# Detecting new miRNAs from deep sequencing data: A field study in worm

David Langenberger

25th TBI Winterseminar

# Outline

- Data
- Pre-Analysis
- miRNA prediction
  - Classifier
  - miRDeep
  - miRanalyzer
- Results

# Data

# C. elegans small RNA sequencing

## Set 1 (MDC - set)

### Data

- SOLiD sequenzer
- 6,080,238 reads
- 18-27 nt in length

### Mapping

- 3,288,430 reads mapped with BWA (54%)
- remapping by segemehl (377,594 tags)

# C. elegans small RNA sequencing

## Set 2 (Bartel - set)

**Data**(GSE5990)

- mixed-stage C. elegans
- 454 sequenzer
- 850,870 reads (181,668 tags)
- 18-36 nt in length

## Mapping

- 138,868 tags mapped by segemehl (76%)

# C. elegans small RNA sequencing

## Set 3 (Berezikov - set)

**Data**(GSE15169)

- mixed-stage C. elegans
- 454 sequenzer
- 181,849 reads (23,327 tags)
- 18-38 nt in length

## Mapping

- 22,277 tags mapped by segemehl (95%)

# Pre-Analysis

# ncRNA mapping results (MDC – set)

| ncRNA type | # of annotated ncRNAs | # of annotated ncRNAs with read support | % of annotated ncRNAs with read support |
|---|---|---|---|
| **all** | 6606 | 4137 | 62.6 |
| **microRNA** | 174 | 165 | 94.8 |
| **tRNA** | 631 | 626 | 99.2 |
| **snoRNA** | 133 | 114 | 85.7 |
| **snRNA** | 90 | 84 | 93.3 |
| **rRNA** | 25 | 25 | 100 |
| **21U-RNA** | 5356 | 2839 | 53.0 |
| **SL2 splice leader** | 8 | 8 | 100 |
| **others** | 189 | 149 | 78.8 |

# ncRNA mapping results (MDC – set)

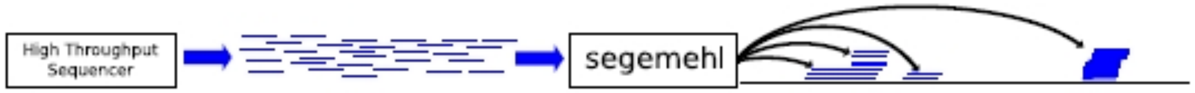
| ncRNA type | # of tags | # of reads | % of all tags | % of all reads |
|---|---|---|---|---|
| **all** | 68,385 | 1,045,428 | 18,1 | 31,8 |
| **miRNAs** | 6,625 | 857,689 | 1,8 | 26,1 |
| **tRNAs** | 10,93 | 93,789 | 2,9 | 2,9 |
| **snoRNAs** | 1,163 | 3,774 | 0,3 | 0,1 |
| **snRNAs** | 2,186 | 5,652 | 0,6 | 0,2 |
| **rRNAs** | 33,519 | 80,152 | 8,9 | 2,4 |
| **21U-RNAs** | 3,524 | 6,181 | 0,9 | 0,2 |
| **others** | 6,946 | 1,626 | 1,8 | 0,0 |

# miRNA prediction

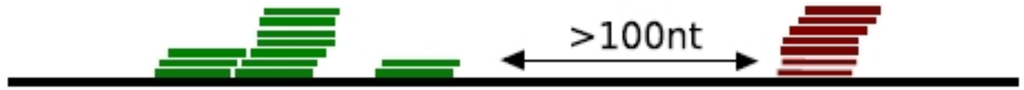
# Classifier

# Data preparation

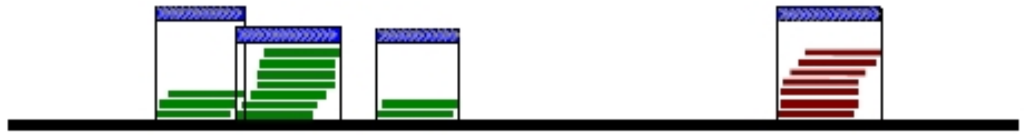
1. Map small RNAs to the human genome, using `segemehl`



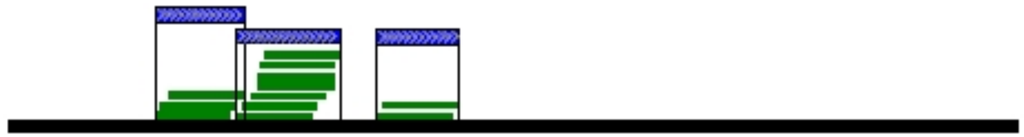
2. Cluster hits based on their genomic location (distance <100nt)



3. Divide consecutive reads into blocks, using `blockbuster`

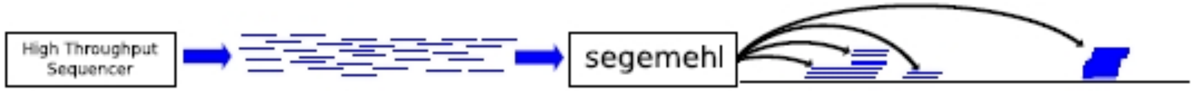


4. Discard clusters with <2 blocks and/or <10 reads (small information content)

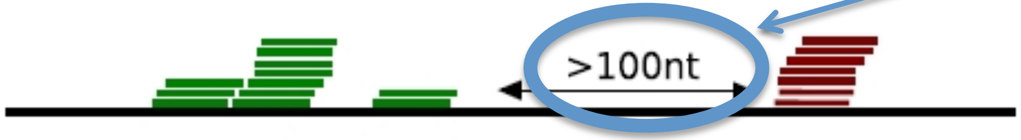


# Data preparation

1. Map small RNAs to the human genome, using `segemehl`

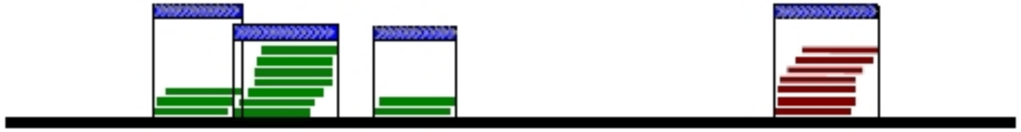


2. Cluster hits based on their genomic location (distance <100nt)

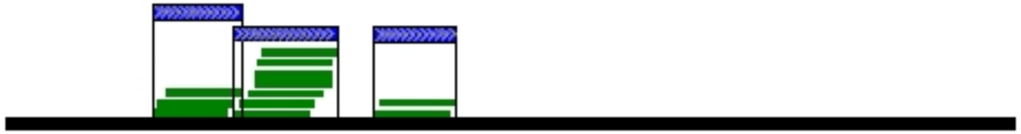


30 nt

3. Divide consecutive reads into blocks, using `blockbuster`

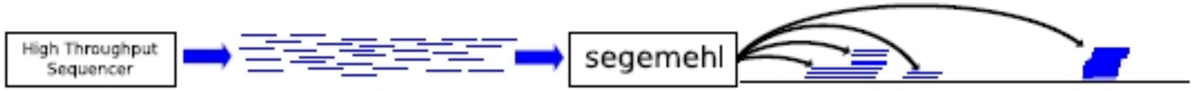


4. Discard clusters with <2 blocks and/or <10 reads (small information content)

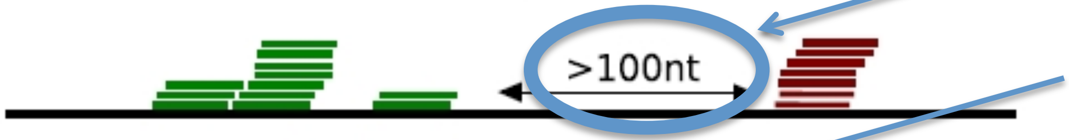


# Data preparation

1. Map small RNAs to the human genome, using `segemehl`



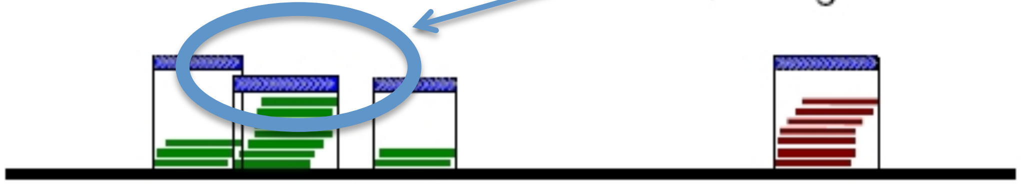
2. Cluster hits based on their genomic location (distance <100nt)



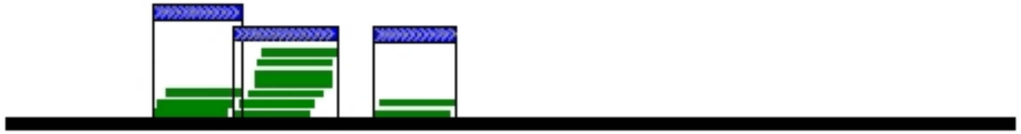
30 nt

Blocks are cut at positions with less than 80% read support

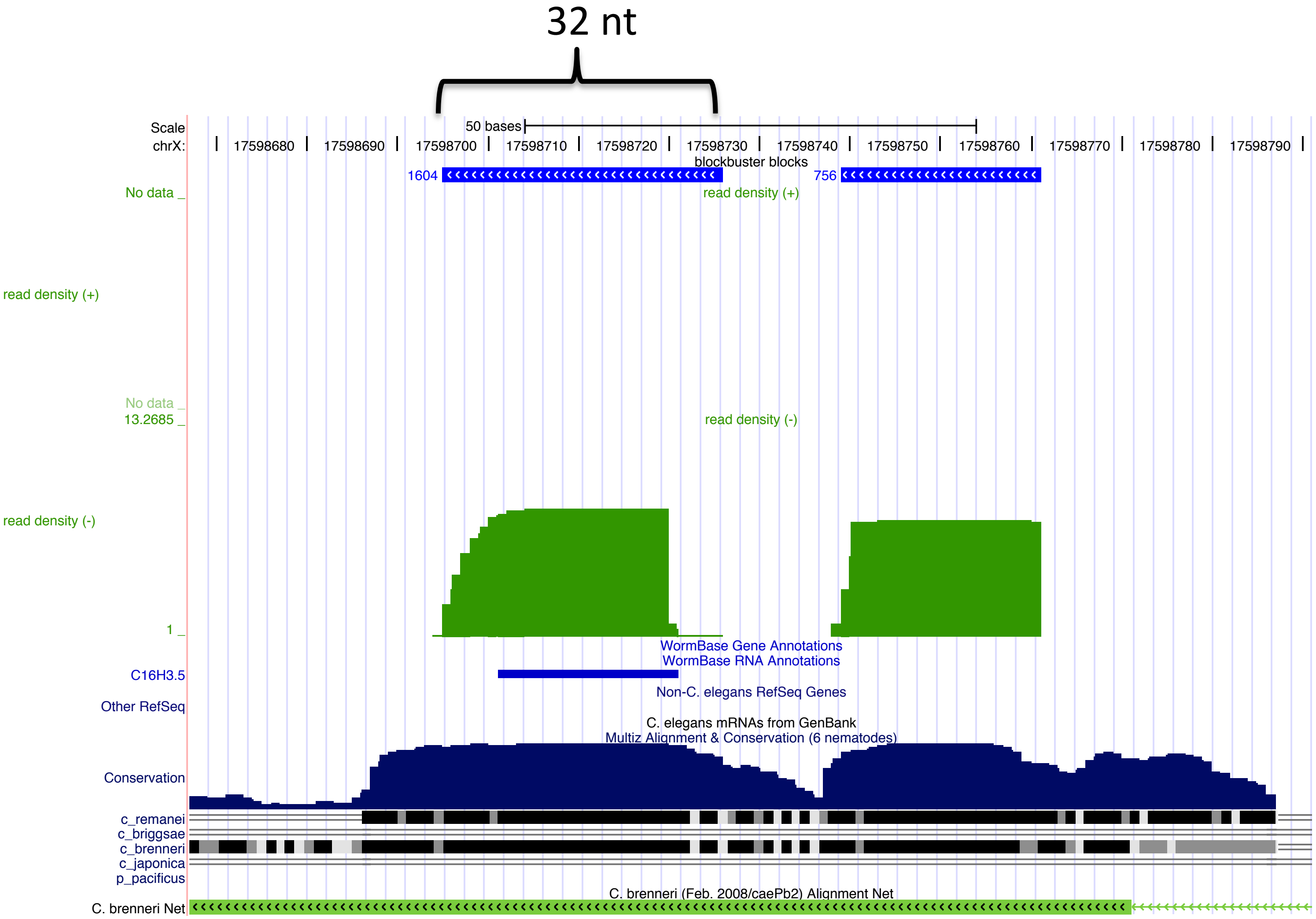
3. Divide consecutive reads into blocks, using `blockbuster`



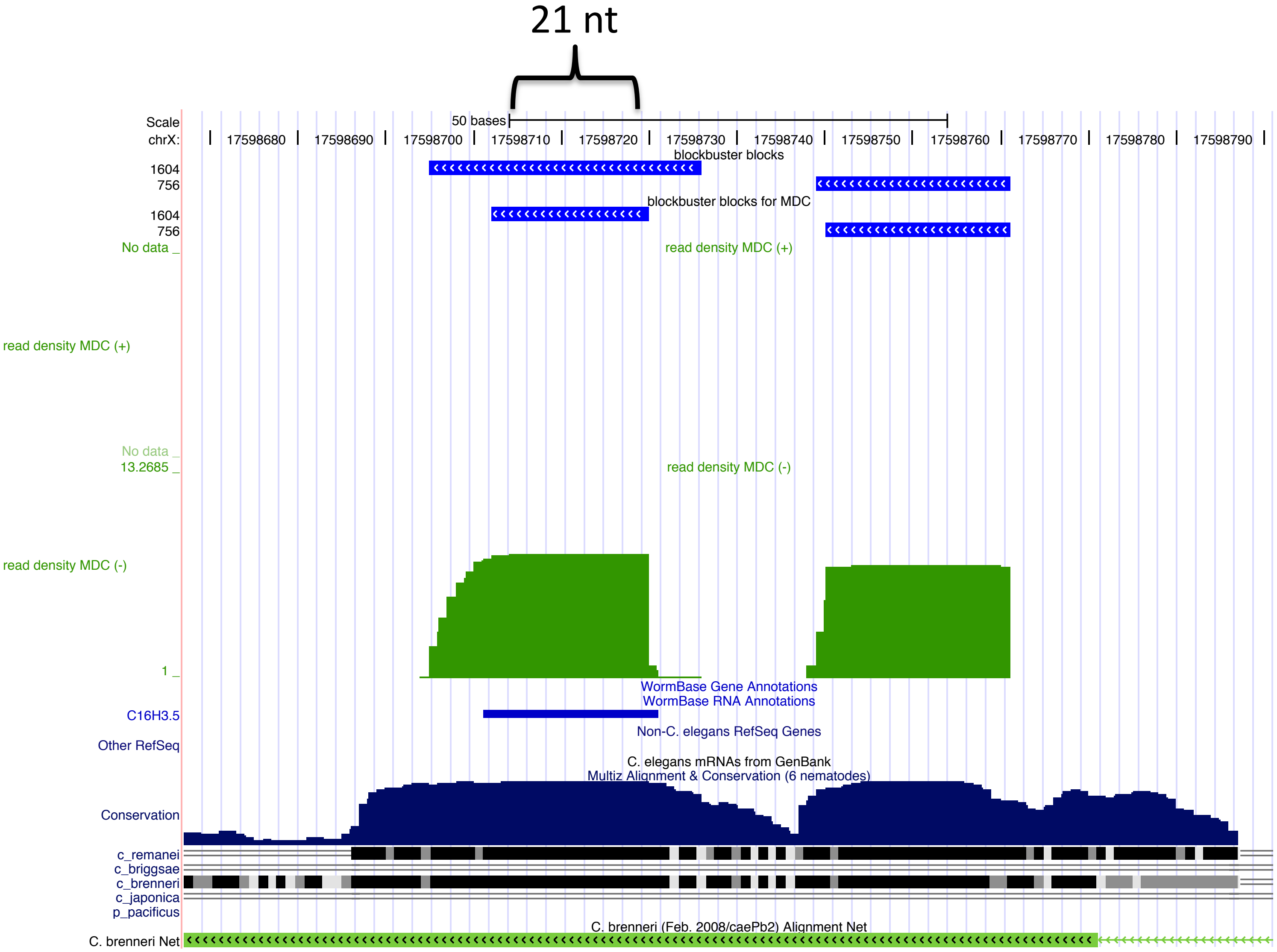
4. Discard clusters with <2 blocks and/or <10 reads (small information content)



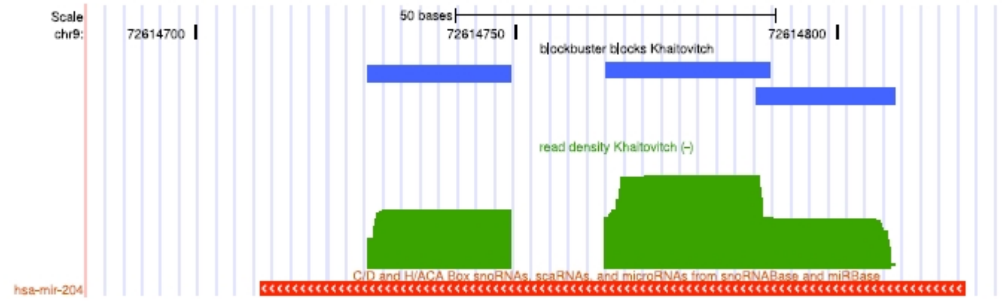
# Data preparation

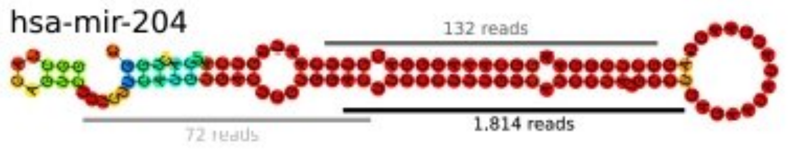


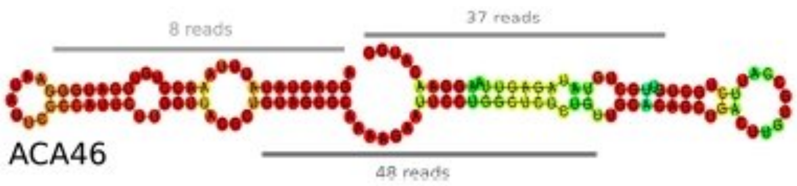
# Data preparation

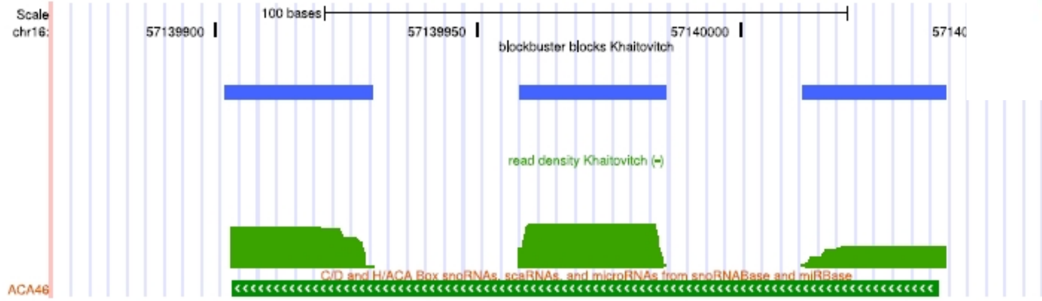


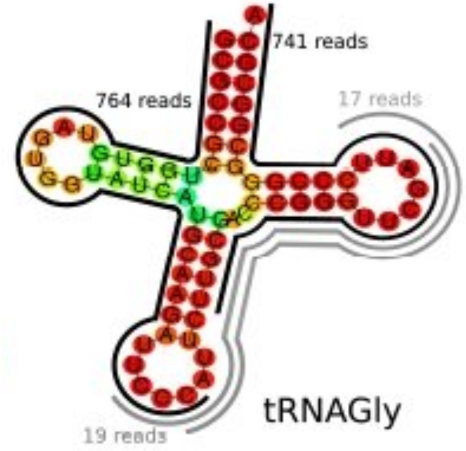
# Different ncRNAs have different read patterns

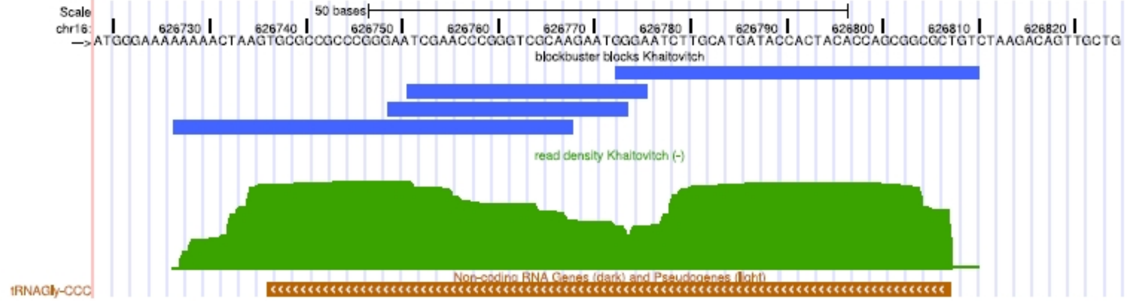

Scale
chr9:
72614700
50 bases
72614750
72614800
blockbuster blocks Khaitovitch
read density Khaitovitch (-)
C/D and H/ACA Box snoRNAs, scaRNAs, and microRNAs from snoRNABase and miRBase
hsa-mir-204






Scale
chr16:
57139900
100 bases
57139950
57140000
5714C
blockbuster blocks Khaitovitch
read density Khaitovitch (-)
C/D and H/ACA Box snoRNAs, scaRNAs, and microRNAs from snoRNABase and miRBase
ACA46




Scale
chr16:
50 bases
626730
626740
626750
626760
626770
626780
626790
626800
626810
626820
ATGGAAAAAAAAACTAAGTGCGCCGCCCGGGAATCGAACCCGGGTCGCAAGAATCGGAATCTTGCATGATACCACTACACCAGCGGCGCTGTCTAAGACAGTTGCTG
blockbuster blocks Khaitovitch
read density Khaitovitch (-)
Non-coding RNA Genes (dark) and Pseudogenes (light)
tRNAGly-CCC

# Classifier (random forest)

- Trained on human reads mapping to human ncRNAs (khaitovitch – brain data – 454 sequenzer)
- Trainingset:
  - 243 miRNAs
  - 19 snoRNAs (H/ACA)
  - 116 snoRNAs (C/D)
  - 336 tRNAs
  - 273 other ncRNAs

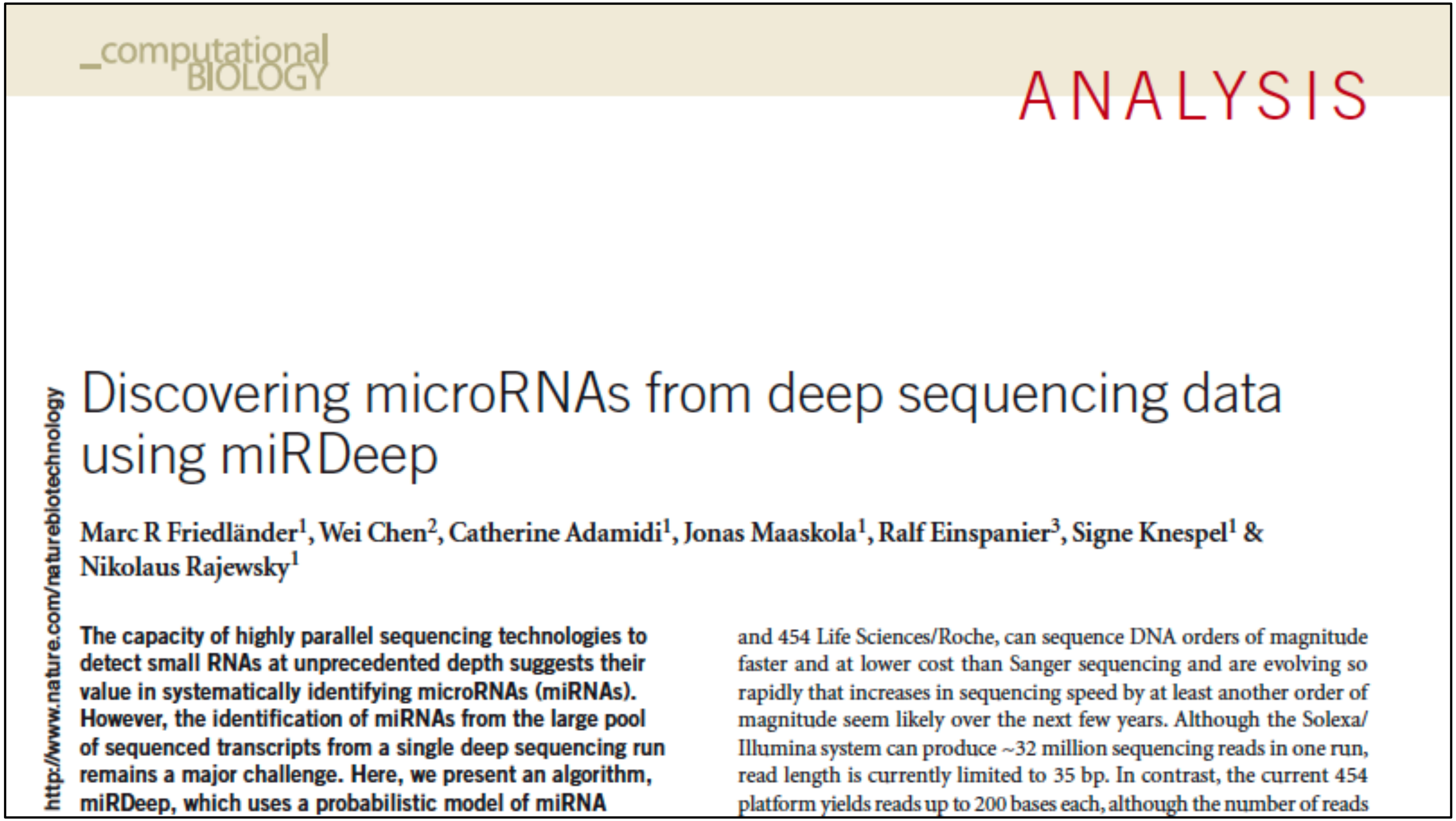
Confusion Matrix (10-fold cross validation)

| **miRNA** | **snoRNA (H/ACA)** | **snoRNA (C/D)** | **tRNA** | **other** | **<- classified as** |
|---|---|---|---|---|---|
| **232 (95%)** | 3 | 2 | 4 | 2 | **miRNA** |
| 8 | **3 (2%)** | 1 | 3 | 4 | **snoRNA (H/ACA)** |
| 2 | 2 | **71 (61%)** | 16 | 25 | **snoRNA (C/D)** |
| 3 | 0 | 5 | **291 (87%)** | 37 | **tRNA** |
| 9 | 1 | 15 | 53 | **195 (71%)** | **other** |

# Classification of worm ncRNAs

| | **Recall** | | | |
|---|---|---|---|---|
| | **miRNA** | **snoRNA** | **tRNA** | **PPV** |
| **MDC set** | **61%**<br>(49 of 80) | **0%**<br>(0 of 28) | **76%**<br>(438 of 573) | 94% |
| **Bartel set** | **85%**<br>(61 of 71) | **5%**<br>(1 of 8) | **76%**<br>(438 of 573) | 95% |
| **Berezikov set** | **79%**<br>(26 of 33) | **0%**<br>(0 of 0) | **31%**<br>(52 of 169) | 96% |

# miRDeep



computational BIOLOGY

ANALYSIS

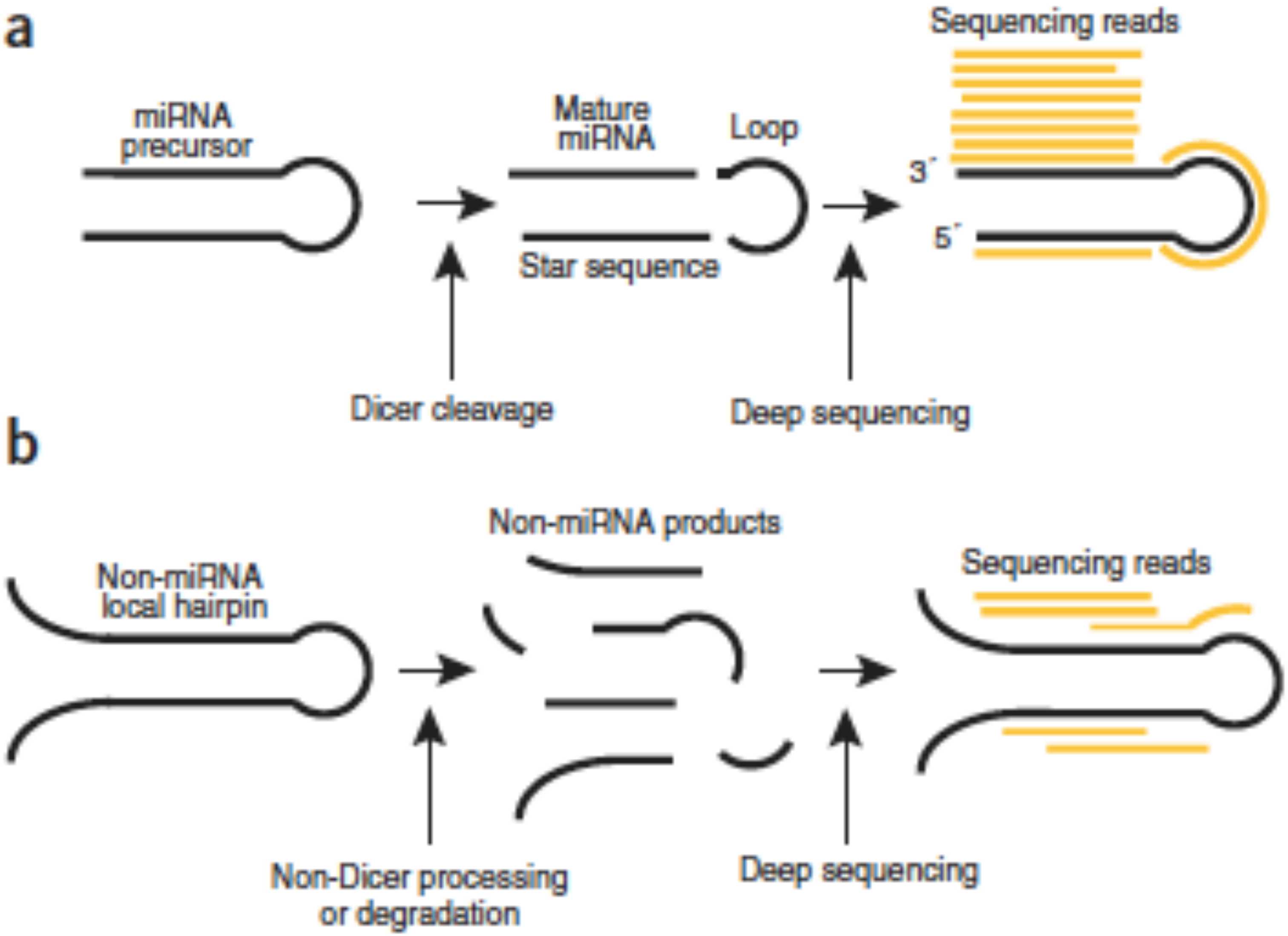
http://www.nature.com/naturebiotechnology

# Discovering microRNAs from deep sequencing data using miRDeep

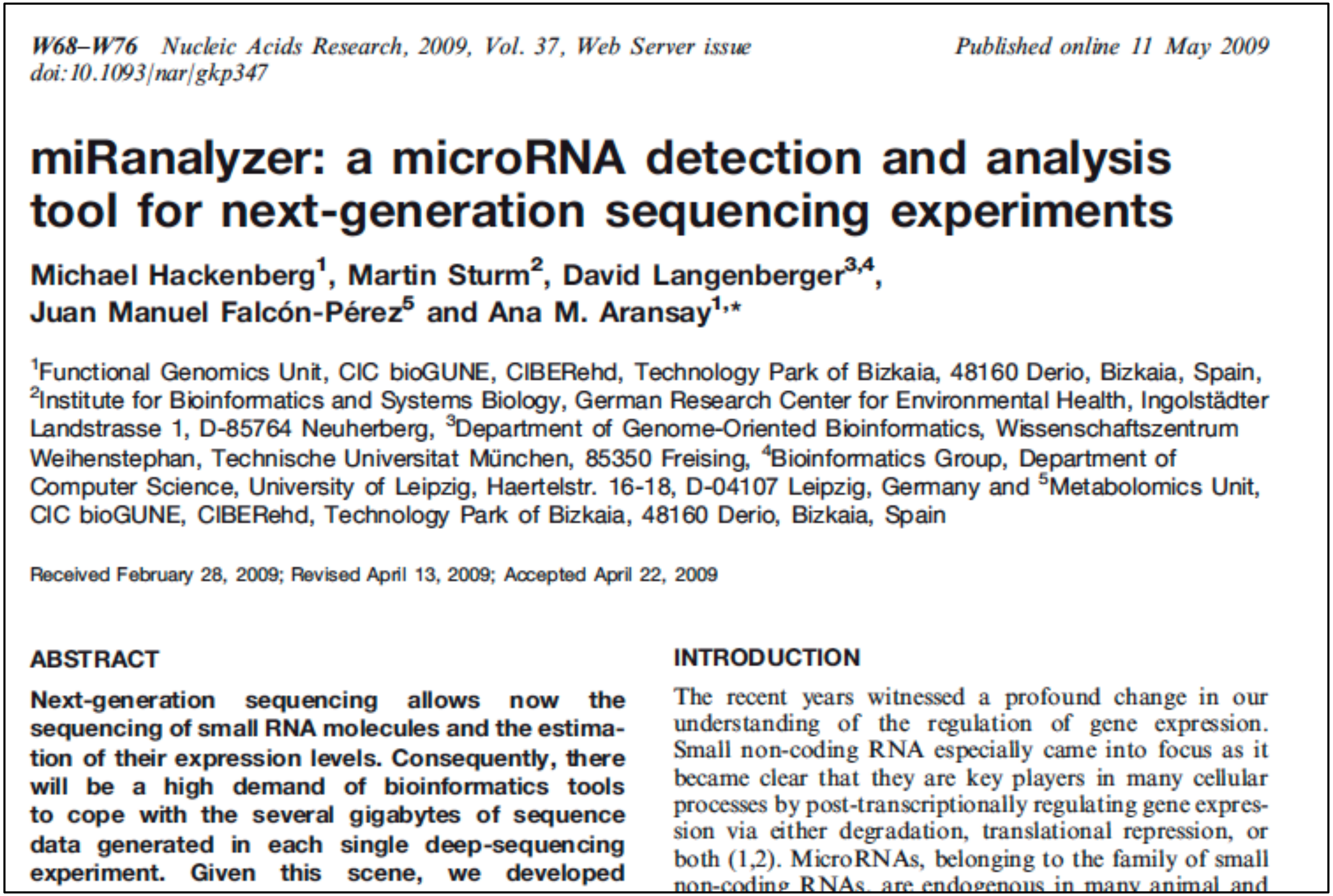
Marc R Friedländer[1], Wei Chen[2], Catherine Adamidi[1], Jonas Maaskola[1], Ralf Einspanier[3], Signe Knespel[1] & Nikolaus Rajewsky[1]

**The capacity of highly parallel sequencing technologies to detect small RNAs at unprecedented depth suggests their value in systematically identifying microRNAs (miRNAs). However, the identification of miRNAs from the large pool of sequenced transcripts from a single deep sequencing run remains a major challenge. Here, we present an algorithm, miRDeep, which uses a probabilistic model of miRNA**

and 454 Life Sciences/Roche, can sequence DNA orders of magnitude faster and at lower cost than Sanger sequencing and are evolving so rapidly that increases in sequencing speed by at least another order of magnitude seem likely over the next few years. Although the Solexa/Illumina system can produce ~32 million sequencing reads in one run, read length is currently limited to 35 bp. In contrast, the current 454 platform yields reads up to 200 bases each, although the number of reads


a
miRNA precursor
Mature miRNA
Loop
Star sequence
Sequencing reads
3´
5´
Dicer cleavage
Deep sequencing
b
Non-miRNA local hairpin
Non-miRNA products
Sequencing reads
Non-Dicer processing or degradation
Deep sequencing

# miRanalyzer



*W68–W76 Nucleic Acids Research, 2009, Vol. 37, Web Server issue* *Published online 11 May 2009*
*doi:10.1093/nar/gkp347*

# miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments

**Michael Hackenberg[1], Martin Sturm[2], David Langenberger[3,4], Juan Manuel Falcón-Pérez[5] and Ana M. Aransay[1,*]**

[1]Functional Genomics Unit, CIC bioGUNE, CIBERehd, Technology Park of Bizkaia, 48160 Derio, Bizkaia, Spain, [2]Institute for Bioinformatics and Systems Biology, German Research Center for Environmental Health, Ingolstädter Landstrasse 1, D-85764 Neuherberg, [3]Department of Genome-Oriented Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universitat München, 85350 Freising, [4]Bioinformatics Group, Department of Computer Science, University of Leipzig, Haertelstr. 16-18, D-04107 Leipzig, Germany and [5]Metabolomics Unit, CIC bioGUNE, CIBERehd, Technology Park of Bizkaia, 48160 Derio, Bizkaia, Spain

Received February 28, 2009; Revised April 13, 2009; Accepted April 22, 2009

## ABSTRACT

**Next-generation sequencing allows now the sequencing of small RNA molecules and the estimation of their expression levels. Consequently, there will be a high demand of bioinformatics tools to cope with the several gigabytes of sequence data generated in each single deep-sequencing experiment. Given this scene, we developed**

## INTRODUCTION

The recent years witnessed a profound change in our understanding of the regulation of gene expression. Small non-coding RNA especially came into focus as it became clear that they are key players in many cellular processes by post-transcriptionally regulating gene expression via either degradation, translational repression, or both (1,2). MicroRNAs, belonging to the family of small non-coding RNAs, are endogenous in many animal and

# Results

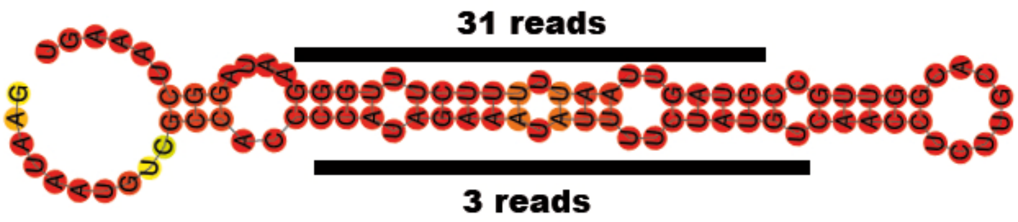
# miRNA prediction in worm

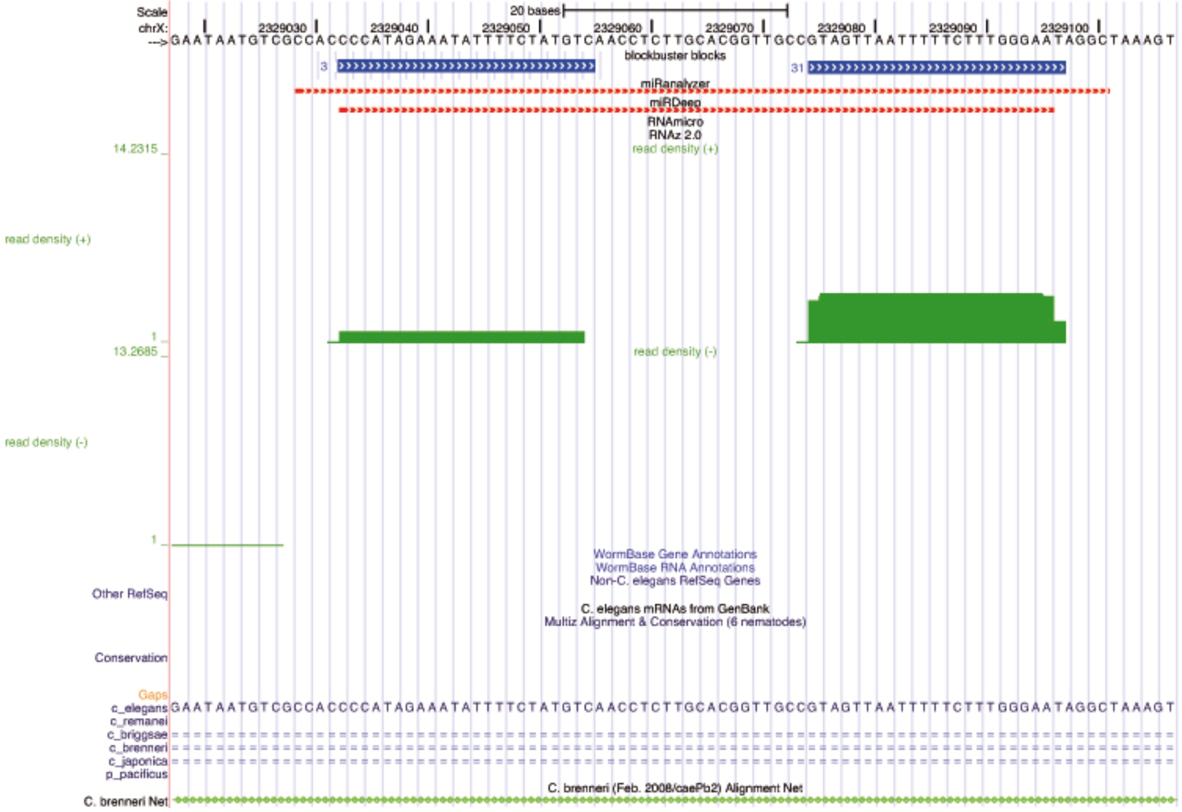
| ***MDC*** | **miRDeep** | **miRanalyzer** | **Classifier** |
|---|---|---|---|
| **miRDeep** | **335** | 50 | 3 |
| **miRanalyzer** | 50 | **650** | 4 |
| **Classifier** | 3 | 4 | **5** |

| ***Bartel*** | **miRanalyzer** | **Classifier** |
|---|---|---|
| **miRanalyzer** | **67** | 1 |
| **Classifier** | 1 | **7** |

| ***Berezikov*** | **miRanalyzer** | **Classifier** |
|---|---|---|
| **miRanalyzer** | **27** | 0 |
| **Classifier** | 0 | **1** |

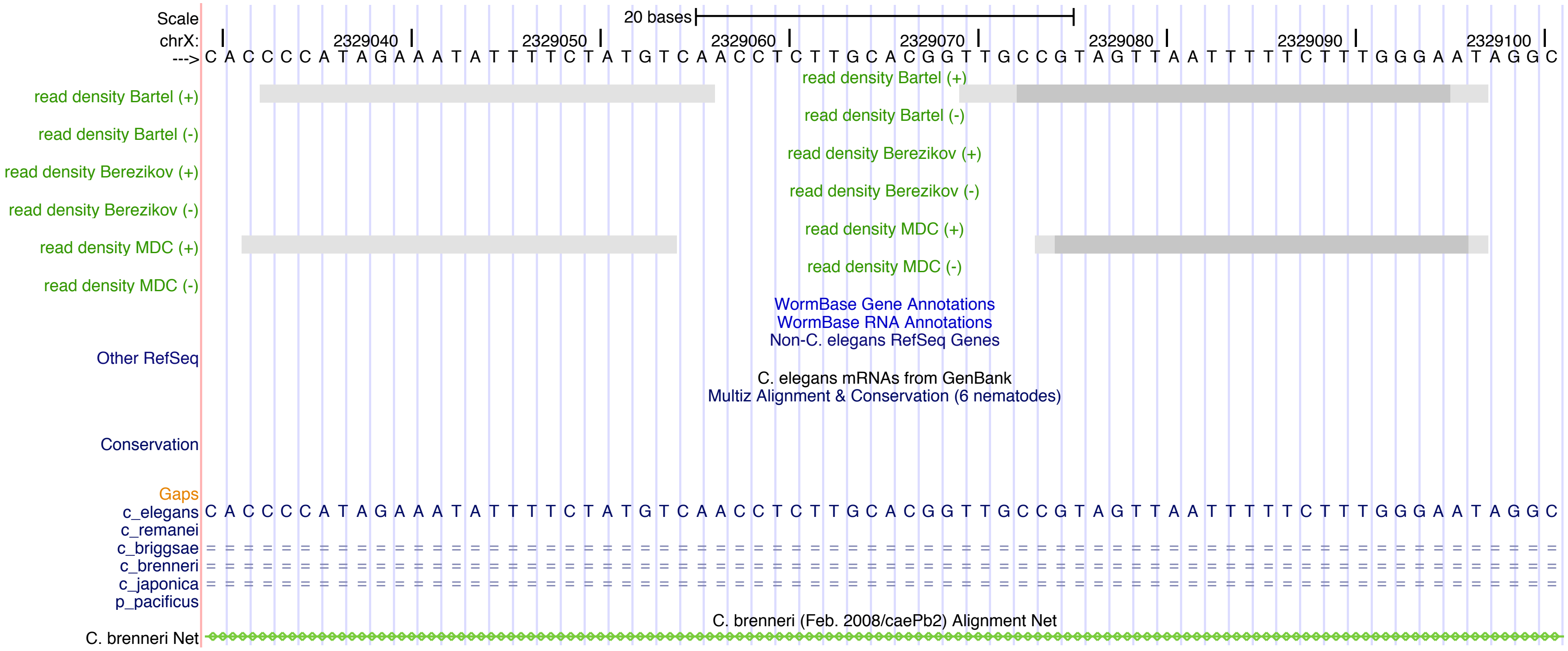
# 3 candidates found by the classifier, miRanalyser and miRDeep (MDC – set)



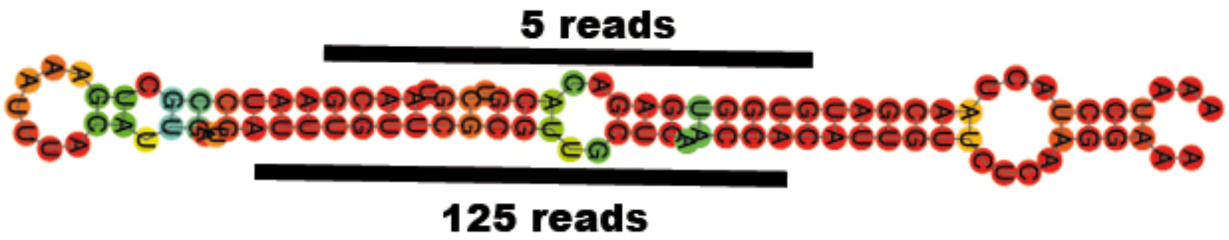

Scale
chrX:
20 bases
2329030
2329040
2329050
2329060
2329070
2329080
2329090
2329100
--->GAATAATGTCGCCACCCCATAGAAATATTTTCTATGTCAACCTCTTGCACGGTTGCCGTAGTTAATTTTTCTTTGGGAATAGGCTAAAGT
blockbuster blocks
3
31
miRanalyzer
miRDeep
RNAmicro
RNAz 2.0
read density (+)
14.2315
read density (+)
1
13.2685
read density (-)
read density (-)
1
WormBase Gene Annotations
WormBase RNA Annotations
Non-C. elegans RefSeq Genes
Other RefSeq
C. elegans mRNAs from GenBank
Multiz Alignment & Conservation (6 nematodes)
Conservation
Gaps
c_elegans GAATAATGTCGCCACCCCATAGAAATATTTTCTATGTCAACCTCTTGCACGGTTGCCGTAGTTAATTTTTCTTTGGGAATAGGCTAAAGT
c_remanei
c_briggsae
c_brenneri
c_japonica
p_pacificus
C. brenneri (Feb. 2008/caePb2) Alignment Net
C. brenneri Net

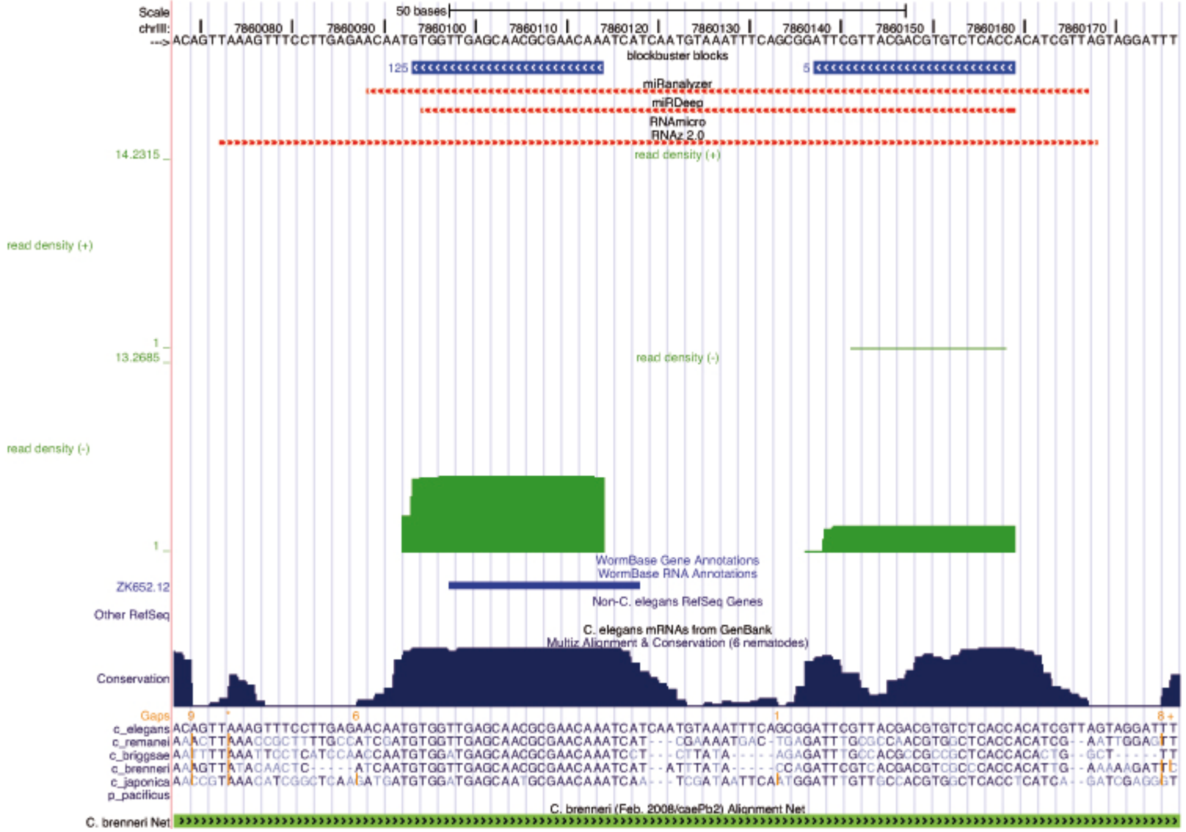
**chrX:2,329,018-2,329,107**

# 3 candidates found by the classifier, miRanalyser and miRDeep (all sets)


Scale
20 bases
chrX:
2329040
2329050
2329060
2329070
2329080
2329090
2329100
---> C A C C C C A T A G A A A T A T T T T C T A T G T C A A C C T C T T G C A C G G T T G C C G T A G T T A A T T T T T C T T T G G G A A T A G G C
read density Bartel (+)
read density Bartel (+)
read density Bartel (-)
read density Bartel (-)
read density Berezikov (+)
read density Berezikov (+)
read density Berezikov (-)
read density Berezikov (-)
read density MDC (+)
read density MDC (+)
read density MDC (-)
read density MDC (-)
WormBase Gene Annotations
WormBase RNA Annotations
Non-C. elegans RefSeq Genes
Other RefSeq
C. elegans mRNAs from GenBank
Multiz Alignment & Conservation (6 nematodes)
Conservation
Gaps
c_elegans C A C C C C A T A G A A A T A T T T T C T A T G T C A A C C T C T T G C A C G G T T G C C G T A G T T A A T T T T T C T T T G G G A A T A G G C
c_remanei
c_briggsae
c_brenneri
c_japonica
p_pacificus
C. brenneri (Feb. 2008/caePb2) Alignment Net
C. brenneri Net

# 3 candidates found by the classifier, miRanalyser and miRDeep (MDC – set)



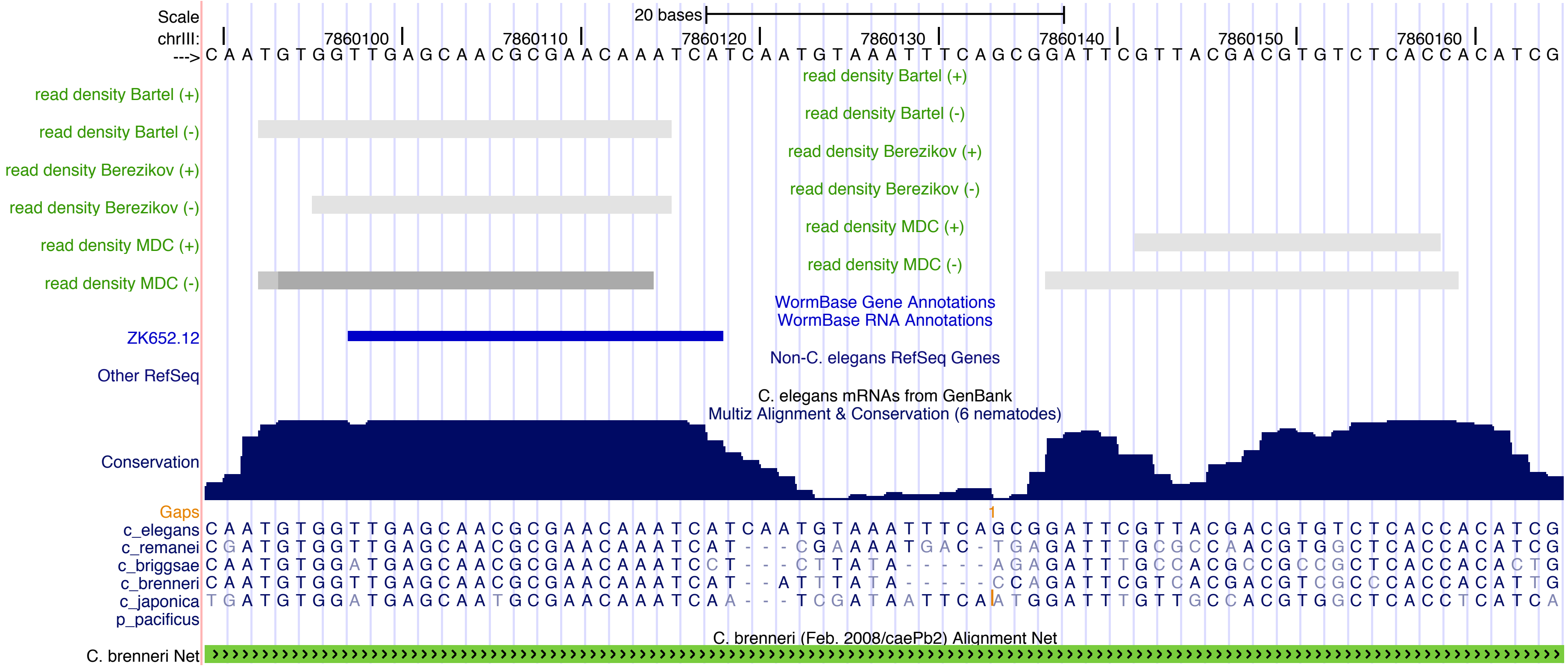


**chrIII:7,860,027-7,860,227**

Antisense to cel-miR-356

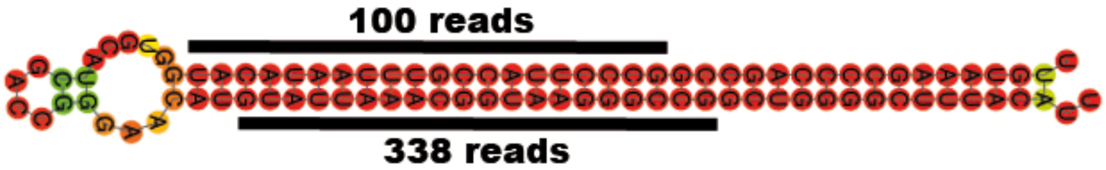
RNAz support

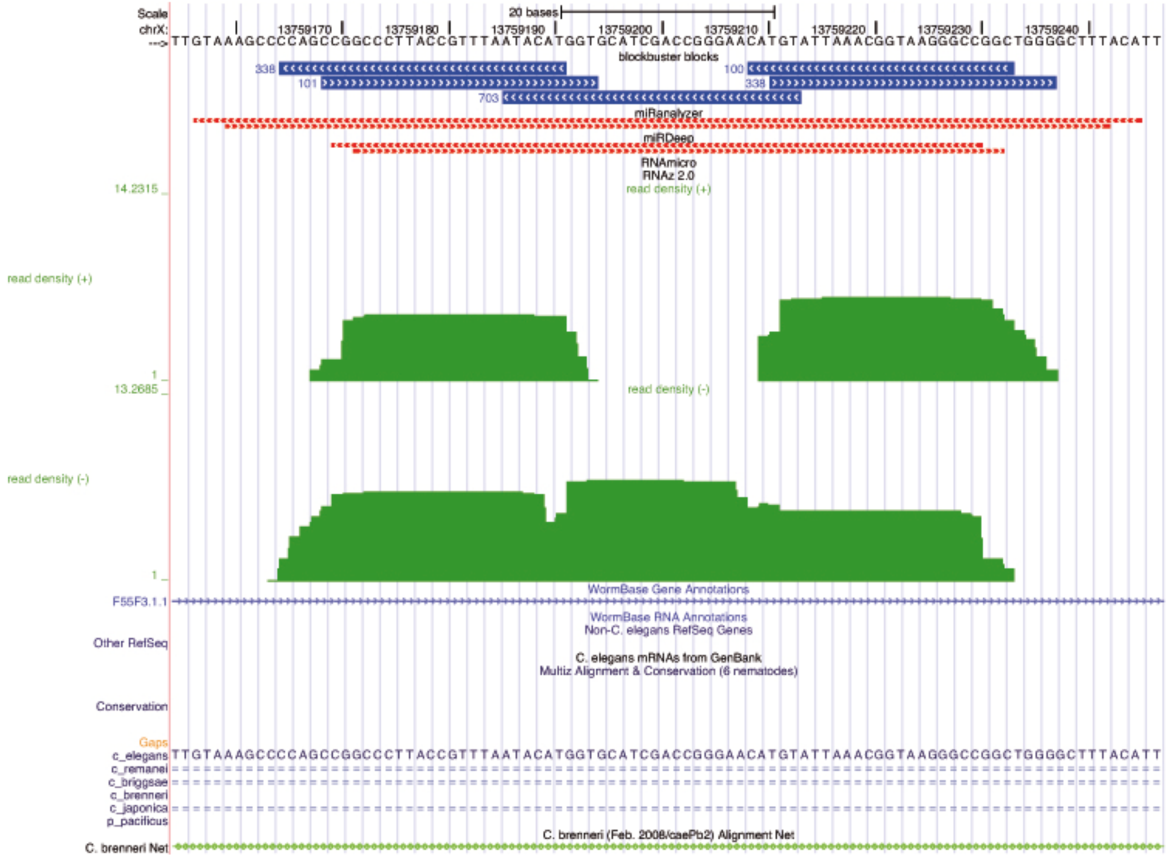
# 3 candidates found by the classifier, miRanalyser and miRDeep (all sets)


Scale
20 bases
chrIII:
7860100
7860110
7860120
7860130
7860140
7860150
7860160
---> C A A T G T G G T T G A G C A A C G C G A A C A A A T C A T C A A T G T A A A T T T C A G C G G A T T C G T T A C G A C G T G T C T C A C C A C A T C G
read density Bartel (+)
read density Bartel (-)
read density Berezikov (+)
read density Berezikov (-)
read density MDC (+)
read density MDC (-)
WormBase Gene Annotations
WormBase RNA Annotations
ZK652.12
Non-C. elegans RefSeq Genes
Other RefSeq
C. elegans mRNAs from GenBank
Multiz Alignment & Conservation (6 nematodes)
Conservation
Gaps
1
c_elegans C A A T G T G G T T G A G C A A C G C G A A C A A A T C A T C A A T G T A A A T T T C A G C G G A T T C G T T A C G A C G T G T C T C A C C A C A T C G
c_remanei C G A T G T G G T T G A G C A A C G C G A A C A A A T C A T - - - C G A A A A T G A C - T G A G A T T T G C G C C A A C G T G G C T C A C C A C A T C G
c_briggsae C A A T G T G G A T G A G C A A C G C G A A C A A A T C C T - - - C T T A T A - - - - - A G A G A T T T G C C A C G C C G C C G C T C A C C A C A C T G
c_brenneri C A A T G T G G T T G A G C A A C G C G A A C A A A T C A T - - A T T T A T A - - - - - C C A G A T T C G T C A C G A C G T C G C C C A C C A C A T T G
c_japonica T G A T G T G G A T G A G C A A T G C G A A C A A A T C A A - - - T C G A T A A T T C A A T G G A T T T G T T G C C A C G T G G C T C A C C T C A T C A
p_pacificus
C. brenneri (Feb. 2008/caePb2) Alignment Net
C. brenneri Net

- miRanalyzer predictions for all datasets
- No classification (only one block in other datasets)

# 3 candidates found by the classifier, miRanalyser and miRDeep (MDC – set)



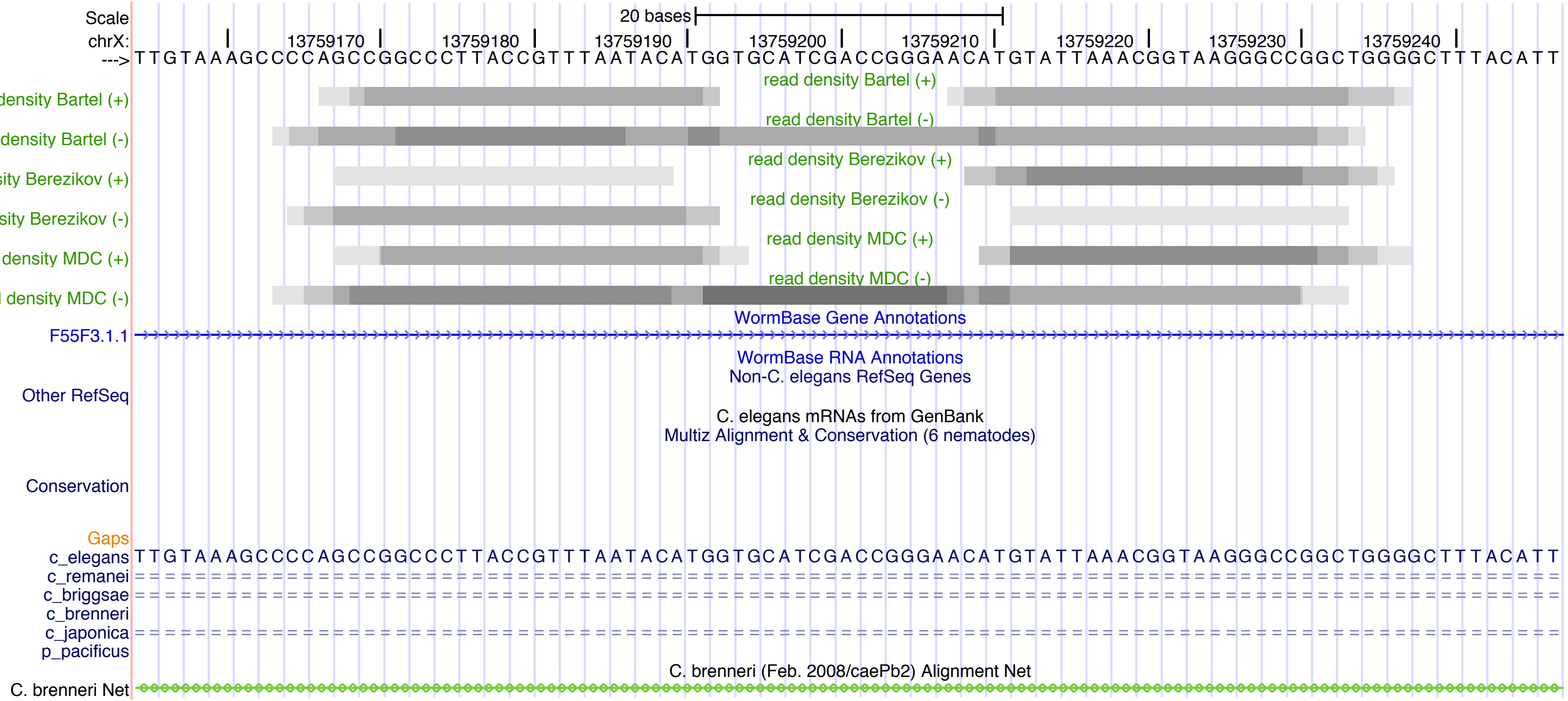


**chrX:13,759,155-13,759,247**

Sense miRNA detected by all three methods

Antisense miRNA only detected by miRanalyzer and miRDeep

# 3 candidates found by the classifier, miRanalyser and miRDeep (all sets)


Scale
20 bases
chrX:
13759170
13759180
13759190
13759200
13759210
13759220
13759230
13759240
---> TTGTAAAGCCCCAGCCGGCCCTTACCGTTTAATACATGGTGCATCGACCGGGAACATGTATTAAACGGTAAGGGCCGGCTGGGGCTTTACATT
read density Bartel (+)
read density Bartel (-)
read density Berezikov (+)
read density Berezikov (-)
read density MDC (+)
read density MDC (-)
WormBase Gene Annotations
F55F3.1.1
WormBase RNA Annotations
Non-C. elegans RefSeq Genes
Other RefSeq
C. elegans mRNAs from GenBank
Multiz Alignment & Conservation (6 nematodes)
Conservation
Gaps
c_elegans TTGTAAAGCCCCAGCCGGCCCTTACCGTTTAATACATGGTGCATCGACCGGGAACATGTATTAAACGGTAAGGGCCGGCTGGGGCTTTACATT
c_remanei
c_briggsae
c_brenneri
c_japonica
p_pacificus
C. brenneri (Feb. 2008/caePb2) Alignment Net
C. brenneri Net

- miRanalyzer predictions for all datasets
- classification for sets from Bartel and MDC (no classification for Berezikov)

# Conclusion

- There are a lot of not annotated ncRNAs (amongst others microRNAs) waiting to be found by using deep sequencing data
- Classifier can be used between species (patterns seem to be conserved)
- Different experiments (454, Solid, Solexa) generate comparable patterns

# Acknowledgements

**University Leipzig**

Peter F. Stadler
Steve Hoffmann
Alexander Donath

**Max-Delbrueck-Center for Molecular Medicine**

Rina Ahmed
Christoph Dieterich