**RNA sequencing reveals diverse and dynamic repertoire of the *Xenopus tropicalis* transcriptome over development**

Meng How Tan, Kin Fai Au, Arielle L. Yablonovitch, Andrea E. Wills, Jason Chuang, Julie C. Baker, Wing Hung Wong, Jin Billy Li

**SUPPLEMENTAL RESULTS**

**Comparison of RNA-seq data with microarray results**

We compared our RNA-seq data with microarray-based expression profiling results that were recently published (Yanai, et al, 2011). For multiple well-known developmental regulators, we found that the transcript levels measured by both methods agreed well with each other (Figure S2). Overall, the correlation between the RNA-seq datasets and the microarray datasets was reasonable (average R2 = 0.446) (Figure S3A) and was similar to previous studies that compared the two techniques (Marioni, et al, 2008,Fu, et al, 2009). Interestingly, we observed that at later developmental stages, a swan neck develops in the plots whereby beyond a threshold, microarray expression values remain relatively unchanged with increasing RNA-seq-derived RPKMs. This is likely because microarrays have a smaller dynamic range than RNA-seq and at high expression levels, the probes on a microarray can become saturated while the read counts from RNA-seq can increase without limit. Our RNA-seq results indicate that there are more highly-expressed genes at later developmental stages (Figure S3B), which may not be accurately quantified using microarrays.

**Analysis of single occurrence splicing events**

We examined the splice junctions that were detected in only one stage. These unique, single occurrence junctions can be found at any stage of development (Figure S9A). Despite our deep sequencing efforts, we found that we have not saturated their discovery, as the total number of such junctions increases linearly with the number of sequencing reads (R2 = 0.6049) (Figure S9B). As expected, when we further divided these single occurrence splicing events into annotated junctions and novel junctions, we only observe the linearly increasing trend for the novel junctions but not for the annotated junctions, since the annotated junctions represent a set with a defined number of members. Although deeper sequencing is likely to result in the recovery of more cryptic and rarely used splice junctions, we note that the number of novel junctions with more than one non-redundant supporting read (nNR > 1) has also not been saturated (Figure S9B). Hence, our analysis suggests that there may exist a number of unannotated stage-specific isoforms or transcripts in *Xenopus tropicalis*, some of which we have identified in this work but some of which probably remain to be uncovered.

**Validations of different types of novel splice junctions**

Novel splicing events identified from our RNA-seq datasets were validated by reverse transcription PCR. We focused on the strongly-supported junctions (nNR > 1), as they are more easily observed on an agarose gel, and selected various types of novel junctions for validations. Out of the ten different transcripts we picked, we were able to confirm the existence of new splice junctions in nine of them (Figure 4 and Figure S10-S13).

We detected exon skipping events in the fibronectin gene, *fn1* (Figure 4A-B). In one instance, exon 18 is spliced directly to exon 20, thereby omitting exon 19 from the transcript. In another instance, exon 23 is spliced directly to exon 25, thereby skipping exon 24. We designed primers flanking each skipped exon and performed PCRs to check the novel splicing events. Indeed, for each PCR, we observed bands of two different sizes. The larger, dominant bands correspond to the annotated transcript, while the lower bands represent the new transcript missing an exon. Interestingly, our PCR results also indicate different isoform usage over development. The transcript skipping exon 19 is most strongly expressed from stage 11 to stage 28, while the transcript skipping exon 24 is most strongly expressed from stage 33 to stage 45.

Next, we attempted to validate intron retention events that were detected in both the metabolic gene, *hpdl*, and the transcription factor, *gata3*. For *hpdl* (Figure 4C), 109 bases within the second exon may be spliced out, resulting in a frameshift and generation of a premature STOP codon. We were able to observe the novel transcript on an agarose gel and it appears to be expressed predominantly at the 2-cell stage and Stage 8. For *gata3* (Figure 4D), the annotated STOP codon and the beginning bases of the 3’UTR may be spliced out, resulting in a novel transcript that encodes a protein with additional amino acids at the C-terminus. A PCR with flanking primers shows that the novel transcript is always expressed concomitantly with the annotated isoform, albeit at a much lower level.

We identified novel 5’ and 3’ splice sites in the intronic region between exon 2 and exon 3 of the myosin gene, *myl6* (Figure 4E). The detected splice sites are consistent with the presence of an extra 27bp exon within the intron, which is also supported by several *Xenopus* ESTs. Using primers corresponding to the flanking exons, we were able to observe the novel transcript bearing the additional exon. Furthermore, this new isoform is only expressed after neurulation and is most strongly transcribed in the tadpole stages. As a negative control, we performed additional PCRs using primers that aligned to exon 4 and exon 5 and observed only a single band on the agarose gel at every developmental timepoint that was tested (Figure S14). Hence, there is no extra exon in the intronic region between exon 4 and exon 5 of *myl6*.

A sizeable percentage of novel splice junctions (27.9%) lie in intergenic regions and we selected two junctions for further validations (Figure 4F and Figure S13). Both junctions are supported by Xenopus ESTs that do not contain matches (by sequence) in the human and mouse genomes (all matches returned by BLAT and BLASTN span less than 5% of the ESTs and have E-values greater than 0.01). Nevertheless, we were able to confirm both junctions by PCR. The expression of the intergenic transcript in scaffold GL172641 is highest at the 2-cell stage and at Stage 8 and then decreases over development such that it cannot be detected at the late tadpole stages (Figure 4F). On the other hand, the intergenic transcript in scaffold GL173067 is present throughout development, although its expression appears to be weaker in the earlier developmental stages (Figure S13).

**Computational pipeline** **to identify novel protein-coding and non-coding genes**

We developed a filtering pipeline to identify unannotated protein-coding and non-coding genes from our RNA-seq data (Figure 5A). First, we used two distinct assemblers, Cufflinks (Trapnell, et al, 2010) and Trinity (Grabherr, et al, 2011), to reconstruct transcripts from our deep sequencing data. Second, for each set of assembled contigs, we used BLAST to filter out the transcripts that are already present in the *Xenopus* RefSeq or Ensembl annotations in the genome. We refer to the remaining contigs as novel (unannotated) transcripts. Third, we sought to identify the protein-coding sequences out of these novel transcripts. To find the contigs that are part of protein-coding genes, we (1) searched for matches to human and mouse protein-coding genes, (2) identified hits in the Pfam database of protein families (Punta, et al, 2012), (3) identified hits in the Uniprot database of protein sequences (UniProt Consortium, 2012), and (4) translated each transcript in all three frames to find the longest possible peptide and flagged those that can give rise to a protein of at least 100 amino acids (see Methods for more details). Those unannotated contigs that show no evidence of being part of a protein-coding gene are hence potential non-coding RNAs. Among the non-coding contigs, we further searched for matches to human and mouse non-coding RNAs as well as identified hits in the NONCODE (Bu, et al, 2012) and Rfam (Gardner, et al, 2011) databases in order to determine which of the unannotated contigs are actually known non-coding transcripts.

**Comparison of transcripts assembled by Cufflinks and Trinity**

On average, Cufflinks assembles fewer but longer contigs than Trinity. In order to determine whether Cufflinks can combine several shorter sequences that are mapped close together into a single long transcript, we examined the 3’end of the cyclin K gene, *ccnK* (Figure 7C, left panel). For Trinity, the corresponding contig extends the annotated 3’UTR by about 700bp. However, almost immediately downstream of this *ccnK*-matching Trinity contig are other shorter contigs. On the other hand, Cufflinks appears to merge these shorter sequences with *ccnK* and reports a single long transcript that extends the annotated 3’UTR by about 1,700bp. To assess whether Cufflinks or Trinity is correct, we performed a PCR using one primer that anneals to the last coding exon of *ccnK* and a second primer that anneals to the end of the long Cufflinks transcript. From the agarose gel, we observed a strong band of the expected size (Figure 7C, right panel), which indicates that the 3’UTR of *ccnK* should indeed be extended by at least 1,700bp in accordance to the Cufflinks assembled transcript.

As another example, we examined the ~40kb intergenic region between *sox10* and *baiap2l2* (ENSXETG00000006351) (Figure S25). While both *sox10* and *baiap2l2* are located on the minus strand, all the transcripts assembled by Cufflinks and Trinity in the intergenic region are derived from the plus strand, indicating that some novel gene exists between *sox10* and *baiap2l2*. However, while Cufflinks reports a single multiply-spliced transcript, Trinity outputs each individual exon as an independent contig. To determine if there are several short independent transcripts derived from the intergenic region or one long, spliced transcript, we performed a PCR using primers that anneal to the first and last exon and detected a band whose size is about 1,000bp and is consistent with a multiply-spliced transcript. Hence, it appears that some of the individual contigs assembled by Trinity may actually belong to the same transcript.

**SUPPLEMENTAL FIGURE LEGENDS**

**Figure S1.** Correlation between biological replicates.

We collected embryos from two separate clutches at different developmental stages and built RNA-seq libraries for them independently. 11 stages are identical between the two clutches and serve as biological replicates. We calculated the expression levels (RPKMs) for each replicate separately and compared them to one another. Since the correlation between the biological replicates was high (R2 ranged from 0.8772 to 0.9493), we then pooled the reads from the two clutches for common developmental stages.

**Figure S2.** Measurements of transcript abundance for some well-known developmental regulators.

We plotted the transcript levels determined by our RNA-seq experiments (in blue) as well as the transcript levels determined by custom-designed Agilent microarrays (in red) (Yanai, et al, 2011) on the same graphs for ten developmental genes. The expression profiles determined by both methods agree well with each other.

**Figure S3.** Global correlation between RNA-seq data and microarray data.

(A) We compared our RNA-seq data with the microarray data generated by Yanai *et al.* (2011) for *Xenopus tropicalis*. Overall, the two sets of gene expression data showed a reasonable correlation with each other (R2 ranged from 0.3235 to 0.5436). Interestingly, we observed the emergence of a “swan neck” feature in later developmental stages, whereby the expression levels given by microarrays appear to saturate beyond a certain point but the expression levels given by RNA-seq can theoretically increase without limit.

(B) At every developmental stage, we determined the maximum RPKM among all the expression values given by RSeq (blue bars) and counted the number of genes with RPKM greater than 5000 (red line). The number of highly transcribed genes shows a general increase with developmental progression.

**Figure S4.** Distribution of RPKMs.

The RPKM values at every developmental stage are displayed in a box-and whisker plot. The top and bottom of each box correspond to the 75th percentile and the 25th percentile respectively, while the black line inside the box corresponds to the 50th percentile (the median). The whiskers (dotted lines) extend to the most extreme data point that is no more than 1.5 times the interquartile range from the box. The open circles correspond to the outliers that are outside the range of the whiskers. The orange boxes represent the RefSeq annotated genes, while the blue boxes represent the Ensembl annotated genes. For the x-axis, R stands for RefSeq, while E stands for Ensembl.

**Figure S5.** Clustered expression profiles for Ensembl genes.

As shown in the heatmap, the expression of most of the 16012 genome-annotated Ensembl genes that were detected in our RNA-seq experiments appeared to be developmentally regulated. The RPKMs were mean-centered and normalized, with each row representing a different gene. The key developmental events, namely the midblastula transition (MBT), gastrulation (Gast), nerulation, and organogenesis, are also labeled below the heatmap.

**Figure S6.** Genes transcribed before the midblastula transition.

(A) Even though the initial stages of development prior to the midblastula transition are believed to be transcriptionally silent, our RNA-seq data indicate that 313 genes may be transcribed based on the following criteria: (i) RPKM at stage 6 is at least 2 fold higher that that at the 2-cell stage, (ii) an average RPKM of 1 at each developmental stage, (iii) no decrease in RPKM by more than 1 or by more than 20% between two consecutive stages. GO analysis of these genes indicate that they are associated with various cell cycle processes, signaling cascades, phosphorylation, and apoptosis.

(B) We selected eight genes for validation by quantitative reverse transcription real-time PCR and were able to confirm six of them. The blue bars represent the average fold change relative to the 2-cell stage as determined by quantitative PCR, while the red lines represent the RPKMs calculated by RSeq based on RefSeq annotations.

**Figure S7.** Detection of splice junctions from RNA-seq reads.

(A) The number of splicing events identified by SpliceMap is dependent on sequencing coverage. Left panel: Total number of splice junctions detected (green crosses) and the number of detected junctions that are annotated in RefSeq or Ensembl (black diamonds) plotted against the number of sequencing reads. Right panel: The number of novel splice junctions with more than one non-redundant supporting read (nNR > 1) (red triangles) and the number of novel splice junctions with only one non-redundant supporting read (nNR = 1) (blue squares) plotted against the number of sequencing reads. Even though there is a general increase in the number of detected splice junctions with deeper sequencing coverage, the identification of annotated splicing events is beginning to saturate. For every additional 10 million reads, the number of annotated junctions that SpliceMap detects increases by only 5.1%, but the number of novel, strongly-supported junctions (nNR > 1) increases by an average of 17.3% and the number of novel, weakly-supported junctions (nNR = 1) increases by 25.2%.

(B) At every developmental stage, the majority of the detected splice junctions can be found in either the RefSeq or Ensembl annotation (unshaded portions). In addition, between 12.0 to 17.8% of the detected junctions at each stage are novel and supported by at least two non-redundant reads (nNR > 1) (red-colored segments), while between 9.2 to 15.5% are novel but supported by only one non-redundant read (nNR = 1) (blue-colored segments).

**Figure S8.** EST evidence for novel splice junctions.

(A) To assess how reliable our splice junction predictions are, we asked how many of our novel junctions are documented in the collection of *Xenopus* ESTs. We focused our attention on the splicing events that are supported by at least two non-redundant reads (red line) because we have previously found that the criterion of nNR > 1 greatly increases the specificity of junction detection by SpliceMap (Au, et al, 2010). Here, we observed that at every developmental stage, the majority of the novel junctions with nNR > 1 have additional backing from ESTs (pink line), while the minority of these junctions have no EST evidence (purple line).

(B) At each stage, between 54.4% to 69.0% of the novel junctions with nNR > 1 are documented in the Xenopus EST collection (pink portions). We note that since the EST database is likely to be incomplete, we expect the specificity of our junction predictions to be considerably higher than these percentages.

**Figure S9.** Discovery of stage-specific splicing events.

(A) Thousands of splice junctions can be detected at only one developmental stage. These single occurrence junctions can be found in all the stages we assayed. With the exception of stage 15, the majority of these junctions are novel and are supported by only one non-redundant read (blue segments), thereby suggesting that these highly stage-specific splicing events belong to rare isoforms, which are not readily detected. In red: novel junctions supported by at least two non-redundant reads; unshaded: RefSeq or Ensembl annotated junctions.

(B) To determine if we have discovered most of the single occurrence splicing events, we plotted their numbers against the sequencing coverage at each stage. While the number of single occurrence junctions that are annotated in either RefSeq or Ensembl seem to have saturated, we observed that the number of novel splicing events that occur in only one developmental stage increases linearly with the number of sequencing reads, indicating that there are likely to be even more stage-specific junctions yet to be discovered despite our sequencing efforts.

**Figure S10.** Novel exon skipping events in the *xbp1* transcription factor.

(A) The annotated gene structure of *xbp1* is shown, which contains five coding exons. The three pairs of arrows (red, orange, and green) represents the primer sets used for PCR validations.

(B) We sought to validate the new splice junctions detected in *xbp1*. According to our RNA-seq data, exon 2 may be spliced directly to exon 4 or exon 5, while exon 3 may be spliced directly to exon 5. However, we were unable to validate the direct splicing of exon 2 to exon 4; we only observed bands of the annotated size when we performed a PCR using primers that anneal to exon 2 and exon 4 (green arrows, PCR 1). Nevertheless, we were able to confirm the other two exon skipping events (PCR2 and PCR3). In addition, we observed different isoform usage over development. The annotated transcript is strongly expressed from stage 11 to stage 45 (mature tadpoles), but the peak expression of the novel transcripts occurs only during gastrulation and neurulation.

(C) The three isoforms of *xbp1* are shown. The annotated isoform contains all five exons, while the new isoforms are missing one or two exons.

**Figure S11.** Extra exon in the *bend5* gene.

(A) Upper panel: The Ensembl annotated gene structure of *bend5* is shown, which contains six coding exons. The arrows represent the primers used for PCR validations, while the red bent lines indicate novel splicing events detected from our RNA-seq data. Lower panel: At every developmental stage, we observed two bands – a stronger band corresponding to the annotated isoform and a weaker band whose size is consistent with the presence of an additional exon predicted from RNA-seq.

(B) The two isoforms of *bend5* are shown. The annotated isoform contains six exons, while the new isoform contains seven exons, with the extra exon occurring between exon 3 and exon 4 of the annotated gene structure.

(C) Additional PCRs were performed to demonstrate that the presence of an extra exon is specific to only one intronic region. Upper panel: In the first negative control, primers that annealed to exon 4 and exon 5 were used. Lower panel: Only a single band was detected on the agarose gel at every developmental stage that was tested.

(D) Upper panel: In the second negative control, primers that annealed to exon 5 and exon 6 were used. Lower panel: Again, only a single band was observed on the gel at every developmental timepoint that was tested.

**Figure S12.** Extra exons in the clathrin gene, *cltb*.

(A) The annotated gene structure of *cltb* is shown, which contains four coding exons. The novel splicing events identified by SpliceMap are indicated as red bent lines, while the PCR primers used for validations are represented as black arrows in the schematic.

(B) Our PCR experiments revealed an unexpected result. We observed PCR products of three different sizes in our gel image, which suggests that there are two new isoforms instead of one. In addition, while the annotated isoform is constitutively expressed, the novel transcripts are only present in the tadpole stages.

(C) We examined the collection of *Xenopus* ESTs to look for clues that might explain our PCR results. As shown in the screenshot of the UCSC Genome Browser, there are numerous ESTs that contain an extra exon between exon 3 and exon 4 of the RefSeq or Ensembl annotation. However, while the length of this extra exon is identical for many ESTs and matches our prediction from RNA-seq data, one of the ESTs (CX794285) contains an additional exon that is longer than expected.

(D) The gene structures of the two new isoforms of *cltb* are shown. Both novel transcripts contain an extra exon between exon 3 and exon 4 of the annotated transcript. However, the additional exon for one of the novel isoforms is shorter than that for the other novel isoform, with both extra exons sharing the same 3’end.

**Figure S13.** Identification of a novel splice junction in the intergenic region between *slc35d1* and *il12rb2*.

(A) A snapshot of the UCSC Genome Browser shows the genomic locus of a splicing event detected between GL173067: 868621 and GL173067: 870321. A few ESTs support the detected splicing event.

(B) As shown in the gel image, we validated the novel splicing event by performing a PCR using primers flanking the novel junction. The bands are of the expected size.

**Figure S14**. A negative control for the myosin gene *myl6*.

Upper panel: The RefSeq annotated gene structure of *myl6* contains six individual exons. The black arrows represent the PCR primers that were used for the negative control experiment. Lower panel: Only a single band of the expected size was observed on the agarose gel for every developmental stage tested, indicating that there is unlikely to be an extra exon in the intronic region between exon 4 and exon 5.

**Figure S15.** Classification of unannotated *Xenopus* contigs into protein-coding transcripts and non-coding RNAs.

(A) The numbers of contigs that matched entries in the Pfam and Uniprot protein databases are shown in the Venn diagram together with the number of contigs that could encode a peptide of at least 100 amino acids.

(B) The numbers of contigs that matched entries in the NONCODE and Rfam non-coding RNA databases are shown in the Venn diagram.

(C) Besides checking against publicly available databases, we also used BLAT to identify *Xenopus* contigs that matched RefSeq-annotated human and mouse genes. The number of contigs that matched mammalian protein-coding genes and non-coding genes are shown.

**Figure S16.** Alignment of reads between *hoxc11* and *hoxc12*.

A screenshot of the Integrative Genomics Viewer (Robinson, et al, 2011) shows the alignment of reads in the genomic locus of *hoxc11* (ENSXETG00000023471) and *hoxc12* (ENSXETG00000023470) for stage 15 and stage 24-26. The red arrows below the screenshot indicate the PCR primers that were used to determine if a putative *HOTAIR* transcript may exist in *Xenopus*.

**Figure S17.** Cumulative distribution functions of intron sizes in *Xenopus* and humans.

(A) The distribution function of *Xenopus* introns based on the RefSeq annotation is shown.

(B) We also plotted the distribution function of the length of introns annotated in the human genome (hg19). As shown by the dotted lines, approximately 95% of human introns are shorter than 25kb.

**Figure S18.** Expression levels of unannotated non-coding Cufflinks transcripts.

(A) We plotted a histogram showing the distribution of confident non-coding contigs based on their maximum expression levels (FPKMs calculated by Cufflinks). The majority of the contigs appear to be lowly expressed and have a maximum FPKM of less than or equal to 5. On the x-axis, a bin that is labeled “10” contains all contigs with a maximum FPKM that is greater than 5 but less than or equal to 10 and so on.

(B) We plotted a heatmap showing how the expression of these non-coding *Xenopus* transcripts changed over all the assayed developmental stages. The FPKMs were mean-centered and normalized, with each row representing a different contig. The key developmental events are also marked below the heatmap.

**Figure S19.** Alignment of Trinity contigs with annotations and the reference genome.

We compared the Trinity assembled transcripts against the RefSeq annotation (purple crosses), Ensembl annotation (green triangles), and the *Xenopus* reference genome (XenTro3, red diamonds). At every developmental stage, more Trinity contigs match Ensembl genes than RefSeq genes (E-value < 1x10-10), since the Ensembl annotation is larger than the RefSeq annotation. On average, 57.1% of the Trinity transcripts can be found in either the RefSeq or the Ensembl annotation (blue squares). Most importantly, about 1.6% of the Trinity assembled contigs cannot be aligned to the current reference genome, highlighting that the *Xenopus tropicalis* genome is still incomplete and may be missing hundreds of genes.

**Figure S20.** Steps for eliminating heterologous sequences.

We designed a pipeline to determine if any of the Trinity contigs that do not contain matches in the XenTro3 reference genome may be derived from contaminating sources. As shown in the flowchart, we checked for bacterial contamination, fungal contamination, and general plasmid contamination. Contigs that passed through all the filters were considered to be genuine *Xenopus* transcripts. Although we focused our checks on the unaligned contigs with no EST evidence, we note that some of the unaligned contigs with EST support may also be heterologous sequences.

**Figure S21.** Protein-coding potential of non-contaminating unaligned Trinity contigs.

(A) The number of unaligned contigs that have protein-coding potential varies from stage to stage. We note that there is an unusually large number of such contigs in stage 11-12 (gastrulation).

(B) At every developmental stage, we calculated the percentage of unaligned Trinity contigs that have protein-coding potential. On average, we found that 13.1% of the contigs may encode proteins.

**Figure S22.** Number of Trinity contigs unmatched to the reference genome as a function of sequencing depth.

(A) The number of EST-supported Trinity contigs that are absent from the current genome was plotted against the sequencing coverage at each stage. We observed that the number of these contigs with EST evidence increases logarithmically with the number of reads and is beginning to saturate.

(B) The number of unmatched Trinity contigs that had never been detected prior to this study was plotted against the number of sequencing reads. We observed that the number of these undocumented contigs increases linearly with sequencing coverage, which suggests that even deeper sequencing is required to uncover more transcripts and that hundreds of transcribed sequences are currently omitted from the *Xenopus tropicalis* genome.

**Figure S23.** Length distributions of assembled contigs

The sizes of *ab inito* assembled and *de novo* assembled contigs at every developmental stage are displayed in a box-and whisker plot. The top and bottom of each box correspond to the 75th percentile and the 25th percentile respectively, while the black line inside the box corresponds to the 50th percentile (the median). The whiskers (dotted lines) extend to the most extreme data point that is no more than 1.5 times the interquartile range from the box. The open circles correspond to the outliers that are outside the range of the whiskers. The orange boxes represent the Cufflinks assembled transcripts, while the blue boxes represent the Trinity assembled transcripts. For the x-axis, C stands for Cufflinks, while T stands for Trinity.

**Figure S24.** Histograms of contig lengths.

At every developmental stage, we plotted the distribution of contigs based on their sizes. The Cufflinks transcripts are shown in blue, while the Trinity contigs are shown in red. We observed that, at each stage, an average of 75.0% of the Trinity contigs are shorter than or equal to 1000bp, while only 45.1% of the Cufflinks transcripts are not longer than 1000bp.

**Figure S25.** A novel gene within the intergenic region of *sox10* and *baiap2l2*.

A screenshot of the UCSC Genome Browser at the genomic locus of *sox10* and *baiap2l2* (ENSXETT00000013905 and ENSXETT00000013906) shows how the contigs assembled by Trinity at the locus differ from the transcript reconstructed by Cufflinks, which is labeled as the BLAT search sequence (“YourSeq”). Trinity outputted each exon as an independent contig, while Cufflinks connected the exons together into a single spliced transcript. We performed a PCR using primers that annealed to the start and end of the Cufflinks sequence (pink arrows) and detected a ~1000bp product (inset), which is consistent with a longer spliced transcript. The PCR product was further sequenced to confirm that it matched the correct genomic locus.

**Figure S26.** Expression of previously reported early transcribed genes.

Prior to our work, a handful of genes had been reported to be transcribed before the midblastula transition (Skirkanich, et al, 2011). Based on our RNA-seq data, we plotted the expression levels (RPKMs) of four of the genes, namely *xnr6*, *mixer*, *sox17a*, and *derrière*, in the beginning stages of development. Contrary to that previous study, we did not observe a gradual increase in transcript levels. Instead, the general trend appeared to be exponential, with a modest amount of transcription before the midblastula transition followed by a strong burst of transcription upon embryonic genome activation.

**SUPPLEMENTARY FILES**

**File S1.** RPKMs of RefSeq annotated genes. The first worksheet contains the overall expression level of each gene, while the second worksheet contains the RPKMs for every isoform.

**File S2.** RPKMs of Ensembl annotated genes. The first worksheet contains the overall expression level of each gene, while the second worksheet contains the RPKMs for every isoform.

**File S3.** GO analysis of the 8 clusters shown in Figure 1D using the Panther classification system (Thomas, et al, 2006).

**File S4.** List of early transcribed genes, as determined by the following three criteria: (i) a two-fold increase in RPKMs between the 2-cell stage and stage 6, (ii) an average RPKM of at least 1, and (iii) the expression level is monotonically increasing. Their expression levels and associated GO terms are also provided. In addition, the genes that are in bold have been previously determined to be transcribed before the midblastula transition in zebrafish (Aanes, et al, 2011) and the corresponding zebrafish gene names are provided in column Y.

**File S5.** List of early transcribed genes, as determined by the following three criteria: (i) a two-fold increase in RPKMs between the 2-cell stage and stage 6, (ii) an average RPKM of at least 1, and (iii) the expression level does not decrease by more than 1 RPKM or by more than 20% between two consecutive stages. Their expression levels and associated GO terms are also provided. In addition, the genes that are in bold have been previously determined to be transcribed before the midblastula transition in zebrafish (Aanes, et al, 2011) and the corresponding zebrafish gene names are provided in column Y.

**File S6.** Breakdown of all the Cufflinks contigs by four categories - annotated transcripts, novel proteins, definite non-coding transcripts, and putative non-coding transcripts.

**File S7.** Lists of non-contaminating, unaligned Trinity contigs that contain matches to the genome or transcriptome of human, mouse, or zebrafish.

**SUPPLEMENTAL TABLES**

**Table S1.** Categories of novel junctions that are supported by at least two non-redundant reads (nNR > 1).

**Table S2.** Primers used to validate genes transcribed before the midblastula transition.

**Table S3.** Primers used to validate novel splice junctions and unannotated transcripts using iQ SYBR Green Supermix (Bio-Rad).

**Table S4.** Primers used to validate unannotated transcribed sequences using Herculase II Fusion DNA Polymerase (Agilent) as well as the annealing temperature, T, for each PCR.

**SUPPLEMENTAL REFERENCES**

Aanes, H., Winata, C. L., Lin, C. H., Chen, J. P., Srinivasan, K. G., Lee, S. G., Lim, A. Y., 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.

Au, K. F., Jiang, H., Lin, L., Xing, Y., and Wong, W. H. 2010. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res.* **38:** 4570-4578.

Bu, D., Yu, K., Sun, S., Xie, C., Skogerbo, G., Miao, R., Xiao, H., Liao, Q., Luo, H., Zhao, G., et al. 2012. NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res.* **40:** D210-5.

Fu, X., Fu, N., Guo, S., Yan, Z., Xu, Y., Hu, H., Menzel, C., Chen, W., Li, Y., Zeng, R., et al. 2009. Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* **10:** 161.

Gardner, P. P., Daub, J., Tate, J., Moore, B. L., Osuch, I. H., Griffiths-Jones, S., Finn, R. D., Nawrocki, E. P., Kolbe, D. L., Eddy, S. R., et al. 2011. Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res.* **39:** D141-5.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29:** 644-652.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18:** 1509-1517.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., et al. 2012. The Pfam protein families database. *Nucleic Acids Res.* **40:** D290-301.

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. 2011. Integrative genomics viewer. *Nat. Biotechnol.* **29:** 24-26.

Skirkanich, J., Luxardi, G., Yang, J., Kodjabachian, L., and Klein, P. S. 2011. An essential role for transcription before the MBT in Xenopus laevis. *Dev. Biol.* **357:** 478-491.

Thomas, P. D., Kejariwal, A., Guo, N., Mi, H., Campbell, M. J., Muruganujan, A., and Lazareva-Ulitsky, B. 2006. Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Res.* **34:** W645-50.

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.

UniProt Consortium. 2012. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **40:** D71-5.

Yanai, I., Peshkin, L., Jorgensen, P., and Kirschner, M. W. 2011. Mapping gene expression in two Xenopus species: evolutionary constraints and developmental flexibility. *Dev. Cell.* **20:** 483-496.