Identification and Classification of Small RNAs in Transcriptome Sequence Data

D. Langenberger[1], C.I. Bermudez-Santana[1,2], P.F. Stadler[1,3*], S. Hoffmann[1]

[1] *University Leipzig*
Chair of Bioinformatics & Interdisciplinary Center for Bioinformatics,
Haertelstrasse 16-18,
D-04107 Leipzig, Germany
E-mail: steve@bioinf.uni-leipzig.de

[2] *Department of Biology*
Universidad Nacional de Colombia
Carrera 45, No. 26-85, Edificio Uriel Gutierrez
D.C., Colombia

[3] *Max-Planck-Insitute for Mathematics in Sciences (MPI-MIS)*
Inselstrasse 22,
D-04103 Leipzig, Germany

Current methods for high throughput sequencing (HTS) for the first time offer the opportunity to investigate the entire transcriptome in an essentially unbiased way. In many species, small non-coding RNAs with specific secondary structures constitute a significant part of the transcriptome. Some of these RNA classes, in particular microRNAs and snoRNAs, undergo maturation processes that lead to the production of shorter RNAs. After mapping the sequences to the reference genome specific patterns of short reads can be observed. These read patterns seem to reflect the processing and thus are specific for the RNA transcripts of which they are derived from. We explore here the potential of short read sequence data in the classification and identification of non-coding RNAs.

Keywords: High throughput sequencing; read patterns; small RNA processing; small RNA classification; machine learning

1. Introduction

Whole-Transcriptome analysis of many species and cell types reveals massive expression of non-coding RNA. It is widely believed that non-coding RNAs act as regulators upon transcription and translation. Recent investigations of whole RNA cDNA-Libraries based on high throughput sequencing (HTS) have shown that these libraries contain both primary and processed transcripts. Over the last years, several classes of small RNAs with a length of about 20nt have been discovered. The most prominent classes are miRNAs, piRNAs, and various variants of endogenous siRNAs.[1,2] In addition, small RNAs have been found to be associated with transcription start and stop sites of mRNAs.[3–5] Several studies reported that well-known ncRNA loci are also processed to give rise to small RNAs. MicroRNA precursor hairpins, for instance, are frequently processed to produce additional "off-set RNAs" (moRNAs) that appear to function like mature miRs. These moRNAs were discovered in *Ciona intestinalis*,[6] where they form an abundant class of processing products. At much lower expression levels they can also be found in the human transcriptome.[7] Specific cleavage and processing of tRNAs was observed in the fungus *Aspergillus fumigatus*[8] and later also found in human short read sequencing data.[9] Small nucleolar RNAs (snoRNAs) are also widely used as a source for specific miRNA-like short RNAs.[10–12] The same holds true for vault RNAs.[13,14] Little is known, however, about the mechanisms of these processing steps and their regulation. Here, we show that the production of short RNAs is correlated with RNA secondary structure and therefore exhibits features that are characteristic for individual ncRNA classes. The specific patterns of mapped HTS reads thus may be suitable to identify and classify the ncRNAs from which they are processed. We explore here to what extent such an approach is feasible in practise.

*P.F.S has external affiliations with the Fraunhofer Institute IZI, the Institute for Chemistry of the University of Vienna, and the Santa Fe Institute.

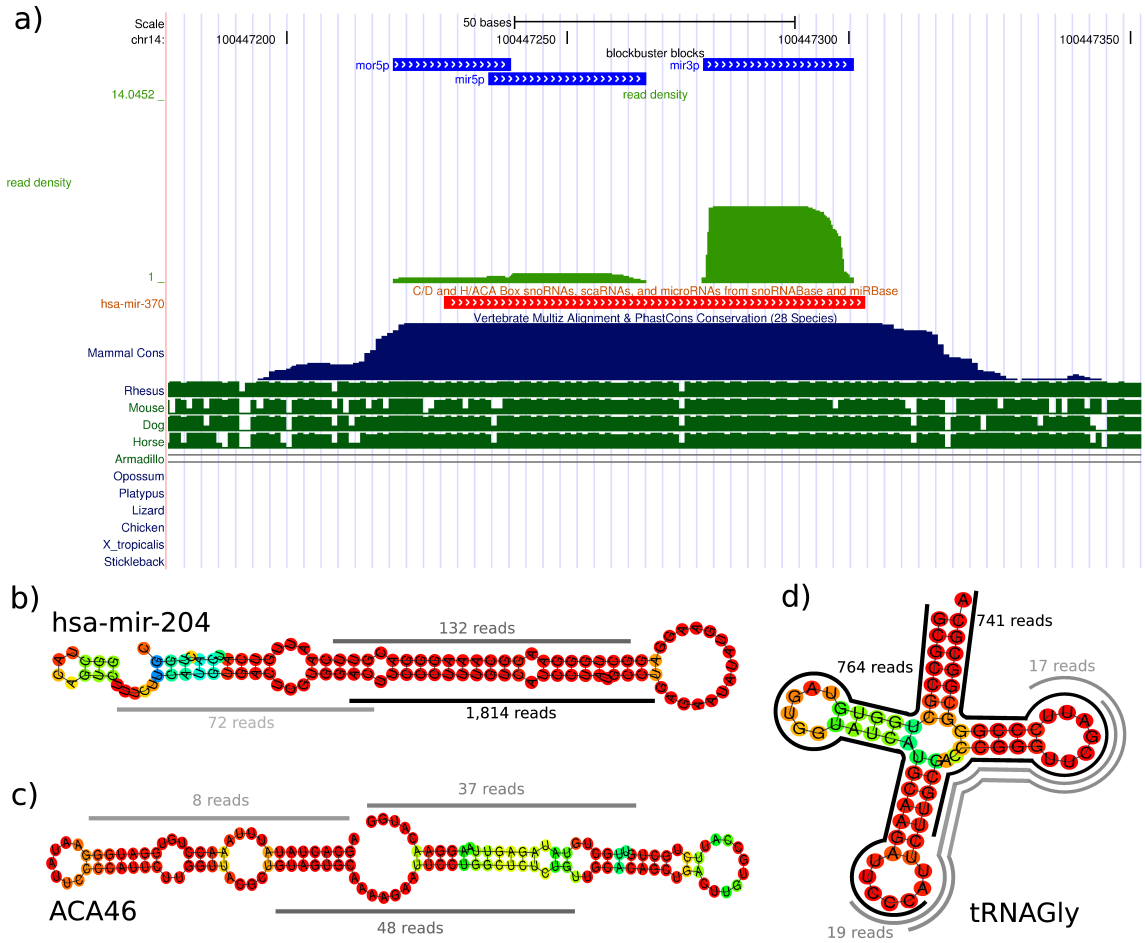

Fig. 1. Non-coding RNAs exhibit specific block patterns. (a) Distribution of short reads at the hsa-mir-370 locus. There are three clearly distinct blocks of reads: they correspond to moR (5'-end), miR* (center) and miR (3'-end) transcripts. The conservation pattern is shown below. (b) The class of miRNAs often shows a block pattern of two or three separated blocks. (c) snoRNAs tend to have miRNA-like mature and star blocks at their 5' and 3' hairpins with minor overlaps, while a series of overlapping blocks is striking for the tRNA class (d).

The first step towards this goal is the identification of ncRNA loci from a collection of mapped HTS reads. We have recently developed the tool `blockbuster`[7] to simplify this task in genome-wide analyses. The program merges mapped HTS reads into *blocks* based on their location in the reference genome (Fig. 1a). After the assembly of blocks, specific block patterns for several ncRNA classes can be observed. For example, miRNAs typically show 2 blocks corresponding to the miR and miR* positions (Fig. 1b). A similar processing can be observed for snoRNAs (Fig. 1c). On the other hand, tRNAs show more complex block patterns with several overlapping blocks (Fig. 1d).

2. Methods

The dataset analyzed here was produced according to standard small RNA transcriptome sequencing protocols in the context of other projects and will be published in that context. In brief, total RNA was isolated from the frozen prefrontal cortex tissue using the TRIzol (Invitrogen, USA) protocol with no modifications. Low molecular weight RNA was isolated, ligated to the adapters, amplified, and sequenced following the Small RNA Preparation Protocol (Illumina, USA) with no modifications. All small RNAs, 17-28nt long, were mapped to the human genome (NCBI36.50 Release of July 2008) using `segemehl`,[15] a method based

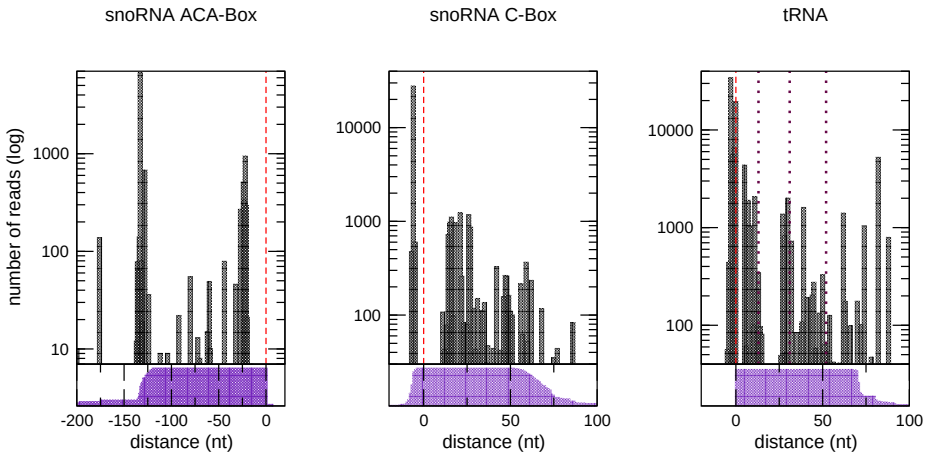

Fig. 2. HTS data reflects structural properties of ncRNAs. Upper panels show the number of 5'-ends of mapped HTS reads (bars) relative to aligned the 5'-ends (dashed vertical lines) of 27 ACA boxes (left), 81 CD boxes (middle) and 87 tRNAs (right). The area in the lower panel represents the number of boxes and tRNAs present at the distance relative to their aligned start sites. In accordance with Taft et al.[12] a sudden and sharp increase of 5'-ends is seen just upstream of the snoRNAs' ACA and C boxes, resp., indicating that read blocks reflect structural properties of snoRNAs. Similarily, the number of 5'-ends increases just upstream of the tRNA and the relative start sites of its three loop regions (dotted lines). Downstream the start sites there is a sudden drop in the number of reads.

on a variant of enhanced suffix arrays that efficiently deals with both mismatches as well as insertions and deletions. We required small RNAs to map with an accuracy of at least 80% and only the best hit was selected. Reads mapping multiple times to the genome with an equivalent accuracy were discarded. After filtering the effective accuracy was $> 97\%$. Subsequently, all hits were sorted by their genomic position. Two reads were assigned to the same putative ncRNA locus, i.e. cluster, if separated by less than 100nt. Clusters consisting of less than 10 reads were discarded because of their low information content.

To detect specific expression patterns, we divided consecutive reads into blocks using `blockbuster`.[7] Here, we used a width parameter of $s = 0.5$, a value that requires blocks to be well separated to be recognizedas distinct. We required a cluster to have at least 2 blocks. In the following we refer to the number of reads comprised in a block as the *block height*. Using `blockbuster`, we identified 852 clusters across the whole human genome. This set comprises 2,538 individual blocks and 85,459 unique reads. 434 clusters were found within annotated ncRNA loci [`miRBase` v12 (727 entries), `tRNAscan-SE` (588 entries) and `snoRNAbase` v3 (451 entries)], see Tab. 1.

We then computed secondary structures (using `RNAfold`[16]) to assess the relationship of reads and structure. For each read, the base pairing probabilities were calculated for the sequences composed of the read itself and 50nt of flanking region both up- and downstream. These data were also collected separately for reads found within annotated miRNA, tRNA, and snoRNA loci, respectively.

In order to investigate whether the short reads patterns carry information on the particular ncRNA class from which they orginate, we selected three distinct ncRNA classes and performed a random forest classification:[17,18] tRNAs ($n = 87$), miRNAs ($n = 218$) and snoRNAs ($n = 129$). Based on a visual inspection of the mapped reads, ten features were selected to train the random forest model: the number of blocks within

Table 1. In total 434 of 852 clusters were found within regions of annotated miRNA, tRNA and snoRNA loci. While the average number of blocks is similar for all three ncRNA classes, the number of reads differs significantly among the classes.

RNA class	source	loci found	blocks/cluster (mean)	reads/cluster (median)
microRNAs	miRBase v12	218	2.42 ± 1.04	4535.33
tRNAs	tRNAscan SE	87	3.22 ± 1.92	183.95
snoRNAs	snoRNAbase v3	129	2.60 ± 1.66	127.5

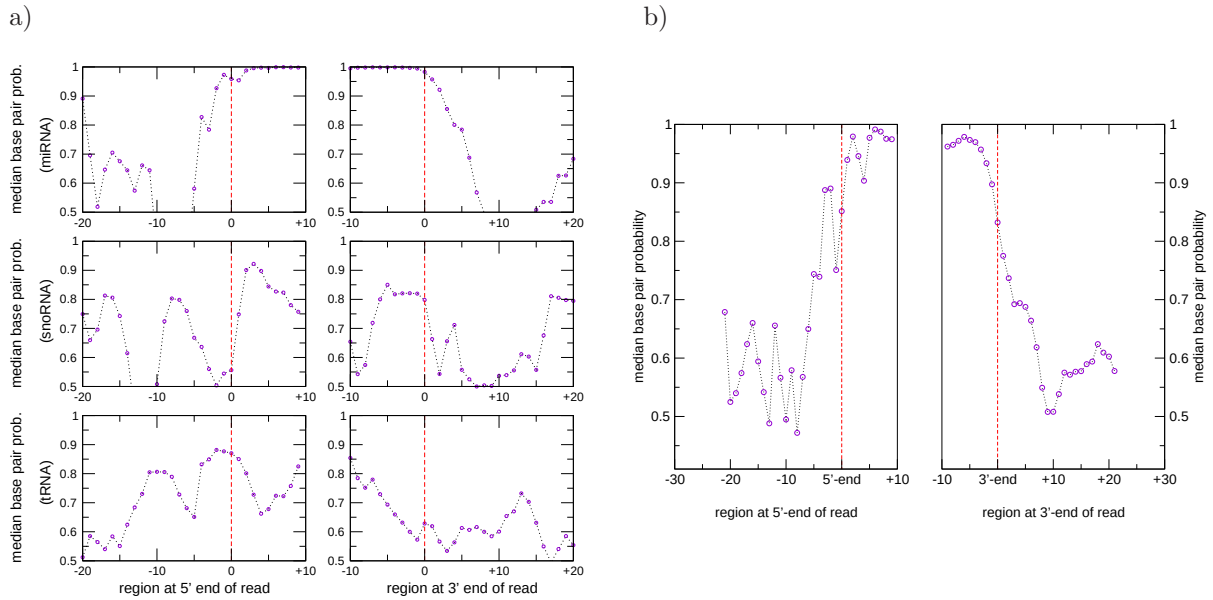

Fig. 3. Base pairing probabilities increase at the 5'-end and decrease at the 3'-end of reads mapped to ncRNA loci. (a) The 3'- and 5'-ends are indicated by dashed lines. The median base pairing probability increases sharply at the 5'-ends (upper left) and drops again at the 3'-ends of reads mapped to miRNA loci (upper right). A similar – but attentuated – effect is observed for snoRNAs (middle panel) and tRNAs (lower panel). (b) The median base pairing probabilities at 5'- (left panel) and 3'- ends (right panel) for all reads within the 852 clusters. The 5'- and 3'-ends are indicated by dashed vertical lines.

a cluster (blocks), the length of a cluster (length), the number of nucleotides covered by at least two blocks (nt overlap), the number of overlapping blocks (block overlap), the maximum, minimum and the mean block height (max, min and mean height) in a cluster as well as the maximum, minimum and the mean distance between consecutive blocks (max, min and mean distance).

3. Results

The 5'-ends of reads arising from known snoRNAs preferentially map just upstream of the C- and ACA-boxes. This indicates the correlation of mapping patterns with processing steps and thus with structural properties of snoRNAs (Fig. 2). Based on earlier findings that miRNA-like products are derived from snoRNAs[12] and the observation that miRNA transcripts tend to have higher blocks (Tab. 1), the two peaks shown in the Figure 2 (left) probably represent small RNAs produced from the 5'- and 3'-hairpins of the HACA (see also Fig. 1c). CD-snoRNAs show, in contrast to the HACA-snoRNAs, only a single prominent peak at the 5'-end (Fig. 2, middle). An increased number of 5'-ends of HTS reads is also observed just upstream of loops of tRNAs (Fig. 2 (right)).

The pairing probabilities of bases covered by HTS reads are significantly increased (Fig. 3b). Just upstream the 5'-end of these reads, the median base pairing probability increases sharply and reaches a level of > 0.9. At the 3'-end the base pairing probability drops again. However, median base pairing probabilities of bases covered by the center of reads drop down to 70%. Although this effect is boosted by reads found within miRNA loci, it can also be observed unambiguously for reads within snoRNA and tRNA loci (Fig. 3a).

The observation that blocks reflect structural properties of ncRNAs was exploited to train a random forest classifier to automatically detect miRNAs, tRNAs and snoRNAs. After visual inspection of block patterns for some representatives of these classes, ten features were selected. Their evaluation reveals significant statistical differences among the chosen ncRNA classes (Fig. 4). As expected, the number of reads mapped to miRNA loci (minimum and maximum block height) clearly distinguishes miRNAs from other ncRNA classes. In contrast to tRNAs and snoRNAs the maximum block distance of miRNAs shows a very narrow distribution

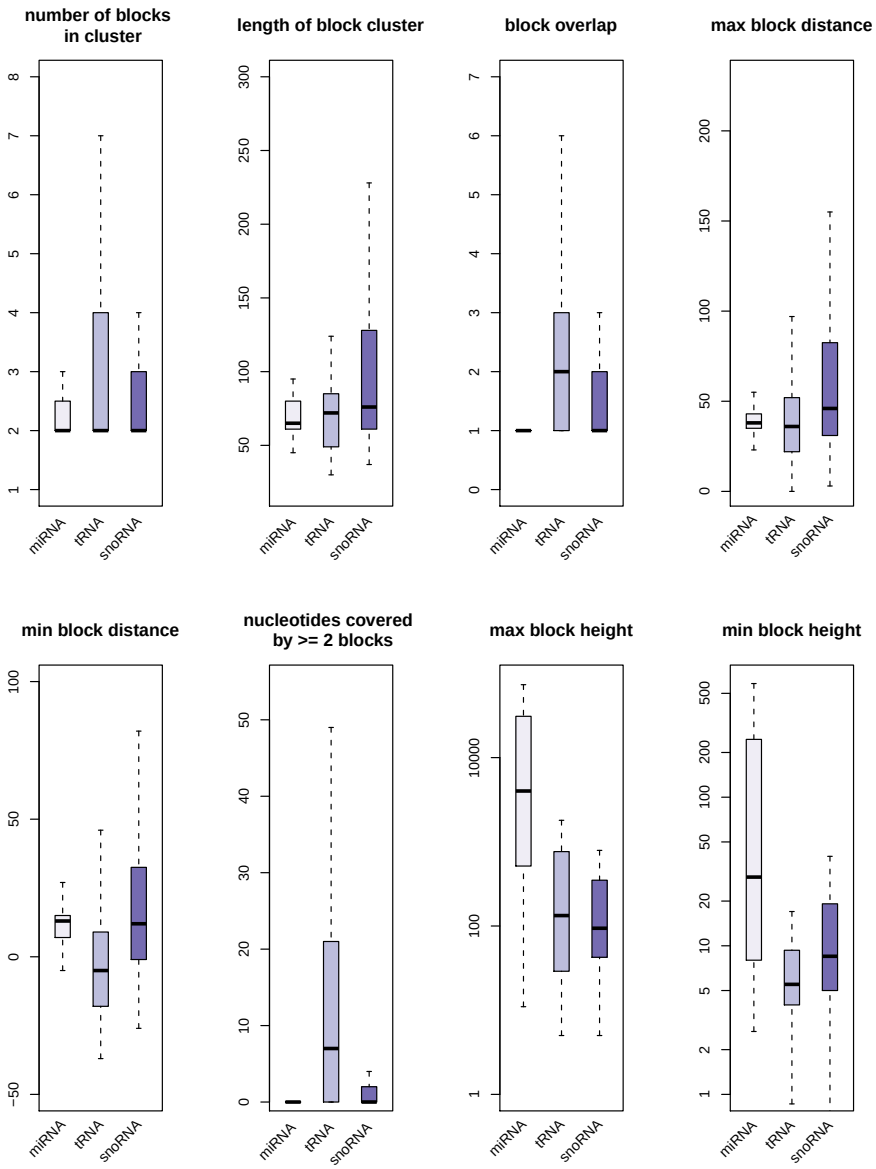

Fig. 4. Box plots for 8 different features selected to train the random forest classifier. The number of reads mapped to miRNA loci alone (max block height and min block height) effectively distinguish miRNAs from other ncRNAs. Likewise, the distribution of block distances seems to be a specific feature for miRNAs. Compared to other regions, tRNA loci frequently show block overlaps of two or more blocks. The minimum block distance shows a median overlap of ≈5nt for blocks in within tRNA loci. SnoRNAs typically have longer block distances than the other classes.

around 40nt, reflecting the distance between miR and miR* transcripts. Furthermore, the class of tRNAs frequently shows more block overlaps than snoRNAs and miRNAs. The distance of blocks is an important feature for snoRNAs: the maximum block as well as the minimum block distance is higher compared to both tRNAs and miRNAs.

The random forest model was repeatedly trained with randomly chosen annotated loci and different training set sizes in order to determine predictive values (PPV) and recall rates. For the training sets comprising 150 clusters the random forest model shows a positive predictive value > 0.7 for all three ncRNA classes. The recall rate for miRNAs is well above 80%. However, with a rate of ≈ 0.55 the recall of snoRNAs

Table 2. Positive predictive values (PPV) and recall rates for training sets of size 150 and 250. For each set size means, medians and standard deviations are calculated from 20 randomly sampled training sets.

		PPV		recall	
	#loci	mean	sdev	mean	sdev
Training size 250					
all	852	0.889	0.015	0.799	0.015
miRNA	227	0.932	0.020	0.918	0.023
tRNA	287	0.860	0.040	0.683	0.046
snoRNA	143	0.819	0.032	0.694	0.060
other	195				
Training size 150					
all	852	0.827	0.020	0.698	0.027
miRNA	236	0.900	0.027	0.847	0.041
tRNA	348	0.755	0.044	0.580	0.062
snoRNA	115	0.733	0.057	0.525	0.071
other	153				

and tRNAs is relatively poor (Tab. 2). For larger training sets containing 250 clusters, the positive predictive value (PPV) is > 0.8 for all classes. For miRNAs the classification achieves recall rates and PPVs of > 0.9. Likewise, the recall rates for snoRNAs and tRNAs rise to 0.7-level. In summary, for both training set sizes and all classes the random forest model achieves PPVs and recall rates of ≈ 0.8.

We applied the classifier to unannotated ncRNA loci. A list of miRNA, snoRNA, and tRNA candidates predicted is available from the supplementary page (`http://www.bioinf.uni-leipzig.de/~david/PSB/`). This resource includes the original reads, their mapping accuracy and their mapping location in machine-readable formats. Furthermore, the page provides links to the UCSC genome browser to visualize the block patterns. For microRNAs and snoRNAs, we also indicate whether the candidates are supported by independent ncRNA prediction tools.

The 29 miRNA predictions contained 3 miRNAs (hsa-mir1978, hsa-mir-2110, hsa-mir-1974) which have already been annotated in the most recent miRBase release (v.14), as well as a novel member of the mir-548 family, and another locus is the human ortholog of the bovine mir-2355. In addition, we found two clusters antisense to annotated miRNA loci (hsa-mir-219-2 and hsa-mir-625). Such antisense transcripts at known miRNA loci have been reported also in several previous publications,[20–22,24] lending further credibility to these predictions.

For the tRNAs and snoRNAs we expect a rather large false positive rate. The 78 tRNA predictions are indeed contaminated by rRNA fragments, but also contain interesting loci, such as sequence on Chr.10 that is identical with the mitochondrial tRNA-Ser. `SnoReport`,[23] a specific predictor for HACA snoRNAs based on sequence and secondary features, recognizes 44 (20%) of our 223 snoRNAs predictions.

Short RNAs are processed from virtually all structured ncRNAs. Complex read patterns are observed, for instance, for the 7SL (SRP) RNA and the U2 snRNA. Y RNAs, which have a panhandle-like secondary structure produce short reads mostly from their 5' and 3' ends, see Fig. 5.

4. Discussion

In extension of previous work establishing that various ncRNA families produce short processing products of defined length,[6,9,12] we show here that these short RNAs are generated from highly specific loci. The dominating majority of reads from short RNAs originates from base paired regions, suggesting that these RNAs are, like miRNAs, produced by Dicer or other specific RNAases. For example, specific cleavage products have recently been reported for tRNAs.[19] In this work we show that the block patterns are characteristic for three different ncRNA classes and thus suitable to recognize additional members of these classes. For instance, the random forest trained with loci annotated in the mirBase v12 predicted five additional miRNAs reported in the mirBase release 14 as well as two "antisense microRNA".

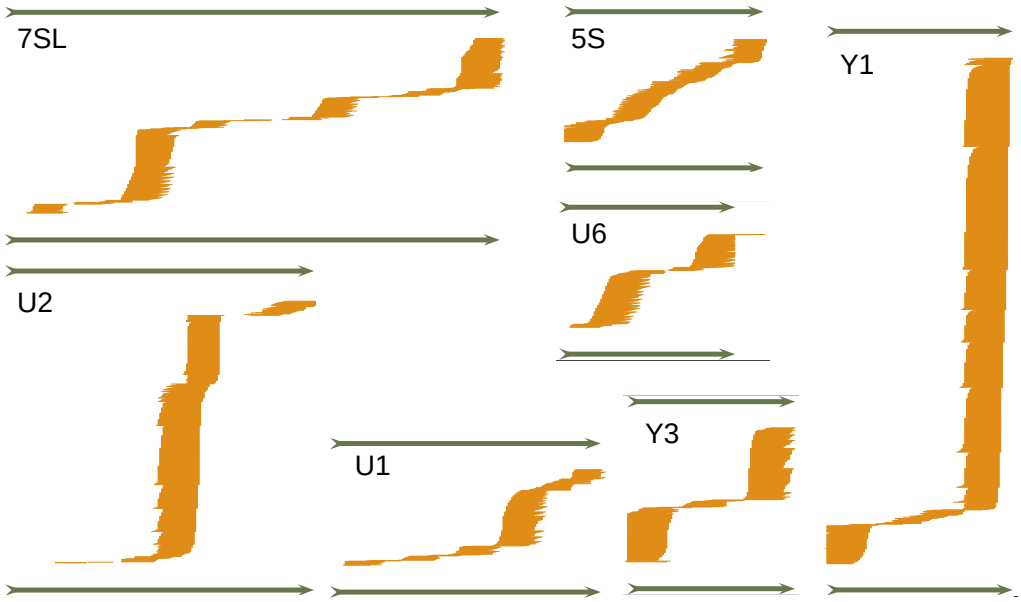

Fig. 5. Short reads are produced from a wide variety of structured ncRNAs. Green arrows indicate the ncRNA gene and its reading direction, individual short reads are shown as orange lines. The same scale is used for all examples.

The block patterns for the evaluated ncRNAs show some interesting characteristics. Although miRNA loci accumulate far more reads than tRNAs and snoRNA loci, the reads are extremely unevenly distributed across the blocks. For tRNAs we observe series of overlapping blocks that are specific enough to separate this class from other classes with high positive predictive values.

However, the successful prediction of miRNAs heavily depends on the height of the blocks, i.e. the number of reads that map to a potential locus. In comparision tRNAs and snoRNAs show significantly lower positive predictive values and recall rates. A relatively large training set is required to achieve PPV's $> 80\%$. Obviously, the selection of appropriate features is crucial for the success of the presented approach. Hence, the random forest classifier is not sufficient as it stands and the identification of other characteristic features is subject to further research. The integration of secondary structure information of cluster regions is likely to enhance the prediction quality.

Beyond the classification by means of soft computing methods, this survey shows that HTS block patterns bear the potential to greatly improve and simplify ncRNA annotation. Given the striking relationship of HTS reads and secondary structure for some ncRNA classes, block patterns may also be used in the future to directly infer secondary structure properties of non-coding RNAs from transcriptome sequencing data. In this context, although not shown here, block patterns may also help to identify new classes of RNAs directly from transcriptome sequencing data.

5. Acknowledgements

This work was supported by the European Union (EDEN, contract 043251), the Deutsche Forschungsgemeinschaft (SPP-1174) (P.F.S), a PhD scolarship of the DAAD-AleCol program (C.B-S.), a formel.1 grant of the Medical Faculty of the University of Leipzig and the Freistaat Sachsen under the auspices of the LIFE project.

References

1. D. Moazed. *Nature* **457**, 413-420 (2009).
2. A. Tanzer *et al.* in *Evolutionary Genomics*, G. Caetano-Anolles (ed.), **in press** (2009).

3. P. Kapranov *et al.*, *Science* **316**, 1484-1488 (2007).
4. R.J. Taft *et al.*, *Nat. Genetics* **41**, 572-578 (2009).
5. R.J. Taft *et al.*, *Cell Cycle* **8**, 2332-2338 (2009).
6. W. Shi *et al.*, *Nat. Struct. Mol. Bio.* **16**, 183-189 (2009).
7. D. Langenberger *et al.*, *Bioinformatics* 10.1093/bioinformatics/btp-419, (2009).
8. C. Jöchl *et al. Nucleic Acids Res.* **36**, 2677-2689 (2008).
9. H. Kawaji *et al.*, *BMC Genomics* **9**, 157 (2008).
10. C. Ender *et al.*, *Mol. Cell* **32**, 519-528 (2008).
11. A. Saraiya *et al.*, *PLoS Pathog.* **4**, e1000224 (2009).
12. R.J. Taft *et al.*, *RNA* **15**, 1233-1240 (2009).
13. P.F. Stadler *et al.*, *Mol. Biol. Evol.* **26**, 1975-1991 (2009).
14. H. Persson *et al.*, *Nat. Cell. Biol.* doi:10.1038/ncb1972 (2009).
15. S. Hoffmann *et al.*, *PLoS Comp. Biol.* **5**, e1000502 (2009).
16. I.L. Hofacker *et al.*, *Bioinformatics* **22**, 1172-1176 (2006).
17. I. Witten *Data Mining: practical machine learning tools and techniques*, 2nd edn., (2005).
18. L. Breiman, *Machine Learning* **45**, 5-32 (2001).
19. D.M. Thompson *et al.*, *Cell* **138**, 215-219 (2009).
20. E.A. Glazov *et al.*, *PLoS One* **4**, e6349 (2009).
21. A. Stark *et al.*, *Genes & Development* **22**, 8-13 (2008).
22. W. Bender *Genes & Development* **22**, 14-19 (2008).
23. J. Hertel *et al.*, *Bioinformatics* **24**, 158-164 (2008).
24. D.M. Tyler *Genes & Development* **22**, 26-36 (2008).