Development 140, 2828-2834 (2013) doi:10.1242/dev.098343

Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs

Guo-Liang Chew[1], Andrea Pauli[1], John L. Rinn[2,3], Aviv Regev[3,4], Alexander F. Schier[1,3,5,*] and Eivind Valen[1,*]

SUMMARY

Large-scale genomics and computational approaches have identified thousands of putative long non-coding RNAs (lncRNAs). It has been controversial, however, as to what fraction of these RNAs is truly non-coding. Here, we combine ribosome profiling with a machine-learning approach to validate lncRNAs during zebrafish development in a high throughput manner. We find that dozens of proposed lncRNAs are protein-coding contaminants and that many lncRNAs have ribosome profiles that resemble the 5′ leaders of coding RNAs. Analysis of ribosome profiling data from embryonic stem cells reveals similar properties for mammalian lncRNAs. These results clarify the annotation of developmental lncRNAs and suggest a potential role for translation in lncRNA regulation. In addition, our computational pipeline and ribosome profiling data provide a powerful resource for the identification of translated open reading frames during zebrafish development.

KEY WORDS: Long non-coding RNAs, Ribosome profiling, Embryogenesis, Zebrafish, ES cells

INTRODUCTION

Long non-coding RNAs (lncRNAs) have emerged as important regulators of gene expression during development (Pauli et al., 2011; Rinn and Chang, 2012). lncRNAs were initially discovered for their essential roles in imprinting (Barlow et al., 1991; Bartolomei et al., 1991; Jinno et al., 1995; Sleutels et al., 2002) and mammalian X chromosome inactivation (Borsani et al., 1991; Brockdorff et al., 1992; Brown et al., 1992). Studies of Hox gene regulation in mammals and of flowering control in plants have identified additional lncRNAs, such as *HOTTIP* (Wang et al., 2011) and *COOLAIR* (Ietswaart et al., 2012; Swiezewski et al., 2009). The past decade has seen an explosion of genome-wide studies that have identified thousands of putative lncRNAs in a range of organisms (Bertone et al., 2004; Cabili et al., 2011; Carninci et al., 2005; Collins et al., 2012; Derrien et al., 2012; Djebali et al., 2012; Birney et al., 2007; Fejes-Toth et al., 2009; Guttman et al., 2009; Guttman et al., 2010; Kapranov et al., 2002; Kapranov et al., 2007; Okazaki et al., 2002; Pauli et al., 2012; Ravasi et al., 2006; Tilgner et al., 2012). Although the developmental roles of the vast majority of these novel transcripts are unknown, recent studies in zebrafish and embryonic stem cells (ESCs) have indicated roles for lncRNAs during embryogenesis, pluripotency and differentiation (Guttman et al., 2011; Ulitsky et al., 2011).

A prerequisite for the functional analysis of lncRNAs is the high-confidence annotation of this class of genes as truly non-coding. The distinction of lncRNAs from coding mRNAs has often relied on the computational classification of expressed transcripts (Dinger et al., 2008; Guttman and Rinn, 2012). These classifiers evaluate transcript features, such as open reading frame (ORF) lengths, coding potential, and protein sequence conservation. Such computational approaches can distinguish between coding RNAs and lncRNAs (Cabili et al., 2011; Carninci et al., 2005; Guttman et al., 2009; Pauli et al., 2012; Ulitsky et al., 2011), but may also give rise to misclassifications: lncRNAs containing short conserved regions might be misclassified as protein-coding (false negatives), whereas protein-coding transcripts containing short or weakly conserved ORFs might be misclassified as non-coding (false positives). For example, two recent zebrafish lncRNA catalogs (Pauli et al., 2012; Ulitsky et al., 2011) share little overlap, suggesting that novel approaches are needed to distinguish coding from non-coding RNAs.

One approach to detect potential coding sequences is ribosome profiling (Ingolia et al., 2009; Ingolia et al., 2012). In this method, mRNA fragments protected from RNaseI digestion by cycloheximide (CHX)-stalled 80S ribosomes are isolated and sequenced. The resultant ribosome-protected fragments (RPFs) correspond to the sites where translating ribosomes resided on mRNA transcripts at the time of isolation, yielding a quantitative, genome-wide snapshot of translation at nucleotide (nt) resolution. Application of this method to mouse embryonic stem cells (mESCs) detected RPFs associated with many previously annotated lncRNAs (Ingolia et al., 2011). This study suggested that the majority of annotated lncRNAs contain highly translated regions comparable to protein-coding genes and might encode proteins. However, translation of a transcript was inferred by measuring localized densities of ribosome profiling reads relative to expression (translational efficiency; TE). As shown below, we find that this approach does not reliably distinguish the main ORFs (coding sequences; CDSs) from upstream ORFs (uORFs). This distinction is important because the vast majority of uORFs are unlikely to code for functional peptide products because their peptide sequences are not conserved, even though their presence in the 5′ leader may be (Hood et al., 2009). Indeed, a recent peptidomics study suggested that most annotated lncRNAs do not generate stable protein products (Bánfai et al., 2012). It has therefore remained unclear what fraction of currently annotated putative lncRNAs are truly non-coding.

[1]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA. [2]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA. [3]The Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA 02142, USA. [4]Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [5]FAS Center for Systems Biology, Harvard University, Cambridge, MA 02138, USA.

*Authors for correspondence (schier@mcb.harvard.edu; eivindvalen@fas.harvard.edu)

Accepted 7 May 2013

Here, we address the issue of lncRNA annotation by combining ribosome profiling during early zebrafish development with a new machine-learning approach. Our study suggests that dozens of previously annotated lncRNAs are protein-coding contaminants. In addition, we find that many lncRNAs in zebrafish and ESCs resemble the 5′ leaders of coding mRNAs, raising the possibility that translation is involved in lncRNA regulation. The methods and datasets provided in this study provide a broad resource for the identification of translated ORFs during zebrafish development.

MATERIALS AND METHODS

Ribosome profiling

Ribosome profiling was adapted from Ingolia et al. (Ingolia et al., 2011) and applied to a zebrafish developmental time course. For each stage [2-4 cells, 256 cells, 1000 cells, dome, shield, bud, 28 hours post fertilization and 5 days post fertilization (Kimmel et al., 1995)], 400-600 embryos were washed with cold PBS, flash-frozen and stored at −80°C. Embryos were lysed by repeated micropipetting in 1.5 ml of cold polysome buffer (20 mM Tris-HCl pH 7.4, 250 mM NaCl, 15 mM $MgCl_2$, 1 mM dithiothreitol, 100 μg/ml CHX) with added 0.5% Triton X-100, 500 μg/ml guanosine 5′-[β,γ-imido]triphosphate (GMP-PNP), 24 U/ml TurboDNase (Ambion AM2238), then incubated with agitation for 10 minutes at 4°C, and clarified by centrifugation at 1300 ***g*** for 10 minutes at 4°C. For ribosome footprinting, 20 μl RNaseI (Ambion AM2294) was added to the 1.5 ml of supernatant and incubated for 30 minutes at 37°C, then stopped by chilling on ice and addition of 40 μl of SuperaseIn (Ambion AM2694). Footprinted samples were pelleted through a sucrose cushion (1 M sucrose in polysome buffer with added 100 U/ml SuperaseIn) by centrifugation at 260,000 ***g*** for 4.5 hours at 4°C, and re-suspended in 800 μl 10 mM Tris pH 7.4 with 1% SDS. RNA was purified by hot acid phenol/chloroform extraction and precipitated by standard ethanol precipitation. From this point, ribosome profiling Illumina-compatible sequencing libraries were prepared as previously described (Ingolia et al., 2011). Supplementary material Table S1 lists the primers and subtractive hybridization oligonucleotides corresponding to the most abundant rRNA contaminants that were determined in a pilot ribosome profiling experiment.

Sequencing and mapping of RPFs

Ribosome profiling libraries were sequenced on an Illumina HiSeq 2000 (one stage per lane, 44 bp reads), resulting in a total of 880 million reads (for an overview, see supplementary material Fig. S1). Following adapter sequence trimming, RPFs were compared with zebrafish rRNAs from the SILVA rRNA database (Quast et al., 2013) using Bowtie2 (Langmead and Salzberg, 2012) (parameters: –N 1; –L 20; –k 20). Reads matching rRNA (~50%) were discarded. The remaining RPFs were mapped by Tophat2 (Trapnell et al., 2009) (parameters: no indels; no novel junctions; –M; –g 10) to a zebrafish developmental transcriptome (Pauli et al., 2012) and the Zv9 genome assembly, resulting in 317 million mapped reads. To obtain near-nucleotide resolution from ribosome profiling (supplementary material Fig. S2), RPFs aligning at annotated start and stop codons of RefSeq genes were subdivided by read length (supplementary material Fig. S2B). Approximate P-site position for each read-length was determined by inspection of coverage and phasing of the read's left-most position relative to annotated start and stop codons. Offsets were determined to be +12 for 27-28 nt RPFs, +13 for 29-31 nt RPFs, and +14 for 32 nt RPFs (supplementary material Fig. S2A). Based on observable phasing over the coding sequences, RPFs between 27 and 32 nts (totaling 220 million) were deemed to be high quality and were used in subsequent analysis. The remaining RPFs were likely to be over- or under-digested, and were discarded. Library sizes between stages were normalized by the number of RPFs in each stage that mapped to annotated coding regions of RefSeq genes. mESC ribosome profiling data was obtained from Ingolia et al. (Ingolia et al., 2011).

Construction of training and lncRNA data sets

The zebrafish training set was constructed from RefSeq genes in the Zv9/danRer7 zebrafish genome assembly. Only genes expressed at fragments per kilobase of exon per million fragments mapped (FPKM) >1 (summed over the developmental transcriptome) (Pauli et al., 2012) were used. Similarly, the mouse training set was based on RefSeq genes in the mm9 mouse genome assembly expressed at FPKM >1 in mESCs (Guttman et al., 2010). ORFs were defined as regions starting with either an ATG or CTG and ending with an in-frame stop codon. Three classes of ORFs were defined: (1) the CDSs in the context of their respective transcripts, (2) all RPF-containing ORFs in transcript leaders in the context of the detached 5′ leaders and (3) all RPF-containing ORFs in the transcript trailers in the context of the detached 3′ trailers (Fig. 1). CDSs with trailers shorter than 100 nt were not included. Owing to the high number of truncated transcripts annotated in zebrafish, all ORFs in the zebrafish set were required to be at least 20 nts from the transcript edge. ORFs in leaders and trailers were filtered to ensure lack of any overlap with annotated RefSeq, Ensembl or XenoRefSeq coding regions.

For classification, lncRNAs were required to be expressed at >1 FPKM over the developmental time course (for zebrafish) and in ESCs (for mouse). As a few transcripts had a clear RPF-covered coding ORF, but lacked start/stop codons (probably owing to truncations in transcript assembly), ORFs were allowed to extend beyond the ends of transcripts. To account for possible transcript truncations, it was assumed that the start/stop of the ORF was at the edge and a pseudo-trailer of 10 nt was added to all transcripts when calculating IO scores (see below).

Classification

For each ORF, we used four metrics designed to distinguish between the three classes and capture the features of protein coding genes:

Translational efficiency (TE) is defined as (density of RPFs within ORF)/(RNA expression). Density is the average sum of normalized RPFs over the embryonic time course within the ORF divided by the length of the ORF. RNA expression is the average FPKM of the locus over the embryonic time course containing this transcript.

Inside versus outside (IO) is defined as (coverage inside ORF)/(coverage outside ORF). Coverage refers to the number of nt positions having any RPF divided by the total number of nts inside or outside the ORF. A pseudo-count of 1 is added to both the inside and outside sums.

Fraction length (FL) is defined as (length of ORF)/(length of transcript), i.e. the fraction of the transcript covered by the ORF.

Disengagement score (DS) is defined as (RPFs over ORF)/(RPFs downstream), i.e. the number of RPFs inside the ORF divided by the number of RPFs downstream. A pseudo-count of one was added to both the ORF and downstream sums.

A random forest classifier (Breiman, 2001) (implemented in the R package randomForest) was trained using these four metrics on the respective training sets. The three classes were weighted according to size, and standard options were used (500 trees, two variables per split). Classes were assigned to loci in order to minimize cross-mapping between coding and non-coding isoforms. If any ORF was classified as coding, the locus was considered to be coding. If not, the locus was considered to be leader-like if at least one ORF was classified as leader-like. Finally, if all ORFs were classified as trailer-like or if no ORF had RPFs, the locus was classified as trailer-like.

Public database accession numbers

The ribosome profiling data are accessible at Gene Expression Omnibus (GEO) with accession number GSE46512. The RNA-seq data was published previously (Pauli et al., 2012) and is available at GEO (accession number GSE32900, subseries GSE32898).

RESULTS AND DISCUSSION

Ribosome profiling outlines translated regions of zebrafish transcripts

To identify ribosome-associated regions in the zebrafish transcriptome, we generated high-depth ribosome profiles over a time course of eight early developmental stages (Fig. 1; supplementary material Fig. S1; for details see Materials and methods). Of 220 million high-quality RPFs, 84.5 million RPFs

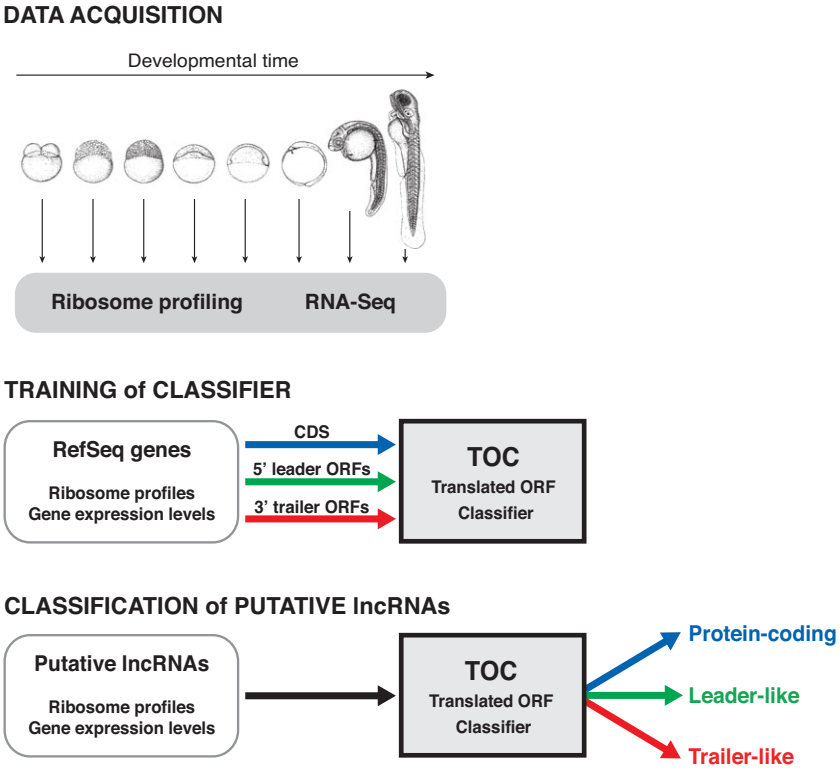

Fig. 1. Overview of lncRNA classification pipeline. (**A**,**B**) High-throughput sequencing data (ribosome profiling and RNA-seq) from eight early developmental stages (A) is used to train a classifier with RefSeq coding sequences (CDSs), 5′ leaders and 3′ trailers (B). (**C**) The translated ORF classifier (TOC) uses ribosome profiles and gene expression levels to classify putative lncRNAs as protein-coding (blue), leader-like (green) or trailer-like (red).

mapped to RefSeq genes (see Fig. 2A for examples of ribosome profiles). Approximately 81% of RefSeq genes that expressed >1 FPKM (12,228 genes) had at least ten normalized RPFs (supplementary material Fig. S3A), and about 68% of genes had reads over at least 10% of their annotated coding sequence (CDS; supplementary material Fig. S3B). Within exons of RefSeq transcripts, 95.7% of RPFs mapped to CDSs (mean density of 3.64 RPFs per nt), 0.54% of RPFs mapped to 3′ transcript trailers (mean density of 0.054 RPFs per nt), and the rest (3.71%) mapped to 5′ transcript leaders (mean density of 1.46 RPFs per nt). This distribution corresponds to a >65-fold enrichment of RPFs associated with CDSs compared with 3′ trailers, and a >25-fold enrichment of RPFs associated with 5′ leaders compared with 3′ trailers, consistent with ribosome profiling data in other systems (Brar et al., 2012; Ingolia et al., 2011). As observed in previous studies, we found triplet phasing of ribosome profiles in the CDSs of coding genes, corresponding to the translocation of translating 80S ribosomes in steps of 3 nts (Fig. 2B).

Consistent with the release of 80S ribosomes at in-frame stop codons, RPFs over 3′ trailers tend to be sparse and randomly distributed (Fig. 2A), and may represent background experimental noise inherent to the ribosome profiling method. As observed in ribosome profiling data in other systems (Brar et al., 2012; Fritsch et al., 2012; Ingolia et al., 2011; Lee et al., 2012), 5′ leaders of coding transcripts are widely associated with ribosomes, showing relatively high densities of RPFs at locations often corresponding, but not limited, to uORFs. The stop codons of annotated ORFs are significantly enriched for RPFs (supplementary material Fig. S2C). We find widespread occurrence of uORFs (49.5% of RefSeq genes have RPF-containing uORFs), as well as many instances of translated, extremely short ORFs that are as small as an AUG followed by a stop (minimal ORFs or minORFs) (supplementary material Fig. S4). These results highlight the power of this approach in identifying translated regions of zebrafish transcripts.

Translated ORF classifier (TOC) distinguishes ORFs in annotated 5′ leaders, CDSs and 3′ trailers

To use the ribosome profiling dataset for the classification of ORFs, we developed a random forest classifier (Breiman, 2001). We tested whether ribosome profiles over RNA subregions might reliably distinguish CDSs from ORFs in 5′ leaders and from ORFs in 3′ trailers. To train the classifier, we used the RefSeq gene sets in zebrafish and mouse (see Materials and methods for details). Our classifier, called TOC (translated ORF classifier), employs four features (Fig. 3A): (1) Translational efficiency (TE) – the density of ribosome profiling reads over an ORF relative to its expression level; (2) inside versus outside (IO) – the ratio of bases covered within an ORF versus outside (upstream and downstream), capturing a distinct feature of coding transcripts for which read coverage tends to be predominantly over a single ORF; (3) fraction length (FL) – the fraction of the transcript covered by the ORF, accounting for the observation that annotated CDSs tend to span a significant portion of the transcript; and (4) disengagement score (DS) – the degree to which RPFs are absent downstream of the ORF, building on prior knowledge that re-initiation after extended translation and stop-codon read-through are rare events (Jackson et al., 2007). These features effectively integrate intrinsic transcript information, such as sequence and location of ORFs, with external data, such as ribosome profiling and expression levels derived from RNA-seq.

Although individual features were able to separate one class of RefSeq ORFs from the other two, the combination of all four was necessary to distinguish reliably ORFs within annotated 5′ leaders, CDSs and 3′ trailers (Fig. 3; supplementary material Fig. S5). Notably, TE distinguished 3′ trailers from 5′ leaders and CDSs, whereas DS helped separate uORFs in 5′ leaders from CDSs (Fig. 3B for zebrafish; supplementary material Fig. S5 for mouse). The combination of IO and FL differentiated CDSs from ORFs in 5′ leaders and 3′ trailers (Fig. 3B; supplementary material Fig. S5).

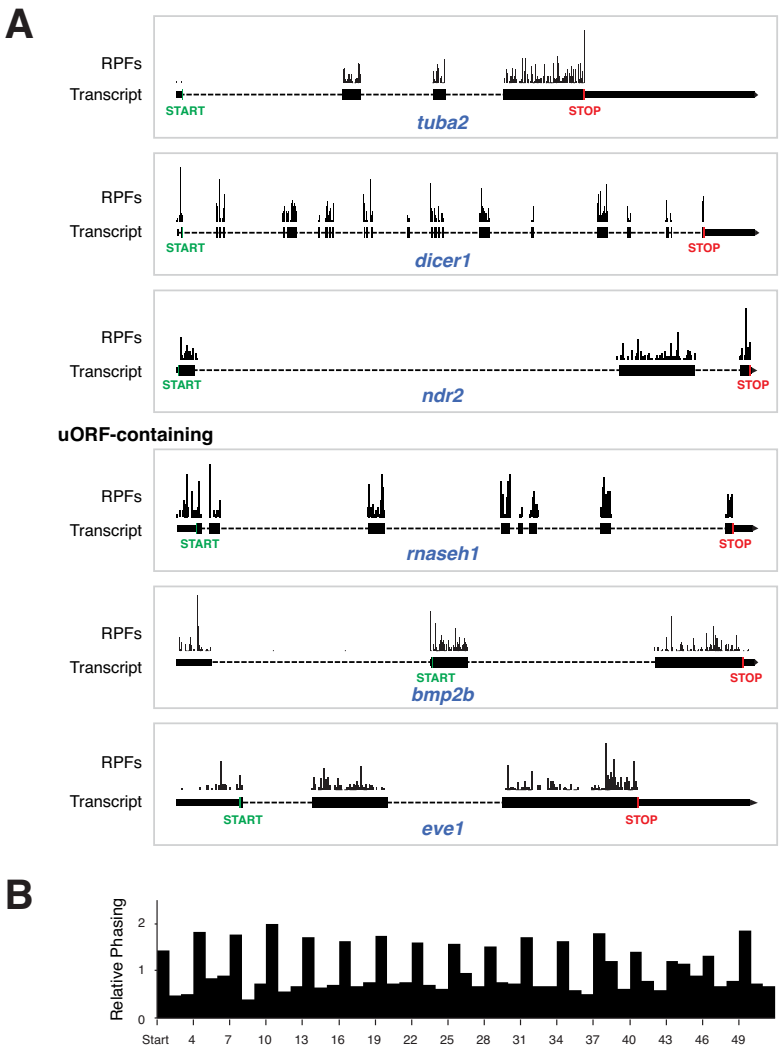

Fig. 2. Ribosome profiles outline translated ORFs of coding genes. (**A**) Representative examples of ribosome-protected fragment (RPF) densities associated with protein-coding genes. Gene structures are depicted as thick bars for the coding sequence (CDS), thin bars for 5′ leaders and 3′ trailers, and dashed lines for introns. Note that the majority of RPFs map within the CDSs and are flanked by the annotated initiation (START, green) and termination codon (STOP, red). The bottom three panels show examples of uORF-containing genes. For these genes, RPF reads map to the CDSs and to short ORFs within the 5′ leaders. (**B**) RefSeq metagene analysis of relative phasing of ribosome P-sites (see Materials and methods). Relative phasing is defined as the number of RPFs at a given position divided by the mean of the number of RPFs at the four adjacent positions. i.e. relative phasing at position i=RPFs at position i/mean (RPFs at positions i-2, i-1, i+1 and i+2). As in previous studies (Ingolia et al., 2011), triplet phasing of ribosome profiles was observed.

The use of all four features in the TOC classifier was highly accurate in distinguishing CDSs from 5′ leader-like ORFs and 3′ trailer-like ORFs even at low RNA expression levels (supplementary material Fig. S6; overall out-of-bag error for zebrafish: 3.25%). These results establish TOC as a powerful classifier to distinguish ORFs in annotated 5′ leaders, CDSs and 3′ trailers.

TOC refines classification of lncRNAs

To refine the classification of putative lncRNAs, we applied TOC to the catalogs recently published for zebrafish embryos (Pauli et al., 2012; Ulitsky et al., 2011) and mESCs (Guttman et al., 2011). The application of TOC to these datasets is justified by the biochemical similarity between coding mRNAs and recently annotated lncRNAs (e.g. both are 5′ capped and 3′ poly-adenylated). Notably, TOC analysis revealed that dozens of putative lncRNAs have the same characteristics as protein-coding mRNAs: a main CDS engaged by ribosomes and few (if any) RPFs downstream (Fig. 4; supplementary material Fig. S7 and Tables S2-S4). Depending on the dataset, we find that 8-45% of previously proposed lncRNAs are likely to be bona fide protein-coding mRNAs (Fig. 4; supplementary material Tables S2-S4). These transcripts will be an interesting resource for identifying previously uncharacterized proteins. By contrast, 18-44% of putative lncRNAs showed little or no association with ribosomes, akin to 3′ trailers of coding transcripts (Fig. 4; supplementary material Tables S2-S4). These transcripts are bona fide lncRNAs and warrant functional characterization.

Strikingly, we found that the ribosome profiles over more than 40% of putative zebrafish and mouse lncRNAs resemble 5′ leaders rather than 3′ trailers (Fig. 4; supplementary material Tables S2-S4). These lncRNAs contain ORFs with a higher TE than 3′ trailer-like lncRNAs, but have shorter and less conserved ORFs than do the CDSs of protein-coding genes (supplementary material Fig. S8). Similar to leaders, RPFs are often distributed over multiple ORFs, none of which stand out as a main CDS of a protein-coding gene. The leader-like class of lncRNAs represents a distinct subset of the previously described short, polycistronic ribosome-associated coding RNAs (sprcRNAs) (Ingolia et al., 2011). Unlike sprcRNAs, which are identified solely by TE, leader-like lncRNAs exclude misannotated protein-coding mRNAs and transcripts with spuriously associated ribosomes.

The association of ribosomes with leader-like lncRNAs raises two important questions: Do the associated ribosomes generate proteins? Are these proteins functional? Several observations suggest that leader-associated ribosomes might generate proteins that are likely to be non-functional. Recent studies have shown that the CHX used in ribosome profiling protocols acts through the E-site of the 60S ribosomal subunit (Schneider-Poetsch et al.,

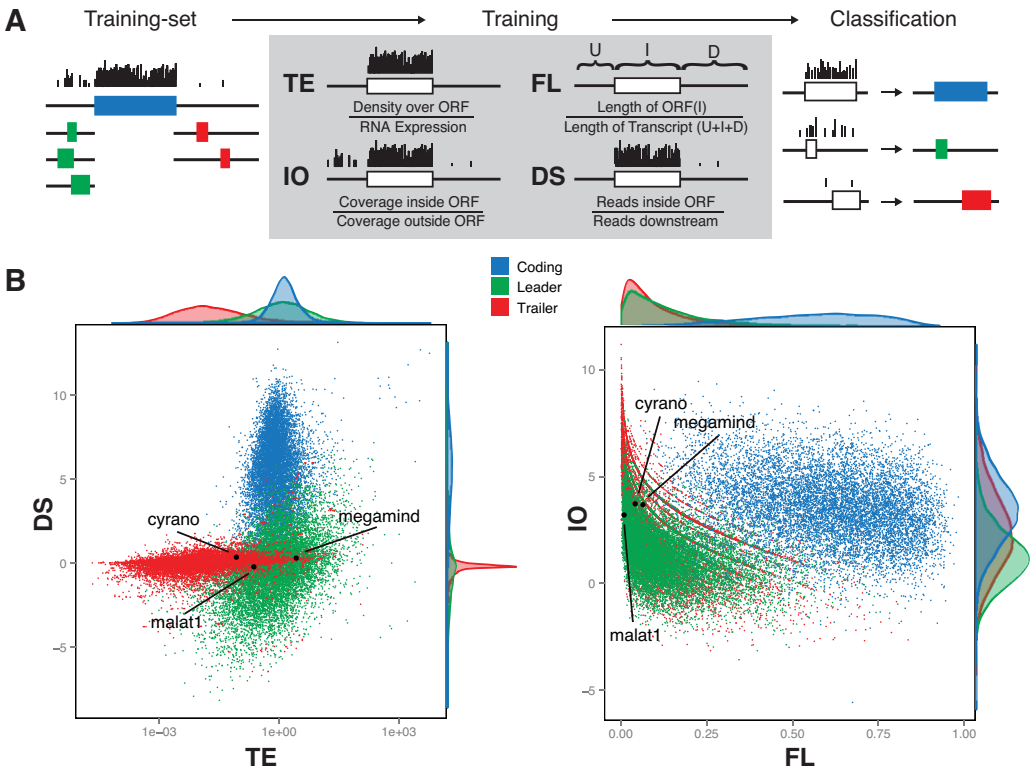

Fig. 3. TOC distinguishes ORFs in 5′ leaders, CDSs and 3′ trailers. (**A**) A training set is constructed from RefSeq genes using (1) annotated CDSs (coding ORFs, blue) in the context of the whole transcript, (2) RPF-containing ORFs in the 5′ leader sequence (green) in the context of the 5′ leader, and (3) RPF-containing ORFs in the 3′ trailer (red) in the context of the 3′ trailer (see Materials and methods). The four metrics used to train the classifier are displayed in the gray box (TE, translational efficiency; IO, inside versus outside; FL, fragment length; DS, disengagement score). After training, TOC uses RPF-covered ORFs to classify transcripts. (**B**) The combination of the four metrics separates coding ORFs, leaders and trailers of the training set. Transcripts lacking a protein-coding ORF cluster with trailers and leaders of the training set, as shown for three validated zebrafish lncRNAs (black). The density of each measure is shown along the axes.

2010), and should only stabilize the translating 80S ribosome during the ribosomal footprinting step. Moreover, the sizes of ribosome footprints isolated in ribosome profiling protocols (~30 nts) correspond to RNA fragments protected by 80S ribosomes (Wolin and Walter, 1988). The translation of ORFs within 5′ leaders is further supported by mass spectrometry data (Slavoff et al., 2013) and by observed enrichment of RPFs over sites of translation initiation in ribosome profiling data from harringtonine-treated (Ingolia et al., 2011), lactimidomycin-treated (Lee et al., 2012) and puromycin-treated (Fritsch et al., 2012) samples. Thus, leader-associated ribosome profiles are likely to represent actual translation of ORFs rather than ribosomal subunits scanning the transcript.

The lack of conservation of most uORFs suggests that the protein product might not be functional (Calvo et al., 2009; Hood et al., 2009; Neafsey and Galagan, 2007). Instead, ribosomal engagement with leader-like lncRNAs might be regulatory. Given the regulatory role of uORFs in some coding transcripts (Arribere and Gilbert, 2013; Calvo et al., 2009; Hinnebusch, 2005; Hood et al., 2009; Johansson and Jacobson, 2010), 5′ leader-like translation might affect lncRNA stability and/or subcellular localization. Translating ORFs within lncRNAs might target the transcript for nonsense-mediated decay (Tani et al., 2013), degrading it in the cytoplasm and/or retaining it in the nucleus (de Turris et al., 2011), resulting in the predominantly nuclear localization of most lncRNAs (Derrien et al., 2012). Prime candidates for such regulation are the minORF-containing lncRNAs for which the single amino acid product of their translation could not conceivably be functional. Alternatively, association of ribosomes with leader-like lncRNAs might be translational noise caused by the cytoplasmic location of 5′-capped and poly-adenylated transcripts. Such spurious translation may only be functional on evolutionary time scales as the source of novel coding genes (Carvunis et al., 2012).

In summary, our ribosome profiling data and translated ORF classifier allow the high-confidence annotation of coding and non-coding RNAs, complementing and extending previous computational approaches such as PhyloCSF. As demonstrated by our previously published pipeline (Pauli et al., 2012), these more traditional computational approaches can exclude the large majority of potential false-positives but misannotate some conserved lncRNAs as coding RNAs (e.g. *cyrano* and *megamind*) (Ulitsky et al., 2011) (Fig. 4). The use of additional approaches such as mass spectrometry will further improve the annotation of coding and non-coding RNAs in zebrafish (Slavoff et al., 2013).

Although our study has focused on the classification of lncRNAs, the accompanying ribosome profiling data will be a rich resource for the discovery of novel protein-coding genes that act during development. Our dataset increases the depth of previous ribosome profiling datasets in zebrafish by an order of magnitude (Bazzini et al., 2012) and expands the temporal coverage to five days of development. The nucleotide resolution of the data allows annotation of translated subregions of transcripts and the

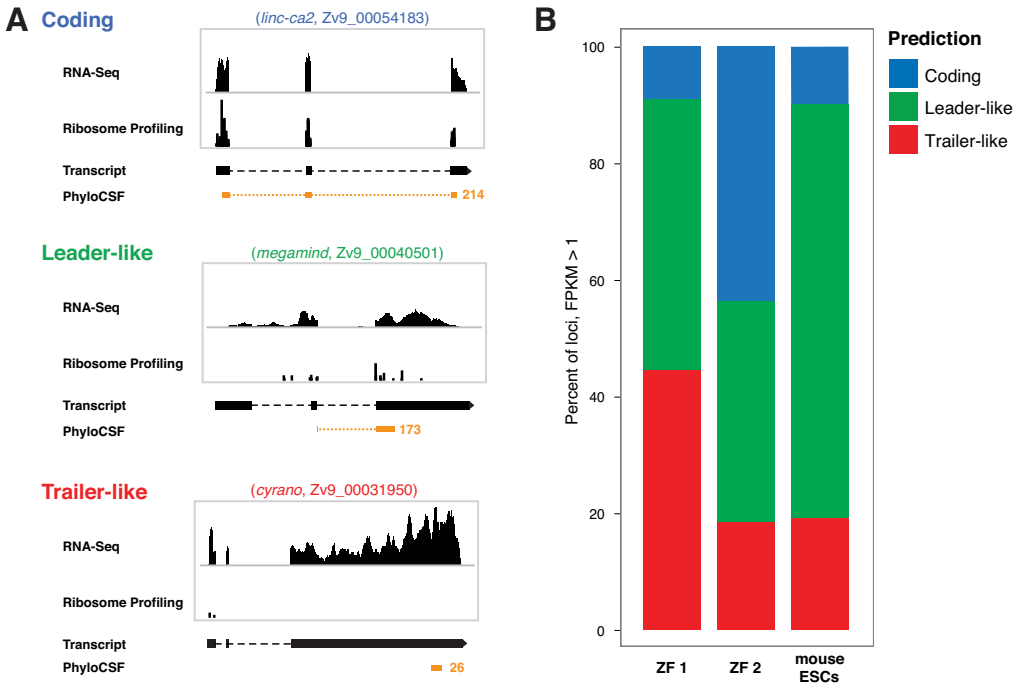

Fig. 4. TOC refines classification of lncRNAs. (**A**) TOC-based classification improves previous lncRNA predictions. Shown are RNA-seq and ribosome profiling read densities associated with three putative lncRNAs (Ulitsky et al., 2011), which had conflicting annotations in published zebrafish lncRNA sets (Pauli et al., 2012; Ulitsky et al., 2011). Transcript structures are shown in black. Introns are indicated as dashed lines. The region scoring highest in PhyloCSF (Lin et al., 2011) is indicated in orange. Whereas TOC reveals the protein-coding nature of *linc-ca2*, it confirms the non-coding nature of the two conserved lncRNAs *megamind* and *cyrano*. These two lncRNAs had been filtered out in the Pauli et al. lncRNA set owing to their relatively high phylogenetic codon substitution frequency scores (PhyloCSF >20). (**B**) Fraction of loci that are classified by TOC as coding (blue), leader-like (green) and trailer-like (red) in three collections of lncRNAs: ZF1 (Pauli et al., 2012), ZF2 (Ulitsky et al., 2011) and mESCs (Guttman et al., 2011).

identification of potential protein isoforms, furthering ongoing efforts to refine zebrafish genome annotation (Kettleborough et al., 2013). Finally, the quantitative nature of ribosome profiling combined with existing RNA-seq data will enable studies of post-transcriptional and translational regulation during zebrafish development.

Acknowledgements
We thank Jonathan Weissman and Nick Ingolia for helpful advice on ribosome profiling and for sharing mouse ESC ribosome profiling data; Mitchell Guttman for sharing results before publication; and Rahul Satija and Schraga Schwartz for discussions and comments on the manuscript.

Funding
G.-L.C. is supported by a Howard Hughes Medical Institute International Student Fellowship. A.P. and E.V. are supported by Human Frontier Science Program (HFSP) postdoctoral fellowships. This work was funded by the National Institutes of Health [grant numbers R01HG005111, R01GM056211 and R01HL109525 to J.L.R., A.R. and A.F.S.]. Deposited in PMC for release after 12 months.

Competing interests statement
The authors declare no competing financial interests.

Author contributions
A.P., A.F.S. and E.V. conceived the study. G.L.C. adapted and applied ribosome profiling to zebrafish, and analyzed resultant sequencing data, with support from A.P., A.F.S. and E.V. E.V. designed and implemented the classifier, with input from G.L.C. and A.P., and discussions with J.L.R., A.R. and A.F.S. G.L.C., A.P., E.V. and A.F.S. wrote the manuscript, with contributions from J.L.R. and A.R.

Supplementary material
Supplementary material available online at
http://dev.biologists.org/lookup/suppl/doi:10.1242/dev.098343/-/DC1

References
Arribere, J. A. and Gilbert, W. V. (2013). Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome Res.* [Epub ahead of print] doi: 10.1101/gr.150342.112

Bánfai, B., Jia, H., Khatun, J., Wood, E., Risk, B., Gundling, W. E., Jr, Kundaje, A., Gunawardena, H. P., Yu, Y., Xie, L. et al. (2012). Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res.* **22**, 1646-1657.

Barlow, D. P., Stöger, R., Herrmann, B. G., Saito, K. and Schweifer, N. (1991). The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus. *Nature* **349**, 84-87.

Bartolomei, M. S., Zemel, S. and Tilghman, S. M. (1991). Parental imprinting of the mouse H19 gene. *Nature* **351**, 153-155.

Bazzini, A. A., Lee, M. T. and Giraldez, A. J. (2012). Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science* **336**, 233-237.

Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., Rinn, J. L., Tongprasit, W., Samanta, M., Weissman, S. et al. (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242-2246.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E. et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816.

Borsani, G., Tonlorenzi, R., Simmler, M. C., Dandolo, L., Arnaud, D., Capra, V., Grompe, M., Pizzuti, A., Muzny, D., Lawrence, C. et al. (1991). Characterization of a murine gene expressed from the inactive X chromosome. *Nature* **351**, 325-329.

Brar, G. A., Yassour, M., Friedman, N., Regev, A., Ingolia, N. T. and Weissman, J. S. (2012). High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* **335**, 552-557.

Breiman, L. (2001). Random forests. *Mach. Learn.* **45**, 5-32.

Brockdorff, N., Ashworth, A., Kay, G. F., McCabe, V. M., Norris, D. P., Cooper, P. J., Swift, S. and Rastan, S. (1992). The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* **71**, 515-526.

Brown, C. J., Hendrich, B. D., Rupert, J. L., Lafrenière, R. G., Xing, Y., Lawrence, J. and Willard, H. F. (1992). The human XIST gene: analysis of a 17

kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* **71**, 527-542.
Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. and Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915-1927.
Calvo, S. E., Pagliarini, D. J. and Mootha, V. K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc. Natl. Acad. Sci. USA* **106**, 7507-7512.
Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C. et al.; FANTOM Consortium; RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) (2005). The transcriptional landscape of the mammalian genome. *Science* **309**, 1559-1563.
Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., Charloteaux, B., Hidalgo, C. A., Barbette, J., Santhanam, B. et al. (2012). Proto-genes and de novo gene birth. *Nature* **487**, 370-374.
Collins, J. E., White, S., Searle, S. M. J. and Stemple, D. L. (2012). Incorporating RNA-seq data into the zebrafish Ensembl genebuild. *Genome Res.* **22**, 2067-2078.
de Turris, V., Nicholson, P., Orozco, R. Z., Singer, R. H. and Mühlemann, O. (2011). Cotranscriptional effect of a premature termination codon revealed by live-cell imaging. *RNA* **17**, 2094-2107.
Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G. et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775-1789.
Dinger, M. E., Pang, K. C., Mercer, T. R. and Mattick, J. S. (2008). Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLOS Comput. Biol.* **4**, e1000176.
Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F. et al. (2012). Landscape of transcription in human cells. *Nature* **489**, 101-108.
Fejes-Toth, K., Sotirova, V., Sachidanandam, R., Assaf, G., Hannon, G. J., Kapranov, P., Foissac, S., Willingham, A. T., Duttagupta, R., Dumais, E. et al.; Affymetrix ENCODE Transcriptome Project; Cold Spring Harbor Laboratory ENCODE Transcriptome Project (2009). Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* **457**, 1028-1032.
Fritsch, C., Herrmann, A., Nothnagel, M., Szafranski, K., Huse, K., Schumann, F., Schreiber, S., Platzer, M., Krawczak, M., Hampe, J. et al. (2012). Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome Res.* **22**, 2208-2218.
Guttman, M. and Rinn, J. L. (2012). Modular regulatory principles of large non-coding RNAs. *Nature* **482**, 339-346.
Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P. 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, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., 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, B. W., Garber, M., Grenier, J. K., Munson, G., Young, G., Lucas, A. B., Ach, R., Bruhn, L. et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**, 295-300.
Hinnebusch, A. G. (2005). Translational regulation of GCN4 and the general amino acid control of yeast. *Annu. Rev. Microbiol.* **59**, 407-450.
Hood, H. M., Neafsey, D. E., Galagan, J. and Sachs, M. S. (2009). Evolutionary roles of upstream open reading frames in mediating gene regulation in fungi. *Annu. Rev. Microbiol.* **63**, 385-409.
Ietswaart, R., Wu, Z. and Dean, C. (2012). Flowering time control: another window to the connection between antisense RNA and chromatin. *Trends Genet.* **28**, 445-453.
Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. and Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218-223.
Ingolia, N. T., Lareau, L. F. and Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789-802.
Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M. and Weissman, J. S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **7**, 1534-1550.
Jackson, R. J., Kaminski, A. and Pöyry, T. A. A. (2007). Coupled termination-reinitiation events in mRNA translation. In *Translational Control in Biology and Medicine* (ed. M. B. Mathews, N. Sonenberg and J. W. B. Hershey), pp. 197-223. Cold Spring Harbor, New York, NY: Cold Spring Harbor Laboratory Press.
Jinno, Y., Ikeda, Y., Yun, K., Maw, M., Masuzaki, H., Fukuda, H., Inuzuka, K., Fujishita, A., Ohtani, Y., Okimoto, T. et al. (1995). Establishment of functional imprinting of the H19 gene in human developing placentae. *Nat. Genet.* **10**, 318-324.
Johansson, M. J. O. and Jacobson, A. (2010). Nonsense-mediated mRNA decay maintains translational fidelity by limiting magnesium uptake. *Genes Dev.* **24**, 1491-1495.
Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P. A. and Gingeras, T. R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916-919.
Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermüller, J., Hofacker, I. L. et al. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484-1488.
Kettleborough, R. N. W., Busch-Nentwich, E. M., Harvey, S. A., Dooley, C. M., de Bruijn, E., van Eeden, F., Sealy, I., White, R. J., Herd, C., Nijman, I. J. et al. (2013). A systematic genome-wide analysis of zebrafish protein-coding gene function. *Nature* **496**, 494-497.
Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B. and Schilling, T. F. (1995). Stages of embryonic development of the zebrafish. *Dev. Dyn.* **203**, 253-310.
Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-359.
Lee, S., Liu, B., Lee, S., Huang, S.-X., Shen, B. and Qian, S.-B. (2012). Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. USA* **109**, E2424-E2432.
Lin, M. F., Jungreis, I. and Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275-i282.
Neafsey, D. E. and Galagan, J. E. (2007). Dual modes of natural selection on upstream open reading frames. *Mol. Biol. Evol.* **24**, 1744-1751.
Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H. et al.; FANTOM Consortium; RIKEN Genome Exploration Research Group Phase I and II Team (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563-573.
Pauli, A., Rinn, J. L. and Schier, A. F. (2011). Non-coding RNAs as regulators of embryogenesis. *Nat. Rev. Genet.* **12**, 136-149.
Pauli, A., Valen, E., Lin, M. F., Garber, M., Vastenhouw, N. L., Levin, J. Z., Fan, L., Sandelin, A., Rinn, J. L., Regev, A. et al. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res.* **22**, 577-591.
Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41 Database issue**, D590-D596.
Ravasi, T., Suzuki, H., Pang, K. C., Katayama, S., Furuno, M., Okunishi, R., Fukuda, S., Ru, K., Frith, M. C., Gongora, M. M. et al. (2006). Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.* **16**, 11-19.
Rinn, J. L. and Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81**, 145-166.
Schneider-Poetsch, T., Ju, J., Eyler, D. E., Dang, Y., Bhat, S., Merrick, W. C., Green, R., Shen, B. and Liu, J. O. (2010). Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. *Nat. Chem. Biol.* **6**, 209-217.
Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., Karger, A. D., Budnik, B. A., Rinn, J. L. and Saghatelian, A. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* **9**, 59-64.
Sleutels, F., Zwart, R. and Barlow, D. P. (2002). The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415**, 810-813.
Swiezewski, S., Liu, F., Magusin, A. and Dean, C. (2009). Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. *Nature* **462**, 799-802.
Tani, H., Torimura, M. and Akimitsu, N. (2013). The RNA degradation pathway regulates the function of GAS5 a non-coding RNA in mammalian cells. *PLoS ONE* **8**, e55684.
Tilgner, H., Knowles, D. G., Johnson, R., Davis, C. A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T. R. and Guigó, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* **22**, 1616-1625.
Trapnell, C., Pachter, L. and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111.
Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H. and Bartel, D. P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**, 1537-1550.
Wang, K. C., Yang, Y. W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B. R., Protacio, A., Flynn, R. A., Gupta, R. A. et al. (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**, 120-124.
Wolin, S. L. and Walter, P. (1988). Ribosome pausing and stacking during translation of a eukaryotic mRNA. *EMBO J.* **7**, 3559-3569.