Supplementary Figures and Tables

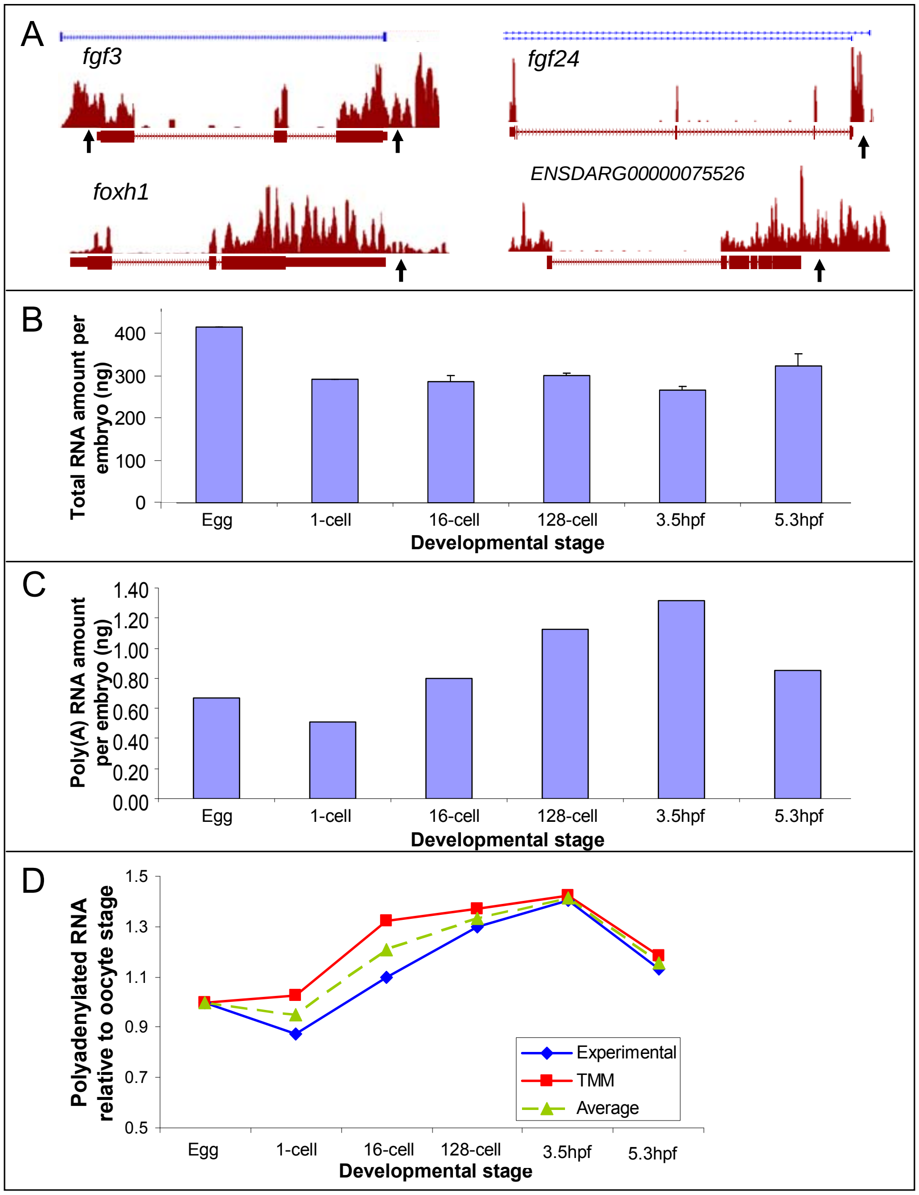

Figure S1 (related to Fig 1 and Supplementary Methods). Data normalization against RNA amount and extension of mapped reads beyond annotated sequences. (A) Extension of mapped reads beyond annotated 5′ or 3′ UTRs observed in several genes. Light brown bars indicate gene structure, blue bars at the top of each panel indicate di-tag mapping. Black arrows indicate the site of extended reads mapping.(B)

The amount of total RNA measured per embryo is reduced from the egg to the 1-cell stage (0 hpf), and remains stable until post-MBT, where it increases again.

(C) The amount of poly(A)$^+$ RNA shows a distinct pattern from the total RNA content. It increases gradually from the 1-cell (0.2%) to the MBT stage (0.5%), whereby it decreases again.

(D) We compared the poly(A) content derived from direct measurement with a global fold change factor (TMM) and found very good correlation (r = 0.90). The average of these values (green dotted line) was used in the normalization procedure.

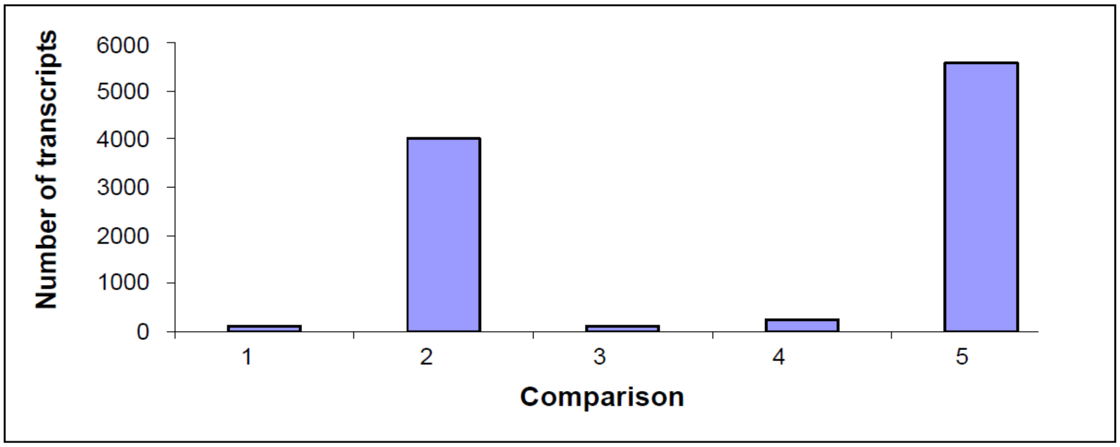

Figure S2 (related to Fig 2). Statistical testing using the R package DEGseq, with the method MARS.

We performed successive pairwise testing (i.e. Egg vs 1-cell, 1-cell vs 16-cell, and so forth). Many transcript appeared twice or more after the initial test. We kept only one observation (with the lowest p-value). This revealed two time points with major changes in poly(A) $^+$ RNA levels (up and down regulation); between the 1-cell and 16-cell (comparison 2) and between 3.5hpf and 5.3hpf (comparison 5). The figure shows the number of transcripts with a significant change between the groups compared.

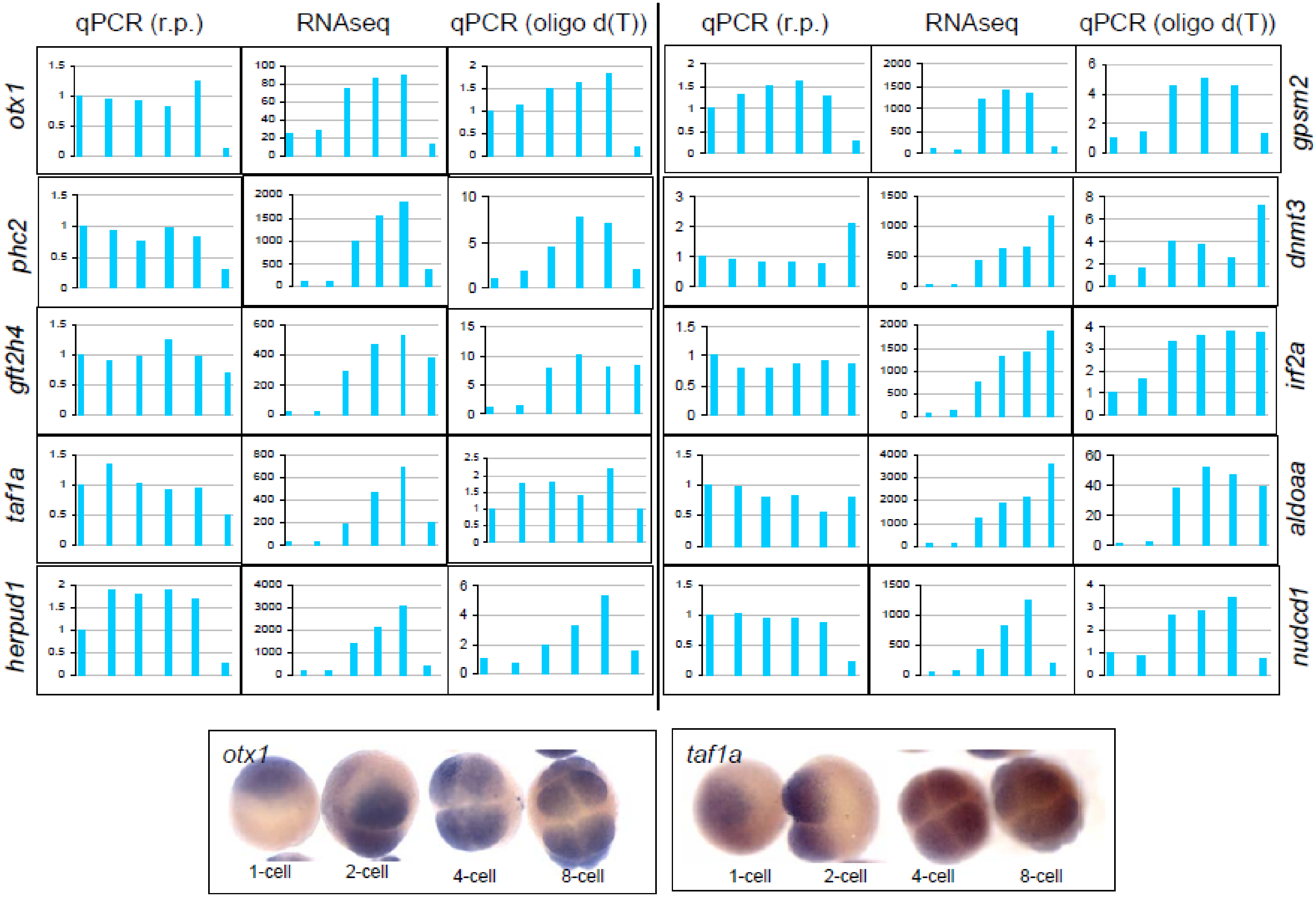
qPCR (r.p.)
RNAseq
qPCR (oligo d(T))
qPCR (r.p.)
RNAseq
qPCR (oligo d(T))
otx1
phc2
gtf2h4
taf1a
herpud1
gpsm2
dnmt3
irf2a
aldoaa
nudcd1
otx1
1-cell
2-cell
4-cell
8-cell
taf1a
1-cell
2-cell
4-cell
8-cell

Figure S3 (related to Fig 3). Real-time PCR validation of pre-MBT super-cluster genes.

Good correlation was observed between mRNA-seq and RT-qPCR using oligo d(T) primers, but not with RT-qPCR using r.p.. Of 29 genes tested, 24 showed such pattern (10 shown here). Y-axis indicates relative enrichment value against egg stage as the baseline ($2^{-\Delta\Delta Ct}$) in RT-qPCR and read counts in mRNA-seq. WISH at 1-cell to 8-cell stages performed on two of the tested pre-MBT transcripts: *otx1* and *taf1a,* showed positive expression.

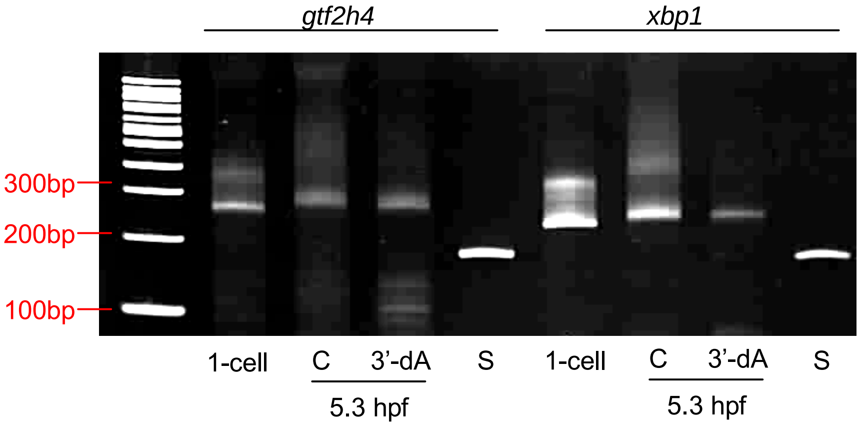

Figure S4 (related to Fig 4). Inhibition of polyadenylation by cordycepin. Poly(A) tails of *gtf2h4* and *xbp1* transcripts from control (C) and 3'-dA-treated (3'-dA) embryos were measured before (at 1-cell stage) and after (at 5.3 hpf) treatment. A longer smear representing longer poly(A) tail could be seen in the control as compared to the treated samples. S – fragment amplified using gene-specific internal primers.

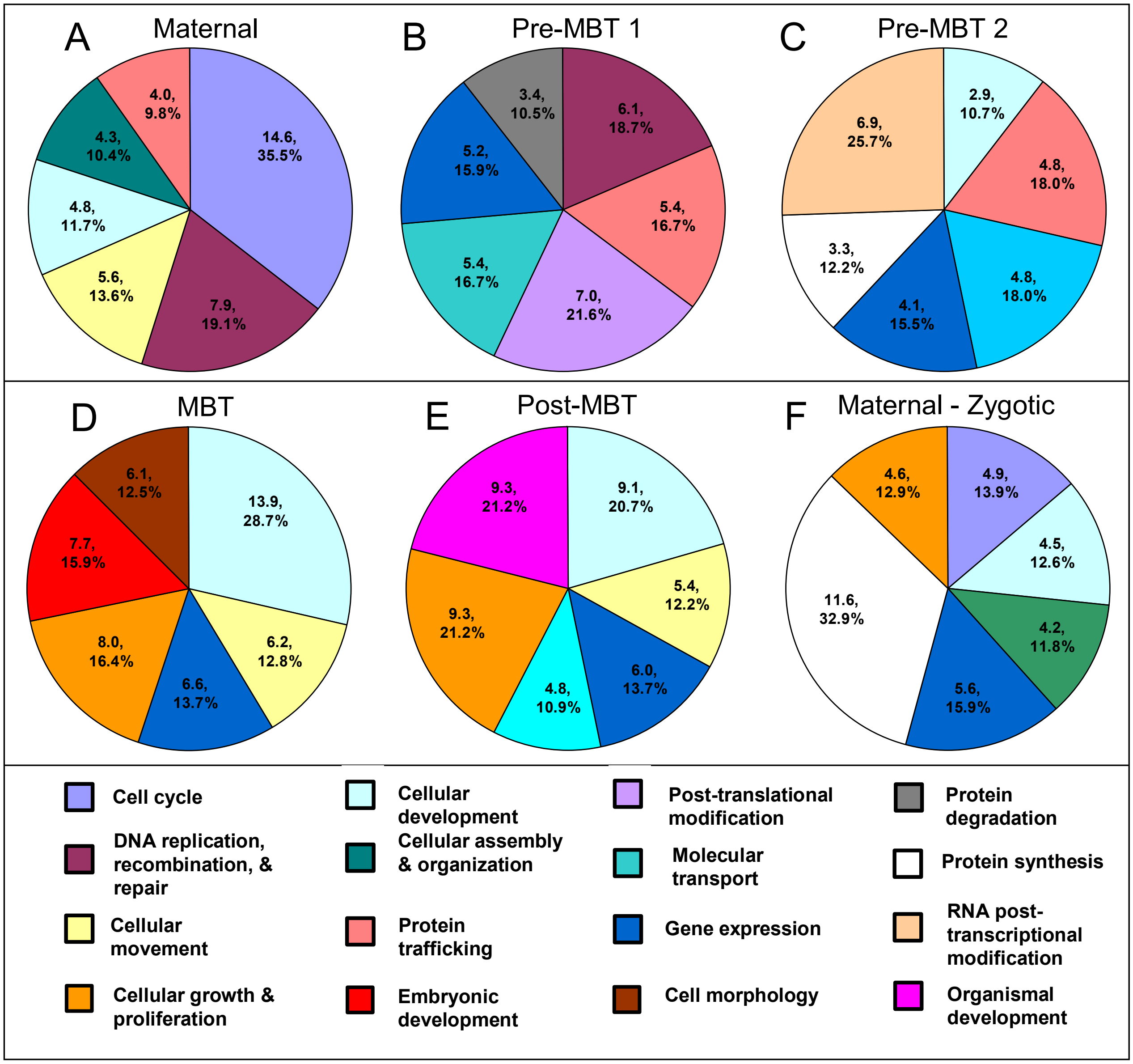

Figure S5. Enrichment of GO terms in individual clusters.

The top-six GO terms based on p-values are depicted. Number in each slice represents –log(p-value) (top) and percentage of genes within cluster (bottom, represents slice size). Cluster names are indicated at the top of each chart. Color representations of different GO terms are collectively indicated below the charts.

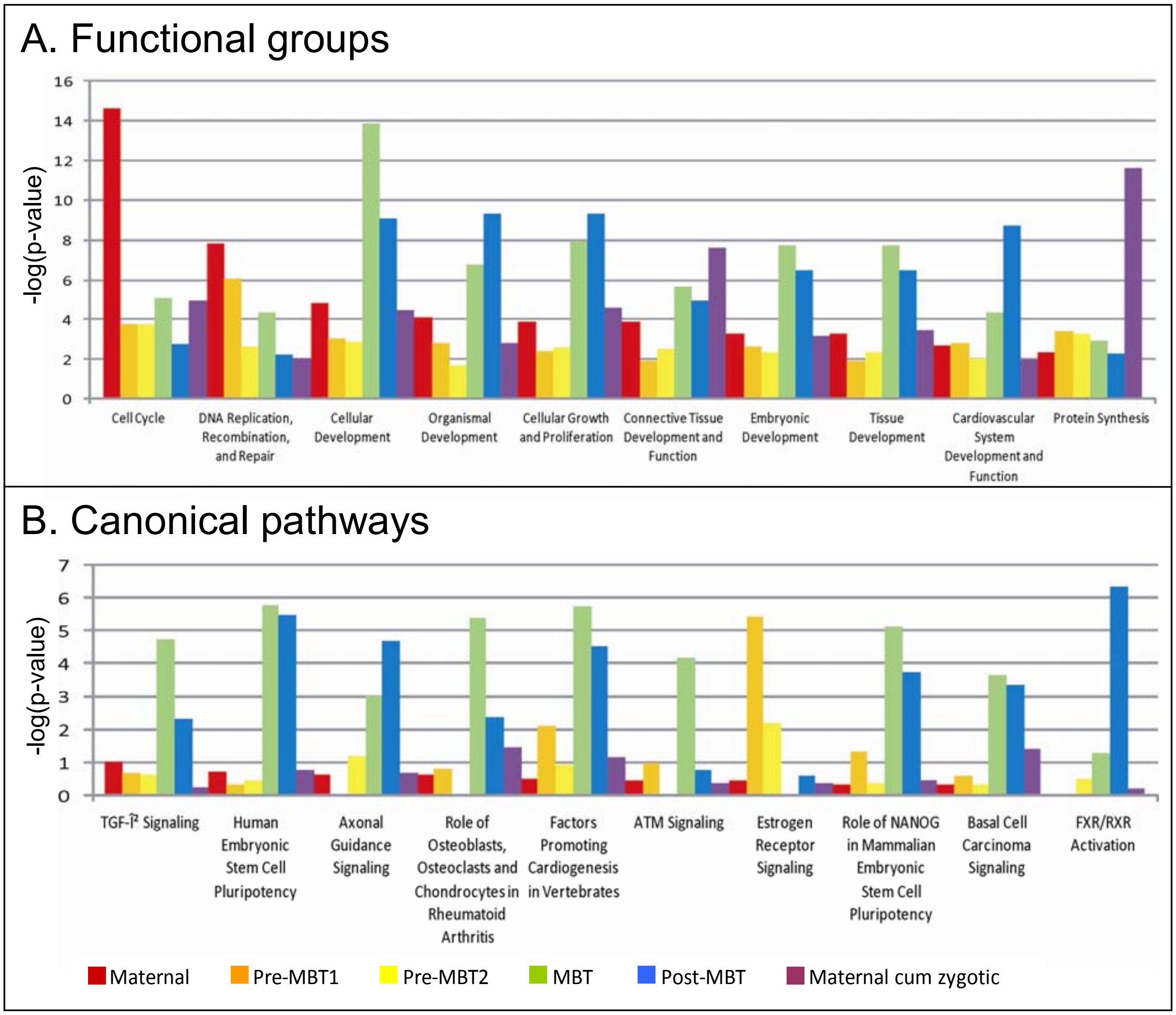

Figure S6. Representation of gene clusters into different functional groups.

(A) Proportion of different clusters enriched in various biological function groups.

(B) Enrichment of clusters in various canonical pathways involved in development.

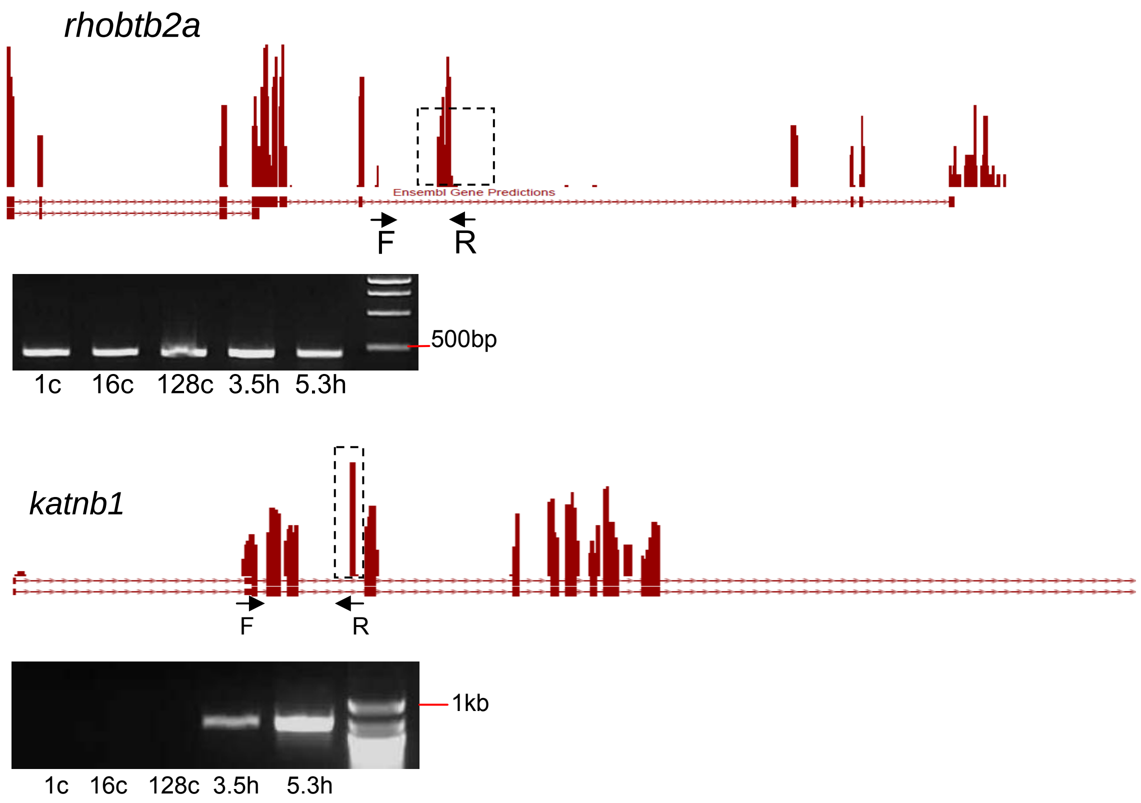

Figure S7 (related to Fig 5). RT-PCR validation of several NTRs located near annotated genes.

Primers (black arrows, F – forward, R – reverse) were designed spanning one annotated exon and the tested NTR (box). RT-PCR analysis was performed in 1-cell (1c), 16-cell (16c), 128-cell (128c), 3.5 hpf (3.5h), and 5.3 hpf (5.3h) samples.

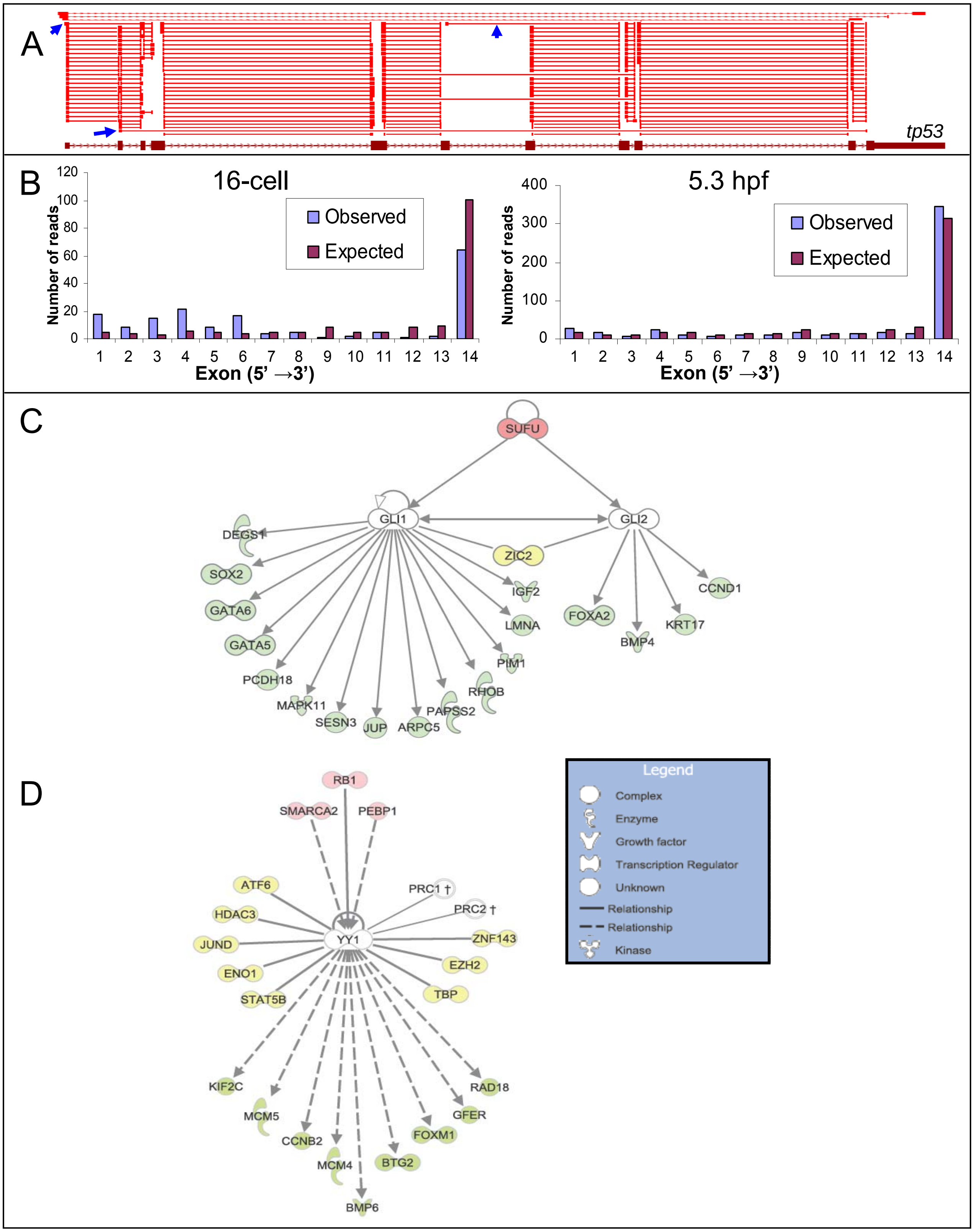
A
tp53
B
16-cell
Number of reads
120
100
80
60
40
20
0
Observed
Expected
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Exon (5' →3')
5.3 hpf
Number of reads
400
300
200
100
0
Observed
Expected
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Exon (5' →3')
C
SUFU
GLI1
GLI2
DEGS1
SOX2
GATA6
GATA5
PCDH18
MAPK11
SESN3
JUP
ARPC5
PAPSS2
RHOB
PIM1
LMNA
IGF2
ZIC2
FOXA2
BMP4
KRT17
CCND1
D
RB1
SMARCA2
PEBP1
ATF6
HDAC3
JUND
ENO1
STAT5B
YY1
PRC1 †
PRC2 †
ZNF143
EZH2
TBP
KIF2C
MCM5
CCNB2
MCM4
BMP6
BTG2
FOXM1
GFER
RAD18
Legend
Complex
Enzyme
Growth factor
Transcription Regulator
Unknown
Relationship
Relationship
Kinase

Figure S8 (related to Fig 6). ***GLI2*** **and** ***YY1*** **gene regulatory networks.**

(A) Splice variant analysis in *tp53* at 5.3 hpf. Di-tags (red bars with arrowheads) are shown on top of each panel. Red horizontal lines indicate junction mapping. Reads spanning exons 1 and 3 (skipping exon 2), 2 and 11 (skipping exons 3-10), 5 and 9 (skipping exons 6-8), as well as exons 6 and 8 (skipping exon 7) imply several isoforms (blue arrows).

(B) Comparison of the observed (red bars) and expected (blue bars) read count for each of the 14 exons of *gli2a* at 16-cell and 5.3 hpf, indicating the presence of two isoforms which differ in the presence of exon 14, encoding a short and a long Gli2a protein variant respectively. A shift towards the longer isoform was observed at 5.3 hpf.

(C) Transcriptional network for *GLI1* and *GLI2*. Both *gli1* and *gli2a* transcripts increase at pre-MBT stages, however *gli1* decreases again post-MBT while *gli2a* continues to increase, suggesting that Gli1 may have a more restrictive role in zebrafish. Their downstream targets are shown in light green (pre-MBT2, MBT and post-MBT cluster genes). All relationships are direct (e.g. physical) and shown to increase the expression of the downstream targets. SUFU is an upstream regulator of the activator/suppressor form of Gli. *sufu* decrease in expression as the 14th exon inclusion isoform increase. ZIC2 which is part of the "Gli-code" and important in cell differentiation interacts with both GLI1 and GLI2a.

(D) Gene regulatory network of Yin and yang 1 (*yy1a* in zebrafish). This gene increased in expression from 3.5 hpf and onwards. Red symbols are upstream inhibitors of *YY1* expression, and their decrease in expression coincides with the onset of *yy1a* expression. Figures in green are possible downstream targets of YY1, and their expression decreases after the onset of *yy1a* expression (pre-MBT1 or Degradation 3 cluster). For better clarity only 9 of 19 IPA predicted downstream targets are shown. Among the excluded genes were *HSPA5*, *SCARB1*, *FDPS* and *SREBF1* (physical interactions, Protein-DNA binding). Figures in yellow are proteins which YY1 has been shown to interact with. The PRC1 and PRC2 interactions are custom made (no color). Our findings suggest that transcription is changed towards more specific inhibition and induction around MBT.

Supplementary Tables

Table S1 (related to Fig 1). Distribution of sequencing reads across all 6 libraries. A: Count of reads with number of mappings in proper range (1 <= N <= 10). B: Count of reads with number of mappings in proper range (1 <= N <= 10) and align score >= 26. C: Count of reads uniquely aligned (minScoreGapToSecondBestAlignment = 4) with align score >= 26.

	Total Reads	Filter	A	B	C	Tags mapping to Ensembl
Egg	31,323,986	105,587	19,462,743	17875474	16,616,142	14,105,191
1-cell	23,506,971	84,840	15,337,433	14246586	13,307,726	11,435,032
16-cell	35,844,766	131,951	23,841,732	21798579	20,475,830	17809,822
128-cell	27,119,026	104,501	18,682,912	17239656	16,213,755	14,132,828
3.5 hpf	37,886,769	255,605	25,330,558	23412248	21,989,690	18,939,525
5.3 hpf	55,560,167	1,384,159	30,068,339	27070756	25,021,063	21,076,918

Table S2 (related to Fig 3). List of all genes validated by RT-qPCR and their primer sequences. Highlighted in yellow are genes from maternal super-cluster, non-highlighted are genes from pre-MBT super-cluster, while those in green are genes from zygotic super-cluster. *bnip3l* was used as reference.

Gene	**Forward primer**	**Reverse primer**
cldnd	CCAGAGCACCGGACAGATG	GGTCTTGAGGTAGAGCCAGCAT
ybx1	AGGAAGATGTCTTTGTGCATCAG	CCCAACGCTACGGAGATATTTC
ccnb1	TCGACTGGCTTGTGCAAGTCCA	GCCTTCCAAAACCAAAGTTGAGGA

cldng	TGCAAGCTCATCTGTAGGTCACA	CAAGAGTGCCTAGTGTTGTTCCA
plk1	GCCAAGCCGTCTGATAGAGACT	CAGGAATACAAGCAGGGTCTTCA
eml2	CATTTGTGTGCGGTTGACG	ACAGTCCTTGGCGTTTAGTGAGA
gtf2h4	GGTAGCCCTGTGGGTTAAAAAA	CCAGAGCCGTAGCCCAGTTA
taf1a	GAAAACGAGGAGCTGAGCAGTAA	AGAGTCACTGACGCCTGTCTGA
otx1	AGGAGAAGGACGGAGTGTTTTG	CCCCGATGTTGCAGTTTGAC
xbp1	TGGAGCTGGAGAATCAGAAACTT	CACTGAGCAGATCACTCGTCTTG
etv5	CCACCCCTGAACCGTGAA	TGAAAGTTGCCGCTGAAACC
dnaja2	TCCCAGACAACAACTGGTTGAG	GAGCATCGGCCCGTGTAG
ncl	AAGGAGGGCCTTGAAATTCAG	CAACGTAGCCGAATTTCTTTGTT
fam53b	TCGCTGGACCTGCTCAAGAT	CCTGGACAGCTGAGACTGTGAA
myst2	CCAGGACGCCAACAGGAA	GGCTGGATATGTCGATGTCTGA
nlk1	GGTGCCAAGTCCTGCTGAA	AAACGCCCCGTAACCGATA
zgc:158138	GGATGTTTGAGAAGATCGAGAAGTC	CAAGGCACGATCCATTAATAGCA
zgc:171708	TCACGACTGAACCTGCTTTCC	CTGGTTAATACCTGGCTCTGTGTCT
irf2a	AGAGAGCACATACAGTGAAGAAGAC	AAAGACAGGTCCTGAGGAAGC
dnmt3	TAATTCCTTCGCTACAGTAATGGC	AACATTCTACTGCGTCTGGAAGC
phf6	TCAGTTACTGGTGGCAGAGAGCC	GTGACCAAGGCTGATGAGAATAGC
aldoaa	CCCTGGCAAAGGCATCCTTG	CGGTTCTCCTCTGTGTTCTCAGC
rtkn2	AGGAAGAAAATAAGGGAAAGCATG	CTCTCATTCGCATCTCAAACTCC
hdlbp	GGACGGATTATCTGGTAACTGACC	TCCCACGTTTACACGGAGTCC
phc2	GAGGGATTCGTGATTCAGGAGG	TCGATCAGCAAGGACGGACG
lrpap1	AACATTGTCATGGACACTGTCAGC	ACATCCCCCTCCAGAGGACTG
herpud1	CTGGACAACCTGCTTATCAGAGATG	GGTGAAGAGTGGGTTTGGTATCTG
asz1	CAACGCTGACCCCAATGTCTG	CGAGCATCAAACACGTCATACTGC
gpsm2	GACATTTAGATATTGCCAGGGAAC	TTCCAAAATTGTAAAGTGCTCTGG
taf5l	GCACAGGAAACACCAGCTCTT	TCCTTCTGAACCGTCCACTGA
sf3b3	GGGTATCGTTGCCATTTCCA	CGGCACCCAGTTTCTCAAGA
bmi1	GAGAAGCAAAATGGATATACCTCC	TAATCCTTCAAAGGCTCATCCTC
gmcl1	TACAGCCACGACGACATCTAGC	ACTGCAAACCACTTTTCCACTGC
zw10	CTCTGTCGCCAGCAAAGCTATCAG	CTTTTCCCAACCTCCGCAGC
nudcd1	GTGTGTACTTTGTGGACAGCAAGG	TCGAAATGAAACATCTCACGAGG
her5	AGAAACTCAAAAACCCTAAAGTGGAA	GCACAACACTTTCCAGGATTTCA
apoeb	GCAGAGAGCTTGACACACTAATTACTG	GGGTTTGGAGATTTTCACTGTATGA
klf4	ACGAAAAGCTCCCATTTGAAAG	TGCAGTGGTAGGGCTTCTCA
sox19a	ACCGCACAGACCTACATGAATG	CTGTGGAGTGCTGCTATAGGACAT
id1	TGTGGAGAACGGCTGTTCAG	CCACCGGGAGTATCTGCTTTAAC
cirbp	TCGCATGATTCGCGTTGAT	GGAACCCCCTCTGAAACCA
hspb1	GGAGATCACTGGCAAACATGAG	TGGTGAAGCATCTGGAAATGAA
atf4	CGCTCTGCTGCCATCGA	GGAGGAGAAGCTGCGGTATTT
anxa1a	CGCCAAAGCCATTGACCTA	CATTTCACAACAGCAATCAAGCA
foxa3	CGCTCGGAGAGGAATAC	AGTGGAATCCTTTCTACAGTGAG
bnip3l	CGCATGGCAGACTGGTCGAGTA	TCCTCATGCTAAGAGTCACTGAACC

Table S3 (related to Fig 3). List of Genbank ID of clones used for WISH probe synthesis.

Gene	Clone number
taf1a	CK693445
otx1	CK692061
xbp1	CK712552.1
etv5	CK690842
apoeb	CK670601
klf4	EH604045
sox19a	CK691764
id1	CK711535
hspb1	CK711069
cldng	CK711236

Table S4 (related to Fig 3). List of gene-specific primers used for poly(A) tail measurements.

Gene	Forward primer	Reverse primer
xbp1	AACGTGACCCAGTATTTTTGTC	ATTATGATTGATGAGAGCATTTAC
otx1	GTGACGCACTCAAGGAATGG	TGTGTAACATCTCACATGAGCTG
taf1a	GTATATGTTCTTGAATATGCCAG	AGTGTGAGGATATACACAGATGC
gtf2h4	GCATTTTACTGCAAGTTGATCTG	ACCTTGGACACAAACGTGTAG
dnaja2	GTCAGCAATCCAGGTGCATG	TTCACAGAAGATAAGCCTAAAGG
ncl	GGATACAGACCGAAAACAGC	TCAAAATGGTTCTTTACATCATG
ybx1	TTGGGGAAACAGCAGATG	AAATGTCAAACTGACCAGATATC

Table S5 (related to Fig. 4). List of motifs enriched in pre-MBT cluster with degradation 1 cluster as discriminant, generated by DREME.

MOTIF	AAAAWAG	2.60E-18
MOTIF	GAAAANA	2.70E-11
MOTIF	CATTNCA	5.70E-10
MOTIF	RAAAAA	1.40E-07
MOTIF	GAACB	5.50E-09
MOTIF	ACAGMT	1.50E-06
MOTIF	TCTTTAMA	2.30E-06
MOTIF	AAAGKA	6.20E-06
MOTIF	STCRA	4.30E-06
MOTIF	GGGRS	1.70E-05

MOTIF	ATTHTG	1.80E-05
MOTIF	CHTAAC	1.10E-05
MOTIF	HTGTAGA	1.20E-04
MOTIF	ACKTCC	9.00E-05
MOTIF	TCCAGCCA	1.40E-04
MOTIF	TCCAGCCA	1.40E-04

Table S6 (related to Fig 5). Number of putative NTRs discovered by the WTAP 1.2 NTR module. Number of putative NTRs reduced significantly after intersection with other known annotations (see Supplementary Methods).

	Egg	1-cell	16-cell	128-cell	3.5 hpf	5.3 hpf
Putative Transcribed Regions	85987	82448	118351	115141	124610	122751
Putative Transcribed Regions with no annotation support	2329	2064	3899	3380	4590	8140

Supplementary Methods

Mapping of sequence tags and data analysis

Reference Genome assemblies and gene annotation were downloaded from Ensembl (Ensembl Genes 60, Zv9). Prior to aligning the sequence reads to the reference genome, reads were filtered against Repbase (Jurka et al. 2005) to remove repeats. mRNA-seq reads were mapped using Bioscope version 1.3.1 (Applied Biosystems) and the Whole Transcriptome Analysis Pipeline version 1.2 (WTAP 1.2) with default settings. Using uniquely mapped reads from Bioscope, read counts per exon were obtained using Bioscope count module and read counts per gene were obtained by summarizing the read counts for all exons within a gene for all Ensembl Gene IDs. Perl scripts were used to generate coverage data for visual display in the UCSC browser (*http://genome.ucsc.edu*). Novel Transcribed Regions (NTRs) were identified by the NTR Finder module in WTAP 1.2 using reads mapped by WTAP. We used a window size of 100bp and minimum window coverage of 4 as parameters. Newly identified NTRs were subsequently subjected to intersection with data from the UCSC genome browser (all mRNA and EST data) using BEDTools (Quinlan and Hall, 2010) to discover exons with no annotation support.

Data normalization and expression clustering

All analyses were performed in the R-environment (http://cran.r-project.org/) unless specified. We modified the normalization strategy presented by Robinson and Oshlack (2010) to obtain data adjusted for the amount of polyA RNA per embryo (see below for detailed description). Statistical tests were performed by successive pair wise comparisons using the R-package DEGseq with the method MARS (MA-plot-based method with random sampling model) (Wang et al. 2008). We applied cut-offs for both absolute change (read count change > 50), q-value (< 0.001) (Storey and Tibshirani 2003) and fold change (>2). Read counts for each transcript were subsequently length normalized by dividing against transcript length in kilobases. Using the genes within the thresholds from statistical testing we designed clusters using a combination of K-means clustering (k=30), subjectively merging similar groups, searching for transcripts with specific gene expression profiles and finally applying cut offs for correlation to the cluster mean (Pearson's correlation coefficient >0.90) and coverage (average of >50 reads across all samples). A total of 5,278 genes were included in the clustering analysis. Genes were subjected to functional annotation using Ingenuity Pathway Analysis (IPA; http://www.ingenuity.com). P-values were calculated using Fischer’s exact test and a threshold p-value of <0.01 was used to define significantly enriched terms.

Splice junction mapping & detection

For detection of splice variants we implemented the strategy proposed by Richard and colleagues (2010) included in the R-package SolaS. Due to the low-throughput layout of this package we reimplemented it in a custom R-script. We selected ~1600 multi-exon transcripts (≥2) with high read coverage in each library (>150 reads) without known isoforms. We removed exons below 50 bp and calculated a z-score for each exon. A low z-score is indicative of a exon skipping event or alternative 5' or 3' splice sites depending on exon position. For genes with known isoforms we used Cufflinks version 0.9.3 (Trapnell et al. 2010) to estimate the relative abundance of each isoform. We inspected the fraction value for a subset of genes expressed at all stages (>50 reads) and with multiple isoforms. We interpreted a low major fraction (<0.80) as evidence for the presence of multiple isoforms.

Motif search for cytoplasmic polyadenylation elements

Using BioMart (http://www.ensembl.org/biomart/), we downloaded the 3' UTRs of transcripts associated with the different genes. For the pre-MBT clusters we applied a stringent cut off (>500 reads increase and >5-fold increase) to reduce the computational load. We used three different programs in the MEME-suite; MEME (Multiple Em for Motif Elicitation) for *de novo* motif discovery, DREME (Discriminative Regular Expression Motif Elicitation) to look for differences between groups, and FIMO (Find Individual motif Occurrences) in the search of specific motifs. Shortest distance between any HEX and CPE motif was found using a custom python script.

Normalization of developmental data

Several approaches have been proposed to normalize RNA-Seq data (e.g. RPKM, median, quantile) (reviewed in Bullard et al., 2010). All of these assume that there is equal amounts of RNA in each experimental unit (per cell, per embryo, per organism). However, it has been pointed out previously that during development there are global shifts in the levels of mRNA (van de Peppel et al, 2003; Gilbert et al., 2009; Thelie et al., 2009; Evsikov and Evsikova, 2009). We measured the levels of total RNA and the polyA RNA fraction per embryo of each stage and observed substantial differences in the polyA content during development (Figure S.1C). A factor for each stage was calculated as follows:

$$Z_j = \sqrt{\frac{Y_j}{Y_r}}$$

Where Z_j is the estimated adjustment factor for sample j, Y_j is the amount of poly(A) RNA in sample j and Y_r is the amount of poly(A) RNA in the reference sample.

Furthermore, Robinson and Oshlack (2010) presented a method to asses global shifts in gene expression based on gene expression values (TMM, Trimmed Mean of M values), which is implemented in the excellent R package edgeR (Robinson et al., 2010). Basically this is a "global fold-change". We found that this estimated factor was in harmony with our experimentally derived factor (Figure S.1D).

Due to the uncertainty in the experimentally derived factor, and because the TMM was estimated only on known genes, we chose a robust middle path, taking the average of the two values. These were calculated relative to the egg levels and gave the following values; 1, 0.95, 1.21, 1.32, 1.38 and 1.15 for the egg, 1-cell, 16-cell, 128-cell, MBT and post-MBT respectively. These values represent a measurement of the poly(A) RNA levels at each stage relative to the egg stage.

The implication of this observation is obvious; we can not compare the respective stages without taking this into consideration. If we make an arbitrary example; Gene A is equally expressed in the egg and at the MBT stage, 10 transcripts (that is per embryo, per cell it is of course far less abundant since there is 999 more cells at the MBT stage). However, the total number of transcripts has increased from 1,000 to 1, 500. For simplicity, lets say that we get 1 read per transcript. If we normalize under the assumption of equal amounts of RNA per embryo we get the expression values (dividing total reads by 1000 for clarity):

Egg: 10/1 = 10

MBT: 10/1.5 = 7.5

The gene seems down regulated. However in a typical experimental design it is a bit more complicated than this. Using microarrays and RNA-Seq it is common to use the same amount of cDNA. However, the problem still persist. Following the example above, if we used 1000 transcripts from both samples these would come from a different number of eggs/embryos, and our gene A would be represented with 10 transcripts and 7.5 transcripts for the egg and MBT sample respectively.

In order to normalize our data we first estimated the relative abundance of each transcript as the number of reads for transcript i divided by the total number of reads in sample j.

$$[E_{ij}] = \frac{R_{ij}}{R_j} \qquad (1)$$

Where $[E_{ij}]$ is the relative abundance (or the probability) of transcript i in sample j , R_{ij} is the number of reads for gene i in sample j, and R_j is the total number of reads in sample j. If the samples are identical, changes of the $[E_{ij}]$ value would represent changes in expression levels. However, since we measured different amounts of poly(A) RNA in our samples we had to adjust for this. We took the average R_j for all samples to generate a common library size:

$$\overline{\mathrm{X}} = \frac{\sum_{j=n}^{j=1} R_j}{n} \qquad (2)$$

And made a pseudo library size by multiplying $\overline{X}$ with the calculated poly(A) content factor Z_j.

$$\mathrm{X}_j = \overline{\mathrm{X}} \times \mathrm{Z}_j \qquad (3)$$

Where Z_j was calculated as described above based on experimentally derived data as well as the TMM factor (for calculation of TMM, see Robinson and Oshlack, 2010).

Furthermore, we reassigned these pseudo read libraries to each gene based on the previously estimated $[E_{ij}]$ value to get the normalized dataset.

$$\mathrm{E}_{ij} = [E_{ij}] \times \mathrm{X}_j \qquad (4)$$

Our modified approach removes technical variation introduced by different amounts of input cDNA, but keeps the biological variation and returns a "per-embryo" estimate of the expression for a certain gene. This is a different implementation of the TMM factor suggested previously (Robinson and Oshlack, 2010), and more suitable for our type of data.

Let us extend the example from above; Gene A still have 10 copies at each of stages and the egg has 1,000 and MBT 1,500 transcripts in total, respectively. For sequencing 800 transcript were used from the egg stage and 1,200 from the MBT stage. Given our simple world conditions this would result in a total of 800 and 1,200 reads in total the egg and MBT samples respectively, and 8 reads for transcript A in both samples. Normalizing by the common RPKM (Reads Per Kilobase per Million) approach:

Egg: 8/0.8= 10

MBT: 8/1.2 = 6.67

The approach proposed by Robinson and Oshlack (2010) with our modification:

(1) Estimation of concentration

8/(800) = 0.01

8/(1,200) = 0.0066

(2 and 3) Calculation of common library size and pseudo library size:

Common library size = (800+1000)/2 = 900

Stage specific pseudo library size:

Egg: 900 x 1 = 900

MBT: 900 x 1.5 = 1350

(4) Generating pseudo-counts

egg: 0.01 x 900 = 9

MBT: 0.0067 x 1,350 = 9

Thus, this approach maintains the per embryo abundance of transcript A. In the original approach it was suggested to implement the TMM factor prior to estimation of the relative abundance of transcripts. This would in our case reinforce the bias at a per-embryo level, but would give a better view of the relative abundance of the transcripts.

It seems most methods of normalization is concentrating on making samples as similar as possible, disregarding the underlying biology of their samples. It is somewhat disturbing that it is only recently that the foundations of these approaches has been questioned (Gilbert et al., 2009; Thelie et al., 2009; Evsikov and Evsikova, 2009).

We validated our data using correlation with recently published microarray data (GSE20137; Lindeman et al., 2010). This were overall robust (r=0.72 and 0.75 for egg and MBT, respectively, for 9,020 genes). Moreover, expression profiles overall followed the same trends. 2,847 genes had a fold- change greater than 2 and a log2 intensity value of >3 in the microarray dataset (2,471 up-regulated genes and 376 down-regulated genes). Nonetheless, 309 genes (~10%) showed opposite expression patterns in our RNA-Seq dataset. Thus, whereas microarray data are consistent with essentially every differentially up-regulated gene (96% concordance), it only detected 45% concordance for down-regulated genes. This is consistent with the implication of assuming the same amounts of RNA across different stages, when there is in fact more poly(A) RNA in the MBT embryos (Evsikov and Evsikova, 2009); All genes at the MBT are underestimated. Importantly, these microarrays were only normalized within groups (quantile normalized). Hence the discrepancies stems from using the same amount of RNA from each sample. Performing a global normalization (e.g. quantile normalization) between groups further increased the bias of differential poly(A) RNA amounts per embryo.

To further validate our data, we examined recently reported RNA polymerase II (RNAPII) occupancy profiles during zebrafish development (Vastenhouw et al., 2010). We examined genes with low RNAPII occupancy at the MBT stage and assessed their expression values (RNA-Seq) post-MBT relative to MBT. With our normalization method, 164/219 genes were detected as down-regulated relative to MBT. Using RPKM normalization, 145/219 genes were detected as down-regulated. Many of the remaining genes were expressed at low levels (27 genes with <50 reads), and/or exhibited minor changes relative to their abundance, or showed discrepancy (14 genes) (data not shown). This discordance may be explained by the relatively long time span between MBT and post-MBT stages (2.3 h). Additional validation of our RNA-Sseq data was provided by quantitative RT-PCR in a developmental context (see main text).

Our approach is simple and do not address technical biases such as transcript length and base content. Or as proposed in the work by (Robinson and Oshlack, 2010), that high abundance gene are taking up sequencing for low abundant transcripts. Further work is needed to sort out these issues.

We hope that this approach will facilitate studies of both development and cross-tissue comparison, and be particular beneficial in studies where the transcriptional activity of specific loci is of importance (e.g. effects of histone modifications).

Paired end di-tag preparation and analysis

The process of Gene Identification Signature (GIS) analysis has been thoroughly decribed (Ng et al., 2005). The complete procedure comprises the construction of GIS full-length cDNA library followed by ditagging which converts the library to a GIS paired-end ditag (PET) library. In principle, 18 nucleotides 5' and 3' signatures of full-length cDNAs were extracted into PETs that are concatenated for efficient sequencing.A residual AA dinucleotide derived from the mRNA poly(A) tail was also included to determine orientation of the ditag. These were subsequently mapped to the genome sequences, allowing a demarcation of transcription boundaries of genes.

Supplementary Text

Cytoplasmic polyadenylation elements are found in the 3'UTR of pre-MBT transcripts

It is generally accepted that regulation of cytoplasmic polyadenylation is mediated through *cis* elements in the 3'UTR of mRNAs. The most well studied include the nuclear polyadenylation element (HEX; A(A/U)UAAA), and the cytoplasmic polyadenylation element (CPE; UUUU(U/A)AU). In addition a dodecauridine (U_{12}) motif, termed embryonic CPE (eCPE), has been described in *Xenopu*s (Simon et al., 1992; Simon and Richter 1994).

We used programs provided in the MEME-suite (Bailey and Elkan, 1994) to find *cis* elements involved in delayed cytoplasmic polyadenylation in pre-MBT transcripts. *De-novo* motif discovery with MEME (Multiple Em for Motif Elicitations) found a long U-rich motif (20-mer) as the most significant (Fig. 4H), which resembles the eCPE reported in *Xenopus* (Simon et al., 1992, Simon and Richter, 1994).

Next, we applied DREME (Discriminative Regular Expression Motif Elicitation), which can find short motifs (≤ 8 bp) enriched relative to a discriminative group. We used the degradation 1 cluster as the discriminator and identified 15 motifs as significant (Table S5) including several A-rich and U-rich motifs. The top ranking motif was AAAA(A/U)AG ($p = 2.6x10^{-18}$).

To test whether the known Hex and CPE motifs were present and at what frequency they occur we applied FIMO (Find Individual Motif Occurrences) on three different groups; pre-MBT, the maternal-zygotic, and a set of randomly chosen genes. Surprisingly, these two elements were present in all, without overrepresentation in any of the groups. We also looked for a 12-mer U-rich motif (allowing 1 mismatch), and found that 33, 24 and 15% of the transcripts harbored this in the pre-MBT, maternal-zygotic and random group, respectively.

Analysis of splice variants

Through splice junction mapping we detected additional splice variants. Tumor protein 53 (*tp53*) has eleven annotated exons and is known to have splice variants (Khoury and Bourdon 2010). We observed reads spanning across exon 2, exons 6, 7 and 8, and exon 7 (Fig. 7C). We found that exon 2, 7 and 9 had low z-values. This implies that the zebrafish *tp53* has several isoforms (human TP53 expresses 9 isoforms) present simultaneously during development, and that both current strategies for splice variant detection are needed to identify all isoforms: splice junction mappings for specificity and exon read counts for sensitivity. Another example of alternative splicing pattern was observed in *gli2a.* Gli proteins activate or repress transcription depending on the length of its C-terminus (Ruel and Therond 2009). We observed during development, a shift in the expression of the *gli2a* 3′ distal exon (exon 14; Fig. S8B) coding for the C-terminus corresponding to its transactivation domain. Inclusion of exon 14 and the extended C-terminus could result in the replacement of *gli2a* transcript encoding the transcriptional repressor by a longer version encoding a transcriptional activator, which might be further processed into repressor post-translationally. Interestingly, the appearance of exon 14 isoform in *gli2a* is paralleled with a decrease in transcripts level of its suppressor *sufu* (Fig. S8C), as well as transcriptional onset of a number its target genes (Fig. S8C), suggesting its activity during MBT in zebrafish.

Supplemental References

Bullard, J.H., Purdom, E., Hansen, K.D., and Dudoit, S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11, 94.

Evsikov, A.V., and Marin do Evsikova, C. (2009) Gene expression during the oocyte-to-embryo transition in mammals. Mol Reprod Dev. 76 (9), 805-18.

Gilbert, I., Scantland, S., Sylvestre, E.L., Gravel, C., Laflamme, I., Sirard, M.A., and Robert, C. (2009) The dynamics of gene products fluctuation during bovine pre-hatching development. Mol Reprod Dev. 76 (8), 762-72.

Jurka, J., V.V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res **110:** 462-467.

Khoury, M.P. and J.C. Bourdon. 2010. The isoforms of the p53 protein. Cold Spring Harb Perspect Biol **2:** a000927.

Ng P, Wei CL, Sung WK, Chiu KP, Lipovich L, Ang CC, Gupta S, Shahab A, Ridwan A, Wong CH, Liu ET, Ruan Y. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. Nat Methods. 2005 Feb;2(2):105-11. Epub 2005 Jan 9.

van de Peppel J, Kemmeren P, van Bakel H, Radonjic M, van Leenen D, Holstege FC. Monitoring global messenger RNA changes in externally controlled microarray experiments. EMBO Rep. 2003 Apr;4(4):387-93.

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26:841-2.

Ruel, L. and P.P. Therond. 2009. Variations in Hedgehog signaling: divergence and perpetuation in Sufu regulation of Gli. Genes Dev 23: 1843-1848.

Storey, J.D. and R. Tibshirani. 2003. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100: 9440-9445.

Thelie, A., Papillier, P., Perreau, C., Uzbekova, S., Hennequet-Antier, C., and Dalbies-Tran, R. (2009) Regulation of bovine oocyte-specific transcripts during in vitro oocyte maturation and after maternal-embryonic transition analyzed using a transcriptomic approach. Mol Reprod Dev. 76, 773-82.