Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells


Many long noncoding RNA (lncRNA) species have been identified in mammalian cells, but the genomic origin and regulation of these molecules in individual cell types is poorly understood. We have generated catalogs of lncRNA species expressed in human and murine embryonic stem cells and mapped their genomic origin. A surprisingly large fraction of these transcripts (>60%) originate from divergent transcription at promoters of active protein-coding genes. The divergently transcribed lncRNA/mRNA gene pairs exhibit coordinated changes in transcription when embryonic stem cells are differentiated into endoderm. Our results reveal that transcription of most lncRNA genes is coordinated with transcription of protein-coding genes.

Results

We describe here catalogs of human and murine ESC lncRNAs that were maintained some redundant species, and these were removed or, if limited coding potential.

The long noncoding RNAs (lncRNAs) are of particular interest because they are known to contribute to gene silencing (3), dosage compensation (2), and epigenetic silencing mechanisms (4), such as X inactivation (4), imprinting (5, 6), and development (7–9), but there is limited understanding of the genomic origin, regulation, and function of lncRNA molecules in individual cell types.

Embryonic stem cells (ESCs) are widely used as a model system to study transcriptional control of cell state during early development (10–13), yet there is no catalog of lncRNAs in human (h) ESCs, and it is not clear how lncRNAs are regulated in these cells. Catalogs of lncRNAs have been recently described in various murine (14, 15) and human cell types (16–19), but the majority were limited to spliced lncRNA species (14–16, 18) and those distant from protein-coding genes (14–17). Because lncRNAs tend to be cell-type–specific (16, 18), these catalogs likely contain only a very small fraction of lncRNAs expressed in hESCs.

We describe here catalogs of human and murine ESC lncRNAs and the genomic regions from which these RNA species arise. We find that the majority of these lncRNAs originate from divergent transcription of lncRNA/mRNA gene pairs and that many such gene pairs are coordinately regulated when ESCs differentiate.

In lncRNAs Expressed in Human ESCs. We compiled a catalog of lncRNA species expressed in hESCs as summarized in Fig. 1A. A total of 21,681 RNA species that have filter transcripts that had a positive protein-coding potential based on the quality of the ORF and the results of BLAST X searches against proteins curated by RefSeq. From this filter, we obtained a set of 21,681 RNA species that have limited coding potential.

Small RNA species were filtered from the pool by requiring that transcripts be at least 100 nt long. The remaining pool contained some redundant species, and these were removed or, if protein-coding RNAs, to be at least 100 nucleotides (nt) long, and to be derived from a unique genomic location (Fig. 1A).

The lncRNAs were required to have a 5′ end that originates from a genomic site for which there is corroborating evidence of active transcription initiation to enhance confidence that these were complete, full-length species. Sequence assembly produces a 5′ end for each putative transcript that may originate from a genuine transcription start site (TSS), from degradation of a longer transcript, or from incomplete sequence assembly. We used the presence of nucleosomes with histone H3 trimethylated at lysine 4 (H3K4me3), as defined by chromatin immunoprecipitation (ChIP) sequencing (ChIP-seq) analysis in hESCs, to provide corroborating evidence of active transcription initiation. Previous studies have shown that the presence of this mark at protein-coding and non-coding RNA loci provides reliable evidence of active transcription initiation at these loci (12, 15, 18, 21–27). Thus, RNA species were removed from the pool if they did not have 5′ ends located at such sites. Furthermore, to enhance confidence that the RNA species are expressed in hESCs, RNA species were removed from the pool if sequence coverage in the hESC RNA sequencing (RNA-seq) data fell below a threshold of 0.07 reads per kilobase of exonic length per million. This threshold was chosen to minimize false negatives (Fig. S1 and SI Materials and Methods) and ensure that using H3K4me3 to define lncRNA 5′ ends did not bias the genomic distribution of selected RNA species (Fig. S1 and SI Materials and Methods). Thus, we obtained a set of 72,406 RNA species that contained a defined, active TSS in hESCs with high confidence of expression.

To distinguish ncRNA species from transcripts that are likely to encode proteins, the coding potential of each of the transcripts was evaluated by using the Coding Potential Calculator (CPC) (28). We filtered transcripts that had a positive protein-coding potential based on the quality of the ORF and the results of BLAST X searches against proteins curated by RefSeq. From this filter, we obtained a set of 21,681 RNA species that have limited coding potential.
The IncRNAs in the hESC catalog range in size from 105 to 10,687,089 nt and have a median size of 1,831 nt (Fig. S2A). The IncRNAs produced as a result of divergent transcription from promoters and that such transcription can generate low-abundance antisense ncRNAs (30–32). To further study the possibility that ESC IncRNAs may be derived from divergent transcription of active protein-coding genes, we investigated whether transcriptionally engaged RNA Pol II occurs in both orientations at active IncRNA/mRNA gene pairs by mapping global nuclear run-on (GRO-seq) data. GRO-seq data provide the positions, relative levels, and orientation of transcriptionally engaged RNA Pol II molecules genome-wide. These data showed that transcriptionally engaged RNA Pol II molecules occurred immediately downstream of the TSS of protein-coding genes, as expected based on prior evidence for RNA Pol II pausing in this region (13) (Fig. 2A). These data also show that RNA Pol II molecules are engaged in active transcription antisense to the protein-coding genes (Fig. 2B).

Further analysis of GRO-seq data revealed that RNA Pol II is engaged in transcription initiation at 66% of protein-coding genes in hESCs (Fig. S2F and Datasets S3 and S4), which is consistent with previous estimates based on other criteria (21). Most (85%) of these active protein-coding genes showed evidence of divergent transcription (Fig. S2G). However, only ~10% of these divergently transcribed genes produced IncRNA species that accumulated to substantial levels. The divergently transcribed IncRNA/mRNA gene pairs that produce detectable IncRNA species tend to be transcribed at higher levels (based on GRO-seq data) than those pairs whose IncRNAs do not accumulate (Fig. S2H-J). Although the number of transcriptionally engaged RNA Pol II molecules was similar for mRNA and IncRNA in divergently transcribed IncRNA/mRNA pairs (Fig. 2A), the steady-state levels of the IncRNAs were much lower than those for mRNAs from the same gene pairs (Fig. 2C). Thus, although divergent transcription occurs at most active protein-coding genes, only a small fraction of IncRNAs produced by divergent transcription are sufficiently stable to be detected under the conditions studied here. The evidence that IncRNAs are less stable than mRNAs is consistent with previous reports that IncRNA transcripts are subjected to xesome-mediated degradation (33, 34).

Association of IncRNA Genes with mRNA Genes in hESCs. Inspection of the genomic positions of IncRNA loci revealed that the majority (89%) are associated with the promoters, enhancers, and bodies of protein-coding genes (Fig. 1B). Most IncRNAs were found to originate within a 2-kb region surrounding the TSS of protein-coding genes (65%), and others originate from antisense transcription of protein-coding genes (5%), enhancers (19%), and other more distant (>2 kb) sites from protein-coding genes (11%) (Fig. 1B).

Fig. 1. Most IncRNAs are associated with active protein-coding genes in hESCs. (A) Schematic diagram of pipeline for identification of IncRNAs in hESCs. An “initial RNA pool” was compiled from transcripts assembled de novo from RNA-seq reads (this study, SI Materials and Methods) and published data (20). Four criteria required for the selection of expressed transcripts from this pool are indicated in red. Transcripts were required to be expressed from a high-confidence start site (occupied by H3K4me3), to be noncoding [lacking features of protein-coding RNAs as defined by the CPC (28)], to be long (>100 nt), and to be nonrepetitive. (B, C) Summary of various types and numbers of IncRNA loci in hESCs, which are listed in Dataset S1. Diagrams at right depict IncRNA loci as red lines, protein-coding genes as blue lines, and an enhancer as an open box. An arrow indicates direction of transcription initiation. Enhancer-associated IncRNAs overlap or originate at genomic regions enriched in nucleosomes with histone 3 acetylated at lysine 27 (H3K27Ac). Enriched regions for H3K27Ac are indicated in red. Transcripts were required to be expressed from a high-confidence start site. Acceptable transcripts were filtered according to length (C) and genomic position (B). (D) Distribution of TSS of lncRNAs relative to the TSS of protein-coding genes, as expected based on prior evidence for RNA Pol II pausing in this region (13) (Fig. 2A). These data also show that RNA Pol II molecules are engaged in active transcription antisense to the protein-coding genes (Fig. 2B). Inspection of the data at individual genes shows that this divergent transcription is associated with IncRNA/mRNA gene pairs (Fig. 2B), consistent with the model that these IncRNAs are produced as a result of divergent transcription from promoters of active protein-coding genes.

overlapping, compiled into loci. Finally, RNA loci were removed if transcripts comprising them mapped to more than one genomic location. We thus compiled a catalog of IncRNA species expressed in hESCs that originated from 3,548 nonredundant loci (Dataset S1).

The IncRNAs in the hESC catalog range in size from 105 to 687,089 nt and have a median size of 1,831 nt (Fig. S2A). These IncRNAs accumulate to levels that are, on average, ~10% that of mRNAs in hESCs (Fig. S2B). The size range and average abundance of IncRNAs in this catalog are similar to those described previously for IncRNAs (16, 29). Approximately half of the IncRNA loci contain spliced transcripts (Dataset S1). The vast majority (73%) of the IncRNA loci in the hESC catalog have not been previously identified (Fig. S2C).
that are divergent from lncRNA locus 3182 generates two alternatively spliced lncRNA transcripts for polyadenylated RNA in the vicinity of H3K4me3-modified DNA are shown separately as indicated on graph. (A) Example of lncRNA locus whose 5′ end occurs within 2 kb of the TSS of a protein-coding gene (promoter-associated lncRNA). Gene tracks represent ChIP-seq data for H3K4me3-modified nucleosomes (48) together with GRO-seq reads and reads for polyadenylated RNA in the vicinity of STAM. Transcription at lncRNA locus 3182 generates two alternatively spliced lncRNA transcripts that are divergent from STAM. The x axis represents the linear sequence of genomic DNA, and the y axis represents the total number of ChIP-seq, GRO-seq, and RNA-seq mapped reads. RNA-seq reads that map to Watson (blue) and Crick (red) strands of genomic DNA are shown separately. GRO-seq reads that map to Watson (purple) and Crick (magenta) strands of genomic DNA are shown separately. The scale is indicated in the upper right. (B) Alignment of RNA-seq reads for the 2,318 protein-coding genes that contain lncRNAs within 2 kb of their TSS. Reads are aligned in 250-bp bins. The x axis indicates the distance from the TSS in kb. The y axis indicates the average number of uniquely mapped RNA-seq counts normalized to reads per genomic bin per million uniquely mapped reads. Reads that map to Watson (black) and Crick (red) strands of genomic DNA are shown separately as indicated on the graph.

Assocation of lncRNA Genes with mRNA Genes in Murine ESCs. Previous studies of murine (m) ESC lncRNAs focused primarily on 226 transcripts that are located some distance from protein-coding genes and are spliced (14). We compiled a catalog of 1,664 lncRNA loci by combining unpublished and published RNA-seq data from mESCs (Fig. S3A and Dataset S1) and by filtering the results with criteria that enhance confidence in this lncRNA population (Fig. S3A and SI Materials and Methods). Inspection of the genomic positions of the murine lncRNA loci revealed that the majority of lncRNAs are transcribed from sites near or within actively transcribed protein-coding genes (Fig. 3A), as in hESCs. Of the 1,664 lncRNA loci in the murine catalog, 62% originate within a 2-kb region surrounding the TSS, 9% originate from antisense transcription of protein-coding genes, and 27% are derived from transcription of enhancer elements (Fig. 3A). As with hESCs, the vast majority of lncRNAs have 5′ ends that occur within 2 kb of the TSS of protein-coding genes, and almost all of these (93%) are transcribed antisense to the protein-coding gene (Fig. 3B and Fig. S3B). These promoter-associated lncRNAs range in size from 204 to 424,645 nt, with a median size of 2,704 nt (Fig. S3C), and have three features of mRNAs—they have a 7-methylguanosine cap and a poly(A) tail (Fig. S3D and Dataset S1). Analysis of published GRO-seq data from mESCs (12) showed similar levels of RNA Pol II engaged in transcription of both the IncRNA and the mRNA at IncRNA/mRNA gene pairs (Fig. 3C). As with hESCs, the IncRNAs from these gene pairs accumulated to lower levels (average 10-fold) than the mRNA species transcribed from the adjacent protein-coding gene (Fig. 3D and Fig. S3E). Thus, as with hESCs, the majority of IncRNAs in mESCs originate from divergent transcription of IncRNA/mRNA gene pairs.

Coordinated Transcription of IncRNA/mRNA Pairs During Differentiation. The observation that >60% of lncRNA genes in ESCs are transcribed divergently from active protein-coding genes suggests that...
transcription of lncRNA/mRNA pairs may be coordinately regulated. To investigate this possibility, we stimulated differentiation of the hESCs into endoderm through activin treatment and studied transcriptional events using RNA-seq and GRO-seq analysis. A catalog of endodermal lncRNA loci was produced by using the criteria for generation of hESC and mESC catalogs; this catalog

![Diagram](image)

**Fig. 4.** Divergent lncRNA/mRNA pairs exhibit coordinated changes in transcription as ESCs differentiate into endoderm. (A) Summary of the genomic distribution of IncRNA loci 48 h after induction of endodermal differentiation in hESCs. Diagrams at right depict IncRNA loci as red lines, protein-coding genes as blue lines, and an enhancer as an open box. An arrow indicates direction of transcription initiation. Enhancer-associated IncRNAs overlap or originate at genomic regions enriched in H3K27Ac. Enriched regions for H3K27Ac are available in Dataset S2. (B) Alignment of GRO-seq reads 48 h after induction of endodermal differentiation in hESCs for the 2,680 protein-coding genes that contain IncRNAs within 2 kb of their TSS. The x axis indicates the distance from the TSS in kilobases, and the y axis indicates the average number of uniquely mapped GRO-seq reads normalized to reads per genomic bin per million uniquely mapped reads. Reads that map to Watson (black) and Crick (red) strands of genomic DNA are shown separately as indicated on the graph. (C) Alignment of RNA-seq reads 48 h after induction of endodermal differentiation in hESCs for the 2,680 protein-coding genes that contain IncRNAs within 2 kb of their TSS. The x axis indicates the distance from the TSS in kilobases, and the y axis indicates the average number of uniquely mapped RNA-seq counts normalized to reads per genomic bin per million uniquely mapped reads. Reads that map to Watson (black) and Crick (red) strands of genomic DNA are shown separately as indicated on graph. (D) Example of IncRNA/mRNA pairs exhibiting coordinated transcriptional induction 48 h after hESCs were differentiated toward the endoderm. Gene tracks represent GRO-seq data in the vicinity of GATA6. Divergent transcription generates antisense IncRNA locus 5689 upstream of GATA6. The x axis represents the linear sequence of genomic DNA, and the y axis represents the number of GRO-seq reads normalized to total number of mapped reads. GRO-seq reads that map to Watson (black) and Crick (red) strands of genomic DNA are shown separately. GRO-seq reads mapped to the Crick (red) strand of genomic DNA are shown flipped/rotated beneath. The scale is indicated in kilobases (kb) above the track. (E) Example of IncRNA/mRNA pairs exhibiting coordinated transcriptional induction 48 h after hESCs were differentiated toward the endoderm. Gene tracks represent GRO-seq data in the vicinity of LHX5. Divergent transcription generates antisense IncRNA locus 4010 upstream of LHX5. The x axis represents the linear sequence of genomic DNA, and the y axis represents the number of GRO-seq reads normalized to total number of mapped reads. (F) Coordinate transcriptional induction of IncRNA/mRNA gene pairs. A total of 683 IncRNA/mRNA pairs were selected, in which the numbers of GRO-seq reads of mRNA increased at least 1.25-fold after 48 h of endodermal differentiation. The average number of uniquely mapped GRO-seq reads from the strands encoding the mRNA transcripts is shown in black (Upper). The average number of uniquely mapped GRO-seq reads from the strands encoding the IncRNA transcripts is shown in red (Lower). Solid lines represent transcription in hESCs, and dashed lines represent transcription 48 h after induction of differentiation toward the endoderm. The x axis indicates the linear distance in kilobases, and the y axis indicates the average reads per genomic bin per million uniquely mapped reads.

Sigova et al.

PNAS | February 19, 2013 | vol. 110 | no. 8 | 2879
contains lncRNA species that originate from 3,986 nonredundant loci (Dataset S1 and Fig. 4). Approximately half of the lncRNAs in the catalog are expressed in hESCs (Fig. S2E). The endodermal lncRNAs—the vast majority (75%) of which have not been described in previous studies (Fig. S2C)—have characteristics similar to the lncRNAs expressed in human and murine ESCs. They range in size from 194 to 6,877,089 nt, have a median size of 2,068 nt, and accumulate to levels that are ~10% of that of mRNAs (Fig. S2B and D). Approximately half of the lncRNAs are spliced (Dataset S1).

Most endodermal lncRNA loci (67%) originate within 2 kb of promoters of protein-coding genes (Fig. 4A). Other lncRNAs are derived from enhancers (23%), antisense transcription of protein-coding genes (5%), and other more distant sites (5%) (Fig. 4A). Similar levels of RNA Pol II are engaged in transcription of lncRNA and mRNA at lncRNA/mRNA gene pairs (Fig. 4B), but lncRNAs typically accumulated to lower levels than mRNAs (Fig. 4C). Thus, as with ESCs, the majority of endodermal lncRNA species are produced from divergent transcription of protein-coding genes.

To determine whether transcription of divergently transcribed lncRNA/mRNA pairs changes coordinately during endodermal differentiation, we compared the density of RNA Pol II molecules at these loci in hESCs and 48 h after differentiation toward endoderm. Strikingly, transcriptional induction of mRNA genes during differentiation was accompanied with induction of associated lncRNAs ($\chi^2$, $P < 4.8 \times 10^{-22}$ (Fig. 4D–F and SI Materials and Methods), whereas repression of mRNA genes was accompanied by repression of associated lncRNAs ($\chi^2$, $P < 2.4 \times 10^{-94}$ (Fig. S4A and SI Materials and Methods). Similar GRO-seq results were obtained only 1 h after activin treatment, indicating that the coordinate changes in transcription of lncRNA/mRNA gene pairs occurred rapidly (Fig. S4 B and C). Thus, changes in transcription of lncRNAs in lncRNA/mRNA gene pairs tend to be coordinated with changes in transcription of neighboring protein-coding genes during differentiation of hESCs into endoderm. These results suggest that coordinated regulation of lncRNA/mRNA gene pairs may be a general feature of differentiation.

Discussion

We have found that the majority of lncRNAs in human and murine ESCs are produced from divergently transcribed protein-coding genes and that the divergent lncRNA/mRNA pairs exhibit coordinated changes in transcription as ESCs differentiate into endoderm. The evidence for these conclusions comes from the assembly of lncRNA catalogs in hESCs and mESCs, and GRO-seq data for hESCs that are described here. These datasets should provide a rich source of information for additional studies into the functions of these lncRNA species and the control of their expression.

Previous studies have described mammalian lncRNAs but have not noted the striking extent to which lncRNA/mRNA gene pairs contribute to the population of lncRNAs that are produced in individual cell types (1, 14, 16, 18, 35–38). Our findings have been made possible by obtaining fuller coverage of lncRNAs in ESCs through use of directional RNA sequencing and by inclusion of both spliced and nonspliced lncRNA species, which, to our knowledge, has not been used in previous studies to define lncRNAs in a single cell type. There is some prior evidence that lncRNAs can originate from regions upstream of coding genes based on studies of a few genes (35). The evidence described here reveals that the majority of lncRNAs are derived from divergent transcription of active protein-coding genes.

The GRO-seq data for hESCs shows that divergent transcription occurs at the vast majority of protein-coding genes where transcription initiation takes place, but the RNA-seq data indicates that only a small fraction of the divergent transcription events produce substantial levels of steady-state lncRNAs. In ESCs, divergent transcripts from the promoters of four protein-coding genes have previously been analyzed for their sites of initiation, presence of 7-methylguanosine cap, and length and level of RNAs per cell (33). The steady-state level of all four divergent transcripts ranged from two to four RNAs per cell, and these levels were found to be controlled, at least in part, by exosome degradation. Thus, divergent transcription most likely generates many lncRNA species that are then degraded by the exosome pathway.

The transcriptional control of ESCs has been the subject of intense study, yet most of this research has been focused on protein-coding genes. An implication of the finding that most lncRNAs are transcribed divergently from active protein-coding genes is that transcription of lncRNAs may often be coregulated with the adjacent protein-coding gene. The coordinate regulation of lncRNA/mRNA gene pairs during differentiation described here supports this model. It is also possible that one or both of the divergent transcripts regulate one another; previous studies have reported that antisense lncRNA can regulate expression of neighboring mRNA genes (5, 39, 40). Future studies of lncRNA/mRNA gene pairs and the lncRNAs described here should provide new insights into the contributions of lncRNAs to the control of cell state and the process of differentiation.

Materials and Methods

Cell Culture Conditions. H1 (WA01) hESCs were grown on matrigel (BD Sciences) by using mTESR1 (Stem Cell Technologies) as described (41). hESCs were differentiated toward endoderm by resting cells in RPMI with B27 supplement for 24 h followed by treatment with activin (R&D Systems). Time points were measured from the time of activin treatment. V6.5 mESCs were grown on irradiated murine embryonic fibroblasts unless otherwise stated. mESCs were grown under conditions as described (42).

RNA-seq and Assembly of Transcripts. Polyadenylated RNA-seq libraries were prepared for directional sequencing according to a modified version of the Illumina paired-end RNA-seq protocol (SI Materials and Methods) and sequenced on Illumina HiSeq 2000 (Table S1). Sequenced reads were aligned to the human (hg18) or mouse (mm9) genomes by using Tophat (Version V1.2.0) (43) using default settings along with “microexon-search” and “coverage-search” parameters. Transcripts were assembled by using Scripture (Version 1.0) (29) and Cufflinks (44). Reads are available at the Gene Expression Omnibus (GEO) database (accession nos. GSE36799 and GSE41009).

ChiP-seq. ChiP of nucleosomes with H3K4me3 and H3K27Ac was performed as described (21) by using H3K4me3 (Millipore; 07-473), H3K27Ac (Abcam; AB4729), and IgG (Millipore; 12-370) antibodies (Table S2). Illumina sequencing was performed on library preparation and sequencing, and quality control were followed as described in SI Materials and Methods. Libraries were sequenced by using Illumina GAII or HiSeq2000. ChiP-seq data for hESC H3K4me3 and H3K27Ac and for mESC H3K27Ac were obtained from GEO (accession nos. GSE33748, GSM466732, and GSE24164, respectively). Reads were aligned to NCBI Build 36 using Bowtie software. Enrichment was determined as described (25). A summary of the enriched genomic regions (P < 10−7) for all H3K4me3 and H3K27Ac datasets is provided (Dataset S2). Data sets generated for this study are available from the GEO database (accession nos. GSM896920, GSM896921, and GSE41009).

GRO-seq. GRO-seq was performed as described (32) by using 5 × 10⁶ cells from biological replicates for each time point. Libraries were sequenced on Illumina Hi-Seq2000 (Table S1). See SI Materials and Methods for further details. GRO-seq reads are available from GEO (accession no. GSE41009).

Pipeline for Generation of lncRNA Catalogs. To derive the initial pool of RNA for hESCs, we integrated transcripts assembled from the RNA-seq reads in this study with the FLI (20) database. For mESCs, we integrated transcripts assembled from the RNA-seq reads in this study with the set of previously assembled transcripts (29) as well as annotated transcripts from FANTOM3 (45) and NIA Gene Index (46, 47). Transcripts from all sources were further processed to filter RefSeq-annotated protein-coding transcripts, pseudogenes, microRNA, tRNA, rRNA, and small nucleolar RNA. Transcripts in the initial pool of RNA are available in Dataset S1. Next, filters were applied to select for transcripts with correct 5′ ends that met minimal read coverage threshold, remove transcripts

2880 | www.pnas.org/cgi/doi/10.1073/pnas.1221904110

Sigova et al.
with positive coding potential, select for long transcripts, and remove repetitive elements (SI Materials and Methods). We combined partially or fully overlapping transcripts into lncRNA loci (Dataset S1) to reduce redundancy that would result from multiple alternatively spliced isoforms originating from the same genomic location or annotation of the same transcripts in multiple databases.

**Expression Abundance and Transcription State of Genes.** The level of expression measured by RNA-seq and GRO-seq was calculated as described in SI Materials and Methods.

### Notes

Human embryonic stem cells. μV6.5 murine embryonic stem cells – Human – 1 of 1
All sequenced reads from hESCs and base was added, and Illumina PE adaptors were ligated onto end) to avoid read-length
fl
2
ends to remove any linker sequence contami-
Coding and noncoding transcripts
hESCs were grown to 80% con-
EST sequences generated previously from
Human ESCs.

Supporting Information

Sigova et al. 10.1073/pnas.1221904110

SI Materials and Methods

Growth Conditions for Cells. Human embryonic stem cells. Human embryonic stem cells (hESCs) (WA01) were maintained in feeder free conditions as described (1). Cells were grown in a monolayer on tissue culture plates coated with matrigel at the dilution rec-
ommended by the manufacturer (BD; 354277). Cells were main-
tained in mTESR1 medium (Stemcell Technologies; 05850) with 100 U/mL penicillin and 100 mg/mL streptomycin. For passaging, mTESR1 was removed, and cells were treated with 2.0 mg/mL

Endodermal differentiation. hESCs were grown to 80% confluence and then cultured for 24 h in RPMI plus B27 supplement (Life Technologies). After 24 h, cells were treated with activin (50 ng/ nL; R&D; 338-AC) in RPMI plus B27. Cells were harvested for analysis at times 0 h (24 h in RPMI plus B27), 1 h (24 h in RPMI plus B27 followed by 1-h treatment with activin in RPMI plus B27), and 48 h (24 h in RPMI plus B27 followed by 48-h treat-
ment with activin in RPMI plus B27).

Murine embryonic stem cells. V6.5 murine embryonic stem cells (mESCs) were grown on irradiated murine embryonic fibroblasts (MEFs). Cells were grown under standard mESC conditions as described (2). Briefly, cells were grown on 0.2% gelatinized (Sigma; G1890) tissue culture plates in ESC medium: DMEMKO (In-
vitrogen; 10829-018) supplemented with 15% (vol/vol) FBS (Atlas Biologicals; F-0505-A), 1000 U/mL LIF (ESGRO; ESG1106), 100 μM nonessential amino acids (Invitrogen; 11140-050), 2 mM l-glutamine (Invitrogen; 25300-081), 100 U/mL penicillin, 100 μg/ mL streptomycin (Invitrogen; 15140-122), and 8 nL/mL 2-mer-
captoethanol (Sigma; M7522). Before total RNA purifica-
tions, mESCs were passaged twice in absence of MEFs.

RNA Isolation and Sample Preparation for Directional RNA-Sequencing. Total RNA was purified by using mirVana mirRNA isolation kit (Life Technologies) following the manufacturer’s instructions. Total RNA was treated with DNA-free DNase I (Ambion) and analyzed on Agilent 2100 Bioanalyzer for integrity. RNA with the RNA Integrity Number (RIN) > 9 was used for library

generation. Polyadenylated RNA was purified from the total RNA by two rounds of selection using Dynabeads mRNA Purification Kit for mRNA Purification from Total RNA (Life Technologies) following the manufacturer’s instructions. Total RNA was depleted of ribosomal content by using the Ribo-Zero rRNA Removal Kit (Epigence). Both kinds of samples were sequenced by using a standard Illumina mRNA-Seq protocol with the fol-
lowing modifications. Briefly, polyadenylated and total RNA de-
pleted of rRNA was fragmented with divalent cations under elevated temperature. First-strand cDNA synthesis was performed with random hexamers and SuperScript III reverse transcriptase.

Second-strand cDNA synthesis was performed by using RNase H and DNA Polymerase I. In the second-strand synthesis re-
tion, dTTP was replaced with dUTP. Following cDNA syn-
thesis, the double-stranded products were end repaired, a single “A” base was added, and Illumina PE adaptors were ligated onto the cDNA products. The ligation reactions with an average size of 300 bp were purified using agarose gel electrophoresis. Fol-
lowing gel purification, the strand of cDNA containing dUTP was selectively destroyed during incubation of purified double-stranded DNA with HK-UNG (Epigence). The adapter ligated single-
stranded cDNA was then amplified with 15 cycles of PCR, and PCR products were purified by using gel electrophoresis. RNA-
sequencing (RNA-seq) libraries were sequenced on Illumina HiSeq 2000.

RNA-seq Read Alignment. All sequenced reads from hESCs and hESCs differentiated toward endoderm were aligned to the human genome (NCBI36, hg18). Sequenced reads from mESCs were aligned to the mouse genome (NCBI 37, mm9). All alignments were performed by using TopHat (Version V1.2.0) (3). TopHat first uses Bowtie (4) to align reads that are mapped to the genome without gaps. It then determines the possible location of gaps in the alignment based on canonical and noncanonical splice sites flanking the aligned reads. Finally, it uses gapped alignments to align the reads that were not aligned by Bowtie in the first step. RNA-seq read statistics are in Table S1.

GRO-seq Read Alignment. All Gro-Seq datasets were 40-bp reads, with the exception of H1 replicate 1 and 2. These two datasets were truncated at 40 bp (starting from 5’ end) to avoid read-length artifacts when comparing datasets. For each dataset, reads were trimmed at the 5’ ends to remove any linker sequence contami-
nation. The linker sequence was ‘TTCGTAITGCCGTCTCTTG’.

A read was trimmed if it ended in part of the linker sequence (requiring at least the first 6 bp). Reads that were <24 bp long after linker trimming were removed. Trimmed reads for each dataset were then grouped by read-length and aligned to the human genome (NCBI36, hg18), using Bowtie with the parameters -e 70 -k 1 -m 10 -n 2 –best –strata, with the additional parameter –l set to the length of the read group being aligned. After alignment of all read-length groups, alignments were combined to produce a complete alignment for each dataset. GRO-seq read statistics are in Table S1.

Expression Abundance and Transcription State of Genes. The level of expression was measured by RNA-seq and GRO-seq as described below. Abundance of transcripts in hESCs and after 48 h of endodermal differentiation was estimated based on the number of RNA-seq reads obtained from sequencing of poly(A)-selected RNA. Abundance of transcripts in mESCs was estimated based on number of RNA-seq reads obtained from sequencing of poly (A)-selected and ribo-depleted RNA. First, reads that uniquely mapped to exons of each transcript were counted toward the raw expression value for that transcript. Then, the raw expression values were further normalized by the combined length in kilobases of all exons in the transcript, and the total number of all reads was uniquely mapped to the human (hg18) or mouse (mm9) genome (RPKM) as described (5).

Strand-specific enriched regions (GRO-Seq Analysis for Di-
vergent Genes and Dataset S3) were determined for each Gro-seq dataset. A gene was considered initiated if there was an enriched region within ± 1 kb of the gene TSS and that enriched region was on the same strand as the gene. A gene was classified as having divergent transcription if there was a Gro-Seq enriched region ± 2 kb of the TSS and that enriched region was on the opposite strand as the gene. The genes considered initiated and/or as having divergent transcription in each condition are listed in Dataset S4.

Initial RNA Pool. Human ESCs. Coding and noncoding transcripts assembled ab initio by Scripture (6) and Cufflinks (3) were used in combination with human EST data (7) containing >1 million 5’ EST sequences generated previously from >60 different human tissues. Transcripts from these sources were pooled, and coding
transcripts were filtered in a two-step process. First, we filtered transcripts that were on the same strand and overlapped annotated protein-coding genes (Dataset S1). Second, because all transcripts filtered in the previous step were considered to be protein coding, we filtered transcripts that were on the same strand and overlapped transcripts removed in the first step. The remaining transcripts of the initial pool of RNA are available in Dataset S1.

**Murine ES.** The initial RNA pool was compiled from the following transcripts:

a. Coding and noncoding transcripts assembled ab initio by Scripture from mESC RNA-seq reads in this study;

b. Coding and noncoding transcripts assembled ab initio by Scripture from mESC RNA-seq reads in published study (6);

c. Noncoding transcripts from FANTOM3 database (8);

d. Coding and noncoding transcripts from NIA Gene Index database (9, 10).

The FANTOM3 database contains sequences of full-length cDNA clones from various mouse tissues, whereas NIA Gene Index database contains EST sequences of cDNA clones from oocytes, blastocysts, and embryonic and adult stem cells. Transcripts from all four sources were pooled, and coding transcripts, pseudogenes, and short RNA were filtered in a two-step process. First, we filtered transcripts that were on the same strand and overlapped annotated protein-coding genes, pseudogenes, microRNA, tRNA, rRNA, or small nuclear RNA (snRNA). Second, because all transcripts filtered in the previous step were considered to be protein coding, pseudogenes, microRNA, tRNA, rRNA, or snRNA, we filtered transcripts that were on the same strand and overlapped transcripts removed in the first step. The remaining transcripts of the initial pool of RNA are available in Dataset S1.

**Selecting Noncoding RNA with Evidence of Transcription Initiation at 5′ End.** To select for transcripts with correct 5′ ends, we screened the initial RNA pool for transcripts whose 5′ ends occurred within 1 kb of enriched regions occupied by H3K4me3 in hESCs, 48-h human endoderm, or mESCs. A summary of the enriched genomic regions (P < 10^-6) for H3K4me3 is provided (Dataset S2). Genomic coordinates for Dataset S2 are in NCBI36/hg18 or in NCBI36/mm8. Coordinates of mESC transcripts in the initial pool of RNA were converted from NCBI36/mm9 to NCBI37/mm9 for all following steps.

**Minimal Read Coverage Threshold (False Negative Rate Estimations).** We established a threshold that would allow us to distinguish expressed transcripts from regions in the genome exhibiting active TSS but no detectable expression. We eliminated transcripts with a maximal coverage <0.065 reads per kilobase of exonic length per million (RPKM) in hESCs, 0.070 RPKM in hESCs at 48 h of differentiation toward the endoderm, and 0.260 RPKM in mESCs. These coverage thresholds were set to minimize false-negative rates as described below. Because filtering of transcripts from the initial RNA pool was directly dependent on whether their 5′ ends were associated with H3K4me3, the false-negative rate was defined as a fraction of correctly annotated noncoding transcripts from RefSeq expressed >0 RPKM in hESCs, 48 h endoderm, or mESCs, which were not enriched for H3K4me3 within 1 kb of their TSS, of the total number of correctly annotated RefSeq transcripts expressed above 0 RPKM in these cells. The false-negative rate was dependent on expression levels of transcripts. To demonstrate this, we plotted the number of correctly annotated noncoding transcripts from RefSeq, whose 5′ ends occurred within 1 kb of enriched regions occupied by H3K4me3 (without H3K27me3 mark), relative to the expression levels (RPKM) of these transcripts in hESCs, 48 h endoderm, or mESCs. Because all RefSeq transcripts chosen for the experiment had correct 5′ ends, H3K4me3 binding at their TSS was directly dependent on their expression levels in hESCs, 48 h endoderm, or mESCs and inversely dependent on false-negative rates (Fig. S1A, C, and E). Thus, transcripts with low expression levels in mESCs had high false-negative rates, whereas highly abundant transcripts had very low false-negative rates.

The coverage threshold was then set to minimize false-negative rates. It corresponded to expression level in RPKM at which at least 50% of noncoding transcripts had H3K4me3 binding within 1 kb of their TSS. False rates increased dramatically as expression of transcripts decreased below the threshold, and they changed little as expression increased above the threshold (Fig. S1 A, C, and E).

We next wished to determine how position of noncoding transcripts from RefSeq relative to TSS of protein-coding genes affected false-negative rates. Transcripts, which were expressed above the threshold and located within 2 kb of TSS of protein-coding genes, exhibited H3K4me3 binding with slightly lower false rates than transcripts located >2 kb away from TSS of protein-coding genes (Fig. S1 B, D, and F).

**Filtering Transcripts with Positive Coding Potential.** Coding potential for transcripts in the initial pool of RNA whose 5′ ends occurred within 1 kb of enriched regions occupied by H3K4me3 and expressed above coverage threshold was estimated by using a support vector machines-based tool: Coding Potential Calculator (CPC) (11). CPC translates transcripts in three possible forward ORFs, and BLASTX analysis compares translated sequences against proteins in the constantly curated RefSeq database. Putative protein-coding transcripts were filtered in a two-step process. First, we filtered transcripts that had positive value for coding potential, which is indicative of high probability for a given RNA to be protein coding. Second, because all transcripts removed in the previous step were considered to be protein coding, we also removed transcripts that were on the same strand and overlapped transcripts removed in the first step.

**Reduction of Redundancy and Size Selection.** Many expressed noncoding transcripts in the initial pool of RNA were annotated by multiple databases or were transcribed into multiple spliced and nonspliced isoforms. Therefore, we reduced redundancy by combining partially or fully overlapping noncoding transcripts into lncRNA loci. Those lncRNA loci, which were on the same strand and whose 5′ ends were within 2 kb of each other, were further combined into unique loci. lncRNA loci of <100 bases were filtered. The expression level of each locus was reported as an average expression of all transcripts within the locus.

**Filtering Repetitive and Protein-Coding Loci.** Some of the loci in the catalog contained repetitive elements. To identify these repetitive loci, we first mapped the genomic origin of each transcript within a locus using BLAT. We then calculated a ratio of BLAT scores for the top two genomic locations corresponding to exonic sequences of the query transcripts. If any transcript within a locus produced a BLAT score ratio of <1.5, the whole locus was considered to be repetitive and excluded from further analysis. Human lncRNA loci that were on the same strand and whose 5′ ends were within 2 kb of each other, were further combined into unique loci. lncRNA loci of <100 bases were filtered. The expression level of each locus was reported as an average expression of all transcripts within the locus.
The nonrepetitive catalogs of lncRNA loci for hESCs, mESCs, and 48 h of human endodermal differentiation are in Dataset S1 (hg18 and mm9).

**Estimation of Sizes of lncRNA Loci.** In estimating the sizes of lncRNA loci, we excluded lncRNA loci that were derived exclusively from the FLJ database for human cells or from NIA Gene Index database for murine cells. Because sequences of many transcripts in these databases were derived from 5' ESTs, their 3' ends might not have been fully characterized. This approach was taken to avoid underestimating the average and median sizes of lncRNA loci. In situations when an lncRNA locus contained transcripts from the FLJ or NIA Gene Index database as well as transcripts from at least one other dataset, the locus was included in size estimations.

**ChiP.** ChiP-seq was performed by using antibodies against H3K4me3, H3K27Ac, and IgG in hESCs that had differentiated toward the endoderm for 48 h and against H3K4me3 in mESCs. Detailed protocols were described (12, 13). Approximately 1 × 10⁶ differentiated hESCs were cross-linked and sonicated for analysis, and samples were split into 1.5 × 10⁶ cells for each ChiP. mESC ChiP experiments were performed with 5–10 × 10⁶ cells. Cells were chemically cross-linked by the addition of one-tenth volume of 10% formaldehyde solution for 15 min at room temperature. Cells were rinsed twice with 1× PBS and harvested using a silicon scraper and flash frozen in liquid nitrogen. Cells were stored at −80 °C before use.

For hESCs ChiPs, cells were sonicated (Misonix Sonicator 3000) in sonication buffer (50 mM Tris·HCl, pH 7.5, 140 mM NaCl, 1 mM EDTA, 1 mM EGTA, 1% Triton X-100, 0.1% Na-deoxycholate, 0.1% SDS) for 10 cycles at 20 s each on ice at 21 °C. The resulting whole-cell extract (WCE) was cleared and then incubated overnight at 4 °C with 100 μL of Dynal Protein G magnetic beads bound by 2 μg of antibody. Beads were washed with sonication buffer, 1× with 20 mM Tris·HCl, pH 8, 500 mM NaCl, 2 mM EDTA, 0.1% SDS, 1% Triton X-100; 1× with 10 mM Tris·HCl, pH 8, 250 mM LiCl, 2 mM EDTA, 1% Nonidet P-40; and 1× with TE containing 50 mM NaCl.

For mESCs ChiPs, the sonication buffer was 10 mM Tris·HCl, pH 8, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% Na-deoxycholate, 0.5% N-lauroylsarcosine. Cells were sonicated at ∼24 W for 10 × 30-s pulses (60-s pause between pulses). Samples were kept on ice at all times. The resulting WCE was incubated overnight at 4 °C with 100 μL of Dynal Protein G magnetic beads that had been preincubated with ∼10 μg of the appropriate antibody. Beads were washed 1× with the sonication buffer; 1× with 20 mM Tris·HCl, pH 8, 500 mM NaCl, 2 mM EDTA, 0.1% SDS, 1% Triton X-100; 1× with 10 mM Tris·HCl, pH 8, 250 mM LiCl, 2 mM EDTA, 1% Nonidet P-40; and 1× with TE containing 50 mM NaCl.

**ChiP-seq Sample Preparation and Analysis.** Purified ChiP DNA from hESCs was used to prepare Illumina multiplexed sequencing libraries. Libraries used the Illumina TruSeq™ DNA Sample Preparation v2 kit protocol with the following exceptions. After end-repair and A-tailing, immunoprecipitated DNA (∼10–50 ng) was ligated to a 1:50 dilution of Illumina Adapter Oligo Mix assigning one unique index in the kit to each sample. Following ligation, libraries were amplified by 18 cycles of PCR using the HiFi NGS Library Amplification kit from KAPA Biosystems. Amplified libraries were size-selected by using a 2% gel cassette in the Pippin PrepTM system from Sage Science to capture fragments between 200 and 400 bp. Libraries were quantified by quantitative PCR using the KAPA Biosystems Illumina Library Quantification kit. Libraries were multiplexed by mixing at equimolar ratios and running together in one lane on the Illumina HiSeq2000.

The concentration of library applied to the flow cell was calibrated such that polonies generated in the bridge amplification step originate from single strands of DNA. Multiple rounds of amplification reagents were flowed across the cell in the bridge amplification step to generate polonies of ∼1,000 strands in 1-μm-diameter spots. Double-stranded polonies were visually checked for density and morphology by staining with a 1:5,000 dilution of SYBR Green I (Invitrogen) and visualizing with a microscope under fluorescent illumination. Validated flow cells were stored at 4 °C until sequencing. Flow cells were removed from storage and subjected to linearization and annealing of sequencing primer on the Cluster Station. Primed flow cells were loaded into the Illumina GAIII. After the first base was incorporated in the sequencing-by-synthesis reaction, the process was paused for a key quality-control checkpoint. A small section of each lane was imaged, and the average intensity value for all four bases was compared with minimum thresholds. Flow cells with low first-base intensities were reprimed, and if signal was not recovered, the flow cell was aborted. Flow cells with signal intensities meeting the minimum thresholds were resubmitted and sequenced for additional 35 cycles.

Purified ChiP DNA from mESCs was prepared for sequencing according to a modified version of the Illumina Genomic DNA protocol. DNA was prepared for ligation of Illumina adapters by repairing the ends and adding a single adenine nucleotide overhang using T4 DNA ligase. For a 1:100 dilution of ChiP DNA, 0.03 μM Oligo Mix (Illumina) was used in the ligation step. A subsequent PCR step with 18 amplification cycles added additional linker sequence to the fragments to prepare them for annealing to the Genome Analyzer flow cell. After amplification, a 2% agarose gel was used to excise bands between 150 and 350 bp (representing fragments between 50 and 250 nt in length and ∼100 bp of primer sequence). The DNA was purified from the agarose gel, and 2–4 pM of the DNA was applied to the flow cell by using the Cluster Station device from Illumina. Sequencing was performed as described above.

**ChiP-seq Data Analysis.** Acquired images were processed through the bundled Illumina image extraction pipeline, which identified polony positions, performed base calling, and generated QC statistics. Sequences were aligned by using Bowtie (4) to NCBI Build 36 (UCSC36, hg18) of the human genome and (UCSC36, mm8) of the murine genome with the parameters -70 -k 2 -m 10 -n 2 --best -strata. A summary of the total number of H3K4me3 ChiP-seq reads that were used in this study is provided in Table S2. ChiP-seq datasets profiling the genomic occupancy of H3K4me3 (14) and H3K27Ac (GSM466732) in hESCs, and H3K27Ac (15) in mESCs were obtained from previous publications and reanalyzed by using the methods described below. Analysis methods were derived from published methods (12, 16–18). For all datasets, each read was extended 200 bp, toward the interior of the sequenced fragment, based on the strand of the alignment. Across the genome, in 25-bp bins, the number of extended ChiP-seq reads was tabulated. The 25-bp genomic bins that contained statistically significant ChiP-seq enrichment were identified by comparison with a Poissonian background model. Assuming background reads are spread randomly throughout the genome, the probability of observing a given number of reads in a 1-kb window can be modeled as a Poisson process in which the expectation can be estimated as the number of mapped reads multiplied by the number of bins (40) into which each read maps, divided by the total number of bins available (we estimated 70%). Enriched bins within 200 bp of one another were combined into regions. The Poissonian background model assumes a random distribution of background reads; however, we have observed significant deviations from this expectation. Some of these nonrandom events can be detected as sites of apparent enrichment in negative control DNA samples and can create many false positives.
in ChIP-seq experiments. To remove these regions, we compared genomic bins and regions that meet the statistical threshold for enrichment to a set of reads obtained from Illumina sequencing of DNA from IgG control ChIP in matched hESCs and WCE in matched mESC samples. We required that enriched bins and enriched regions have fivefold greater ChIP-seq density in the specific IP sample, compared with the control sample, normalized to the total number of reads in each dataset. This served to filter out genomic regions that are biased to having a greater than expected background density of ChIP-seq reads. Enriched genomic regions (P < 10^-9) for H3K4me3 and H3K27Ac in hESCs, H3K4me3 and H3K27Ac after 48 h of differentiation toward the endoderm, and H3K4me3 and H3K27Ac in mESCs are in Dataset S2.

**Rapid Amplification of cDNA Ends.** The 5’ rapid amplification of cDNA ends (RACE) was performed with the following procedures. Large (>200 nt) RNA was prepared by using the mirVana miRNA Isolation Kit (Ambion). Large fractionated RNA (10 μg) was treated with the Turbo DNA-Free kit (Ambion). The First-Choice RLM-RACE kit (Ambion) was used with the following modifications. First, T4 RNA Ligase 1 (ssRNA ligase) was heat-inactivated for 15 min at 65 °C. Second, SuperScript III reverse transcriptase (Invitrogen) was used for cDNA synthesis according to the manufacturer’s instructions. Reverse transcription was performed with a target-specific primer, gsp1, at final concentration of 0.25 μM. Two subsequent nested PCR reactions using HotStarTaq polymerase (Qiagen) were performed with two forward primers, Po and Pi, 5’ RACE adaptor-specific primers (Ambion), and two reverse target-specific primers, gsp-2 and gsp-3, at final concentration of 0.4 μM. PCR products were run on a 2% agarose gel, extracted by using a QIAquick Gel Extraction Kit (Qiagen), cloned into a TOPO TA kit (Invitrogen) sequencing plasmid, and sequenced by using Sanger methods. The 3’ RACE was performed with the following procedures. Poly(A)-selected RNA was obtained from 60 μg of total RNA by using the Dynabeads Oligo(dt) kit (Invitrogen). The GeneRacer kit (Ambion) was used for 3’ RACE according to the manufacturer’s instructions. Two nested PCRs using HotStarTaq polymerase (Qiagen) were performed with two forward target-specific primers, gsp-1 and gsp-2, and a reverse oligo-dT specific primer (GeneRacer kit) at final concentration of 0.4 μM. Sequences of all Sanger sequenced RACE products and target-specific RACE primers are in Dataset S5.

**Gro-seq Analysis and GO Term Analysis for Divergent Genes.** For Gro-seq data, enriched regions were determined with the same model as described in ChIP-seq data analysis with the following differences. Reads were first separated by strand and treated individually. Unique reads of a particular strand were then extended 10 bp and allocated into 10-bp bins. Finally, because no WCE data are available for Gro-seq, we could not use the fivefold over background filter. Instead, we raised the required P value to 1 × 10^-15 and used the resulting enriched regions for analysis. These regions are listed in Dataset S3.

To quantify the amount of Gro-seq signal at a gene’s TSS, the number of same-strand Gro-seq reads (normalized per million total mapped reads) mapping within ±500 bp was calculated. The value for each gene in each dataset is available in Dataset S4. For Fig. 4F and Fig. S4, a gene was considered changing if there was at least a 125% change in the density of Gro-seq reads around its TSS between specified conditions.

Gro-seq gene tracks are a visualization of all mapped reads, shown as total reads or as reads per million mapped reads as described in the respective figure legends. Gro-seq metagenes are a quantification of unique reads, where only one read is allowed to map to any particular genomic location, and any replicate reads mapping to that location are ignored. This was done to ensure that any highly repetitive reads did not skew the mean and cause a misinterpretation of the average profile. χ^2 tests were performed to calculate the significance of coordinate change in transcription for lncRNA and mRNA in lncRNA/mRNA pairs during differentiation. mRNA and lncRNA levels were considered to be changing if there was at least 1.25-fold change in the density of Gro-seq reads around the TSS of the lncRNA or mRNA between hESCs and 48 h of differentiation. Then, 2 × 2 contingency tables were prepared and used to calculate the P value for coordinate induction (P < 4.9 × 10^-24) or coordinate repression (P < 2.5 × 10^-99) of lncRNA/mRNA gene pairs. For these contingency tables, mRNA or lncRNA not changing in the specified direction tested (i.e., induced or repressed, respectively) were included in the unchanged category. In addition, a 3 × 3 contingency table was prepared and used to calculate the P value for coordinate regulation of lncRNA/mRNA pairs, where induced, unchanged, or repressed states were treated as three categories (P < 1.5 × 10^-108).

We conducted Gene Ontology term analysis for 2,100 divergently transcribed mRNA genes in lncRNA/mRNA pairs as well as 18,032 divergently transcribed mRNA genes without detectable lncRNAs. Gene lists for each category were submitted to the Database for Annotation, Visualization and Integrated Discovery (DAVID; http://david.abcc.ncifcrf.gov/), and the top 10 enriched Gene Ontology Biological Process terms (ranked by P value) are displayed in Fig. S4D.
Fig. S1. Estimation of false-negative rates for transcripts expressed in hESCs, 48-h endoderm, and mESCs. (A) Estimation of false-negative rates for transcripts expressed at various expression levels in hESCs. The 1,073 lncRNAs from the current RefSeq gene table with expression above 0 RPKM in hESCs were used for analysis. The left y axis represents the fraction of RefSeq lncRNAs whose promoters are associated with H3K4me3 (but not with H3K27me3) in hESCs; the x axis represents expression levels of these transcripts in hESCs, and the right y axis represents estimated false-negative rates. Rates are shown separately for all correctly annotated RefSeq lncRNAs expressed above the threshold, transcripts within 2 kb of TSS of protein-coding genes, and transcripts that are more than 2 kb away from such genes. (B) Estimation of false-negative rates for transcripts expressed above the threshold of 0.065 RPKM in hESCs. The x axis represents expression levels of these transcripts in hESCs, and the right y axis represents estimated false-negative rate. Abundance of each transcript is expressed as number of RNA-seq reads per kilobase of exonic length per million (RPKM). Transcripts are binned in groups representing various ranges of expression above 0 RPKM in hESCs. The 1,073 lncRNAs from the current RefSeq gene table with expression above 0 RPKM in hESCs were used for analysis. The left y axis represents the fraction of RefSeq lncRNAs whose promoters are associated with H3K4me3 (but not with H3K27me3) in hESCs; the x axis represents expression levels of these transcripts in hESCs, and the right y axis represents estimated false-negative rates. Rates are shown separately for all correctly annotated RefSeq lncRNAs expressed above the threshold, transcripts within 2 kb of TSS of protein-coding genes, and transcripts that are more than 2 kb away from such genes. (C) Estimation of false-negative rates for transcripts expressed above the threshold of 0.260 RPKM in hESCs. The 635 correctly annotated lncRNAs from the current RefSeq gene table with expression above 0 RPKM in hESCs were used for analysis. The left y axis represents the fraction of RefSeq lncRNAs whose promoters are associated with H3K4me3 (but not with H3K27me3) in hESCs; the x axis represents expression levels of these transcripts in hESCs, and the right y axis represents estimated false-negative rates. Rates are shown separately for all correctly annotated RefSeq lncRNAs expressed above the threshold, transcripts within 2 kb of TSS of protein-coding genes, and transcripts that are more than 2 kb away from such genes. (D) Estimation of false-negative rates for transcripts expressed at various expression levels in the 48-h endoderm. The 840 lncRNAs from the current RefSeq gene table with expression above 0 RPKM in the 48-h endoderm were used for analysis. The left y axis represents the fraction of RefSeq lncRNAs whose promoters are associated with H3K4me3 (but not with H3K27me3) in the 48-h endoderm; the x axis represents expression levels of these transcripts in the 48-h endoderm, and the right y axis represents estimated false-negative rate. Abundance of each transcript is expressed as number of RNA-seq reads per kilobase of exonic length per million (RPKM). Transcripts are binned in groups representing various ranges of expression above 0 RPKM in the 48-h endoderm. The 840 lncRNAs from the current RefSeq gene table with expression above 0 RPKM in the 48-h endoderm were used for analysis. The left y axis represents the fraction of RefSeq lncRNAs whose promoters are associated with H3K4me3 (but not with H3K27me3) in the 48-h endoderm; the x axis represents expression levels of these transcripts in the 48-h endoderm, and the right y axis represents estimated false-negative rates. Rates are shown separately for all correctly annotated RefSeq lncRNAs expressed above the threshold, transcripts within 2 kb of TSS of protein-coding genes, and transcripts that are more than 2 kb away from such genes. (E) Estimation of false-negative rates for transcripts expressed at various expression levels in mESCs. The 635 correctly annotated lncRNAs from the current RefSeq gene table with expression above 0 RPKM in mESCs were used for analysis. The left y axis represents the fraction of RefSeq lncRNAs whose promoters are associated with H3K4me3 (but not with H3K27me3) in mESCs, the x axis represents expression levels of these transcripts in mESCs, and the right y axis represents estimated false-negative rate. Abundance of each transcript is expressed as number of RNA-seq reads per kilobase of exonic length per million (RPKM). Transcripts are binned in groups representing various ranges of expression above 0 RPKM in mESCs. The 635 correctly annotated lncRNAs from the current RefSeq gene table with expression above 0 RPKM in mESCs were used for analysis. The left y axis represents the fraction of RefSeq lncRNAs whose promoters are associated with H3K4me3 (but not with H3K27me3) in mESCs, the x axis represents expression levels of these transcripts in mESCs, and the right y axis represents estimated false-negative rates. Rates are shown separately for all correctly annotated RefSeq lncRNAs expressed above the threshold, transcripts within 2 kb of TSS of protein-coding genes, and transcripts that are more than 2 kb away from such genes.
Fig. S2. Characteristics of lncRNA loci in the hESC and endodermal lncRNA catalogs as well as fraction of protein-coding genes with evidence of divergent transcription initiation. (A) Size distribution of lncRNA loci in hESCs. The black lines indicate median and average sizes of lncRNA loci. (B) Expression levels of lncRNA loci and mRNAs in hESCs and in the 48-h endoderm, as measured by directional RNA-seq and expressed as reads per kilobase of exonic length per million (RPKM) (1). A horizontal black line in the colored region indicates median expression. (C) Venn diagram depicting overlap between the combined catalogs of lncRNA loci expressed in hESCs and 48-h endoderm (this study) and lncRNAs identified in the current RefSeq table, by global analysis of large intergenic noncoding RNAs (lincRNAs) in multiple human cell types (2), and GENCODE v7 (3) (other studies*). (D) Size distribution of lncRNA loci in the 48-h endoderm. The black lines indicate median and average sizes of lncRNA loci. (E) Venn diagram depicting overlap between the catalog of lncRNA loci expressed in hESCs and the catalog of the 48-h endoderm generated in this study. (F) Fraction of protein-coding genes with evidence of transcription initiation. Coding genes with significant GRO-seq signal at the TSS were classified as initiated, whereas the remaining genes were classified as silent (SI Materials and Methods). (G) Fraction of protein-coding genes with evidence of divergent transcription initiation. GRO-seq data were used as in F to determine if there was evidence of divergent transcription at each start site. (H) Expression levels of divergently initiated mRNAs in lncRNA/mRNA pairs and those without detectable lncRNAs as measured by directional RNA-seq and expressed as reads per kilobase of exonic length per million (RPKM). A horizontal black line in the colored region indicates median expression. (I) Alignment of GRO-seq reads for the 12,620 divergently initiated protein-coding genes that do not contain lncRNAs or other protein-coding genes within 2 kb of their TSS. The x axis indicates the distance from the TSS in kilobases. The y axis indicates the average number of uniquely mapped GRO-seq counts normalized to reads per genomic bin per million uniquely mapped reads. Reads that map to Watson (black) and Crick (red) strands of genomic DNA are shown separately as indicated on the graph. (J) Alignment of RNA-seq reads for the 12,620 divergently initiated protein-coding genes that do not contain lncRNAs or other protein-coding genes within 2 kb of their TSS. Reads are aligned in 250-bp bins. The x axis indicates the distance from the TSS in kilobases. The y axis indicates the average number of uniquely mapped RNA-seq counts normalized to reads per genomic bin per million uniquely mapped reads. Reads that map to Watson (black) and Crick (red) strands of genomic DNA are shown separately as indicated on the graph.


Sigova et al. www.pnas.org/cgi/content/short/1221904110 6 of 11
Fig. S3. Discovery and characterization of the mESC IncRNA catalog. (A) Schematic diagram of experimental and computational methods used to create a catalog of IncRNAs in mESCs. An “initial RNA pool” of RNA transcripts was compiled from transcripts assembled de novo from RNA-seq reads using Scripture (this study; SI Materials and Methods), Functional Annotation of Mouse 3 (FANTOM3) library (1), transcripts from the NIA mouse cDNA project (2, 3), and previous ab initio reconstruction of mESC transcripts (4). Four criteria required for the selection of expressed transcripts from this pool are indicated in red. Transcripts were required to be expressed from a high-confidence start site (occupied by H3K4me3), to be noncoding (lacking features of protein-coding RNAs as defined by the Coding Potential Calculator (CPC) (5)), to be long (>100 nt), and to be nonrepetitive. (B) Distribution of TSS of lncRNAs relative to the TSS of protein-coding genes. Coding regions are normalized to equal length and the regions upstream of associated promoters are divided into one hundred 100-bp bins. Distance between TSS of protein-coding gene and 5’ end of IncRNA is indicated on the x axis and expressed in kilobases (kb). Antisense lncRNA loci are indicated in red. Sense lncRNA loci are indicated in blue. (C) Size distribution of lncRNA loci in mESCs. The black lines indicate median and average sizes of lncRNA loci. (D) Example of an IncRNA locus. The 5’ ends of IncRNA loci are indicated by red boxes, and the 3’ ends by blue boxes. (E) Log2 of expression, RPKM. mRNA and IncRNA expression levels are shown for each gene. The black lines indicate median and average expression levels.
lncRNA loci. (D) The promoter-associated lncRNA locus 625 identified in our screen originates from a locus shown previously to produce divergent transcripts (6). ChIP-seq binding profile of nucleosomes with H3K4me3 is represented. GRO-seq reads for nascent transcripts and RNA-seq reads for polyadenylated RNA are shown relative to gene model for the lncRNA upstream of Aen. The 5′ end of the lncRNA model corresponds to the 5′ end of previously published divergent transcripts at Aen (Isg20l1) (6), whereas the 3′ end was determined by 3′ RACE optimized for detecting low abundant polyadenylated lncRNA species (SI Materials and Methods and Dataset S5). Primers used for 3′ RACE of the Aen promoter-associated lncRNA are in Dataset S5. 3′ RACE products were run on a 2% agarose gel, cloned into the sequencing plasmid, and sequenced using the Sanger method to determine the identity. Sequences of 3′ RACE products as well as the sequence of the full-length lncRNA are in Dataset S5. Black arrows depict TSS and direction of transcription for respective genes. Aen mRNA and the promoter-associated lncRNA are shown in blue and red, respectively. Genomic DNA sequence corresponding to the 3′ end of the cloned lncRNA is shown in black at the bottom of the gene track. Sequence of the 3′ end of the cloned lncRNA is shown in red. Black rectangle points to the sequence and position of nontemplated nucleotides. Light blue rectangle depicts upstream antisense polyadenylation signal sequence (PAS). (E) Expression levels of lncRNA loci and mRNAs in mESCs as measured by directional RNA-seq and expressed as reads per kilobase of exon length per million (RPKM) (7). A horizontal black line in the colored region indicates median expression.

Fig. S4. Coordinate transcriptional induction and repression of lncRNA/mRNA gene pairs during differentiation of hESCs toward the endoderm. 

(A) Coordinate transcriptional repression of lncRNA/mRNA gene pairs. The 283 lncRNA/mRNA pairs where the numbers of GRO-seq reads within mRNA genes decreased at least 1.25-fold after 48 hr of endodermal differentiation were selected. The average number of GRO-seq reads per million of mapped reads within 2 kb of TSS of mRNA half of the pair (black) is shown on top. The average number of GRO-seq reads per million of mapped reads for the lncRNA half of the pair (red) within 2 kb of TSS of the lncRNA is shown flipped/rotated beneath. Solid lines represent average transcription in hESCs; dashed lines represent average transcription at 48 hr after induction of differentiation toward the endoderm. 

(B) Coordinate transcriptional induction of lncRNA/mRNA gene pairs. The 364 lncRNA/mRNA pairs where the numbers of GRO-seq reads within mRNA genes increased at least 1.25-fold after 1 hr of endodermal differentiation were selected. The average number of GRO-seq reads per million of mapped reads within 2 kb of TSS of mRNA half of the pair (black) is shown on top. The average number of GRO-seq reads per million of mapped reads for the lncRNA half of the pair (red) within 2 kb of TSS of the lncRNA is shown flipped/rotated beneath. Solid lines represent average transcription in hESCs; dashed lines represent average transcription at 1 hr after induction of differentiation toward the endoderm. 

Legend continued on following page
endoderm. (C) Coordinate transcriptional repression of lncRNA/mRNA gene pairs. The 189 lncRNA/mRNA pairs where the numbers of GRO-seq reads within mRNA genes decreased at least 1.25-fold after 1 h of endodermal differentiation were selected. The average number of GRO-seq reads per million of mapped reads within 2 kb of TSS of mRNA half of the pair (black) is shown on top. The average number of GRO-seq reads per million of mapped reads for the lncRNA half of the pair (red) within 2 kb of TSS of the lncRNA is shown flipped/rotated beneath. Solid lines represent average transcription in hESCs, dashed lines represent average transcription at 1 h after induction of differentiation toward the endoderm. (D) GO term analysis results for 2,100 divergently transcribed mRNA genes in lncRNA/mRNA pairs as well as 18,032 divergently transcribed mRNA genes without detectable lncRNAs. Gene lists for each category were submitted to DAVID (http://david.abcc.ncifcrf.gov/), and the top 10 enriched Gene Ontology Biological Process terms (ranked by P value) are shown.

### Table S1. RNA-seq and GRO-seq datasets generated in this study

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Build</th>
<th>Sequencing platform</th>
<th>Read length</th>
<th>No. of aligned reads</th>
<th>GEO accession nos.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poly(A)-selected RNA</td>
<td>MM8</td>
<td>HiSeq. 2000</td>
<td>80 (SE)</td>
<td>70,415,559</td>
<td>GSE36799</td>
</tr>
<tr>
<td>Ribo-depleted RNA</td>
<td>MM8</td>
<td>HiSeq. 2000</td>
<td>80 (PE)</td>
<td>90,049,846</td>
<td>GSE36799</td>
</tr>
<tr>
<td>RNA-Seq_ES</td>
<td>HG18</td>
<td>HiSeq. 2000</td>
<td>80 (PE)</td>
<td>224,358,543</td>
<td>GSE41009</td>
</tr>
<tr>
<td>RNA-Seq_48hr</td>
<td>HG18</td>
<td>HiSeq. 2000</td>
<td>80 (PE)</td>
<td>246,268,203</td>
<td>GSE41009</td>
</tr>
<tr>
<td>Gro_hESC</td>
<td>HG18</td>
<td>HiSeq. 2000</td>
<td>80 (PE)</td>
<td>157,309,428</td>
<td>GSE41009</td>
</tr>
<tr>
<td>Gro_48hr</td>
<td>HG18</td>
<td>HiSeq. 2000</td>
<td>80 (PE)</td>
<td>159,207,580</td>
<td>GSE41009</td>
</tr>
<tr>
<td>Gro_1hr</td>
<td>HG18</td>
<td>HiSeq. 2000</td>
<td>80 (PE)</td>
<td>157,309,428</td>
<td>GSE41009</td>
</tr>
<tr>
<td>Gro_0hr</td>
<td>HG18</td>
<td>HiSeq. 2000</td>
<td>80 (PE)</td>
<td>60,869,930</td>
<td>GSE41009</td>
</tr>
</tbody>
</table>

*SE, single end; PE, paired end.*

### Table S2. ChIP-seq datasets generated in this study

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sequencing platform</th>
<th>Read length, bp</th>
<th>No. of aligned reads</th>
<th>GEO accession nos.</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3K4me3</td>
<td>GAII</td>
<td>36, SE</td>
<td>2,557,663</td>
<td>GSM8969 GSM896921</td>
</tr>
<tr>
<td>H3K27Ac</td>
<td>HiSeq. 2000</td>
<td>40, SE</td>
<td>17,451,219</td>
<td>GSE41009</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>HiSeq. 2000</td>
<td>40, SE</td>
<td>15,989,330</td>
<td>GSE41009</td>
</tr>
<tr>
<td>IgG</td>
<td>HiSeq. 2000</td>
<td>40, SE</td>
<td>13,831,604</td>
<td>GSE41009</td>
</tr>
</tbody>
</table>

*SE, single end.*

### Table S3. Publicly available annotations

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Download date</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein-coding transcripts from RefSeq</td>
<td>02/21/2012</td>
<td>1</td>
</tr>
<tr>
<td>microRNA, tRNA, rRNA, and snoRNAs from Ensembl</td>
<td>01/04/2012</td>
<td>2</td>
</tr>
<tr>
<td>Pseudogenes from RefSeq</td>
<td>02/21/2012</td>
<td>1</td>
</tr>
<tr>
<td>Noncoding transcripts from RefSeq</td>
<td>02/21/2012</td>
<td>1</td>
</tr>
<tr>
<td>Ab initio reconstructed transcripts from previously published study</td>
<td>09/01/2011</td>
<td>3</td>
</tr>
<tr>
<td>FANTOM3</td>
<td>02/21/2012</td>
<td>4</td>
</tr>
<tr>
<td>mm9 tBLASTn alignments of peptides from genes identified in the hg18</td>
<td>03/15/2012</td>
<td>5</td>
</tr>
<tr>
<td>Yale Pseudogenes based on Ensembl Release 60</td>
<td>03/15/2012</td>
<td>6</td>
</tr>
<tr>
<td>NIA Gene Index</td>
<td>02/23/2012</td>
<td>7, 8</td>
</tr>
<tr>
<td>FLJ</td>
<td>04/13/2012</td>
<td>9</td>
</tr>
</tbody>
</table>

Dataset S1.  Initial pools of RNA and catalogs of IncRNA loci expressed in hESCs, mESCs, and 48-h human endoderm

Dataset S2.  Enriched regions for H3K4me3 ($P < 10^{-9}$) and H3K27Ac ($P < 10^{-9}$)

Dataset S3.  GRO-seq enriched regions ($P < 10^{-15}$)

Dataset S4.  Evidence of divergent transcription initiation

Dataset S5.  Sequences of cloned RACE product, full-length IncRNA, and primers used for analysis of the IncRNA shown in Fig. S3D