RED CELLS, IRON, AND ERYTHROPOIESIS

Lineage and species-specific long noncoding RNAs during erythro-megakaryocytic development

Vikram R. Paralkar,1 Tejaswini Mishra,2,3 Jing Luan,4 Yu Yao,4 Andrew V. Kossenkov,5 Stacie M. Anderson,6 Margaret Dunagin,7 Maxim Pimkin,4 Meghneel Gore,8 Diana Sun,4 Neeraja Konuthula,6 Arjun Raj,7 Xiuli An,10 Narla Mohandas,11 David M. Bodine,12 Ross C. Hardison,2,3 and Mitchell J. Weiss4

1Division of Hematology/Oncology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; 2Center for Comparative Genomics and Bioinformatics, and 3Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA; 4Division of Hematology, The Children’s Hospital of Philadelphia, Philadelphia, PA; 5Center for Systems and Computational Biology, The Wistar Institute, Philadelphia, PA; 6National Human Genome Research Institute Flow Cytometry Core Facility, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD; 7Department of Bioengineering, University of Pennsylvania, Philadelphia, PA; 8Salesforce, San Francisco, CA; 9Icahn School of Medicine at Mount Sinai, New York, NY; 10Laboratory of Membrane Biology, and 11Red Cell Physiology Laboratory, New York Blood Center, New York, NY; and 12Genetics and Molecular Biology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD

Key Points
• Hundreds of lineage-specific IncRNAs are expressed during mouse and human erythropoiesis.
• Most mouse erythroid IncRNAs are not expressed in human erythroblasts and vice versa, yet some appear to be functional.

Mammals express thousands of long noncoding (Inc) RNAs, a few of which are known to function in tissue development. However, the entire repertoire of IncRNAs in most tissues and species is not defined. Indeed, most IncRNAs are not conserved, raising questions about function. We used RNA sequencing to identify 1109 polyadenylated IncRNAs expressed in erythroblasts, megakaryocytes, and megakaryocyte-erythroid precursors of mice, and 594 in erythroblasts of humans. More than half of these IncRNAs were unannotated, emphasizing the opportunity for new discovery through studies of specialized cell types. Analysis of the mouse erythro-megakaryocytic polyadenylated IncRNA transcriptome indicates that ~75% arise from promoters and 25% from enhancers, many of which are regulated by key transcription factors including GATA1 and TAL1. Erythroid IncRNA expression is largely conserved among 8 different mouse strains, yet only 15% of mouse IncRNAs are expressed in humans and vice versa, reflecting dramatic species-specificity. RNA interference assays of 21 abundant erythroid-specific murine IncRNAs in primary mouse erythroid precursors identified 7 whose knockdown inhibited terminal erythroid maturation. At least 6 of these 7 functional IncRNAs have no detectable expression in human erythroblasts, suggesting that lack of conservation between mammalian species does not predict lack of function. (Blood. 2014;123(12):1927-1937)

Introduction

Long noncoding (Inc) RNAs, defined as a subclass of noncoding RNAs that exceeds 200 nucleotides, are highly diverse, with thousands identified in many cell types and species.1,2 LncRNAs play a variety of roles in eukaryotic development by regulating gene expression transcriptionally and posttranscriptionally. However, the vast majority of IncRNAs remain uncharacterized, and there is debate over what proportion of expressed IncRNAs actually has biological function.3 This issue is fueled by observations that orthologous IncRNAs expressed across species typically exhibit low nucleotide sequence conservation and that some IncRNAs transcribed in 1 species are not transcribed in others.

LncRNAs in hematopoiesis are just beginning to be defined.4,5 The human IncRNAs EGO and HOTAIR1 regulate eosinophil granule protein expression and myeloid differentiation genes, respectively.6,7 LincRNA-EPS promotes the survival of mouse erythroblasts by suppressing Pycard, a pro-apoptotic gene.8 Imprinting of the IncRNA H19 regulates mouse hematopoietic stem cell quiescence, at least in part via its processing into the microRNA miR-675.9 Loss of Xist, an IncRNA that coordinates X chromosome inactivation in females, causes hematologic cancers in mice.10 A recent study defined IncRNAs expressed during murine erythropoiesis and used RNA interference (RNAi) to identify 12 with potential functions.11 However, comprehensive identification and analyses of IncRNAs expressed during all stages of hematopoiesis and in multiple species are lacking.

The current study uses deep sequencing of polyA+ RNA to examine IncRNA expression in purified murine megakaryocyte-erythroid precursors (MEPs), megakaryocytes, and erythroblasts as well as in human erythroblasts. We defined hundreds of IncRNAs unique to
each cell type and focused on the erythroid lineage to investigate several questions related to lncRNA biology. First, do erythroid precursors express novel and/or lineage-specific lncRNAs? Second, how are erythroid lncRNAs regulated and to what extent is their transcription conserved between species? Finally, does conserved lncRNA expression (or lack thereof) predict whether they are biologically active? Our findings identify distinct populations of lncRNAs in mouse and human erythroblasts, very few of which have cross-species orthologs, and suggest developmental roles for multiple lncRNAs in terminal mouse erythropoiesis despite this lack of conservation.

Materials and methods

Cell purification

Animal studies were performed under approved ethical protocols at The Children’s Hospital of Philadelphia and The National Institutes of Health. Murine hematopoietic populations were derived from embryonic day 14.5 day CD-1 fetal livers and bone marrow of C57BL/6 adult animals as described in supplemental Methods on the Blood Web site and Pimkin et al., 2013. G1E-ER4 cells were cultured as described previously, and RNA was extracted at 0, 3, 7, 14, 24, and 30 hours. Cultured human cord blood erythroblasts at distinct maturational stages were isolated by flow cytometry, as described. Total RNA was extracted using Qiagen RNeasy Kits.

RNA-sequencing analysis

Construction of complementary DNA libraries, sequencing, and bioinformatic analyses including lncRNA pipelines are provided in supplemental Methods.

ChIP-seq

The chromatin immunoprecipitation and massively parallel DNA sequencing (ChIP-seq) assay was performed as described.

Mouse strain RNA-seq comparison

BAM files for mouse splenic erythroblast RNA sequencing (RNA-seq) from 8 different strains were obtained from the authors of Hosseini et al, 2013. We assembled the transcriptomes using the TopHat and Cuffdiff tools in distinguishing between coding and noncoding genes. In total, we identified 1109 potential lncRNA genes (high stringency), including Xist, Tug1, Malat1, and LincRNA-EPS into the low-stringency group. This group likely represents a mixture of coding and lncRNA genes, reflecting the limitations of current bioinformatic tools in distinguishing between coding and noncoding genes. In total, we identified 1109 potential lncRNA genes (high + low stringency) transcribed during murine erythro-megakaryopoiesis (683 transcribed in erythroblasts), which constitute about 8.4% of genes producing polyA+ RNA transcripts in these cell types (Figure 1C). More than half of the identified lncRNAs were not annotated in public databases, including noncoding transcripts listed in RefSeq, UCSC, and Ensembl annotations (Figure 1D). Coding messenger RNAs (mRNAs) were expressed at higher levels than lncRNAs, as reflected by median fragments per kilobase per million mapped reads (FPKM) values (Figure 1E). As with coding mRNAs, lncRNAs showed clustered expression patterns of lineage specific protein coding genes confirmed the purity of input cells (supplemental Figure 1B).

Results

Novel cell-specific lncRNAs are expressed during mouse erythroid-megakaryopoiesis

We used RNA-seq to identify and compare lncRNAs in primary murine E14.5 fetal liver erythroblasts, megakaryocytes cultured from murine fetal liver progenitors, and MEPs from mouse bone marrow. Strand-specific, paired-end, deep sequencing was performed on polyA+ RNA from biological replicates of each sample (Figure 1A; supplemental Figure 1A). We assembled the transcriptomes using the TopHat and Cufflinks packages and generated a high-confidence transcriptome of 13131 genes expressed in at least 1 of the 3 cell types. Appropriate expression patterns of lineage specific protein coding genes confirmed the purity of input cells (supplemental Figure 1B).

We identified coding genes, pseudogenes, and potential small RNA precursors based on genomic overlap with existing annotations. We investigated the coding potential of the remaining genes using 3 independent algorithms: (1) protein database homology with BlastX; (2) codon conservation with PhyloCSF; and (3) presence of long open reading frames >100 amino acids with EMBOSS GetORF. Genes that had no evidence of coding potential based on all 3 criteria were classified as “high-stringency lncRNA” candidate genes and those that failed any 1 or more of them were classified as “low-stringency lncRNA” candidate genes. The coding potential tools had limited correlation with each other (Figure 1B), with numerous false positives and negatives occurring. For example, our pipeline assigned several known lncRNAs, including Xist, Tug1, Malat1, and LincRNA-EPS into the low-stringency group. This group likely represents a mixture of coding and lncRNA genes, reflecting the limitations of current bioinformatic tools in distinguishing between coding and noncoding genes. In total, we identified 1109 potential lncRNA genes (high + low stringency) transcribed during murine erythro-megakaryopoiesis (683 transcribed in erythroblasts), which constitute about 8.4% of genes producing polyA+ RNA transcripts in these cell types (Figure 1C). More than half of the identified lncRNAs were not annotated in public databases, including noncoding transcripts listed in RefSeq, UCSC, and Ensembl annotations (Figure 1D). Coding messenger RNAs (mRNAs) were expressed at higher levels than lncRNAs, as reflected by median fragments per kilobase per million mapped reads (FPKM) values (Figure 1E). As with coding mRNAs, lncRNAs showed clustered expression based on cell of origin (supplemental Figure 1D). However, ~30% of high-stringency lncRNAs were detectable in just 1 of the 3 cell types examined, contrasting with coding mRNAs, of which ~90% were detected in all 3 lineages (Figure 1F). This suggests that lncRNAs are more lineage-restricted than coding mRNAs and that RNA-seq analysis of previously uncharacterized cell types will identify unappreciated lncRNAs. The hematopoietic lncRNAs defined in this study segregated roughly...
equally into divergent transcripts with transcription start sites (TSS) within 300 bp of a coding gene TSS, antisense transcripts to coding genes, and intergenic transcripts (Figure 1G).

Because the erythroid RNA-seq was performed in mouse fetal liver erythroblasts, we compared lncRNA expression between fetal and adult erythropoiesis. Using a custom microarray with probes against the standard coding transcriptome and erythroid lncRNAs identified in this study, we probed gene expression in adult mouse bone marrow hematopoietic progenitors purified by fluorescent-activated cell sorting (supplemental Figure 2A). More than 85% of fetal liver erythroid lncRNAs were detected in adult erythroblasts. This rose to more than 90% when transcripts at the lowest quartile of expression (FPKM < 1) were excluded (supplemental Figure 2B). Thus, most mouse erythroid lncRNAs are transcribed from conventional gene promoters regulated by key hematopoietic transcription factors.

It is unclear whether most lncRNAs are transcribed from genes with conventional promoters or whether they are transcribed from enhancers. We investigated this by performing genome-wide ChIP-seq for H3K4me1, H3K4me3, and H3K36me3 histone marks on fetal liver erythroblasts and megakaryocytes (see Pimkin et al and Mouse ENCODE Consortium, manuscript in revision) and mapped these chromatin signatures to the genomic loci of coding and lncRNA genes. We classified gene TSSs with H3K4me3-high/H3K4me1-low signatures as promoters and those with H3K4me1-high/H3K4me3-low signatures as enhancers (Figure 2A). About 25% of the high-stringency lncRNA TSSs exhibited enhancer-like signatures, which was significantly greater ($P = 10^{-12}$ for erythroblasts,
for megakaryocytes by Fisher’s exact test) than coding genes (10%) (Figure 2B). Nonetheless, 75% of erythro-megakaryocytic lncRNAs were transcribed from regions with promoter-like signatures. Low-stringency lncRNAs arose from enhancers less frequently than high-stringency lncRNAs, but more frequently than coding genes, consistent with our interpretation that this group contains a mixture of protein coding and lncRNA genes.

LncRNA genes are likely regulated by cell-type–specific TFs, similar to coding genes, although this has yet to be demonstrated on a transcriptome-wide level in primary differentiated cells. We performed genome-wide ChIP-seq for GATA1 and TAL1 in erythroblasts and for GATA1, GATA2, TAL1, and FLI1 in megakaryocytes (see Pimkin et al., Wu et al., and Mouse ENCODE Consortium, in revision). We compared the binding of these TFs with genomic loci of coding and lncRNA genes that were up- or downregulated in erythroblasts or megakaryocytes compared with MEPs, as measured by RNA-seq. The expression of erythroid coding and noncoding genes in the Gata1+ mouse erythroid line G1E-ER4 after activation of an estrogen-induced form of GATA1.

To investigate further whether GATA1 regulates erythroid lncRNA genes directly, we examined their expression in G1E-ER4 cells, an immortalized Gata1+ erythroblast line expressing an estrogen-activated form of GATA1. Treatment of G1E-ER4 cells with β-estradiol triggers a GATA-1–regulated program of gene expression and terminal erythroid maturation. We performed RNA-seq at 0, 3, 7, 14, 24, and 30 hours after β-estradiol treatment and assessed the expression kinetics of genes that were upregulated in erythroblasts as compared with coding genes. Both coding genes and lncRNAs, genes that are upregulated during megakaryopoiesis and either downregulated or unchanged during erythropoiesis show higher levels of genomic locus heptad occupancy in progenitor cells. Dn, downregulated; N, no change; Up, upregulated.

Transcriptional regulation. High-stringency lncRNA genes downregulated during erythropoiesis showed a higher level of GATA1 and TAL1 occupancy compared with coding genes. The biological significance of this is unclear and will require further investigation.

To investigate further whether GATA1 regulates erythroid lncRNA genes directly, we examined their expression in G1E-ER4 cells, an immortalized Gata1+ erythroblast line expressing an estrogen-activated form of GATA1. Treatment of G1E-ER4 cells with β-estradiol triggers a GATA-1–regulated program of gene expression and terminal erythroid maturation. We performed RNA-seq at 0, 3, 7, 14, 24, and 30 hours after β-estradiol treatment and assessed the expression kinetics of genes that were upregulated in erythroblasts as compared with coding genes. Both coding genes and lncRNAs, genes that are upregulated during megakaryopoiesis and either downregulated or unchanged during erythropoiesis show higher levels of genomic locus heptad occupancy in progenitor cells. Dn, downregulated; N, no change; Up, upregulated.

10−14 for megakaryocytes by Fisher’s exact test) than coding genes (<10%) (Figure 2B). Nonetheless, ~75% of erythro-megakaryocytic lncRNAs were transcribed from regions with promoter-like signatures. Low-stringency lncRNAs arose from enhancers less frequently than high-stringency lncRNAs, but more frequently than coding genes, consistent with our interpretation that this group contains a mixture of protein coding and lncRNA genes.

LncRNA genes are likely regulated by cell-type–specific TFs, similar to coding genes, although this has yet to be demonstrated on a transcriptome-wide level in primary differentiated cells. We performed genome-wide ChIP-seq for GATA1 and TAL1 in erythroblasts and for GATA1, GATA2, TAL1, and FLI1 in megakaryocytes (see Pimkin et al., Wu et al., and Mouse ENCODE Consortium, in revision). We compared the binding of these TFs with genomic loci of coding and lncRNA genes that were up- or downregulated during erythropoiesis. Overall, the genomic loci of all 3 sets of genes were occupied by key hematopoietic TFs at similar frequencies, indicating shared modes of transcriptional regulation. High-stringency lncRNA genes downregulated during erythropoiesis showed a higher level of GATA1 and TAL1 occupancy compared with coding genes. The biological significance of this is unclear and will require further investigation.
We recently demonstrated that megakaryocyte-enriched coding genes, but not erythroid ones, are “primed” in hematopoietic stem/progenitor cells (HSPCs). Specifically, these genes are occupied in HSPCs by a TF heptad (Gata2, Runx1, Tal1, Fli1, Lmo2, Ldb1, Lyl1, and Erg) that facilitates low-level expression. These genes are subsequently activated during megakaryopoiesis and repressed during erythropoiesis, with concomitant lineage-specific alterations in TF occupancy (Figure 2E and Pimkin et al). To determine whether this mechanism applies to hematopoietic lncRNAs, we grouped them according to their expression patterns (megakaryocyte vs MEP and erythroblast vs MEP) and assessed occupancy of the TF heptad at the corresponding genes in the HSPC-like cell line HPC-7. For both coding genes and lncRNAs, the highest frequency of heptad occupancy occurred in those that are induced during megakaryopoiesis and either repressed or unchanged during erythropoiesis (Figure 2F). Thus, the model that some megakaryocyte-enriched genes are specifically primed in HSPCs extends to lncRNA loci, emphasizing common regulatory mechanisms with coding genes.

Overall, our findings demonstrate that most erythroid and megakaryocytic polyA-lncRNAs are transcribed from H3K4me3-rich gene promoters and are regulated by the same key TF networks as coding genes.

Mouse erythroid lncRNAs are poorly conserved in nucleotide sequence and expression in humans

LncRNAs expressed in multiple species are less conserved in primary nucleotide sequence than coding genes. We compared the erythroid transcriptome to a phastCons annotation of 5% of the mouse genome most conserved in vertebrates. More than half of the spliced mRNA length of coding genes overlapped with phastCons elements, reflecting DNA sequence conservation in multispecies alignments. The degree of overlap for lncRNA genes is greater than random, but lower than for coding genes. (B) PhastConsElement overlap of lncRNA genes subclassified by orientation relative to coding genes (see Figure 1G). (C) Percentages of mouse erythroid lncRNA genes are expressed in humans despite identified transmapping orthologous regions for most. (D) Percentages of human erythroid genes transmapping to the mouse genome and expressed in mouse erythroblasts showing low levels of conserved lncRNA expression. HS, high stringency; LS, low stringency.
most (~85%) coding genes expressed in mouse erythroblasts were also expressed in humans. In contrast, <20% of erythroid lncRNA genes expressed in mouse erythroblasts were expressed in matched human cells, even though orthologous regions were identified for ~90% of these genes (Figure 3C). Transmapping of the human erythroid transcriptome to the mouse genome showed similar results, with most lncRNA genes expressed in humans not being transcribed in mice (Figure 3D).

To investigate the possibility that this discrepancy between coding and noncoding genes is due to the low abundance of lncRNA transcripts compared with coding mRNAs, we separately assessed cross-species expression of transcripts with high or low abundance (supplemental Figure 3C-D). Even the most highly expressed lncRNAs (FKPM > 5) were largely not detected in the other species, even though orthologous loci could be identified. These findings demonstrate that most mouse erythroid lncRNA genes are not transcribed in human erythroblasts, and vice versa.

**Murine erythroid lncRNAs are conserved between different mouse strains**

We investigated the evolution of murine erythroid lncRNAs further by comparing RNA-seq data from splenic erythroblasts of 8 different mouse strains (A/J, AKR/J, BALB/c, C3H/HeJ, C57BL/6J, CBA/J, DBA/2J, LP/J). We quantified the erythroid transcriptome of each strain and examined the cross-strain expression of coding and lncRNAs. Most lncRNAs (~75%) were present in all mouse strains examined, though at a lower percentage than coding genes (~90%) (Figure 4A). On excluding transcripts with the lowest quartile of expression, more than 95% of high-stringency lncRNAs were expressed across mouse species, similar to that of coding genes (Figure 4B). In addition, the expression levels of all transcript types were highly concordant between different mouse strains (2-tailed Spearman’s R = 0.95 for coding mRNAs, >0.89 for low-stringency lncRNAs, and >0.80 for high-stringency lncRNAs).

**Potential roles for lncRNAs in murine erythroblast maturation**

We used RNAi to assess the functions of murine erythroid lncRNAs during in vitro maturation of primary murine erythroblasts. We examined candidate lncRNAs that were abundant (FKPM > 4) and erythroid-enriched (at least 4-fold greater expression in fetal liver erythroblasts than MEPs). We included low-stringency lncRNAs as candidates because this group contains some known lncRNAs. We selected 11 conserved (orthologous transcripts detected in human erythroblasts) and 16 nonconserved (orthologous transcripts not detected) lncRNA candidates (supplemental Table 2). Expression analysis using a custom microarray (supplemental Figure 2A-C, above) confirmed that these genes were also enriched in adult bone marrow erythroblasts and upregulated a median of 41-fold (range;
Figure 5. Mouse IncRNAs selected for RNAi studies. Twenty-seven IncRNAs were selected according to criteria described in the text. The relative expression of these IncRNAs in purified hematopoietic lineages from adult mouse bone marrow was determined using a custom oligonucleotide array and is shown graphically. “Conserved” refers to the presence of assembled orthologous transcripts detected in human erythroblasts (Figure 3C). Candidate erythroid IncRNAs shown here are upregulated 2- to 2054-fold (median 41-fold) in erythroblasts compared with multipotent progenitors. Names assigned to IncRNAs that potentially function in erythropoiesis according to RNAi studies in Figure 6 are indicated. Linc051 was previously named “Lincred1”. Ggbrp2os was not detectable in primary human erythroblasts, but transcription cannot be ruled out in human K562 erythroleukemia cells (supplemental Figure 8). CMP, common myeloid progenitor; EE, early erythroblast; Exp, absolute expression level by microarray; GMP, granulocyte macrophage progenitor; GR, granulocyte; HSC, hematopoietic stem cell; LE, late erythroblast.

<table>
<thead>
<tr>
<th>Knockdown</th>
<th>HSC</th>
<th>CMP</th>
<th>MEP</th>
<th>EE</th>
<th>LE</th>
<th>GMP</th>
<th>GR</th>
<th>Exp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conserved</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inc014</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inc013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inc023</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inc029</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inc015</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inc021</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inc009</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inc016</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inc017</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inc007</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Non-Conserved</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Inc036</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inc032</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inc040</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inc034</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inc031</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Expression (log2) | 15 | 13 | 11 | 9 | 7 | 5 | 3 | 1 | ne

Erythroid development is accompanied by downregulation of surface CD44 and reduced cell size (assessed by forward scatter). At 48 hours, knockdown of 2 IncRNAs (Erytha and Scarletl) showed reduced proportions of late stage (CD44low forward scatterlow) erythroblasts consistently using 2 different shRNAs that target different regions of the same transcript (Figure 6C-D; supplemental Figure 5). Thus, we have identified 7 previously uncharacterized IncRNAs with potential roles in terminal erythroid maturation.

We investigated the cellular localization of IncRNAs with the strongest knockdown phenotypes using RNA fluorescent in situ hybridization. Probe sets against 4 of 5 IncRNAs examined (Bloodlinc, Galont, Redrum, Lincred1) produced unambiguous signals in primary mouse fetal liver erythroblasts (Figure 6E). Single-slice confocal images were used to identify subcellular localization and composite images to assess overall IncRNA abundance. Probes against CyclinA2 mRNA used as a positive control showed cytoplasmic localization. Redrum RNA was largely cytoplasmic, whereas Bloodlinc and Galont were pancellular. Lincred1 was largely nuclear and displayed occasional bright sites of active transcription or accumulation. Our observation that individual IncRNAs exhibit distinct localization patterns suggests different mechanisms of action.

Functional mouse erythroid IncRNAs are not expressed in human erythroblasts

Six of the 7 erythroid IncRNAs with potential functions identified by RNAi were not expressed in human erythroblasts, even though orthologous genomic loci were identified by transmapping. We compared the transcriptional and epigenetic features at these loci in mouse and human erythroblasts (Figure 7; supplemental Figures 6-8). The mouse loci were marked by robust RNA-seq signals, indicating gene expression both by our erythroblast RNA-seq and ENCODE mouse fetal liver RNA-seq. They also showed marks of active chromatin and, in 5 of 7 cases, GATA1/TAL1 occupancy. In contrast, for 6 of these loci, no RNA-seq transcripts were evident in the mapped orthologous human loci by de novo RNA-seq assembly or by visual inspection of mapped reads of 5 distinct stages of human erythroblasts defined by flow cytometry. In some of these transcriptionally silent human loci, no marks of active chromatin were observed (see Figure 7A-B for Redrum, supplemental Figure 6A-B for Galont), whereas for others, histone marks and TF occupancy peaks were present without any mapped RNA-seq reads (see supplemental Figure 6C-D for Erytha, supplemental Figure 7A-B for Lincred1, supplemental
For Ggnbp2os (antisense to the coding gene Ggnbp2), no assembled transcript was identifiable in human erythroblasts, although human K562 RNA-seq tracks showed some reads mapping to the strand antisense to Ggnbp2, so human transcription cannot be ruled out. Thus, for 6 of 7 lncRNAs with potential functions in erythropoiesis indicated by RNAi, we found no evidence for expression of the orthologous human loci in multiple RNA-seq datasets derived from primary human erythroblasts, K562 erythroleukemia cells, or any other nonerythroid cell line or primary tissue analyzed by ENCODE (not shown). Our mouse and human RNA-seq datasets were of similar depth (more than 100 million reads per sample), and the lack of evidence for expression of these lncRNAs in human erythroblasts suggests that their functions in mouse erythroid development may be specific to the mouse species.
the lncRNAs were expressed relatively robustly in mouse erythroblasts (median FPKM 13, range 4 to 312). Therefore, failure to detect orthologous human transcripts cannot be attributed to differences in the depth of sequencing or low abundance.

**Discussion**

**RNA profiling of cell lines and primary tissues** has discovered thousands of mammalian lncRNAs. We identified 1109 candidate lncRNAs in erythroblasts, megakaryocytes, and MEPs, more than half of which were not annotated in gene databases. Given the exquisite lineage specificity of lncRNAs (this report and Calibi et al), it is likely that additional new ones will be discovered through transcriptome analysis of other specialized cell types. Our current findings identify a cohort of lncRNAs that potentially regulate erythropoiesis and also generate new insights into lncRNA biology in general.

Orthologous lncRNA genes that are expressed in multiple species are known to have less conservation of primary nucleotide sequence compared with coding genes. By directly comparing transcriptomes, we demonstrate an even more radical lack of evolutionary conservation. Specifically, more than 80% of mouse erythroid lncRNAs, including the most highly expressed ones, are not detected in human erythroblasts of similar developmental and maturational stages, and most human erythroid lncRNAs are not detected in mice. This feature of lncRNA evolution is noted in a few studies, but not currently appreciated or tested systematically. Moreover, our findings are consistent with findings from the Mouse ENCODE Consortium (manuscript in revision) that lncRNAs are far more species-specific than coding genes. The ENCODE group compared RNA-seq datasets predominantly derived from murine primary cells against human cell lines. Our approach refines and extends these general comparisons by focusing on a set of matched cell types generated from primary tissues and demonstrates that human and mouse tissues produce largely distinct repertoires of IncRNAs.

Few studies compare the transcriptomes of matched cells from different species to investigate the actual expression of orthologous lncRNA genes. To our knowledge, ours is the first study to perform de novo assembly of IncRNAs in matched hematopoietic cell types (erythroblasts) of mice and humans, compare the transcriptomes, and demonstrate lack of cross-species transcription definitively for most IncRNAs.

It is possible that different lncRNAs have converged in evolution to regulate common sets of genes in different species. However, recent studies reveal a surprising number of differences in the levels and stage specificity of protein-coding mRNAs between human and mouse erythroblasts (Patrick Gallagher, X.A., and N.M., manuscript submitted January 2014). These findings may arise at least partly from interspecies differences in IncRNAs that regulate gene expression. It should be possible to test this as sets of murine and human IncRNA regulated genes become defined in the future.

There is lively debate over what proportion of the lncRNA transcriptome mediates biological functions as opposed to representing inert “transcriptional noise.” Our finding that knockdown of 7 lncRNAs impairs the maturation of primary erythroblasts provides preliminary evidence that at least a subset may have biological functions, in agreement with recent studies. Indeed, 4 of the lncRNAs implicated in erythroid maturation by our RNAi studies (Bloodline, Galont, Redrum, Lincred1; Figure 6) were also found by Alvarez-Dominguez et al to be functional using independent sets of shRNAs in a similar assay. These well-controlled RNAi experiments identify overlapping sets of candidate lncRNAs that may facilitate erythropoiesis. However, their biological functions must now be determined definitively through additional
investigations, including in vivo gene ablation and identification of relevant interacting genes and/or proteins.

Six of the 7 lncRNAs identified by the current study to have potential roles in erythropoiesis showed no evidence of transcription in human erythroblasts, indicating that conserved interspecies expression of lncRNAs does not necessarily correlate with function. In agreement, Braveheart, an lncRNA required for cardiac development in mice, is not expressed in rats or humans.\textsuperscript{30} We extend this example by reporting multiple other lncRNAs that potentially function in tissue development and yet are not expressed across mammalian species. New RNAs form constantly during evolution\textsuperscript{40} and can arise even from a single nucleotide replacement.\textsuperscript{41} The evolution of novel proteins from these transcripts likely requires a relatively stringent set of criteria to allow the formation of a functional peptide chain. In contrast, functional lncRNAs likely evolve relatively rapidly through the formation of nucleotide permutations that form protein- and nucleic acid–interacting motifs.\textsuperscript{32} Our data, if combined with RNA-seq from appropriately matched erythroblasts of other mammals, may illuminate further the evolution of these lncRNAs and stratify their ancestry along the 70-million-year timeline since human and mouse ancestors diverged.

Acknowledgments

The authors thank Moran Cabili for custom PfamAnalysis scripts, Doug Higgs for helpful access and discussions to unpublished data, Mike Lin for assistance with PhyloCSF, Merav Socolovsky for guidance with flow cytometry analysis, and Harvey Lodish and members of his laboratory, Wenqian Hu and Juan Alvarez-Dominguez, for helpful discussions.

This work was supported by the American Society of Hematology (ASH) Research Training Award for Fellows (V.R.P.); ASH Scholar Award (V.R.P.); and P30DK090969 (M.J.W.)

References


Lineage and species-specific long noncoding RNAs during erythro-megakaryocytic development

Vikram R. Paralkar, Tejaswini Mishra, Jing Luan, Yu Yao, Andrew V. Kossenkov, Stacie M. Anderson, Margaret Dunagin, Maxim Pimkin, Megheel Gore, Diana Sun, Neeraja Konuthula, Arjun Raj, Xiuli An, Narla Mohandas, David M. Bodine, Ross C. Hardison and Mitchell J. Weiss