Identification of Novel Long Non-Coding RNAs Underlying Vertebrate Cardiovascular Development

Running title: Kurian et al.: Transcriptome of human cardiovascular development

Leo Kurian, PhD\textsuperscript{1,2}\textdagger; Aitor Aguirre, PhD\textsuperscript{1,2}\textdagger; Ignacio Sancho-Martinez, PhD\textsuperscript{1}\textdagger; Christopher Benner, PhD\textsuperscript{3}; Tomoaki Hishida, PhD\textsuperscript{1}; Thai B. Nguyen, BsC\textsuperscript{1,2}; Pradeep Reddy, PhD\textsuperscript{1}; Emmanuel Nivet, PhD\textsuperscript{1,5}; Marie N. Krause, MD\textsuperscript{1}; David A. Nelles, BsC\textsuperscript{2}; Concepcion Rodriguez Esteban, PhD\textsuperscript{1}; Josep M. Campistol, MD\textsuperscript{4}; Gene W. Yeo, PhD\textsuperscript{2}; Juan Carlos Izpisua Belmonte, PhD\textsuperscript{1}

\textsuperscript{1}Gene Expression Laboratory, Salk Institute for Biological Studies, La Jolla, CA; \textsuperscript{2}University of California San Diego, Dept of Cellular and Molecular Medicine, Stem Cell Program, Institute for Genomic Medicine, Sanford Consortium for Regenerative Medicine, La Jolla, CA; \textsuperscript{3}Integrative Genomics Core, Salk Institute for Biological Studies, La Jolla, CA; \textsuperscript{4}Hospital Clinic of Barcelona, Barcelona, Spain; \textsuperscript{5}Present address: Laboratory for Developmental and Regenerative RNA Biology, University of Cologne, Center for Molecular Medicine Cologne, CECAD and Institute for Neurophysiology, Cologne, Germany; \textsuperscript{6}Present address: University of California San Diego, School of Medicine, Dept of Pediatrics, Genome Information Sciences Division, La Jolla, CA; \textsuperscript{7}Present address: Aix Marseille Université, CNRS, NICN UMR 7259, Marseille, France

*These authors contributed equally

Address for Correspondence:
Juan Carlos Izpisua Belmonte, PhD
Salk Institute for Biological Studies
10010 North Torrey Pines Road
La Jolla, CA 92037
Tel: 858-453-4100 x1130
Fax: 858-453-2573
E-mail: belmonte@salk.edu

Journal Subject Codes: Basic science research:\textsuperscript{[129]} Angiogenesis, Vascular biology:\textsuperscript{[97]}
Other vascular biology, Vascular biology:\textsuperscript{[98]} Other research
Abstract

**Background**—Long non-coding RNAs (lncRNAs) have emerged as critical epigenetic regulators, with important functions in development and disease. Here we sought to identify and functionally characterize novel lncRNAs critical for vertebrate development.

**Methods and Results**—By relying on human pluripotent stem cell differentiation models, we investigated lncRNAs differentially regulated at key steps during human cardiovascular development with a special focus on vascular endothelial cells. RNA-sequencing led to the generation of large datasets that serve as a gene expression roadmap highlighting gene expression changes during human pluripotent cell differentiation. Stage-specific analyses led to the identification of three previously uncharacterized lncRNAs: TERMINATOR, ALIEN and PUNISHER specifically expressed in undifferentiated pluripotent stem cells, cardiovascular progenitors and differentiated endothelial cells, respectively. Functional characterization including localization studies, dynamic expression analyses, epigenetic modification monitoring and knock-down experiments in lower vertebrates, as well as murine embryos and human cells, confirmed a critical role for each lncRNA specific for each analyzed developmental stage.

**Conclusions**—We have identified and functionally characterized three novel lncRNAs involved in vertebrate and human cardiovascular development, and provide a comprehensive transcriptomic roadmap that sheds new light on the molecular mechanisms underlying human embryonic development, mesodermal commitment and cardiovascular specification.

**Key words:** long-non coding RNAs, cardiovascular, transcriptomic, vertebrate, developmental biology
Introduction

The advent of novel sequencing technologies has revealed that less than 2% of the human genome encodes for proteins\textsuperscript{1–3}. Interestingly, despite not being translated into proteins, a large fraction of the mammalian genome is transcribed into what is known as non-coding RNAs (ncRNAs)\textsuperscript{1–3}. The consistent observation of pervasive transcription originating from these non-coding sites suggests that ncRNAs might play a key role in fundamental biological functions\textsuperscript{1–3}. To date, thousands of ncRNAs have been putatively described. Yet, the precise functional roles of the vast majority remain unclear\textsuperscript{4–6}.

Among the different classes of ncRNAs, long non-coding RNAs (lncRNAs), broadly defined as those non-coding transcripts larger than 200 nucleotides\textsuperscript{1}, represent one of the largest and least understood nucleic acid molecules in vertebrates. Overall, more than 9,000 genomic loci are predicted to code for these transcript subclasses in the human genome\textsuperscript{5,7,8}. Recent reports describing important developmental roles for certain lncRNAs in various vertebrates have appeared over the last few years\textsuperscript{9–11}. Low expression levels and poor sequence conservation have limited the identification and functional characterization of novel lncRNAs\textsuperscript{12}. Combining novel sequencing and analytical technologies with the powerful platforms that pluripotent stem cell (PSC) models provide for the study of human embryonic development, these limitations can now be circumvented.

By using a methodology recently developed in our laboratory that allows for the highly efficient generation of human cardiovascular progenitor cells of mesodermal origin and terminally differentiated vascular endothelial cells\textsuperscript{13}, here we report on the identification and functional characterization of three novel human lncRNAs indispensable for the development of the cardiovascular system in different vertebrate species.
Materials and methods

Cell culture

Human Embryonic Stem Cells (hESCs), H1 (WA1, WiCell) (passage 25-40) were used for differentiation studies. Terminally differentiated primary Human Umbilical Vein Endothelial Cells (HUVECs) were used to investigate the transcriptome and functional roles of lncRNAs candidates in unmodified vascular endothelial cells. Briefly, hESCs were cultured in chemically defined growth media, mTeSR (StemCell technologies), on growth factor reduced matrigel (BD biosciences) coated plates. 70-80% confluent hESCs were treated with dispase (Invitrogen) for 7 minutes at 37°C and the colonies were dispersed to small clusters and lifted carefully using a 5ml glass pipette, at a ratio of ~1:6. HUVECs were purchased from Promocell and cultured in Endothelial Basal Medium (EBM) supplemented with Endothelial Growth Medium (EGM) (Lonza). hESC-derived endothelial cells were cultivated in EGM-2 bullet kit media (Lonza). Endothelial cells were grown on collagen I coated plates (BD biosciences). All cell lines were maintained in an incubator (37°C, 5% CO2) with media changes every day (hESCs) or every second day (HUVEC).

Cell lineages analyzed

Pluripotent stem cells were differentiated into early mesoderm-derived cells including c-Kit+ and KDR+ cardiovascular progenitors (day 2 to day 4 of differentiation) and committed CD34+CD31+ bi-potent vascular progenitor cells (day 4 to day 8 of differentiation) as previously described (please refer to reference 13 by Kurian et al., for detailed characterization of the differentiation methodologies and cell types generated). Terminally differentiated primary HUVECs were also included in our analyses as positive controls and no significant differences between differentiated endothelial cells and HUVECs were found13. Briefly, for the generation of
early cardiovascular progenitors (day 2 to day 4) and committed vascular progenitor cells (day 4
to day 8) undifferentiated hESCs were freshly split on to matrigel-coated plates, making sure the
sub-colonies were of small size (~300-500 cells/colony). Cells were differentiated to different
progenitor stages using a chemically defined mesodermal induction media (MIM) as previously
described\(^\text{13}\) (DMEM:F12, 15mg/ml stem cell grade BSA (MP biomedicals), 17.5μg/ml Human
Insulin (SAFC bioscience), 275μg/ml Human holo transferrin (Sigma Aldrich), 20ng/ml bFGF
(Humanzyme), 50ng/ml Human VEGF-165 aa (Humanzyme), 25ng/ml human BMP4
(Stemgent), 450μM monothioglycerol (Sigma Aldrich), 2.25mM each L-Glutamine and Non-
essential amino acids (Invitrogen). Switching day 8 MIM-differentiated cells to EGM2 media as
described\(^\text{13}\) induced endothelial cell differentiation. HUVEC cells are of primary origin and were
therefore not modified during our studies.

**RNA-sequencing**

Total RNA from roughly 1x10\(^7\) cells of each of the above-indicated groups was isolated with
TRIzol (Invitrogen). Intact total RNA samples were treated with DNase1 and sent to the Beijing
Genomics Institute (BGI) for sequencing. Prior to sequencing, all samples were subjected to
quality control processes to ensure the lack of contaminating DNA and integrity of the RNA. All
the RNA samples met the following RNA quality threshold- OD260/280 = 2-2.2; OD260/230 ≥ 2.0;
28S:18S> 1.0, RIN>7. Whole transcriptome sequencing was then performed. Briefly,
TruSeq Stranded Total RNA with Ribo-Zero Human kit (Illumina) was used to remove
ribosomal RNA and prepare strand specific paired end RNA seq libraries. 90 million 2x 100 bp
paired end reads were sequenced on an Illumina HiSeq2000 instrument for each library. Reads
were aligned to the hg19/GRCh37 version of the human genome by using STAR (\textit{v}2.1.4a). Only
reads that aligned to a single unique location were kept for further analysis. Quantification of
reads on each strand in 10kb windows across the entire genome and comparisons with DNA sequencing data were utilized to further evaluate RNA purity and integrity (Figure 1 in the online-only Data Supplement).

**Primers used in this study**

Please refer to Table 1 in the online-only Data Supplement for a list of primers and their respective sequences.

**Morpholino embryo injections**

Morpholinos (Gene Tools) were designed against highly conserved regions in the lncRNAs, or alternatively to block putative splice sites (Table 2 in the online-only Data Supplement). A pair of morpholinos was generated and tested per lncRNA. For zebrafish injection, morpholinos were dissolved in water at a 2mM stock concentration and diluted to a 2ng/ml working concentration in PBS/phenol red solution. Embryo injections were performed by injecting ~1nl morpholino solution at the 1-cell stage using a FemtoJet (Eppendorf). Morphants were evaluated at 24, 48, 72 and 96 hours in a StereoLumar stereoscope (Zeiss). For murine experiments, CD-1 female mice were superovulated by injecting pregnant mares serum gonadotropin (PMSG) followed by human chorionic gonadotropin (hCG) 48 hours later. The female mice were housed with males after hCG injection. 1-cell stage embryos were collected the next day after 20 hours and were injected with 10pl of 2nM stock Terminator morpholino using a FemtoJet. The embryos were cultured until the blastocyst stage in K-modified simplex optimized medium (KSOM) and RNA was extracted using RNeasy mini kit (Qiagen).

**Data availability**

All datasets in this study are available in the NCBI Gene Expression Omnibus under the following accession number: GSE54969.
Statistical evaluation

Statistical analyses of all endpoints were performed by statisticians at the Salk Institute and UCSD by using Excel, SPSS and/or GraphPad software. All data presented a normal distribution. Statistical significance was evaluated using standard unpaired Student’s t test (two-tailed, p<0.05) when appropriate. For multiple comparison analysis, one-way Analysis of the Variance (ANOVA) with Dunnett's correction post-test was applied when appropriate (p<0.05). Comparisons of groups with small sample size (n<6) were performed as follows: i) Mann-Whitney test (two-sided, 95% confidence level; p<0.05) was used; ii) Kruskal-Wallis test with Dunn’s post-test (p<0.05) has been applied for multiple comparisons. All data are presented as mean ± standard deviation (s.d) and represent a minimum of three independent experiments with at least three technical replicates, unless otherwise stated.

Results

Non-coding RNAs account for half of the transcriptome during human vascular differentiation.

With the goal of identifying and functionally validating novel genetic elements underlying cardiovascular vertebrate development we first decided to perform an RNA-sequencing-based screening of transcripts differentially regulated during stem cell differentiation. We focused on our recently reported methodology for the efficient differentiation of human pluripotent cells to early c-Kit+ cardiovascular progenitors, committed CD34+CD31+ vascular progenitors and terminally differentiated functional vascular cells13. We selected five different stages: undifferentiated human embryonic stem cells (hESCs), mesoderm committed cells (day 2, KDR+), early cardiovascular progenitor cells (day 4, c-Kit+), committed vascular progenitor
cells (day 8, CD34+CD31+) and vascular endothelial cells (VE-Cadherin+Endoglin+) (please refer to Kurian et al.\textsuperscript{13} for detailed characterization of the generated cell lineages) (Figure 1A in the online-only Data Supplement). After ribosomal RNA depletion and size removal of species whose length was smaller than 150 nucleotides, we generated datasets with a sequencing depth of 90 million paired end reads of 100 base pairs per sample (Figure 1A,B; Figure 2B,C and Table 3 in the online-only Data Supplement). To provide an overall estimation of the transcriptome changes occurring during the differentiation of pluripotent cells to vascular progenitor cells we first decided to collectively assess global expression changes in all cell lineages, including early c-Kit+ cardiovascular cells appearing from day 2 to 4 during the course of differentiation as well as committed vascular progenitor cells (day 4 to 8)\textsuperscript{13}. In order to describe the global transcriptomic changes across the genome we investigated the genomic coverage of RNA-seq data (at least one read), all different cell types indicated that ~56% of the genome was transcriptionally activated during vascular progenitor cell differentiation\textsuperscript{14-17} (Figure 2B in the online-only Data Supplement). Over 25,000 transcripts were expressed upon differentiation (Reads Per Kilobase per Million [RPKM] > 0.1), with 13,796 transcripts exhibiting at least 3-fold changes in expression. Among these 13,796 transcripts, a total of 44% represented non-coding RNAs (Figure 1A and Figure 2B in the online-only Data Supplement). After excluding repetitive sequences, we identified a total of 406 novel transcripts that were not annotated previously in any database (83 were promoter anti-sense, 188 were intergenic and 135 were anti-sense gene sequences)\textsuperscript{14-17}. By considering both, annotated and non-annotated transcripts, we observed that one of the largest groups differentially regulated during differentiation was that of lncRNAs (~1924). These observations therefore suggest that lncRNAs have a functional involvement during mesoderm development and cell fate specification\textsuperscript{5,9,10,18}. 


Non-coding RNA regulation accounts for the majority of transcriptomic changes regulating lineage specification.

Next, we focused our attention on the analyses of each respective cell lineage generated during differentiation. Similar to recent reports\textsuperscript{19-20}, we validated the robustness of our system by focusing our attention on genes characteristic of each of the different stages analyzed: pluripotent cells (day 0) (\textit{POU5F1 (Oct4), NANOG, SOX2, ZFP42})\textsuperscript{13,19}, mesoderm-committed cardiovascular progenitor cells (day 2 to day 4) (\textit{T (Brachyury), MSX1, GSC, EOMES, WNT3A, SNAI2, EVX1})\textsuperscript{9,10,13}, committed vascular progenitor cells (day 4 to day 8) (\textit{GATA1, GATA2, LMO2, ETS1, HOXB4})\textsuperscript{13,21} and endothelial cells (\textit{CDH5, VWF, EPHB2, NRF2F2, ENG})\textsuperscript{13} (Figure 1C). As expected, all genes appeared upregulated at the corresponding developmental stages (Figure 1C). Gene Ontology analysis highlighted the expression of key genetic circuitries driving cardiovascular commitment (upregulated at day 2) as well as blood vessel and heart development (upregulated from day 2 to day 8) (Figure 1D). These results are in agreement with previous reports and further demonstrate that early cardiac and vascular developmental programs share common genetic pathways prior to further cell type specification in humans\textsuperscript{21,22}. Chromatin modification analysis identified an increasing number of bivalent sites as differentiation progressed, for both protein-coding as well as for lncRNA sites (Figure 1E and Figure 2D in the online-only Data Supplement). Differentiation to endothelial cells resulted in the up-regulation of >1000 genes and the downregulation of ~2500 genes (Figure 1B). Genome-wide methylation profiling highlighted significant hypomethylation at lncRNA sites (19%, Figure 2E in the online-only Data Supplement). Together, these results suggested an important regulatory role for lncRNAs during lineage specification (as compared to, for example, 5% for microRNAs, Figure 2E in the online-only Data Supplement).
Identification of non-coding RNAs differentially regulated during human Pluripotent Stem Cell differentiation to vascular endothelial cells.

To identify enriched DNA motifs acting as regulatory elements, we made use of HOMER, an analytical tool suitable for promoter enrichment analyses\(^\text{15}\). Promoter motif enrichment (-300, +50 bp from the transcriptional start site) revealed that transcripts differentially expressed during differentiation harbor binding sites for the major transcriptional networks regulating vascular development as compared to non-regulated promoters (Figure 1F). Key developmental drivers differentially regulated in a time-dependent manner included members of the homeobox gene family, mesodermal regulators (EOMES, GATA, SMAD) and vascular transcription factors (ETS, HIF) (Figure 1F). Interestingly, IncRNAs comprised the largest class of non-coding transcripts subjected to strict temporal regulation patterns during differentiation. These results are in line with the notion that IncRNAs are more tightly controlled in terms of timing and cell type specificity than protein-coding transcripts\(^\text{18,23}\). We then selected 4 distinct groups of IncRNAs, corresponding to the 4 top-level clusters, highlighted by hierarchical clustering (Figure 3A in the online-only Data Supplement). These 4 distinct groups had the highest average dissimilarity between one another and correlated with the different developmental stages analyzed, with ~300 IncRNAs expressed specifically in pluripotent stem cells, ~100 in mesodermal progenitors (day 0 to day 2), ~250 in early cardiovascular progenitor cells (day 2 to day 4), ~550 in late vascular progenitors (day 4 to day 8) and lastly, ~200 in differentiated endothelial cells (Figure 3A in the online-only Data Supplement). Furthermore, anti-sense IncRNAs (AS-IncRNAs) expression demonstrated a positive correlation with the respective sense protein-coding counterparts in many cases (Figure 3B in the online-only Data Supplement). We next employed a stringent five-step filtering process at each indicated differentiation stage to identify novel IncRNAs, whose
sequence and functionality might be conserved across different vertebrate species (Figure 3A in the online-only Data Supplement). To do this we focused on the following parameters: a) cell stage specific patterns of expression distinguishing pluripotent cells, cardiovascular progenitor cells, and terminally differentiated endothelial cells in multiple biological replicates (at least 3); b) significant sequence conservation across different vertebrate species; c) exon-intron structure data; d) availability of Expressed Sequenced Tags (ESTs) in human, mouse and zebrafish to evaluate coding potential and facilitate expression analyses; e) physical location in close proximity to known cardiovascular regulatory elements. Our analyses highlighted a total of 75 lncRNAs that successfully fulfilled the first 3 criteria (i.e.: specific expression, exon-intron structure, partial sequence conservation across vertebrates) and at least one of the latter (e.g. EST data in lower vertebrates). Next, we randomly selected three previously uncharacterized lncRNAs specifically expressed in pluripotent stem cells (TERMINATOR), vascular progenitors (ALIEN) and mature endothelial cells (PUNISHER) for further characterization (names were retrospectively assigned based on the phenotypes elicited in zebrafish). Non-coding RNA identification relies, to a great extent, on known sequence-defining gene characteristics (e.g.: well-defined promoters, exon-intron structure codon conservation and ribosomal footprints, among others), as well as a lack of coding potential⁶,²⁴. None of these transcripts possessed coding potential, as determined by GeneID analysis²⁴ (Figure 3C in the online-only Data Supplement), as well as a lack of identifiable amino acid domains and lack of codon conservation across evolution or ribosomal footprints (data not shown).

**Characterization of three novel lncRNAs differentially expressed during endothelial cell differentiation.**

Next, we sought to determine the functionality of the three different lncRNAs selected. To do so,
we first focused on localization studies in the different cell lineages in where the three lncRNAs were identified as differentially regulated (pluripotent cells, cardiovascular progenitors and endothelial cells). Fluorescence In-Situ Hybridization (FISH) experiments revealed preferential expression of \textit{TERMINATOR} in the nucleus of pluripotent stem cells, whereas nuclear, perinuclear and cytosolic localization of \textit{ALIEN and PUNISHER} was found in cardiovascular progenitors and endothelial cells, respectively (Figures 2A,B). lncRNAs have been demonstrated to play important roles in gene regulation during cell fate specification and development\textsuperscript{3,9,10,25,26}. To comprehensively characterize the gene networks associated with the different lncRNAs, we performed Pearson correlation analyses where protein-coding mRNA expression was systematically evaluated and associated with each of the three different lncRNAs. Upon the coefficients of correlation we next focused exclusively on those transcripts with a correlation coefficient $> 0.85$ for further gene ontology analyses. Expression of \textit{TERMINATOR} correlated positively with genes involved in cell cycle, DNA repair and chromatin assembly while negatively with genes involved in cell death and regulation of proliferation (Figure 2C and Figure 4A,B in the online-only Data Supplement). \textit{ALIEN} expression correlated positively with transcripts involved in skeletal muscle development, heart morphogenesis and tube formation and correlated inversely with cell adhesion, membrane transport and neural function related genes (Figure 2C and Figure 4A,B in the online-only Data Supplement), suggesting that \textit{ALIEN} might possess a functional role during early cardiovascular development prior to vascular specialization. Last, \textit{PUNISHER} demonstrated a positive correlation with genes participating in definitive vascular development while negatively correlating with cell cycle regulators, chromatin modifiers and DNA damage response genes (Figure 2C and Figure 4A,B in the online-only Data Supplement). Next, we performed RNA
immunoprecipitation coupled to mass spectrometry analysis (Table 4 in the online-only Data Supplement). Protein complex analysis upon pull-down led to the identification of proteins involved in RNA binding\textsuperscript{27}, post-transcriptional control and epigenetic remodeling\textsuperscript{28} (Figure 4C and Table 4 in the online-only Data). Spring-embedded algorithms (Figure 2D) revealed hits closely correlating with each lncRNA. TERMINATOR expression associated with POU5F1 (Oct4), SOX2, ZIC5 and REX1, all of them known regulators of pluripotency. ALIEN demonstrated a high degree of correlation with pivotal drivers of mesoderm and cardiovascular commitment, including \textit{T} (Brachyury), EOMES, MIXL1, and GATA4 (Figure 2D). Last, expression of transcription factors essential for endothelial cells, such as TAL1 and FOXC1, correlated with PUNISHER (Figure 2D). Together, these results indicate a stage-specific function for each of the different lncRNAs in regulating gene expression. Additionally, none of the identified lncRNAs were physically associated to polypeptides of the ribosome translational machinery. Thus, this confirms the non-coding nature of the selected transcripts. Novel lncRNAs functionally control pluripotency, cardiovascular commitment and endothelial cell identity.

In order to gain insights into the physiological relevance of the identified lncRNAs, we first evaluated their expression pattern during mouse and zebrafish embryonic development. In accordance with their expression during human vascular differentiation, \textit{Terminator} expression was maximal at the blastocyst stage in mouse and at 6 hours post-fertilization (hpf) in zebrafish, closely correlating with the expression of Pou5f1 (Oct4) and Nanog (Figure 3A and Figure 2C and Figure 5 in the online-only Data Supplement). \textit{terminator} levels were undetectable in 1-cell stage zebrafish embryos immediately after fertilization, suggesting that Terminator was not already present in the oocyte prior to fertilization (Figure 3A). \textit{Alien} was expressed in the
allantois and lateral plate mesoderm of E8.5 mouse embryos and 12 hpf zebrafish embryos (Figure 3B), correlating with the expression of T (Brachyury), Meox1 and Mix1 (Figure 5 in the online-only Data Supplement). The endothelial cell specific lncRNA Punisher showed the highest levels of expression in E12.5 mouse embryos and at 72 hpf in zebrafish, once the vasculature was formed (Figure 3C and Figure 5 in the online-only Data Supplement). PhyloP analysis for detailed evolutionary conservation revealed relatively short, highly conserved sequence regions of 250-500 bp (>90-95%) across vertebrates29 (Figure 3D and Table 5 in the online-only Data Supplement). Additionally, as previously mentioned, there were described ESTs transcribed from the conserved loci in mouse as well as in zebrafish. TERMINATOR was found to be an intergenic lncRNA located next to ZNF281 in human and mouse (in both cases in chromosome 1). It presents H3K27Ac sites in start and end regions (marking actively transcribed chromatin) and its genomic location is conserved in sense. ALIEN represents a subclass of lncRNAs, a lincRNA (long interspersed non-coding RNA). It was found located in an intergenic region with low gene density. It is encoded in sense with exons and its location is conserved next to FOXA2 in mouse, human and zebrafish. Last, PUNISHER was found to be an antisense lncRNA covering exonic and intronic sequences of the AGAP2 gene. Its location is conserved in mouse and human and it is transcribed as an antisense to AGAP2 and it presents H3K27Ac sites. Taken together, our results indicate that these three lncRNAs are conserved from zebrafish to human, and show similar stage-specific expression during development.

To further characterize the functional roles of the different lncRNAs during vertebrate development we designed a series of experiments in different vertebrate systems including zebrafish and mammalian murine embryos. Anti-sense morpholino (MO) oligonucleotides, synthetic molecules used to sterically block RNA binding sites in the absence of degradation,
targeting highly conserved regions or putative splice sites in the identified lncRNAs were designed to induce specific knockdown or functional blockade\textsuperscript{30,31}. Loss-of-function experiments were first performed in the transgenic zebrafish strains \textit{flil}:GFP (vascular reporter) and \textit{cmlc2}:GFP (cardiac reporter) (Figure 4A) with specific anti-sense MO, and compared to the corresponding non-targeting sequences used as negative controls. MO injections targeting \textit{terminator} compromised development at the gastrulation stage and resulted in $>50\%$ lethality whereas the surviving embryos exhibited developmental arrest and severe cardiovascular defects (Figure 4B-D; Tables 6 and 7 in the online-only Data Supplement). In addition, delayed epiboly stages and detachment of the cell mass from the yolk at 5-7 hpf could be observed early during development (Figure 6 in the online-only Data Supplement) in line with \textit{terminator}'s peak of expression at 6 hpf. MO-mediated loss of function of \textit{alien} led to severe impairment in multiple anatomical structures, highlighting the specific role of Alien in mesodermal specification. Among other mesoderm-related defects, \textit{alien} inhibition resulted in defective vascular patterning, with pronounced branching defects abrogating the correct formation of dorsal and intersegmental blood vessels and defective cardiac chamber formation, demonstrating that Alien specifically functions in an early developmental progenitor stage common to both the vascular and cardiac lineages (Figure 4B-D; Tables 6 and 7 in the online-only Data Supplement). Inhibition of \textit{punisher} also resulted in severe defects in the vasculature, including defective branching and compromised vessel formation (Figure 4B-D; Tables 6 and 7 in the online-only Data Supplement). Abrogation of Punisher activities demonstrated significant changes in vessel tube number, length and severely impaired cardiac development and function (Figures 4E-G; Tables 6 and 7 in the online-only Data Supplement). To confirm the \textit{in vivo} role of Terminator and Punisher, we performed rescue experiments. Co-injection of mature human lncRNA sequences
alongside the respective MOs sufficed for rescuing the observed phenotypes with a significant effect in embryo survival, animal morphology as well as branching and development of the cardiovascular system (Figure 7 and Table 8 in the online-only Data Supplement). It should be noted that zebrafish MOs could not efficiently target the human lncRNA counterparts because of sequence divergence. Therefore, these rescue experiments further highlighted a qualitatively specific functional role for each of the different lncRNAs. Together, our results demonstrated a specific function for the identified lncRNAs in the differentiation of pluripotent cells to mesoderm and further specification of cardiovascular followed by vascular and endothelial cell specification in vertebrates.

Identified lncRNAs demonstrate a critical role at specific stages during mammalian development.

Last, to provide an overview on the role that the identified lncRNAs play in mammalian development, we evaluated the role of all three lncRNAs in mammals by loss-of-function experiments in mouse embryos and human cells. To do so, three independent biological replicates were pooled and prior microarray hybridization. Further analysis indicated that TERMINATOR blockade in mouse embryos and human ESCs resulted in the downregulation of the pluripotency factors POU5F1 (Oct4), SOX2 and NANOG at the blastocyst stage and led to significant cell death (Figure 5A-E). Downregulation of TERMINATOR resulted in the upregulation of 506 genes and the downregulation of 185 different genes involved in cell-cell interactions and chromatin remodeling necessary for the maintenance of a pluripotent state (Figure 6). ALIEN knockdown resulted in the significant upregulation of 738 genes involved in cell adhesion and extracellular matrix remodeling while downregulation of 503 genes related to angiogenesis and blood vessel development could be observed (Figure 6). Last, PUNISHER
knockdown resulted in profound gene expression changes in endothelial cells, with 802 genes involved in mitosis and cell division being downregulated and 831 genes involved in cell adhesion and extracellular interactions being upregulated (Figure 6). PUNISHER knockdown resulted in decreased histone H3 phosphorylation (Figure 7A), a marker indicative of mitosis, and impaired human vessel maturation (Figures 7B,C). PUNISHER knockdown also impaired acetyl-LDL uptake, a hallmark of endothelial cell functionality, therefore indicating severely impaired endothelial cell function as compared to scrambled shRNA controls and unmodified endothelial cells (Figure 7D).

Discussion

Recent discoveries indicating that the mammalian transcriptome is comprised of a large majority of non-coding transcripts (~60%) as opposed to coding RNAs (~7%) have brought about the question as to whether non-coding sequences can play a role in controlling cellular fate and functionality, ultimately affecting biological diversity25,32–34. Despite major efforts, little information is available regarding the actual cellular functions of ncRNAs in the context of human development, physiology and disease9,10,19,35,36. Among the existing information, lncRNAs identified in pluripotent cells have been mainly ascribed to gene-network interactions, regulation of chromatin and the control of the pluripotency network19,33,35. Similarly, other recent reports have focused on the identification of lncRNAs during cardiomyocyte generation and identified Braveheart and Fenderr as critical players during the development of the early cardiovascular system and ultimately, the murine heart9,10.

Pluripotent cell differentiation has been demonstrated to be a suitable platform for recapitulating key developmental stages in a dish, as well as for the establishment of cellular
disease models. Therefore, and with the goal of comprehensively investigating human cardiovascular development, here we analyzed human ESCs during the course of differentiation to mesoderm, cardiovascular progenitor cells and differentiated vascular endothelial cells. By relying on our recently reported methodologies for the differentiation of human stem cells to cardiovascular progenitors and terminally differentiated vascular endothelial cells, here we report on the identification and characterization of 3 novel human lncRNAs (for a summary of findings for each lncRNA, see Table 9 in the online-only Data Supplement). First, we found a total of 75 novel lncRNAs with different expression patterns depending on the differentiation stage. Among those lncRNAs fulfilling our pre-selection criteria, we next selected 3 novel non-coding transcripts (TERMINATOR, ALIEN and PUNISHER) for further characterization in different vertebrate models. We found that TERMINATOR specifically controls pluripotent stem cell identity, ALIEN impairs cardiovascular development and PUNISHER compromises endothelial cell function. The differentiation platform utilized does however present several limitations. As for most protocols on pluripotent stem cell differentiation, the generated cells generally present a fetal-like signature. Therefore, potentially obviating lncRNAs that are specifically expressed in adult cells. In addition, in vitro differentiation protocols and genetic profiling suffers from the inherent disadvantage of providing snapshots of what is otherwise a continuous process. Thus, we cannot rule out that, despite collecting samples every second day, other more tightly regulated lncRNAs are underrepresented in our datasets. Despite these limitations and considering the role that the identified lncRNAs play during cardiovascular development, it is tempting to speculate that deregulation of lncRNAs affecting early embryonic development might be one of the causes for congenital cardiovascular diseases and malformations. Indeed, our work indicates that deregulation of lncRNAs and the resulting changes in gene expression have
profound implication in embryonic development, cardiovascular system formation and function across different vertebrate species. Whereas the implications that IncRNAs play in human disease remain obscure, and only now we start unraveling the complex role that non-coding genetic elements play in the context of disease, it seems clear that leveraging next generation sequencing technologies with patient data collection might shed new light onto how aberrant IncRNA expression functionality correlates with disease. Understanding the precise role of ncRNAs during development and disease might ultimately open new avenues for the development of human therapeutics.

Altogether, our work establishes an analytic pipeline for the systematic study of IncRNAs in cardiovascular development and demonstrates that appropriate in vitro systems can be used to identify novel players controlling lineage commitment and human development. Ultimately, broad-scale application of genomic strategies based on the use of pluripotent cells might shed new light on the fundamental mechanisms underlying human development and disease.

Acknowledgments: We thank M. Schwarz, P. Schwarz, I. Dubova, M. Marti, C. Fabregat and D. Okamura for administrative and technical support. We thank Ruoping Chen, and the Beijing Genomics Institute (BGI) for sequencing services and Nasun Hah for valuable advice.

Funding Sources: Work in the laboratory of G.W.Y. was supported by grants from the National Institutes of Health (U54 HG007005, R01 HG004659 and R01 NS075449) and the California Institute for Regenerative Medicine (RB3-05009). Work in the laboratory of J.C.I.B. was supported by grants from NIH (5U01HL107442), CIRM, the G. Harold and Leila Y. Mathers Charitable Foundation, The Leona M. and Harry B. Helmsley Charitable Trust (2012-PG-MED002). L.K. was partially supported by the California Institute for Regenerative Medicine (CIRM). E.N. was partially supported by a F.M. Kirby Foundation postdoctoral fellowship. I.S.M. was partially supported by a Nomis Foundation postdoctoral fellowship.
Conflict of Interest Disclosures: None.

References:


**Figure Legends**

**Figure 1.** Transcriptome kinetics during human Embryonic Stem Cell differentiation to endothelial cells. **A,** Annotation of transcripts expressed >0.1 RPKM during differentiation of human H1 embryonic stem cells into endothelial cells. **B,** Total genes regulated more than 3-fold between key stages of vascular differentiation. Fractions attributed to lncRNA regulation are shown in dark red and dark blue. **C,** Dynamics of key lineage-restricted markers during vascular differentiation from hESCs to endothelial cells. **D,** Gene Ontology enrichment for biological processes regulated at each specific stage of differentiation. **E,** Contour plots depicting the relative fraction of genes presenting bivalent methylation marks (H3K4me3 and H3K27me3) at promoters of each gene (ChIP-Seq reads [log2] from -2kb,+2kb from the GENCODE TSS). Bivalent genes are defined as those with greater than 5 (log2) normalized ChIP-Seq reads. Pie chart represents the summary of genes activated during vascular differentiation which are driven by bivalent promoters, including lncRNAs. **F,** Regulatory motif enrichment in the promoters [-300 bp,+50 bp] of critical genes at each stage of differentiation.

**Figure 2.** Characterization of TERMINATOR, ALIEN and PUNISHER, three novel developmentally regulated lncRNAs. **A,** Representative images of subcellular localization of TERMINATOR in human ES cells, ALIEN in vascular progenitors and PUNISHER in primary endothelial cells as determined by RNA *in situ* hybridization using specific LNA-probes. A scrambled control LNA-probe has been tested in all 3 different cell types with similar results (representative pictures on the left). **B,** Quantification of nuclear and cytosolic lncRNA foci (n≥5). **C,** On the left, RNA-Seq read density coverage along lncRNA loci. On the right, gene
ontology functional enrichment analysis from all genes exhibiting similar expression profiles to *TERMINATOR* (upper), *ALIEN* (middle) or *PUNISHER* (lower). **D,** Network depicting correlated gene expression profiles of the uncharacterised IncRNAs and key developmental transcription factors. A Pearson correlation threshold of 0.85 was used to define edges in the network. Thick, red lines indicate higher correlation relative to black lines. Data are represented as mean +/- s.d. Scale bars: 25μm (A, upper panels) and 5μm (A, lower panels).

**Figure 3.** IncRNA sequence conservation and expression profiles through vertebrate evolution and development. **A-C,** Dynamic expression pattern in developing mouse and zebrafish embryos for *Terminator* (A); *Alien* (B) and *Punisher* (C), as determined by qRT-PCR (n=5/group). **D,** Vertebrate genomic alignments and PhyloP conservation scores for each uncharacterized IncRNA across 44 species. See Supplemental Table 5 for the full list of animals used for PhyloP conservation analysis (1-7). Data are represented as mean +/- s.d.

**Figure 4.** *In vivo* functional evaluation of conserved IncRNAs during zebrafish development by morpholino mediated loss of function. **A,** Morpholinos were designed to block specific highly conserved regions or putative splice sites and were injected in cardiovascular reporter zebrafish embryos. **B,** Representative phenotypes observed for the different knockdowns during vascular (*fli1*:GFP) and cardiac (*cmlc2*:GFP) development at 48 hpf. **C,D,** Quantification of vascular (C) and cardiac (D) phenotypes obtained for the different morpholinos (n>70). **E,F,** analysis of defects in blood vessel formation by quantification of tube number (E) and tube length (F) (n>70). **G,** Heart function measured as heartbeats/min for the different morphants (n>10). Data are represented as mean +/- s.d., *P< 0.05.* hpf: hours post-fertilization. Scale bars: 5μm (B)
**Figure 5.** Terminator is essential for pluripotent stem cells survival. **A,** Morpholino mediated loss of function of Terminator in mouse blastocyst stage embryos led to significant downregulation of pluripotency factors Oct4, Nanog and Sox2, indicating a fundamental role during very early development (n>8). **B, C,** Massive cell death in human embryonic stem cell lines (H1 and H9) was induced by loss of TERMINATOR as indicated by representative brightfield pictures of sh-scrambled and sh-TERMINATOR treated cell culture (**B**). In **C,** treated cells were stained by crystal violet. **D,** TERMINATOR knockdown induced apoptosis in human pluripotent stem cells as indicated by TUNEL staining. **E,** Downregulation of essential pluripotency transcription factors Oct4 and Sox2 upon TERMINATOR knockdown (n=5). Data are represented as mean +/- s.d., * P < 0.05. Scale bars: 200μm (B), 75μm (D).

**Figure 6.** Gene expression profiling of key molecular networks influenced by lncRNA knockdown during human vascular differentiation. Heat map depicting genetic networks differentially regulated upon shRNA-mediated knockdown of TERMINATOR (human ES cells), ALIEN (vascular progenitors) and PUNISHER (HUVECs), respectively. Enriched biological processes are indicated on the right.

**Figure 7.** PUNISHER knockdown results in severe functional aberrations in human primary endothelial cells. **A,** Loss of PUNISHER in endothelial cells led to significant reduction in cell proliferation indicated by loss of phosphorylated histone H3 48 hours after knockdown. **B,** Representative brightfield pictures showing well organized capillary-like structures formed by scramble-treated endothelial cells (left) as opposed to PUNISHER knockdown endothelial cells (right). **C,** Quantitative analysis after matrigel tube formation assay upon knockdown of
PUNISHER indicated severe defects in branching (left), tubularity (middle) and mean tube length (right) (n≥5). D, acLDL functional assay showed significantly reduced uptake upon PUNISHER knockdown (n=5). Note that the control group represents primary endothelial cells. Data are represented as mean +/-s.d., * P< 0.05. Scale bars: 100μm (A) and 1mm(B).
Figure 1
Figure 3
Figure 4
Figure 5
Figure 7