

eScholarship@UMassChan

Flnc: Machine Learning Improves the Identification of Novel Long Noncoding RNAs from Stand-Alone RNA-Seq Data

Item Type	Journal Article
Authors	Li, Zixiu;Zhou, Peng;Kwon, Euijin;Fitzgerald, Katherine A;Weng, Zhiping;Zhou, Chan
Citation	Li Z, Zhou P, Kwon E, Fitzgerald KA, Weng Z, Zhou C. Flnc: Machine Learning Improves the Identification of Novel Long Noncoding RNAs from Stand-Alone RNA-Seq Data. Noncoding RNA. 2022 Oct 13;8(5):70. doi: 10.3390/ncrna8050070. PMID: 36287122; PMCID: PMC9607125.
DOI	10.3390/ncrna8050070
Journal	Non-coding RNA
Rights	Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).; Attribution 4.0 International
Download date	2026-05-18 16:19:00
Item License	http://creativecommons.org/licenses/by/4.0/
Link to Item	https://hdl.handle.net/20.500.14038/51448

Article

Finc: Machine Learning Improves the Identification of Novel Long Noncoding RNAs from Stand-Alone RNA-Seq Data

Zixiu Li ¹, Peng Zhou ¹, Euijin Kwon ^{1,2}, Katherine A. Fitzgerald ³, Zhiping Weng ²  and Chan Zhou ^{1,2,4,5,*}

¹ Division of Biostatistics and Health Services Research, Department of Population and Quantitative Health Sciences, University of Massachusetts Chan Medical School, Worcester, MA 01605, USA

² Program in Bioinformatics and Integrative Biology, University of Massachusetts Chan Medical School, Worcester, MA 01605, USA

³ Program in Innate Immunity, Division of Infectious Disease and Immunology, Department of Medicine, University of Massachusetts Chan Medical School, Worcester, MA 01605, USA

⁴ The RNA Therapeutics Institute, University of Massachusetts Chan Medical School, Worcester, MA 01605, USA

⁵ UMass Cancer Center, University of Massachusetts Chan Medical School, Worcester, MA 01605, USA

* Correspondence: chan.zhou@umassmed.edu; Tel.: +1-508-856-8972

Abstract: Long noncoding RNAs (lncRNAs) play critical regulatory roles in human development and disease. Although there are over 100,000 samples with available RNA sequencing (RNA-seq) data, many lncRNAs have yet to be annotated. The conventional approach to identifying novel lncRNAs from RNA-seq data is to find transcripts without coding potential but this approach has a false discovery rate of 30–75%. Other existing methods either identify only multi-exon lncRNAs, missing single-exon lncRNAs, or require transcriptional initiation profiling data (such as H3K4me3 ChIP-seq data), which is unavailable for many samples with RNA-seq data. Because of these limitations, current methods cannot accurately identify novel lncRNAs from existing RNA-seq data. To address this problem, we have developed software, *Finc*, to accurately identify both novel and annotated full-length lncRNAs, including single-exon lncRNAs, directly from RNA-seq data without requiring transcriptional initiation profiles. *Finc* integrates machine learning models built by incorporating four types of features: transcript length, promoter signature, multiple exons, and genomic location. *Finc* achieves state-of-the-art prediction power with an AUROC score over 0.92. *Finc* significantly improves the prediction accuracy from less than 50% using the conventional approach to over 85%. *Finc* is available via GitHub platform.

Keywords: lncRNA; machine learning; RNA-seq; tool; unannotated



Citation: Li, Z.; Zhou, P.; Kwon, E.; Fitzgerald, K.A.; Weng, Z.; Zhou, C. *Finc*: Machine Learning Improves the Identification of Novel Long Noncoding RNAs from Stand-Alone RNA-Seq Data. *Non-Coding RNA* **2022**, *8*, 70. <https://doi.org/10.3390/ncrna8050070>

Academic Editors: Ling Yang and Laura Poliseno

Received: 12 September 2022

Accepted: 6 October 2022

Published: 13 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Only approximately 1% of the human genome and less than 3% of RNA transcripts encode proteins [1]. Long noncoding RNAs (lncRNAs) are a subset of noncoding RNAs longer than 200 nucleotides. Similar to messenger RNAs (mRNAs), most lncRNAs contain 5' caps and 3' polyadenylation (polyA) tails [2]. lncRNAs play regulatory roles in numerous biological processes, including human stem cell development and immunity [2–9], and they have been implicated in many diseases, including neurological disorders, cardiovascular, lung, and liver diseases, infectious diseases, and cancer [9–11]. lncRNAs are also associated with complex genetic traits, playing important roles in gene regulation by functioning as protein scaffolds, guiding DNA–protein interactions, controlling post-transcriptional regulation, and functioning as cis-regulatory elements at enhancers [12]. However, because evolutionary constraints in noncoding regions are lower than in coding regions, lncRNAs are less conserved and evolve faster than coding genes [13]. Increasingly, studies have shown that lncRNAs exhibit striking disease-specific expression patterns and are potential drivers and modifiers of disease [2,14]. Current databases, such as GENCODE [15], NON-CODE [16], and LNCipedia [17], lack annotations for many disease- or cell-type-specific

lncRNAs. For example, we previously found that approximately 40% of lncRNAs expressed in human hepatic stellate cells are not yet included in the database [18]. Therefore, it is critical to create a universal tool for the identification of novel lncRNAs which have not been annotated in the current databases, and this tool can be applied across various diseases and developmental processes.

Advances in high-throughput sequencing techniques have produced a large amount of publicly available RNA sequencing data from various tissues, cell lines, and disease models. To date, there are more than 12,700 study series in the NCBI Gene Expression Omnibus (GEO) database with available RNA-seq datasets. This large amount of data offers great opportunities to identify novel lncRNAs expressed in specific biological samples. In addition, the rapid development of computational methods for analyzing sequencing data, including methods for transcript assembly (e.g., Scripture [19], Trinity [20], Cufflink [21], StringTie [22], Strawberry [23], and TransComb [24]) and methods for examining the coding abilities of transcripts (e.g., CPAT [25], LGC [26], PLEK [27], and CPPred [28]), have allowed researchers to generate catalogs of putative lncRNAs—sembled transcripts without protein-coding potential.

However, these methods cannot distinguish between true lncRNAs and false lncRNAs among the putative lncRNAs. Most false lncRNAs are nonfunctional transcribed fragments and transcriptional noise. Unlike the false lncRNAs, true lncRNAs are high-confidence full-length lncRNA transcripts that include transcriptional start sites (TSSs). Additionally, true lncRNAs cannot be distinguished from false lncRNAs based purely on expression levels, because some known functional lncRNAs have low expression levels, such as *VELUCT* [29].

Several approaches have been developed which improve the accuracy of true lncRNAs identification, including examining mammalian conservation, selecting multiple-exon transcripts, or integrating additional transcriptional initiation data to determine TSSs. Each of these suffers from drawbacks. Approaches that identify lncRNA using mammalian conservation [30,31] can fail to detect human-specific lncRNAs, whereas selecting only multiple-exon transcripts to identify multiple-exon lncRNAs have improved accuracy, but will fail to detect single-exon lncRNAs [32]. As over 3000 single-exon lncRNA transcripts have been curated in the GENCODE database, excluding single-exon transcripts in lncRNA identification will overlook many important single-exon lncRNAs. For example, *MALAT1* is a single-exon lncRNA with critical functions in various diseases, ranging from diabetes to cancer [33]. The trimethylation of histone H3 at lysine 4 (H3K4me3) is a chromatin modification known to mark transcription start sites of active genes [34], including lncRNA genes. Therefore, H3K4me3 profiling data (ChIP-seq or CUT-RUN seq data) are commonly used to identify lncRNAs. However, this approach relies on the existence of matched H3K4me3 profiling data and over 96% of the study series with available RNA-seq data in the GEO database lack matched H3K4me3 profiling data. Generating matched H3K4me3 data is often impractical due to limitations on cost, time, and sample availability. This is especially true for patient-derived clinical samples. Therefore, we sought to improve the accuracy of novel lncRNA identification directly from stand-alone (lacking transcription initiation profiles) RNA-seq data.

We developed a machine learning (ML)-based method to distinguish between true and false lncRNAs. Incorporating this new method, we developed *Flnc*, a software program that can directly identify true lncRNAs, including novel and annotated lncRNAs, from stand-alone RNA-seq data. We assessed *Flnc* by comparing the results produced using only RNA-seq data to our previous method [18], which uses RNA-seq and H3K4me3 ChIP-seq data. In five independent test datasets, the *Flnc* pipeline significantly reduced the rate of false positives and achieved over 85% prediction accuracy, significantly outperforming the prediction accuracy of the conventional RNA-seq only method (50%). The *Flnc* pipeline integrates pre-processing, mapping, transcript assembly, evaluation of protein-coding ability, and evaluation of transcripts using our machine-learning algorithm. The *Flnc* pipeline, which is implemented on the user-friendly Singularity platform, has minimal prerequisites and is easily portable. The *Flnc* can also run with parallelize tasks to optimizes

computing resources. The *FInc* is accessible from <https://github.com/CZhouLab/FInc> (accessed on 10 September 2022).

2. Results

2.1. Generation of a Benchmark Dataset of True and False lncRNAs

To benchmark our new tool, we first updated our previous computational pipeline to use the latest available tools [18,35] (Figure 1A, see Section 4). This pipeline distinguishes between true and false lncRNAs by first identifying putative lncRNAs from raw RNA-seq data, then examining if putative lncRNAs have H3K4me3 peaks near their 5' ends. When we applied this pipeline to the 46 publicly available datasets with matched RNA-seq and ChIP-seq data, we identified hundreds to thousands of true and false lncRNAs in each individual dataset (Supplementary Figure S1A). Each putative lncRNA found within every dataset was counted as one data point, and this resulted in a total of 244,412 putative lncRNA data points. These putative lncRNAs include 95,265 true lncRNA data points and 149,147 false lncRNA data points. These true and false lncRNA data points constitute our benchmark dataset.

Among all putative lncRNAs in the benchmark dataset, 39% are true lncRNAs, which are supported by both RNA-seq and H3K4me3 ChIP-seq data; these transcripts, with clearly defined transcription start sites (Figure 1C,D), we call true lncRNA. Among these true lncRNAs, 43,941 (over 46%) lncRNAs have not been annotated in the GENCODE database and are called novel lncRNAs.

For the remaining 61% of the putative lncRNAs, we could not locate a transcription start site (Figure 1E–G and Supplementary Figure S1B). These transcripts would be identified as false positive lncRNAs by an RNA-seq only pipeline. Using these criteria, we determined that 30–75% of the putative lncRNAs in each individual dataset were not true lncRNAs (Figure 1E–G and Supplementary Figure S1B). Therefore, without H3K4me3 ChIP-seq data, 30–75% of putative lncRNAs identified from stand-alone RNA-seq can be expected to be false hits.

For each lncRNA track, H3K4me3 peaks mark the site of transcription initiation (black, top). RNA-seq reads supporting true lncRNA transcripts are shown in red (RNA-seq sense). RNA-seq reads supporting annotated genes in the sense strand are shown in green (RNA-seq sense). Antisense transcripts are shown in blue. RNA-seq reads supporting false lncRNAs are shown in gray or in green when these reads continue the RNA-seq reads of supporting protein-coding transcripts. The genomic structure for each gene is shown below the RNA-seq tracks. True lncRNAs are shown in red, false lncRNA are shown in grey, annotated genes in the antisense strand are shown in blue, and annotated genes in the sense strand are shown in green. Boxes represent exons, lines represent introns, and arrows represent the start and direction of transcription. Two annotated genes, *LOC100287042* and *PURPL*, have multiple isoforms. True lncRNAs identified in this study are named with the prefix “lncRNA” followed by the locus number assigned during assembly. False lncRNAs are named with the prefix “false-lnc” followed by the locus number assigned during assembly.

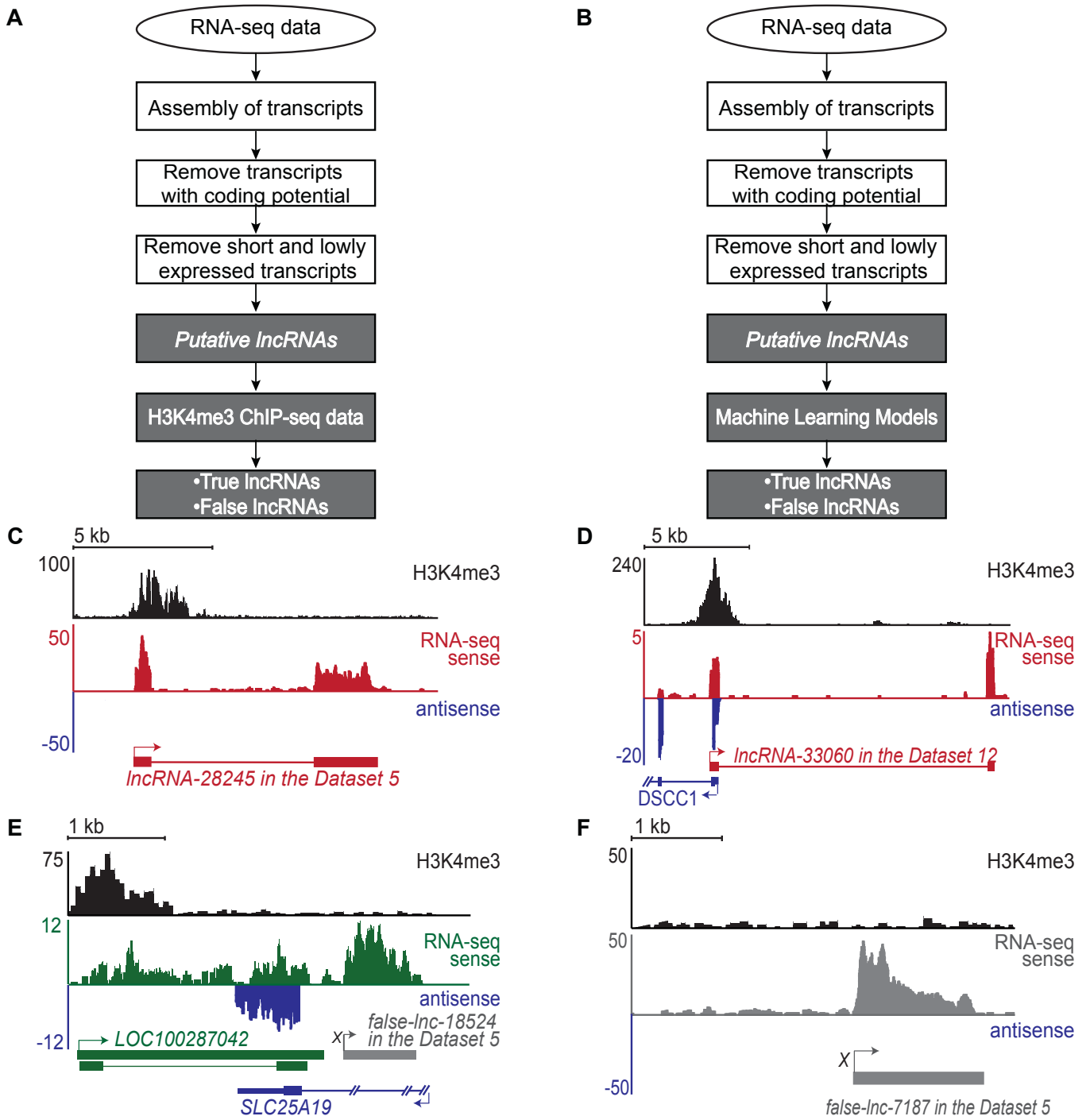


Figure 1. Cont.

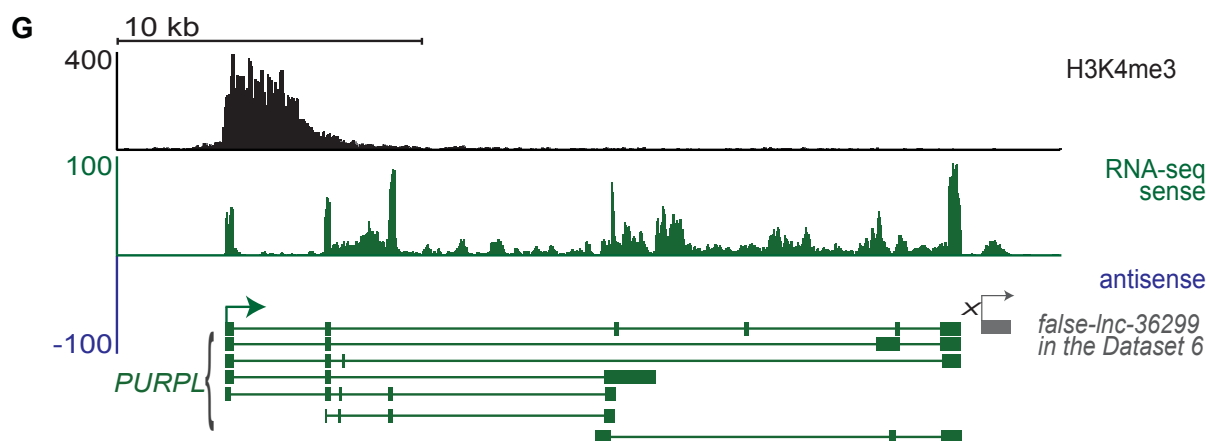


Figure 1. Identification of true and false lncRNAs. (A) Flowchart of the standard computational pipeline for identifying true and false lncRNAs (see Section 4 for details). The identified true lncRNAs include both novel and annotated lncRNAs. The standard pipeline requires both RNA-seq data and H3K4me3 ChIP-seq data generated from the same sample. (B) Flowchart of *Flnc* pipeline. *Flnc* integrates machine learning models to identify true and false lncRNAs directly from RNA-seq data, without matched H3K4me3 profiles. The *Flnc* pipeline includes two steps: first, *Flnc* identifies putative lncRNAs from RNA-seq data (see Section 4 for details). Then, *Flnc* classifies putative lncRNAs as true or false based on built-in machine learning models. The true lncRNAs predicted by *Flnc* include both novel and annotated lncRNAs. (C) A novel true lncRNA detected in Dataset 5. *lncRNA-28245* (red) is located in an intergenic region on chromosome 3. The characteristic H3K4me3 peak identifies it as a true lncRNA. (D) A novel true lncRNA identified in Dataset 12. *lncRNA-33060* (red) is located on the sense strand and transcribed divergently from the promoter of the protein-coding *DSCC1* (blue) gene. (E) A false lncRNA identified in Dataset 5. *False-lnc-18524.1* (grey) is located downstream of *LOC100287042* (green) and antisense to *SLC25A19* (blue). This false lncRNA is supported by an abundance of RNA-seq reads but lacks H3K4me3 peaks at the 5' end. The false-lncRNA could be a fragment of an isoform transcript of the *LOC100287042* gene. (F) An intergenic false lncRNA identified in Dataset 5. This lncRNA is supported by an abundance of RNA-seq reads but lacks a H3K4me3 peak at the 5' end. (G) A false lncRNA identified in Dataset 6. *False-lnc-36299* (grey) is downstream of the protein-coding gene *PURPL* (green). Compared to *PURPL*, *false-lnc-36299* is expressed at relatively low levels; therefore, it may be the result in RNA polymerase continuing beyond the polyA signal sequences when transcribing *PURPL* (transcriptional noise) [36,37].

2.2. Four Genomic Features Can Be Used to Distinguish True and False lncRNAs

To build a new pipeline (Figure 1B) which can identify lncRNAs from RNA-seq data lacking matched H3K4me3 ChIP-seq data, we first needed to determine which features could distinguish between true and false lncRNA. We hypothesized that a combination of four types of features, transcript length, promoter signature, multiple exons, and genomic location, would provide enough information to distinguish between true and false lncRNAs. Therefore, we used our benchmark dataset to examine these features in true and false lncRNAs.

Most false lncRNAs are transcript fragments or transcriptional noise. Therefore, we hypothesized that true lncRNAs would tend to be longer than false lncRNAs. To test this hypothesis, we examined the transcript length of all putative lncRNAs in the benchmark dataset. Read length and sequencing depth differ across various RNA-seq datasets and these factors affect transcript assembly [38–40]. To control for these differences, for each of the 46 datasets, we normalized the transcript length of putative lncRNAs to a value between 0 and 1 (see Section 4). After normalization, the transcript lengths of true lncRNAs were significantly longer than those of false lncRNAs (Figure 2A and Supplementary Figure S2A).

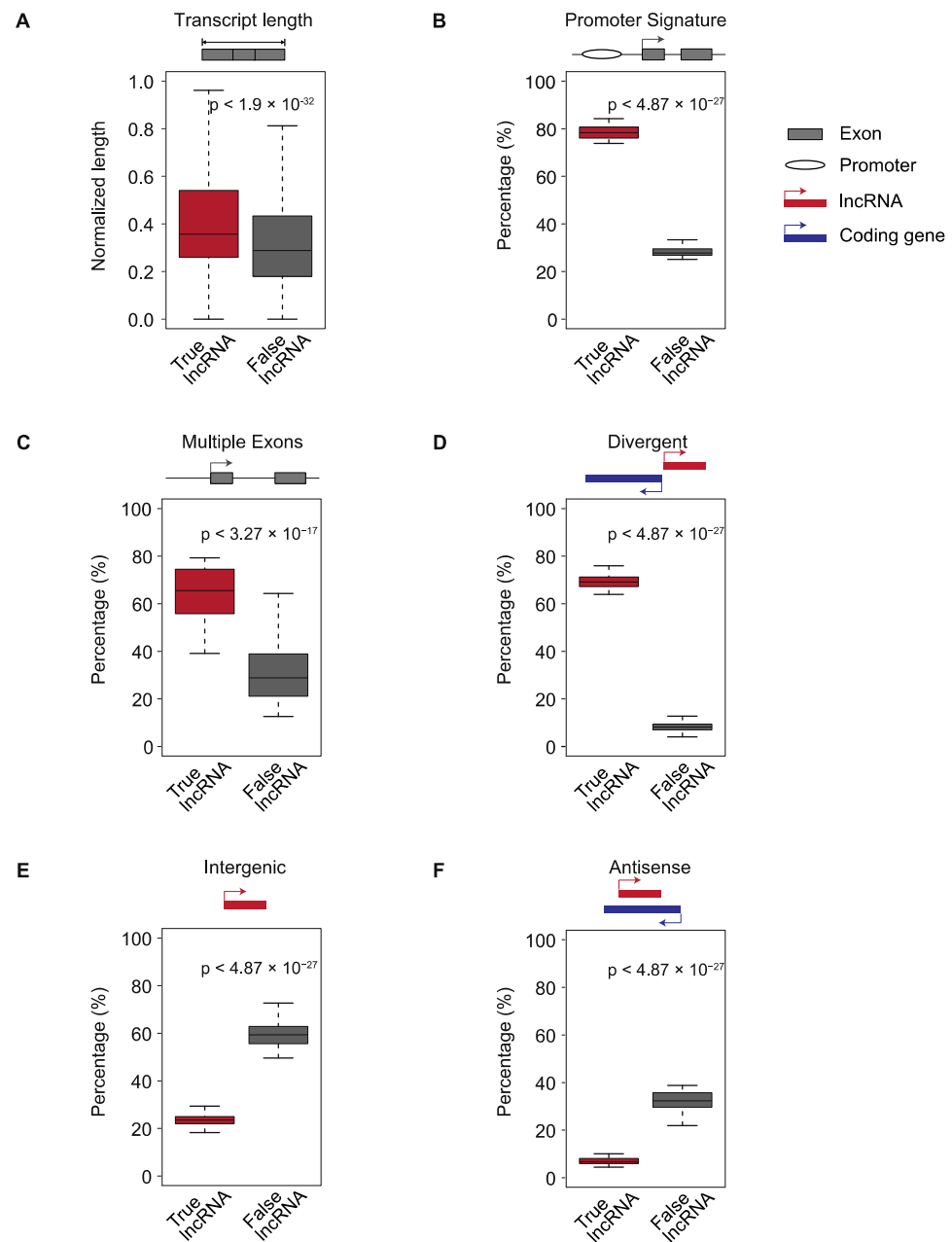


Figure 2. Four features exhibit significant differences between true and false lncRNAs in the 46 benchmark datasets. True lncRNAs can be distinguished from false lncRNAs because they tend to be longer (A), to have a predicted upstream promoter signature (B), and to be more likely to have multiple exons (C). They can also be distinguished by genomic location and are more likely to be divergently transcribed from the promoter of a protein-coding gene (D). True lncRNAs are less likely to be in intergenic regions (E) or antisense (F) to a protein-coding gene. For the graph (A), the boxplot represents the scaled transcript lengths of true and false lncRNAs. For the graphs (B–F), each boxplot represents the percentage of a feature of true and false lncRNAs among the identified putative lncRNAs for each of the 46 benchmark datasets. The p -values were calculated by a two-sided Wilcoxon–Mann–Whitney test.

Whereas RNA fragments and transcripts resulting from noisy transcription lack promoters and TSSs, true lncRNAs should have upstream promoters. Therefore, we used TSSG software [41] to identify putative promoters upstream of true and false lncRNAs and examined the percentage of true and false lncRNAs with upstream promoter signatures.

Almost 80% of true lncRNAs had putative upstream promoters. In contrast, we find this feature in only about 20% of false lncRNAs (Figure 2B and Supplementary Figure S2B).

Because the vast majority of single-exon transcripts result from transcriptional and alignment noise [42,43], we hypothesized that false lncRNAs (transfrags or noise) would have fewer exons than true lncRNAs. To test this hypothesis, we examined the percentage of multiple exon transcripts among true and false lncRNAs. We found that true lncRNAs are significantly more likely to have multiple exons than false lncRNAs (p -value $< 3.27 \times 10^{-17}$; Figure 2C and Supplementary Figure S2C).

Over 60% of lncRNAs have been shown to be divergently transcribed from the promoter regions of protein-coding genes [18,35]. This suggests that genomic context could be used to distinguish between true and false lncRNAs. We examined the true and false lncRNAs in our benchmark dataset, classifying them into three categories—divergent, antisense, and intergenic—based on their genomic locations. Consistent with previous findings, more than 60% of true lncRNAs were divergent transcripts of protein-coding genes, while less than 20% of false lncRNAs in each of the 46 datasets were divergent transcripts (Figure 2D and Supplementary Figure S2D). Additionally, a small fraction of true lncRNAs originate from intergenic regions or are antisense to coding genes. In contrast, approximately 60% of false lncRNAs are in intergenic regions and 30% are antisense to coding genes (Figure 2E,F and Supplementary Figure S2E,F). Therefore, true and false lncRNAs show significantly different patterns in terms of genomic location.

In conclusion, true and false lncRNAs show significant differences in transcript length, promoter signature, multiple exons, and genomic location. Therefore, we integrated these features into ML models to distinguish between true lncRNAs and other noncoding transcripts.

2.3. Training ML Models to Distinguish between True and False lncRNAs

We developed a new computational tool, *Flnc* (Figure 1B) which incorporates the four types of genomic features detailed above into ML models to identify true lncRNA from stand-alone RNA-seq data. We trained seven of the most common ML algorithms using our training set, which comprised 81,420 true and 123,764 false lncRNA that were identified from 41 datasets submitted to the GEO database before 2019. The seven ML algorithms include: logistic regression (LR), k-nearest neighbors (KNN), decision tree (DT), random forest (RF), naïve Bayes (NB), linear kernel support vector machine (SVM), and radial-basis-function (RBF) kernel SVM.

To fit and select the best model for each ML algorithm, we first used the 10-fold cross validation approach to train and evaluate each model using all possible combinations of hyperparameters (Figure 3A). We divided the putative lncRNAs in the training set randomly into 10 non-overlapping subsets (folds). For each ML algorithm, we held one of the 10 subsets of putative lncRNAs aside and trained a model using each set of hyperparameter values on the other 9 subsets. We then evaluated the performance of the model on the remaining held-aside data. We repeated the process 10 times, each time holding aside a different subset of data. We considered the hyperparameter values that defined the model with the best mean F1 score—the harmonic mean value of the precision and sensitivity—the optimal model architecture. Next, we trained the optimal model architecture on the entire training set to build the models for *Flnc*.

On the training set, the random forest, decision tree, and KNN models had the best overall prediction performance based on the F1 score and Area Under the Receiver Operating Characteristic (ROC) Curve (AUROC) score (Figure 3B and Supplementary Table S3). The random forest model resulted in F1 = 0.84 and AUROC = 0.94 values, the decision tree model resulted in F1 = 0.83 and AUROC = 0.93 values, and the KNN model resulted in F1 = 0.82 and AUROC = 0.92 values. Compared to the other models, these three models also have better accuracy, precision, and specificity. In contrast, the linear SVM was the worst performing of all models with respect to the F1 score, sensitivity, and accuracy.

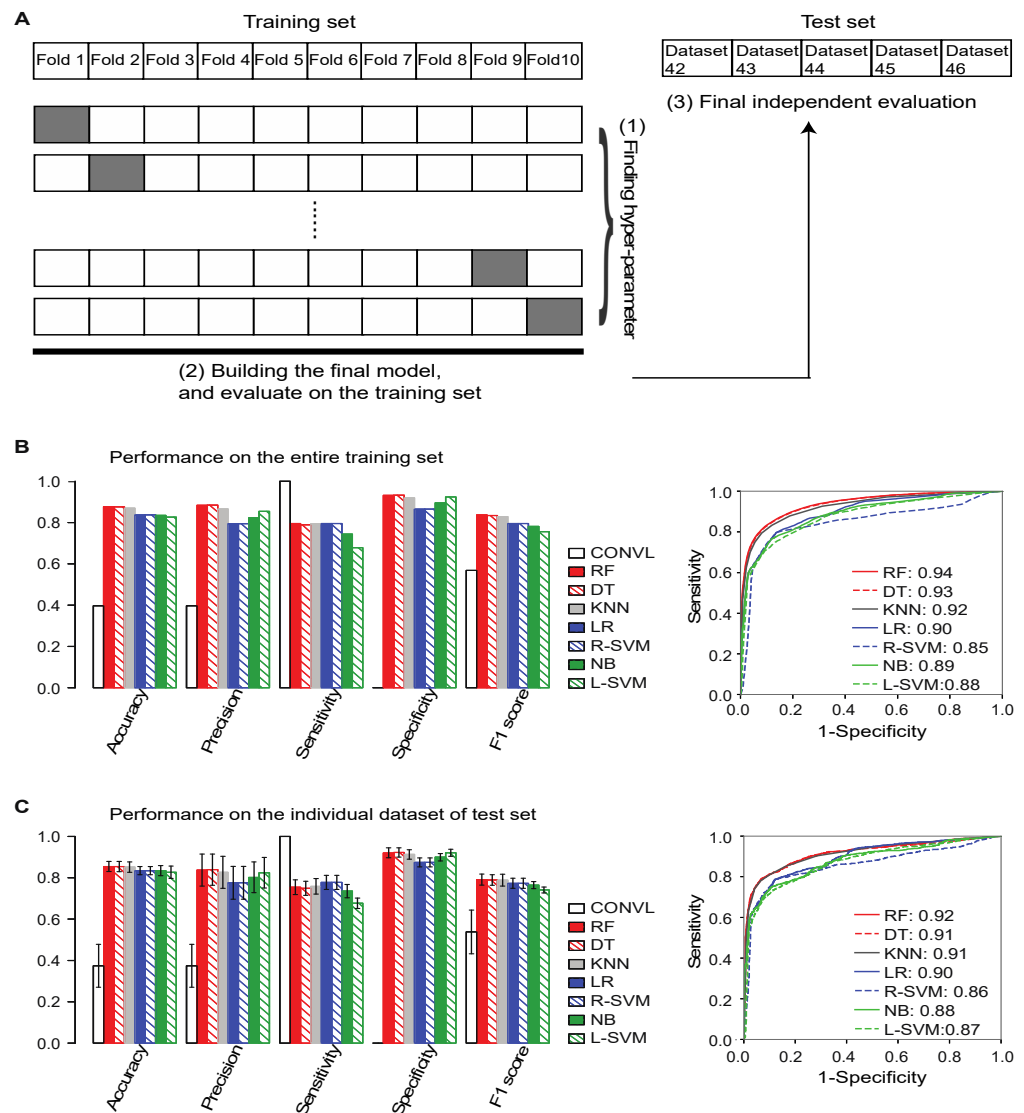


Figure 3. ML model construction and evaluation. (A) The overall architecture of *Flnc* training and testing. The best hyperparameters, specifically the hyperparameters yielding the best mean F1 score, were chosen using the 10-fold cross-validation approach as follows. The training set was randomly split into 10 even folds, 9 of which were used in training the models. The 10th fold was used for validation. We repeated this process holding each individual fold out and training with the other 9 folds. From this, we derived 10 models for each hyperparameter setting. For each of the 10 models, we calculated the mean F1 score for each hyperparameter setting. After selecting the best hyperparameter setting, the final model was built using the entire training set and we examined the performance of the final model on the entire training set. Finally, we evaluated the performance of the final model on the test set, which was composed of the five independent datasets generated in 2019 and 2020 (Datasets 42–46). (B) The seven final models of *Flnc* outperformed the conventional method (without models) on the entire training set, and (C) on the five individual datasets of test set. The left bar graphs of (B,C) show the performance metrics accuracy, precision, sensitivity, specificity and F1 score. The right graphs (B,C) show the ROC curve for each model. The AUROC score is shown next to each ML model. The ROC curve of (C) is the ROC curve of the seven models on the Dataset 46. Please see Supplementary Figure S3B for the ROC curves of the four additional test datasets (Dataset 42–25). The seven models are ranked by the F1 score from largest to smallest. Abbreviations: CONVL, conventional approach; RF, random forest; DT, decision tree; R-SVM, RBF support vector machine; L-SVM, linear support vector machine; LR: linear regression; NB, naïve Bayes; KNN, and k-nearest neighbors.

2.4. *Flnc* Identifies True lncRNAs with up to 87% Prediction Precision

After training the models, we evaluated *Flnc*'s ability to predict true lncRNAs from stand-alone RNA-seq data by testing *Flnc* on a test set composed of five independent datasets that were released to the GEO database after 2019 (Figure 3C). These five datasets were generated from multiple biological samples, including the MOLM-13 human myeloid leukemia cell line, the HUDEP-2 erythroid cell line, Jurkat leukemia cells, and the H1299 non-small cell lung cancer cell line. We evaluated the performance of *Flnc* both on the entire test set (Supplementary Figure S3A and Supplementary Table S3) and on the five individual datasets within the test set (Figure 3C and Supplementary Figure S3B). Consistent with the training set results, the random forest, decision tree, and KNN models have the best overall prediction abilities, as indicated by the F1 and AUROC scores (Figure 3C and Supplementary Figure S3B), although the seven models achieve 72–87% consistency in lncRNA prediction (Supplementary Figure S3C). The three best models achieve 93–96% consistency in lncRNA prediction (Supplementary Figure S3D) with the accuracy of 0.85 or greater, and precision of 0.83 or more. As with the training set, based on the F1 score and accuracy, the linear SVM and naïve Bayes models had the worst performance. Variations in performance between datasets were small (Figure 3C and Supplementary Table S3) with standard deviations in F1 and AUROC scores of less than 0.03 and less than 0.02, respectively (Supplementary Table S3). Furthermore, an ensemble approach, which consists of combining the results from all seven models such that a lncRNA is considered true if all seven models predict it, can further improve prediction precision from 83% to 87% and specificity from 92% to 95% at the cost of reducing the sensitivity (Supplementary Figure S3E).

2.5. Many lncRNAs Identified by *Flnc* Are Novel and Are Supported by H3K4me3 Profiles

Because true lncRNAs include both novel and annotated lncRNAs, we examined the novel lncRNAs among the true lncRNAs predicted by *Flnc* in the five independent test datasets. We found that up to 60% of true lncRNAs predicted by *Flnc* have not yet been annotated in the GENCODE database (Figure 4A). To determine if these novel lncRNAs are true lncRNAs, we examined H3K4me3 ChIP-seq data to determine if these lncRNAs have TSSs; we found that 65–90% of these novel lncRNAs are supported by H3K4me3 ChIP-seq data (Figure 4B). Indeed, a novel lncRNA identified by *Flnc* is almost twice as likely to be confirmed by H3K4me3 than one identified using conventional RNA-seq only methods (Figure 4B). Taking the ensemble approach improved the chance of being confirmed even further to between 74% and 94% (Figure 4B).

Next, we examined the genomic features of novel lncRNAs predicted by *Flnc* in the five independent test datasets. Because the random forest model had the best F1 and AUROC scores, we compared the genomic features of novel lncRNAs predicted with the random forest model with those of novel lncRNAs identified by the H3K4me3 ChIP-seq data (Figure 4C,D). The novel lncRNAs predicted by *Flnc* and those identified by H3K4me3 profiles were similar in terms of transcript length (Figure 4C) and multiple exons (Figure 4D), but the novel lncRNAs predicted by *Flnc* were more likely to have promoter signatures and to be divergent transcripts (Figure 4D). We observed similar trends among all the true lncRNAs predicted by *Flnc* and those determined by H3K4me3 profiles (Supplementary Figure S4A,B).

Additionally, we note that 16–40% of novel lncRNAs predicted by *Flnc* have multiple-exon (Figure 4D) across the five independent test datasets. This means that 60–74% of novel lncRNAs predicted by *Flnc* are single-exon lncRNAs. Similarly, among all the true lncRNAs predicted by *Flnc*, 37–42% of true lncRNAs are single-exon lncRNAs (Supplementary Figure S4D). This indicates that *Flnc* will not miss the single-exon lncRNAs with potential functions.

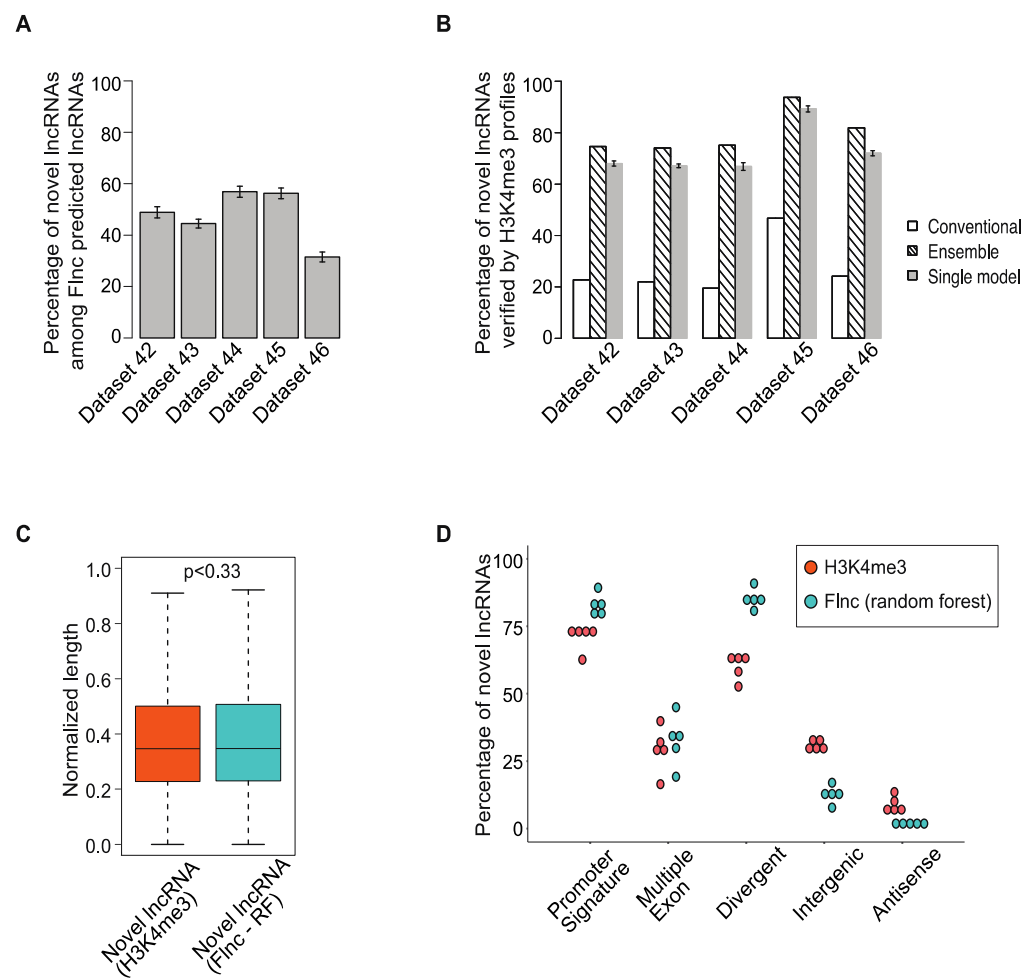


Figure 4. Most novel lncRNAs predicted by *Flnc* are supported by H3K4me3 profiles. **(A)** A large proportion of true lncRNAs predicted by *Flnc* are novel lncRNAs in the five individual test datasets. **(B)** The percentage of novel lncRNAs predicted by the conventional method without models and *Flnc* methods are verified by H3K3me3 profiles. Most novel lncRNAs predicted by the conventional method cannot be verified by H3K4me3 profiles, whereas most of the novel lncRNAs predicted by *Flnc* methods (ensemble approach or each of the seven single models) can be verified by H3K4me3 profiles. The error bars of the single model (gray) bars represent the standard deviations for the results of the seven ML models. **(C)** The novel lncRNAs predicted by *Flnc* (with random forest model) show similar normalized transcript length distribution as the novel lncRNAs determined by H3K4me3 ChIP-seq data. **(D)** The novel lncRNAs predicted by *Flnc* (with the random forest model) includes significantly more divergent transcripts and transcripts with promoter signatures than the novel lncRNAs determined by H3K4me3 ChIP-seq data, whereas multiple exon features exhibit similar percentage between these two groups of novel lncRNAs. Each dot represents one of the five independent datasets in the test set.

2.6. *Flnc* Predicts True lncRNAs in Multiple Types of RNA-Seq Samples

For both the training and test data, we used stranded RNA-seq data generated from polyA-selected RNA. However, many RNA-seq datasets are generated not from polyA RNA, but from ribosomal-RNA (rRNA)-depleted RNA and some are sequenced without strand information. To evaluate the performance of *Flnc* in other types of RNA-seq data, we tested *Flnc* on two stranded RNA-seq datasets that were generated using rRNA depletion (Figure 5A) and on two unstranded polyA RNA-seq datasets (Figure 5B).

On both stranded rRNA-depleted RNA-seq and unstranded polyA-selected RNA-seq data, four models—naïve Bayes, linear SVM, RBF SVM, and logistic regression—had similar F1 scores and based on the F1 scores, these four models performed better than other

three models (Figure 5). On these types of data, the linear SVM model achieved the best prediction precision and accuracy; the naïve Bayes model ranks as the second best by the precision metric. Although the random forest, decision tree, and KNN models ranked as among the best models when run on polyA-selected RNA-seq data, their performance on rRNA-depleted RNA-seq data ranked as the worst by the metrics of accuracy, precision, sensitivity, and F1 score (Figure 5).

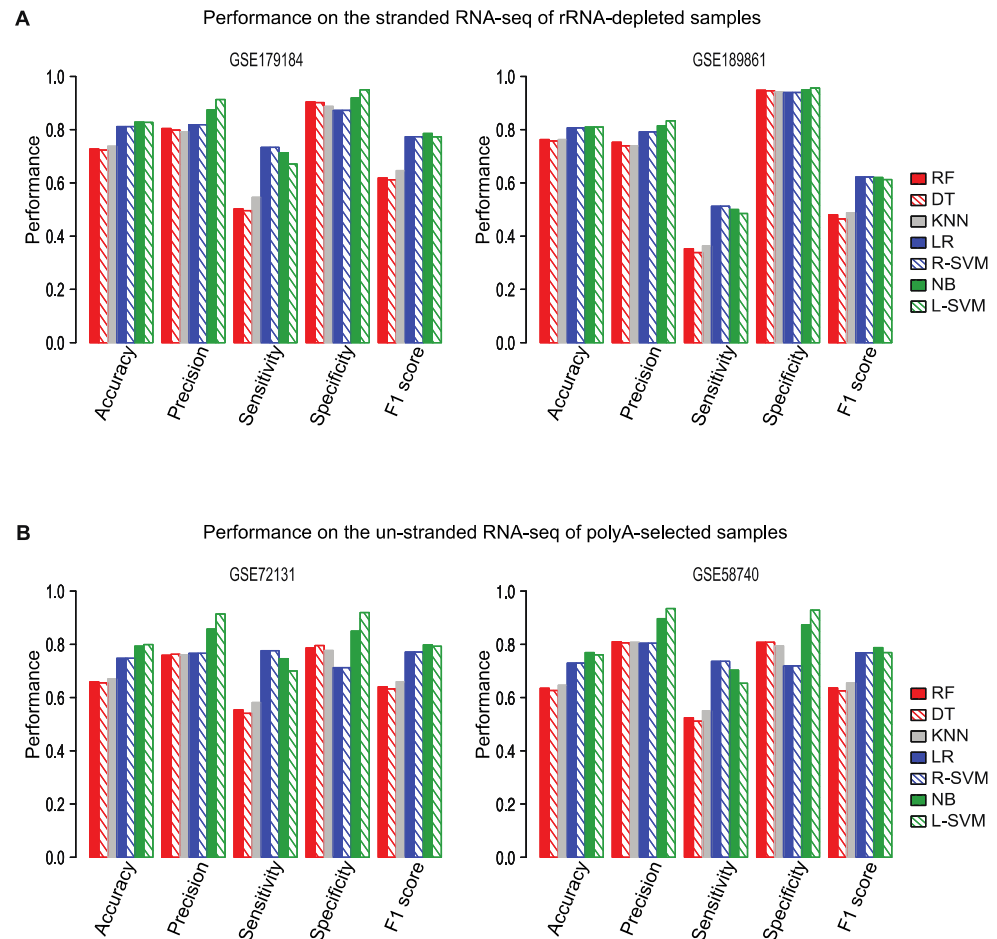


Figure 5. *Flnc* accurately identifies true lncRNAs from rRNA-depleted and unstranded RNA-seq datasets. (A) Performance of *Flnc* on two stranded RNA-seq datasets generated from ribosomal RNA (rRNA)-depleted samples (GSE179184 on the left, GSE189861 on the right). (B) Performance of *Flnc* on two un-stranded polyA-selected RNA-seq datasets (GSE72131 on the left, GSE58740 on the right).

2.7. Divergent Transcription Is the Most Important Feature for Predicting True lncRNAs

We evaluated the significance of the selected genomic features for each ML model. The feature importance score indicates the relative importance of each genomic feature (transcript length, promoter signature, multiple exons, and genomic location) for *Flnc* in detecting true lncRNAs from RNA-seq data. Because the genomic location feature includes three categories (divergent, antisense, and intergenic), we evaluated the feature importance for six categories: transcript length, promoter signature, multiple exons, divergent transcript, antisense transcript, and intergenic transcript. We found that divergent transcript feature is the most important category across all ML models (Figure 6). For three models—random forest, decision tree, and KNN—promoter signature ranks as the second most important feature. In contrast, the intergenic and antisense transcript features are the least important in all models (Figure 6).

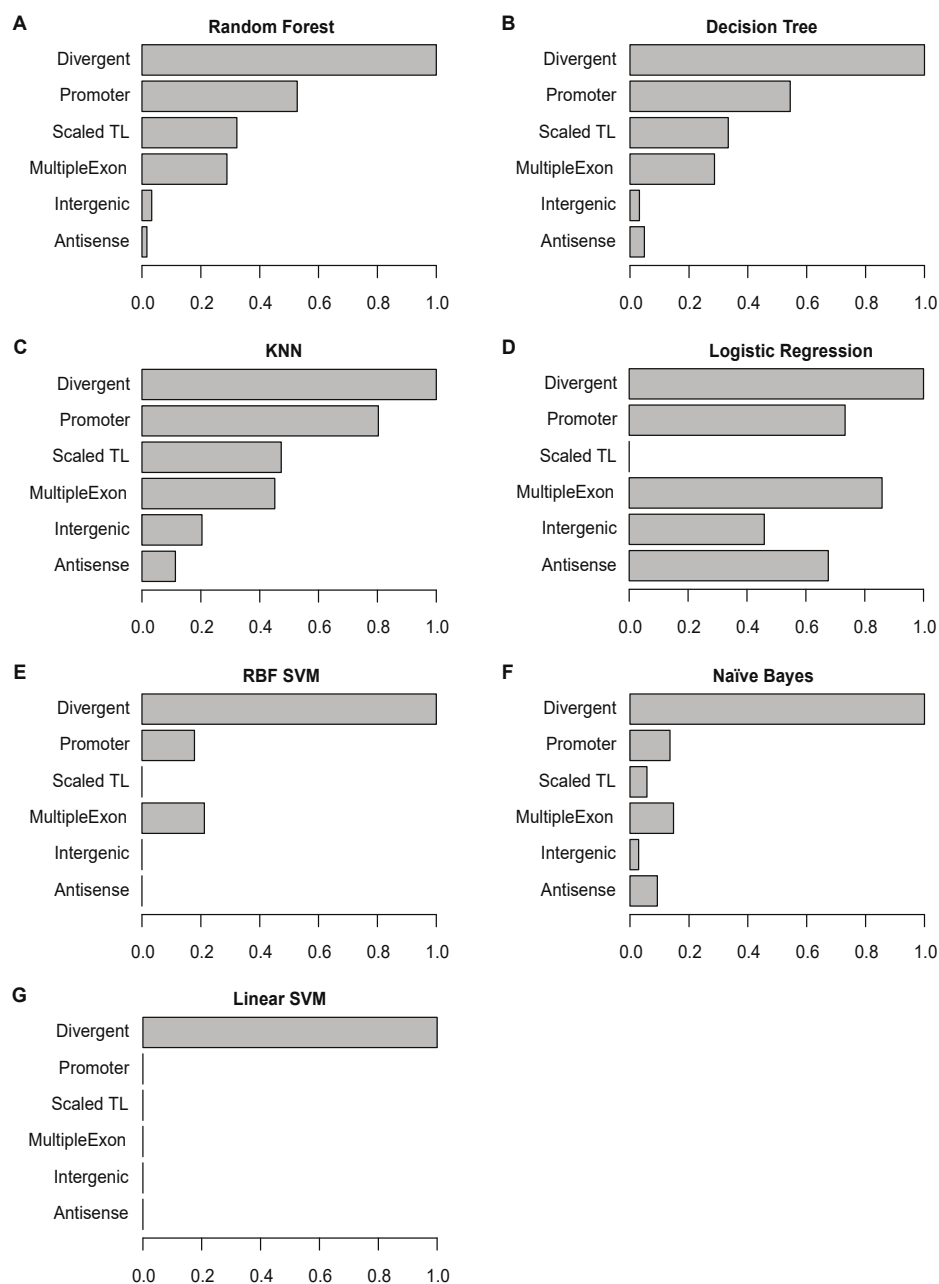


Figure 6. Feature importance for each of the seven final models. Each graph shows the scaled importance scores for each feature for the given model: random forest (A), decision tree (B), KNN (C), logistic regression (D), support vector machine (SVM) with RBF kernel (E), Naïve Bayes (F), and SVM model with linear kernel (G) models. For each model, the importance scores of all features were scaled to (0.1) by dividing the original score of the most significant feature (see Section 4).

2.8. *Flnc* Achieves Similar Performance at the lncRNA Gene Locus Level as at the Transcript Level

More than half (57%) of the true lncRNAs in the benchmark dataset have multiple exons. Among them, 43% of multi-exon lncRNA genes have undergone alternative splicing events, generating multiple lncRNA isoforms from the same lncRNA gene locus. *Flnc* has been trained and tested at the transcript level for each RNA-seq dataset. To examine the performance of *Flnc* at the locus level, we further tested *Flnc* on the lncRNA gene loci which encode only lncRNA transcripts. This test showed that *Flnc* can achieve similar performance at the lncRNA gene locus level as at the lncRNA transcript level (Supplementary Figure S5), although *Flnc* performs better at the transcript level than at the locus level.

3. Conclusions and Discussion

We have developed a comprehensive pipeline and software *Finc*, which uses ML models to accurately identify true lncRNAs from RNA-seq data. *Finc* does not require matched transcription initiation profiles which are usually marked by H3K4me3 histone modifications. We trained the ML models on four types of features, transcript length, promoter signature, multiple exons, and genomic location, which show a high degree of discrepancy between true lncRNAs and false lncRNAs in our benchmark datasets. We benchmarked *Finc* on both the transcript level, using all 46 datasets, and the dataset level, using the 46 RNA-seq datasets individually. *Finc* can achieve up to 85% prediction accuracy using only RNA-seq data. This software improves prediction of true lncRNAs from samples, such as clinical samples, where ChIP-seq data are unavailable. *Finc* can save time and resources by making the generation of H3K4me3 unnecessary and allowing the identification of lncRNAs from publicly available RNA-seq data.

The predictive power of *Finc* relies on the quality of the benchmark datasets and the identification of true and false lncRNA within them. Therefore, the prediction power of *Finc* may be limited by factors that affect the reliability of the benchmark datasets. These factors include RNA-seq and ChIP-seq data quality, accuracy of transcript assembly and ChIP-seq peak calling, and sequencing depth. To minimize the effects of these factors, we examined the data quality and sequencing depths of both RNA-seq and ChIP-seq data in the benchmark datasets and only keep the data with good data quality and adequate sequencing depths in this study. To improve the accuracy of transcript assembly, we integrated multiple transcript-assembly methods. It has been shown that the accuracy of ChIP-seq peak calling can be improved by using different approaches for long (≥ 70 bp) and short (< 70 bp) reads studies [44,45], and we, therefore, used this approach.

Compared to the true lncRNAs identified by H3K4me3 profiles, true lncRNAs predicted by *Finc* (with random forest) are more likely to be divergent transcripts of nearby protein-coding genes and to have promoter signatures (Supplementary Figure S4B). This result suggests that *Finc* might overestimate divergent putative lncRNAs and putative lncRNAs with promoter signatures as true lncRNAs, but underestimate the true lncRNAs that are in intergenic regions. One possible explanation for this is that divergent transcription and promoter signature are the two most important features for the three best models (random forest, decision tree, and KNN).

Because different methods of generating and sequencing RNA-seq libraries result in slightly different collections of transcripts, feature importance may determine model performance for different types of RNA-seq data. We found that the random forest, decision tree, and KNN models outperform than other models for stranded RNA-seq data generated from polyA-selected RNA (Figure 3 and Supplementary Figure S3). In contrast, the linear SVM and naïve Bayes models, which performed poorly for stranded polyA-selected RNA-seq data, achieve better prediction accuracy and precision than other models for rRNA-depleted and unstranded RNA-seq data. This may be because the random forest, decision tree, and KNN models rely on three major features: divergent transcript, promoter signature, and transcript length (Figure 6), but the transcript length feature may not be applicable to nascent lncRNAs, which are included in rRNA-depleted RNA-seq data. In contrast, for unstranded RNA-seq data, the promoter signature feature may not be applicable because without accurate strand information, the promoters of these putative lncRNA cannot be inferred. Unlike the random forest, decision tree, and KNN models, the linear SVM and naïve Bayes models mainly rely on the divergent genomic structure feature (Figure 6), which is applicable for both nascent and unstranded lncRNAs.

Finc includes seven embedded ML models (random forest, decision tree, logistic regression, naïve Bayes, KNN, linear-SVM, and RBF-SVM) and an ensemble approach, which makes *Finc* suitable for identifying lncRNAs from different types of RNA-seq data. Model performance depends somewhat on the type of RNA-seq data that is used as input. The best three models (random forest, decision tree, and KNN) for the stranded polyA-selected RNA-seq data achieve very high consistency (93–96%) in true lncRNA prediction

(Supplementary Figure S3D). Therefore, we recommend that users select any of these three models when using *Flnc* to identify lncRNAs from stranded polyA-selected RNA-seq data. Precision can be further improved by the ensemble approach (Supplementary Figure S3E). Therefore, we recommend using the *ensemble* setting of the *Flnc* software, which provides users with the common set of true lncRNAs predicted by all models.

Although *Flnc* was trained on the benchmark datasets of lncRNAs identified from stranded polyA-selected RNA-seq data and matched H3K4me3 ChIP-seq data, the *Flnc* pipeline can be applied to identify lncRNAs from other types of RNA-seq data, such as stranded rRNA-depleted RNA-seq data and unstranded polyA-selected data. Based on the performance of the seven models on stranded rRNA-depleted RNA-seq data and unstranded polyA-selected data, we recommend using the linear KNN or naïve Bayes models for these two types of RNA-seq data.

The *Flnc* software can identify novel lncRNAs directly from RNA-seq data, as well as evaluate whether a transcript is a lncRNA. *Flnc* can take three types of input files, including raw RNA-seq data in the FASTQ format, transcript coordinates in the BED format, and transcript sequences in the FASTA format. For maximum portability and usability, *Flnc* is implemented in Python 2 and the entire pipeline is wrapped within a Singularity container. The current version of *Flnc* is applicable to data of human samples, but the concept level of *Flnc* is applicable to other organisms. As a follow-up study, we plan to extend *Flnc*'s method to be applicable to other organisms.

4. Material and Methods

4.1. Collection of Sequencing Datasets to Generate Benchmark lncRNAs

Because most mature lncRNAs have polyA tails [2] and clear transcriptional start sites, generating a benchmark dataset of true and false lncRNAs requires stranded polyA-selected RNA-seq and sample-matched H3K4me3 ChIP-seq. We identified 388 data series with the keywords "RNA seq," "H3K4me3," and "*Homo sapiens*" from the NCBI GEO database through February of 2021. From the 388 GEO studies, we selected 61 datasets with available sample-matched, stranded polyA-selected RNA-seq and H3K4me3 ChIP-seq data.

For the 61 datasets, we examined the quality and sequencing depths of RNA-seq and ChIP-seq data and removed 15 datasets, which had poor RNA-seq or ChIP-seq data quality or extremely low sequencing depths. We considered the quality of RNA-seq data poor if more than half of the reads had quality control scores less than 35. For ChIP-seq data, we considered their quality poor if the number of called peaks in the dataset was an outlier among peak numbers found in the 61 datasets (extremely small < 2000 or extremely large > 200,000). This process resulted in selection of 46 datasets. The RNA-seq and ChIP-seq data within each of the selected 46 high-quality datasets were generated from the same type of sample. The datasets are heterogeneous, including both single-end (50%) and paired-end (50%) RNA-seq data and each dataset has one to four RNA-seq replicates. We indexed these datasets 1 to 46 in chronological order by submission date (Supplementary Tables S1–S3) and used them to identify true and false lncRNAs in each dataset as benchmark data.

After identifying true and false lncRNAs in each dataset (see below), we split the benchmark dataset into training and testing sets based on the submission dates of the RNA-seq data into the NCBI GEO database [46]. The training set included the 41 datasets that were submitted to the GEO database before 2019; we used the five datasets generated in 2019 and 2020 as the testing set.

4.2. Identification of Putative lncRNAs

First, we mapped the RNA-seq data to the human reference genome (hg38) using HISAT2 v 2.0.5 [47] and assembled transcripts using both StringTie v1.3.4 [22] and Strawberry v1.1.2 [23]. For each sample, we then merged assembled transcripts of replicates with the StringTie merge function. Next, we examined the coding potential of the assembled transcripts using CPAT v1.2.4 [25], LGC v1.0 [26], PLEK v1.2. [27], and CPPred [28]. We

excluded the transcripts with coding potential defined in any of the above tools. Next, we excluded the resulting transcripts, as well as any transcripts that overlapped protein-coding genes or pseudogenes on the same strand. We also excluded transcripts overlapping other annotated noncoding RNAs, including snoRNA, rRNA, tRNA, and microRNAs, on the same strand. From the remaining transcripts, we selected expressed transcripts that were over 200 nucleotides long as putative lncRNAs (see Supplemental Methods for details).

4.3. Identification of H3K4me3 Peaks Using H3K4me3 ChIP-Seq Data

We aligned H3K4me3 ChIP-seq reads to the human reference genome (hg38/GRCh38) using the BWA v0.7.5a toolkit [44]. For reads with mean read length ≥ 70 bp we used BWA-MEM and for short reads we used BWA-aln. Next, we used the MACS2 v2.2.7.1 [48] peakcall function to identify H3K4me3, as described previously [18], with the following settings: `-q 0.01 -broad -broad-cutoff=0.01 -nomodel -extsize 300`. For H3K4me3 ChIP-seq data without matched control data, MACS2 called peaks based on the H3K4me3 input ChIP-seq data. For H3K4me3 ChIP-seq data with matched control data, MACS2 called peaks by comparing the bam files to the matched background control. We used the H3K4me3 broad peaks called by MACS2 as transcription initiation markers [49,50].

4.4. Identification of True and False lncRNAs Based on H3K4me3 ChIP-Seq Data

We used our previously established approach [18] to identify true and false lncRNAs among putative lncRNAs. Because H3K4me3 chromatin modification is well known as transcriptional initiation marks of active genes [34], including lncRNA genes, we examined the distance between the 5' ends of putative lncRNAs and the matched H3K4me3 peaks. If the H3K4me3 peak was within 1 kb of the 5' end of a putative lncRNA, we considered it a true lncRNA; otherwise, we considered it false.

4.5. Normalization of Transcript Lengths

We calculated the length of putative lncRNAs by summing the length of all the transcript's exons. Next, we log-transformed the lengths. We calculated the upper and lower limits as mean plus and minus $3 \times$ standard deviation, respectively. Then, we set the outlier data points of the log-transformed lengths as the value of upper or lower limits. Finally, we scaled the log-transformed values of transcript lengths for each putative lncRNA into the range between 0 and 1 using the min-max normalization technique.

4.6. Identification of Promoter Regions

We used TSSG software [41] to identify the potential promoters in the genomic regions ± 1 kb of the 5' end of putative lncRNAs. TSSG detects promoters by scanning for transcription factor binding sites and is considered to be one of the most accurate mammalian promoter prediction programs with the fewest false positive predictions [51].

4.7. Classification of Putative lncRNAs by Genomic Location

We used the genomic locations of protein-coding genes annotated by the GENCODE Project release 29 [52], and compared them to the genomic locations of putative lncRNAs using the BEDTools suite [53] and GffCompare tool [54]. We classified putative lncRNAs—true and false—into three categories by genomic locations [18,35]: divergent, antisense, and intergenic. A putative lncRNA was classified as divergent if the 5' end of its locus was within ± 2 kb of the TSS of a protein-coding gene on the opposite strand [18,35]. The remaining putative lncRNAs that were antisense to protein-coding genes and overlapped the gene by at least one base pair were classified as antisense. All remaining putative lncRNAs were classified as intergenic.

4.8. Calculation of the Exon Numbers of Putative lncRNAs

The number of exons of putative lncRNAs were counted based on the assembled transcripts in GTF format, which include the information of exon numbers and coordinates of each exon.

4.9. Measurement Feature Importance

To determine the feature importance for all machine learning models, we used the permutation approach [55]. We scaled the importance scores into the range between 0 and 1 by dividing each importance score by the importance score of the most significant feature.

4.10. Collection of Other RNA-Seq Datasets and Matched H3K4me3 ChIP-Seq Data

To evaluate *Flnc*'s performance on other types of RNA-seq data, we collected four additional datasets, each corresponding to one sample from the GEO database (Supplementary Table S4). These four datasets all contained RNA-seq data and matched H3K4me3 ChIP-seq. Two datasets included stranded rRNA-depleted RNA-seq (accession numbers: GSE179184 and GSE189861) and two included unstranded polyA-selected RNA-seq data (accession numbers: GSE72131 and GSE58740).

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/ncrna8050070/s1>. Figure S1: True and false lncRNAs in each of the 46 benchmark datasets. Figure S2: Genomic features of true and false lncRNAs in each of the 46 benchmark datasets. Figure S3: Performance of the seven final ML models used in *Flnc*. Figure S4: The genomic features of true lncRNAs predicted by *Flnc* and true lncRNAs determined by H3K4me3 profiles in the test set. Figure S5: *Flnc* achieves similar performance at the lncRNA gene locus level as at the transcript level. Table S1: Information of the collected 46 datasets, each of which includes both of stranded polyA-selected RNA-seq data and matched H3K4me3 ChIP-seq data from the same human sample. Table S2: The mapping information for the RNA-seq data and ChIP-seq data among the 46 datasets used in this study. Table S3: Performance of the conventional approach and the seven ML models of *Flnc* in the training set and test datasets. Table S4: Information of four datasets with other types of RNA-seq data.

Author Contributions: Formal analysis, Z.L.; Methodology, Z.L., P.Z. and C.Z.; Software, Z.L. and E.K.; Supervision, C.Z.; Writing—original draft, Z.L. and C.Z.; Writing—review and editing, Z.L., K.A.F., Z.W. and C.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Institutes of Health [UL1TR001453] through The University of Massachusetts Center for Clinical and Translational Sciences (C.Z.) and supported by the Defense Advanced Research Projects Agency, contracted via the Department of Navy, Office of Naval Research under the Federal Award Number [N660011924036] as part of the Preemptive Expression of Protective Alleles and Response Elements (PREPARE) program (to C.Z. and K.F.). This work was also supported by NIH R03DE032455-01 (to C.Z.).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The datasets analyzed in this study are available in GEO with accession numbers listed in Supplementary Table S1. These datasets were downloaded from the NCBI GEO database: <https://www.ncbi.nlm.nih.gov/geo/>, accessed on 10 September 2022. The constructed benchmark dataset of true and false lncRNAs are available at https://zhoulab.umassmed.edu/Flnc_data/, accessed on 10 September 2022. The analyses were performed with *Flnc* version 1. The *Flnc* software is freely available on GitHub: <https://github.com/CZhouLab/Flnc>, accessed on 10 September 2022 along with a tutorial.

Acknowledgments: We thank Qian Qi for the helpful discussions about the ChIP-seq analysis. We thank Alan C. Mullen for review of the manuscript and insightful suggestions. We thank Feifan Liu for helpful discussion about the ML models. We thank Edith Pfister for revising the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dunham, I.; Kundaje, A.; Aldred, S.F.; Collins, P.J.; Davis, C.A.; Doyle, F.; Epstein, C.B.; Frietze, S.; Harrow, J.; Kaul, R.; et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**, *489*, 57–74.
2. Statello, L.; Guo, C.J.; Chen, L.L.; Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* **2021**, *22*, 96–118. [[CrossRef](#)]
3. Daneshvar, K.; Pondick, J.V.; Kim, B.-M.; Zhou, C.; York, S.R.; Macklin, J.A.; Abualteen, A.; Tan, B.; Sigova, A.A.; Marcho, C.; et al. DIGIT Is a Conserved Long Noncoding RNA that Regulates GSC Expression to Control Definitive Endoderm Differentiation of Embryonic Stem Cells. *Cell Rep.* **2016**, *17*, 353–365. [[CrossRef](#)] [[PubMed](#)]
4. Daneshvar, K.; Ardehali, M.B.; Klein, I.A.; Hsieh, F.K.; Kratkiewicz, A.J.; Mahpour, A.; Cancelliere, S.O.L.; Zhou, C.; Cook, B.M.; Li, W.; et al. lncRNA DIGIT and BRD3 protein form phase-separated condensates to regulate endoderm differentiation. *Nat. Cell Biol.* **2020**, *22*, 1211–1222. [[CrossRef](#)]
5. Cabili, M.N.; Trapnell, C.; Goff, L.; Koziol, M.; Tazon-Vega, B.; Regev, A.; Rinn, J.L. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **2011**, *25*, 1915–1927. [[CrossRef](#)]
6. Li, G.; Kryczek, I.; Nam, J.; Li, X.; Li, S.; Li, J.; Grove, S.; Vatan, L.; Zhou, J.; Du, W.; et al. LIMIT is an immunogenic lncRNA in cancer immunity and immunotherapy. *Nat. Cell Biol.* **2021**, *23*, 526–537. [[CrossRef](#)]
7. Guo, C.J.; Ma, X.K.; Xing, Y.H.; Zheng, C.C.; Xu, Y.F.; Shan, L.; Zhang, J.; Wang, S.; Wang, Y.; Carmichael, G.G.; et al. Distinct Processing of lncRNAs Contributes to Non-conserved Functions in Stem Cells. *Cell* **2020**, *181*, 621–636.e22. [[CrossRef](#)] [[PubMed](#)]
8. Quinn, J.J.; Chang, H.Y. Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet.* **2016**, *17*, 47–62. [[CrossRef](#)] [[PubMed](#)]
9. de Goede, O.M.; Nachun, D.C.; Ferraro, N.M.; Gloudemans, M.J.; Rao, A.S.; Smail, C.; Eulalio, T.Y.; Aguet, F.; Ng, B.; Xu, J.; et al. Population-scale tissue transcriptomics maps long non-coding RNAs to complex disease. *Cell* **2021**, *184*, 2633–2648.e19. [[CrossRef](#)] [[PubMed](#)]
10. Huarte, M. The emerging role of lncRNAs in cancer. *Nat. Med.* **2015**, *21*, 1253–1261. [[CrossRef](#)]
11. Fernandes, J.C.R.; Acuña, S.M.; Aoki, J.I.; Floeter-Winter, L.M.; Muxel, S.M. Long non-coding RNAs in the regulation of gene expression: Physiology and disease. *Non-Coding RNA* **2019**, *5*, 17. [[CrossRef](#)]
12. Zhang, X.; Wang, W.; Zhu, W.; Dong, J.; Cheng, Y.; Yin, Z.; Shen, F. Mechanisms and Functions of Long Non-Coding RNAs at Multiple Regulatory Levels. *Int. J. Mol. Sci.* **2019**, *20*, 5573. [[CrossRef](#)]
13. Hezroni, H.; Koppstein, D.; Schwartz, M.G.; Avrutin, A.; Bartel, D.P.; Ulitsky, I. Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. *Cell Rep.* **2015**, *11*, 1110–1122. [[CrossRef](#)]
14. Du, Z.; Fei, T.; Verhaak, R.G.W.; Su, Z.; Zhang, Y.; Brown, M.; Chen, Y.; Liu, X.S. Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat. Struct. Mol. Biol.* **2013**, *20*, 908–913. [[CrossRef](#)]
15. Frankish, A.; Diekhans, M.; Jungreis, I.; Lagarde, J.; Loveland, J.E.; Mudge, J.M.; Sisu, C.; Wright, J.C.; Armstrong, J.; Barnes, I.; et al. GENCODE. *Nucleic Acids Res.* **2021**, *49*, D916–D923. [[CrossRef](#)]
16. Fang, S.; Zhang, L.; Guo, J.; Niu, Y.; Wu, Y.; Li, H.; Zhao, L.; Li, X.; Teng, X.; Sun, X.; et al. NONCODEV5: A comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res.* **2018**, *46*, D308–D314. [[CrossRef](#)]
17. Volders, P.-J.; Anckaert, J.; Verheggen, K.; Nuytens, J.; Martens, L.; Mestdagh, P.; Vandesompele, J. LNCipedia 5: Towards a reference set of human long non-coding RNAs. *Nucleic Acids Res.* **2019**, *47*, D135–D139. [[CrossRef](#)] [[PubMed](#)]
18. Zhou, C.; York, S.R.; Chen, J.Y.; Pondick, J.V.; Motola, D.L.; Chung, R.T.; Mullen, A.C. Long noncoding RNAs expressed in human hepatic stellate cells form networks with extracellular matrix proteins. *Genome Med.* **2016**, *8*, 31. [[CrossRef](#)] [[PubMed](#)]
19. Guttman, M.; Garber, M.; Levin, J.Z.; Donaghey, J.; Robinson, J.; Adiconis, X.; Fan, L.; Koziol, M.J.; Gnirke, A.; Nusbaum, C.; et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **2010**, *28*, 503–510. [[CrossRef](#)] [[PubMed](#)]
20. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; et al. Trinity: Reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* **2013**, *29*, 644–652. [[CrossRef](#)]
21. Trapnell, C.; Williams, B.P.; Pertea, G.; Mortazavi, A.; Kwan, G.; van Baren, M.J.; Salzberg, S.L.; Wold, B.J.; Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **2010**, *28*, 511–515. [[CrossRef](#)]
22. Pertea, M.; Pertea, G.M.; Antonescu, C.M.; Chang, T.C.; Mendell, J.T.; Salzberg, S.L. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **2015**, *33*, 290–295. [[CrossRef](#)] [[PubMed](#)]
23. Liu, R.; Dickerson, J. Strawberry: Fast and accurate genome-guided transcript reconstruction and quantification from RNA-Seq. *PLoS Comput. Biol.* **2017**, *13*, e1005851. [[CrossRef](#)]
24. Liu, J.; Yu, T.; Jiang, T.; Li, G. TransComb: Genome-guided transcriptome assembly via combing junctions in splicing graphs. *Genome Biol.* **2016**, *17*, 213. [[CrossRef](#)]
25. Wang, L.; Park, H.J.; Dasari, S.; Wang, S.; Kocher, J.-P.; Li, W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **2013**, *41*, e74. [[CrossRef](#)] [[PubMed](#)]
26. Wang, G.; Yin, H.; Li, B.; Yu, C.; Wang, F.; Xu, X.; Cao, J.; Bao, Y.; Wang, L.; Abbasi, A.A.; et al. Characterization and identification of long non-coding RNAs based on feature relationship. *Bioinformatics* **2019**, *35*, 2949–2956. [[PubMed](#)]

27. Li, A.; Zhang, J.; Zhou, Z. PLEK: A tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinform.* **2014**, *15*, 311. [[CrossRef](#)]
28. Tong, X.; Liu, S. CPPred: Coding potential prediction based on the global description of RNA sequence. *Nucleic Acids Res.* **2019**, *47*, e43. [[CrossRef](#)] [[PubMed](#)]
29. Danis, J.; Széll, M. VELUCT, a long non-coding RNA with an important cellular function despite low abundance. *J. Thorac. Dis.* **2017**, *9*, 3638–3640. [[PubMed](#)]
30. Ulitsky, I.; Shkumatava, A.; Jan, C.H.; Sive, H.; Bartel, D.P. Conserved Function of lincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution. *Cell* **2011**, *147*, 1537–1550. [[PubMed](#)]
31. Washietl, S.; Findeiß, S.; Müller, S.A.; Kalkhof, S.; von Bergen, M.; Hofacker, I.L.; Stadler, P.F.; Goldman, N. RNAcode: Robust discrimination of coding and noncoding regions in comparative sequence data. *RNA* **2011**, *17*, 578–594. [[CrossRef](#)]
32. Ruan, X.; Li, P.; Chen, Y.; Shi, Y.; Pirooznia, M.; Seifuddin, F.; Suemizu, H.; Ohnishi, Y.; Yoneda, N.; Nishiwaki, M.; et al. In vivo functional analysis of non-conserved human lincRNAs associated with cardiometabolic traits. *Nat. Commun.* **2020**, *11*, 45. [[CrossRef](#)] [[PubMed](#)]
33. Qiao, F.-H.; Tu, M.; Liu, H.-Y. Role of MALAT1 in gynecological cancers: Pathologic and therapeutic aspects (Review). *Oncol. Lett.* **2021**, *21*, 333. [[CrossRef](#)]
34. Guenther, M.G.; Levine, S.S.; Boyer, L.A.; Jaenisch, R.; Young, R.A. A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells. *Cell* **2007**, *130*, 77–88. [[CrossRef](#)]
35. Sigova, A.A.; Mullen, A.C.; Molinie, B.; Gupta, S.; Orlando, D.A.; Guenther, M.G.; Almada, A.E.; Lin, C.; Sharp, P.A.; Giallourakis, C.C.; et al. Divergent transcription of long noncoding RNA / mRNA gene pairs in embryonic stem cells. *Proc. Natl. Acad. Sci. USA.* **2013**, *110*, 2876–2881. [[CrossRef](#)]
36. Luo, W.; Bentley, D. A Ribonucleolytic Rat Torpedoes RNA Polymerase II. *Cell* **2004**, *119*, 911–914. [[CrossRef](#)]
37. Eaton, J.D.; West, S. An end in sight? Xrn2 and transcriptional termination by RNA polymerase II. *Transcription* **2018**, *9*, 321–326. [[CrossRef](#)]
38. Babarinde, I.A.; Hutchins, A.P. The effects of sequencing depth on the assembly of coding and noncoding transcripts in the human genome. *BMC Genom.* **2022**, *23*, 4487. [[CrossRef](#)]
39. Keel, B.N.; Oliver, W.T.; Keele, J.W.; Lindholm-Perry, A.K. Evaluation of transcript assembly in multiple porcine tissues suggests optimal sequencing depth for RNA-Seq using total RNA library. *Anim. Gene* **2020**, *17–18*, 200105. [[CrossRef](#)]
40. Chang, Z.; Wang, Z.; Li, G. The impacts of read length and transcriptome complexity for de novo assembly: A simulation study. *PLoS ONE* **2014**, *9*, e94825.
41. Solovyev, V.; Salamov, A. The Gene-Finder computer tools for analysis of human and model organisms genome sequences. *Proc. Int. Conf. Intell. Syst. Mol. Boil.* **1997**, *5*, 294–302.
42. Chen, J.; Shishkin, A.A.; Zhu, X.; Kadri, S.; Maza, I.; Guttman, M.; Hanna, J.H.; Regev, A.; Garber, M. Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. *Genome Biol.* **2016**, *17*, 19. [[CrossRef](#)]
43. Yang, Y.; Zhou, R.; Zhu, S.; Li, X.; Li, H.; Yu, H.; Li, K. Systematic Identification and Molecular Characteristics of Long Noncoding RNAs in Pig Tissues. *BioMed Res. Int.* **2017**, *2017*, 6152582. [[CrossRef](#)]
44. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows—Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [[CrossRef](#)]
45. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **2013**, arXiv:1303.3997.
46. Edgar, R.; Domrachev, M.; Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **2002**, *30*, 207–210. [[CrossRef](#)]
47. Kim, D.; Langmead, B.; Salzberg, S.L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **2015**, *12*, 357–360. [[CrossRef](#)]
48. Zhang, Y.; Liu, T.; Meyer, C.A.; Eeckhoute, J.; Johnson, D.S.; Bernstein, B.E.; Nusbaum, C.; Myers, R.M.; Brown, M.; Li, W.; et al. Model-based analysis of CHIP-Seq (MACS). *Genome Biol.* **2018**, *9*, R137. [[CrossRef](#)]
49. Chen, K.; Chen, Z.; Wu, D.; Zhang, L.; Lin, X.; Su, J.; Rodriguez, B.; Xi, Y.; Xia, Z.; Chen, X.; et al. Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes. *Nat. Genet.* **2015**, *47*, 1149–1157. [[CrossRef](#)]
50. Dincer, A.; Gavin, D.P.; Xu, K.; Zhang, B.; Dudley, J.T.; Schadt, E.E.; Akbarian, S. Deciphering H3K4me3 broad domains associated with gene-regulatory networks and conserved epigenomic landscapes in the human brain. *Transl. Psychiatry* **2015**, *5*, e679. [[CrossRef](#)]
51. Liu, R.; States, D.J. Consensus Promoter Identification in the Human Genome Utilizing Expressed Gene Markers and Gene Modeling. *Genome Res.* **2002**, *12*, 462–469. [[CrossRef](#)]
52. Harrow, J.; Frankish, A.; Gonzalez, J.M.; Tapanari, E.; Diekhans, M.; Kokocinski, F.; Aken, B.L.; Barrell, D.; Zadissa, A.; Searle, S.; et al. GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* **2012**, *22*, 1760–1774. [[CrossRef](#)]
53. Quinlan, A.R.; Hall, I.M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26*, 841–842. [[CrossRef](#)]
54. Pertea, G.; Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Research* **2020**, *9*, 1–19. [[CrossRef](#)]
55. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: A corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340–1347. [[CrossRef](#)]