

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

Alla A. Sigova^{a,1}, Alan C. Mullen^{a,b,1}, Benoit Molinie^{b,1}, Sumeet Gupta^{a,1}, David A. Orlando^{a,1}, Matthew G. Guenther^{a,1}, Albert E. Almada^{c,d}, Charles Lin^{a,c}, Phillip A. Sharp^{c,d}, Cosmas C. Giallourakis^{b,e,2}, and Richard A. Young^{a,c,2}

^aWhitehead Institute for Biomedical Research, Cambridge, MA 02142; ^bGastrointestinal Unit, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114; ^cDepartment of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139; ^dDavid H. Koch Institute for Integrative Cancer Research, Cambridge, MA 02140; and ^eHarvard Stem Cell Institute, Cambridge, MA 02138

Contributed by Richard A. Young, December 17, 2012 (sent for review November 30, 2012)

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.

development | expression

The non-protein-coding portion of the mammalian genome is transcribed into a vast array of RNA species (1), some of which play important roles in cellular regulation, development, and disease (2). The long noncoding RNAs (lncRNAs) are of particular interest because they are known to contribute to gene silencing (3), 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.

Results

lncRNAs Expressed in Human ESCs. We compiled a catalog of lncRNA species expressed in hESCs as summarized in Fig. 1A. An initial pool of RNA candidates was generated by sequencing polyadenylated RNA species from hESCs and supplementing these with EST data from the full-length long Japan (FLJ) collection of sequenced human cDNAs, which contains transcripts expressed in >60 human tissues, including embryonal tissue (20). An initial pool of 170,162 ncRNA candidates (Dataset S1) was obtained after removing protein-coding transcripts based on the National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq). This pool was further filtered by using multiple criteria to identify lncRNAs. The RNA species were required to have a 5' end that originates from a genomic site where there is corroborating evidence of active transcription initiation, to lack features of

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 BLASTX 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

Author contributions: A.A.S., A.C.M., S.G., P.A.S., C.C.G., and R.A.Y. designed research; A.A.S., A.C.M., B.M., M.G.G., and A.E.A. performed research; A.A.S., A.C.M., S.G., D.A.O., C.L., and C.C.G. analyzed data; and A.A.S., A.C.M., M.G.G., P.A.S., and R.A.Y. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession nos. GSE36799, GSE41009, GSM896920, and GSM896921).

¹A.A.S., A.C.M., B.M., S.G., D.A.O., and M.G.G. contributed equally to this work.

²To whom correspondence may be addressed. E-mail: young@wi.mit.edu or CGIALLOURAKIS@partners.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1221904110/-DCSupplemental.

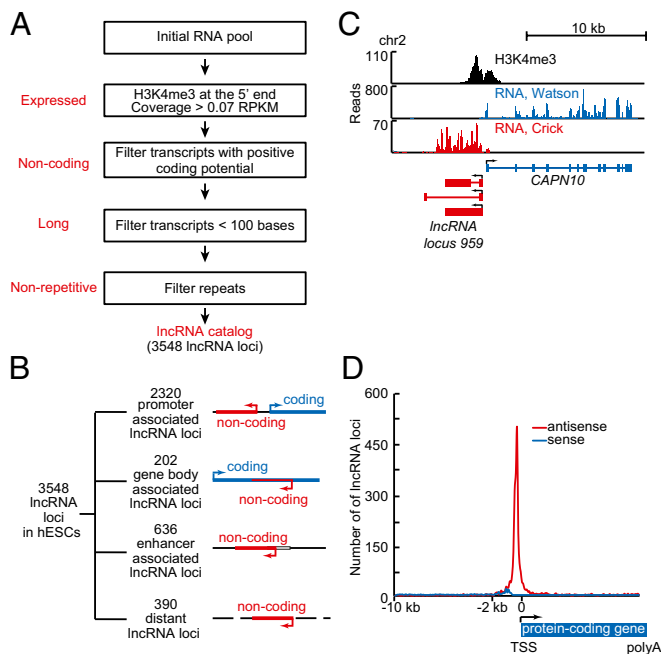


Fig. 1. Most lncRNAs are associated with active protein-coding genes in hESCs. (A) Schematic diagram of pipeline for identification of lncRNAs in hESCs. An “initial RNA pool” was compiled from transcripts assembled de novo from RNA-seq reads (this study; *S1 Materials and Methods*) and published data (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) Summary of various types and numbers of lncRNA loci in hESCs, which are listed in *Dataset S1*. Diagrams at right depict lncRNA 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 lncRNAs overlap or originate at genomic regions enriched in nucleosomes with histone 3 acetylated at lysine 27 (H3K27Ac). Enriched regions for H3K27Ac are available in *Dataset S2*. (C) 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 reads for polyadenylated RNA in the vicinity of *CAPN10*. Transcription at *lncRNA locus 959* generates three alternatively spliced lncRNA transcripts that are divergent from *CAPN10*. The x axis represents the linear sequence of genomic DNA, and the y axis represents the total number of ChIP-seq and RNA-seq mapped reads. RNA-seq reads that map to Watson (blue) and Crick (red) strands of genomic DNA are shown separately. The scale is indicated in the upper right. (D) 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 lncRNA is indicated on x axis and expressed in kilobases (kb). Antisense lncRNA loci are indicated in red. Sense lncRNA loci are indicated in blue.

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 lncRNA species expressed in hESCs that originated from 3,548 nonredundant loci (*Dataset S1*).

The lncRNAs 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 lncRNAs accumulate to levels that are, on average, ~10% that of mRNAs in hESCs (*Fig. S2B*). The size range and average abundance of lncRNAs in this catalog are similar to those described previously for lncRNAs (16, 29). Approximately half of the lncRNA loci contain spliced transcripts (*Dataset S1*). The vast majority (73%) of the lncRNA loci in the hESC catalog have not been previously identified (*Fig. S2C*).

Association of lncRNA Genes with mRNA Genes in hESCs. Inspection of the genomic positions of lncRNA loci revealed that the majority (89%) are associated with the promoters, enhancers, and bodies of protein-coding genes (*Fig. 1B*). Most lncRNAs 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*).

Visual inspection of tracks at individual genes suggested that many of the lncRNA species are transcribed divergently from regions near the promoters of protein-coding genes. At the *CAPN10* locus, for example, lncRNAs originate ~3 bp upstream of the *CAPN10* protein-coding transcript, and these lncRNAs are transcribed in the antisense orientation relative to the protein-coding transcript (*Fig. 1C*). Analysis of the entire population of lncRNAs revealed that the vast majority have 5' ends that occur within 2 kb of the TSS of protein-coding genes, and almost all of these (95%) are transcribed antisense to the protein-coding gene (*Fig. 1D*). The median distance between the TSS for the lncRNAs and the TSS for adjacent protein-coding mRNAs was 329 bp. We refer to these closely associated lncRNA and mRNA genes as lncRNA/mRNA gene pairs.

Previous studies have shown that mammalian RNA polymerase II (Pol II) often initiates transcription divergently at active promoters and that such transcription can generate low-abundance antisense ncRNAs (30–32). To further study the possibility that ESC lncRNAs 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 lncRNA/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. 2A*). Inspection of the data at individual genes shows that this divergent transcription is associated with lncRNA/mRNA gene pairs (*Fig. 2B*), consistent with the model that these lncRNAs are produced as a result of divergent transcription from promoters of active protein-coding genes.

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 lncRNA species that accumulated to substantial levels. The divergently transcribed lncRNA/mRNA gene pairs that produce detectable lncRNA species tend to be transcribed at higher levels (based on GRO-seq data) than those pairs whose lncRNAs do not accumulate (*Fig. S2H–J*). Although the number of transcriptionally engaged RNA Pol II molecules was similar for mRNA and lncRNA in divergently transcribed lncRNA/mRNA pairs (*Fig. 2A*), the steady-state levels of the lncRNAs were much lower than those for mRNAs from the gene pairs (*Fig. 2C*). Thus, although divergent transcription occurs at most active protein-coding genes, only a small fraction of lncRNAs produced by divergent transcription are sufficiently stable to be detected under the conditions studied here. The evidence that lncRNAs are less stable than mRNAs is consistent with previous reports that lncRNA transcripts are subjected to exosome-mediated degradation (33, 34).

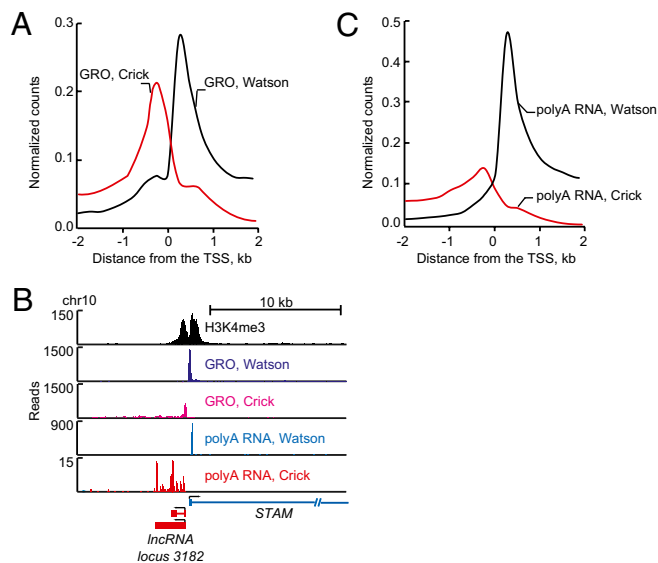


Fig. 2. IncRNAs are derived from divergent transcription of active protein-coding genes in hESCs. **(A)** Alignment of GRO-seq reads for the 2,318 protein-coding genes that contain IncRNAs within 2 kb of their TSS. Reads are aligned in 250-bp bins. The x axis indicates the distance from the TSS in kilobases. The y axis indicates the average number of uniquely mapped 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 graph. **(B)** Example of IncRNA locus whose 5' end occurs within 2 kb of the TSS of a protein-coding gene (promoter-associated IncRNA). 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 *IncRNA locus 3182* generates two alternatively spliced IncRNA 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. **(C)** Alignment of RNA-seq reads for the 2,318 protein-coding genes that contain IncRNAs 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 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.

Association of IncRNA Genes with mRNA Genes in Murine ESCs. Previous studies of murine (m) ESC IncRNAs 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 IncRNA 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 IncRNA population (Fig. S3A and SI Materials and Methods). Inspection of the genomic positions of the murine IncRNA loci revealed that the majority of IncRNAs are transcribed from sites near or within actively transcribed protein-coding genes (Fig. 3A), as in hESCs. Of the 1,664 IncRNA 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 IncRNAs 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 IncRNAs 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 S5), and 30% have been subjected to splicing (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 IncRNA genes in ESCs are transcribed divergently from active protein-coding genes suggests that

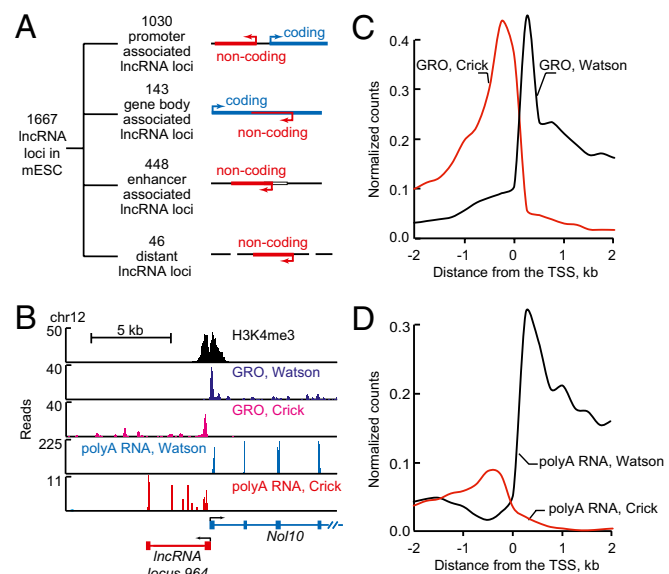


Fig. 3. Most IncRNAs are divergently transcribed from protein-coding genes in mESCs. **(A)** Summary of various types and numbers of IncRNA loci in the mESC catalog, which are listed in Dataset S1. Diagrams at right depict IncRNA loci as red lines, protein-coding genes as blue lines, and an enhancer as an open box. An arrow indicates direction of transcription initiation. Enhancer-associated IncRNAs overlap or originate at genomic regions enriched in H3K27Ac (49). Enriched regions for H3K27Ac are available in Dataset S2. **(B)** Example of IncRNA locus whose 5' end occurs within 2 kb of the TSS of a protein-coding gene (promoter-associated IncRNA). Gene tracks represent ChIP-seq data for H3K4me3 modified nucleosomes (this study), together with GRO-seq reads and reads for polyadenylated RNA in the vicinity of *Nol10*. Transcription at *IncRNA locus 1160* generates IncRNA transcripts that are divergent from *Nol10*. 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 above the track. **(C)** Alignment of GRO-seq reads for the 1,030 protein-coding genes that contain IncRNAs 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 represents 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 graph. **(D)** Alignment of RNA-seq reads for the 1,030 protein-coding genes that contain IncRNAs 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.

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

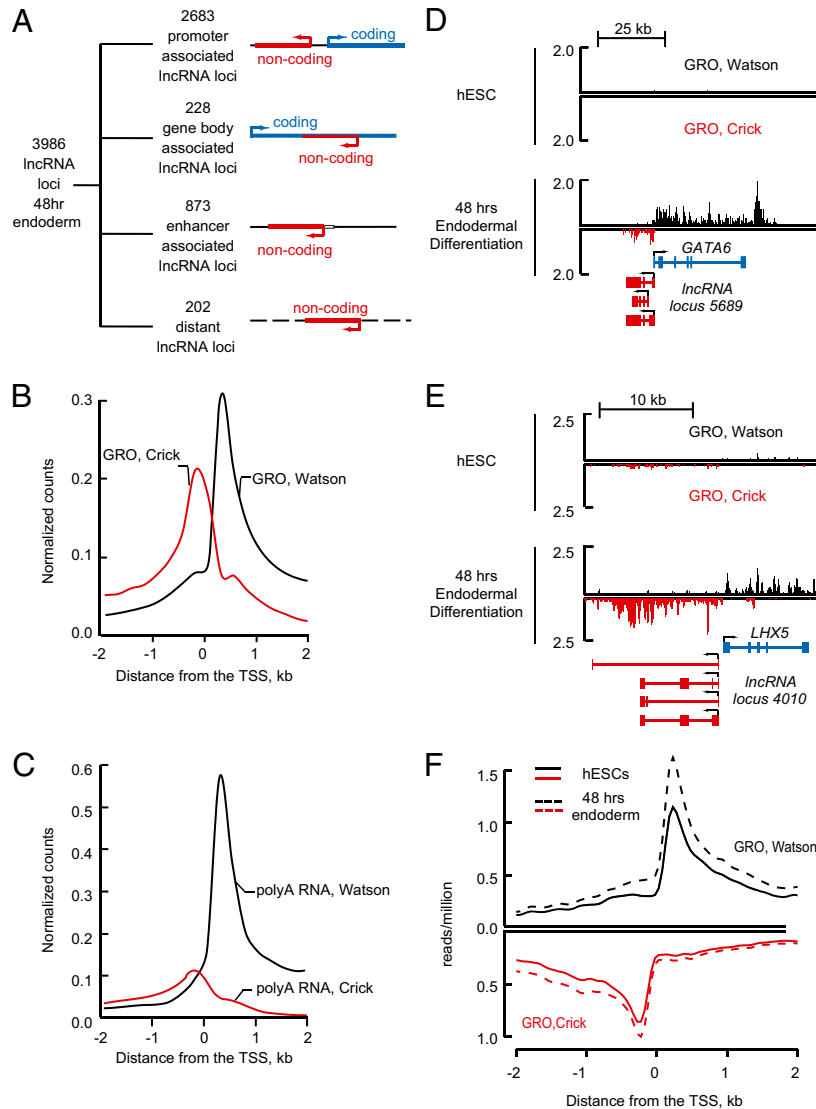


Fig. 4. Divergent lncRNA/mRNA pairs exhibit coordinated changes in transcription as ESCs differentiate into endoderm. (A) Summary of the genomic distribution of lncRNA loci 48 h after induction of endodermal differentiation in hESCs. Diagrams at right depict lncRNA loci as red lines, protein-coding genes as blue lines, and an enhancer as an open box. An arrow indicates direction of transcription initiation. Enhancer-associated lncRNAs 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 lncRNAs 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 lncRNAs 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 lncRNA/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 *lncRNA 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 lncRNA/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 *lncRNA 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. 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. (F) Co-ordinate transcriptional induction of lncRNA/mRNA gene pairs. A total of 683 lncRNA/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 lncRNA 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.

contains lncRNA species that originate from 3,986 nonredundant loci (Dataset S1 and Fig. 4). Approximately half of the lncRNAs in the catalog are also expressed in hESCs (Fig. S2E). The endodermal lncRNAs—the vast majority (73%) 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 687,089 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 coordinated with induction of associated lncRNAs (χ^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 (χ^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. S4B 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 mRNA-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 (48 h) was performed as described (21) by using H3K4me3 (Millipore; 07-473), H3K27Ac (Abcam; AB4729), and IgG (Millipore; 12-370) antibodies (Table S2). Illumina protocols for library preparation, 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. GSM733748, 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 ($P < 10^{-9}$) 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 5×10^6 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 FLJ (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

with positive coding potential, select for long transcripts, and remove repetitive elements (*SI Materials and Methods*). We combined partially or fully overlapping transcripts into lincRNA 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*.

- Djebali S, et al. (2012) Landscape of transcription in human cells. *Nature* 489(7414): 101–108.
- Esteller M (2011) Non-coding RNAs in human disease. *Nat Rev Genet* 12(12):861–874.
- Yu W, et al. (2008) Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature* 451(7175):202–206.
- Zhao J, Sun BK, Erwin JA, Song JJ, Lee JT (2008) Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* 322(5902):750–756.
- Pandey RR, et al. (2008) Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol Cell* 32(2):232–246.
- Sluetsels F, Zwart R, Barlow DP (2002) The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* 415(6873):810–813.
- Rinn JL, Bondre C, Gladstone HB, Brown PO, Chang HY (2006) Anatomic demarcation by positional variation in fibroblast gene expression programs. *PLoS Genet* 2(7):e119.
- Heo JB, Sung S (2011) Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science* 331(6013):76–79.
- Young TL, Matsuda T, Cepko CL (2005) The noncoding RNA taurine upregulated gene 1 is required for differentiation of the murine retina. *Curr Biol* 15(6):501–512.
- Young RA (2011) Control of the embryonic stem cell state. *Cell* 144(6):940–954.
- Hawkins RD, et al. (2010) Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* 6(5):479–491.
- Min IM, et al. (2011) Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev* 25(7):742–754.
- Adelman K, Lis JT (2012) Promoter-proximal pausing of RNA polymerase II: Emerging roles in metazoans. *Nat Rev Genet* 13(10):720–731.
- Guttman M, et al. (2011) lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* 477(7364):295–300.
- Guttman M, et al. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458(7235):223–227.
- Cabili MN, et al. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 25(18):1915–1927.
- Khalil AM, et al. (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci USA* 106(28):11667–11672.
- Derrien T, et al. (2012) The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res* 22(9): 1775–1789.
- Jia H, et al. (2010) Genome-wide computational identification and manual annotation of human long noncoding RNA genes. *RNA* 16(8):1478–1487.
- Ota T, et al. (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat Genet* 36(1):40–45.
- Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA (2007) A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* 130(1):77–88.
- Bernstein BE, et al. (2002) Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc Natl Acad Sci USA* 99(13):8695–8700.
- Schübeler D, et al. (2004) The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev* 18(11): 1263–1271.
- Rahl PB, et al. (2010) c-Myc regulates transcriptional pause release. *Cell* 141(3): 432–445.
- Marson A, et al. (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* 134(3):521–533.
- Santos-Rosa H, et al. (2002) Active genes are tri-methylated at K4 of histone H3. *Nature* 419(6905):407–411.
- Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP (2011) Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* 147(7):1537–1550.
- Kong L, et al. (2007) CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* 35(Web Server Issue):W345–W349.
- Guttman M, et al. (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* 28(5):503–510.
- Seila AC, et al. (2008) Divergent transcription from active promoters. *Science* 322(5909): 1849–1851.
- Kapranov P, et al. (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316(5830):1484–1488.
- Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* 322(5909):1845–1848.
- Flynn RA, Almada AE, Zamudio JR, Sharp PA (2011) Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. *Proc Natl Acad Sci USA* 108(26):10460–10465.
- Preker R, et al. (2008) RNA exosome depletion reveals transcription upstream of active human promoters. *Science* 322(5909):1851–1854.
- Hung T, et al. (2011) Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nat Genet* 43(7):621–629.
- Dinger ME, et al. (2008) Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res* 18(9):1433–1445.
- Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS (2008) Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci USA* 105(2):716–721.
- Ponjavic J, Oliver PL, Lunter G, Ponting CP (2009) Genomic and transcriptional colocalization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS Genet* 5(8):e1000617.
- Wang ET, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456(7221):470–476.
- Feng J, et al. (2006) The Evi-2 noncoding RNA is transcribed from the Dlx-5/6 ultra-conserved region and functions as a Dlx-2 transcriptional coactivator. *Genes Dev* 20(11):1470–1484.
- Mullen AC, et al. (2011) Master transcription factors determine cell-type-specific responses to TGF- β signaling. *Cell* 147(3):565–576.
- Boyer LA, et al. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122(6):947–956.
- Trapnell C, Pachter L, Salzberg SL (2009) TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* 25(9):1105–1111.
- Trapnell C, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28(5):511–515.
- Carninci P, et al.; FANTOM Consortium; RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) (2005) The transcriptional landscape of the mammalian genome. *Science* 309(5740):1559–1563.
- Sharov AA, et al. (2003) Transcriptome analysis of mouse stem cells and early embryos. *PLoS Biol* 1(3):E74.
- Sharov AA, Dudekula DB, Ko MS (2005) Genome-wide assembly and analysis of alternative transcripts in mouse. *Genome Res* 15(5):748–754.
- Ernst J, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473(7345):43–49.
- Creyghton MP, et al. (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci USA* 107(50):21931–21936.