Computational Modeling of miRNA Biogenesis

Brian Caffrey and Annalisa Marsico

Abstract Over the past few years it has been observed, thanks in no small part to high-throughput methods, that a large proportion of the human genome is transcribed in a tissue- and time-specific manner. Most of the detected transcripts are non-coding RNAs and their functional consequences are not yet fully understood. Among the different classes of non-coding transcripts, microRNAs (miRNAs) are small RNAs that post-transcriptionally regulate gene expression. Despite great progress in understanding the biological role of miRNAs, our understanding of how miRNAs are regulated and processed is still developing. High-throughput sequencing data have provided a robust platform for transcriptome-level, as well as gene-promoter analyses. *In silico* predictive models help shed light on the transcriptional and post-transcriptional regulation of miRNAs, including their role in gene regulatory networks. Here we discuss the advances in computational methods that model different aspects of miRNA biogeneis, from transcriptional regulation to post-transcriptional processing. In particular, we show how the predicted miRNA promoters from PROmiRNA, a miRNA promoter prediction tool, can be used to identify the most probable regulatory factors for a miRNA in a specific tissue. As differential miRNA post-transcriptional processing also affects gene-regulatory networks, especially in diseases like cancer, we also describe a statistical model proposed in the literature to predict efficient miRNA processing from sequence features.

Keywords Mirna regulation • Promoter prediction • Mirna processing • Gene regulatory networks

1 The Role of miRNAs in Gene-Regulatory Networks

In biological research, diverse high-throughput techniques enable the investigation of whole systems at the molecular level. One of the main challenges for computational biologists is the integrated analysis of gene expression, interactions between

B. Caffrey • A. Marsico (✉)
Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany
e-mail: caffrey@molgen.mpg.de; marsico@molgen.mpg.de

V. Zazzu et al. (eds.), *Mathematical Models in Biology*,
DOI 10.1007/978-3-319-23497-7_6

genes and the associated regulatory mechanisms. The two most important types of regulators, Transcription Factors (TFs) and microRNA (miRNAs) often cooperate in complex networks at the transcriptional level and at the post-transcriptional level, thus enabling a combinatorial and highly complex regulation of cellular processes [1].

While TFs regulate genes at the transcriptional level by binding to proximal or distal regulatory elements within gene promoters [1], microRNAs (miRNAs) act at the post-transcriptional level on roughly half of the human genes. These short non-coding RNAs of 18–24 nucleotides in length which can bind to the 3′-untranslated regions (3′ UTRs) or coding regions of target genes, leading to the degradation of target mRNAs or translational repression [2].

MiRNAs are associated with an array of biological processes, such as embryonic development and stem cell functions in mammals [3], and a crucial role of miRNAs in gene regulatory networks has been recognized in the last decade in the context of cancer and other diseases [4, 5]. Altered miRNA expression profiles have often been associated with cancer development, progression and prognosis [6]. MiRNAs which negatively regulate tumor suppressor genes can be amplified in association with cancer development. On the other hand, deletions or mutations in miRNAs targeting oncogenes can lead to the over-expression of their targets [5, 6].

MiRNAs also affect several aspects of the immune system response [7]. For example, cells of the hematopoietic system can be distinguished from other tissues by their miRNA expression profiles, including, among the others, the highly expressed miRNA hsa-miR-155 [7]. Other immune system-related miRNAs are activated in response to viral or bacterial infections (e.g. hsa-miR-146a) and they affect the expression of several cytokines downstream [8].

Given the growing prevalence of miRNA functions in contributing to the control of gene expression, gene regulatory networks have been expanded to become rather complex incorporating the involvement of miRNAs. The general framework for inferring gene regulatory networks involving Transcription Factors (TFs) and miRNAs is usually built using the following steps:

- **1:** When expression data are available under a certain condition, the first step is to identify those genes which are mostly expressed in that particular condition or de-regulated compared to a control experiment.
- **2:** miRNAs responsible for the observed co-expression or de-regulation of a set of genes are identified by identifying enriched miRNA binding sites in the 3′-UTRs of such genes. This is usually done by mining publicly available databases for miRNA-target interactions [9, 10].
- **3:** MiRNA-target interactions are filtered based on the miRNA expression level (when available) or by using a cutoff score indicating the reliability of the predicted interaction. In addition, it is expected that when a miRNA regulates a gene, the miRNA and the gene show typical correlated expression patterns across multiple samples. This can be used as a criterium to further filter miRNA-gene interactions which do not show any such correlation [9].

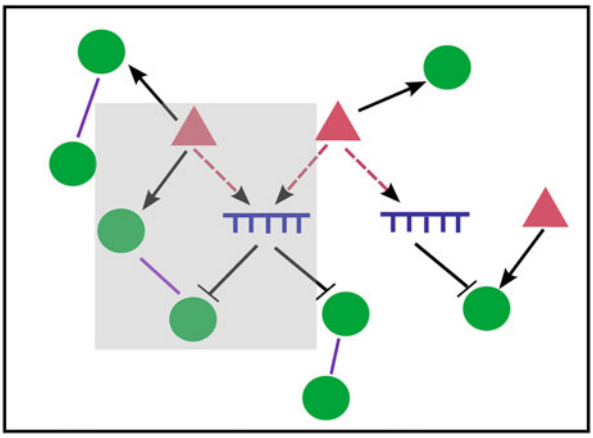

Fig. 1 Cooperative action of miRNAs and TFs in gene regulatory networks. miRNAs are colored in *blue*, TFs in *red* and their regulated target genes, as well as genes involved in potential protein–protein interactions are colored in *green*. *Dotted red arrows* indicate potential regulators of miRNAs and *purple lines* indicate protein–protein interactions extracted from databases. A typical feedback loop is highlighted in *grey*

- **4:** TFs regulating this set of genes can be inferred by means of prediction algorithms which scan for known TF binding sites in the proximal gene promoter regions using Position Weight Matrices (PWMs) [11].
- **5:** Protein–protein interaction databases, such as STRING, BioGrid and KEGG can be inspected to find possible interactors of such genes and the cellular pathways that they affect.

These steps give rise to a network as depicted in Fig. 1. In this schematic representation nodes represent the significant set of genes, miRNAs and transcription factors in the process under study and the links between them represent predicted regulatory interactions.

It is well known that miRNAs are involved in negative regulation and/or positive feedback loops which can also involve the transcription factors that regulate their activity [12]. The knowledge of the transcription factors which regulated the miRNAs in question often provide the missing links in the aforementioned regulatory network (Fig. 1, red dotted arrows). The identification of TF-miRNA interactions remains a difficult task without which a full understanding of the underlying processes is hampered. In recent years there has been an increase in the development of computational methods to predict miRNA promoters and their regulating TFs in order to unravel the TF-miRNA interactions missing in such typical regulatory networks.

2 MiRNA Transcriptional Regulation

2.1 Challenges of in silico miRNA Promoter Identification

MiRNA promoter recognition is a crucial step towards the understanding of miRNA regulation. Knowing the location of the miRNA transcription start site (TSS) enables the location of the core promoter, the region upstream of the TSS which contains

the TFs binding sites necessary to initiate and regulate transcription. Predictions of binding sites in the core promoter elements can enable the identification of regulatory factors for a specific miRNA (or a class of miRNAs), greatly improving our understanding of miRNA function, and their role in tissue-specific feedback loops.

Genome-wide identification of miRNA promoters has been hindered for many years by two main factors. The first reason is the deficit in mammalian promoter sequence characterization, which makes promoter prediction a challenging task in general [13]. Although promoter regions contain short conserved sequence features that distinguish them from other genomic regions, they lack overall sequence similarity, preventing detection by sequence-similarity-based search methods such as BLAST. Promoter recognition methods in the early 90s exploited the fact that promoters contain few specific sequence features or TF binding sites that distinguish them from other genomic features [13]. This observation could be used to build a consensus model, such as Position Weight Matrices (PWMs) or Logos to search for new promoters in the genome. It soon became clear that such methods could not be generalized to all existing promoters and more advanced strategies for pattern recognition utilized machine learning models trained on sequence k-mers.

The second reason for the lack of knowledge in miRNA transcriptional regulation is due to the complexity of the miRNA biogenesis pathway: miRNAs, whether they are located in intergenic regions or within introns of protein-coding genes, often referred to as host genes, are generally generated from long primary transcripts which are rapidly cleaved in the nucleus by the enzyme Drosha [2]. This presents a technical barrier for large-scale identification of miRNA TSSs as they can be located in regions far away from the mature miRNA and cannot be inferred simply from the annotation of the processed mature miRNA, as done for stable protein coding gene transcripts [14]. In addition, the situation is further complicated by the fact that recent studies indicate that several alternative miRNA biogenesis pathways exist, especially for intragenic miRNAs. Indeed, if co-transcription with the host gene were the only mechanism to generate intragenic miRNAs, then the mirna and hostgene expression should be highly correlated among different tissues or conditions. Many recent studies, however, show many instances of poor correlation between mirna and host gene, pointing to an independent regulation of the mirna, utilizing an alternative intronic promoter [15]. There is evidence that intragenic miRNAs may act as negative feedback regulatory elements of their hosts interactomes [16] but the contribution of host gene promoter versus intronic miRNA promoters, and the mechanisms that control intronic promoter usage are still interesting open questions in the biology of miRNA biogenesis.

Although overall similarity in promoters is not a general phenomena, it does exist in the form of phylogenetic footprinting. Based on this observation, one of the first methods for miRNA promoter detection identifies about 60 miRNA transcriptional start regions by scanning for highly conserved genomic blocks within 100 kb of each mature miRNA and searching for a core promoter element in the consensus

sequence regions extracted from these blocks [17]. Although this method proved to be valid in the identification of evolutionary conserved promoters, the sensitivity of such an evolutionary approach is very low, given the high number of non-conserved miRNAs annotated in MiRBase [18].

2.2 *Next Generation Sequencing (NGS) Technology Leads to Significant Advances in miRNA Promoter Prediction*

Recently, thanks to the advent of next-generation sequencing technologies combined with Chromatin Immunoprecipitation (CHIP-Seq technology [19]) and nascent transcript capturing methods, such as Cap Analysis of Gene Expression coupled to NGS sequencing (deepCAGE) [20] or Global run on sequencing (GRO-seq) [21], several computational methods for miRNA promoter prediction genome-wide have been developed, providing valuable understanding in the detailed mechanisms of miRNA transcriptional regulation. For example, the epigenetic mark H3K4me3 has been identified as a hallmark of active promoters, and computational methods for promoter recognition have begun exploiting this information systematically.

The deepCAGE technique enables the mapping of the location of TSSs genome-wide. In the FANTOM4 Consortium this technique was applied across various different tissues and conditions in order to profile transcriptional activities and promoter usage among different libraries.

GRO-seq is a technique to capture nascent RNAs genome-wide by quantifying the signal of transcriptionally engaged PolII at gene promoters. Both deepCAGE and GRO-seq read density is sharply peaked around transcripts TSS and it can be successfully used to locate the TSSs of miRNA primary transcripts [14, 22]. Finally, recent RNA-seq studies with increased sequencing depth can also be used to identify the transient and lowly expressed pri-miRNA transcripts [22].

2.3 *Classification and Comparison of miRNA Promoter Prediction Methods*

A limited number of miRNA promoter recognition methods have been developed in the past few years and can be classified either according to the methodology used, supervised versus unsupervised learning approaches, or based on the nature of their predictions, tissue specific versus general promoter predictions and intergenic versus all predicted promoters, including intronic promoters. The main features of existing miRNA promoter prediction methods can be summarized in Table 1.

According to the model used to describe the data one can distinguish two categories of miRNA promoter recognition methods:

Table 1 Comparison of different methods for miRNA promoter prediction

	Fujita [17]	Ozsolak [25]	Marson [23]	Barski [24]	S-Peaker: Megraw [27]	miRStart: Chien [26]	PROmiRNA: Marsico [14]	microTSS: Georgakilas [22]
Cell line	–	UACC62, MALME, MCF cells	mESC, hESC cells	CD4+ T cells	–	36 different tissues	33 different tissues	mESC, hESC, IMR90 cells
Data used	Blastz genomic alignments from UCSC	Nucleosome positions from ChIP-chip data	H3K4me3 ChIP-seq data	H3K4me3, H3K9ac, H2AZ, and PolII ChIp-seq data	CAGE data (FANTOM4)	36 deepCAGE libraries (FANTOM4) and 14 TSS-seq libraries	33 deepCAGE libraries (FANTOM4)	RNA-seq data, HeK4me2, PolII ChIP-seq and DNAse-seq
Methodology	Unsupervised approach: identification of conserved blocks upstream of miRNAs	Unsupervised approach: empirical score of nucleosome-free regions based on sequence features (TFBs)	Unsupervised approach: empirical score based on HeK4me3, conservation and proximity to the mature miRNA	Unsupervised approach: score accounting for evidence of peaks from four ChIP-seq signals, plus EST evidence	Supervised approach: L1-logistic regression model trained on gene promoter TFBs	Supervised model: SVM trained on protein coding genes	Semi-supervised mixture model built on CAGE data and sequence features	Supervised model: SVM trained on chromatin features at gene promoters and then used to score miRNA RNA-seq enriched regions
Intergenic promoters	yes	yes	yes	yes	yes	yes	yes	yes
Intronic promoters	Not reported	yes	no	no	no	not reported	yes	no

- *De novo approaches*, which identify and score miRNA TSS in an unsupervised manner. These include models based on experimentally determined histone mark profiles [23, 24] or nucleosome positioning patterns [25]. For example Marson [23] and Barski [24] consider regions enriched in H3K4me3 signal as putative promoters and assign them a score. Ozsolak [25] combine nucleosome positioning patterns with ChIP-chip screens to score putative transcription initiation regions upstream of active miRNAs.
- *Supervised methods*, based on the evidence that miRNA promoters present the typical characteristics of Polymerase II transcription and therefore trained on protein coding gene promoter features and subsequently used to predict miRNA promoters. Such methods include mirStart [26] and microTSS [22] (Table 1). MirStart trains a SVM model on protein coding gene features (CAGE tags, TSS-Seq and HeK4me3 ChIP-Seq data), and uses the trained model to identify putative miRNA promoters [26]. microTSS also uses a combination of three SVM models trained on HEK4me3 and PolII occupancy at protein-coding gene promoters to score putative initial miRNA TSSs candidates derived from deeply sequenced RNA-Seq data [22].

One of the latest miRNA promoter prediction tools, PROmiRNA, is a method in between these two categories [14]. PROmiRNA is based on a semi-supervised classification model which does not make any assumption about the nature of miRNA promoters and their similarities to protein-coding genes. On the contrary, PROmiRNA tries to learn the separation between putative miRNA promoters and transcriptional noise based on few features, such as CAGE tag clusters upstream of annotated miRNAs and sequence features.

Each of the described methods has advantages and disadvantages. Methods for miRNA promoter recognition based solely on sequence features, such as the evolutionary framework proposed by Fujita [17] or S-Peaker [27], based on TF binding sites and proposed by Megraw et al., are very accurate in identifying putative promoter regions. MiRNAs are, however, known to mediate gene regulation in a highly tissue-specific manner, therefore it is expected that their regulation also happens in a tissue-specific way. Such methods cannot distinguish between promoters potentially active in different tissues, given that sequence features are invariant features, but merely suggest possible locations for miRNA promoters. On the other hand, methods based on chromatin features have been designed for specific cell lines, therefore providing a snapshot of the active promoters. Histone mark-based methods provide a broad view of promoter regions, rather than high-resolution predictions, hampering the detection of multiple TSSs close to each other in the genome. In addition, most chromain-based methods can predict the promoters of independently transcribed intergenic miRNAs, but lack sensitivity in discovering alternative or intronic promoters.

MicroTSS overcomes the problem of the non-informative broad predictions by making use of deep-coverage RNA-seq data and pre-selecting RNA-seq islands of transcription upstream of intergenic pre-miRNAs at single-nucleotide resolution. Such initial miRNA promoter candidates are then given as input to the SVM model

which returns the predictions. Due to the nature of the RNA-seq used to pre-select candidate TSSs, microTSS works well for intergenic miRNAs but is not suitable for identifying intronic promoters due to the difficulty in discriminating transcription initiation events from the read coverage signal corresponding to the host transcript.

The method from Ozsolak [25] and PROmiRNA [14] are the only two methodologies which report predictions of intronic promoters. In particular, in PROmiRNA miRNA promoter predictions are derived from multiple high-coverage deepCAGE libraries, and correspond to highly expressed, as well as lowly expressed tissue-specific intronic promoters.

Figure 2 shows seven predicted TSSs for hsa-miR-155, six of which are intronic promoters in a leukemia cell line, indicating that alternative promoters are able to drive the expression of this miRNA in this cell line. However, due to the low expression of alternative intronic promoters, compared to intergenic promoters, and to the difficulty of validating such promoter predictions (a gold standard data-set for miRNA promoters is missing), predictions of intronic promoters may suffer from higher false discovery rates compared to intergenic promoters.

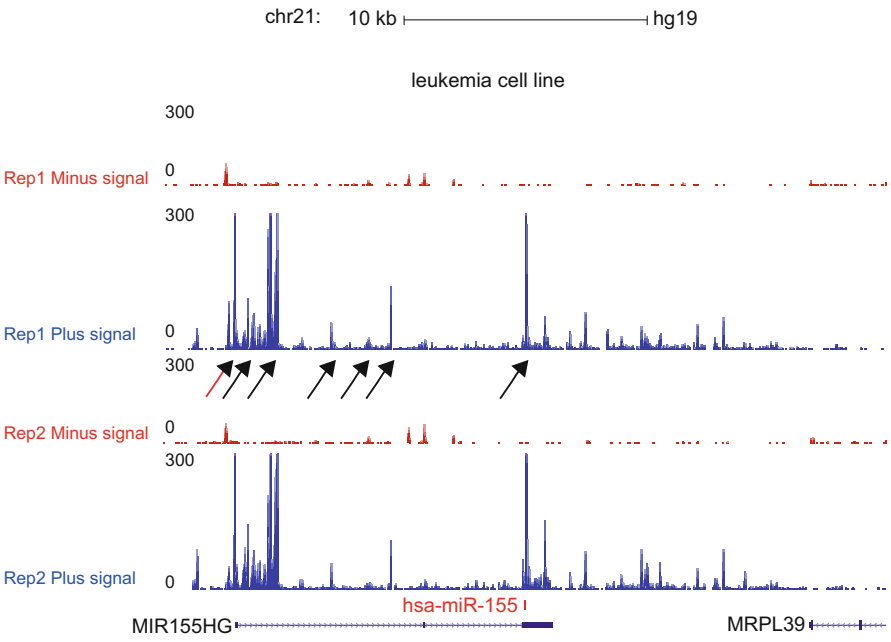

Fig. 2 PROmiRNA predicted promoters for hsa-miR-155, a human miRNA located in the non-coding BIC host transcript (also called MIR155HG). The *red arrow* indicates the TSS of the host gene and the *other arrows* point to the predicted alternative intronic promoters located in the genomic range between 677 bp and 12 kb upstream of the mature miRNA. The promoter predictions where consistent in two different CAGE replicates

2.3.1 miRNA-Mediated Regulatory Network Reconstruction

Here we show an application of PROmiRNA to the derivation of the tissue-specific miRNA-mediated regulatory network in three immune cell line libraries from the FANTOM4 CAGE data. For simplicity we include in this network only two of the main human miRNAs which are highly expressed and known to play a role in the Immune System: hsa-miR-146a and hsa-miR-155. MiR-146a is an intergenic miRNA known to be involved in regulation of inflammation and other processes related to innate immune response [8]. Mir-155 resides in the non-coding host gene MIR155HG and is known to play a role in cancer, as well as viral and bacterial infection processes [28]

PROmiRNA predicts two alternative promoters in the leukemia cell line for hsa-miR-146a, one located 17 kb upstream of the mature miRNA and the other 16.6 kb, and six alternative intronic promoters (in addition to the host gene promoter) for hsa-miR-155 (as already shown in Fig. 2). Starting from these predictions, we scanned the 1000 bp regions around each predicted miRNA TSSs for putative transcription factor binding sites with the TRAP tool [29]. Given a database of TFBs motif models, TRAP computes the affinity of each factor for a certain genomic sequence. For each predicted promoter we ranked the TFs based on their computed binding affinities. The top ten factors regulating each miRNA are selected and included in the network if they are expressed in Immune System cell lines, according to the Human Protein Atlas database [30]. Potential regulatory factors are connected by means of edges to the corresponding miRNA (Fig. 3). Also potential miRNA targets extracted from TargetScan and other miRNA target databases [9], as well as interactions between gene–gene and gene-TF are extracted from the STRING database [31] and, if expressed in the Immune system, added to the network (Fig. 3). This partial reconstruction of the regulatory network involving hsa-mir-146a and hsa-mir-155 in the Immune System shows that a portion of the top target predictions is shared between the two miRNAs, while other targets are specific to one or the other miRNA. Also, hsa-miR-146a and hsa-miR-155 seem to be targeted by a set of common transcription factors, among which we find the NFKB1, a well known Immune System factor.

3 Predictive Models of miRNA Processing

Global mature miRNA expression is not only regulated at transcriptional level, but several post-transcriptional steps influence the final miRNA expression level and contribute to define a particular phenotype. In detail, miRNA initially generated in the nucleus as long primary transcripts are processed by the Microprocessor complex (Drosha/DGCR8) to produce stem-loop structured precursors which are then further processed in the cytoplasm by Dicer [32]. While signatures of miRNA expression may be used as biomarkers for cancer diagnosis and stratification in several cancers, it has become clear in recent years that specifically aberrant processing,

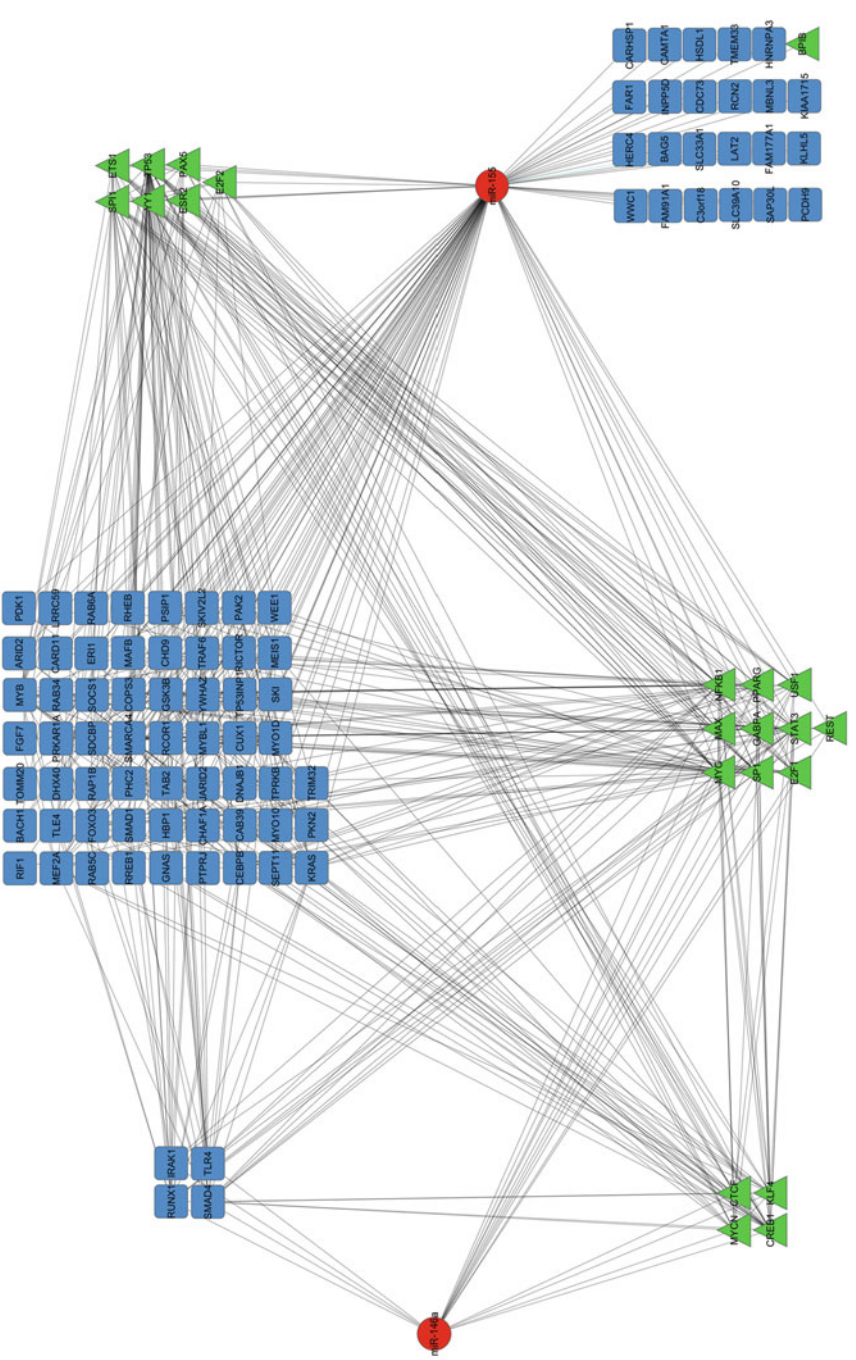

Fig. 3 Regulatory network mediated by miRNAs hsa-miR-155 and hsa-miR-146a, including experimentally validated (mirTarbase) and predicted (overlap of miranda & TargetScan) miRNA-target interactions, transcription factor-miRNA interactions detected by means of the TRAP tool in the PROmiRNA predicted miRNA regions, and gene–gene interactions extracted from the STRING database

rather than altered transcription, correlates with cell invasion or progression of inflamation. The method by which the Microprocessor is able to distinguish miRNA hairpins from random hairpin structures along the genome and efficiently process them is still a subject of investigation. Recent studies have shown that sequence motifs flanking precursor miRNAs play a significant role in primary transcript cleavage [33].

In a recent study [34] we have quantified the effect of different sequence motifs on the Microprocessor activity in an endogenous setting. We have performed high-throughput RNA sequencing experiments of nascent transcripts associated to the chromatin fraction in different cell lines. Since processing of primary miRNA transcripts occurs co-transcriptionally, while the transcript is still associated to chromatin, the read coverage pattern at miRNA loci shows the typical Microprocessor signature, where Drosha cleavage is reflected in a significant drop in the read coverage in the precursor region. We have defined a quantitative measure of processing efficiency called Microprocessing Index (MPI), as the logarithm of the ratio between the read density adjacent to the pre-miRNA and the read density in the precursor region. On the basis of MPI values, miRNAs could be divided into *efficiently processed* (Fig. 4a $MPI <= -1.0$, also called positive examples) and *non-efficiently processed* (Fig. 4b $MPI >= -0.4$, also called negative examples).

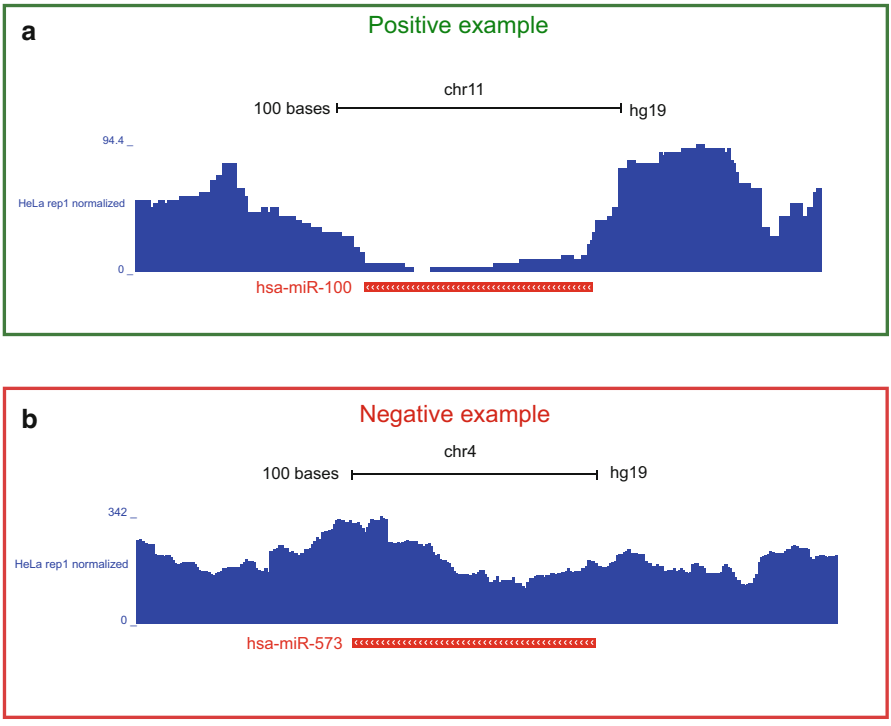

Fig. 4 Genomic regions around miRNAs hsa-miR-100 (**a**) and hsa-miR-573 (**b**), respectively, and normalized read coverage at the miRNA loci. The significant drop in read coverage at the miR-100 precursor indicates that this miRNA is efficiently processed in HeLa cells (**a**), while miR-573 is not

A classification model based on sequence features was built in order to discriminate between these two classes. We used L1-regularized logistic regression for training and classification of the miRNA in positives and negatives. In detail, given a binary variable Y, where $y_i = 0$ or $y_i = 1$ for each data point i, the probability of the outcome of Y, given the data **X**, is given by the following sigmoid function:

$$P(y = 1|x, \theta) = \frac{1}{1 + exp(-\theta^T x)} \tag{1}$$

where θ is the parameter vector of the logistic regression model. The optimization problem (Maximum Likelihood Estimate of θ) in the case of L1-regularization is formulated as the following:

$$min_\theta(\sum_{i=1}^{M} -logP(y_i|x_i, \theta) + \beta\|\theta\|_1) \tag{2}$$

In our case the features used in the model were either dinucleotide counts (dinucleotide-based model) or counts of short motifs (motif-based model) in the regions upstream and downstream of miRNA precursors. L1-regularized logistic regression performs automatic feature selection penalizing dinucleotides or motifs which are not significant in distinguishing efficiently processed miRNAs from non-efficiently processed. We found that the most important features associated with enhanced processing are: a GNNU motif (N indicates any nucleotide) directly upstream of the 5′ of the miRNA, a CNNC motif between 17 and 21 positions downstream of 3′ of the miRNAs and dinucleotides GC and CU enriched at the base of the miRNA stem loop.

4 Conclusions

In silico methods for studying miRNA biogenesis, ranging from statistical models of promoter recognition and transcription factor binding site prediction to predictive models of miRNA processing, enable a better understanding of miRNA-mediated regulation in tissue-specific networks. Recent progress in the field of NGS resulted in a plethora of high-throughput and high-quality datasets in the last few years. This enabled the development of data-driven computational approaches which make use of such data and combine them with traditional sequence signals, in order to get more accurate prediction of miRNA promoters. Although the basics of the miRNA biogenesis pathway are known, there are still many unsolved questions. For example, several regulatory factors might be involved in miRNA regulation at different levels. Although some regulators of miRNA transcription and processing

have been predicted and experimentally validated, more sophisticated *in silico* methods are needed to discover more of these factors and predict how they affect miRNA biogenesis.

RNA binding proteins interact with both pri-miRNAs in addition to intermediate miRNA products at different stages of their regulation. High-throughput sequencing of RNA sites bound by a particular protein will reveal more aspects about miRNA regulation, as well as enable more reliable identification of targets which are physiologically relevant.

Although observations from different sources need to be unified in a coherent framework, it is clear that targeted computational approaches can help linking different evidence from several genomic datasets and give a significant contribution to discover additional details about miRNA-mediated regulation.

References

1. Guo, Z.: Genome-wide survey of tissue-specific microRNA and transcription factor regulatory networks in 12 tissues. Sci. Rep. **4**, 5150 (2014)
2. Davis, B.N.: Regulation of MicroRNA biogenesis: a miRiad of mechanisms. Cell Commun. Signal **10**, 7–18 (2014)
3. Bartel, D.P.: MicroRNAs: target recognition and regulatory functions. Cell **136**, 215–233 (2009)
4. Plaisier, C.L.: A miRNA-regulatory network explains how dysregulated miRNAs perturb oncogenic processes across diverse cancers. Genome Res. **22**, 2302–2314 (2012)
5. Esquela-Kerscher, A.: Oncomirs - microRNAs with a role in cancer. Nat. Rev. Cancer **6**, 259–269 (2006)
6. Takahashi, R.U.: The role of microRNAs in the regulation of cancer stem cells. Front Genet **4**, 295 (2014)
7. Davidson-Moncada, J.: MicroRNAs of the immune system: roles in inflammation and cancer. Ann. N. Y. Acad. Sci. **1183**, 183–194 (2010)
8. Ma, X.: MicrorNAs in NF-kappaB signaling. J. Mol. Cell Biol. **3**, 159–166 (2011)
9. Lewis, B.P.: Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell **120**, 15–20 (2005)
10. Betel, D.: Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol. **11**, R90 (2010)
11. Sandelin, A.: JASPAR: an open-access database for eukaryotic transcription factor binding profiles Nucl. Acids Res. **32** D91–D94 (2004)
12. Krol, J.: The widespread regulation of microRNA biogenesis, function and decay. Nat. Rev. Genet. **11**, 597–610 (2010)
13. Fickett, J.: Eukaryotic promoter recognition. Genome Res. **7**, 861–878 (1997)
14. Marsico, A.: PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs. Genome Biol. **14**, R84 (2013)
15. Monteys, A.M.: Structure and activity of putative intronic miRNA promoters. RNA **16**, 495–505 (2010)
16. Hinske, L.C.: A potential role for intragenic miRNAs on their hosts' interactome. BMC Genomics **11**, 533 (2010)
17. Fujita, S.: Putative promoter regions of miRNA genes involved in evolutionarily conserved regulatory systems among vertebrates. Bioinformatics **24**, 303–308 (2008)

18. Kozomara, A.: miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. **39**, D152-D157 (2011)
19. Kozomara, A.: ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. **10**, 669–680 (2009)
20. de Hoon, M.: Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. Biotechniques **44**, 627–628 (2008)
21. Core, L.J.: Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science **322**, 1845–1848 (2008)
22. Georgakilas, G.: microTSS: accurate microRNA transcription start site identification reveals a significant number of divergent pri-miRNAs. Nat. Commun. **10**, 5700 (2014)
23. Barski, A.: Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell **134**, 521–533 (2008)
24. Barski, A.: Chromatin poises miRNA- and protein-coding genes for expression. Genome Res. **19**, 1742–1751 (2009)
25. Ozsolak, F.: Chromatin structure analyses identify miRNA promoters. Gene Dev. **22**, 3172–3183 (2008)
26. Chien, C.H.: Identifying transcriptional start sites of human microRNAs based on high-throughput sequencing data. Nucleic Acids Res. **39**, 9345–9356 (2011)
27. Megraw, M.: A transcription factor affinity-based code for mammalian transcription initiation. Genome Res. **19**, 644–656 (2009)
28. Eis, P.: Accumulation of miR-155 and BIC RNA in human B cell lymphomas. Proc. Natl. Acad. Sci. **102**, 3627–3632 (2003)
29. Thomas-Chollier, M.: Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. Nat. Protoc. **102**, 3627–3632 (2003)
30. Uhlen, M.: Tissue-based map of the human proteome. Science **347**, 1260419 (2015)
31. Szklarczyk, D.: The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acid Res. **39**, D561–568 (2011)
32. Ha, M.: Regulation of microRNA biogenesis. Nat. Rev. Mol. Cell Biol. **15**, 509–524 (2014)
33. Auyeung, V.C.: Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. Cell **152**, 844–858 (2013)
34. Conrad, T.: Microprocessor activity controls differential miRNA biogenesis in vivo. Cell Rep. **9**, 542–554 (2014)