HHS Public Access

Author manuscript

Cell Rep. Author manuscript; available in PMC 2016 November 23.

Published in final edited form as:

Cell Rep. 2016 October 4; 17(2): 353–365. doi:10.1016/j.celrep.2016.09.017.

DIGIT is a conserved long noncoding RNA that regulates *GSC* expression to control definitive endoderm differentiation of embryonic stem cells

Kaveh Daneshvar[1], **Joshua V. Pondick**[1], **Byeong-Moo Kim**[1], **Chan Zhou**[1], **Samuel R. York**[1], **Jillian A. Macklin**[1], **Ameed Abualteen**[1,2], **Bo Tan**[1,3], **Alla A. Sigova**[4], **Chelsea Marcho**[5], **Kimberly D. Tremblay**[5], **Jesse Mager**[5], **Michael Choi**[1], and **Alan C. Mullen**[1,2,6,†]

[1]Gastrointestinal Unit, Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts 02114 USA

[2]Harvard Stem Cell Institute, Cambridge, MA 02138 USA

[3]School of Fundamental Medical Sciences, Guangzhou University of Chinese Medicine, Guangzhou 510405, China

[4]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA

[5]Department of Veterinary and Animal Sciences, University of Massachusetts at Amherst, Amherst, MA 01003 USA

Abstract

Long noncoding RNAs (lncRNAs) exhibit diverse functions, including regulation of development. Here we combine genome-wide mapping of SMAD3 occupancy with expression analysis to identify lncRNAs induced by activin signaling during endoderm differentiation of human embryonic stem cells (hESCs). We find that *DIGIT* is divergent to *Goosecoid* (*GSC*) and expressed during endoderm differentiation. Deletion of the SMAD3-occupied enhancer proximal to *DIGIT* inhibits *DIGIT* and *GSC* expression and definitive endoderm differentiation. Disruption

† Corresponding author. Thier 306B, 55 Fruit Street, Massachusetts General Hospital, Boston, MA, 02114, (617) 726-6342, acmullen@mgh.harvard.edu.

[6]Lead Contact

Supplemental Information

Supplemental information includes Extended Experimental Procedures, four supplemental figures and three supplemental tables.

Author contribution

K.D. and A.C.M. conceived the study and designed the experiments. K.D., J.V.P., B.K., S.R.Y., J.A.M., A.A., B.T., C.M., K.D.T., J.M., and M.C. performed and analyzed the experiments. C.Z. and J.V.P. performed the computational analysis. K.D., and A.C.M. wrote the manuscript with input from A.A.S.

Accession Numbers

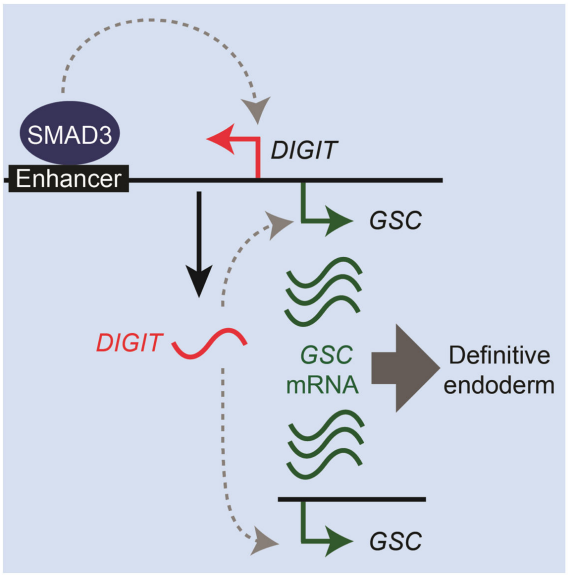
The ChIP-seq and RNA-seq data produced for this study are deposited in GEO under the accession number: GSE75297

In Brief (eTOC)

Daneshvar et al. identify *DIGIT* as a conserved lncRNA that is induced by activin signaling and controls definitive endoderm differentiation of human and mouse embryonic stem cells. *DIGIT* is transcribed divergently relative to *Goosecoid* (*GSC*) and regulates *GSC* expression to control differentiation.

of the gene encoding *DIGIT* and depletion of the *DIGIT* transcript reveal that *DIGIT* is required for definitive endoderm differentiation. In addition, we identify the mouse ortholog of *DIGIT* and show that it is expressed during development and promotes definitive endoderm differentiation of mouse ESCs. *DIGIT* regulates *GSC in trans,* and activation of endogenous *GSC* expression is sufficient to rescue definitive endoderm differentiation in *DIGIT*-deficient hESCs. Our study defines *DIGIT* as a conserved noncoding developmental regulator of definitive endoderm.

Graphical abstract



Keywords

Definitive endoderm; lncRNA; *GSC*; *DIGIT*; SMAD3

Introduction

Definitive endoderm (DE) differentiation is one of the earliest steps in lineage commitment, leading to development of the gastrointestinal organs, lungs and thymus (Zorn and Wells, 2009). Activin or Nodal signaling, through the transforming growth factor beta (TGF-β) receptors, is the primary event initiating DE differentiation of embryonic stem cells (ESCs) (D'Amour et al., 2005; Kubo, 2004). Activation of the TGF-β receptors leads to phosphorylation of the transcription factors SMAD2 and SMAD3, which translocate from the cytoplasm to the nucleus to regulate gene expression (Massagué et al., 2005). SMAD2/3 tend to occupy unique enhancers in different cell types by associating with key transcription factors that determine cell identity (Mullen et al., 2011). As a result, SMAD2/3 switch locations to occupy many new enhancers during DE differentiation (Brown et al., 2011; Kim et al., 2011). Key mesendoderm and endoderm transcriptional regulators including EOMES, MIXL1, SOX17, and FOXA2 are each direct targets of the activin/SMAD2/3 pathway during differentiation (Brown et al., 2011; D'Amour et al., 2005; Kim et al., 2011).

Long noncoding RNAs (lncRNAs) are increasingly recognized as regulators of development and differentiation. lncRNAs are greater than 200 nucleotides (nt) in length, are often polyadenylated, have the same features as messenger (m) RNAs (Guttman et al., 2009), and can localize to the nucleus or cytoplasm without being translated into proteins (Cabili et al., 2015). Recent studies have identified lncRNAs as key contributors to the specification of lineages derived from all three germ layers including neural (Sauvageau et al., 2013), epidermal (Kretz et al., 2012), cardiovascular (Grote et al., 2013; Klattenhoff et al., 2013), dendritic (Wang et al., 2014), skeletal muscle (Gong et al., 2015), and lung (Herriges et al., 2014).

Although many lncRNAs have been identified and cataloged in human cells (Derrien et al., 2012; Xie et al., 2014), in most cases, their functions have yet to be defined. Two studies have identified lncRNAs induced in human endoderm differentiation (Jiang et al., 2015; Sigova et al., 2013). However, the lncRNAs directly targeted by activin signaling that contribute to regulation of DE differentiation remain unknown. We mapped the genome-wide occupancy of SMAD3 during endoderm differentiation and identified *DIGIT* (**Di**vergent to ***GSC***, **I**nduced by **T**GF-β family signaling) as an lncRNA regulated by an enhancer bound by SMAD3 following activin signaling. We find that *DIGIT* is required for productive DE differentiation of both human and mouse ESCs. *DIGIT* is divergently transcribed from the mesendoderm regulator *Goosecoid* (*GSC*), and *DIGIT* and *GSC* transcripts are induced coordinately in the same cells during differentiation. *GSC* expression is regulated by *DIGIT in trans*, and the defect in DE differentiation present in *DIGIT*-deficient hESCs can be rescued by induction of *GSC*. These results identify *DIGIT* as an early regulator of DE differentiation for both human and mouse ESCs.

Results

DIGIT is induced by activin signaling

We first defined the lncRNAs that were induced by activin signaling during endoderm differentiation. Chromatin immunoprecipitation and sequencing (ChIP-seq) identified 252 SMAD3 enhancers activated with induction of endoderm (Figures 1A and 1B) (Table S1). Steady-state levels of 1387 lncRNAs were elevated at least two fold (Sigova et al., 2013) with endoderm differentiation, and 14 of these lncRNAs were in close proximity to SMAD3 enhancers, suggesting that they may be direct targets of activin signaling. Of these 14 candidates, four showed transcriptional activation during endoderm differentiation as measured by global run on sequencing (GRO-seq, p value <0.01) (Sigova et al., 2013). Only one lncRNA had an ortholog annotated in the mouse genome (GRCm38/mm10), suggesting possible functional conservation. Transcription of this lncRNA was activated during endoderm differentiation as measured by GRO-seq, together with the divergently transcribed developmental transcription factor *GSC* (Figure 1C). Formaldehyde-Assisted Isolation of Regulatory Elements (Faire) (Giresi et al., 2007; Simon et al., 2012) showed that *DIGIT* and *GSC* are transcribed from bidirectional promoters (Scruggs et al., 2015), characterized by nucleosome depletion between the two transcription start sites (TSSs) during endoderm differentiation (Figure S1A).

To validate the genomic features of *DIGIT*, we generated complementary (c) DNA using polyadenylated (polyA) RNA isolated from differentiating hESCs and performed rapid amplification of cDNA ends (RACE) to define the location of the 5' cap and the 3' terminus of the *DIGIT* transcript (Figure S1B; the full sequence is contained in Supplemental Experimental Procedures). We then cloned the full-length transcript from polyA RNA to confirm expression during endoderm differentiation. This analysis showed that *DIGIT* is encoded by two exons, as predicted by RNA sequencing (RNA-seq) (Sigova et al., 2013), and defined the 3' end of exon 2 to be downstream of the 3' end predicted by RNA-seq (Figure 1C, cloned). We then analyzed *DIGIT* expression during endoderm differentiation using quantitative reverse transcription PCR (qRT-PCR). Both *DIGIT* and *GSC* are induced during the first four days of endoderm differentiation (Figure 1D), and both *DIGIT* and *GSC* are highly restricted to endoderm compared to other human tissues (Figures S1C-S1E).

We performed single molecule RNA fluorescent in-situ hybridization (FISH) and found that *DIGIT* transcripts are primarily retained in the nucleus while *GSC* mRNA transcripts are most abundant in the cytoplasm (Figure 1E). Quantification of this distribution across 50 cells revealed that about 90% of *DIGIT* transcripts are retained in the nucleus compared to 15% of *GSC* transcripts. Nuclear localization of *DIGIT* provides further support that *DIGIT* is not an mRNA, and this analysis also revealed that *DIGIT* and *GSC* are induced in the same cells during differentiation.

DIGIT and *GSC* are coordinately induced during endoderm differentiation, and we asked if the proximal enhancer occupied by SMAD3 (Figures 1B and S1F), which is located 5 kb downstream of the *DIGIT* TSS, regulates *DIGIT* and *GSC* activation. We used the CRISPR system (Cong et al., 2013) to delete both copies of the enhancer occupied by SMAD3 during hESC differentiation (Figure 1F). hESCs in which this enhancer was deleted maintain expression of the ESC markers *OCT4* and *NANOG* (Figures 1G and 1H), but show a defect in activation of *DIGIT* and *GSC* upon endoderm differentiation (Figures 1I and S1G). Furthermore, deletion of the enhancer occupied by SMAD3 was also associated with a significant reduction in mRNA and protein expression of SOX17, FOXA2 and CXCR4 (Figures 1I, 1J, S1G and S1H), which together identify DE (D'Amour et al., 2005; Green et al., 2011; Loh et al., 2014; Ogawa et al., 2013). Passage-matched hESCs and hESCs containing a green fluorescent protein (GFP) expression system (Sim et al., 2015) were used as controls. We used the CRISPR system to insert the components of the GFP expression system into the *AAVS1* loci to create a control that had undergone the same manipulations used to create the SMAD3 enhancer deletions. These findings show that the enhancer newly-occupied by SMAD3 during endoderm differentiation regulates *DIGIT* and *GSC* expression and suggest that depletion of *DIGIT* may inhibit DE differentiation.

DIGIT is a regulator of definitive endoderm differentiation

If *DIGIT* is required for DE differentiation, we would expect depletion of the *DIGIT* transcript to inhibit expression of genes that mark the DE fate. We used the CRISPR system to create hESC lines with constitutive expression of short hairpin (sh) RNAs against *DIGIT* by inserting the shRNA expression cassette along with a drug resistance cassette into the *AAVS1* locus. Knockdown of *DIGIT* expression resulted in a defect in induction of *SOX17,*

FOXA2 and *CXCR4* after four days of differentiation compared to controls (Figure 2A), indicating that DE differentiation is compromised. In addition, depletion of *DIGIT* using locked nucleic acids (LNAs) demonstrated a similar defect in endoderm differentiation (Figure S2A). The LNA experiments were performed on day 2 of endoderm differentiation using two LNA constructs that targeted *DIGIT*. Transient transfection with LNAs was less effective in depleting *DIGIT* compared to stable integration of shRNAs, but still showed that reduction in the *DIGIT* transcript was associated with decreased expression of *SOX17* and *FOXA2*. *CXCR4* expression was not assessed because it is not induced on day 2 of endoderm differentiation.

To further investigate the role of *DIGIT* in DE differentiation, we asked how disruption of the gene encoding *DIGIT* affects differentiation. We used the CRISPR system to insert a sequence encoding GFP followed by a polyA signal 44 base pairs (bp) downstream of the *DIGIT* TSS (Figure 2B). This sequence was inserted without deleting endogenous DNA to avoid inadvertently removing regulatory elements that might affect nearby genes. Two drug selection markers were used to identify colonies with GFP-polyA insertions in both *DIGIT* alleles. hESC lines were expanded from single colonies and then transfected with a plasmid expressing Cre recombinase to remove the drug resistance cassettes (Figures S2B and S2C). Single cells were sorted and expanded to establish *DIGIT*-deficient ($DIGIT^{gfp/gfp}$) hESC lines. The insertion of GFP-polyA allows activation of transcription of GFP at the *DIGIT* locus during endoderm differentiation but leads to termination at the polyA signal, preventing expression of the full-length *DIGIT* transcript. Analysis of RNA expression by PCR and RNA-FISH demonstrates that *GFP* transcripts are induced during endoderm differentiation (Figures 2C and 2D) and are translated into protein (Figure 2E). The presence of GFP transcripts and the dramatic reduction of *DIGIT* (Figure 2D) indicate that transcription is activated at the *DIGIT* locus with little production of the *DIGIT* transcript.

$DIGIT^{gfp/gfp}$ hESC lines maintain expression of hESC master regulators (Figure S2D) (Boyer et al., 2005). However, these cells show significantly reduced levels of *DIGIT* when differentiated towards endoderm and also demonstrate a defect in the induction of *SOX17*, *FOXA2* and *CXCR4* mRNA (Figure 2F). Reduced *DIGIT* expression also inhibited activation of *DEANR1*, an lncRNA recently described to regulate *FOXA2* expression (Figure S2E) (Jiang et al., 2015). The primers detecting the *DIGIT* transcript are located downstream of the GFP-polyA insertion, and the low level of *DIGIT* still detected in $DIGIT^{gfp/gfp}$ cells (Figure 2F) is likely the result of a low frequency of read through of the polyA sequence.

hESCs were also analyzed by flow cytometry after four days of endoderm differentiation (Figure 2G). Forty-eight percent of wild-type hESCs expressed SOX17, FOXA2 and CXCR4, which together identify DE. In contrast, only four and twenty-seven percent of two independent $DIGIT^{gfp/gfp}$ cell lines co-express the DE markers (Figure 2G, right). $DIGIT^{gfp/gfp}$ hESCs that do differentiate into DE maintain expression of SOX17 and FOXA2 at levels comparable to wildtype cells (Figure S2F). Thus, *DIGIT* appears to regulate the fraction of cells differentiating into DE but not the level of FOXA2 and SOX17 proteins in DE cells. We also quantified DE by another set of markers and found that reduced expression of *DIGIT* led to reduced expression of DE (58-64% reduced to 3-7%), as

quantified by cells co-expression of c-KIT and CXCR4 (Figure S2G) (Green et al., 2011; Jiang et al., 2013; Nostro et al., 2011; Ogawa et al., 2013). In contrast to genes that together identify DE, reduced *DIGIT* levels did not have a significant effect on expression of *SOX7*, a marker of primitive and visceral endoderm (Kanai-Azuma et al., 2002), *PLAT(T-PA)*, a marker of parietal endoderm (Cheng and Grabel, 1997), *or MEOX1,* a marker of mesoderm (Candia et al., 1992) (Figure S2H), suggesting that the activity of *DIGIT* is specific for DE.

We performed RNA-seq to determine the effect of depletion of *DIGIT* on the transcriptome during endoderm differentiation. Compared to wild-type controls, 225 genes were repressed and 45 genes were induced in $DIGIT^{gfp/gfp}$ cells after four days of differentiation (Figure 2H and Table S1). Sequencing was performed in duplicate for $DIGIT^{gfp/gfp}$ 1 and $DIGIT^{gfp/gfp}$ 2 cells and compared to control hESCs. The genes that were repressed with depletion of *DIGIT* tend to be induced in endoderm differentiation compared to differentiation into ectoderm or mesoderm lineages (Figure 2I, top). The small number of genes induced with loss of *DIGIT* were enriched in neural progenitor cells (NPCs) (Figure 2I, bottom, p-value < 0.0018).

To assess the role of *DIGIT* in non-directed differentiation (Osafune et al., 2008), we created embryoid bodies (EBs) (Figure S2I) using $DIGIT^{gfp/gfp}$ 1 and $DIGIT^{gfp/gfp}$ 2 hESC lines and one control hESC line matched to the same passage number. EBs were allowed to spontaneously differentiate for 12 days. *OCT4* expression was similar between all groups on day 0 and decreased dramatically by day 12. The endoderm markers *SOX17* and *Albumin* are induced in wild-type cells on day 12, but this induction was inhibited in $DIGIT^{gfp/gfp}$ hESCs (Figure 2J). Expression of *PAX6* and *MEOX1*, which mark ectoderm and mesoderm, respectively were unchanged between all groups (Figure S2J). Overall, these results reveal that the *DIGIT* transcript is required for both directed and spontaneous differentiation of hESCs towards DE.

Definitive endoderm differentiation is regulated by an ortholog of *DIGIT* in mice

We next asked if *DIGIT* is conserved in mammalian development. We identified a 3002 nt mouse transcript divergent to *Gsc* that was isolated from day 9.5 embryos (*Gm10000*) and annotated to contain a 357 nt open reading frame (ORF). We performed RACE-PCR and then cloned the full-length lncRNA from polyA RNA harvested from mESCs that were differentiated towards endoderm for five days. This analysis confirmed a single exon transcript that contained a 5' cap and 3' polyA signal (Figure 3A; the full sequence is contained in Supplemental Experimental Procedures).

It was unclear if this transcript was induced during endoderm differentiation, if it was a coding or noncoding transcript, or if it shared conserved function with *DIGIT.* We found that mouse *Digit* was induced after 5 days of endoderm differentiation compared to mESCs (Day 0) (Figure 3B). Similar to human transcripts, mouse *Digit* and *Gsc* are also highly restricted in expression to endoderm compared to other tissues (Figures S3A and S3B). Analysis of nucleic acid sequence conservation between the human and mouse orthologs also revealed a 335 nucleotide sequence with 88% identity (Figure 3C). These results show that both *DIGIT* and *GSC* are induced during endoderm differentiation of hESCs and mESCs. We then isolated RNA from mouse embryos from embryonic day (E) 5.5 to 8.5 (Figure S3C) to

quantify *Digit* and *Gsc* expression during *in vivo* development. RT-PCR results show that both *Digit* and *Gsc* are induced on E6.5, which corresponds to formation of the anterior primitive streak (Lawson et al., 1991). The highest levels of *Digit* expression are reached on E7.5, corresponding to formation of definitive endoderm (Tam and Beddington, 1987) and continue to be expressed on E8.5 (Figure 3D).

We then asked if the *Digit* transcript encoded a protein. First, we prepared nuclear and cytoplasmic extracts from mESCs differentiated towards endoderm for 5 days and found that *Digit* was retained in the nucleus (Figure 3E), which suggests that *Digit* is not an mRNA despite containing an annotated ORF. Next, we assessed the coding potency of the annotated ORF. CPAT analysis (Wang et al., 2013) provided further evidence that *Digit* is not a protein-coding transcript ($p < 0.08$). Finally, we tried to detect protein encoded by the annotated ORF. We generated a transgene in which sequence encoding a hemagglutinin (HA) tag was inserted in frame within the annotated ORF immediately upstream of the annotated stop codon. As a control, a transgene was created in which sequence encoding an HA-tag was inserted at 3'end of mouse Hemoglobin (Hb) cDNA immediately upstream of the stop codon. The *Hb* cDNA was chosen because it encodes a protein of similar molecular weight to that predicted for *Digit*.

Ectopic expression of these transgenes in HEK 293T cells resulted in production of RNA for both transgenes (Figure S3D), but Western blot analysis detected the HA tag only in cells expressing the *Hb* transgene and not in cells expressing the *Digit* transgene (Figure 3F). It is possible that HEK 293T cells may not express all the factors necessary to promote translation of a mouse transcript, so we transfected mESCs with the same constructs. We could detect the Hb-HA protein by Immunofluorescence (IF) microscopy, but not the digit-HA product (Figure 3G). These results establish *Digit* as an lncRNA that shares a conserved genomic location and pattern of expression with *DIGIT*.

We applied the GFP-polyA knock-in strategy previously used for premature termination of *DIGIT* transcription in hESCs to determine if *Digit* is required for DE differentiation in mESCs (Figure S3E). Mouse *Digit* was also required for DE differentiation, as $Digit^{gfp/gfp}$ mESCs were deficient in induction of Sox17, Foxa2 and Cxcr4 compared to controls (Figures 3H and 3I). Thus, *DIGIT* is an lncRNA with conserved function in differentiation of mammalian ESCs.

DIGIT contributes to definitive endoderm differentiation through regulation of *GSC* expression

Divergent lncRNAs can positively and negatively regulate neighboring protein-coding genes (Guil and Esteller, 2012), and we asked if *GSC* expression is dependent on *DIGIT*. We quantified expression of *GSC* in hESCs during endoderm differentiation with depletion of *DIGIT* by shRNAs and LNAs and found that depletion of *DIGIT* was also associated with reduced *GSC* expression (Figures 4A left and S4A). Furthermore, $DIGIT^{gfp/gfp}$ hESCs (Figure 4A, right) and $Digit^{gfp/gfp}$ mESCs (Figure S4B) also showed reduced induction of *GSC* during endoderm differentiation.

GSC is a homeobox gene induced by activin/Nodal signaling in early gastrulation in *Xenopus* and the primitive streak in mammals (Blum et al., 1992; Cho et al., 1991), but its role in hESC differentiation is not fully understood. We used an shRNA to deplete *GSC* (Figure 4B) and determine how depletion of *GSC* affected differentiation of DE. Depletion of *GSC* expression inhibited *SOX17*, as previously described (Kalisz et al., 2012) as well as *FOXA2* and *CXCR4*, which together mark DE. The results suggest that *GSC* is required for formation of DE during hESC differentiation.

The *DIGIT*$^{gfp/gfp}$ phenotype is rescued by induction of endogenous *GSC*

If *DIGIT* contributes to DE differentiation by regulating *GSC*, then induction of *GSC* in *DIGIT*-deficient hESCs should rescue the defect observed during endoderm differentiation. We first transfected hESCs with a plasmid encoding a dead Cas9 fused to the VP64-P65-Rta transcription activation domain (dCas9-VPR) (Chavez et al., 2015) and a plasmid encoding a gRNA that directs dCas9-VPR to the *GSC* promoter (Figure 4C) to determine if recruiting dCas9-VPR to the GSC promoter could activate *GSC* expression. hESCs transfected with both plasmids showed a nearly 200-fold induction of *GSC* after 48 hours (Figure 4D, left). dCas9-VPR was directed to bind 102 bp upstream of the GSC TSS, but this was also 376 bp upstream of the *DIGIT* TSS (Figure 4C), and targeting dCas9-VPR to this location also increased *DIGIT* expression, but to a lower extent (Figure 4D, right). These results showed that dCas9-VPR primarily activated *GSC*, but also increased expression of *DIGIT*. *DIGIT*$^{gfp/gfp}$ hESCs rely on a polyA signal to terminate transcription and prevent production of *DIGIT*. We were concerned that increased transcription at the *DIGIT* locus in *DIGIT*$^{gfp/gfp}$ cells would lead to a proportional increase in the small number of transcripts reading through the polyA and complicate interpretation of the experiment. To avoid this effect, we used the CRISPR system to generate hESC lines with homozygous deletion of the second exon of the *DIGIT* (Figure 4E and S4C). We then confirmed that these *DIGIT*$^{-/-}$ hESCs do not express *DIGIT* during endoderm differentiation (Figure S4D) and are also deficient in DE differentiation (Figure 4F). We next transiently transfected *DIGIT*$^{-/-}$ hESCs with plasmids expressing dCas9-VPR and *GSC* gRNA. We cultured cells in hESCs media for two days to allow expression of GSC prior to induction of differentiation. Because these were transient transfections, we analyzed gene expression after 2 days of endoderm differentiation and found that the DE markers *SOX17* and *FOXA2* were induced in *DIGIT*$^{-/-}$ cells with activation of GSC (Figure 4G). We repeated the experiments in *DIGIT*$^{gfp/gfp}$ hESCs with the same results (Figures S4E and S4F). These findings demonstrate that activation of endogenous *GSC* can rescue the defect in DE differentiation created by loss of *DIGIT* expression and suggest that *DIGIT* controls DE differentiation of hESCs, at least in part, through regulation of *GSC*.

DIGIT regulates *GSC in trans*

Depletion of *DIGIT* by shRNAs, LNAs, and insertion of polyA termination sequences shows that it is the *DIGIT* transcript and not transcription at the *DIGIT* locus that is required to regulate *GSC* expression. To provide further evidence that the *DIGIT* transcript regulates GSC expression, we asked if ectopic expression of *DIGIT* could induce *GSC* expression in *DIGIT*-deficient cells. We first transfected hESCs with a plasmid expressing *DIGIT* behind a PGK promoter and performed RNA-FISH to confirm that *DIGIT* expressed from the

plasmid is localized to the nucleus (Figure 4H). Next, we transiently transfected *DIGIT* $^{-/-}$ hESCs with a plasmid expressing *DIGIT* or a scrambled *DIGIT* sequence along with GFP. One day after transfection, hESCs were differentiated towards endoderm, and GFP+ cells were sorted after 2 days of differentiation. We found that *DIGIT* $^{-/-}$ hESCs expressing ectopic *DIGIT* (Figure 4I, left) showed increased expression of *GSC* as well as *FOXA2* and *SOX17* (Figure 4I, right). *DIGIT* $^{-/-}$ cells were used for this experiment instead of *DIGIT*$^{gfp/gfp}$ hESCs because they did not already express GFP, and expression analysis was performed on day 2 of differentiation because these were transient transfections. *CXCR4* expression was not assessed, as it is not induced on day 2 of endoderm differentiation. These results show that *DIGIT* does not need to be transcribed adjacent to *GSC* to promote *GSC* expression and provides further evidence that *DIGIT* regulates *GSC* expression.

The results from ectopic expression of *DIGIT* suggest that the *DIGIT* transcript may function in *trans* to regulate *GSC*. To further evaluate this possibility, we created *DIGIT heterozygous (*$^{+/gfp}$*)* hESCs and asked if *GSC* was expressed from both alleles or only the allele adjacent to the wildtype copy of *DIGIT*. We performed RNA-FISH to analyze *GSC* expression in wildtype and *DIGIT*$^{+/gfp}$ hESCs undergoing endoderm differentiation (Figure 4J). Sites of *GSC* transcription are identified by bright foci of *GSC* RNA (Figure 1E) (Levesque and Raj, 2013) and can be observed at one or two sites in differentiating cells. The distribution of cells demonstrating one or two foci to mark active *GSC* transcription are not statistically different between wildtype and *DIGIT*$^{+/gfp}$ cells, and *DIGIT*$^{+/gfp}$ cells show a similar frequency of cells expressing *GSC* from two alleles compared to wildtype cells (Figure 4J). These results suggest that transcription can occur at both *GSC* alleles even when one allele is no longer producing *DIGIT* (Figure 4K). Together, the results from ectopic expression of *DIGIT* in *DIGIT* $^{-/-}$ cells and the expression of *GSC* from both alleles in *DIGIT* heterozygous cells suggest that the *DIGIT* transcript can function *in trans* to regulate *GSC* expression.

Discussion

In this study we performed genome-wide analysis of SMAD3 occupancy to identify enhancers targeted by activin signaling during endoderm differentiation. This approach combined with previous analysis of transcriptional responses (Sigova et al., 2013) allowed us to identify *DIGIT* out of over 1000 lncRNAs induced during endoderm differentiation. Deletion of the SMAD3-occupied enhancer proximal to *DIGIT* showed a profound reduction in *DIGIT* and *GSC* expression and inhibition of DE differentiation. These results show that the enhancer is required for normal activation of *DIGIT* and *GSC*, but targeted disruption of SMAD3 binding elements within this enhancer will be required to determine whether or not SMAD3 occupancy is required for activation. Our analysis defined 252 sites activated for SMAD3 binding within 48 hours of endoderm differentiation, and it is likely that investigation of these sites will identify additional enhancers that are also required for endoderm differentiation. Applying this concept more broadly suggests that focusing on lncRNAs that are direct targets of the signaling pathways that control differentiation will identify the lncRNAs that are likely to regulate development.

DIGIT is expressed divergently from *GSC* and is coordinately induced with *GSC* during endoderm differentiation when analyzed at the population level by RNA expression (Figure 1D) (Sigova et al., 2013), and expression of both genes are highly restricted to endoderm differentiation (Figures S1C-S1E and S3A-S3B). We also observe both *DIGIT* and *GSC* transcripts in the same cells during differentiation (Figure 1E). In other cell types, divergent genes are not always associated with co-expression (Cabili et al., 2015). This coordinated induction of divergent lncRNA and coding gene pairs may be more common with activation of loci during differentiation (Lepoivre et al., 2013; Ponjavic et al., 2009; Sigova et al., 2013) where sets of genes are turned on in a coordinated fashion to regulate changes in cell identity and may be less fixed in cells maintaining homeostasis.

DIGIT is not only divergently transcribed from the gene encoding *GSC*, it also regulates *GSC* expression. Depletion of the *DIGIT* transcript inhibited induction of *GSC* during endoderm differentiation of both human and mouse ESCs (Figures 4A, S4A-B), and ectopic expression of *DIGIT* in *DIGIT* $^{-/-}$ cells was also sufficient to induce *GSC* expression (Figure 4I). These findings, coupled with the observation that *GSC* can be expressed from both alleles in *DIGIT* heterozygous cells, show that *DIGIT* regulates *GSC in trans* rather than by being transcribed in close proximity to *GSC*.

GSC is a homeobox gene that is induced in early gastrulation with formation of the dorsal lip of the blastopore in *Xenopus* and the primitive streak in mammals, and is activated as a target of activin/Nodal signaling (Blum et al., 1992; Cho et al., 1991). Surprisingly, *GSC*$^{-/-}$ mice do not show a defect in gastrulation and survive to gestation with craniofacial defects (Rivera-Pérez et al., 1995). Loss of *GSC* expression in hESCs was previously show to be associated with significant reduction of *SOX17* (Kalisz et al., 2012), which is expressed in both definitive and visceral endoderm (Shimoda et al., 2007). By quantifying the effects of GSC on additional genes, our results provide more convincing data that *GSC* is required for DE differentiation of hESCs. Thus, while loss of *GSC* does not appear to be required for *in vivo* gastrulation in mice, we demonstrate that *GSC* is required for DE of ESCs. Furthermore, the defect in DE differentiation that occurs with loss of *DIGIT* expression can be rescued by induction of *GSC* (Figure 4G), showing that *DIGIT* controls DE differentiation by promoting expression of *GSC*. Further experiments will be required to determine the function of *Digit* during *in vivo* development.

There are increasing examples of lncRNAs that positively regulate proximal protein-coding genes, including those divergently transcribed from developmental genes (Arnes et al., 2016; Herriges et al., 2014; Jiang et al., 2015; Luo et al., 2016), making it necessary to have genomic tools to disrupt lncRNAs without genomic deletions that could affect regulatory elements of neighboring genes. We have established a method for efficient loss of lncRNA function via insertion of a GFP-polyA sequence into the lncRNA gene. Whether transcription initiation at divergently transcribed genes is mediated through bidirectional or adjacent unidirectional promoters (Core et al., 2014; Duttke et al., 2015; Grzechnik et al., 2014), the ability to allow transcription initiation and elongation at the endogenous lncRNA TSS while preventing production of the lncRNA transcript provides an essential tool to dissect lncRNA function. This system can be applied to many divergently transcribed lncRNAs or those in close proximity to protein-coding genes and allows preservation of

transcriptional activation at the lncRNA locus while preventing production of the lncRNA product. It also avoids the need to delete any endogenous sequences that might act as a regulatory element of nearby genes. This approach has led to the identification of *DIGIT* as a key regulator of human and mouse DE differentiation and can be applied to investigate the function of lncRNAs in any developmental system.

Experimental Procedures

ESC culture and differentiation

hESCs were cultured and differentiated as previously described (Sigova et al., 2013). mESCs were cultured on irradiated MEFs and differentiated with 100 ng/ml Activin A. Please see Supplemental Experimental Procedures for additional details.

Sequencing Analysis

ChIP-seq libraries were prepared and analyzed as described (Mullen et al., 2011). GRO-seq (Sigova et al., 2013) analysis was performed using HTseq (Anders et al., 2014) and DESeq2 (Love et al., 2014). RNA-seq analysis of wild-type and $DIGIT^{gfp/gfp}$ hESCs and comparison to other data sets of hESC differentiation was performed using HTseq and DESeq2. Please see Supplemental Experimental Procedures for additional details.

Genome editing

Genome editing was performed by transfecting cells with the px330 plasmid (Cong et al., 2013) containing the indicated gRNAs along with the indicated homology plasmids. The sequences of all gRNAs are listed in Table S2. Please see Supplemental Experimental Procedures for additional information including maps of the homology plasmids.

RNA in-situ hybridization

Custom Stellaris® FISH probes (Biosearch Technologies) were designed against *DIGIT*, *GSC* and *GFP* using the Stellaris® FISH Probe Designer (www.biosearchtech.com/stellarisdesigner) (Table S2). The samples were hybridized with Stellaris FISH Probe sets labeled with Quasar570 or Quasar670.

Cell sorting and flow cytometry analysis

hESCs and differentiated cells were separated into single cells by Accutase treatment prior to cell sorting using a FacsAria II (BD) or flow cytometry using a using an Accuri C6 (BD). Antibodies used for FACS: SOX17-PerCP (BD 562387) FOXA2-PE antibody (BD 561589) CXCR4-APC antibody (R&D Systems FAB173A). Please see Supplemental Experimental Procedures for additional details.

Analysis of RNA expression in developing embryos

Female CD1 mice were mated with male CD1 mice and checked daily for vaginal plugs indicating embryonic day 0. Pregnant mice were euthanized 5-8 days later, and embryos were dissected, carefully staged, and imaged (stages indicated in Figure S3C).

Total RNA was isolated using the Roche High Pure RNA Isolation Kit (Roche 11828665001) following the manufacturer's protocol with the addition of a second DNAse treatment following the first wash. cDNA synthesis was performed using SuperScript II reverse transcriptase (Invitrogen 18064014) with both oligo(dT) primers (Promega C1181) and random primers (Promega C1101). RT-PCR was performed though 35 cycles at 94 degrees C for 30 seconds, 60 degrees C for 30 seconds, and 72 degrees C for 30 seconds. Products were electrophoresed on 2% agarose gel.

Immunofluorescence microscopy

IF was performed using the following antibodies: Oct4 (Abcam, ab19857), Cxcr4 (BD, 551852), FoxA2 (Abcam, ab40874), Sox17 (R&D, AF1924), HA (Abcam, ab9110). Please see Supplemental Experimental Procedures for additional details.

Induction of endogenous *GSC* expression.

A plasmid expressing dCas9-VPR (Chavez et al., 2015) was transiently transfected along with a plasmid expressing either a gRNA specific to *LACZ* or a gRNA specific to promoter of *GSC*. Cells were maintained in mTESR1 for 48 hours after transfection and were then differentiated for an additional 48 hours.

Previously Published RNA Datasets used in this Study

Human tissue datasets from GTEx (The GTEx Consortium, 2013) were obtained through dbGAP (Table S3). Additional datasets of differentiating hESCs were obtained from GSE63935, GSE52657, and GSE44875. GRO-seq data for hESCs and endoderm differentiation were obtained from GSE41009. Mouse tissue datasets were obtained from GSE36025 and datasets for differentiating mESCs were from GSE36114.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

We would like to thank MGH/HSCI flow cytometry core and Meredith Weglarz for technical support, MGH/PBM microscopy core, Biosearch Technologies, Marc Beal for assistance with RNA-FISH, and Kate Jeffrey, Konrad Hochedlinger, Igor Ulitsky, Abid Khan and Jennifer Chen for helpful discussion. This work was supported by NIH grants DK090122 and DK104009 (A.C.M.).

References

Anders S, Pyl PT, Huber W. HTSeq A Python framework to work with high-throughput sequencing data. Bioinformatics. 2014; 31:166–169. [PubMed: 25260700]

Arnes L, Akerman I, Balderes DA, Ferrer J, Sussel L. β linc1 encodes a long noncoding RNA that regulates islet β -cell formation and function. Genes Dev. 2016:502–507. [PubMed: 26944677]

Blum M, Gaunt SJ, Cho KW, Steinbeisser H, Blumberg B, Bittner D, De Robertis EM. Gastrulation in the mouse: the role of the homeobox gene goosecoid. Cell. 1992; 69:1097–1106. [PubMed: 1352187]

Boyer L, Lee TI, Cole M, Johnstone S, Levine S, Zucker J, Guenther M, Kumar R, Murray H, Jenner R, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell. 2005; 122:947–956. [PubMed: 16153702]

Brown S, Teo A, Pauklin S, Hannan N, Cho CH-H, Lim B, Vardy L, Dunn NR, Trotter M, Pedersen R, et al. Activin/Nodal signaling controls divergent transcriptional networks in human embryonic stem cells and in endoderm progenitors. Stem Cells. 2011; 29:1176–1185. [PubMed: 21630377]

Cabili MN, Dunagin MC, McClanahan PD, Biaesch A, Padovan-Merhar O, Regev A, Rinn JL, Raj A. Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. Genome Biol. 2015; 16:20. [PubMed: 25630241]

Candia, a F.; Hu, J.; Crosby, J.; Lalley, P. a; Noden, D.; Nadeau, JH.; Wright, CV. Mox-1 and Mox-2 define a novel homeobox gene subfamily and are differentially expressed during early mesodermal patterning in mouse embryos. Development. 1992; 116:1123–1136. [PubMed: 1363541]

Chavez A, Scheiman J, Vora S, Pruitt BW, Tuttle M, Iyer PR, Lin S, Kiani S, Guzman CD, Wiegand DJ, et al. Highly efficient Cas9-mediated transcriptional programming. Nat. Methods. 2015; 12:2–6. E.

Cheng L, Grabel LB. The involvement of tissue-type plasminogen activator in parietal endoderm outgrowth. Exp Cell Res. 1997; 230:187–196. [PubMed: 9024778]

Cho KW, Blumberg B, Steinbeisser H, De Robertis EM. Molecular nature of Spemann's organizer: the role of the Xenopus homeobox gene goosecoid. Cell. 1991; 67:1111–1120. [PubMed: 1684739]

Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini L. a, et al. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013; 339:819–823. [PubMed: 23287718]

Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat. Genet. 2014; 46:1311–1320. [PubMed: 25383968]

D'Amour, K. a; Agulnick, AD.; Eliazer, S.; Kelly, OG.; Kroon, E.; Baetge, EE. Efficient differentiation of human embryonic stem cells to definitive endoderm. Nat. Biotechnol. 2005; 23:1534–1541. [PubMed: 16258519]

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res. 2012; 22:1775–1789. [PubMed: 22955988]

Duttke SHC, Lacadie SA, Ibrahim MM, Glass CK, Corcoran DL, Benner C, Heinz S, Kadonaga JT, Ohler U. Human Promoters Are Intrinsically Directional. Mol. Cell. 2015; 57:674–684. [PubMed: 25639469]

Giresi PG, Kim J, Mcdaniell RM, Giresi PG, Kim J, Mcdaniell RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. 2007:877–885.

Gong C, Li Z, Ramanujan K, Clay I, Zhang Y, Lemire-Brachat S, Glass DJ. A Long Non-coding RNA, LncMyoD, Regulates Skeletal Muscle Differentiation by Blocking IMP2-Mediated mRNA Translation. Dev. Cell. 2015; 34:181–191. [PubMed: 26143994]

Green MD, Chen A, Nostro M-C, d'Souza SL, Schaniel C, Lemischka IR, Gouon-Evans V, Keller G, Snoeck H-W. Generation of anterior foregut endoderm from human embryonic and induced pluripotent stem cells. Nat. Biotechnol. 2011; 29:267–272. [PubMed: 21358635]

Grote P, Wittler L, Hendrix D, Koch F, Währisch S, Beisaw A, Macura K, Bläss G, Kellis M, Werber M, et al. The Tissue-Specific lncRNA Fendrr Is an Essential Regulator of Heart and Body Wall Development in the Mouse. Dev. Cell. 2013; 24:206–214. [PubMed: 23369715]

Grzechnik P, Tan-Wong SM, Proudfoot NJ. Terminate and make a loop: Regulation of transcriptional directionality. Trends Biochem. Sci. 2014; 39:319–327. [PubMed: 24928762]

Guil S, Esteller M. Cis-acting noncoding RNAs: friends and foes. Nat. Struct. Mol. Biol. 2012; 19:1068–1075. [PubMed: 23132386]

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009; 458:223–227. [PubMed: 19182780]

Herriges MJ, Swarr DT, Morley MP, Rathi KS, Peng T, Stewart KM, Morrisey EE. Long noncoding RNAs are spatially correlated with transcription factors and regulate lung development. Genes Dev. 2014; 28:1363–1379. [PubMed: 24939938]

Jiang W, Zhang D, Bursac N, Zhang Y. WNT3 is a biomarker capable of predicting the definitive endoderm differentiation potential of hESCs. Stem Cell Reports. 2013; 1:46–52. [PubMed: 24052941]

Jiang W, Liu Y, Liu R, Zhang K, Zhang Y. The lncRNA DEANR1 Facilitates Human Endoderm Differentiation by Activating FOXA2 Expression. Cell Rep. 2015; 11:137–148. [PubMed: 25843708]

Kalisz M, Winzi M, Bisgaard HC, Serup P. Even-Skipped Homeobox 1 controls human ES cell differentiation by directly repressing Goosecoid expression. Dev. Biol. 2012; 362:94–103. [PubMed: 22178155]

Kanai-Azuma M, Kanai Y, Gad JM, Tajima Y, Taya C, Kurohmaru M, Sanai Y, Yonekawa H, Yazaki K, Tam PPL, et al. Depletion of definitive gut endoderm in Sox17-null mutant mice. Development. 2002; 129:2367–2379. [PubMed: 11973269]

Kim SW, Yoon S-J, Chuong E, Oyolu C, Wills AE, Gupta R, Baker J. Chromatin and transcriptional signatures for Nodal signaling during endoderm formation in hESCs. Dev. Biol. 2011; 357:492–504. [PubMed: 21741376]

Klattenhoff, C. a; Scheuermann, JC.; Surface, LE.; Bradley, RK.; Fields, P. a; Steinhauser, ML.; Ding, H.; Butty, VL.; Torrey, L.; Haas, S., et al. Braveheart, a Long Noncoding RNA Required for Cardiovascular Lineage Commitment. Cell. 2013:1–14.

Kretz M, Webster DE, Flockhart RJ, Lee CS, Zehnder A, Lopez-Pajares V, Qu K, Zheng GXY, Chow J, Kim GE, et al. Suppression of progenitor differentiation requires the long noncoding RNA ANCR. Genes Dev. 2012; 26:338–343. [PubMed: 22302877]

Kubo A. Development of definitive endoderm from embryonic stem cells in culture. Development. 2004; 131:1651–1662. [PubMed: 14998924]

Lawson KA, Meneses JJ, Pedersen RA. Clonal analysis of epiblast fate during germ layer formation in the mouse embryo. Development. 1991; 113:891–911. [PubMed: 1821858]

Lepoivre C, Belhocine M, Bergon A, Griffon A, Yammine M, Vanhille L, Zacarias-Cabeza J, Garibal M-A, Koch F, Maqbool MA, et al. Divergent transcription is associated with promoters of transcriptional regulators. BMC Genomics. 2013; 14:914. [PubMed: 24365181]

Levesque MJ, Raj A. Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. Nat. Methods. 2013; 10:246–248. [PubMed: 23416756]

Loh KM, Ang LT, Zhang J, Kumar V, Ang J, Auyeong JQ, Lee KL, Choo SH, Lim CYY, Nichane M, et al. Efficient Endoderm Induction from Human Pluripotent Stem Cells by Logically Directing Signals Controlling Lineage Bifurcations. Cell Stem Cell. 2014; 14:237–252. [PubMed: 24412311]

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15:550. [PubMed: 25516281]

Luo S, Lu JY, Liu L, Yin Y, Chen C, Han X, Wu B, Xu R, Liu W, Yan P. Divergent lncRNAs Regulate Gene Expression and Lineage Differentiation in Pluripotent Cells. Cell Stem Cell. 2016:1–16.

Massagué J, Seoane J, Wotton D. Smad transcription factors. 2005:2783–2810.

Mullen AC, Orlando D. a, Newman JJ, Lovén J, Kumar RM, Bilodeau S, Reddy J, Guenther MG, DeKoter RP, Young R. a. Master transcription factors determine cell-type-specific responses to TGF-β signaling. Cell. 2011; 147:565–576. [PubMed: 22036565]

Nostro MC, Sarangi F, Ogawa S, Holtzinger A, Corneo B, Li X, Micallef SJ, Park I-H, Basford C, Wheeler MB, et al. Stage-specific signaling through TGFβ family members and WNT regulates patterning and pancreatic specification of human pluripotent stem cells. Development. 2011; 138:861–871. [PubMed: 21270052]

Ogawa S, Surapisitchat J, Virtanen C, Ogawa M, Niapour M, Sugamori KS, Wang S, Tamblyn L, Guillemette C, Hoffmann E, et al. Three-dimensional culture and cAMP signaling promote the maturation of human pluripotent stem cell-derived hepatocytes. Development. 2013; 140:3285–3296. [PubMed: 23861064]

Osafune K, Caron L, Borowiak M, Martinez RJ, Fitz-Gerald CS, Sato Y, Cowan C. a, Chien KR, Melton D. a. Marked differences in differentiation propensity among human embryonic stem cell lines. Nat. Biotechnol. 2008; 26:313–315. [PubMed: 18278034]

Ponjavic J, Oliver PL, Lunter G, Ponting CP. Genomic and transcriptional colocalization of protein-coding and long non-coding RNA pairs in the developing brain. PLoS Genet. 2009; 5:e1000617. [PubMed: 19696892]

Rivera-Pérez JA, Mallo M, Gendron-Maguire M, Gridley T, Behringer RR. Goosecoid is not an essential component of the mouse gastrula organizer but is required for craniofacial and rib development. Development. 1995; 121:3005–3012. [PubMed: 7555726]

Sauvageau M, Goff LA, Lodato S, Bonev B, Groff AF, Gerhardinger C, Sanchez-Gomez DB, Hacisuleyman E, Li E, Spence M, et al. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. Elife. 2013; 2:1–24.

Scruggs BS, Gilchrist DA, Nechaev S, Muse GW, Burkholder A, Fargo DC, Adelman K. Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. Mol. Cell. 2015; 58:1101–1112. [PubMed: 26028540]

Shimoda M, Kanai-Azuma M, Hara K, Miyazaki S, Kanai Y, Monden M, Miyazaki J. Sox17 plays a substantial role in late-stage differentiation of the extraembryonic endoderm in vitro. J. Cell Sci. 2007; 120:3859–3869. [PubMed: 17940068]

Sigova AA, Mullen AC, Molinie B, Gupta S, Orlando DA, Guenther MG, Almada AE, Lin C, Sharp PA, Giallourakis CC, et al. Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc. Natl. Acad. Sci. U. S. A. 2013; 110:2876–2881. [PubMed: 23382218]

Sim X, Cardenas-Diaz FL, French DL, Gadue P. A Doxycycline-Inducible System for Genetic Correction of iPSC Disease Models. Methods Mol. Biol. 2015

Simon JM, Giresi PG, Davis IJ, Lieb JD. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. Nat. Protoc. 2012; 7:256–267. [PubMed: 22262007]

Tam PP, Beddington RS. The formation of mesodermal tissues in the mouse embryo during gastrulation and early organogenesis. Development. 1987; 99:109–126. [PubMed: 3652985]

The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat. Genet. 2013; 45:580–585. [PubMed: 23715323]

Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res. 2013; 41:e74–e74. [PubMed: 23335781]

Wang P, Xue Y, Han Y, Lin L, Wu C, Xu S, Jiang Z, Xu J, Liu Q, Cao X. The STAT3-binding long noncoding RNA lnc-DC controls human dendritic cell differentiation. Science. 2014; 344:310–313. [PubMed: 24744378]

Xie C, Yuan J, Li H, Li M, Zhao G, Bu D, Zhu W, Wu W, Chen R, Zhao Y. NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res. 2014; 42:D98–D103. [PubMed: 24285305]

Zorn AM, Wells JM. Vertebrate Endoderm Development and Organ Formation. Annu. Rev. Cell Dev. Biol. 2009; 25:221–251. [PubMed: 19575677]

Highlights

- The lncRNA *DIGIT* is induced by activin signaling during endoderm differentiation
- Depletion of *DIGIT* inhibits definitive endoderm differentiation in culture
- *DIGIT* regulates differentiation of human and mouse embryonic stem cells
- *DIGIT* regulates expression of *Goosecoid* to control endoderm differentiation

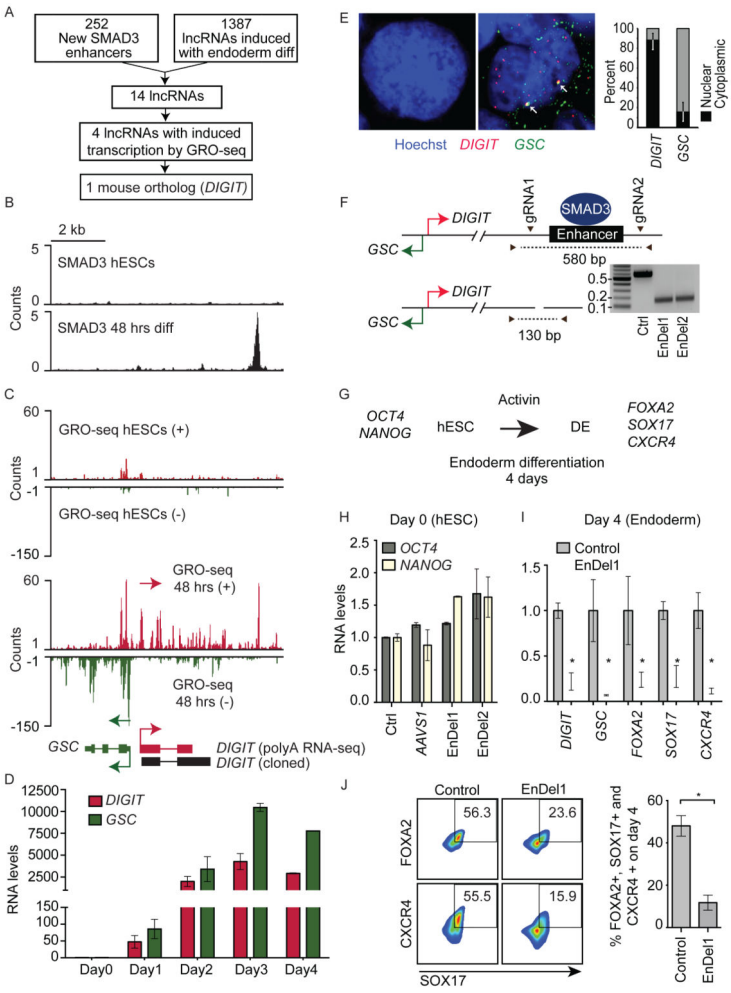


Figure 1. *DIGIT* is divergently transcribed from *GSC* and is activated by an enhancer bound by SMAD3 during endoderm differentiation.

A) Schematic showing the identification of *DIGIT* as a candidate lncRNA that regulates endoderm differentiation. ChIP-seq, RNA-seq and GRO-seq analysis were combined to identify four lncRNAs that were directly targeted by activin signaling, and only one lncRNA had a mouse ortholog.

B) ChIP-seq was performed to identify sites of SMAD3 occupancy in hESCs and after 48 hours of endoderm differentiation. The x-axis represents the linear sequence of genomic DNA, and the y-axis represents the relative number of mapped reads. The genomic scale in kilobases (kb) is indicated above the tracks. The site of SMAD3 occupancy is located 5 kb upstream of the *DIGIT* TSS. The SMAD3 site is enriched for H3K27ac (Tsankov et al. 2015, Figure S1F), which marks active enhancers. The locations of *GSC* and *DIGIT* are shown at the bottom of (C).

C) GRO-Seq was analyzed from Sigova el al., 2013, for hESCs (day 0, top) and hESCs differentiated toward endoderm for 48 hours (bottom). Transcription of the Watson (+) strand is indicated in red and transcription of the Crick (−) strand is indicated in green. Arrows show the direction of transcription. The structure of the *GSC* gene and the predicted structure of the *DIGIT* gene (labeled polyA RNA-seq) are shown below the tracks. *DIGIT*

was cloned after RACE-PCR to define the 5' and 3' ends of the *DIGIT* transcript (Figure S1B), and the structure of the gene encoding this transcript is shown in black (labeled cloned). The cloned *DIGIT* transcript is shown for remainder of the manuscript.

D) *DIGIT* (red) and *GSC* expression (green) were analyzed by qRT-PCR in hESCs (Day 0) and for the first four days of endoderm differentiation. Fold enrichment is indicated on the y-axis, and error bars represent standard deviation.

E) Single-molecule RNA-FISH was performed for hESCs (Day 0, left) and on day 4 of endoderm differentiation (center). Red probes identify *DIGIT* and green probes identify *GSC* mRNA. Nuclei are stained with Hoechst (blue). Each dot represents a transcript, and white arrows indicate two foci of overlapping dots at sites of transcription (Levesque and Raj, 2013). The percentage of transcripts (y-axis) in the nucleus (black) and cytoplasm (white) is shown for *DIGIT* and *GSC* (far right).

F) The positions of two gRNAs flanking the enhancer occupied by SMAD3 (black box) are shown. The TSS of *DIGIT* (red) and *GSC* (green) are indicated on the left. Arrows connected by dotted lines indicate the location of PCR primers. Following deletion of the region occupied by SMAD3, the PCR product decreases from 580 bp to 130 bp (bottom). Genomic PCR was performed on two independent hESC lines with deletion of the SMAD3 enhancer (EnDel1, EnDel2) and is compared to wild-type hESCs (WT).

G) *OCT4* and *NANOG* are markers of undifferentiated hESCs, while *SOX17*, *FOXA2* and *CXCR4* together are markers of DE.

H) Expression of *OCT4* (gray) and *NANOG* (white) was quantified in two hESC controls (Ctrl, *AAVS1*) and two SMAD3 enhancer deletions (EnDel1, EnDel2). Wild-type hESCs are shown as the first control. hESCs that have undergone genome editing to insert a GFP construct into the *AAVS1* locus were used as a second control. Tukey's multiple comparison statistical test showed no significant difference between the means of any pair.

I) hESCs were differentiated towards endoderm for 4 days. RNA levels (y-axis) were quantified for the indicated genes (x-axis) in the *AAVS1* control cells and in hESCs in which the SMAD3 enhancer is deleted (EnDel1). * indicates $p < 0.05$. Analysis of EnDel2 is shown in Figure S1G.

J) Flow cytometry was performed to quantify protein expression of SOX17 (x-axis), FOXA2 and CXCR4 (y-axis) in *AAVS1* control hESCs and hESCs containing the SMAD3 enhancer deletion (EnDel1) after 4 days of endoderm differentiation. The percentage of double positive cells is indicated in the gated areas. The percentage of cells expressing all three markers is quantified for control hESCs and hESCs with the SMAD3 enhancer deletion on the right. * indicates $p < 0.05$. Analysis of EnDel2 is shown in Figure S1H.

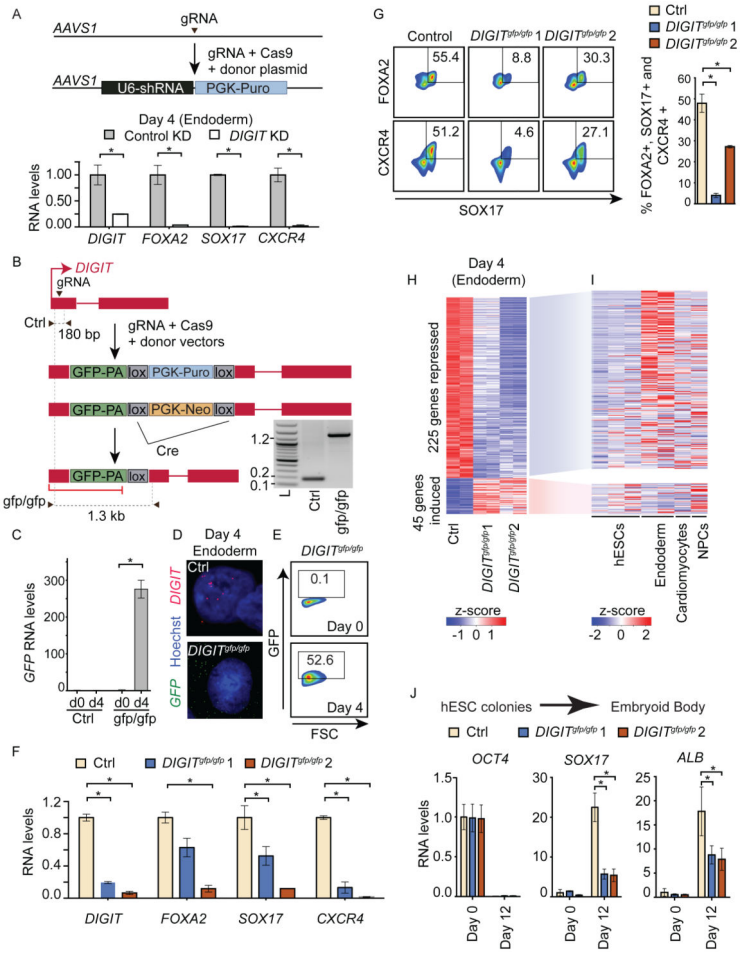


Figure 2. Loss of *DIGIT* expression inhibits definitive endoderm differentiation

A) The CRISPR system was used to insert a vector into the *AAVS1* locus that expresses an shRNA targeting *LACZ* (control KD) or an shRNA targeting *DIGIT* (*DIGIT* KD). hESCs lines were established from single colonies. Control and *DIGIT* KD hESCs were differentiated towards endoderm for 4 days prior to RNA analysis. * indicates $p < 0.05$.

B) Insertion of a GFP-polyA (pA) sequence into the gene encoding *DIGIT* allows transcription at the *DIGIT* locus while inhbiting production of the *DIGIT* transcript. Genomic PCR was performed (lower right) using wild-type hESCs (Ctrl) and cells with the GFP cassette knocked into both copies of *DIGIT* (gfp/gfp) and demonstrates loss of the 180 bp wild-type product and insertion of the GFP-pA cassette (1.3kb). The red line (bottom left) indicates that transcription is terminated after the GFP-pA sequence. Please see Figure S2B and S2C for additional details.

C) GFP mRNA levels (y-axis) were quantified in wild-type hESCs (Ctrl) and *DIGIT*$^{gfp/gfp}$ (gfp/gfp) hESCs prior to endoderm differentiation (d0) and after 4 days of differentiation (d4). * indicates $p < 0.05$.

D) RNA-FISH was performed after 4 days of endoderm differentiation for wild-type hESCs (Ctrl, top) and *DIGIT*$^{gfp/gfp}$ hESCs (bottom). Cells were probed for RNA encoding *GFP* (green), *DIGIT* (red). Nuclei were stained with Hoechst (blue).

E) GFP expression was quantified by flow cytometry in *DIGIT*$^{gfp/gfp}$ hESCs prior to endoderm differentiation (Day 0, top) and after 4 days of differentiation (Day 4, bottom). Green fluorescence is shown on the y-axis and forward scatter (FSC) is shown on the x-axis. The percentage of GFP^{+} cells in each condition is indicated.

F) Wild-type hESCs (Ctrl) and two independently derived *DIGIT*$^{gfp/gfp}$ lines were differentiated towards endoderm for 4 days prior to RNA analysis. RNA levels (y-axis) are shown for the indicated genes (x-axis). * indicates $p < 0.05$.

G) Wildtype (Control), *DIGIT*$^{gfp/gfp}$ 1 and *DIGIT*$^{gfp/gfp}$ 2 cells were differentiated towards endoderm for 4 days prior to analysis by flow cytometry. The percentage of double positive cells is indicated for each condition. The percentage of cells expressing all three endoderm markers is shown on the far right. * indicates $p < 0.05$.

H) Wildtype hESCs, *DIGIT*$^{gfp/gfp}$ 1 and *DIGIT*$^{gfp/gfp}$ 2 hESCs were differentiated towards endoderm for 4 days prior to preparation of RNA-seq libraries. Two replicates for each condition were analyzed. Heat maps display genes that change in expression by at least two fold (FDR <0.05).

I) RNA-seq data were analyzed to quantify expression of the genes repressed (top) and activated (bottom) with loss of *DIGIT* expression (Figure 2H). RNA-seq datasets from hESCs, hESCs undergoing endoderm differentiation, hESCs differentiated towards cardiomyocytes, and hESCs differentiated towards neural progenitor cells (NPCs) are displayed (Jiang et al., 2015; Loh et al., 2014; Palpant et al., 2015; Schwartz et al., 2015).

J) hESCs were harvested after EB formation (Day 0) and 12 days later. RNA levels (y-axis) were quantified for *OCT4* (marker of pluripotency) and *SOX17* and Albumin (markers of endoderm). hESCs that underwent CRISPR targeting to insert a GFP cassette into the *AAVS1* locus were used as controls. * indicates $p < 0.05$.

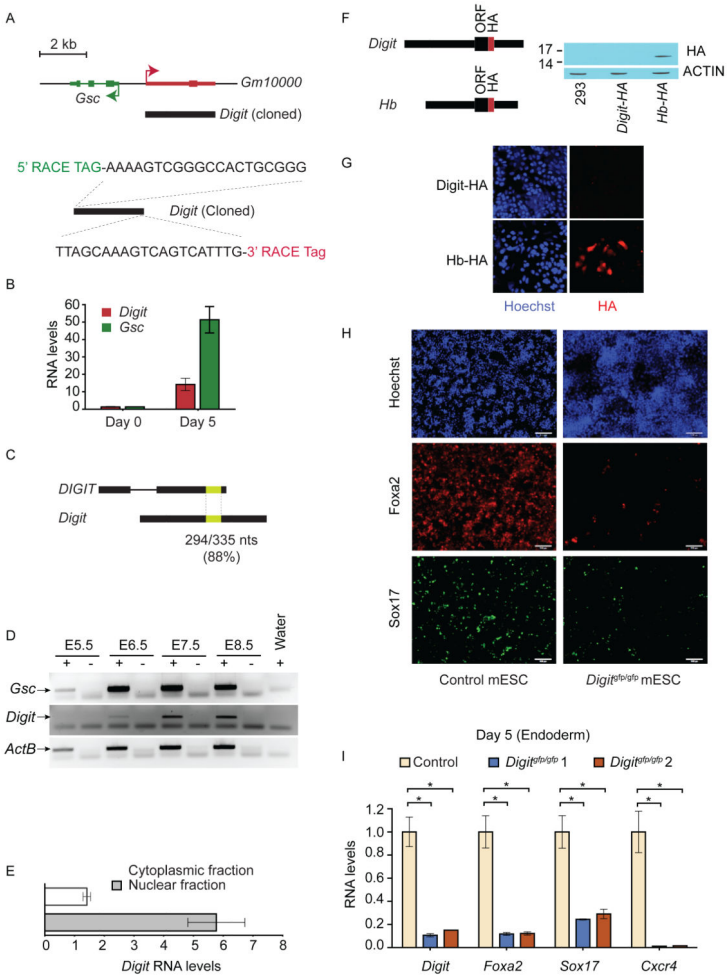


Figure 3. An ortholog of *DIGIT* is expressed in mouse and regulates definitive endoderm differentiation

A) *Digit* (*Gm10000*) is divergent to *Gsc* in mouse. The 5' and 3' ends of the *Digit* transcript were defined by RACE-PCR, and the sequence of the 5' and 3' ends of the transcript are shown below. The full-length transcript cloned from differentiating mESCs using polyA RNA is shown in black.

B) *Digit* levels were quantified in mESCs (Day 0) and mESCs that were differentiated towards endoderm for 5 days.

C) BLAST alignment of the sequences from the human and mouse *DIGIT* orthologs was performed. Yellow areas mark the region of homology in which 294 of 335 nt in *DIGIT* are conserved with *Digit*.

D) RNA was extracted from mouse embryos at the indicated developmental stage. RT-PCR was performed to detect *Gsc* and *Digit*. *ActB* was used as a loading control. A water blank, serving as a negative control is on the far right.

E) RNA was extracted from nuclear (gray) and cytoplasmic fractions (white) on day 5 of differentiation. *Digit* expression in each fraction was quantified in comparison to *Gapdh* mRNA.

F) *Digit* is annotated with a 357 nt ORF. A sequence encoding hemagglutinin (HA) was inserted in frame in the annotated *Digit* and hemoglobin (Hb) ORFs. Plasmids containing the HA-tagged transgenes were transiently transfected into HEK 293T cells. HA expression was quantified by Western blot. The *Digit* product is predicted to be 17kD. ACTIN is shown as a loading control. RNA levels of *Digit* and *Hb* were comparable after transfection (Figure S3D).

G) Plasmids encoding *Digit*-HA and *Hb*-HA were transfected into mESCs 48 hrs prior to analysis by IF microscopy. HA is shown in red, and nuclear staining is shown in blue (Hoechst).

H) Wildtype (control mESC) and $Digit^{gfp/gfp}$ mESCs were differentiated towards endoderm for 5 days prior to analysis by IF microscopy. Images show nuclear staining (HOECHST, top) and expression of Foxa2 (middle) and Sox17 (bottom). $Digit^{gfp/gfp}$ mESCs were created by inserting a GFP-pA sequence to disrupt *Digit* transcription, and these experiments were performed before drug selection markers were removed (Figure S3E).

I) Wildtype mESCs (Ctrl) and two independently derived $Digit^{gfp/gfp}$ lines were differentiated towards endoderm for 5 days prior to RNA analysis. RNA levels (y-axis) are shown for the indicated genes (x-axis). These experiments were performed after transient transfection with a plasmid expressing Cre recombinase to remove the drug resistance cassettes.

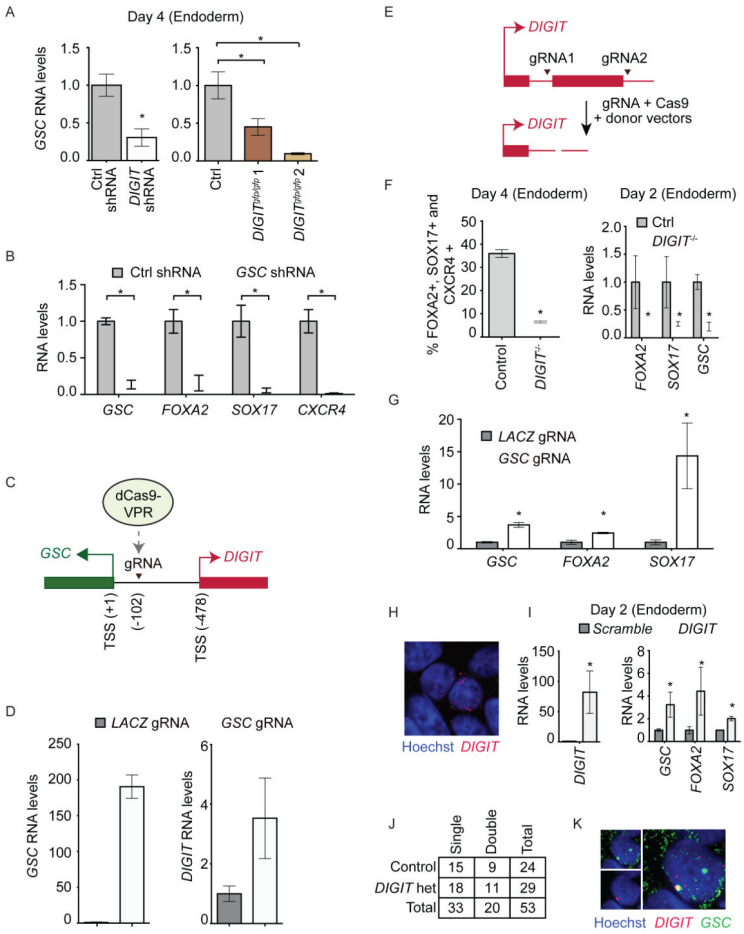


Figure 4. *DIGIT* regulates definitive endoderm differentiation through control of *GSC* in hESCs

A) *GSC* mRNA levels were quantified in hESCs expressing shRNA targeting *GFP* (Ctrl shRNA) and shRNA targeting *DIGIT* (*DIGIT* shRNA) after 4 days of endoderm differentiation (left). *DIGIT* expression is shown in Figure 2A. *GSC* mRNA levels were quantified in $DIGIT^{gfp/gfp}$ 1 and $DIGIT^{gfp/gfp}$ 2 hESC lines and compared to expression in wild-type hESCs after 4 days of differentiation (right). *DIGIT* expression is shown in Figure 2F. * indicates $p < 0.05$.

B) RNA levels were quantified for the indicated genes after 4 days of endoderm differentiation in hESCs containing an shRNA recognizing *LACZ* inserted into the *AAVS1* locus (Ctrl shRNA) and hESCs containing an shRNA targeting *GSC* inserted into the *AAVS1* locus (*GSC* shRNA). * indicates $p < 0.05$.

C) The location of the gRNA used to induce *GSC* expression with dCas9-VPR.

D) hESCs were transfected with a plasmid expressing the gRNA (in C) or a control gRNA recognizing *LACZ* along with a second plasmid expressing dCas9-VPR. *GSC* (left) and *DIGIT* expression (right) was quantified by qRT-PCR after 48 hours. * indicates $p < 0.05$.

E) The positions of two gRNAs flanking the second exon of *DIGIT* and used to create $DIGIT^{-/-}$ hESCs are shown.

F) Control (gray) and $DIGIT^{-/-}$ hESCs (white) were differentiated towards endoderm for 4 days and DE differentiation is quantified by co-expression of FOXA2, SOX17 and CXCR4

by flow cytometry (left). Control and *DIGIT*$^{-/-}$ hESCs were differentiated towards endoderm for 2 days prior to analysis of RNA expression (right). * indicates $p < 0.05$.

G) *DIGIT* $^{-/-}$ hESCs were transiently transfected with dCas9-VPR and the gRNA targeting the *GSC* promoter (white) or a *LACZ* gRNA, which does not target the *GSC* promoter (gray). hESCs were maintained in mTESR1 for 2 days and then differentiated towards endoderm for 2 days before analysis of RNA expression. * indicates $p < 0.05$.

H) hESCs were transfected with a plasmid encoding *DIGIT*. RNA-FISH was performed after 2 days and shows nuclear localization (Hoechst, blue) of *DIGIT* (red) with transient transfection.

I) *DIGIT*$^{-/-}$ hESCs were transfected with a plasmid expressing GFP and either *DIGIT* or a scrambled sequence of *DIGIT* (scrambled). hESCs were transfected and GFP^{+} cells were sorted after 2 days of endoderm differentiation. RNA expression was analyzed for the indicated genes. * indicates $p < 0.05$.

J) RNA-FISH was performed after 4 days of endoderm differentiation for wild-type hESCs (control) and *DIGIT*$^{+/gfp}$ hESCs (*DIGIT* het). Cells were probed for RNA encoding *GSC*. The transcription sites of *GSC* were counted in nuclei with either single or double sites of transcription for both wild-type and *DIGIT*$^{+/gfp}$ hESCs. Fisher exact test showed no significant difference between the ratio of single to double transcription sites between the two genotypes (P-value = 1).

K) RNA-FISH was performed after 4 days of endoderm differentiation for *DIGIT*$^{+/gfp}$ hESCs. Cells were probed for RNA encoding *GSC* (green), *DIGIT* (red). Nuclei were stained with Hoechst (blue).