Bioinformatics, 31(5), 2015, 665–673
doi: 10.1093/bioinformatics/btu696
Advance Access Publication Date: 22 October 2014
Original Paper

OXFORD

Sequence analysis

FlaiMapper: computational annotation of small ncRNA-derived fragments using RNA-seq high-throughput data

Youri Hoogstrate, Guido Jenster and Elena S. Martens-Uzunova*

Department of Urology, Erasmus University Medical Center, Be 362a, PO Box 2040, 3000 CA Rotterdam, The Netherlands

*To whom correspondence should be addressed.
Associate Editor: Ivo Hofacker

Received on May 2, 2014; revised on September 18, 2014; accepted on October 20, 2014

Abstract

Motivation: Recent discoveries show that most types of small non-coding RNAs (sncRNAs) such as miRNAs, snoRNAs and tRNAs get further processed into putatively active smaller RNA species. Their roles, genetic profiles and underlying processing mechanisms are only partially understood. To find their quantities and characteristics, a proper annotation is essential. Here, we present FlaiMapper, a method that extracts and annotates the locations of sncRNA-derived RNAs (sncdRNAs). These sncdRNAs are often detected in sequencing data and observed as *fragments* of their precursor sncRNA. Using small RNA-seq read alignments, FlaiMapper is able to annotate fragments primarily by peak detection on the start and end position densities followed by filtering and a reconstruction process.

Results: To assess performance of FlaiMapper, we used independent publicly available small RNA-seq data. We were able to detect fragments representing putative sncdRNAs from nearly all types of sncRNA, including 97.8% of the annotated miRNAs in miRBase that have supporting reads. Comparison of FlaiMapper-predicted boundaries of miRNAs with miRBase entries demonstrated that 89% of the start and 54% of the end positions are identical. Additional benchmarking showed that FlaiMapper is superior in performance compared with existing software. Further analysis indicated a variety of characteristics in the fragments, including sequence motifs and relations with RNA interacting factors. These characteristics set a good basis for further research on sncdRNAs.

Availability and implementation: The platform independent GPL licensed Python 2.7 code is available at: https://github.com/yhoogstrate/flaimapper. Corresponding Linux-specific scripts and annotations can be found in the same repository.

Contact: e.martens@erasmusmc.nl

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

1 Introduction

Sequencing of small non-coding RNAs (sncRNAs) aiming at the quantification and discovery of microRNAs (miRNAs), small nucleolar RNAs (snoRNAs), transfer RNAs (tRNAs) and vault RNAs (vtRNAs) has revealed that most types of sncRNAs get processed into smaller RNAs (Chen and Heard, 2013). Initially, it was suggested that these smaller RNAs are degradation products of the turnover of their precursors. Nevertheless, evidence accumulating over the last years demonstrates that some RNA fragments are

functional and have specific maturation mechanisms indicating their importance and novelty (Sobala and Hutvagner, 2011; Scott *et al.*, 2012; Chen and Heard, 2013). Such fragments find their origin in tRNAs, vtRNAs and snoRNAs and are assumed to have a variety of functions. Most importantly, deregulation and involvement of different types of fragments have been demonstrated in different types of cancer (Martens-Uzunova *et al.*, 2013). A description of commonly detected fragments and their precursors is given below:

- A *pre-miRNA* is an approximately 75 nt long RNA molecule produced from its primary precursor transcript (pri-miRNA) by Drosha (Valen *et al.*, 2011). Pre-miRNAs adopt a hairpin structure recognized by Dicer that cleaves the terminal loop to release an approximately 22-nt long double-stranded miRNA duplex. One of the strands (miRNA) is loaded into AGO to generate the functional miRISC complex (Lin He, 2004; Fabian and Sonenberg, 2012). The remaining strand (miRNA*) is usually degraded. Often both strands are found as fragments in small RNA-seq (Friedlander *et al.*, 2008).
- Fragments originating from mature tRNAs are commonly classified into two subgroups (Sobala and Hutvagner, 2011), tRNA halves and tRNA-derived RNA fragments (tRFs):
  - tRNA halves are most probably produced by angiogenin that cleaves the tRNA near its anticodon, resulting into halve tRNAs (∼ 35 nt). It is believed that some tRNA halves contribute to translational repression and cell stress response (Thompson and Parker, 2009; Yamasaki *et al.*, 2009).
  - The smaller (∼ 20 nt) tRFs are derived from the tRNAs 5′- and 3′-end and from the pre-tRNAs 3′-end. It is not completely understood which proteins are involved in the production of tRFs, although evidence for associations with both Dicer and RNaseZ are reported (Sobala and Hutvagner, 2011; Chen and Heard, 2013). Although the putative functions of the majority of tRFs are unclear, evidence suggests that some are involved in RNA interference, with effects on cell proliferation and gene regulation (Chen and Heard, 2013).
- snoRNA are (60–250 nt) small RNAs found in the nucleolus. They comprise the subtypes H/ACA-box, C/D-box and small Cajal body-specific RNAs (scaRNAs) (Henras *et al.*, 2004). Putative functions such as regulation of alternative splicing, post-transcriptional regulation of gene expression and associations with cancer have been proposed for their fragments (Ender *et al.*, 2008; Dong *et al.*, 2009; Kishore *et al.*, 2010; Askarian-Amiri *et al.*, 2011; Brameier *et al.*, 2011; Ono *et al.*, 2011; Mei *et al.*, 2012; Scott *et al.*, 2012).

Currently, studies on fragments other than miRNA and miRNA* are restricted to (often visual) interpretation of alignments. Consequently, the data are inspected only at a global ncRNA level. Not making use of the annotation of exact fragment coordinates is a shortcoming, since it restricts analysis at the level of individual fragments. Additional benefit of such analysis is the gained statistical power.

Here, we describe Fragment Location Annotation Mapper (*FlaiMapper*) that predicts the locations of sncRNA fragments in small RNA-seq alignments. Prediction is based on the densities of start and end positions of aligned reads. It is important to state that the goal is not to predict any particular subtype of fragment but to annotate data for subsequent quantitative analysis, by making use of sequencing data only. Therefore, FlaiMapper does not use 2D structure prediction or classification based on heuristics of previous discoveries as often is used for the prediction of pre-miRNAs (Friedlander *et al.*, 2008).

2 Methods

Fragments are measured with small RNA-seq, where the corresponding variable-sized sequences, called reads, are aligned back to a reference sequence. The reference sequence is used to determine the reads origin. This reference can be the genome, the transcriptome or specific regions (e.g. miRNA or tRNA databases). The library used for our analysis was manually composed (Supplementary information). Pre-processing and alignment for each dataset are further discussed in the Supplementary information.

Analysis was applied to two different publicly available datasets with SRA accession numbers `SRP002175` (Stark *et al.*, 2010) and `SRP006788` (Valen *et al.*, 2011). Dataset `SRP002175` contains 12 small RNA-seq samples, taken from human pigment cells. The reads are 18–23-nt long and processed on the Illumina's Genome Analyzer II platform. Dataset `SRP006788`, processed on the same platform, contains 18–30-nt long reads, taken from six samples from a HeLa cell line. In this dataset, the samples have undergone the following treatments (Valen *et al.*, 2011):

- `SRR207111` Total cellular RNA was extracted from HeLa cells.
- `SRR207112` Total cellular RNA was extracted after RRP40 core subunit depletion; RRP40 has 3′ → 5′ exonucleolytic activity.
- `SRR207113` RNA pool-down obtained from non-treated HeLa cells after AGO1 and AGO2 immunoprecipitation.
- `SRR207114` RNA pool-down obtained from RRP40-depleted HeLa cells after AGO1 and AGO2 immunoprecipitation.
- `SRR207115` Total cellular RNA was extracted after XRN1 and XRN2 depletion; XRN has 5′ → 3′ exonucleolytic activity.
- `SRR207116` RNA was extracted from the nucleus.

2.1 Formal problem

For convenience, we use the term *boundary* to describe either a start or an end position of a fragment, without being specific to one of them. If an alignment of a fragment is inspected in more detail, its boundaries are indicated by the corresponding start and end positions of the aligned reads. Fragment boundaries are variable, as indicated by the aligned reads (Fig. 1). Read starts and ends are located at variable positions, but close to the boundaries. This results in peaks in the densities of aligned start and end positions, near the boundaries. Therefore, it seems more convenient to estimate fragments using the most common start and end positions instead of the most common read. The number of read starts or ends at a certain position in the sequence (*intensity*) decreases rapidly and symmetrically with respect to the position with the highest intensity. As a result, the peaks have characteristics compatible to a normal distribution, with its expected value being the position with the highest intensity. Because of the variability in the alignments and the limited sequencing depth, the data contain noise (Fig. 2). In FlaiMapper, a fragment is defined as:

1. *The region in a precursor ncRNA in-between the most common start and most common end position, as defined by aligned reads.*

Fig. 1. 1468 reads aligned to SNORD74 (black line) in dataset `SRP006788` (total RNA). The contours of the aligned reads (grey) form three separate clusters. This may be an indicator for the presence of multiple fragments

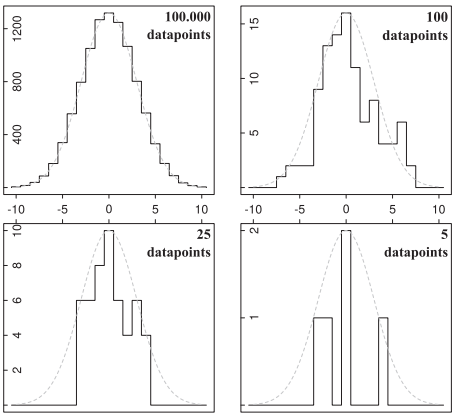

Fig. 2. Relation between noise and sequencing depth. Under the assumption that the variability of reads near a boundary is normally distributed with a standard deviation of 3, we illustrate effects of noise by binned sampling of this distribution at different resolutions. The horizontal axes give the offset to a fragment's true position and the vertical axes the number of times an (artificial) intensity is sampled from the distribution. The dashed line represents the distribution used for sampling. A sequencing technology with an infinite resolution (top left; by approximation) would result into one single peak in the vertical axis. As the sequencing depth decreases, the sampled distribution deviates further from the true distribution and more peaks in the vertical axis may appear by chance (top right and bottom). Each illustration belongs to the same simulated fragment boundary, derived from the same distribution. Because FlaiMapper expects only one peak per boundary, the remaining peaks, caused by the deviation from their original distribution, are referred to as noise

Consequently, the problem of finding such fragments is defined as:

2. *Given a set of aligned reads to a precursor ncRNA, the challenge is to estimate a fragment by: (i) finding the correct candidate start and end positions, (ii) taking the optimal proportion of noise into account and (iii) relating the corresponding start and end positions that belong to the same fragment back to each other.*

2.2 Algorithm

The FlaiMapper algorithm is divided into five sequential steps (Fig. 3): (**i**) *parsing*, (**ii**) *metrics*, (**iii**) *peak detection*, (**iv**) *filtering* and (**v**) *reconstruction*.

2.2.1 Parsing

For every ncRNA, alignments are parsed from input files. There is no preference towards a specific alignment algorithm as long as its output is in BAM format.

2.2.2 Metrics

Given an ncRNA with a length of n nt, the following corresponding vectors are determined:

- Start and stop position densities
  1. $\mathbf{p}^{5'} = (p_1^{5'}, p_2^{5'}, \ldots, p_n^{5'})$; here, $p_i^{5'}$ is the total number of reads that have their start position (5′-end) aligned to position i of the precursor ncRNA, where $1 \leq i \leq n$.
  2. $\mathbf{p}^{3'} = (p_1^{3'}, p_2^{3'}, \ldots, p_n^{3'})$; here, $p_i^{3'}$ is the total number of reads that have their end position (3′-end) aligned to position i of the precursor ncRNA, where $1 \leq i \leq n$.
- Read lengths
  1. $\mathbf{l}^{5'} = (l_1^{5'}, l_2^{5'}, \ldots, l_n^{5'})$; here, $l_i^{5'}$ is the average read length of reads that have their start position (5′-end) aligned to position i of the precursor ncRNA, where $1 \leq i \leq n$. If no reads have their start position aligned to nucleotide i, $l_i^{5'} = 0$.
  2. $\mathbf{l}^{3'} = (l_1^{3'}, l_2^{3'}, \ldots, l_n^{3'})$; here, $l_i^{3'}$ is the average read length of reads that have their end position (3′-end) aligned to position i of the precursor ncRNA, where $1 \leq i \leq n$. If no reads have their end position aligned to nucleotide i, $l_i^{3'} = 0$.

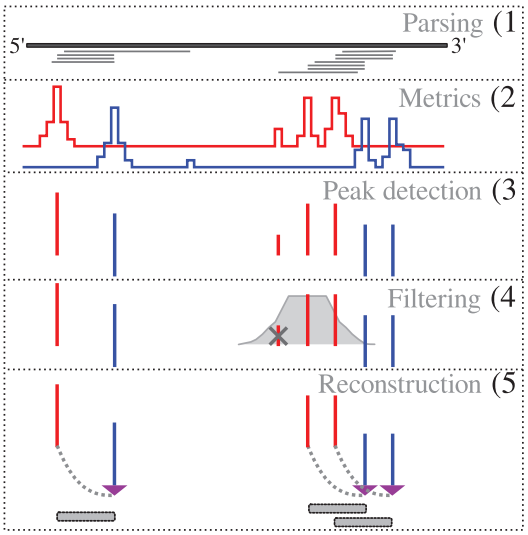

Fig. 3. Schematic overview of the five steps that FlaiMapper performs per sncRNA. (**i**) *Parsing*: alignment file is parsed; reads (thin lines) are aligned to a sncRNA (bold line). (**ii**) *Metrics*: acquire alignment statistics; for all positions in the sncRNA: I: find the number of aligned start (red) and end (blue) positions (referred to as intensity) and II: find the average length of mapped reads (not illustrated). (**iii**) *Peak detection*: predict candidate start and end positions (vertical lines) upon the intensity vectors using peak detection. (**iv**) *Filtering*: remove candidate start and end positions expected to be detected due to noise. In the example above, a candidate start position is discarded (grey cross) because it is an artefact of the noise of its neighbour. The remaining positions are considered as actual start and ends. (**v**) *Reconstruction*: reconstruct predicted fragments (grey bars) by finding corresponding start and end positions using a balance (purple triangle) between expected distance and intensity

2.2.3 Peak detection

Candidate start and end positions are characterized by peaks in the intensity vectors. Therefore, candidate positions are estimated independently in directions $d = 5'$ (start) and $d = 3'$ (end) on vector $\mathbf{p}^d$. The methodology of independence between start and end positions used by FlaiMapper is different from methods that rely on (i) the most common read or (ii) distributions of read density. Because of this, candidate start and end positions loose their one-to-one relationship. The purpose of peak detection is to find all positions that have an intensity higher than its adjacent positions. Because of noise in the intensities, the difference in intensity with respect to the adjacent values must be above a certain threshold. For direction d, the algorithm detects peaks upon corresponding vector $\mathbf{p}^d$ of length n. Vector $\mathbf{p}^d$ should be extended with a 0 at the end, to ensure that a peak at the very last position can be called. To avoid confusion about the lengths, we denote $\mathbf{q}^d = \{\mathbf{p}^d, 0\}$ and as consequence its length $n' = n + 1$. For every ith position in the vector, the intensity q_i^d is compared with the previous highest value.

- If the intensity is larger, it becomes the highest value, and is therefore the (new) candidate to become a peak.
- If the intensity is smaller, a drop in intensity is observed. If the drop is more than 90%, the jth peak is called by putting the

location in c_j^d. Subsequently, the candidate position will be reset and iterator j is increased with 1.

The formal description of peak detection is given in algorithm 1 and per ncRNA, the following vectors are added:

1. $\mathbf{c}^{5'} = (c_1^{5'}, c_2^{5'}, \ldots, c_k^{5'})$; for a number of k candidate start positions, the ith start position is located at nucleotide $c_i^{5'}$ of the ncRNA, where $1 \leq i \leq k$ and $1 \leq c_i^{5'} \leq n$.
2. $\mathbf{c}^{3'} = (c_1^{3'}, c_2^{3'}, \ldots, c_m^{3'})$; for a number of m candidate end positions $c_i^{3'}$, the ith end position is located at nucleotide $c_i^{3'}$ of the ncRNA, where $1 \leq i \leq m$ and $1 \leq c_i^{3'} \leq n$.

2.2.4 Filtering

Per fragment, multiple candidate start and end positions are frequently found due to noise. A target peak may be derived from the same fragment as surrounding peaks (Fig. 2). For each target peak at position i, a filter tests whether the remaining peaks at i', are indeed noise of the target. The intensity around a boundary has characteristics of a normal distribution and decreases as the distance to the true start or stop position increases. Peaks caused by noise will have similar characteristics and therefore their intensity is expected to be (i) a function of the distance (between the positions i and i') and (ii) proportional to the targets intensity, p_i^d. The filter uses these characteristics to separate peaks derived from noise, from peaks derived from other fragments. The distance Δ (in nt) between a target and noise candidate position is defined in Equation (1), where $|\ldots|$ is the absolute value operator. Δ will always be larger than 0 because a target is not compared with itself.

$$\Delta = |i - i'|, \quad \text{if } i \neq i'. \tag{1}$$

Because intensities of noise artefacts are proportional to the intensity of the target, a weight matrix is used to define the area border (Fig. 4). The weights are derived from the probability density function of a normal distribution with a standard deviation of 3, for all integer values $0 \leq x \leq 15$. To rescale densities to weights, the densities were divided through the density for $x = 0$ (0.1329808). To improve performance for peaks with a very low number of corresponding reads, the densities for a Δ of 1, 2, 3 and 4 were changed to 1.0. The complete weight matrix ω is available in the source code.

For each target peak, the filter evaluates whether any other peaks fall within the range that can be expected by noise in both directions ($d = 5'$ or $d = 3'$) as follows (Fig. 4):

A Sort $\mathbf{c}^d$ on corresponding intensities in descending order.

B For each $i \in \mathbf{c}^d$ target peak, remove corresponding noise artefacts:

B.i For all $i' \in \mathbf{c}^d_{i' \neq i}$ noise candidate peaks, find ω_Δ and define whether the candidate is noise or belongs to a separate fragment by evaluating Equation (2). If the equation is *true*, the candidate peak is considered to be a noise artefact of the target; immediately discard candidate $\mathbf{c}_{i'}^d$. If it is *false*, the candidate peak is not considered to be a noise artefact and must be retained.

$$p_{i'}^d \leq (\omega_\Delta \times p_i^d). \tag{2}$$

As a result of the filter, $\mathbf{c}^{5'}$ and $\mathbf{c}^{3'}$ may have shrunk and their respective lengths k and m may have become smaller.

2.2.5 Reconstruction

The peaks are expected to be the actual boundaries of fragments. Because start and stop positions do not have a direct one-to-one relationship with each other, a trace back is required to reconstruct the fragments. Because the number of predicted start (k) and end (m) positions is not necessarily equal, it is convenient to start reconstruction from direction d with $min(k, m)$ candidate positions, and

Algorithm 1 Peak detection

```
q^d ← {p^d, 0}                                   ▷ Input
n' = n + 1
α ← 0.1                                          ▷ Noise threshold
val_previous, val_max, pos_max ← 0               ▷ Init
c^d ← {}                                         ▷ Output
j ← 1                                            ▷ Output iterator
for 1≤i≤n' do
    if q_i^d > val_previous then
        if q_i^d > val_max then
            pos_max ← i
            val_max ← q^d_pos_max
        end if
    else if q_i^d < val_previous then
        if pos_max > 0 and (α × q_i^d) < val_max then
            c_j^d ← pos_max                      ▷ Call peak
            val_max ← 0                          ▷ Reset for next peak
            j ← j + 1
        end if
    end if
    val_previous ← q_i^d
end for
```

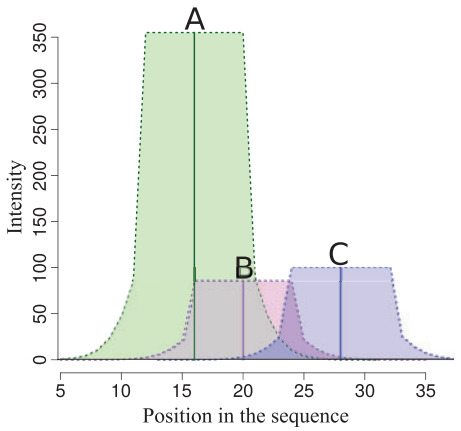

Fig. 4. Illustration of the filter. Vector $\mathbf{c}^d = \{16, 20, 28\}$ contains three predicted peaks referred to as peak A, B and C. For each peak, the intensity is indicated with vertical solid lines, at positions 16, 20 and 28 in green, purple and blue, respectively. Peak A (350 corresponding reads) has the highest intensity, followed by C (100) and B (85). Using a top-down approach in terms of intensity, the algorithm starts with filtering the noise artefacts that belong to peak A. The borders that separate noise from true fragments are indicated with dashed lines. Peaks within the coloured areas are marked as noise. For peak A, this is the light-green area. Any other peak within this region (peak B; solid purple line) gets discarded. Thus, B is expected to noise of peak A. In the next iteration, the noise of next top peak C is taken into account. Because no other peaks fall in its corresponding blue area, none will be discarded. Since peak B is discarded already, only peak A and C remain

find for each position the most likely corresponding position d'. Direction d is defined in Equation (3), and d' is its complement.

$$d = \begin{cases} 5'(start\ positions) & \text{if } k \leq m \\ 3'(end\ positions) & \text{if } m > k \end{cases}. \quad (3)$$

Important information required for reconstruction is the expected length of reads that were used for detecting a peak, given in $l^{5'}$ and $l^{3'}$.

Indeed:

- A fragment that starts at position i is expected to have its end i^* close to: $i^* \approx i + l_i^{5'}$.
- A fragment that ends at position i is expected to have its start i^* close to: $i^* \approx i - l_i^{3'}$.

The number of reads that correspond to a start position is expected to be close to the number of reads that define the end position: $p_i^d \approx p_{i'}^{d'}$. Thus, the reconstruction process needs a balance between (i) the expected position and (ii) the expected intensity of the counter position. This is achieved by conjoining an associated start and end position into a fragment as follows:

A Sort $\mathbf{c}^d$ based on corresponding intensities in descending order.
B For all $i \in \mathbf{c}^d$ candidate positions find expected counter position i^*.

B.i For all candidate counter positions $i' \in \mathbf{c}^{d'}$, the goal is to determine the counter position which has the optimal trade-off between a small distance with the expected counter position and a small difference in intensity. This is achieved by solving of Equation (4). In the equation 0.09 is an arbitrary chosen weight that forms the linear balance between distance and intensity. A predicted fragment is determined with its start position: min (i, j) and end position: max (i, j). After reconstruction, positions i and j are discarded from $\mathbf{c}^d$ and $\mathbf{c}^{d'}$, respectively.

$$j = \max_{i'} ((1 - 0.09 \times |i^* - i'|) \times p_{i'}^{d'}), \quad \text{for all } i' \in \mathbf{c}^{d'}. \quad (4)$$

3 Results

3.1 Validation of FlaiMapper performance

3.1.1 miRBase

To get an impression of FlaiMapper's performance, its predictions for corresponding miRNAs detected in dataset SRP002175 were compared with miRNA annotations in miRBase 20 (Kozomara and Griffiths-Jones, 2011). Because all experiments in this dataset are generated under the same conditions, alignments to the same ncRNA from all 12 experiments were merged, to maximize resolution. Of the 1037 miRNAs annotated in miRBase, 169 lacked supporting reads, and were not included in the quality assessment (because they would influence the outcome negatively without assessing the algorithm itself). Of the remaining 868 miRNAs, FlaiMapper was not able to predict a fragment that overlaps an annotated miRNA only 21 times, with a corresponding sensitivity of $847/868 = 0.98$.

A detailed assessment was performed by measuring the offset between a predicted fragment and a miRNA annotation in miRBase (Fig. 5, top). We assume that miRBase provides the 'ground truth' in terms of miRNA annotations. The results show that the majority of FlaiMapper predictions are identical to miRBase annotations. Also, the decrease of the offset bars (Fig. 5, top) is symmetrical, indicating no systematic inconsistency. 89% of the predicted start positions are identical to the reference. When an offset of 1 nt is allowed, the ratio correctly predicted start positions increases to 95%. In contrast, 54% of the end positions are predicted identical to the reference. When an offset of 1 nt is allowed, this increases to 82%. In addition, their offset-bars descend slower. This indicates that estimation of start positions is more precise.

To get an impression of the influence of sequencing depth on accuracy of start and end positions corresponding to miRNA and miRNA* predictions, the number of corresponding reads (*intensity*) was plotted as a function of the offset for dataset SRP002175. Figure 6 illustrates that with the increase of sequencing depth, the offset for both start and end positions decreases. However, at identical intensity, end positions have higher offset than start positions and require deeper sequencing to achieve the same accuracy.

In addition, we analysed the performance of FlaiMapper on three supplementary datasets generated on other sequencing platforms. Performance was similar to the performance described above.

3.1.2 Existing software

Previous research reported a comparable method (Langenberger *et al.*, 2009). Its goal is to detect miRNA-offset-RNAs (moRNAs), fragments adjacent to pre-miRNAs. The authors also used the method to demonstrate its ability to discover miRNAs. The

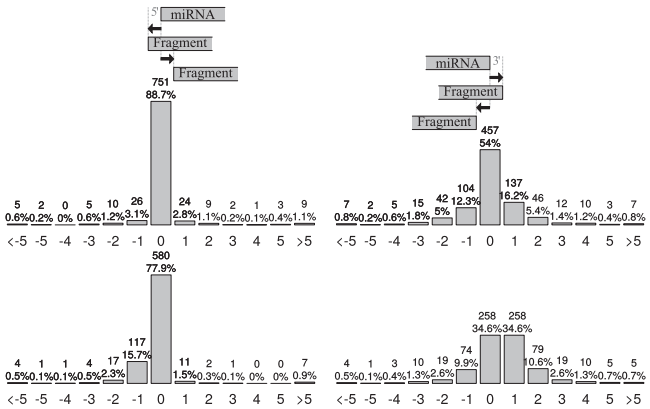

Fig. 5. Comparison between the predictions of miRBase 20 and FlaiMapper (top) and between miRBase and BlockBuster (bottom) on dataset SRP002175, indicating the offset of the start positions (left) and end positions (right). The vertical axes reflect the amount of predictions that correspond to a particular offset. The horizontal axes represent the offset between a predicted fragment and an annotated miRNA; exact matches are located at 0, offsets < 0 are predicted upstream the miRNA's boundaries and offsets > 0 –downstream

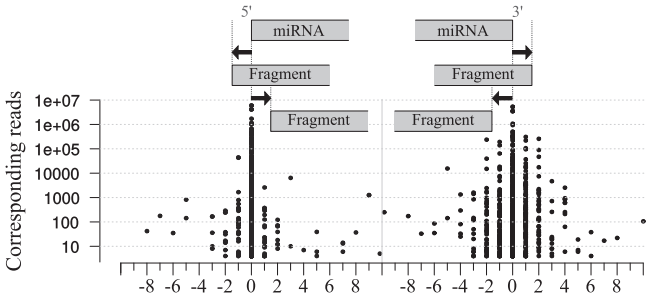

Fig. 6. Relation between sequencing depth and offset in predicted miRBase annotations for dataset SRP002175. The horizontal axis represents the offset for the start (left) and end (right) positions between a miRNA annotation and a predicted fragment. Predictions with exact matches are located at 0, offsets < 0 are predicted upstream the miRNA's boundaries and offsets > 0 –downstream. The vertical axis represents the number of reads corresponding to a start (left) or end (right) position. The figure indicates that the higher the number of corresponding reads, the lower the offset. Overall, the predicted start positions have a lower offset than the end positions, for the same sequencing depth

algorithm has no restrictions to a certain type of fragment, so its outcome should be comparable to FlaiMapper and therefore their performances can be compared with each other.

In contrast to our method, BlockBuster relies on the overall aligned read density of a fragment and transforms this into a normal distribution. Consequently, the prediction of the start and end positions are dependent on each other since they are derived from the same distribution. This implies that the alignments near a fragment's start and end position should have a symmetrical shape. BlockBuster's performance was tested on dataset SRP002175 and the alignments of the 12 corresponding experiments were merged and converted into the BED format for compatibility. BlockBuster was used with a variety of parameters where its *scale* parameter of 0.05 and *distance* of 26 were found to be rough estimates for the optimum. Optimum is defined by the lowest amount of root squared error of all predicted miRNAs where error is defined as the offset between a predicted miRNA and its miRBase annotation. The following is observed (Fig. 5, bottom):

- A lower sensitivity: $745/868 = 0.86$ (compared with 0.98 in FlaiMapper).
- A lower accuracy:
  - The number of start positions identical to miRBase is 78% compared with FlaiMapper's 89%. When an offset of 1 nt is allowed, both tools show comparable accuracy of 95%.
  - The number of end positions identical to miRBase is 34% compared with FlaiMapper's 54%. When an offset of 1 nt is allowed, 79% is predicted correctly compared with 82% using FlaiMapper.
- The offset bars of the start position decrease asymmetrically.
- The predictions are shifted; for the start positions, there is an overhang towards the pre-miRNAs 5′-end and for the end positions there is an overhang towards the pre-miRNAs 3′-end, indicating that the predicted fragments are on both sides systematically longer than the miRBase annotations.

3.2 Fragment analysis

We used FlaiMapper to detect fragments originating from sncRNAs other than pre-miRNAs on datasets SRP002175, SRP006788 and supplementary datasets SRP028959, SRP034013 and SRP041082. To maximize resolution, alignments of dataset SRP002175, SRP034013 and SRP041082 were merged. For SRP006788 and SRP028959, experiments were analysed individually to investigate possible influence of specific RNA processing-related treatments. The numbers of predicted fragments, categorized per type of precursor, are given in Table 1. The ratios of predicted fragments per precursor type were used for principal component analysis (Fig. 7). The largest difference between fragment profiles was observed between datasets SRP002175, SRP034013 and SRP041082. Since they are from different tissues and experiments, this is expected. Sub-conditions within dataset SRP006788 that are taken from AGO pool-downs showed nearly identical fragment profiles. In addition, samples of which nuclear RNA was extracted from independent HeLa experiments processed on different sequencers, also show very similar fragment profiles.

Table 1 shows that the AGO-pool down samples of dataset SRP006788 have a relatively high proportion of fragments derived from pre-miRNAs compared with the other samples in the dataset. This observation is consistent with the known association of miRNAs with AGO proteins (Fabian and Sonenberg, 2012). On the same time, it also suggests that fragments derived from other precursor types than pre-miRNAs are not associated with AGO to the same extend. Taken together, this supports the biological context of the FlaiMapper-derived fragment profiles.

3.3 Sequence logos

To show that the outcome of FlaiMapper can be used to explore characteristics of sncdRNAs like sequence motifs, fragments were analysed for over-represented pre- or suffixes using sequence logo plots (Schneider and Stephens, 1990) (Fig. 8). The analysis on pre-miRNA-derived fragments did not indicate over-represented motifs. Although it must be stated that the number of predicted fragments derived from C/D-box snoRNAs is lower than for pre-miRNAs, the analysis confirms that the C-box is over-represented (Brameier *et al.*, 2011). It also shows that sequences of H/ACA-box snoRNA-derived fragments located at the 3′-half of the precursor, most often contain the suffix ACANNN, where ACA is the precursor's ACA-box and N can represent any nucleotide. On the 5′-half of the H/ACA box the fourth is preferentially occupied by a G/C. However, due to the mild bit score and the low number of used fragments, this observation should be interpreted with caution. Because of the highly conserved sequences and the high number of genomic copies, tRNAs were excluded from motif analysis.

4 Discussion

We set out a method able to extract and annotate ncRNA fragments, because such annotations can be helpful in further high-throughput

Table 1. Summary of predicted fragments on datasets SRP002175 (Stark *et al.*, 2010), SRP006788 (Valen *et al.*, 2011), SRP028959 (Bai *et al.*, 2014), SRP034013 (Contrant *et al.*, 2014) and SRP041082 (Selth *et al.*, 2014)

Dataset	Type	pre-miR	SNORD	SNORA	tRNA	SCARNA	MISC
Total (in ncRNA reference)	ncRNAs	1386	264	106	451	23	39
SRP002175 (Pigment): RNA extracted Total cellular RNA (12 merged experiments)	ncRNAs	645	141	51	381	16	29
	fragments	947 (37.2%)	202 (7.9%)	56 (2.2%)	1210 (47.5%)	26 (1.0%)	107 (4.2%)
SRP006788 (HeLa) Total cellular RNA	ncRNAs	463	92	20	359	8	28
	fragments	680 (26.1%)	140 (5.4%)	24 (0.9%)	998 (38.3%)	11 (0.4%)	755 (28.9%)
SRP006788 (HeLa): Total cellular RNA after RRP40 core subunit depletion	ncRNAs	455	108	29	367	16	32
	fragments	686 (24.2%)	181 (6.4%)	34 (1.2%)	1104 (38.9%)	24 (0.8%)	806 (28.4%)
SRP006788 (HeLa): RNA pool-down after AGO immunoprecipitation	ncRNAs	415	19	14	151	6	15
	fragments	560 (39.7%)	30 (2.1%)	15 (1.1%)	208 (14.7%)	9 (0.6%)	590 (41.8%)
SRP006788 (HeLa): pool-down from RRP4 core depleted cells after AGO immunoprecipitation	ncRNAs	393	30	16	155	4	18
	fragments	517 (38.6%)	42 (3.1%)	17 (1.3%)	209 (15.6%)	6 (0.4%)	550 (41.0%)
SRP006788 (HeLa): RNA pool-down after after XRN immunoprecipitation	ncRNAs	513	129	65	301	18	33
	fragments	738 (25.9%)	294 (10.3%)	112 (3.9%)	760 (26.6%)	47 (1.6%)	903 (31.6%)
SRP006788 (HeLa) RNA was extracted from the nucleus	ncRNAs	451	129	57	281	16	33
	fragments	649 (27.3%)	280 (11.8%)	87 (3.7%)	519 (21.9%)	32 (1.3%)	806 (34.0%)
SRP028959 (HeLa): SRR954957 total cell small RNA preparation	ncRNAs	168	10	0	167	2	4
	fragments	203 (30.8%)	15 (2.3%)	0 (0%)	229 (34.7%)	2 (0.3%)	210 (31.9%)
SRP028959 (HeLa): SRR9558 nuclear small RNA preparation	ncRNAs	146	28	0	104	3	6
	fragments	169 (33.3%)	33 (6.5%)	0 (0%)	126 (24.8%)	4 (0.8%)	176 (34.6%)
SRP028959 (HeLa): SRR954959 cytoplasmic small RNA preparation	ncRNAs	162	2	0	183	1	6
	fragments	197 (29.1%)	2 (0.3%)	0 (0.0%)	274 (40.4%)	1 (0.1%)	204 (30.1%)
SRP034013 (B cells): Total cellular RNA (three merged experiments)	ncRNAs	1012	213	99	409	23	39
	fragments	2230 (42.4%)	670 (12.7%)	384 (7.3%)	1478 (28.1%)	105 (2%)	391 (7.4%)
SRP041082 (Prostate): Total cellular RNA (two merged experiments)	ncRNAs	803	230	90	381	21	38
	fragments	1454 (37.9%)	552 (14.4%)	201 (5.2%)	1222 (31.9%)	99 (2.6%)	304 (7.9%)

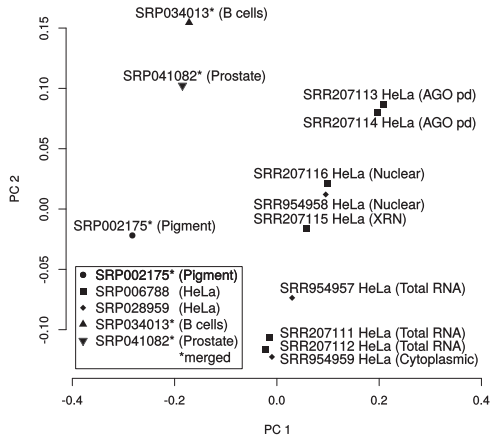

Fig. 7. The first two components of principal component analysis applied on percentages of predicted fragments (Table 1) comprise 92% of the variance. The circle represents the profile of merged dataset SRP002175 (pigment cells, Illumina GA II); squares individual experiments of dataset SRP006788 (HeLa cells, Illumina GA II); diamonds individual experiments of SRP028959 (HeLa cells, Ion Torrent PGM); triangle pointing up dataset SRP034013 (B cells, Illumina HiSeq2000) and the triangle pointing down dataset SRP041082 (prostate cells, Illumina HiSeq2000). The last three datasets are addressed in the Supplementary information. Datasets corresponding to separate tissue types demonstrate unique fragment profiles. Experiments with HeLa cells taken under similar circumstances group together; total RNA samples from independent datasets SRP006788 and SRP028959 cluster closely together

research. We designed FlaiMapper, a computer program to predict ncRNA fragments using small RNA-seq alignments. Benchmarking indicated that FlaiMapper is able to predict 97.8% of the miRNAs with corresponding reads. 95% of the miRNAs 5′-end and 82% of the 3′-end predictions were concordant with miRBase annotations. For this analysis, data from the Illumina Genome Analyser II was used. A similar accuracy was observed for sequencing data derived from the Ion Torrent PGM and Illumina HiSeq2000 (Supplementary information), indicating FlaiMapper can perform well on data from different platforms. We demonstrated that FlaiMapper performs better than existing similar software (Langenberger *et al.*, 2009).

FlaiMapper predicts fragments by looking at the most common start and end positions in alignments. It can be argued whether the most common start and end positions should indeed provide the evidence for the prediction of a fragment, since the most common read could be used instead. However, the most common start and end positions should usually be covered by a higher number of reads. This corresponds to a higher resolution, which is especially advantageous for the prediction of fragments with a low read coverage, and should therefore also be more robust towards noise. Together with the demonstrated high performance, this implies that predictions based on start and end position densities provide a more appropriate solution for fragment annotation.

The weights used in the filtering step are based on a normal distribution with an arbitrary chosen σ. These parameters probably find their optimum in relation with sequencing protocols, 5′/3′-end-specific processing factors or different families of fragments. Therefore, once there is a better understanding of the processing of

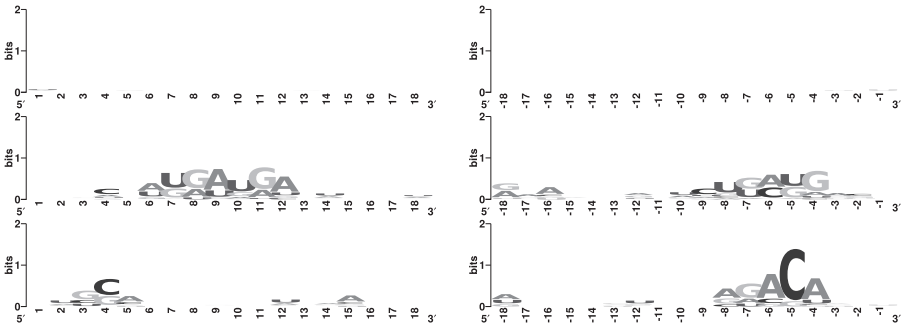

Fig. 8. Sequence logos (Schneider and Stephens, 1990) of the fragments located on the 5′- (*left*) and the 3′-end (*right*) of the precursor ncRNA from the predictions on dataset SRP002175. Only fragments with their centre located at ≤40% of the precursor are used for prefix analysis and fragments located at ≥60% for suffix analysis. (top) Pre-miRNA fragments: no common prefix or suffix motif (based on 520 prefixes and 427 suffixes). (middle) C/D-box snoRNA fragments: (middle left) the prefix motif UGAUGA is found more often around the sixth and seventh nucleotide (based on 130 prefixes). (middle right) The suffix UGAUG is found more often around the −eighth nucleotide (based on 72 suffixes). (bottom) H/ACA-box snoRNA fragments: (left) the fourth nucleotide of the prefix appears to be G/C enriched, although the bit score is mild (based on 37 prefixes). (bottom right) The suffixes are enriched with the motif ACA from the −sixth until the −fourth nucleotide (based on 19 suffixes)

fragments, it is recommended to spend effort in optimizing these parameters.

Additional analysis indicated that FlaiMapper's performance in miRNA annotation is positively correlated to sequencing depth. Predicted 3′-ends of miRNAs have a larger offset compared with miRBase annotations than the 5′-ends, even for the same sequencing depth. The higher variability of the miRNAs 3′-ends has earlier been reported (Kozomara and Griffiths-Jones, 2011). In addition, research on the classification of sncRNAs indicated that metrics corresponding to the variability in the alignment are indeed higher for the 3′-end in miRNAs (Leung *et al.*, 2013). They were able to indicate that different levels of variability correspond to specific types of sncRNAs. Possible explanations could be RNA post-processing or RNA editing. This means that alignments over the entire fragment can be asymmetrical because of a larger variation observed at miRNAs 3′-ends. Since BlockBuster assumes reads to be symmetrically distributed over a fragment, this might explain why (i) its accuracy is lower, (ii) its predictions are longer and (iii) shifted with respect to miRBase.

Although it seems counter-intuitive that an ncRNA can produce different fragments that originate from an overlapping region, there are situations where overlapping fragments can be expected. This can be stressed by recalling the not fully understood tRNA processing mechanism(s), where tRNA halves and tRFs spanning similar regions have been reported. Therefore, FlaiMapper has no restriction to the prediction of overlapping fragments, similar to the method of Langenberger *et al.* (2009).

Sequence logos indicated that the ACA box, as part of the the ACANNN suffix, is over-represented and position specific in fragments derived from the 3′-halve of the H/ACA-box snoRNAs. The analysis also confirmed that the C-box of C/D box snoRNAs is over-represented in corresponding fragments. Yet, this result may be biased by the existence of multiple, highly homologous, genomic copies of certain C/D-box snoRNAs such as HBII-52 and HBII-85.

Fragment characteristics can play an important role in finding associations with their processing mechanism. For example, although the larger variability of the alignments at the 3′-end of miRNAs affects performance, it clearly indicates a difference in the processing of miRNAs ends. Characteristics such as 5′- and 3′-end entropy have been successfully used in the classification of sncRNAs (Leung *et al.*, 2013; Yuan and Sun, 2013). Using such characteristics on the fragment level, for example for clustering or classification, might provide new insights into the processes of production, functioning or degradation of fragments or indicate a possible sub-grouping. The future in-depth analysis of sncdRNAs will require more comprehensive datasets with higher sequencing depth and more statistical power.

5 Conclusion

The lack of a sncdRNAs annotation is a short coming in small RNA-seq analysis. To overcome this, we designed the computer program FlaiMapper. FlaiMapper has a high performance in predicting miRNA boundaries, but can be used for the annotation of any type of sncdRNA. Examination of FlaiMapper-predicted sncdRNAs indicated different type specific characteristics: 5′/3′-end-specific variability in miRNAs, associations between AGO and relative fragment profiles in dataset SRP006788 and a position-specific sequence motif in a subset of the H/ACA-box fragments. These characteristics indicate that FlaiMapper is a good starting point for the downstream analysis of small RNA sequencing experiments.

Acknowledgement

The authors would also like to thank Bas Pigmans for his work on sequencing alignment methodology.

Funding

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n°201438 and from the research programme ALW-VENI Grant 863.12.014 financed by the Netherlands Organisation for Scientific Research (NWO).

Conflict of interest: none declared.

References

Askarian-Amiri,M.E. *et al.* (2011) Snord-host RNA zfas1 is a regulator of mammary development and a potential marker for breast cancer. *RNA*, **17**, 878–891.

Bai,B. *et al.* (2014) Small RNA expression and deep sequencing analyses of the nucleolus reveal the presence of nucleolus-associated microRNAs. *FEBS Open Bio.*, **4**, 441–449.

Brameier,M. *et al*. (2011) Human box c/d snoRNAs with miRNA like functions: expanding the range of regulatory RNAs. *Nucleic Acids Res.*, **39**, 675–686.

Chen,C.-J. and Heard,E. (2013) Small RNAs derived from structural non-coding RNAs. *Methods*, **63**, 76–84. [Diversity of the non-coding transcriptomes revealed by RNA-seq technologies.]

Contrant,M. *et al*. (2014) Importance of the RNA secondary structure for the relative accumulation of clustered viral microRNAs. *Nucleic Acids Res.*, **42**, 7981–7996.

Dong,X.-Y. *et al*. (2009) Implication of snoRNA {U50} in human breast cancer. *J. Genet. Genomics*, **36**, 447–454.

Ender,C. *et al*. (2008) A human snoRNA with microRNA-like functions. *Mol. Cell*, **32**, 519–528.

Fabian,M.R. and Sonenberg,N. (2012) The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC. *Nat. Struct. Mol. Biol.*, **19**, 586–593.

Friedlander,M.R. *et al*. (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotech.*, **26**, 407–415.

Henras,A.K. *et al*. (2004) RNA structure and function in C/D and H/ACA s(no)RNPs. *Curr. Opin. Struct. Biol.*, **14**, 335–343.

Kishore,S. *et al*. (2010) The snoRNA mbii-52 (snord 115) is processed into smaller RNAs and regulates alternative splicing. *Hum. Mol. Genet.*, **19**, 1153–1164.

Kozomara,A. and Griffiths-Jones,S. (2011) mirbase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**(Suppl. 1), D152–D157.

Langenberger,D. *et al*. (2009) Evidence for human microRNA-offset RNAs in small RNA sequencing data. *Bioinformatics*, **25**, 2298–2301.

Leung,Y.Y. *et al*. (2013) Coral: predicting non-coding RNAs from small RNA-sequencing data. *Nucleic Acids Res.*, **41**, E137.

Lin He,G.J.H. (2004) MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.*, **5**, 522–531.

Martens-Uzunova,E.S. *et al*. (2013) Beyond microRNA novel RNAs derived from small non-coding RNA and their implication in cancer. *Cancer Lett.*, **340**, 201–211.

Mei,Y.-P. *et al*. (2012) Small nucleolar RNA 42 acts as an oncogene in lung tumorigenesis. *Oncogene*, **31**, 2794–2804.

Ono,M. *et al*. (2011) Identification of human miRNA precursors that resemble box c/d snoRNAs. *Nucleic Acids Res.*, **39**, 3879–3891.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Scott,M.S. *et al*. (2012) Human box c/d snoRNA processing conservation across multiple cell types. *Nucleic Acids Res.*, **40**, 3676–3688.

Selth,L.A. *et al*. (2014) Human seminal fluid as a source of prostate cancer-specific microRNA biomarkers. *Endocr. Relat. Cancer*, **21**, L17–L21.

Sobala,A. and Hutvagner,G. (2011) Transfer RNA-derived fragments: origins, processing, and functions. *Wiley Interdiscip. Rev. RNA*, **2**, 853–862.

Stark,M.S. *et al*. (2010) Characterization of the melanoma miRNAome by deep sequencing. *PLoS One*, **5**, e9685.

Thompson,D.M. and Parker,R. (2009) Stressing out over tRNA cleavage. *Cell*, **138**, 215–219.

Valen,E. *et al*. (2011) Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes. *Nat. Struct. Mol. Biol.*, **18**, 1075–1082.

Yamasaki,S. *et al*. (2009) Angiogenin cleaves tRNA and promotes stress-induced translational repression. *J. Cell Biol.*, **185**, 35–42.

Yuan,C. and Sun,Y. (2013) RNA-code: a noncoding RNA classification tool for short reads in NGS data lacking reference genomes. *PLoS One*, **8**, e77596.