Small RNA mapping Pipeline (*C. elegans*)

We performed multiple 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 between 17 and 1 nucleotides. Reads were mapped to the genome with IRM.par8 using an edit distance of \(-2\). 2,821,452 reads (46%) mapped uniquely to the genome. Allowing multiple hits at most (at 30%) 1,195,87 reads (53%) could be mapped.

### Small RNA 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 nematophytes and young adult males. Fisher’s exact test values were smaller than the 5.964 x 10^-11.

### De novo miRNA prediction using miDeep2

Novel miRNA prediction was performed using miDeep2. miDeep2 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 high homology 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.

### Prefix mapping based miRNA prediction strategy

Reads were binned using a prefix of k=18 nucleotides. Every 1/miRNA bucket was assigned a read count \( t \) by summing the amount of reads belonging to this bucket (Figure 5b). 1/miRNA buckets which perfectly matched with mature miRNA sequences (Figure 5c) were considered as true positives (TP). Non-matched 1/miRNA 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 true positive rates (TPR). The area under the curve (AUC) 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 recall (TP/TP+FN).

Our method could be improved by filtering out tRNAs and tRNAs predominantly by taking into account the complementary sequence distributions and the count distribution of all sequences within a miRNA bucket.

### References

2. 1 Chen K. and Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs.
3. 1 Chen K. and Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs.
4. 1 Chen K. and Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs.
5. 1 Chen K. and Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs.
6. 1 Chen K. and Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs.
7. 1 Chen K. and Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs.
8. 1 Chen K. and Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs.
9. 1 Chen K. and Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs.
10. 1 Chen K. and Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs.