Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis

Andrea Pauli,[1,7,8] Eivind Valen,[2,7] Michael F. Lin,[3,4] Manuel Garber,[4]
Nadine L. Vastenhouw,[1] Joshua Z. Levin,[4] Lin Fan,[4] Albin Sandelin,[2]
John L. Rinn,[4,5] Aviv Regev,[3,4,6,8] and Alexander F. Schier[1,4,8]

[1]Department of Molecular and Cellular Biology (MCB), Harvard University, Cambridge, Massachusetts 02138, USA; [2]The Bioinformatics Centre, Department of Biology and the Biotech, Research and Innovation Centre (BRIC), University of Copenhagen, Copenhagen DK-2200, Denmark; [3]Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts 02139, USA; [4]The Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA; [5]Department of Stem Cell and Regenerative Biology (SCRB), Harvard University, Cambridge, Massachusetts 02138, USA; [6]Howard Hughes Medical Institute (HHMI), Chevy Chase, Maryland 20815, USA

Long noncoding RNAs (lncRNAs) comprise a diverse class of transcripts that structurally resemble mRNAs but do not encode proteins. Recent genome-wide studies in humans and the mouse have annotated lncRNAs expressed in cell lines and adult tissues, but a systematic analysis of lncRNAs expressed during vertebrate embryogenesis has been elusive. To identify lncRNAs with potential functions in vertebrate embryogenesis, we performed a time-series of RNA-seq experiments at eight stages during early zebrafish development. We reconstructed 56,535 high-confidence transcripts in 28,912 loci, recovering the vast majority of expressed RefSeq transcripts while identifying thousands of novel isoforms and expressed loci. We defined a stringent set of 1133 noncoding multi-exonic transcripts expressed during embryogenesis. These include long intergenic ncRNAs (lincRNAs), intronic overlapping lncRNAs, exonic antisense overlapping lncRNAs, and precursors for small RNAs (sRNAs). Zebrafish lncRNAs share many of the characteristics of their mammalian counterparts: relatively short length, low exon number, low expression, and conservation levels comparable to that of introns. Subsets of lncRNAs carry chromatin signatures characteristic of genes with developmental functions. The temporal expression profile of lncRNAs revealed two novel properties: lncRNAs are expressed in narrower time windows than are protein-coding genes and are specifically enriched in early-stage embryos. In addition, several lncRNAs show tissue-specific expression and distinct subcellular localization patterns. Integrative computational analyses associated individual lncRNAs with specific pathways and functions, ranging from cell cycle regulation to morphogenesis. Our study provides the first systematic identification of lncRNAs in a vertebrate embryo and forms the foundation for future genetic, genomic, and evolutionary studies.

[Supplemental material is available for this article.]

Large-scale genomic studies have identified a significant number of transcripts that do not code for proteins (Kapranov et al. 2002, 2007; Bertone 2004; Carninci et al. 2005; ENCODE Project Consortium et al. 2007; Ponjavic et al. 2007; Fejes-Toth et al. 2009; Guttman et al. 2009, 2010; Cabili et al. 2011). Such noncoding RNAs (ncRNAs) can be broadly classified as either small (<200 nucleotides [nt]; sRNAs) or large (>200 nt; lncRNAs) based on the size of their mature transcripts. While miRNAs (microRNAs), the best-studied class of sRNAs, regulate their mRNA targets post-transcriptionally (Bartel 2009), mRNA-like lncRNAs act by a range of mechanisms (for reviews, see Koziol and Rinn 2010; Pauli et al. 2011; Wang and Chang 2011). For example, several lncRNAs have been shown to interact with and modulate the activity of the chromatin modifying machinery (Rinn et al. 2007; Nagano et al. 2008; Pandey et al. 2008; Zhao et al. 2008, 2010; Khalil et al. 2009; Huarte et al. 2010; Tian et al. 2010; Tsai et al. 2010; Guttman et al. 2011; Wang et al. 2011). Other lncRNAs may act as decoys in the sequestration of miRNAs (Poliseno et al. 2010), transcription factors (Hung et al. 2011), or other proteins (Tripathi et al. 2010). Yet others may serve as precursors for the generation of sRNAs (Kapranov et al. 2007; Wilusz et al. 2008; Fejes-Toth et al. 2009).

Although most lncRNAs have not been functionally characterized, an emerging theme is their role in the regulation of gene expression in either *cis* or *trans*. Several *trans*-acting lncRNAs have been identified, including *HOTAIR* (Rinn et al. 2007), *TP53COR1* (also known as *lincRNA-p21*) (Huarte et al. 2010), and *PANDA* (Hung et al. 2011). Moreover, knockdown of more than 100 individual long intergenic ncRNAs (lincRNAs) in mouse embryonic stem cells led to widespread changes in gene expression that could not be explained by a *cis*-acting mechanism (Guttman et al. 2011). Other well-described lncRNAs act in *cis*. For example, mammalian X chromosome inactivation and allelic imprinting depend on lncRNAs that mediate the silencing of neighboring genes by recruiting repressive chromatin modifiers (Sleutels et al. 2002; Mancini-Dinardo et al. 2006; Nagano et al. 2008; Pandey et al. 2008; Zhao et al. 2008). Additional recently identified *cis*-acting RNAs activate the expression of neighboring genes (Kim et al. 2010; Ørom et al. 2010; Wang et al. 2011). Collectively, these studies have demonstrated that lncRNAs can have a profound impact on gene regulation in both *cis* and *trans*.

[7]These authors contributed equally to this work.
[8]Corresponding authors.
E-mail pauli@fas.harvard.edu.
E-mail schier@fas.harvard.edu.
E-mail aregev@broad.mit.edu.

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.133009.111.

22:577–591 © 2012 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/12; www.genome.org

Existing annotations of mammalian lncRNAs are derived from large-scale studies of cultured cells (Kapranov et al. 2002; Rinn et al. 2003; Carninci et al. 2005; ENCODE Project Consortium et al. 2007; Dinger et al. 2008; Guttman et al. 2009, 2010) or adult tissue samples (Ponjavic et al. 2009; Cabili et al. 2011). Such relatively homogenous and abundant samples have facilitated the identification of low abundance, cell type–specific transcripts. However, this strategy is likely to miss lncRNAs that are only expressed during narrow developmental time windows. To fully characterize vertebrate lncRNAs, it is therefore necessary to systematically search for lncRNAs that are expressed during specific developmental stages.

Here, we report the systematic identification and characterization of developmental lncRNAs. We leveraged the ability to obtain large numbers of developmentally synchronous zebrafish embryos in order to perform a time-series of eight RNA-seq experiments (200–300 million reads per stage) from shortly after fertilization to early larval stages. As a measure of quality of our data set, we were able to reconstruct the vast majority of annotated zebrafish RefSeq genes and a large fraction of Ensembl gene models (Flicek et al. 2011). In contrast to recent smaller-scale RNA-seq studies that focused on protein-coding genes (Aanes et al. 2011; Vesterlund et al. 2011), we annotated and analyzed lncRNAs at high temporal resolution. We combined RNA-seq–based de novo transcript identification with a stringent filtering of putative protein-coding transcripts to define a high-confidence set of 1133 multi-exonic noncoding transcripts. Our lncRNA catalog includes 397 intergenic, 184 intronic overlapping, and 566 antisense exonic overlapping transcripts, many of which are expressed in a developmentally regulated manner. We characterized each lncRNA by diverse features, including transcript structure, evolutionary conservation, developmental expression, and associated chromatin marks. Our expression pattern data revealed several intriguing properties of zebrafish lncRNAs. Notably, lncRNAs are expressed in particularly narrow developmental windows and in specific cell types. Moreover, lncRNAs are particularly numerous in the very early embryo. Computational analysis of expression correlation with functional gene sets associated subsets of lncRNAs with developmental processes ranging from cell cycle regulation to morphogenesis. Collectively, the systematic annotation and characterization of lncRNAs expressed during zebrafish embryogenesis opens the way for future genetic, genomic, and evolutionary studies.

Results

Assembly of a high-confidence embryonic transcriptome

To systematically discover noncoding transcripts with potential functions during early vertebrate development, we performed large-scale cDNA sequencing experiments across zebrafish embryogenesis. We chose eight time-points that mark important developmental stages (Fig. 1A): (1) shortly after fertilization (two- to four-cell stage); (2) at the time when zygotic transcription of the genome is initiated (1000-cell stage); (3–5) during blastula and gastrula stages (dome, shield, and bud stages), when cell fates are specified and large-scale cell movements occur; and (6–8) at late embryonic and early larval developmental stages, when organs are forming (28 h post fertilization [hpf], 48 hpf, and 120 hpf) (see overview in Fig. 1A). Polyadenylated RNA was purified from approximately 1000 embryos per time-point and converted into cDNA libraries for strand-specific, paired-end 76 bp sequencing on Illumina's HiSeq platform (see Methods; Parkhomchuk et al. 2009; Levin et al. 2010). On average, we obtained about 200–300 million reads per stage (more than two billion reads in total) (Supplemental Table 1). Eighty-eight percent of the reads passed initial quality thresholds, of which ~80% could be aligned to the latest assembly (Zv9) of the zebrafish genome sequence (Methods; Supplemental Table 1).

We assembled transcripts using a step-wise protocol (Methods; Fig. 1A). Briefly, we used TopHat (Trapnell et al. 2009) to align all reads per time-point, including those that span splice junctions. We then reconstructed transcripts using two assemblers—Cufflinks (Trapnell et al. 2010) and Scripture (Guttman et al. 2010), resulting in the assembly of a total number of 316,373 nonredundant transcript isoforms from 143,626 loci across all embryonic stages.

We defined a "high-confidence" set of 56,535 embryonic transcripts, following a similar strategy as described by Cabili et al. (2011). Specifically, we developed a filtering pipeline aimed at reducing the number of transcripts that might be erroneously assembled or below significance thresholds (Supplemental Fig. 1A). We first required a transcript be assembled at least twice: either identified by both assemblers or in at least two developmental stages. Next, we removed transcripts there were likely to be assembly artifacts or run-on fragments or that did not pass our high-confidence thresholds (Methods; Supplemental Fig. 1A). This resulted in a final set of 56,535 embryonic transcripts from 28,912 loci (on average, 1.95 transcripts per locus) (Supplemental Fig. 1B,C), of which 50,904 were multi-exonic and 5631 were single exon transcripts. We will henceforth refer to this set as the "embryonic transcriptome," and all subsequent analyses are based on it.

The embryonic transcriptome has high coverage, quality, and depth

To estimate the quality and coverage of our embryonic transcriptome, we compared it to the current RefSeq and Ensembl gene annotations. Compared to the 15,175 zebrafish RefSeq genes, our embryonic transcriptome provides more than a threefold increase in the number of identified transcripts (56,535) from nearly twice as many loci (28,912), suggesting that the increase in the number of individual transcripts is due to both novel isoforms of known genes and novel loci (Fig. 1B). Notably, of the 13,942 RefSeq genes that are expressed (FPKM [fragments per kilobase of exon per million fragments mapped] > 0) during the stages covered by our data set, 90% (12,527/13,942) have transcript evidence (exonic overlapping transcripts) in our embryonic transcriptome, and 70% of those (8751/12,527) are identical to RefSeq isoforms (Supplemental Fig. 2A). In addition, 3532 of our transcripts are variants of known RefSeq genes (novel isoforms and partial transcripts), many of which extend the existing exon–intron structures with additional 5′ or 3′ exons (Supplemental Fig. 2A).

Compared to the most recent Ensembl Zv9 gene models (52,873 transcripts in 31,711 loci), our embryonic transcriptome is of similar size (Supplemental Fig. 2B). The Ensembl gene set integrates transcript annotations from several sources and includes RNA-seq transcript models built from a total of 376 million reads derived from embryonic, larval, and adult zebrafish (Sanger Institute). Thus, the comparable transcriptome size and larger read numbers in our purely embryonic RNA-seq experiments again suggest that our embryonic transcriptome is of high depth. Moreover, we have transcript evidence for ~74% of Ensembl gene loci of comparable (>160 nt) transcript sizes, corresponding to ~68% of our embryonically identified loci (Fig. 1B). This high degree of overlap provides independent confirmation for a large fraction of our transcriptome.

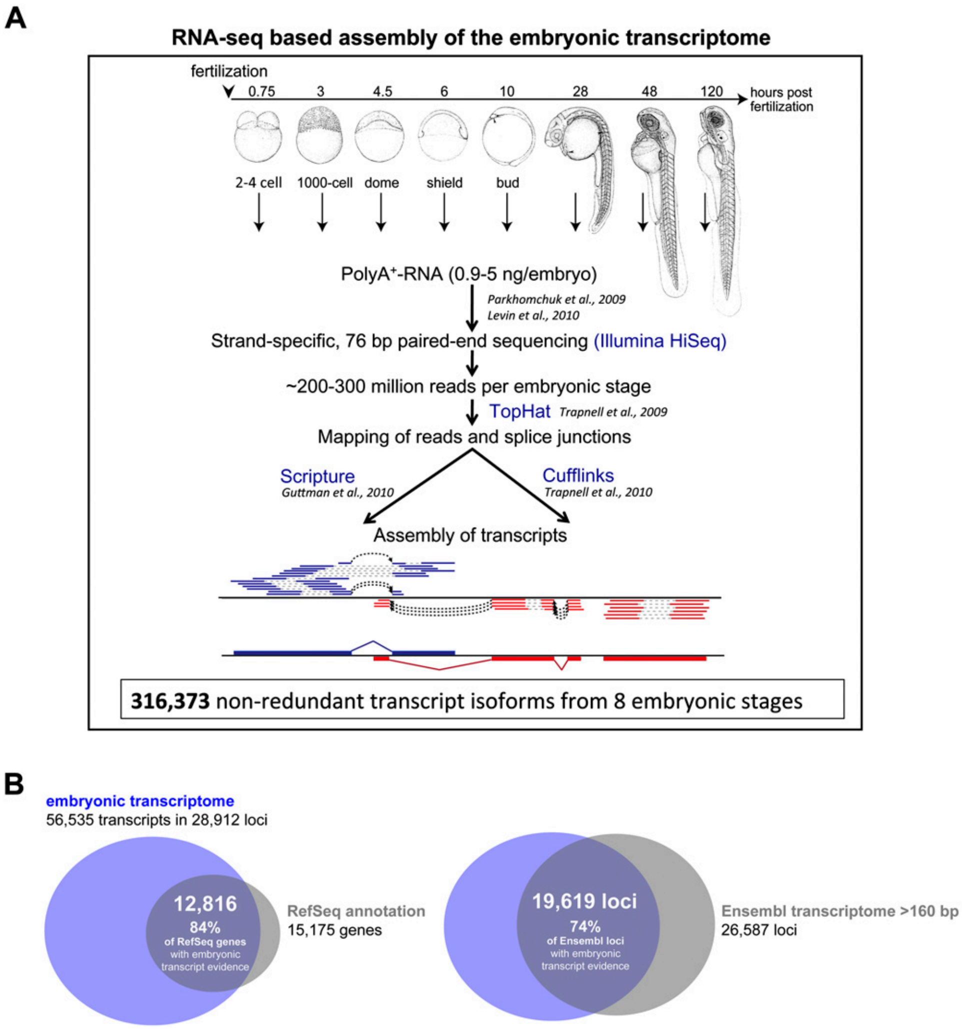


Figure 1. Overview of the RNA-seq–based embryonic transcriptome assembly. (*A*) Overview of the RNA-seq–based transcript reconstruction pipeline that was employed to identify embryonically expressed transcripts in zebrafish. Stage-specific transcriptomes were reconstructed from a time-series of eight embryonic stages: two to four cell, 1000 cell, dome, shield, bud, 28 h post fertilization (hpf), 48 hpf, and 120 hpf. Stage-specific drawings of representative embryos are adapted from Kimmel et al. (1995) (with permission from Wiley © 1995). A schematic outline of the process of transcriptome reconstruction is shown at the *bottom* for three genes. Reads were mapped to either the + (blue) or – (red) strand using TopHat. Gaps inferred from mapping each of the two paired-end reads are indicated as dashed gray lines; dashed black arrows indicate splice-junctions inferred from a gap in mapping of a single read; and the deduced final transcript structures reconstructed by Scripture or Cufflinks are depicted at the *bottom*. (*B*) Overlap between loci from the RNA-seq–based embryonic transcriptome assembly (blue) and previously annotated genes (gray): RefSeq genes (*left*) and Ensembl loci >160 bp (*right*). The majority of known loci (84% of RefSeq loci and 74% of Ensembl loci >160 bp) are recovered in the embryonic transcriptome. Note that the number of loci in the Ensembl transcriptome is based on comparison with loci of the embryonic transcriptome (which were used as reference), which reduced the number of 27,751 Ensembl loci (>160 bp) to 26,587.

Compared to two recent RNA-seq–based transcriptome studies in zebrafish embryos (Aanes et al. 2011; Vesterlund et al. 2011), our data set is of significantly higher depth: a total of about 220 million (Vesterlund et al. 2011) and about 100 million (Aanes et al. 2011) mapped reads from four and six embryonic stages, respectively, versus about 1.5 billion mapped reads in our study (Supplemental Table 1). Moreover, we identify many more known and novel transcribed loci: about 4000 "novel transcribed regions" reported by Vesterlund et al. (2011) and Aanes et al. (2011) versus more than 9000 novel loci in our embryonic transcriptome with no previous annotations in RefSeq or Ensembl. This suggests that our embryonic transcriptome provides a highly comprehensive and more complete assembly than that previously available.

Our transcript assemblies are also consistent with chromatin marks known to be associated with promoters (Zhou et al. 2011) (see also below). The fraction of marked protein-coding loci of our embryonic transcriptome (44% for H3K4me3 only, 19% for H3K4me3 and H3K27me3) is nearly identical to the fraction of marked RefSeq loci (46% for H3K4me3 only, 16% for H3K4me3 and H3K27me3) (see Fig. 5). This suggests that (1) our embryonic transcriptome is of a quality comparable to RefSeq genes, and (2) many of our RNA-seq–based transcript structures contain complete 5′ ends.

Identification of a stringent set of embryonic lncRNAs

To identify mRNAs that exert their biological function as lncRNAs, we developed a highly stringent filtering pipeline aimed at removing transcripts with evidence for protein-coding potential (Methods; Fig. 2). We identified putative lncRNAs by considering their phylogenetic conservation across species, homology with known proteins and protein domains, and potential ORFs.

Four filters were used. First, we used PhyloCSF (phylogenetic coding substitution frequency; see Methods), to score the coding potential of transcripts using phylogenetic alignments (Lin et al. 2011). PhyloCSF exploits the fact that protein-coding sequences—but not lncRNAs and other sequences—tend to have a higher rate of synonymous versus nonsynonymous substitutions (Supplemental Figs. 3, 4A). We chose a PhyloCSF threshold of less than 20 because it retained the majority of RefSeq ncRNAs (Supplemental Figs. 3, 4A) but removed 96.2% of protein-coding RefSeq transcripts. This filter retained 4867 putative noncoding transcripts (Fig. 2).

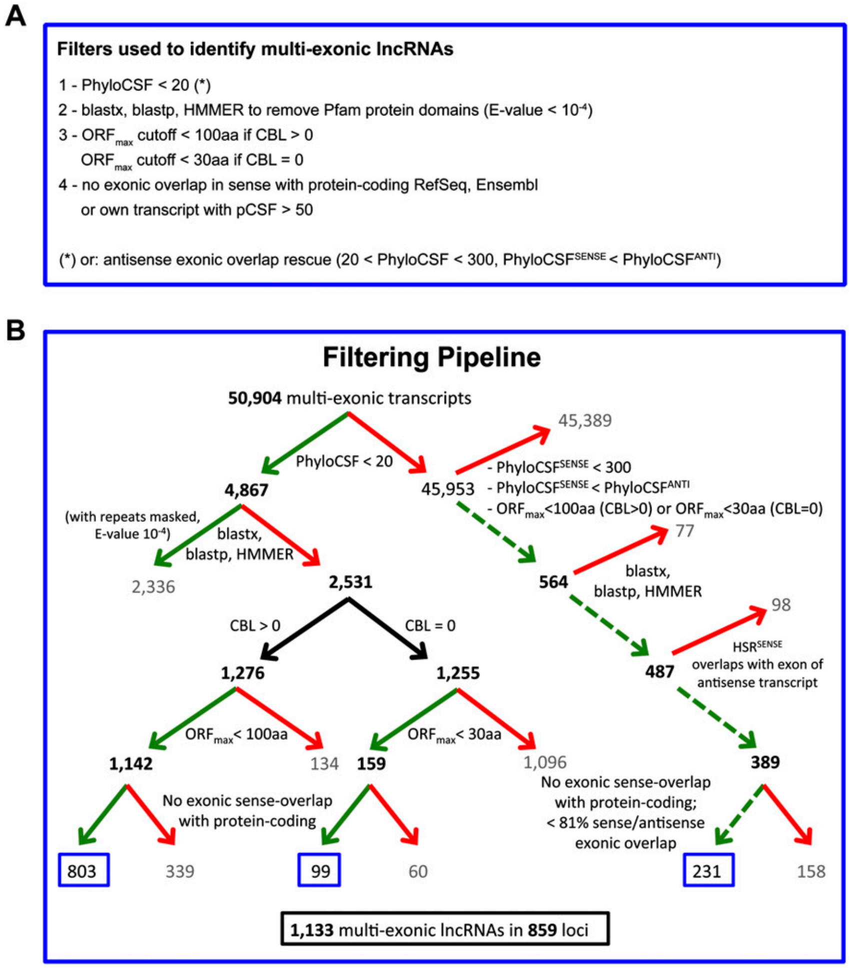

A
Filters used to identify multi-exonic lncRNAs
1 - PhyloCSF < 20 (*)
2 - blastx, blastp, HMMER to remove Pfam protein domains (E-value < 10-4)
3 - ORFmax cutoff < 100aa if CBL > 0
ORFmax cutoff < 30aa if CBL = 0
4 - no exonic overlap in sense with protein-coding RefSeq, Ensembl
or own transcript with pCSF > 50
(*) or: antisense exonic overlap rescue (20 < PhyloCSF < 300, PhyloCSFSENSE < PhyloCSFANTI)
B
Filtering Pipeline
50,904 multi-exonic transcripts
PhyloCSF < 20
4,867
45,953
45,389
- PhyloCSFSENSE < 300
- PhyloCSFSENSE < PhyloCSFANTI
- ORFmax<100aa (CBL>0) or ORFmax<30aa (CBL=0)
(with repeats masked,
E-value 10-4)
blastx,
blastp, HMMER
2,336
2,531
564
77
blastx,
blastp, HMMER
98
CBL > 0
CBL = 0
HSRSENSE
overlaps with exon of
antisense transcript
1,276
1,255
487
ORFmax< 100aa
ORFmax< 30aa
1,142
134
159
1,096
389
No exonic sense-overlap
with protein-coding
No exonic sense-overlap
with protein-coding;
< 81% sense/antisense
exonic overlap
803
339
99
60
231
158
1,133 multi-exonic lncRNAs in 859 loci

Figure 2. Overview of the stringent filtering pipeline that defined a conservative set of 1,133 lncRNAs. (*A*) Filters at a glance: overview of classification criteria used to define noncoding transcripts. (*B*) Detailed outline of the filtering pipeline that defined a conservative set of 1133 multi-exonic, embryonically expressed lncRNAs. The following filtering criteria were used: (1) Phylogenetic Codon Substitution Frequency (PhyloCSF) score <20 (*left* branch of the *top* node) or rescue by the antisense pipeline (*right* branch of the *top* node [dashed lines]: $\text{PhyloCSF}^{\text{sense}} < 300$ and $\text{PhyloCSF}^{\text{sense}} < \text{PhyloCSF}^{\text{anti}}$ and highest scoring region [HSR] overlapping with an exon on the opposite strand); (2) no known protein homologs based on blastx, blastp, and HMMER; (3) maximal ORF (ORFmax) <100 aa (transcripts with alignments [complete branch length (CBL) > 0]) or <30 aa (transcripts without alignments [CBL = 0]); and (4) no sense-overlap with any protein-coding transcript. At each step, a green arrow denotes the transcripts that passed the filter; a red arrow, those that were removed. Black bold numbers indicate the number of transcripts that passed the filter. Blue boxes highlight the number of transcripts that passed all filters and are considered noncoding (1133 lncRNAs in 859 loci).

Second, we removed transcripts that had similarity to known proteins or protein domains based on blastx, blastp, and HMMER (Pfam domains) (Eddy 2009). This filter retained 2531 putative noncoding transcripts (Fig. 2B). The excluded transcripts had not been captured by the PhyloCSF filter because they typically received low PhyloCSF scores due to poorly aligned sequences (complete branch lengths [CBLs] of zero) (Methods; Supplemental Fig. 4B).

Third, we removed any remaining transcript of uncertain coding potential by applying a maximal ORF filter. Consistent with the traditional cutoff for protein-coding transcripts (Okazaki et al. 2002), we excluded any transcript with a maximal ORF > 100 amino acids (aa). For transcripts that were not scored by PhyloCSF due to lacking sequence alignments (CBL = 0), we used a more stringent maximal ORF cutoff of 30 aa. The ORF filter retained 1301 transcripts.

Finally, to exclude potentially incomplete transcript structures, we removed any transcript that had sense exonic overlap with a protein-coding transcript. The resulting set contained 902 lncRNAs (mean PhyloCSF score of 5) (Fig. 2B; Supplemental Fig. 3).

Identification of antisense overlapping embryonic lncRNAs

Some putative noncoding transcripts had antisense exonic overlap with protein-coding genes. Examination of the range of PhyloCSF scores obtained for antisense strands of sense-coding transcripts revealed that transcripts with a high-scoring sense strand also tended to score relatively high on the antisense strand (Supplemental Fig. 4C). Thus, PhyloCSF scores of antisense exonic overlapping transcripts can be confounded by high coding potential on the opposite strand.

To address this issue and "rescue" noncoding antisense transcripts, we employed a modified filtering pipeline with four additional criteria (Fig. 2; for details, see Methods): (1) The putative noncoding transcript had a lower PhyloCSF score than the overlapping coding transcript; (2) its highest PhyloCSF score was obtained in the region of overlap (e.g., Supplemental Fig. 4D); (3) its PhyloCSF score was less than 300; and (4) the sense/antisense exonic overlap did not exceed 81% of the sense strand. This approach retained 231 multi-exonic antisense transcripts and resulted in a final stringent set of 1133 lncRNAs (Fig. 2B).

Genomic characterization of embryonic lncRNAs

According to their genomic location, our 1133 embryonic lncRNAs are partitioned into 397 lincRNAs without overlap with any genes, 184 intronic overlapping lncRNAs, and 566 antisense exonic overlapping lncRNAs (Fig. 3). Intronic overlapping lncRNAs are defined as loci that have no exon–exon overlap with another locus, i.e., there is no overlap between the mature lncRNA with exons of the overlapping locus. Intronic overlapping lncRNAs are in either sense or antisense orientation with respect to the overlapping gene and can be further partitioned into 105 intronic contained lncRNAs (incs; the lncRNA is contained within the transcribed region of another locus), 60 completely overlapping lncRNAs (concs; the other locus is contained within the transcribed region of the lncRNA locus), and 19 partially overlapping lncRNAs (poncs; neither incs nor poncs but with at least one exon of the lncRNA contained within an intron of another locus).

Some lncRNAs may function as precursors for the generation of sRNAs (ENCODE Project Consortium et al. 2007; Wilusz et al. 2008). To identify sRNA-precursor lncRNAs, we compared our lncRNA transcripts to a set of sRNAs present in 2-d-old zebrafish (Methods; Cifuentes et al. 2010). We identified 41 lncRNAs that appear to function as precursors for the production of miRNAs (16), snoRNAs (nine), or sRNAs of unknown categories (20) (Supplemental Table 2). Four lncRNAs of the latter category contained a vast number of sRNAs throughout the entire transcript. For example, the zebrafish ortholog of the abundant nuclear lncRNA *MALAT1* (also called *NEAT2*) was cleaved throughout its transcript and gave rise to multiple sRNAs (Supplemental Fig. 5). Consistent with this observation, *MALAT1* has previously been shown to be associated with Ago2 (also known as EIF2C2), a known component of the sRNA processing machinery (Weinmann et al. 2009). This analysis indicates that the large majority of our lncRNAs are not processed into sRNAs.

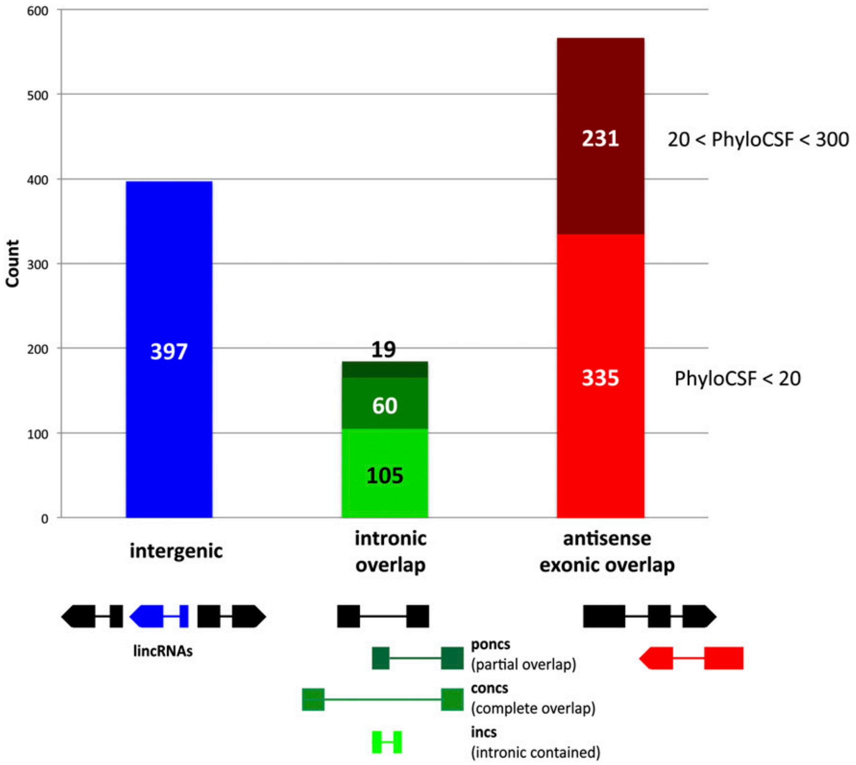


Figure 3. Classification of lncRNAs. Numbers of lncRNAs in each of the three main classes, as defined by their genomic location relative to neighboring or overlapping genes. Intergenic lncRNAs (blue; lincRNAs) have no overlap with any gene. lncRNAs with intronic overlap (green) are defined as loci that have overlap with another transcribed locus but no exon–exon overlap (no overlap between the mature lncRNA transcript with exons of the overlapping locus). They are on either the same or the opposite strand relative to the overlapping gene and can be partitioned into intronic contained lncRNAs (incs, light green; the lncRNA is contained within the transcribed region of another locus), completely overlapping lncRNAs (concs, green; the other locus is contained within the transcribed region of the lncRNA locus), and partially overlapping lncRNAs (poncs, dark green; neither inc nor conc, but at least one exon of the lncRNA has overlap with an intron of another locus). LncRNAs with antisense exonic overlap (red) have at least one exon that overlaps with an exon of a protein-coding transcript on the opposite strand; they can be partitioned into those identified via the general pipeline (PhyloCSF < 20, light red) and those rescued via the antisense pipeline (20 < PhyloCSF < 300, dark red). A scheme of the position of the lncRNA gene (in color) relative to neighboring or overlapping gene(s) (black) is shown at the *bottom*.

Zebrafish lncRNAs are shorter, less conserved and expressed at lower levels than are protein-coding genes

Previous studies in mammals have shown that lncRNAs are shorter, less conserved, and expressed at significantly lower levels than are protein-coding genes (Guttman et al. 2010; Cabili et al. 2011). To determine whether embryonic lncRNAs have similar features, we analyzed the structure, expression level, and conservation of our lncRNAs (Fig. 4). We found that zebrafish lncRNAs were on average about one-third of the length of protein-coding transcripts (mean length of 1113 nt for lncRNAs versus 3352 nt for coding transcripts) (Fig. 4Aa). Moreover, lncRNAs had fewer exons per transcript (about 2.8) than the average protein-coding gene (about 11) (Fig. 4Ab). These properties are comparable to the estimated transcript length and exon number of human lincRNAs (on average, ~1 kb and 2.9 exons, respectively) (Cabili et al. 2011). Notably, zebrafish embryonic lncRNAs were expressed on average at about 10-fold lower levels than protein-coding genes (Fig. 4B), consistent with the low expression levels of their mammalian counterparts (Guttman et al. 2010; Cabili et al. 2011).

To assess the level of conservation of lncRNAs, we used the CBL score, a measure of the fraction of phylogenetic teleost alignments present over the region of interest (Methods). In agreement with signatures of conservation in mammalian lncRNAs (Ponjavic et al. 2007, 2009; Guttman et al. 2009, 2010; Ørom et al. 2010), a few lncRNAs were clearly conserved across fish species (Fig. 4C; for two conserved examples, see Supplemental Fig. 6A). However, the majority of zebrafish lncRNA loci had low CBL scores, indicating a lack of sequence alignments over many noncoding regions (Fig. 4C). The conservation of zebrafish lncRNAs as reflected by CBL scores was substantially lower than the conservation of protein-coding genes and was comparable to the conservation of intronic sequences (Supplemental Fig. 6B).

lncRNA genes carry chromatin marks associated with developmental regulators

To assess to which extent lncRNA genes carry chromatin marks that are known to

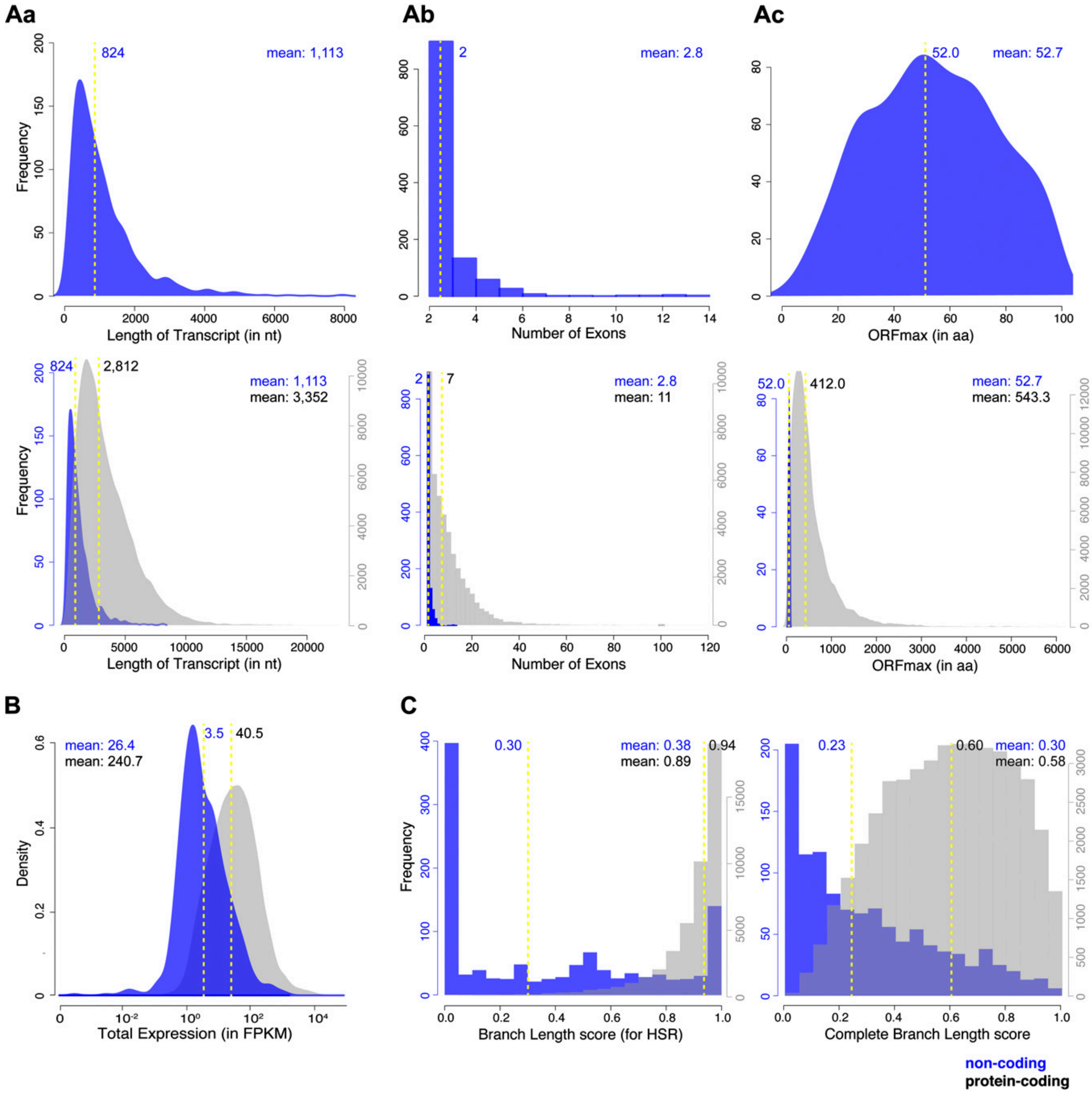


Figure 4. LncRNAs are shorter, less conserved, and expressed at lower levels than protein-coding genes. (*A*) Transcript length (*a*), number of exons (*b*), and maximum ORF length (ORFmax) (*c*) of the 1133 lncRNAs (*top* row) and of the 1133 lncRNAs (blue) in comparison to protein-coding transcripts (44,810 transcripts with PhyloCSF > 50; gray; *bottom* row). LncRNAs are generally shorter, have fewer exons, and contain shorter ORFs than protein-coding transcripts. Note that this might be an underestimation of the actual size of lncRNAs due to a potentially more incomplete assembly of low-expressed transcripts. (*B*) Comparison of the expression levels of lncRNA loci (859) and protein-coding loci (19,592 loci with PhyloCSF >50), plotted as fragments per kilobase of exon per million fragments mapped (FPKM). LncRNA loci are expressed at approximately 10-fold lower levels than the majority of protein-coding loci. (*C*) Comparison of the alignment quality across the locus of interest, assessed by two alternative measurements of the branch lengths present in the alignment. Branch lengths are measured on a scale from 0 to 1, where 0 indicates no alignments over the region of interest and 1 indicates the presence of 100% of sequence alignments. The branch length (BL) score refers to the alignment quality of the region that scores highest in PhyloCSF (the highest scoring region [HSR]; *left*). The complete branch length (CBL) score refers to the alignment quality over the entire length of the transcript (*right*). In the case of noncoding genes, alignments are poorer for the HSRs than for the entire gene length (BL scores < CBL scores). The reverse is true for protein-coding genes, which tend to have the best alignments over the HSRs (BL scores close to one). The values of the median (yellow dashed line) and mean are indicated in all panels.

be associated with protein-coding genes (Vastenhouw et al. 2010; Zhou et al. 2011), we performed chromatin immunoprecipitation assays in shield stage embryos followed by deep sequencing (ChIP-seq). We tested for the presence of trimethylated lysine 4 on histone 3 (H3K4me3), a known marker of promoters, and trimethylated lysine 27 on histone H3 (H3K27me3), a repressive histone modification. We restricted our analysis to lincRNAs and intronic overlapping lncRNAs since unambiguous assignment of marks to antisense exonic overlapping transcripts can be confounded by the overlapping genes.

Of all lncRNA promoter regions that were assessed, 29% were marked with H3K4me3 (both H3K4me3-only and H3K4me3/H3K27me3) (Fig. 5A). Notably, the fraction of H3K4me3-marked zebrafish lncRNA genes was similar to the 24% of human lincRNA genes that have a K4-K36 domain (Cabili et al. 2011), but was smaller than the fraction (63%) of marked zebrafish protein-coding genes (Fig. 5A).

To consider the possibility that the discrepancy between the fraction of H3K4me3-marked lncRNA and protein-coding loci could be due to the lower expression levels of lncRNA loci, we restricted our analysis to protein-coding genes expressed at shield stage and at expression levels similar to lncRNAs. Even under these conditions, the discrepancy between H3K4me3-positive noncoding (34%) and coding (74%) loci remained (Fig. 5B). This suggests that (1) the different expression levels of noncoding and protein-coding loci are not the primary cause of the different fractions of H3K4me3-marked loci, and (2) similarly to protein-coding genes (Vastenhouw et al. 2010), noncoding loci are marked with H3K4me3 largely independently of their expression status.

Interestingly, 7% of lincRNA and intronic overlapping lncRNA loci were marked by both H3K4me3 and H3K27me3 at shield stage (Fig. 5). Since Gene Ontology (GO)–term analysis of protein-coding genes marked with both H3K4me3 and H3K27me3 at shield stage revealed enrichment for developmental and regulatory functions (Supplemental Table 3), lncRNA loci may be important developmental regulators.

Nearest neighbor analysis of lncRNA genes

Previous studies have shown that mammalian lncRNAs are preferentially located next to genes with developmental functions (Dinger et al. 2008; Mercer et al. 2008; Guttman et al. 2009; Ponjavic et al. 2009; Cabili et al. 2011). We therefore analyzed the GO terms of genes that overlap with or are neighbors of zebrafish lncRNAs. We found significant enrichments ($P < 0.05$) of transcription factor activity, fate specification, and embryonic development and morphogenesis for genes that overlap with antisense exonic lncRNAs (Supplemental Fig. 7A; Supplemental Table 4) but not for neighbors of lincRNAs and intronic overlapping lncRNAs (Supplemental Table 4).

The mere physical proximity of lncRNAs and genes with developmental functions does not necessarily imply a functional link between the protein-coding gene and the lncRNA. For example, recent studies in the mouse did not detect a strong correlation between the expression levels of most lncRNAs and their neighbors (Guttman et al. 2011). Consistent with this study and with data from human lincRNAs (Cabili et al. 2011), we did not detect a higher degree of expression correlation for the majority of lncRNAs and their neighbors (or overlapping genes) than for protein–protein gene pairs or randomly assigned gene pairs (Supplemental Fig. 7B). The only exceptions were sense intronic overlapping lncRNAs, which tended to positively correlate in expression with the overlapping genes (Supplemental Fig. 7B). Such overlapping lncRNAs might resemble enhancer-associated lncRNAs (De Santa et al. 2010; Kim et al. 2010; Ørom et al. 2010; Wang et al. 2011).

To test whether lncRNA genes are preferentially located near protein-coding genes of certain evolutionary ages, we analyzed the phylostratographic classes (Domazet-Lošo and Tautz 2010) of genes that neighbor or overlap lncRNAs. Comparison with enrichments observed in a control set of protein–protein gene neighbors revealed no significant enrichment of particular evolutionary age groups for our lncRNA neighbors (Supplemental Fig. 7C).

Collectively, our analysis suggests that the neighbors of zebrafish lncRNAs belong to various classes of protein-coding genes of both ancient and more recent evolutionary origin and generally do not correlate in their expression with the neighboring lncRNAs.

Temporal expression profiles of lncRNAs

Our high-resolution time-series of RNA-seq experiments allowed us to follow the expression dynamics of lncRNAs and protein-coding genes as development proceeds. Comparison of independently clustered expression profiles of noncoding and protein-coding loci (Methods) revealed that both types of loci could be grouped into three broad classes (Fig. 6A): (1) loci whose transcripts were parentally supplied—these transcripts were present in the two- to four-cell-stage embryo (cleavage stages) and rapidly decayed after the first few hours of embryogenesis; (2) loci whose expression peaked during blastula and gastrula stages (dome, shield, and bud stages)—these transcripts were absent or only present at low levels during the early cleavage stages and were zygotically transcribed; and (3) loci that were only induced 1 d after fertilization during the process of organogenesis.

We discovered two differences between the expression patterns of protein-coding and noncoding loci. First, lncRNAs were more likely to be parentally supplied than were protein-coding mRNAs (see Fig. 6A). Any locus was classified as "parentally provided" for which at least 10% of its total expression across all eight embryonic stages was derived from the two- to four-cell stage. Of all transcripts present in our catalog, ~42% of lncRNAs classified as parentally provided, compared with only ~34% of protein-coding transcripts (Fisher's exact test, $P < 10^{-05}$). These observations suggest that parentally provided transcripts are specifically enriched in lncRNAs.

Second, the changes in a transcript's expression level between two consecutive stages were more pronounced for lncRNAs than for protein-coding genes. This observation suggests that lncRNAs have a more restricted temporal expression than do coding RNAs. To further test this hypothesis, we calculated a Shannon entropy-based specificity score per locus as a measure of expression level divergence during embryogenesis (Methods). All three classes of lncRNAs (lincRNAs, intronic overlapping, and antisense exonic overlapping lncRNAs) showed an increased temporal specificity compared with protein-coding genes (Fig. 6B). To rule out that this effect was caused by an increase in noise due to the lower expression levels of lncRNAs, we also sampled protein-coding loci from the same expression quantiles as lncRNAs (Methods). Although protein-coding loci that were expressed at low levels tended to be more restricted in time than were highly expressed protein-coding loci, they were significantly less restricted than were lncRNAs ($P < 10^{-4}$) (Fig. 6B). Together, these analyses reveal high temporal specificity during development as a novel property of lncRNAs.

Assigning function through expression correlation

The lack of annotated features makes the assignment of functions to lncRNAs a more challenging task than for proteins. Therefore, functional predictions for mammalian lncRNAs have often been based on "guilt-by-association" analyses (Dinger et al.

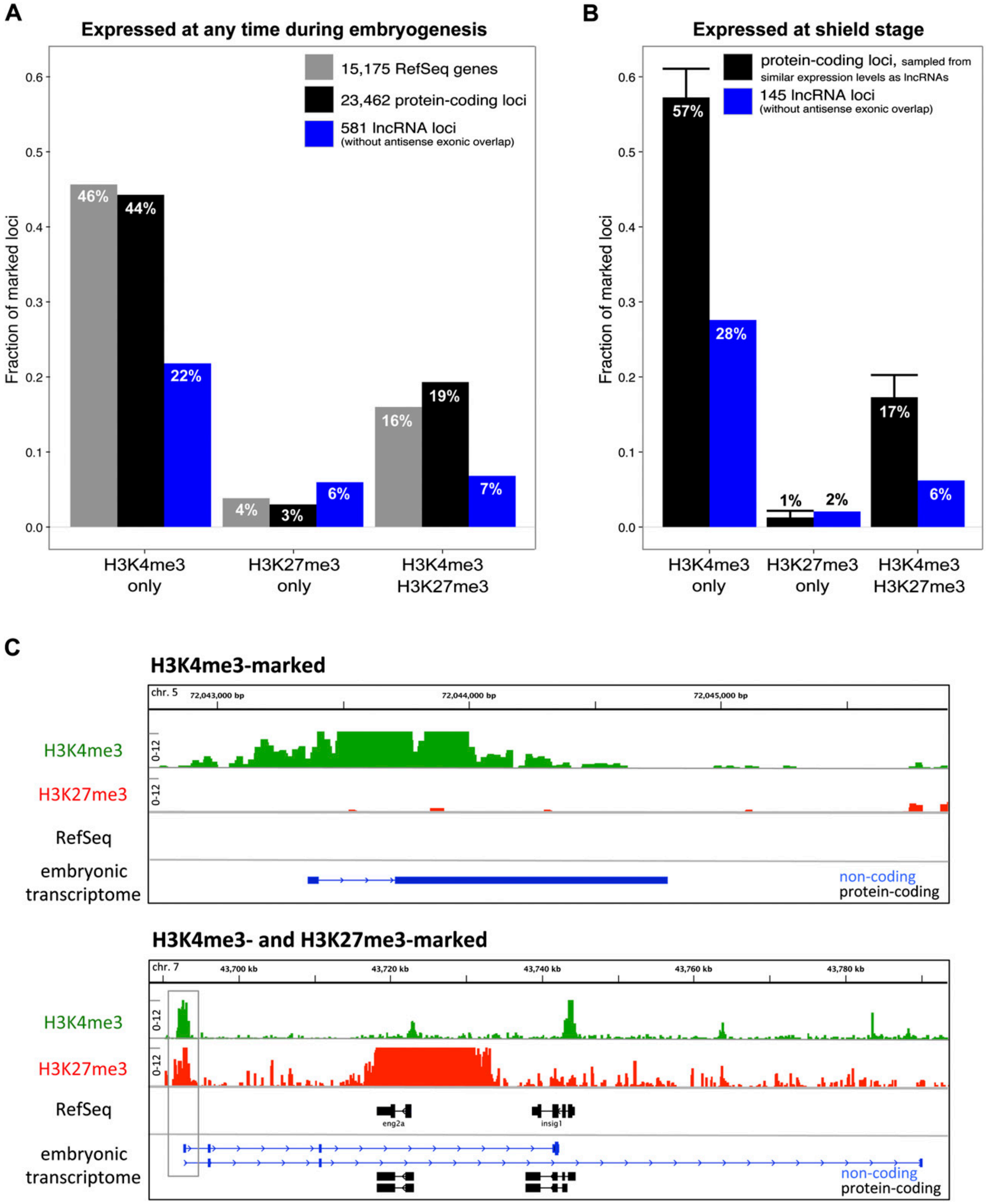
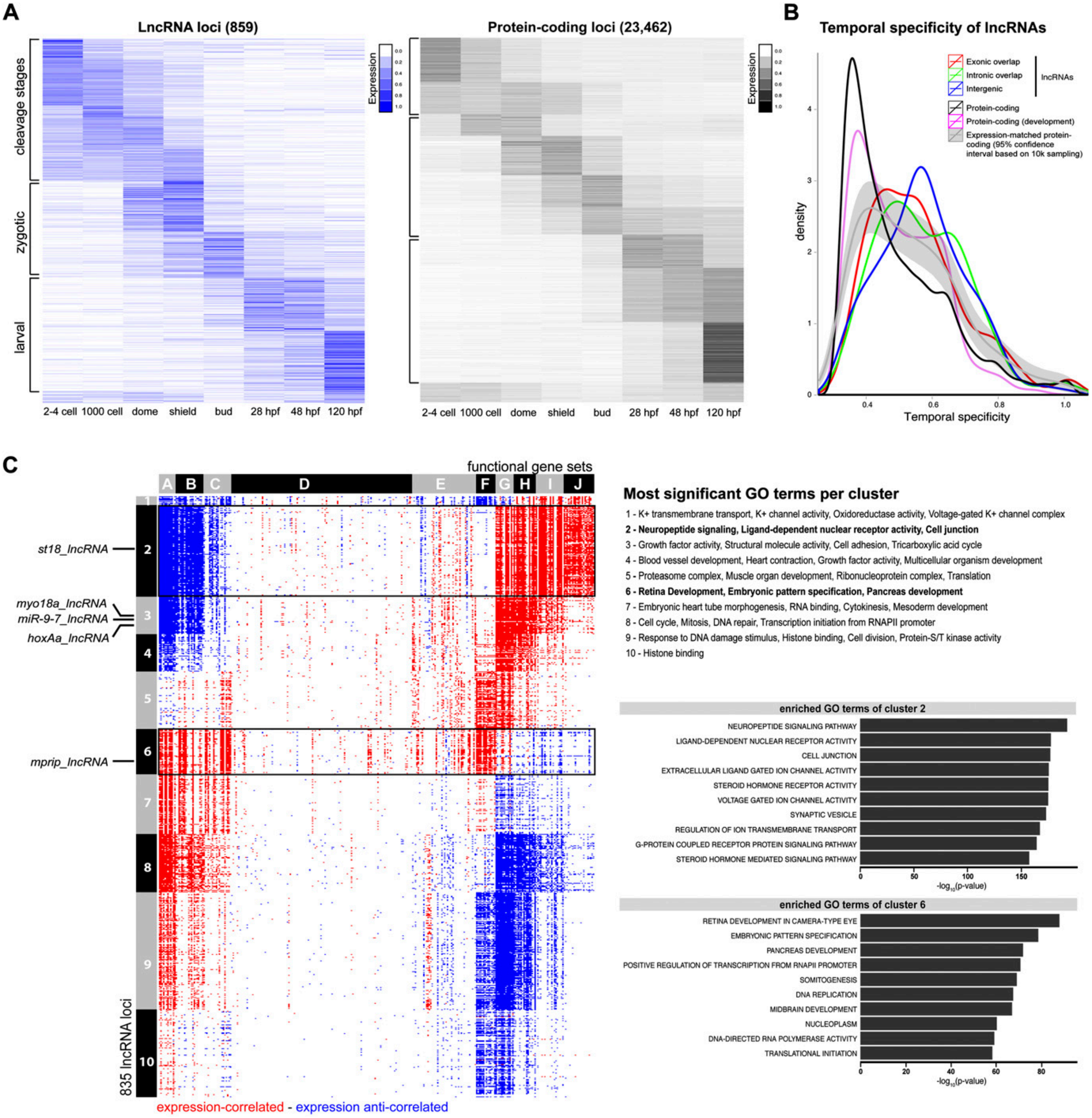


Figure 5. LncRNA genes carry chromatin marks associated with developmental regulators. Shown are the fractions of promoters (±500 bp relative to the transcription start site [TSS]) that are marked by a specific histone modification at shield stage. Histone marks were assessed by ChIP-seq experiments and analyzed for the presence of H3K4me3 only, H3K27me3 only, and both H3K4me3 and H3K27me3. RefSeq genes (gray bars); protein-coding loci (black bars); lncRNA loci (blue bars). (*A*) Marked fractions of promoters considering all loci. (*B*) Marked fractions of promoters only considering loci expressed at shield stage. In *B*, protein-coding loci were sampled from expression levels comparable to the set of 145 lncRNA loci expressed at shield (see Methods). Error bars, 1 SD of 10,000-times sampling. (*C*) Example chromatin profiles for a shield-expressed lincRNA gene marked by H3K4me3 (*top*) and for a lncRNA locus (overlapping the protein-coding genes *eng2a* and *insig1*) marked by both H3K4me3 and H3K27me3 (*bottom*). Signals are shown as the number of ChIP-seq reads that aligned overlapping in a 5-bp window (note that the *y*-axis ranges from 0–12).



Most significant GO terms per cluster

1 - K+ transmembrane transport, K+ channel activity, Oxidoreductase activity, Voltage-gated K+ channel complex

2 - Neuropeptide signaling, Ligand-dependent nuclear receptor activity, Cell junction

3 - Growth factor activity, Structural molecule activity, Cell adhesion, Tricarboxylic acid cycle

4 - Blood vessel development, Heart contraction, Growth factor activity, Multicellular organism development

5 - Proteasome complex, Muscle organ development, Ribonucleoprotein complex, Translation

6 - Retina Development, Embryonic pattern specification, Pancreas development

7 - Embryonic heart tube morphogenesis, RNA binding, Cytokinesis, Mesoderm development

8 - Cell cycle, Mitosis, DNA repair, Transcription initiation from RNAPII promoter

9 - Response to DNA damage stimulus, Histone binding, Cell division, Protein-S/T kinase activity

10 - Histone binding

enriched GO terms of cluster 2
NEUROPEPTIDE SIGNALING PATHWAY
LIGAND-DEPENDENT NUCLEAR RECEPTOR ACTIVITY
CELL JUNCTION
EXTRACELLULAR LIGAND GATED ION CHANNEL ACTIVITY
STEROID HORMONE RECEPTOR ACTIVITY
VOLTAGE GATED ION CHANNEL ACTIVITY
SYNAPTIC VESICLE
REGULATION OF ION TRANSMEMBRANE TRANSPORT
G-PROTEIN COUPLED RECEPTOR PROTEIN SIGNALING PATHWAY
STEROID HORMONE MEDIATED SIGNALING PATHWAY
0
50
100
150
-log10(p-value)
enriched GO terms of cluster 6
RETINA DEVELOPMENT IN CAMERA-TYPE EYE
EMBRYONIC PATTERN SPECIFICATION
PANCREAS DEVELOPMENT
POSITIVE REGULATION OF TRANSCRIPTION FROM RNAPII PROMOTER
SOMITOGENESIS
DNA REPLICATION
MIDBRAIN DEVELOPMENT
NUCLEOPLASM
DNA-DIRECTED RNA POLYMERASE ACTIVITY
TRANSLATIONAL INITIATION
0
20
40
60
80
-log10(p-value)

Figure 6. Temporal expression profiles of lncRNA genes compared to protein-coding genes. (*A*) Dynamic changes in expression profiles of loci (rows) across eight embryonic stages (columns). Heatmaps of 859 lncRNA loci (blue; *left*) and 23,462 protein-coding loci (gray; *right*) show normalized expression values (the sum of expression across all stages per locus is set to one). Three main expression patterns can be distinguished: ''cleavage stages'' (transcripts present in two- to four-cell-stage embryos), ''zygotic'' (transcripts enriched during blastula and gastrula stages and absent/only present at low levels at the two- to four-cell stage), and ''larval'' (transcripts induced only 1 d after fertilization). Note that the fraction of parentally provided (cleavage stage) transcripts is higher for lncRNAs than for protein-coding transcripts. (*B*) Temporal restriction of expression. Shown are distributions of Shannon entropy-based temporal specificity scores that were calculated for distinct classes of lncRNA loci and protein-coding loci (see Methods): exonic overlapping antisense lncRNAs (red), intronic overlapping lncRNAs (green), intergenic lncRNAs (blue), all protein-coding loci (black), and protein-coding loci of similar expression levels as lncRNA loci (gray; 95% confidence interval based on 10,000-times sampling). All classes of lncRNA loci display higher temporal specificity than protein-coding loci. (*C*) Expression-based association matrix of 835 lncRNA loci (rows) and functional gene sets (columns), derived from gene set enrichment analysis (GSEA). (Red) Positive correlation; (blue) negative correlation; (white) no correlation. Rows corresponding to lncRNAs whose RNA expression pattern is shown by in situ hybridization in Figure 7 are indicated on the *left*. Black boxes highlight two clusters associated with functions in signaling (cluster 2) and development (cluster 6). (*Top right*) The most enriched GO terms per cluster in comparison to all other clusters. (*Bottom right*) The 10 most enriched GO terms in the two boxed clusters in comparison to all other clusters, ranked by their –$\log_{10}$(*P*-values).

2008; Guttman et al. 2009; Cabili et al. 2011). We therefore analyzed the correlation between the expression dynamics of each protein-coding gene with the expression dynamics of each lncRNA locus. We performed gene set enrichment analysis (GSEA) (Methods; Mootha et al. 2003; Subramanian et al. 2005; Guttman et al. 2009) and associated GO terms and lncRNAs. This analysis identified several groups of lncRNAs associated with protein-coding gene sets of distinct functional categories such as signaling (cluster 2), development (cluster 6), and cell cycle (cluster 8) (Fig. 6C; for a complete list of enriched GO terms in each cluster, see also Supplemental Table 5). Interestingly, about one-third of our lncRNAs were associated with clusters enriched in developmental functions (clusters 4–7). These results indicate that many embryonic lncRNAs are putative developmental regulators.

LncRNAs show tissue-specific and subcellularly restricted expression patterns

To determine whether embryonic lncRNAs were expressed in specific tissues, we performed RNA in situ hybridization for a selected set of lncRNAs. Thirty-two lncRNAs were amplified from cDNA. While the majority of lncRNAs did not reveal strong or tissue-specific expression (Supplemental Fig. 8), several lncRNAs showed both expression in specific cell types and distinct subcellular RNA localization patterns (Fig. 7 and data not shown). Examples for cell type–specific expression patterns included lncRNAs that were loaded into the fertilized embryo by cytoplasmic streaming (e.g., *hoxAa_lncRNA*) (Fig. 7Ai), a lncRNA expressed in developing somites (*myo18a-lncRNA*) (Fig. 7Aii), and lncRNAs with distinct expression patterns in the developing nervous system (Fig. 7Aiii,iv).

Several zebrafish lncRNAs showed distinct subcellular localization patterns, supporting and extending previous localization studies in the mouse brain (Mercer et al. 2008; Ponjavic et al. 2009). For example, we observed nuclear enrichment of some lncRNAs in early cleavage stage embryos, including chromatin association in mitotically dividing cells (e.g., *hoxAa_lncRNA*) (Fig. 7Bi). Other lncRNAs such as *mprip_lncRNA* were found to accumulate at the cytoplasmic side of yolk syncytial layer nuclei at the bud stage (Fig. 7Bii). A particular striking example for a subcellularly localized lncRNA was *myo18a-lncRNA*, which was enriched specifically at the

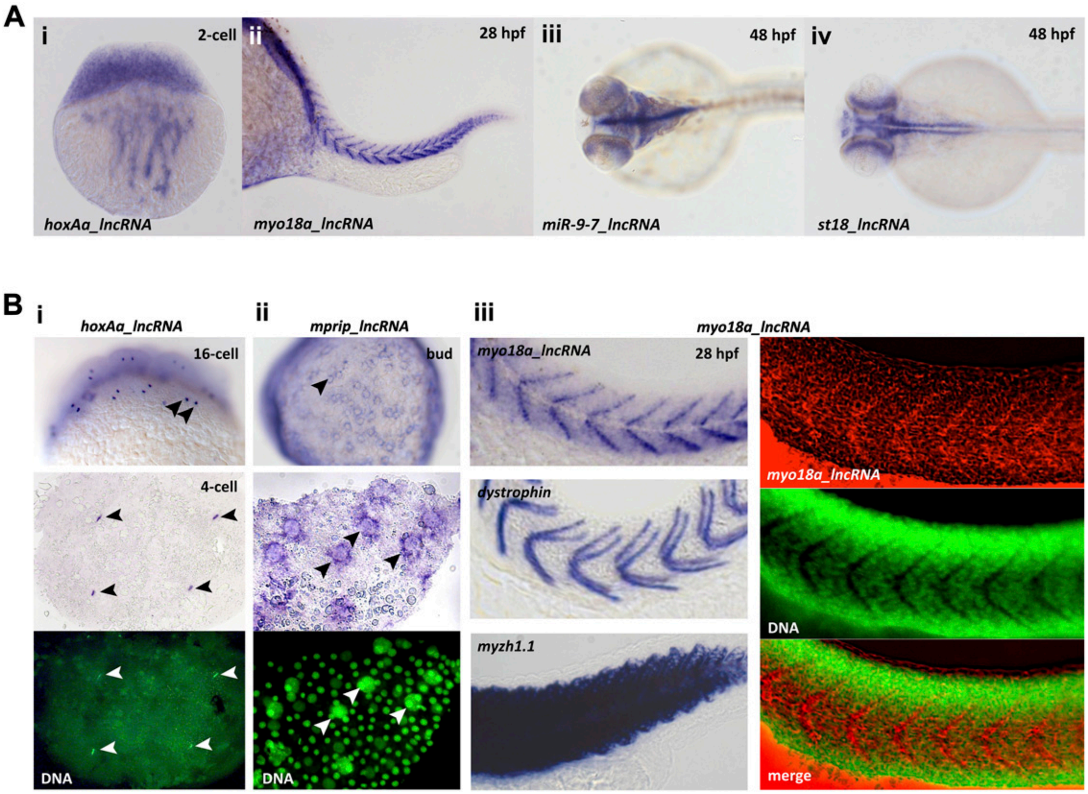


Figure 7. LncRNAs show tissue-specific and subcellularly restricted expression patterns. (*A*) Examples of lncRNAs with cell type–specific expression patterns at different stages of embryogenesis. Shown are in situ hybridization images with probes specific to the indicated lncRNAs. Expression is observed (*i*) in a two-cell stage embryo (cytoplasmic streaming from the yolk), (*ii*) in developing muscles, and (*iii,iv*) in distinct cells in the developing nervous system. (*i,ii*) Lateral views (anterior toward the *left* in *ii*); (*iii,iv*) dorsal views, anterior toward the *left*. (*B*) Examples of subcellularly localized lncRNAs. *Bottom* panels in *i* and *ii* (*middle* panel in *iii*, *right*) show a counterstain of the in situ image with the DNA-dye OliGreen (green). Black arrowheads point to subcellularly localized RNAs; white arrowheads point to the same position in the OliGreen-stained images. (*i*) Nuclear enrichment and association with chromatin *(hoxAa-lncRNA)*; (*top*) 16-cell stage embryo with mitotically dividing nuclei; (*middle*, *bottom*) four-cell stage embryo. (*ii*) Enrichment at the nuclear periphery *(mprip_lncRNA)*: (*top*) overview of a bud-stage embryo, showing accumulation of the lncRNA around nuclei of the yolk syncytial layer (YSL); (*middle*, *bottom*) close-up view of a dissected portion of the embryo shown in the *top* panel. Note that the lncRNA is specifically enriched around the large nuclei of the YSL but not around the small nuclei of the overlying cell-sheet. (*iii*) Enrichment at the myoseptum, the boundary between two adjacent myotubes *(myo18a-lncRNA*; *top left*, *right*); *dystrophin* mRNA (*middle left*) is a known marker of the myoseptum (Bassett 2003); *myzh1.1* (*myosin heavy chain*) mRNA (*bottom left*) is detected throughout the somites (not subcellularly localized); and (*right*) *myo18a-lncRNA* (red, in situ) is enriched at the myoseptum, which is characterized by the absence of nuclei (regions of no green in the OliGreen-stained panel). Note that there is no overlap between red and green in the merge panel.

myoseptum, the boundary where myotubes of adjacent somites meet (Fig. 7Aii,Biii). The *myo18a-lncRNA* localization pattern was distinct from many muscle-specific mRNAs that are ubiquitously expressed in somites (e.g., *myzh1.1* [*myosin heavy chain*]), but resembled the mRNA localization of the protein-coding gene *dystrophin*, a known marker of the myoseptum (Fig. 7Biii, left panel; Bassett 2003). Dystrophin is a key component of the protein complex that connects the cytoskeleton of the muscle fiber to the extracellular matrix, and its deficiency causes severe myopathies (Koenig et al. 1988). Intriguingly, a potential function of *myo18a_lncRNA* in cell–cell contact formation was supported by our expression-based GSEA approach, which associated *myo18a_lncRNA* with functions in "cell adhesion" and "structural molecule activity" (Fig. 6C, cluster 3). Collectively, these results reveal that several embryonic lncRNAs are expressed not only in specific tissues but also in specific subcellular domains.

Discussion

We have generated a systematic annotation of the zebrafish embryonic transcriptome, focusing specifically on the identification and characterization of lncRNAs. Large-scale RNA-seq experiments at eight embryonic stages allowed us to reconstruct 56,535 high-confidence coding and noncoding transcripts from 28,912 loci. We recovered the vast majority of expressed RefSeq transcripts, identified thousands of novel expressed loci and novel isoforms, and also captured the dynamic changes in expression levels of each transcript as development proceeds. Our data set is of about three- to fourfold higher depth than two recent zebrafish RNA-seq studies (Aanes et al. 2011; Vesterlund et al. 2011). This higher sequencing depth also translated into a significant increase in the number of identified expressed genes and was particularly important for the detection of lncRNAs that are expressed at relatively low levels. While both previous studies report read coverage across about 11,000 annotated genes, we have transcript evidence for 12,816 RefSeq genes (Fig. 1B) and 19,668 Ensembl loci (Supplemental Fig. 2B). In addition, we identified and reconstructed high-confidence (two-times evidence) transcripts expressed from more than 9000 novel loci with no previous annotations in RefSeq or Ensembl—almost twice as many as the number of "novel transcribed regions" reported by Aanes et al. (2011) and Vesterlund et al. (2011). Thus, our data set provides the, to date, most comprehensive annotation of the zebrafish embryonic transcriptome.

We defined a stringent set of 1133 multi-exonic noncoding transcripts, which includes lincRNAs, intronic overlapping lncRNAs, exonic antisense overlapping lncRNAs, and precursors for sRNAs. Our lncRNAs—the first long noncoding transcript catalog in a vertebrate embryo and in the zebrafish—share many of the characteristics of their mammalian counterparts (Dinger et al. 2008; Guttman et al. 2009, 2010, 2011; Ponjavic et al. 2009; Cabili et al. 2011): relatively short length, low exon number, relatively low expression, and conservation levels comparable to introns. Several observations indicate that zebrafish lncRNAs are likely to have diverse functions: They are associated with chromatin marks characteristic of genes with developmental functions (Bernstein et al. 2006; Vastenhouw et al. 2010), several are expressed in spatially and temporally restricted domains, and functional "guilt-by-association" analyses predict roles in processes ranging from cell cycle regulation to morphogenesis. Thus, zebrafish lncRNAs will be an excellent model system for functional studies that are difficult to perform in mammals.

Analysis of the developmental in vivo expression profile of our data set highlighted two novel properties of lncRNAs. First, the fraction of parentally biased transcripts is higher for lncRNAs than for protein-coding genes. Because there is no de novo transcription from the zygotic genome at this stage, these lncRNAs must be either maternally or paternally provided. The vast majority of protein-coding mRNAs and proteins present in the early embryo are of maternal origin and stored in the oocyte. This might also apply to lncRNAs, but in light of the striking testis enrichment of lincRNAs in humans (Cabili et al. 2011), it is intriguing to speculate that some of the early lncRNAs may belong to the yet poorly characterized small class of sperm-provided RNAs (Lalancette et al. 2008).

Second, lncRNAs are expressed in narrower time windows than are protein-coding genes. Thus, in addition to being highly tissue-specific (Cabili et al. 2011), lncRNA expression is highly temporally restricted. The association of specific sets of lncRNAs with well-defined developmental stages, together with their chromatin state and GSEA predictions, suggests diverse roles in development. For example, lncRNAs present during the early cleavage cycles may function in the still mysterious process that orchestrates the ubiquitous repression of zygotic transcription. This is an intriguing possibility in light of the fact that numerous lncRNAs have been shown to interact with repressive chromatin modifying complexes (Rinn et al. 2007; Khalil et al. 2009; Huarte et al. 2010; Schmitz et al. 2010; Tsai et al. 2010; Zhao et al. 2010; Guttman et al. 2011). In addition, early embryonic lncRNAs might regulate transcription of cell-cycle genes, a function recently suggested for a subset of cell-cycle promoter–associated human lincRNAs (Hung et al. 2011). LncRNAs expressed during blastula and gastrula stages might have important roles in cell fate decisions, differentiation, and cell migration. Indeed, recent large-scale knockdown analyses in mouse ESCs revealed key roles for lincRNAs in cell fate specification and maintenance of pluripotency (Guttman et al. 2011).

LncRNAs expressed during later embryonic and early larval stages are candidates for functioning in specific tissues and cell types during organogenesis. Potential roles during organogenesis are also supported by the tissue-specific expression of several lncRNAs. For example, specific lncRNAs are enriched in muscles and distinct subsets of neurons. Intriguingly, we also found lncRNAs with specific subcellular localization patterns. These patterns range from nuclear accumulation during the early cleavage stages to the enrichment at the boundary between adjacent myotubes. Studies of mRNA localization patterns of protein-coding genes in yeast (e.g., *ASH1*) (Long et al. 1997) and flies (e.g., *bicoid, oskar, gurken*) (for review, see Johnstone and Lasko 2001) have shown that the subcellular localization of specific RNAs is essential for normal development. Thus, enrichment of lncRNAs in specific subcellular compartments may be of fundamental importance for the regulatory functions of ncRNAs.

In summary, our study provides the first catalog of lncRNAs in a developing vertebrate. It suggests numerous roles of lncRNAs in vertebrate development and provides a high-quality resource for future genetic, evolutionary, and genomic studies.

Methods

RNA-seq of embryonic time course

Wild-type zebrafish embryos (TLAB) were staged according to standard procedures. About 1000 embryos were collected per stage (two to four cell, 1000 cell, dome, shield, bud, 28 hpf, 48 hpf, and 120 hpf) within a tight time window of ~10 min; it was ensured that all embryos were at the same developmental stage. Total RNA was isolated using the standard TRIzol (Invitrogen) protocol. Genomic DNA was removed by DNase treatment and confirmed by qPCR assay. Two rounds of PolyA^{+}-RNA purification were performed for each sample, using the PolyA(Purist)-MAG kit (Ambion). The quality of the RNA and lack of contaminating ribosomal RNA were confirmed using the Agilent 2100 Bioanalyzer. Strand-specific libraries for 76-bp paired-end sequencing were prepared according to a modified UTP-method (Parkhomchuk et al. 2009), as detailed by Levin et al. (2010). Libraries were sequenced on the GA-analyzer (shield stage library) and on the Illumina HiSeq 2000 (all stages), at a depth of 200–300 million reads per library (for statistics on read counts, see Supplemental Table 1).

Transcriptome assembly

RNA-seq–derived reads were aligned independently for each developmental stage with TopHat (version 1.2.1) (Trapnell et al. 2009). To aid these alignments, all known transcript annotations (Ensembl, RefSeq, and mRNAs from UCSC danRer7 [Zv9]) were pooled and used as an additional junction set (AJS) for each TopHat run. The junction outputs from individual stage-specific TopHat runs were pooled and added to the AJS (augmented AJS) to allow TopHat to use junction information from all stages. TopHat was rerun on each of the stages using the augmented AJS. The output of this second run comprised the final alignment and junction set for transcript assembly.

Transcriptomes were assembled with two different assemblers: Cufflinks (version 1.0.3) (Trapnell et al. 2010) and Scripture (version R4) (Guttman et al. 2010). The resulting transcripts were pooled, and a transcript was only considered reliable if it had support either from both assemblers or from at least two stages. Transcripts <160 bp were excluded, as these were most likely sequencing or assembly artifacts.

Multi-exonic transcripts

Multi-exonic transcripts were merged with Cuffcompare, and all transcripts classified as repeat were discarded. Scripture's strategy is to call all possible isoforms, including some that are most likely wrong and have no Cufflinks support. Therefore, all Scripture-only isoforms lacking Cufflinks support were excluded whenever Cufflinks had an assembled transcript for this locus. Furthermore, any di-exonic antisense transcript only supported by Scripture was removed since these transcripts are likely artifacts due to Scripture's lack of strand-aware library support.

Single exon transcripts

Single exon transcripts were subjected to additional scrutiny: They had to be significantly enriched in read coverage by Scripture (multiple testing corrected $P < 0.01$) (Guttman et al. 2010) and had to have at least one supporting transfrag from Cufflinks. Cufflinks uses library strand information and can therefore correctly assign the strand for single exon transcripts, while Scripture relies only on splice junctions and therefore cannot determine the strand-orientation of single exons. Transcripts classified by Cuffcompare (Trapnell et al. 2010) as contained (c), exon–intron fragment (e), exonic overlap (o), RNA pol II run-on (p), and repeat (r) were removed. Moreover, any single exon that was within a range of 500 bp in the sense direction relative to a multi-exonic transcript was removed.

Finally, single exon and multi-exonic transcripts were merged with Cuffcompare, discarding all contained and redundant isoforms.

ncRNA classification

Classification of each transcript as either coding or noncoding was determined using a step-wise filtering pipeline. First, all candidates were scored with PhyloCSF (Lin et al. 2011) to determine their coding potential (see PhyloCSF section). All transcripts that scored less than 20 were retained as potential noncoding candidates, and transcripts with PhyloCSF scores greater than 50 were considered proteins. The remaining transcripts (20 < PhyloCSF < 50) were initially classified as an ambiguous "gray" set. Second, the putatively noncoding transcripts and the "gray" set transcripts were repeat-masked and subjected to blastx, blastp, and HMMER (versus Pfam-A and Pfam-B) (Eddy 2009). For blastp and HMMER, the transcripts were translated (stop to stop codon, due to possible incomplete assemblies that could result in incomplete ORFs lacking the ATG start codon) in all three sense frames. Any transcript with an E-value less than 10^{-4} in any of the three search algorithms was considered as protein-coding.

Not all candidates were alignable to regions in the other four fish species. PhyloCSF-based coding potential predictions are less reliable for transcripts with no alignments over their entire region (see Supplemental Fig. 4B). Therefore, a maximal ORF cutoff was imposed. For candidates without alignments (CBL = 0, see Comparative Genomics Analysis of Conservation and Coding Potential) this cutoff was set to 30 aa, for all remaining transcripts (CBL > 0) it was set to 100 aa.

Finally, transcripts were removed that have exonic sense overlap with either Ensembl or RefSeq protein-coding genes or with the protein-coding gene set of the embryonic transcriptome.

Antisense rescue pipeline

PhyloCSF scores of antisense transcripts can be confounded by high-scoring protein-coding genes on the opposite strand, necessitating an alternative strategy for this set. The antisense-rescue pipeline was similar to the general pipeline (see above), with the exception that the PhyloCSF threshold was set to 300. In addition, the highest scoring region (HSR) for the putative antisense lncRNA had to have overlap with a protein-coding transcript on the opposite strand, and the PhyloCSF score for the protein-coding gene had to be higher than the PhyloCSF score for the lncRNA (see Supplemental Fig. 4D). Finally, after manual inspection, a threshold of maximal 81% was set for the sense/antisense exonic overlap with the protein-coding gene to remove a small number of likely artifactual transcripts that had substantial overlap and likely stemmed from errors in strand-calling during assembly.

Classification of lncRNAs

The resulting set of lncRNAs was subdivided into three categories: (1) lncRNAs without any overlap with other loci classify as lincRNAs (intergenic lncRNAs); (2) lncRNAs with intronic overlap are expressed from loci that have overlap (exon–intron or intron–intron but not exon–exon) with another transcribed locus (i.e., there is no overlap between the mature lncRNA with exons of the overlapping locus). They can be in either sense or antisense orientation with respect to the overlapping gene and can be further partitioned into intronic contained lncRNAs (incs; the lncRNA is contained within the transcribed region of another locus), completely overlapping lncRNAs (concs; the other locus is contained within the transcribed region of the lncRNA locus), and partially overlapping lncRNAs (poncs; neither incs nor poncs, but with at least one exon of the lncRNA contained within an intron of another locus). (3) Exonic antisense overlapping lncRNAs have exonic overlap with an exon of a protein-coding transcript on the opposite strand.

Comparative genomics analysis of conservation and coding potential

The RNA-seq transcripts were analyzed in MULTIZ whole-genome alignments of zebrafish with four other fish species (Tetraodon, Fugu, Stickleback, Medaka), generated by UCSC (http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=danRer7&g=cons8way).

Analysis of conservation by branch length analysis

Due to the large phylogenetic distances separating these species, sequences that can be aligned with BLASTZ/MULTIZ can generally be assumed to have evolved under negative selection to some extent. Therefore, the branch length (BL) score, a measure of alignment coverage that accounts for phylogenetic distances separating the species, was used as a simple conservation score for each transcript. The BL score is based on a phylogenetic tree with neutral BLs relating the species under analysis. For a single alignment column, the BL score is the ratio of the total BL of the tree relating only the species that aligned (not gapped) in that position, to the total BL of the tree relating all five species. The CBL score for a transcript is the average of the column-wise BL scores across the transcript.

Analysis of coding potential by PhyloCSF

PhyloCSF (Lin et al. 2011) was used to assess coding potential in the transcripts based on evolutionary signatures in the five-fish genome alignment. The alignment of each transcript was extracted from the genome alignments ("stitching" the alignments of individual exons as needed), and PhyloCSF was applied using the settings "–strategy=omega -f3–orf=StopStop3–minCodons=25." This command causes the program to enumerate complete and partial regions between stop codons, in three frames, and report the best scoring. Due to the limited completeness and reliability of existing zebrafish gene annotations, PhyloCSF was run in the simplified "–strategy=omega" mode that estimates evidence for a reduced d_N/d_S ratio, rather than performing a full empirical codon model comparison (which requires extensive training data).

A transcript was classified as potentially protein-coding if PhyloCSF reported a score of 20 or above, corresponding to a likelihood ratio of (10^(20/10)):1 in favor of reduced d_N/d_S. Furthermore, each transcript was scored on both the "sense" and "antisense" strands.

sRNA analysis

sRNAs expressed in 2-d-old wild-type zebrafish larvae (two biological replicates) were obtained from Cifuentes et al. (2010) and mapped to Zv9 using Bowtie (Langmead et al. 2009). The number of sRNAs overlapping lncRNA loci were counted. Transcripts with at least five uniquely mapped overlapped sRNAs were annotated according to known sRNA classes (miRNA precursor, snoRNA precursor, *MALAT1*-like transcripts, transcripts of unknown sRNA types).

Expression analysis

The expression level of each locus was assessed using Cuffdiff (Trapnell et al. 2010) in its time-series mode with upper quantile normalization. To visualize developmental expression profiles via heatmaps, expression levels were normalized to get relative expression levels over the developmental time-course (sum to an expression of one over all stages). LncRNAs and protein-coding loci were clustered separately using k-means (10 clusters) with a distance matrix constructed from the Pearson correlation.

The temporal specificity score over N time-points ($N = 8$ embryonic stages) was defined as 1 − H(g)/log2(N), where H(g) is the Shannon entropy expressed in bits of the expression vector of gene g. To compare lncRNA loci to protein-coding loci of similar expression levels, the expression levels of each locus were summed over all time-points and sorted into 100 quantiles. The Shannon entropy was then calculated for samples of proteins equal in number to the lncRNAs and from the same expression bins (10,000 repetitions to estimate the dispersion and to calculate the *P*-value).

In situ expression analysis

To analyze the expression pattern and localization of RNAs, 300- to 800-nt-long partial or full sequences of mRNAs were amplified from cDNA by PCR and cloned into the pSC-vector (Strataclone) according to standard procedures (primer sequences are available upon request). Thirty-two lncRNAs were cloned and tested by in situ hybridization experiments for their expression patterns from shortly after fertilization to 2-d-old larvae. Digoxigenin (DIG)-labeled antisense RNA probes were generated by in vitro transcription with T3 or T7 RNA polymerases, using plasmid-encoded T3 or T7 polymerase binding sites. In situ hybridization of zebrafish embryos of different embryonic stages was performed according to standard procedures (Thisse and Thisse 2008), using immunohistochemical detection of the DIG-labeled RNA–RNA hybrids by an anti-DIG Alkaline-Phosphatase coupled antibody, followed by nonfluorescent detection with BCIP/NBT. DNA was visualized by incubation of stained embryos with the DNA-dye OliGreen (Invitrogen, used at 1/400 in PBST). Images were processed with Photoshop and ImageJ.

Chromatin mark analysis: ChIP-seq for H3K4me3 and H3K27me3 at shield stage

ChIP was performed as previously described (Vastenhouw et al. 2010). Antibodies used were H3K4me3 (Millipore no. 07-473) and H3K27me3 (Millipore no. 07-449). For analysis on the Illumina Hi-Seq platform, sequencing libraries were prepared according to Illumina protocols.

Peak calling for chromatin marks was done using Scripture's ChIP-seq module (Guttman et al. 2010). This module scans fixed-size windows across the genome and computes read coverage and a multiple hypothesis corrected *P*-value for the observed coverage. For both H3K4me3 and H3K27me3, 500- and 1000-bp windows were scanned to account for both short regions with high read coverage and for larger regions with lower read coverage. All windows that were covered at a significant level ($P < 0.01$) were merged into "peaks." The ends of the peaks were finally trimmed until coverage at the ends is at least the average peak coverage. To account for systematic biases—e.g., due to open chromatin—peaks were filtered using input genomic DNA sequence by requiring that every peak called contained a 500-bp window with a library score at least threefold higher than the input genomic sequence score.

Peaks were intersected with promoter regions (±500 bp relative to the transcriptional start site [TSS]) of our transcripts. To obtain protein-coding loci of similar expression levels as lncRNA loci at shield stage, the same strategy was used as for expression analysis (see above), except that ranking of protein-coding loci was based exclusively on their expression levels at shield stage.

Gene set enrichment analysis

The expression level of each lncRNA locus was correlated with all protein-coding loci, similar to (Guttman et al. 2009). For each lncRNA locus, a list of correlation-based ranked protein-coding loci was constructed and subjected to GSEA (Mootha et al. 2003; Subramanian et al. 2005). An association matrix between lncRNA loci and GO terms was constructed, using a false-discovery rate threshold of 0.01. Rows (lncRNA loci) and columns (GO terms) were clustered (k-means, 10 clusters), resulting in distinct subsets of lncRNAs associated with functional GO terms. To determine the enrichment level of positively associated GO terms for each cluster with respect to other clusters, positively correlated GO terms were ranked according to a binominal test.

Nearest neighbor analysis

For each lincRNA locus the nearest protein-coding neighbor within <10 kb was identified. For antisense overlapping and intronic overlapping lncRNAs, overlapping gene(s) were identified. This resulted in a list of lncRNA loci/protein-coding loci pairs. Similar to the method described by Cabili et al. (2011), Pearson correlation was used to explore the expression-based relationship between these pairs. The results were subdivided based on (1) the class of the lncRNA and (2) the orientation of the lncRNA locus relative to the neighbor/overlapping protein-coding locus. The list of pairs (lncRNA loci/protein-coding loci) also formed the basis for GO term enrichment analysis using GOstat (Beissbarth and Speed 2004) and phylostratographic analysis of nearest neighbors (see below).

Phylostratographic analysis of nearest neighbors

Phylostratographic classes for zebrafish genes were obtained from Domazet-Lošo and Tautz (2010). LncRNA neighboring/overlapping protein-coding loci (see above) were tested for enrichment in certain phylostratographic classes by a sampling procedure that used the protein population as a null model.

Data access

The RNA-seq and ChIP-seq data have been submitted to the NCBI Gene Expression Omnibus (GEO) under accession no. GSE32900, containing the Subseries GSE32898 (RNA-seq) and GSE32899 (ChIP-seq). All data will also be accessible for downloading and convenient viewing on our website Z-Seq (http://www.broadinstitute.org/software/z-seq/).

Acknowledgments

We thank Cole Trapnell for helpful advice and continuous support with Cufflinks; Geo Pertea for help with Cuffcompare; UCSC and Ann Zweig for providing the teleost sequence alignments; the Broad Sequencing Platform for all sequencing work; Sara Chauvin for Project Management at the Broad Institute; Dongkeun Jang, Michael Reich, and Jill Mesirov (Broad Institute) for help in creating our website Z-Seq; Xian Adiconis for assistance in library preparation; Moran Cabili for sharing the human lincRNA data set prior to publication and for helpful discussions and comments on the manuscript; and Guo-Liang Chew and James Gagnon for discussions and helpful comments on the manuscript. A.P. was supported by an EMBO Long-term postdoctoral fellowship and a Human Frontier Science Program (HFSP) postdoctoral fellowship. E.V. was supported by a grant from FNU, Denmark. N.L.V. was supported by the Human Frontier Science Program (HFSP), Charles A. King Trust postdoctoral fellowships, and an NIH grant (1K99HD067220-01). A.S. was supported by the EU Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement 204135. J.L.R. is a Runyon Rachleff Innovation and Searle Scholar supported by the NIH (1R01ES02026). A.F.S. is supported by the NIH (5RO1 GM056211). This work was supported by the NHGRI grant 1RO1HG005111-01.

References

Aanes H, Winata CL, Lin CH, Chen JP, Srinivasan KG, Lee SGP, Lim AYM, Hajan HS, 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.

Bartel DP. 2009. MicroRNAs: target recognition and regulatory functions. *Cell* **136:** 215–233.

Bassett DI. 2003. Dystrophin is required for the formation of stable muscle attachments in the zebrafish embryo. *Development* **130:** 5851–5860.

Beissbarth T, Speed TP. 2004. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20:** 1464–1465.

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125:** 315–326.

Bertone P. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* **306:** 2242–2246.

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25:** 1915–1927.

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309:** 1559–1563.

Cifuentes D, Xue H, Taylor DW, Patnode H, Mishima Y, Cheloufi S, Ma E, Mane S, Hannon GJ, Lawson ND, et al. 2010. A novel miRNA processing pathway independent of dicer requires Argonaute2 catalytic activity. *Science* **328:** 1694–1698.

De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, Muller H, Ragoussis J, Wei C-L, Natoli G. 2010. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol* **8:** e1000384. doi: 10.1371/journal.pbio.1000384.

Dinger ME, Amaral PP, Mercer TR, Pang KC, Bruce SJ, Gardiner BB, Askarian-Amiri ME, Ru K, Solda G, Simons C, et al. 2008. Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res* **18:** 1433–1445.

Domazet-Lošo T, Tautz D. 2010. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468:** 815–818.

Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. *Genome Inform* **23:** 205–211.

ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

Fejes-Toth K, Sotirova V, Sachidanandam R, Assaf G, Hannon GJ, Kapranov P, Foissac S, Willingham AT, Duttagupta R, Dumais E, et al. 2009. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* **457:** 1028–1032.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, et al. 2011. Ensembl 2011. *Nucleic Acids Res* **39:** D800–D806.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458:** 223–227.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, et al. 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28:** 503–510.

Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L, et al. 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477:** 295–300.

Huarte M, Guttman M, Feldser D, Garber M, Koziol MJ, Kenzelmann-Broz D, Khalil AM, Zuk O, Amit I, Rabani M, et al. 2010. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142:** 409–419.

Hung T, Wang Y, Lin MF, Koegel AK, Kotake Y, Grant GD, Horlings HM, Shah N, Umbricht C, Wang P, et al. 2011. Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nat Genet* **43:** 621–629.

Johnstone O, Lasko P. 2001. Translational regulation and RNA localization in *Drosophila* oocytes and embryos. *Annu Rev Genet* **35:** 365–406.

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SPA, Gingeras TR. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296:** 916–919.

Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, et al. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316:** 1484–1488.

Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, et al. 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci* **106:** 11667–11672.

Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465:** 182–187.

Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF. 1995. Stages of embryonic development of the zebrafish. *Dev Dyn* **203:** 253–310.

Koenig M, Monaco AP, Kunkel LM. 1988. The complete sequence of dystrophin predicts a rod-shaped cytoskeletal protein. *Cell* **53:** 219–228.

Koziol MJ, Rinn JL. 2010. RNA traffic control of chromatin complexes. *Curr Opin Genet Dev* **20:** 142–148.

Lalancette C, Miller D, Li Y, Krawetz SA. 2008. Paternal contributions: new functional insights for spermatozoal RNA. *J Cell Biochem* **104:** 1570–1579.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25. doi: 10.1186/gb-2009-10-3-r25.

Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7:** 709–715.

Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27:** i275–i282.

Long RM, Singer RH, Meng X, Gonzalez I, Nasmyth K, Jansen RP. 1997. Mating type switching in yeast controlled by asymmetric localization of ASH1 mRNA. *Science* **277:** 383–387.

Mancini-Dinardo D, Steele SJS, Levorse JM, Ingram RS, Tilghman SM. 2006. Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. *Genes Dev* **20:** 1268–1282.

Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS. 2008. Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci* **105:** 716–721.

Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, et al. 2003. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34:** 267–273.

Nagano T, Mitchell JA, Sanz LA, Pauler FM, Ferguson-Smith AC, Feil R, Fraser P. 2008. The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* **322:** 1717–1720.

Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420:** 563–573.

Ørom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q, et al. 2010. Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143:** 46–58.

Pandey RR, Mondal T, Mohammad F, Enroth S, Redrup L, Komorowski J, Nagano T, Mancini-Dinardo D, Kanduri C. 2008. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol Cell* **32:** 232–246.

Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A. 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* **37:** e123. doi: 10.1093/nar/gkp596.

Pauli A, Rinn JL, Schier AF. 2011. Non-coding RNAs as regulators of embryogenesis. *Nat Rev Genet* **12:** 136–149.

Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. 2010. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465:** 1033–1038.

Ponjavic J, Ponting CP, Lunter G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* **17:** 556–565.

Ponjavic J, Oliver PL, Lunter G, Ponting CP. 2009. Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS Genet* **5:** e1000617. doi: 10.1371/journal.pgen.1000617.

Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, Hartman S, Harrison PM, Nelson FK, Miller P, Gerstein M, et al. 2003. The transcriptional activity of human chromosome 22. *Genes Dev* **17:** 529–540.

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, et al. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129:** 1311–1323.

Schmitz K-M, Mayer C, Postepska A, Grummt I. 2010. Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev* **24:** 2264–2269.

Sleutels F, Zwart R, Barlow DP. 2002. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415:** 810–813.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* **102:** 15545–15550.

Thisse C, Thisse B. 2008. High-resolution in situ hybridization to whole-mount zebrafish embryos. *Nat Protoc* **3:** 59–69.

Tian D, Sun S, Lee JT. 2010. The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. *Cell* **143:** 390–403.

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

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28:** 511–515.

Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, Freier SM, Bennett CF, Sharma A, Bubulya PA, et al. 2010. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell* **39:** 925–938.

Tsai M-C, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY. 2010. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329:** 689–693.

Vastenhouw NL, Zhang Y, Woods IG, Imam F, Regev A, Liu XS, Rinn J, Schier AF. 2010. Chromatin signature of embryonic pluripotency is established during genome activation. *Nature* **464:** 922–926.

Vesterlund L, Jiao H, Unneberg P, Hovatta O, Kere J. 2011. The zebrafish transcriptome during early development. *BMC Dev Biol* **11:** 30. doi: 10.1186/1471-213X-11-30.

Wang KC, Chang HY. 2011. Molecular mechanisms of long noncoding RNAs. *Mol Cell* **43:** 904–914.

Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA, et al. 2011. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472:** 120–124.

Weinmann L, Höck J, Ivacevic T, Ohrt T, Mütze J, Schwille P, Kremmer E, Benes V, Urlaub H, Meister G. 2009. Importin 8 is a gene silencing factor that targets argonaute proteins to distinct mRNAs. *Cell* **136:** 496–507.

Wilusz JE, Freier SM, Spector DL. 2008. 3′ end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell* **135:** 919–932.

Zhao J, Sun BK, Erwin JA, Song J-J, Lee JT. 2008. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322:** 750–756.

Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ, Sarma K, Song J-J, Kingston RE, Borowsky M, Lee JT. 2010. Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol Cell* **40:** 939–953.

Zhou VW, Goren A, Bernstein BE. 2011. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* **12:** 7–18.

Received October 12, 2011; accepted in revised form November 21, 2011.