A comprehensive catalog of LncRNAs expressed in T-cell acute lymphoblastic leukemia

Yasmina Kermezli, Wiam Saadi, Mohamed Belhocine, E. L. Mathieu, Marc-Antoine Garibal, Vahid Asnafi, Mourad Aribi, Salvatore Spicuglia, Denis Puthier

Title: A comprehensive catalog of LncRNAs expressed in T-Cell Acute Lymphoblastic Leukemia.

Authors: Yasmina Kermezli[1-3], Wiam Saadi[1-3], Mohamed Belhocine[1,2,5], Eve-Lyne Mathieu[1,2], Marc-Antoine Garibal[1,6], Vahid Asnafi[4], Mourad Aribi[3,#], Salvatore Spicuglia[1,2*] and Denis Puthier[1,2*]

AUTHOR AFFILIATIONS

[1]Aix-Marseille University, Inserm, TAGC, UMR1090, Marseille, France.
[2]Equipe Labéllisée Ligue Nationale Contre le Cancer
[3]Tlemcen University, The Laboratory of Applied Molecular Biology and Immunology, Algeria.
[4]Université Paris Descartes Sorbonne Cité, Institut Necker-Enfants Malades (INEM), Institut National de la Santé et de la Recherche Médicale (Inserm) U1151, and Laboratory of Onco-Haematology, Assistance Publique-Hôpitaux de Paris (AP-HP), Hôpital Necker Enfants-Malades, Paris, France
[5]Present adress: Molecular Biology and Genetics Laboratory, Dubai, United Arab Emirates.
[6]Present adress: Aix Marseille University, Inserm, MMG, Marseille France
#Senior author

***Correspondence**: salvatore.spicuglia@inserm.fr, denis.puthier@univ-amu.fr

Abstract

Several studies have demonstrated that LncRNAs can play major roles in cancer development. The creation of a catalogue of LncRNAs expressed in T cell acute lymphoblastic leukemia (T-ALL) is thus of particular importance. However, this task is challenging as LncRNA expression is highly restricted in a time and space manner and may thus greatly differ between samples. We performed a systematic transcript discovery in RNA-Seq data obtained from T-ALL primary cells and cell lines. This led to the identification of 2560 novel LncRNAs. After the integration of these transcripts into a large compendium of LncRNAs (n=30478) containing both known LncRNAs and those previously described in T-ALLs, we then performed a systematic genomic and epigenetic characterization of these transcript models demonstrating that these novel LncRNAs share properties with known LncRNAs. Finally, we provide evidences that these novel transcripts could be enriched in LncRNAs with potential oncogenic effects and identified a subset of LncRNAs coregulated with T-ALL oncogenes. Overall, our study represents a comprehensive resource of LncRNAs expressed in T-ALL and might provide new cues on the role of lncRNAs in this type of leukemia.

Keywords: Large non-coding RNA, LncRNA, T-cell acute leukemia, T-ALL, oncogenes

Introduction

Long non-coding RNAs (LncRNAs) are a novel class of untranslated RNA species defined as transcripts with poor coding potential and size above 200 nucleotides[1,2]. They can lie in both sense and antisense direction of exonic or intronic elements, or in intergenic regions (a subclass termed 'long intergenic non coding RNAs', LincRNA), or even in the promoter regions of coding [3]. LncRNAs are transcribed by RNA polymerase II and mirror the features of protein-coding genes, such as polyadenylation and splicing, without containing a functional open reading frame. They are often transcribed at lower abundance than coding genes and in a more tissue-specific manner. In this regard, the function of these transcripts is suggested to be particularly important to shape cell identity. Several studies have demonstrated that lncRNAs are functional and regulate both the expression of neighboring genes and distant genomic sequences by a variety of mechanisms [4]. A growing number of examples also demonstrated that LncRNAs play a major role in cancer development by acting on different levels of regulation to disrupt cellular regulatory networks including proliferation, immortality and motility [5].

T-cell acute lymphoblastic leukemia (T-ALL) is an aggressive hematological cancer arising from the transformation of T cell [6,7]. Cytogenetic and global transcriptomic analyses led to the classification of T-ALL into molecular groups characterized by the abnormal expression of specific transcription factors (TAL; LMO1/2; TLX1/3; LYL; HOXA; MEF2c, respectively) and their block of differentiation at specific stages [8,9]. Although, the outcome of T-ALLs has globally improved by modern poly-chemotherapy, T-ALL remains of poor prognosis notably in relapsing cases. A major obstacle to understanding the mechanisms of T-ALL oncogenesis is the heterogeneous cellular and molecular nature of the disease, which is driven by a complex interplay of multiple oncogenic events. In this context, LncRNA signatures have been shown to define oncogenic subtypes [10] and several LncRNAs regulated by key T-ALL oncogenes have been identified [11–14].

The time and space restricted expression of LncRNAs makes it challenging to envision the creation of a complete catalog of LncRNAs. Yet such a catalog appears as a prerequisite to better characterize LncRNAs involved in pathological processes. We thus performed systematic transcripts discovery in a collection of T-ALL samples [15] and integrate previously created catalogs into a non-redundant set. The subsequent list of LncRNAs was thoroughly characterized regarding genomic structure and epigenetic features. Finally, we set up a strategy to prioritize LncRNAs having potential oncogenic effects. This approach allowed us to point out LncRNA candidates potentially relevant in the leukemia pathogenesis.

Material and methods

De novo LncRNAs discovery

RNA-Seq experiment from Atak *et al.* [15], were retrieved in BAM format from European Genome-phenome Archive under accession number EGAS00001000536. The alignments (BAM files) were provided to cufflinks (v2.2.1) which aims at assembling reads into transcript models. Cufflinks was used with default settings, except for arguments "-j/--pre-mrna-fraction" (set to 0.6) and "-a/--junc-alpha" (set to 0.00001) in order to reduce the number of intronic transcripts (that may correspond to fragments of immature transcripts) and to include well-supported exons into transcript models [16]. The transcript models obtained from the 50 samples (31 primary T-ALL patients, 18 T-ALL cell lines and 1 pool of 5 thymuses) were subjected to a cleaning procedure using bedtools (v2.17.0) in order to remove all transcripts described in hg19 RefSeq annotation (Illumina iGenomes web site) [17]. Cuffcompare v2.1.1 was then used to merge all files and remove transcript model redundancy [18]. The subsequent gtf file was then filtered to eliminate any transcript model defined in RefSeq, Gencode V19 and new lincRNAs discovered by Trimarchi and his coworkers [11]. Transcripts expression levels were estimated using cufflink ('-G' option) and only those with FPKM greater than 1 in at least one sample were kept. Filters on transcripts size (at least 200 nucleotides as defined for LncRNAs), number of exons (at least 2 exons) and poor coding potential (CPAT score lower than 0.2) as expected from LncRNAs [19] were subsequently applied.

Genomic annotation of transcripts

The subsequent gtf, including all transcripts categories, (**Dataset 1**) was then used to perform genomic annotation. All analyses were done using R software or Python scripts.

Assessment of LncRNA Tissue Specificity

We processed a set of fastq files corresponding to 20 human tissues (SRA accession number SRP056969). After read mapping (tophat2), the genes expression levels were quantified using Cuffdiff [18]. The gene expression specificity of each gene was computed across all tissues using the tau score [20]:

$$tau = \sum_{i=0}^{n} \frac{(1-\hat{x_i})}{n-1} \; ; \hat{x_i} = \frac{x_i}{max1 \leq i \leq n(x_i)}$$

Where *(i)* n corresponds to the number of samples, *(ii)* x_i corresponds to the expression level (log2-transformed FPKM values) in condition *i (iii) and max(x$_i$) corresponds to the* maximum expression level through all tissues.

Epigenetic characterization of LncRNA

The ChIP-Seq datasets obtained from thymus and 3 T-ALL cell lines (DND41, Jurkat and RPMI-8402) and corresponding to H3K4me3 and H3K27ac were obtained from ENCODE and GEO databases. The H3K4me3 and H3K27ac ChIP-Seq in RPMI-8402 cell line were sequenced in our laboratory (SRA accession numbers SRX3437292 and SRX3437293, see Table S1). To measure the ChIP-Seq signal around the TSS ([-3000,

+3000] pb) we focused on genes with FPKM above 1. Coverage analyses were performed using a Python script making calls to the pyBigWig python library.

Search for potential oncogene

We computed the variance of Log2-transformed FPKM values as a score to find genes displaying high dispersion of expression levels across samples. Pearson's correlation coefficients were computed, using R software, between the top 10% variants LncRNAs and known coding T-ALL oncogenes to bring out coding-noncoding pairs.

LncRNA expression analyses

Total RNA was extracted using TRIzol (Invitrogen) according to the manufacturer's instruction. 1 µg of RNA was treated with 1 U of DNase I (Ambion) and incubated at 37˚C for 30 min. DNAse I was then inactivated (15mM of EDTA and incubation at 75˚C for 10 minutes). DNase-treated RNAs were reversed transcribed using SuperScript II (Invitrogen) and oligo (dT) or random primers according to the manufacturer's instruction. Control genomic DNA was purified from RPMI-8402 cells using the DNeasy Kit (Qiagen) according to the manufacturer's specifications. Sequences of primers used for PCR of LncRNA *XLOC_00017544_Atak,* XLOC*_00009269_Atak and* XLOC_00012823_Atak *are provided in* **Table S2**. PCR using 1 µL of cDNA was performed with Herculase II Fusion kit (Agilent, Waldbronn, Germany) following manufacturer instructions. Amplifications were carried out with 40 cycles (95˚C for 1 minute, denaturation at 95˚C for 20 seconds, annealing for 20 seconds, extension at 68˚C for 1 minutes), followed by a final extension step (68˚C for 4 minutes).

Quantitative reverse transcriptase polymerase chain reaction (qRT-PCR)

The qPCR with Power SYBR green mix (Thermo Fisher) was performed on a Mx3000P real-time PCR system. Each reaction was performed with 2µl of cDNA. *GAPDH* was used as reference for normalization.

Results

Building a catalog of LncRNAs expressed in T-ALLs

In order to get an exhaustive catalog of LncRNAs expressed in T-ALLs we first performed a systematic transcript discovery on 50 RNA-Seq samples (a pool of 5 normal thymuses, 31 T-ALLs primary blasts and 18 T-ALLs cell lines) previously described by Atak *et al* [15] (**Figure 1A**). *De novo* transcripts were filtered and only multi-exonic transcripts with size greater than 200 bp, FPKM greater than 1 and coding potential lower than 0.2 were kept. This transcript models were then merged with known LncRNAs obtained from GENCODE version 19 [21] and T-ALL LncRNAs described by Trimarchi et *al* [11]. The final catalog contains a non-redundant list of 30478 LncRNAs. This encompasses 26092 from GENCODE (LncRNA_Known), 1826 LncRNAs from the Trimarchi dataset (LncRNA_Trimarchi) and 2560 new LncRNAs from the Atak dataset (LncRNA_Atak). The corresponding GTF file is provided as supplementary (**Dataset 1**).

Genomic characterization

We next performed a thorough genomic comparison of the three different sets of LncRNA transcripts obtained from GENCODE, Trimarchi et *al* and our own analysis of Atak *et al* RNA-Seq data. They will be denoted hereafter as LncRNA_Known, LncRNA_Trimarchi and LncRNA_Atak respectively. Throughout the analysis, these three sets of LncRNAs were compared to coding transcripts (mRNA) in order to underlie their specific properties. **Figure 1B** shows that the number of exons differs between mRNA transcripts (which mainly contain 5 exons or more) and the three sets of LncRNAs. This underscores the unusual exonic structure of LncRNAs that tends to be limited to two exons as already reported by others [21].Note that the *de novo* LncRNA_Atak dataset lack mono-exonic transcripts as they were discarded during the filtering process. Regarding transcript size, the reference LncRNAs were found to be shorter than mRNAs (average size of 0.8 KB compared to 2.5 KB) as observed by Derrien *et al* [21]. The mean size of *de novo* LncRNA_Atak was close to that of the reference (LncRNA_Known) validating our procedure of transcript reconstruction (**Figure 1C**). In contrast, the mean size of LncRNAs defined by Trimarchi et *al* were greater (4 kb) than the mean size of mRNAs (2.5 kb), which may point out an intrinsic difference in the procedure used for reconstruction of underlying transcript models. Concerning chromosomal distribution, the LncRNAs tend to be similarly spread throughout the chromosomes while several differences were observed (**Figure 1D**). LncRNA_Trimarchi dataset is enriched in transcripts from chromosome 13 and Y while depleted of transcripts located on chromosome 19. Such a result may probably highlight the representation of some particular tumor karyotypes and gender distribution in the samples used by Trimarchi *et al*. In contrast, the proportion of LncRNAs from LncRNAs_Atak dataset is very similar throughout the chromosomes although their representation is slightly increased on chromosome 21 and 22. LncRNAs are generally classified based on their location with respect to protein-coding genes. We defined five types of LncRNAs (**Figure 1E**): *(i)* 'Intergenic', transcribed outside of any known coding gene; *(ii) 'D*ivergent',

produced in promoter regions of coding genes on opposite strand; *(iii)* 'Convergent' whose transcription ends in 3' regions of coding genes on opposite strand *(iv)* "Sense" and *(v) "Antisense"* whose transcription takes place inside the gene body of a coding gene in sense or antisense direction, respectively. Regarding these five classes, the composition of LncRNA_Atak dataset was rather close to the reference with 55.38% and 55.56% of intergenic transcripts respectively although less transcripts were classified as divergent (6.57% versus 14.44% for GENCODE) and more transcripts were labelled as antisense (32.30% versus 21.34%) (**Figure 1F**). In contrast, the Trimarchi dataset was found to be mainly composed of intergenic transcripts (90.47%) and divergent transcripts (8.12%) since the other classes were discarded during the building steps of the catalog [11].

LncRNAs expression in T-ALL and thymus

We next intended to assess the expression of these LncRNAs in several sample groups including normal thymus, T-ALL cell lines and patient samples. We computed, for each transcript, its median expression level across each sample group (**Figure 2A**). In agreement with the weak expression level of LncRNAs reported earlier [21], the mRNAs were more highly expressed than any of the three LncRNA sets. Transcripts from LncRNA_Atak were found to be more highly expressed than LncRNA_Known in Thymus and T-ALL patients while slightly less expressed in cell lines. Of note, however, weaker expression was observed in the LncRNA_Trimarchi dataset suggesting that the lack of accuracy in transcript reconstruction step may also impair proper quantification of these transcripts. LncRNAs are known to be highly tissue specific compared to mRNAs. To verify this, we computed the *tau* tissue-specificity score [20] using a public RNA-Seq dataset encompassing 20 human tissues [22]. This score ranges from 0 for housekeeping genes to 1 for highly tissue-specific genes. As expected, mRNAs displayed a bimodal signal with a major fraction of genes behaving as ubiquitous genes and a minor fraction having high. In contrast, a clear shift toward high tissue-specificity scores were observed for LncRNAs regardless of the underlying groups (**Figure 2B**). Moreover, assessment of expression in individual tissues demonstrated a strong bias for thymus-specific LncRNAs in the LncRNA_Trimarchi and LncRNA_Atak dataset (**Supplementary Figure S1**). This also underlines that while LncRNA_Atak were selected against LncRNA_Trimarchi, numerous LncRNAs with strong expression bias in the thymus remained to be discovered. Altogether, these results underscore the enrichment that exists in our catalog for LncRNA biases toward tissue-specificity.

Functional annotation

Many LncRNAs have been shown to regulate the expression of neighbor genes in cis [11,23,24]. Therefore, to characterize the functional relevance of the LncRNA sets we performed a functional annotation of their closest neighbor genes using GREAT [25]. For the LncRNA_Known class, significant enrichment was observed for annotation terms related to ubiquitous processes including 'genes expression' and 'metabolism' (**Figure 3A**). In contrast, for the LncRNA_Trimarchi set, annotation terms related to immune system were

significantly enriched, including: 'regulation of interleukin 4 production', 'leukocyte activation' and 'T cell receptor V(D)J recombination' (**Figure 3B**). In the same way, closest genes for LncRNA_Atak dataset were related to 'negative regulation of Notch signaling pathway' or 'regulation of leukocyte degranulation' for instance (**Figure 3C**). This indicates that both LncRNA_Atak and LncRNA_Trimarchi datasets are enriched for LncRNAs located close to coding genes having major role in normal immune processes and leukemia development. As some LncRNA have been shown to regulate protein-coding genes in cis, this would suggest that some of our newly discovered transcripts could potentially act on key genes regulating immune response and oncogenic processes.

Epigenetic features of LncRNAs

LncRNAs are known to share epigenetic features with coding genes. Both H3K4me3 and H3K27ac have been described as epigenetic marks strongly associated to the promoter region of expressed genes. In order to compare epigenetic features across all transcript sets, we used H3K4me3 and H3K27ac ChIP-Seq obtained from normal thymus and three T-ALL cell lines (DND41, RPMI-8402 and Jurkat). We filtered LncRNAs and mRNAs based on their expression in the corresponding samples by selecting transcripts with FPKM above 1. Using ChIP-seq datasets for H3K4me3 and H3K27ac, we then computed the number of reads falling in binned regions around the promoter (defined as [-3000, 3000] pb around the TSS) for each gene. The mean number of reads for each bin across all genes of a class was used to compute the meta profile shown in Figure 4. Although the results slightly differ between the samples, the LncRNAs and mRNAs sets displayed consistent epigenetic profiles. A striking difference is observed for the LncRNA_Trimarchi set, which display high levels of H3K27ac likely indicating a location bias toward enhancer regions [26].

Experimental validation expression of three LncRNAs in RPMI-8402, and Jurkat cell lines

We next aimed at validating the expression of *de novo* identified LncRNAs from the LncRNA_Atak dataset. We selected three LncRNAs (*XLOC_00017544_Atak, XLOC_00009269_Atak and XLOC_00012823_Atak)* located on chromosome 9, 2 and 3 and containing 5, 8 and 2 exons respectively (see Dataset 1 for coordinates). RNA-Seq and ChIP-Seq signals indicated that all three LncRNAs were expressed in RPMI-8402 and Jurkat cell lines (**Figure 5A**). This result was confirmed for the 3 candidates by RT-PCR performed on RNA isolated from RPMI-8402, and Jurkat cell lines. PCR products corresponding to DNA fragments of expected sizes were observed in all three cases (**Figure 5B**).

Variability of gene expression among T-ALL samples predict potential oncogenic LncRNAs

T cell transformation is related to many genomic and chromosomal abnormalities, which can lead to aberrant gene transcription [6,27]. Many of the described oncogenes are not expressed in normal T cell development [28–30] and only restricted to a subset of T-ALL

samples. Therefore, it is expected that the expression of these oncogenes should be associated with a high variance across leukemic samples. Based on this hypothesis, we aimed at mining our LncRNA dataset for potentially new oncogenes using the variance as a proxy.

We first computed the variance of coding genes across T-ALL cell lines and patient samples and check our ability to recover known T-ALL oncogenes [28]. As depicted in **Figure 6A**, typical leukemia oncogenes (TAL1, TLX1, TLX3, HOXA9, NKX3-1, LMO2) were ranked within the top 10 % of genes with the highest variance in the cell lines and patients. A statistical analysis demonstrated that both in cell lines and patients, leukemia oncogenes have significantly higher variance compared to non-oncogenic genes or to a random list of genes matched for expression distribution (**Figure 6B**). The same prioritizing strategy was applied to LncRNAs in order to identify potential oncogene candidates. Strikingly, numerous LncRNAs known for their implication in cancer (*e.g.* H19, XIST, LUNAR1, MIAT and NEAT1) were ranked within the top 10% of LncRNAs displaying the highest variance in both cell lines and patients. Interestingly, several *de novo* LncRNAs from the Atak dataset, including the 3 LncRNAs validated in **Figure 5**, were found among the highest variable transcripts (**Figure 6C**). Moreover, the variance of LncRNA_Atak list was found to be significantly higher when compared to the two other sets, suggesting that it may be enriched for potential oncogenic LncRNAs (**Figure 6D**). As an example, we validated the variable expression of *XLOC_00000871_Atak*, one of the LncRNAs with the highest variance in both T-ALL patients and cell lines (**Figure 6B and S2**), by RT-qPCR across a panel of T-ALL cell lines (**Figure 6E).**

Correlation between T-ALL oncogenes and LncRNAs

To address the possibility that some of the highly variable LncRNAs might be associated with the regulation of key oncogenes, we computed the expression correlation between the 10% of LncRNAs with highest variance and the set of known T-ALL oncogenes and retrieved the correlated gene-lncRNA pairs. 60.5 % (1146) and 59% (1116) of these LncRNAs were correlated ($r > 0,5$) with at least one oncogene in T-ALL cell lines and patients, respectively (**Supplementary dataset 2-3**). For instances, the expression of LUNAR-1 (XLOC_LNC_TALL01_Trimarchi), a Notch1-regulated LncRNA in T-ALL (11), was highly correlated with *NOTCH1* ($r=0.75$; **Supplementary dataset 2-3**). About 14% (patients) to 17% (cell lines) of the correlated gene-lncRNA pairs were located within the same chromosome, while 10% were separated by less than 1 Mb (**Supplementary** Figure S3), suggesting potential cis-regulation for a substantial number of oncogenes. Thirty percent of correlated LncRNAs come from the LncRNA_Atak dataset, with some being highly correlated with T-ALL oncogenes (**Figure 7A**). Two examples are shown in Figure 7B-C. Interestingly, some T-ALL oncogenes, such as EML1 and OLIG2 (Figure 7B-C), correlated with several LncRNAs and form complex regulatory networks (**Supplementary** Figure S4). Altogether, these results suggest that some LncRNA from the LncRNA_Atak set might be potential regulators of oncogenic genes in T-ALL and pave the way for more detailed studies.

Discussion

LncRNAs are transcribed weakly in a large fraction of the genome and display remarkably restricted expression in a space and time-dependent manner [2], making challenging to draw up an exhaustive list of transcripts expressed in all cellular conditions. This is especially true in the case of tumor cells were genomic alterations are expected to alter transcriptional programs, generally leading to large heterogeneous cancer subtypes [6]. In order to establish a comprehensive catalogue of LncRNAs expressed in T-ALL, we analyzed a set of 50 RNA-Seq samples produced by Atak *et al.* [15] (31 primary T-ALL patients, 18 T-ALL cell lines and 1 pool of 5 thymuses) and performed *de novo* transcript discovery in order to systematically identify transcript models. This approach led to the discovery of 2560 novel LncRNAs. Subsequently, we perform a deep characterization of the genomic and epigenetic properties of these transcripts and showed they are comparable to previously identified LncRNAs.

Several approaches have been suggested to identify functionally relevant LncRNAs, including guilty-by-association or correlation-based approaches [31]. Master oncogenes in T-ALL are generally ectopically expressed in a restricted number of patients resulting in highly variable expression among tumor samples. Indeed, genes displaying high variance throughout the T-ALL samples were demonstrated to be significantly enriched in known T-ALL oncogenes. We thus used the expression variance as a proxy to estimate oncogenic potential of the LncRNA expressed in T-ALL. We observed that LncRNAs with known implication in cancer (e.g. LUNAR1) were ranked among those with the highest variance. Interestingly, many newly identified LncRNAs were found to have highly variable expression. Combined variance and correlation analysis also suggest that a fraction of these LncRNAs could have oncogenic properties by functionally interacting with known oncogenes.

One of the key features of LncRNAs is that their expression pattern is highly tissue and cell type specific [2]. This is consistent with our finding that *de novo* LncRNAs discovered in T-ALL demonstrated high tissue-specificity (**Figure 2**) and that many LncRNAs found in T-ALL are expressed in few leukemic samples and (**Figure 6**). Consequently, molecules targeting either their expression or their interactions with chromatin or protein complexes would represent therapeutic targets able to kill cancer cells while sparing normal cells [5]. Additionally, correlation of expression patterns with leukemia progression and outcome could led to novel prognosis markers and help classification and stratification of the patients.

T-ALL comprises several molecular subgroups characterized by the aberrant expression of distinct oncogenic transcription factors, unique gene expression signatures, and different prognoses [6]. While the existence of specific molecular subtypes of T-ALL has long been established, therapeutic strategies are applied uniformly across subtypes, leading to variable responses between patients coupled with high toxicity. Our comprehensive resource of LncRNAs expressed in T-ALL should allow further exploration of LncRNAs potentially involved in leukemia and provide new rationales for patients/risk stratification.

Acknowledgements

We thank the Transcriptomics and Genomics Marseille-Luminy (TGML) platform for sequencing of ChIP samples and the Marseille-Luminy cell biology platform for management of cell culture. Work in the TAGC laboratory was supported by recurrent funding from INSERM and Aix-Marseille University and by the Foundation for Cancer Research ARC (ARC PJA 20151203149) and A*MIDEX (ANR-11-IDEX-0001-02), the Cancéropôle PACA, Plan Cancer 2015 (C15076AS) and Ligue Nationale contre le Cancer, of which the TAGC lab is an "Equipe Labellisée". Y.K. and W.S. were supported, by the Franco-Algerian partnership Hubert Curien (PHC) Tassili (15MDU935).

Legends to figures

Figure 1: Genomic characterization of LncRNA transcripts. (A) Schematic illustration of the procedure used to create our LncRNA catalog. (**B)** Bar plots displaying the number of exons per transcript set. **(C)** Distribution of transcript sizes. Each vertical line indicates the mean transcript sizes of the corresponding set. **(D)** Chromosomal distribution of transcript sets. **(E)** Schematic illustration of LncRNAs categories. LncRNA exons appear as red and coding genes as blue. **(F)** Pie chart representing the fraction of LncRNA categories across the transcripts sets.

Figure 2: Expression levels and tissue-specificity of LncRNAs classes. (A) Violin plot showing expression levels of transcripts. For each transcript, expression level was computed by calculating the median of its expression values among 31 T-ALL blasts, 18 T-ALL cell lines and a pool of 5 normal thymus. Wilcoxon test was used to assess differences. **(B)** Representative human tissues [22] were used to compute gene expression and assess tissue-specificity of each transcript. The density plots shows the distributions of the tissue-specificity score (see material and method section). Each vertical line indicates the mean tissue specificity score of the corresponding class.

Figure 3: Functional annotation of neighboring genes for the LncRNAs. (**A**) LncRNA_Known. (**B**) LncRNA_Trimarchi. (**C**) LncRNA_Atak. The *x*-axis is corresponding to $-log_{10}$*(corrected p-value)* and the *y*-axis shows the corresponding biological processes.

Figure 4: TSS coverage plot of ChIP-Seq signal for H3K4me3 and. Signals are shown for the four transcript in one tissue (total thymus) and three human cell lines (DND41, RPMI-8402 and Jurkat).

Figure 5: Experimental validation of expression for three lncRNAs. (A) Integrated genomics viewer (IGV) screenshots displaying H3K4me3 ChIP-Seq signals as well as RNA-Seq signals for genomic regions corresponding to *XLOC_00017544_Atak, XLOC_00009269_Atak and XLOC_00012823_Atak* in RPMI-8402, and Jurkat cell lines. **(B)** PCR validation of *XLOC_00017544_Atak, XLOC_00009269_Atak and XLOC_00012823_Atak* in RPMI-8402 and Jurkat cell lines. *MALAT1* was used as positive control.

Figure 6: **New candidate oncogenes identification by variance. (A)** Variance for coding genes in T-ALL cell lines and patients. (**B**) Box plots showing the distribution of variance of

coding genes in T-ALL cell lines and patients. Wilcoxon test was used to assess differences. **(C)**Variance of non-coding genes in T-ALL cell lines and patients. Arrows highlight leukemic oncogenes. (**D**) Box plots showing the distribution of variance of non-coding in T-ALL cell lines and patients. Wilcoxon test was used to assess differences. (**E**) Variability of XLOC_*00000871_Atak* expression normalized against the GAPDH gene (n = 20) in thymus and cell lines.

Figure 7: Co-expression between oncogenes and LncRNAs from Atak dataset. (**A**): Heatmaps showing the correlation between oncogenes (columns) and the most correlated transcript from the LncRNA_Atak dataset (rows). Correlation are shown both for cell lines (left panel) and patients (right panel). (**B**) Screenshots obtained from IGV displaying two examples of strong co-expression between an oncogene and a LncRNA from the Atak dataset. Tracks corresponding to cell lines are shown in red while patients are shown in green. (**C**) Scatter plots showing the expression of pairs oncogene-LncRNA_Atak in cell lines (red) and patients (green).

References

[1] Schadt EE, Edwards SW, GuhaThakurta D, et al. A comprehensive transcript index of the human genome generated using microarrays and computational approaches. Genome Biol. 2004;5:R73.

[2] Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. Nature. 2012;482:339–346.

[3] Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22:1760–1774.

[4] Bonasio R, Shiekhattar R. Regulation of transcription by long noncoding RNAs. Annu. Rev. Genet. 2014;48:433–455.

[5] Schmitt AM, Chang HY. Long Noncoding RNAs in Cancer Pathways. Cancer Cell. 2016;29:452–463.

[6] Ferrando AA, Neuberg DS, Staunton J, et al. Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. Cancer Cell. 2002;1:75–87.

[7] Soulier J, Clappier E, Cayuela J-M, et al. HOXA genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL). Blood. 2005;106:274–286.

[8] Asnafi V. HiJAKing T-ALL. Blood. 2014;124:3038–3040.

[9] Ntziachristos P, Abdel-Wahab O, Aifantis I. Emerging concepts of epigenetic dysregulation in hematological malignancies. Nat. Immunol. 2016;17:1016–1024.

[10] Wallaert A, Durinck K, Van Loocke W, et al. Long noncoding RNA signatures define oncogenic subtypes in T-cell acute lymphoblastic leukemia. Leukemia. 2016;30:1927–1930.

[11] Trimarchi T, Bilal E, Ntziachristos P, et al. Genome-wide mapping and characterization of Notch-regulated long noncoding RNAs in acute leukemia. Cell. 2014;158:593–606.

[12] Wang Y, Wu P, Lin R, et al. LncRNA NALT interaction with NOTCH1 promoted cell proliferation in pediatric T cell acute lymphoblastic leukemia. Sci Rep. 2015;5:13749.

[13] Ngoc PCT, Tan SH, Tan TK, et al. Identification of novel lncRNAs regulated by the TAL1 complex in T-cell acute lymphoblastic leukemia. Leukemia. 2018;

[14] Wallaert A, Durinck K, Taghon T, et al. T-ALL and thymocytes: a message of noncoding RNAs. J Hematol Oncol. 2017;10:66.

[15] Atak ZK, Gianfelici V, Hulselmans G, et al. Comprehensive analysis of transcriptome variation uncovers known and novel driver events in T-cell acute lymphoblastic leukemia. PLoS Genet. 2013;9:e1003997.

[16] Roberts A, Pimentel H, Trapnell C, et al. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics. 2011;27:2325–2329.

[17] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–842.

[18] Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7:562–578.

[19] Wang L, Park HJ, Dasari S, et al. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res. 2013;41:e74.

[20] Kryuchkova-Mostacci N, Robinson-Rechavi M. A benchmark of gene expression tissue-specificity metrics. Brief. Bioinformatics. 2017;18:205–214.

[21] Derrien T, Johnson R, Bussotti G, 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.

[22] Duff MO, Olson S, Wei X, et al. Genome-wide identification of zero nucleotide recursive splicing in Drosophila. Nature. 2015;521:376–379.

[23] Pandey RR, Mondal T, Mohammad F, et al. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol. Cell. 2008;32:232–246.

[24] Zhao J, Sun BK, Erwin JA, et al. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science. 2008;322:750–756.

516 [25] McLean CY, Bristor D, Hiller M, et al. GREAT improves functional interpretation of cis-
517 regulatory regions. Nat. Biotechnol. 2010;28:495–501.

518 [26] Ørom UA, Shiekhattar R. Long noncoding RNAs usher in a new era in the biology of
519 enhancers. Cell. 2013;154:1190–1193.

520 [27] Van Vlierberghe P, Ferrando A. The molecular basis of T cell acute lymphoblastic
521 leukemia. J. Clin. Invest. 2012;122:3398–3406.

522 [28] Xia Y, Brown L, Yang CY, et al. TAL2, a helix-loop-helix gene activated by the
523 (7;9)(q34;q32) translocation in human T-cell leukemia. Proc. Natl. Acad. Sci. U.S.A.
524 1991;88:11416–11420.

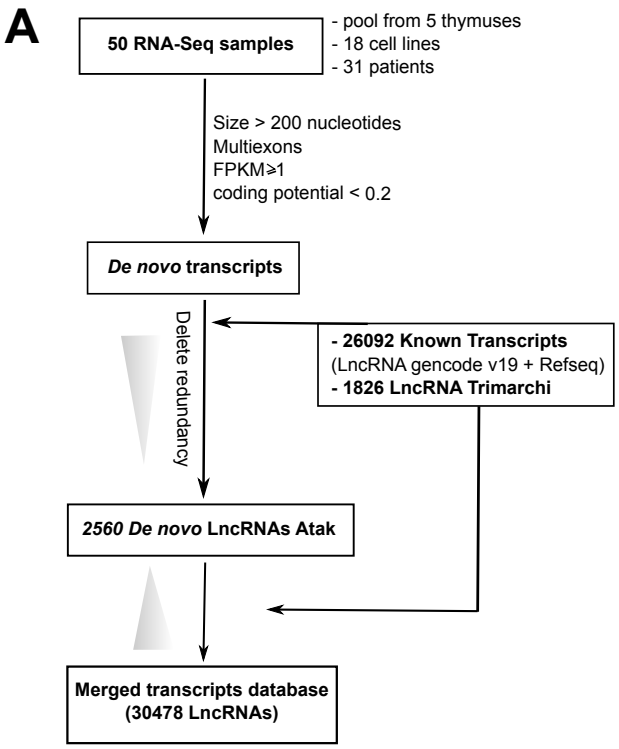
525 [29] Brown L, Cheng JT, Chen Q, et al. Site-specific recombination of the tal-1 gene is a
526 common occurrence in human T cell leukemia. EMBO J. 1990;9:3343–3351.

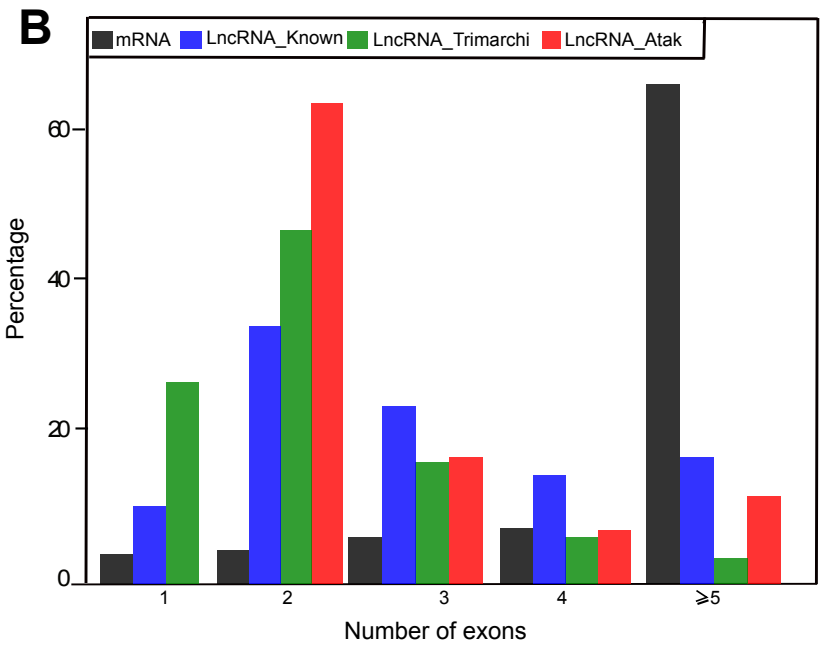
527 [30] Armstrong SA, Staunton JE, Silverman LB, et al. MLL translocations specify a distinct
528 gene expression profile that distinguishes a unique leukemia. Nat. Genet.
529 2002;30:41–47.

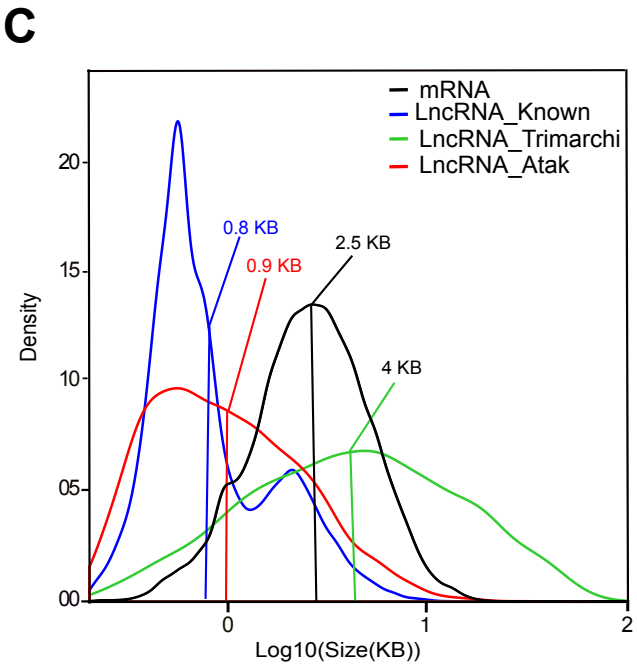
530 [31] Liao Q, Liu C, Yuan X, et al. Large-scale prediction of long non-coding RNA functions in
531 a coding-non-coding gene co-expression network. Nucleic Acids Res. 2011;39:3864–
532 3878.

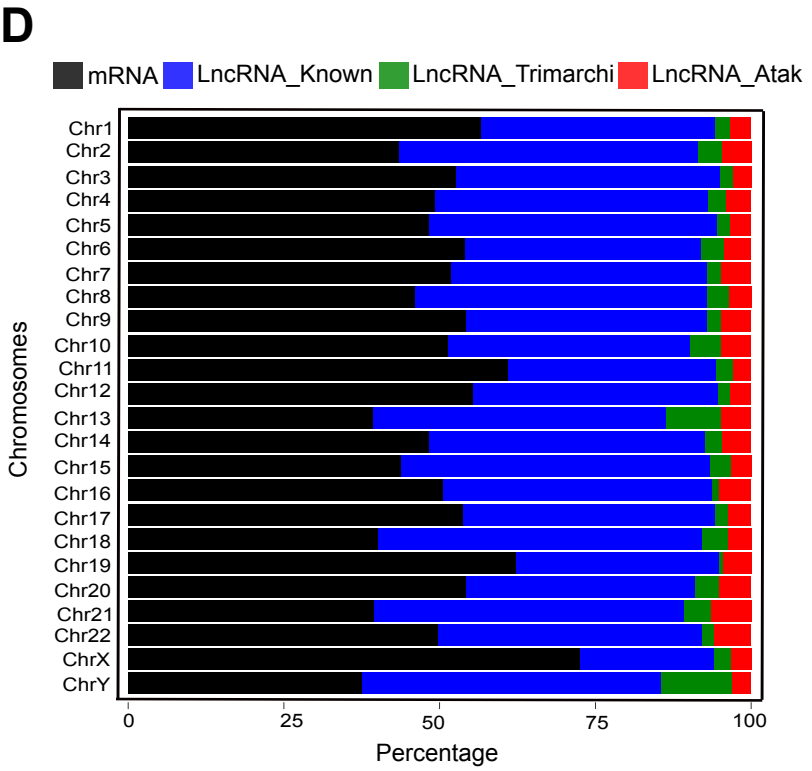
533

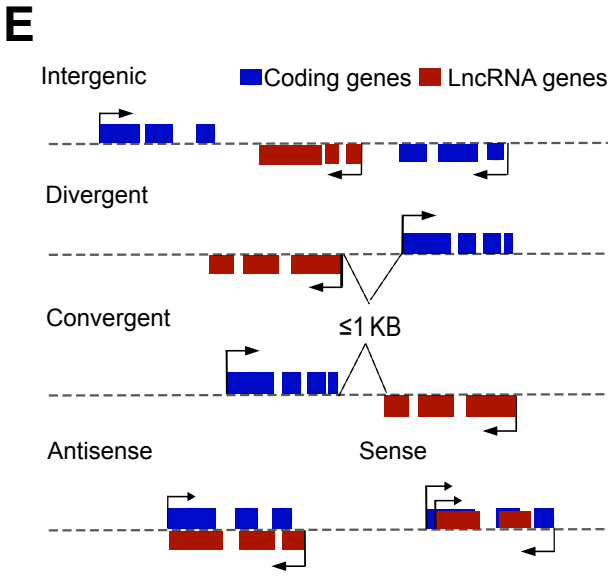
Figure 1


A
50 RNA-Seq samples
- pool from 5 thymuses
- 18 cell lines
- 31 patients
Size > 200 nucleotides
Multiexons
FPKM≥1
coding potential < 0.2
De novo transcripts
Delete redundancy
- 26092 Known Transcripts
(LncRNA gencode v19 + Refseq)
- 1826 LncRNA Trimarchi
2560 De novo LncRNAs Atak
Merged transcripts database
(30478 LncRNAs)


B
mRNA
LncRNA_Known
LncRNA_Trimarchi
LncRNA_Atak
Percentage
0
20
40
60
1
2
3
4
≥5
Number of exons


C
mRNA
LncRNA_Known
LncRNA_Trimarchi
LncRNA_Atak
0.8 KB
0.9 KB
2.5 KB
4 KB
Density
20
15
10
05
00
0
1
2
Log10(Size(KB))


D
mRNA
LncRNA_Known
LncRNA_Trimarchi
LncRNA_Atak
Chromosomes
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
Chr22
ChrX
ChrY
0
25
50
75
100
Percentage


E
Intergenic
Coding genes
LncRNA genes
Divergent
Convergent
≤1 KB
Antisense
Sense

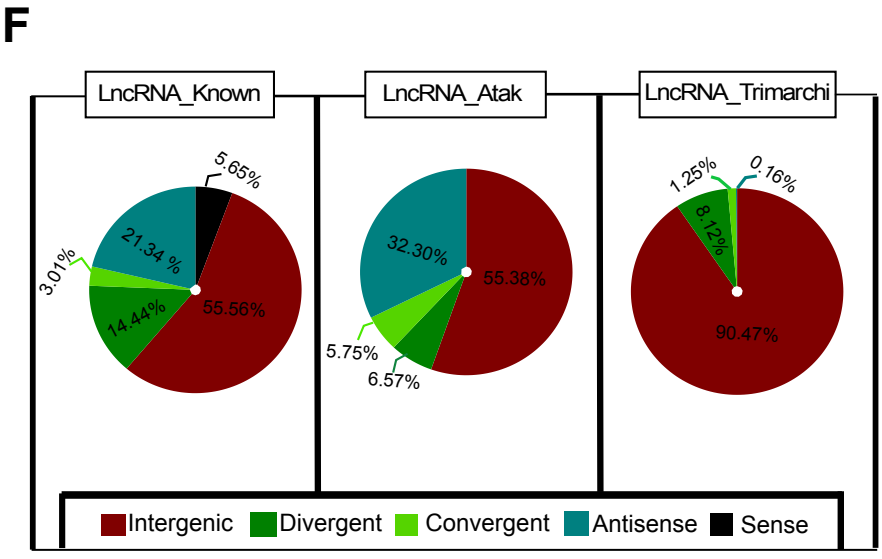
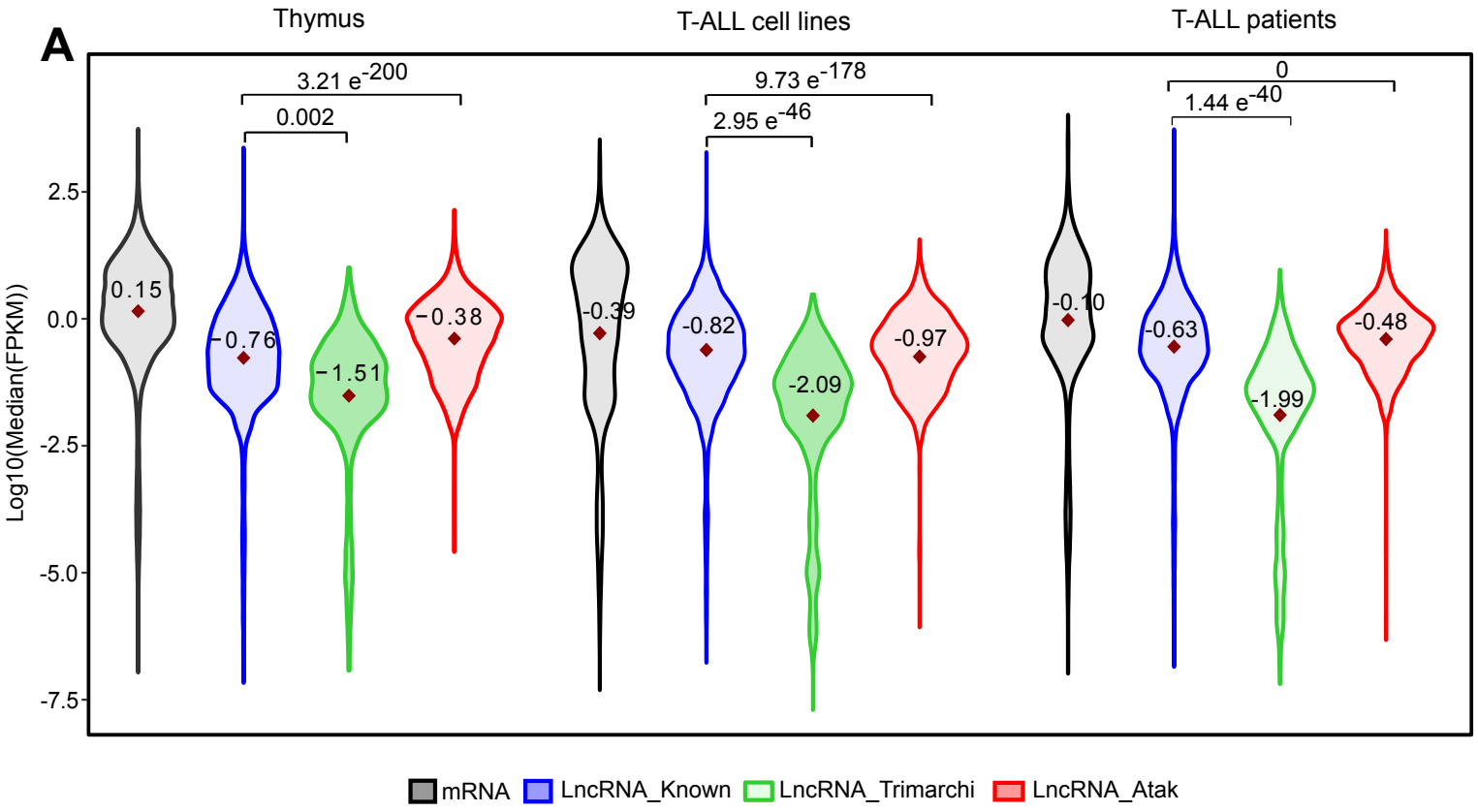

F
LncRNA_Known
LncRNA_Atak
LncRNA_Trimarchi
5.65%
21.34 %
3.01%
14.44%
55.56%
32.30%
55.38%
5.75%
6.57%
1.25%
0.16%
8.12%
90.47%
Intergenic
Divergent
Convergent
Antisense
Sense

Figure 2


A
Thymus
T-ALL cell lines
T-ALL patients
3.21 e-200
0.002
9.73 e-178
2.95 e-46
0
1.44 e-40
Log10(Median(FPKM))
2.5
0.0
-2.5
-5.0
-7.5
0.15
-0.76
-1.51
-0.38
-0.39
-0.82
-2.09
-0.97
-0.10
-0.63
-1.99
-0.48
mRNA
LncRNA_Known
LncRNA_Trimarchi
LncRNA_Atak

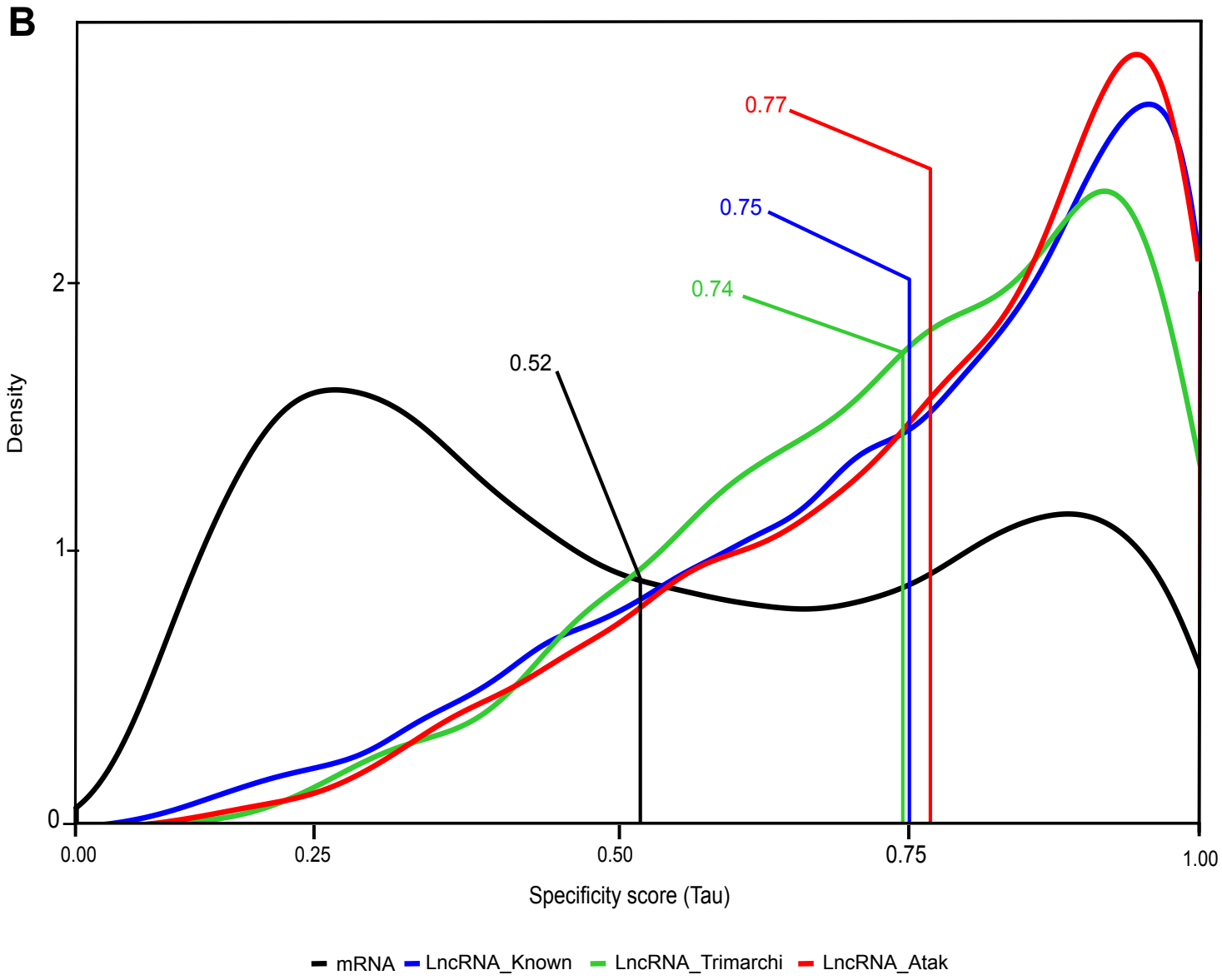
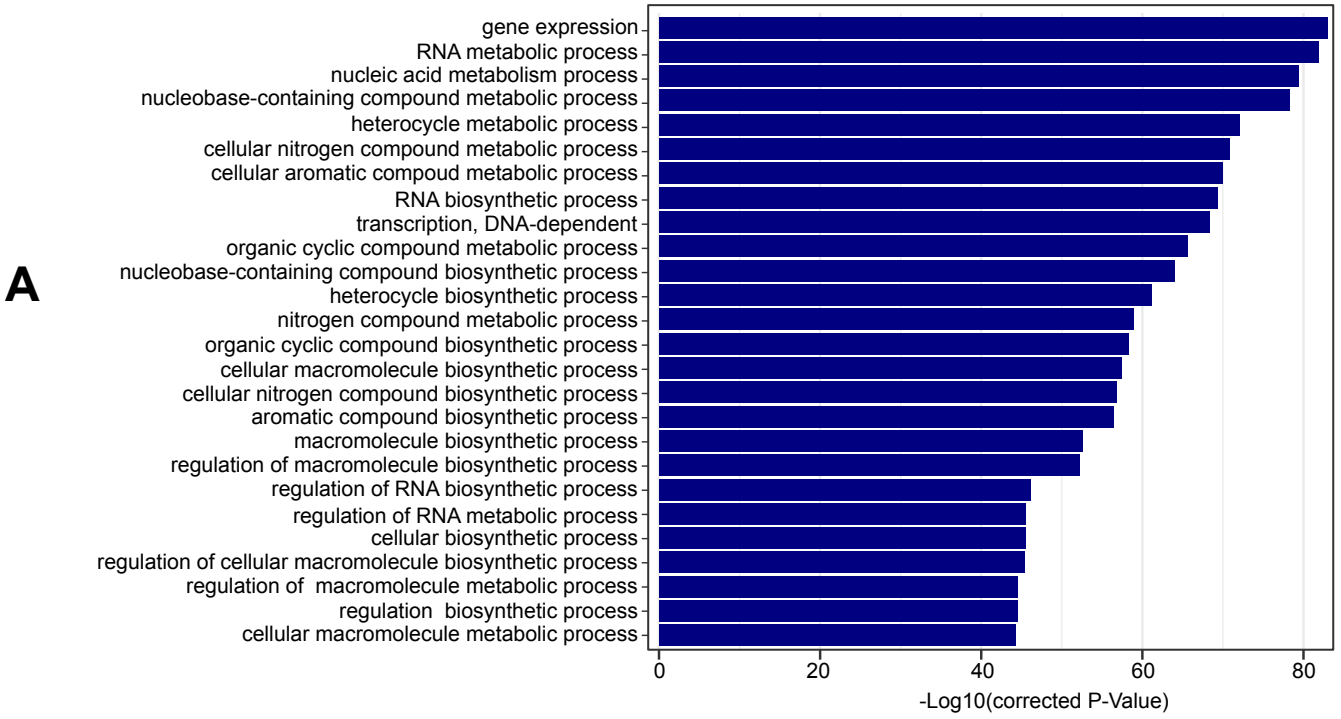
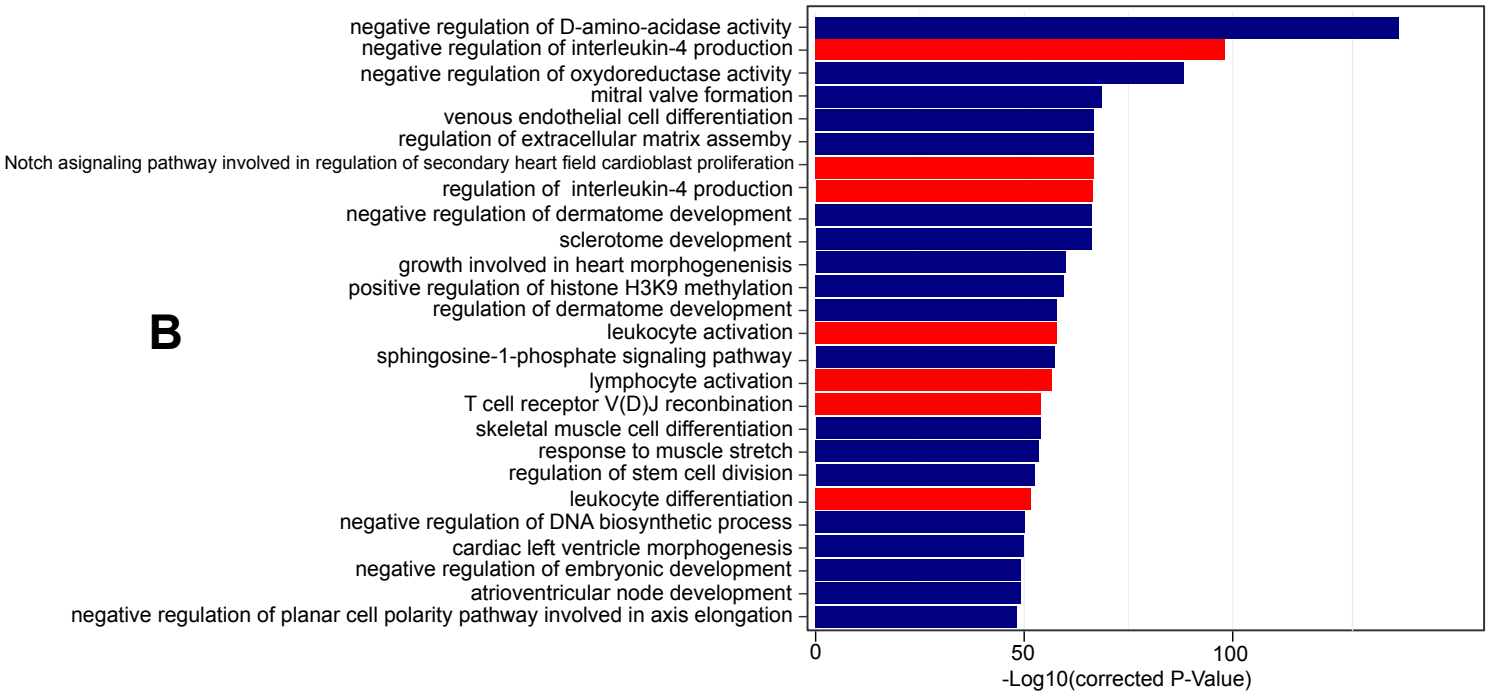

B
0.77
0.75
0.74
0.52
Density
2
1
0
0.00
0.25
0.50
0.75
1.00
Specificity score (Tau)
mRNA
LncRNA_Known
LncRNA_Trimarchi
LncRNA_Atak

Figure 3


A
gene expression
RNA metabolic process
nucleic acid metabolism process
nucleobase-containing compound metabolic process
heterocycle metabolic process
cellular nitrogen compound metabolic process
cellular aromatic compoud metabolic process
RNA biosynthetic process
transcription, DNA-dependent
organic cyclic compound metabolic process
nucleobase-containing compound biosynthetic process
heterocycle biosynthetic process
nitrogen compound metabolic process
organic cyclic compound biosynthetic process
cellular macromolecule biosynthetic process
cellular nitrogen compound biosynthetic process
aromatic compound biosynthetic process
macromolecule biosynthetic process
regulation of macromolecule biosynthetic process
regulation of RNA biosynthetic process
regulation of RNA metabolic process
cellular biosynthetic process
regulation of cellular macromolecule biosynthetic process
regulation of macromolecule metabolic process
regulation biosynthetic process
cellular macromolecule metabolic process
0
20
40
60
80
-Log10(corrected P-Value)


B
negative regulation of D-amino-acidase activity
negative regulation of interleukin-4 production
negative regulation of oxydoreductase activity
mitral valve formation
venous endothelial cell differentiation
regulation of extracellular matrix assemby
Notch asignaling pathway involved in regulation of secondary heart field cardioblast proliferation
regulation of interleukin-4 production
negative regulation of dermatome development
sclerotome development
growth involved in heart morphogenesis
positive regulation of histone H3K9 methylation
regulation of dermatome development
leukocyte activation
sphingosine-1-phosphate signaling pathway
lymphocyte activation
T cell receptor V(D)J reconbination
skeletal muscle cell differentiation
response to muscle stretch
regulation of stem cell division
leukocyte differentiation
negative regulation of DNA biosynthetic process
cardiac left ventricle morphogenesis
negative regulation of embryonic development
atrioventricular node development
negative regulation of planar cell polarity pathway involved in axis elongation
0
50
100
-Log10(corrected P-Value)

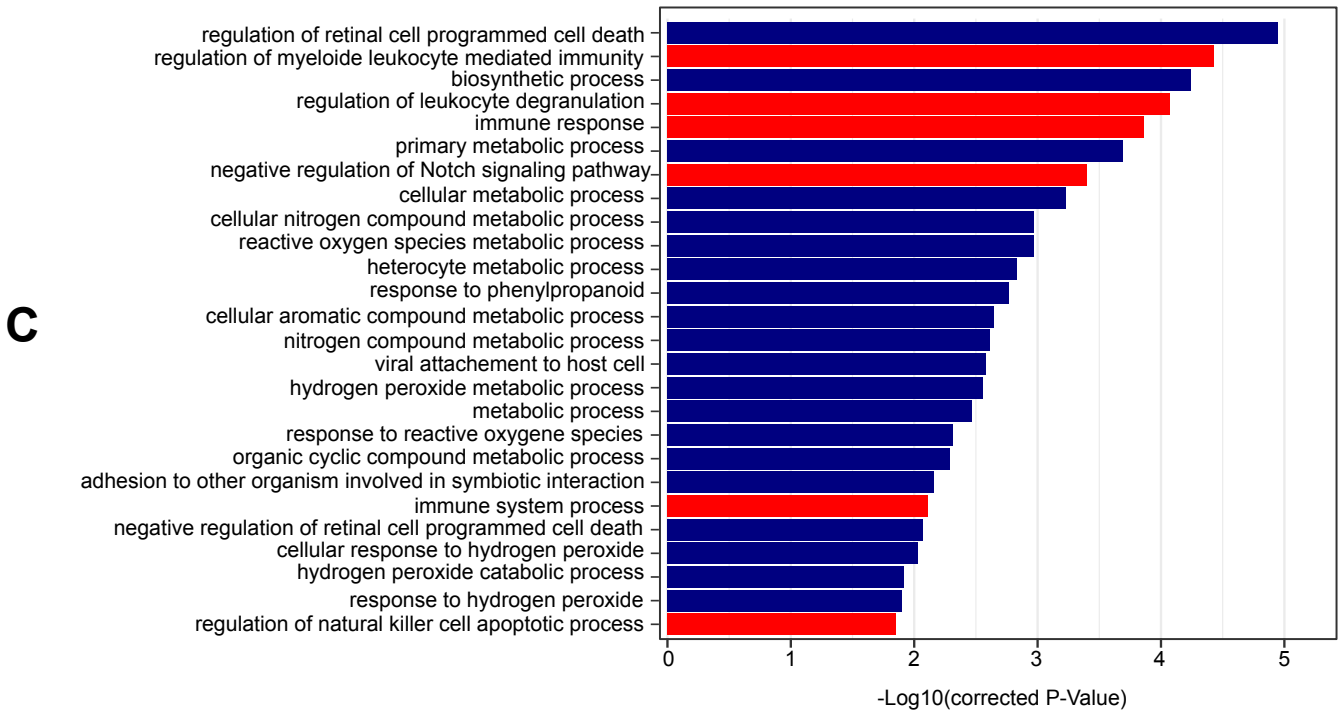

C
regulation of retinal cell programmed cell death
regulation of myeloide leukocyte mediated immunity
biosynthetic process
regulation of leukocyte degranulation
immune response
primary metabolic process
negative regulation of Notch signaling pathway
cellular metabolic process
cellular nitrogen compound metabolic process
reactive oxygen species metabolic process
heterocyte metabolic process
response to phenylpropanoid
cellular aromatic compound metabolic process
nitrogen compound metabolic process
viral attachement to host cell
hydrogen peroxide metabolic process
metabolic process
response to reactive oxygene species
organic cyclic compound metabolic process
adhesion to other organism involved in symbiotic interaction
immune system process
negative regulation of retinal cell programmed cell death
cellular response to hydrogen peroxide
hydrogen peroxide catabolic process
response to hydrogen peroxide
regulation of natural killer cell apoptotic process
0
1
2
3
4
5
-Log10(corrected P-Value)

Figure 4

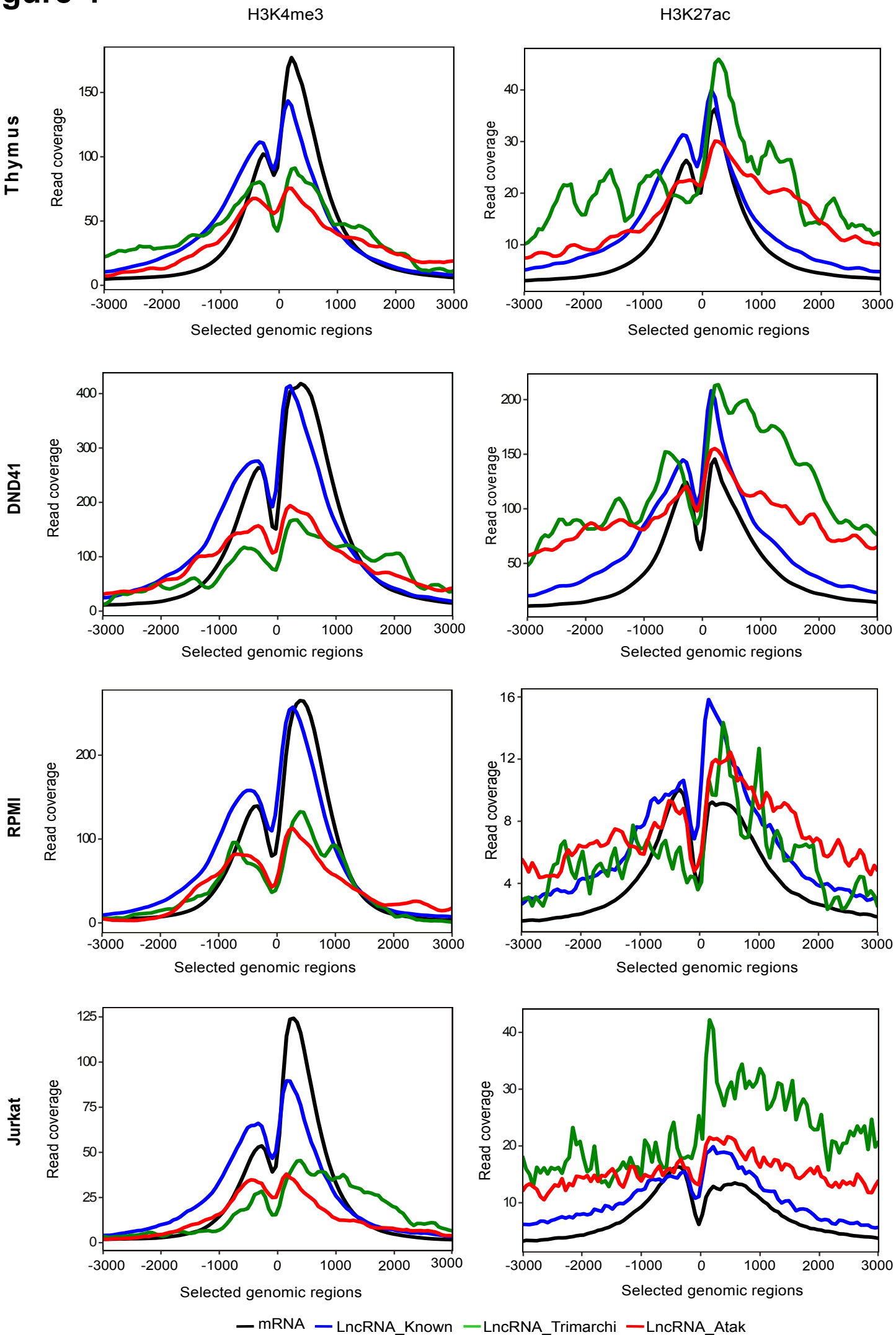
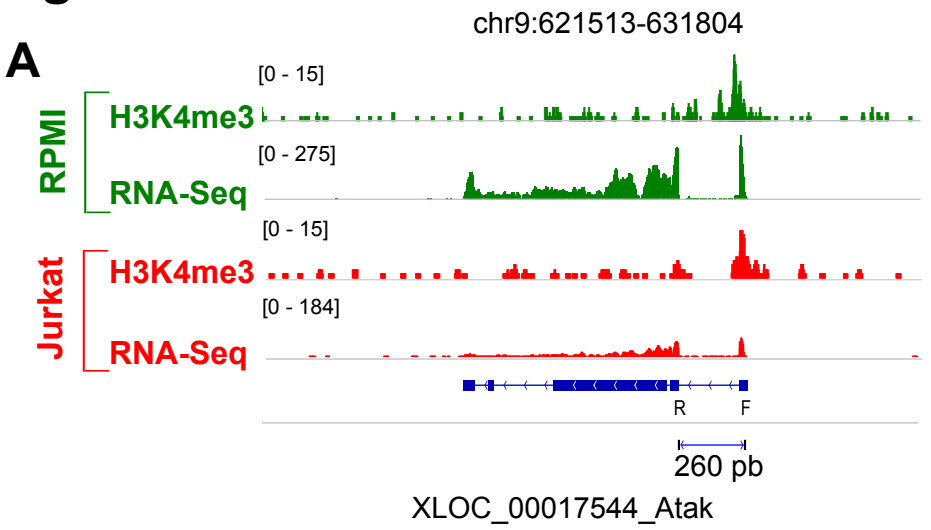
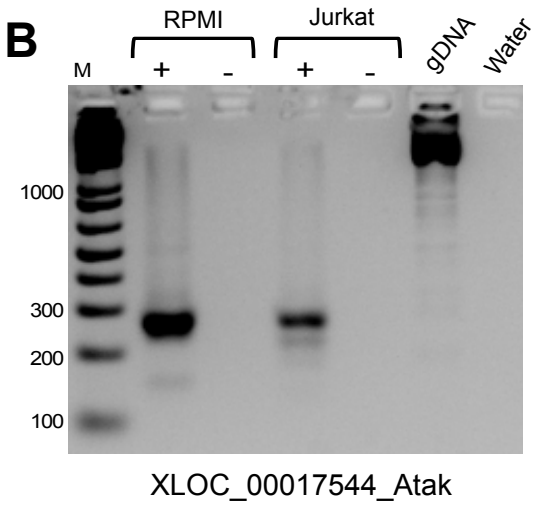
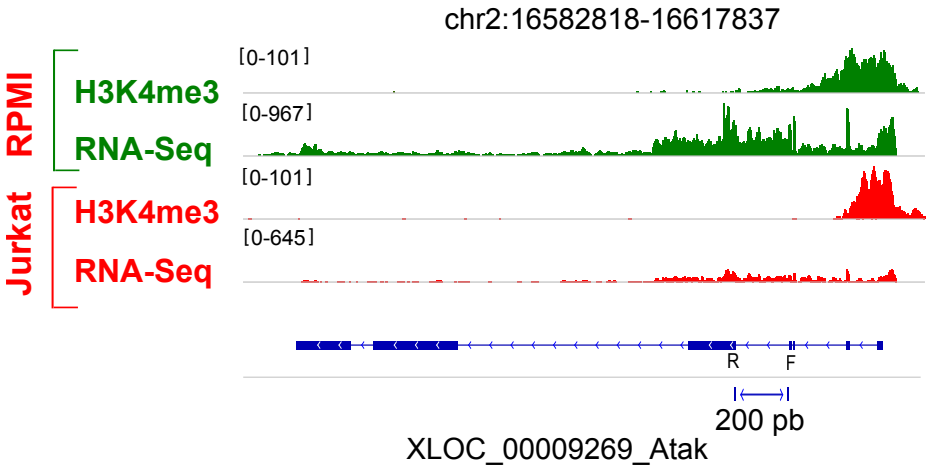
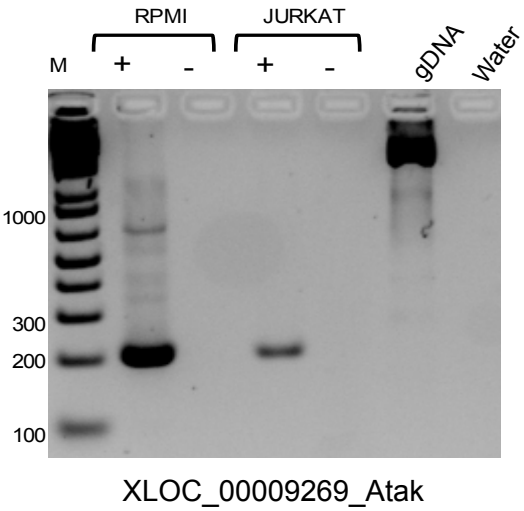

H3K4me3
H3K27ac
Thymus
DND41
RPMI
Jurkat
Read coverage
Selected genomic regions
-3000
-2000
-1000
0
1000
2000
3000
mRNA
LncRNA_Known
LncRNA_Trimarchi
LncRNA_Atak

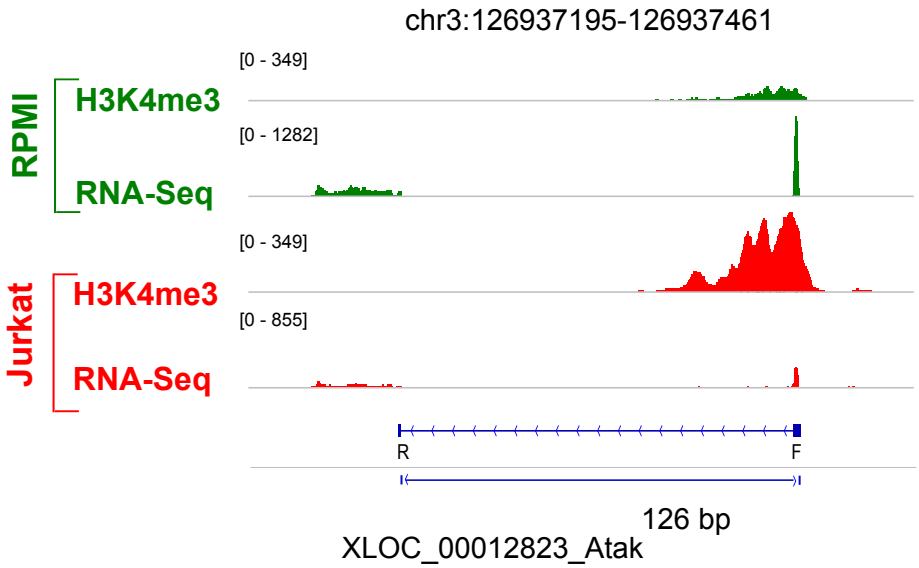
Figure 5

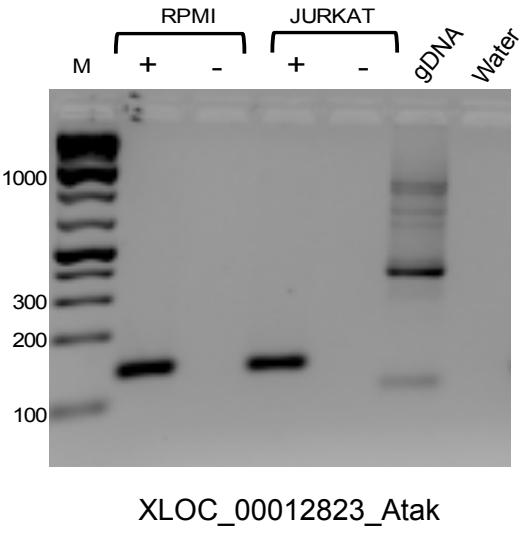

A
chr9:621513-631804
RPMI
H3K4me3
[0 - 15]
RNA-Seq
[0 - 275]
Jurkat
H3K4me3
[0 - 15]
RNA-Seq
[0 - 184]
R
F
260 pb
XLOC_00017544_Atak


B
RPMI
Jurkat
M
+
-
+
-
gDNA
Water
1000
300
200
100
XLOC_00017544_Atak


chr2:16582818-16617837
RPMI
H3K4me3
[0-101]
RNA-Seq
[0-967]
Jurkat
H3K4me3
[0-101]
RNA-Seq
[0-645]
R
F
200 pb
XLOC_00009269_Atak


RPMI
JURKAT
M
+
-
+
-
gDNA
Water
1000
300
200
100
XLOC_00009269_Atak


chr3:126937195-126937461
RPMI
H3K4me3
[0 - 349]
RNA-Seq
[0 - 1282]
Jurkat
H3K4me3
[0 - 349]
RNA-Seq
[0 - 855]
R
F
126 bp
XLOC_00012823_Atak


RPMI
JURKAT
M
+
-
+
-
gDNA
Water
1000
300
200
100
XLOC_00012823_Atak


200
100
MALAT1

Figure 6

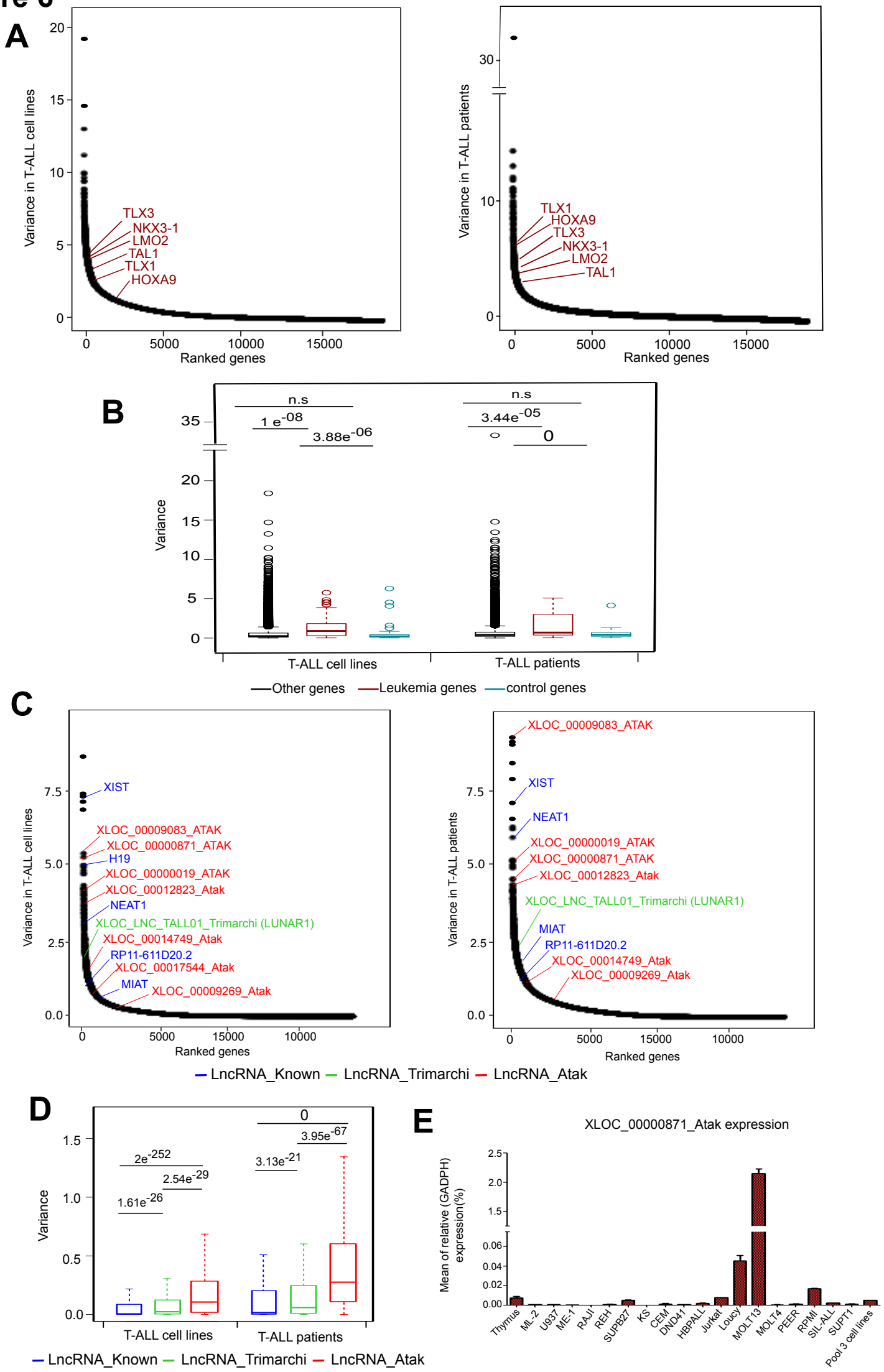
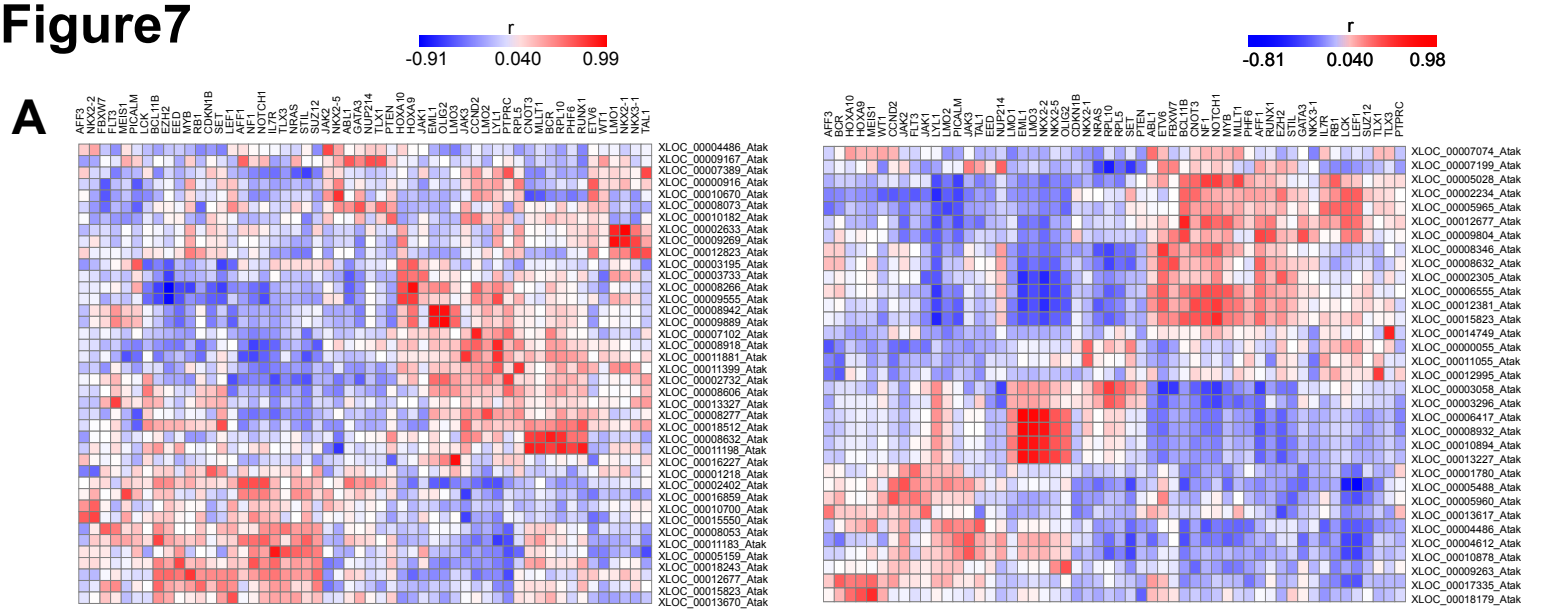
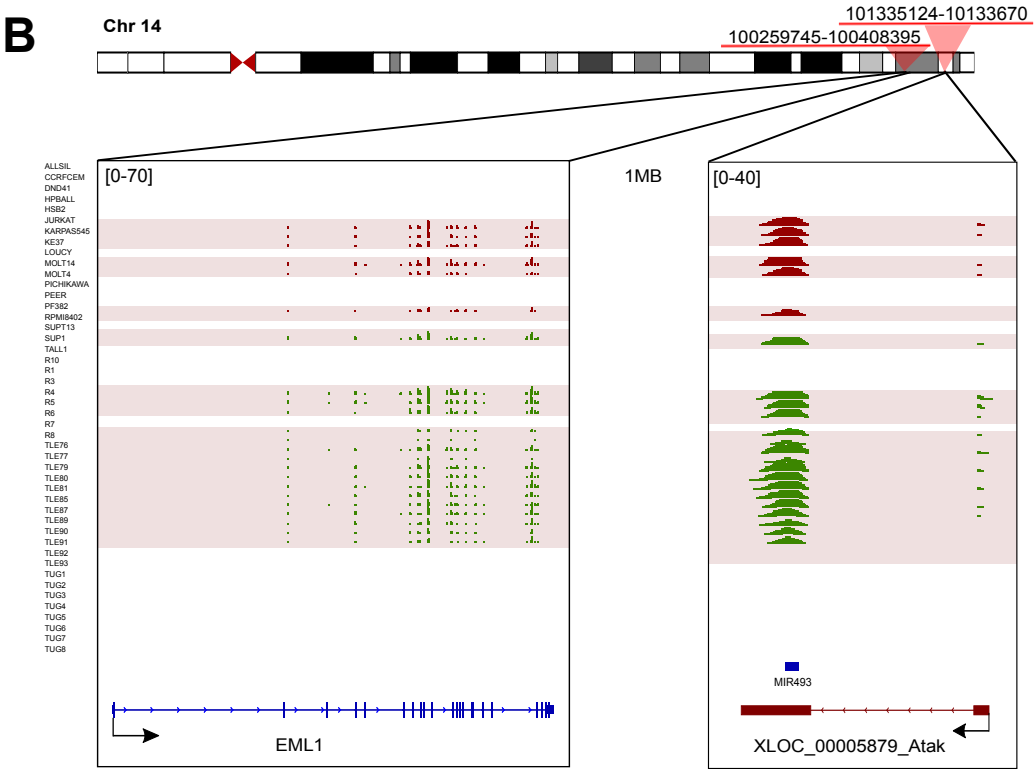
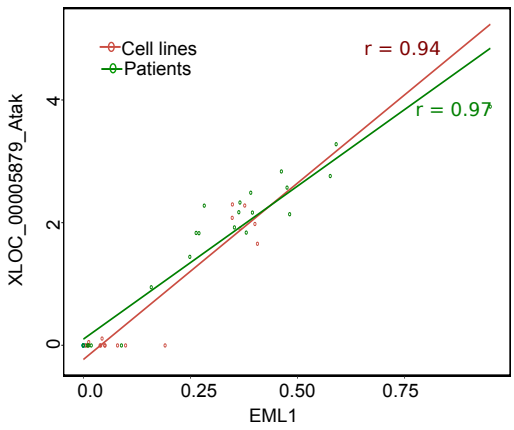
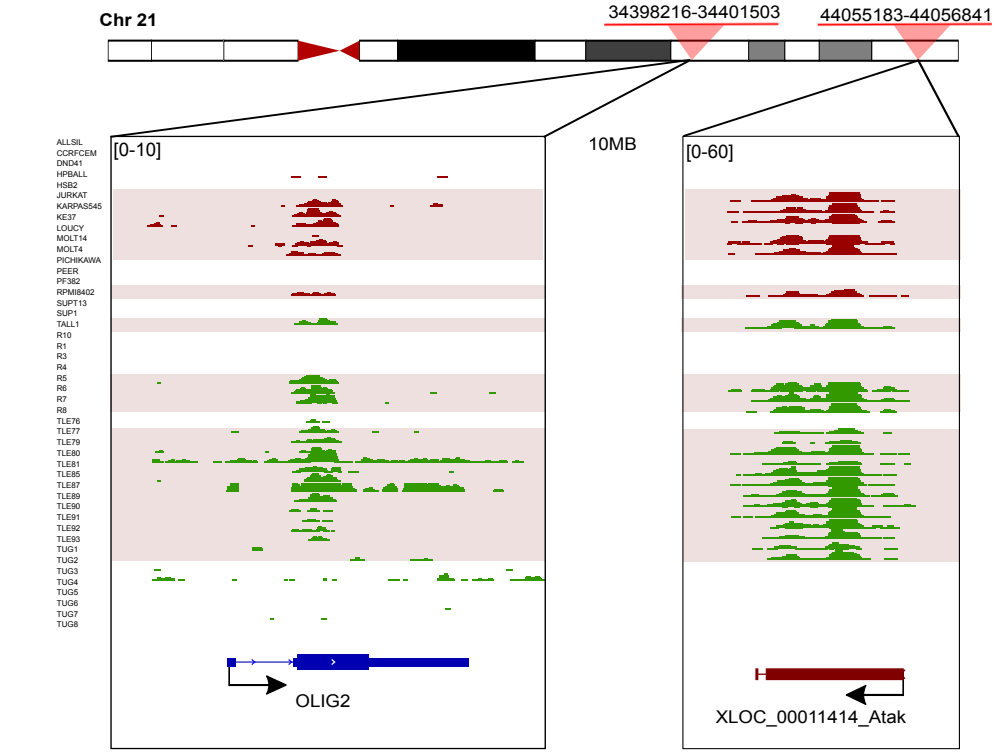

A
Variance in T-ALL cell lines
Ranked genes
TLX3
NKX3-1
LMO2
TAL1
TLX1
HOXA9
Variance in T-ALL patients
Ranked genes
TLX1
HOXA9
TLX3
NKX3-1
LMO2
TAL1
B
Variance
n.s
1 e-08
3.88e-06
n.s
3.44e-05
0
T-ALL cell lines
T-ALL patients
Other genes
Leukemia genes
control genes
C
Variance in T-ALL cell lines
Ranked genes
XIST
XLOC_00009083_ATAK
XLOC_00000871_ATAK
H19
XLOC_00000019_ATAK
XLOC_00012823_Atak
NEAT1
XLOC_LNC_TALL01_Trimarchi (LUNAR1)
XLOC_00014749_Atak
RP11-611D20.2
XLOC_00017544_Atak
MIAT
XLOC_00009269_Atak
Variance in T-ALL patients
Ranked genes
XLOC_00009083_ATAK
XIST
NEAT1
XLOC_00000019_ATAK
XLOC_00000871_ATAK
XLOC_00012823_Atak
XLOC_LNC_TALL01_Trimarchi (LUNAR1)
MIAT
RP11-611D20.2
XLOC_00014749_Atak
XLOC_00009269_Atak
LncRNA_Known
LncRNA_Trimarchi
LncRNA_Atak
D
Variance
2e-252
2.54e-29
1.61e-26
0
3.95e-67
3.13e-21
T-ALL cell lines
T-ALL patients
LncRNA_Known
LncRNA_Trimarchi
LncRNA_Atak
E
XLOC_00000871_Atak expression
Mean of relative (GADPH) expression(%)
Thymus
ML-2
U937
ME-1
RAJI
REH
SUPB27
KS
CEM
DND41
HBPALL
Jurkat
Loucy
MOLT13
MOLT4
PEER
RPMI
SIL-ALL
SUPT1
Pool 3 cell lines

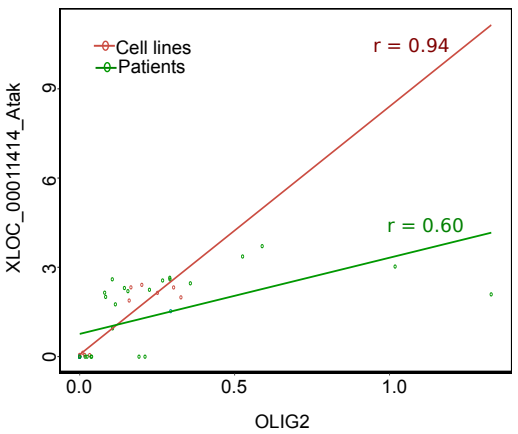
Figure7


A
r
-0.91
0.040
0.99
r
-0.81
0.040
0.98


B
Chr 14
101335124-10133670
100259745-100408395
[0-70]
1MB
[0-40]
EML1
MIR493
XLOC_00005879_Atak


C
Cell lines
Patients
r = 0.94
r = 0.97
XLOC_00005879_Atak
EML1


Chr 21
34398216-34401503
44055183-44056841
[0-10]
10MB
[0-60]
OLIG2
XLOC_00011414_Atak


Cell lines
Patients
r = 0.94
r = 0.60
XLOC_00011414_Atak
OLIG2