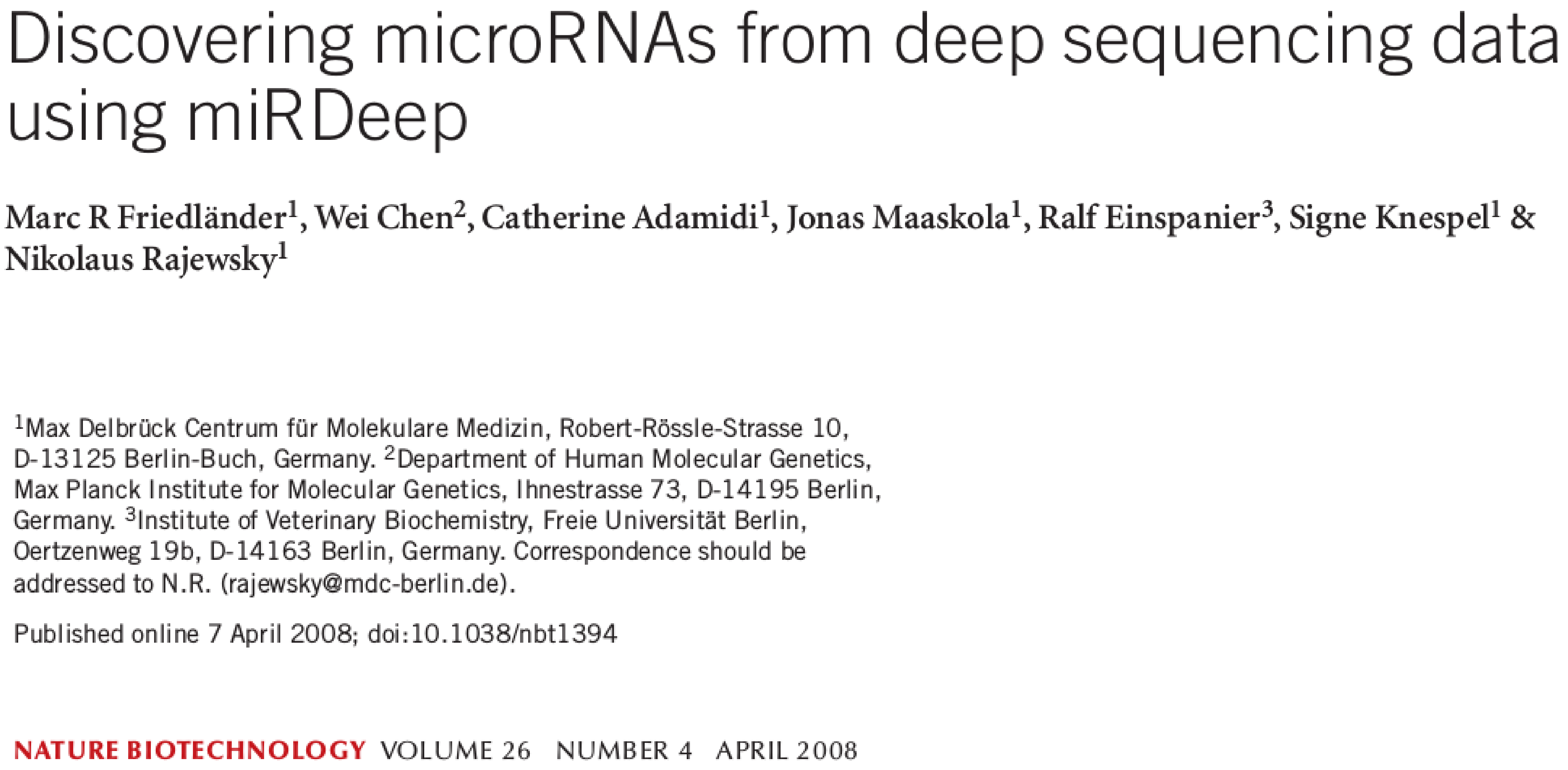


Discovering microRNAs from deep sequencing data using miRDeep

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

[1]Max Delbrück Centrum für Molekulare Medizin, Robert-Rössle-Strasse 10, D-13125 Berlin-Buch, Germany. [2]Department of Human Molecular Genetics, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, D-14195 Berlin, Germany. [3]Institute of Veterinary Biochemistry, Freie Universität Berlin, Oertzenweg 19b, D-14163 Berlin, Germany. Correspondence should be addressed to N.R. (rajewsky@mdc-berlin.de).

Published online 7 April 2008; doi:10.1038/nbt1394

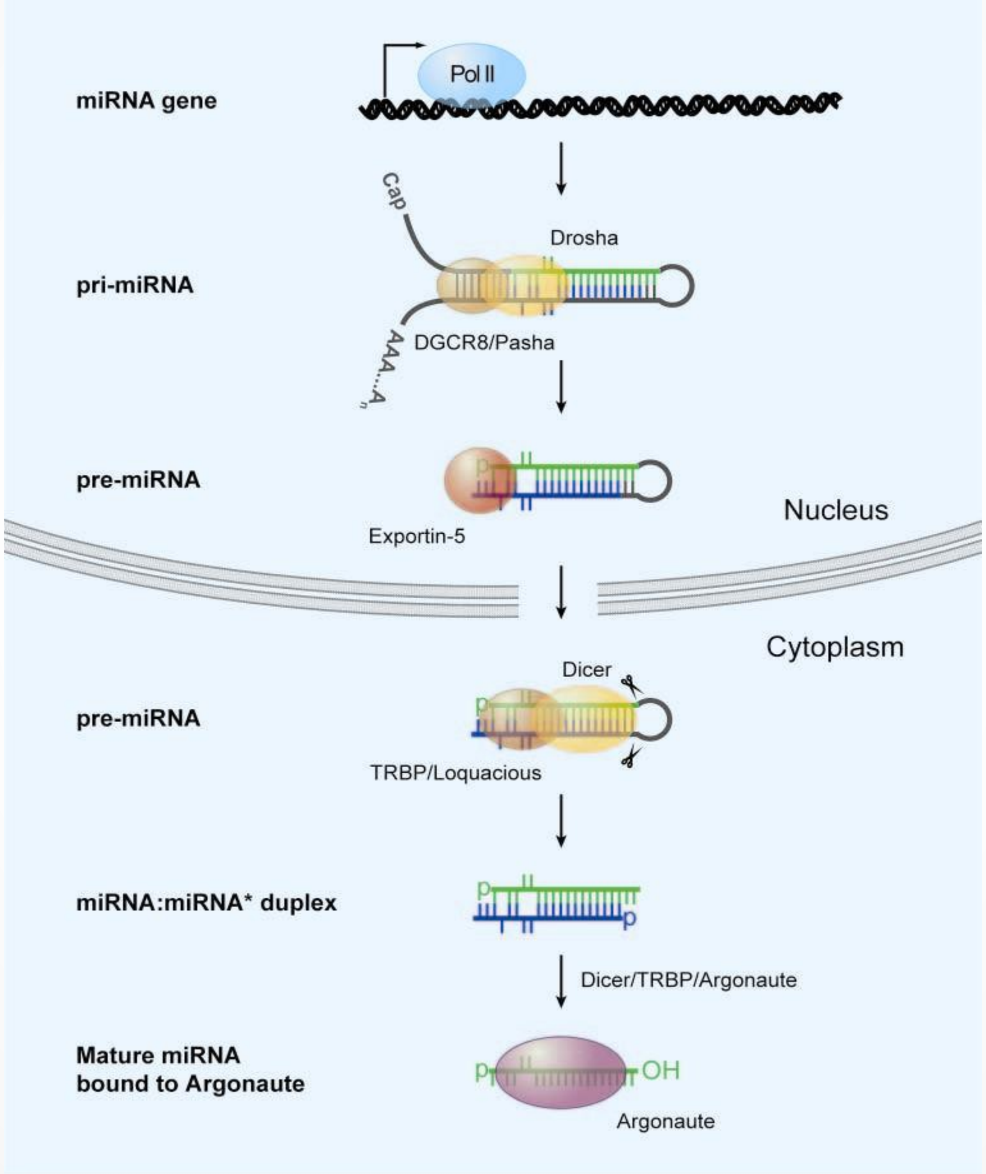
NATURE BIOTECHNOLOGY VOLUME 26 NUMBER 4 APRIL 2008

miRNA

- small non coding RNA (18-24 bp)
- post transcriptional regulation
  - inhibition
  - destabilization
  - cleave

- miRBase (Release 21: June 2014)
  - 28645 hairpin precursor
    - 35828 miRNA in 223 species
    - 2813 miRNA in hg38

miRNA biogenesis



Bushati N, Cohen SM. 2007.
Annu. Rev. Cell Dev. Biol. 23:175–205

Problems

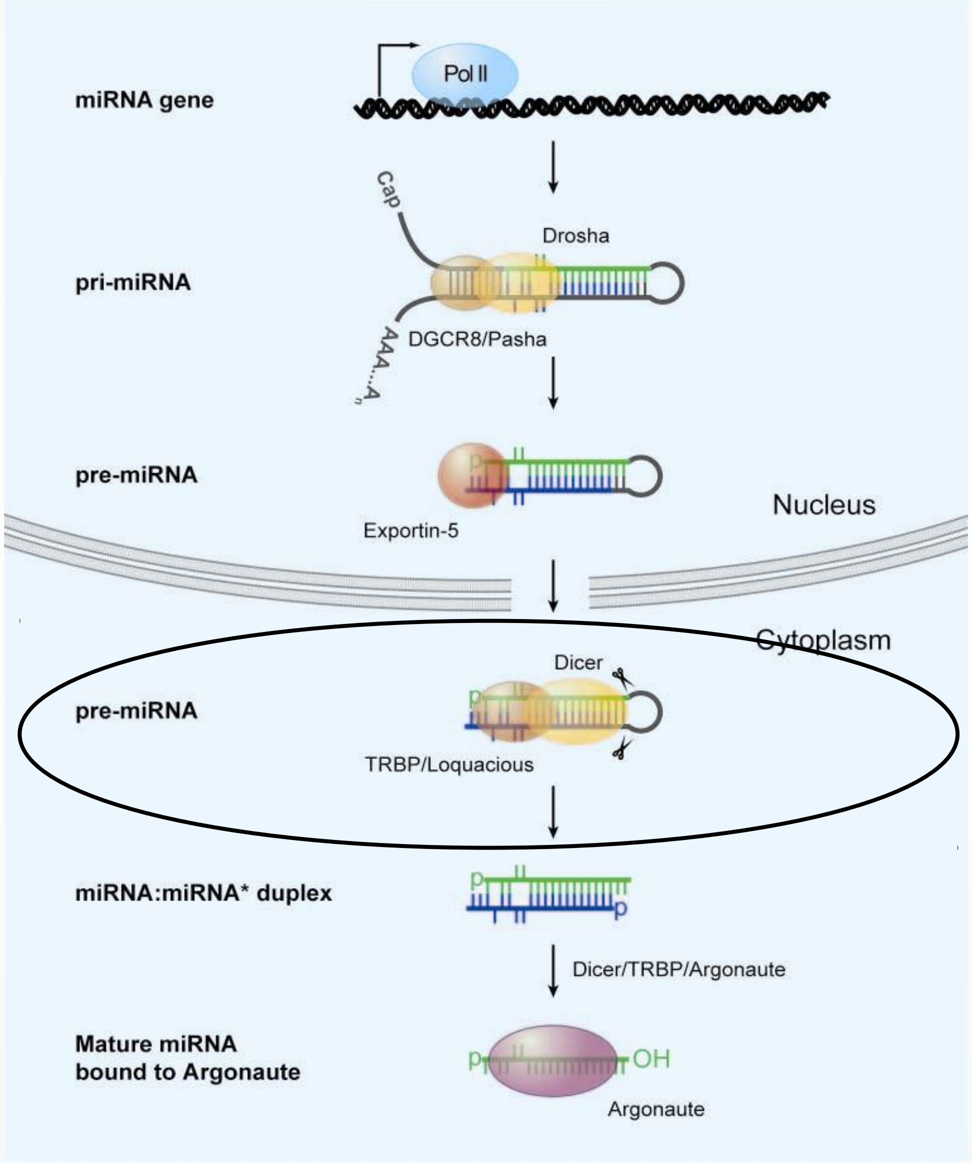
- traditional detection
  - Sanger
  - low expressed miRNA

- genomic DNA
  - sensitivity & FP

Idea

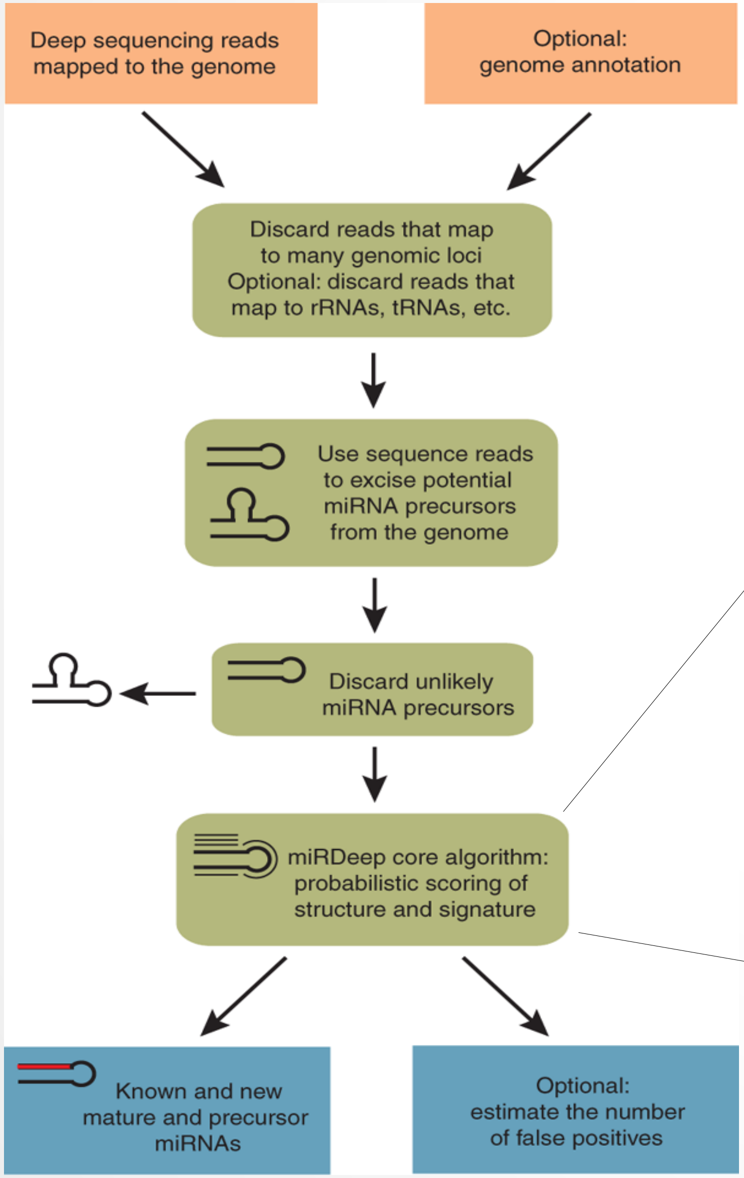
- NGS
- Compare sequenced miRNA with precursor processing

miRNA biogenesis



Bushati N, Cohen SM. 2007.
Annu. Rev. Cell Dev. Biol. 23:175–205

miRDeep



Deep sequencing reads mapped to the genome

Optional: genome annotation

Discard reads that map to many genomic loci
Optional: discard reads that map to rRNAs, tRNAs, etc.

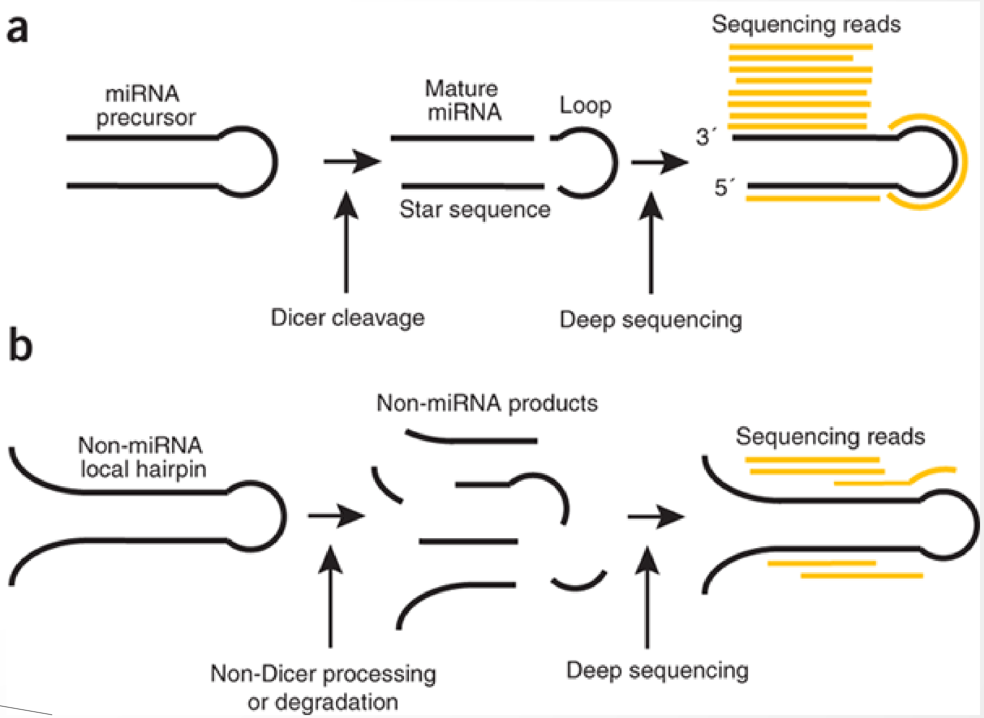
Use sequence reads to excise potential miRNA precursors from the genome

Discard unlikely miRNA precursors

miRDeep core algorithm: probabilistic scoring of structure and signature

Known and new mature and precursor miRNAs

Optional: estimate the number of false positives



Scoring

Scoring (log-odds) of candidate miRNA

miRNA precursor vs. background hairpin

$$score = \frac{P(pre|data)}{P(bgr|data)}$$

Scoring

$$P(pre|data)=\frac{\boxed{P(abs|pre)*P(rel|pre)*P(sig|pre)*P(star|pre)}*P(pre)}{P(data)}$$

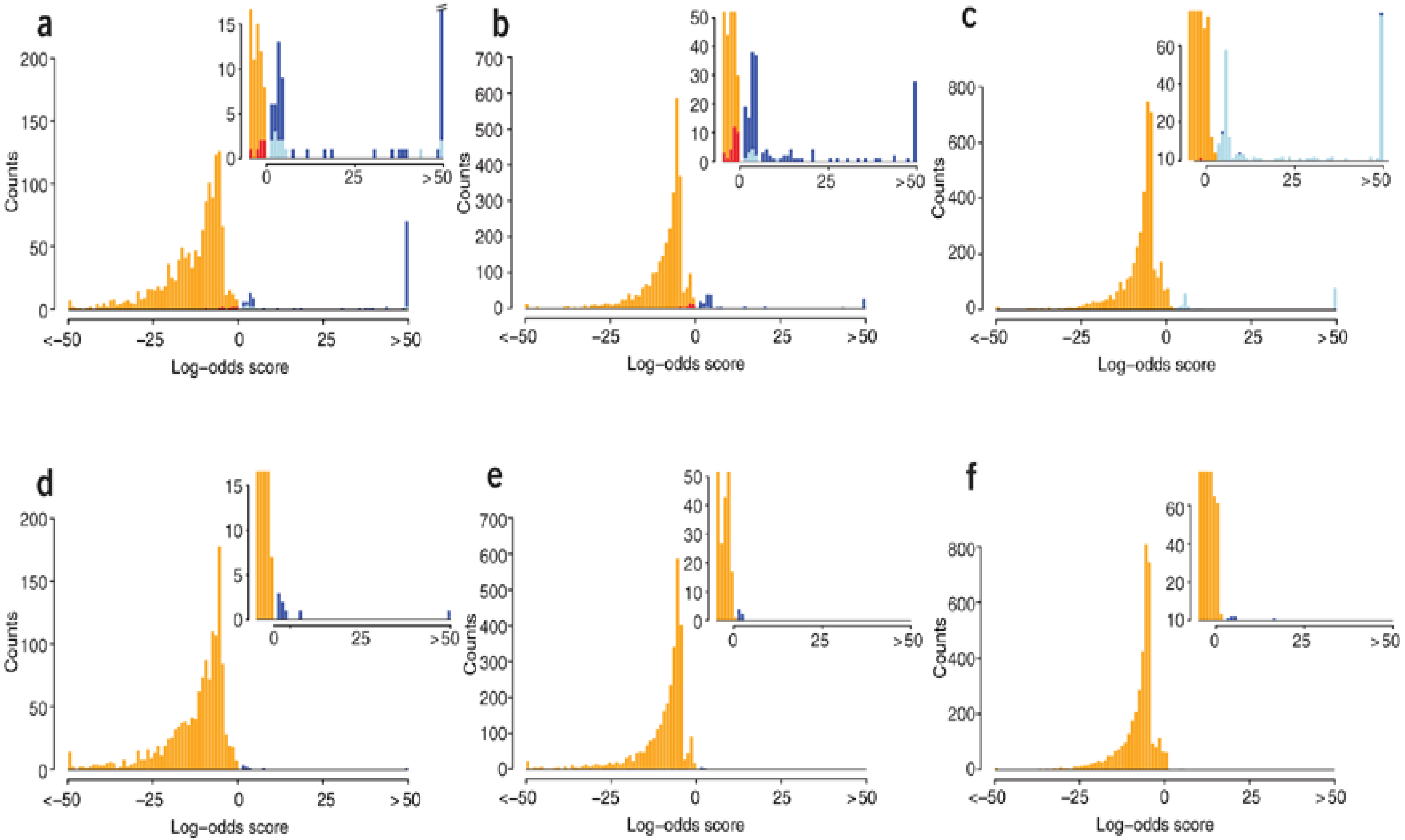
- abs = min. free energy
- rel = stable or unstable
- sig = mapped reads
- star = presence of star sequence

some statistics

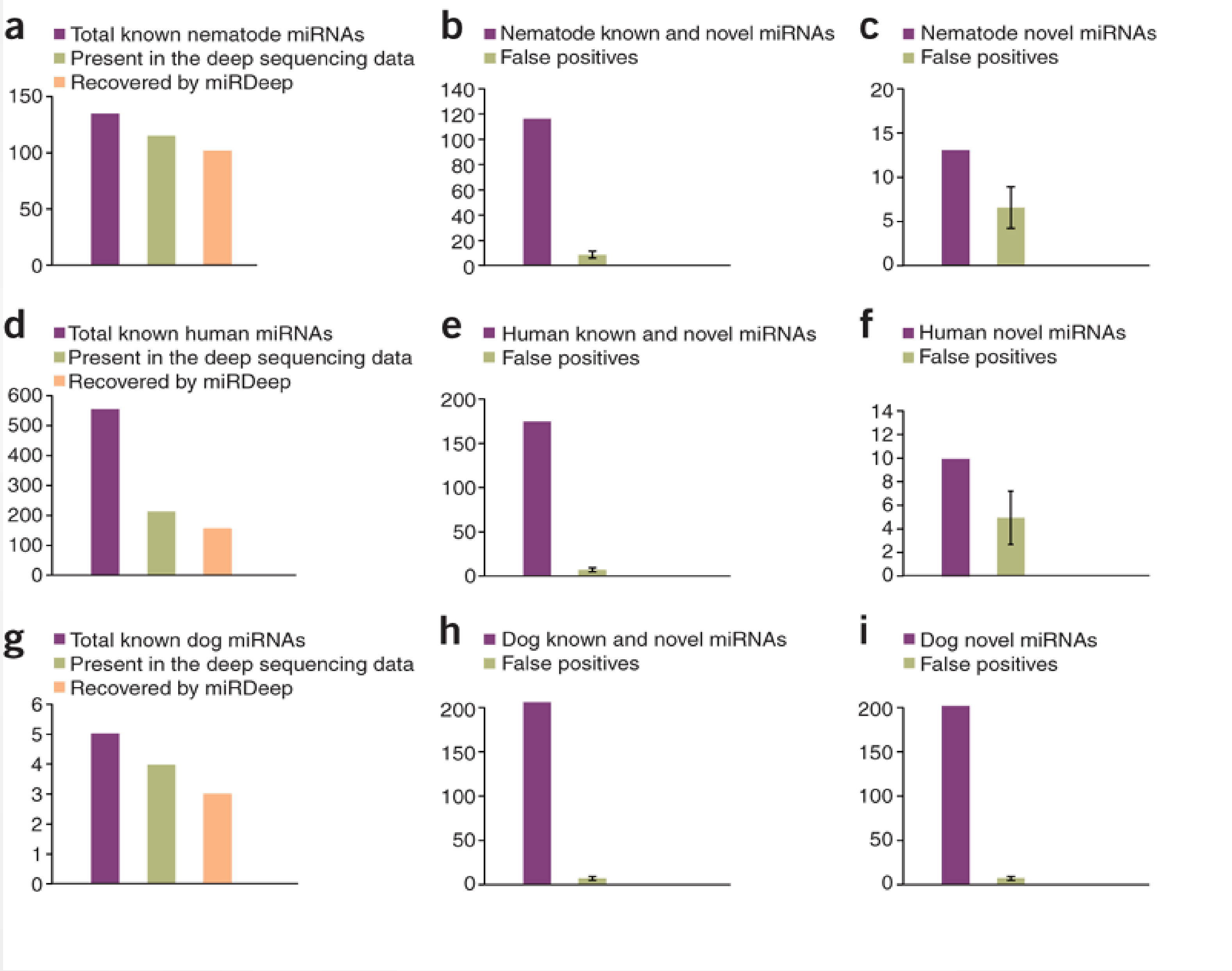
$$sensitivity = \frac{miRNA\ recovered}{reported\ miRNA\ \mathrm{in}\ dataset}$$

- FPR : permutation of structure & signature pairings

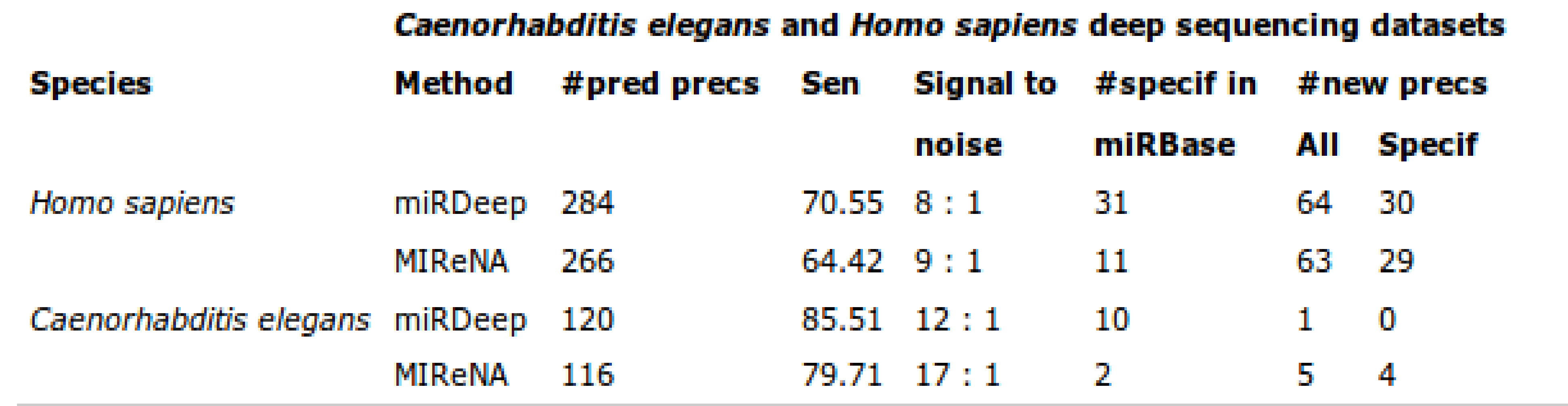
Results


a
Counts
200
150
100
50
0
<−50
−25
0
25
>50
Log−odds score
15
10
5
0
0
25
>50
b
Counts
700
600
500
400
300
200
100
0
<−50
−25
0
25
>50
Log−odds score
50
40
30
20
10
0
0
25
>50
c
Counts
800
600
400
200
0
<−50
−25
0
25
>50
Log−odds score
60
40
20
10
0
25
>50
d
Counts
200
150
100
50
0
<−50
−25
0
25
>50
Log−odds score
15
10
5
0
0
25
>50
e
Counts
700
600
500
400
300
200
100
0
<−50
−25
0
25
>50
Log−odds score
50
40
30
20
10
0
0
25
>50
f
Counts
800
600
400
200
0
<−50
−25
0
25
>50
Log−odds score
60
40
20
10
0
25
>50

Results


a
Total known nematode miRNAs
Present in the deep sequencing data
Recovered by miRDeep
150
100
50
0
b
Nematode known and novel miRNAs
False positives
140
120
100
80
60
40
20
0
c
Nematode novel miRNAs
False positives
20
15
10
5
0
d
Total known human miRNAs
Present in the deep sequencing data
Recovered by miRDeep
600
500
400
300
200
100
0
e
Human known and novel miRNAs
False positives
200
150
100
50
0
f
Human novel miRNAs
False positives
14
12
10
8
6
4
2
0
g
Total known dog miRNAs
Present in the deep sequencing data
Recovered by miRDeep
6
5
4
3
2
1
0
h
Dog known and novel miRNAs
False positives
200
150
100
50
0
i
Dog novel miRNAs
False positives
200
150
100
50
0

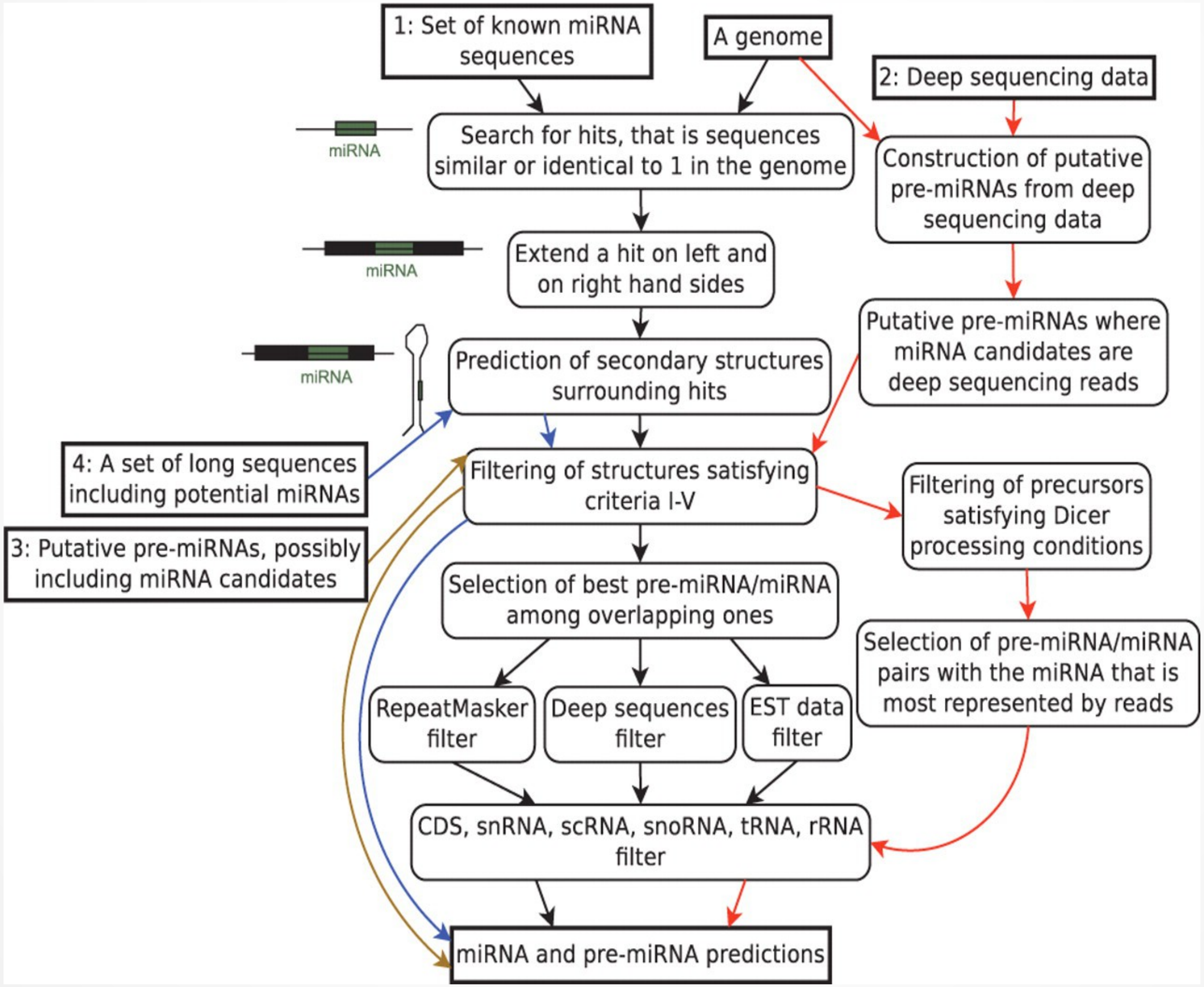
MIReNA



	***Caenorhabditis elegans* and *Homo sapiens* deep sequencing datasets**						
Species	**Method**	**#pred precs**	**Sen**	**Signal to noise**	**#specif in miRBase**	**#new precs**	
						All	**Specif**
Homo sapiens	miRDeep	284	70.55	8 : 1	31	64	30
	MIReNA	266	64.42	9 : 1	11	63	29
Caenorhabditis elegans	miRDeep	120	85.51	12 : 1	10	1	0
	MIReNA	116	79.71	17 : 1	2	5	4

- MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data
  - Anthonny Mathelier and Alessandra Carbone

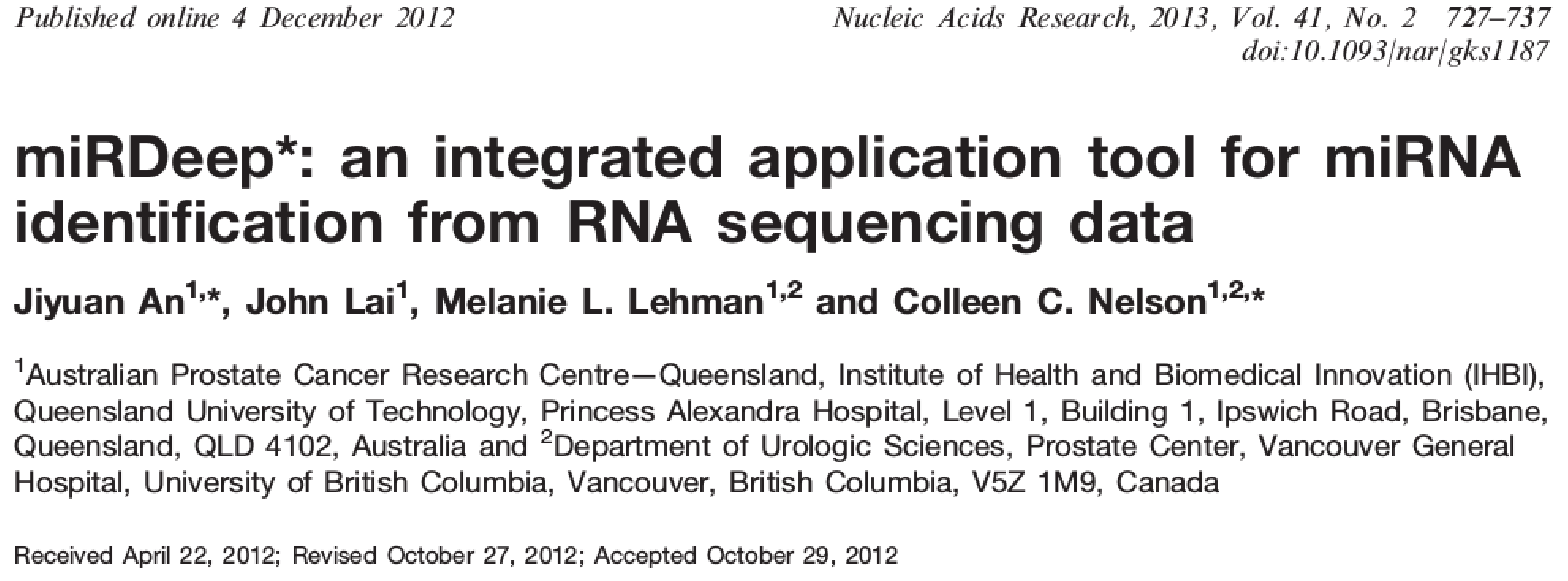
alternative to miRDeep



miRDeep vs. MIReNA

- miRDeep favours read abundant candidates
- MIReNA favours Criteria I-V candidates
- probabilistic model versus combinatorial rules

miRDeep*



Nucleic Acids Research, 2013, Vol. 41, No. 2 ***727–737***
doi:10.1093/nar/gks1187

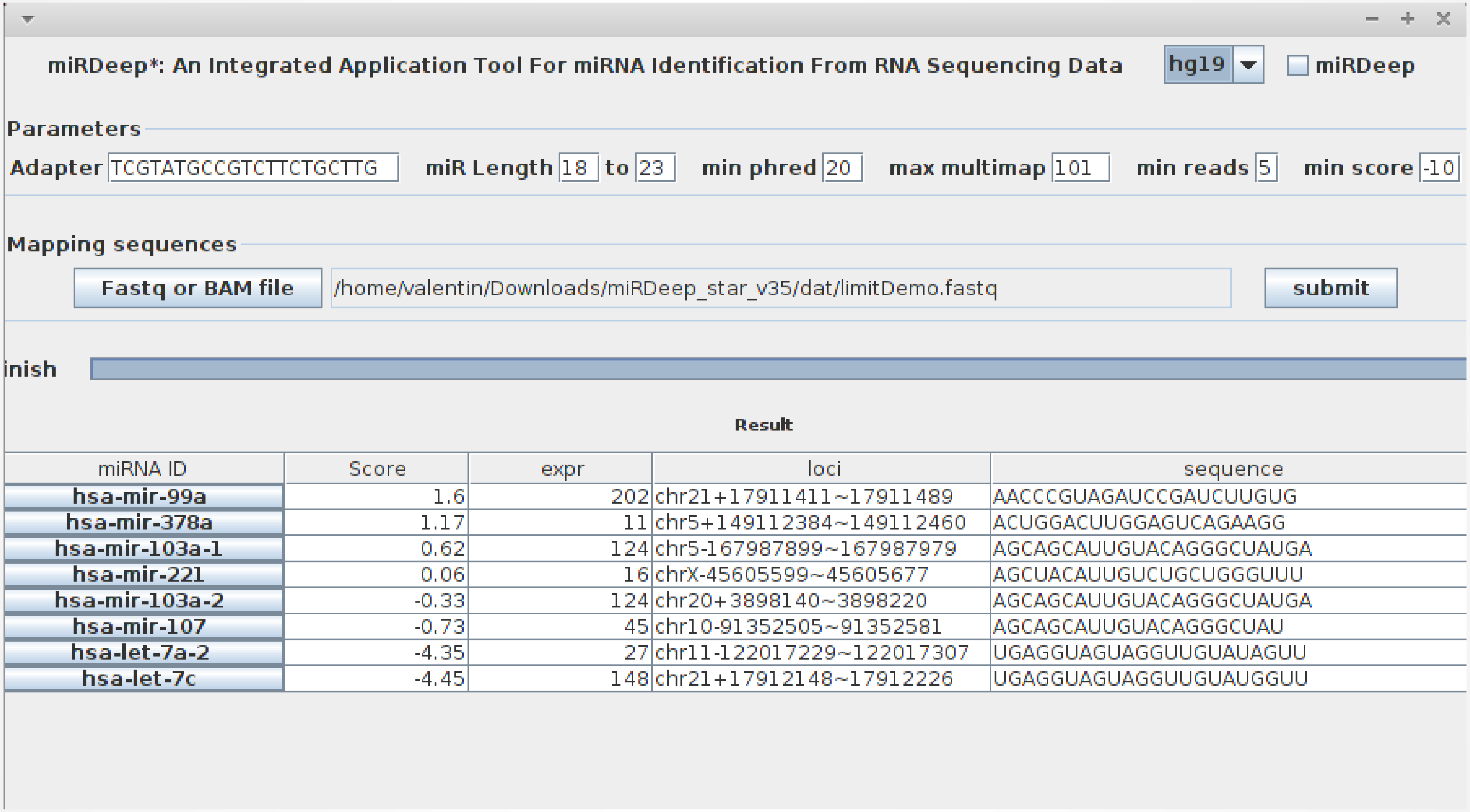
miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data

Jiyuan An[1,*], John Lai[1], Melanie L. Lehman[1,2] and Colleen C. Nelson[1,2,*]

[1]Australian Prostate Cancer Research Centre—Queensland, Institute of Health and Biomedical Innovation (IHBI), Queensland University of Technology, Princess Alexandra Hospital, Level 1, Building 1, Ipswich Road, Brisbane, Queensland, QLD 4102, Australia and [2]Department of Urologic Sciences, Prostate Center, Vancouver General Hospital, University of British Columbia, Vancouver, British Columbia, V5Z 1M9, Canada

Received April 22, 2012; Revised October 27, 2012; Accepted October 29, 2012

miRDeep*



miRDeep*: An Integrated Application Tool For miRNA Identification From RNA Sequencing Data hg19 ☐ miRDeep

Parameters

Adapter TCGTATGCCGTCTTCTGCTTG **miR Length** 18 **to** 23 **min phred** 20 **max multimap** 101 **min reads** 5 **min score** -10

Mapping sequences

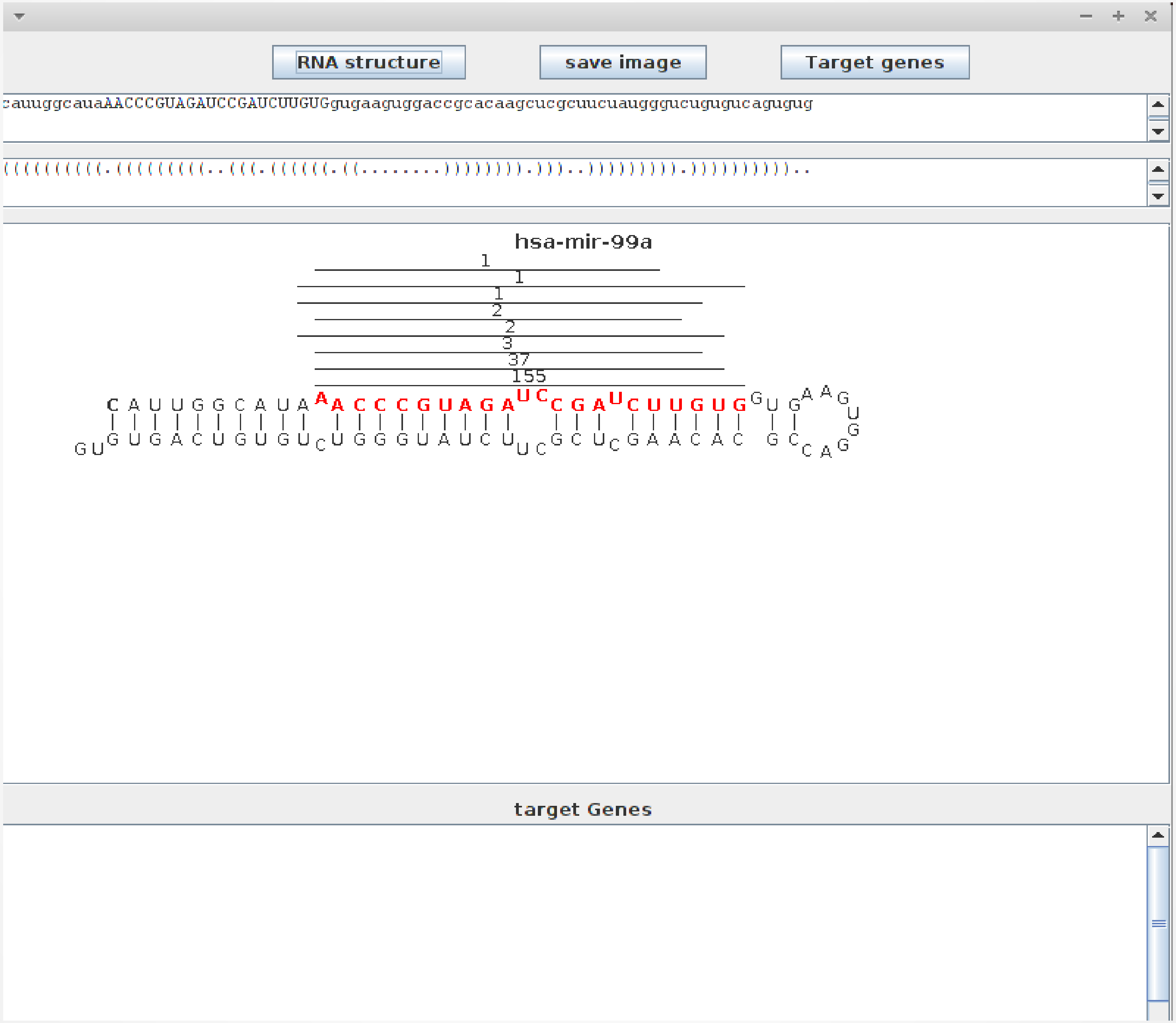
/home/valentin/Downloads/miRDeep_star_v35/dat/limitDemo.fastq

inish

Result

miRNA ID	Score	expr	loci	sequence
hsa-mir-99a	1.6	202	chr21+17911411~17911489	AACCCGUAGAUCCGAUCUUGUG
hsa-mir-378a	1.17	11	chr5+149112384~149112460	ACUGGACUUGGAGUCAGAAGG
hsa-mir-103a-1	0.62	124	chr5-167987899~167987979	AGCAGCAUUGUACAGGGCUAUGA
hsa-mir-221	0.06	16	chrX-45605599~45605677	AGCUACAUUGUCUGCUGGGUUU
hsa-mir-103a-2	-0.33	124	chr20+3898140~3898220	AGCAGCAUUGUACAGGGCUAUGA
hsa-mir-107	-0.73	45	chr10-91352505~91352581	AGCAGCAUUGUACAGGGCUAU
hsa-let-7a-2	-4.35	27	chr11-122017229~122017307	UGAGGUAGUAGGUUGUAUAGUU
hsa-let-7c	-4.45	148	chr21+17912148~17912226	UGAGGUAGUAGGUUGUAUGGUU

miRDeep*



```
cauuggcauaAACCCGUAGAUCCGAUCUUGUGgugaaguggaccgcacaagcucgcuucuaugggucugugucagugug
```

```
((((((((((.(((((((((..(((.((((((.((........)))))))).)))..))))))))).))))))))))..
```

Discussion

- simple model for miRNA detection
- basis for new tools
- more sensitive than genomic based algorithms

- hard to run
- not sensitive for SNPs

➔ use MIReNA/miRDeep*

Thank you for the attention!