Bioinformatics, 32(16), 2016, 2403–2410
doi: 10.1093/bioinformatics/btw171
Advance Access Publication Date: 19 April 2016
Original Paper

Genome analysis

Identifying cell-specific microRNA transcriptional start sites

Xu Hua[1,2], Luxiao Chen[1], Jin Wang[1,*], Jie Li[1,2,*] and Edgar Wingender[2,*]

[1]The State Key Laboratory of Pharmaceutical Biotechnology and Jiangsu Engineering Research Center for MicroRNA Biology and Biotechnology, NJU Advanced Institute for Life Sciences (NAILS), School of Life Science, Nanjing University, Nanjing 210093, China and [2]Department of Bioinformatics, Medical School, George August University of Göttingen, Goldschmidtstrasse 1, Göttingen D-37077, Germany

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on March 10, 2015; revised on February 9, 2016; accepted on March 27, 2016

Abstract

Motivation: Identification of microRNA (miRNA) transcriptional start sites (TSSs) is crucial to understand the transcriptional regulation of miRNA. As miRNA expression is highly cell specific, an automatic and systematic method that could identify miRNA TSSs accurately and cell specifically is in urgent requirement.

Results: A workflow to identify the TSSs of miRNAs was built by integrating the data of H3K4me3 and DNase I hypersensitive sites as well as combining the conservation level and sequence feature. By applying the workflow to the data for 54 cell lines from the ENCODE project, we successfully identified TSSs for 663 intragenic miRNAs and 620 intergenic miRNAs, which cover 84.2% (1283/1523) of all miRNAs recorded in miRBase 18. For these cell lines, we found 4042 alternative TSSs for intragenic miRNAs and 3186 alternative TSSs for intergenic miRNAs. Our method achieved a better performance than the previous non-cell-specific methods on miRNA TSSs. The cell-specific method developed by Georgakilas *et al.* gives 158 TSSs of higher accuracy in two cell lines, benefitting from the employment of deep-sequencing technique. In contrast, our method provided a much higher number of miRNA TSSs (7228) for a broader range of cell lines without the limitation of costly deep-sequencing data, thus being more applicable for various experimental cases. Analysis showed that upstream promoters at −2 kb to −200 bp of TSS are more conserved for independently transcribed miRNAs, while for miRNAs transcribed with host genes, their core promoters (−200 bp to 200 bp of TSS) are significantly conserved.

Availability and implementation: Predicted miRNA TSSs and promoters can be downloaded from supplementary files.

Contact: jwang@nju.edu.cn or jlee@nju.edu.cn or edgar.wingender@bioinf.med.uni-goettingen.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

MicroRNAs (miRNAs) are small (∼22 nt), non-coding RNAs that negatively regulates gene expression by binding to the 3’-untranslated region of mRNAs. Studies so far have revealed that miRNAs are involved in almost all the key biological processes, such as development, differentiation, apoptosis, cell proliferation and carcinogenesis (Gama-Carvalho *et al.*, 2014; He *et al.*, 2007; Hermeking, 2010; Lee and Dutta, 2007). To understand the underlying mechanism of the miRNA-mediated regulatory pathways, it is important to know where the miRNAs are produced from in the genome and which genes they are targeting. While the latter aspect has been intensely studied for both animals and plants system in the recent years (Betel *et al.*, 2010; Friedman *et al.*, 2009; Hsu *et al.*, 2014; Wang and El Naqa, 2008), the picture of transcription of miRNAs is far from complete, especially for the transcriptional start sites (TSSs) of miRNAs.

In general, miRNA genes are firstly transcribed into primary miRNAs (pri-miRNAs) by RNA polymerase (Pol) II or III in nucleus (Borchert *et al.*, 2006; Cai *et al.*, 2004; Lee *et al.*, 2004). Then, the RNase Drosha processes pri-miRNAs into precursor miRNAs (pre-miRNAs) (Han *et al.*, 2006; Lee *et al.*, 2003, 2006), which later are exported into the cytosol and turn to mature miRNAs with the help of Dicer (Lee *et al.*, 2002). The process of miRNA transcription is basically the same as that of protein-coding genes; however, both 5'- and 3'-ends of pri-miRNA are subject to quick degradation. The full-length transcript of a miRNA gene is very difficult to be retained (Chen and Shyu, 2011; Houseley *et al.*, 2006). Therefore, it is hard to accurately locate the TSS of miRNA.

The database miRT collected predicted miRNA TSSs from different sources (Bhattacharyya *et al.*, 2012) and proved helpful for experimental validation (Hurst *et al.*, 2014). The methods of the identification of miRNA TSSs available so far are mainly based on the cap analysis of gene expression (CAGE), RNA polymerase II (Pol II) data or histone modification data. For example, CAGE tags (Kodzius *et al.*, 2006) were used to identify miRNA TSSs by considering its possibility to capture the 5' cap (Chien *et al.*, 2011; Marsico *et al.*, 2013; Saini *et al.*, 2008). However, the rapid processing of the primary miRNA transcript by Drosha produces uncapped pre-miRNAs that cannot be captured by CAGE technology. Wang *et al.* (2010) predicted miRNA TSSs by modeling the RNA Pol II binding patterns. Corcoran *et al.* (2009) developed a method to predict the core promoters of miRNAs with Support Vector Machine models based on RNA Pol II data. More methods were developed by integrating the histone modification data. Ozsolak *et al.* (2008) identified miRNA TSSs by combining the information of RNA Pol II, H3K4me3 and H3K9/14A data. However, these studies were limited to small amount of miRNAs due to the insufficiency of Pol II data and ignored the case of Poly III transcription in one aspect. In further aspect, they discarded the cell-specific information of miRNA TSSs by pooling the Pol II data from experiments.

Annotating the cell-specific information of miRNA TSSs is important because it has been proved that miRNA is of higher tissue specificity than protein-coding genes (Kawaji *et al.*, 2009; Roux *et al.*, 2012). Most recently, Georgios *et al.* developed an algorithm 'microTSS' for integrating RNA-Sequencing data with H3K4me3, Pol II and DNase-seq data to identify the tissue-specific TSSs of intergenic miRNAs in high resolution (Georgakilas *et al.*, 2014). To overcome the effect of Drosha processing and capture the pri-miRNAs of low abundance, microTSS required high deep sequencing data which seems not easy to obtain for the large number of cell lines and various cellular environmental conditions. So far, miRNA TSSs of very few cell lines have been predicted in a cell-specific way. An automatic and systematic method that could identify miRNA TSSs accurately and cell specifically in a broad range of cell lines is still in urgent requirement.

Here, we developed a workflow to identify miRNA TSSs in 54 cell lines based on H3K4me3 and DNase I-hypersensitive sites (DHSs), which combined the features of species conservation and sequences. The proximal H3K4me3 region to each pre-miRNA in each cell line is considered as the potential region containing the TSS of the miRNA. DHSs are potential transcription factor binding sites, and their existence of them in the promoter region can assure the function of the identified TSSs. The accuracy of our predicted miRNA TSSs is verified in several ways, including comparing the results with the miRNA TSSs that are identified by previous studies, showing expressed sequence tag (EST) evidence for them and analyzing their global conservation. Known CAGE tags (Kawaji *et al.*, 2009, 2011) showed a higher signal at the TSSs that were identified using our approach; however, the signal levels were not comparable with those of protein-coding genes, which indicates the effect of the rapid degradation of the 5′ cap of miRNA transcript and proves the previous viewpoint that CAGE data is inadequate for the prediction of miRNA TSSs. Current models of miRNA biogenesis were also confirmed by our finding that the CAGE signal detected in nucleus at the identified miRNA TSSs was much higher than that detected in cytosol, whereas the nuclear and cytosolic CAGE signals for protein-coding genes are comparable. In addition, the analysis on the predicted TSSs showed the promoters upstream the independent TSSs of either intragenic or intergenic miRNAs are more conserved than that of the host gene TSSs by which some intragenic miRNAs are transcribed. In cases that miRNAs were broadly activated in various cell lines, alternative TSSs and promoters were used in different cell lines. Then they could be regulated in a promoter-specific manner.

2 Methods

2.1 Human miRNAs and Ensembl-annotated genes

A total of 1523 human miRNAs, including 821 intergenic miRNAs and 702 intragenic miRNAs, were collected from miRBase release 18 (Kozomara and Griffiths-Jones, 2011). In total, 44 762 Ensembl-annotated genes in hg19 were included in this research and were derived from Ensembl through BioMart.

2.2 Cell-specific H3K4me3 modifications

As known from previous reports, H3K4me3 modification marks transcription initiation sites in chromatin (Guenther *et al.*, 2007). Using these sites to identify miRNA TSSs can avoid the problem of rapid pri-miRNA processing. We collected data for 54 cell lines from the UCSC track 'UW Histone'. These data are a part of the ENCODE project and were generated by ChIP-seq high-throughput sequencing. Overall, we obtained 401 058 peak tags and merged those tags that were less than 250 bp apart. The tags were merged based on the intrinsic feature of the data by observing that the interval distance between tags was most likely to be smaller than 250 bp, and this distance was also similar to the length of one nucleosome. So tags less than 250 bp apart may be from the same nucleosome.

The DNA sequences of these tags, including flanking regions from 200 bp upstream to 100 bp downstream, were downloaded from the human genome (GRCh37/hg19) and were considered to be regions that comprised real TSSs. The flanking regions allowed us to predict in a broader peak but did not affect our result very much.

2.3 Conservation score

The conservation score of each genomic position was scored from 0 to 1 based on the data derived from UCSC track 'Mammal Cons (phastCons46wayPlacental)'. This track shows multiple alignments across 46 placental mammalian species based on the phastCons method (Pollard *et al.*, 2010; Siepel *et al.*, 2005).

2.4 DNase I-hypersensitive sites

A DHS is a chromatin-accessible region in the genome that marks regulatory regions of DNA. We collected data for 54 cell lines from the UCSC track 'UW Histone'. On average, 98.5% of ENCODE ChIP-seq regions overlapped with DHSs (Bernstein *et al.*, 2012). We used these data to ensure the function of miRNA TSSs, resulting in more reliable annotation of TSSs.

2.5 Determination of miRNA TSS

To combine different sources of data, the following formula was introduced to determine miRNA TSS:

$$S_i = \mathrm{H3K4me3}_i * \mathrm{DHS}_i * \max(\mathrm{CONS}(X_i) + \mathrm{EPO}(X_i)) \qquad (1)$$

This formula is a generalization of our workflow for predicting the candidate TSSs. X_i denotes a predicted site for a miRNA in a cell line. $\mathrm{H3K4me3}_i = 1$, if X_i lies in the nearest H3K4me3 region or 0 otherwise. H3K4me3 is a good marker for TSSs of protein coding genes, so $\mathrm{H3K4me3}_i$ ensures the satisfied X_i must locate in an H3K4me3 region. $\mathrm{DHS}_i = 1$, if there exists any DHS on the upstream of X_i, or 0 otherwise. DHS_i ensures the satisfied X_i must have some potential regulatory elements in its upstream sequences. CONS denotes the conservation score of X_i; EPO denotes the Eponine (Down and Hubbard, 2002) score of $X_{i.}$ Conservation is closely related with functional importance and Eponine score is an evaluation for TSS sequence feature. By selecting the X_i of $\max(S_i)$, we integrated all these features related with TSS together and identified the candidate TSS for a miRNA. The identified TSS should meet $S_i > 0$.

2.6 Collection of validated TSSs

We searched the literature to collect TSSs of miRNAs that were experimentally validated as benchmark data. These TSSs were verified using different methods, such as retrieving complete reverse transcription polymerase chain reaction transcripts by inhibiting Drosha (Cai *et al.*, 2004; Chang *et al.*, 2007; Chien *et al.*, 2011; Fukao *et al.*, 2007; Ribas *et al.*, 2012; Taganov *et al.*, 2006), validating the activity of promoters by a luciferase reporter system (Cai *et al.*, 2004; Chang *et al.*, 2007; Chien *et al.*, 2011; Fukao *et al.*, 2007; Lee *et al.*, 2004; Ozsolak *et al.*, 2008) or annotating a transcript as it was transcribed from its host gene (Kluiver *et al.*, 2005). Genomic coordinates of the validated TSSs in the human genome were collected from the literature, and then all coordinates were converted to hg19 by 'liftover' tools available on UCSC genome browser (Supplementary File S1). Ultimately, this benchmark data contained 17 validated TSSs of 25 miRNAs.

2.7 Cell-specific validated TSSs in human embryonic stem cell

Cell-specific validated TSSs in human embryonic stem cell (hESC) have been derived from microTSS (Georgakilas *et al.*, 2014). Totally, validated TSSs of 67 miRNAs in miRBase 18 were listed in Supplementary Table S2 in Supplementary File S2 and all the coordinates were based on hg19 genome assembly.

2.8 EST data

An EST is a short sub-sequence of a gene transcript, thus it represents a portion of an expressed gene. If the predicted primary miRNAs were indeed transcribed, there should be ESTs derived from them. In this study, all human ESTs in GeneBank (Benson *et al.*, 2013) were downloaded from UCSC, including 8 685 808 records.

3. Results

3.1 Workflow chart for the cell-specific identification of miRNA TSSs

Here, we developed a workflow (see Supplementary Fig. S1 in Supplementary File S2) to identify the miRNA TSSs in a cell-specific way. First, we selected those H3K4me3 regions that were proximal to each pre-miRNA in each cell line. For each H3K4me3 region, we considered the score calculated by Eponine (Down and Hubbard, 2002), which can evaluate the similarity of the sequence signal to the environment of a TSS and the conservation score. By employing Eponine, we obtained several predicted sites with an Eponine score in one region. To consider the conservation, we used the phastCons score to calculate the conservation of each site that was predicted by Eponine. To combine the sequence features with conservation, we sorted the Eponine and phastCons scores in descending order and then selected those sites with the minimal sum of these two ranks as putative TSSs. To assure the function of these putative TSSs, we also checked the regions 1 kb upstream of the miRNA TSSs using cell-specific DHS data. The promoter of 1 kb region was selected in a very strict way, because most transcription factor binding sites are within 1 kb region (Hurst *et al.*, 2014). Generalized as Equation (1), we integrated H3K4me3 and DHS data with the score of conservation and promoter sequence features and obtained candidate TSSs.

We subsequently dealt with these candidate TSSs differently based on the genomic context of the respective miRNA. For the candidate TSSs of intragenic miRNAs, the TSS of its host gene was analyzed. If the distance between the candidate TSS and the host-gene TSS is small enough (i.e. less than a certain threshold), it is considered that the candidate TSS and the host-gene TSS are identical, and the host gene TSS is taken as the candidate TSS. Here the threshold is set as 150 bp, because the general length of H3K4me3 peaks in our data is ~150 bp and the TSSs within a same peak region could be identical one. In the case that there was no candidate TSSs close to the host gene TSS, we checked whether there were candidate TSSs located between the host gene TSS and the pre-miRNA. If it was, the candidate TSS was selected as the independent TSS of the miRNA; if not, no TSS was predicted for the miRNA. For the candidate TSS of an intergenic miRNA, we checked whether any Ensembl-annotated genes were present between the candidate TSS and pre-miRNA. If there was an Ensembl-annotated gene, the candidate TSS might be the TSS of another transcript and was therefore discarded.

After these two processes, our workflow yielded the TSSs of miRNAs with 1 bp resolution in a cell-specific manner. Therefore, for each miRNA in specific cell line, there are three possible types of TSSs, i.e. the independent TSS for intergenic miRNA, the independent for intragenic miRNA and the host-gene TSS for intragenic miRNA.

3.2 Evaluation of identified TSSs

3.2.1 Comparisons with collection of validated TSSs

We collected 25 experimentally verified TSSs (see Supplementary Table S1 in Supplementary File S2) as benchmark data and compared our prediction with other predictions. In the case there were multiple predictions for one miRNA, we selected the best prediction which is the closest one to the validated TSS. If we consider predictions located closer than 1 kb from the validated TSSs as true positives (TPs) and the others as false positives, then we can compare the sensitivity and precision among different methods. Sensitivity is defined as TPs/(all validated TSSs), and precision is defined as TPs/(TPs + false positives). Since this comparison is non-cell specific, we excluded the cell-specific method of Georgakilas *et al.* (2014). Our method achieved 84%/91.3% in sensitivity and precision. Marsico *et al.* (2013) achieved 88%/88%, Chien *et al.* (2011) achieved 28%/46.7% and Marson *et al.* (2008) achieved 36%/39.1% in sensitivity and precision. Apparently, our method performs better than Chien's and Marson's work. However, comparing with Marsico's method, our approach was superior with regard to precision, whereas Marsico *et al.* (2013) performed better in sensitivity.

3.2.2 Comparison with validated TSSs in hESC

To evaluate the accuracy of predicted cell-specific TSSs, we used 67 experimentally validated TSSs in hESC (Georgakilas *et al.*, 2014) as benchmark data and compared our predictions with other methods. Since PROmiRNA (Marsico *et al.*, 2013) provide multiple predictions for one miRNA, we selected the prediction with the highest score and took it into comparison. As shown in Supplementary Table S2 in Supplementary File S2, we highlighted the correct predictions of five works which are closer than 1 kb from validated TSSs.

By considering predictions located closer than 1 kb from the validated TSSs as TPs, we also calculated the sensitivity and precision for each method. Our method achieved 55.2%/77.1% in sensitivity and precision. Georgakilas *et al.* (2014) achieved 94.0%/97.0%, Marsico *et al.* (2013) achieved 52.2%/53.8%, Chien *et al.* (2011) achieved 13.4%/31.0% and Marson *et al.* (2008) achieved 16.4%/40.7% in sensitivity and precision. For Marsico *et al.*, Chien *et al.* and Marson *et al.*, these non-cell-specific methods showed their disadvantage in performance and were incapable of predicting cell-specific TSSs. Our method showed a better performance than those non-cell-specific methods in both sensitivity and precision. Georgakilas *et al.* (2014) outperforms all the other methods in this cell-specific comparison. The main reason is that they integrated the high-quality RNA deep-sequencing data with DNase-Seq/Digital Genomic Footprinting (DGF), H3K4me3 and Pol II ChIP-Seq data. The high-quality sequencing data provide a 10 times (~200 million reads) higher resolution than the normal RNA sequencing data (~20 million reads), which apparently benefits the accuracy of Georgakilas's prediction a lot. However, such deep-sequencing data are costly and rare at present, which limits the application of Georgakilas's method. Our method, without requiring such expensive sequencing data, can give more resourceful miRNA TSSs for a broad range of cell lines. The sensitivity of our prediction is relatively low (55.2%), but the precision is acceptable (77.1%).

3.2.3 EST support for identified pri-miRNAs and TSSs

As shown in previous comparisons, the number of experimentally validated TSSs is quite limited. To evaluate predictions in a general way, ESTs were used as support. First, we mapped ESTs to the predicted pri-miRNA genes, which were defined as the genomic regions from an identified TSS to the end of a pre-miRNA. We calculated the number of ESTs which could overlap with identified pri-miRNAs. It was found that 133/221 ESTs were averagely mapped on each intergenic/intragenic miRNA. In the comparison with other methods (Table 1), all the results showed the same order of magnitude for the number of ESTs on each pri-miRNA and the intragenic miRNAs got more ESTs than the intergenic ones as expected. The distribution of EST coverage for pri-miRNAs (see Supplementary Fig. S2a in Supplementary File S2) also showed that most pri-miRNAs had a high EST coverage. The considerable number of overlapping ESTs supported the existence of the pri-miRNAs originating from the thus identified TSSs.

Theoretically, only the sequences downstream to TSSs could be detected as ESTs. Therefore, the differences of EST coverage depth between upstream and downstream to predicted TSSs were checked as an evaluation of biological authenticity for predictions. By averaging the EST coverage depth around all identified TSSs in each work, both the results of Georgakilas *et al.* (2014), Marsico *et al.* (2013) and of our work showed obvious differences of EST coverage depth in the TSS neighboring region, which supports the authenticity of these TSS predictions and the difference of our result was larger than the other methods (see Supplementary Fig. S3 in Supplementary File S2). To further evaluate the accuracy of each method, for each predicted TSS we made a *T*-test for EST coverage depth between upstream and downstream 5 kb regions. The TSS with significant (P value < 0.05) difference was considered as authentic. Our method achieved 74.2% (5365/7228), Georgakilas *et al.* achieved 77.2% (122/158) and Marsico *et al.* achieved 55.0% (4956/9018) EST supported TSSs (details can be seen in Supplementary Table S3 of Supplementary Files S2 and File S3). In comparison, Georgakilas *et al.* got the highest accuracy and our method yielded the highest number of EST supported TSSs with accuracy comparable to Georgakilas's method.

3.2.4 Global feature of conservation

Although the conservation information in a limited region has been considered to predict TSSs, the conservation feature in a larger region was still worth to observe (Fig. 1). There was a clear peak of the conservation score at the location of the identified TSSs for all three types of TSS, which confirms the functional importance of TSSs. The conservation level of the regions upstream of the predicted miRNA TSSs was also high, which might indicate the existence of orthologous promoters. Notably, such a highly conserved region of independently transcribed miRNA genes (blue and red curves in Fig. 1) is much more extensive than that for the host genes of intragenic miRNAs (black curve in Fig. 1), suggesting that the promoter regions of the independently transcribed miRNA genes were longer than those of protein-coding genes. We split the TSS upstream regions into windows of 200 bp and made *T*-tests for the mean conservation of each window between different types of TSSs. We found that for all the windows in the region of [−2000 bp, −200 bp] upstream of TSS, the mean conversation for the

Table 1. Number of ESTs mapped on each intergenic and intragenic pri-miRNAs with EST evidence from different studies

Studies	Mean (median) number of ESTs for each intergenic pri-miRNA	Number of intergenic pri-miRNA	Mean (median) number of ESTs for each intragenic pri-miRNA	Number of intragenic pri-miRNA
This study	133 (19)	3186	213 (101)	4042
Georgakilas *et al.* (2014)	93 (14)	125	146 (56)	33
Marsico *et al.* (2013)	198 (14)	4187	325 (85)	4831
Chien *et al.* (2011)	114 (21)	363	43 (45)	15
Marson *et al.* (2008)	131 (51)	245	293 (96)	245

independent TSS of either intergenic miRNA or intragenic miRNA is significantly higher than that for the host-gene TSS of intragenic miRNA (see Supplementary Table S4 in Supplementary File S2). The observation that the promoters of intergenic miRNAs were more conserved than those of the protein-coding genes was previously made (Mahony *et al.*, 2007). Here, our result suggests that it is the type of miRNA transcription (independent or host-gene-dependent transcription), instead of miRNA type (intergenic or intragenic miRNA), that matters the conservation of promoters. However, at the core promoter region which is [−200 bp, 200 bp] around TSS, the conservation for host-gene TSS is significantly higher than that for independent TSS (see Supplementary Table S4 in Supplementary File S2). This observation showed another different pattern of conservation between the promoters of host-gene-dependently and independently transcribed miRNA genes. A random creation model proposed that hairpin structures are generally abundant in eukaryotic genomes, so the creation of a new miRNA is somehow involved with its appropriate transcription (Chen and Rajewsky, 2007). Therefore, the different pattern of conservation on the core promoters and upstream promoters may indicate a different evolution process for host-gene-dependently and independently transcribed miRNA genes.

For the regions downstream of the TSSs, the conservation of miRNA host genes (black curve) decreases much more slowly than those of independently transcribed miRNA genes (blue and red curve), because the coding regions of host genes are much more conserved than intergenic or intronic regions (Siepel *et al.*, 2005). In conclusion, the observed conservation distribution further supports the reliability of the TSSs that we identified by this workflow.

3.3 Statistics of identified TSSs

We successfully identified TSSs for 663 intragenic miRNAs and 620 intergenic miRNAs, which cover 84.2% (1283/1523) of all miRNAs in miRBase 18. From the data for 54 different cell lines, we identified 4042 alternative TSSs (including 3758 independent intragenic TSSs and 284 host TSSs) for intragenic miRNAs and 3186 alternative TSSs for intergenic miRNAs, which correspond to 6.1 alternative TSSs per intragenic miRNA and 5.1 alternative TSSs per intergenic miRNA (Supplementary File S4). After obtaining the cell-specific TSSs, we constructed a global set of promoters for miRNAs (Supplementary File S5), which can be helpful for obtaining a general picture of the transcriptional regulation of miRNA genes and to construct a transcriptional regulatory network. For each identified TSS, we considered the region 1 kb upstream of the TSS as the promoter of each miRNA. Then, the alternative promoters of one miRNA were merged if they overlapped by at least 1 bp. In this manner, we obtained 2540 promoters for 1283 miRNAs, which corresponded to 2.0 alternative promoters for each miRNA.

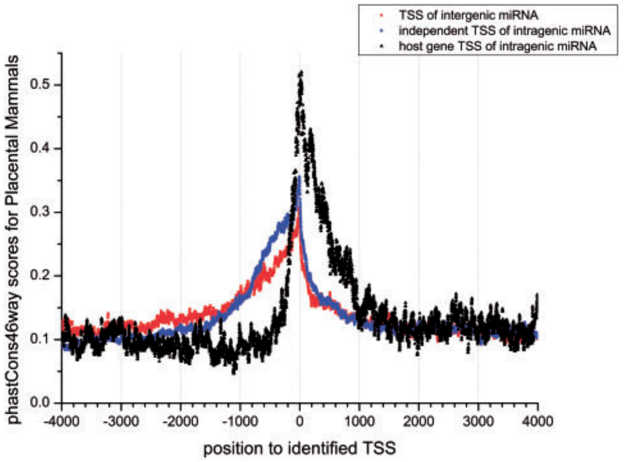

Fig. 1. Conservation features around the predicted TSSs of miRNA genes. Blue/black curves represent intragenic miRNAs in cases of independent or concomitant transcription with its host gene, respectively; the red curve represents intergenic miRNA genes (Color version of this figure is available at *Bioinformatics* online.)

3.4 Distance from TSS to pre-miRNA

As reported previously, the distance between miRNA TSSs and the encoded respective pre-miRNAs can range from a few hundred bases to hundreds of thousands bases (Bhattacharyya *et al.*, 2012). The distance is very much longer than that of protein-coding genes, of which the average *5′-untranslated region* length of human genes is 210 bp, with a maximum of 2800 bp (Mignone *et al.*, 2002). Such extremely long 'non-coding' sequences upstream to the pre-miRNA seem to be unique for miRNAs. For intragenic miRNAs (see Supplementary Fig. S4a in Supplementary File S2), in this study, the highest percentage (28.5%) of them had a predicted TSS within 5 kb upstream of the pre-miRNA and a small percentage (6.4%) of them had a predicted TSS at a distance greater than 200 kb. For intergenic miRNAs (see Supplementary Fig. S4b in Supplementary File S2), the highest percentage (35.2%) of TSSs were also located within 5 kb upstream of the pre-miRNA and showed a long-tail distribution. Such long-tail distribution of distance suggested that although most TSSs are located close to their pre-miRNAs, a significant portion of TSSs are dispersed widely in distant regions.

To further compare the distances from the TSSs to the pre-miRNAs in our prediction with previous predictions, we selected those miRNAs that were common in all studies. Even though our method has not imposed an artificial upper limit for the TSS - pre-miRNA distance, most of our predicted TSSs were within 20 kb range as the cases in the other predictions (Fig. 2), supporting again the solidity of our predictions.

3.5 Long non-coding RNA and pri-miRNA

Recent studies revealed that the human genome encodes many long non-coding RNA (lncRNAs), and these non-coding transcription

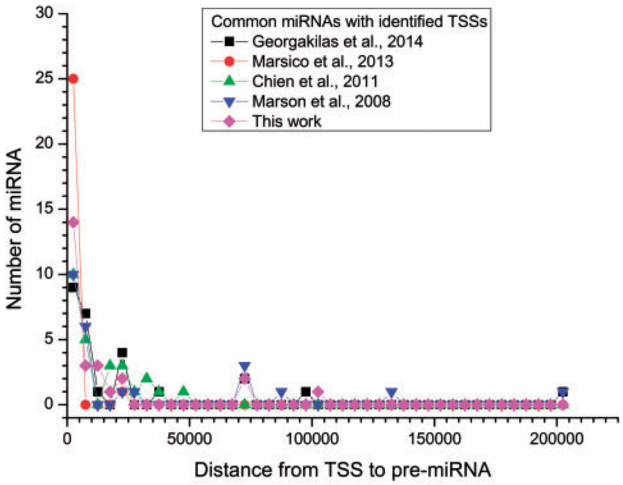

Fig. 2. Distribution of the distance from TSS to the pre-miRNA of common miRNAs. Twenty-six common miRNAs from this and other studies (Chien *et al.*, 2011; Georgakilas *et al.*, 2014; Marsico *et al.*, 2013; Marson *et al.*, 2008) were involved in the plot

units might also harbor miRNA genes (Rodriguez *et al.*, 2004). To check for potential overlaps of the pri-miRNAs in our prediction with other non-coding transcripts, we mapped them to known lncRNAs in Ensemble (GRCh37.p13/hg19) (Flicek *et al.*, 2013). We found that a considerable portion (23.4%) of intergenic pri-miRNAs overlap with lncRNAs, which is significantly higher than the percentage of intragenic pri-miRNAs whenever it is transcribed independently (12.0%) or host-gene-dependently (9.9%). This indicates that some intergenic pri-miRNAs significantly overlap, or are even identical, with known lncRNAs, which might be a partial explanation for their extended 5′-leader sequences.

3.6 CAGE tags around predicted TSSs of expressed miRNA indicate a decrease of 5′ cap in cytosol

We used the 'CSHL Small RNA-seq' dataset (GSE24565) (Djebali *et al.*, 2012) to select the expressed miRNAs. We checked the miRNAs expressed in the cell line A549, for which we considered pre-miRNAs with an reads per million (RPM) value > 0 to be expressed.

CAGE experiments were designed to capture the 5′-cap of transcripts and the first ~27 bp of its cDNA were usually sequenced. Pri-miRNA is initially transcribed as a long transcript with a 5′-cap and a polyA tail and then loses most of its 5′-leader sequence, including the cap, before being transported into the cytosol. In this study, the CAGE signal dataset (GSE34448) (Djebali *et al.*, 2012) was derived from UCSC track 'RIKEN CAGE Loc Track'. Two kinds of CAGE signals are included. The CAGE signals that are detected in nucleus is termed as the nuclear CAGE signals, whereas the signals detected in cytosol are the cytosolic CAGE signals. We mapped the nuclear and cytosolic CAGE signals separately to our predicted TSSs of the miRNAs expressed in A549 cells.

For the host-gene TSSs, we observed a strong CAGE signal at the TSS location and its nearby downstream region (Fig. 3a). This is consistent with the finding that protein-coding genes are usually highly expressed and keep their 5′ cap intact. For the TSSs of intergenic miRNAs, we noticed that the absolute level of the CAGE signal was much lower than that for host-gene TSSs (Fig. 3b). This might be resulted from a lower expression level of miRNAs and was certainly influenced by the rapid 5′-processing of pri-miRNAs. Nevertheless, we observed a clear CAGE signal at the predicted TSSs and in the associated downstream regions. This signal was stronger in the nucleus than in the cytosol where most processed pre-miRNAs were localized. For the independent TSSs of intragenic miRNAs, they showed a signal distribution similar to that of intergenic miRNAs in two aspects: (i) an absolute lower level of signal when compared with host genes and (ii) a stronger signal in the nucleus than in the cytosol (Fig. 3c). To further evaluate the difference of mean CAGE signal at the predicted TSSs between nucleus and cytosol, we made paired sample Wilcoxon signed rank tests for the three types of TSSs. The mean value of CAGE signal 'peak' is calculated for each miRNA (Supplementary File S8), and the 'peak' is the region [−10 bp, 50 bp] around TSS as shown in Figure 3. We found that for the independent TSSs of both intragenic and intergenic miRNAs, their mean CAGE signals at TSSs are significantly (*P* value = 0.007 and 9.490×10^{-5}) higher in nucleus than cytosol, while the host-gene TSSs are not (*P* value = 0.563). This observation indicates a decrease of 5' cap for independently transcribed miRNAs in cytosol.

In conclusion, clear CAGE signals could be observed near the predicted TSSs, providing further support for the viability of our TSS predictions. Additionally, the signal differences between the nucleus and cytosol confirmed the current model of miRNA biogenesis.

4. Discussion

In this study, we developed a method for miRNA TSS prediction by integrating the information of experimental data of H3K4me3 and DHS, the species conservation and sequence features. Compared with other methods that are based on histone modification (Marson *et al.*, 2008; Ozsolak *et al.*, 2008; Vrba *et al.*, 2011), we used DHSs data to assure the function of promoters upstream of the predicted TSSs. In addition, instead of pooling data of different cell lines (Chien *et al.*, 2011; Marsico *et al.*, 2013; Vrba *et al.*, 2011), the information from 54 cell lines was processed separately in our workflow. This proves useful for conducting experiments and analyzing the cell specificity of miRNA TSSs. To evaluate the performance of our method, a set of experimentally verified miRNA TSSs were collected from literature as benchmark data and our method performs best comparing with other predictions. To evaluate the performance of our method in a cell-specific way, a set of cell-specific validated miRNA TSSs in hESC cells were used to make an independent comparison. Our method performed better than those non-cell-specific methods but worse than microTSS. However, the input data for our method were more easily acquirable than microTSS, which enabled us to make predictions in a broader range of 54 cell lines with an acceptable precision (77.1%).

In our method, we set a distance threshold to discriminate a candidate TSS from its nearby annotated host TSS. Only the candidate TSS that is more than this threshold distance away from its nearby host TSS was considered as independent TSS. We set this distance threshold to 150 bp based on the length of an H3K4me3 peak, assuming that there is only one TSS in one H3K4me3 peak region.

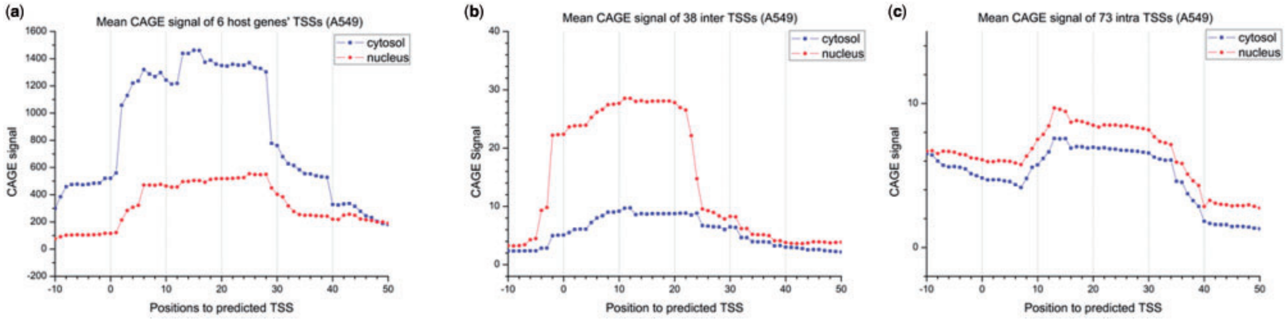

Fig. 3 The CAGE signal distribution around the predicted TSSs of pri-miRNAs in the A549 cell line. The blue curve represents the signal in the cytosol and the red curve represents the signal in the nucleus. The predicted TSS is located at '0' on the *x* axis. (**a**) Mean CAGE signal around host gene TSSs. (**b**) Mean CAGE signal around intergenic TSSs. (**c**) Mean CAGE signal around intragenic TSSs (Color version of this figure is available at *Bioinformatics* online.)

The influence of this threshold on the prediction was checked. The number of intergenic and intragenic miRNAs stays at about 620 and 663, respectively, as the variation of threshold (Supplementary Fig. S5 in Supplementary File S2). For the number of TSSs, on average, every intergenic/intragenic miRNA loses about 0.0014/0.0026 TSSs as the threshold increases by 1 bp (Supplementary Fig. S6 in Supplementary File S2). The accuracy evaluated by experimentally validated TSSs is not influenced by the change. Therefore, the influence of changing the threshold to TSS prediction is limited.

It has always been a dispute whether CAGE data are able to efficiently capture the miRNA TSS because of the transient nature of miRNA primary transcripts. To clarify this issue, we analyzed the CAGE signal at predicted miRNA TSS and found the nuclear CAGE signal was much higher than the cytosolic CAGE signal at TSS location. It is evident that the CAGE signal at miRNA TSS would be reduced if the nuclear and cytosolic CAGE signals were pooled together. To deal with this problem, a deep-sequencing approach was employed to increase the CAGE signal at miRNA TSS by increasing the total count of reads (Georgakilas *et al.*, 2014; Marsico *et al.*, 2013). However, the deep sequencing data may also increase the 'noise' signal in the background region and promoter region of protein-coding gene. To properly and efficiently use CAGE data, maintaining the full miRNA primary transcript is necessary. Such as in *Caenorhabditis elegans* (Saito *et al.*, 2013), full transcripts could be maintained when RNAs were isolated from nucleolus. Alternatively, inhibition of Drosha could help to maintain full transcripts for deep sequencing (Georgakilas *et al.*, 2014). However, this kind of data are very limited at present time.

In our result, we noticed that some of the miRNAs had quite long primary transcripts. For example, hsa-mir-548as is embedded in its host gene, *GPC5*, and its host transcript is over 1000 kb. The long pri-miRNA transcript might provide the possibility of more alternative TSSs.

On the basis of the analysis of species conservation of miRNA promoter, we found that for the upstream promoters of TSS which were regions of [−2000 bp, −200 bp], the mean conservation in the case that miRNAs are independently transcribed was significantly higher than that in the case that miRNAs are co-transcribed with the host genes. However, at the core promoter region which is [−200 bp, 200 bp], the conservation behaves in opposite. This observation showed a different pattern of conservation between the promoters of host-gene-dependently and independently transcribed miRNA genes. Mahony *et al.* (2007) found the promoters of intergenic miRNAs were even more conserved than those of the intragenic miRNAs that are co-transcribed with the host genes in opossum. Godnic *et al.* (2013) found only 27 conserved host-intragenic miRNA pairs from 849 intragenic miRNAs in humans, indicating the poor conservation for the promoters of host-genes-dependent miRNAs. Our result is consistent with these reports and draws a better explanation for these observations. The different pattern of conservation on the core promoters and upstream promoters may indicate a different evolution process for host gene-dependently and independently transcribed miRNA genes.

MiRNAs are mainly expressed in a cell-specific manner (Lagos-Quintana *et al.*, 2002; Landgraf *et al.*, 2007; Liang *et al.*, 2007), which could originate from the use of selectively activated TSSs and specific promoters. However, it should be noticed that there is no unambiguous correspondence between expression of a gene and its active TSS, neither for protein coding nor for miRNA genes. As reported for protein-coding genes, many genes do not produce full length transcripts, nonetheless experience transcript initiation (Guenther *et al.*, 2007). As reported from microTSS, for 118 predicted TSSs, 57 miRNAs were expressed in mESC. Similarly in our result, for 653 predicted TSSs, 146 miRNAs were expressed in A549 (GSE24565). So current methods could predict the possibly active TSSs, but there still needs more efforts to fill in the gap between the transcription and expression of miRNAs. Anyway, to know the active transcription start site is the first step.

In conclusion, we predicted the miRNA TSSs for 54 cell lines and several observations were made about the cell line specificity of different types of TSSs, the conservation of miRNA promoters and the long non-coding RNAs related to pri-miRNAs. These results from this research could benefit the understanding of the specific expression of miRNAs and their exact regulation network in different cells.

Acknowledgement

The authors would acknowledge the Center of High Performance Computation of Nanjing University for the support of computational resources.

Funding

This work was supported by China Scholarship Council, DZHK (German Center for Cardiovascular Research), National Science Foundation of China (No. 81250044 and No. 31500674) and NJU-Yangzhou Institute of Optoelectronics.

Conflict of Interest: none declared.

References

Benson,D.A. *et al.* (2013) GenBank. *Nucleic Acids Res.*, **41**, D36–D42.

Bernstein,B.E. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Betel,D. *et al.* (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.*, **11**, R90.

Bhattacharyya,M. *et al.* (2012) miRT: a database of validated transcription start sites of human microRNAs. *Genomics Proteomics Bioinformatics*, **10**, 310–316.

Borchert,G.M. *et al.* (2006) RNA polymerase III transcribes human microRNAs. *Nat. Struct. Mol. Biol.*, **13**, 1097–1101.

Cai,X. *et al.* (2004) Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA*, **10**, 1957–1966.

Chang,T.C. *et al.* (2007) Transactivation of miR-34a by p53 broadly influences gene expression and promotes apoptosis. *Mol. Cell*, **26**, 745–752.

Chen,C.Y. and Shyu,A.B. (2011) Mechanisms of deadenylation-dependent decay. *Wiley Interdiscip. Rev. RNA*, **2**, 167–183.

Chen,K. and Rajewsky,N. (2007) The evolution of gene regulation by transcription factors and microRNAs. *Nat. Rev. Genet.*, **8**, 93–103.

Chien,C.H. *et al.* (2011) Identifying transcriptional start sites of human microRNAs based on high-throughput sequencing data. *Nucleic Acids Res.*, **39**, 9345–9356.

Corcoran,D.L. *et al.* (2009) Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data. *PLoS One*, **4**, e5279.

Djebali,S. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.

Down,T.A. and Hubbard,T.J. (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.*, **12**, 458-461.

Flicek,P. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–D55.

Friedman,R.C. *et al.* (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.

Fukao,T. *et al.* (2007) An evolutionarily conserved mechanism for microRNA-223 expression revealed by microRNA gene profiling. *Cell*, **129**, 617–631.

Gama-Carvalho,M. *et al.* (2014) Regulation of cardiac cell fate by microRNAs: implications for heart regeneration. *Cells*, **3**, 996–1026.

Georgakilas,G. *et al.* (2014) microTSS: accurate microRNA transcription start site identification reveals a significant number of divergent pri-miRNAs. *Nat. Commun.*, **5**, 5700.

Godnic,I. *et al.* (2013) Genome-wide and species-wide in silico screening for intragenic MicroRNAs in human, mouse and chicken. *PLoS One*, **8**, e65165.

Guenther,M.G. *et al.* (2007) A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, **130**, 77–88.

Han,J. *et al.* (2006) Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell*, **125**, 887–901.

He,L. *et al.* (2007) A microRNA component of the p53 tumour suppressor network. *Nature*, **447**, 1130–1134.

Hermeking,H. (2010) The miR-34 family in cancer and apoptosis. *Cell Death Differ.*, **17**, 193–199.

Houseley,J. *et al.* (2006) RNA-quality control by the exosome. *Nat. Rev. Mol. Cell Biol.*, **7**, 529–539.

Hsu,S.D. *et al.* (2014) miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res.*, **42**, D78–D85.

Hurst,L.D. *et al.* (2014) A simple metric of promoter architecture robustly predicts expression breadth of human genes suggesting that most transcription factors are positive regulators. *Genome Biol.*, **15**, 413.

Kawaji,H. *et al.* (2009) The FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. *Genome Biol.*, **10**, R40.

Kawaji,H. *et al.* (2011) Update of the FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. *Nucleic Acids Res.*, **39**, D856–D860.

Kluiver,J. *et al.* (2005) BIC and miR-155 are highly expressed in Hodgkin, primary mediastinal and diffuse large B cell lymphomas. *J. Pathol.*, **207**, 243–249.

Kodzius,R. *et al.* (2006) CAGE: cap analysis of gene expression. *Nat. Methods*, **3**, 211–222.

Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.

Lagos-Quintana,M. *et al.* (2002) Identification of tissue-specific microRNAs from mouse. *Curr. Biol.*, **12**, 735–739.

Landgraf,P. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401–1414.

Lee,Y. *et al.* (2002) MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J.*, **21**, 4663–4670.

Lee,Y. *et al.* (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature*, **425**, 415–419.

Lee,Y. *et al.* (2004) MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.*, **23**, 4051–4060.

Lee,Y. *et al.* (2006) Drosha in primary microRNA processing. *Cold Spring Harb. Symp. Quant. Biol.*, **71**, 51–57.

Lee,Y.S. and Dutta,A. (2007) The tumor suppressor microRNA let-7 represses the HMGA2 oncogene. *Genes Dev.*, **21**, 1025–1030.

Liang,Y. *et al.* (2007) Characterization of microRNA expression profiles in normal human tissues. *BMC Genomics*, **8**, 166.

Mahony,S. *et al.* (2007) Regulatory conservation of protein coding and microRNA genes in vertebrates: lessons from the opossum genome. *Genome Biol.*, **8**, R84.

Marsico,A. *et al.* (2013) PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs. *Genome Biol.*, **14**, R84.

Marson,A. *et al.* (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, **134**, 521–533.

Mignone,F. *et al.* (2002) Untranslated regions of mRNAs. *Genome Biol.*, **3**, REVIEWS0004.

Ozsolak,F. *et al.* (2008) Chromatin structure analyses identify miRNA promoters. *Genes Dev.*, **22**, 3172–3183.

Pollard,K.S. *et al.* (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.

Ribas,J. *et al.* (2012) A novel source for miR-21 expression through the alternative polyadenylation of VMP1 gene transcripts. *Nucleic Acids Res.*, **40**, 6821–6833.

Rodriguez,A. *et al.* (2004) Identification of mammalian microRNA host genes and transcription units. *Genome Res.*, **14**, 1902–1910.

Roux,J. *et al.* (2012) Comparative analysis of human and mouse expression data illuminates tissue-specific evolutionary patterns of miRNAs. *Nucleic Acids Res.*, **40**, 5890–5900.

Saini,H.K. *et al.* (2008) Annotation of mammalian primary microRNAs. *BMC Genomics*, **9**, 564.

Saito,T.L. *et al.* (2013) The transcription start site landscape of C. elegans. *Genome Res.*, **23**, 1348–1361.

Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

Taganov,K.D. *et al.* (2006) NF-kappaB-dependent induction of microRNA miR-146, an inhibitor targeted to signaling proteins of innate immune responses. *Proc. Natl. Acad. Sci. USA*, **103**, 12481–12486.

Vrba,L. *et al.* (2011) Epigenetic regulation of normal human mammary cell type-specific miRNAs. *Genome Res.*, **21**, 2026–2037.

Wang,G. *et al.* (2010) RNA polymerase II binding patterns reveal genomic regions involved in microRNA gene regulation. *PLoS One*, **5**, e13798.

Wang,X. and El Naqa,I.M. (2008) Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics*, **24**, 325–332.