Cell Reports
Report

The Earliest Transcribed Zygotic Genes Are Short, Newly Evolved, and Different across Species

Patricia Heyn,[1] Martin Kircher,[2,4] Andreas Dahl,[3] Janet Kelso,[2] Pavel Tomancak,[1,*] Alex T. Kalinka,[1,6] and Karla M. Neugebauer[1,5,*]

[1]Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108, 01307 Dresden, Germany
[2]Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany
[3]Deep Sequencing Group SFB 655, DFG Research Center for Regenerative Therapies (CRTD), Biotechnology Center (Biotec), Technische Universität Dresden, Fetscherstrasse 105, 01307 Dresden, Germany
[4]Department of Genome Sciences, University of Washington, 3720 15th Avenue NE, Seattle, WA 98195, USA
[5]Present address: Molecular Biophysics and Biochemistry, Yale University, 333 Cedar St, New Haven, CT 06520, USA
[6]Present address: Institute of Population Genetics, Vetmeduni Vienna, Veterinärplatz 1, 1210 Vienna, Austria
*Correspondence: tomancak@mpi-cbg.de (P.T.), karla.neugebauer@yale.edu (K.M.N.)
http://dx.doi.org/10.1016/j.celrep.2013.12.030

SUMMARY

The transition from maternal to zygotic control is fundamental to the life cycle of all multicellular organisms. It is widely believed that genomes are transcriptionally inactive from fertilization until zygotic genome activation (ZGA). Thus, the earliest genes expressed probably support the rapid cell divisions that precede morphogenesis and, if so, might be evolutionarily conserved. Here, we identify the earliest zygotic transcripts in the zebrafish, *Danio rerio*, through metabolic labeling and purification of RNA from staged embryos. Surprisingly, the mitochondrial genome was highly active from the one-cell stage onwards, showing that significant transcriptional activity exists at fertilization. We show that 592 nuclear genes become active when cell cycles are still only 15 min long, confining expression to relatively short genes. Furthermore, these zygotic genes are evolutionarily younger than those expressed at other developmental stages. Comparison of fish, fly, and mouse data revealed different sets of genes expressed at ZGA. This species specificity uncovers an evolutionary plasticity in early embryogenesis that probably confers substantial adaptive potential.

INTRODUCTION

In all metazoans, the fertilized embryo is provided with proteins and RNAs deposited by the mother during oogenesis. These maternal stores support early embryonic cell divisions before the onset of transcription at zygotic genome activation (ZGA), which occurs after a stereotypical number of cell cycles (Baroux et al., 2008; Tadros and Lipshitz, 2009). The mechanisms regulating ZGA are not fully understood in any species, though recent advances in *Drosophila melanogaster* and *Danio rerio* have begun to identify the maternally provided transcription factors governing the onset of at least some zygotically activated genes (Baroux et al., 2008; De Renzis et al., 2007; Lee et al., 2013; Leichsenring et al., 2013; Tadros and Lipshitz, 2009). In the zebrafish, these studies have focused on a period of robust zygotic transcription that follows the midblastula transition (MBT) when cell cycles begin to lengthen (Kane and Kimmel, 1993). However, some genes are transcribed much earlier. For example, zebrafish MBT occurs at the 1,000-cell stage (1K), yet several studies indicate that first gene transcription might occur as early as the 64-cell stage (Giraldez et al., 2006; Lindeman et al., 2011; Mathavan et al., 2005; O'Boyle et al., 2007). These first zygotic transcripts may play crucial roles in the overall process of ZGA, which continues for several hours during the period of rapid cell division that precedes morphogenesis. Therefore, determining the precise timing and identity of the earliest genes transcribed is central for understanding developmental mechanisms.

The gene expression machinery is confronted by a unique set of challenges during embryogenesis, because the production of functional RNA products is limited by the time it takes to transcribe and process RNA (Swinburne and Silver, 2008). Because transcription shuts down at mitosis and incomplete transcripts are "aborted," short early embryonic cell cycles are thought to impose a time limit on the genes that can be expressed (Shermoen and O'Farrell, 1991; Tadros and Lipshitz, 2009). The notion that cell-cycle length restricts the potential for gene expression to short genes received support from seminal work in fly development (McHale et al., 2011; McKnight and Miller, 1976; Rothe et al., 1992). Thus, we would assume that early embryos may only be able to express short genes with relatively simple exon-intron architecture. Alternatively, early embryos may have evolved mechanisms for overriding transcript abortion; if so, expression of longer genes might occur. Indeed, one study in zebrafish did identify long genes with many introns among the earliest expressed genes (Mathavan et al., 2005). Thus,

Cell PRESS
Open ACCESS

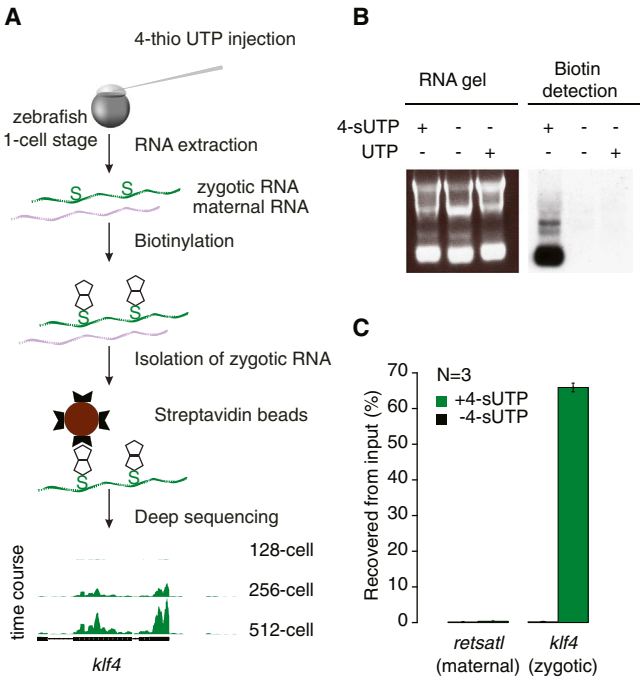


Figure 1. Metabolic Labeling with 4-sUTP Identifies Zygotically Transcribed Genes

(A) Schematic of the method for positive selection of zygotic transcripts. Embryos are microinjected with 4-sUTP and grown to the desired developmental stage. Total RNA is extracted and biotinylated in a thiol-specific manner. Biotinylated RNA is captured on magnetic streptavidin beads, extensively washed, and used for downstream applications; here, deep sequencing at three developmental stages and the outcome showing gene coverage for *klf4* is depicted.

(B) Northern blot from 4-sUTP-injected, UTP-injected, or uninjected wild-type control embryos (3 hr postfertilization [hpf]). The panel on the right shows "Northwestern" detection of the biotinylated 4-sUTP incorporated in the RNA.

(C) Quantification of RNA isolation efficiency by qRT-PCR. Plotted are average values (±SD) for percent recovered from input for one maternal gene and one zygotic gene from 3hpf 4-sUTP (+4-sUTP)-microinjected and wild-type control embryos (−4-sUTP) (n = 3).

See also Figures S1 and S2.

comprehensive identification of transcribed genes in early embryos provides an opportunity to interrogate cellular and developmental mechanisms.

We sought to address early zygotic transcription in the zebrafish, where the full complement of genes transcribed at the very beginning of ZGA remains unknown. Several studies have performed total steady-state RNA sequencing (RNA-seq) from early embryonic samples (Aanes et al., 2011; Pauli et al., 2012; Vesterlund et al., 2011). However, uncovering which zygotic genes are transcribed earliest is hampered by the presence of massive amounts of maternally loaded RNA. Not only are newly transcribed RNAs relatively less abundant, but also "maternal-zygotic" RNAs—those that are maternally provided and are additionally synthesized at ZGA—cannot be deduced from total pools (Supplemental Discussion). To overcome this, recent SNP analysis of maternal and paternal genomes was used to demonstrate the zygotic expression of 3,342 protein-coding genes from MBT onwards (Harvey et al., 2013). However, SNP analysis is not a comprehensive method, because only ~25% of all genes contained diagnostic SNPs. Previously, other specialized methods have been developed to identify zygotically transcribed genes in flies and plants (De Renzis et al., 2007; Lott et al., 2011; Nodine and Bartel, 2012). Here, we identify the earliest zygotic transcripts in the zebrafish through the establishment of a method to isolate newly transcribed RNA by metabolic labeling of RNA from staged embryos. This is a comprehensive method that allows us to extend robust analysis to vertebrates. We used this data set to compare the first zygotically transcribed genes in fish, fly, and mouse and address whether the set of genes expressed at ZGA is evolutionarily conserved. The results shed light on both the regulation of transcription onset and the evolution of early animal development.

RESULTS AND DISCUSSION

In order to identify the full complement of the earliest zygotic transcripts, we established a protocol based on RNA metabolic labeling and positive selection of newly synthesized transcripts (Figure 1A; Figure S1A). Contamination from maternal RNA was minimal (Figures 1B and 1C; Figures S1B and S1C), and recovery of zygotic transcripts (e.g., the known zygotic gene *klf4*) was highly efficient (Figure 1C). Therefore, embryos injected at the one-cell stage with 4-thio-UTP (4-sUTP) were allowed to develop to the 128-, 256-, and 512-cell stages before extraction, biotinylation, purification of labeled RNA, and deep sequencing (Figure 1A). We focused on this early sequence of cell divisions because several genes are known to be transcribed before MBT and active chromatin marks have been detected at the 256-cell stage (Lindeman et al., 2011; Mathavan et al., 2005). Coverage of many detected transcripts, such as *klf4*, *vox*, and *cited3*, increased during the cell cycles examined (Figures 1A and 2A), indicating that transcription begins during this time window. Therefore, accumulation of metabolically labeled transcripts over time was used as a signature for robust identification of early zygotic transcripts.

Although it is believed that zygotic genomes are transcriptionally silent before ZGA, RNAs transcribed from the mitochondrial genome were abundant at all three developmental time points (Figure 2B; Figure S2). These data suggest that mitochondrial RNA polymerase is active before the 128-cell stage and that expression has reached steady-state levels. Indeed, independent quantitative RT-PCR (qRT-PCR) experiments show that mitochondrial RNA precursors are clearly detectable from fertilization onward (Figures S3A–S3F). Constitutive mitochondrial transcriptional activity is likely related to the fact that the mitochondrial genome is not packaged into chromatin or subject to DNA methylation (Holt et al., 2007; Potok et al., 2013). Therefore, mitochondrial genome activity is independent of nuclear genome regulation, and the dogma that the early zygote is transcriptionally silent (Baroux et al., 2008; De Renzis et al., 2007; Kane and Kimmel, 1993; Lee et al., 2013; Leichsenring et al., 2013; O'Boyle et al., 2007; Tadros and Lipshitz, 2009) is not strictly correct.

Nuclear genome activity was detected between the 128- and 512-cell stages, with clear increases in transcript abundance observed for 592 nuclear genes, comprising 670 transcript

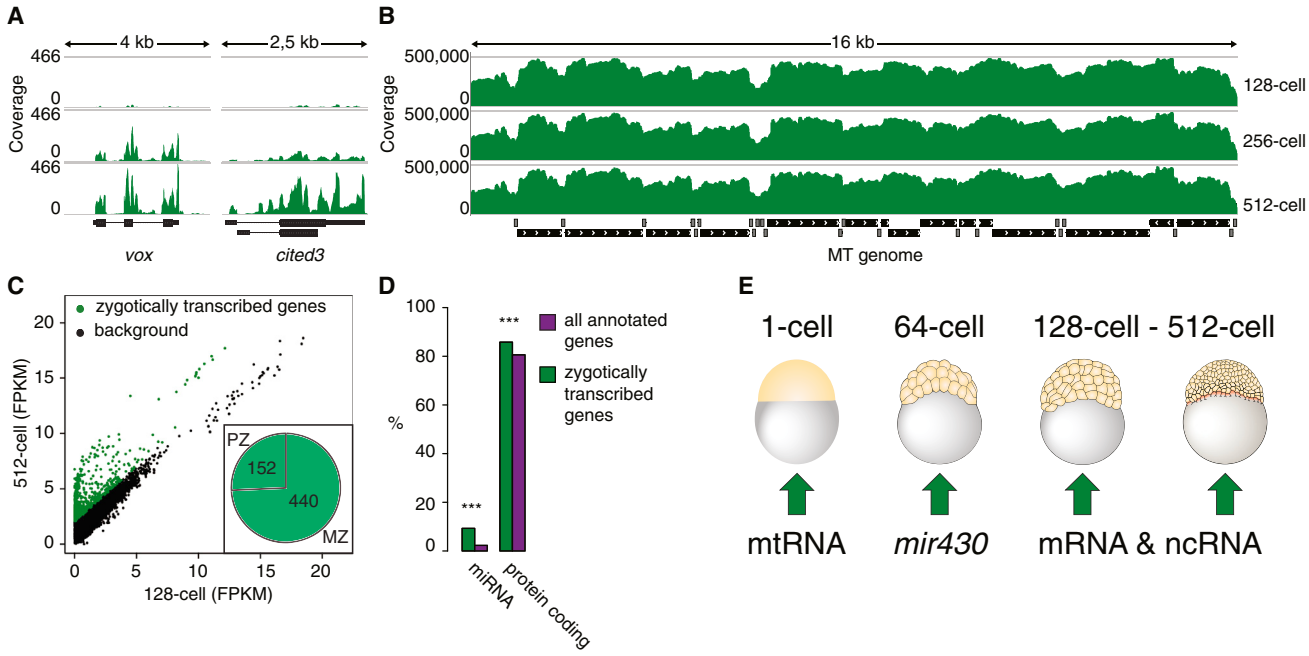


Figure 2. Timing of Protein-Coding, miRNA, and Mitochondrial Gene Transcription during ZGA

(A) Example of increasing read coverage over the zygotically transcribed genes (*vox* and *cited3*). The architecture of the respective genes is displayed below.

(B) Display of constant high read coverage over the mitochondrial genome. The architecture of the zebrafish mitochondrial genome is displayed below. Protein-coding genes are colored in black and tRNAs in gray.

(C) Scatterplot of abundance for zygotically transcribed genes (green, n = 592). Genes in black are unchanged between the developmental time points. To avoid negative values in the log2 scale, 1 was added to all FPKM values before transformation. The inset shows the proportion of purely zygotically transcribed genes (PZ) and maternal-zygotic expressed genes (MZ).

(D) Identification of overrepresented transcript types in zygotically transcribed genes compared to the overall distribution in all annotated zebrafish genes. The proportions of each biotype among the zygotically transcribed genes (green bars) and among all annotated genes (purple) are plotted in percent. The stars indicate significant enrichment over random distribution (hypergeometric test, *** equals p value $\leq$ 0.001; Table S3). For other biotypes present, see Figure S5A.

(E) Schematic showing developmental stages and the detection of different types of newly transcribed genes. See also Figures S3–S5 and Tables S1, S2, and S3.

isoforms (Figure 2C; Tables S1 and S2; Figures S2B and S4A–S4C). Transcription inhibition by α-amanitin showed that purification of metabolically labeled RNA is due to incorporation of 4-sUTP during transcription (Figure S4D). Because our RNA-seq data were normalized to constant mitochondrial RNA levels, the identified changes are specific for each gene and not the result of global shifts in RNA populations. Comparison of this data set with the recent study based on SNP analysis (Harvey et al., 2013) revealed that our method detects 350 zygotically transcribed genes that lack informative SNP loci. In addition, 84 SNP-containing genes overlapped with our data set. Interestingly, Pou5f1 was recently identified as a major regulator of ZGA in zebrafish embryos (Lee et al., 2013; Leichsenring et al., 2013). We find enrichment of Pou5f1 and Sox2 peaks among ~50% of our genes (data not shown), making it plausible that Pou5f1 drives expression of at least some of these early zygotic genes.

The zygotically expressed genes identified in our study primarily encode proteins and microRNAs (miRNAs) transcribed by RNA polymerase II (Figure 2D; Figures S5A and S5B; Table S3). As expected, we detected *mir430*, which has demonstrated roles in embryonic RNA regulation (Giraldez et al., 2006), as well as *mir19a*, a newly identified zygotic transcript (Figure S4C). In addition, noncoding RNAs transcribed by Pol III were robustly identified (Table S1). Of the 592 genes detected, 152 were "purely zygotic" (PZ) and not detected in the maternal pool (Figure 2C, inset). Increased detectability of RNA precursors over the time course is consistent with de novo transcription (Figure 2A; Figures S4A–S4C). Note that many introns from robustly transcribed genes are poorly covered (see Figure 2A), likely because pre-mRNA splicing is a fast cotranscriptional process (Brugiolo et al., 2013). We conclude that early ZGA begins with transcription of *mir430* genes by the 64-cell stage, followed by nuclear protein-coding genes and other noncoding RNAs at the 256- to 512-cell stage (Figures 2E and S4C). The remaining 74% (440) of genes were "maternal-zygotic" (MZ; Figure 2C, inset), meaning that transcripts were maternally provided and then synthesized again at early ZGA. MZ genes account for only 3% of the total number of maternally expressed genes (14,500) previously identified (Pauli et al., 2012). This suggests that the re-expression of selected maternal genes at early ZGA may be specifically required.

Comparison of maternal and zygotic protein-coding genes reveals dramatic differences in function, reflecting a trend toward RNA regulation for maternal transcripts and DNA regulation

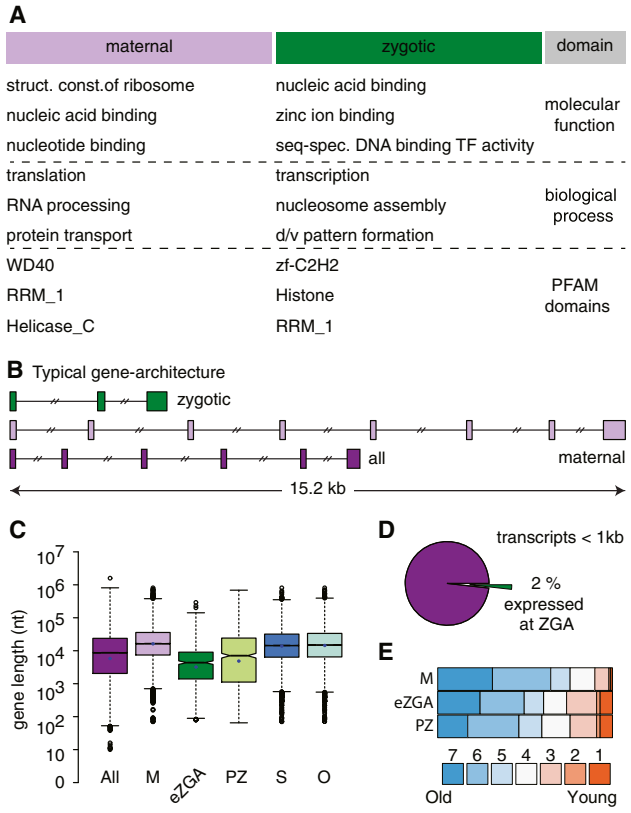


Figure 3. Zygotically Transcribed Genes Are Short and Intron Poor

(A) Gene Ontology term and protein families (PFAM) domain analysis of maternally expressed and zygotically transcribed genes. The three most enriched terms or domains are shown.

(B) Stick diagram of typical gene architecture for zygotic, maternal, and all annotated zebrafish transcripts. Drawn to scale is the median length of the genes (zygotically transcribed median = 3,874 bp; maternally expressed median = 15,202 bp and all annotated genes median = 8,644bp) and the first and last exons. For internal exons, the population median for all exons per transcript is drawn. Introns are not to scale; median numbers of introns (zygotic: 2, maternal: 7; and all: 5) are shown.

(C) Distribution and median gene length for different developmental gene categories for *D. rerio*. The categories correspond to all annotated genes (all), maternal genes (M), early zygotic genes (eZGA), purely zygotic genes expressed after MBT (PZ), genes expressed at somitogenesis (S), and genes expressed at organogenesis (O). For p values of all pair-wise comparisons, see Table S6. Table S11 lists all data sets used.

(D) Pie chart showing the proportion of zygotic transcripts that are shorter than 1kb among all (n = 5,822).

(E) Stacked bar graphs showing the normalized distribution of the inferred age of zebrafish genes expressed at maternal stage (M) and at early ZGA (eZGA) and purely zygotic genes expressed after MBT (PZ). Gene age is retrieved from Ensembl protein trees depicted schematically in Figure S8C. Different colors correspond to numbers depicted underneath the bar graph and indicate the node (clade) in the protein tree in which the genes emerged, where smaller numbers represent more recent emergence.

See also Figures S5–S8 and Tables S4, S5, S6, and S7.

for zygotic transcripts (Figure 3A; Tables S4 and S5). For example, ribosomes and spliceosomes are abundantly contributed maternally to facilitate high levels of translation and splicing during early embryogenesis (Amsterdam et al., 2004; Strzelecka et al., 2010). Interestingly, 143 of 144 genes annotated for "structural constituent of the ribosome" are detected among maternal transcripts, while only 13 of these genes were detected among zygotic transcripts. Because zygotic rRNA and small nuclear RNA expression were also not detected (Figure S5B), early embryos may rely exclusively on maternally provided ribosomes and spliceosomes. In contrast, zygotic RNAs encode DNA binding proteins, histones and histone variants, and chromatin modifiers (Figure 3A; Tables S4 and S5). A number of transcription factors with later roles in patterning and organogenesis, such as Forkhead and Dickkopf proteins, are among the purely zygotic transcripts; roles for these factors prior to the first steps of morphogenesis are thus far unknown (Sprague et al., 2003). On the other hand, the observed early zygotic expression of cell-cycle regulators, such as c-*fos* (*fos*) and cyclin-dependent kinase inhibitor (*cdkn1a*), could reflect their involvement in establishing gap phases in the cell cycle at MBT.

Gene transcription and RNA processing take time, and transcription elongation is usually aborted at mitosis. Cell-cycle lengths in the zebrafish are only 15 min long prior to MBT, suggesting that zygotic gene lengths may also be limited. However, work in zebrafish suggested that very long genes can indeed be expressed at early time points (Mathavan et al., 2005). Analysis of the architecture of our 592 zygotically expressed genes shows that early zygotic genes are four times shorter than maternal genes (Figure 3B and 3C; Figures S5C and S5D; Tables S6 and S7). They are relatively intron poor or intronless (Figure 3B; Figures S5E and S5F). These findings suggest that many of the long genes identified in the previous study might reflect contaminating levels of maternal RNA. This highlights the utility of a positive selection method, such as the one presented here, to identify all classes of newly transcribed genes. Orthologs of the zebrafish early zygotic genes in the very compact *Fugu* as well as the enlarged *Coelacanth* genomes are also short (Figure S6; Table S6). To address whether gene shortness is specific for ZGA or might coincidentally reflect the types of genes enriched in the zygotic pool, the gene architecture of all transcription factors was analyzed. We show that transcription factor genes are not short in general, yet those that are zygotically expressed are significantly shorter (Figure S7A; Table S6; Mann-Whitney *U* test, $p < 0.001$). Furthermore, if transcription factors are eliminated from the zygotic pool, we still observe an enrichment of short genes (Figure S7B). In contrast, maternally expressed genes as well as genes expressed during differentiation and organogenesis (Aanes et al., 2011; Pauli et al., 2012) are significantly longer than the genome-wide median (Figures 3B and 3C; Figures S5C, S5D, S7A; Tables S6 and S7; Mann-Whitney *U* test, $p < 0.001$). Therefore, zygotic genes are short, and maternal genes are long.

The shortness of early zygotic genes in fruit fly (Figure S6B) and mosquito has been taken as evidence that cell-cycle length may determine gene expression in early embryos (Biedler et al., 2012; De Renzis et al., 2007; McKnight and Miller, 1976; Rothe et al., 1992). Moreover, one model proposes that gene length actively determines which genes are expressed at MBT; in this model, transcription activation is widespread but transcript abortion at mitosis prevents long genes from being expressed (Swinburne and Silver, 2008; Tadros and Lipshitz, 2009). We

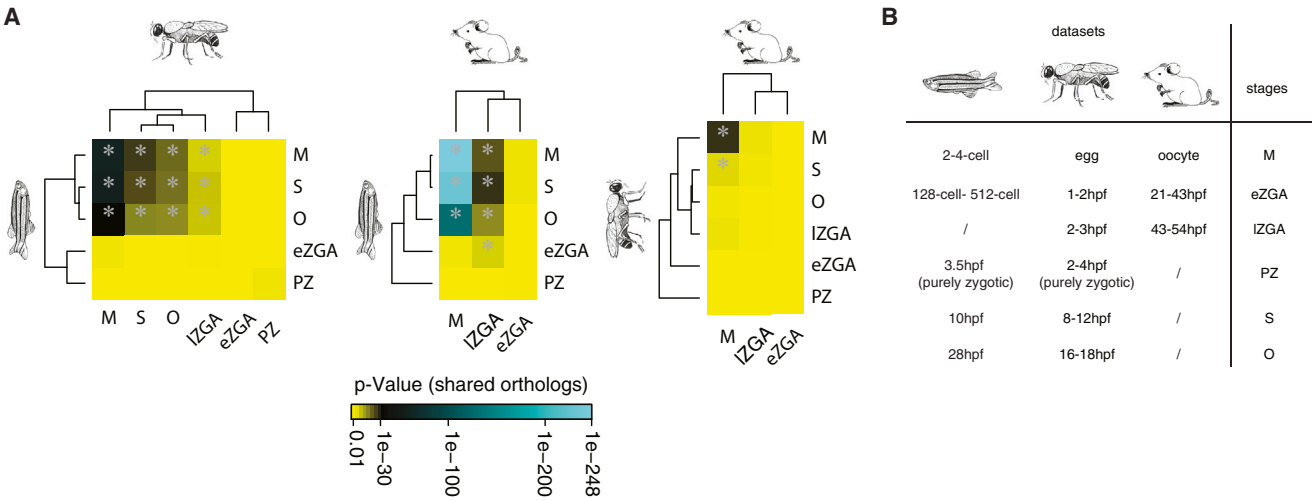


Figure 4. Expression of Early Zygotic Genes Is Evolutionarily Divergent

(A) Clustered heatmaps for pairwise ortholog comparisons between species (fly, fish, and mouse) at and between respective developmental stages. See Table S11 for a list of the data sets used. Depicted are color-coded p values describing the significance of shared ortholog enrichment between species. Enrichments were considered significant with a p value smaller than the Bonferroni cutoff of 0.05/63 = 0.00079, indicated by gray asterisks. Hierarchical clustering is solely for visualization purposes.

(B) Overview of developmental stages used for analysis of shared orthologs in the three species (fly, fish, and mouse).

See also Figure S9 and Tables S8, S9, S10, and S11.

tested this hypothesis by examining all genes less than 1 kb long (less than one-third the length of observed zygotic genes) in order to objectively evaluate expression of short genes. Only 2% of >5,000 short genes are expressed zygotically in zebrafish (Figure 3D). Therefore, our data rule out the possibility that zygotic transcription is determined by short gene length alone. Interestingly, the presence of introns in many of these zygotic genes, though short, may be an important factor in gene activation; it was recently shown that intron content positively regulates transcriptional output and quality (Bieberstein et al., 2012).

Short genes tend to be evolutionarily young (Grzybowska, 2012; Neme and Tautz, 2013; Shabalina et al., 2010), suggesting that early zygotic genes may have evolved recently. To test this, we compared the evolutionary age of genes expressed at different stages of zebrafish development. Genes expressed at ZGA were indeed younger than genes expressed maternally and later during tissue specification (Figure 3E; Figures S8A–S8D). Extending this analysis to published fruit fly data (De Renzis et al., 2007; Graveley et al., 2011; Hamatani et al., 2004; Xue et al., 2013; Zeng et al., 2004) (Figures S8A and S8C), an even tighter correlation regarding gene age was revealed (Figure S8A). The results show that genes expressed at early ZGA in the fly and the fish are significantly younger than expected (Figure S8D; χ^2 test, $p < 2.22 \times 10^{-16}$). This was not the case in the mouse (Figure S8). Mice and fish are both vertebrates and share a common ancestor more recently than they do with flies. However, fish and mice differ in the dynamics of early embryogenesis; for example, early cell division times differ by almost two orders of magnitude. Interestingly, cellular dynamics are more similar between fish and fly embryos, where cell-cycle lengths are only 8–15 min before ZGA. Therefore, cellular dynamics may be a major determinant of gene expression at ZGA across phyla and species.

Fly, fish, and mouse early embryos clearly differ in many ways. For example, in the syncytial blastoderm of the fly mitosis takes place without cytokinesis, while partial cleavages of the egg take place in fish; cell-cycle lengths and patterning strategies also differ. However as ZGA precedes cellular differentiation in all metazoans, one might expect specific mechanisms to have evolved in the last common ancestor and to have been retained in the animals we study today. The observation that early zygotic genes tend to be evolutionarily younger, while maternal genes are more ancient, prompted us to ask whether the sets of genes expressed in the early embryo at different time points are evolutionarily conserved. We conducted a comparative study using our data set and published data on fruit fly and mouse (De Renzis et al., 2007; Hamatani et al., 2004; Xue et al., 2013) to specifically interrogate the sets of genes in different species for shared expression of orthologs. This analysis reveals a strong enrichment of shared orthologs among maternally provided genes (Figure 4; Figure S9A; Tables S8 and S9; hypergeometric test, $p < 0.00079$). Between fruit fly and zebrafish, 2,773 shared orthologs were identified among the maternal genes, representing 19.1% of the zebrafish maternal gene pool (Table S9). The same trend was observed among the maternal genes of fruit fly and mouse (Figure 4; Figure S9B; Table S8). Maternal genes are also highly enriched for essential genes in all three species (Table S10; one-tailed hypergeometric test, $p < 0.005$). These data indicate that maternally contributed RNAs and proteins, and the essential functions they provide, constitute evolutionarily conserved features of early embryogenesis.

If early zygotic gene products also play a role in the awakening of the rest of the genome, then we would expect to reveal shared orthologs among the early zygotic genes in flies, fish, and mice as well. The recent identification of so-called hub genes, shared in

expression between mouse and human embryos, supports this notion (Xue et al., 2013). However, the identification of shared orthologs expressed among the three zygotic gene pools yielded a poor level of overlap that would be expected by chance (Figure 4; Figures S9A–S9C; Table S8; hypergeometric test, p = 1 [fly-fish early ZGA], p = 0.1027935 [fish-mouse early ZGA], p = 1 [fly-mouse early ZGA]). Note that our analysis only considers orthologs that are expressed in both species at specific developmental stages (shared orthologs). In contrast to the zygotic genes, adult traits associated with similar small numbers of genes showed strong enrichment across the same species (Figure S9D; Table S8). Remarkably, there are no shared orthologous early zygotic genes expressed in both fruit fly and zebrafish. Moreover, purely zygotic genes were depleted of essential genes (Table S10), consistent with the observation that young genes are less likely to be essential (Chen et al., 2012). Therefore, in stark contrast to maternal genes, early zygotic genes are completely different between arthropods and vertebrates as well as among vertebrates, suggesting species-specific functions.

Many classes of developmental events, such as signaling by growth factors, establishment of cell polarity, and the execution of differentiation programs, show a high degree of evolutionary conservation in the genes and gene products involved. On the other hand, mechanisms underlying other key events like sex determination and dosage compensation differ widely. We show that maternal transcripts are long, evolutionarily older, and tend to be shared across distantly related species. This suggests that the complement of maternal RNA and protein provided during oogenesis is critically important for the regulation of early embryogenesis by primarily RNA-driven mechanisms such as translation. In contrast, early zygotic genes are short, evolutionarily younger than other developmental genes, and not orthologous to early zygotic genes in other species. This indicates that ZGA occurs during a remarkably flexible period of development in which the order of gene activation is likely sculpted by species- or lineage-specific cellular mechanisms, such as morphology or the usage of yolk or trophectoderm to support the zygote (Kalinka and Tomancak, 2012). Such a correspondence between early, species-specific aspects of embryogenesis and the species specificity of the underlying molecular components of ZGA is consistent with the hourglass model of developmental evolution, in which the nature of early development reflects the diverse reproductive and ecological strategies of individual species (Duboule, 1994; Raff, 1996). However, the expression of shared sets of orthologs at the maternal stage does not necessarily imply that these components of early development are being used in identical ways in these different species. Rather, our results suggest that the unique constraints acting during ZGA provide a window of opportunity for the expression of evolutionarily younger, short genes that are capable of adding new functions to the zygotic gene expression program.

EXPERIMENTAL PROCEDURES

Metabolic Labeling

Fertilized zebrafish embryos were obtained according to approved protocols at the MPI-CBG zebrafish facility. Embryos were injected at the one-cell stage with 1 nl 50 mM 4-sUTP (Ambion, Trilink) and collected at the desired stage. Biotinylation and purification of labeled RNA was performed as described previously (Zeiner et al., 2008) with minor modifications.

RNA-Seq Analysis

Isolated RNA was amplified with the WT-Ovation One-Direct RNA Amplification System (NuGEN), converted into TruSeq sequencing libraries (Illumina), and sequenced on a HiSeq2000 instrument (Illumina). For a list of primers, see Table S12. Raw reads were trimmed and mapped with TopHat 1.3.3 (Trapnell et al., 2009) to the zebrafish genome assembly Zv9/GCA_000002035.2. Quantification of gene/transcript expression was performed with Cufflinks 2.0.2 (Trapnell et al., 2010). To normalize between different time points in our study, the sum of FPKM values for protein-coding mitochondrial transcripts/genes was assumed to be constant.

Gene Age and Ortholog Analysis

The age of each gene was determined by the age of the oldest node in its protein phylogeny (Piasecka et al., 2013) using protein trees. Protein trees, generated by EnsemblCompara (Vilella et al., 2009), a phylogeny-aware, gene-tree-building pipeline based on BLASTP, were downloaded from Ensembl for mouse and fish, and from Ensembl Metazoa for the fly. Enrichment of shared orthologs was tested for each pair of species and for genes expressed at several developmental stages. One to two orthologs were allowed for fish comparisons due to the whole-genome duplication in teleosts (1:2, fly:fish, mouse:fish), and a variant of the hypergeometric test was applied to test for enrichment (Kalinka, 2013). Enrichments were considered significant with a p value smaller than the Bonferroni cutoff of 0.05/63 = 0.00079.

ACCESSION NUMBERS

Raw and processed RNA-seq data sets have been deposited into NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE47709.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Discussion, Supplemental Experimental Procedures, nine figures, and twelve tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2013.12.030.

AUTHOR CONTRIBUTIONS

P.H. and K.M.N. designed the study; P.H. performed all experiments; A.D. provided expertise for library preparation; M.K. and J.K. developed and implemented analysis tools for the primary data; M.K., P.H., and A.T.K. performed secondary analyses; A.T.K. developed methods for ortholog analysis; and P.H. and K.M.N. wrote the manuscript with support from P.T., M.K., J.K., and A.T.K.

ACKNOWLEDGMENTS

We thank F. Carrillo Oesterreich, M. Hiller, J. Howard, A. Oates, N. Vastenhouw, and members of our lab for helpful discussions and comments on the manuscript. Sylvia Schimpke contributed valuable expertise to the deep sequencing, Holger Brandl contributed to data analysis, and Olivia Howard and Anja Rudnick provided artwork. We are grateful for funding from the Max Planck Society (to K.N., J.K., and P.T.), the Dresden International Graduate School for Biomedicine and Bioengineering (DIGS-BB fellowship to P.H.), and Deutsche Forschungsgemeinschaft (NE22 to K.N. and SFB 655 to the Deep Sequencing Group, CRTD). A.T.K. and P.T. were funded by Human Frontier Science Program Young Investigator grant (RGY0093/2012) and PT additionally received support from European Research Council Community's Seventh Framework Programme (FP7/2007-2013) grant agreement 260746.

Received: October 11, 2013
Revised: December 4, 2013
Accepted: December 17, 2013
Published: January 16, 2014

REFERENCES

Aanes, H., Winata, C.L., Lin, C.H., Chen, J.P., Srinivasan, K.G., Lee, S.G.P., Lim, A.Y.M., Hajan, H.S., Collas, P., Bourque, G., et al. (2011). Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. Genome Res. *21*, 1328–1338.

Amsterdam, A.A., Nissen, R.M.R., Sun, Z.Z., Swindell, E.C.E., Farrington, S.S., and Hopkins, N.N. (2004). Identification of 315 genes essential for early zebrafish development. Proc. Natl. Acad. Sci. USA *101*, 12792–12797.

Baroux, C., Autran, D., Gillmor, C.S., Grimanelli, D., and Grossniklaus, U. (2008). The maternal to zygotic transition in animals and plants. Cold Spring Harb. Symp. Quant. Biol. *73*, 89–100.

Bieberstein, N.I., Carrillo Oesterreich, F., Straube, K., and Neugebauer, K.M. (2012). First exon length controls active chromatin signatures and transcription. Cell Rep. *2*, 62–68.

Biedler, J.K., Hu, W., Tae, H., and Tu, Z. (2012). Identification of early zygotic genes in the yellow fever mosquito Aedes aegypti and discovery of a motif involved in early zygotic genome activation. PLoS ONE *7*, e33933.

Brugiolo, M., Herzel, L., and Neugebauer, K.M. (2013). Counting on co-transcriptional splicing. F1000Prime Rep. *5*, 9.

Chen, W.-H., Trachana, K., Lercher, M.J., and Bork, P. (2012). Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. Mol. Biol. Evol. *29*, 1703–1706.

De Renzis, S., Elemento, O., Tavazoie, S., and Wieschaus, E.F. (2007). Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo. PLoS Biol. *5*, e117.

Duboule, D. (1994). Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. Dev. Suppl. *1994*, 135–142.

Giraldez, A.J., Mishima, Y., Rihel, J., Grocock, R.J., Van Dongen, S., Inoue, K., Enright, A.J., and Schier, A.F. (2006). Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. Science *312*, 75–79.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W., et al. (2011). The developmental transcriptome of Drosophila melanogaster. Nature *471*, 473–479.

Grzybowska, E.A. (2012). Human intronless genes: functional groups, associated diseases, evolution, and mRNA processing in absence of splicing. Biochem. Biophys. Res. Commun. *424*, 1–6.

Hamatani, T., Carter, M.G., Sharov, A.A., and Ko, M.S.H. (2004). Dynamics of global gene expression changes during mouse preimplantation development. Dev. Cell *6*, 117–131.

Harvey, S.A., Sealy, I., Kettleborough, R., Fényes, F., White, R., Stemple, D., and Smith, J.C. (2013). Identification of the zebrafish maternal and paternal transcriptomes. Development *140*, 2703–2710.

Holt, I.J., He, J., Mao, C.-C., Boyd-Kirkup, J.D., Martinsson, P., Sembongi, H., Reyes, A., and Spelbrink, J.N. (2007). Mammalian mitochondrial nucleoids: organizing an independently minded genome. Mitochondrion *7*, 311–321.

Kalinka, A.T. (2013). The probability of drawing intersections: extending the hypergeometric distribution. arXiv, arXiv:1305.0717, http://arxiv.org/abs/1305.0717.

Kalinka, A.T., and Tomancak, P. (2012). The evolution of early animal embryos: conservation or divergence? Trends Ecol. Evol. *27*, 385–393.

Kane, D.A., and Kimmel, C.B. (1993). The zebrafish midblastula transition. Development *119*, 447–456.

Lee, M.T., Bonneau, A.R., Takacs, C.M., Bazzini, A.A., DiVito, K.R., Fleming, E.S., and Giraldez, A.J. (2013). Nanog, Pou5f1 and SoxB1 activate zygotic gene expression during the maternal-to-zygotic transition. Nature *503*, 360–364.

Leichsenring, M., Maes, J., Mössner, R., Driever, W., and Onichtchouk, D. (2013). Pou5f1 transcription factor controls zygotic gene activation in vertebrates. Science *341*, 1005–1009.

Lindeman, L.C., Andersen, I.S., Reiner, A.H., Li, N., Aanes, H., Østrup, O., Winata, C., Mathavan, S., Müller, F., Aleström, P., and Collas, P. (2011). Prepatterning of developmental gene expression by modified histones before zygotic genome activation. Dev. Cell *21*, 993–1004.

Lott, S.E., Villalta, J.E., Schroth, G.P., Luo, S., Tonkin, L.A., and Eisen, M.B. (2011). Noncanonical compensation of zygotic X transcription in early Drosophila melanogaster development revealed through single-embryo RNA-seq. PLoS Biol. *9*, e1000590.

Mathavan, S., Lee, S.G.P., Mak, A., Miller, L.D., Murthy, K.R.K., Govindarajan, K.R., Tong, Y., Wu, Y.L., Lam, S.H., Yang, H., et al. (2005). Transcriptome analysis of zebrafish embryogenesis using microarrays. PLoS Genet. *1*, 260–276.

McHale, P., Mizutani, C.M., Kosman, D., MacKay, D.L., Belu, M., Hermann, A., McGinnis, W., Bier, E., and Hwa, T. (2011). Gene length may contribute to graded transcriptional responses in the Drosophila embryo. Dev. Biol. *360*, 230–240.

McKnight, S.L., and Miller, O.L., Jr. (1976). Ultrastructural patterns of RNA synthesis during early embryogenesis of Drosophila melanogaster. Cell *8*, 305–319.

Neme, R., and Tautz, D. (2013). Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. BMC Genomics *14*, 117.

Nodine, M.D., and Bartel, D.P. (2012). Maternal and paternal genomes contribute equally to the transcriptome of early plant embryos. Nature *482*, 94–97.

O'Boyle, S., Bree, R.T., McLoughlin, S., Grealy, M., and Byrnes, L. (2007). Identification of zygotic genes expressed at the midblastula transition in zebrafish. Biochem. Biophys. Res. Commun. *358*, 462–468.

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A., and Schier, A.F. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res. *22*, 577–591.

Piasecka, B., Lichocki, P., Moretti, S., Bergmann, S., and Robinson-Rechavi, M. (2013). The hourglass and the early conservation models—co-existing patterns of developmental constraints in vertebrates. PLoS Genet. *9*, e1003476.

Potok, M.E., Nix, D.A., Parnell, T.J., and Cairns, B.R. (2013). Reprogramming the maternal zebrafish genome after fertilization to match the paternal methylation pattern. Cell *153*, 759–772.

Raff, R.A. (1996). The Shape of Life: Genes, Development, and the Evolution of Animal Form (Chicago: University of Chicago Press).

Rothe, M., Pehl, M., Taubert, H., and Jäckle, H. (1992). Loss of gene function through rapid mitotic cycles in the Drosophila embryo. Nature *359*, 156–159.

Shabalina, S.A., Ogurtsov, A.Y., Spiridonov, A.N., Novichkov, P.S., Spiridonov, N.A., and Koonin, E.V. (2010). Distinct patterns of expression and evolution of intronless and intron-containing mammalian genes. Mol. Biol. Evol. *27*, 1745–1749.

Shermoen, A.W., and O'Farrell, P.H. (1991). Progression of the cell cycle through mitosis leads to abortion of nascent transcripts. Cell *67*, 303–310.

Sprague, J., Clements, D., Conlin, T., Edwards, P., Frazer, K., Schaper, K., Segerdell, E., Song, P., Sprunger, B., and Westerfield, M. (2003). The Zebrafish Information Network (ZFIN): the zebrafish model organism database. Nucleic Acids Res. *31*, 241–243.

Strzelecka, M., Trowitzsch, S., Weber, G., Lührmann, R., Oates, A.C., and Neugebauer, K.M. (2010). Coilin-dependent snRNP assembly is essential for zebrafish embryogenesis. Nat. Struct. Mol. Biol. *17*, 403–409.

Swinburne, I.A., and Silver, P.A. (2008). Intron delays and transcriptional timing during development. Dev. Cell *14*, 324–330.

Tadros, W., and Lipshitz, H.D. (2009). The maternal-to-zygotic transition: a play in two acts. Development *136*, 3033–3042.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Vesterlund, L., Jiao, H., Unneberg, P., Hovatta, O., and Kere, J. (2011). The zebrafish transcriptome during early development. BMC Dev. Biol. *11*, 30.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. *19*, 327–335.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.-Y., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y.E., et al. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature *500*, 593–597.

Zeiner, G.M., Cleary, M.D., Fouts, A.E., Meiring, C.D., Mocarski, E.S., and Boothroyd, J.C. (2008). RNA analysis by biosynthetic tagging using 4-thiouracil and uracil phosphoribosyltransferase. Methods Mol. Biol. *419*, 135–146.

Zeng, F., Baldwin, D.A., and Schultz, R.M. (2004). Transcript profiling during preimplantation mouse development. Dev. Biol. *272*, 483–496.

SUPPLEMENTAL INFORMATION

The earliest transcribed zygotic genes are short, newly evolved and different across species

Patricia Heyn, Martin Kircher, Andreas Dahl, Janet Kelso, Pavel Tomancak, Alex T. Kalinka, and Karla M. Neugebauer

Supplemental Information includes 9 figures, 12 tables, Supplemental Discussion, Supplemental Experimental Procedures and references.

Figure S1

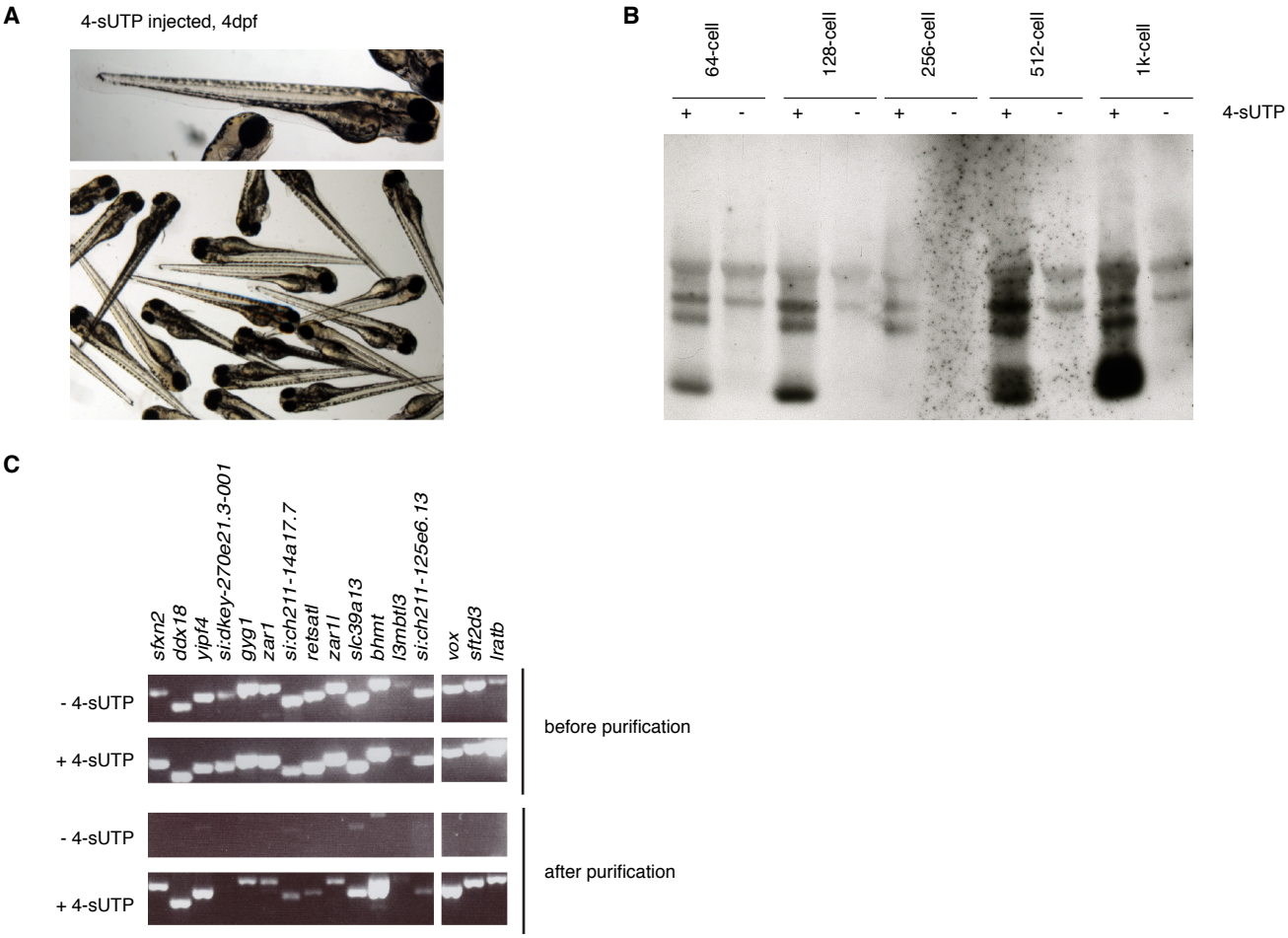

A
4-sUTP injected, 4dpf
B
64-cell
128-cell
256-cell
512-cell
1k-cell
+ - + - + - + - + -
4-sUTP
C
sfxn2
ddx18
yipf4
si:dkey-270e21.3-001
gyg1
zar1
si:ch211-14a17.7
retsatl
zar1l
slc39a13
bhmt
l3mbtl3
si:ch211-125e6.13
vox
sft2d3
lratb
- 4-sUTP
+ 4-sUTP
before purification
- 4-sUTP
+ 4-sUTP
after purification

Figure S1. Zygotically transcribed genes in zebrafish can be labeled with 4-sUTP and isolated, Related to Figure 1. (**A**) Zebrafish embryos injected at the 1-cell stage with 1nl 50 mM 4-sUTP at 4dpf. (**B**) Total RNA from 4-sUTP microinjected embryos (+4-sUTP) and uninjected wildtype control embryos (-4-sUTP) were loaded in equal amounts. The panel shows the membrane after transfer of the RNA and detection of biotinylated 4-sUTP, incorporated into RNA, with streptavidin-alkaline phosphatase. A developmental time course from 64-cell stage to 1k-cell stage is shown (indicated on top). An air-bubble impaired the transfer for the 256-cell stage (**C**) Agarose gel showing RT-PCR products from RNA isolated from 8-9hpf zebrafish embryos injected with 4-sUTP (+ 4-sUTP) or wildtype control embryos (- 4-sUTP). The upper two lanes show PCR products from the input material (before purification). The lower two lanes show PCR products from isolated newly transcribed RNA (after purification). Sixteen different genes (indicated on top) were analyzed from both known maternal and zygotically transcribed genes. Five µl PCR product were loaded per lane.

Figure S2

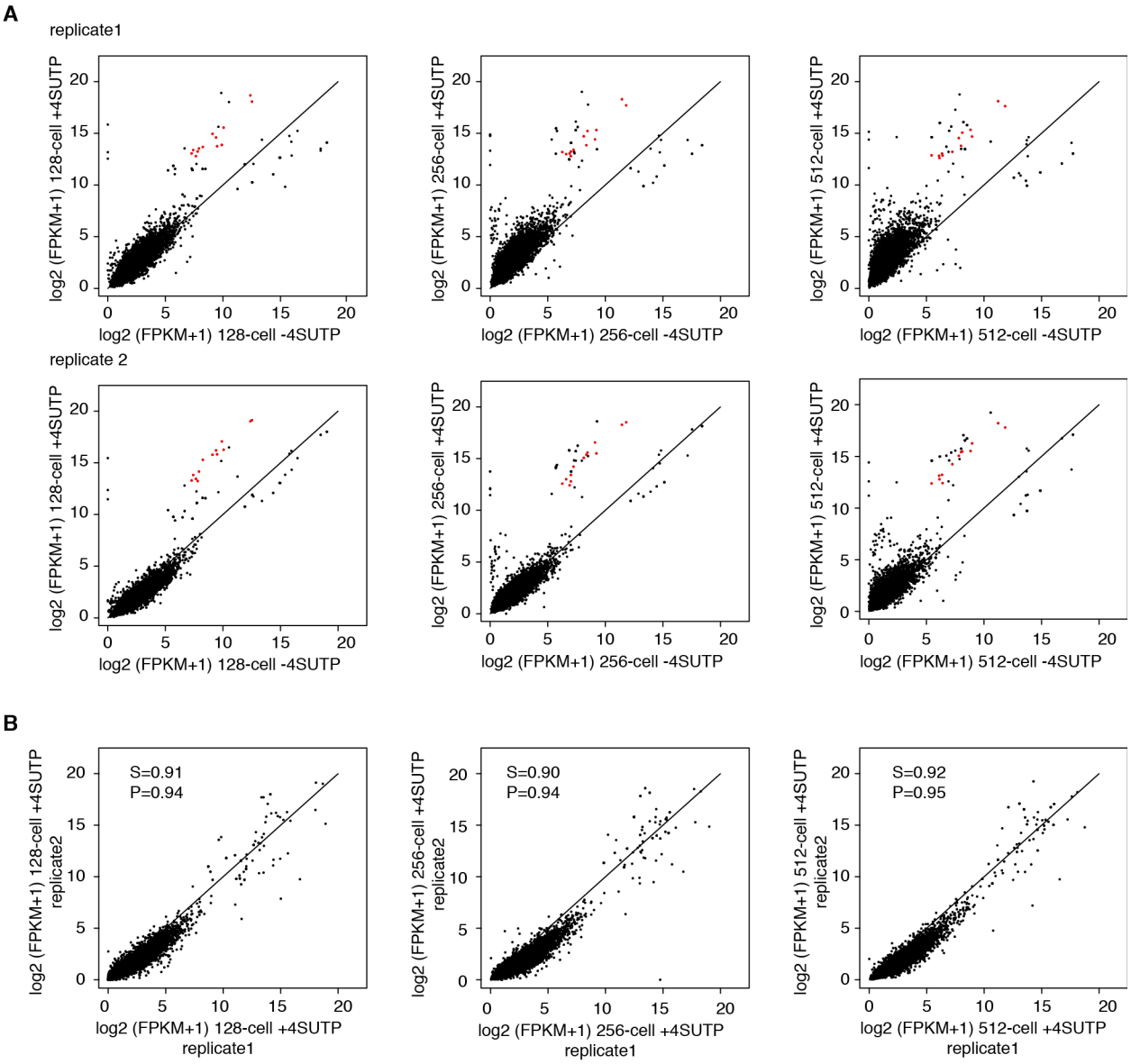

A
replicate1
log2 (FPKM+1) 128-cell +4SUTP
log2 (FPKM+1) 128-cell -4SUTP
log2 (FPKM+1) 256-cell +4SUTP
log2 (FPKM+1) 256-cell -4SUTP
log2 (FPKM+1) 512-cell +4SUTP
log2 (FPKM+1) 512-cell -4SUTP
replicate 2
B
S=0.91
P=0.94
S=0.90
P=0.94
S=0.92
P=0.95
replicate2
replicate1

Figure S2. High enrichment of mitochondrial genes over background and quality of technical replicates, Related to Figure 1. (**A**) Scatterplots of abundance for each gene identified in the deep sequencing experiment are shown. FPKM (Fragments Per Kilobase of exon per Million fragments mapped) values per gene are plotted in log2 scale from the two 4-sUTP-labeled time course replicates (+4-sUTP) on the y-axis against the FPKM values from the uninjected time course samples (-4-sUTP) on the x-axis. Each dot represents a gene. To avoid negative values in the log2 scale, 1 was added to all FPKM values before transformation. Protein-coding genes from the mitochondrial genome are highlighted in red. For all samples, the read number was sampled down to 8,775,274 reads (corresponds to the read number in the sample with the lowest coverage) before the FPKM values were calculated. (**B**) Scatterplots of abundance from the genes identified in the deep sequencing experiment are shown. FPKM values per gene are plotted in log2 scale for the two 4-sUTP time course replicates against each other for each time point (replicate 1 on x-axis and replicate 2 on y-axis). Each dot represents a gene. To avoid negative values in the log2 scale, 1 was added to all FPKM values before transformation. Inside the plots the correlation coefficients are given. Spearman correlation coefficient is indicated with S and Pearson correlation coefficient indicated with P.

Figure S3

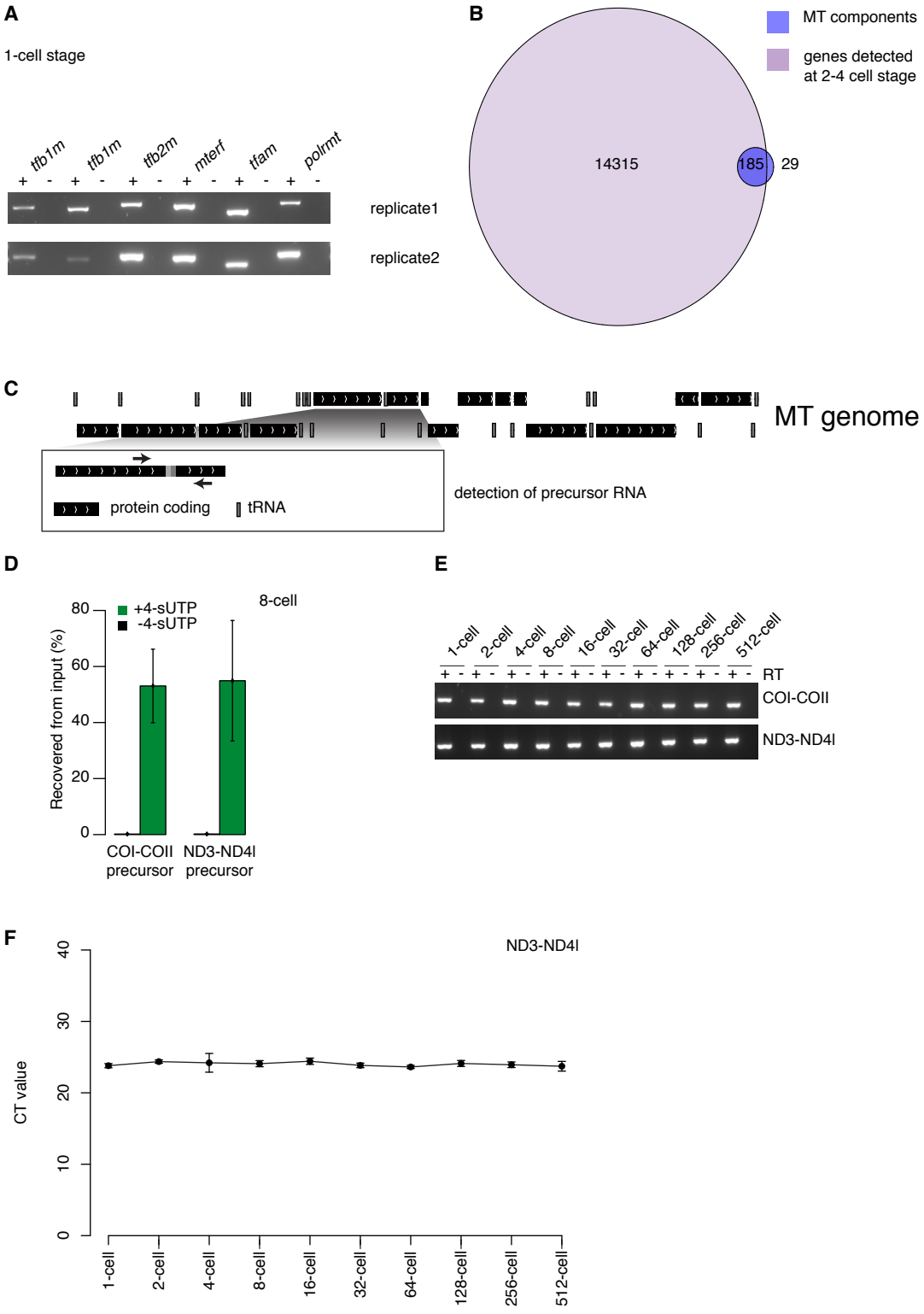

A
1-cell stage
tfb1m
tfb1m
tfb2m
mterf
tfam
polrmt
replicate1
replicate2
B
MT components
genes detected at 2-4 cell stage
14315
185
29
C
MT genome
protein coding
tRNA
detection of precursor RNA
D
8-cell
+4-sUTP
-4-sUTP
Recovered from input (%)
COI-COII precursor
ND3-ND4l precursor
E
1-cell
2-cell
4-cell
8-cell
16-cell
32-cell
64-cell
128-cell
256-cell
512-cell
RT
COI-COII
ND3-ND4l
F
ND3-ND4l
CT value

Figure S3. Mitochondrial genomes are transcriptionally active from the beginning of the development, Related to Figure 2. (**A**) RT-PCR to test for the presence of mitochondrial transcription machinery components at the 1-cell stage. RNA was reverse transcribed with oligo(dT) for two different pools of maternal RNA. (**B**) Venn diagram showing the overlap between genes expressed at the 2-4 cell stage and of genes annotated by GO-terms as part of mitochondria. (**C**) Scheme of RT-PCR method to detect precursor RNA from mitochondrial genomes (MT genomes). (**D**) Quantification of newly transcribed mitochondrial precursor RNAs by RT-qPCR from 8-cell stage zebrafish embryos. Average values for percent recovered from input are shown for two different primer-pairs detecting the mitochondrial precursor RNA (N=2). Error bars indicate SD. RNA from embryos microinjected with 4-sUTP (+4-sUTP, green) is compared to uninjected wildtype control embryos (-4-sUTP, black). (**E**) Detection of precursor RNA with two different primer pairs in a RT-PCR from the 1-cell stage to the 512-cell stage. PCR products from reactions with and without reverse transcriptase added (+ and – RT respectively) are shown. For each lane 5µl PCR product were loaded. (**F**) RT-qPCR over a developmental time course for precursor mitochondrial RNA. At each time point the RNA from exact 3 staged embryos was used. Plotted is the CT value over the time course. Error bars indicate the SD. N is 3 for 1-cell to 256-cell and 2 for 512-cell stage.

Figure S4

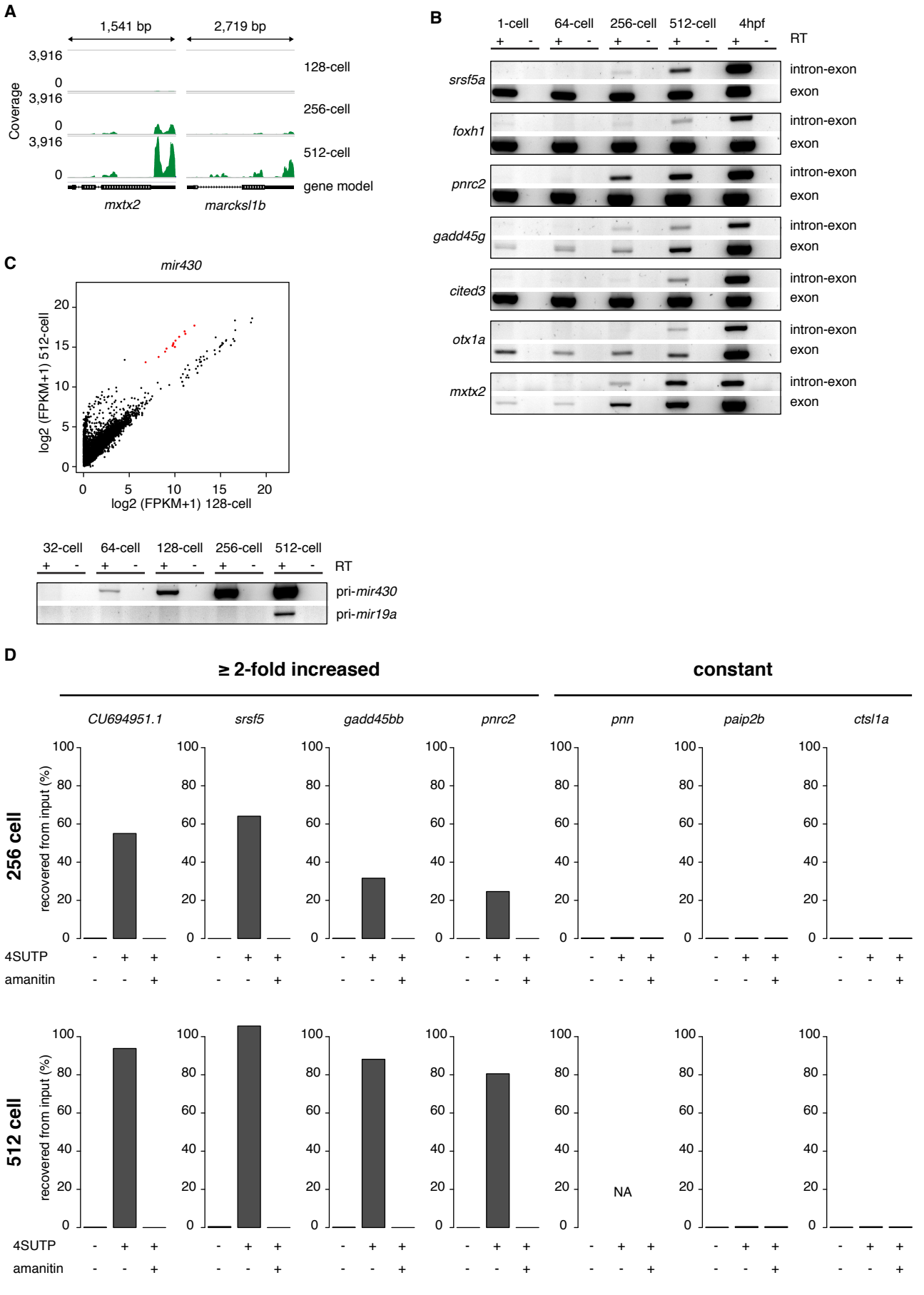
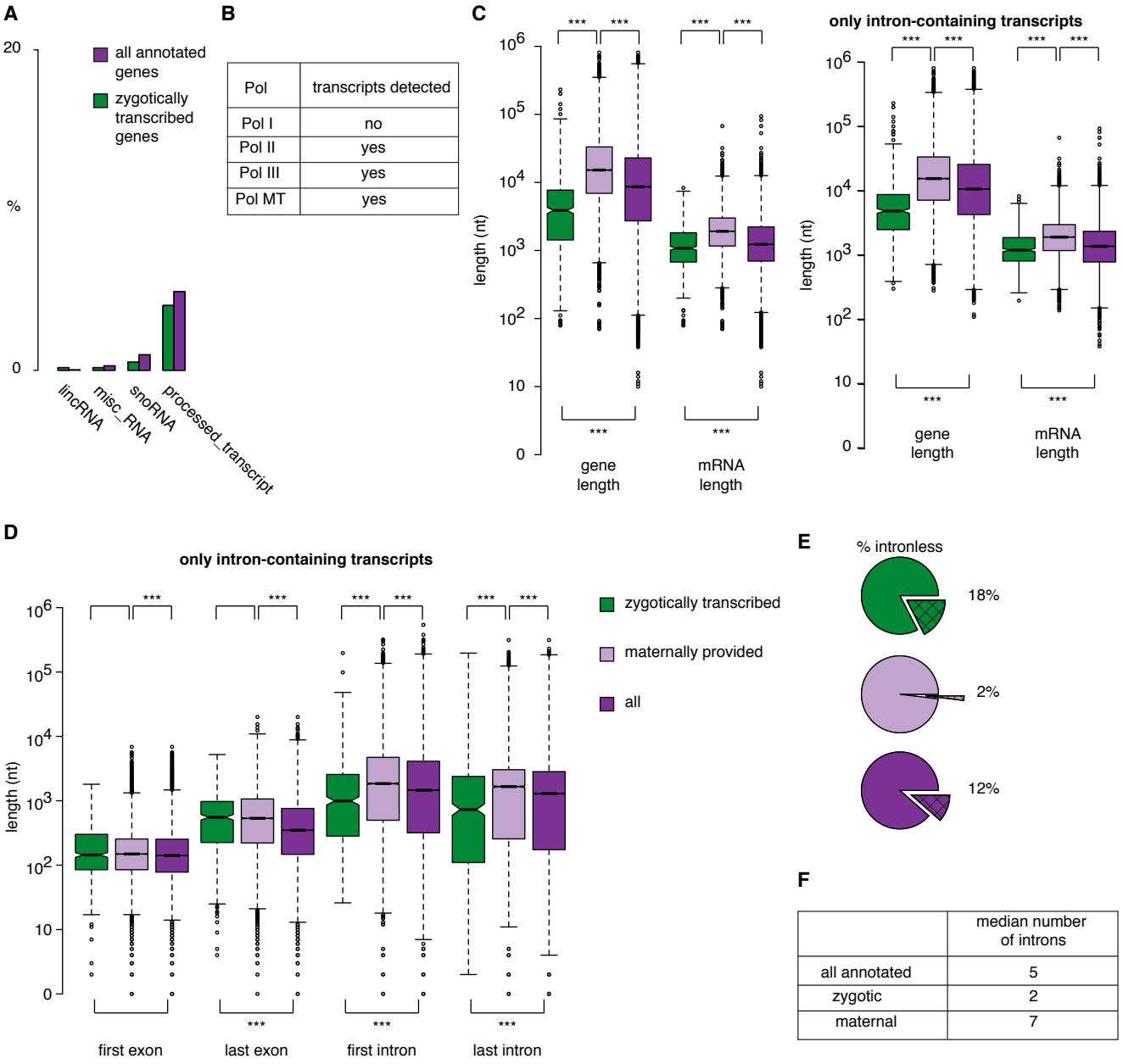

A
1,541 bp
2,719 bp
Coverage
3,916
0
3,916
0
3,916
0
128-cell
256-cell
512-cell
gene model
mxtx2
marcksl1b
B
1-cell
64-cell
256-cell
512-cell
4hpf
RT
srsf5a
foxh1
pnrc2
gadd45g
cited3
otx1a
mxtx2
intron-exon
exon
C
mir430
log2 (FPKM+1) 512-cell
log2 (FPKM+1) 128-cell
32-cell
64-cell
128-cell
256-cell
512-cell
RT
pri-mir430
pri-mir19a
D
≥ 2-fold increased
constant
CU694951.1
srsf5
gadd45bb
pnrc2
pnn
paip2b
ctsl1a
256 cell
512 cell
recovered from input (%)
4SUTP
amanitin
NA

Figure S4. Validation of deep sequencing results, Related to Figure 2. (**A**) Example of increasing read coverage over the zygotically transcribed genes (*mxtx2* and *marcksl1b*) in the three developmental time points investigated (i.e. 128-cell, 256-cell and 512-cell stage). The y-axis indicates coverage per nucleotide. The architecture of the respective genes is displayed below. (**B**) Detection of introns in newly transcribed RNA by RT-PCR over a developmental time course (stages indicated on top). Analyzed were wildtype uninjected embryos. For each gene (indicated on the left) one primer-pair was placed flanking an intron-exon boundary and one primer-pair inside of an exon. PCR products from reactions with and without reverse transcriptase added (+ and – RT respectively) are shown. For each lane 14µl PCR product were loaded. Levels of the whole image were adjusted in photoshop for visibility. (**C**) Scatterplot of abundance from the genes identified in the deep sequencing experiment are shown. FPKM values per gene are plotted in log2 scale for the 4-sUTP time course samples. The earliest investigated time point is plotted on the x-axis (128-cell stage) and the latest investigated developmental time point is plotted on the y-axis (512-cell stage). Each dot represents a gene. To avoid negative values in the log2 scale, 1 was added to all FPKM values before transformation. Genes of the *mir430* family, which are identified as zygotically transcribed, are highlighted in red. The panel underneath shows a RT-PCR detection of pri-*mir430* and pri-*mir19a* over a developmental time course (time points indicated on top). PCR products from reactions with and without reverse transcriptase added (+ and – RT respectively) are shown. For each lane 14µl PCR product were loaded. Levels of the whole image were adjusted in photoshop for visibility. (**D**) RT-qPCR for genes that were at least 2-fold increased in abundance at 512-cell stage compared to the 128-cell stage (≥2-fold increased) and genes that did not change in abundance over the time course (constant) in the deep sequencing experiment (see Figure 2C). Names of the analyzed genes are shown on top. Plotted are percent recovered from input from a single experiment at two developmental time points. Samples from uninjected wildtype control samples (- 4-sUTP, -amanitin), from 4-sUTP injected embryos (+4-sUTP, - amanitin) and from 4-sUTP and α-amanitin injected embryos (+4-sUTP, + amanitin) are compared. NA indicates missing values.

Figure S5



Pol	transcripts detected
Pol I	no
Pol II	yes
Pol III	yes
Pol MT	yes

	median number of introns
all annotated	5
zygotic	2
maternal	7

Figure S5. Characteristics of zygotically transcribed genes, Related to Figure 2 and 3. (**A**) Classes of biotypes present, but not overrepresented in zygotically transcribed genes are compared to the overall distribution in all annotated zebrafish genes. The proportions of each biotype among the zygotically transcribed genes (green bars) and among all annotated genes (purple) are plotted in percent. (**B**) Table showing each RNA polymerase in zebrafish and whether transcripts synthesized by the respective RNA polymerase were detected. (**C**) Distribution and median of gene length and mRNA length for zygotic, maternal and all annotated transcripts. Plotted is the length in nt of the gene or mRNA in a log10 scale for all detected isoforms. The panel on the right shows intron-containing transcripts only. Solid lines in the box indicate the median (zygotic = 3,874 bp, maternal = 15,202 bp, all = 8,644 bp). Significant differences of the median for the different transcript classes are indicated by stars (Mann-Whitney U test p<=0.05 *; p<=0.01 **; p<=0.001 ***). For p-values of pairwise comparison please refer to Table S6. (**D**) Distribution and median of first exon, last exon, first intron and last intron length for intron-containing transcripts among zygotic, maternal and all annotated transcripts. Plotted is the length in nt of the respective feature in a log10 scale for all intron-containing transcripts. Solid lines in the box indicate the median. Significant differences between the medians for the different transcript classes are indicated by stars (Mann-Whitney U test p<=0.05 *; p<=0.01 **; p<=0.001 ***). For p-values of pairwise comparison please refer to Table S7. (**E**) Pie charts showing the proportion of intronless transcripts among zygotic, maternal and all annotated transcripts. (**F**) Table with median number of introns per transcript for all transcripts per class.

Figure S6

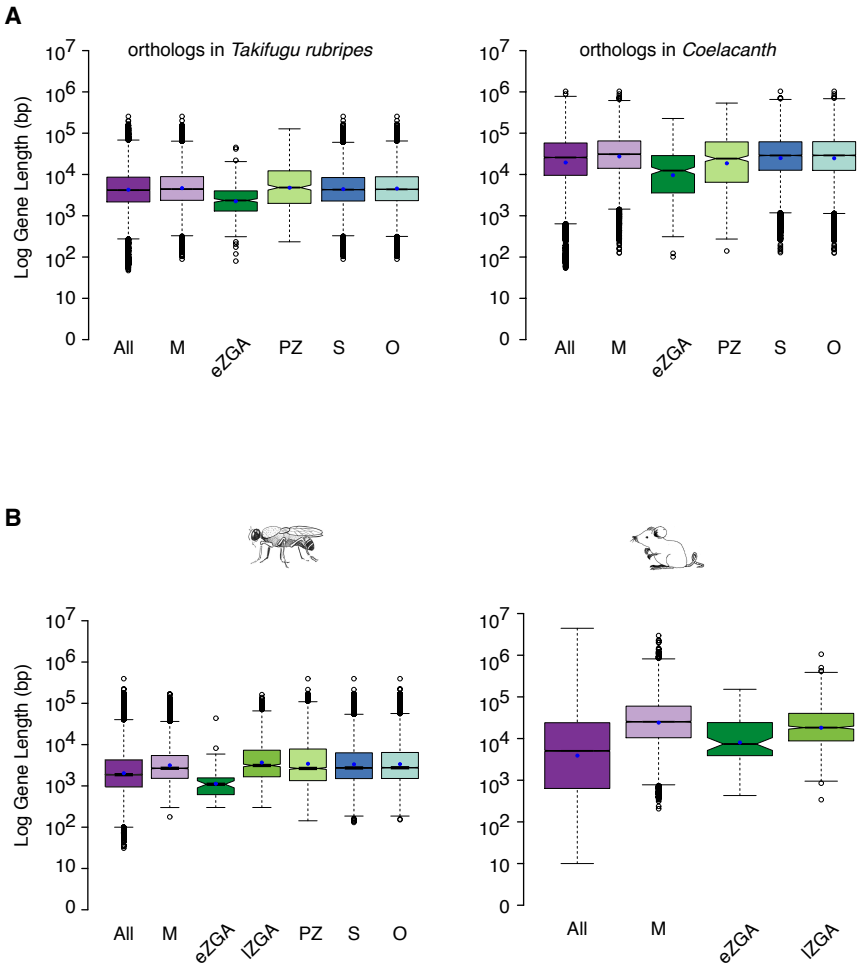

A
orthologs in Takifugu rubripes
orthologs in Coelacanth
Log Gene Length (bp)
All
M
eZGA
PZ
S
O
B
lZGA

Figure S6. Orthologs of zebrafish zygotic genes are short in *Fugu* and *Coelacanth*, Related to Figure 3. (A) Orthologs of zebrafish maternally expressed genes (M), early zygotic genes (eZGA), purely zygotic genes expressed after MBT (PZ), genes expressed at somitogenesis (S) and genes expressed at organogenesis (O) were identified in *Takifugu rubripes* and *Coelacanth* and their gene length distribution analyzed. Plotted is the gene length in nt of the gene in a log10 scale for all genes per category. For p-values of pairwise comparison please refer to Table S6. (**B**) Distribution and median of gene length for different developmental gene categories for *D. melanogaster* and *M. musculus*. Plotted is the length in nt of the gene in a log10 scale for all genes per category. The categories correspond to all= all annotated genes/ M=maternal genes/ eZGA=early zygotic genes/ lZGA= late zygotic genes/ PZ=purely zygotic genes expressed after MBT/ S=somitogenesis/segmentation / O=organogenesis.

Figure S7

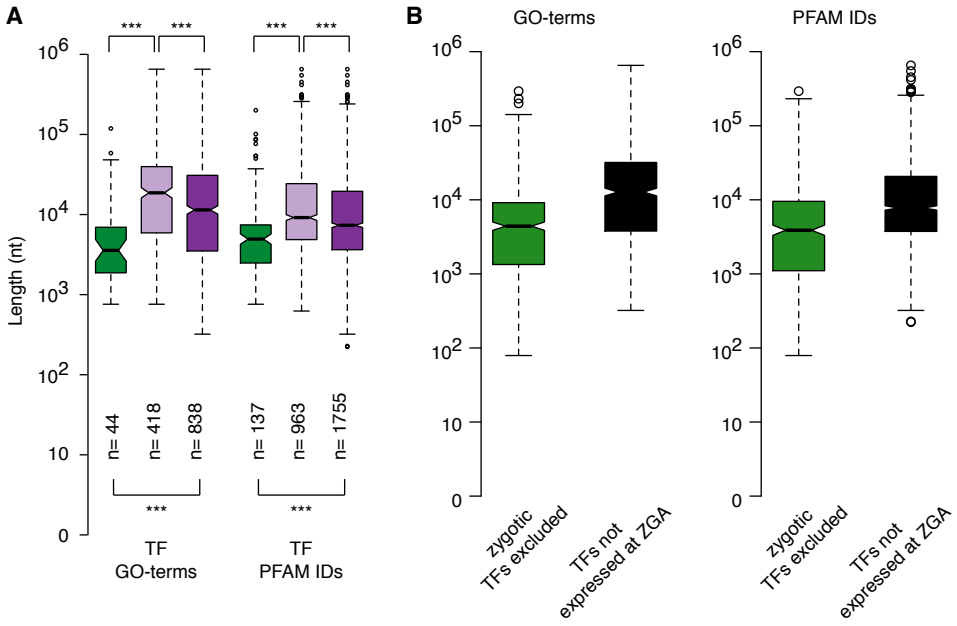

A
B
GO-terms
PFAM IDs
Length (nt)
***
n= 44
n= 418
n= 838
n= 137
n= 963
n= 1755
TF
GO-terms
TF
PFAM IDs
zygotic
TFs excluded
TFs not
expressed at ZGA

Figure S7. Zygotic transcription factors are short, Related to Figure 3. (**A**) Distribution and median of gene length for zygotically transcribed, maternally expressed or all annotated transcription factors as annotated by GO-terms or PFAM-IDs. Plotted is the length in nt of the gene in a log10 scale. Significant differences of the median for the different transcript classes are indicated by stars (Mann-Whitney U test p<=0.05 *; p<=0.01 **; p<=0.001 ***). For p-values of pairwise comparison please refer to Table S6. (**B**) Distribution and median of gene length for zygotically transcribed genes after removal of genes identified to be transcription factors (green) and for transcription factors not expressed at ZGA (black) based on GO-terms or PFAM-IDs. Plotted is the length in nt of the gene in a log10 scale. Solid lines in the box indicate the median.

Figure S8

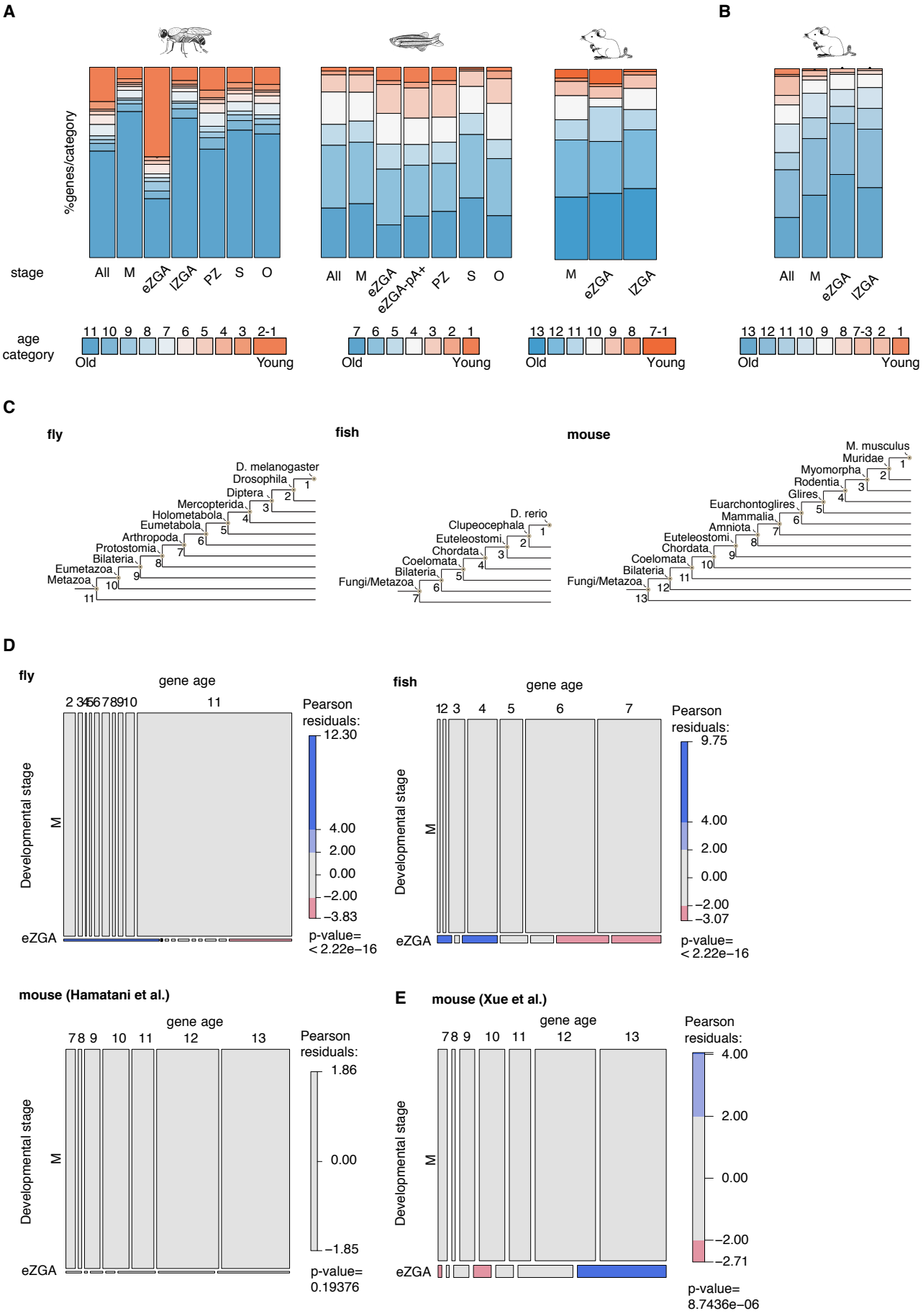

A
B
%genes/category
stage
All M eZGA IZGA PZ S O
All M eZGA eZGA-pA+ PZ S O
M eZGA IZGA
All M eZGA IZGA
age category
11 10 9 8 7 6 5 4 3 2-1
7 6 5 4 3 2 1
13 12 11 10 9 8 7-1
13 12 11 10 9 8 7-3 2 1
Old
Young
C
fly
D. melanogaster
Drosophila
Diptera
Mercopterida
Holometabola
Eumetabola
Arthropoda
Protostomia
Bilateria
Eumetazoa
Metazoa
fish
D. rerio
Clupeocephala
Euteleostomi
Chordata
Coelomata
Bilateria
Fungi/Metazoa
mouse
M. musculus
Muridae
Myomorpha
Rodentia
Glires
Euarchontoglires
Mammalia
Amniota
Euteleostomi
Chordata
Coelomata
Bilateria
Fungi/Metazoa
D
fly
gene age
2 3 4 5 6 7 8 9 10 11
Developmental stage
M
eZGA
Pearson residuals:
12.30
4.00
2.00
0.00
−2.00
−3.83
p-value= < 2.22e−16
fish
gene age
12 3 4 5 6 7
Developmental stage
M
eZGA
Pearson residuals:
9.75
4.00
2.00
0.00
−2.00
−3.07
p-value= < 2.22e−16
mouse (Hamatani et al.)
gene age
7 8 9 10 11 12 13
Developmental stage
M
eZGA
Pearson residuals:
1.86
0.00
−1.85
p-value= 0.19376
E
mouse (Xue et al.)
gene age
7 8 9 10 11 12 13
Developmental stage
M
eZGA
Pearson residuals:
4.00
2.00
0.00
−2.00
−2.71
p-value= 8.7436e−06

Figure S8. Protein-coding early zygotic genes from fly and fish are evolutionary younger compared to all other developmental stages, Related to Figure 3. (**A**) Stacked bar graphs showing the normalized distribution of the inferred age of the genes expressed at the respective developmental stage in the 3 species (fly, fish and mouse). Maternal (M) and zygotic (eZGA, lZGA, PZ) and to control for genes not detected in polyA(+) datasets (eZGA-poly(A)+) stages are compared with all annotated genes (all) and genes expressed at later developmental stages (somitogenesis/segmentation (S) and organogenesis (O)). Gene age is retrieved from Ensembl protein-trees depicted schematically in panel S8C. Different colors correspond to numbers depicted underneath the bar graph and indicate the node (clade) in the protein-tree in which the genes emerged, where smaller numbers represent more recent emergence. Please note that "young" and "old" are used as relative terms. (**B**) Stacked bar graphs showing the normalized distribution of the inferred age of the genes expressed at the respective developmental stage in the mouse. Zygotic (eZGA) are derived from Xue et al. 2013 and compared with all annotated genes (all), maternal (M) and zygotic genes at later stages (lZGA) (Xue et al., 2013). Gene age is retrieved from Ensembl protein-trees depicted schematically in panel S8C. Different colors correspond to numbers depicted underneath the bar graph and indicate the node (clade) in the protein-tree in which the genes emerged, where smaller numbers represent more recent emergence. Please note that "young" and "old" are used as relative terms. (**C**) Schematic of the nodes in the protein-coding trees used for age analysis and the corresponding names of the clades representing the nodes. (**D**) Gene age tests to determine if early zygotic genes are significantly younger than maternally expressed genes. Mosaic plots for the three species are shown. Blue indicates an excess of genes in a particular age category (more than expected given the sums in the rows and columns) and red indicates a paucity of genes. The P-value indicates if the two variables (gene age and developmental stage) are independent (Chi-squared test). Fish and fly both fail the test, but mouse passes showing that there is a strong interaction between age and stage for these two species (fish and fly). (**E**) Gene age tests to determine if early zygotic mouse genes identified in Xue et al. 2013 are significantly younger than maternally expressed genes (Xue et al., 2013). A mosaic plot is shown. Blue indicates an excess of genes in a particular age category (more than expected given the sums in the rows and columns) and red indicates a paucity of genes. The P-value indicates if the two variables (gene age and developmental stage) are independent (Chi-squared test). The test indicates an excess of older genes among early zygotic genes, which could indicate that the differential gene expression analysis identified false positive zygotic genes (see also supplemental text).

Figure S9

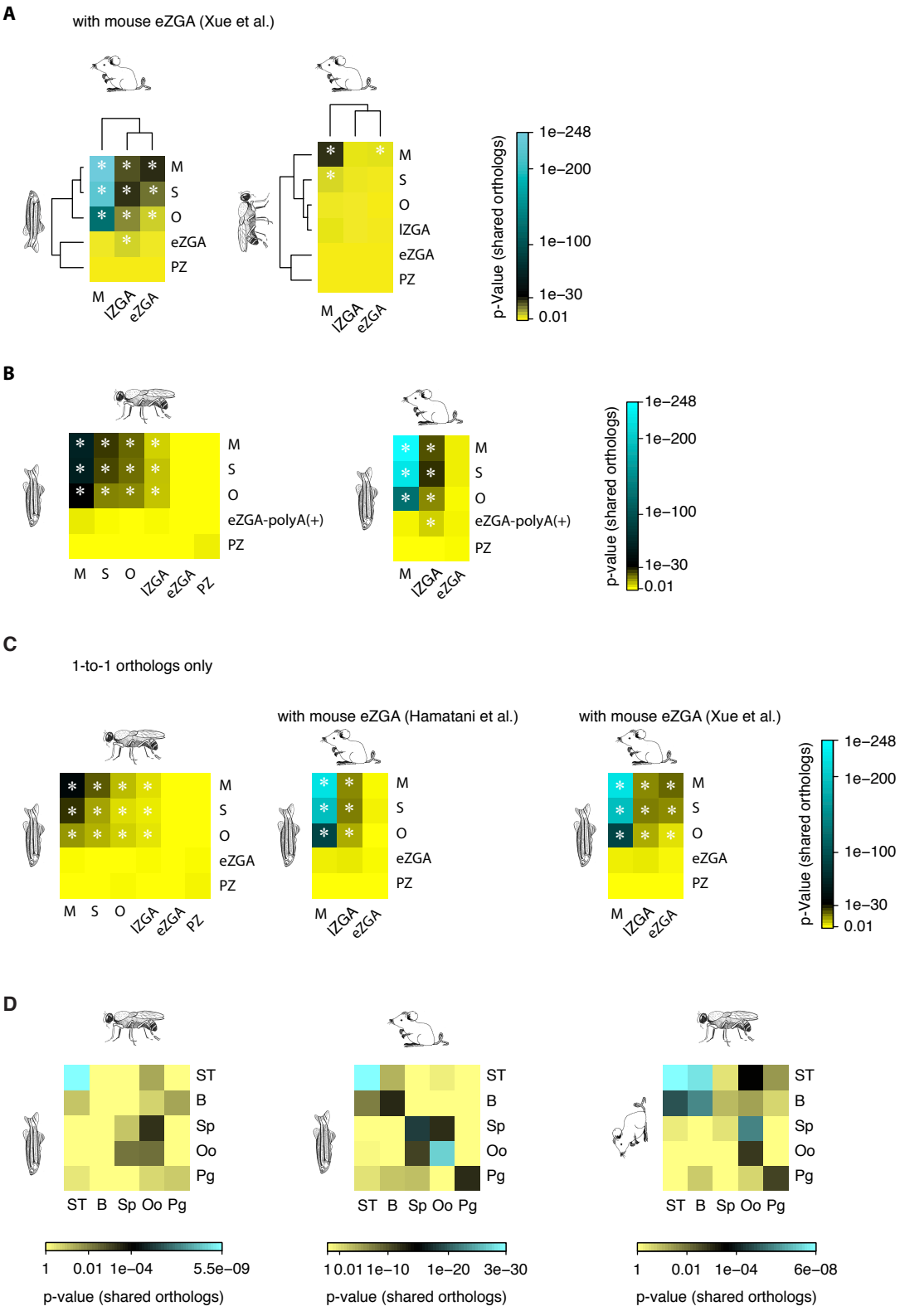

A
with mouse eZGA (Xue et al.)
M
S
O
eZGA
PZ
M IZGA eZGA
M
S
O
IZGA
eZGA
PZ
M IZGA eZGA
p-Value (shared orthologs)
1e−248
1e−200
1e−100
1e−30
0.01
B
M
S
O
eZGA-polyA(+)
PZ
M S O IZGA eZGA PZ
M
S
O
eZGA-polyA(+)
PZ
M IZGA eZGA
p-value (shared orthologs)
1e−248
1e−200
1e−100
1e−30
0.01
C
1-to-1 orthologs only
M
S
O
eZGA
PZ
M S O IZGA eZGA PZ
with mouse eZGA (Hamatani et al.)
M
S
O
eZGA
PZ
M IZGA eZGA
with mouse eZGA (Xue et al.)
M
S
O
eZGA
PZ
M IZGA eZGA
p-Value (shared orthologs)
1e−248
1e−200
1e−100
1e−30
0.01
D
ST
B
Sp
Oo
Pg
ST B Sp Oo Pg
1 0.01 1e−04 5.5e−09
p-value (shared orthologs)
ST
B
Sp
Oo
Pg
ST B Sp Oo Pg
1 0.01 1e−10 1e−20 3e−30
p-value (shared orthologs)
ST
B
Sp
Oo
Pg
ST B Sp Oo Pg
1 0.01 1e−04 6e−08
p-value (shared orthologs)

Figure S9. Early zygotic genes have fewest shared orthologs compared to all other developmental stages, Related to Figure 4. (**A**) Clustered heatmaps for pairwise ortholog comparisons between species (fly, fish and mouse) using the mouse eZGA genes identified in Xue et al. 2013 (Xue et al., 2013). Depicted are color-coded p-values describing the significance of shared ortholog enrichment between species. Enrichments were considered significant with a p-value smaller than the Bonferroni cutoff of 0.05/63= 0.00079 indicated by white asterisks. (**B**) Heatmaps for pairwise ortholog comparisons between species (fly, fish and mouse) at and between respective developmental stages. Depicted are color-coded p-values describing the significance of shared ortholog enrichment between species. eZGA-polyA(+) dataset contains only genes with polyA tails. Enrichments were considered significant with a p-value smaller than the Bonferroni cutoff of 0.05/63= 0.00079 indicated by white asterisks. (**C**) Heatmaps for pairwise ortholog comparisons between species (fly, fish and mouse) considering only 1:1 orthologs. Both sets of eZGA genes identified in Hamatani et al. 2004 and in Xue et al. 2013 are shown (Hamatani et al., 2004; Xue et al., 2013). Depicted are color-coded p-values describing the significance of shared ortholog enrichment between species. Enrichments were considered significant with a p-value smaller than the Bonferroni cutoff of 0.05/63= 0.00079 indicated by white asterisks. (**D**) Heatmaps for pairwise ortholog comparisons between species (fly, fish and mouse) for different adult traits (synaptic transmission (ST), behavior (B), spermatogenesis (Sp), oogenesis (Oo), pigmentation (PG)). Depicted are color-coded p-values describing the significance of shared ortholog enrichment between species. P-value scales are different for each heatmap. For fly-fish and mouse-fish comparison 1:2 orthologs were considered, to account for the whole genome duplication in teleosts. For fly-mouse comparison, 1:1 orthologs were considered. For gene group size and p-values of pairwise comparisons please see Table S8.

Supplemental Tables:

Table S1. FPKM values for zygotically expressed genes, Related to Figure 2. Table S1 shows Fragments Per Kilobase of exon per Million fragments mapped (FPKM) levels for all identified zygotically expressed zebrafish genes (rows) for 3 different developmental stages (columns).

Table S2. FPKM values for zygotically expressed transcripts, Related to Figure 2. Table S2 shows Fragments Per Kilobase of exon per Million fragments mapped (FPKM) levels for all identified zygotically expressed zebrafish transcripts (rows) for 3 different developmental stages (columns).

Table S3. P-values for biotype enrichment analysis, Related to Figure 2. Table S3 contains the p-values of the hypergeometric test for biotype enrichment analysis.

Table S4. Significantly enriched GO-terms, Related to Figure 3. Table S4 contains all significant GO-terms for zygotically expressed and maternally deposited genes for the three GO-domains as well as the p-values.

Table S5. Significantly enriched PFAM-IDs, Related to Figure 3. Table S5 contains all significant PFAM-IDs and the Bonferroni corrected p-values.

Table S6. Median values of gene lengths, Related to Figure 3. Table S6 contains median values of gene lengths for *D. rerio, D. melanogaster, M. musculus, T. rubripes* and *Coelacanth* genes and *D. rerio* transcription factors, as well as p-values for pairwise comparisons of the medians (Mann-Whitney U test).

Table S7. Median values of gene-architecture features, Related to Figure 3. Table S7 contains median values of gene-architecture features for zygotic, maternal and all annotated *D. rerio* transcripts, as well as p-values for pairwise comparisons of the medians (Mann-Whitney U test).

Table S8. P-values for the enrichment of shared orthologs, Related to Figure 4. Table S8 contains p-values for the enrichment of shared orthologs between *D. rerio*, *D. melanogaster* and *M. musculus* for developmental stages and controls (eZGApolyA(+) only, 1:1 orthologs only and adult traits).

Table S9. Percentages of shared orthologs, Related to Figure 4. Table S9 contains numbers of genes expressed relative to the total numbers of genes expressed for both orthologs and non-orthologs.

Table S10. P-values for the enrichment of essential genes, Related to Figure 4. Table S10 contains p-values for the enrichment of essential genes.

Table S11. Overview of publically available datasets used, Related to Figure 3 and 4. Table S11 gives an overview of publically available datasets used in this study and their designation.

Table S12. PCR and qPCR primer pairs, Related to Figure 1 and Figures S1, S3 and S4. Table S12 contains PCR and qPCR primer pairs used for development of the method and deep sequencing validation.

SUPPLEMENTAL DISCUSSION

Identification of zygotically transcribed genes

To identify zygotically transcribed genes we reasoned that metabolically labeled and directly isolated transcripts of actively transcribed genes should increase in abundance – a hallmark of transcription. Without direct labeling and isolation this accumulation can be masked by complex post-transcriptional regulation taking place in the embryo (Aanes et al., 2011; De Renzis et al., 2007) leading to a balance of production and degradation of the transcript. Simple time course experiments, utilizing microarray technology or RNAseq, were performed before to characterize the embryonic transcriptome and identify zygotically transcribed genes (Aanes et al., 2011; Baldessari et al., 2005; Graveley et al., 2011; Hamatani et al., 2004; Mathavan et al., 2005; Tan et al., 2013; Vesterlund et al., 2011; Zeng et al., 2004).Transcriptionally active genes were identified by subtractive methods, based on the observation of increased expression levels in later developmental stages as compared to expression levels in the maternal RNA pool. For the vertebrate model zebrafish, one of those studies claimed early transcription for 125 genes from the 64-cell stage onwards (Mathavan et al., 2005). However, gene-lists produced in those experiments are unreliable, since studies showed that analysis of steady-state transcriptomes can be misleading during early embryo development (De Renzis et al., 2007; Robert, 2010). In particular, complex post-transcriptional regulation of maternally deposited RNAs can lead to an artificial apparent increase in transcript abundance over time. Indeed, the early transcription was not confirmed in another study from the same group (Aanes et al., 2011). Further, it was recently demonstrated that the underlying assumption for expression data normalization that total RNA amounts produced per cell are similar between sources could be wrong (Lovén et al., 2012). Certainly, this assumption does not hold in a developing organism, where maternally deposited RNA is massively degraded and transcription of a large number of genes is newly activated at the ZGA. However, since we can prove that mitochondrial RNAs are produced and detected at a constant level over the investigated time course (Figures S3C-F) we can use them to calibrate and normalize the abundance of the detected genes/transcripts over the developmental time course.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Zebrafish.

Zebrafish WT AB strain was maintained and raised under standard conditions. Embryos were microinjected at the 1-cell stage with 1 nl of 50 mM 4-thio-UTP in 10 mM Tris-HCl pH 7.4 (Ambion, Trilink) or 2 nl of 25mM 4-thio-UTP with α-amanitin (0.2mg/ml) and allowed to develop until the desired stage at 28°C. Embryos were precisely staged under a stereomicroscope as described before (Kimmel et al., 1995).

Metabolic labeling and zygotic RNA enrichment.

Staged embryos were quickly washed 3-times with PCR-grade water, homogenized in Trizol and total RNA was extracted according to manufacturers instructions. Residual DNA was removed with TurboDNAse. Subsequently RNA was extracted twice with acidic Phenol-Chloroform to remove proteins. Biotinylation of 4-sUTP and subsequent isolation was done as described before (Zeiner et al., 2008) with minor modifications. In brief, per developmental stage an aliquot of 10 µg total RNA from 125 pooled embryos was biotinylated and precipitated to recover RNAs as recommended (Zeiner et al., 2008). For Biotin-detection in Northern blots the BrightStar BioDetect Kit (Ambion) was used. Biotinylated RNA was purified with 50 µl magnetic streptavidin-coated beads (Dynabeads MyOne Streptavidin C1, Invitrogen) according to manufacturer instructions. Bioinylated RNA was bound to beads under rotation in B&W buffer (5mM Tris-HCl pH7.5/ 0.5mM EDTA/ 1M NaCl) supplemented with 0.025% Tween20 at room temperature (RT) for 30 min. Supernatant was removed and beads were washed 5 times with B&W buffer supplemented with 0.025% Tween20 at 50°C at 1400 rpm. RNA was eluted with 50 µl 5% β-mercaptoethanol (BME) for 5 min at RT and in a 2nd elution with 50 µl 5% BME for 10 min at 50°C. RNA was precipitated and processed immediately.

RT-PCR and qPCR.

Labeled (4-sUTP) or unlabeled RNA was converted into cDNA with SSIII including a no Reverse Transcriptase control. Conventional PCR was carried out with Phusion polymerase (Biozyme). For quantification of RNA quantitative RT-PCRs were carried out on the MxPro3000, Stratagene system with the Absolute qPCR SYBR Green Mix. All primers used (listed in Table S12) were optimized for a working concentration of 100% (+/- 5%) efficiency of amplification. To calculate the amount of RNA recovered from input in the isolation of 4-thio-UTP labeled RNA, the following equation was used: $\%input=100*2^{(Ctinput-Ctsample)}$ where Ct_{input} was corrected for dilutions used. For data display the mean and standard deviations were calculated with R and plotted with R (R Development Core Team, 2011).

RNAseq library preparation.

Isolated metabolically labeled RNA was amplified in a sub-exponential manner with the WT-Ovation™ One-Direct RNA Amplification System (NuGEN). The reaction was stopped after the SPIA amplification. To generate double-stranded cDNA, 1µg of amplified material was converted with random Hexamers (Roche). After purification double-stranded cDNA was sheared with the Covaris S2 system to ~200-300 bp fragment size. The cDNA was end-repaired with the NEBNext End Repair Module (NEB) and subsequently purified with the MinElute PCR purification Kit (Qiagen). The resulting fragments were adenylated at the 3'-end with the NEBNext dA-Tailing Module (NEB) and purified with the MinElute PCR purification Kit (Qiagen). To convert the cDNA into an Illumina deep sequencing library, TruSeq adapters were ligated. At first universal short adapters were ligated (sequences see Table S12; P5short and P7short). In a subsequent PCR the full length adapter was generated and a barcode introduced for multiplexing different libraries and sequencing them at once (sequences see Table S12; P5, P7 and indexingprimer). The concentration of molecules with adapters ligated in the final library was determined by qPCR with the KAPA Library Quant Kit (Kapa Biosystems). Nine different libraries (two technical replicates and one background sample for each of the 3 stages investigated) were prepared and sequenced on an Illumina HiSeq2000 machine as single-end reads of 50- or 75-bp.

RNAseq primary data analysis.

Raw reads were demultiplexed and FastQ converted with Illumina OLB 1.9 / CASAVA 1.7 and quality assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Over-represented sequences originating from adapters and SPIA-amplification primers were identified from the FastQ output and then trimmed (Kircher, 2011). Remaining sequences were mapped with TopHat 1.3.3 (Trapnell et al., 2009) to the zebrafish genome assembly Zv9/GCA_000002035.2. We used a sensitive search for spliced reads, defined a non-strand-sensitive RNAseq library preparation protocol, allowed up to one mismatch of the read sequence with the canonical splice site bases and reported up to 100 equally good mappings per read with parameters: --butterfly-search --library-type=fr-unstranded --splice-mismatches=1 --max-multihits=100. Expression of genes and transcripts of the Ensembl release 64 were quantified using Cufflinks 2.0.2 (Trapnell et al., 2010) with multi-read correction, increased maximum number of read fragments per cufflinks transcribed bundle for a non-strand-sensitive RNAseq library preparation protocol by the following parameters: --multi-read-correct --max-bundle-frags=20000000 --library-type=fr-unstranded. After initial inspection of the correlation of the technical replicates (Spearman rho 0.90-0.92), we pooled reads of the two replicates per developmental time point and then remapped, sampled to the same number of reads between developmental time points and re-quantified the pooled data.

Publically available deep sequencing sets for poly(A)+ mRNA for *D. rerio* (SRR372787, SRR372788, SRX107392, SRX107393, SRX107394, SRX107395 (Pauli et al., 2012) and SRR062661 (Aanes et al., 2011) and *D. melanogaster* (SRX008015, SRX008010, SRX008005, SRX008006) (Graveley et al., 2011) were analyzed as

described for the RNAseq data generated in this study. Fly sequences were mapped to the fly genome assembly BDGP 5 and gene/transcript annotations retrieved from Ensembl release 69. For analysis using the early zygotic genes from *M. musculus* (Figure S9A), the list of differentially expressed genes as identified in Xue et al. 2013 was used (personal communication).

Analysis of mitochondrial transcript levels over a developmental time course.

For each time point three staged, dechorionated embryos were collected into 1.5ml reaction tubes and 250µl Trizol (Invitrogen) was added and was stored at -80°C until further processing. RNA was extracted with Trizol according to manufacturers instructions. The samples were subsequently treated with TurboDNase to remove DNA. After DNase treatment the samples were extracted with acidic Phenol-Chloroform and precipitated. The final precipitate was air dried and dissolved in 10µl DEPC-H2O. Care was taken to process always the same volume in order to be able to compare the samples. Each sample was split in half and each half was used for reverse transcription with or without reverse transcriptase respectively. For each sample 1µl of a 1:10 cDNA dilution was used for PCR and qPCR.

Analysis of maternal mitochondrial components.

We identified all zebrafish genes annotated with the GO-term "Mitochondrion" (GO:0005739) as annotated in Ensembl release 64 and computationally determined whether they are expressed in a publicly available maternal RNAseq dataset (Pauli et al., 2012).

Normalization of RNAseq data.

The abundance of genes or transcripts was reported in Fragments Per Kilobase of exon per Million fragments mapped (FPKM) and only those transcripts or genes which had the cufflinks flag status OK and with a FPKM value larger than the width of the 95% confidence interval (FPKM > Δ95 FPKM) considered to be expressed (Nagaraj et al., 2011). To normalize between different time points in our study, the sum of FPKM values for protein-coding mitochondrial transcripts/genes was assumed to be constant. Downstream analysis was performed in R (R Development Core Team, 2011) based on log2-transformed FPKM values. To avoid negative values in the log2 scale, 1 was added to all FPKM values before transformation. A gene or transcript is considered differentially expressed, if the log2 fold-change exceeded 2 between 128-cell and 512-cell stage.

Functional analysis (Gene Ontology, PFAM ID enrichment, Biotype analysis).

Gene Ontology (GO) analysis was performed using the R package topGO version 2.4.0 (Alexa et al., 2006). Enrichment of GO-terms was tested with the "weight" algorithm. The p-value was determined with Fisher's exact test. GO term annotations,

PFAM IDs and Ensembl biotypes were retrieved from Ensembl release 64 Biomart. Enrichments (PFAM domains and biotypes) were tested with the hypergeometric test. The p-values for PFAM ID enrichments are Bonferroni corrected.

Gene architecture.

Gene architecture features for all annotated transcripts from zebrafish were extracted from a gtf file (Ensembl 64) with a custom script. Statistical significance was tested pairwise with an unpaired Wilcoxon rank sum test (Mann Whitney U). Gene lengths for other species (*D. melanogaster*, *M. musculus*, *T. rubripes* and *Coelacanth*) were retrieved from Ensembl Biomart.

Gene age analysis.

The age of each gene was determined by the age of the oldest node in its protein phylogeny (Piasecka et al., 2013). Protein trees, generated by EnsemblCompara (Vilella et al., 2009), a phylogeny-aware, gene tree building pipeline based on BLASTP, were downloaded from Ensembl for mouse and fish, and from Ensembl Metazoa for the fly. Genes were then placed into age categories separately for each species. Gene ages cannot be compared across species since there are differing numbers of age categories in each species; hence, statements about gene age are always made in relation to different stages in the same species. In other words, we ask whether genes tend to be younger or older in two stages in the same species, not whether genes are absolutely young or old across species.

To test whether sample size biases might be responsible for differences in gene age between maternal and eZGA, we sub-sampled the number of genes at eZGA from the maternal set for each species separately and asked whether the median of this reduced maternal set was equal to or lower than that of the observed eZGA median gene age. In 100,000 replicates, no single sub-sample had a median gene age equal to or lower than eZGA, which in both fly and fish is 5. Maternal and eZGA median gene ages are equal in the mouse (median gene age = 12). In addition, we tested whether gene age and developmental stage (maternal and eZGA) are independent variables for all three species and display the results as mosaic plots (Meyer et al., 2006) (Figure S8D and S8E).

Ortholog analysis.

To test for enrichment of orthologs expressed at different developmental stages for the fish relative to the mouse and the fly separately, we downloaded all homologs for each comparison from Ensembl and filtered these to allow only one-to-one and one-to-two orthologs (1:2 fly:fish, mouse:fish) because of the whole genome duplication in teleosts. For fly-mouse comparison, 1:1 orthologs were considered. Note that shared expression of histone transcripts could not be detected by this analysis, because the fly and mouse datasets used were generated using polyA+ selection. Please note further that enrichment of shared orthologs does not imply conservation of DNA or protein sequence. Enrichment of shared orthologs also does not imply that two species share a large fraction

of the total number of genes they express at a given stage since we restrict our tests of enrichment to expressed genes that are orthologs (percentages of shared orthologs relative to the total numbers of genes expressed – both orthologs and non-orthologs - are given in Table S9). Hence, it is possible for two species to both express a large fraction of orthologs at a given stage, but to have a small overlap between these orthologs (a small number of shared orthologs), and vice versa.

Tests of enrichment were based on a variant of the hypergeometric distribution (Kalinka, 2013) that describes the probability of drawing intersections (υ) of differing size, given that there are n orthologs, when sampling from two separate urns independently and when one of the urns contains duplicates in q of the n ortholog categories:

$$P(X = \upsilon) = \sum_{m=0}^{\alpha} \sum_{l=0}^{\beta} \sum_{j=0}^{l} \binom{n-q}{\upsilon-l} \binom{q}{l} \binom{q-l}{m} \binom{n-\upsilon-q+l}{a-\upsilon-m} \binom{l}{j} \binom{n+q-a-m-j}{b-\upsilon} \Big/ \binom{n}{a} \binom{n+q}{b}$$

A derivation and a detailed description of this distribution is provided in (Kalinka, 2013). When sampling from 1:1 orthologs (flymouse), the distribution of intersection sizes follows the classic hypergeometric distribution (Kalinka, 2013). All P-values are one-tailed tests for enrichment of shared orthologs. Functions for calculating these probabilities are available in the R package 'hint' (http://cran.r-project.org/web/packages/hint/index.html).

Hierarchical clustering was applied to P-values in the heat maps (Figure 4) using the Canberra distance and a joint between-within clustering metric (Szekely and Rizzo, 2005). Genes associated with adult traits (synaptic transmission (ST), behavior (B), spermatogenesis (Sp), oogenesis (Oo), pigmentation (PG)) used for Figure S9D were retrieved from GO-term annotations.

We used two different datasets for early zygotic genes from *M. musculus* identified by differential gene expression (Hamatani et al., 2004; Xue et al., 2013). The gene set used for the analysis in the main Figure 4 (Hamatani et al., 2004) was validated by experiments blocking transcription and is considered to be a high quality dataset. For control reasons, we used a second dataset (shown in Figure S9). This second dataset was generated with more recent technology, however not verified by transcription block experiments and therefore might contain potential false positively identified zygotic genes. This is also indicated by a poor overlap of these two dataset, with only 2 out of 66 possible genes overlapping.

Analysis of publicly available microarray data.

Data for *D. melanogaster* maternal, early and late zygotic genes was retrieved from (De Renzis et al., 2007). Gene IDs were converted to the latest versions. Data for *M. musculus* maternal, early and late zygotic genes (used in Figure 4) were retrieved from (Hamatani et al., 2004; Zeng et al., 2004). Probes were BLASTed against the mouse RefSeq RNA database using the NCBI online BLAST tool (http://blast.ncbi.nlm.nih.gov/Blast.cgi), and we recovered 66 genes from the 98 early transcripts given in the original microarray study (Hamatani et al., 2004).

Analysis of transcription factor gene length.

We identified transcription factors either by using GO-ontology ("sequence-specific DNA binding transcription factor activity"; GO:0003700) or PFAM-domains (zf-C2H2/PF00096, bHLH/PF00010, zf-C4/PF00105, SRF-TF/PF00319, bZIP_1/PF00170, IRF/PF00605, Homeobox/PF00046, HTH_11/PF08279).

Gene essentiality analysis.

Known essential genes were downloaded from the OGEE database (Chen et al., 2012) (http://ogeedb.embl.de/) for each species, and enrichment of essential genes was tested using a one-tailed hypergeometric test (Table S10).

Statistical analyses and graphics.

The R statistical computing base package was used for analysis and generation of graphics (R Development Core Team, 2011). Read density plots were generated with the count function of IGV 2.2 (bin size of 1) (Thorvaldsdóttir et al., 2013).

SUPPLEMENTAL REFERENCES

Aanes, H., Winata, C.L., Lin, C.H., Chen, J.P., Srinivasan, K.G., Lee, S.G.P., Lim, A.Y.M., Hajan, H.S., Collas, P., Bourque, G., Gong, Z., Korzh, V., Alestrom, P., Mathavan, S., 2011. Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. Genome Research 21, 1328–1338.

Alexa, A., Rahnenfuhrer, J., Lengauer, T., 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22, 1600–1607.

Baldessari, D., Shin, Y., Krebs, O., König, R., Koide, T., Vinayagam, A., Fenger, U., Mochii, M., Terasaka, C., Kitayama, A., Peiffer, D., Ueno, N., Eils, R., Cho, K.W., Niehrs, C., 2005. Global gene expression profiling and cluster analysis in Xenopus laevis. Mech. Dev. 122, 441–475.

Chen, W.-H., Trachana, K., Lercher, M.J., Bork, P., 2012. Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. Mol. Biol. Evol. 29, 1703–1706.

De Renzis, S., Elemento, O., Tavazoie, S., Wieschaus, E.F., 2007. Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo. Plos Biol 5, e117.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W., Brown, J.B., Cherbas, L., Davis, C.A., Dobin, A., Li, R., Lin, W., Malone, J.H., Mattiuzzo, N.R., Miller, D., Sturgill, D., Tuch, B.B., Zaleski, C., Zhang, D., Blanchette, M., Dudoit, S., Eads, B., Green, R.E., Hammonds, A., Jiang, L., Kapranov, P., Langton, L., Perrimon, N., Sandler, J.E., Wan, K.H., Willingham, A., Zhang, Y., Zou, Y., Andrews, J., Bickel, P.J., Brenner, S.E., Brent, M.R., Cherbas, P., Gingeras, T.R., Hoskins, R.A., Kaufman, T.C., Oliver, B., Celniker, S.E., 2011. The developmental transcriptome of Drosophila melanogaster. Nature 471, 473–479.

Hamatani, T., Carter, M.G., Sharov, A.A., Ko, M.S.H., 2004. Dynamics of global gene expression changes during mouse preimplantation development. Dev. Cell 6, 117–131.

Kalinka, A.T., 2013. The probability of drawing intersections: extending the hypergeometric distribution. arXiv:1305.0717.

Kimmel, C.B., Ballard, W.W., Kimmel, S.R., Ullmann, B., Schilling, T.F., 1995. Stages of embryonic development of the zebrafish. Dev. Dyn. 203, 253–310.
Kircher, M., 2011. Analysis of High-Throughput Ancient DNA Sequencing Data, in: Methods in Molecular Biology, Methods in Molecular Biology. Humana Press, Totowa, NJ, pp. 197–228.
Lovén, J., Orlando, D.A., Sigova, A.A., Lin, C.Y., Rahl, P.B., Burge, C.B., Levens, D.L., Lee, T.I., Young, R.A., 2012. Revisiting global gene expression analysis. Cell 151, 476–482.
Mathavan, S., Lee, S.G.P., Mak, A., Miller, L.D., Murthy, K.R.K., Govindarajan, K.R., Tong, Y., Wu, Y.L., Lam, S.H., Yang, H., Ruan, Y., Korzh, V., Gong, Z., Liu, E.T., Lufkin, T., 2005. Transcriptome analysis of zebrafish embryogenesis using microarrays. PLoS Genet 1, 260–276.
Meyer, D., Zeileis, A., Hornik, K., 2006. The strucplot framework: visualizing multi-way contingency tables with VCD. Journal of Statistical Software 17, 1–48.
Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., bo, S.P.A.A., Mann, M., 2011. Deep proteome and transcriptome mapping of a human cancer cell line. Molecular Systems Biology 7, 1–8.
Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A., Schier, A.F., 2012. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Research 22, 577–591.
Piasecka, B., Lichocki, P., Moretti, S., Bergmann, S., Robinson-Rechavi, M., 2013. The hourglass and the early conservation models-co-existing patterns of developmental constraints in vertebrates. PLoS Genet 9, e1003476.
R Development Core Team, 2011. R: A language and environment for statistical computing. (R Foundation for Statistical Computing, Vienna, Austria, 2011).
Robert, C., 2010. Microarray analysis of gene expression during early development: a cautionary overview. Reproduction 140, 787–801.
Szekely, G.J., Rizzo, M.L., 2005. Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. Journal of classification 22, 151–183.
Tan, M.H., Au, K.F., Yablonovitch, A.L., Wills, A.E., Chuang, J., Baker, J.C., Wong, W.H., Li, J.B., 2013. RNA sequencing reveals a diverse and dynamic repertoire of the Xenopus tropicalis transcriptome over development. Genome Research 23, 201–216.
Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14, 178–192.
Trapnell, C., Pachter, L., Salzberg, S.L., 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105–1111.
Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 28, 511–515.
Vesterlund, L., Jiao, H., Unneberg, P., Hovatta, O., Kere, J., 2011. The zebrafish transcriptome during early development. BMC Developmental Biology 11, 30.
Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., Birney, E., 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Research 19, 327–335.
Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.-Y., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y.E., Liu, J.-Y., Horvath, S., Fan, G., 2013. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature 500, 593–597.
Zeiner, G.M., Cleary, M.D., Fouts, A.E., Meiring, C.D., Mocarski, E.S., Boothroyd, J.C., 2008. RNA analysis by biosynthetic tagging using 4-thiouracil and uracil phosphoribosyltransferase. Methods Mol. Biol. 419, 135–146.
Zeng, F., Baldwin, D.A., Schultz, R.M., 2004. Transcript profiling during preimplantation mouse development. Dev. Biol. 272, 483–496.