Nematode microRNA evolution

Rina Ahmed[1,2], Zisong Chang[1], Claudia Langnick[1], Wei Chen[1] and Christoph Dieterich[1]

[1]Berlin Insitute for Medical Systems Biology at the Max Delbrück Center for Molecular Medicine Berlin, Robert-Roessle-Str. 10, 13125 Berlin, Germany
[2]BIMSB/NYU International PhD Programme

Abstract

Background: A major goal in evolutionary research is to understand the genetic changes and variations in the genomes of all species that underlie the emergence of evolutionary novelties. Comparative studies have revealed that morphological diversification requires evolutionary modifications of gene regulatory mechanisms[1]. Small non-coding RNAs, including microRNAs (miRNAs), play important role in post-transcriptional regulation by complementary binding to the 3' untranslated regions of target genes. miRNAs are important for a broad range of biological processes, such as organism development and disease[2]. To get further insight into the process of adaptation, genus-wide small RNA analyses in terms of conservation and species-specificity are essential.

Results: We performed deep sequencing of small non-coding RNAs with SOLiD SREK in 10 different nematodes: *Caenorhabditis elegans*, *Koeneria sudhausi*, *Diplogasteroides magnus* and 7 species belonging to the genus *Pristionchus*. These nematodes diverged up to 430 million years ago[3]. In *C.elegans*, we identified ~90% of previously known miRNAs. In addition, we found 16 novel expressed miRNA gene candidates based on miRDeep2[4] analysis on the sequenced genome. Comparing our list of detected miRNAs with data from Kato *et al.*[5] showed a significant overlap with all six development stages of hermaphrodites and young adult males.

Similar analyses are planned using the *P.pacificus* genome. The remaining 8 nematode species have not been sequenced to date. Therefore, an alternative procedure needs to be established. A simple approach would be to identify conserved miRNAs based on related species. Since, we are also interested in species-specific miRNA genes, we developed a kmer based miRNA prediction strategy for species without genome data based on our *C.elegans* small RNA data set.

Conclusion: Our results showed that our prediction method works with an area under the ROC curve (AUC) of 0.8788. However, the precision decreases with an exponential decay. Therefore, our method could be improved by taking into account the complementary kmer sequences and the count distribution of all sequences per kmer bucket. Our analyses confirmed that SREK libraries enrich miRNAs. Therefore, our prediction strategy for miRNAs seems applicable to species without a genome sequence.

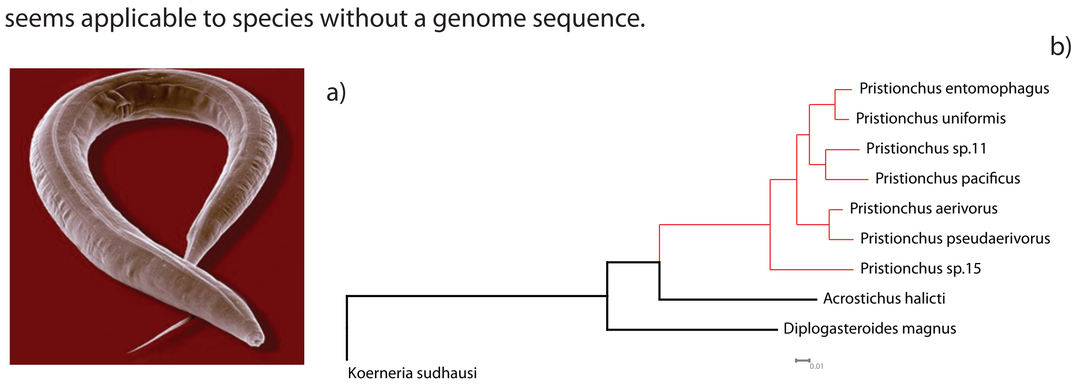


Fig. 1: a) *C.elegans* picture is courtesy of Jürgen Berger (MPI Tübingen). a) Phylogeny of 10 nematode species. *C.elegans* and *Pristionchus* diverged 280-430 million years ago.

miRNA expression profile

We detected 156 out of 174 known miRNA genes in our *C.elegans* library. The ranked list of expressed miRNAs in decreasing order shows an exponential decay in read numbers. Comparing the list of expressed miRNAs with data from Kato *et al.* showed a significant overlap with all six development stages of hermaphrodites and young adult males. Fisher's exact test p-values were smaller than 5.946×10^{-7}.

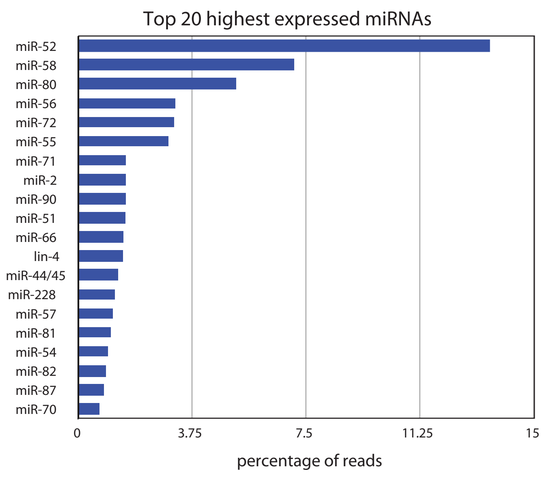


Fig. 2: Top 20 highest expressed miRNAs considering mapping with multiple hits. The percentage represents the fraction of reads of each miRNA compared to all reads mapped to the genome.

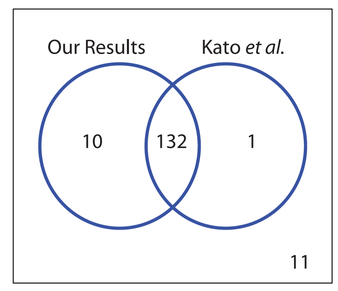


Fig. 3: Venn Diagram of expressed miRNAs in our data compared to Kato *et al.* (total of expressed miRNAs in all stages) based on miRBase v11. P-value of 1.973×10^{-7}.

De novo miRNA prediction using miRDeep2

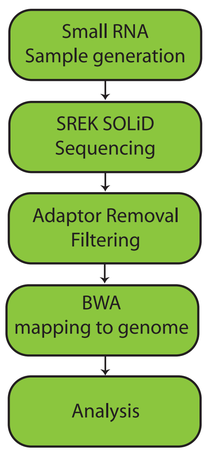
Novel miRNA prediction was performed using miRDeep2. miRDeep2 initially reported 26 candidate miRNAs (score cutoff of 1 recovering known miRNAs present in the data with 83% sensitivity). These candidates were manually curated to remove highly palindromic precursors or redundant sequences resulting in a refined set of 16 miRNA candidates. Figure 4 shows the structure of a candidate miRNA precursor. This miRNA fell exactly into the intron of a gene, strongly suggesting that this miRNA is a mirtron.

Fig. 4: Folded structure of predicted mirtron pre-miRNA. Red corresponds with the mature, purple with the star and yellow with the hairpin sequence.

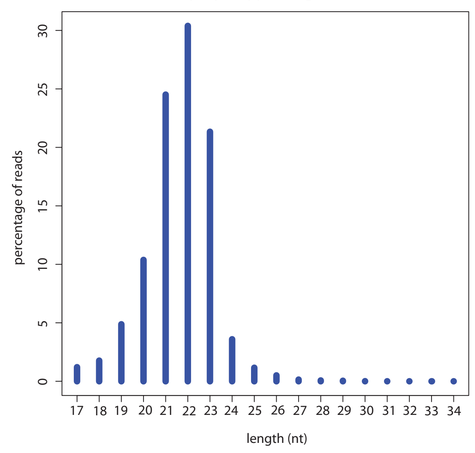
Small RNA mapping Pipeline (*C.elegans*)

We performed multiplex sequencing of 10 different nematodes on a 3/4 flowcell. We initially analyzed all *C. elegans* reads from 1/4 of a flowcell. 6,080,238 reads were left after the removal of 3' adaptor sequences, quality filtering and selecting reads of length >17 nucleotides. Reads were mapped to the genome with BWA[6] using an edit distance <= 2. 2,821,432 reads (46%) mapped uniquely to the genome. Allowing multiple hits (at most 10) 3,195,87 reads (53%) could be mapped.

Flowchart of small RNA mapping pipeline



Length distribution of *C.elegans* small RNA reads uniquely mapped to genome



Tab. 1: Reads mapped to non-coding RNAs.

	Uniquely		Multiple hits	
Category	No. of reads	No. of detected features	No. of reads	No. of detected features
mature (174)	2.422.498 (85.86%)	151 (86.78%)	2.548.472 (61.64%)	156 (89.66%)
star (21)	656 (0.02%)	10 (47.62%)	661 (0.02%)	12 (57.14%)
hairpin (174)	13.049 (0.46%)	123 (70.69%)	68.685 (1.66%)	130 (74.713%)
21U-RNA (15.341)	10.720 (0.38%)	4.667 (69.88%)	13.447 (0.33%)	5.514 (35.94%)
rRNA (23)	40.880 (1.45%)	8 (34.783%)	107.939 (2.61%)	19 (82.61%)
SL1 (12)	0	0	30 (0.0007%)	10 (83.33%)
SL2 (19)	167 (0.006%)	17 (89.47%)	227 (0.006%)	17 (89.47%)
snoRNA (133)	4.544 (0.16%)	90 (67.67%)	4.886 (0.12%)	100 (75.19%)
snRNA (90)	172 (0.006%)	32 (35.56%)	2.761 (0.067%)	56 (62.22%)
tRNA (630)	4.550 (0.16%)	139 (22.06%)	30.763 (0.74%)	424 (67.30%)
Rest	324.196 (11.49%)	-	1.356.729 (32.81%)	-
Total	2.821.432	-	4.134.600	-

Prefix based miRNA prediction strategy

Reads were binned using a prefix of *k=18* nucleotides. Every *18mer* bucket was assigned a read count *t* by summing the amount of reads belonging to this bucket (Figure 5b). *18mers* which perfectly matched with mature miRNA sequences (Figure 5c) were considered as true positives (TP). Non-matched *18mers* were considered as true negatives (TN). We constructed a Receiver Operator Characteristic curve (ROC) (Figure 6a) from evaluating 98 different read count thresholds, which were defined by the observed 98 different frequencies for known miRNAs. The ROC with an AUC of 0.8788 indicates a good performance of our prediction strategy. However, the number of negative examples (354.191) largely exceeds the number of positive examples (235). This is reflected in the exponential precision drop in the precision-recall curve (Figure 6b). The difference between the two curves is explained by the difference between precision (TP/TP+FP) and specificity (TN/TN+FP).

Our method could be improved by filtering out rRNAs and tRNAs primarily and by taking into account the complementary *kmer* sequences and the count distribution of all sequences within a *kmer* bucket.

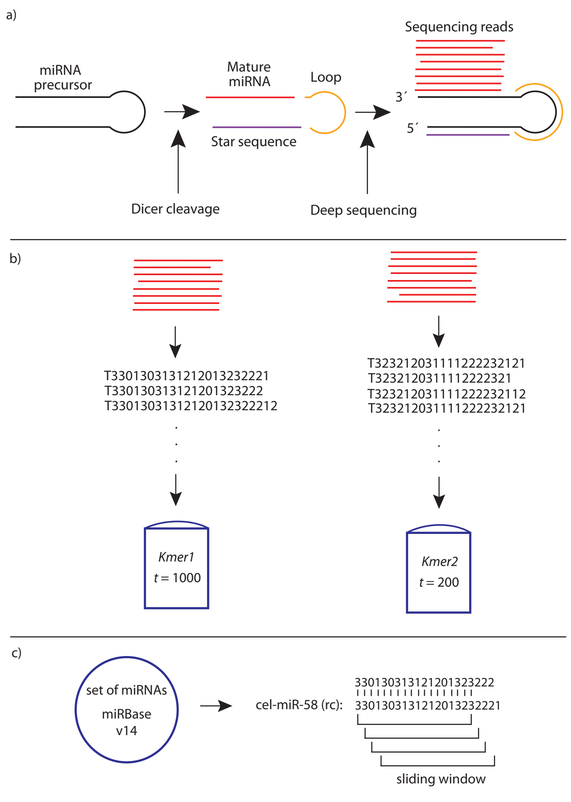


Fig. 5: a) Mature miRNAs are cleaved from hairpin precursors by Dicer, releasing three products (mature (red), star (purple) and loop (yellow) sequence). Deep sequencing detects one or more of these products. Resulting sequencing reads will map to specific positions on the hairpin precursor (modified picture taken from Friedländer *et al.*[7]). b) Binning of reads into *kmer* buckets using a prefix of *k=18* nucleotides. Each *kmer* bucket was assigned a summed up read count *t*. c) Perfectly aligning *18mers* with miRNAs using a sliding window of length *k*.

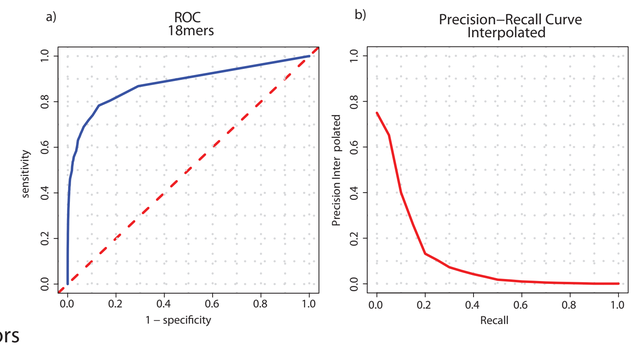


Fig. 6: a) ROC curve (AUC=0.8788) and b) precision-recall curve of *18mers* evaluating 98 different miRNA read count thresholds.

References

1 Chen K. and Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs. *Nat. Rev. Genet.* 8, 93-103 (2007)
2 Okamura K., Phillips M.D., Tyler D.M., Duan H., Chou Y. and Lai E.C. The regulatory activity of microRNA* species has substantial influence on microRNA and 3' UTR evolution. *Nat. Struct. Mol. Biol.* 15, 354-363 (2008)
3 Dieterich C. *et al.* The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. *Nat. Genetics* 40, 1193-1198 (2008)
4 Friedländer M., Mackowiak S. and Rajewsky N. miRDeep2 discovers novel mouse miRNAs from numerous and diverse genomic sources. Unpublished
5 Kato M., de Lencastre A., Pincus Z. and Slack F.J. Dynamic expression of small non-coding RNAs, including novel microRNAs and piRNAs/21U-RNAs, during *Caenorhabditis elegans* development. *Genome Biol.* 10, R54 (2009)
6 Li H. and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754-1760 (2009)
7 Friedländer M.R., Chen W. , Adamidi C., Maaskola J., Einspanier R., Knespel S. and Rajewsky N. Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol.* 26, 407-415 (2008)

Acknowledgements

We would like to thank Marc Friedländer and Sebastian Mackowiak for their help with miRDeep2.