SUPPLEMENTAL FIGURES

Figure S1. Chromatin Maps and Properties of Zebrafish lincRNAs, Related to Figure 1.

(A) H3K4me3 ChIP-Seq read density around the transcription start site at 24 hpf. Each read was extended by 76–109 bp using MACS, and the number of extended reads covering each bin of 10 bp was tallied and divided by the total number of mapped reads. Ensembl zebrafish genes were divided into four quartiles based on their RNA-Seq read density at 24 hpf, and the aggregate density is plotted for each quartile.

(B) H3K36me3 ChIP-Seq read density around 5' and 3' ends of genes at 24 hpf. Each read was extended 600 bp downstream of the 3' end and 400 bp upstream of the 5' end. Otherwise, as in (A).

(C) Overlap between H3K4me3 peak sets detected at different developmental stages.

(D) Distributions of the H3K36me3 reads at 24 hpf in lincRNA domains and in random intergenic regions. RPK is the number of reads per kilobase.

(E) A/U content of zebrafish lincRNAs and different regions of protein-coding genes.

(F) A/U Content of mouse lincRNAs and different regions of protein-coding genes. Colors are as in panel (E).

(G) Fraction of homopolymeric sequence in lincRNAs and different regions of protein-coding genes of zebrafish. Sequences were evaluated using a 5-nt sliding window and recording the fraction of 5-mers that were homopolymers (i.e., five identical nucleotides). Colors are as in panel (E).

(H) Fraction of homopolymeric sequence in lincRNAs and different regions of protein-coding genes of mouse. Otherwise, as in (G).

(I) Fraction of transcripts overlapping at least one base of a repetitive element (left) and fraction of all transcript sequence overlapping repeats (right). With the exception of simple repeats, all the repetitive elements annotated by RepeatMasker in the UCSC genome browser (November 2011) were used.

(J) Distances from the closest protein-coding gene. The controls are random intergenic regions size- and chromosome-matched to the lincRNA set.

(K) Relative orientation with respect to the closest protein-coding gene within 100,000 bases. Error bars indicate 95% confidence intervals, based on 1,000 cohorts of control regions, which were each size- and chromosome-matched to the lincRNA set.

(L) Enrichments of Gene Ontology categories for the protein-coding neighbors of lincRNAs in different species. Error bars indicate 95% confidence intervals, based on 1,000 cohorts of random size- and chromosome-matched intergenic controls.

Figure S2. Whole-mount *in situ* hybridizations using sense probes for selected lincRNAs, Related to Figure 2.

Figure S3. Detailed Genomic Characterization of *linc-oip5* (*cyrano*) and *linc-birc6* (*megamind*) genes, Related to Figure 4.

(A) The zebrafish *cyrano* locus. The RNA-Seq track is a composite of reads from ten stages and tissues (SRA Accession ERP000016).

(B) The mouse *cyrano* locus. The isoforms shown were the ones annotated in Ensembl (ENSMUST00000147425 and ENSMUST00000153581). The longer isoform was used in rescue experiments. Neural progenitor cell (NPC) RNA-Seq data was from (Guttman et al., 2010), otherwise, as in (A).

(C) The human *cyrano* locus. Brain RNA-Seq data was from (Wang et al., 2008), otherwise as in (A). The presented gene structure corresponds to the one that was annotated in Ensembl (ENSMUST00000153581) and the one used in our rescue experiments. Dashed line indicates a likely longer isoform inferred from RNA-Seq data.

(D) Conservation of the *cyrano* conserved site. The sequence logo based on all 45 homologous sequences is shown above representative examples from the indicated genomes.

(E) Zebrafish *megamind* locus. The black box indicates the conserved region. Otherwise as in (A).

(F) Zebrafish *linc-hhipl1* locus. Black box indicates the region homologous with *megamind*. Otherwise as in (A).

(G) An additional locus on chr17 with homology to *megamind* and evidence of transcription. Otherwise as in (F). This locus, named *linc-bdkrb2*, was manually constructed after recognizing its homology to *megamind*, and is not represented by the 567 lincRNAs in Table S2.

(H) Conservation of the *megamind* conserved site. The sequence logo based on all 75 homologous sequences is shown above representative examples. Bases conserved in all the representative examples are in bold and in black.

Figure S4. Characterization of linc-*oip5* (cyrano) and NeuroD expression in zebrafish embryos, Related to Figure 5.

(A) *In situ* hybridization showing cyrano expression in the brain and notochord of zebrafish embryos at 28 hpf.

(B) Reduced cyrano levels in embryos injected with the *cyrano* splice-site MO measured by qRT-PCR.

(C) *In situ* hybridization of cyrano in embryos injected with the *cyrano* splice-site MO.

(D) Embryos at 48 hpf that had been injected with the indicated reagents. Bottom panel shows subsets of neurons expressing GFP driven by the *neurod* promoter. Arrows point at NeuroD-positive neurons in the retina and tectum at 48 hpf.

(E) Embryo co-injected with mock RNA (*RFP*) and the indicated MOs. Bottom panel shows RFP expression.

Figure S5. Characterization of linc-*birc6* (*megamind*) expression in zebrafish embryos, Related to Figure 6.

(A) *In situ* hybridization showing megamind expression in the brain of zebrafish embryos at 72 hpf.

(B) Reduced megamind levels in embryos injected with the *megamind* splice-site MOs measured by qRT-PCR.

(C) *In situ* hybridization of megamind expression in zebrafish embryos injected with the splice site MOs.

(D) Embryo at 28 hpf that had been injected with the indicated reagents. Bottom panel shows RFP expression. Arrow points at defects in brain development.

SUPPLEMENTAL TABLES

Table S1. Statistics for High Throughput Sequencing Datasets, Related to Experimental Procedures, Related to Experimental Procedures.

Library	Total reads	≤4 alignments	0 alignments	>4 alignments
H3K4me3 IP 24 hpf	12,189,261	8,449,068	1,832,467	1,907,726
H3K4me3 IP 72 hpf	16,657,288	10,559,648	1,358,270	4,739,370
H3K4me3 IP Adult	9,146,789	5,487,668	1,114,291	2,544,830
H3K36me3 IP 24 hpf	12,939,306	6,834,681	1,008,475	5,096,150
H3K36me3 IP 72 hpf	19,956,175	11,589,103	1,813,807	6,553,265
H3K36me3 IP Adult	13,666,859	7,521,044	1,429,574	4,716,241
Input 24 hpf (control for H3K4me3)	9,961,209	4,877,619	1,762,600	3,320,990
Input 24 hpf (control for H3K36me3)	10,886,048	5,726,839	924,820	4,234,389
Input 72 hpf (control for H3K4me3)	17,892,954	9,105,274	1,891,952	6,895,728
Input 72 hpf (control for H3K36me3)	19,069,127	10,872,021	1,459,677	6,737,429
Input Adult	14,041,883	7,523,101	1,537,844	4,980,938
Strand-specific RNA-Seq 24 hpf	37,203,901	18,307,186	5,337,724	13,558,991
Strand-specific RNA-Seq 72 hpf	37,004,726	15,410,622	3,787,831	17,806,273
3P-Seq 24 hpf	22,448,667	13,179,736	8,524,042	744,889
3P-Seq 72 hpf	20,914,473	17,818,974	1,984,988	1,110,511
3P-Seq Adult	18,445,044	14,743,023	2,493,766	1,208,255

Table S2. H3K4me3 Peaks Identified Using MACS, Poly(A) Sites Identified by 3P-Seq and lincRNA Exon-Intron Structures, Related to Figure 1.

MACS score, fold-enrichment and FDR computations are as described in the MACS manual (http://liulab.dfci.harvard.edu/MACS/00README.html). These peaks were detected using MACS applied to H3K4me3 tags from 24 hpf (part A), 72 hpf (part B) and adult (part C) fish. Genomic coordinates are zero-based and refer to the danRer7 assembly. 3P tags often fall into clusters because of micro-heterogeneity of 3' ends. For each cluster, the poly(A) site specified is the one used most often, as inferred by the number of supporting tags. For lincRNA structures, exon positions are relative to the start coordinate of the transcript (as used in BED file formats for the UCSC browser).

Table S3. Human and Mouse lincRNAs, Related to Figure 3.

(Part A) Human lincRNAs used in this study. Genomic positions refer to the hg18 assembly.

(Part B) Mouse lincRNAs used in this study. Genomic positions refer to the mm9 assembly.

Table S4. Spatial Expression of lincRNAs Analyzed Using *in situ* Hybridization, Related to Figure 2. n/d, not detected.

lincRNA	Genomic location	Expression at 24 hpf	Expression at 72 hpf
malat1	chr14:48566202-48573730	ubiquitous, enriched in the brain, mucous cells	brain, mucous cells
linc-*mipep1*	chr10:40425061-40428902	enriched in the brain, spinal cord, blood vessels	enriched in the brain
linc-*bin2a*	chr23:33944861-33952782	lens	n/d
linc-*cldn7a*	chr7:23796995-23800090	pronephros, cranial ganglia	n/d
linc-*gtf2f2b*	chr9:19526783-19529739	cranial ganglia	n/d
linc-*epb4.1l4*	chr10:1734816-1745335	enriched in CNS	Brain
linc-*srd5a2a*	chr1:51582321-51592093	cranial ganglia, nose	n/d
linc-*prr14*	chr3:32992137-32996372	n/d	n/d
linc-*agpat3*	chr1:47327616-47330238	n/d	n/d
linc-*mettl3*	chr7:23027784-23042820	n/d	specific hindbrain neurons
linc-*csnk1a1*	chr14:40176987-40183099	n/d	cartilage of the jaw, nose epithelium
cyrano	chr13:33484735-33491213	brain, notochord,	brain, notochord, spinal cord
linc-*loc100001135*	chr7:4334540-4358502	n/d	n/d
linc-*onecut1*	chr18:37186817-37191791	n/d	n/d
linc-*pou3f3b-2*	chr6:14538090-14542184	n/d	n/d
linc-*meis1*	chr13:5245604-5250645	n/d	n/d
linc-*arid4a*	chr17:11281895-11282994	brain, eye, spinal cord	n/d
linc-*setd1ba*	chr10:43348128-43351841	specific neurons	specific neurons
megamind	chr17:22517187-22519802	brain, eye	brain
linc-*trpc7*	chr14:1668370-1672051	brain, notochord	brain, notochord
linc-*elovl1a*	chr2:19346942-19351962	specific face neurons	n/d
linc-*plcb2*	chr17:2208928-2210522	n/d	n/d
linc-*tbx2b*	chr15:26704323-26717885	dorsal retina, ear	n/d
linc-*rogdi*	chr3:36765129-36766960	n/d	n/d

Table S5. Number of Embryos in *cyrano* and *megamind* Experiments, Related to Figures 5 and 6.

Experiment	wild type	mutant
Control *cyrano* MO1	32	0
Control *cyrano* MO2	45	0
Conserved site *cyrano* MO	6	69
Splice *cyrano* MO	25	157
Splice *cyrano* MO + RFP RNA	7	53
Splice *cyrano* MO + zebrafish cyrano RNA	50	54
Splice *cyrano* MO + mouse cyrano RNA	21	11
Splice *cyrano* MO + human cyrano RNA	27	34
Splice *cyrano* MO + cyrano_mut_a RNA	19	45
Splice *cyrano* MO + cyrano_mut_b RNA	22	57
Splice *cyrano* MO + cyrano_mut_a+b RNA	13	80
Splice *cyrano* MO + cyrano conserved site RNA	5	26
Splice *cyrano* MO + hybrid 1 RNA	16	65
Control *megamind* MO1	46	0
Control *megamind* MO2	61	0
Conserved site *megamind* MO	10	83
Splice *megamind* MOs	19	174
Splice *megamind* MOs + RFP megamind RNA	5	39
Splice *megamind* MOs + zebrafish megamind RNA	53	60
Splice *megamind* MOs + mouse megamind RNA	32	15
Splice *megamind* MOs + human megamind RNA	35	43
Splice *megamind* MOs + megamind_stop RNA	38	38
Splice *megamind* MOs + megamind_frameshift RNA	25	29
Splice *megamind* MOs + megamind_mut_a RNA	20	35
Splice *megamind* MOs + megamind_mut_b RNA	27	32
Splice *megamind* MOs + megamind_mut_a+b RNA	8	72
Splice *megamind* MOs + megamind conserved site RNA	9	47
Splice *megamind* MOs + hybrid 2 RNA	16	107

Table S6. Morpholino Sequences and Concentrations, Related to Experimental Procedures.

Morpholino	Sequence (5'→3')	Targeting description
cyrano e1i1 MO (5.5 ng)	AACACTCATCCCGCACTTACCGTCA	*cyrano* intron 1 5' splice site
cyrano e2i2 MO (5.5 ng)	TGCTGTTTTTGATGACCTACCTGGT	*cyrano* intron 2 5' splice site
cyrano i2e3 MO (5.5 ng)	TCATCTGCACAGAATGGACATTTGA	*cyrano* intron 2 3' splice site
cyrano conserved site MO (5 ng)	ATTGGTGATTTTGTTGTTTTTGCGA	*cyrano* conserved site in exon 3
cyrano control MO1 (4 ng)	ATTGGT**C**ATTTT**C**TT**C**TTT**A**T**C**CGA	Same as *cyrano* conserved site MO but with five mismatches (underlined)
cyrano control MO2 (4 ng)	ACTAGGAATAATCTACCCACAGCTC	Non-conserved region in *cyrano* exon 3
megamind e1i1 MO (1.6 ng)	GTAGAAAAACTGGCCCCCACCTTCT	*megamind* intron 1 5' splice site
megamind i2e3 MO (1.6 ng)	ATGAAAATAGGGAGTCTTACCCTAC	*megamind* intron 2 3' splice site
megamind conserved site MO (5 ng)	TGATCCCCAGAAGGGCCAATATGGA	*megamind* conserved site in exon 3
megamind control MO1 (4 ng)	TG**T**T**G**CCCA**C**AAGG**C**CCAATAT**C**GA	Same as *megamind* conserved site MO but with five mismatches (highlighted)
megamind control MO2 (4 ng)	GCATTTTCCTTTGCACAGAAACAAC	Non-conserved region in *megamind* exon 3

Table S7. Oligonucleotide Sequences related to Experimental Procedures

Amplicon	PCR primer pairs (5'→3').
Insert for *in situ* probe template	
malat1	GACGTTTTCCGTTGGTTATACAAAGGTT AGTTGTAACACATTTACATTATAGCTGGC
linc-*mipep*	GCTCAACACAGTGTCGACTGTTTTTTCAGCGT TCAGAACGCTTTACAACTAAAGAGATC
linc-*bin2a*	GGTCATCGCCCTGATCCTGCTGACCCT ACAAGGAACATAATATTGTAACCCTGCACAAAACAC
linc-*cldn7a*	CTTCCGACTAGCGCCGAACAAACCGACACAGA AATGTCAAGGTAGACTCCAGTTACCAAG
linc-*gtf2f2b*	TCGAAGAATAGCTTGAAGAAACAGACGCAATCCCTG ACTGCAGCATTCATGGTTCGGGTGCTC
linc-*epb4.1l4*	GACTTTAATCTGCTCCTTGGTAAGGAAGCTCAG TGCTCCGACCGTCTTGGATTTCTGAGTTTCGC
linc-*srd5a2a*	AGGACCCAAAATGGTGGCGGCGTGAGTGAAAAC TAATACGACTCACTATAGGAGCCGGCGCAGGCTGAGCGACGTACGACAC
linc-*prr14*	GACACTGTGAAACTGTTTATGACT AATGAATGCCTTAATACTCTCAGGATGGC
linc-*agpat3*	GAAGTCGTTACACAAACCGTCTGTCCAAGCAGA ATTACACAGTGATGCCATAATCAATTCAAC
linc-*mettl3*	GCTGAACGAGTCTCTCTACATCACCAGTGA ACACTGGCCAATGCCTACTTTGCACACTG
linc-*csnk1a1*	CAACACCTGCTGAGTTTCCCACTCTAAACTCGCTCA TGGCATATTTATGGTTATTAGTTGTATTGACTGGACAGC
cyrano	GGTAATCACTATTAGTTGATGATAACGTCATAGCATGCT AGTCACAACACTGGTCCACTCATAGATTTAGTGTC
linc-*loc100001135*	CTCGAAGCCTGTCTTATTCATCTATCTCCTCACTTACGGT TCTCACAGTTGATATAAACAGAGTGCCATTGTGC
linc-*onecut1*	GCACGGATAACAGAATCTAGAGGCGAGAGACAAGCA ATTGTTGTGCTATTAAGAGTAACGAACCAAGCCATC
linc-*pou3f3b-2*	CAGGGAGAGGGGCCTCCTTTCTACACTGGACCCA ATGACCGTACATGAAAGAAGAGGGTGGAGAACAGGTTC
linc-*meis1*	GCTGTGGTTCAGAAGTCAAACGGAGGTCATCCTTTAT TGAAATGCAAATTCCGTTACTTAAACTTTC
linc-*arid4a*	CCTATATGACTGCTTCAGCTCAGCATTCGACTAGGTTGCA TCCACAGAAACGACATAAAGACGCCATTACCGTTGC
linc-*setd1ba*	GCCTTTAAGTACAATTATTGTTTCCTCACTGTGT

	ACTACACAAAATAGATTAGAATCCACATTTATTATAG
megamind	GCACATCTGTAGGGCTTCTACACCCACAGAAAAAGCGGA AGATGGGTTTCAGAGTCTAACATTCTTCTCTGTTAATTC
linc-*trpc7*	GTGAGGATACTGCGAGCCGTCATGCGCGCTCGCTGATCTTC TCAGTATATACAGACAAATCAGCGAGTTTCAGTCCACGCTGGTC
linc-*elovl1a*	GGTTGGGAAGCGCAACATAAACTGCATGCTAACAAACAAT TCAGAAAATGTAAAGACTACAGTGAAAACAGGTGTGGAC
linc-*plcb2*	GATATTTGACCTCAGAAACATCTCAGTCTTCA AGCCCATGCAAGCTTTGTATTTTCTTAAAAAGCTCGCGTC
linc-*rogdi*	GCGGGACCATTCTGACTGAAGTCATGGACAAAAGCA TCATACAAATAATGTCATCCATTCAAAACATTCCAAC
linc-*tbx2b*	GATACACAGACAACCAAGACTTAAGATTATTGACGTGTA TAAACGCACACATGATGTTTCGCAGTGCAGTTTGTGGAAC
Insert for expression constructs	
Zebrafish *megamind* full-length	GCAATGCACGGCGCTCTCAGGCTCCGAGACGGGACCTATA GTATGTAATCATGTATCAATACAAAAGCATTTTCCTTTGCA
Zebrafish *cyrano* full-length	GACCGAAATGGCGTAACGCGCAGTCGAGCACCGCAGCAGCGCA TACAAAACCATGCGGGACGCTTCTGTAGTGCATAGATCA
cyrano conserved site for the hybrid 1 construct	GGAAGATCTGTATATTGTACAAACAAGTGACAAGTTGTTCGCA GGAAGATCTACCCTAAAGCAAGCACATGAAACTATACATC
megamind conserved site for the hybrid 2 construct	ACTAGTAGGCCTGTAAAGAGGAGCGAGAGGAGTCCATA ACTAGTAGGCCTTTGTGTAGATGTAAACAAACACAATGACGAAGAG
T7 *in vitro* transcription templates	
cyrano conserved site	TAATACGACTCACTATAGGGGTATATTGTACAAACAAGTGACAAGTTGTTCGCA ACCCTAAAGCAAGCACATGAAACTATACATC
megamind conserved site	TAATACGACTCACTATAGGGGTAAAGAGGAGCGAGAGGAGTCCATA TTGTGTAGATGTAAACAAACACAATGACGAAGAG
cyrano RNA-blot probe	CCTCAATGACTGGAATGCAA TTCTAATACGACTCACTATAGGAAGAGCAAAAGCCCTGCATA
qRT-PCR	
Zebrafish cyrano	ACAAACCAAGACAGGCAGTGGCA TGCAACTCAATAGCACCCCGCT
Zebrafish megamind	GCAATGCACGGCGCTCTCAGGCTCCGAGACGGGACCTATA GCATTTTCCTTTGCACAGAAACAACGTGTCTGACACTGCACT
β-actin1	CTCTTCCAGCCTTCCTTCCT CTTCTGCATACGGTCAGCAA

EXTENDED EXPERIMENTAL PROCEDURES

Data Sources

Zebrafish genome assembly Zv9 was used throughout the study. Only the 25 chromosomes (and not contigs or scaffolds) were used for lincRNA discovery. Zebrafish Ensembl and RNA-Seq-based gene structures were obtained from Ensembl version 60. RNA-Seq from ten developmental stages and tissues, and strand-specific RNA-Seq data from six early development stages were obtained from NCBI SRA database (accessions ERP000016 and SRP003165). Genome alignments of RefSeq transcripts from zebrafish and other organisms, GenBank mRNAs and ESTs, as well as annotations of repetitive elements, whole-genome alignments and PhastCons scores were obtained from the UCSC genome browser (July 2011). Mouse and human orthologs (including putative orthologs) of zebrafish protein-coding genes were obtained from Ensembl.

ChIP-Seq

Zebrafish were maintained and staged using standard procedures (Kimmel et al., 1995). Anesthetized 24 hpf and 72 hpf decorioneted embryos were washed three times in PBS (137 mM NaCl, 2.7 mM KCl, 1.5 mM KH_2PO_4, 8 mM Na_2HP0_4, pH 7.4) and suspended in PBS containing 1% freshly added formaldehyde. Embryos were transferred to a dounce homogenizer, dounced several times and incubated at room temperature for 15 min. Formaldehyde was quenched by adding 1/20 volume 2.5 M glycine. Cells were pelleted at 400 x g for 5 min. The supernatant was removed, and pellets were rinsed twice with PBS, flash frozen in liquid nitrogen and stored at –80^{0}C. For adult fish, anesthetized fish were homogenized in the TissueRuptor (Qiagen), followed by the same crosslinking protocol. ChIP-Seq was carried out as described (Guenther et al., 2008) using α-H3K4me3 and α-H3K36me3 antibodies (Abcam), and input DNA was sequenced to assess the FDR. Illumina reads were aligned to the zebrafish genome using Bowtie (Langmead et al., 2009), allowing for up to one mismatch and carrying forward only reads that mapped to no more than four genomic positions (Table S1). MACS 1.4 with default parameters (Zhang et al., 2008) was used to identify peaks of H3K4me3 enrichment and to assess the FDR using sequencing reads from input DNA. Peaks from an FDR cutoff of 0.1 (Table S2) were used in subsequent analyses. H3K36me3 reads were extended 600 bp downstream and 400 bp upstream prior to analyses.

3P-Seq

Total RNA was isolated using TRI Reagent (Ambion). The 3P-Seq protocol and the mapping of reads to the genome was as described (Jan et al., 2011). Only poly(A) sites supported by at least four reads and multiple distinct tags within 30 bp of each other were considered. Distinct tags were reads ending at different positions, sequenced in different libraries or with different numbers of untemplated adenosines. Sites that appeared within 30 bp from each other were grouped into clusters, and within each cluster the poly(A) site supported by the most 3P tags was carried forward (Table S2).

RNA-Seq

Poly(A)-selected RNA was amplified using RiboAmp Plus RNA Amplification kit (Applied Biosystems) and strand-specific RNA-Seq of 24- and 72-hpf embryos (Table S1) was performed as described (Guo et al., 2010).

Identification of lincRNA Domains

Poly(A) sites overlapping mRNA introns or exons, and those appearing up to 5 kb downstream of 3'UTR ends (annotated in Ensembl or RefSeq) with connectivity to the annotated transcript supported by RNA-Seq were excluded from the lincRNA discovery pipeline. The use of our chromatin map and 3P-Seq data to systematically annotate the promoters and poly(A) sites of proten-coding genes and small-RNA precursors will be described elsewhere. The remaining poly(A) sites, which could not be assigned to transcripts of known genes, were used as the starting points for lincRNA discovery. These were first filtered to exclude those that mapped to repetitive regions (taken from the UCSC genome browser, November 2010, excluding simple repeats). For each remaining poly(A) site, the three closest upstream H3K4me3 peaks were identified at each stage, restricting the search to the 100 kb region upstream of the poly(A) site. Only cases in which the H3K4me3 peak began at least 750 bp upstream of the poly(A) site were carried forward. Overlapping H3K4me3 peaks from the three stages were merged. The 5' end of the H3K4me3 peak was used as a putative start of the lincRNA domain and the poly(A) site as its putative end point. To exclude protein-coding genes and precursors of annotated small RNAs, a collection of known or predicted protein-coding genes and ncRNAs shorter than 200 bp was compiled from both Ensembl and RefSeq, and from alignments of RefSeq transcripts from other organisms to the zebrafish genome. To this collection we added gene structures predicted using RNA-Seq (obtained from Ensembl) or mRNAs from GenBank that were predicted to have significant protein-coding potential (due to a long ORF or significant similarity to an annotated protein-coding gene) using Coding Potential Calculator (CPC) (Kong et al., 2007). In order to avoid transcripts potentially misclassified as protein-coding, CPC was constrained to ignore predicted protein-coding structures in RefSeq “XP” accessions. Any transcript with a positive CPC score (“coding” and “weakly coding” categories) was classified as an mRNA. lincRNA domains that overlapped the coding sequence of an mRNA, or overlapped on the sense strand any exon of an mRNA, were excluded from further consideration. To avoid discarding lincRNAs that share promoters with protein-coding genes, only regions spanning the 3' ends of the H3K4me3 peaks and the poly(A) sites were tested for overlap with coding sequences.

Prediction of lincRNA Exonic Structures

To assign exon-intron structures to the predicted lincRNA domains, we combined transcript information from GenBank (mRNA and EST alignments) and RNA-Seq–based gene models identified by Exonerate (Slater and Birney, 2005) and deposited in Ensembl (“Ensembl RNA-Seq” structures). In addition, we predicted RNA-Seq models using Cufflinks (Trapnell et al., 2010), starting with RNA-Seq reads from 10 stages and tissues (SRA Accession ERP000016) aligned to the genome using TopHat (Trapnell et al., 2009) v1.1.4 (retaining only transcripts ≥100 nt and with RPKM >0.5). To assign spliced exonic structures, we first sought primary-transcript structures that spanned the entire predicted domain from the H3K4me3 mark to the poly(A) site. If no such structure was found, the structure with the longest exonic sequence was identified, requiring that it spans at least 25% of the predicted domain. If two possible exonic structures overlapped by more than 80%, only the structure with the longer exonic sequence was retained. Remaining structures were trimmed based on the 5' and 3' ends of the domain, and only exons falling within the boundaries of the domain were retained.

Filtering of Predicted lincRNAs Transcripts

We first filtered the data using RNA-Seq data from ten developmental stages/tissues and H3K36me3 chromatin maps. The spliced exonic structures were filtered requiring RNA-Seq coverage across at least 200 nt, RNA-Seq coverage across at least 50% of the predicted exonic nucleotides, at least 300 RNA-Seq reads mapping to the predicted spliced structure, and H3K36me3 coverage across at least 50% of the predicted transcribed domain in at least one stage. To confirm that the predicted lincRNAs were transcribed predominantly in the predicted orientation, we used our strand-specific RNA-Seq reads from 24- and 72-hpf embryos (Table S1) and strand-specific RNA-Seq reads from six early developmental stages acquired by Aanes et al. (2011) using the SOLiD platform (GEO Accession GSE22830). Using these 284 million reads, we estimated that over 95% of the domains that did not overlap UTRs were transcribed mostly from the predicted strand; domains that did not meet these criteria were excluded from further analysis. Finally, an additional filtering for coding potential using CPC removed all transcripts with a positive coding score.

Human and Mouse lincRNA Collections

A set of 2,458 human lincRNAs (Table S3) was obtained by using all the RefSeq long (>200 bp) noncoding transcripts, Ensembl genes ("lincRNA" category only) and UCSC genes, after excluding transcripts overlapping protein-coding genes from Ensembl, RefSeq, coding RefSeq transcripts from other species mapped to the human genome in the UCSC genome browser, pseudogenes and "other RNAs" annotated in Ensembl. When evaluating overlap with protein-coding genes, transcripts were excluded that overlapped either any part of the pre-mRNA in the sense orientation or exons in the antisense orientation. A set of 3,345 mouse lincRNAs (Table S3) was obtained from (Guttman et al., 2010) (clustering overlapping transcripts and using the union of their exons), RefSeq long (>200 bp) noncoding transcripts and Ensembl genes ("lincRNA" category only), after excluding transcripts overlapping either protein-coding or pseudogenes annotated in RefSeq and Ensembl, or with known small RNA genes.

Additional Bioinformatic Analyses

For each lincRNA locus, a computational control was generated by random sampling of a length-matched region from intergenic space of the same chromosome. Within this control region, exons were assigned to the same relative positions as in the authentic lincRNA locus. To estimate confidence intervals, 200–1000 cohorts of computational controls were used. RPKM values were computed from RNA-Seq using Cufflinks (Trapnell et al., 2010). NCBI BLASTN was used to find sequence-similar lincRNAs using the parameters "-task blastn -word_size 6 -

evalue 0.01 -strand plus" and an E-value cutoff of 10^{-5}. Synteny blocks around the lincRNA with conserved neighbors were defined as the maximal *n* for which *n* of the *n*+5 genes closest to the lincRNA were orthologs of the *n+5* genes closest to the corresponding mammalian lincRNAs. Zebrafish lincRNAs and controls were compared in this way to the human and the mouse genomes, and the human and mouse block sizes were averaged and reported with the geometric mean of the p-values.

In situ **Hybridizations**

In situ hybridization in whole-mount embryos was carried out as described (Thisse and Thisse, 2008). Digoxygenin-labeled anti-sense riboprobes (DIG Labeling Kit, Roche) to zebrafish lincRNAs were generated from DNA fragments (typically at least 750 bp) that had been amplified from a mixed-stage cDNA library (containing cDNA from 24-hpf, 72-hpf and 10-day stages, primers listed in Table S7).

Morpholino Injections and Rescue Experiments

MOs were designed based on Gene Tools LLC recommendations and computationally evaluated for specificity using BLASTN against zebrafish RefSeq mRNAs (requiring no more than 14 consecutive base pairs). Commercially synthesized MOs (Gene Tools, LLC) were dissolved in water and injected into one-cell embryos, using 1.6–5.0 ng per embyro (amounts and sequences listed in Table S6). For rescue experiments, lincRNAs were *in vitro* transcribed (mMessage mMachine kit, Ambion), purified (RNeasy kit, Qiagen) and precipitated (0.3M NaOAc pH 5.2 and 2.5 volumes ethanol). Mouse and human cyrano and megamind RNAs were *in vitro* polyadenylated [Poly(A) tailing kit, Ambion]. (*In vitro* polyadenylation was not required for zebrafish wild-type and mutant lincRNAs because they were transcribed from a vector that added an SV40 polyadenylation signal, which promotes polyadenylation in the embryo.) The following amounts of RNA were injected, in a volume of 1 nL, into one-cell embryos: 100 pg of zebrafish cyrano; 150 pg of mouse or human cyrano; 150 pg of zebrafish megamind; 200 pg of mouse or human megamind.

RNA Blots

3 μg of total RNA isolated from 48 hpf fish was treated with glyoxal using NorthernMax-Gly Sample Loading Dye (Ambion), heated to 60°C for 30 min and loaded on a 1% agarose gel prepared using NorthernMax-Gly Gel Running Buffer (Ambion) according to the manufacturer's instructions. The RNA was blotted onto a Nytran membrane in 20X SSC (175.3 g NaCl, 88.2 g sodium citrate in 1.0 l water adjusted to pH 7) using the TurboBlotter System (Whatman). After UV crosslinking RNA to the membrane, glyoxal treatment was reversed by incubating the membrane in 10 mM Tris-HCl pH 8 for 20 minutes at room temperature. The membrane was incubated in 12 ml QuickHyb solution (Stratagene) for 30 minutes at 65°C, and then radio-labeled RNA probe was added. Body-labeled antisense riboprobe for cyrano was *in vitro* transcribed (MaxiScript kit, Ambion) from a PCR product (primers listed in Table S7). After an overnight hybridization at 65°C, the membrane was washed twice in 2X SSC, 0.1% SDS for 5 minutes and once in 0.2X SSC, 0.1% SDS for 30 minutes. The membrane was exposed to phosphorimager plates and analyzed using the Bass Phosphorimager system (FujiFilm).

qRT-PCR Analysis

Total RNA from MO-injected embryos was isolated using TRI Reagent (Ambion). For cyrano MO injections, 100 ng of total RNA was used in reverse transcription reactions using oligo-dT primers (IDT) and SuperScript III Reverse Transcriptase (Invitrogen). For *megamind*, polyadenylated RNA was selected using Dynabeads (Invitrogen) and 100 ng of RNA was reversed transcribed with equal molar ratios of gene-specific primers for megamind and β-actin.

Gene-specific and β-actin primers were used in the real-time PCR using a 7900HT Fast Real-Time PCR system (Applied Biosystems). ΔΔCt values were calculated for each gene, and relative transcript levels were derived from these values.

DNA Constructs

Full-length zebrafish *cyrano* and *megamind* cDNAs were amplified from the mixed-stage cDNA library (primers listed in Table S7) and cloned into the pCS2+ vector (Rupp and Weintraub, 1991). Mouse cDNAs were obtained from the RIKEN cDNA clone collection through DNAFORM (clone IDs 2810011L19 and M5C1004K09). Human cDNAs were obtained from imaGenes (clone IDs DKFZp686E0352Q and IMAGp998O103712Q). Hybrid 1 construct (cyrano conserved site in the megamind backbone) was obtained by digesting a pGEM-T Easy vector with the full-length zebrafish megamind with BglII, which removed a region containing the conserved site of megamind. The conserved site of cyrano was amplified by PCR (primers listed in Table S7) and cloned into the megamind backbone using BglII restriction sites. Hybrid 2 construct (megamind conserved site in the cyrano backbone) was obtained by digesting a pGEM-T Easy vector with the full-length zebrafish cyrano with StuI, which removed a region containing the conserved site of cyrano. The conserved site of megamind was amplified by PCR (primers listed in Table S7) and cloned into the cyrano backbone using StuI restriction sites.

SUPPLEMENTAL REFERENCES

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

Guenther, M.G., Lawton, L.N., Rozovskaia, T., Frampton, G.M., Levine, S.S., Volkert, T.L., Croce, C.M., Nakamura, T., Canaani, E., and Young, R.A. (2008). Aberrant chromatin at genes encoding stem cell regulators in human mixed-lineage leukemia. Genes Dev *22*, 3403-3408.

Guo, H., Ingolia, N.T., Weissman, J.S., and Bartel, D.P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature *466*, 835-840.

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.

Jan, C.H., Friedman, R.C., Ruby, J.G., and Bartel, D.P. (2011). Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. Nature *469*, 97-101.

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.

Kong, L., Zhang, Y., Ye, Z.Q., Liu, X.Q., Zhao, S.Q., Wei, L., and Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res *35*, W345-349.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Rupp, R.A., and Weintraub, H. (1991). Ubiquitous MyoD transcription at the midblastula transition precedes induction-dependent MyoD expression in presumptive mesoderm of X. laevis. Cell *65*, 927-937.

Slater, G.S., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics *6*, 31.

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

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

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

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470-476.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., *et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.