a transcript that had constant expression across stages as measured by RNAseq would have mean expression levels measured by RT-qPCR at each stage as shown in Figure 3J.

Identifying transcripts predicted to have approximately constant expression across oocyte and preimplantation embryo stages when measured by RT-qPCR

The conversion factors between measurements of transcript expression by RNAseq and RT-qPCR derived as described above and shown in Figure 3J were designated here by z_s, where s indicates the oocyte or embryo stage. If P_s is the expression of a transcript measured by RT-qPCR and R_s is its expression measured by RNAseq (both normalized to the

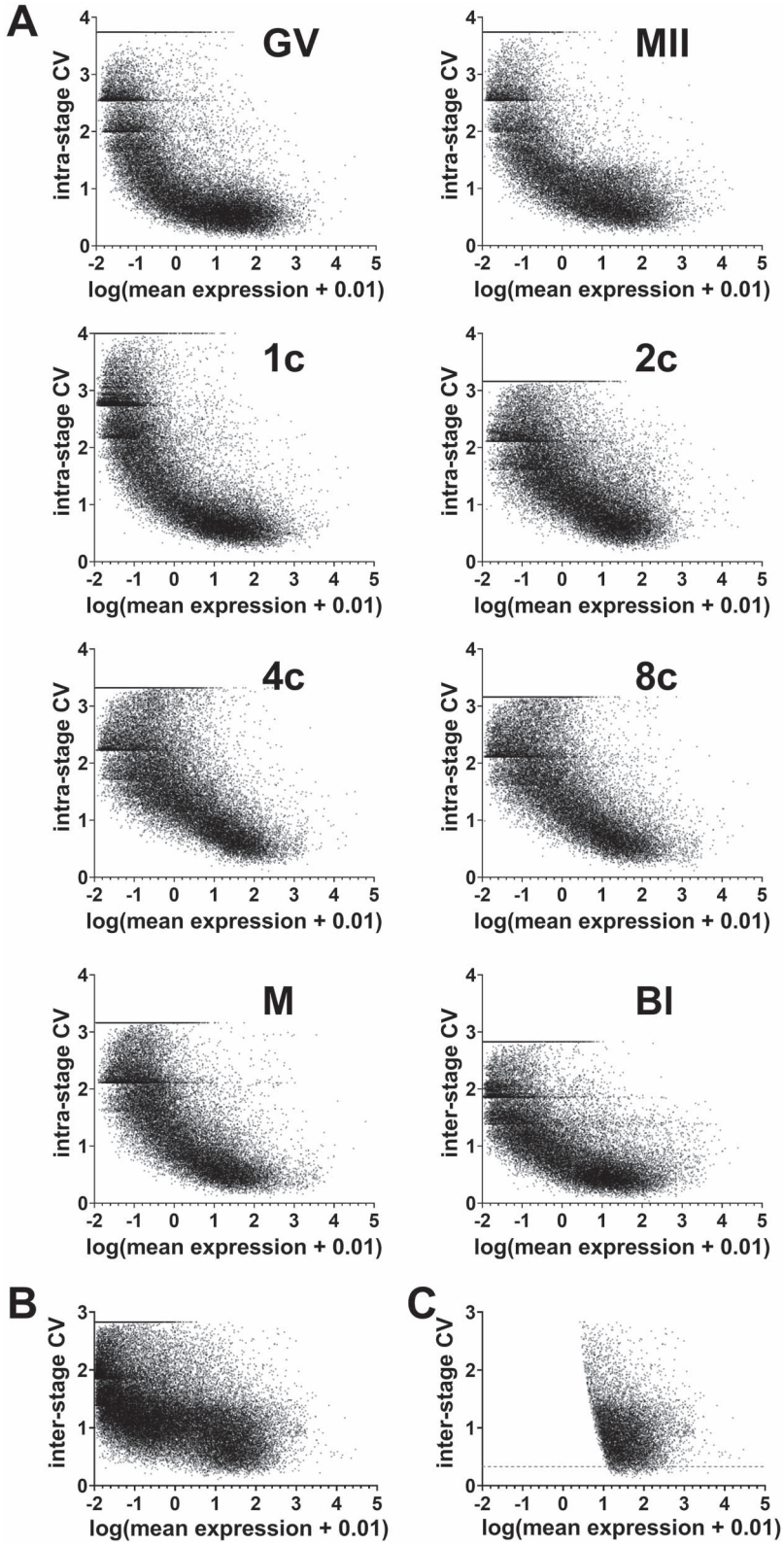


Figure 1. Coefficients of variation as a function of expression levels. Coefficients of variation were calculated and expression in TPM was converted to log [mean expression $+0.01] = \log_{10}(\mu_{s,t} + 0.01)$ as described in the text for graphing. (**A**) The intra-stage CV ($CV_{s,t}$) within each stage of oocyte, egg, or embryo is shown vs. the expression levels for all 54 347 mapped transcripts at each stage. Stages are indicated as GV oocyte (GV), mature MII egg (MII), 1-cell (1c), 2-cell (2c), 4-cell (4c), and 8-cell (8c) embryos, morulae (M), and blastocyst (Bl). (**B**) The inter-stage CV (CV_t) is shown plotted against the mean expression across all stages (μ_t) as defined in the text. Expression is transformed as described in (A) for graphing. (**C**) Transcripts where all stages exhibited expression of <20 TPM were eliminated and the resulting data plotted as in (B). The dotted horizontal line indicates a $CV_t = 0.33$. Transcripts where expression was 0 for all stages are not plotted. The quantization that appears in the data at higher CV values is because of the effect of having zeroes in the values used to calculate CV for transcripts with high values of CV. It can be shown that, if the total number of values used to calculate CV is N and of those, the number that are zero (i.e., expression = 0) is Z, then the minimum value that CV can have is given by $\left(\frac{NZ}{(N-1)(N-Z)}\right)^{\frac{1}{2}}$, which results in there being discreet minimum values that CV can have when $Z \neq 0$ and which become further apart as $Z \rightarrow N$. This did not affect the selection of transcripts here since transcripts with higher expression values and low CV were chosen, as indicated in panel (C) below the dotted line.

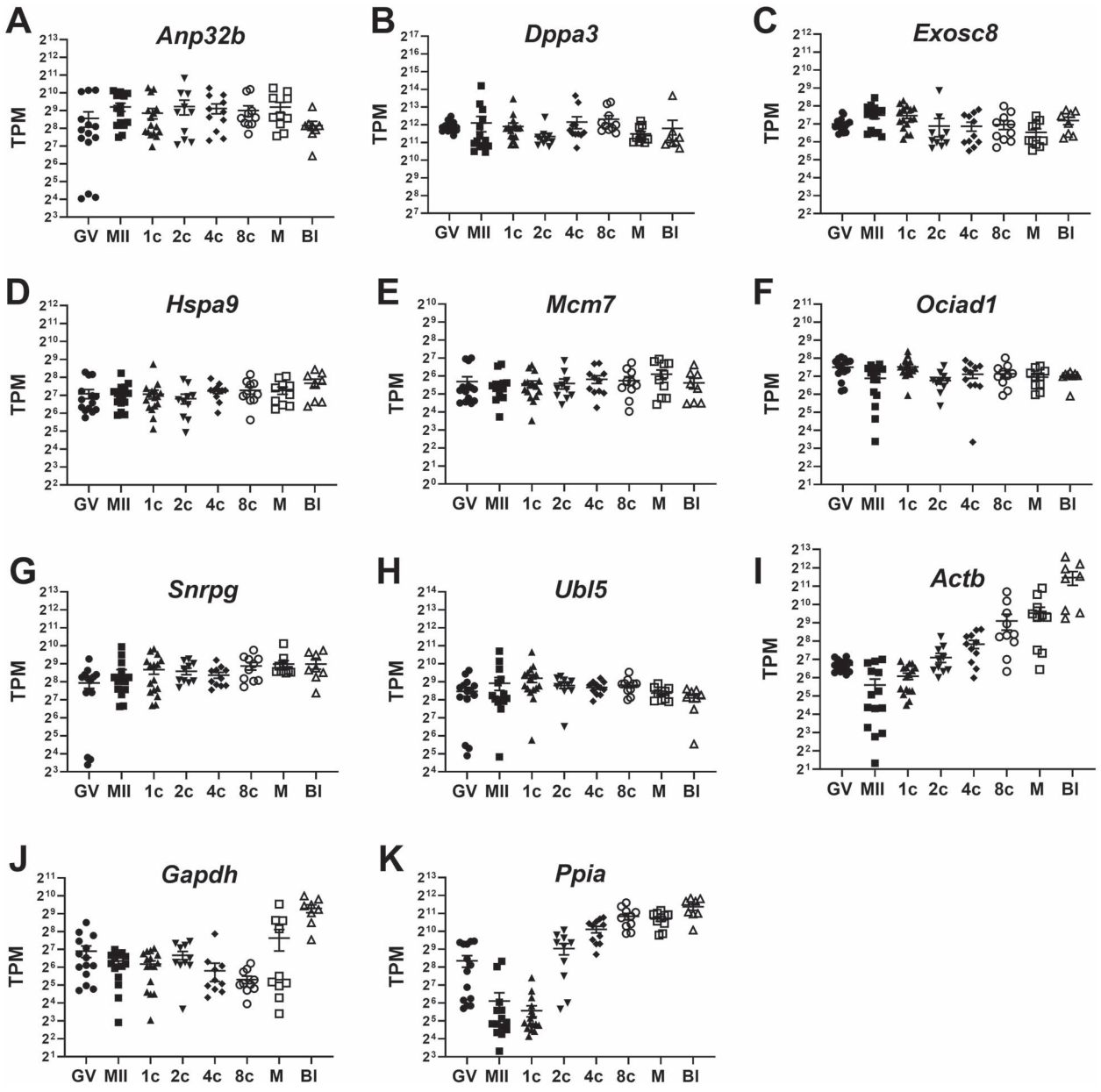


Figure 2. Transcripts with nearly constant expression across stages by RNAseq. (**A**–**H**) The eight transcripts chosen for their low CV and MAX/MIN ratios in the RNAseq data sets have nearly constant expression across stages. (**I**–**K**) Three examples of transcripts previously used as reference genes exhibited greater variation across stages. Expression of each transcript is plotted in TPM for each of the RNAseq data sets at each stage on a $\log_2$ scale. Stages are indicated at the bottom of each graph as GV oocyte (GV), mature MII egg (MII), 1-cell (1c), 2-cell (2c), 4-cell (4c), and 8-cell (8c) embryos, morulae (M), and blastocyst (Bl). Means ± SEM are indicated by a horizontal line with error bars. Statistical analysis is described in the text.

GV stage arbitrarily set to 1), then the relationship between the two methods of measurement is given by $P_s = z_s R_s$. If a hypothetical transcript has constant expression across all stages when measured by RT-qPCR, then $P_s = 1$ for all stages. Therefore, the expression of the same transcript when measured by RNAseq would be $R_s = 1/z_s$. Thus, a transcript predicted to have constant expression when measured by RT-qPCR would have expression levels measured by RNAseq of $R_{GV} = 1.00$, $R_{MII} = 1.59$, $R_{1c} = 2.12$, $R_{2c} = 8.16$, $R_{4c} = 7.65$, $R_{8c} = 6.44$, $R_M = 2.72$, and $R_{Bl} = 1.33$. We therefore sought transcripts that most closely approximated this idealized expression pattern.

To identify transcripts that most closely approximated this pattern of expression, we used the mean expression in the RNAseq data set at each stage normalized to the GV stage for every transcript. Each transcript was then represented by a vector, v_t, with eight elements corresponding to the stages from GV to blastocyst. The idealized expression pattern was defined by the vector $v_{ideal} = [1.00, 1.59, 2.12, 8.16, 7.65, 6.44, 2.72, 1.33]$. The Euclidean vector distance (EVD_t) between v_t and v_{ideal} was calculated for each transcript. These were sorted by Euclidean vector distance to identify those transcripts with the smallest distances from the idealized pattern. The 500 transcripts with the smallest EVD_t from the idealized pattern are shown in Supplemental File S5.

As a second method of selection, we used the idealized expression pattern (Figure 3J) to transform the RNAseq data

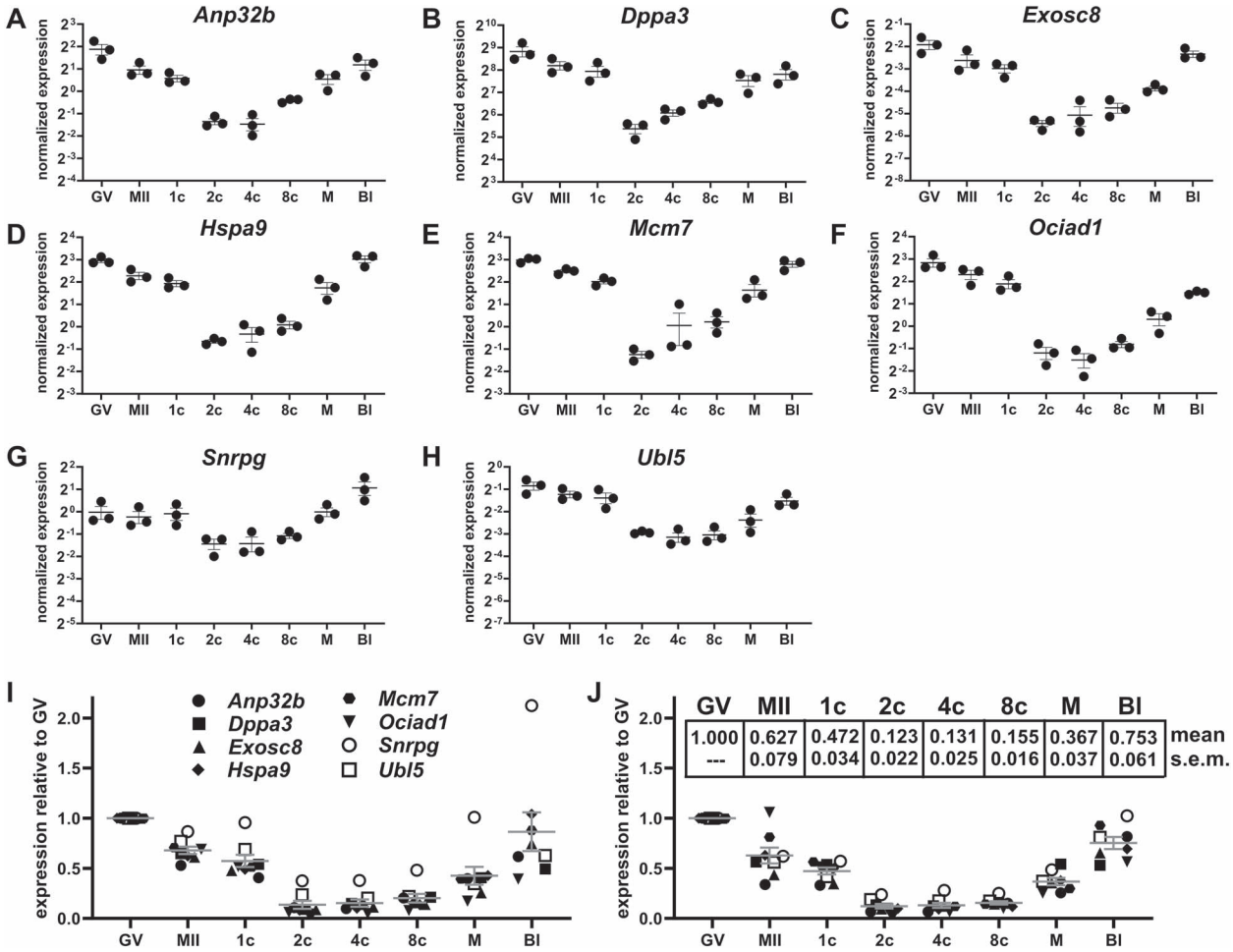


Figure 3. RT-qPCR measurement of expression of transcripts with nearly constant expression by RNAseq. Transcripts that were identified as having nearly constant expression across stages by RNAseq (Figure 2A–H) were measured by RT-qPCR ($N = 3$ independent repeats) as described in the text. (**A**–**H**) The patterns of expression were similar for each of the eight transcripts, with expression reaching a minimum within the 2-cell to 8-cell stages. Expression of each transcript is plotted as the ratio of its expression to that of *pTRI-Xef* in the same sample on a $\log_2$ scale. Means ± SEM are indicated by a horizontal line with error bars. Statistical analysis is described in the text. (**I**) The expression of each of the eight transcripts was normalized to its expression at the GV stage (set to 1) to allow direct comparison of their relative expression patterns. (**J**) The data shown in (I) were corrected for the deviations from constant expression in the corresponding RNAseq expression (Figure 2A–H) as described in the text. This generally decreased the variability within each stage (except MII). The inset table shows the mean (±SEM) expression at each stage relative to the GV stage, which equal z_s for each stage (see text). Stages are indicated as GV oocyte (GV), mature MII egg (MII), 1-cell (1c), 2-cell (2c), 4-cell (4c), and 8-cell (8c) embryos, morulae (M), and blastocyst (Bl).

by multiplying the expression in TPM by the factor (z_s) for each stage for each of the 500 transcripts identified above as having the least Euclidean vector distances from the idealized expression pattern (Supplemental File S6). This should result in constant values across stages for any transcript with the desired expression pattern of R_s described above (Figure 3J) since those would have relative expression proportional to $1/z_s$ at each stage. We then calculated the CV for the expression across stages for each of the 500 transcripts (Supplemental File S6). To select transcripts likely to have sufficient expression for RT-qPCR, we restricted the candidates to those with expression levels of ≥50 TPM and then eliminated mitochondrial genes and transcripts without defined MGI gene symbols. Finally, we selected those with CV ≤ 0.55. From those, we retained the top 10 as ranked by smallest Euclidean vector distances (Supplemental File S7), which were *Polr2f*, *Taf1d*, *Crb3*, *Hmgb3*, *Wdr43*, *Rb1cc1*, *Ccna2*, *Nsmce4a*, *Cd320*, and *Mpc1*. Their expression by RNAseq compared to the desired expression pattern of R_s proportional to $1/z_s$ is shown in Figure 4. All of these transcripts were highly significantly different across stages by Welch ANOVA ($P < 0.0001$ for all except *Hmgb3*, $P = 0.0003$). From this set of transcripts, we chose *Polr2f*, *Taf1d*, *Hmgb3*, *Wdr43*, *Rb1cc1*, *Ccna2*, and *Cd320* (Figure 4A, B, D, E, F, G, I) for testing by RT-qPCR, rejecting *Crb3*, *Nsmce4a*, and *Mpc1* (Figure 4C, H, J) based on their excessive intra-stage variation.

It was possible that the transcripts differed between RNAseq data sets where the reverse transcription step had used polyA primers versus mixed primers since both types were included in the data sets used here (Table 1). To examine this, we plotted the data for the seven selected transcripts at the GV and MII stages, where global polyadenylation status differs substantially, separately for the RNASeq data sets that used polyA versus mixed primers. The relationship between expression of each transcript at the GV and MII stages was similar for both types of primer (Supplemental Figure S5). Thus, there did not appear to be a substantial bias because of

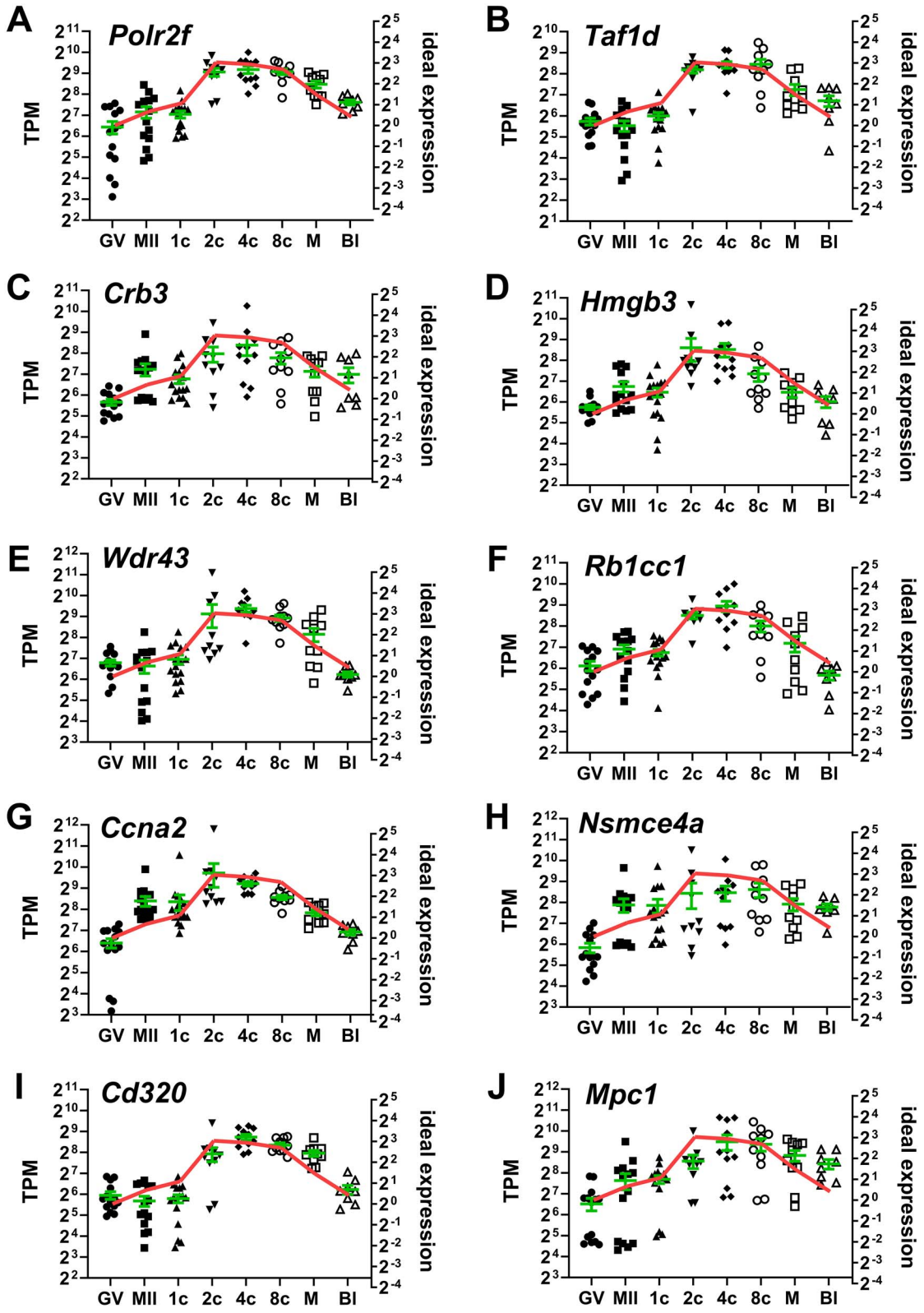


Figure 4. Transcripts predicted to have constant expression by RT-qPCR. Transcripts were identified by analysis using Euclidean vector distance analysis from the predicted pattern described by v_{ideal} and by having low CV values after transformation as described in the text and Supplemental Files S5, S6, and S7. Expression of each from the RNAseq data set is plotted here in TPM on a $\log_2$ scale (left axis). For each transcript, the vector elements ($1/z_s$) from v_{ideal} are plotted as the line, also on a $\log_2$ scale (right axis). The extent of the left axes and vertical placement of the right axes have been adjusted to facilitate visual comparison (not used for quantitative analysis). Of the 10 transcripts, seven (**A**, **B**, **D**, **E**, **F**, **G**, and **I**) were chosen for further analysis, whereas three, *Crb3* (**C**), *Nsmce4a* (**H**), and *Mpc1* (**J**), were rejected based on excessive intra-stage variation. The means ± SEM are shown as a horizontal line and error bars. Stages are indicated as GV oocyte (GV), mature MII egg (MII), 1-cell (1c), 2-cell (2c), 4-cell (4c), and 8-cell (8c) embryos, morulae (M), and blastocyst (Bl). Statistical analysis is described in the text.

polyadenylation status in these data that could have affected transcript selection.

RT-qPCR of candidate transcripts predicted to have approximately constant expression across oocyte and preimplantation embryo stages

The expression patterns of the seven selected transcripts were determined by RT-qPCR (Figure 5A–G) along with three transcripts commonly used as reference genes, *Actb*, *Gapdh*, and *Ppia* (Figure 5H–J). These seven genes chosen for RT-qPCR each showed single bands on conventional RT-PCR gels (Supplemental Figure S4B). For each of the seven, most stages fell within a factor of two of the overall means, with the only means falling outside this range being *Polr2f*, *Wdr43*, and *Cdc320* each at the 2-cell stage and *Taf1d* at the blastocyst stage. The means across stages for each transcript were, however, still significantly different ($P < 0.001$ by one-way ANOVA) for all except *Hmgb3* ($P = 0.12$). The means of the three commonly used reference genes (Figure 5H–J) were highly significantly different ($P \leq 0.0001$). Thus, we have identified a set of transcripts with nearly constant expression across mouse oocyte and preimplantation embryo stages.

To confirm the reproducibility of these expression patterns, we carried out three additional independent collections of GV oocytes, 2-cell embryos, and morulae and measured expression of the same seven transcripts by RT-qPCR. To allow a direct comparison between the two sets of collections, the expression levels for each transcript were normalized to the mean expression at the GV stage for that collection, arbitrarily set to 1.0 (Figure 6). None of the repeat measurements at the 2-cell and morula stages were significantly different ($P > 0.05$ by unpaired *t*-tests) from the previous measurements, except for the 2-cell stages of *Polr2f* and *Taf1d* (Figure 6A and B), which differed ~2-fold between the repeats (0.29 ± 0.06 vs. 0.59 ± 0.07, $P = 0.03$ for *Polr2f* and 0.63 ± 0.09 vs. 1.17 ± 0.11, $P = 0.02$, for *Taf1d*). Thus, the patterns of the seven transcripts we identified as having approximately constant expression across mouse oocyte and embryo stages were confirmed.

Ranking transcripts for least variation across stages for RT-qPCR and RNAseq

Finally, we ranked transcripts by their CV_t at each stage for the transcripts assessed by RT-qPCR in both the first and second sets. The means used for RT-qPCR were of the data in Figures 3 and 5, whereas for RNAseq the means are of the data in Figures 2 and 4. As expected, in both cases, the commonly used reference genes (*Actb*, *Gapdh*, and *Ppia*) had the highest CV_t (Figure 7A and B). For RT-qPCR, the initial set of eight transcripts that were chosen for nearly constant expression by RNAseq had higher CV_t than the second set of seven that were predicted to have constant expression in RT-qPCR (Figure 7A). In contrast, the opposite was seen for the same genes assessed by RNAseq (Figure 7B). For RT-qPCR, the least variation across stages was exhibited by *Hmgb3* followed by *Rb1cc1* and *Ccna2*. By RNAseq, the transcripts with the least variation were *Mcm7* and then *Hspa9* and *Ociad1*.

For normalizing expression data, the accepted standard is that more than one reference gene be used [12] and it is recommended that the geometric mean of several genes be used for normalization [18]. We therefore calculated the geometric means of the seven transcripts that approximated constant expression across stages in all combinations of two or three transcripts (Supplemental File S8). Combining either pairs or triplets of the transcripts resulted in only minimal improvements in the calculated values for CV_t. For single transcripts, the lowest CV_t was 0.28 for *Hmgb3*, whereas among all pairs the lowest was 0.27 for *Hmgb3* paired with either *Rb1cc1* or with *Cd320*, and for triplets, the lowest was 0.25 for *Hmgb3*, *Rb1cc1*, and *Taf1d* followed by 0.27 for *Hmgb3*, *Rb1cc1*, and *Cd320* (Supplemental File S8). Combining *Hmgb3*, *Rb1cc1*, and *Taf1d* did, however, reduce MAX_t/MIN_t to 2.00 compared to 2.33 for *Hmgb3* and *Rb1cc1* as a pair or 2.47 for *Hmgb3* alone. The means or geometric means at each stage are shown in Figure 7C for the single transcripts, pairs, and triplets with the two lowest calculated CV values.

Dividing by reference genes should generally preserve the expression patterns of a given transcript across stages of development. We therefore plotted the expression of transcripts from the second set plus previously used housekeeping genes (Figures 5 and 6) as their non-normalized expression, normalized to pTriXEF, and normalized to the geometric mean of *Hmgb3*, *Rb1cc1*, and *Taf1d* (Supplementary Figure S6). This confirmed that the expression patterns were not substantially changed by division by the geometric mean that was proposed as a suitable reference.

Discussion

Transcripts were identified that have approximately constant expression by RT-qPCR across mouse GV oocytes, MII eggs, and 1-cell, 2-cell, 4-cell, 8-cell, morula, and blastocyst embryos (Figure 5). The transcripts exhibiting the least variation when assessed by RT-qPCR were *Hmgb3* and *Rb1cc1* (Figure 7A), whose mean expression at each stage fell well within a factor of two in either direction from the overall inter-stage mean (Figure 7C). Furthermore, the inter-stage variability in expression of these transcripts was comparable to the variation between replicates within the same experiment (Figure 5) and is certainly comparable to the somewhat larger variation between independent repeats (Figure 6). Therefore, we propose that *Hmgb3* and *Rb1cc1* may be appropriate reference genes for mouse oocytes, eggs, and preimplantation embryos.

A small further decrease in variability was achieved by combining three transcripts, *Hmgb3*, *Rb1cc1*, and *Taf1d*, by their geometric mean. It is highly recommended that at least three reference genes be used whenever possible and that the genes be from pathways with different biological functions to minimize the chance that each will be similarly affected by any physiological perturbations [12, 18]. *Hmgb3* is High Mobility Group Box 3, which binds DNA and functions in DNA replication and repair [40]. *Rb1cc1* is RB1 Inducible Coiled-Coil 1 (a.k.a. FIP200, FAK-family Interacting Protein of 200 kDa), which is involved in binding protein kinases and autophagy [41]. *Taf1d* is TATA-Box Binding Protein Associated Factor, RNA Polymerase 1 Subunit D (a.k.a. TAF(I)41), that is a component of RNA polymerase 1 [42]. *Cd320* (transcobalamin receptor CD320) which functions in vitamin B_{12} uptake into cells [43] could be substituted for *Taf1d* with only a slight increase in variability (Figure 7C). These should provide enough diversity in physiological pathways to avoid excess correlation in changes induced by perturbations.

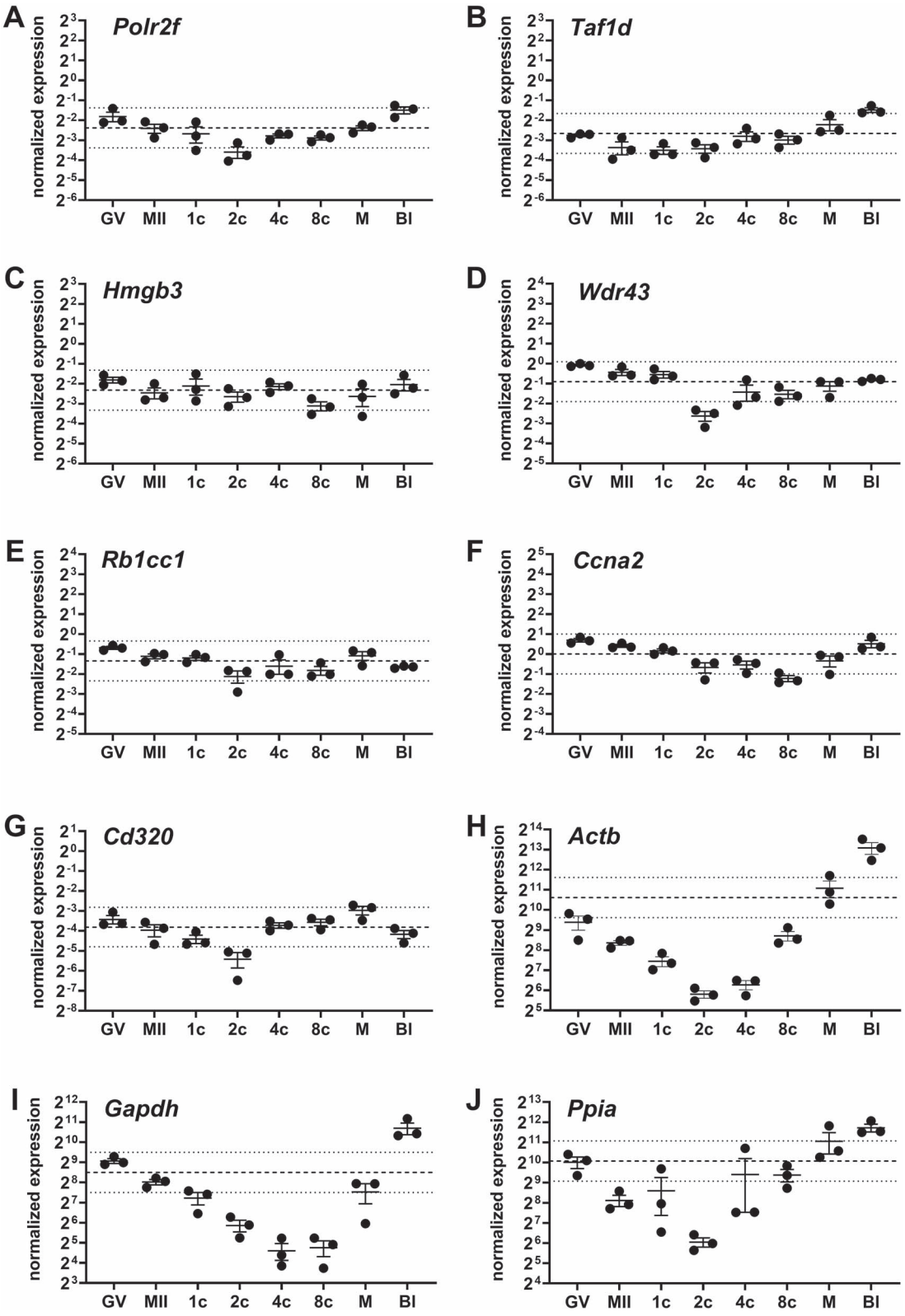


Figure 5. RT-qPCR measurement of expression of transcripts predicted to have nearly constant expression. Transcripts that were predicted to have nearly constant expression across stages by RT-qPCR (Figure 4A–G) were measured by RT-qPCR ($N = 3$ independent repeats) as described in the text. (**A**–**G**) Expression of each transcript is plotted as the ratio of its expression to that of *pTRI-Xef* in the same sample on a $\log_2$ scale. The patterns of expression were similar for each of the seven transcripts, with the means at most stages lying within a factor of two of the overall mean except as noted (see text). (**H**–**J**) Three genes commonly used as reference genes were also assessed. Means ± SEM are indicated by a horizontal line with error bars. The overall mean calculated as the mean of the means at each stage is indicated by horizontal dashed lines with ± one factor of two above or below shown by dotted lines. Stages are indicated as GV oocyte (GV), mature MII egg (MII), 1-cell (1c), 2-cell (2c), 4-cell (4c), and 8-cell (8c) embryos, morulae (M), and blastocyst (Bl). Statistical analysis is described in the text.

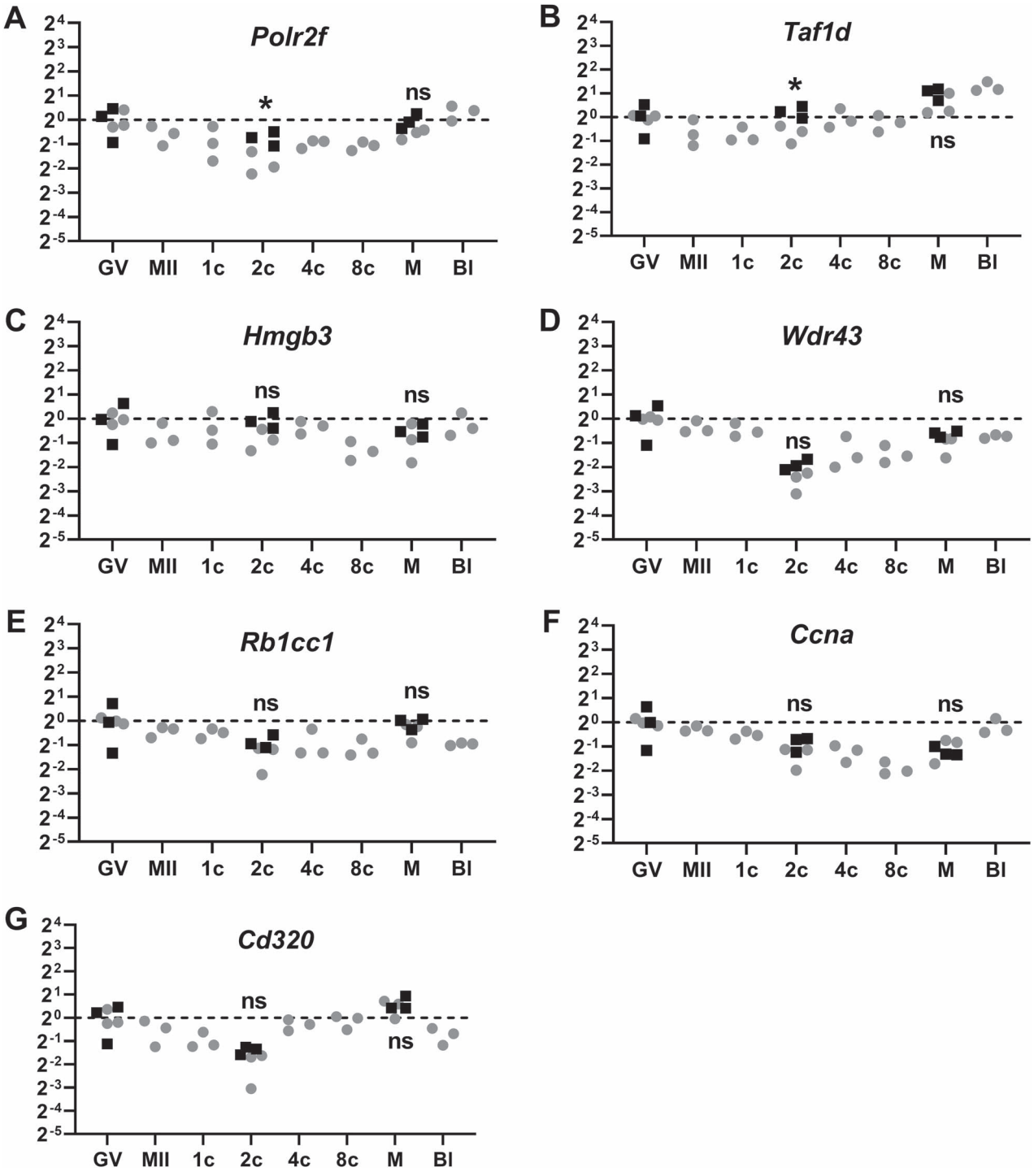


Figure 6. Reproducibility of RT-qPCR. Three independent collections at selected stages (GV, 2-cell, and morula) were performed and assessed by RT-qPCR for each of the seven chosen transcripts. The data from the new repeats are shown as black squares, whereas the previous data from Figure 5A–G are shown as gray circles for comparison. To allow direct comparison, the data were expressed relative to *pTRI-Xef* in the same samples as previously done, and then each repeat was further normalized to the expression at its GV stage for each transcript. There were no significant differences (ns) between the two sets of RT-qPCR results except where indicated (∗, see text). The small deviations from constant expression across stages exhibited by most transcripts were again evident, indicating that these expression patterns were reproducible. Stages are indicated as GV oocyte (GV), mature MII egg (MII), 1-cell (1c), 2-cell (2c), 4-cell (4c), and 8-cell (8c) embryos, morulae (M), and blastocyst (Bl).

Previous attempts to identify "stable" reference genes for preimplantation embryos using published algorithms and software packages [17–19] failed to find transcripts that approached constant expression across embryo stages [13–16]. This is likely because of the mathematical models used to predict transcript stability. For example, the commonly used NormFinder models the expression of each transcript as a linear combination of the actual gene expression level, the amount of mRNA in the sample, and random variation [17]. Because the second term—the amount of mRNA in

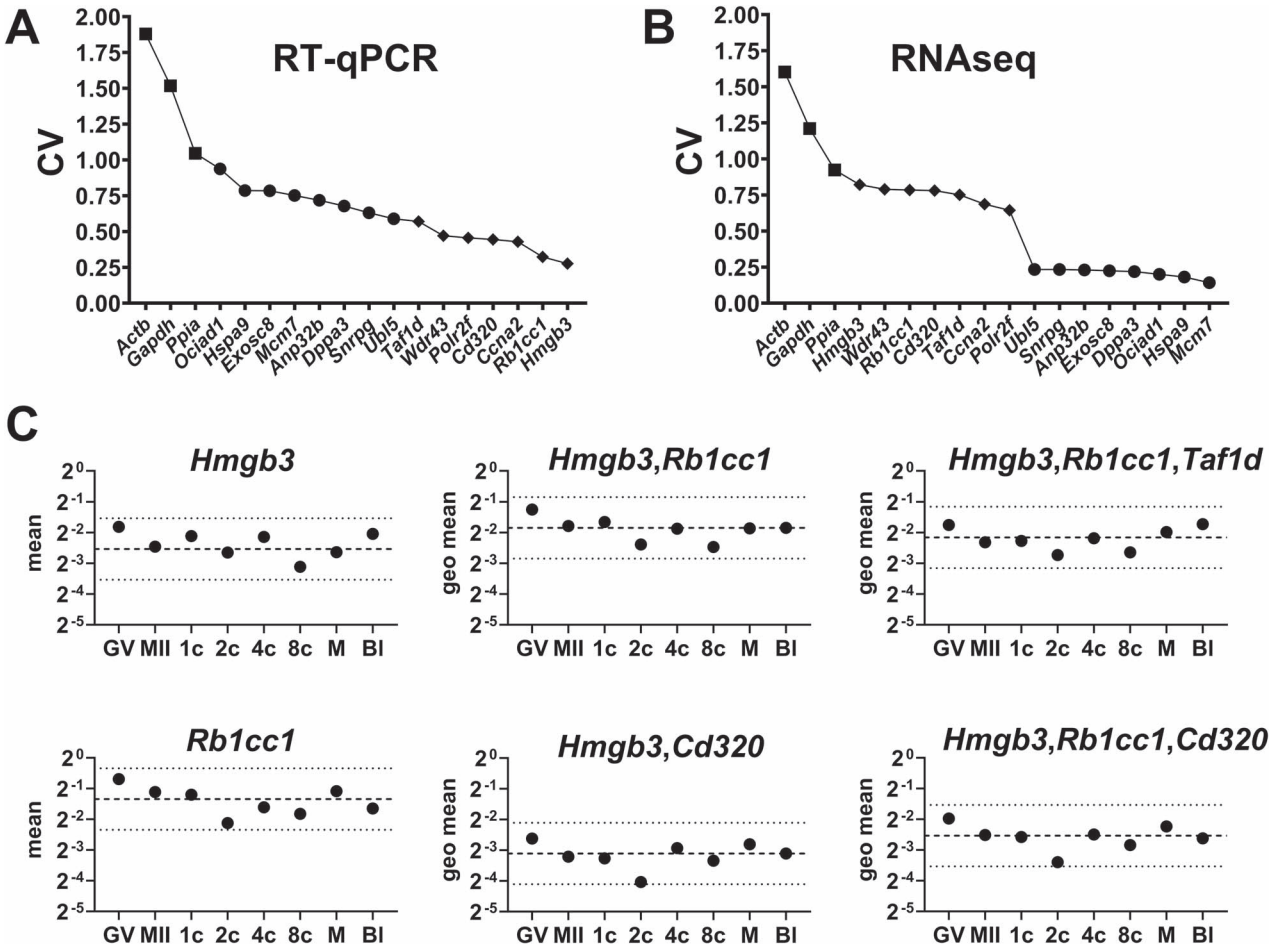


Figure 7. Transcripts ranked by coefficient of variation between stages. The coefficients of variation (CV_t) between the mean expression levels at each stage were calculated for each transcript from both sets and for three previously used reference genes for (**A**) RT-qPCR and (**B**) RNAseq. Ranking is from most variable to least, left to right. (**C**) Mean expression of the two genes with the lowest CV values (*Hmgb3* and *Rb1cc1*) are compared to the two geometric means of pairs or triplets with the lowest CV values. Means among stages are shown by the horizontal dashed line with ± factor of 2 indicated by dotted lines. The symbols represent the means at each stage. For the single genes, these are the means of the three independent repeats normalized to the *pTRI-Xef* spike-ins at each stage and are the same means shown in Figure 5C and E. Geometric means were calculated from the means (normalized to *pTRI-Xef*) of the individual genes at each stage. Stages are indicated as GV oocyte (GV), mature MII egg (MII), 1-cell (1c), 2-cell (2c), 4-cell (4c), and 8-cell (8c) embryos, morulae (M), and blastocyst (BI).

each sample—is determined by fitting the model to RT-qPCR data for the set of genes assessed, it will not distinguish a physiological decrease in total mRNA such as occurs at the 2-cell stage from a random variation in the amount of mRNA loaded in the sample. Thus, physiological changes in transcript levels that follow a pattern that is common among transcripts (e.g., the decrease in total mRNA at the 2-cell to 8-cell stages; Figure 3J) would be disregarded. Indeed, running NormFinder on all of the RT-qPCR data for the 18 transcripts (first set of eight, second set of seven, and three previously used reference genes; Figure 7A) identified *Anp32b* as the "most stable" gene (not shown), even though this transcript clearly varies substantially between stages (Figure 3A). The algorithms used by other programs, such as GeNorm which ranks by similarity of the ratio of a gene to each of the others in the set [18], would suffer from similar problems. This is likely why previous attempts to find appropriate reference genes for preimplantation embryos were unsuccessful and indicates that programs for identifying "stable" transcripts must not be used for sets of samples where the total mRNA varies because of physiological differences rather than chance, such as oocytes and preimplantation embryos. For that purpose, calculating the CVs to identify transcripts with the lowest CV among stages or minimizing Euclidean vector distance from constant expression across stages would instead be appropriate. Such programs that identify stable genes could, however, be used to identify whether transcripts are resistant to perturbation by various experimental conditions, such as gene knockouts or knockdowns, gene overexpression, or variations in culture conditions, as examples.

Transcripts with constant expression across stages were also identified for RNAseq (Figure 2). Of these, the least variable were *Mcm7* and *Hspa9* (Figure 7B), which not only showed nearly constant expression across all stages but also showed no outliers at each stage (Figure 2D and E). While reference genes are not customarily used for RNAseq analysis, these may prove useful for quality control when few RNAseq replicates are performed.

For this project, we compiled a standardized and merged set of RNAseq data that covers mouse GV oocytes, MII eggs, and 1-cell, 2-cell, 4-cell, 8-cell, morula, and blastocyst stage embryos (Supplemental File S2, Table 1). By using the mean

expression for each transcript at each stage, the variability inherent in RNAseq when one or few repeats are obtained at each stage can be avoided since this merged data set has 8–16 repeats at each stage. Thus, this merged and standardized RNAseq data set may prove useful for assessing the expression patterns of genes of interest in mouse oocytes or early embryos. The conversion factors we derived here (Figure 3J) should furthermore allow prediction of the expected RT-qPCR expression patterns from RNAseq data.

Identification of a set of genes with nearly constant expression across stages by RNASeq provided the basis for comparing the expression of such genes when measured by RT-qPCR. This revealed that, after correcting for small variations from constant expression in the RNAseq data, all eight transcripts exhibited very similar patterns of expression across stages (Figure 3J). Since constant expression with RNAseq means that the transcript is expressed as the same fraction of total mRNA at each stage while RT-qPCR yields the absolute amount of the same transcript, the relationship between the two is equivalent to the relative amount of total mRNA at each stage. We found (Figure 3J) that GV oocytes had the largest amount of total mRNA, which decreased to about 63% in mature MII eggs and dropped further to about 47% in 1-cell stage embryos relative to total mRNA in GV oocytes. There was then a further decrease to an essentially constant minimum from the 2-cell through 8-cell stages, which these calculations indicate have only ~10–15% as much total mRNA as in GV oocytes. The total mRNA then increases by about 5-fold from this minimum through the morula stage to the blastocyst stage to reach a level about 75% that in GV oocytes.

As far as we are aware, total mRNA has not been precisely measured in mouse oocytes, eggs, and embryos. However, total RNA and poly(A) mRNA contents have been determined by classical chemical and spectroscopic methods. Total RNA was determined by optical density measurements of unmodified [44], dye (Azur B)-reacted [24], or radiolabeled [45] RNA, which indicated that total RNA dropped from the MII egg to about 60–70% in 2-cell to 4-cell embryos before increasing to very high levels at the blastocyst stage (~2.5–7-fold relative to GV oocytes, depending on blastocyst stage). Poly(A) RNA, which represents a portion of total mRNA, decreases steadily from the GV stage to the 2-cell stage which has a level of poly(A) RNA ~27% that of the GV stage. This then increases to reach about 150% in blastocysts [45]. This pattern is similar to that for total mRNA that we derived here, although amount of poly(A) mRNA relative to the GV stage would appear to be ~2-fold higher than total mRNA at each stage.

As described above, we have identified transcripts that are nearly constantly expressed by RT-qPCR and proposed *Hmgb3* and *Rb1cc1* as reference genes or using the geometric mean of *Hmgb3*, *Rb1cc1*, and *Taf1d* or *Cd320*. Although these were reproducible in our hands (Figure 6), it has not been rigorously tested whether their expression is stable against physiological perturbations or between strains of mice. There is some evidence for stability, however, afforded by the RNAseq data (Figure 4). There, it was evident that similar expression patterns were found among the independent repeats in the merged data set, which comprised different strains of mice, in vivo- and in vitro-derived eggs and preimplantation embryos, and with or without superovulation (Table 1). However, the stability of reference genes should be confirmed under the particular conditions to be used before employing them [19], which should be carried out as part of any set of experiments in which reference genes are used with oocytes or embryos.

In addition to any reference genes, a spike-in of exogenous cRNA such as the *pTri-XEF* used here should also be included. The spike-in cRNA allows correction for differences in the efficiency of reverse transcription and PCR, as is evident in Supplemental Figure S6. In contrast, reference genes control mainly for biological variation between samples or handling errors that result in loss of oocytes or embryos. The most informative practice may be to present the data for transcripts of interest normalized by a spike-in cRNA rather than normalizing by reference genes, while separately presenting data for reference genes also normalized to the spike-in, since this allows assessment of variability between the biological samples separate from any variability because of the RT-qPCR procedures themselves.

Acknowledgment

The authors thank Gareth Palidwor of the Ottawa Bioinformatics Core Facility for processing, standardizing, and merging the oocyte, egg, and embryo transcriptomics data sets. We also thank Samuel Baltz (Massachusetts Institute of Technology) for conceiving the analysis by Euclidean vector distance.

Supplementary material

Supplementary material is available at *BIOLRE* online.

Conflict of Interest

The authors have declared that no conflict of interest exists.

Data availability

All data are included in the article or in the supplementary material. The original RNAseq data sets that were used are publicly available in GEO as described in the text.

References

1. Moore GP, Lintern-Moore S, Peters H, Faber M. RNA synthesis in the mouse oocyte. *J Cell Biol* 1974; **60**:416–422.
2. Svoboda P. Mammalian zygotic genome activation. *Semin Cell Dev Biol* 2018; **84**:118–126.
3. Zeng F, Schultz RM. RNA transcript profiling during zygotic gene activation in the preimplantation mouse embryo. *Dev Biol* 2005; **283**:40–57.
4. Telford NA, Watson AJ, Schultz GA. Transition from maternal to embryonic control in early mammalian development: a comparison of several species. *Mol Reprod Dev* 1990; **26**:90–100.
5. Zeng F, Baldwin DA, Schultz RM. Transcript profiling during preimplantation mouse development. *Dev Biol* 2004; **272**: 483–496.
6. Israel S, Ernst M, Psathaki OE, Drexler HCA, Casser E, Suzuki Y, Makalowski W, Boiani M, Fuellen G, Taher L. An integrated genome-wide multi-omics analysis of gene expression dynamics in the preimplantation mouse embryo. *Sci Rep* 2019; **9**:13356.
7. Su YQ, Sugiura K, Woo Y, Wigglesworth K, Kamdar S, Affourtit J, Eppig JJ. Selective degradation of transcripts during meiotic maturation of mouse oocytes. *Dev Biol* 2007; **302**:104–117.
8. Paynton BV, Rempel R, Bachvarova R. Changes in state of adenylation and time course of degradation of maternal mRNAs during

oocyte maturation and early embryonic development in the mouse. *Dev Biol* 1988; **129**:304–314.
9. Posfai E, Tam OH, Rossant J. Mechanisms of pluripotency in vivo and in vitro. *Curr Top Dev Biol* 2014; **107**:1–37.
10. Duncan FE, Schultz RM. Gene expression profiling of mouse oocytes and preimplantation embryos. *Methods Enzymol* 2010; **477**:457–480.
11. Ki A, Yamamoto R, Franke V, Cao M, Suzuki Y, Suzuki MG, Vlahovicek K, Svoboda P, Schultz RM, Aoki F. The first murine zygotic transcription is promiscuous and uncoupled from splicing and 3′ processing. *EMBO J* 2015; **34**:1523–1537.
12. Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T, Pfaffl MW, Shipley GL, Vandesompele J, Wittwer CT. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem* 2009; **55**:611–622.
13. Gu Y, Shen X, Zhou D, Wang Z, Zhang N, Shan Z, Jin L, Lei L. Selection and expression profiles of reference genes in mouse preimplantation embryos of different ploidies at various developmental stages. *PloS One* 2014; **9**:e98956.
14. Jeong J-K, Kang M-H, Gurunathan S, Cho S-G, Park C, Seo HG, Kim J-H. Evaluation of reference genes in mouse preimplantation embryos for gene expression studies using real-time quantitative RT-PCR (RT-qPCR). *BMC Res Notes* 2014; **7**:675.
15. Mamo S, Gal AB, Bodo S, Dinnyes A. Quantitative evaluation and selection of reference genes in mouse oocytes and embryos cultured in vivo and in vitro. *BMC Dev Biol* 2007; **7**:14.
16. Jeong YJ, Choi HW, Shin HS, Cui XS, Kim NH, Gerton GL, Jun JH. Optimization of real time RT-PCR methods for the analysis of gene expression in mouse eggs and preimplantation embryos. *Mol Reprod Dev* 2005; **71**:284–289.
17. Andersen CL, Jensen JL, Ørntoft TF. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res* 2004; **64**:5245–5250.
18. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* 2002; **3**:research0034.1.
19. Kozera B, Rapacz M. Reference genes in real-time PCR. *J Appl Genet* 2013; **54**:391–406.
20. Goossens K, Van Poucke M, Van Soom A, Vandesompele J, Van Zeveren A, Peelman LJ. Selection of reference genes for quantitative real-time PCR in bovine preimplantation embryos. *BMC Dev Biol* 2005; **5**:27.
21. Llobat L, Marco-Jiménez F, Peñaranda DS, Saenz-de-Juano MD, Vicente JS. Effect of embryonic genotype on reference gene selection for RT-qPCR normalization. *Reprod Domest Anim* 2012; **47**: 629–634.
22. Ginzinger DG. Gene quantification using real-time quantitative PCR: an emerging technology hits the mainstream. *Exp Hematol* 2002; **30**:503–512.
23. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. A survey of best practices for RNA-seq data analysis. *Genome Biol* 2016; **17**:13.
24. Moore GP, Lintern-Moore S, Scott-Murphy L. Relative changes in the RNA contents of oocytes and preimplantation embryos of the mouse. *Reprod Nutr Dev* 1980; **21**:505–512.
25. Hogan B, Beddington R, Constantini F, Lacy E. *Manipulating the Mouse Embryo: A Laboratory Manual*. Plainview, NY: Cold Spring Harbor Press; 1994.
26. Lawitts JA, Biggers JD. Culture of preimplantation embryos. *Methods Enzymol* 1993; **225**:153–164.
27. Sha QQ, Yu JL, Guo JX, Dai XX, Jiang JC, Zhang YL, Yu C, Ji SY, Jiang Y, Zhang SY, Shen L, Ou XH, *et al.* CNOT6L couples the selective degradation of maternal transcripts to meiotic cell cycle progression in mouse oocyte. *EMBO J* 2018; **37**:37.
28. Seah MKY, Wang Y, Goy PA, Loh HM, Peh WJ, Low DHP, Han BY, Wong E, Leong EL, Wolf G, Mzoughi S, Wollmann H, *et al.* The KRAB-zinc-finger protein ZFP708 mediates epigenetic repression at RMER19B retrotransposons. *Development* 2019; **146**:dev170266.
29. Qiao Y, Ren C, Huang S, Yuan J, Liu X, Fan J, Lin J, Wu S, Chen Q, Bo X, Li X, Huang X, *et al.* High-resolution annotation of the mouse preimplantation embryo transcriptome using long-read sequencing. *Nat Commun* 2020; **11**:2653.
30. Xue Z, Huang K, Cai C, Cai L, Jiang C-y, Feng Y, Liu Z, Zeng Q, Cheng L, Sun YE, Liu J-y, Horvath S, *et al.* Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* 2013; **500**:593–597.
31. Pfender S, Kuznetsov V, Pasternak M, Tischer T, Santhanam B, Schuh M. Live imaging RNAi screen reveals genes essential for meiosis in mammalian oocytes. *Nature* 2015; **524**:239–242.
32. Liu W, Liu X, Wang C, Gao Y, Gao R, Kou X, Zhao Y, Li J, Wu Y, Xiu W, Wang S, Yin J, *et al.* Identification of key factors conquering developmental arrest of somatic cell cloned embryos by combining embryo biopsy and single-cell sequencing. *Cell Discov* 2016; **2**:16010.
33. Wang F, Shin J, Shea JM, Yu J, Bošković A, Byron M, Zhu X, Shalek AK, Regev A, Lawrence JB, Torres EM, Zhu LJ, *et al.* Regulation of X-linked gene expression during early mouse development by Rlim. *Elife* 2016; **5**:5.
34. Kim J, Singh AK, Takata Y, Lin K, Shen J, Lu Y, Kerenyi MA, Orkin SH, Chen T. LSD1 is essential for oocyte meiotic progression by regulating CDC25B expression in mice. *Nat Commun* 2015; **6**:10116.
35. Franke V, Ganesh S, Karlic R, Malik R, Pasulka J, Horvat F, Kuzman M, Fulka H, Cernohorska M, Urbanova J, Svobodova E, Ma J, *et al.* Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes. *Genome Res* 2017; **27**:1384–1394.
36. Wang C, Liu X, Gao Y, Yang L, Li C, Liu W, Chen C, Kou X, Zhao Y, Chen J, Wang Y, Le R, *et al.* Reprogramming of H3K9me3-dependent heterochromatin during mammalian embryo development. *Nat Cell Biol* 2018; **20**:620–631.
37. Madissoon E, Damdimopoulos A, Katayama S, Krjutškov K, Einarsdottir E, Mamia K, De Groef B, Hovatta O, Kere J, Damdimopoulou P. Pleomorphic adenoma gene 1 is needed for timely zygotic genome activation and early embryo development. *Sci Rep* 2019; **9**:8411.
38. McIntyre LM, Lopiano KK, Morse AM, Amin V, Oberg AL, Young LJ, Nuzhdin SV. RNA-seq: technical variability and sampling. *BMC Genomics* 2011; **12**:293.
39. Corbett HE, Dubé CD, Slow S, Lever M, Trasler JM, Baltz JM. Uptake of betaine into mouse cumulus-oocyte complexes via the SLC7A6 isoform of y+L transporter. *Biol Reprod* 2014; **90**:81.
40. Wen B, Wei YT, Zhao K. The role of high mobility group protein B3 (HMGB3) in tumor proliferation and drug resistance. *Mol Cell Biochem* 2021; **476**:1729–1739.
41. Yeo SK, Wang C, Guan JL. Role of FIP200 in inflammatory processes beyond its canonical autophagy function. *Biochem Soc Trans* 2020; **48**:1599–1607.
42. Gorski JJ, Pathak S, Panov K, Kasciukovic T, Panova T, Russell J, Zomerdijk JC. A novel TBP-associated factor of SL1 functions in RNA polymerase I transcription. *EMBO J* 2007; **26**: 1560–1568.
43. Kozyraki R, Cases O. Vitamin B12 absorption: mammalian physiology and acquired and inherited disorders. *Biochimie* 2013; **95**: 1002–1007.
44. Olds PJ, Stern S, Biggers JD. Chemical estimates of the RNA and DNA contents of the early mouse embryo. *J Exp Zool* 1973; **186**: 39–45.
45. Pikó L, Clegg KB. Quantitative changes in total RNA, total poly(a), and ribosomes in early mouse embryos. *Dev Biol* 1982; **89**: 362–378.