Identification of novel lncRNAs regulated by the TAL1 complex in T-cell acute lymphoblastic leukemia

Original Article

Identification of novel lncRNAs regulated by the TAL1 complex in T-cell acute lymphoblastic leukemia

Phuong Cao Thi Ngoc,[1,#] Shi Hao Tan,[1,#] Tze King Tan,[1] Min Min Chan,[1] Zhenhua Li,[2] Allen E.J. Yeoh,[2,3] Daniel G. Tenen,[1,4,5] and Takaomi Sanda[1,5,*]

[1]Cancer Science Institute of Singapore, National University of Singapore, 117599, Singapore; [2]Centre for Translational Research in Acute Leukaemia, Department of Paediatrics, Yong Loo Lin School of Medicine, National University of Singapore, 117599, Singapore; [3]VIVA–University Children's Cancer Centre, Khoo Teck Puat–National University Children's Medical Institute, National University Hospital, National University Health System, 119228, Singapore; [4]Harvard Medical School, Boston, Massachusetts 02215, USA; [5]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, 117599, Singapore.

[#] These authors contributed equally to this work.

[*] Corresponding author:
Takaomi Sanda, MD, PhD
Cancer Science Institute of Singapore, National University of Singapore
14 Medical Drive, Centre for Translational Medicine, #12-01, Singapore, 117599
Tel.: +65-6516-1016 Fax: +65-6873-9664
Email: takaomi_sanda@nus.edu.sg

Running title: lncRNA expression profile in T-ALL

Abstract

TAL1/SCL is one of the most prevalent oncogenes in T-cell acute lymphoblastic leukemia (T-ALL). TAL1 and its regulatory partners (GATA3, RUNX1 and MYB) positively regulate each other and coordinately regulate the expression of their downstream target genes in T-ALL cells. However, long non-coding RNAs (lncRNAs) regulated by these factors are largely unknown. Here, we established a bioinformatics pipeline and analyzed RNA-seq datasets with deep coverage to identify lncRNAs regulated by TAL1 in T-ALL cells. Our analysis predicted 57 putative lncRNAs that are activated by TAL1. Many of these transcripts were regulated by GATA3, RUNX1 and MYB in a coordinated manner. We identified two novel transcripts that were activated in multiple T-ALL cell samples but were downregulated in normal thymocytes. One transcript near the *ARID5B* gene locus was specifically expressed in *TAL1*-positive T-ALL cases. The other transcript located between the *FAM49A* and *MYCN* gene locus was also expressed in normal hematopoietic stem cells and T-cell progenitor cells. Additionally, we identified a subset of lncRNAs that were negatively regulated by TAL1 and positively regulated by E-proteins in T-ALL cells. This included a known lncRNA (lnc-OAZ3-2:7) located near the *RORC* gene, which was expressed in normal thymocytes but repressed in *TAL1*-positive T-ALL cells.

Introduction

Rapidly accumulating evidence demonstrates crucial roles for long non-coding RNAs (lncRNAs) in normal tissue development and cancer.[1-6] lncRNAs are generally defined as non-coding transcripts longer than 200 nucleotides that are encoded in the intergenic or intronic regions or overlapping with protein-coding genes.[1] lncRNAs are expressed in a tissue-specific manner, and their expression levels are generally lower than those of protein-coding genes.[7] The lncRNA sequences are less conserved among species compared to protein-coding genes.[8] Importantly, a number of lncRNAs are functional and contribute to various aspects of cell homeostasis.[5, 6] The lncRNAs can regulate gene expression through interactions with genomic DNA, mRNA or protein in the nucleus or cytoplasm,[1, 3, 4] thus participating in the regulatory network in normal and malignant cells.

T-cell acute lymphoblastic leukemia (T-ALL) is a malignant disorder of thymic T-cell precursors.[9] One of the most prevalent oncogenes in T-ALL is the transcription factor *TAL1/SCL*, which is an essential regulator of hematopoiesis.[10, 11] TAL1 is normally expressed in hematopoietic stem cells (HSCs), progenitor cells and erythro-megakaryocytic lineages, but it is silenced during T-cell development.[12, 13] In contrast, TAL1 is ectopically overexpressed in 40-60% of T-ALL cases due to chromosomal translocation, intra-chromosomal rearrangement, or mutations in the non-coding element.[14, 15] In both normal hematopoietic cells and T-ALL cells, TAL1 forms a heterodimer complex with a class I bHLH protein, known as E-proteins (E2A, HEB), and makes a large transcriptional complex with LMO protein (LMO1 or LMO2), LDB1 and GATA protein (GATA1, GATA2 or GATA3). We previously reported that in T-ALL cells, TAL1 and several hematopoietic transcription factors (GATA3, RUNX1 and MYB) regulate the expression of downstream target genes in a coordinated manner.[16] These four factors also co-occupy their own regulatory elements and

positively regulate each other, forming an interconnected auto-regulatory loop.[15, 16] The same structure has been reported in normal HSCs.[17, 18]

In normal developing thymocytes, E-proteins form a homo- or heterodimer and regulate genes required for T-cell differentiation, such as *RAG1*, *RAG2* and *PTCRA*.[19, 20] Deficiency in E-proteins leads to blocked differentiation and development of T-cell malignancies in mice.[21] TAL1 can inhibit the formation of E-protein dimers by forming a more stable transcriptional complex,[22, 23] resulting in the deregulation of E-protein target genes. Together, TAL1 disrupts the transcriptional regulatory program in developing thymocytes by inducing stem cell-like transcriptional circuitry and by blocking T-cell differentiation program. Thus, it is important to identify downstream factors controlled by TAL1 and E-proteins. We have previously identified several protein-coding genes and microRNAs that are directly activated by the TAL1 complex and contribute to T-ALL pathogenesis.[16, 24-27] However, lncRNAs that are regulated by TAL1 and E-proteins in T-ALL cells are entirely unknown.

In this study, we established a bioinformatics pipeline for the prediction of lncRNAs and analyzed the RNA-seq dataset with deep coverage after knockdown of each member of the TAL1 complex in T-ALL cells. This analysis identified 57 putative lncRNAs that are directly regulated by TAL1, including novel transcripts. Many were coordinately regulated by TAL1 and its regulatory partners. Some of them were repressed in normal thymocytes. Additionally, we identified lncRNAs that are differentially controlled by TAL1 and E-proteins.

Materials and Methods

Cell culture and gene knockdown

All T-ALL cell lines were stocked in our laboratory[24, 25] and cultured in RPMI-1640 medium (BioWest) supplemented with 10% fetal bovine serum (FBS; BioWest). All cell lines were confirmed by DNA fingerprinting using the PowerPlex 1.2 system (Promega) in January 2013 and were used from the original stock. Cell lines were regularly tested for mycoplasma contamination. See Supplementary Table 1 for the expression of transcription factor genes defining subgroups of T-ALL. The mouse lymphohematopoietic progenitor cell line EML was cultured in IMDM medium supplemented with 20% FBS and 200 ng/ml Recombinant Murine SCF (Peprotech). Lentiviruses encoding short-hairpin RNAs (shRNAs) targeting each transcription factor were produced and infected the T-ALL cells, as described previously.[24] Detailed procedures are described in the Supplementary Information.

RNA-seq analysis

Total RNA was extracted using a miRNeasy kit (Qiagen) followed by DNase treatment (Turbo DNA-freeTM kit, Ambion). After the depletion of ribosomal RNAs, strand-specific library construction and sequencing of paired-end, 100-bp-long reads by the Illumina HiSeq4000 platform were performed at BGI Biotech Solutions Co., Ltd. (Hong Kong). See Supplementary Table 2 for the number of sequencing reads and quality scores for each sample. All RNA-seq data have been deposited in the NCBI GEO database (GSE97514 and GSE103046).[27] The RNA-seq datasets for human hematopoietic cells reported by Casero et al.[28] and for mouse hematopoietic cells reported by Sun et al.[29] were obtained from the GEO under accession numbers GSE69239 and GSE47819, respectively. Twenty-seven primary T-ALL cases from the Ma-Spore ALL 2003 study (DSRB ref number

2004/00275) were analyzed in this study (Supplementary Table 3). Primary cells were obtained with informed consent after approval by the institutional ethical committee. The RNA-seq dataset for primary samples have been reported[30] and deposited by Jun J. Yang in the EGA database (EGAD00001002151). Genomic sequence for the selected lncRNA retrieved from human (hg19) genome were aligned to mouse (mm10) genome using the UCSC BLAT tool. Details are described in the Supplementary Information.

ChIP-seq datasets

The Jurkat ChIP-seq datasets for TAL1, GATA3, RUNX1, MYB and H3K4me1 have been reported by us and the Young lab[16, 26, 31] and can be downloaded from the GEO (GSE29181, GSE68976 and GSE59657). The Jurkat ChIP-seq dataset for H3K4me3[32] was obtained from the GEO (GSM945267 and GSM945268). H3K27ac ChIP-seq datasets for Jurkat, RPMI-8402, CCRF-CEM, MOLT-4[15] and normal thymus[33] were obtained from the GEO (GSM1296384, GSM1442003, GSM2037781, GSM2037790 and GSM1013125). H3K27ac ChIP-seq datasets for Loucy[34] and DND-41[35] were downloaded from the GEO (GSE74312, and GSE54380). ChIP-seq data and super-enhancers were analyzed as described previously.[24, 31, 36, 37] Details are shown in the Supplementary Information.

Identification of lncRNAs

RNA-seq datasets of T-ALL cell lines were aligned to the hg19 human genome using STAR 2.5.2a and Ensembl database.[38] Alignment files were used for ab initio transcriptome assembly with Cufflinks v2.2.1.[39] A RefSeq GTF file was applied for Cufflinks to guide the assembly. Transcriptome assemblies from each cell line were combined into one annotation

using Cuffmerge from Cufflinks v2.2.1. The novel lncRNA transcripts were included when calculating the FPKM matrix. Details are described in the Supplementary Information.

Correlation analysis with neighboring gene

Two protein-coding genes or a pair of lncRNA and protein-coding gene as neighbors were defined by a minimal distance of less than 100 kb. To calculate the Pearson correlation of two neighbors, FPKM (cut off 1 in at least one T-ALL cell line) was log2 transformed after addition of 0.05.

Gene set enrichment analysis (GSEA)

GSEA[40] was performed for the normalized genes using the log2_Ratio-of-Classes as a metric for ranking genes. TAL1 target genes that were downregulated by knockdown were defined using the RNA-seq dataset and used as a gene set.

Quantitative reverse transcription PCR (qRT-PCR) analysis

Total RNA was harvested using a NucleoSpin RNA kit (Macherey-Nagel) and reverse-transcribed into cDNA using QuantiTect (QIAGEN). Quantitative PCR was performed on a QuantStudio 3 Real-Time PCR System (Thermo Fisher Scientific) using Power SYBR Green PCR Master Mix (Roche). The PCR primer sequences can be found in the Supplementary Information.

Rapid Amplification of cDNA Ends (RACE)

RACE was performed using the SMARTer RACE 5’/3’ Kit (Takara) according to the manufacturer’s instructions. Briefly, total RNA was first extracted from Jurkat cells and

subjected to the SMARTer first-strand cDNA synthesis to generate either 5'-RACE or 3'-RACE ready cDNA. The 5'-RACE and 3'-RACE PCR was then performed using touchdown PCR using specific gene primers (see Supplementary Information), and the generated PCR fragments were purified and cloned into the pRACE vector for subsequent Sanger sequencing using In-Fusion HD cloning provided by the kit.

In vitro transcription and translation assay

The *in vitro* transcription and translation assay was performed using a TNT Quick Coupled Transcription/Translation System (Promega) according to manufacturer's instructions. Briefly, we first cloned the sense or antisense sequence of *XLOC_005968*, the full-length mRNA of *C10ORF107*/*CABCOCO1* or *GATA3* into the expression plasmid pCS2. See Supplementary Information for primer sequences. A total of 500 ng of plasmid DNA was mixed with reagents for the *in vitro* transcription and translation assay and incubated at 30°C for 90 minutes. Western Immunoblotting was performed using the lysates to detect the presence of translated protein. Details are described in the Supplementary Information.

Statistical analysis

The Wald test was used to analyze differences in gene expression among groups by RNA-seq analysis. Two-tailed student t-tests were used to analyze differences in gene expression by qRT-PCR among groups. Adjusted P-values less than 0.05 were considered statistically significant.

Results

Prediction of lncRNAs expressed in T-ALL cells

To identify lncRNAs expressed in T-ALL cells, we performed RNA-seq analysis in eight T-ALL cell lines, which represent several major subgroups of T-ALL (Supplementary Table 1). For RNA-seq analysis, we depleted ribosomal RNAs and included both polyA+ and polyA- RNAs, because many lncRNAs have been known to lack polyA sequences. From this analysis, we obtained over 300 million reads of high quality for each sample (Supplementary Table 2). Using this dataset, we established an analytical pipeline to predict lncRNAs (Figure 1A).

We first mapped all sequence reads to the human genome (hg19), assembled them into transcripts and defined the annotations (i.e., "*XLOC_000001*"). After removing short transcripts, we applied the Slncky pipeline developed by Chen et al.[41] to search for high-confidence lncRNAs ("putative lncRNAs 1"). Because this tool potentially filters out lncRNAs that overlap with protein-coding genes, we included an additional pipeline to predict lncRNAs that share genomic DNA sequences with protein-coding genes. From transcripts overlapping with protein-coding genes which that were rejected from Slncky pipeline, we identified 1,031 additional transcripts ("putative lncRNAs 2"). Finally, we filtered out low-expression lncRNAs for which the FKPM value was below the cut-off of 0.5. From this analysis, we identified a total of 2,236 putative lncRNA loci including 4,857 transcripts (Supplementary Table 4).

To confirm the validity of our pipeline, we compared our result to two public databases and the results for T-ALL samples reported by others. Out of 2,236 putative lncRNA loci, we identified 1,321 (59%) that were annotated in the GENCODE version 25 (total 15,875 loci); 1,055 (47%) were annotated in the LNCipedia version 4.0 (total 48,044 loci).

Importantly, several lncRNAs were previously reported to be differentially expressed among the different subgroups of T-ALL in the dataset by Wallaert et al.[34] For example, *lnc-FAM160A1-6* was shown to be highly expressed in the *TAL1*-positive subgroup, while *lnc-DUSP6-2* and *lnc-TOMM20-2* were reported to be highly expressed in the ETP/immature T-ALL and *TLX*-positive subgroups, respectively.[34] Consistently, our pipeline identified the *lnc-FAM160A1-6* that was highly expressed in four *TAL1*-positive T-ALL cell lines (Jurkat, RPMI-8401, CCRF-CEM and MOLT-4) (Figure 1B, top; and Supplementary Table 1). In contrast, *lnc-DUSP6-2* was specifically expressed in DND-41 cells (Figure 1B, middle), which represents the *TLX*-positive subgroup. Additionally, *lnc-TOMM20-2* was specifically expressed in Loucy cells (Figure 1B, bottom), which represents the ETP/immature T-ALL subgroup.[42] These results indicated that our pipeline accurately predicted lncRNAs known to be expressed in T-ALL.

Characterization of lncRNAs expressed in T-ALL cells

We next analyzed the expression levels of the predicted lncRNAs and positional overlap with protein-coding genes. As generally reported, the RNA expression level of putative lncRNAs was significantly lower than that of protein-coding genes, for which the median log10 maximum FPKM value was 0.17 (Figure 1C). Classification of lncRNAs based on the positional relationship with annotated protein-coding genes indicated that approximately half were located in intergenic regions; the other half were found within the protein-coding genes ("sense", "antisense" or "intronic") or in the opposite direction to nearby protein-coding genes (< 1,000 bp) sharing the same transcriptional start site ("divergent"/bidirectional) (Figure 1D; and see Supplementary Figure 1A for the definitions).

We then analyzed the correlation in RNA expression levels between putative lncRNAs and mRNAs transcribed from neighboring protein-coding genes. This analysis demonstrated a higher positive correlation in the lncRNA-mRNA pairs (blue) than the random pairs (purple: Figure 1E). This level was similar to that in the mRNA-mRNA pairs (red). These results support previous findings reported by others.[7, 43, 44] We did not observe notable differences between the "putative lncRNAs 1" and "putative lncRNAs 2" (Supplementary Figure 1B).

Regulation of lncRNAs by TAL1 and its regulatory partners in T-ALL cells

We next aimed to identify lncRNAs that are directly regulated by TAL1 and its regulatory partners (E2A, HEB, GATA3, RUNX1 and MYB) in T-ALL cells. For this purpose, we performed RNA-seq analysis after lentiviral shRNA knockdown of each of these transcription factors in one of the *TAL1*-positive cell lines (Jurkat). We independently validated the efficiency of knockdown using qRT-PCR analysis (Supplementary Figure 2A). We used the same annotations defined earlier (i.e., "*XLOC_000001*") to compare the expression level of each lncRNA between the control and knockdown samples.

We first selected the putative lncRNAs for which the TAL1 protein binds near or at the lncRNA loci using our ChIP-seq dataset in Jurkat cells.[16] We applied the same cut-off (within 12 kb of the lncRNA loci) that we previously used for identifying protein-coding genes and microRNAs targeted by TAL1.[16, 26] We then filtered transcripts that were significantly downregulated after shRNA knockdown of *TAL1* in the same cells (adjusted p-value<0.05 & log2-fold change < -0.2). A total of 57 putative lncRNAs were selected by these criteria (Figure 2A and Supplementary Table 5). We next analyzed the effect of the regulatory partners of TAL1 (GATA3, RUNX1 and MYB) on the expression of TAL1-regulated lncRNAs

in the same cells. The heatmap analysis illustrated that many of the transcripts were also downregulated after knockdown of *GATA3, RUNX1* or *MYB* (Figure 2B). The gene set enrichment analysis (GSEA) demonstrated a positive correlation in the RNA expression changes after knockdown of each transcription factor (Figure 2C). This trend was essentially similar to that observed for protein-coding genes using microarray analysis in our previous study[16] and by RNA-seq analysis in our current study (Supplementary Figures 2B-D). This indicated that TAL1 coordinately regulates lncRNAs with its regulatory partners in T-ALL cells.

Additionally, we evaluated the overlap of target lncRNAs that are directly regulated by each transcription factor (Figure 2D). As expected, we observed a high level of overlap between TAL1 and other transcription factors, where 49 (86.0%) out of 57 putative lncRNAs were also directly regulated by at least one transcription factor (GATA3, RUNX1 or MYB). In contrast, RUNX1 possessed many other target lncRNAs that were not directly regulated by TAL1, GATA3 or MYB. This result was similar to the previous observation for protein-coding genes,[16] suggesting that RUNX1 could also regulate lncRNA expression independent of the TAL1 complex in T-ALL cells.

Two novel lncRNAs aberrantly activated by the TAL1 complex in T-ALL cells

Next, we sought lncRNAs that are normally downregulated in thymocytes and are activated by TAL1 in T-ALL cells. Among 57 putative lncRNAs selected as direct targets of TAL1 (Figure 2A), we filtered the ones which were also significantly downregulated after knockdown of *GATA3, RUNX1* and *MYB* and were bound by all these factors (Figure 3A). We further narrowed down this list to select lncRNAs associated with super-enhancers, which are defined as clusters of enhancer that exhibit high levels of the H3K27Ac active

histone mark,[31, 36, 37] in the same T-ALL cells but not normal thymus in ChIP-seq analysis (Figures 3B, 3C and Supplementary Figures 3A and 3B). This selection fielded lncRNAs that might be abnormally activated in T-ALL cells. We independently validated the expression of each candidate by qRT-PCR in eight T-ALL cell lines (Supplementary Figures 3C and 3D) and after *TAL1* knockdown in Jurkat cells (Supplementary Figure 3E). We excluded one transcript which was not significantly downregulated by qRT-PCR. Lastly, we focused on the transcripts which have not been previously annotated. These criteria selected two putative lncRNAs (*XLOC_030252* and *XLOC_005968*) (Figures 3B and 3C).

XLOC_030252 was identified as a "putative lncRNAs 1" (Figure 1A) and was located in the intergenic region between the *MYCN* and *FAM49A* gene locus (Figure 3B). The expression was observed in four *TAL1*-positive T-ALL cell lines (Jurkat, RPMI-8402, CCRF-CRM and MOLT-4) and one *TAL1*-negative cell line (Loucy)(Figure 3B and Supplementary Table 3). This transcript was expressed in many of *TAL*-positive subgroup of primary T-ALL, whereas it was also found in *TLX*-positive or other cases (Figures 3B, 3D and Supplementary Figure 3F). On the other hand, *XLOC_005968* was identified as a "putative lncRNAs 2" in our pipeline (Figure 1A), which could not be predicted by the Slncky model because the genomic position of *XLOC_005968* overlapped with the last 3 exons (exons 5-7) of a protein-coding gene *CABCOCO1/C10ORF107* (Figure 3C). We detected the expression of this particular transcript in two *TAL1*-positive T-ALL cell lines (Jurkat and MOLT-4) and two *TAL1*-positive primary leukemia samples (Patients 1 and 2)(Figures 3C and 3E). This transcript was not expressed in *TAL1*-negative T-ALL cases above the cut-off value (Figure 3E). There was no statistical significance observed due to small size of the cohort. Importantly, the full-length of *CABCOCO1* was not expressed in any T-ALL samples based on the RNA-seq analysis (Figure 3C). To verify this result, we designed

specific primers targeting the exon 3 of *CABCOCO1,* which was not shared by *XLOC_005968*, or the exon 7*,* which are shared by *XLOC_005968*, and measured the expression using qRT-PCR analysis. This result revealed that only exon 7 was expressed in Jurkat and MOLT-4 cells (Supplementary Figures 3C and 3D).

Critically, *XLOC_030252* and *XLOC_005968* loci were bound by TAL1 and its regulatory partner proteins (GATA3, RUNX1 and MYB)(Figures 3B and 3C). The expression of both transcripts was downregulated after knockdown of each of these factors (Figures 3F and 3G). Those loci were associated with super-enhancers in *TAL1*-positive T-ALL cell lines but not normal thymus (Figures 3B and 3C, red bars). Together, our results suggested that *XLOC_030252* and *XLOC_005968* are normally repressed in developing thymocytes but are highly activated by TAL1 under super-enhancers in T-ALL cells.

Expression of predicted lncRNAs in normal hematopoietic cells

We next examined whether *XLOC_030252* and *XLOC_005968* were expressed in normal human hematopoietic cells, using a dataset reported by Casero et al.[28] Notably, *XLOC_030252* was expressed in human HSCs and early double-negative (DN) stage thymocyte/progenitor cells ("thy1-2": see figure legend for definition). But, it was downregulated in double-positive (DP) stage ("thy4") and single-positive (SP) stage thymocytes ("thy5, 6")(Figure 4A and 4B), which consist the vast majority of thymocytes. This result was consistent with the H3K27ac ChIP-seq data (Figure 3B). Interestingly, chromosomal position of *XLOC_030252* was conserved between human and mouse (Figure 4C). Using the RNA-seq dataset for mouse HSCs reported by Sun et al.,[29] we were able to detect RNA expression between the *Mycn* and *Fam49a* gene locus (Figure 4C), which corresponded to the human *XLOC_030252* locus. We validated this result by qRT-PCR

analysis using freshly isolated samples. Although the expression pattern was slightly different from human cells, we found that the transcript was expressed in mouse HSCs and progenitor cells and downregulated after DN1 stage thymocytes (Figure 4D), which is similar to the expression pattern of mouse *Tal1* (Supplementary Figure 4A). Notably, the expression of this transcript was significantly downregulated after knockdown of *Tal1* in the mouse hematopoietic progenitor cell line (EML) (Figure 4E). These results indicated that *XLOC_030252* is expressed in normal HSCs and progenitor cells but is downregulated during T-cell development. It is noted that one *TAL1*-negative cell line (Loucy), which represents the ETP subtype of T-ALL,[42] also expressed *XLOC_030252* (Figure 3B). Because this transcript is expressed in T-cell progenitors ("thy1")(Figures 4A and B), which correspond to ETP that is characterized by the absence of *CD1a* and *CD8* expression, activation of this lncRNA in Loucy cells may reflect the stage of normal T-cell development.

In contrast, the expression of *XLOC_005968* was found to be unique to human *TAL1*-positive T-ALL cells. This transcript was not detected in normal human hematopoietic cells in the dataset by Casero et al.[28] (data not shown). The mouse genome possesses the *Cabcoco1* gene near the *Arid5b* gene locus (Supplementary Figure 4B), similar to the human genome, where *XLOC_005968* is located (Figure 3C). However, RNA expression was detected in all exons of *Cabcoco1*, in marked contrast to human T-ALL cells where only *XLOC_005968,* but not the full-length *CABCOCO1*, was expressed (Figure 3C). According to qRT-PCR analysis, we independently confirmed that exon 1 of the *Cabcoco1* gene was expressed in mouse HSCs (Supplementary Figure 4C). Thus, these results indicated that the expression of *XLOC_005968* was specific to *TAL1*-positive human T-ALL cells.

lncRNAs differentially controlled by TAL1 and E-proteins in T-ALL cells

We next analyzed the lncRNAs that are differentially controlled by TAL1 and E-proteins. TAL1 has been reported to inhibit E-protein function by sequestering E-protein dimers, leading to deregulation of E-protein target genes.[22, 23] Other group previously reported that the mutant form of TAL1, which cannot bind to DNA, could still induce T-cell leukemia/lymphoma in mice,[45] indicating that TAL1 can also exert its oncogenic ability through inhibition of E-protein functions. Hence, it would be of interest to identify lncRNAs that are positively regulated by E-proteins and opposed by TAL1 in T-ALL cells. From our current RNA-seq analysis, we confirmed that *RAG1, RAG2* and *PTCRA*, which are known E-protein targets,[19, 20] were upregulated by knockdown of *TAL1* and were further downregulated by knockdown of *E2A* and *HEB* in Jurkat cells (Supplementary Figures 5A-C). Although basal expression levels of these genes are low, this cell line is still useful for analysis of transcripts activated by E-proteins.

First, we selected putative lncRNAs (n=199) that were upregulated by *TAL1* knockdown (Figure 5A and Supplementary Table 6). We next analyzed the expression changes of these transcripts after knockdown of *E2A* and *HEB* in the same cell line. Interestingly, 81 putative lncRNAs were downregulated by *E2A* and/or *HEB* knockdown (top), showing that these are positively regulated by E-proteins and are opposed by TAL1. In contrast, 49 putative lncRNAs were also upregulated by *E2A* and/or *HEB* knockdown (bottom), indicating that these transcripts are cooperatively regulated by TAL1 and E-proteins, possibly as the TAL1-E-protein heterodimer. GSEA showed a mixed pattern of gene regulation, including both positively and negatively correlated genes (Figure 5B).

For example, lncRNAs that were differentially controlled by TAL1 and E-proteins included the *XLOC_001561*, which is transcribed from the antisense strand encoding the *RORC* gene on chromosome 1 (Figure 5C). This transcript has been annotated as *lnc-*

374 *OAZ3-2:7* in the LNCipedia. *RORC* has been reported as an E2A target.[46] Both
375 *XLOC_001561* and *RORC* expression were downregulated by *E2A* and *HEB* knockdown
376 and were upregulated by *TAL1* knockdown in Jurkat cells (Figure 5D and Supplementary
377 Figure 5D). We independently validated this result by qRT-PCR analysis in eight T-ALL cell
378 lines (Supplementary Figure 5E) and after *TAL1* knockdown in Jurkat cells (Supplementary
379 Figure 5F).

380 Notably, DNA binding of TAL1 was not observed at this locus (Supplementary Figure
381 5G), suggesting that TAL1 negatively regulates the expression of *XLOC_001561* and *RORC*
382 in a DNA binding-independent manner, possibly through the inhibition of E-protein dimers.
383 Both *XLOC_001561* and *RORC* were more highly expressed in *TAL1*-negative cell lines
384 (HPB-ALL and TALL-1) than in *TAL1*-positive cases (Figure 5C). Importantly, this putative
385 lncRNA was upregulated in the DP stage ("thy4") and $CD4^+$ or $CD8^+$ single-positive stages
386 of thymocytes ("thy5" and "thy6") (Figure 5E). A similar pattern was observed for the *RORC*
387 gene (Supplementary Figure 5H). These results indicated that *XLOC_001561* and *RORC*
388 are activated by E-proteins in developing T-cells and are repressed by TAL1 in malignant
389 T-cells.

390

391 **Cloning of a novel lncRNA located near the *ARID5B* gene locus**

392 Finally, we validated the protein-coding potential of the candidate lncRNA. We selected
393 *XLOC_005968* (Figure 3C) for this purpose because this transcript was uniquely identified
394 by our pipeline and specifically expressed in human T-ALL cells. The expected size of the
395 transcript was also less than 1 kb, and thus it could be inserted into an expression vector,
396 while the other novel lnRNA (*XLOC_030252*) was too large to be cloned.

We first tried to detect the full-length sequence of *XLOC_005968* by the RACE method using specific primers (Figure 6A and Supplementary Figure 6A). We successfully identified 802-bp nucleotide sequences for this transcript (Supplementary Figure 6B). We then performed an *in vitro* transcription assay followed by an *in vitro* translation analysis to analyze whether it can encode any proteins. We included full-length *CABCOCO1* and *GATA3* cDNA as positive controls. This analysis revealed that the sense or anti-sense strand of *XLOC_005968* did not produce any proteins, while *CABCOCO1* and *GATA3* showed the expected sizes of translated protein (Figure 6B). This result clearly indicated that *XLOC_005968* is a lncRNA and validated our bioinformatics pipeline. Importantly, *XLOC_005968* has not been previously annotated in public databases. Therefore, *XLOC_005968* is a novel lncRNA.

Discussion

Recent advances in RNA research have demonstrated pivotal roles of lncRNAs in normal tissue development and various cancers.[1-5] Several studies have implicated lncRNAs in T-ALL pathogenesis. Trimarchi and Aifantis et al. showed that the lncRNA *LUNAR* is induced by NOTCH1, which is another prevalent oncogene in T-ALL, and is required for T-ALL growth due to its ability to enhance *IGF1R* mRNA expression.[44] More recently, Wallaert and Speleman et al. demonstrated that primary T-ALL cases can be classified into four subgroups based on the lncRNA expression profiles,[34] similar to classification based on expression of protein-coding genes. Additionally, here we identified lncRNAs regulated by TAL1 in T-ALL cells.

For this purpose, we performed RNA-seq analysis with deep coverage. Using this dataset, we first established a pipeline for the prediction of lncRNAs. We used the Slncky model,[41] which is less confounded by evolutional conservation than codon substitution models such as PhyloCSF[47] implemented in previous studies. However, this tool requires a cut-off for filtering out transcripts that overlap with protein-coding genes on the same strand (sense lncRNAs). Hence, we additionally applied four stringent steps to predict sense lncRNAs. For example, we identified a novel lncRNA (*XLOC_005968*) overlapping with the last 3 exons of *CABCOCO1* gene as a sense lncRNA, which could not be found by the Slncky using various minimum overlap values. Importantly, we have shown that *XLOC_005968* does not have the potential to encode a protein, whereas full-length *CABCOCO1* can produce a protein, proving that *XLOC_005968* is indeed a novel lncRNA. Hence, our pipeline accurately predicted lncRNA and is more comprehensive than previously reported approaches. Additionally, we confirmed that expression levels of lncRNAs are lower than for protein-coding genes. Correlation analysis of the expression of

neighboring protein-coding genes showed high correlations in these pairs. All these results support earlier findings reported for other cell types.[7, 43, 44] Notably, among 2,236 putative lncRNA loci identified in our study, 59% and 47% have been already annotated in GENCODE version 25 and the LNCipedia version, respectively. Additionally, 50% of lncRNAs were previously reported to be more highly expressed in *TAL1*-positive T-ALL subgroup than other subgroups.[34] Although these numbers are relatively low, this could possibly be due to differences in the analytical platform (microarray vs RNA-seq), library preparation (e.g., depletion of ribosomal RNAs, inclusion of polyA- RNAs), sequencing depth, annotation or tissue specificity.

Critically, we identified putative lncRNAs that are transcriptionally activated by TAL1 in T-ALL cells. Many were also positively regulated by GATA3, RUNX1 and MYB. Thus, our study demonstrates that TAL1 and its regulatory partners coordinately regulate the expression of both protein-coding genes and lncRNAs. In particular, we highlighted two novel lncRNAs, *XLOC_030252* and *XLOC_005968,* in this study. Both loci were highly activated in T-ALL cells under super-enhancers but were downregulated in normal DP stage thymocytes. Interestingly, *XLOC_030252* was also expressed in HSCs where *TAL1* is physiologically expressed. Our result suggests that these transcripts are normally repressed in developing thymocytes but could be aberrantly activated in T-cells upon ectopic expression of TAL1. Because *XLOC_030252* is also expressed in normal HSCs and progenitor cells, this putative lncRNA may also play a role in normal hematopoiesis. Notably, both *XLOC_030252* and *XLOC_005968* were located near the TAL1-bound region, and the expression of *XLOC_030252* and *XLOC_005968* was significantly downregulated after *TAL1* knockdown. Hence, a possible explanation is that those lncRNAs might be transcribed as a result of enhancer activation. Alternatively, those lncRNAs may act as enhancer RNAs

that control the expression of neighboring genes, as reported for *LUNAR*, which controls *IGF1R* expression. In particular, *XLOC_005968* is located near the *ARID5B* gene, which was recently reported by us as a critical downstream target of TAL1 in T-ALL.[27] ARID5B positively regulates the expression of the TAL1 complex and the *MYC* oncogene in T-ALL cells. ARID5B also promotes the growth and survival of T-ALL cells *in vitro* and *in vivo*. This study implicated ARID5B as a pro-oncogenic factor in the context of *TAL1*-positive T-ALL. Hence, *XLOC_005968* is possibly involved in the regulatory mechanism and in T-cell leukemogenesis. Further investigation is necessary to elucidate their molecular functions.

In this study, we also focused on the regulation of lncRNAs by TAL1 and E-proteins (E2A and HEB). E-protein is a critical heterodimerization partner of TAL1 but also functions as an E-protein dimer to regulate genes required for lymphocyte development.[19, 20] A number of studies have shown that E-proteins serve as tumor suppressors in the context of T-ALL.[21, 48] TAL1 counteracts E-protein functions in T-ALL cells. Therefore, it is of great interest to identify lncRNAs that are differentially controlled by TAL1 and E-proteins. We found that many putative lncRNAs are negatively regulated by TAL1 and positively regulated by E2A and HEB. This included a putative lncRNA *XLOC_001561* near the *RORC* locus. Both the *XLOC_001561* and *RORC* genes were activated by E-proteins and repressed by TAL1 in T-ALL cells, whereas those transcripts are induced during thymocyte development. The *RORC* gene has been reported as an E2A target and has been implicated in normal thymus development.[46, 49, 50] Although further investigations are necessary, our study suggests that *XLOC_001561* is involved in T-cell differentiation and/or may contribute to tumorigenesis when repressed by *TAL1* overexpression.

Acknowledgments

We thank Toshio Suda's laboratory for providing the EML cell line. The research is supported by the National Research Foundation (NRF) Singapore and the Singapore Ministry of Education (MOE) under its Research Centres of Excellence initiative. The research is also supported by the NRF under its Competitive Research Programme (NRF-NRFF2013-02) and the RNA Biology Center at CSI Singapore, NUS, from funding by the Singapore MOE's Tier 3 grants (MOE2014-T3-1-006). A.E.J.Y. is supported by the National Medical Research Council, Singapore (NMRC/CSA/0053/2013).

Authorship contributions

P.C.T.N., S.H.T., T.K.T. and T.S. analyzed the results; S.H.T. and M.M.C. performed the experiments; Z.L. and A.E.J.Y. analyzed the data for primary samples; P.C.T.N., S.H.T., D.G.T. and T.S. designed the research; P.C.T.N. and T.S. wrote the paper.

Conflict of interest

The authors declare no competing financial interests.

References

1. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. *Annu Rev Biochem* 2012; **81:** 145-166.
2. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D*, et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 2009 Mar 12; **458**(7235)**:** 223-227.
3. Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. *Nature reviews Genetics* 2009 Mar; **10**(3)**:** 155-159.
4. Fatica A, Bozzoni I. Long non-coding RNAs: new players in cell differentiation and development. *Nature reviews Genetics* 2014 Jan; **15**(1)**:** 7-21.
5. Huarte M. The emerging role of lncRNAs in cancer. *Nat Med* 2015 Nov; **21**(11)**:** 1253-1261.
6. Schmitt AM, Chang HY. Long Noncoding RNAs in Cancer Pathways. *Cancer Cell* 2016 Apr 11; **29**(4)**:** 452-463.
7. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H*, et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* 2012 Sep; **22**(9)**:** 1775-1789.
8. Ulitsky I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nature reviews Genetics* 2016 Oct; **17**(10)**:** 601-614.
9. Pui CH, Robison LL, Look AT. Acute lymphoblastic leukaemia. *Lancet* 2008 Mar 22; **371**(9617)**:** 1030-1043.
10. Porcher C, Chagraoui H, Kristiansen MS. SCL/TAL1: a multifaceted regulator from blood development to disease. *Blood* 2017 Apr 13; **129**(15)**:** 2051-2060.
11. Sanda T, Leong WZ. TAL1 as a master oncogenic transcription factor in T-cell acute lymphoblastic leukemia. *Exp Hematol* 2017 Jun 24.
12. Mouthon MA, Bernard O, Mitjavila MT, Romeo PH, Vainchenker W, Mathieu-Mahul D. Expression of tal-1 and GATA-binding proteins during human hematopoiesis. *Blood* 1993 Feb 01; **81**(3)**:** 647-655.
13. Herblot S, Steff AM, Hugo P, Aplan PD, Hoang T. SCL and LMO1 alter thymocyte differentiation: inhibition of E2A-HEB function and pre-T alpha chain expression. *Nat Immunol* 2000 Aug; **1**(2)**:** 138-144.
14. Look AT. Oncogenic transcription factors in the human acute leukemias. *Science* 1997 Nov 7; **278**(5340)**:** 1059-1064.
15. Mansour MR, Abraham BJ, Anders L, Berezovskaya A, Gutierrez A, Durbin AD*, et al.* Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* 2014 Dec 12; **346**(6215)**:** 1373-1377.
16. Sanda T, Lawton LN, Barrasa MI, Fan ZP, Kohlhammer H, Gutierrez A*, et al.* Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia. *Cancer Cell* 2012 Aug 14; **22**(2)**:** 209-221.
17. Cole MF, Young RA. Mapping key features of transcriptional regulatory circuitry in embryonic stem cells. *Cold Spring Harb Symp Quant Biol* 2008; **73:** 183-193.
18. Moignard V, Woodhouse S, Fisher J, Gottgens B. Transcriptional hierarchies regulating early blood cell development. *Blood Cells Mol Dis* 2013 Dec; **51**(4)**:** 239-247.
19. Kee BL. E and ID proteins branch out. *Nat Rev Immunol* 2009 Mar; **9**(3)**:** 175-184.

20. Murre C. Helix-loop-helix proteins and lymphocyte development. *Nat Immunol* 2005 Nov; **6**(11)**:** 1079-1086.
21. Bain G, Engel I, Robanus Maandag EC, te Riele HP, Voland JR, Sharp LL*, et al.* E2A deficiency leads to abnormalities in alphabeta T-cell development and to rapid development of T-cell lymphomas. *Mol Cell Biol* 1997 Aug; **17**(8)**:** 4782-4791.
22. Goldfarb AN, Lewandowska K. Inhibition of cellular differentiation by the SCL/tal oncoprotein: transcriptional repression by an Id-like mechanism. *Blood* 1995 Jan 15; **85**(2)**:** 465-471.
23. El Omari K, Hoosdally SJ, Tuladhar K, Karia D, Hall-Ponsele E, Platonova O*, et al.* Structural basis for LMO2-driven recruitment of the SCL:E47bHLH heterodimer to hematopoietic-specific transcriptional targets. *Cell Rep* 2013 Jul 11; **4**(1)**:** 135-147.
24. Liau WS, Tan SH, Ngoc PC, Wang CQ, Tergaonkar V, Feng H*, et al.* Aberrant activation of the GIMAP enhancer by oncogenic transcription factors in T-cell acute lymphoblastic leukemia. *Leukemia* Jan 27.
25. Tan SH, Yam AW, Lawton LN, Wong RW, Young RA, Look AT*, et al.* TRIB2 reinforces the oncogenic transcriptional program controlled by the TAL1 complex in T-cell acute lymphoblastic leukemia. *Leukemia* 2016 Apr; **30**(4)**:** 959-962.
26. Mansour MR, Sanda T, Lawton LN, Li X, Kreslavsky T, Novina CD*, et al.* The TAL1 complex targets the FBXW7 tumor suppressor by activating miR-223 in human T cell acute lymphoblastic leukemia. *J Exp Med* 2013 Jul 29; **210**(8)**:** 1545-1557.
27. Leong WZ, Tan SH, Ngoc PCT, Amanda S, Yam AWY, Liau WS*, et al.* ARID5B as a critical downstream target of the TAL1 complex that activates the oncogenic transcriptional program and promotes T-cell leukemogenesis. *Genes Dev* 2017 Dec 1; **31**(23-24)**:** 2343-2360.
28. Casero D, Sandoval S, Seet CS, Scholes J, Zhu Y, Ha VL*, et al.* Long non-coding RNA profiling of human lymphoid progenitor cells reveals transcriptional divergence of B cell and T cell lineages. *Nat Immunol* 2015 Dec; **16**(12)**:** 1282-1291.
29. Sun D, Luo M, Jeong M, Rodriguez B, Xia Z, Hannah R*, et al.* Epigenomic profiling of young and aged HSCs reveals concerted changes during aging that reinforce self-renewal. *Cell Stem Cell* 2014 May 1; **14**(5)**:** 673-688.
30. Qian M, Zhang H, Kham SK, Liu S, Jiang C, Zhao X*, et al.* Whole-transcriptome sequencing identifies a distinct subtype of acute lymphoblastic leukemia with predominant genomic abnormalities of EP300 and CREBBP. *Genome Res* 2017 Feb; **27**(2)**:** 185-195.
31. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA*, et al.* Super-enhancers in the control of cell identity and disease. *Cell* 2013 Nov 7; **155**(4)**:** 934-947.
32. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E*, et al.* The accessible chromatin landscape of the human genome. *Nature* 2012 Sep 06; **489**(7414)**:** 75-82.
33. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A*, et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* 2010 Oct; **28**(10)**:** 1045-1048.
34. Wallaert A, Durinck K, Van Loocke W, Van de Walle I, Matthijssens F, Volders PJ*, et al.* Long noncoding RNA signatures define oncogenic subtypes in T-cell acute lymphoblastic leukemia. *Leukemia* 2016 Sep; **30**(9)**:** 1927-1930.

35. Knoechel B, Roderick JE, Williamson KE, Zhu J, Lohr JG, Cotton MJ*, et al.* An epigenetic mechanism of resistance to targeted therapy in T cell acute lymphoblastic leukemia. *Nat Genet* 2014 Apr; **46**(4)**:** 364-370.
36. Kwiatkowski N, Zhang T, Rahl PB, Abraham BJ, Reddy J, Ficarro SB*, et al.* Targeting transcription regulation in cancer with a covalent CDK7 inhibitor. *Nature* 2014 Jul 31; **511**(7511)**:** 616-620.
37. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH*, et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 2013 Apr 11; **153**(2)**:** 307-319.
38. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S*, et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013 Jan 01; **29**(1)**:** 15-21.
39. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ*, et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010 May; **28**(5)**:** 511-515.
40. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005 Oct 25; **102**(43)**:** 15545-15550.
41. Chen J, Shishkin AA, Zhu X, Kadri S, Maza I, Guttman M*, et al.* Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. *Genome biology* 2016 Feb 02; **17:** 19.
42. Chonghaile TN, Roderick JE, Glenfield C, Ryan J, Sallan SE, Silverman LB*, et al.* Maturation stage of T-cell acute lymphoblastic leukemia determines BCL-2 versus BCL-XL dependence and sensitivity to ABT-199. *Cancer Discov* 2014 Sep; **4**(9)**:** 1074-1087.
43. Kornienko AE, Dotter CP, Guenzl PM, Gisslinger H, Gisslinger B, Cleary C*, et al.* Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome biology* 2016 Jan 29; **17:** 14.
44. Trimarchi T, Bilal E, Ntziachristos P, Fabbri G, Dalla-Favera R, Tsirigos A*, et al.* Genome-wide mapping and characterization of Notch-regulated long noncoding RNAs in acute leukemia. *Cell* 2014 Jul 31; **158**(3)**:** 593-606.
45. O'Neil J, Billa M, Oikemus S, Kelliher M. The DNA binding activity of TAL-1 is not required to induce leukemia/lymphoma in mice. *Oncogene* 2001 Jun 28; **20**(29)**:** 3897-3905.
46. Miyazaki M, Rivera RR, Miyazaki K, Lin YC, Agata Y, Murre C. The opposing roles of the transcription factor E2A and its antagonist Id3 that orchestrate and enforce the naive fate of T cells. *Nat Immunol* 2011 Aug 21; **12**(10)**:** 992-1001.
47. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* 2011 Jul 01; **27**(13)**:** i275-282.
48. O'Neil J, Shank J, Cusson N, Murre C, Kelliher M. TAL1/SCL induces leukemia by inhibiting the transcriptional activity of E47/HEB. *Cancer Cell* 2004 Jun; **5**(6)**:** 587-596.
49. Kurebayashi S, Ueda E, Sakaue M, Patel DD, Medvedev A, Zhang F*, et al.* Retinoid-related orphan receptor gamma (RORgamma) is essential for lymphoid

organogenesis and controls apoptosis during thymopoiesis. *Proc Natl Acad Sci U S A* 2000 Aug 29; **97**(18)**:** 10132-10137.

50. Sun Z, Unutmaz D, Zou YR, Sunshine MJ, Pierani A, Brenner-Morton S*, et al.* Requirement for RORgamma in thymocyte survival and lymphoid organ development. *Science* 2000 Jun 30; **288**(5475)**:** 2369-2373.

643 **Figure Legends**

644 **Figure 1. Identification of lncRNAs expressed in T-ALL cells. (A)** Schematic diagram
645 of the analytical pipeline. See Materials and Methods for details. **(B)** RNA-seq gene tracks
646 showing the expression and genomic loci of three representative lncRNAs expressed in T-
647 ALL cell lines: *XLOC_040800/lnc-FAM160A1-6 (top), XLOC_012951/lnc-DUSP6-2* (middle)
648 and *XLOC_005308/lnc-TOMM20-2* (bottom). The x-axis indicates the linear sequence of
649 genomic DNA, and the y-axis indicates the number of mapped reads per million. The black
650 horizontal bar indicates the genomic scale in kilobases (kb). Black boxes in the gene map
651 represent exons, and arrows indicate the location and direction of the transcriptional start
652 site. **(C)** Violin plots showing expression levels of protein-coding genes and lncRNAs in T-
653 ALL. The y-axis indicates the maximum FPKM values (log10) of protein-coding genes and
654 lncRNAs among eight T-ALL cell lines. Black boxes represent median and quartile
655 expression levels. **(D)** Pie chart showing the percentage of each class of lncRNA. See
656 Supplementary Figure 1A for details of classification. **(E)** Density distributions showing
657 pairwise Pearson correlations between loci. Pairs were selected for lncRNAs or protein-
658 coding genes (mRNA) with their neighboring protein-coding genes. As a control, 1,000
659 mRNA-mRNA pairs were randomly selected. The y-axis indicates density at each
660 correlation value (x-axis).

661

662 **Figure 2. Regulation of lncRNAs by TAL1 and its regulatory partners in Jurkat cells.**
663 **(A, B)** Expression of TAL1 target lncRNAs after knockdown (KD) of transcription factor
664 genes in Jurkat cells. Jurkat cells were transduced with control shRNA (*GFP* or *Luc*) or
665 shRNA targeting a transcription factor gene (*TAL1, GATA3, RUNX1* or *MYB*) by lentivirus
666 infection. Two different shRNAs (#1 and #2) were used for each gene. Heatmap images

represent expression levels of 57 selected lncRNAs in two controls and two knockdown samples. **(C)** GSEA was performed to determine the correlation of lncRNAs positively regulated by TAL1 (n=57) and gene expression changes upon KD of *GATA3*, *RUNX1* and *MYB*. GSEA plot indicates the degree to which TAL1 target lncRNAs are overrepresented at the extreme left (downregulated by KD) or right (upregulated by KD) of the entire ranked list. Each solid bar represents one lncRNA gene within the gene set. Normalized enrichment scores (NES) and p-values are shown. **(D)** Venn diagram represents the number of lncRNAs directly regulated by each transcription factor in Jurkat cells.

Figure 3. lncRNAs aberrantly activated by TAL1 in T-ALL cells. (A) Schematic diagram of selection criteria. **(B,C)** ChIP-seq gene tracks showing binding locations of TAL1, its regulatory partners (GATA3, RUNX1 and MYB) and RNA polymerase 2 (RNAP2) at *XLOC_030252* (B) and *XLOC_005968* (C) loci in Jurkat cells. ChIP-seq data for H3K27ac in six T-ALL cell lines and normal thymus (bulk) and RNA-seq data in eight T-ALL cell lines and two primary samples (Patients 1 and 2) are also shown. The y-axis indicates the total number of mapped ChIP-seq reads per million. +, positive strand; -, negative strand, analyzed by strand-specific RNA-seq. See Figure 1B legend for details of RNA-seq gene tracks. Two scales were included for the positive strand in Figure 3C, because expression levels of *XLOC_005968* and protein-coding genes were highly different. **(D,E)** RNA expressions of *XLOC_030252* (D) and *XLOC_005968* (E) in 27 primary T-ALL cases reported by Qian et al[30]. T-ALL samples were classified into *TAL* (*TAL1* or *TAL2* positive), *TLX* (*TLX1* or *TLX3* positive) or others. Values are shown by the transcripts per kilobase million (TPM). There was no statistical significance observed due to small size of the cohort. **(F,G)** RNA expression changes of *XLOC_030252* (F) and *XLOC_005968* (G) after

691 knockdown of *TAL1, GATA3, RUNX1* and *MYB* in Jurkat cells analyzed by RNA-seq. Values
692 are shown in log2 fold-change (FC).
693
694 **Figure 4. Expression of *XLOC_030252* in human and mouse hematopoietic cells. (A,**
695 **B)** Expression of *XLOC_030252* in different stages of human hematopoietic cells. RNA-seq
696 gene tracks showing RNA expression in human hematopoietic cells at *XLOC_030252* locus
697 (A). Each cell type was defined by Casero et al.[28] as follows: HSC, $CD34^{+}CD38^{neg}lin^{neg}$;
698 lymphoid-primed multipotent progenitors (LMPP),
699 $CD34^{+}CD38^{+}CD10^{neg}CD45RA^{+}CD62L^{high}lin^{neg}$; common lymphoid progenitor (CLP),
700 $CD34^{+}CD38^{+}CD10^{+}CD45RA^{+}lin^{neg}$; Thy1, $CD34^{+}CD7^{neg}CD1a^{neg}CD4^{neg}CD8^{neg}$; Thy2,
701 $CD34^{+}CD7^{+}CD1a^{neg}CD4^{neg}CD8^{neg}$; Thy3, $CD34^{+}CD7^{+}CD1a^{+}CD4^{neg}CD8^{neg}$; Thy4,
702 $CD4^{+}CD8^{+}$; Thy5, $CD3^{+}CD4^{+}CD8^{neg}$; and Thy6, $CD3^{+}CD4^{neg}CD8^{+}$. Two scales were
703 included because expression levels of *XLOC_030252* and protein-coding genes were
704 different. Expression level of *XLOC_030252* in duplicate samples reported by Casero et al.
705 [28] were shown by the TPM values (B). **(C)** RNA-seq dataset reported by Sun et al.[29] was
706 used to analyze RNA expression in mouse HSCs harvested at different ages (4 and 24
707 months) at the genomic locus conserved with human *XLOC_030252* sequence. **(D)** The
708 RNA expression of putative paralog of *XLOC_030252* in different stages of mouse
709 hematopoietic cells were measured by qRT-PCR and normalized to *β2M* expression. Long-
710 term HSC (LT-SHC), short-term HSC (ST-HSC), multi potent progenitor (MPP), common
711 lymphoid progenitor (CLP), DN1-4, DP, CD4 single positive and CD8 single positive cells.
712 Fold-change values compared with LT-HSC are shown as the mean ±standard deviation
713 (SD) of duplicate samples. See Supplementary Materials for the definition of each
714 population. **(E)** Mouse lymphoid progenitor cell line EML was transduced with control

715 shRNA (sh*Luc*) or shRNA targeting mouse *Tal1* (sh1 and sh2) by lentivirus infection. RNA
716 expression of putative paralog of human *XLOC_030252* was measured by qRT-PCR and
717 normalized to *β2M* expression. Fold-change values compared with the control are shown
718 as the mean ± SD of duplicate samples. **p<0.01 by two-sample, two-tailed t-test.
719
720 **Figure 5. lncRNAs differentially controlled by TAL1 and E-proteins in T-ALL cells. (A)**
721 Jurkat cells were transduced with control shRNA (sh*GFP* or sh*Luc*) or shRNA targeting a
722 transcription factor gene (*TAL1, E2A* or *HEB*) by lentivirus infection. Two different shRNAs
723 (#1 and #2) were used for each gene. Heatmap images represent expression levels of 199
724 selected putative lncRNAs in two controls and two knockdown (KD) samples. **(B)** GSEA
725 was performed to determine the correlation of lncRNAs negatively regulated by TAL1
726 (n=199) and gene expression changes on KD of *E2A* and *HEB*. See Figure 2C legend for
727 details. **(C)** RNA-seq gene track showing expression and genomic loci of *XLOC_001561/*
728 *lnc-OAZ3-2:7* and *RORC* gene in T-ALL cell lines and normal hematopoietic cells. See
729 Figure 1B legend for details. **(D)** RNA expression changes of *XLOC_001561/lnc-OAZ3-2:7*
730 after knockdown of *TAL1, E2A* and *HEB* in Jurkat cells. **(E)** RNA expression levels of
731 *XLOC_001561/lnc-OAZ3-2:7* in different stages of normal hematopoietic cells. See Figure
732 4A legend for details.
733
734 **Figure 6. *XLOC_005968* is a novel lncRNA expressed in T-ALL cells. (A)** The 5' and 3'
735 RACE primers were designed to target exons 3 and 1 of the *XLOC_005968* transcript,
736 respectively. **(B)** *In vitro* transcription and translation assay was performed using expression
737 vectors encoding sense or anti-sense strands of *XLOC_005968*. The cDNA for full-length

738 *CABCOCO1* and *GATA3* were used as positive controls. Proteins were visualized by
739 Western blot analysis using Strepavidin-HRP.

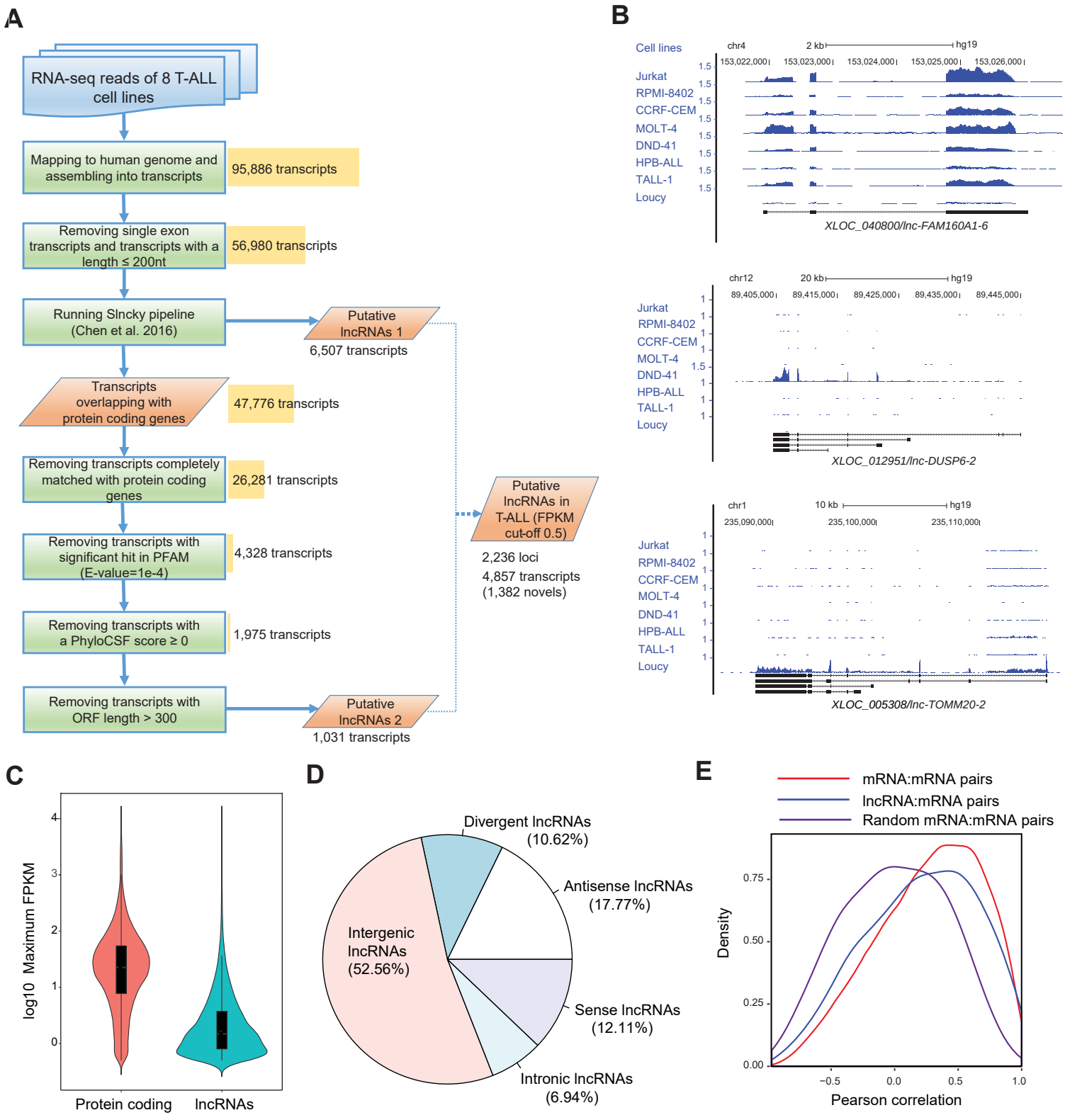

A
RNA-seq reads of 8 T-ALL cell lines
Mapping to human genome and assembling into transcripts
95,886 transcripts
Removing single exon transcripts and transcripts with a length ≤ 200nt
56,980 transcripts
Running SIncky pipeline (Chen et al. 2016)
Putative lncRNAs 1
6,507 transcripts
Transcripts overlapping with protein coding genes
47,776 transcripts
Removing transcripts completely matched with protein coding genes
26,281 transcripts
Removing transcripts with significant hit in PFAM (E-value=1e-4)
4,328 transcripts
Removing transcripts with a PhyloCSF score ≥ 0
1,975 transcripts
Removing transcripts with ORF length > 300
Putative lncRNAs 2
1,031 transcripts
Putative lncRNAs in T-ALL (FPKM cut-off 0.5)
2,236 loci
4,857 transcripts
(1,382 novels)
B
Cell lines
Jurkat
RPMI-8402
CCRF-CEM
MOLT-4
DND-41
HPB-ALL
TALL-1
Loucy
chr4
2 kb
hg19
153,022,000
153,023,000
153,024,000
153,025,000
153,026,000
XLOC_040800/lnc-FAM160A1-6
chr12
20 kb
89,405,000
89,415,000
89,425,000
89,435,000
89,445,000
XLOC_012951/lnc-DUSP6-2
chr1
10 kb
235,090,000
235,100,000
235,110,000
XLOC_005308/lnc-TOMM20-2
C
log10 Maximum FPKM
Protein coding
lncRNAs
D
Divergent lncRNAs (10.62%)
Antisense lncRNAs (17.77%)
Intergenic lncRNAs (52.56%)
Sense lncRNAs (12.11%)
Intronic lncRNAs (6.94%)
E
mRNA:mRNA pairs
lncRNA:mRNA pairs
Random mRNA:mRNA pairs
Density
Pearson correlation

Figure 1

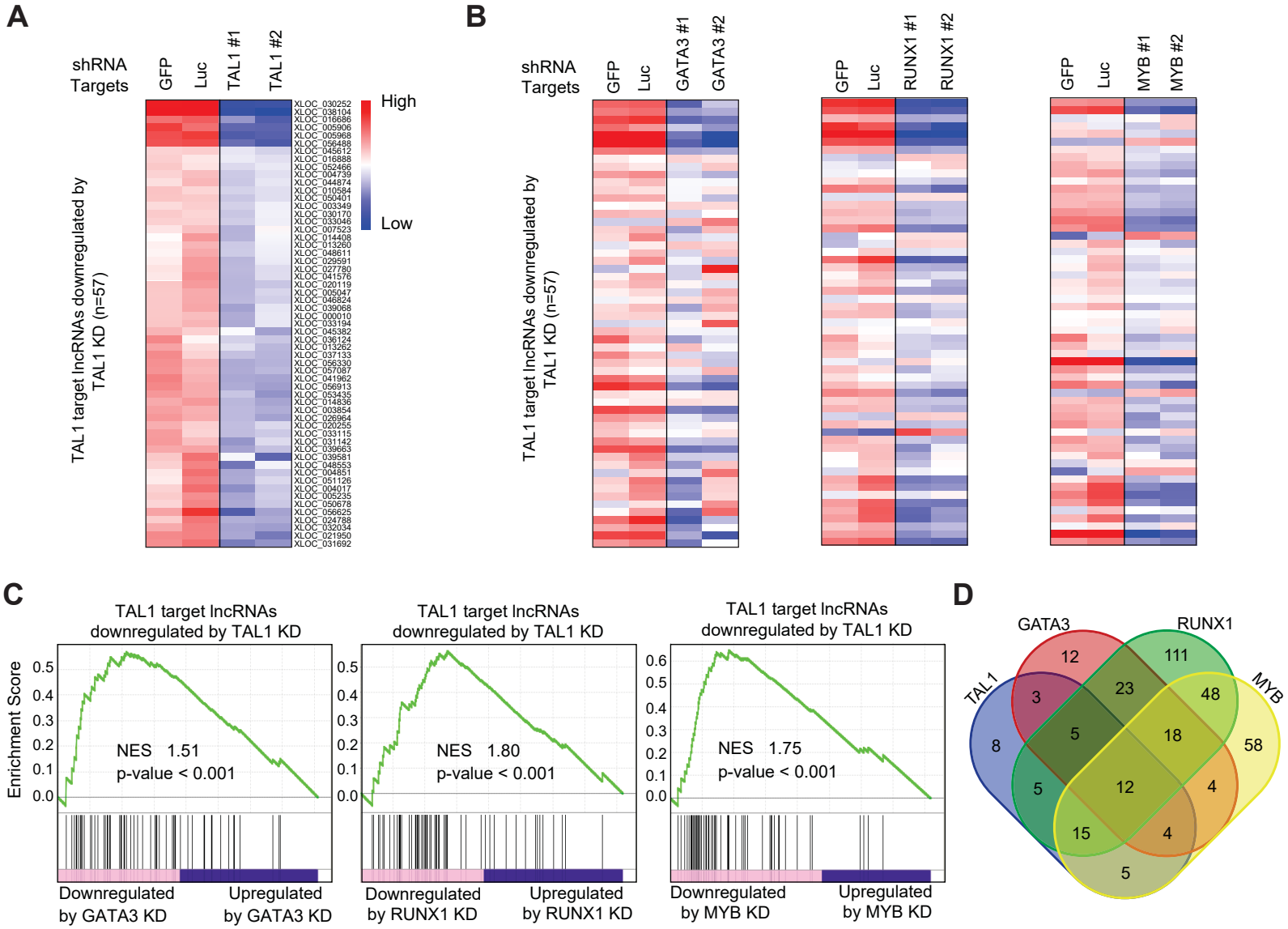

A
shRNA Targets
GFP
Luc
TAL1 #1
TAL1 #2
High
Low
TAL1 target lncRNAs downregulated by TAL1 KD (n=57)
B
shRNA Targets
GFP
Luc
GATA3 #1
GATA3 #2
RUNX1 #1
RUNX1 #2
MYB #1
MYB #2
TAL1 target lncRNAs downregulated by TAL1 KD (n=57)
C
TAL1 target lncRNAs downregulated by TAL1 KD
Enrichment Score
NES 1.51
p-value < 0.001
Downregulated by GATA3 KD
Upregulated by GATA3 KD
NES 1.80
p-value < 0.001
Downregulated by RUNX1 KD
Upregulated by RUNX1 KD
NES 1.75
p-value < 0.001
Downregulated by MYB KD
Upregulated by MYB KD
D
GATA3
RUNX1
TAL1
MYB
12
111
3
23
48
8
5
18
58
5
12
4
15
4
5

Figure 2

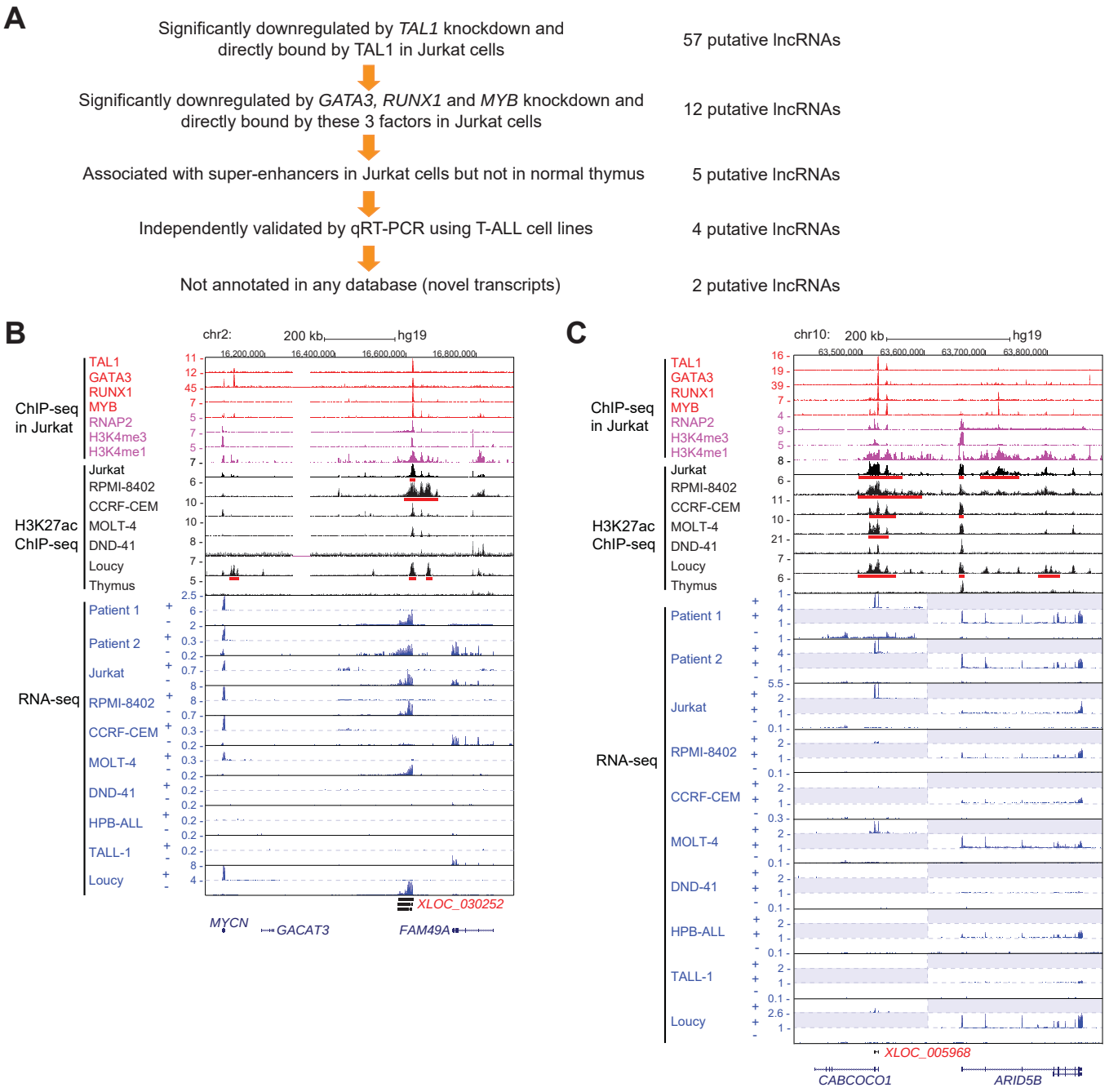

A
Significantly downregulated by TAL1 knockdown and directly bound by TAL1 in Jurkat cells
57 putative lncRNAs
Significantly downregulated by GATA3, RUNX1 and MYB knockdown and directly bound by these 3 factors in Jurkat cells
12 putative lncRNAs
Associated with super-enhancers in Jurkat cells but not in normal thymus
5 putative lncRNAs
Independently validated by qRT-PCR using T-ALL cell lines
4 putative lncRNAs
Not annotated in any database (novel transcripts)
2 putative lncRNAs
B
chr2:
200 kb
hg19
ChIP-seq in Jurkat
TAL1
GATA3
RUNX1
MYB
RNAP2
H3K4me3
H3K4me1
H3K27ac ChIP-seq
Jurkat
RPMI-8402
CCRF-CEM
MOLT-4
DND-41
Loucy
Thymus
RNA-seq
Patient 1
Patient 2
Jurkat
RPMI-8402
CCRF-CEM
MOLT-4
DND-41
HPB-ALL
TALL-1
Loucy
XLOC_030252
MYCN
GACAT3
FAM49A
C
chr10:
200 kb
hg19
XLOC_005968
CABCOCO1
ARID5B
D
XLOC_030252
TPM
TAL
TLX
Others
E
XLOC_005968
TPM
TAL
TLX
Others
F
XLOC_030252
log2FC
TAL1 GATA3 RUNX1 MYB
Knockdown
G
XLOC_005968
log2FC
TAL1 GATA3 RUNX1 MYB
Knockdown

Figure 3

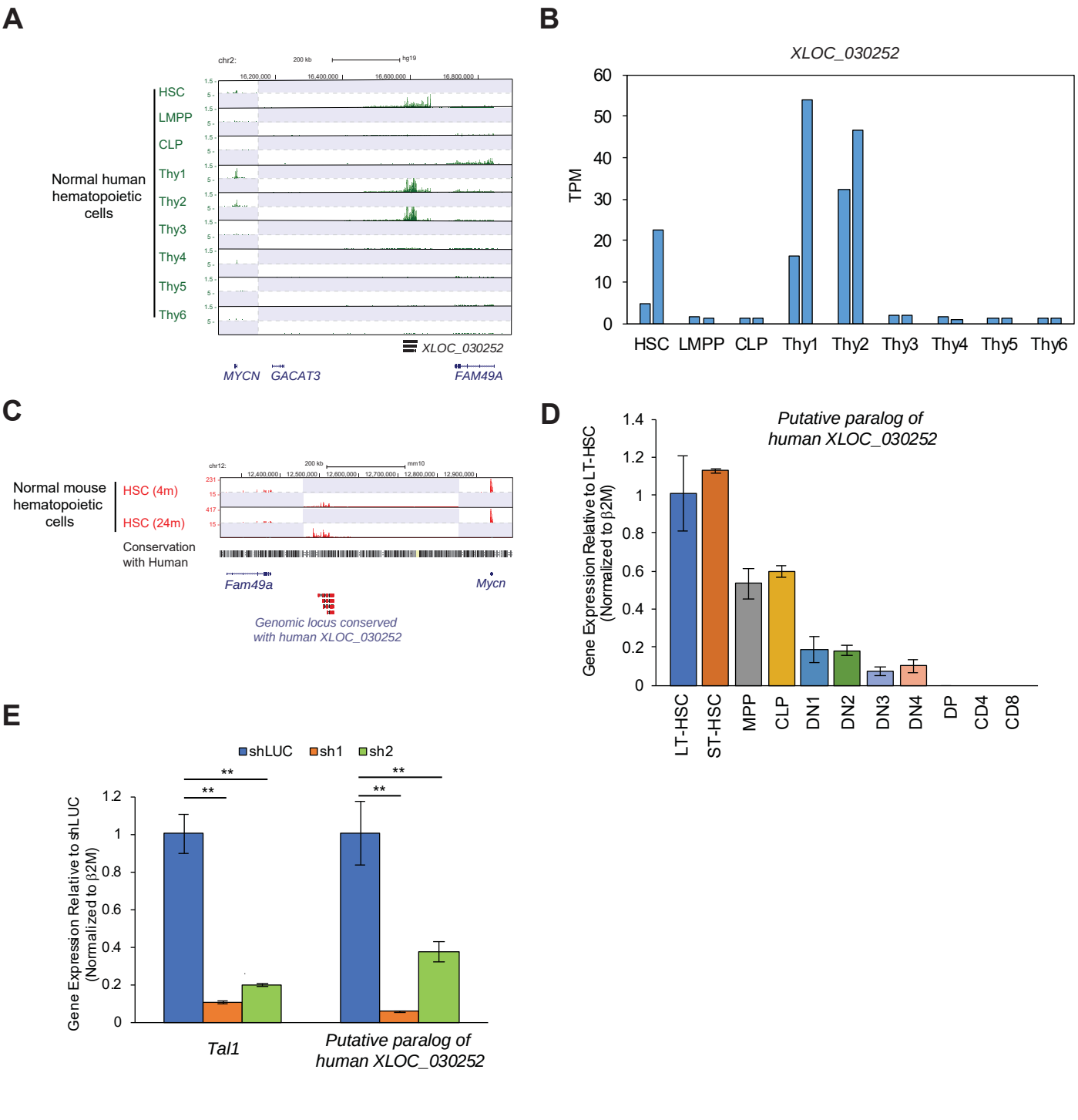

A
Normal human hematopoietic cells
HSC
LMPP
CLP
Thy1
Thy2
Thy3
Thy4
Thy5
Thy6
XLOC_030252
MYCN
GACAT3
FAM49A
B
XLOC_030252
TPM
HSC
LMPP
CLP
Thy1
Thy2
Thy3
Thy4
Thy5
Thy6
C
Normal mouse hematopoietic cells
HSC (4m)
HSC (24m)
Conservation with Human
Fam49a
Mycn
Genomic locus conserved with human XLOC_030252
D
Putative paralog of human XLOC_030252
Gene Expression Relative to LT-HSC (Normalized to β2M)
LT-HSC
ST-HSC
MPP
CLP
DN1
DN2
DN3
DN4
DP
CD4
CD8
E
shLUC
sh1
sh2
Gene Expression Relative to shLUC (Normalized to β2M)
Tal1
Putative paralog of human XLOC_030252

Figure 4

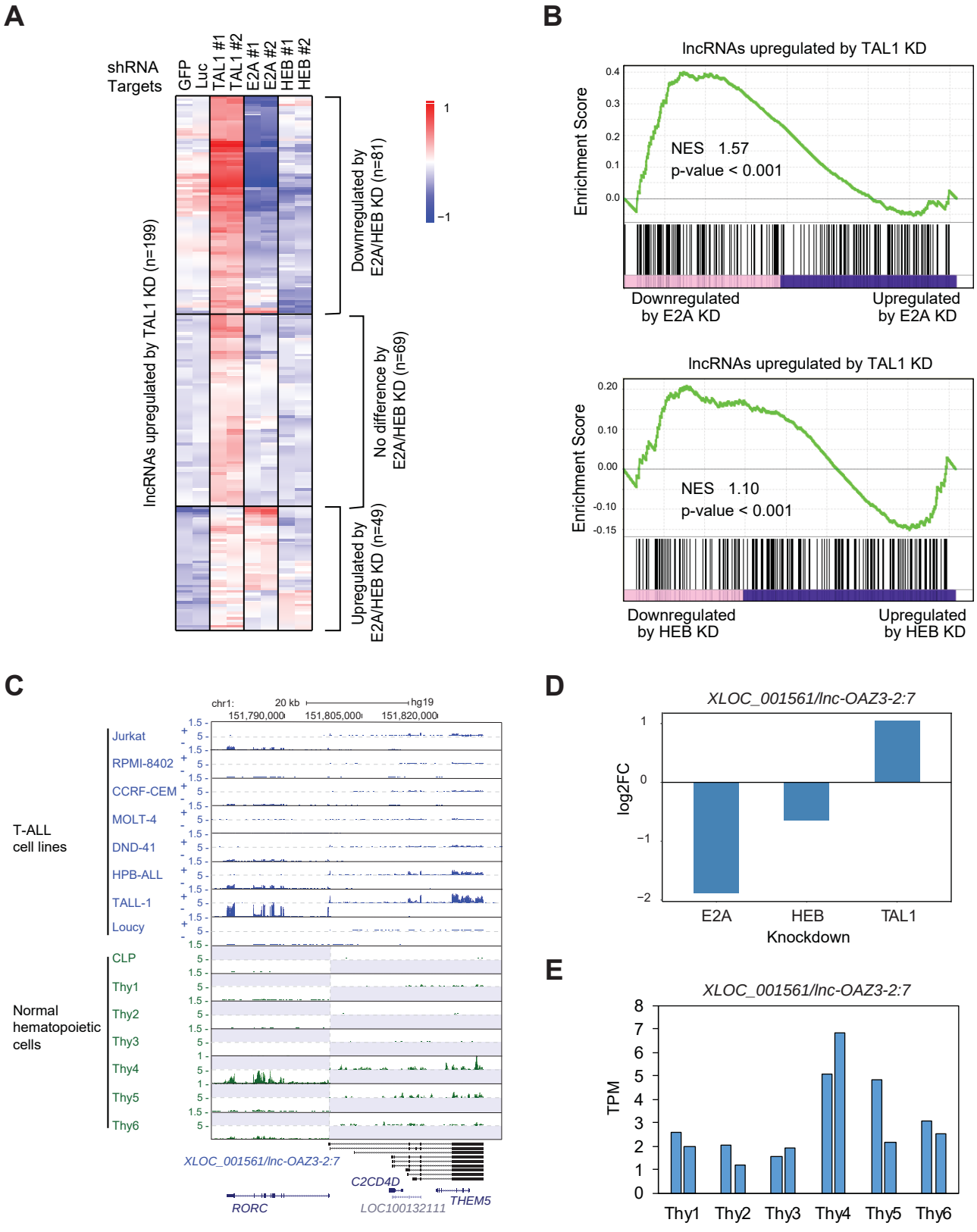

A
shRNA Targets
GFP
Luc
TAL1 #1
TAL1 #2
E2A #1
E2A #2
HEB #1
HEB #2
lncRNAs upregulated by TAL1 KD (n=199)
Downregulated by E2A/HEB KD (n=81)
No difference by E2A/HEB KD (n=69)
Upregulated by E2A/HEB KD (n=49)
1
−1
B
lncRNAs upregulated by TAL1 KD
Enrichment Score
NES 1.57
p-value < 0.001
Downregulated by E2A KD
Upregulated by E2A KD
lncRNAs upregulated by TAL1 KD
Enrichment Score
NES 1.10
p-value < 0.001
Downregulated by HEB KD
Upregulated by HEB KD
C
chr1:
20 kb
hg19
151,790,000
151,805,000
151,820,000
T-ALL cell lines
Jurkat
RPMI-8402
CCRF-CEM
MOLT-4
DND-41
HPB-ALL
TALL-1
Loucy
Normal hematopoietic cells
CLP
Thy1
Thy2
Thy3
Thy4
Thy5
Thy6
XLOC_001561/lnc-OAZ3-2:7
C2CD4D
RORC
LOC100132111
THEM5
D
XLOC_001561/lnc-OAZ3-2:7
log2FC
E2A
HEB
TAL1
Knockdown
E
XLOC_001561/lnc-OAZ3-2:7
TPM
Thy1
Thy2
Thy3
Thy4
Thy5
Thy6

Figure 5

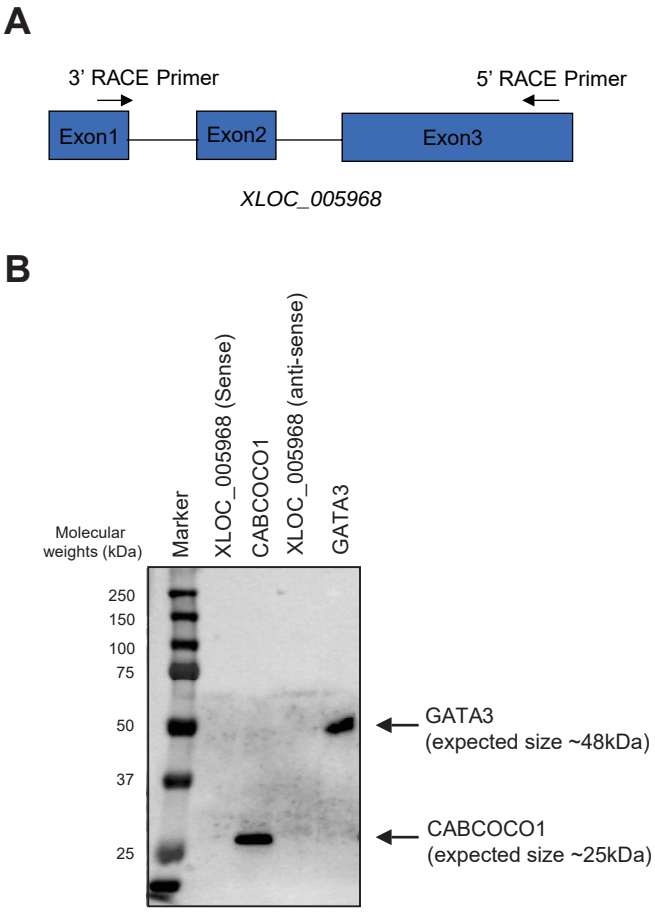

A
3' RACE Primer
5' RACE Primer
Exon1
Exon2
Exon3
XLOC_005968
B
Molecular weights (kDa)
Marker
XLOC_005968 (Sense)
CABCOCO1
XLOC_005968 (anti-sense)
GATA3
250
150
100
75
50
37
25
GATA3
(expected size ~48kDa)
CABCOCO1
(expected size ~25kDa)

Figure 6

Supplementary Information

Supplementary Materials and Methods

Gene knockdown and RNA extraction

Lentiviruses encoding short-hairpin RNAs (shRNAs) targeting each transcription factor were produced by co-transfecting individual shRNA constructs with the packaging plasmids pMDLg/pRRE and pRSV-Rev and the envelope plasmid pMD2.G into 293T cells by using FuGENE 6 transfection reagent (Roche). Supernatants containing lentivirus particles were collected and filtered through a 0.45 μm filter (Thermo). T-ALL cells were infected with lentivirus in the presence of polybrene (8 μg/ml; Millipore) by centrifugation at 1,300 rcf for 1.5 hr. The infected cells were selected with 0.7 μg/ml of puromycin (Sigma) in RPMI-1640 medium for at least 36 hr after infection. The mouse lympho-hematopoietic progenitor cell line EML was infected as described above and the infected cells that are GFP positive were then sorted out using flow cytometry 72 hours after infection. RNA was extracted using a miRNeasy kit (Qiagen) followed by DNase treatment (Turbo DNA-free™ kit, Ambion). For each sample, 10 million cells were collected and pelleted for RNA extraction. Sequences of shRNAs used for knockdown experiments are shown below:

Target Genes	shRNA IDs	shRNA sequences
GFP	shGFP	ACAACAGCCACAACGTCTATA
LUC	shLUC	CTTCGAAATGTCCGTTCGGTT
TAL1	shTAL1	GCTCAGCAAGAATGAGATCCT
GATA3	shGATA3	CATCCAGACCAGAAACCGAAA
RUNX1	shRUNX1	CGGCAGAAACTAGATGATCAG
MYB	shMYB	CCAGATTGTAAATGCTCATTT

TCF3/E2A	shE2A	GCAGCCTCTCTTCATCCACAT
TCF/12HEB	shHEB	GCAATCATTCAGTCCTGTCTA
Tal1 (mouse)	shTal1-1	ACTGACCCTAGGCTGGCATTT
	shTal1-2	TCCGCCTTGCCATGAAGTACA

RNA-seq analysis

Strand-specific library construction and sequencing of paired-end, 100-bp-long reads by the Illumina HiSeq4000 platform were performed at BGI Biotech Solutions Co Ltd (Hong Kong). See Supplementary Table 1 for the number of sequencing reads and quality scores for each cell sample. All RNA-seq data have been deposited in the GEO database (GSE97514).

Bioinformatic analysis and code availability

Identification of lncRNAs- The first dataset containing eight T-ALL cell lines (Jurkat, RPMI-8402, CCRF-CEM, MOLT-4, HPB-ALL, DND-41, TALL-1 and LOUCY) without the knockdown conditions were used to establish a bioinformatics pipeline. All RNA-seq datasets were aligned to the hg19 human genome using STAR 2.5.2a with the parameter outFilterMultimapNmax set to 1 and Ensembl database for annotated splice junctions.[1] Alignment files from STAR were used for *ab initio* transcriptome assembly with Cufflinks v2.2.1.[2] A RefSeq GTF file was applied for Cufflinks to guide the assembly. We then combined transcriptome assemblies from each cell line into one annotation using Cuffmerge from Cufflinks v2.2.1. After we removed single exon transcripts and any transcripts with a length less than or equal 200 nucleotides, we applied the SIncky pipeline[3] with --no_collapse to obtain lncRNAs that show no overlap with protein-coding genes on the same strand. From transcripts overlapping with protein-coding genes which were rejected by the SIncky

pipeline, we predicted lncRNAs overlapping protein-coding genes by four steps: (1) we removed transcripts that completely matched protein coding genes; (2) we removed transcripts with significant hits in PFAM[4] using the script pfam_scan.pl with hidden Markov models (HMMs) in Pfam-A and an evalue cut-off 1e-4; (3) we used PhyloCSF[5] with –orf ATGStop to identify putative Open Reading Frames (ORFs) with a minimum ORF length of 50 amino acids and removed any transcript with a PhyloCSF score greater than 0; and (4) we removed any transcript with an ORF length greater than 300 nucleotides. The transcripts passing these steps were considered to have no potential to encode functional proteins and were classified as sense lncRNAs. We selected putative lncRNAs that were expressed with an FPKM cut-off more than 0.5 in at least one T-ALL cell line. lncRNAs were annotated based on GenCode version 25[6] and LNCipedia version 4.0.[7, 8].

Differential expression analysis - To analyze lncRNAs regulated after gene knockdown, RNA-seq datasets of two control samples and two knockdown samples for each member of the TAL1 complex were aligned to the hg19 human genome using STAR 2.5.2a with the parameter outFilterMultimapNmax set to 1. We used the tool htseq-count v0.6.1 with the intersection-strict mode for the mapped reads in.bam files to generate count tables based on the Ensembl gene annotation of hg19. From these tables, the Bioconductor package DESeq2 v1.12.4 was used to analyze differential gene expression.

Quantification of transcript abundance in primary samples - We retrieved the dataset from EMBL-EBI (Accession_ID: EGAD00001002151) and analyzed by Kallisto software[9] using Ensembl annotation and lncRNA sequence as an index file. Transcript per million (TPM) value of each lncRNA in each primary sample was calculated and normalized using Sleuth software.[10]

Comparative analysis of genomic sequences - Genomic sequence for the selected lncRNA retrieved from human (hg19) genome was aligned to mouse (mm10) genome using the UCSC BLAT tool. Conserved human lncRNAs in mouse detected and gene track generated for further analysis.

Selection of lncRNAs directly targeted by TAL1 - We first selected genes that are bound by TAL1 within 12kb of the lncRNA loci based on our ChIP-seq data. The same cut-off (12kb) was used in our previous studies to identify protein-coding genes[11] and microRNAs[12] targeted by TAL1. We then filtered genes were significantly downregulated after knockdown of TAL1 (adjusted p-value < 0.05, log2-fold change < -0.2).

Primers for qRT-PCR analysis

Target Genes	Direction	PCR primer sequences (5' to 3')
TAL1	Forward	TTCCCTATGTTCACCACCAA
	Reverse	AAGATACGCCGCACAACTTT
TCF3/E2A	Forward	TCATCCTGAACTTGGAGCAG
	Reverse	CAACCACACCTGACACCTTT
GATA3	Forward	TTCAGTTGGCCTAAGGTGGT
	Reverse	CGCCGGACTCTTAGAAGCTA
RUNX1	Forward	CTGGTGTCTTCAGCCAGATG
	Reverse	CGACTGTGTACCGTGGACTG
GAPDH	Forward	CTCCTCTGACTTCAACAGCGACAC
	Reverse	TGCTGTAGCCAAATTCGTTGTCAT
TCF12/HEB	Forward	TCTCCAGTTTCCATTGTTGG

	Reverse	TTCTTGCTGCTGGTTGAAAA
MYB	Forward	TGTTGCATGGATCCTGTGTT
	Reverse	AGTTCAGTGCTGGCCATCTT
XLOC_005968	Forward	GAAAGCATAAAATGGCTTGGA
	Reverse	TGTAAAACAAGCCAATGCTATCA
XLOC_030252	Forward	GCCTCTTACTAGGTTGTCAAAATC
	Reverse	GAGTTATTCCTCACTCAGCTCACTC
XLOC_001561	Forward	AAGTTACTTTAGCTGTGTGCATCTC
	Reverse	ATATGATATTATGGGCCTATGTGTG
XLOC_024788	Forward	CAAGAGAAAGTCCGCAGCAG
	Reverse	CACATGGTGTCTGAACTGCC
XLOC_045612	Forward	TGTCATGGGGAGCTCTTCTG
	Reverse	CATCAACAGCTGGCCAAACA
XLOC_056913	Forward	TGCTAACAGAGATGGGTGCA
	Reverse	CAGAGCTGAGAGTACCCCAG
Mouse Tal1	Forward	CTCACTAGGCAGTGGGTTCTT
	Reverse	GGACCATCAGAAATCTCCATCT
Mouse B2M	Forward	ACGTAGCAGTTCAGTATGTTCG
	Reverse	GGTCTTTCTGGTGCTTGTCT
Mouse Cabcoco1	Forward	CTTCCTCTAAAGAAAACGTGTCAG
	Reverse	TATCTGGGAGACAGAATCCTTTCATT
Mouse XLOC 030252	Forward	CTTTTATTCAGAGTTTCCACATCGT
	Reverse	ACTGTCTTGGTGAGACTCTATTGCT

71

Primers for RACE

The gene specific primers used for 5' and 3' RACE were designed following the manufacturer's instructions: for 5' RACE, 5'-GAT TAC GCC AAG CTT TGT TCC AGT AAC GTG GCA GTC CGT TTG GG-3'; for 3' RACE, 5'- GAT TAC GCC AAG CTT TGT CAA GTC TGC ATG TGG CCC TTT CCC A-3'.

Primers for cDNA cloning

Primer sequences for cloning of sense *XLOC_005968*: Forward, 5'- GCT CAC CAA AGG CTC TGT CAT AAT T-3'; Reverse 5'-TGT TCC AGT AAC GTG GCA GTC-3'. Primer sequences for cloning of antisense *XLOC_005968:* Forward, 5'-TGT TCC AGT AAC GTG GCA G-3'; Reverse, 5'-GCT CAC CAA AGG CTC TGT C-3'. Primer sequences for cloning of the full-length *CABCOCO1*: Forward, 5'-ATG GAT TTC TCT ATT ATT CAG TAT TCA A-3'; Reverse, 5'-TTA GGC CTT TTT CAA TTT TTC TAT TCG-3'.

Isolation of mouse hematopoietic cells

Mouse studies strictly adhered to the recommendations of the Institutional Animal Care and Use Committee, and all protocols were approved by the Committee at the National University of Singapore. C57BL/6 mice were maintained, and bone marrow (BM) cells from 8-10 weeks-old inbred male mice were flushed from the long bones. Flow cytometry sorting was performed for isolating mouse hematopoietic cells using FACSAria (BD Biosciences) and a standard method. Briefly, after non-specific binding was blocked using mouse serum, antibody staining was performed on ice for 30 minutes. After washing, cells were suspended in phosphate-buffered saline (PBS) with Hoechst 33258 for dead cell discrimination. On the flow-cytometry screen, populations that were homogeneous in cell size [forward scatter–

side scatter (FSC-SSC) window] and viable cells were gated. Positive selection for the markers c-Kit and Sca-1 and negative selection for markers of mature hematopoietic cell lineages (CD3, CD4, CD8, B220, Gr-1, Mac-1, Ter-119 and IL-7Rα) were used to identify c-Kit^{+}Sca-1^{+}Lineage^{-} (KSL) cells. Antibodies against CD34 and Flt3 were used to further fractionate the KSL population into long-term hematopoietic stem cell (LT-HSC), short-term hematopoietic stem cell (ST-HSC) and multipotent progenitor (MPP) compartments. Long-term hematopoietic stem cells (LT-HSCs) were identified as c-Kit^{+}Sca-1^{+}Lineage(CD3, CD4, CD8, B220, Gr-1, Mac-1, Ter-119 and IL-7Rα)$^{-}$CD34^{-}Flt3^{-}, short-term hematopoietic stem cells (ST-HSCs) were identified as c-Kit^{+}Sca-1^{+}Lineage^{-}CD34^{+}Flt3^{-}, multipotent progenitors (MPPs) were identified as c-Kit^{+}Sca-1^{+}Lineage^{-}CD34^{+}Flt3^{+} and common lymphoid progenitors (CLPs) were identified as c-Kit^{low}Sca-1^{low} IL-7Rα $^{+}$. T-cell precursors were isolated from mouse thymus and identified as follows: double-negative 1 (DN1), CD4^{-}CD8^{-}CD44^{+}CD25^{-}, double-negative 2 (DN2), CD4^{-}CD8^{-}CD44^{+}CD25^{+}; DN3, CD4^{-}CD8^{-}CD44^{-}CD25^{+}; DN4, CD4^{-}CD8^{-}CD44^{-}CD25^{-}; double-positive (DP), CD4^{+}CD8^{+}; and single-positive (SP), CD4^{+}CD8^{-} or CD4^{-}CD8^{+}. All monoclonal antibodies were purchased from BD Pharmingen or eBioscience.

Supplementary Tables

Supplementary Table 1: RNA expressions of transcription factor genes in 8 T-ALL cell lines

Supplementary Table 2. The number of sequencing reads and quality scores for each RNA-seq sample

Supplementary Table 3: Sample information for primary T-ALL cases

Supplementary Table 4. List of lncRNAs expressed in T-ALL cell lines

Supplementary Table 5. List of lncRNAs directly activated by TAL1 in T-ALL cells

Supplementary Table 6. List of lncRNAs negatively regulated by TAL1 in T-ALL cells

Supplementary Figure Legends

Supplementary Figure 1. (A) Schematic representation of the classification of lncRNAs. Intergenic lncRNA is located between two protein coding genes. Divergent lncRNA is located within 1,000bp of promoter in the opposite direction from protein coding gene. Intronic lncRNA is located in an intron of protein coding gene. Antisense lncRNA is located in the antisense strand of protein coding gene and overlapped with one or several introns and exons of the sense strand. Sense lncRNA is overlapped with one or several introns and exons of protein coding gene on the same strand. **(B)** Density distributions showing pairwise Pearson correlations between loci for lncRNAs 1 and lncRNAs 2. See Figure 1E legend for details.

Supplementary Figure 2. Regulation of protein-coding genes by TAL1 and its regulatory partners in Jurkat cells. (A) The shRNAs targeting each transcription factor or control (shGFP) were transduced by lentiviral infection into Jurkat cells. RNAs were harvested at day 3 after infection. The expression levels of each factor were analyzed by qRT-PCR using specific primers and normalized to *GAPDH* expression. Fold change values compared with the control are shown as the mean ±standard deviation (SD) of triplicate samples. *p<0.05, **p<0.01 by two-sample, two-tailed t-test. **(B, C)** Expression of TAL1 target genes after knockdown of *TAL1* (B) or its regulatory partners (*GATA3, RUNX1, MYB*; C) in Jurkat cells. A total of 480 protein-coding genes that were directly bound by TAL1 and were also significantly downregulated by *TAL1* knockdown were selected. Heatmap images represent the expression levels of 480 genes in two controls and two knockdown samples. See the Figure 2A and B legends for details. **(D)** GSEA was performed to determine the correlation of genes positively regulated by TAL1 (n=480) and

gene expression changes upon knockdown (KD) of *GATA3*, *RUNX1* and *MYB*. See the Figure 2C legends for details.

Supplementary Figure 3. lncRNAs differentially expressed between T-ALL cell lines and normal human thymocytes. (A-B) ChIP-seq and RNA-seq gene tracks showing the binding of transcription factors, H3K27ac signals and RNA expressions at *XLOC_045612* (A) and *XLOC_056913* (B) loci in T-ALL cells. See Figure 3B legend for details. **(C)** Validation of a PCR primer set targeting exon 3 of *CABCOCO1.* qRT-PCR analysis was performed using an expression plasmid encoding the full-length of *CABCOCO1* (pCS2-CABCOCO1). Standard curve is shown. **(D)** Expression of predicted lncRNAs in eight T-ALL cell lines were measured by qRT-PCR analysis. The expressions were normalized to *GAPDH* expression. Fold change values compared with Jurkat cells were shown as the mean ±SD of triplicates samples. ND, not detected. **(E)** Validation of RNA expressions of putative lncRNAs after *TAL1* knockdown in Jurkat cells. Jurkat cells were transduced with a control shRNA (shGFP) or TAL1 shRNA by lentiviral infection. Total RNA was harvested at day 3 after infection. The expressions of each lncRNA were measured by qRT-PCR and normalized to *GAPDH* expression. Fold change values compared with the control are shown as the mean ±SD of triplicates samples. **(F)** RNA expressions of transcription factor genes (*TAL1, TAL2, TLX3, HOXA9, LMO1, LMO2, LYL1, NKX2-1* and *MEF2C)* in 27 primary T-ALL samples analyzed by RNA-seq. Values are shown by the transcripts per kilobase million (TPM). Note that *TLX3* was not expressed in any of T-ALL samples in this cohort.

Supplementary Figure 4. Expression of lncRNAs in mouse hematopoietic cells. (A) The RNA expression of mouse *Tal1* in different stage of mouse hematopoietic cells were

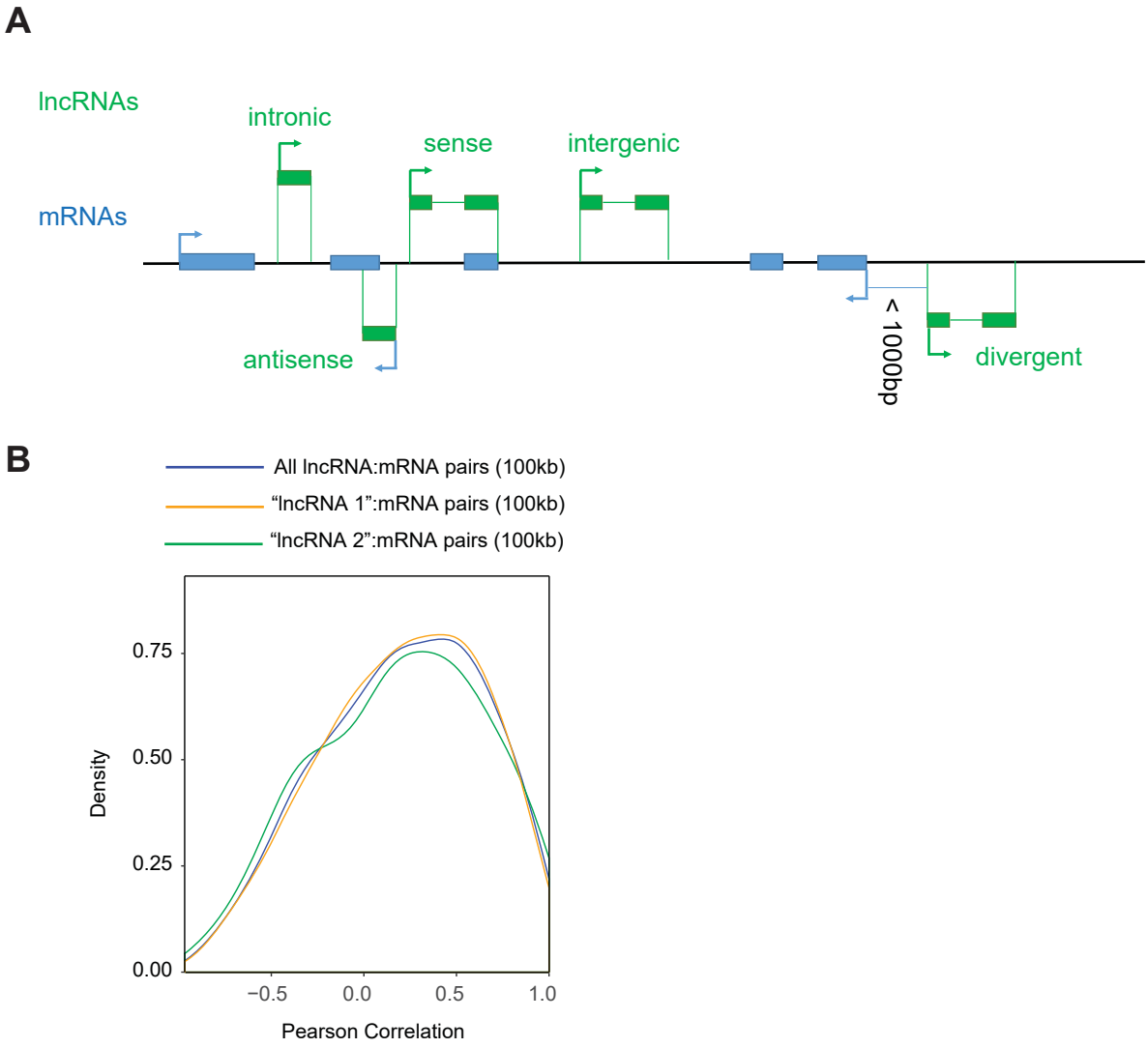
measured by qRT-PCR and normalized to *β2M* expression. See Figure 4D legend for details and Supplementary Materials for the definition of each population. **(B)** RNA expression around the genomic locus conserved with human *XLOC_005968* sequence in mouse HSCs. See Figure 4C legend for details. **(C)** The RNA expression of mouse *Cabcoco1* in different stage of mouse hematopoietic cells were measured by qRT-PCR and normalized to *β2M* expression. See Figure 4D legend for details.

Supplementary Figure 5. Genes differentially controlled by TAL1 and E-proteins in T-ALL cells. (A-D) RNA expression levels of *RAG1* (A)*, RAG2* (B), *PTCRA* (C) and *RORC* (D) genes after knockdown of *TAL1, E2A* and *HEB* in Jurkat cells were measured by RNA-seq. The values are shown in log2 fold-change (FC). **(E)** RNA expressions of *XLOC_001561/ lnc-OAZ3-2:7* in eight T-ALL cell lines were measured by qRT-PCR analysis. Fold change values compared with Jurkat cells are shown as the mean ±SD of triplicates samples. **(F)** RNA expression of *XLOC_001561* in Jurkat cells after knockdown of TAL1, E2A or HEB. See Supplementary Figure 2A legend for details. **(G)** ChIP-seq gene tracks showing the binding locations of TAL1, its regulatory partners (GATA3, RUNX1 and MYB) and RNA polymerase 2 (RNAP2) and various histone marks (H3K4me3, H3K4me1 and H3K27ac) at *XLOC_001561/lnc-OAZ3-2:7 locus* in Jurkat cells. See Figure 3B legend for details. **(H)** RNA expression levels of *RORC* in different stages of normal hematopoietic cells. See Figure 4A legend for details.

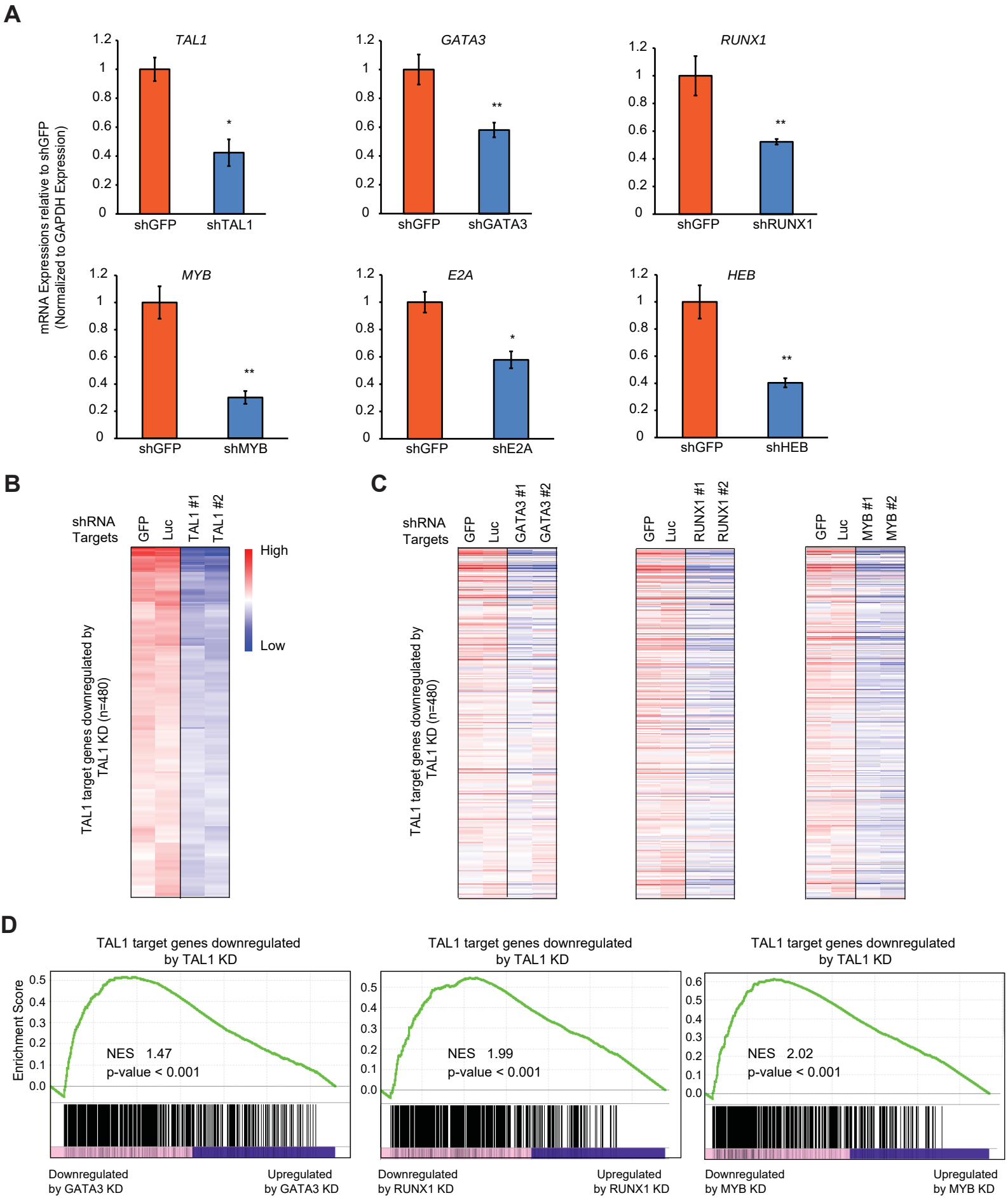
Supplementary Figure 6. *XLOC_005968* is a novel lncRNA expressed in T-ALL cells. (A) The 5' and 3' RACE were performed and the products were analyzed by DNA electrophoresis. **(B)** The full-length DNA sequence of *XLOC_005968*.

Supplementary References

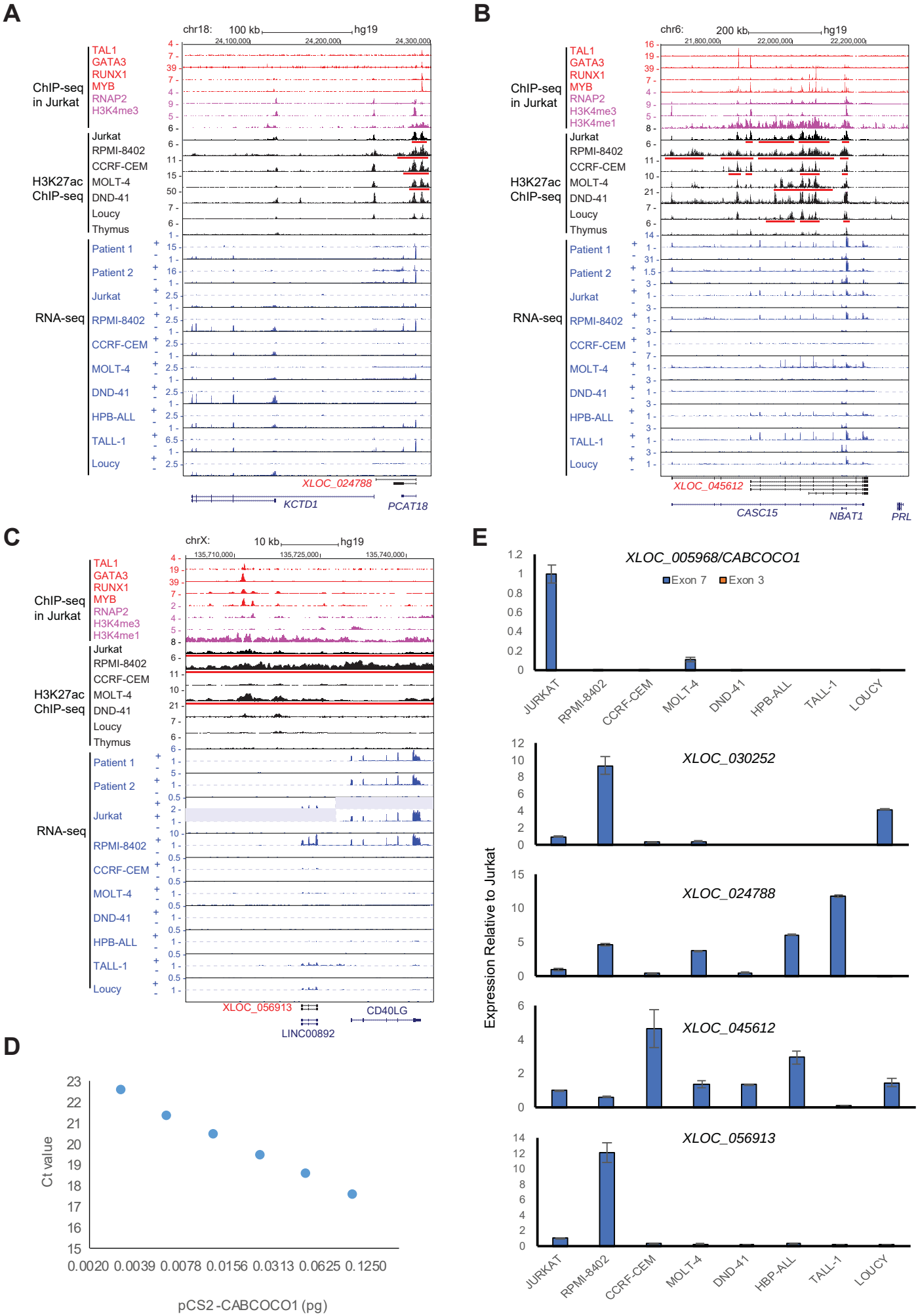
1. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S*, et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013 Jan 01; **29**(1)**:** 15-21.
2. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ*, et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010 May; **28**(5)**:** 511-515.
3. Chen J, Shishkin AA, Zhu X, Kadri S, Maza I, Guttman M*, et al.* Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. *Genome biology* 2016 Feb 02; **17:** 19.
4. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR*, et al.* Pfam: the protein families database. *Nucleic Acids Res* 2014 Jan; **42**(Database issue)**:** D222-230.
5. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* 2011 Jul 01; **27**(13)**:** i275-282.
6. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F*, et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* 2012 Sep; **22**(9)**:** 1760-1774.
7. Volders PJ, Helsens K, Wang X, Menten B, Martens L, Gevaert K*, et al.* LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res* 2013 Jan; **41**(Database issue)**:** D246-251.
8. Volders PJ, Verheggen K, Menschaert G, Vandepoele K, Martens L, Vandesompele J*, et al.* An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res* 2015 Jan; **43**(Database issue)**:** D174-180.
9. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 2016 May; **34**(5)**:** 525-527.
10. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature methods* 2017 Jul; **14**(7)**:** 687-690.
11. Sanda T, Lawton LN, Barrasa MI, Fan ZP, Kohlhammer H, Gutierrez A*, et al.* Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia. *Cancer Cell* 2012 Aug 14; **22**(2)**:** 209-221.
12. Mansour MR, Sanda T, Lawton LN, Li X, Kreslavsky T, Novina CD*, et al.* The TAL1 complex targets the FBXW7 tumor suppressor by activating miR-223 in human T cell acute lymphoblastic leukemia. *J Exp Med* 2013 Jul 29; **210**(8)**:** 1545-1557.



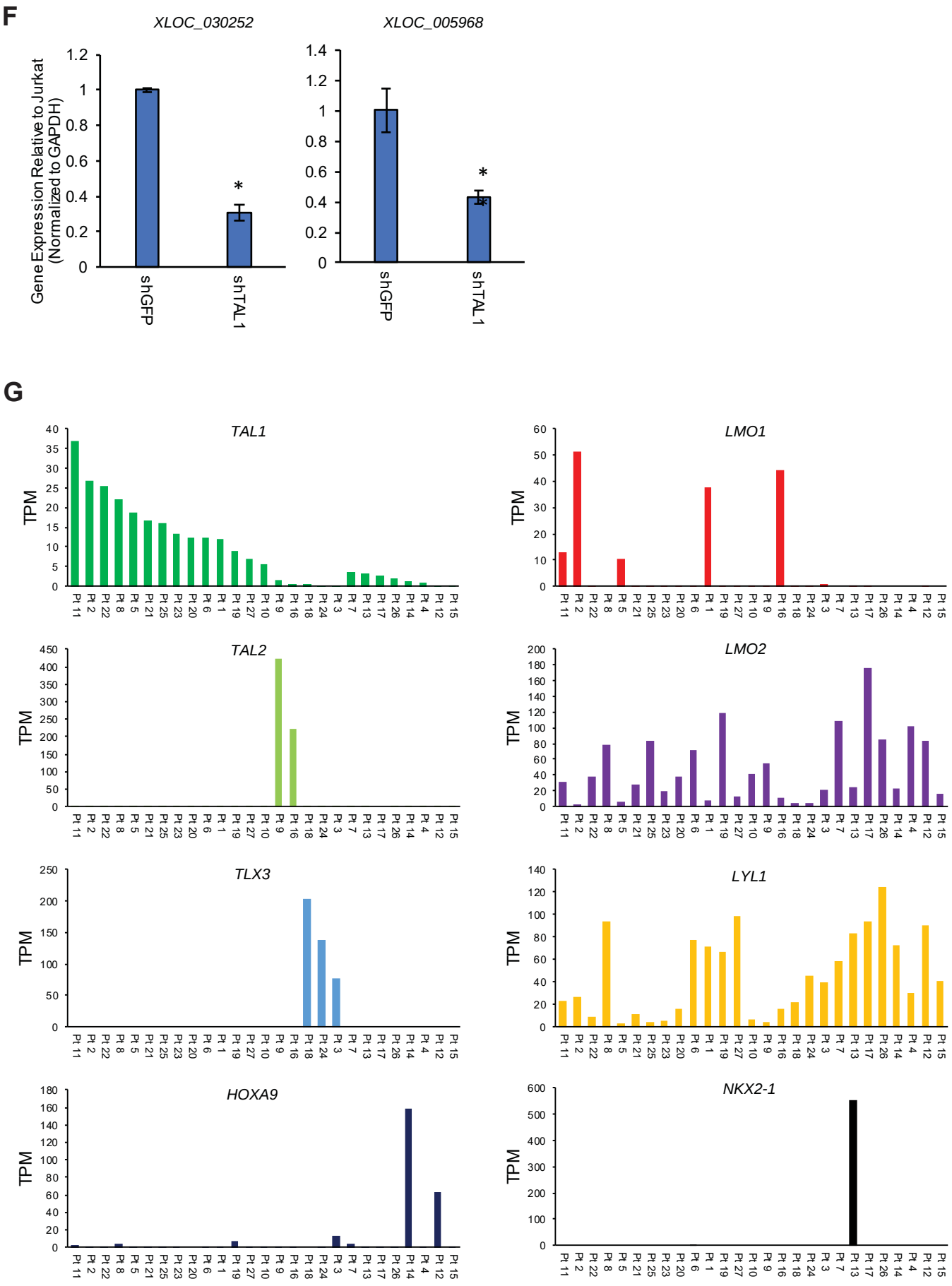
Supplementary Figure 1



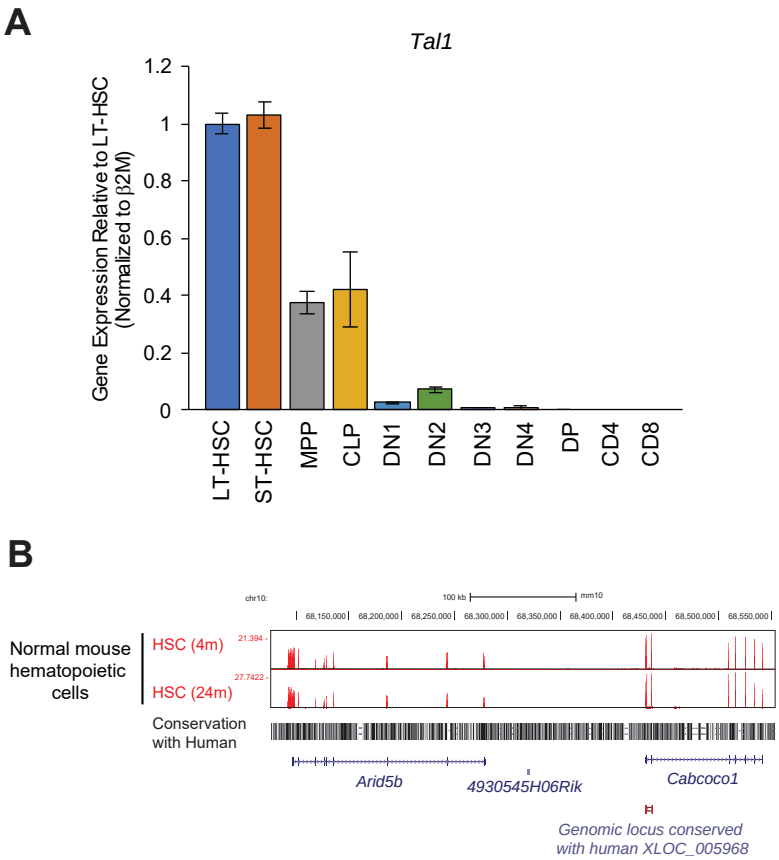
Supplementary Figure 2



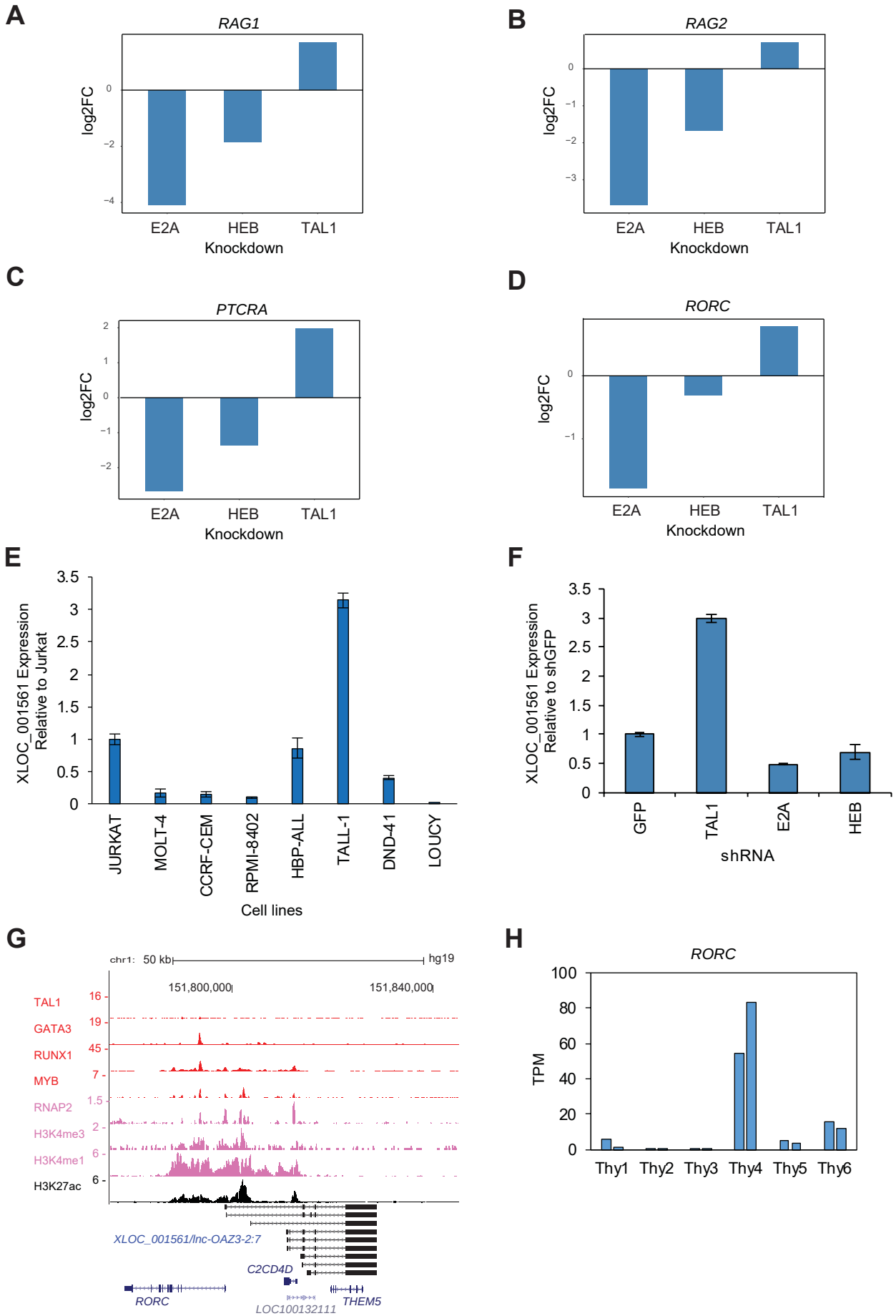
Supplementary Figure 3



Supplementary Figure 3

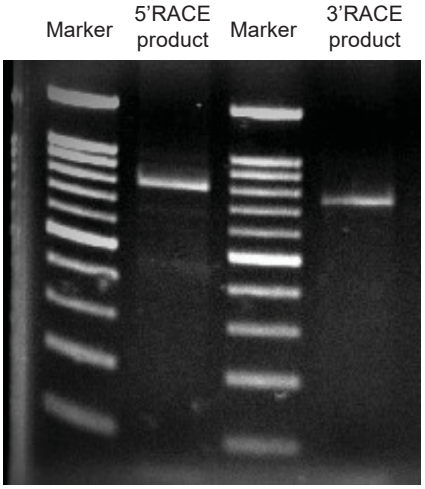


Supplementary Figure 4



Supplementary Figure 5

A



B

GCTCACCAAAGGCTCTGTCATAATTCAACTTTTTTCTTCCCATCAGCAAGTGATAG
AGGTTGTCAAGTCTGCATGTGGCCCTTTCCCAAATCCTCTGGAAGAAGGAATCTC
ATTTGATATTTATTCAACATTCATAGAGCCCCCCACAATATTGGATACGGAAATGAA
GAGGTTGGATCAAGAACAAGGCCCTGAGGAGTCTCAGCCAGAAACTGACACCT
CAGACATGGATCCTTTAGTTGGTTTCACCATTGAAGATGTAAAATCAGTGCTGGAT
CAAGTCACAGATGACATTTTAATCGGCATTCAGACCGAGATAAACGAAAAACTGC
AAATACAGGAAGAGGCCTTTAATGCACGAATAGAAAAATTGAAAAAGGCCTAAG
GACTTGGTACAAGGAGAGTGATGCTAAACTTCACAGAAACAAACAAAACCAGCC
AGAATAACAGACTCTCTGGAGCGTCGTGTCTCCATCACTTAGTTGTGAAAGGAAA
ACCAAGCCCCACTTTTTATTTTCCTAAGTAATTAGAAAAGTATTGGTCCCAATTTTG
CTATCTCCCATCCCATAACAGCCTCTGAATTTATTGCACCAAGTGTAATGAGAACA
TTTTGTATACAAGAGTTTGAAAAATTATATATAAAAATACAATTACTTTTATGATAGT
CCTTTGACGTTTTGACTAATAAATCCTATTCATCTTCAAGGAGAATGAAGTCAACT
TTTTAAATAACCAGAAATAAAGAAAGATGATAATAACTTCATCCATTTTGACCCAAA
CGGACTGCCACGTTACTGGAACA

XLOC_005968 (802bp)

Supplementary Figure 6