Briefings in Bioinformatics, 22(1), 2021, 380–392

doi: 10.1093/bib/bbz178
Advance Access Publication Date: 31 January 2020
Articles

Computational annotation of miRNA transcription start sites

Saidi Wang†, Amlan Talukder†, Mingyu Cha, Xiaoman Li*, Haiyan Hu*

*Corresponding authors: Xiaoman Li, Burnett School of Biomedical Science, University of Central Florida, Orlando, FL-32816, United States. E-mail: xiaoman@mail.ucf.edu; Haiyan Hu, Computer Science, University of Central Florida, Orlando, FL-32816, United States. Tel: 1-407-823-4811; Fax: 1-407-823-5835; E-mail: haihu@cs.ucf.edu.
†These authors contributed equally to this work.

Abstract

Motivation: MicroRNAs (miRNAs) are small noncoding RNAs that play important roles in gene regulation and phenotype development. The identification of miRNA transcription start sites (TSSs) is critical to understand the functional roles of miRNA genes and their transcriptional regulation. Unlike protein-coding genes, miRNA TSSs are not directly detectable from conventional RNA-Seq experiments due to miRNA-specific process of biogenesis. In the past decade, large-scale genome-wide TSS-Seq and transcription activation marker profiling data have become available, based on which, many computational methods have been developed. These methods have greatly advanced genome-wide miRNA TSS annotation.
Results: In this study, we summarized recent computational methods and their results on miRNA TSS annotation. We collected and performed a comparative analysis of miRNA TSS annotations from 14 representative studies. We further compiled a robust set of miRNA TSSs (RSmirT) that are supported by multiple studies. Integrative genomic and epigenomic data analysis on RSmirT revealed the genomic and epigenomic features of miRNA TSSs as well as their relations to protein-coding and long non-coding genes.
Contact: xiaoman@mail.ucf.edu, haihu@cs.ucf.edu
Supplementary information: Supplementary data are available at RSmiRT website http://hulab.ucf.edu/research/projects/RSMIRT/index.html.

Key words: miRNA; intergenic miRNA; intragenic miRNA; miRNA TSS.

Introduction

MicroRNAs (miRNAs) are small (~22 nucleotides), single-stranded endogenous non-coding RNAs derived from hairpin precursors [1–3]. Binding with their targeted mRNAs in the 3′UTR regions, miRNAs often induce gene silencing and therefore serve as the post-transcriptional gene regulators [4–6]. miRNAs can be simply categorized into two categories: intergenic miRNAs

Saidi Wang is a graduate student from the department of computer science, University of Central Florida. He mainly works on gene transcriptional regulation.
Amlan Talukder is a graduate student from the department of computer science, University of Central Florida. He mainly works on miRNAs and epigenomics.
Mingyu Cha is a graduate student from the department of computer science, University of Central Florida. He mainly works on TSS-seq and CAGE data analysis.
Xiaoman Li is an associate professor from Burnett School of Biomedical Science, University of Central Florida. He works on chromatin interactions and metagenomics.
Haiyan Hu is an associate professor from the department of computer science, University of Central Florida. She works on miRNAs, epigenomics and gene transcriptional regulation.
Submitted: 04 November 2019; **Received (in revised form):** 13 December 2019

when they are located between genes and intragenic miRNAs when they are found overlapping exons and introns on the same strand of annotated genes.

Since the first discovery of miRNAs in Caenorhabditis elegans in 1993 [7, 8], a large number of miRNAs have been discovered in metazoan, plants and viruses [9–11]. Today, miRNAs are known to express ubiquitously in almost all cell types, evolutionarily conserved in most metazoan and plant species and potentially regulate more than 30% of mammalian gene products [12–14]. miRNAs are also implicated in critical processes such as developmental timing control, hematopoietic cell differentiation, apoptosis, cell proliferation and organ development [1, 15]. Although misexpression of miRNAs has been linked to cancer and many other diseases [16–20], little is known about the mechanism of how the expression of miRNA genes is regulated under different phenotypic conditions, majorly due to current limited knowledge of miRNA transcription initiation in various types of cells and tissues [21].

miRNA biogenesis often begins with RNA polymerase II (Pol II) transcription of primary miRNA (pri-miRNA). Unlike protein-coding genes, miRNA transcription start sites (TSSs) can be more than several kilobases (kb) upstream of the location of mature miRNAs [2]. Following the transcription, the long pri-miRNAs are then precisely cropped around a hairpin-shaped region by a microprocessor consisting of the nuclear RNase III Drosha and its cofactor protein DGCR8 to generate precursor miRNAs (pre-miRNA) [22]. Pre-miRNAs are then exported to the cytoplasm where they are cleaved by Dicer, another RNase III protein. The cleavage of pre-miRNA results in a small RNA duplex that is subsequently loaded onto an Argonaute (AGO) protein to form a complex called RNA-induced silencing complex (RISC), where the mature miRNA is then generated after the passenger strand of the miRNA duplex is removed [23]. During the miRNA biogenesis, the pri-miRNA processing takes place so fast that the conventional mRNA-Seq experiments cannot capture much of the pri-miRNAs. As a consequence, conventional mRNA-Seq experiments cannot be used to identify TSSs of miRNAs directly. Note that, this is very different from protein-coding genes whose TSSs can be detected from mRNA-Seq experiments given a reasonable sequencing depth.

With the rapid accumulation of high-throughput next-generation sequencing data such as chromatin immunoprecipitation followed by massive parallel sequencing (ChIP-Seq) and TSS-Seq data in the last decade, dozens of computational methods have been developed to predict miRNA TSSs at the genome scale. In this study, we first briefly summarized these computational methods and their results for miRNA TSS annotation. We then performed a comparative analysis of the genome-wide miRNA TSS annotation compiled from 14 latest publications. We illustrated the differences and similarities of miRNA TSS annotations across surveyed studies and defined a robust set of miRNA TSS (RSmiRT)annotation supported by multiple studies. Investigating the RSmiRT, we learned genomic and epigenomic features of miRNA TSSs in comparison to protein-coding and long non-coding genes. In the last part, we discussed the limitations of the current computational annotation of miRNA TSSs.

Brief overview of recent high-throughput computational methods for miRNA TSS prediction

Early methods for genome-wide TSS prediction focus on the use of expressed sequence tags (ESTs) and sequence features such as over-represented k-mers, transcription factor binding site (TFBS) distribution, sequence conservation score and CpG content to scan upstream regions of mature miRNAs to predict their corresponding TSSs [24–30]. These studies have provided initial insight into miRNA TSS properties. Since these sequence features are often summarized from a limited number of annotated coding RNA promoters [24, 25], the prediction results often not only are non-condition-specific, but also contain high-rate of false-positives and are limited to a handful of miRNAs [31].

In the past decade, ChIP-Seq experiments discovered that a number of chromatin modifications can be markers for gene transcription activation [32–34]. For example, trimethylation of Lys 4 of histone 3 (H3K4me3) and acetylation of Lys 9/14 of histone 3 (H3K9/14Ac), Pol II and DNase I hypersensitive sites sequencing (DNase-Seq) measurements [35]. Dozens of computational studies have been developed to predict miRNA TSSs using these chromatin markers [31, 36–40]. For example, in the work of Marson et al., H3K4me3-enriched regions pooled from multiple cell lines [32, 41, 42] were first utilized to systematically identify putative miRNA TSSs in human and mouse [14]. The putative TSSs corresponding to each miRNA were then scored based on their genomic distances to the mature miRNA sequence, EST evidence and cross-species conservation levels. Similarly, in the work of Ozsolak et al., active genes were observed to exhibit nucleosome depletion in the 100–130 base pairs (bp) window surrounding TSSs [35]. Chromatin markers of H3K4me3, H3K9/14Ac, Pol II and/or Pol III were subsequently integrated with sequence features such as conservation, CpG island and transcription factor (TF) binding motif occurrence in nucleosome-depleted regions. This led to the identification of 175 human miRNA promoters in two melanoma cell lines and one breast cancer cell line. In addition, based on Pol II ChIP-ChIP experiments in A549 lung epithelial cells, Corcoran et al. scanned 50 kb upstream of 531 known miRNA genes and identified 1 kb windows near the 5′ end of these known miRNAs that exhibit Pol II signals with statistical significance [37]. The Pol II scanning resulted in TSS predictions for 34 intergenic and 43 intragenic miRNAs. They also used 3015 verified core promoters of protein-coding genes as training data to create a support vector machine (SVM) model for TSS prediction based on sequence features such as TF binding profiles, n-mer frequency and GC content. The trained SVM model was able to identify TSSs for 29 out of the 34 intergenic miRNAs. Barski et al. also used chromatin markers in T cells to define miRNA TSSs [36]. After first identifying statistically significant peaks from H3K4me3, H2A.Z and Pol II data using Gaussian kernel density estimation profiles, they defined miRNA TSSs as regions where at least two out of three marker peaks co-localized. In total, they identified TSSs for 234 human miRNAs including 129 intragenic and 105 intergenic miRNAs.

Recent genome-wide TSS-Seq experiments such as cap analysis gene expression (CAGE), 5′-SAGE (serial analysis of gene expression), PET (paired-end tags) and GRO-cap [43–49] further facilitated high-throughput miRNA TSS annotation. TSS-Seq experiments measuring global transcriptional events have been performed on thousands of cell lines and tissues [50–54]. For example, FANTOM (Functional Annotation of the Mammalian Genome) consortium has published CAGE data for more than 1000 human and mouse primary cells, cell lines and tissue samples [55]. Marsico et al. reported a computational tool PROmiRNA to predict miRNA TSSs, especially those of intragenic miRNAs by combining sequence features with all the available CAGE measurements from FANTOM (Version 4) [39]. PROmiRNA was motivated by the hypothesis that intronic miRNAs might have different promoters from their host genes.

Table 1. Statistics of predicted TSSs from 14 resources

Index	Reference	# miRNA	# TSS	# Intergenic miRNA	# Intragenic miRNA	TSS size [min, max]	Cell-specific
P0	Corcoran et al., 2009	82	50	61	21	[26, 15309]	N
P1	Landgraf et al., 2007	145	95	88	57	[71, 64182693]	N
P2	Marson et al., 2008	431	332	231	200	[193, 46725386]	N
P3	Fujita et al., 2007	73	55	36	37	[2, 2]	N
P4	Ozsolak et al., 2008	165	172	78	87	[2, 2]	N
P5	Saini et al., 2007	25	14	22	3	[574, 6468]	N
P6	MiRStart (Chien et al., 2011)	295	203	215	80	[1, 1]	N
P7	PROmiRNA (Marsico et al., 2013)	1333	7133	529	804	[17, 749]	N
P8	MicroTSS (Georgakilas et al., 2014)	132	106	116	16	[2, 2]	Y
P9	Hua et al., 2016	1263	6012	468	795	[1, 1]	Y
P10	MirSTP (Liu et al., 2017)	475	3669	428	47	[1, 1]	Y
P11	FANTOM (de Rie et al., 2017)	1157	1029	203	954	[1, 1]	N
P12	Bouvy-Liivrand et al., 2017	1014	1033	310	704	[1, 1]	N
P13	MiRGen (Georgakilas et al., 2015)	224	426	194	30	[1, 2]	N

The EM algorithm underlying PROmiRNA aimed to distinguish a true promoter region from a CAGE tag-enriched region that might be the background noise. Together with sequence features such as CpG content, conservation and TATA box affinity score, PROmiRNA made predictions of 7244 TSSs corresponding to 1228 miRNAs. Several other computational studies attempted to integrate chromatin markers with CAGE data as well as available TSS-Seq libraries derived from high-throughput ChIP-Seq experiments. For example, Chien et al. combined the H3K4me3 signal with TSS-Seq distributions to identify miRNA TSSs [38]. They created a software miRStart that used SVM to model CAGE tags from FANTOM (V4), TSS-Seq from DBTSS (V7) and H3K4me3 ChIP-Seq data from CD4+ T cells [41]. Utilizing annotated TSSs of 7286 protein-coding genes from DBTSS as the training data, they identified 847 human miRNA TSSs. Based solely on CAGE experimental data from 396 human samples, recently FANTOM (version 5) also inferred 1357 human miRNA promoters as follows. For a given miRNA, they defined pri-miRNA candidates as transcripts whose TSSs located upstream of the corresponding pre-miRNAs and 3′ ends downstream of the 5′ end of the corresponding pre-miRNAs. They then defined promoter candidates as CAGE peaks that were located within the regions ranging from 500 bp upstream of these candidate pri-miRNAs to the 5′ end of the corresponding pre-miRNAs and then predicted miRNA promoters as those candidate promoters whose averaged expression levels over all the FANTOM CAGE samples were the highest. The predicted miRNA promoters were further validated by ENCODE RAMPAGE (RNA annotation and mapping of promoters for the analysis of gene expression) sequencing data and RACE experiments.

It has been shown miRNA expression is condition-specific [56, 57]. However, the aforementioned studies in general predicted miRNA TSSs in a non-cell-specific manner by pooling data from multiple experimental conditions. Aiming for condition-specific miRNA TSS identification, Georgakilas et al. created the computational tool microTSS to identify miRNA TSSs directly from deeply sequenced RNA-Seq data [31]. Scanning 30 bp window in the 400 kb upstream of each pre-miRNA, they identified putative miRNA TSSs. These putative TSSs were further classified as true or false TSSs by three SVM models that were pre-trained on the H3K4me3, Pol II, digital genomic footprinting DNase-Seq, TFBSs of 8740 protein-coding genes, respectively. In total, they identified 70 intergenic miRNA TSSs (for 118 pre-miRNAs) in mESCs, 63 TSSs (for 86 pre-miRNAs) in hESCs and 50 (for 82 pre-miRNAs) in IMR90 cells. Taking advantage of various types of cell-specific data available from the ENCODE project, Hua et al. predicted cell-specific miRNA TSSs in 54 cell lines by linearly combining H3K4me3, DNase-Seq, phastCons conservation and Eponine TSS scores [25]. Their method identified TSSs for 663 intragenic miRNAs and 620 intergenic miRNAs [58].

Beyond using CAGE data as an important feature for miRNA TSS prediction, mirSTP took advantage of global nuclear run-on sequencing (GRO-Seq) and precision nuclear run-on sequencing (PRO-Seq) experimental data [59]. Unlike CAGE experiments that capture only the 5′ end of the transcripts, GRO/PRO-Seq experiments are able to measure the transcriptional activities over the whole transcripts. By identifying the sharp peaks of GRO/PRO-Seq profiles with a likelihood ratio test, mirSTP was able to detect high-resolution condition-specific miRNA TSSs. mirSTP was applied to 183 GRO-Seq and 28 PRO-Seq experiments in 27 human cell lines and identified TSSs corresponding to 480 intergenic miRNAs. In addition, a recent study from Bouvy-Liivrand et al. integrated 92 GRO-Seq data with FANTOM CAGE data and identified 305 intergenic and 1242 intragenic miRNA TSSs [60].

So far, computational methods have generated TSS predictions for a large portion of all known miRNAs. Such genome-wide miRNA TSS annotation is able to provide insight into the miRNA transcription initiation mechanisms and help further elucidation of the miRNA gene regulation. Therefore, several miRNA TSS databases have been created [38, 61, 62]. For example, miRGen has compiled 276 miRNA TSSs corresponding to 428 pre-miRNAs [61]. However, a comparative study is needed to facilitate the usage and interpretation of current miRNA TSS annotation.

Comparative study of predicted miRNA TSSs led to a robust miRNA TSS annotation set

To better understand the current status of miRNA TSS annotation, we performed a comparative analysis of the most recent high throughput computational annotation of miRNA TSSs. We collected miRNA TSS annotations from 14 studies and resources (Table 1). The compilation resulted in a total of 20 329 TSS loci that correspond to 1801 of the total 1881 miRNAs in miRBASE (v21). Since the genomic coordinates of the different data sets vary, we used the liftOver program from the UCSC Genome browser [27] to convert all data sets into the hg19 version.

Initial summarization of miRNA TSS annotations from these studies showed that, although the majority of known miRNA TSSs (both intergenic and intragenic) were predicted, most of

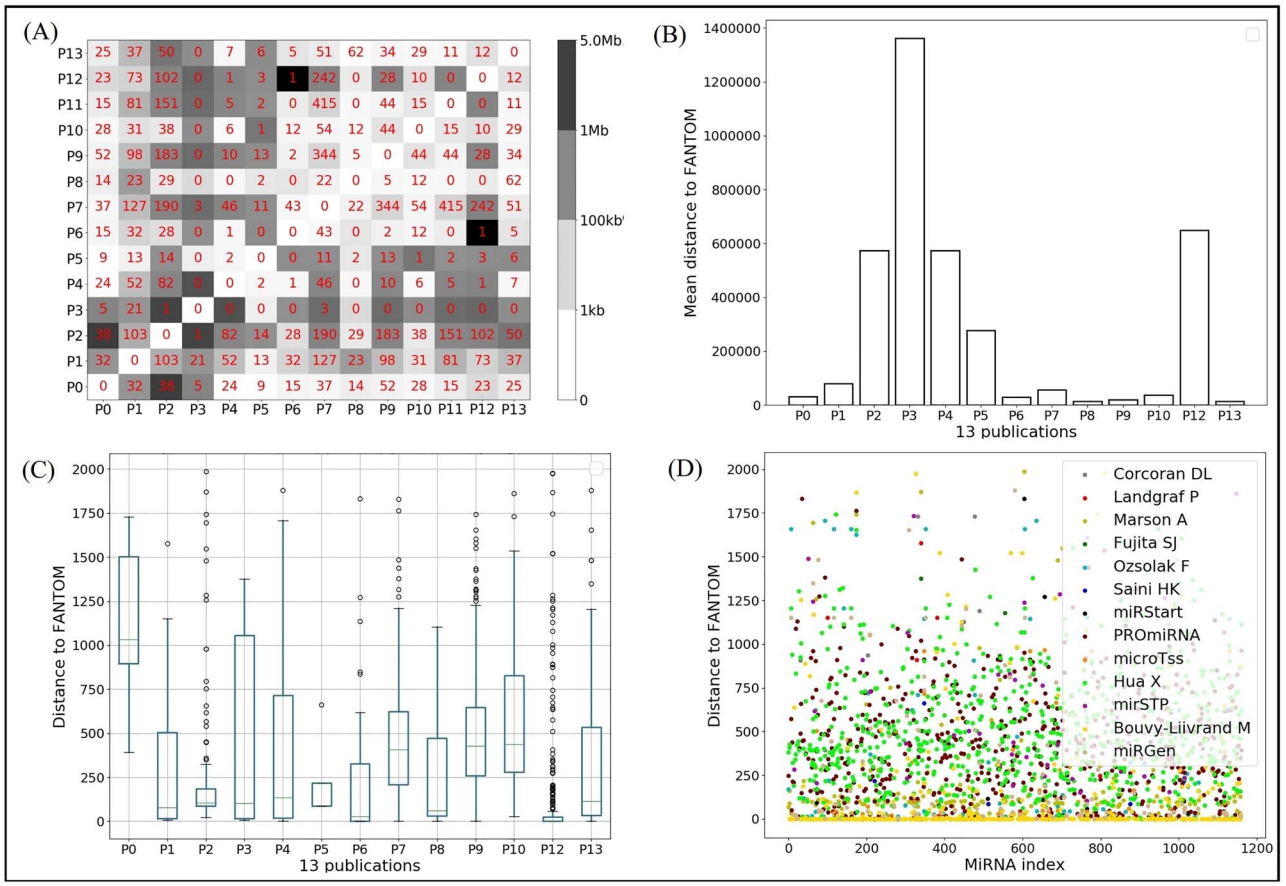

FIG. 1. **(A)** Average TSS distance heat map annotated by the 13 studies. The numbers inside represent the number of miRNAs for which each study has consistent TSS annotations with other studies. **(B)** Mean distances to the FANTOM annotations from all miRNA TSSs in 13 studies, respectively. **(C)** Mean distances to the FANTOM annotations of the miRNA TSSs in 13 studies within 2 kb of their corresponding FANTOM annotations. **(D)** All distances to the FANTOM annotations from the 13 studies for only miRNA TSSs that are located within 2 kb of their corresponding FANTOM annotations.

these annotations were predicted based on features pooled from multiple cell lines and thus are not cell-specific (Table 1). There are also cases where multiple TSS annotated for one miRNA, resulting in a much larger total number of TSSs than the number of miRNAs in some studies (P7, P9, P10). For example, PROmiRNA reported 7133 TSSs corresponding to 1333 miRNAs. This is consistent with recent discoveries on alternative TSSs that illustrates the complexity of cell transcription initialization [63–65]. Besides, the resolution of TSS prediction often varies by studies. For example, a TSS prediction can be a genomic region up to a million base pair long [15, 34]. Several CAGE-based methods have reported single base pair resolution TSSs [38, 58, 66]. In addition, we observed that a subset of methods based on FANTOM CAGE experiments such as PROmiRNA and Bouvy-Liivrand et al. have predicted for a largely common set of known miRNAs. However, the predicted TSSs are often inconsistent with each other (located ¿+/-100 bp away from each other). In fact, 1094 out of the 1801 (61%) miRNAs predicted by them have at least two TSS loci that are at least 100 bp apart [39, 60].

Because of the lack of ground truth annotations for miRNA TSSs, it is challenging to directly compare different studies. Therefore, instead of making a direct comparison of the 14 studies in terms of the prediction accuracy, we attempted to assess the consistency of these annotations. Initial investigation shows that the TSSs supported by the 14 studies are not consistent for all studies (Figure 1A). In fact, we found only 40 TSSs supported by half of the 14 studies (Supplemetary Figure S1). To further evaluate the consistency between different studies, we decided to use one study as a reference. We reasoned it was not essential which study we would choose to use as the reference if our goal was to evaluate the consistency of all studies by measuring the distance relative to this specific reference. If the studies were truly not consistent, they would not become consistent due to certain choice of a reference study. We finally decided to use FANTOM miRNA TSS annotation (P11 in Table 1) as the reference for the comparison, since FANTOM (version 5) published the most recent and largest number of validated miRNA TSSs. Note that, choosing other studies as reference would also work and would not affect our conclusion of inconsistency of the 14 studies. So, we calculated the distances between the predicted miRNA TSSs from each study and their corresponding annotations from FANTOM (Figure 1B and C). Also, when calculating distance from FANTOM, if a miRNA has multiple TSSs annotated in one study, we treated each TSS separately. Also, when a miRNA TSS is annotated as a region, we calculated the distance using the central point of the region. Averaging all distances, we found that nearly half (45%) of the annotations (excluding those from FANTOM) have their averaged distances greater than 2 kb. For example, Fujita et al. (P3) annotations have their averaged distance 1.4 Mb and thus differ most from FANTOM TSS loci, whereas miRGen (P13) and microTSS (P8) have the smallest averaged distance, but both are around 13 kb (Figure 1C).

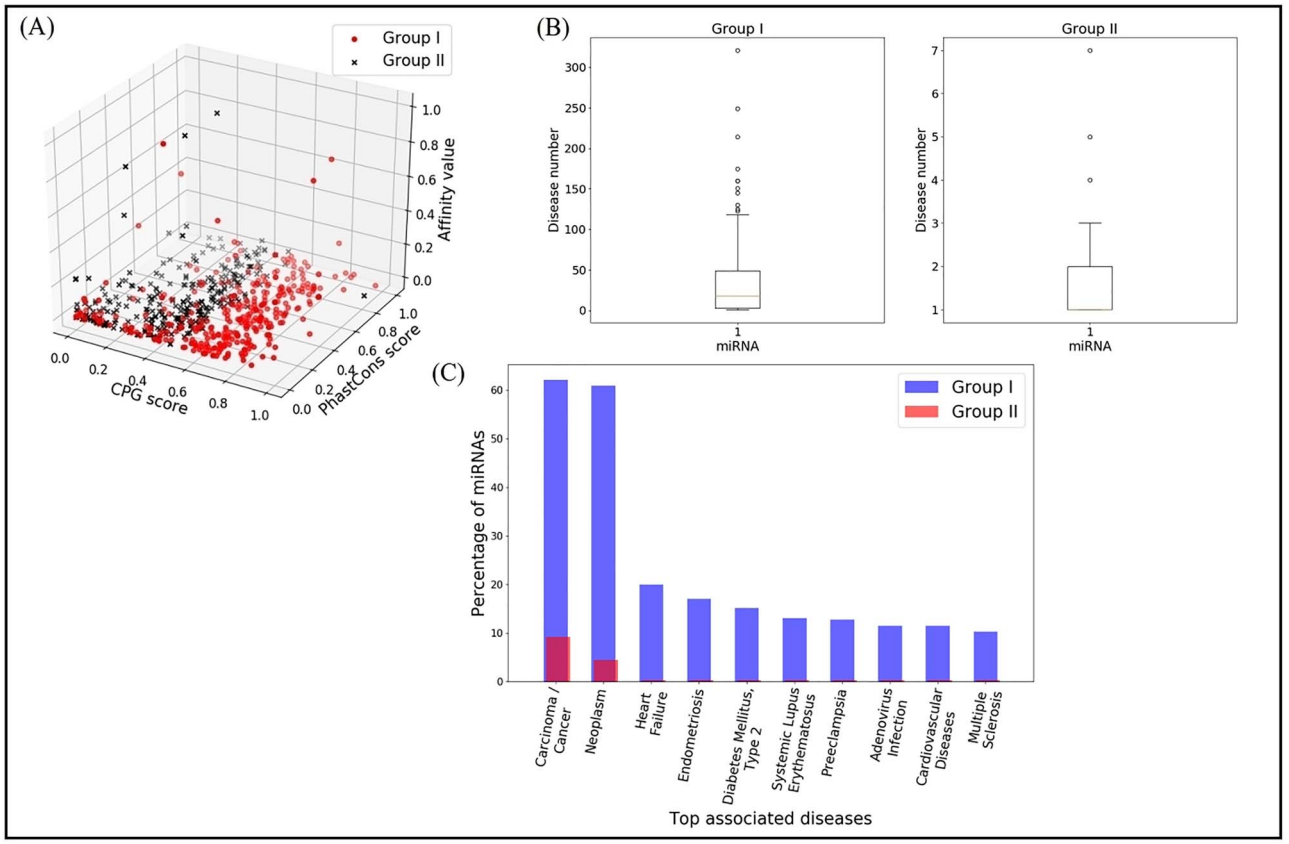

FIG. 2. **(A)** CpG content scores, PhastCons conservation scores across 46 vertebrates and affinity scores for the TATA box-binding motifs, for TSSs in groups I and II. **(B)** Disease number distributions for miRNAs in groups I and II. **(C)** The top 10 diseases associated with the miRNAs in groups I and II.

Although the averaged distances from different studies are quite large, the majority of their predictions (55%) are actually located within 2 kb of FANTOM annotations. However, on a close investigation, these predictions within the 2 kb range are not always consistent with each other. By calculating the averaged distance to the FANTOM annotations for only the predictions that are located within the 2 kb range of their corresponding FANTOM annotations, we found that the largest averaged distance is 1087 bp [37] and the smallest averaged distance is 108 bp [60]. In general, CAGE-based studies are most consistent with FANTOM predictions in comparison to those not based on CAGE experiments for TSS annotations (Figure 1D).

To understand the differences of the TSS annotation in terms of the number of literature support, we divided the miRNA TSSs into two groups: group I contains miRNA TSSs annotated consistently in at least four studies, and group II contains miRNA TSSs annotated only in one study. Note that, although intuitively, miRNAs in earlier studies are more likely to be selected in group I, we also observed many miRNAs annotated in recent years in group I. No obvious bias toward those miRNAs in earlier studies were noticed perhaps due to the frequency of recent genome-scale TSS annotation studies. Here, we defined the consistency of annotations from multiple studies as annotations within +/- 100 bp of each other. We evaluated the GO categories of miRNAs in groups I and II using the miR2GO tool (*P*-value cutoff = 1e-16) [67]. We found that group I miRNAs are more likely to be annotated to more general GO categories such as DNA-binding TF activity (GO:0000981) and binding (GO:0005488), while group II miRNAs' annotations are likely to be more specific such as neurogenesis (GO:0022008) and positive regulation of cellular process (GO:0048522). We further calculated the genomic features including CpG content, conservation across vertebrates (PhastCons score) and affinity score of the TATA box-binding motifs for TSSs in groups I and II, respectively (Figure 2A). The genomic features were calculated along the 1 kb region centered around a miRNA TSS. We used the same scoring strategy, parameters and data used by PROmiRNA to calculate these three features [39].

We found the distributions of these genomic features are similar for both groups, although miRNA TSSs from the group I tend to have a higher CpG score. We then investigated whether the inconsistency can be caused by the functional significance of certain miRNAs. We examined the associations of miRNAs and diseases in the Human MiRNA & Disease Database (HMDD) [68]. HMDD stored manually retrieved associations of miRNAs and diseases from literature. There are currently 472 miRNAs, 351 diseases and 4489 miRNA-disease associations in HMDD (version 3.0). We found that miRNAs with their TSS annotations consistent in multiple studies tend to be associated with a larger number of diseases comparing with those having annotations from only one study. The 76% of group I miRNAs were associated with at least one disease, while the number was only 16% for group II. On average, 130 diseases per miRNA in group I and 1 disease per miRNA in group II were observed (Figure 2B). The top 10 associated diseases also showed a much higher association of group I miRNAs than group II (Figure 2C). For example, hsa-miR-21 in group I is associated with 160 diseases including granulosa cell tumor, allergic asthma, idiopathic pulmonary fibrosis and so

on. In contrast, the miRNA associated with the largest number of diseases in group II is hsa-mir-302f that is found to be associated with only seven diseases including head and neck neoplasms, breast neoplasms, gastric neoplasms and so on.

Because of the possibility of alternative TSS usage under different cellular conditions, the inconsistency of TSS annotations might be due to the dynamic usage of alternative TSSs. We investigated the transcriptional activities of different TSS loci that correspond to the same miRNA based on CAGE experimental peak data from FANTOM. If a predicted TSS has a CAGE peak in its +/-ω bp surrounding region, we defined this TSS as ω-CAGE-supported. Among all the 20 329 TSSs associated with 1801 miRNAs, 7252 (36%) were identified 100-CAGE-supported, and 11 565 (57%) are 500-CAGE-supported, suggesting that the multiple TSS annotations corresponding to the same miRNA could be alternative TSSs activated under different cellular conditions. For example, hsa-miR-770 has four TSSs spread over 1 kb, all of which are supported by CAGE peaks, but are found activated under different cellular conditions such as testicular germ cell, embryonal carcinoma cell, hippocampus cell and so on. Similarly, hsa-miR-1227 has four TSSs activated in acute lymphoblastic leukemia, monocytes and fibroblast cell lines. Among all the 8404 intergenic miRNA TSSs, 2409 (29%) were 100-CAGE-supported, and 4083 (49%) were 500-CAGE-supported. Among all the 12 028 intragenic miRNA TSSs, 4882 (41%) were identified as 100-CAGE-supported, and 7543 (63%) were 500-CAGE-supported. Therefore, both intergenic and intragenic miRNAs could have alternative TSS annotations in different cell lines.

In summary, the miRNA TSS annotation from different studies is often inconsistent. In general, CAGE-based studies are most consistent with the FANTOM predictions in comparison to those not based on CAGE experiments. miRNAs with their TSS annotation consistent in multiple studies are often associated with a larger number of diseases than those annotated in only one study. We also showed that dynamic usage of alternative TSSs under different cellular conditions might be one of the causes of miRNA TSS annotation inconsistency.

Characterization of a RSmiRT

Robust TSS annotations for miRNAs benefit the characterization of miRNA TSS usage, transcription initiation mechanisms and de-novo prediction validations. We attempted to create a RSmiRT following the steps below based on the collection of current annotations. We first clustered TSS annotations that are within +/-100 bp of each other, and we then redefined the centers of these clusters as unified TSS loci (UTLs), we finally kept only the UTLs that are supported by at least four surveyed studies. This procedure resulted in 311 UTLs associated with 2064 predicted TSSs and 330 miRNAs (202 intragenic and 128 intergenic miRNAs). These miRNAs and their UTLs were included in the RSmiRT for further study of genomic and epigenomic features of miRNA TSSs.

Intragenic miRNA genes have independent TSSs from their host gene TSSs

Even though earlier studies often assumed intragenic miRNAs are co-transcribed with their host genes [69], large-scale miRNA TSS annotation has shown that intragenic miRNAs and their host genes can have independent transcription initiation mechanisms [35, 37, 39]. For example, by identifying transcription initiation regions of 175 miRNAs in three human cancer cell lines, Ozsolak et al. discovered 32 of the 88 intronic miRNAs having TSSs that are different from their host genes. Observations have also been made that intragenic miRNAs might not always have a good expression correlation with their host genes. Take hsa-miR-32 for example, it has been shown to often have a negative correlation with its host gene C9orf5 [70].

Our study on RSmiRT also shows the existence of independent promoters of intragenic miRNAs. For the 311 UTLs corresponding 330 miRNAs, we have 213 UTLs corresponding to a total of 202 intragenic miRNAs. For these intragenic miRNAs, we identified their host genes using the miRIAD database [71]. The distance distribution between the intragenic miRNA TSSs and their host gene TSSs is shown in Figure 3. We observed that miRNA-host gene TSSs can be hundreds of bp and even millions of bp away from each other. For example, hsa-miR-2276 has its UTL 180 kb away from its host gene SPATA13 and similarly, hsa-miR-24-1 has a UTL 278 kb away its host gene C9orf3. In fact, 81 out of the 202 intragenic miRNAs (~40%) have their miRNA UTLs more than 1 kb away from their host gene TSSs, implying the existence of independent promoters. This is consistent with the previous understanding that intragenic miRNAs with their own promoters are more likely to be far away from host gene TSSs [35].

We also investigated the expression correlation between intragenic miRNAs and their host genes using the miRNA and mRNA measurements in 18 samples (9 disease samples and 9 normal samples) corresponding to 9 tissues in Lu et al. and Ramaswamy et al. [20, 72]. This data contains expression profiles corresponding to 164 miRNAs and 10 991 genes and has been frequently used to study miRNA–mRNA co-expression [73–75]. Using this expression data, we were able to identify the expression profiles of 30 intragenic miRNAs of RSmiRT along with the transcripts of their host genes annotated by miRIAD. We calculated the Spearman correlation coefficient between expression profiles of these miRNAs and host genes. We found the majority (~73%) of the miRNA-host gene transcript correlations were below 0.2 suggesting the likelihood of independent promoters owned by these miRNAs (Figure 3B and C).

In summary, our study on RSmiRT confirmed intragenic miRNAs may have independent TSSs from their host genes, which is supported by literature and our study on RSmiRT. We also showed that the expression correlation of intragenic miRNAs and their host genes together with the genomic distance between their annotated TSSs can help identify independent promoters of intragenic miRNAs.

Genomic, epigenomic features and functional annotations of the RSmiRT

For the 330 miRNAs involved in the RSmiRT, we investigated their genomic features (Figure 4A–C). We calculated the scores regarding CpG content, PhastCon conservation and TATA-binding affinity, respectively, along the 1 kb region centered around the miRNA TSS as described in PROmiRNA [39]. We found that host genes and intergenic miRNAs in RSmiRT in general have similar genomic feature distributions. This is consistent with the previous observation that protein-coding gene and intergenic miRNA promoters often share similar sequence features [37]. In fact, we did not notice sequence features that are substantially different among the three different promoter classes although intergenic miRNA promoters tended to have lower CpG scores than intragenic miRNAs and host genes (Figure 4A–C).

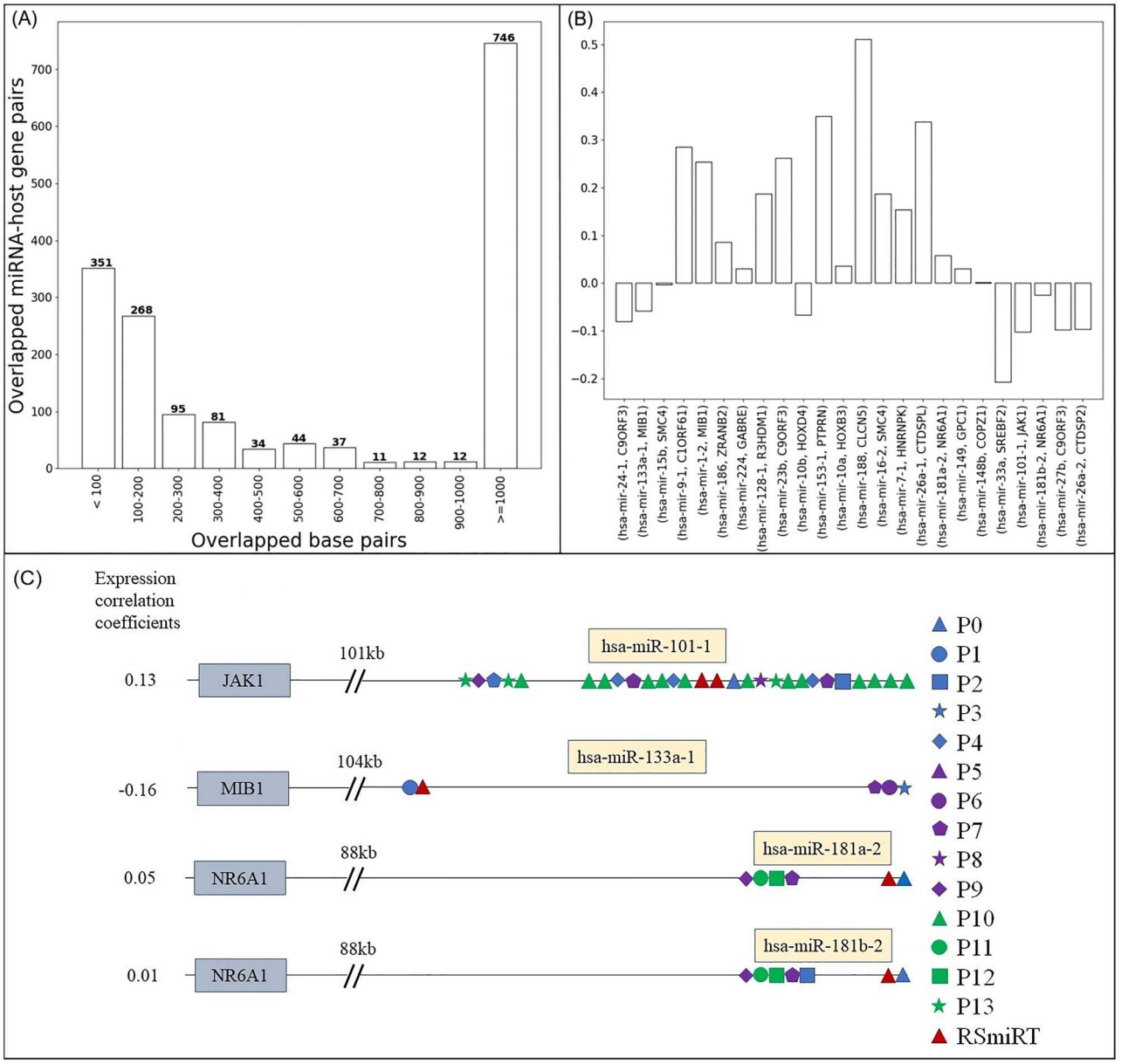

FIG. 3. **(A)** Distance between the intragenic miRNA TSSs and their host gene TSSs. **(B)** Expression correlation between the intragenic miRNAs and their host genes. **(C)** Examples of possible independent TSS of intragenic miRNAs and their host genes. The host gene TSSs are on the left side of the grey box, which are far from the predicted intragenic miRNA TSSs.

Epigenetic features such as H3K4me1, H3K4me3 and H3K27ac have often been considered as effective chromatin markers for protein-coding and miRNA gene promoter predictions. We thus investigated the H3K4me1, H3K4me3 and H3K27ac in the 1 kb regions surrounding the TSSs of the intergenic, intragenic miRNAs and their host genes. To focus on the TSSs that are active under a given condition, we downloaded CAGE data in seven cell lines: A540, GM12878, H1, HeLa, HEPG2, K562 and MCF7 from the FANTOM project. We further identified a subset of active TSSs in the seven cell lines by requiring the UTLs were 500-CAGE supported. This means, each active UTL needs to have at least one CAGE peak in its 1 kb surrounding region. We then downloaded the corresponding H3K4me1, H3K4me3 and H3K27ac ChIP-Seq data from the ENCODE project (Table 2). For each active TSS in the seven cell lines, we divided its 1 kb surrounding region into 100 bins (each of size 10 bp). Considering the maximum read number in the region of a bin as the read coverage of that bin, we calculated the average read coverages in these 100 bins and obtained their H3K4me1, H3K4me3 and H3K27ac coverage profiles in the seven cell lines (Figure 4D–F). We observed similar distributions of these chromatin markers over the three types of promoter regions, suggesting such epigenetic features themselves might not be able to effectively distinguish the miRNA TSSs from those of other RNA transcripts.

We further studied the function of RSmiRT miRNAs in terms of their disease associations. We found the majority of these miRNAs had disease associations in literature. Out of the 330 miRNAs in the RSmiRT, we identified 201 miRNAs including 90 intergenic miRNAs and 111 intragenic miRNAs associated with at least one of the top 10 diseases in the HMDD database. For

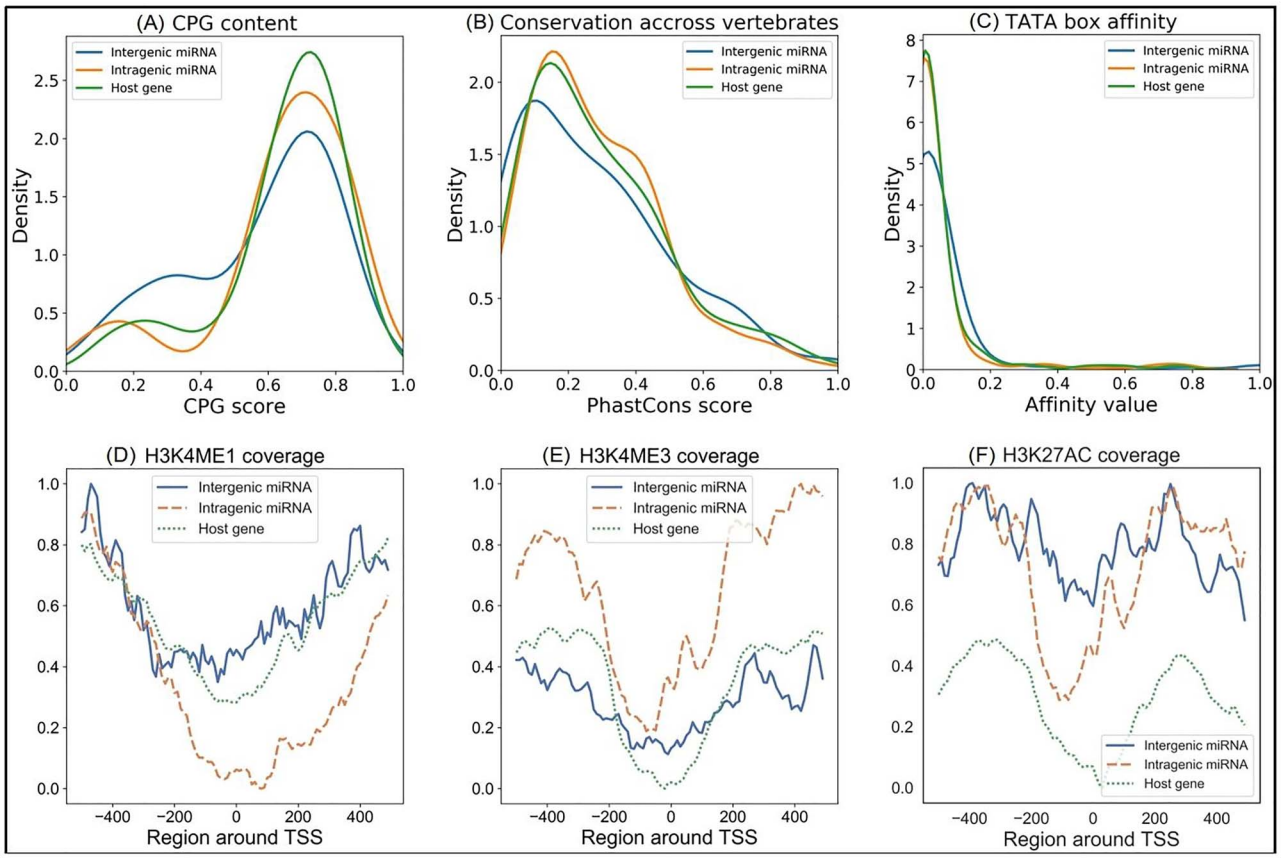

FIG. 4. Density plot of **(A)** CpG score, **(B)** conservation score and **(C)** TATA box affinity of the 1 kb region around intergenic, intragenic miRNA TSS and the host gene TSSs of the intragenic miRNAs. Distribution of **(D)** H3K4me1, **(E)** H3K4me3 and **(F)** H3K27ac signals across the 1 kb region around the intergenic, intragenic miRNA TSSs and the host gene TSSs of the intragenic miRNAs.

Table 2. ENCODE experiment IDs for the histone data sets

	GM12878	HELA	K562	MCF7	H1	HEPG2	A549
H3K4ME1	ENCSR000AKF	ENCSR000APW	ENCSR000EWC	ENCSR493NBY	ENCSR000ANA	ENCSR000APV	ENCSR000AVH
H3K4ME3	ENCSR057BWO	ENCSR340WQU	ENCSR000DWD	ENCSR985MIB	ENCSR814XPE	ENCSR575RRX	ENCSR000AST
H3K27AC	ENCSR000AKC	ENCSR000AOC	ENCSR000AKP	ENCSR752UOD	ENCSR000ANP	ENCSR000AMO	ENCSR000AVF

example, 169 miRNAs are associated with hepatocellular carcinoma and 145 miRNAs are associated with breast neoplasms. We observed that many of these miRNAs are associated with multiple diseases. For example, 82.1% of the 201 miRNAs are associated with at least 2 of the top 10 diseases. Nearly 20% of the 201 miRNAs including hsa-miR-195 and hsa-miR-27b are associated with 10 diseases. hsa-miR-195 and hsa-miR-27b are associated with all top 10 diseases. These miRNAs involved in many diseases often have consistent TSS annotations. For example, hsa-miR-195 TSS annotation is supported by 10 of the 14 surveyed studies. Among them, four studies have their annotations within +/-100 bp of each other. For hsa-miR-155, 9 of the 14 surveyed studies have their TSS annotations, and 6 of the 9 studies have their annotations located within +/-100 bp of each other.

In summary, the TSSs of intergenic miRNAs, intragenic miRNAs and their host genes have shown similar genomic and epigenomic feature patterns. Additional functional anaylsis of miRNAs in RSmiRT shown their involvement in multiple diseases.

Existence of alternative TSSs

Both computational and experimental studies have reported the existence of alternative TSSs. For example, mirSTP found that the miR200b cluster uses alternative TSSs under different cellular conditions [59]. Hua et al. also identifiedsix alternative TSSs per intragenic miRNA and five alternative TSSs per intergenic miRNA [58]. Among the 330 miRNAs in the RSmiRT, we found 26 out of 128 intergenic miRNAs (20.3%) have multiple UTLs. For example, hsa-miR-940 has two UTLs that are 508 bp away. Similarly, hsa-miR-503 has two UTLs that are 400 bp away; 11 out of 202 intragenic miRNAs (5.4%) have multiple UTLs, such as hsa-miR-1303 and hsa-miR-101-1. The majority of these miRNAs have their UTLs supported by CAGE data. The 82 (70.6%) of the intergenic miRNA UTLs and 169 (86.7%) of the intragenic miRNA UTLs are 100-CAGE supported. For example, chr13:92000048 UTL of the intergenic miRNA hsa-miR-19b-1 overlaps with a cage peak chr13:92001257-92001286 in K562 cell line and chr11:75062740 UTL of the intragenic miRNA hsa-miR-326 has an overlapping cage peak in chr11:

75062730-75062775 in NEC15 cell line. The 96 (82.8%) of the intergenic miRNA TSSs and 174 (89.2%) of the intragenic miRNA TSSs are 500-CAGE supported. For example, chr10:98592616 UTL of the intergenic miRNA hsa-miR-607 overlaps with a cage peak chr10:98592674-98592724 in NCI H82 cell line and chr19:52202963 UTL of the intragenic miRNA hsa-let-7e overlaps with a cage peak in chr19:52196579-52196584 in donor cell line. miRNAs with multiple UTLs are often found as 100-CAGE-supported or 500-CAGE-supported in different cell lines. For example, hsa-miR-940 has two different UTLs in chromosome 16 that are 100-CAGE-supported in embryonic kidney and adipose cell lines, respectively. Similarly, hsa-miR-29b-2 has two UTLs that are 100-CAGE-supported in locus coeruleus and extra skeletal myxoid chondrosarcoma cell lines.

In summary, we identified alternative TSSs by clustering consecutive TSSs based on their mutual distances. The presence of multiple clusters located apart by a large distance (¿100 bp) indicated the possible alterative TSSs of a miRNA. These alternative TSSs were supported by CAGE data in different cell lines.

Overlapping between miRNAs and long non-coding RNAs

Long non-coding RNAs(lncRNAs) can overlap with miRNA loci. Studies have shown lncRNA can encode miRNAs and pri-miRNAs can also participate in lncRNA biosynthesis [76, 77]. Computational studies on miRNA TSS annotation also identified many lncRNA-miRNA overlapping pairs. For example, microTSS found that the pri-miRNA of hsa-mir-675 overlaps with the H19 lncRNA gene. In fact, it has been reported more than 20% of intergenic pri-miRNAs overlap with an annotated lncRNA [58]. To investigate the overlap between lncRNAs and the pri-miRNAs involved in RSmiRT, we downloaded 48 260 lncRNA sequences from GENCODE (version 19). Because most of the 3′ ends of pri-miRNA sequences are not reliably annotated, we only consider the part of a pri-miRNA between its annotated TSS and the 3′ end of its pre-miRNA for the calculation of lncRNA overlap.

Interestingly, we observed 288 (87%) of the miRNAs in RSmiRT overlapping with at least one of GENCODE lncRNAs. These overlapped miRNAs consisted of 95% of the intergenic miRNAs (121 out of 128) and 82% of the intragenic miRNA (167 out of 202). Note that, one miRNA can have multiple TSSs, which allows it to have multiple pri-miRNA and each pri-miRNA can overlap with multiple lncRNAs. This led to a total of 5593 pri-miRNA-lncRNA overlapping pairs. On closer inspection, 265 of 288 pri-miRNAs embedded at least one lncRNA inside their regions. The percentages of embedding miRNAs were similar for both intergenic (82%) and intragenic miRNAs (79%). For example, the pri-miRNA transcript of hsa-miR-148b itself embedded 38 lncRNAs. In the meantime, we found 73 lncRNAs containing seven pri-miRNA transcripts, many of these lncRNA regions were overlapped by each other. For example, MIR210HG lncRNA family (MIR210HG-201 to MIR210HG-205) contains hsa-miR-210 pri-miRNA transcript.

We also considered a non-overlapping pri-miRNA-lncRNA pair as ‘adjacent’ if their distance is no larger than 100 bp. This resulted in 29 miRNAs being adjacent to lncRNAs. Among them 27 were intergenic. For example, hsa-miR-22 pri-miRNA is adjacent to MIR22-HG lncRNA family. We also investigated this overlap statistic in terms of pre-miRNA locations of the corresponding 330 miRNAs. We found 47 pre-miRNAs (17 intergenic and 30 intragenic) overlapping with a lncRNA. In only one case, we found a pre-miRNA embedding a lncRNA (hsa-miR-101-2 embedded AL158147.1-201 lncRNA). All other cases showed complete embedding of a pre-miRNA by a lncRNA, which is understandable given the much longer sizes of lncRNAs compared to the pre-miRNAs. We found only one case where the pre-miRNA was adjacent to a lncRNA (hsa-miR-196b was less than 100 bp upstream of HOXA-AS3-201 lncRNA).

Several overlapped and adjacent (pri-miRNA, lncRNA) pairs we found are involved in various disease pathways according to the experimentally supported lncRNA-disease association data downloaded from LncRNADisease database [78]. For example, in the case of (hsa-miR-155, MIR155HG) pair, the lncRNA MIR155HG is adjacent in the downstream of hsa-miR-155 pri-miRNA transcript. This pair is associated with chronic lymphocytic leukemia (CLL) where MIR155HG transcriptionally regulates the hsa-miR-155 host gene. The overlapped (hsa-miR-16-1/hsa-miR-15a, DLEU1/DLEU2) pairs are also involved in CLL. DLEU1 and DLEU2 co-regulate several tumor suppressor genes, including the miRNA genes hsa-miR-16-1 and hsa-miR-15a. These miRNA genes are downregulated in multiple tumor types and are frequently deleted in CLL, myeloma and mantle cell lymphoma. Also, DLEU2 overexpression blocks cellular proliferation and inhibits the colony-forming ability of tumor cell lines in a miR-15a/miR-16-1-dependent way. In the case of (hsa-miR-31, MIR31HG) pair, hsa-miR-31 is fully embedded by the MIR31HG family, where both hsa-miR-31 and its host gene lncRNA LOC554202 (MIR31HG) are downregulated in breast cancer. The hsa-miR-17∼92 cluster contains several pri-miRNAs: hsa-miR-17, hsa-miR-18a, hsa-miR-19a, hsa-miR-20a, hsa-miR-19b-1 and hsa-miR-92a-1. This cluster of miRNAs is adjacent to the upstream of the lncRNA MIR17HG. According to the LncRNADisease database, germline deletion of MIR17HG encodes the miR-17∼92 polycistronic miRNA cluster in individuals with microcephaly, short stature and digital abnormalities.

In summary, we analyzed the location of the miRNA transcripts with respect to that of annotated lncRNAs. We found that a large percentage of RSmiRT pri-miRNA transcripts overlap with or are adjacent to lncRNAs. These pri-miRNA-lncRNA pairs may be co-involved in different disease pathways.

Contribution of miRNA TSS annotations to gene regulatory network construction

The annotation of miRNA TSSs is essential to the study of miRNA gene regulation and can thus help the miRNA-TF association and gene regulatory network construction. Ozsolak et al. identified dozens of miRNA promoters potentially bound by TF MITF, whose corresponding gene is known to be involved in melanoma development [35]. mirSTP found TF binding regions within the surrounding regions of the predicted miRNA TSSs based on the ENCODE TF ChIP-Seq data analysis [59]. PROmiRNA also assigned TFs to miRNA promoters by performing enrichment analysis of TFBSs from the JASPAR database [35].

To identify potential TF-miRNA interactions regarding the RSmiRT data, we employed TRAP [79] to identify TFBSs enriched in the TSS-surrounding regions corresponding to the RSmiRT. Using the multiple sequence webtool of TRAP, we searched for the TF motifs from the JASPAR vertebrates database [35] along the 1 kb region surrounding the miRNA TSS, with human promoters as background model and the Benjamini–Hochberg as multiple test correction method. We considered combined P-value cutoff ≤ 0.01 to consider only the most significant TFs. We obtained three sets of TFs for intergenic, intragenic miRNA and host gene TSSs, respectively. In order to scan for TF motifs, we categorized the regions around TSSs into six groups:

Table 3. Unique and common TF motifs for the six region groups

	Transcription factors
TFs found only around intergenic miRNA TSSs	CREB1, Spz1, Stat3, EWSR1-FLI1
TFs common around host gene and intragenic miRNA TSSs, but not present around intergenic miRNA TSSs	ELK4, Myb, Mafb, USF1, Tcfcp2l1, NHLH1, Myf
TFs common around the TSSs in the three groups	Zfx, MZF1_5-13, GABPA, REL, NF-kappaB, Arnt, MIZF, Egr1, ELK1, PLAG1, MZF1_1-4, NFYA, Zfp423, Klf4, INSM1, SP1, NFKB1, Pax5, Myc, RELA, TFAP2A, RREB1, E2F1, CTCF, EBF1, Mycn
TFs found only around intragenic miRNA TSSs near their host genes	CTCF, GABPA, MZF1_1-4, EBF1, Tcfcp2l1, Myb, REL, MIZF, Myf, Mafb
TFs found only around intragenic miRNA TSSs far from their host genes	Spz1, NHLH1
TFs found only around host gene TSSs far from their host genes	RREB1
TFs common around intragenic miRNA TSSs but not present around the host gene TSSs that are located far	ELK4, ELK1, E2F1, Zfx
TFs common around intragenic miRNA TSSs (near) and intragenic miRNA host gene TSSs (far) but not present around corresponding intragenic miRNA TSSs (far)	INSM1, NFKB1, Zfp423, RELA, NF-kappaB, PLAG1, MZF1_5-13, Myc, NFYA, Mycn, Pax5
Common TFs around all intragenic miRNA TSSs and their host gene TSSs	Klf4, Egr1, TFAP2A, SP

intergenic miRNA TSSs (116), intragenic miRNA TSSs (195), host gene TSSs (202), intragenic miRNA TSSs that are located near ($\leq$ 100 bp) their host gene TSSs (91), intragenic miRNA TSSs that are located far ($\geq$ 10 kb) from their host gene TSSs (52) and the host gene TSSs of the intragenic miRNAs that are far ($\geq$ 10 kb) from the corresponding miRNA TSSs (52). According to these six categories, we generated different sets of motifs that are unique or common to different subgroups (Table 3). We observed that the TFs around the intergenic, intragenic miRNA TSSs and their host genes TSSs largely overlapped, indicating the potential correlated gene regulation between intergenic, intragenic miRNAs and their host genes. However, ELK4, Myb, Mafb, USF1, Tcfcp2l1, NHLH1 and Myf motifs occurred only around the host gene TSSs of the intragenic genes in RSmiRT but not around the intragenic and intergenic miRNA TSSs, whereas CREB1, Spz1, Stat3 and EWSR1-FLI1 were found only around the intergenic miRNA TSSs. Although different parameters could affect the results, these results show that intergenic, intragenic miRNAs and host genes can both share common regulatory mechanisms and have their different regulatory pathways.

In summary, by studying the most significant TFs around the RSmiRT TSSs, we found a large set of common TFs for the intergenic miRNAs, intragenic miRNAs and their host genes. We also observed that several TFs are specific to the intergenic miRNAs, intragenic miRNAs and their host genes.

Discussion

miRNA TSS identification is important to the understanding of gene regulation. One miRNA can regulate thousands of genes' transcription expression by target binding. miRNAs can also interact with other miRNAs or other RNA species such as lncRNAs to regulate gene expression. Because of the miRNA-specific biogenesis, miRNA TSS identification is more challenging than coding gene TSS detection. However, recent large-scale TSS-Seq data such as CAGE experiments enable dozens of computational studies to predict miRNA TSSs.

The comparative study presented above shows that the current computational prediction of miRNA TSSs is still at the early stage. TSS annotations are largely inconsistent between studies. The observations made by the studies might be affected by this inconsistency and the limited amount of labeled training and testing data. Therefore, a benchmark data set for computational miRNA TSS prediction is highly desirable. We here presented a set of miRNA TSSs that are consistently predicted by at least four studies, i.e. the RSmiRT data set. Initial study of this RSmiRT revealed their genomic, epigenomic and functional features. Further study of the miRNA–lncRNA relationship as well as miRNA gene regulatory network construction has the promise to provide a glance at the complex RNA world.

It is also possible that dynamic usage of alternative miRNA TSSs had some role to play behind the current TSS annotation inconsistency. For example, a study using epigenomic markers as features in a specific cell line perhaps can only detect the miRNA TSSs activated in that cell line. Computational methods based on CAGE-experiments also need to be aware of the capping noise inside the CAGE measurements. Recently large-scale TSS annotations specifically for miRNAs identified from TSS-Seq, GRO-Seq, GRO-cap, as well as DROSHA-inhibited RNA-Seq experiments have become available [43–54, 60, 80]. Efficient computational methods that can learn and model condition-specific miRNA TSSs based on these annotation data are highly desirable. Besides, we have identified the lack of correlation between many intragenic miRNAs and their host genes. With their TSSs being thousands of nucleotides apart from each other, their low correlated expression levels are very likely to indicate their promoter independency. However, we also like to point out that correlation alone is not sufficient to conclude the promoter independency of two genes considering the complex steps involved in the miRNA biogenesis.

Although miRNA TSSs in RSmiRT are supported by a reasonable number of (at least four) studies, they are not necessarily active under a given experimental condition. Cell-specific predictions are still a challenging problem due to the data availability for a large number of cell lines. A couple of recent studies attempted the cell-specific miRNA TSS prediction problem. Hua et al. integrated H3K4me3 and DNase-Seq data together with sequence features to score candidate TSSs detected from CAGE experiments, and microTSS additionally required deeply sequenced, high-coverage, cell-specific RNA-Seq data [31, 58]. While trying to predict condition-specific miRNA TSSs, these

methods might face data unavailability and/or noisy data issue for many cellular conditions.

Based on our investigation, we also found genomic and epigenomic features are largely similar between intergenic, intragenic and host genes. Thus, current computational approaches relying on these features are not expected to distinguish TSSs of miRNAs from those of other RNA species. Further studies on the miRNA biogenesis provide insight into miRNA TSS identification and prediction. When conducting an integrative analysis of the data from different studies, the TSS prediction resolution can also affect the robust TSS annotation. We simply considered the middle point of the predicted TSS regions for data comparison here. However, more sophisticated methods can be developed to compare consistency between two different annotations.

Key Points

- MicroRNA (miRNA) transcription start sites (TSS) identification is essential to understand gene regulation.
- Recent computational methods have generated genome-wide TSS annotations for a large number of miRNAs by integrating large-scale high throughput sequencing data.
- Current computational annotations of miRNA TSS are often inconsistent with each other.
- Current features used for computational annotation of miRNA TSSs cannot distinguish TSSs of miRNAs from those of other RNA species.
- A robust set of miRNA TSS annotations is needed to characterize genomic and epigenomic features of miRNA TSSs.

Authors' contributions

X.L. and H.H. conceived and designed the experiments. S.W., A.T., M.C. and H.H. analyzed data. H.H. wrote this manuscript. All authors read and approved the final manuscript.

Funding

National Science Foundation (grant number 1356524, 1149955, 1661414), National Institutes of Health (grant number R15GM-123407).

Supplementary Data

Supplementary data are available online at http://hulab.ucf.edu/research/projects/RSMIRT/index.html

References

1. Bartel DP. MicroRNAs. *Cell* 2004; **116**(2): 281–97.
2. Lee Y, Kim M, Han J, *et al*. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J* 2004; **23**(20): 4051–60.
3. Amy E, Pasquinelli SH, Bracht J. MicroRNAs: a developing story. *Curr Opin Genet Dev* 2005; **15**(2): 200–5.
4. Ding J, Li X, Haiyan H. MicroRNA modules prefer to bind weak and unconventional target sites. *Bioinformatics* 2014; **31**(9): 1366–74.
5. Huntzinger E, Izaurralde E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet* 2011; **12**(2): 99–110.
6. Liu W, Wang X. Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. *Genome Biol* 2019; **20**(1).
7. Lee RC, Feinbaum RL, Ambros V. The c. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 1993; **75**(5): 843–54.
8. Wightman B, Ha I, Ruvkun G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in c. elegans. *Cell* 1993; **75**(5): 855–62.
9. Eric C, Lai PT, Williams RW, *et al*. Computational identification of drosophila microrna genes. *Genome Biol* 2003; **4**(7): R42.
10. Lim LP. The microRNAs of caenorhabditis elegans. *Genes Dev* 2003; **17**(8): 991–1008.
11. Rajagopalan R, Vaucheret H, Trejo J, *et al*. A diverse and evolutionarily fluid set of microRNAs in arabidopsis thaliana. *Genes Dev* 2006; **20**(24): 3407–25.
12. Benjamin PL, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are MicroRNA targets. *Cell* 2005; **120**(1): 15–20.
13. Lee PL, Lau NC, Garrett-Engele P, *et al*. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* 2005; **433**(7027): 769–73.
14. Friedman RC, Farh KK-H, Burge CB, *et al*. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* 2008; **19**(1): 92–105.
15. Marson A, Levine SS, Megan F. Cole, et al. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* 2008; **134**(3): 521–33.
16. Garzon R, Fabbri M, Cimmino A, *et al*. MicroRNA expression and function in cancer. *Trends Mol Med* 2006; **12**(12): 580–7.
17. Sassen S, Miska EA, Caldas C. MicroRNA–implications for cancer. *Virchows Arch* 2007; **452**(1): 1–10.
18. Paul SM. Small RNAs with big impacts. *Nature* 2005; **435**(7043): 745–6.
19. Rottiers V, Najafi-Shoushtari SH, Kristo F, *et al*. MicroRNAs in metabolism and metabolic diseases. *Cold Spring Harb Symp Quant Biol* 2011; **76**(0): 225–33.
20. Lu J, Getz G, Miska EA, *et al*. MicroRNA expression profiles classify human cancers. *Nature* 2005; **435**(7043): 834–8.
21. Brian CS, Li X. Transcriptional regulation of mammalian miRNA genes. *Genomics* 2011; **97**(1): 1–6.
22. Han J. The drosha-DGCR8 complex in primary microRNA processing. *Genes Dev* 2004; **18**(24): 3016–27.
23. Hutvagner G. A cellular function for the RNA-interference enzyme dicer in the maturation of the let-7 small temporal RNA. *Science* 2001; **293**(5531): 834–8.
24. Zhou X, Ruan J, Wang G, *et al*. Characterization and identification of MicroRNA core promoters in four model species. *PLoS Comput Biol* 2007; **3**(3):e37.
25. Down TA. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res* 2002; **12**(3): 458–61.
26. Saini HK, Griffiths-Jones S, Enright AJ. Genomic analysis of human microRNA transcripts. *Proc Natl Acad Sci U S A* 2007; **104**(45): 17719–24.
27. Smalheiser NR. Est analyses predict the existence of a population of chimeric microrna precursor-mrna transcripts expressed in normal human and mouse tissues. *Genome Biol* 2003; **4**(7): 403.
28. Jin G, He T, Pei Y, *et al*. Primary transcripts and expressions of mammal intergenic microRNAs detected by mapping ESTs to their flanking sequences. *Mamm Genome* 2006; **17**(10): 1033–41.

29. Fujita S, Iba H. Putative promoter regions of miRNA genes involved in evolutionarily conserved regulatory systems among vertebrates. *Bioinformatics* 2007; **24**(3): 303–8.
30. Megraw M, Pereira F, Jensen ST, *et al*. A transcription factor affinity-based code for mammalian transcription initiation. *Genome Res* 2009; **19**(4): 644–56.
31. Georgakilas G, Vlachos IS, Paraskevopoulou MD, *et al*. microTSS: accurate microRNA transcription start site identification reveals a significant number of divergent pri-miRNAs. *Nat Commun* 2014; **5**(1): 5700–5710.
32. Matthew GG, Levine SS, Boyer LA, *et al*. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* 2007; **130**(1): 77–88.
33. Heintzman ND, Stuart RK, Hon G, *et al*. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 2007; **39**(3): 311–8.
34. Landgraf P, Rusu M, Sheridan R, *et al*. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* 2007; **129**(7): 1401–14.
35. Ozsolak F, Poling LL, Wang Z, *et al*. Chromatin structure analyses identify miRNA promoters. *Genes Dev* 2008; **22**(22): 3172–83.
36. Barski A, Jothi R, Cuddapah S, *et al*. Chromatin poises miRNA- and protein-coding genes for expression. *Genome Res* 2009; **19**(10):1742–51.
37. David LC, Pandit KV, Gordon B, *et al*. Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data. *PLoS ONE* 2009; **4**(4): e5279.
38. Chien C-H, Sun Y-M, Chang W-C, *et al*. Identifying transcriptional start sites of human microRNAs based on high-throughput sequencing data. *Nucleic Acids Res* 2011; **39**(21): 9345–56.
39. Marsico A, Huska MR, Lasserre J, *et al*. PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs. *Genome Biol* 2013; **14**(8): R84.
40. Burnham C, Cha M, Li X, *et al*. Application of deep learning models to microrna transcription start site identification. *Unpublished,* 2019.
41. Barski A, Cuddapah S, Cui K, *et al*. High-resolution profiling of histone methylations in the human genome. *Cell* 2007; **129**(4): 823–37.
42. Tarjei SM, Ku M, Jaffe DB, *et al*. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 2007; **448**(7153): 553–60.
43. Shiraki T, Kondo S, Katayama S, *et al*. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A* 2003; **100**(26): 15776–81.
44. Kodzius R, Kojima M, Nishiyori H, *et al*. CAGE: cap analysis of gene expression. *Nat Methods* 2006; **3**(3): 211–22.
45. Hashimoto S, Suzuki Y, Kasai Y, *et al*. 5′-end SAGE for the analysis of transcriptional start sites. *Nat Biotechnol* 2004; **22**(9): 1146–9.
46. Wei C-L, Ng P, Chiu KP, *et al*. 5′ long serial analysis of gene expression (LongSAGE) and 3′ LongSAGE for transcriptome characterization and genome annotation. *Proc Natl Acad Sci U S A* 2004; **101**(32): 11701–6.
47. Ng P, Wei C-L, Sung W-K, *et al*. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat Methods* 2005; **2**(2): 105–11.
48. Salimullah M, Mizuho S, Plessy C, *et al*. NanoCAGE: a high-resolution technique to discover and interrogate cell transcriptomes. *Cold Spring Harb Protoc* 2011; **2011**(1): pdb.prot5559–9.
49. Leighton JC, Martins AL, Danko CG, *et al*. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet* 2014; **46**(12): 1311–20.
50. de Hoon M, Hayashizaki Y. Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. *BioTechniques* 2008; **44**(5): 627–32.
51. Valen E, Pascarella G, Chalk A, *et al*. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res* 2008; **19**(2): 255–65.
52. Carninci P, Sandelin A, Lenhard B, *et al*. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* 2006; **38**(6): 626–35.
53. Djebali S, Davis CA, Merkel A, *et al*. Landscape of transcription in human cells. *Nature* 2012; **489**(7414): 101–8.
54. Yamashita R, Wakaguri H, Sugano S, *et al*. DBTSS provides a tissue specific dynamic view of transcription start sites. *Nucleic Acids Res* 2009; **38**(suppl_1): D98–104.
55. Yu NY-L, Hallström BM, Fagerberg L, *et al*. Complementing tissue characterization by integrating transcriptome profiling from the human protein atlas and from the FANTOM5 consortium. *Nucleic Acids Res* 2015; **43**(14): 6787–98.
56. Roux J, Gonzàlez-Porta M, Robinson-Rechavi M. Comparative analysis of human and mouse expression data illuminates tissue-specific evolutionary patterns of miRNAs. *Nucleic Acids Res* 2012; **40**(13): 5890–900.
57. Olive V, Minella AC, Lin H. Outside the coding genome, mammalian microRNAs confer structural and functional complexity. *Sci Signal* 2015; **8**(368): re2–2.
58. Xu H, Chen L, Wang J, *et al*. Identifying cell-specific microRNA transcriptional start sites. *Bioinformatics* 2016; **32**(16): 2403–10.
59. Liu Q, Wang J, Zhao Y, *et al*. Identification of active miRNA promoters from nuclear run-on RNA sequencing. *Nucleic Acids Res* 2017; **45**(13): e121–1.
60. Bouvy-Liivrand M, de Sande AH, Pölönen P, *et al*. Analysis of primary microRNA loci from nascent transcriptomes reveals regulatory domains governed by chromatin architecture. *Nucleic Acids Res* 2017; **45**(17): 9837–49.
61. Georgakilas G, Vlachos IS, Zagganas K, *et al*. DIANA-miRGen v3.0: accurate characterization of microRNA promoters and their regulators. *Nucleic Acids Res* 2015; **44**(D1): D190–5.
62. Bhattacharyya M, Das M, Bandyopadhyay S. miRT: a database of validated transcription start sites of human MicroRNAs. *Genomics Proteomics Bioinformatics* 2012; **10**(5): 310–6.
63. Forrest AR, Kawaji H, Rehli M, *et al*. A promoter-level mammalian expression atlas. *Nature* 2014; **507**(7493): 462–70.
64. de Klerk E, 't Hoen PA. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends Genet* 2015; **31**(3): 128–139.
65. Batut P, Dobin A, Plessy C, *et al*. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res* 2012; **23**(1): 169–80.
66. de Rie D, Alam T, AbugessaisaI, *et al*. An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nat Biotechnol* **35**(9): 872–878.
67. Bhattacharya A, Cui Y. miR2go: comparative functional analysis for microRNAs: Fig. 1. *Bioinformatics* 2015; **31**(14): 2403–5.

68. Huang Z, Shi J, Gao Y, *et al*. HMDD v3.0: a database for experimentally supported human microRNA–disease associations. *Nucleic Acids Res* 2018; **47**(D1): D1013–7.
69. Rodriguez A. Identification of mammalian microRNA host genes and transcription units. *Genome Res* 2004; **14**(10a): 1902–10.
70. Baskerville S. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* 2005; **11**(3): 241–7.
71. Gupta S, Ross KE, Tudor CO, *et al*. miRiaD: a text mining tool for detecting associations of microRNAs with diseases. *J Biomed Semant* 2016; **7**(1).
72. Ramaswamy S, Tamayo P, Rifkin R, *et al*. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A* 2001; **98**(26): 15149–54.
73. Wang Z, Xu W, Liu Y. Integrating full spectrum of sequence features into predicting functional microRNA–mRNA interactions. *Bioinformatics* 2015; **31**(21): 3529–36.
74. Muniategui A, Nogales-Cadenas R, Vázquez M, *et al*. Quantification of miRNA–mRNA interactions. *PLoS ONE* 2012; **7**(2): e30766.
75. Jim CH, Morris QD, Frey BJ. Bayesian inference of MicroRNA targets from sequence and expression data. *J Comput Biol* 2007; **14**(5): 550–63.
76. He S, Hua S, Liu C, *et al*. MicroRNA-encoding long non-coding RNAs. *BMC Genom* 2008; **9**(1): 236.
77. Sun Q, Tripathi V, Yoon J-H, *et al*. MIR100 host gene-encoded lncRNAs regulate cell cycle by modulating the interaction between HuR and its target mRNAs. *Nucleic Acids Res* 2018; **46**(19): 10405–16.
78. Bao Z, Yang Z, Huang Z, *et al*. LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic Acids Res* 2018; **47**(D1): D1034–7.
79. Warnatz H-J, Querfurth R, Guerasimova A, *et al*. Functional analysis and identification of cis-regulatory elements of human chromosome 21 gene promoters. *Nucleic Acids Res* 2010; **38**(18): 6112–23.
80. Kim B, Jeong K, Kim VN. Genome-wide mapping of DROSHA cleavage sites on primary MicroRNAs and noncanonical substrates. *Mol Cell* 2017; **66**(2): 258–69.