Plateforme Bioinformatique Midi-Pyrénées

small RNAseq data analysis

miRNA detection

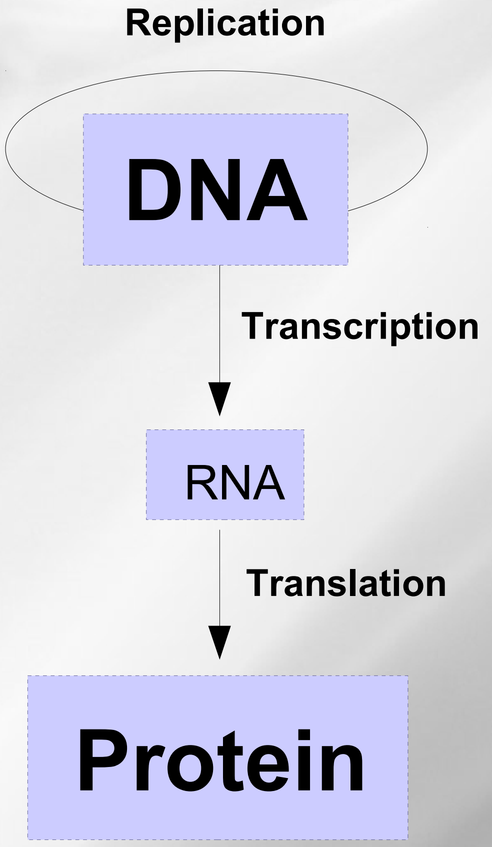
P. Bardou, C. Gaspin, S. Maman, J. Mariette & O. Rué

Introduction ncRNA

Central dogma of molecular biology

- **Evolution of the dogma : 1950-1970**

DNA structure discovery.



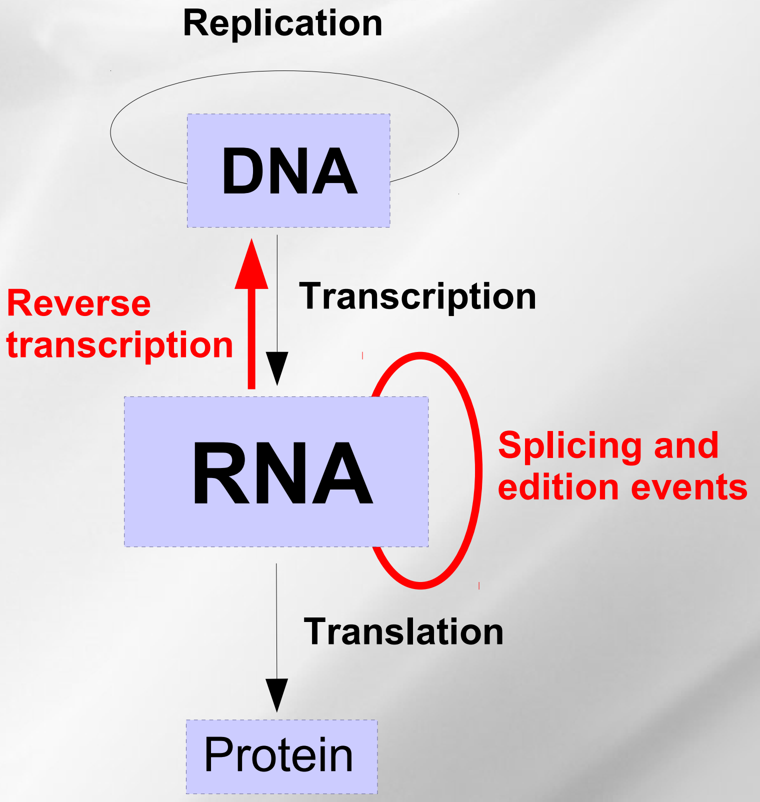
One gene = one function

Adapted from Dujon B., Toulouse, 2009

Central dogma of molecular biology

- **Evolution of the dogma : 1970-1980**

Genome analysis

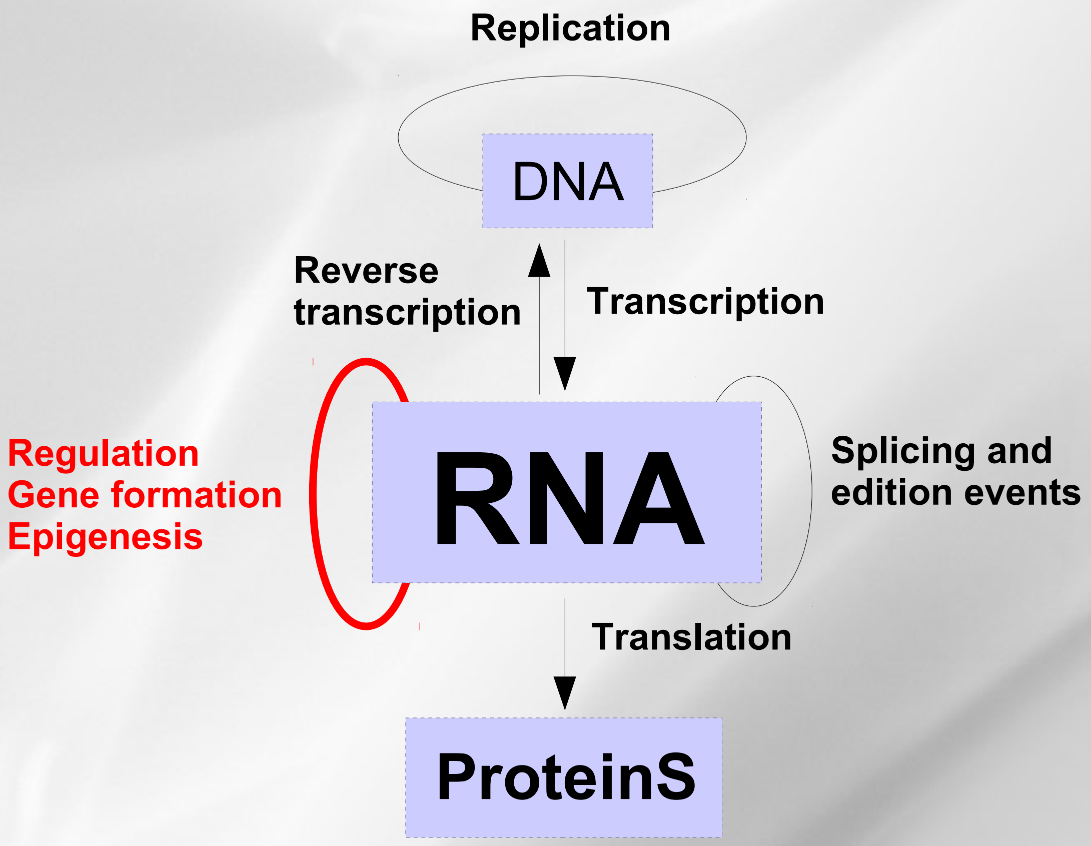


Adapted from Dujon B., Toulouse, 2009

Central dogma of molecular biology

- **Evolution of the dogma : aujourd'hui**

Genome analysis + Sequencing



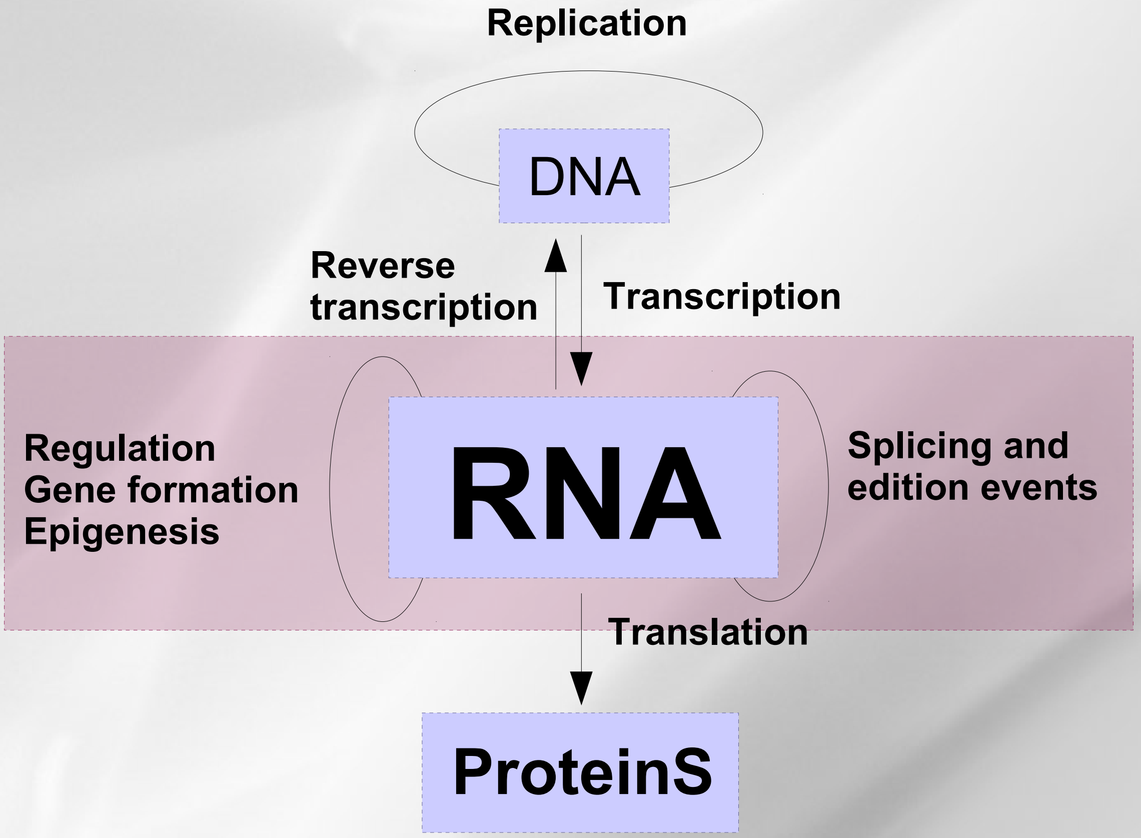
Many genes = one functionnel complex

Adapted from Dujon B., Toulouse, 2009

Central dogma of molecular biology

- **Evolution of the dogma : aujourd'hui**

Genome analysis + Sequencing



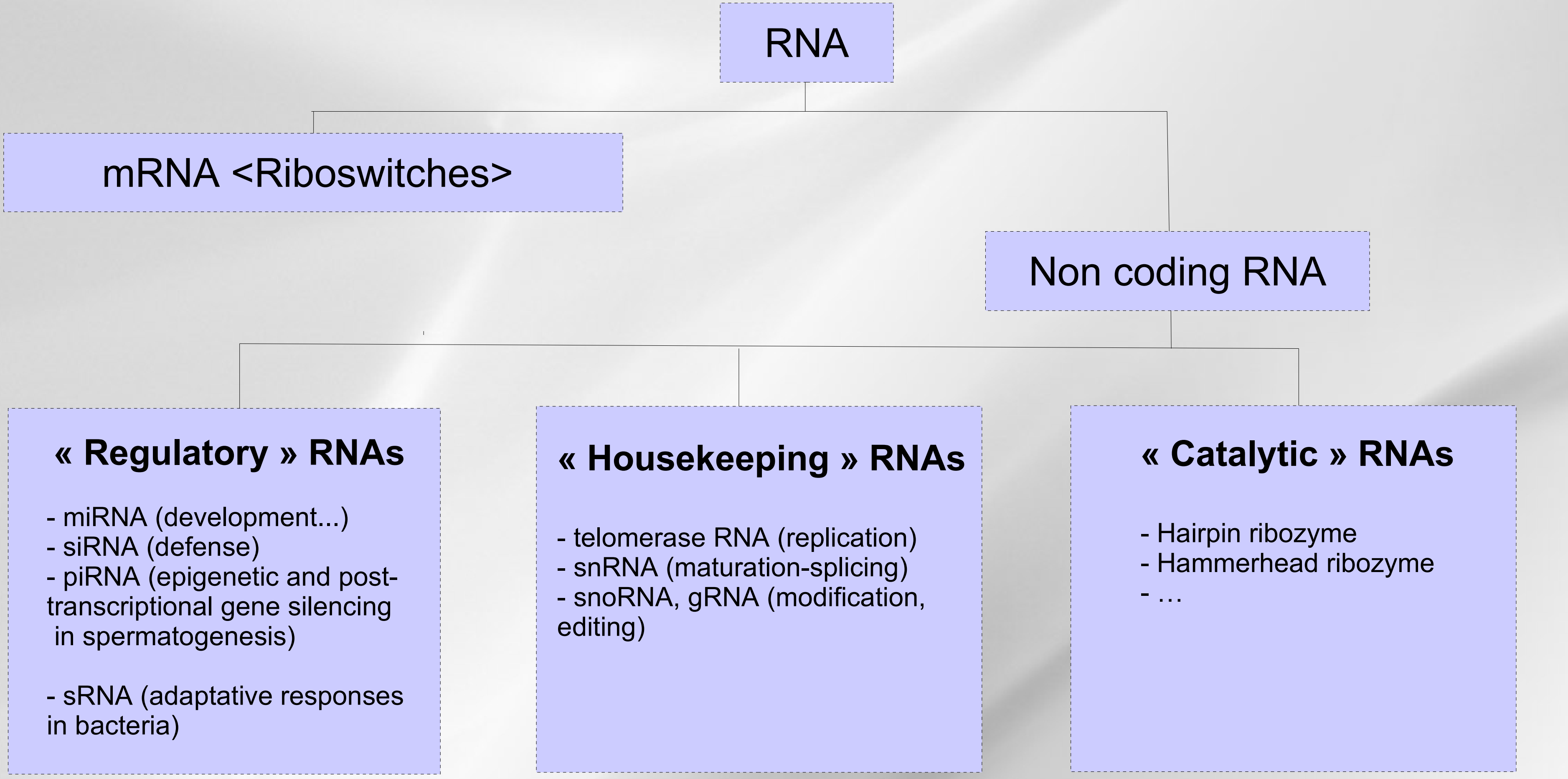
Many genes = one functionnel complex

Adapted from Dujon B., Toulouse, 2009

Plateforme Bioinformatique Midi-Pyrénées

The RNA world

- **An expending universe of RNA**



RNA

mRNA <Riboswitches>

Non coding RNA

« Regulatory » RNAs

- miRNA (development...)
- siRNA (defense)
- piRNA (epigenetic and post-transcriptional gene silencing in spermatogenesis)

- sRNA (adaptative responses in bacteria)

« Housekeeping » RNAs

- telomerase RNA (replication)
- snRNA (maturation-splicing)
- snoRNA, gRNA (modification, editing)

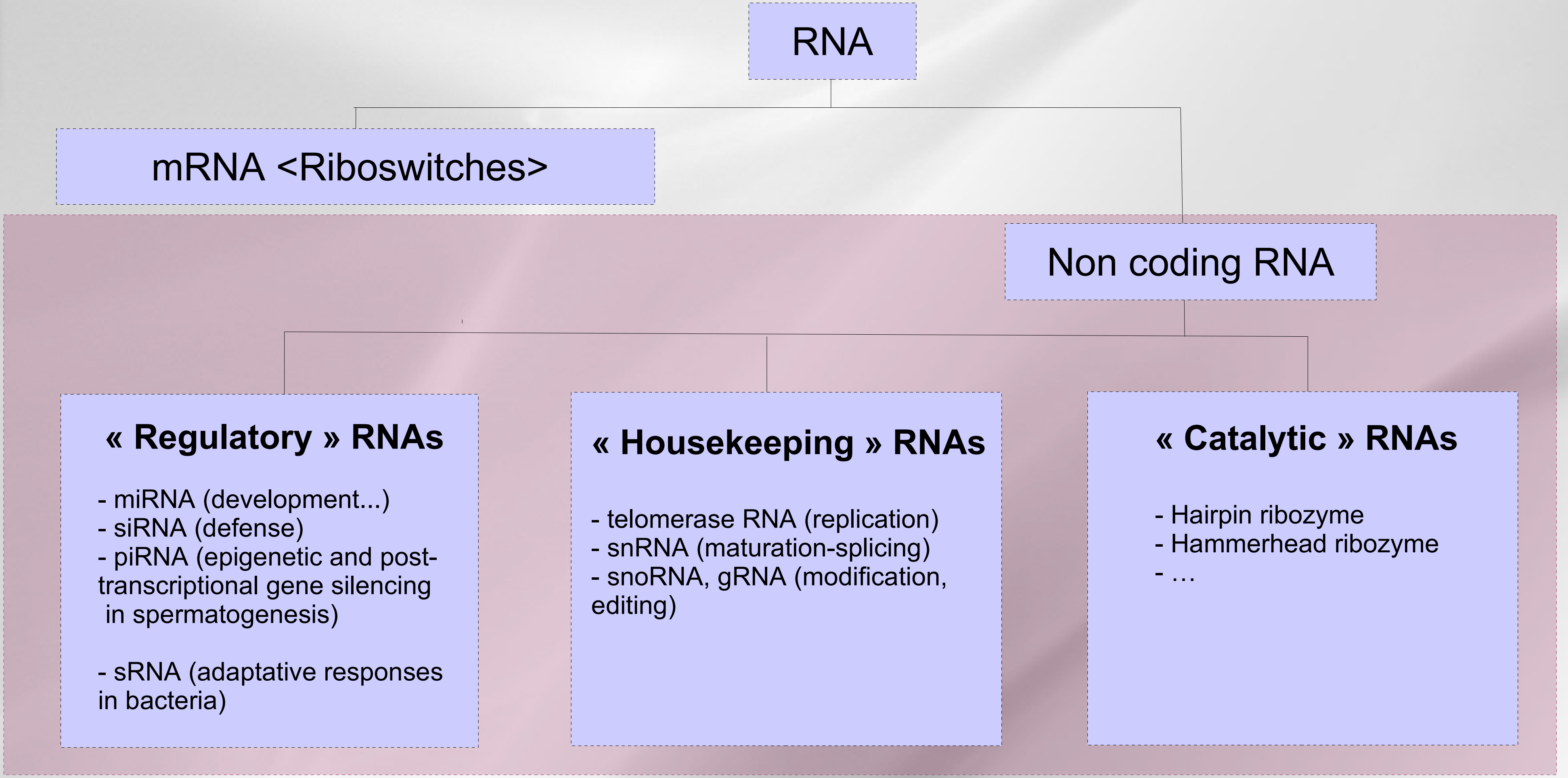
« Catalytic » RNAs

- Hairpin ribozyme
- Hammerhead ribozyme
- …

→ **Multiple roles of RNA in genes regulation**

The RNA world

- **An expending universe of RNA**



RNA

mRNA <Riboswitches>

Non coding RNA

« Regulatory » RNAs

- miRNA (development...)
- siRNA (defense)
- piRNA (epigenetic and post-transcriptional gene silencing in spermatogenesis)

- sRNA (adaptative responses in bacteria)

« Housekeeping » RNAs

- telomerase RNA (replication)
- snRNA (maturation-splicing)
- snoRNA, gRNA (modification, editing)

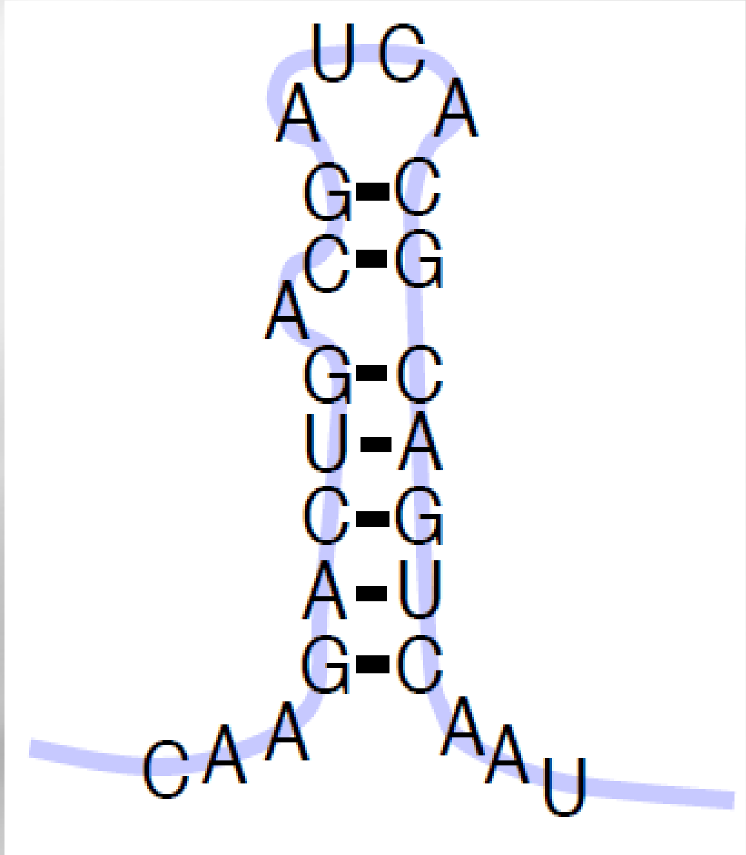
« Catalytic » RNAs

- Hairpin ribozyme
- Hammerhead ribozyme
- …

→ Multiple roles of RNA in genes regulation

RNA background

- RNA folds on itself by base pairing :
  - A with U : A-U, U-A
  - C with G : G-C, C-G
  - Sometimes G with U : U-G, G-U
- Folding = Secondary structure
- Structure related to function : ncRNA of the same family have a conserved structure
- Sequence less conserved



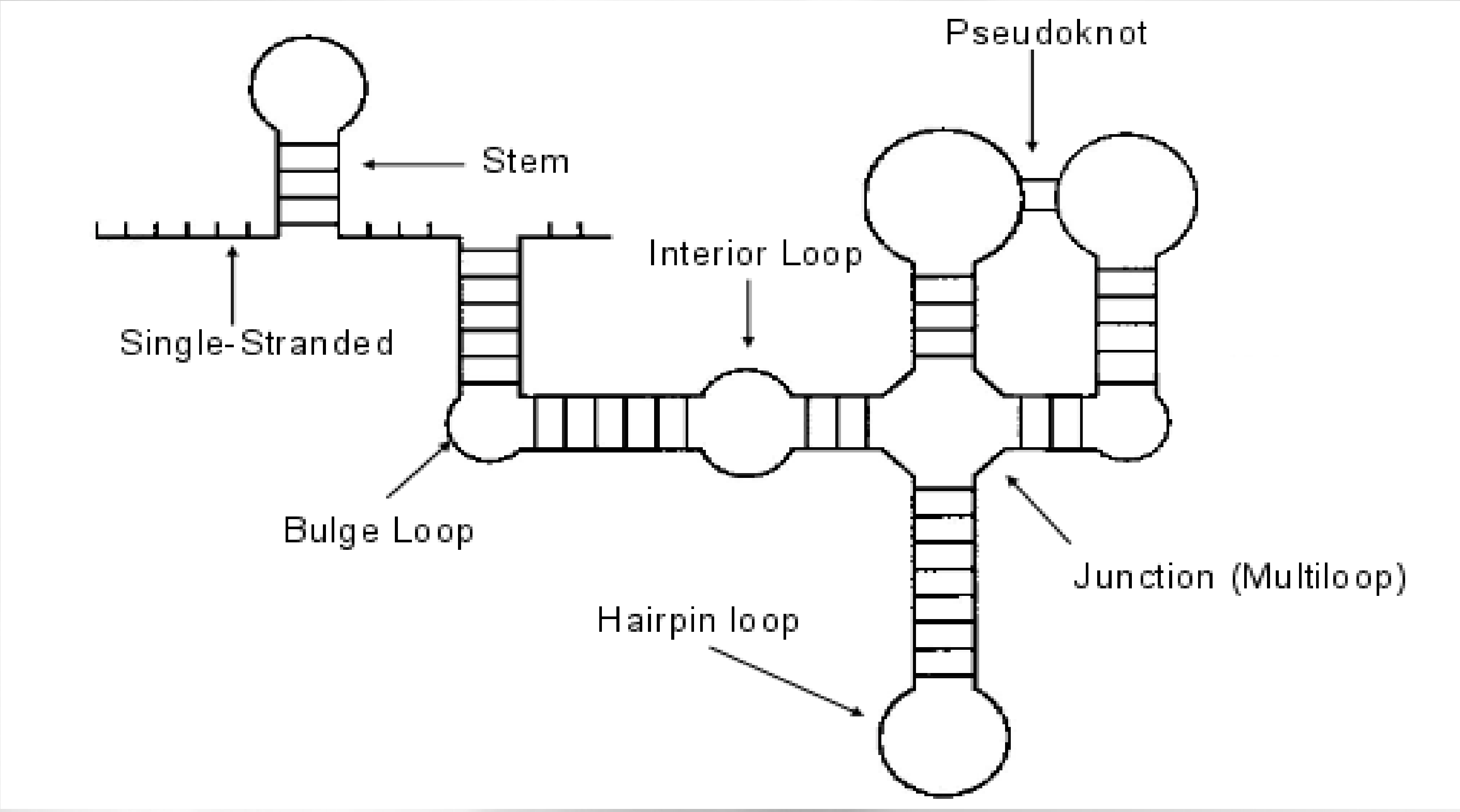
The non coding protein RNA world

- **Not predicted by gene prediction**
  - No specific signal (start, stop, splicing sites...)
  - Multiple location (intergenic, intronic, coding, antisens)
  - Variable size
  - No strong sequence conservation in general

- **A variety of existing approaches not always easy to integrate**
  - Known family: Homology prediction
  - New family: *De novo* prediction

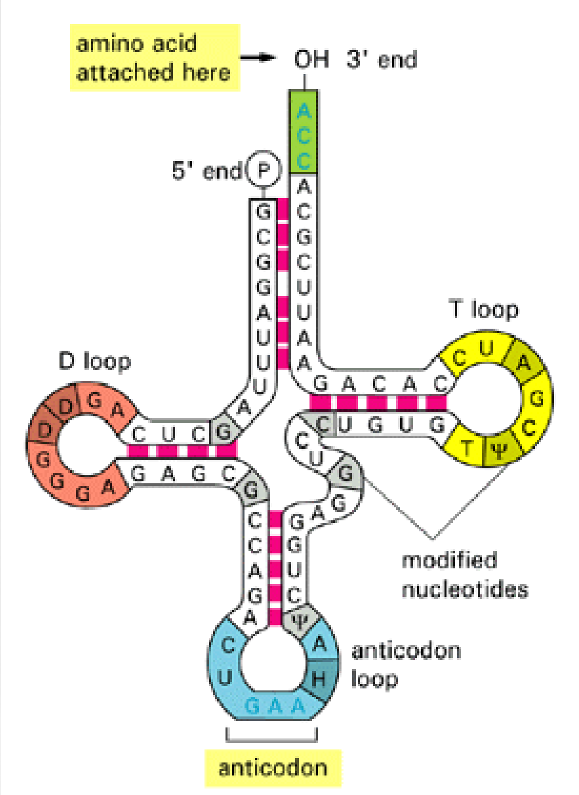
RNA background

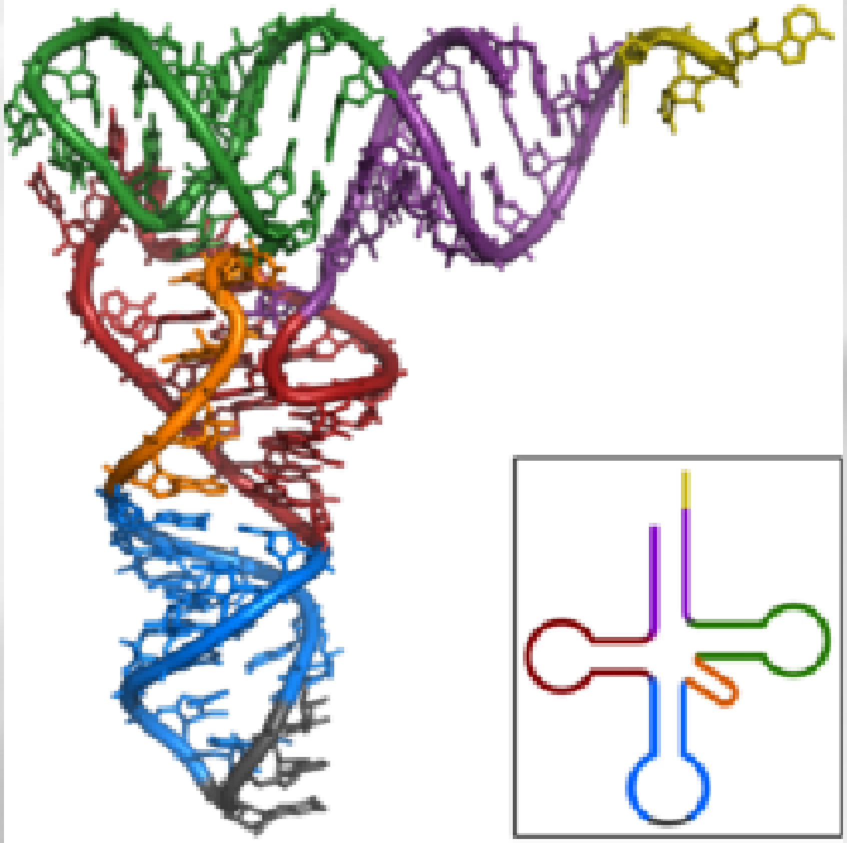
Different elementary motifs



RNA background

Example: tRNA structure





The non coding protein RNA world

- **Large non coding protein RNA**
  - >300 nt
  - rRNA, tRNA, Xist, H19, ...
  - Genome structure & expression
- **Small non coding protein RNA**
  - >30 nt
  - snoRNA, snRNA...
  - mRNA maturation, translation
- **Micro non coding protein RNA**
  - 18-30 nt
  - miRNA, hc-siRNA, ta-siRNA, nat-siRNA, piRNA...
  - PTGS, TGS, Genome stability, defense...

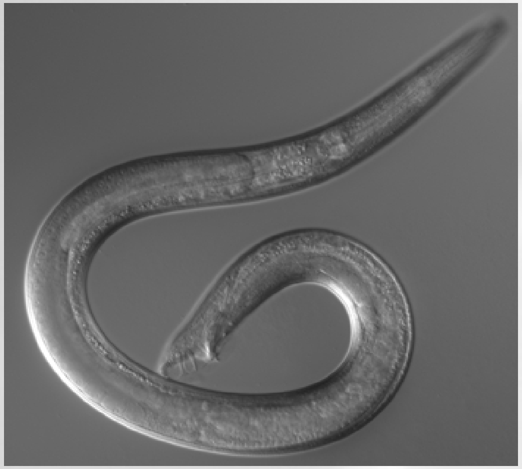
The non coding protein RNA world

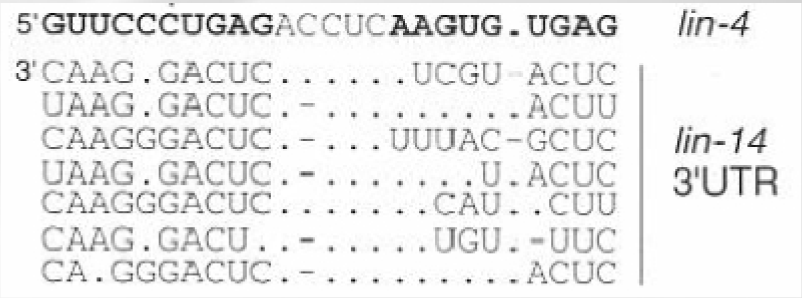
- **Large non coding protein RNA**
  - >300 nt
  - rRNA, tRNA, Xist, H19, ...
  - Genome structure & expression
- **Small non coding protein RNA**
  - >30 nt
  - snoRNA, snRNA...
  - mRNA maturation, translation
- **Micro non coding protein RNA**
  - 18-30 nt
  - miRNA, hc-siRNA, ta-siRNA, nat-siRNA, piRNA...
  - PTGS, TGS, Genome stability, defense...

Introduction to miRNA world and sRNAseq

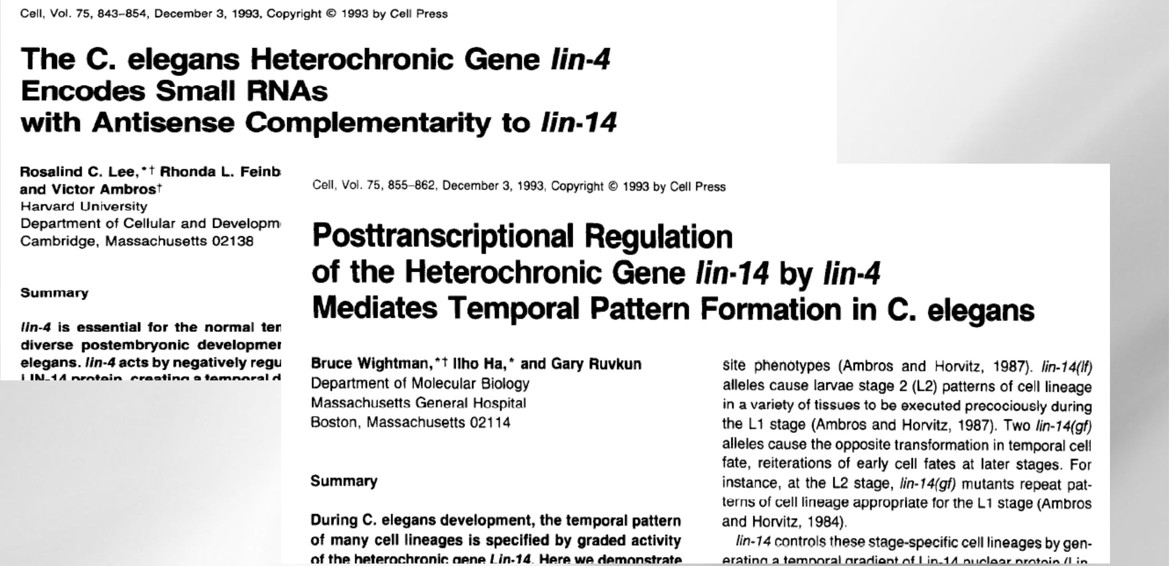
The miRNA world

- **Discovery of lin-4 in C. elegans in 1993**





```
5'GUUCCCUGAGACCUCAAGUG.UGAG         lin-4

3'CAAG.GACUC......UCGU-ACUC
  UAAG.GACUC.-..........ACUU
  CAAGGGACUC.-...UUUAC-GCUC         lin-14
  UAAG.GACUC.-........U.ACUC        3'UTR
  CAAGGGACUC........CAU..CUU
  CAAG.GACU..-......UGU.-UUC
  CA.GGGACUC.-..........ACUC
```



Cell, Vol. 75, 843–854, December 3, 1993, Copyright © 1993 by Cell Press

The C. elegans Heterochronic Gene *lin-4* Encodes Small RNAs with Antisense Complementarity to *lin-14*

Rosalind C. Lee,*† Rhonda L. Feinb
and Victor Ambros†
Harvard University
Department of Cellular and Developm
Cambridge, Massachusetts 02138

Summary

lin-4 is essential for the normal ter
diverse postembryonic developmer
elegans. *lin-4* acts by negatively regu
LIN-14 protein, creating a temporal d

Cell, Vol. 75, 855–862, December 3, 1993, Copyright © 1993 by Cell Press

Posttranscriptional Regulation of the Heterochronic Gene *lin-14* by *lin-4* Mediates Temporal Pattern Formation in C. elegans

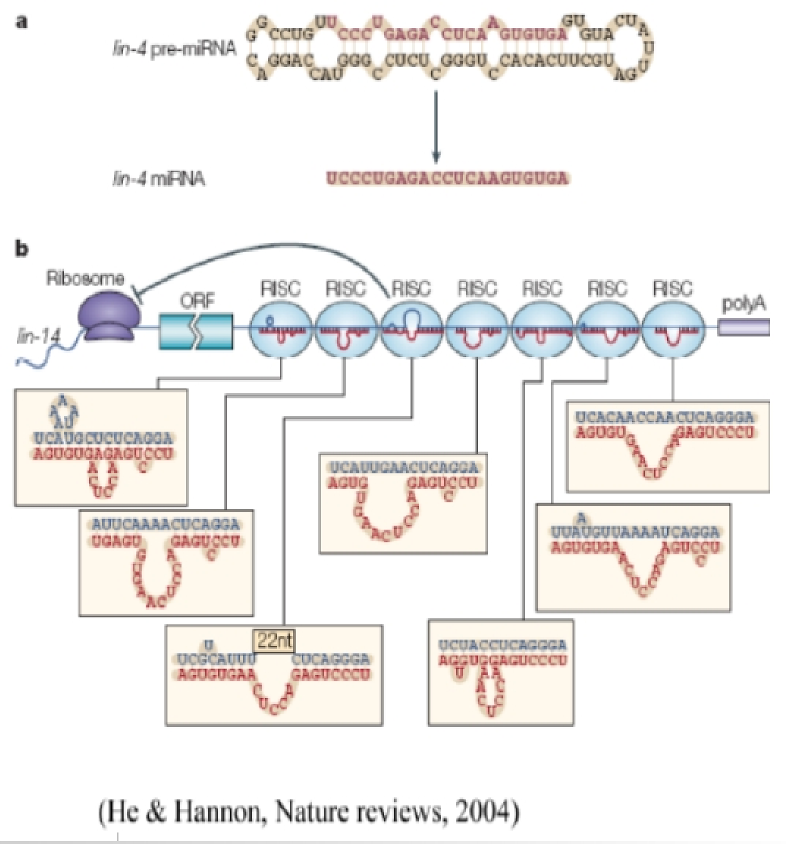
Bruce Wightman,*† Ilho Ha,* and Gary Ruvkun
Department of Molecular Biology
Massachusetts General Hospital
Boston, Massachusetts 02114

Summary

During C. elegans development, the temporal pattern of many cell lineages is specified by graded activity of the heterochronic gene *Lin-14*. Here we demonstrate

site phenotypes (Ambros and Horvitz, 1987). *lin-14(lf)* alleles cause larvae stage 2 (L2) patterns of cell lineage in a variety of tissues to be executed precociously during the L1 stage (Ambros and Horvitz, 1987). Two *lin-14(gf)* alleles cause the opposite transformation in temporal cell fate, reiterations of early cell fates at later stages. For instance, at the L2 stage, *lin-14(gf)* mutants repeat patterns of cell lineage appropriate for the L1 stage (Ambros and Horvitz, 1984).

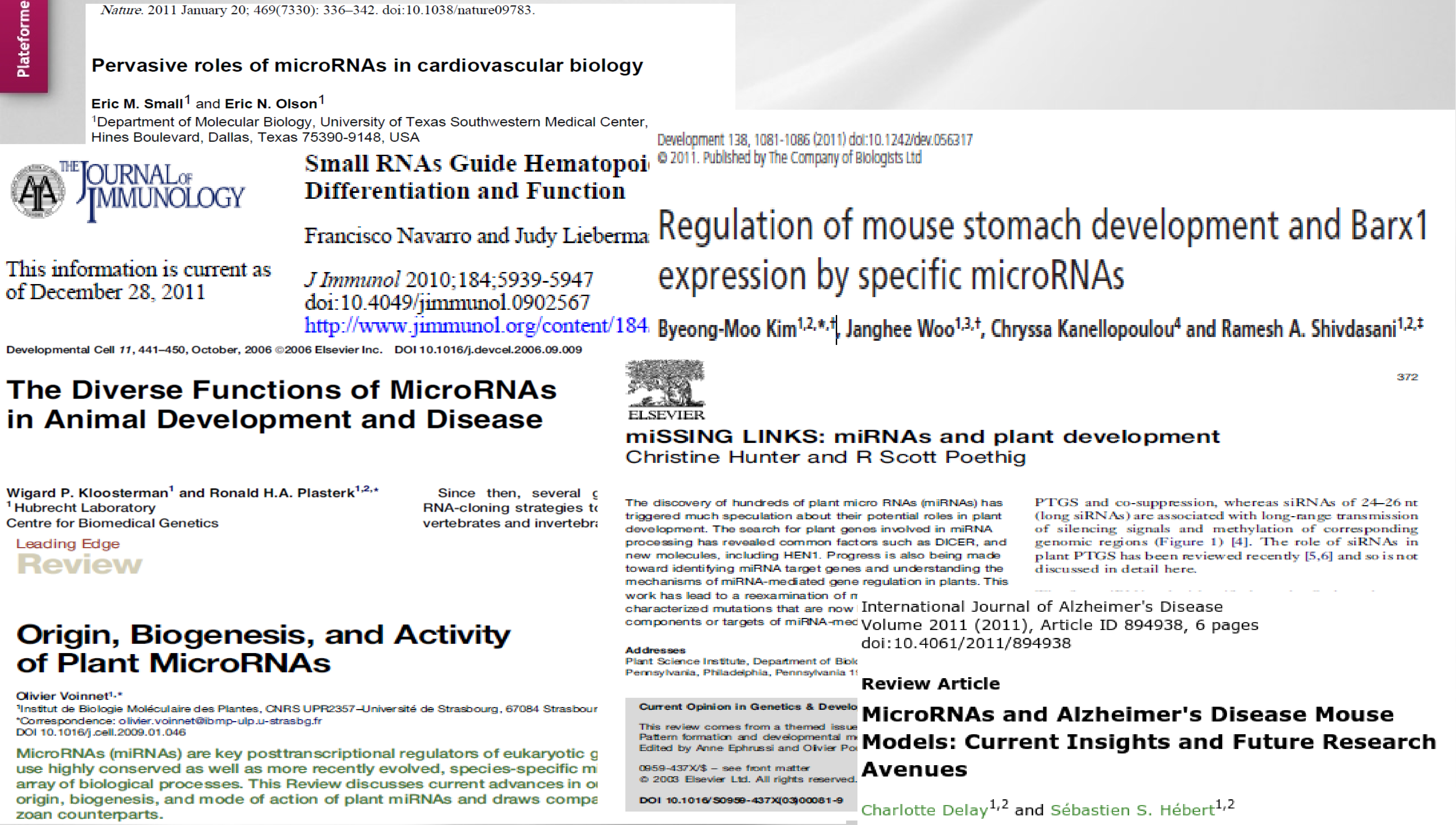
lin-14 controls these stage-specific cell lineages by generating a temporal gradient of Lin-14 nuclear protein (Lin



(He & Hannon, Nature reviews, 2004)

The miRNA world

- **A key regulation function**



Nature. 2011 January 20; 469(7330): 336–342. doi:10.1038/nature09783.

Pervasive roles of microRNAs in cardiovascular biology

Eric M. Small[1] and **Eric N. Olson**[1]
[1]Department of Molecular Biology, University of Texas Southwestern Medical Center, Hines Boulevard, Dallas, Texas 75390-9148, USA

THE JOURNAL OF IMMUNOLOGY

Small RNAs Guide Hematopoi Differentiation and Function

Francisco Navarro and Judy Lieberma

This information is current as of December 28, 2011

J Immunol 2010;184;5939-5947
doi:10.4049/jimmunol.0902567
http://www.jimmunol.org/content/184

Development 138, 1081-1086 (2011) doi:10.1242/dev.056317
© 2011. Published by The Company of Biologists Ltd

Regulation of mouse stomach development and Barx1 expression by specific microRNAs

Byeong-Moo Kim[1,2,*,†], Janghee Woo[1,3,†], Chryssa Kanellopoulou[4] and Ramesh A. Shivdasani[1,2,‡]

Developmental Cell *11*, 441–450, October, 2006 ©2006 Elsevier Inc. DOI 10.1016/j.devcel.2006.09.009

The Diverse Functions of MicroRNAs in Animal Development and Disease

Wigard P. Kloosterman[1] and **Ronald H.A. Plasterk**[1,2,*]
[1]Hubrecht Laboratory
Centre for Biomedical Genetics

Since then, several g RNA-cloning strategies to vertebrates and invertebra

Leading Edge
Review

Origin, Biogenesis, and Activity of Plant MicroRNAs

Olivier Voinnet[1,*]
[1]Institut de Biologie Moléculaire des Plantes, CNRS UPR2357–Université de Strasbourg, 67084 Strasbour
*Correspondence: olivier.voinnet@ibmp-ulp.u-strasbg.fr
DOI 10.1016/j.cell.2009.01.046

MicroRNAs (miRNAs) are key posttranscriptional regulators of eukaryotic g use highly conserved as well as more recently evolved, species-specific mi array of biological processes. This Review discusses current advances in o origin, biogenesis, and mode of action of plant miRNAs and draws compa zoan counterparts.

ELSEVIER

372

miSSING LINKS: miRNAs and plant development
Christine Hunter and R Scott Poethig

The discovery of hundreds of plant micro RNAs (miRNAs) has triggered much speculation about their potential roles in plant development. The search for plant genes involved in miRNA processing has revealed common factors such as DICER, and new molecules, including HEN1. Progress is also being made toward identifying miRNA target genes and understanding the mechanisms of miRNA-mediated gene regulation in plants. This work has lead to a reexamination of m characterized mutations that are now components or targets of miRNA-med

Addresses
Plant Science Institute, Department of Biol
Pennsylvania, Philadelphia, Pennsylvania 1

Current Opinion in Genetics & Develo

This review comes from a themed issue
Pattern formation and developmental m
Edited by Anne Ephrussi and Olivier Po

0959-437X/$ – see front matter
© 2003 Elsevier Ltd. All rights reserved.

DOI 10.1016/S0959-437X(03)00081-9

PTGS and co-suppression, whereas siRNAs of 24–26 nt (long siRNAs) are associated with long-range transmission of silencing signals and methylation of corresponding genomic regions (Figure 1) [4]. The role of siRNAs in plant PTGS has been reviewed recently [5,6] and so is not discussed in detail here.

International Journal of Alzheimer's Disease
Volume 2011 (2011), Article ID 894938, 6 pages
doi:10.4061/2011/894938

Review Article

MicroRNAs and Alzheimer's Disease Mouse Models: Current Insights and Future Research Avenues

Charlotte Delay[1,2] and Sébastien S. Hébert[1,2]

The miRNA world

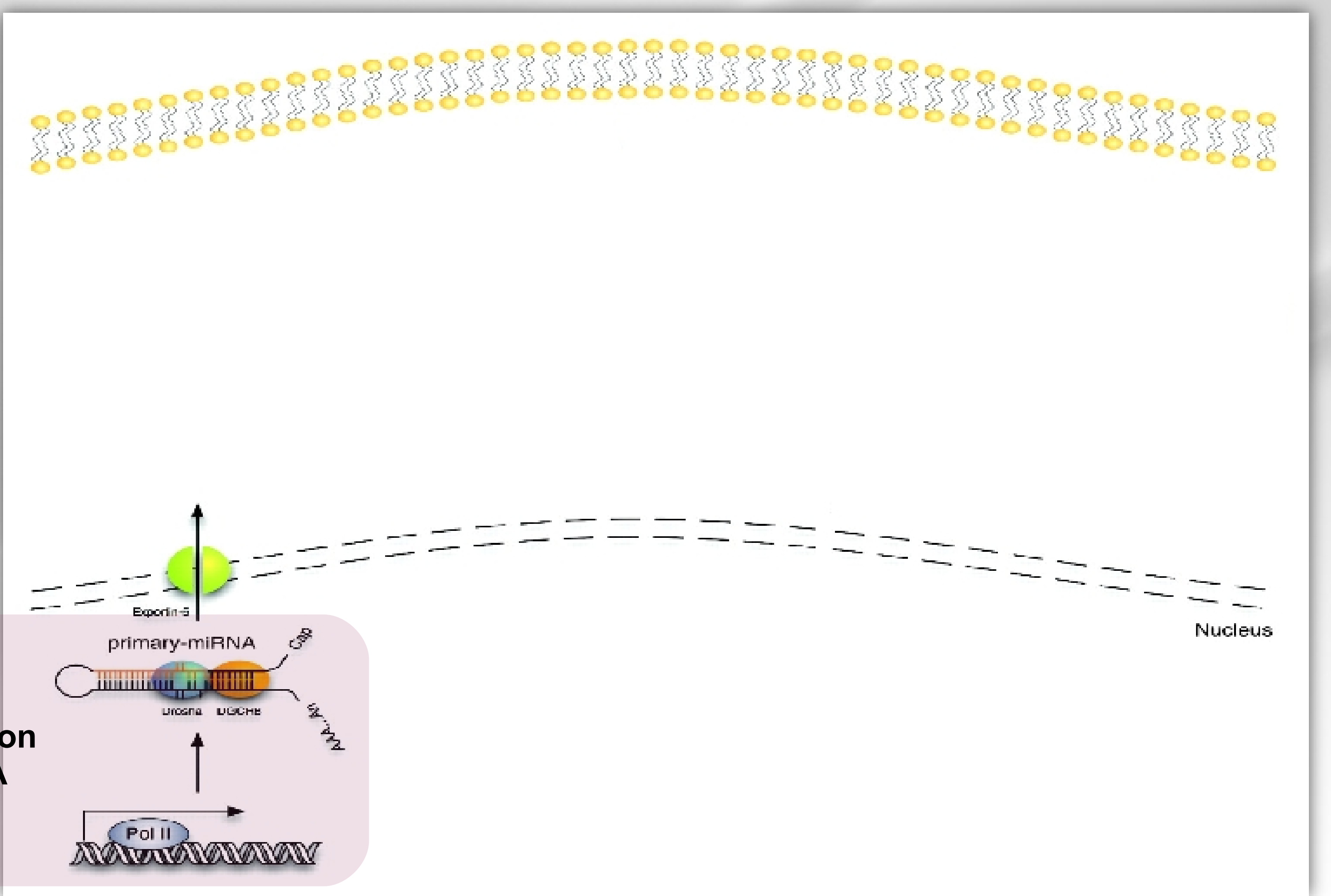
Regulatory functions

- **Animals**
  - **Developmental timing (C. elegans): lin-4, let-7**
  - **Neuronal left/right asymetry (C. elegans): Lys-6, mir-273**
  - **Programmed cell death/fat metabolism (D. melanogaster): mir-14**
  - **Notch signaling (D. malanogaster): mir-7**
  - **Brain morphogenesis (Zebrafish): mir-430**
  - **Myogeneses and cardiogenesis: mir-1, miR-181, miR-133**
  - **Insulin secretion: miR-375**
  - **…**

  1600 precursors in Human !!! (ref: miRBase, August 2012)

- **Plants**
  - **Floral timing and leaf development: miR-156**
  - **Organ polarity, vascular and meristen development: mir-165, miR-166**
  - **Expression of auxin response genes: miR-160**
  - **...**

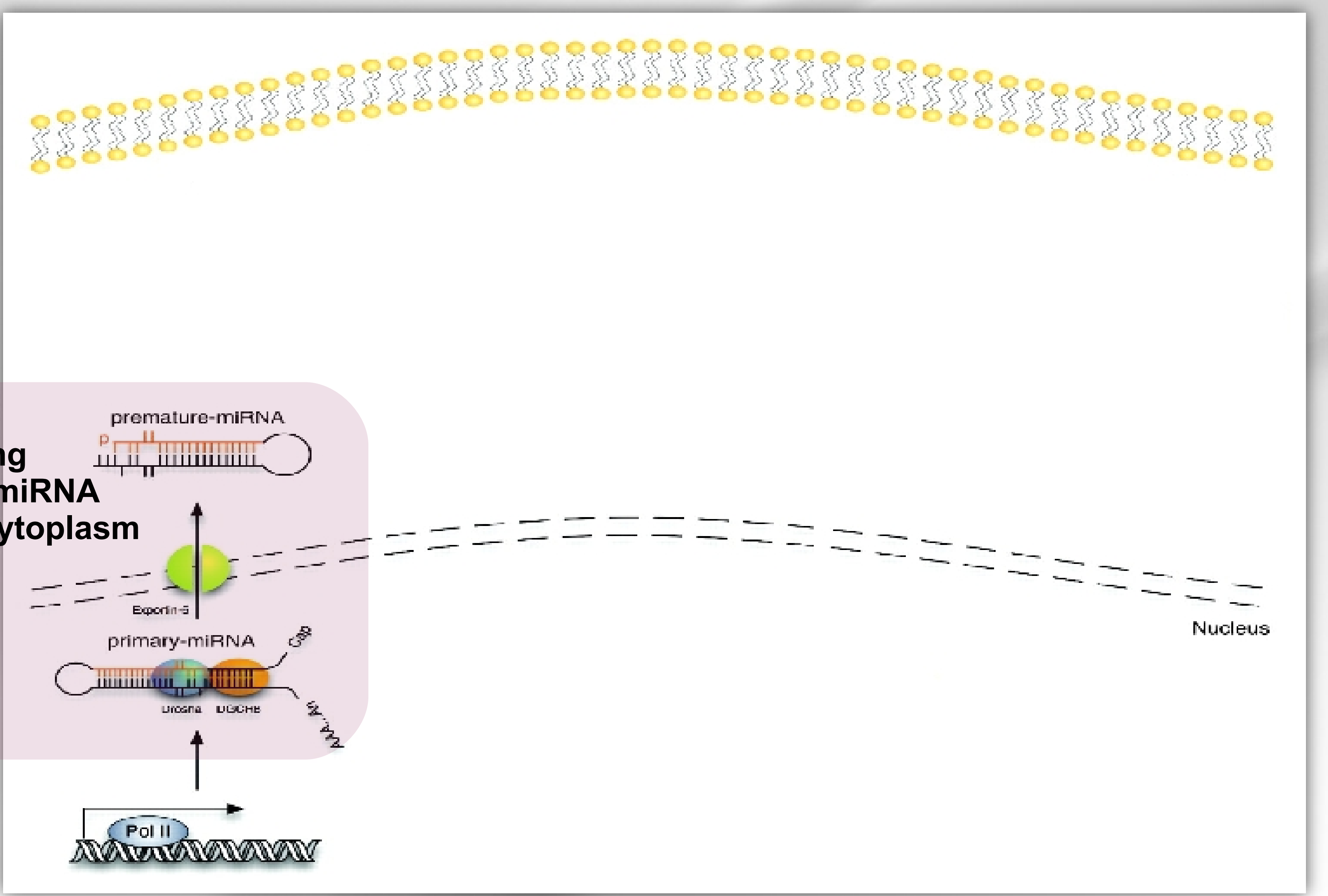
The miRNA biogenesis



Pol II transcription
Into a pri-miRNA

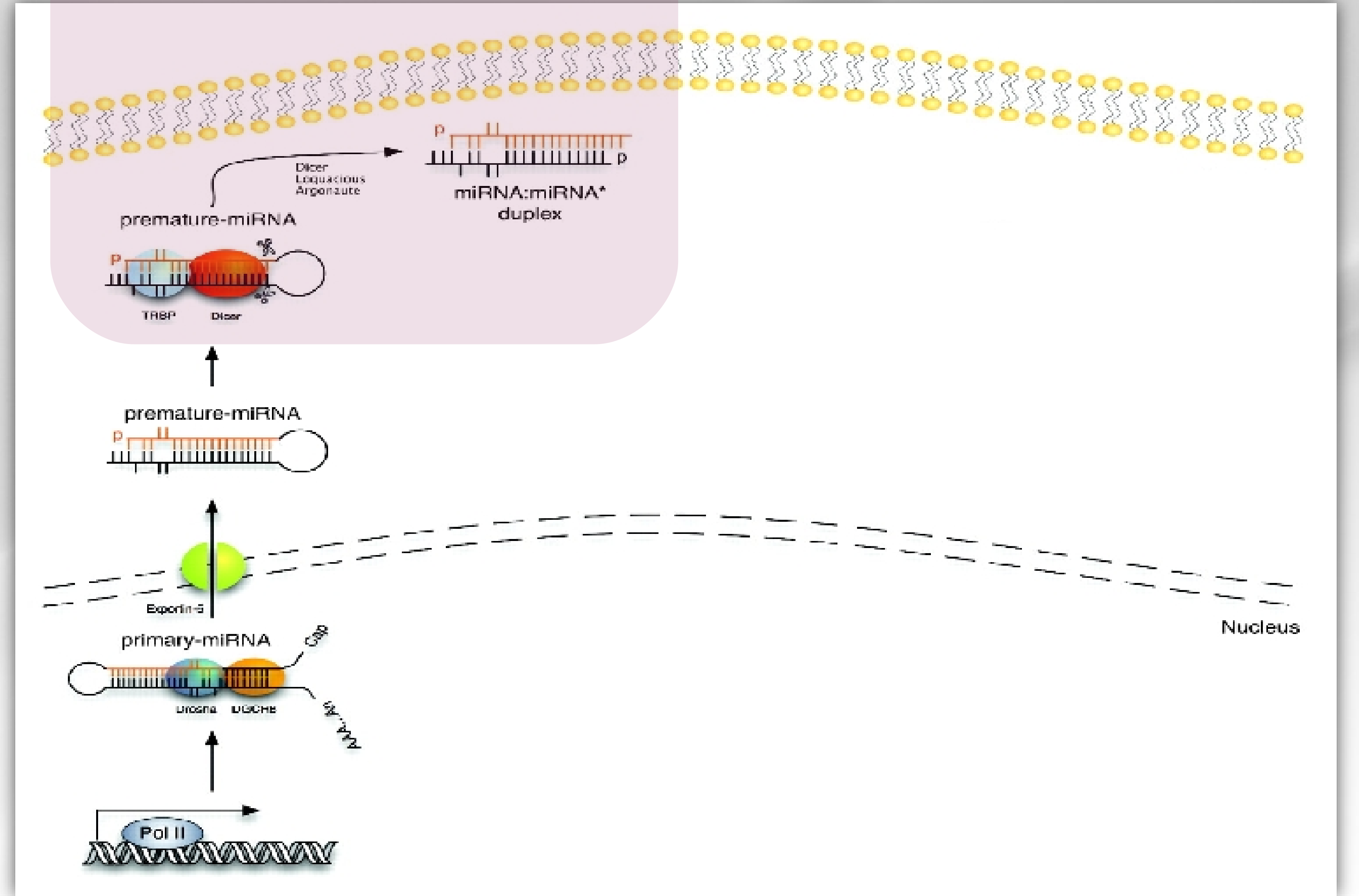
The miRNA biogenesis

Drosha processing
one or more pre-miRNA
Exported in the cytoplasm

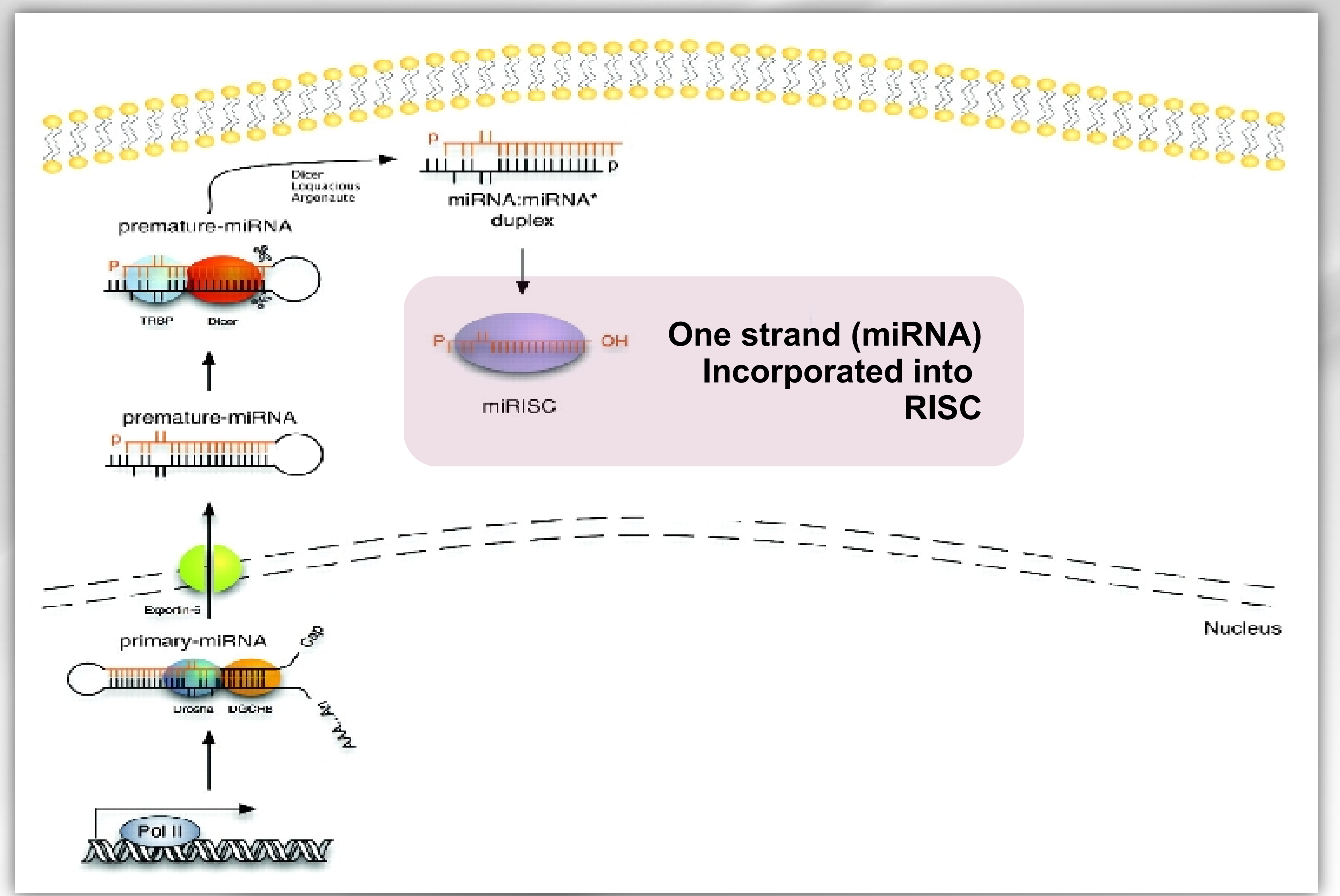


The miRNA biogenesis

Dicer processing
Into a duplex miRNA
Structure



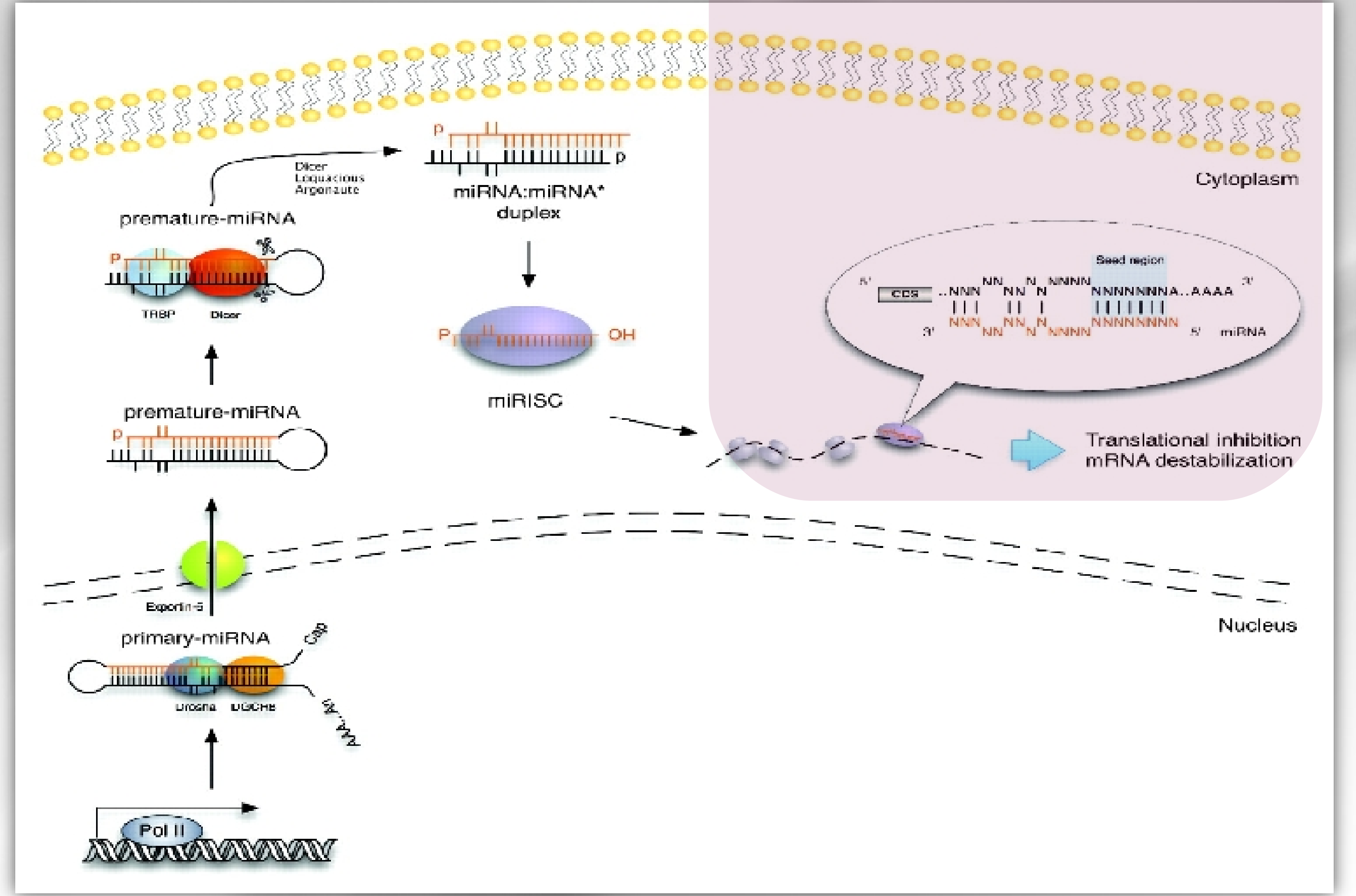
The miRNA biogenesis



One strand (miRNA)
Incorporated into
RISC

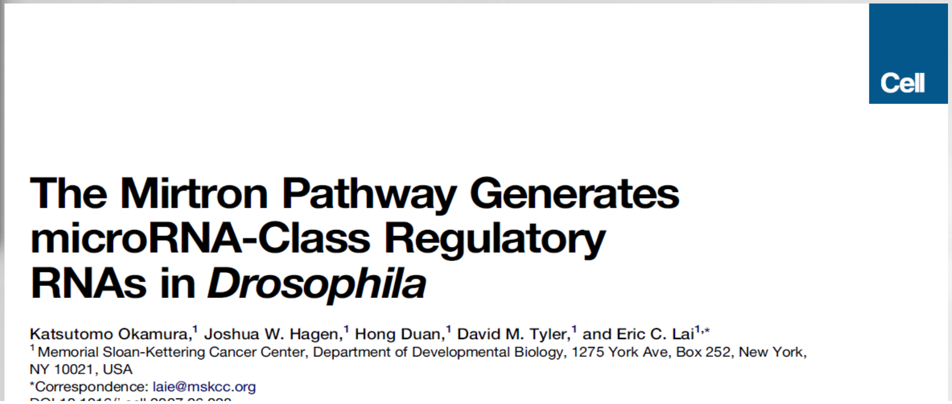
The miRNA biogenesis

target mRNA translationally repressed



Plateforme Bioinformatique Midi-Pyrénées

The miRNA biogenesis



Cell

The Mirtron Pathway Generates microRNA-Class Regulatory RNAs in *Drosophila*

Katsutomo Okamura,[1] Joshua W. Hagen,[1] Hong Duan,[1] David M. Tyler,[1] and Eric C. Lai[1,*]
[1] Memorial Sloan-Kettering Cancer Center, Department of Developmental Biology, 1275 York Ave, Box 252, New York, NY 10021, USA
*Correspondence: laie@mskcc.org

DOI 10.1016/j.cell.2007.06.028

Molecular Cell

Cell PRESS

Volume 28, Issue 2, 26 October 2007, Pages 328–336

Resource

Mammalian Mirtron Genes

Eugene Berezikov[1], Wei-Jen Chung[2], Jason Willis[2], Edwin Cuppen[1], Eric C. Lai[2]
[1] Hubrecht Institute, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands
[2] Sloan-Kettering Institute, 1275 York Avenue, Box 252, New York, NY 10021, USA

http://dx.doi.org/10.1016/j.molcel.2007.09.028, How to Cite or Link Using DOI

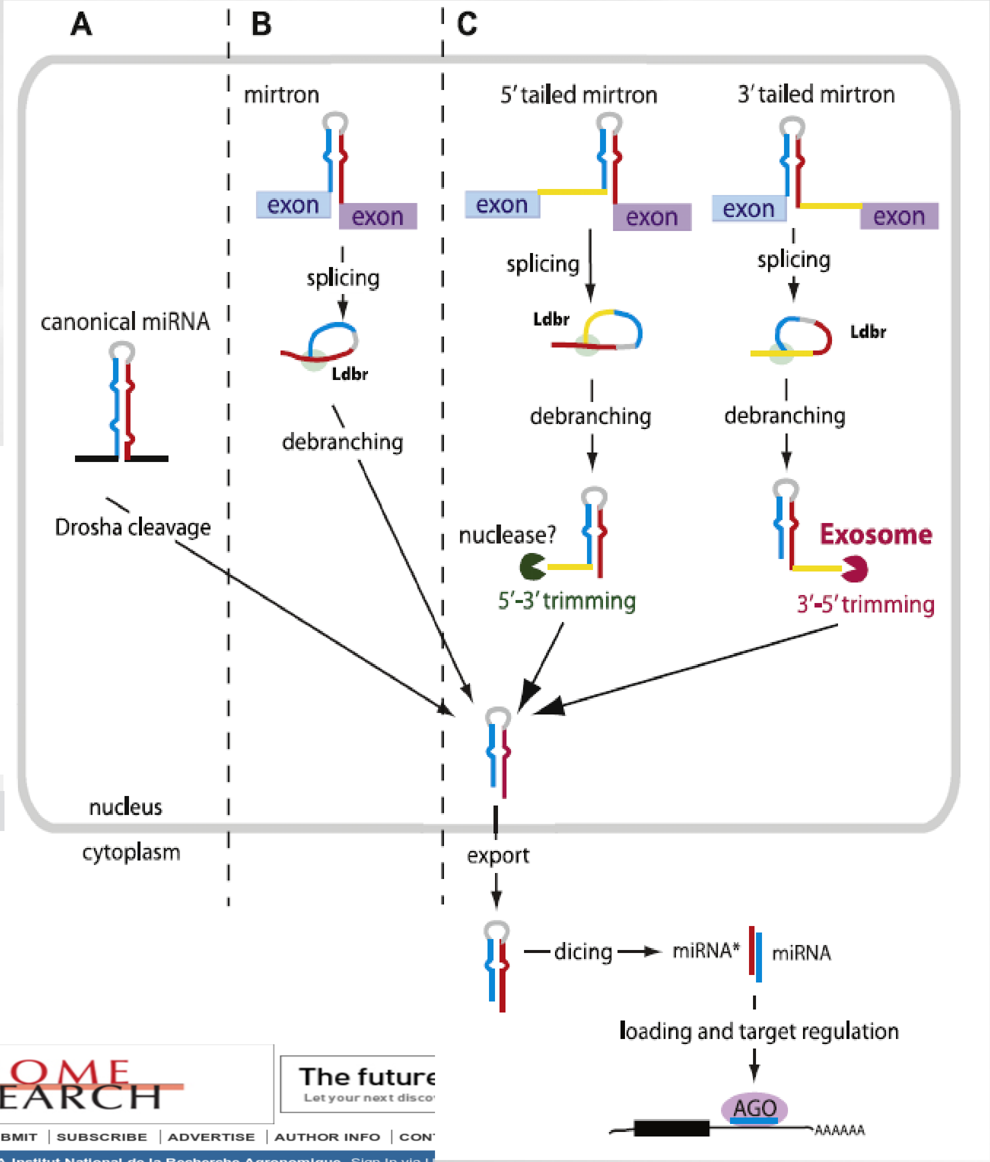
Vol 448 | 5 July 2007 | doi:10.1038/nature05983

nature

LETTERS

Intronic microRNA precursors that bypass Drosha processing

J. Graham Ruby[1,2*], Calvin H. Jan[1,2*] & David P. Bartel[1,2]



CSH PRESS

Genome Research

The future
Let your next disco

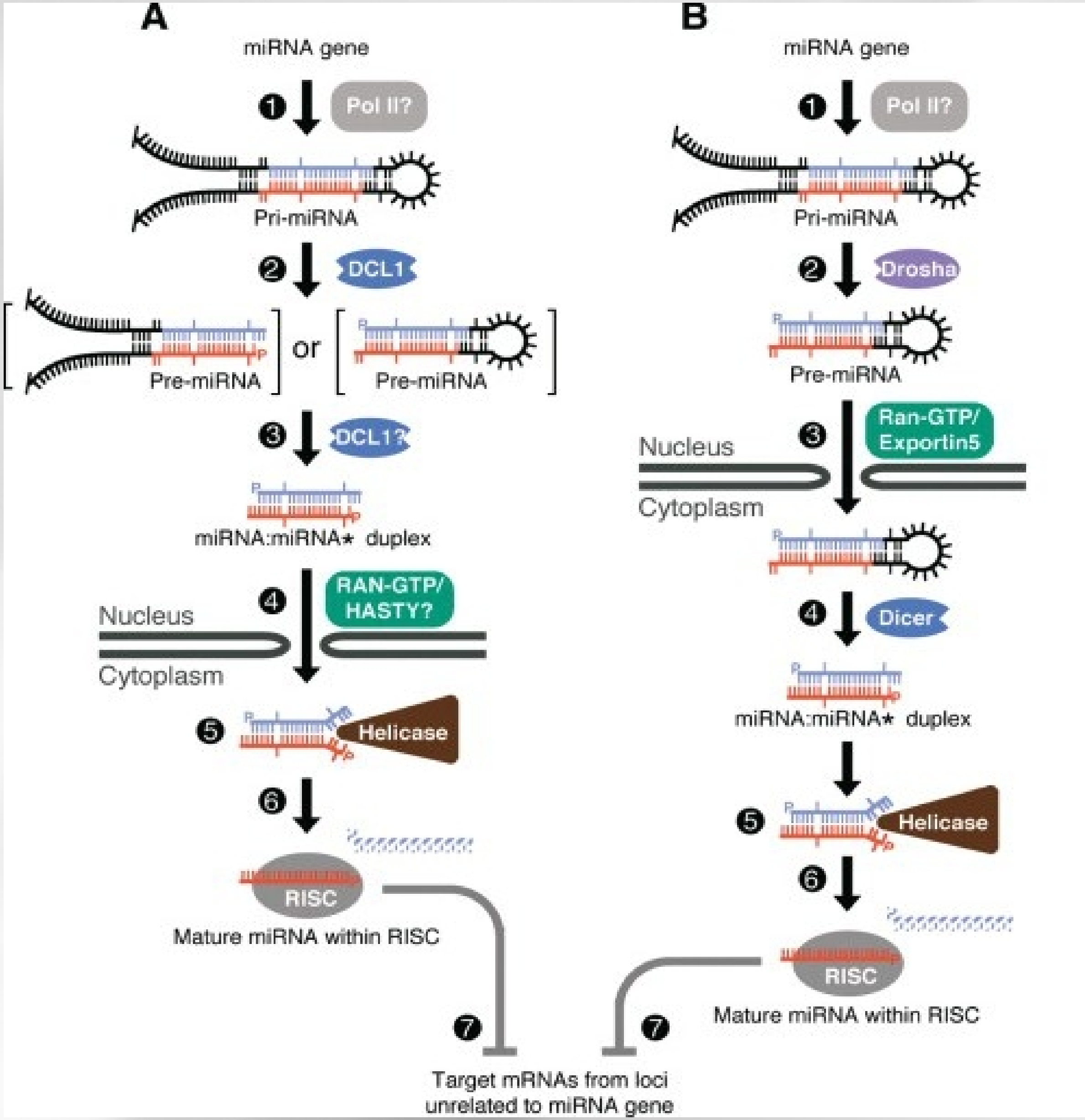
HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CON

Institution: INRA Institut National de la Recherche Agronomique Sign In via U

Discovery of hundreds of mirtrons in mouse and human small RNA data

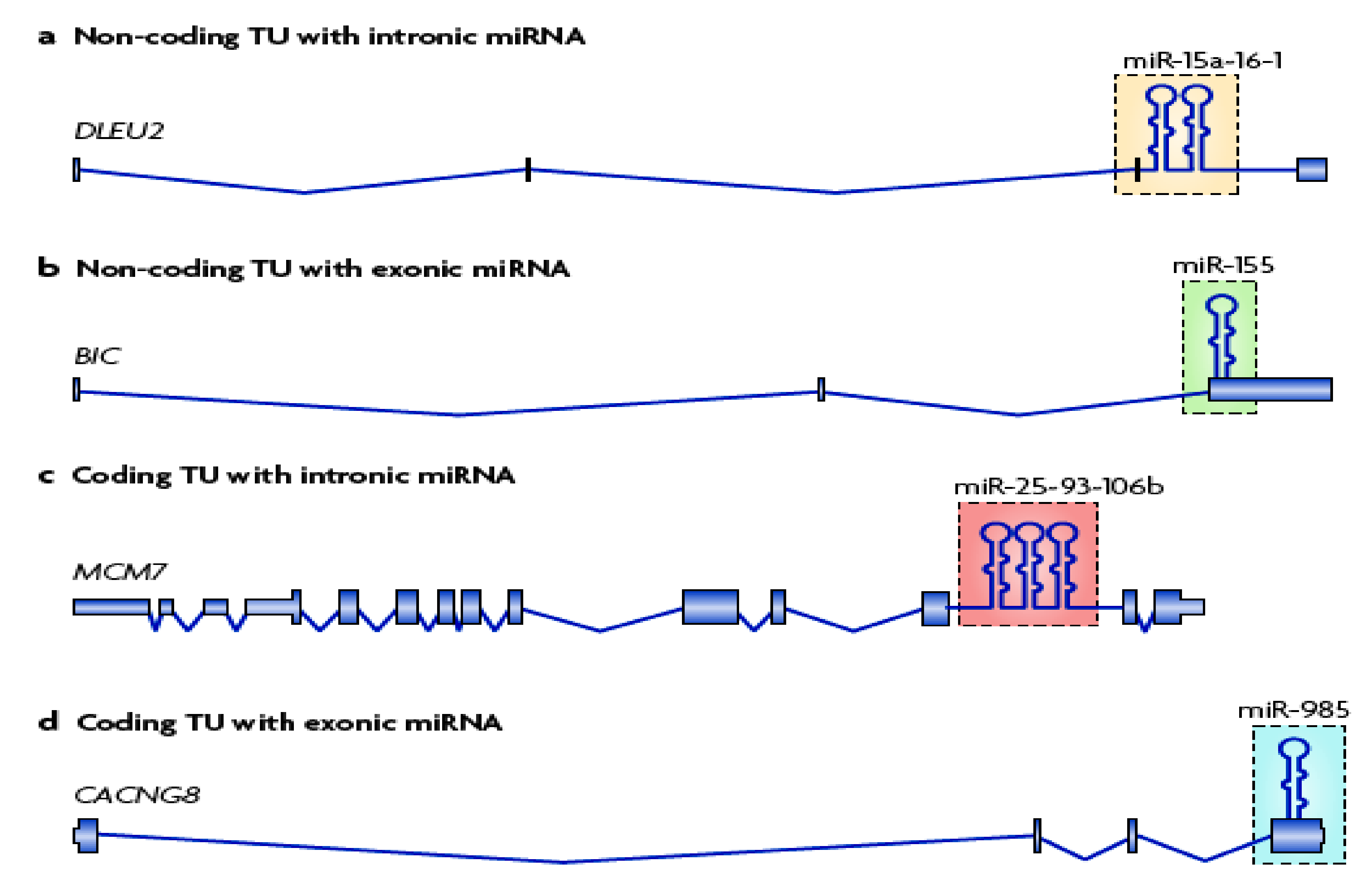
Erik Ladewig[1], Katsutomo Okamura[1,2], Alex S. Flynt[1], Jakub O. Westholm[1] and Eric C. Lai[1,3]

The miRNA biogenesis



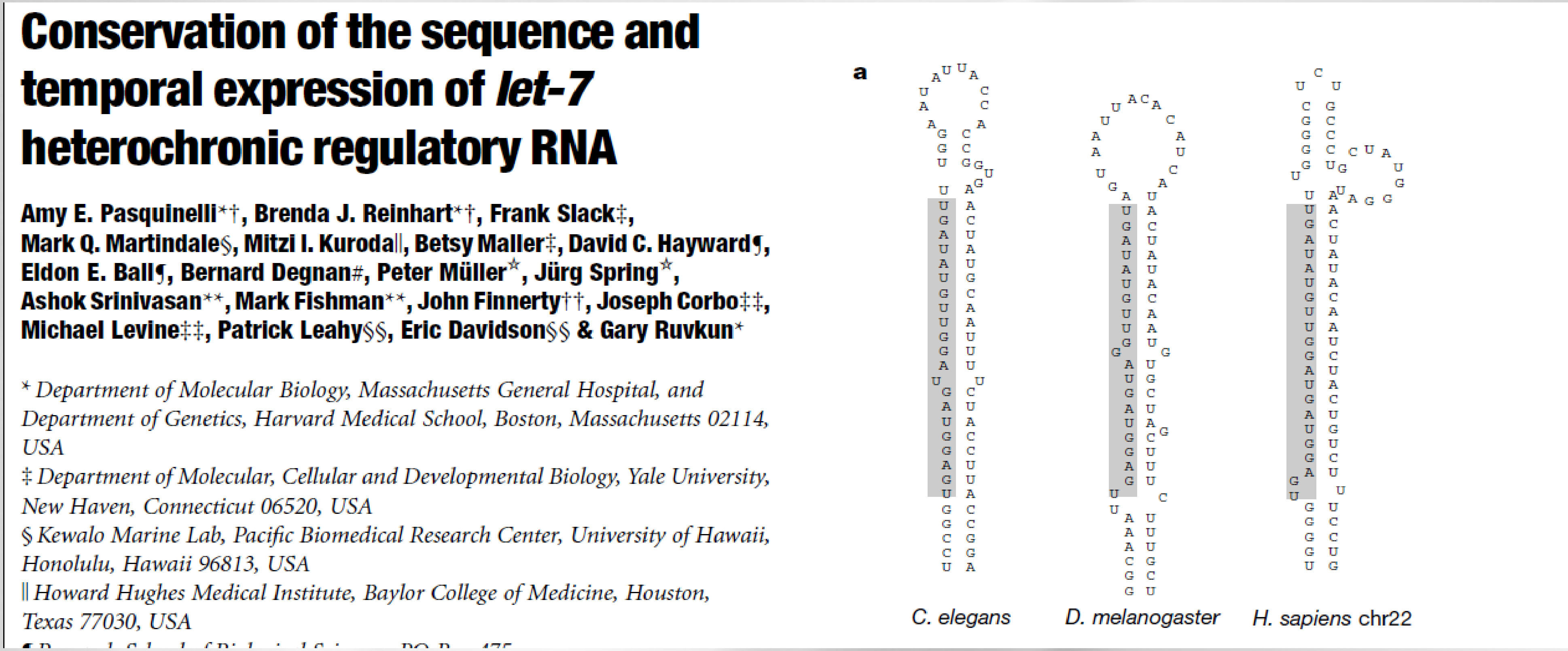
D. Bartel, Cell, 2004

The miRNA location


a Non-coding TU with intronic miRNA
miR-15a-16-1
DLEU2
b Non-coding TU with exonic miRNA
miR-155
BIC
c Coding TU with intronic miRNA
miR-25-93-106b
MCM7
d Coding TU with exonic miRNA
miR-985
CACNG8

→ **Cluster organisation**

miRNA conservation



Conservation of the sequence and temporal expression of *let-7* heterochronic regulatory RNA

Amy E. Pasquinelli*†, Brenda J. Reinhart*†, Frank Slack‡, Mark Q. Martindale§, Mitzi I. Kuroda‖, Betsy Maller‡, David C. Hayward¶, Eldon E. Ball¶, Bernard Degnan#, Peter Müller☆, Jürg Spring☆, Ashok Srinivasan, Mark Fishman**, John Finnerty††, Joseph Corbo‡‡, Michael Levine‡‡, Patrick Leahy§§, Eric Davidson§§ & Gary Ruvkun***

** Department of Molecular Biology, Massachusetts General Hospital, and Department of Genetics, Harvard Medical School, Boston, Massachusetts 02114, USA*

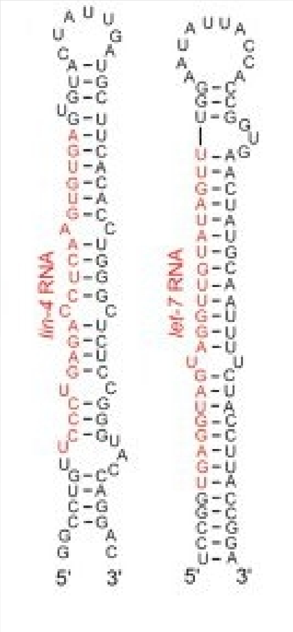
‡ Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut 06520, USA

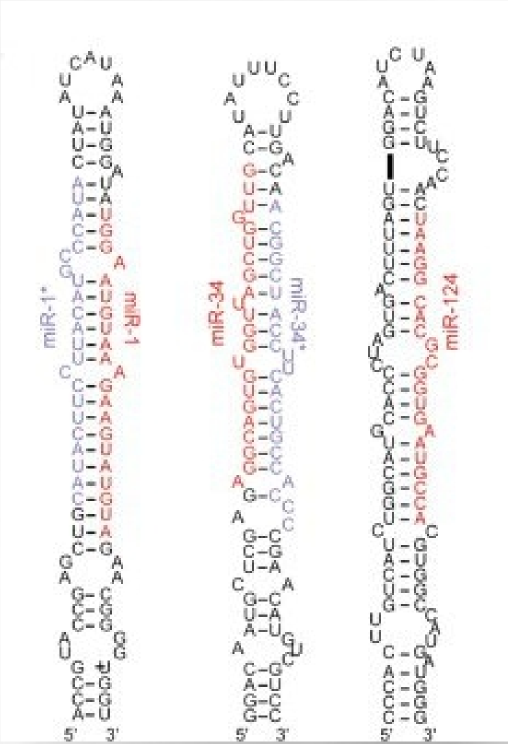
§ Kewalo Marine Lab, Pacific Biomedical Research Center, University of Hawaii, Honolulu, Hawaii 96813, USA

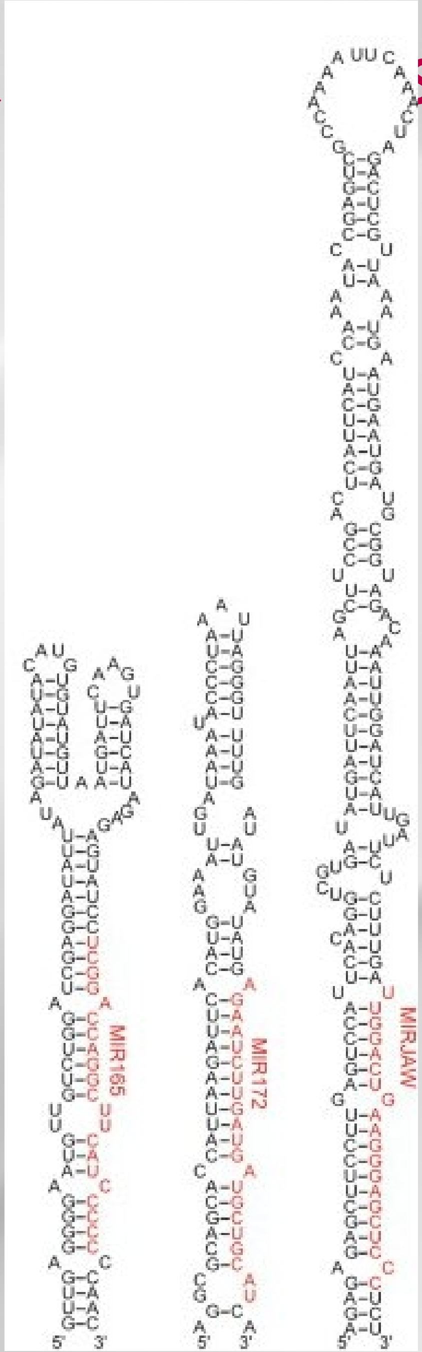
‖ Howard Hughes Medical Institute, Baylor College of Medicine, Houston, Texas 77030, USA

A. E. Pasquinelli et al., Nature 408, 86-9 (2000)

miRNA: pl s animal







Initial miRNA

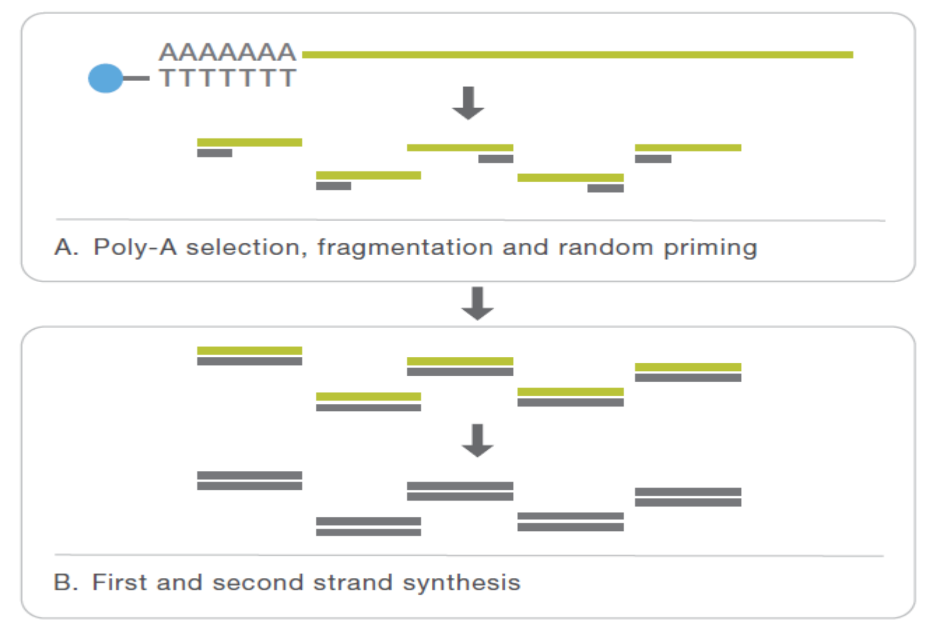
Animal miRNA

Plant miRNA

A. E. Pasquinelli et al., Nature 408, 86-9 (2000)

How can we study miRNA ?

- RNAseq not suited for miRNA (protocol and size)



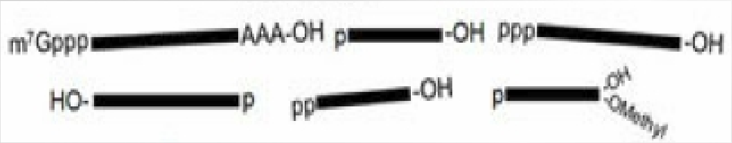
A. Poly-A selection, fragmentation and random priming

B. First and second strand synthesis

- small RNAseq: ability of high throughput sequencing to
    - Interrogate known and new small RNAs
    - Quantify them
    - Profile them on a large number of samples
    - Cost-effective

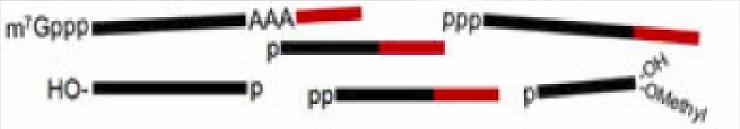
small RNA-Seq library preparation

- Monophosphate presence in 5' extremity and OH presence in 3' extremity



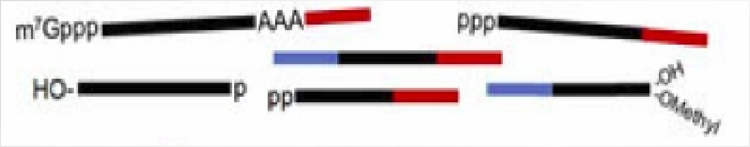
Total RNA: contain all kinds of RNA species including miRNA, mRNA, tRNA, rRNA...

Ligate with 3' adapter



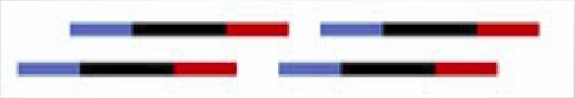
RNA with modified 3'-end will not ligate with 3' adapters. Only RNA with OH in 3'-end will ligate.

Ligate with 5' adapter



Only RNA with monophosphate in 5'-end will ligate with 5' adapters.

RT-PCR and Size Selection



MicroRNA sequencing library

CDNA containing both adapter sequences will be amplified. MicroRNA will be enriched from PCR and gel size selection.

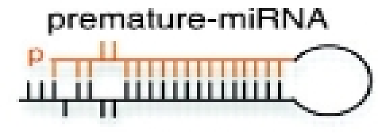
What are we looking for ?

- List of known miRNA
- List of new miRNA
- miRNA target(s)
- miRNA quantification
- Differential expression

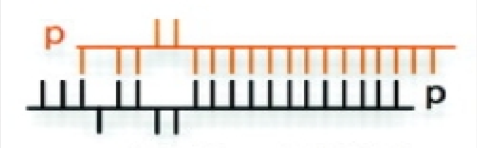
small RNAseq data analysis

What should we retain for data analysis ?

- Pre-miRNA information:



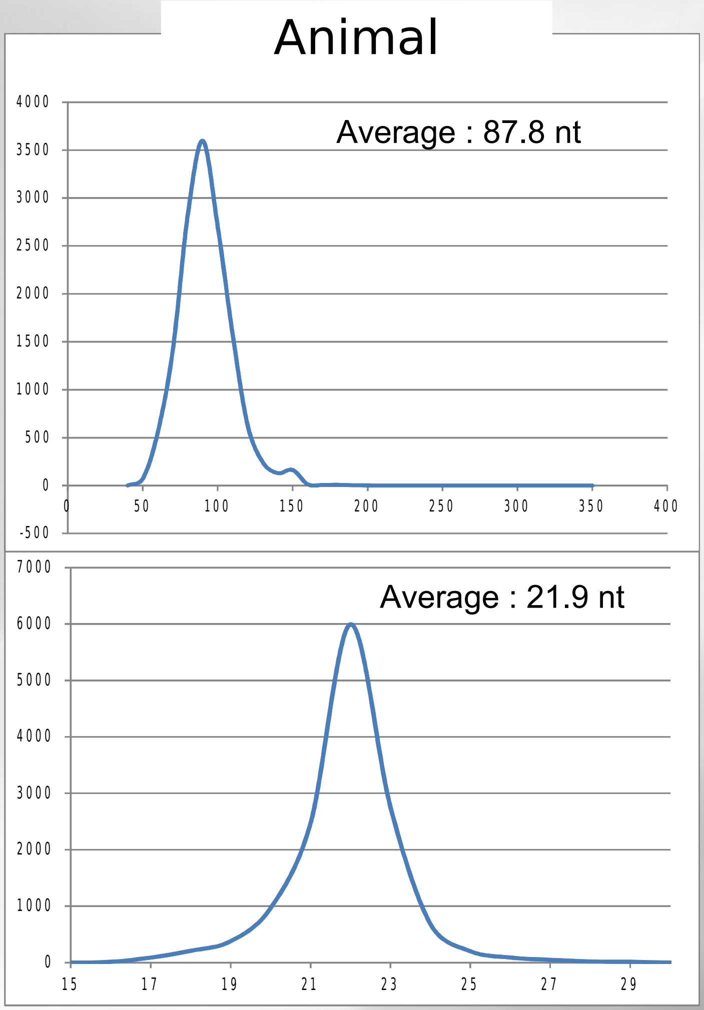
  - Hairpin structure of the pre-miRNA
  - Pre-miRNA localisation (coding/non coding TU intronic/exonic)
  - Presence of cluster
  - Size of the pre-miRNA
- miRNA-5p and miRNA-3p information:

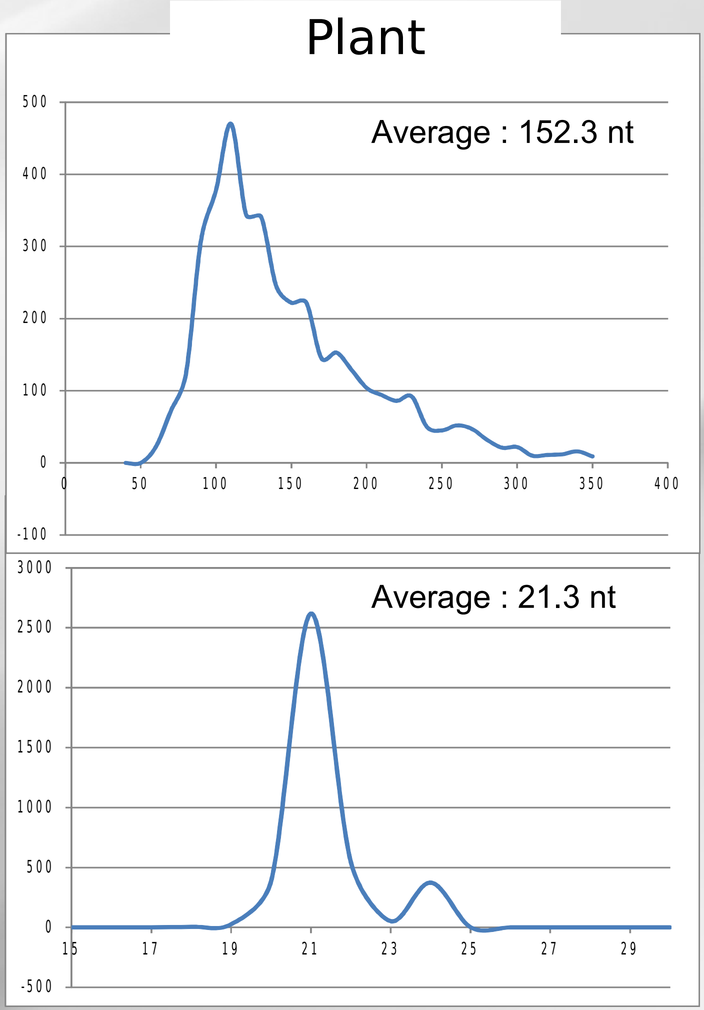


  - Existence of both miRNA-5p and miRNA-3p
  - Sequence conservation
  - Overhang (around 2 nt) related to drosha and Dicer cuts
  - Size of miRNA-5p and miRNA-3p
  - Overexpression of one of the miRNA-5p and miRNA-3p
- Existence of other products in sRNAseq data

What should we retain for data analysis ?

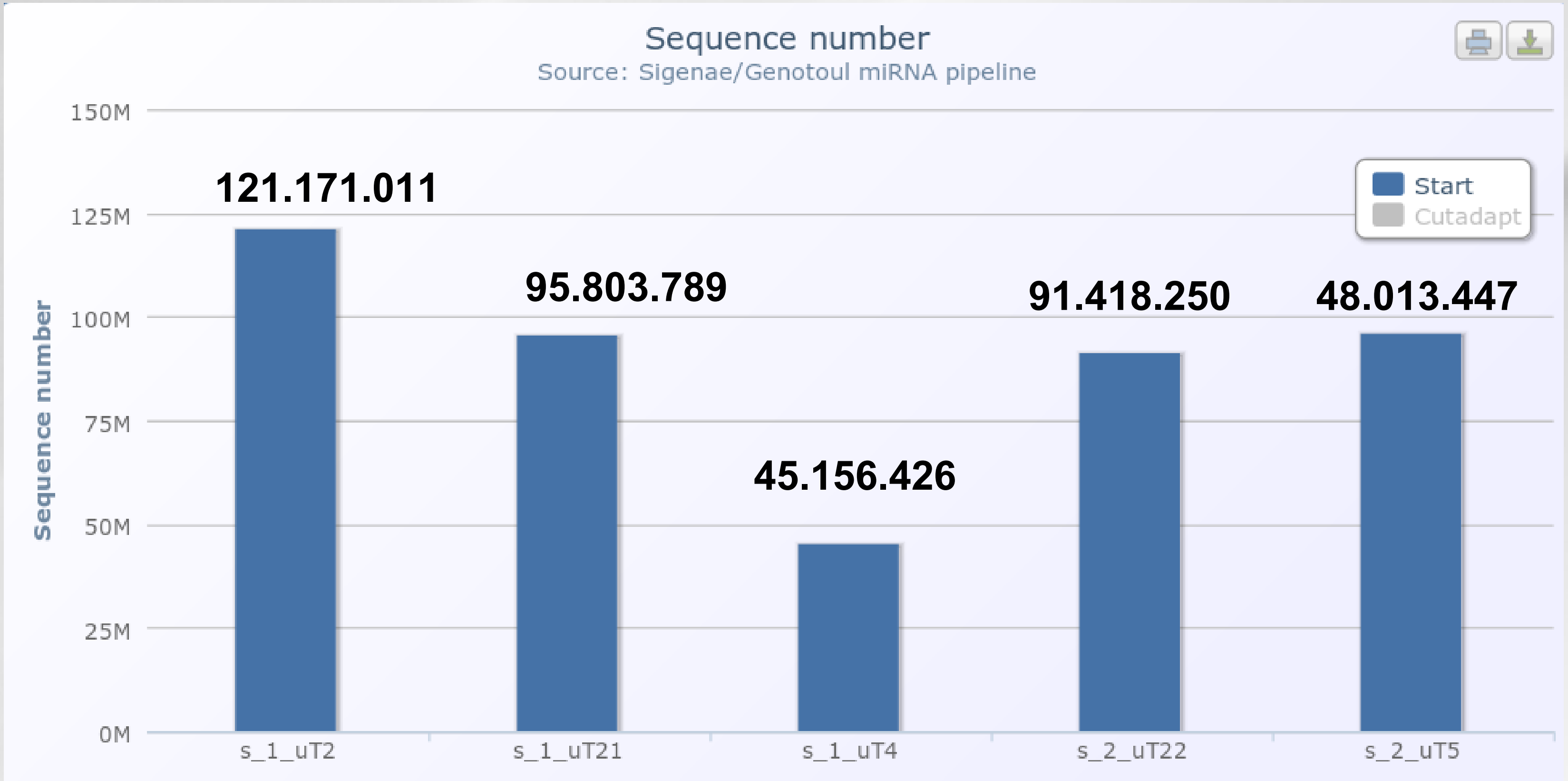
miRbase data on pre-miRNA / mature


Animal

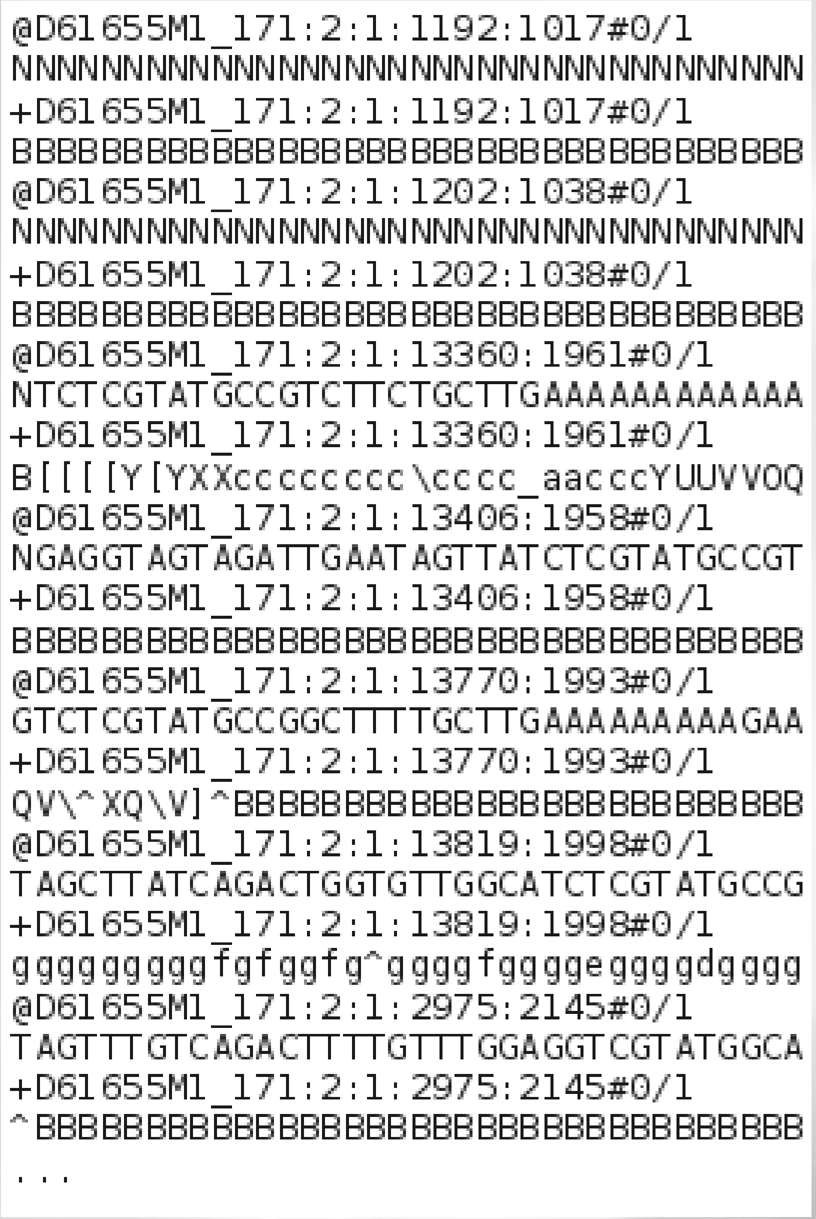

Plant

Description of the dataset

- 5 experiments (5 lanes, no multiplexing)
  - Different tissues, different stages
- No reference genome
  - Only scaffolds

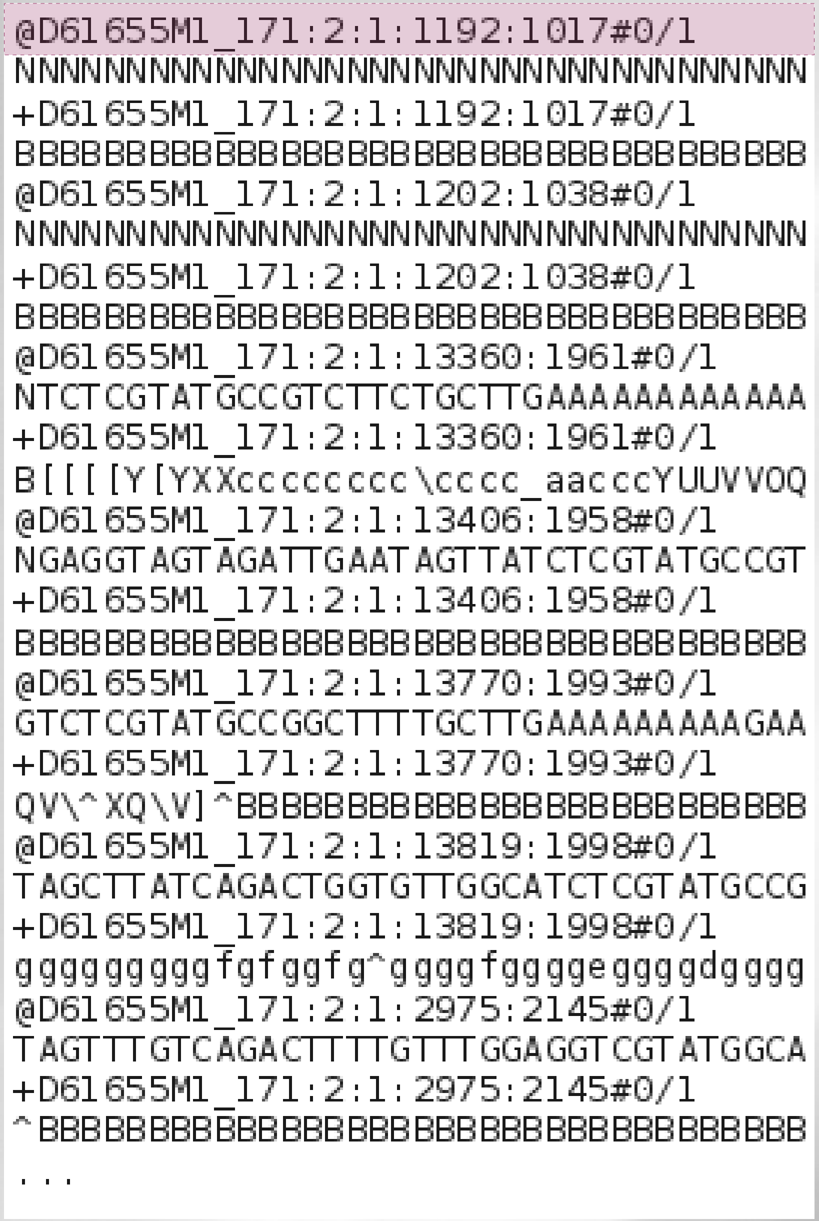


Fastq format



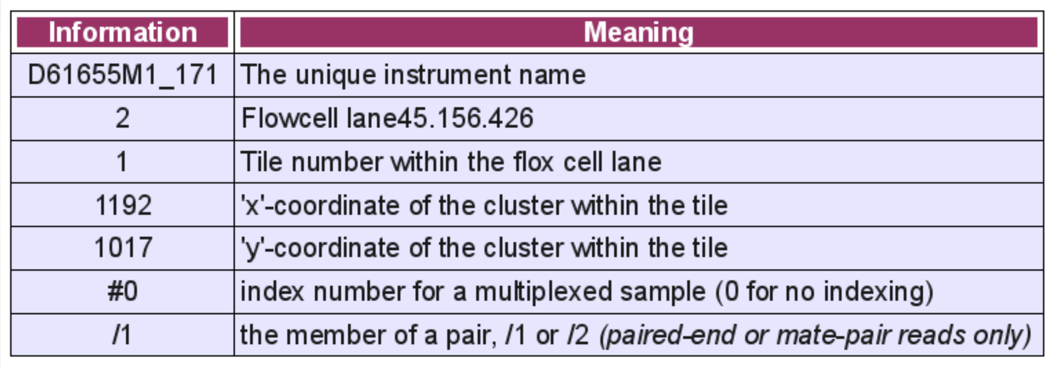
```
@D61655M1_171:2:1:1192:1017#0/1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+D61655M1_171:2:1:1192:1017#0/1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:1202:1038#0/1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+D61655M1_171:2:1:1202:1038#0/1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:13360:1961#0/1
NTCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAA
+D61655M1_171:2:1:13360:1961#0/1
B[[[[Y[YXXcccccccc\cccc_aacccYUUVVOQ
@D61655M1_171:2:1:13406:1958#0/1
NGAGGTAGTAGATTGAATAGTTATCTCGTATGCCGT
+D61655M1_171:2:1:13406:1958#0/1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:13770:1993#0/1
GTCTCGTATGCCGGCTTTTGCTTGAAAAAAAAAGAA
+D61655M1_171:2:1:13770:1993#0/1
QV\^XQ\V]^BBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:13819:1998#0/1
TAGCTTATCAGACTGGTGTTGGCATCTCGTATGCCG
+D61655M1_171:2:1:13819:1998#0/1
gggggggggfgfggfg^ggggfggggeggggdgggg
@D61655M1_171:2:1:2975:2145#0/1
TAGTTTGTCAGACTTTTGTTTGGAGGTCGTATGGCA
+D61655M1_171:2:1:2975:2145#0/1
^BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
...
```

Fastq format



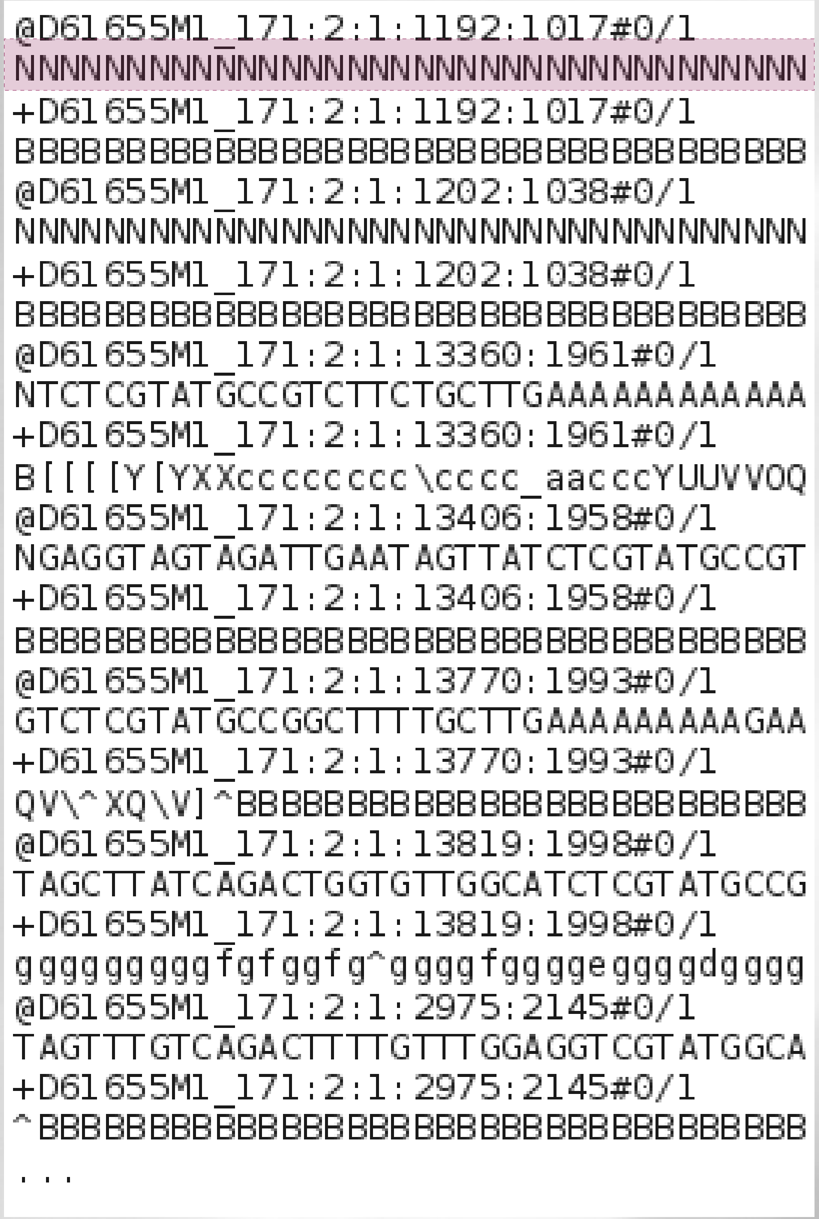
```
@D61655M1_171:2:1:1192:1017#0/1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+D61655M1_171:2:1:1192:1017#0/1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:1202:1038#0/1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+D61655M1_171:2:1:1202:1038#0/1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:13360:1961#0/1
NTCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAA
+D61655M1_171:2:1:13360:1961#0/1
B[[[[Y[YXXcccccccc\cccc_aacccYUUVVOQ
@D61655M1_171:2:1:13406:1958#0/1
NGAGGTAGTAGATTGAATAGTTATCTCGTATGCCGT
+D61655M1_171:2:1:13406:1958#0/1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:13770:1993#0/1
GTCTCGTATGCCGGCTTTTGCTTGAAAAAAAAAGAA
+D61655M1_171:2:1:13770:1993#0/1
QV\^XQ\V]^BBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:13819:1998#0/1
TAGCTTATCAGACTGGTGTTGGCATCTCGTATGCCG
+D61655M1_171:2:1:13819:1998#0/1
gggggggggfgfggfg^ggggfggggeggggdgggg
@D61655M1_171:2:1:2975:2145#0/1
TAGTTTGTCAGACTTTTGTTTGGAGGTCGTATGGCA
+D61655M1_171:2:1:2975:2145#0/1
^BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
...
```

Line 1 starts with @



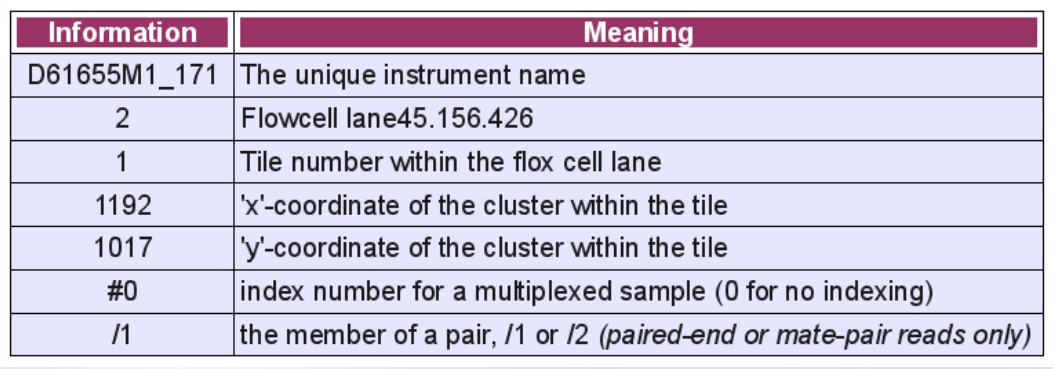
Information	Meaning
D61655M1_171	The unique instrument name
2	Flowcell lane45.156.426
1	Tile number within the flox cell lane
1192	'x'-coordinate of the cluster within the tile
1017	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*

Fastq format



```
@D61655M1_171:2:1:1192:1017#0/1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+D61655M1_171:2:1:1192:1017#0/1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:1202:1038#0/1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+D61655M1_171:2:1:1202:1038#0/1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:13360:1961#0/1
NTCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAA
+D61655M1_171:2:1:13360:1961#0/1
B[[[[Y[YXXcccccccc\cccc_aacccYUUVVOQ
@D61655M1_171:2:1:13406:1958#0/1
NGAGGTAGTAGATTGAATAGTTATCTCGTATGCCGT
+D61655M1_171:2:1:13406:1958#0/1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:13770:1993#0/1
GTCTCGTATGCCGGCTTTTGCTTGAAAAAAAAAGAA
+D61655M1_171:2:1:13770:1993#0/1
QV\^XQ\V]^BBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:13819:1998#0/1
TAGCTTATCAGACTGGTGTTGGCATCTCGTATGCCG
+D61655M1_171:2:1:13819:1998#0/1
gggggggggfgfggfg^ggggfggggeggggdgggg
@D61655M1_171:2:1:2975:2145#0/1
TAGTTTGTCAGACTTTTGTTTGGAGGTCGTATGGCA
+D61655M1_171:2:1:2975:2145#0/1
^BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
...
```

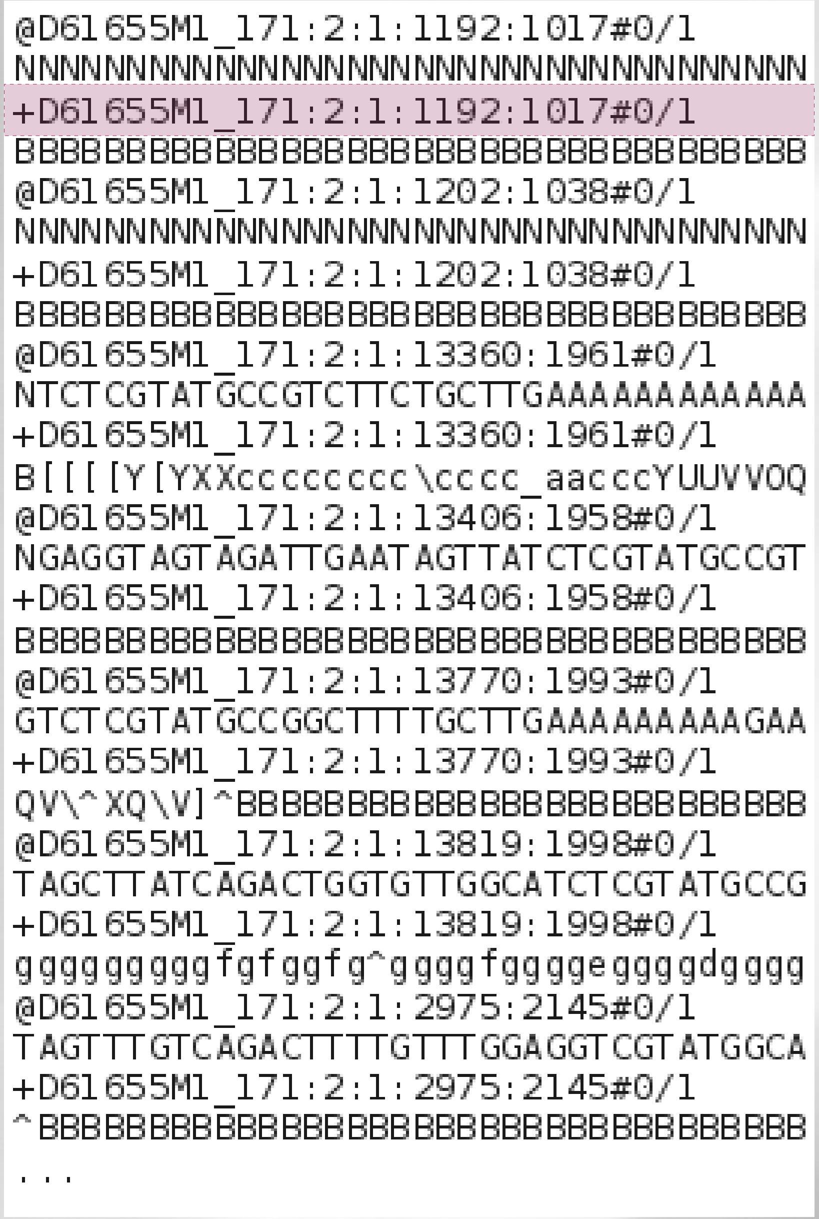
Line 1 starts with @



Information	Meaning
D61655M1_171	The unique instrument name
2	Flowcell lane45.156.426
1	Tile number within the flox cell lane
1192	'x'-coordinate of the cluster within the tile
1017	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*

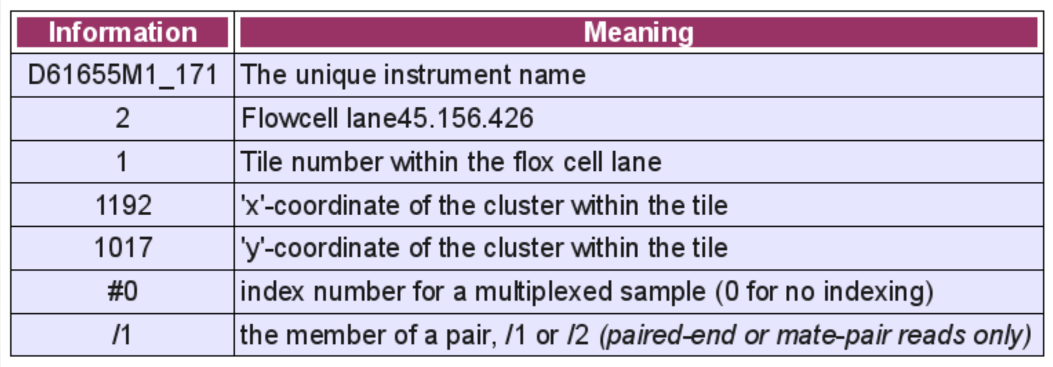
Line 2 Raw sequence of 36 nt (36 cycles in sequencing)

Fastq format



```
@D61655M1_171:2:1:1192:1017#0/1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+D61655M1_171:2:1:1192:1017#0/1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:1202:1038#0/1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+D61655M1_171:2:1:1202:1038#0/1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:13360:1961#0/1
NTCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAA
+D61655M1_171:2:1:13360:1961#0/1
B[[[[Y[YXXcccccccc\cccc_aacccYUUVVOQ
@D61655M1_171:2:1:13406:1958#0/1
NGAGGTAGTAGATTGAATAGTTATCTCGTATGCCGT
+D61655M1_171:2:1:13406:1958#0/1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:13770:1993#0/1
GTCTCGTATGCCGGCTTTTGCTTGAAAAAAAAAGAA
+D61655M1_171:2:1:13770:1993#0/1
QV\^XQ\V]^BBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:13819:1998#0/1
TAGCTTATCAGACTGGTGTTGGCATCTCGTATGCCG
+D61655M1_171:2:1:13819:1998#0/1
gggggggggfgfggfg^ggggfggggeggggdgggg
@D61655M1_171:2:1:2975:2145#0/1
TAGTTTGTCAGACTTTTGTTTGGAGGTCGTATGGCA
+D61655M1_171:2:1:2975:2145#0/1
^BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
...
```

Line 1 starts with @

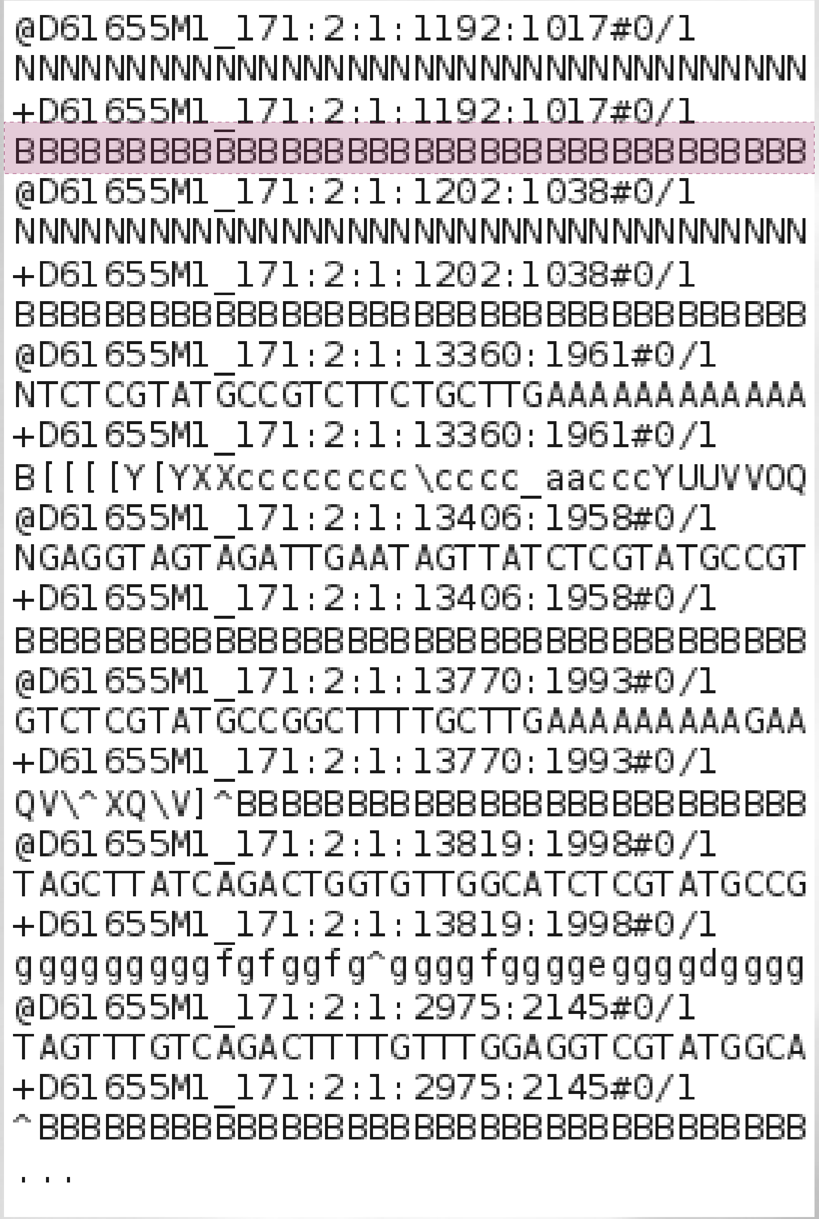


Information	Meaning
D61655M1_171	The unique instrument name
2	Flowcell lane45.156.426
1	Tile number within the flox cell lane
1192	'x'-coordinate of the cluster within the tile
1017	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*

Line 2 Raw sequence of 36 nt (36 cycles in sequencing)

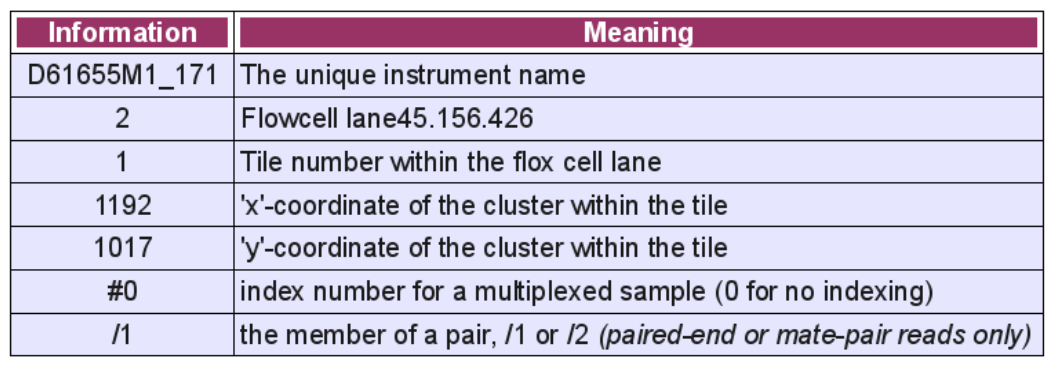
Line 3 starts with a '+' character and is optionally followed by the same sequence identifier (and any description) again.

Fastq format



```
@D61655M1_171:2:1:1192:1017#0/1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+D61655M1_171:2:1:1192:1017#0/1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:1202:1038#0/1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+D61655M1_171:2:1:1202:1038#0/1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:13360:1961#0/1
NTCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAA
+D61655M1_171:2:1:13360:1961#0/1
B[[[[Y[YXXcccccccc\cccc_aacccYUUVVOQ
@D61655M1_171:2:1:13406:1958#0/1
NGAGGTAGTAGATTGAATAGTTATCTCGTATGCCGT
+D61655M1_171:2:1:13406:1958#0/1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:13770:1993#0/1
GTCTCGTATGCCGGCTTTTGCTTGAAAAAAAAAGAA
+D61655M1_171:2:1:13770:1993#0/1
QV\^XQ\V]^BBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:13819:1998#0/1
TAGCTTATCAGACTGGTGTTGGCATCTCGTATGCCG
+D61655M1_171:2:1:13819:1998#0/1
gggggggggfgfggfg^ggggfggggeggggdgggg
@D61655M1_171:2:1:2975:2145#0/1
TAGTTTGTCAGACTTTTGTTTGGAGGTCGTATGGCA
+D61655M1_171:2:1:2975:2145#0/1
^BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
...
```

Line 1 starts with @



Information	Meaning
D61655M1_171	The unique instrument name
2	Flowcell lane45.156.426
1	Tile number within the flox cell lane
1192	'x'-coordinate of the cluster within the tile
1017	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*

Line 2 Raw sequence of 36 nt (36 cycles in sequencing)

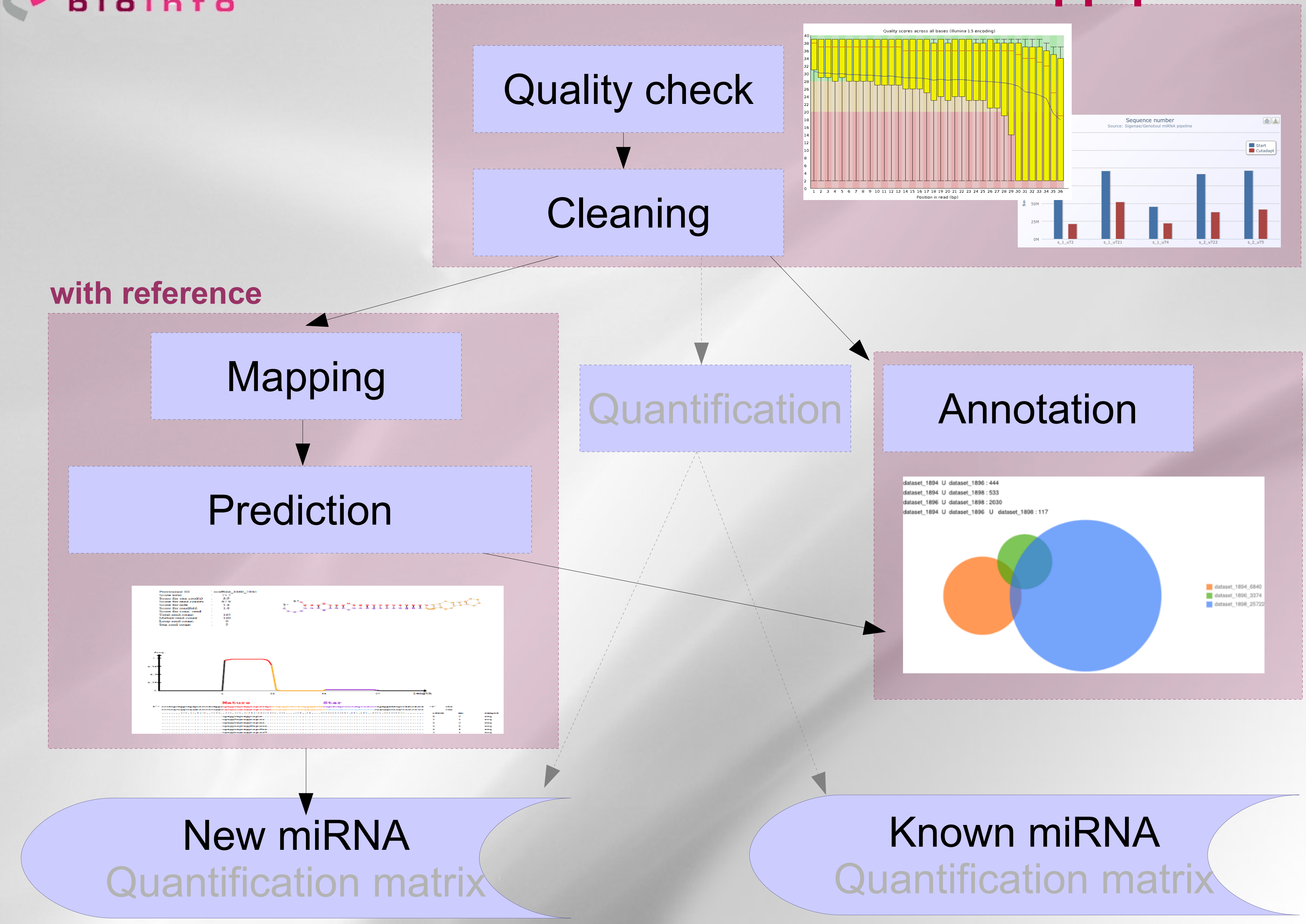
Line 3 starts with a '+' character and is optionally followed by the same sequence identifier (and any description) again.

Line 4 Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

Plateforme Bioinformatique Midi-Pyrénées

genotoul bioinfo

small RNAseq pipeline



Quality check

Cleaning

with reference

Mapping

Prediction

Quantification

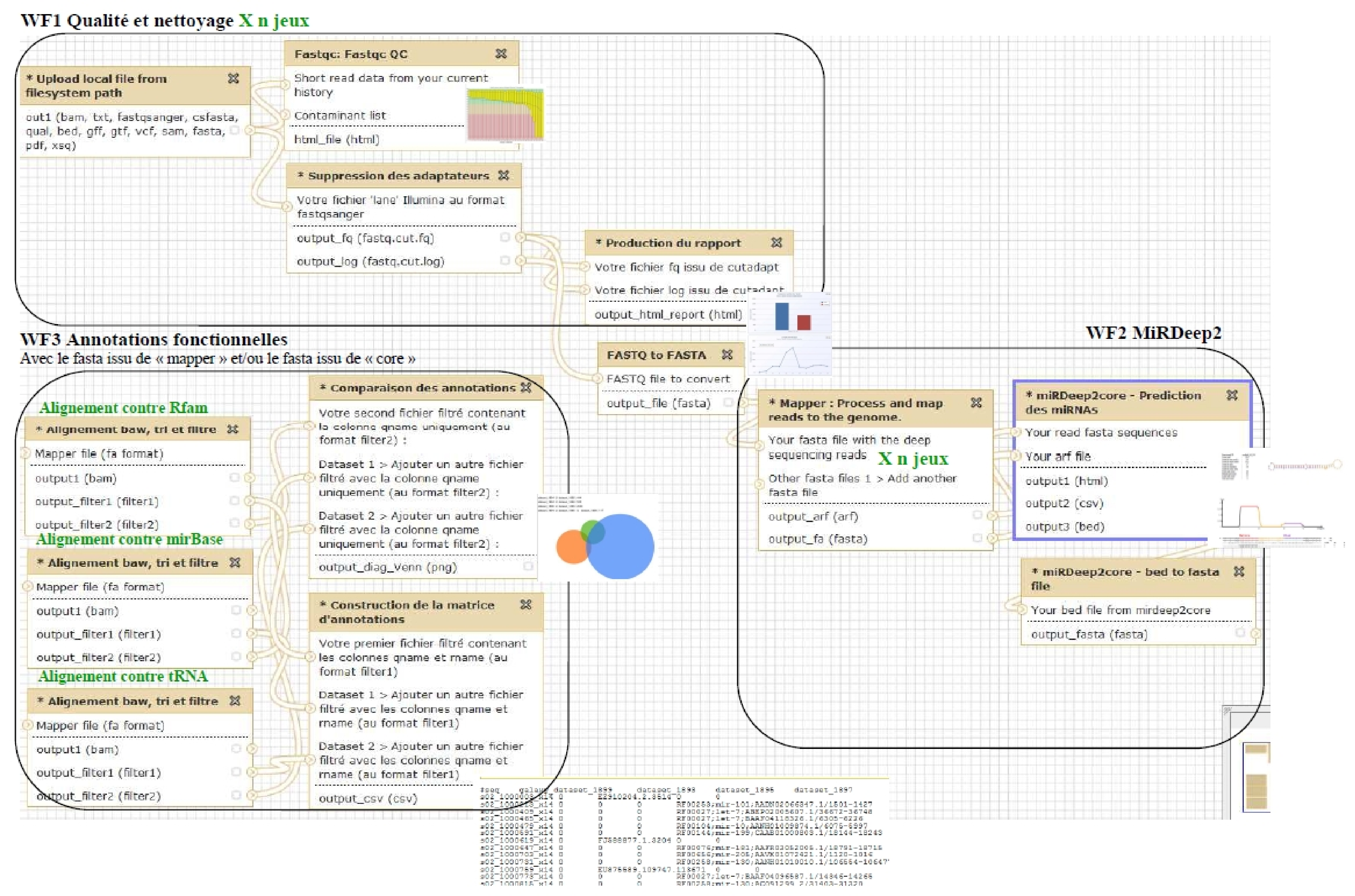
Annotation

dataset_1894 U dataset_1896 : 444
dataset_1894 U dataset_1898 : 533
dataset_1896 U dataset_1898 : 2030
dataset_1894 U dataset_1896 U dataset_1898 : 117

New miRNA
Quantification matrix

Known miRNA
Quantification matrix

small RNAseq pipeline



WF1 Qualité et nettoyage X n jeux

* Upload local file from filesystem path
out1 (bam, txt, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq)

Fastqc: Fastqc QC
Short read data from your current history
Contaminant list
html_file (html)

* Suppression des adaptateurs
Votre fichier 'lane' Illumina au format fastqsanger
output_fq (fastq.cut.fq)
output_log (fastq.cut.log)

* Production du rapport
Votre fichier fq issu de cutadapt
Votre fichier log issu de cutadapt
output_html_report (html)

FASTQ to FASTA
FASTQ file to convert
output_file (fasta)

WF2 MiRDeep2

* Mapper : Process and map reads to the genome.
Your fasta file with the deep sequencing reads X n jeux
Other fasta files 1 > Add another fasta file
output_arf (arf)
output_fa (fasta)

* miRDeep2core - Prediction des miRNAs
Your read fasta sequences
Your arf file
output1 (html)
output2 (csv)
output3 (bed)

* miRDeep2core - bed to fasta file
Your bed file from mirdeep2core
output_fasta (fasta)

WF3 Annotations fonctionnelles
Avec le fasta issu de « mapper » et/ou le fasta issu de « core »

Alignement contre Rfam

* Alignement baw, tri et filtre
Mapper file (fa format)
output1 (bam)
output_filter1 (filter1)
output_filter2 (filter2)

Alignement contre mirBase

* Alignement baw, tri et filtre
Mapper file (fa format)
output1 (bam)
output_filter1 (filter1)
output_filter2 (filter2)

Alignement contre tRNA

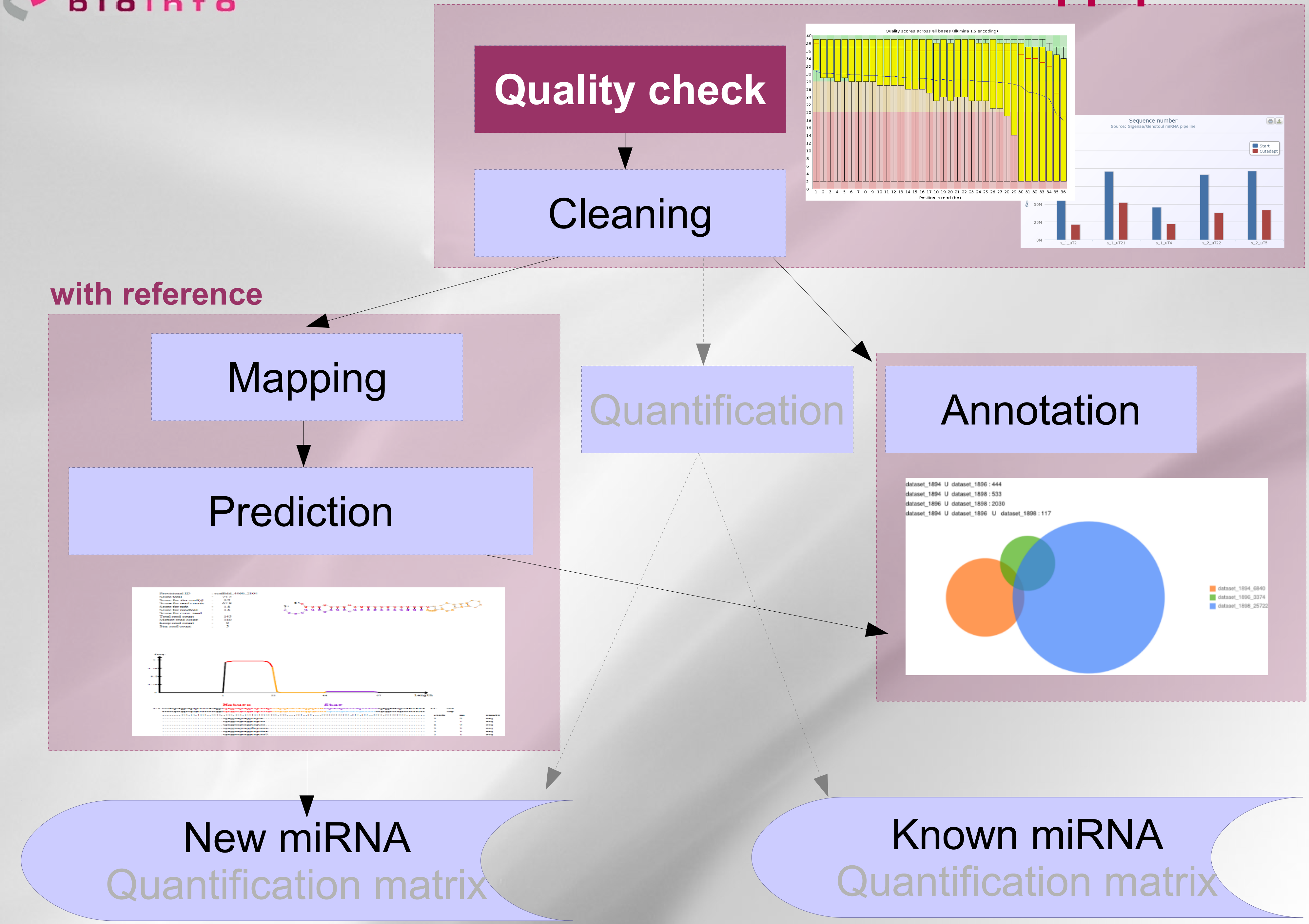
* Alignement baw, tri et filtre
Mapper file (fa format)
output1 (bam)
output_filter1 (filter1)
output_filter2 (filter2)

* Comparaison des annotations
Votre second fichier filtré contenant la colonne qname uniquement (au format filter2) :
Dataset 1 > Ajouter un autre fichier filtré avec la colonne qname uniquement (au format filter2) :
Dataset 2 > Ajouter un autre fichier filtré avec la colonne qname uniquement (au format filter2) :
output_diag_Venn (png)

* Construction de la matrice d'annotations
Votre premier fichier filtré contenant les colonnes qname et rname (au format filter1)
Dataset 1 > Ajouter un autre fichier filtré avec les colonnes qname et rname (au format filter1)
Dataset 2 > Ajouter un autre fichier filtré avec les colonnes qname et rname (au format filter1)
output_csv (csv)

```
#seq    galaxy dataset_1899     dataset_1898    dataset_1896    dataset_1897
s02_1000008_x14 0      EZ910204.2.3519 0       0
s02_1000323_x14 0      0       0       RF00258;mir-101;AADN02066847.1/1501-1427
s02_1000405_x14 0      0       0       RF00027;let-7;ABEP02005607.1/36672-36748
s02_1000465_x14 0      0       0       RF00027;let-7;BAAF04115326.1/6305-6226
s02_1000479_x14 0      0       0       RF00104;mir-10;AANH01009874.1/6075-5997
s02_1000691_x14 0      0       0       RF00144;mir-199;CAAB01000803.1/18144-18243
s02_1000619_x14 0      FJ388877.1.3204 0       0
s02_1000647_x14 0      0       0       RF00076;mir-181;AAFR03062005.1/18791-18716
s02_1000703_x14 0      0       0       RF00466;mir-205;AAVX01072421.1/1120-1016
s02_1000781_x14 0      0       0       RF00258;mir-130;AANH01010010.1/106564-10647'
s02_1000769_x14 0      EU876689.109747.113671  0       0
s02_1000773_x14 0      0       0       RF00027;let-7;BAAF04036587.1/14346-14266
s02_1000815_x14 0      0       0       RF00258;mir-130;AC091299.2/31403-31320
```

small RNAseq pipeline



Quality check

Cleaning

with reference

Mapping

Prediction

Quantification

Annotation

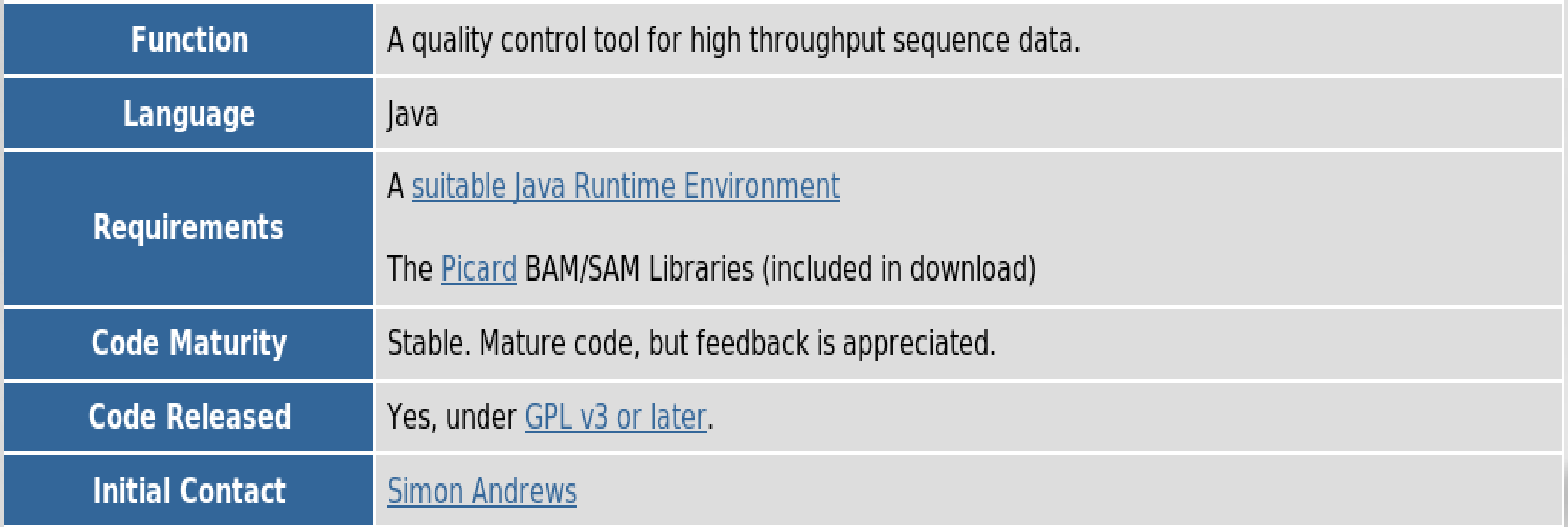
New miRNA
Quantification matrix

Known miRNA
Quantification matrix

Quality scores across all bases (Illumina 1.5 encoding)
Position in read (bp)
Sequence number
Source: Sigenae/Genotoul miRNA pipeline
Start
Cutadapt
50M
25M
0M
s_1_uT2
s_1_uT21
s_1_uT4
s_2_uT22
s_2_uT5
Mature
Star
dataset_1894 U dataset_1896 : 444
dataset_1894 U dataset_1898 : 533
dataset_1896 U dataset_1898 : 2030
dataset_1894 U dataset_1896 U dataset_1898 : 117
dataset_1894_6840
dataset_1896_3374
dataset_1898_25722

Quality control

- **FastQC** (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/)



Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later.
Initial Contact	Simon Andrews

A simple way to do quality control. It provides a modular set of analyses to give a quick impression of whether data has any problems of which you should be aware before doing any further analysis. The main functions of FastQC are:
- Import of data from BAM, SAM or FastQ files (any variant)
- Provide a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

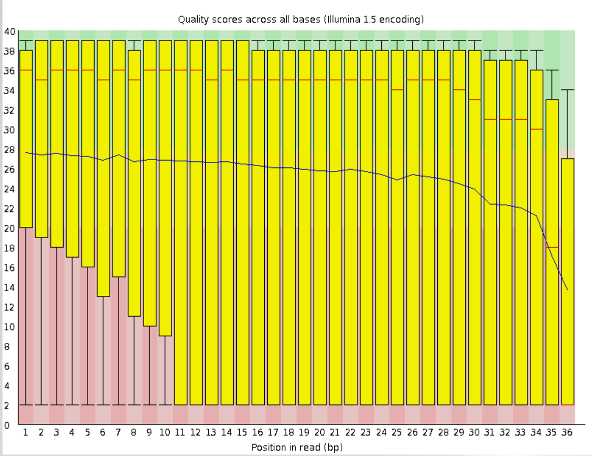
```
Fastqc -o nf.out nf_in.fastq
```

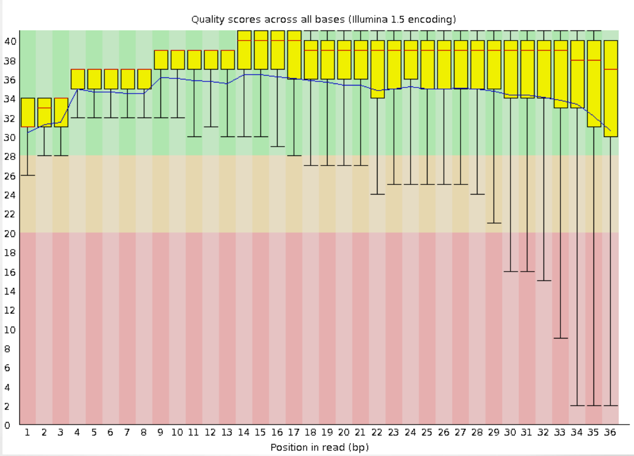
Plateforme Bioinformatique Midi-Pyrénées

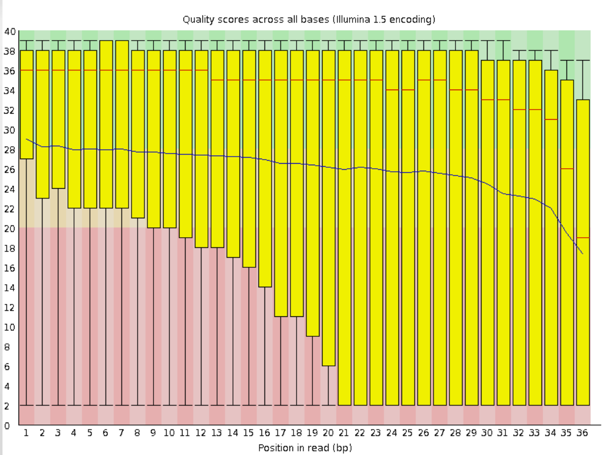
genotoul bioinfo

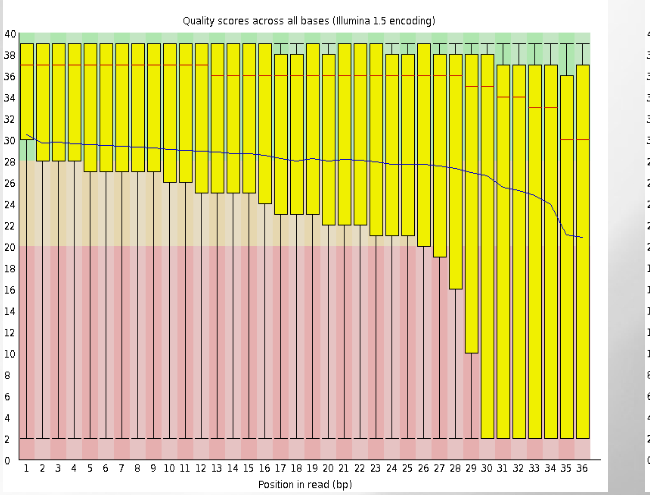
Quality control

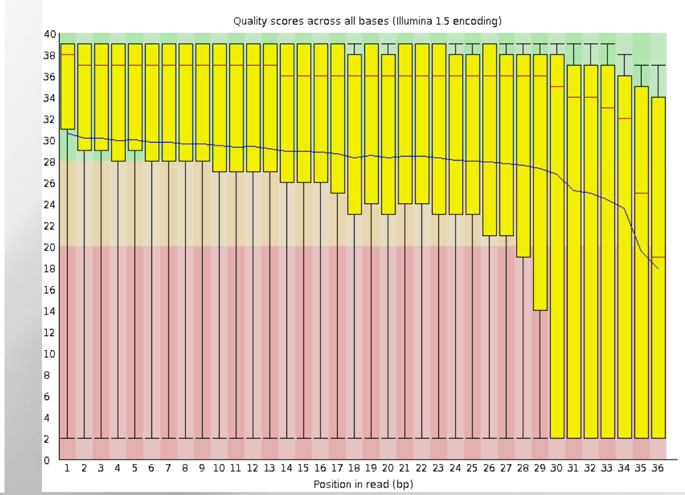
- **Per base quality**





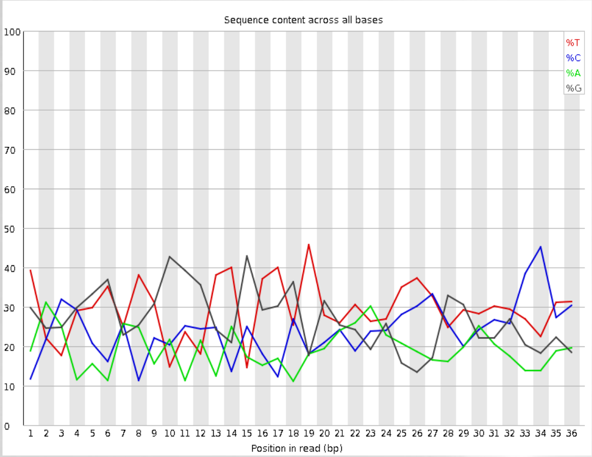


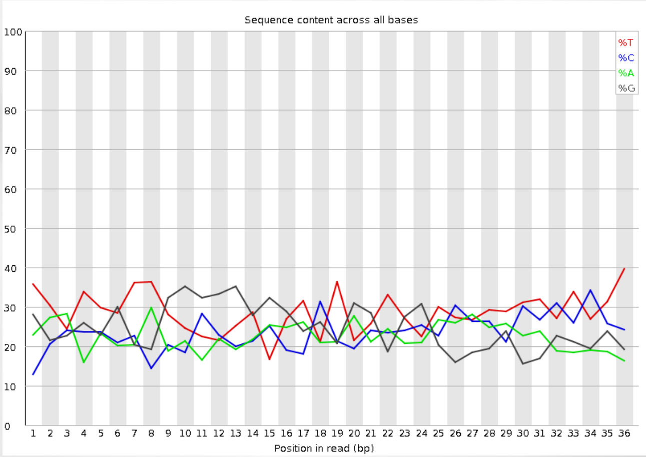


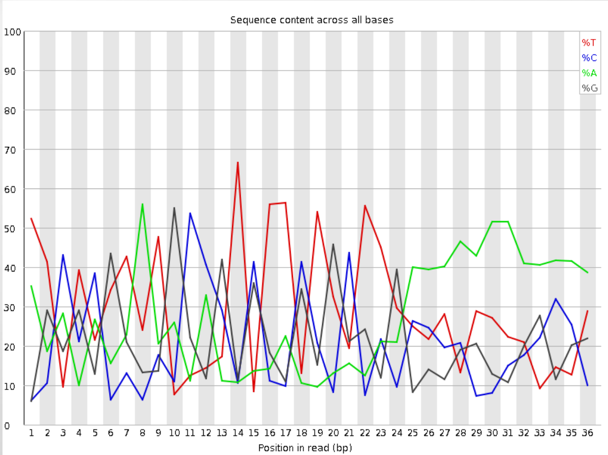


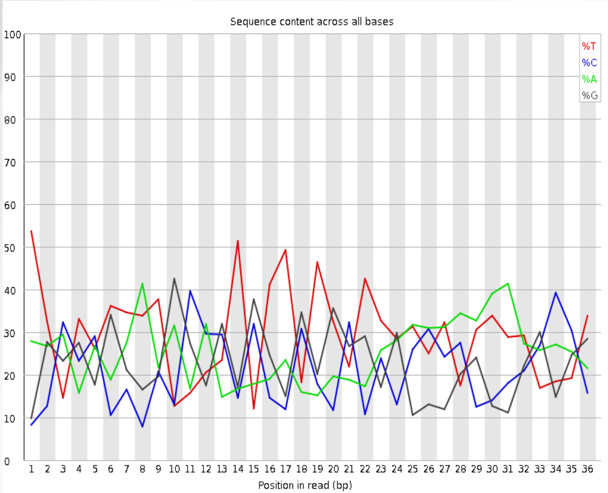
Quality control

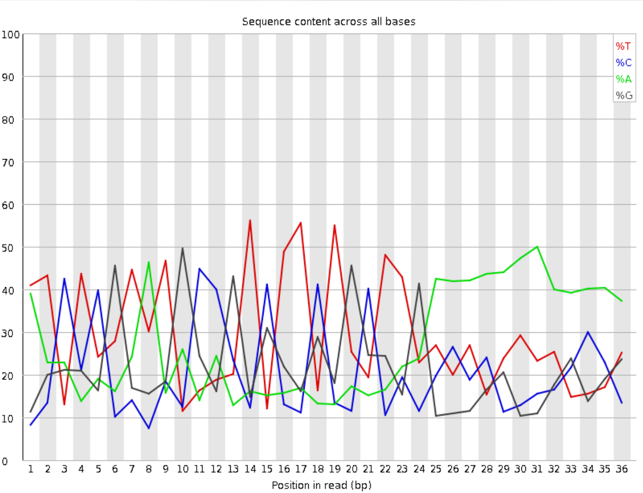
- **Sequences content in nucleotides**



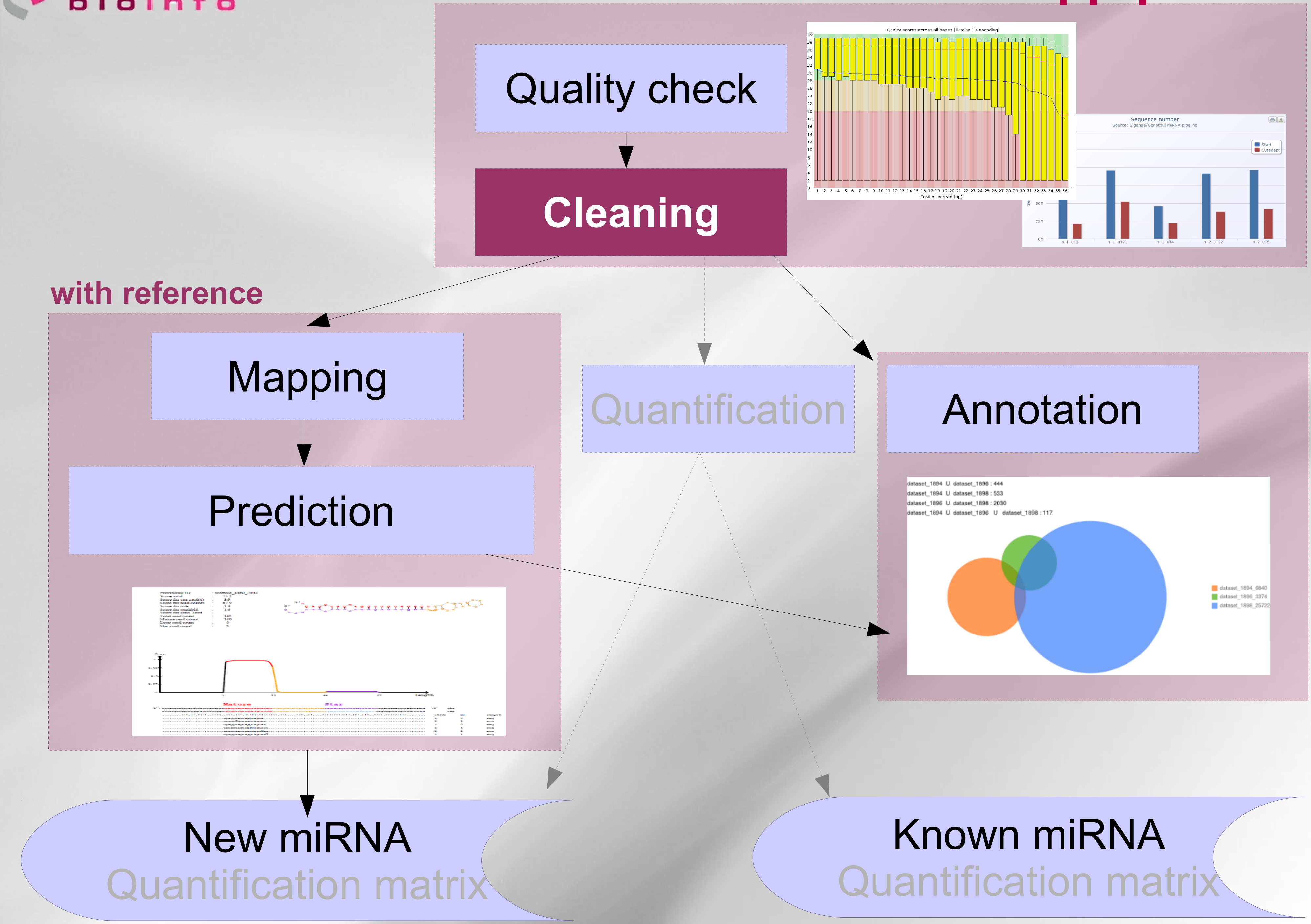








small RNAseq pipeline



Quality check

Cleaning

with reference

Mapping

Prediction

Quantification

Annotation

New miRNA
Quantification matrix

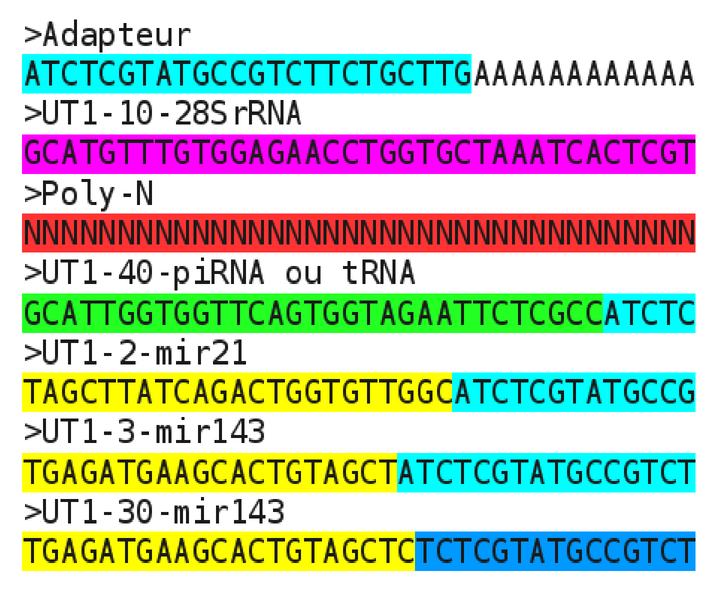
Known miRNA
Quantification matrix

Quality scores across all bases (Illumina 1.5 encoding)
Position in read (bp)
Sequence number
Source: Sigenae/Genotoul miRNA pipeline
Start
Cutadapt
50M
25M
0M
s_1_uT2
s_1_uT21
s_1_uT4
s_2_uT22
s_2_uT5
dataset_1894 U dataset_1896 : 444
dataset_1894 U dataset_1898 : 533
dataset_1896 U dataset_1898 : 2030
dataset_1894 U dataset_1896 U dataset_1898 : 117
dataset_1894_6840
dataset_1896_3374
dataset_1898_25722
Mature
Star

Plateforme Bioinformatique Midi-Pyrénées

Why cleaning ?

Output reads

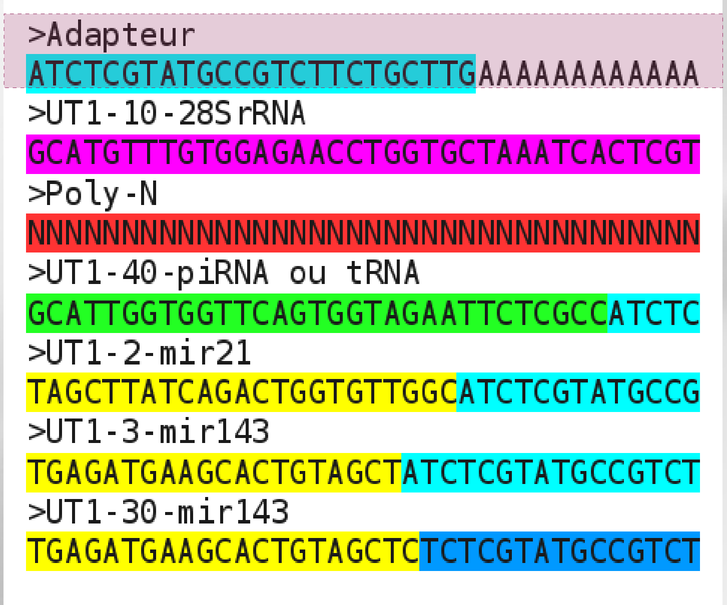


```
>Adapteur
ATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAA
>UT1-10-28SrRNA
GCATGTTTGTGGAGAACCTGGTGCTAAATCACTCGT
>Poly-N
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
>UT1-40-piRNA ou tRNA
GCATTGGTGGTTCAGTGGTAGAATTCTCGCCATCTC
>UT1-2-mir21
TAGCTTATCAGACTGGTGTTGGCATCTCGTATGCCG
>UT1-3-mir143
TGAGATGAAGCACTGTAGCTATCTCGTATGCCGTCT
>UT1-30-mir143
TGAGATGAAGCACTGTAGCTCTCTCGTATGCCGTCT
```

Why cleaning ?

Output reads

- Some sequences contain only adapters

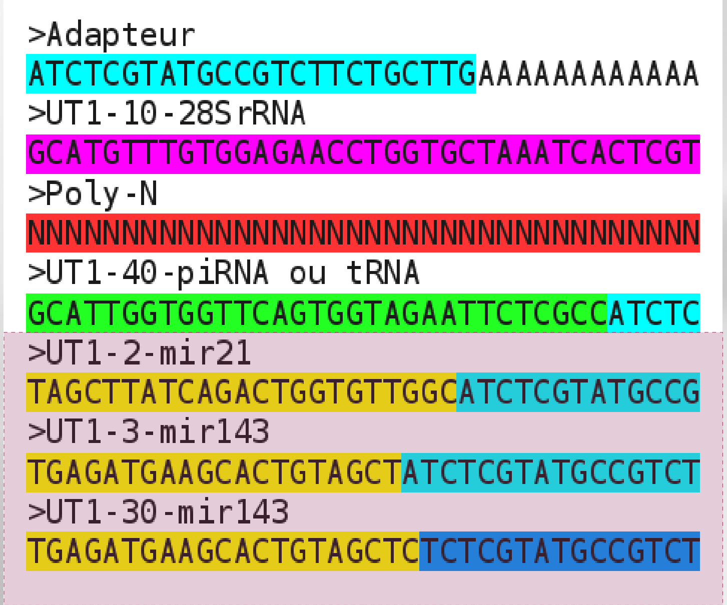


```
>Adapteur
ATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAA
>UT1-10-28SrRNA
GCATGTTTGTGGAGAACCTGGTGCTAAATCACTCGT
>Poly-N
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
>UT1-40-piRNA ou tRNA
GCATTGGTGGTTCAGTGGTAGAATTCTCGCCATCTC
>UT1-2-mir21
TAGCTTATCAGACTGGTGTTGGCATCTCGTATGCCG
>UT1-3-mir143
TGAGATGAAGCACTGTAGCTATCTCGTATGCCGTCT
>UT1-30-mir143
TGAGATGAAGCACTGTAGCTCTCTCGTATGCCGTCT
```

Why cleaning ?

Output reads

- Some sequences contain only adapters
- Some sequences contain sequences of interest flanked by the beginning of adapters:
    - Some of them are miRNA (yellow).

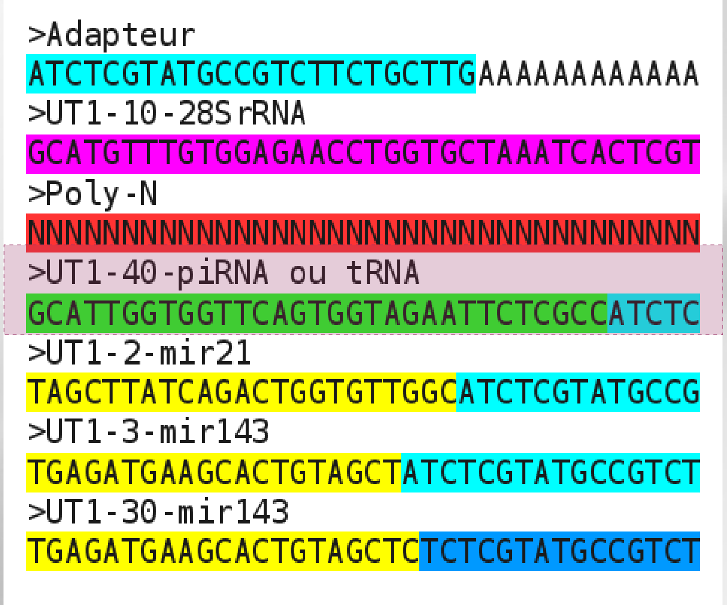


```
>Adapteur
ATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAA
>UT1-10-28SrRNA
GCATGTTTGTGGAGAACCTGGTGCTAAATCACTCGT
>Poly-N
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
>UT1-40-piRNA ou tRNA
GCATTGGTGGTTCAGTGGTAGAATTCTCGCCATCTC
>UT1-2-mir21
TAGCTTATCAGACTGGTGTTGGCATCTCGTATGCCG
>UT1-3-mir143
TGAGATGAAGCACTGTAGCTATCTCGTATGCCGTCT
>UT1-30-mir143
TGAGATGAAGCACTGTAGCTCTCTCGTATGCCGTCT
```

Why cleaning ?

Output reads

- Some sequences contain only adapters
- Some sequences contain sequences of interest flanked by the beginning of adapters:
    - Some of them are miRNA (yellow).
    - Some of them are other type of ncRNAs (green).

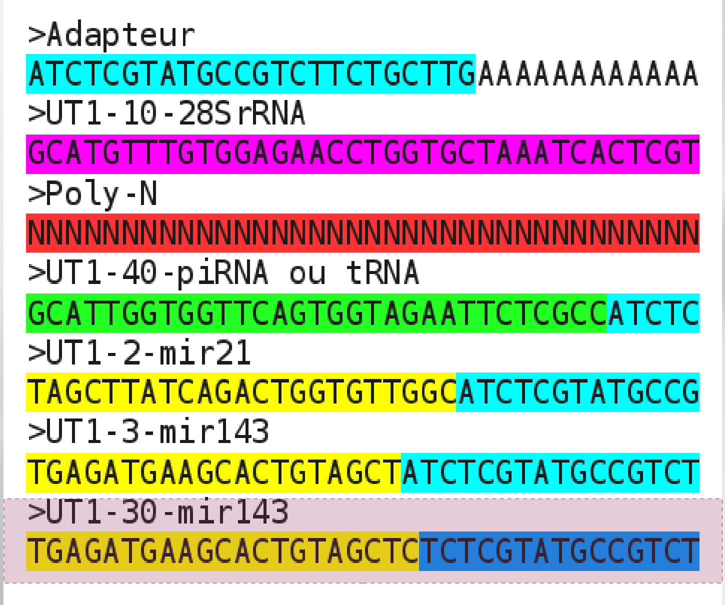


```
>Adapteur
ATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAA
>UT1-10-28S rRNA
GCATGTTTGTGGAGAACCTGGTGCTAAATCACTCGT
>Poly-N
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
>UT1-40-piRNA ou tRNA
GCATTGGTGGTTCAGTGGTAGAATTCTCGCCATCTC
>UT1-2-mir21
TAGCTTATCAGACTGGTGTTGGCATCTCGTATGCCG
>UT1-3-mir143
TGAGATGAAGCACTGTAGCTATCTCGTATGCCGTCT
>UT1-30-mir143
TGAGATGAAGCACTGTAGCTCTCTCGTATGCCGTCT
```

Why cleaning ?

Output reads

- Some sequences contain only adapters
- Some sequences contain sequences of interest flanked by the beginning of adapters:
    - Some of them are miRNA (yellow).
    - Some of them are other type of ncRNAs (green).
    - Some adapters contain errors (blue).

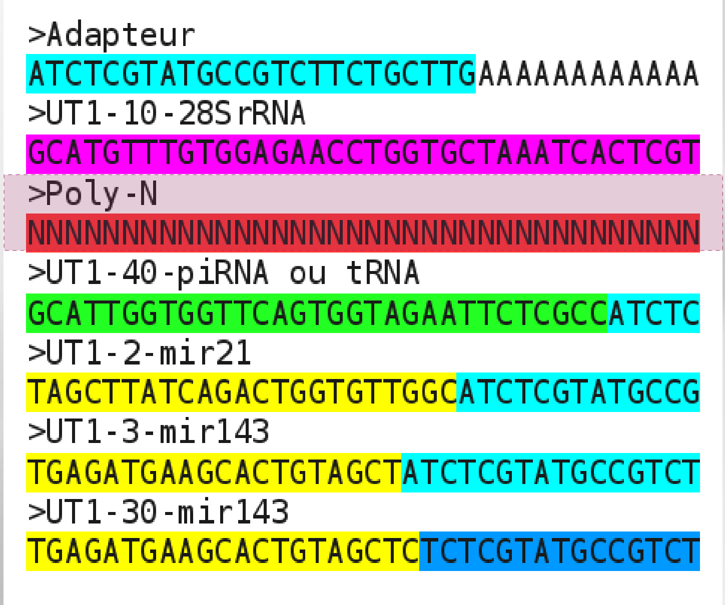


```
>Adapteur
ATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAA
>UT1-10-28SrRNA
GCATGTTTGTGGAGAACCTGGTGCTAAATCACTCGT
>Poly-N
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
>UT1-40-piRNA ou tRNA
GCATTGGTGGTTCAGTGGTAGAATTCTCGCCATCTC
>UT1-2-mir21
TAGCTTATCAGACTGGTGTTGGCATCTCGTATGCCG
>UT1-3-mir143
TGAGATGAAGCACTGTAGCTATCTCGTATGCCGTCT
>UT1-30-mir143
TGAGATGAAGCACTGTAGCTCTCTCGTATGCCGTCT
```

Why cleaning ?

Output reads

- Some sequences contain only adapters
- Some sequences contain sequences of interest flanked by the beginning of adapters:
    - Some of them are miRNA (yellow).
    - Some of them are other type of ncRNAs (green).
    - Some adapters contain errors (blue).
- Some sequences contain polyN (red)

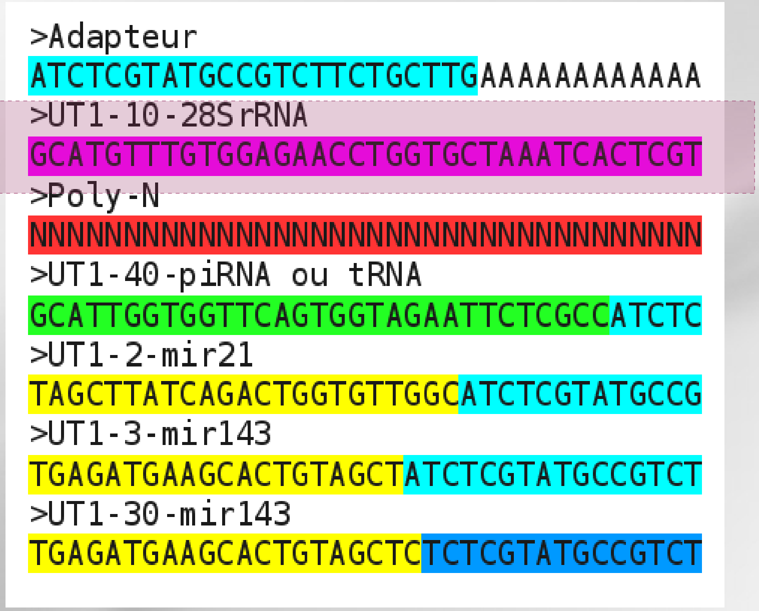


```
>Adapteur
ATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAA
>UT1-10-28S rRNA
GCATGTTTGTGGAGAACCTGGTGCTAAATCACTCGT
>Poly-N
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
>UT1-40-piRNA ou tRNA
GCATTGGTGGTTCAGTGGTAGAATTCTCGCCATCTC
>UT1-2-mir21
TAGCTTATCAGACTGGTGTTGGCATCTCGTATGCCG
>UT1-3-mir143
TGAGATGAAGCACTGTAGCTATCTCGTATGCCGTCT
>UT1-30-mir143
TGAGATGAAGCACTGTAGCTCTCTCGTATGCCGTCT
```

Why cleaning ?

Output reads

- Some sequences contain only adapters
- Some sequences contain sequences of interest flanked by the beginning of adapters:
    - Some of them are miRNA (yellow).
    - Some of them are other type of ncRNAs (green).
    - Some adapters contain errors (blue).
- Some sequences contain polyN (red)
- Some sequences contain other type of ncRNA (pink)



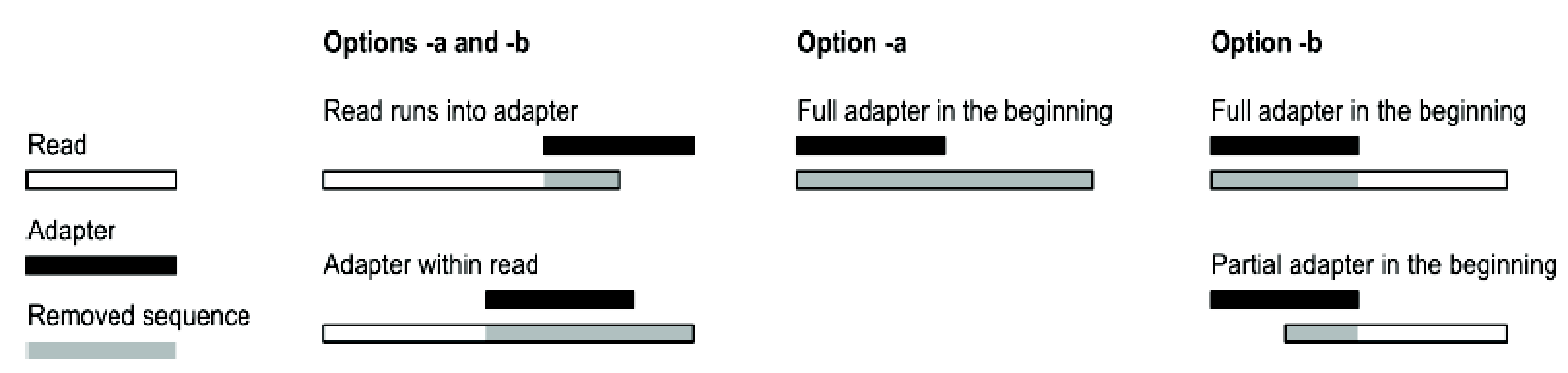
```
>Adapteur
ATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAA
>UT1-10-28S rRNA
GCATGTTTGTGGAGAACCTGGTGCTAAATCACTCGT
>Poly-N
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
>UT1-40-piRNA ou tRNA
GCATTGGTGGTTCAGTGGTAGAATTCTCGCCATCTC
>UT1-2-mir21
TAGCTTATCAGACTGGTGTTGGCATCTCGTATGCCG
>UT1-3-mir143
TGAGATGAAGCACTGTAGCTATCTCGTATGCCGTCT
>UT1-30-mir143
TGAGATGAAGCACTGTAGCTCTCTCGTATGCCGTCT
```

Cleaning

- **Adapters removing and length filtering**

Cutadapt http://code.google.com/p/cutadapt/.

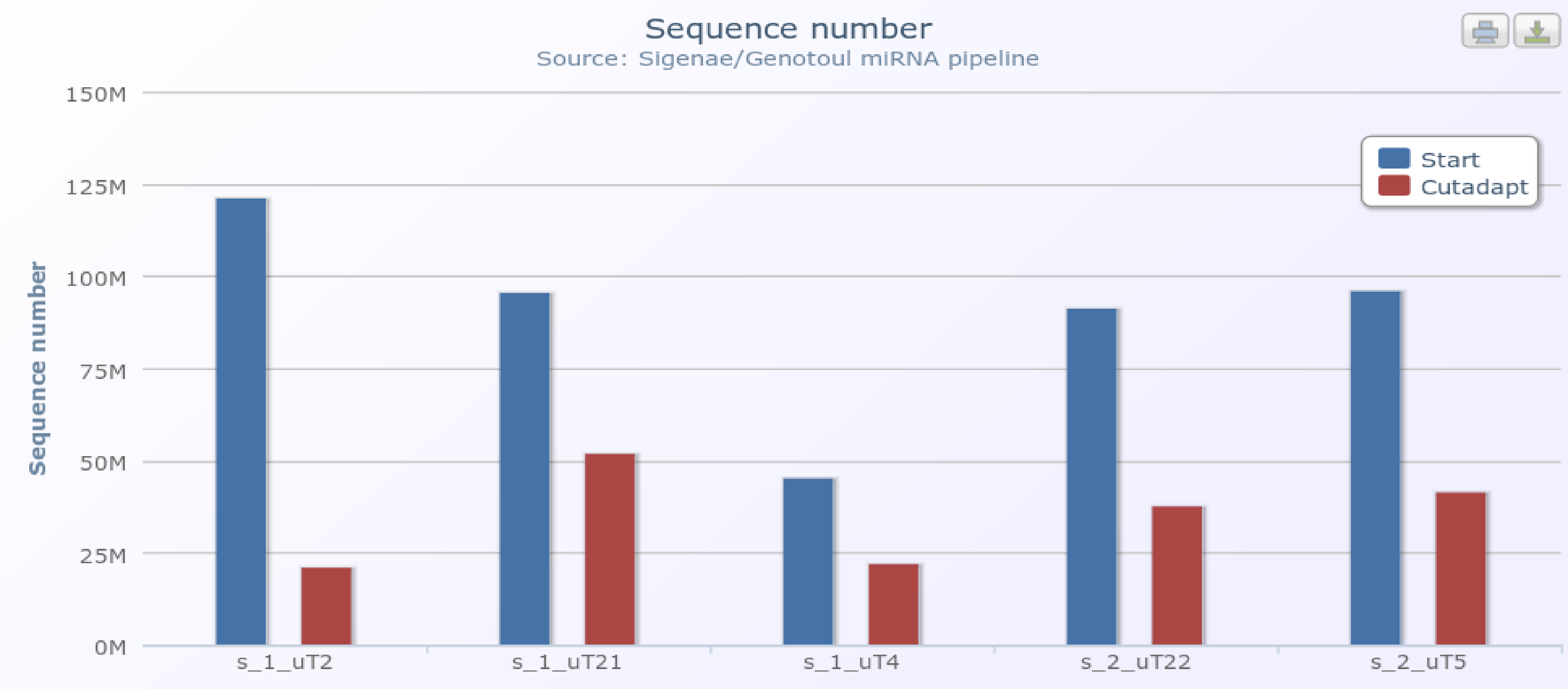
Cutadapt removes adapter sequences from high-throughput sequencing data. Indeed, reads are usually longer than the RNA, and therefore contain parts of the 3’ adapter. It also allows to keep only sequences of desired length (15<length<29).



```
cutadapt -a ATCTCGTATGCCGTCTTCTGCTTG -m -M 29 -o nf_out.fg nf_in.fq
```

Cleaning

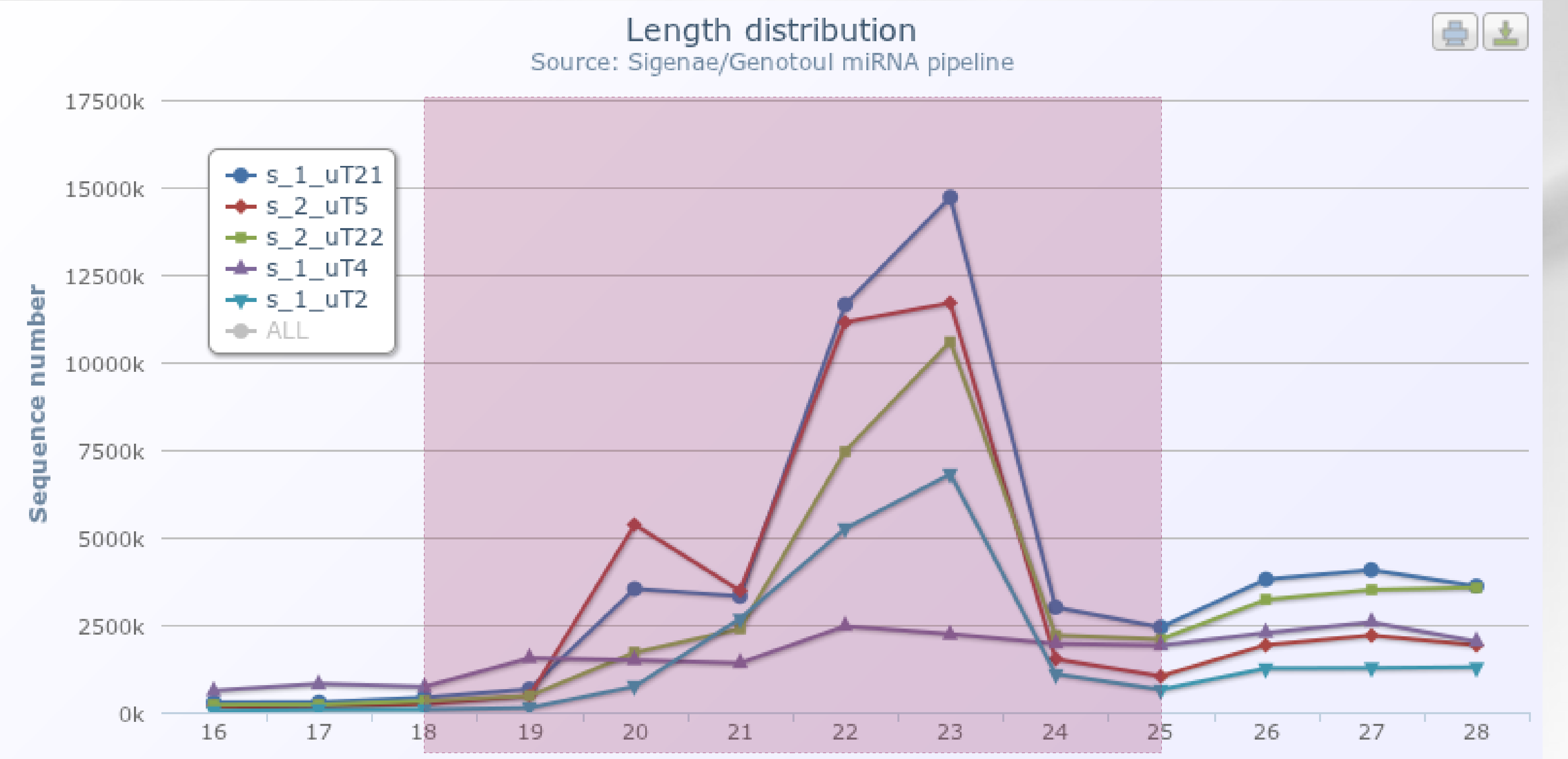
- **56 % of reads discarded**



Cleaning

- **Size in between 18bp:24bp**

  → miRNA ?



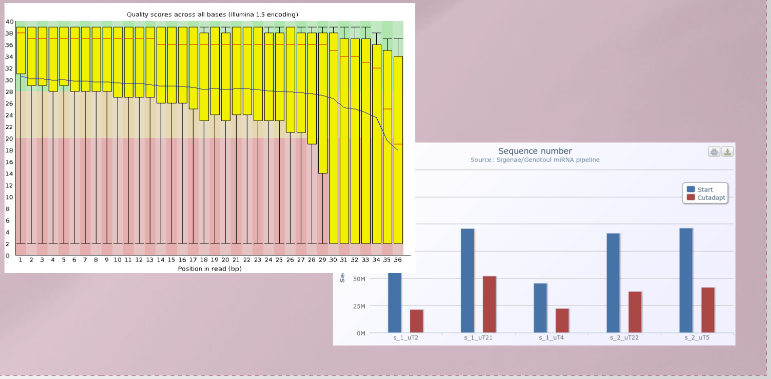
Exercices:

- WF1:
  - Quality control
  - Cleaning

small RNAseq pipeline

Quality check

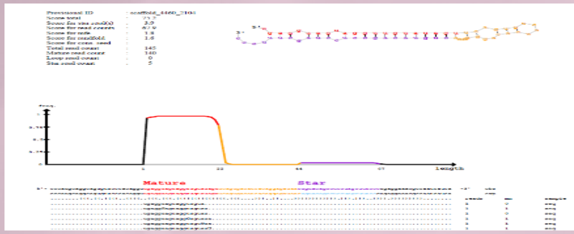
Cleaning



with reference

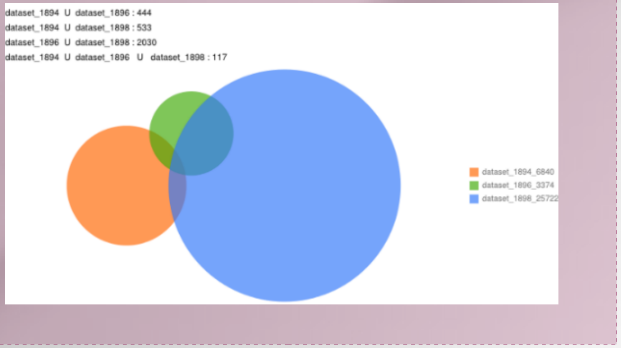
Mapping

Prediction



Quantification

Annotation



New miRNA
Quantification matrix

Known miRNA
Quantification matrix

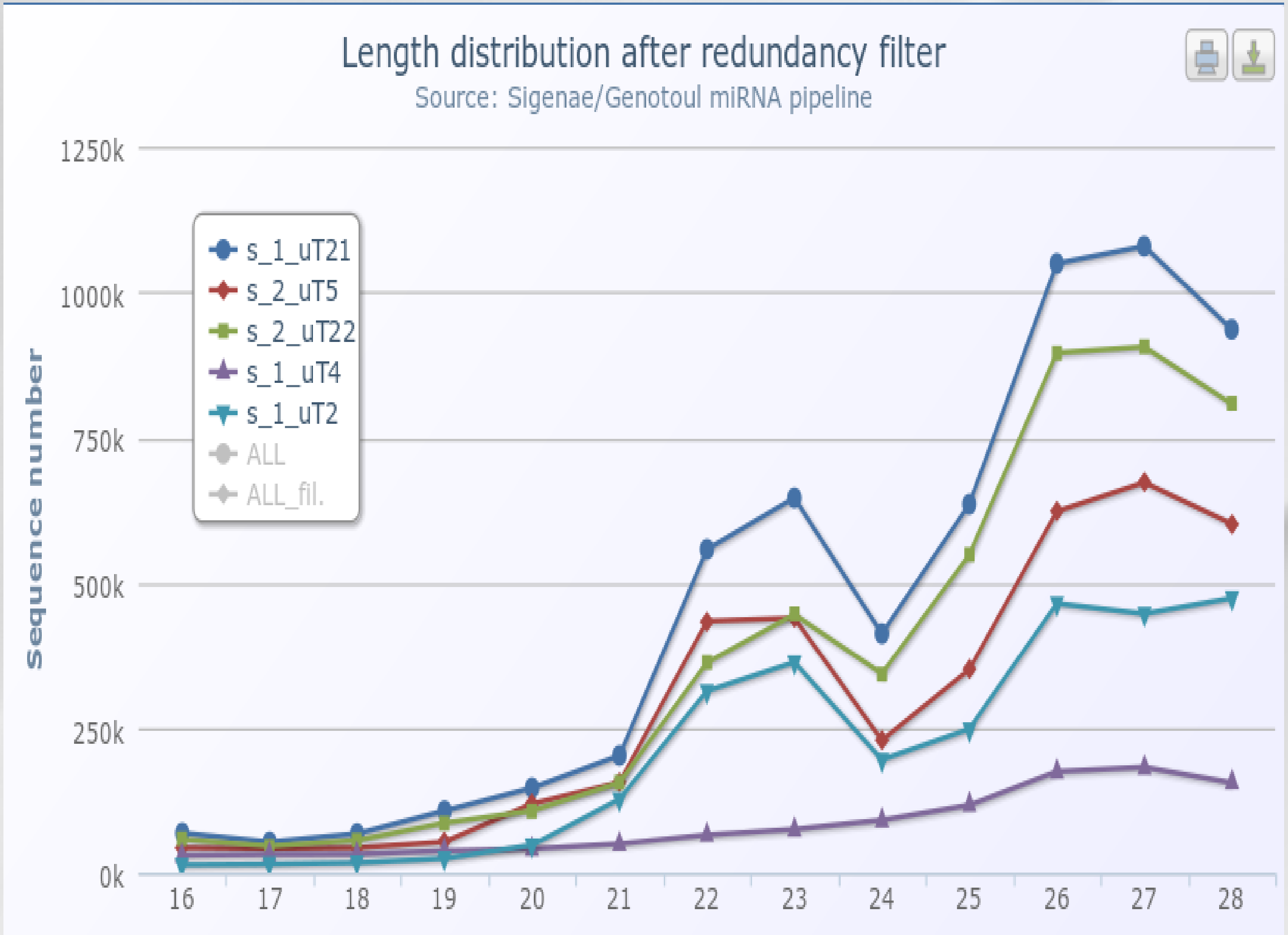
Mapping

Before mapping : Remove redundancy

- **Removing identical reads**
  - **save computational time**
  - **useless to keep all the read**
  - **Keep the number of occurrence for each read**

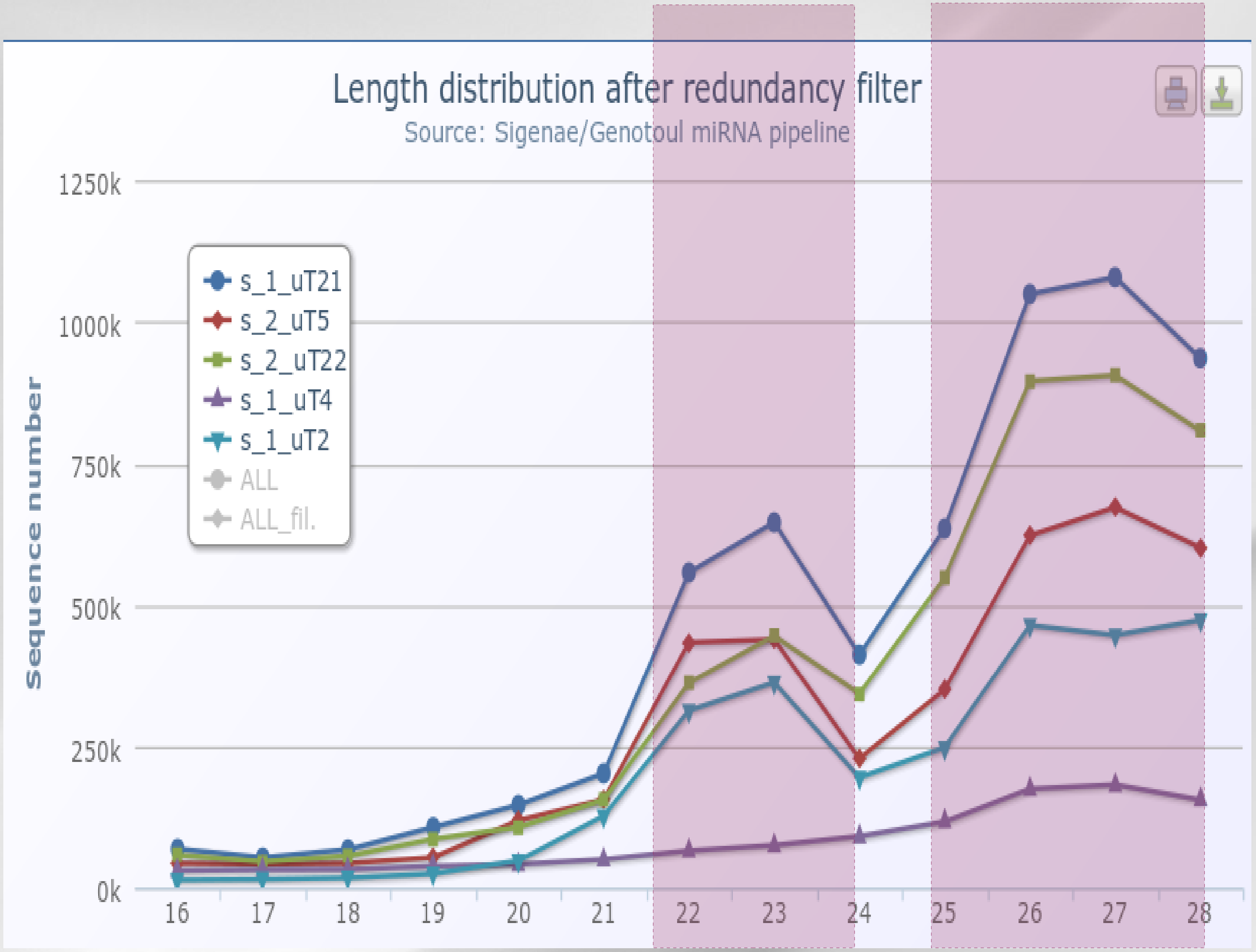
```
…
AAATGAATGATCTATGGACAGCA           2
AAATGAATGATCTATGGACAGCAG          38
AAATGAATGATCTATGGACAGCAGA         2
AAATGAATGATCTATGGACAGCAGAAAG      1
AAATGAATGATCTATGGACAGCAGC         51
AAATGAATGATCTATGGACAGCAGCA        82
AAATGAATGATCTATGGACAGCAGCAA       5
AAATGAATGATCTATGGACAGCAGCAAA      2
AAATGAATGATCTATGGACAGCAGCAAC      3
AAATGAATGATCTATGGACAGCAGCAAG      57
AAATGAATGATCTATGGACAGCAGCAG       2
AAATGAATGATCTATGGACAGCCGC         1
AAATGAATGATCTATGGACGGCAGCA        1
...
```

Remove redundancy



Remove redundancy

miRNA ? piRNA ?



- **More differencies between piRNAs than with miRNAs ?**

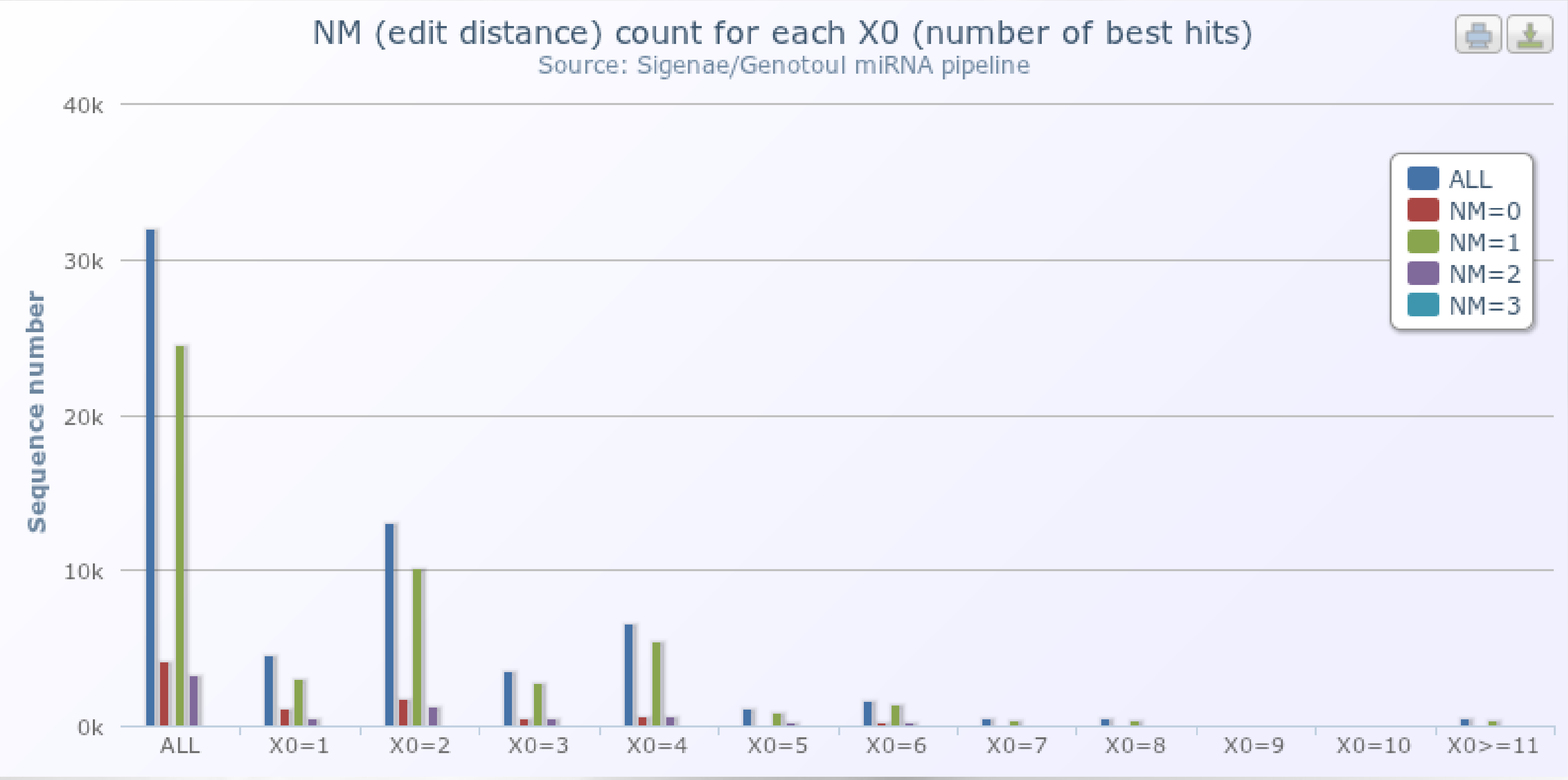
Plateforme Bioinformatique Midi-Pyrénées

Mapping the reads

- Blat http://genome.ucsc.edu/cgi-bin/hgBlat
- Blast http://blast.ncbi.nlm.nih.gov/Blast.cgi
- Gmap http://www.gene.com/share/gmap/
- **Bowtie http://bowtie-bio.sourceforge.net/index.shtml**
- **BWA http://bio-bwa.sourceforge.net**
- …

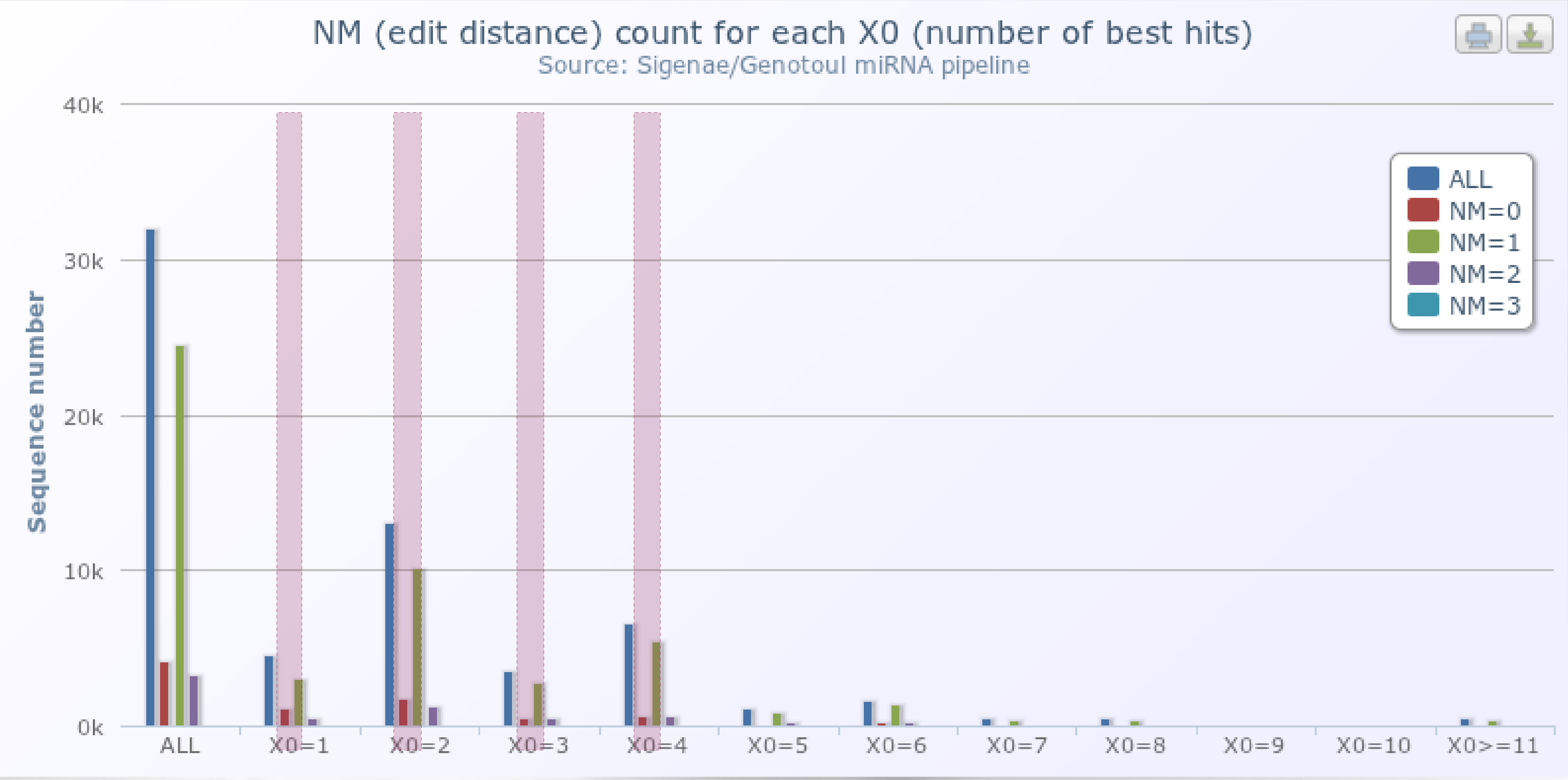
Mapping the reads

- **Alignement of annotated reads**



Mapping the reads

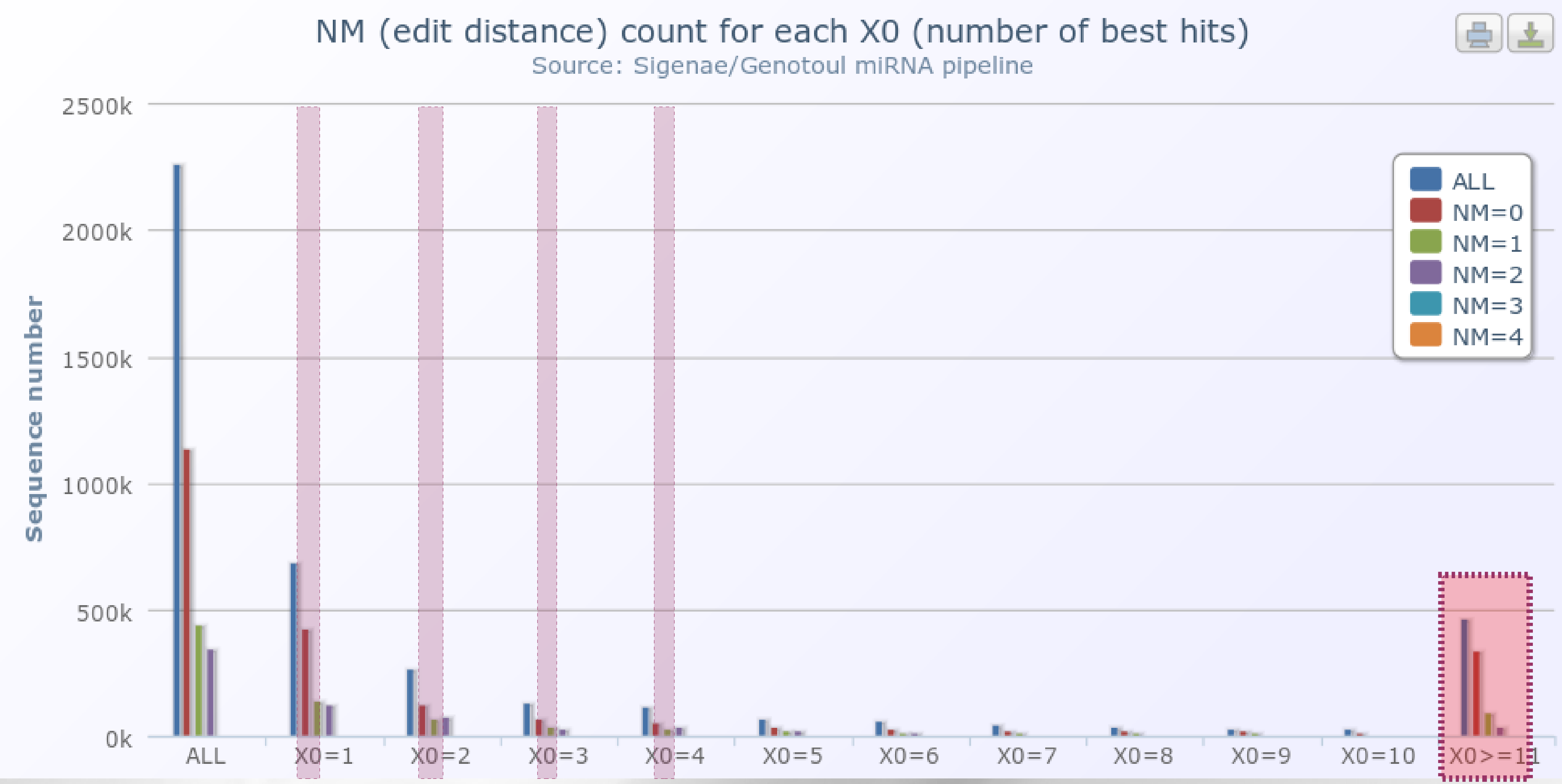
- **Alignement of annotated reads**



→ keep reads aligned the most at 4 positions with 0 or 1 error

Mapping the reads

- **Alignement of all reads**

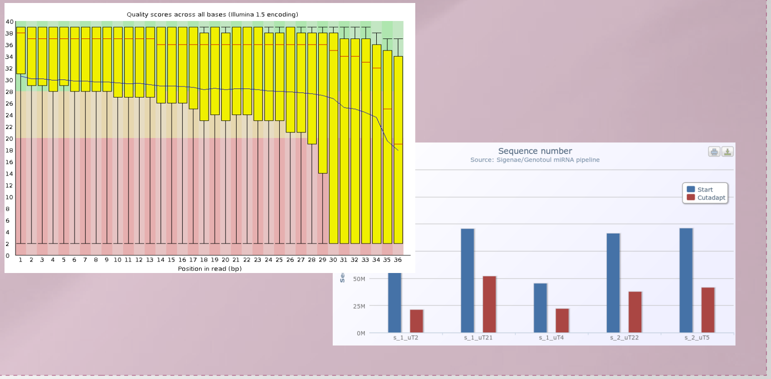


→ keep reads aligned the most at 4 positions with 0 or 1 error

small RNAseq pipeline

Quality check

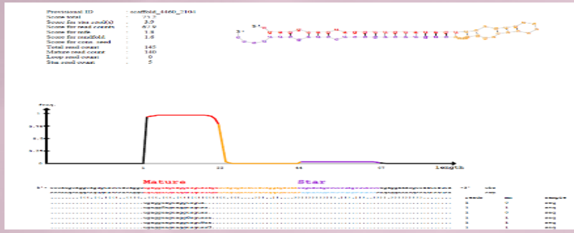
Cleaning



with reference

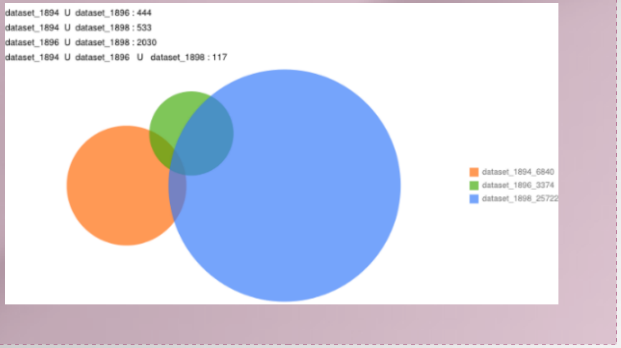
Mapping

Prediction



Quantification

Annotation


dataset_1894 U dataset_1896 : 444
dataset_1894 U dataset_1898 : 533
dataset_1896 U dataset_1898 : 2030
dataset_1894 U dataset_1896 U dataset_1898 : 117

New miRNA
Quantification matrix

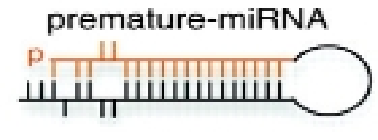
Known miRNA
Quantification matrix

Exercices:

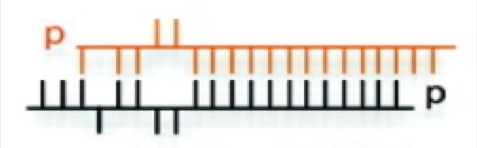
- Mapping the reads with miRDeep2
  - Using Bowtie for mapping
  - miRDeep2-core for miRNA identification

What should we retain for data analysis ?

- Pre-miRNA information:



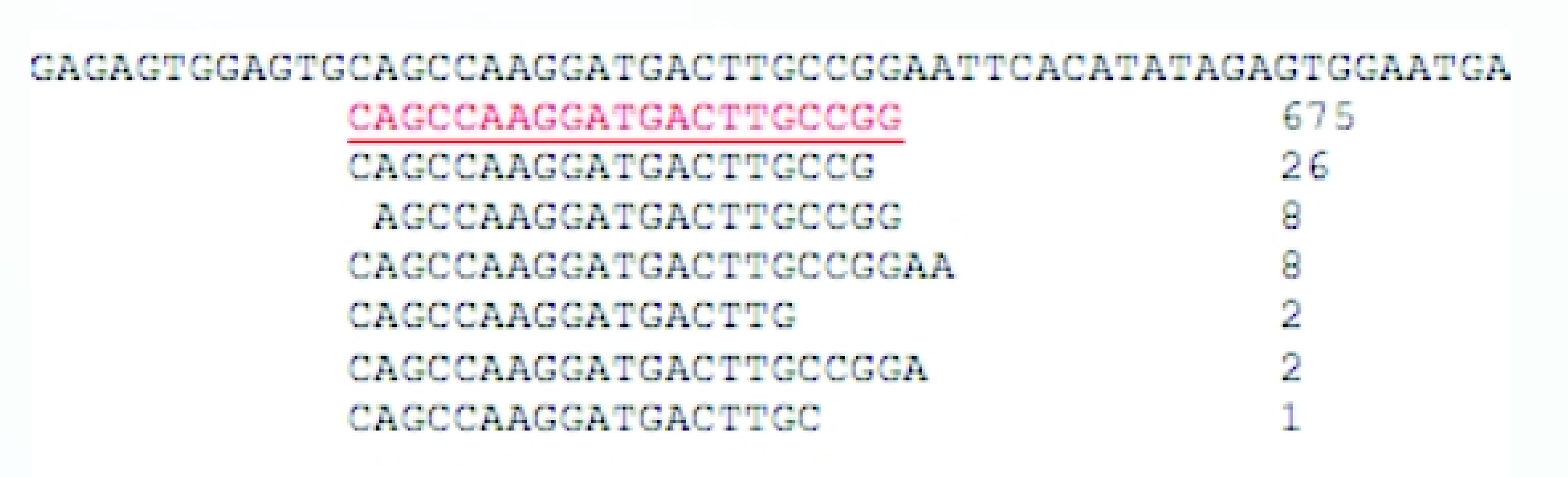
  - Hairpin structure of the pre-miRNA
  - Pre-miRNA localisation (coding/non coding TU intronic/exonic)
  - Presence of cluster
  - Size of the pre-miRNA
- miRNA-5p and miRNA-3p information:



  - Existence of both miRNA-5p and miRNA-3p
  - Sequence conservation
  - Overhang (around 2 nt) related to Drosha and Dicer cuts
  - Size of miRNA-5p and miRNA-3p
  - Overexpression of one of the miRNA-5p and miRNA-3p
- Existence of other products in sRNAseq data

Prediction

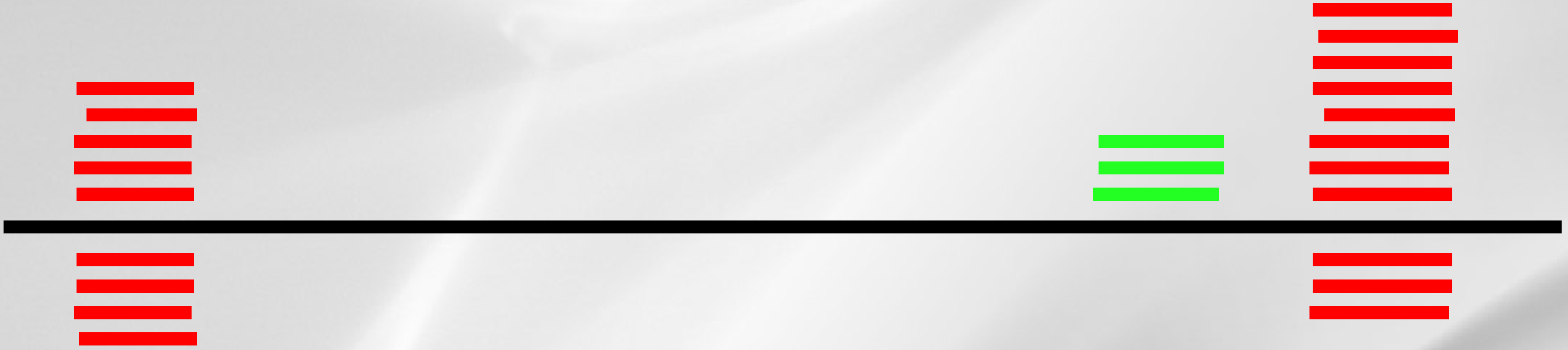
- Precise excision of a 21-22mer is typical of microRNA
  - less represented reads are products of Dicer errors and sequencing/sample preparation artifacts



```
GAGAGTGGAGTGCAGCCAAGGATGACTTGCCGGAATTCACATATAGAGTGGAATGA
            CAGCCAAGGATGACTTGCCGG              675
            CAGCCAAGGATGACTTGCCG               26
             AGCCAAGGATGACTTGCCGG              8
            CAGCCAAGGATGACTTGCCGGAA            8
            CAGCCAAGGATGACTTG                  2
            CAGCCAAGGATGACTTGCCGGA             2
            CAGCCAAGGATGACTTGC                 1
```

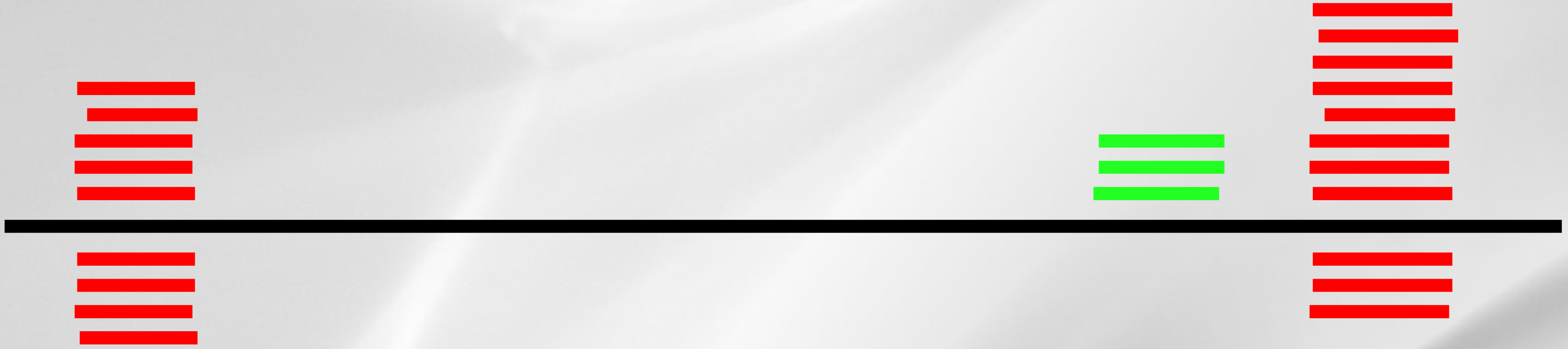
Prediction

- Once the reads mapped



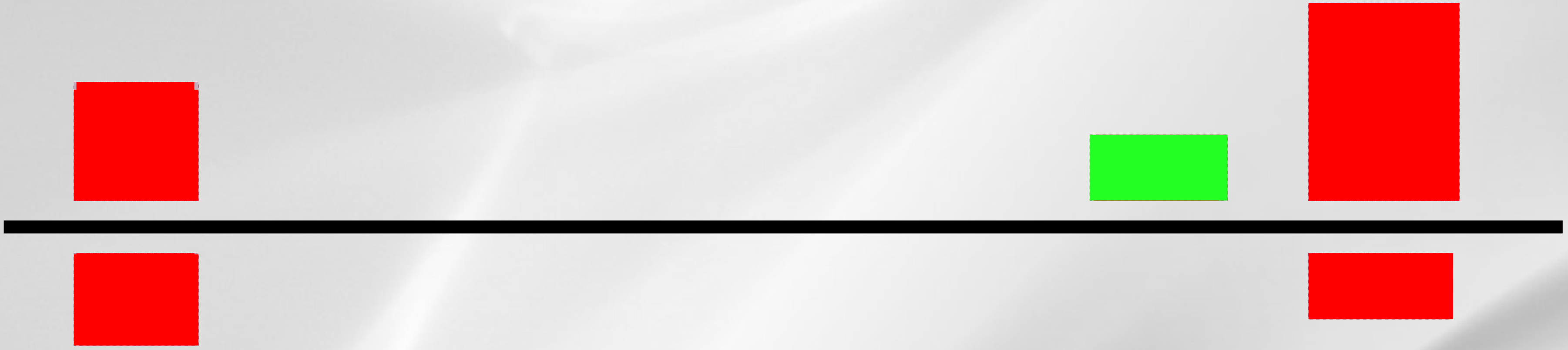
Prediction

- Identify all contiguous read regions



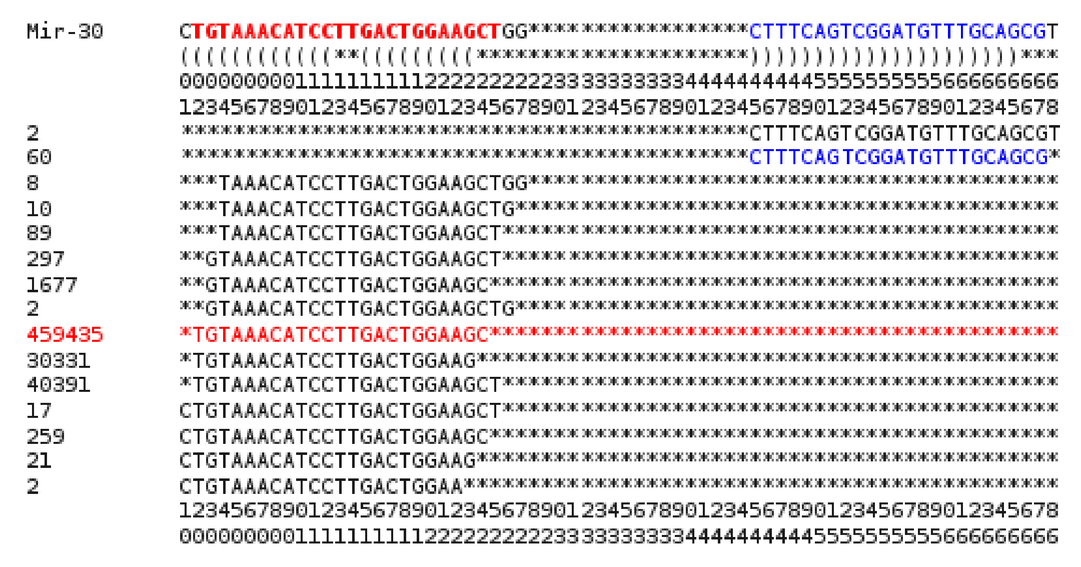
Prediction

- Identify all contiguous read regions

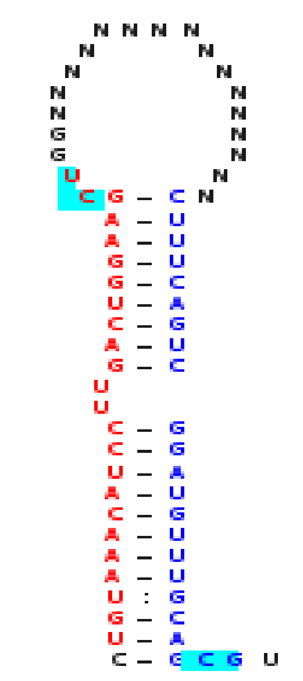


Prediction

- miRNA precursors have a characteristic secondary structure
  - The detection of a microRNA* sequence, opposing the most frequent read in a stable hairpin (but shifted by 2 bases), is sufficient to diagnose a microRNA.

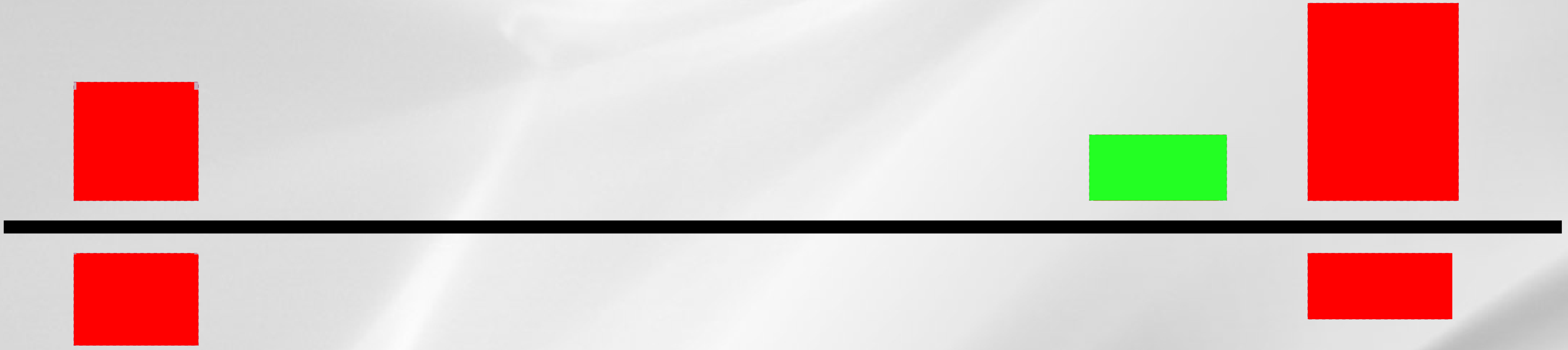


```
Mir-30      CTGTAAACATCCTTGACTGGAAGCTGG*****************CTTTCAGTCGGATGTTTGCAGCGT
            ((((((((((((**(((((((((*********************)))))))))))))))))))))***
            00000000011111111112222222222333333333344444444445555555555666666666
            12345678901234567890123456789012345678901234567890123456789012345678
2           ********************************************CTTTCAGTCGGATGTTTGCAGCGT
60          ********************************************CTTTCAGTCGGATGTTTGCAGCG*
8           ***TAAACATCCTTGACTGGAAGCTGG*****************************************
10          ***TAAACATCCTTGACTGGAAGCTG******************************************
89          ***TAAACATCCTTGACTGGAAGCT*******************************************
297         **GTAAACATCCTTGACTGGAAGCT*******************************************
1677        **GTAAACATCCTTGACTGGAAGC********************************************
2           **GTAAACATCCTTGACTGGAAGCTG******************************************
459435      *TGTAAACATCCTTGACTGGAAGC********************************************
30331       *TGTAAACATCCTTGACTGGAAG*********************************************
40391       *TGTAAACATCCTTGACTGGAAGCT*******************************************
17          CTGTAAACATCCTTGACTGGAAGCT*******************************************
259         CTGTAAACATCCTTGACTGGAAGC********************************************
21          CTGTAAACATCCTTGACTGGAAG*********************************************
2           CTGTAAACATCCTTGACTGGAA**********************************************
            12345678901234567890123456789012345678901234567890123456789012345678
            00000000011111111112222222222333333333344444444445555555555666666666
```



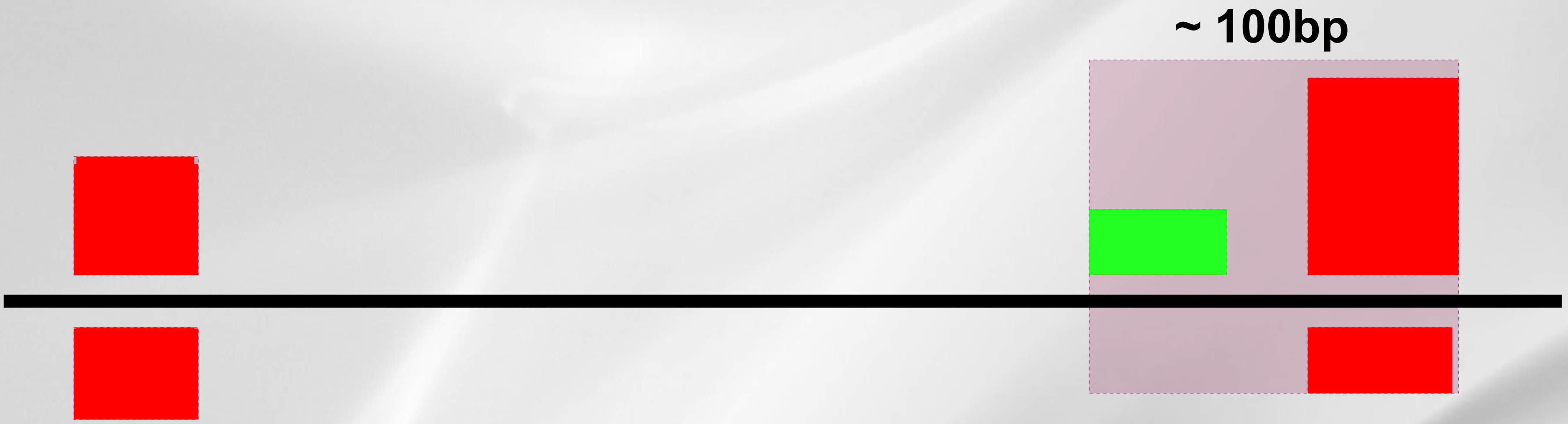
Prediction

- Extend and fold read regions



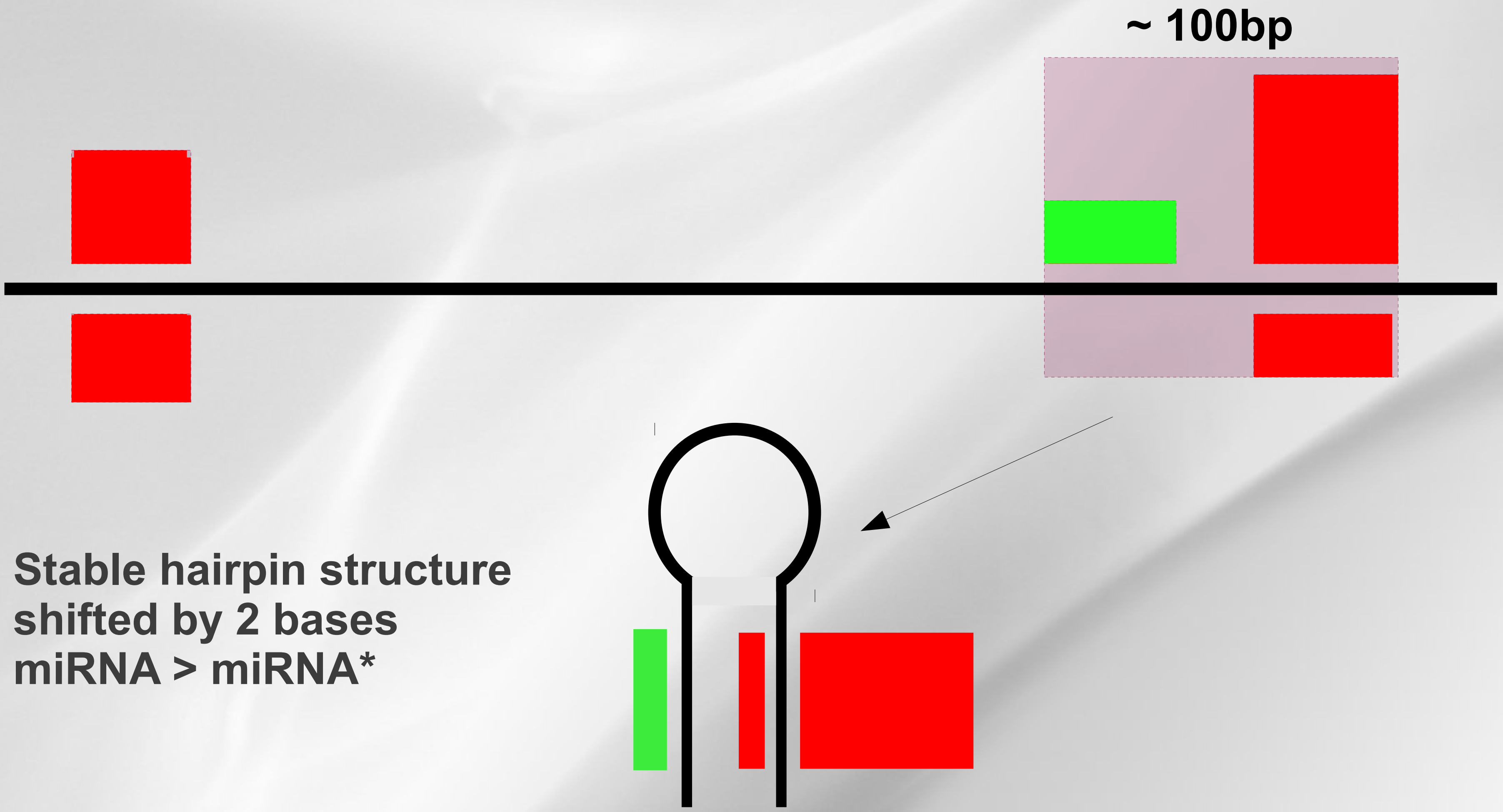
Prediction

- Extend and fold read regions



Prediction

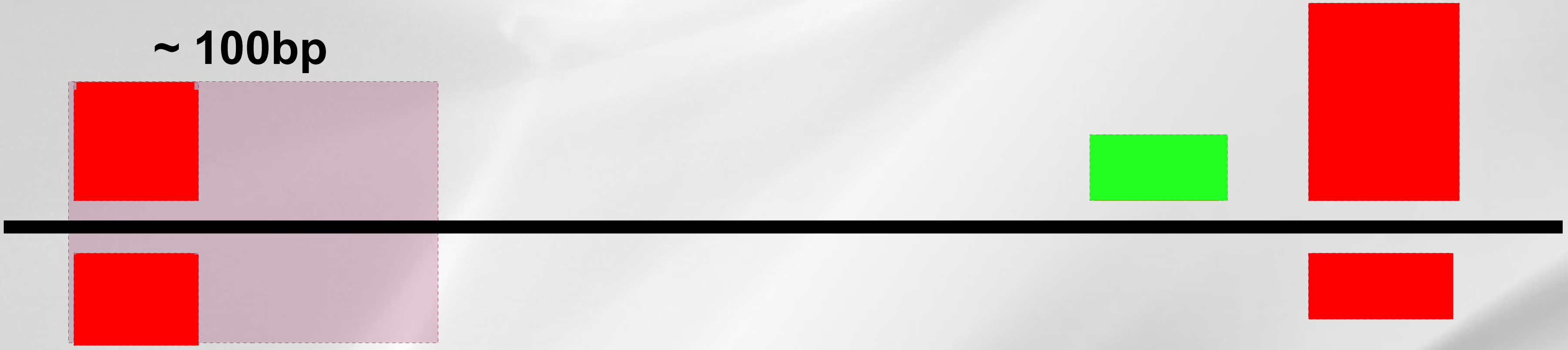
- Extend and fold read regions



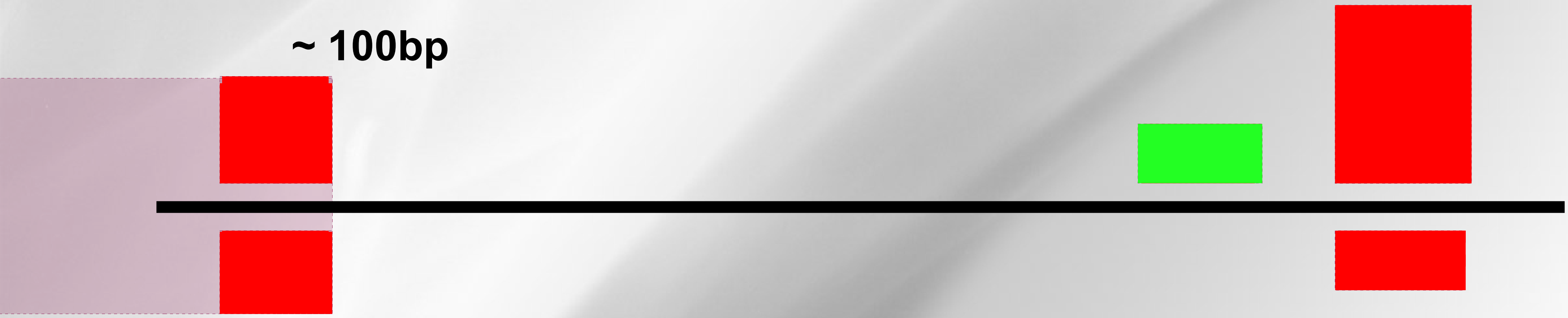
- Stable hairpin structure shifted by 2 bases
- miRNA > miRNA*

Prediction

- Extend and fold read regions

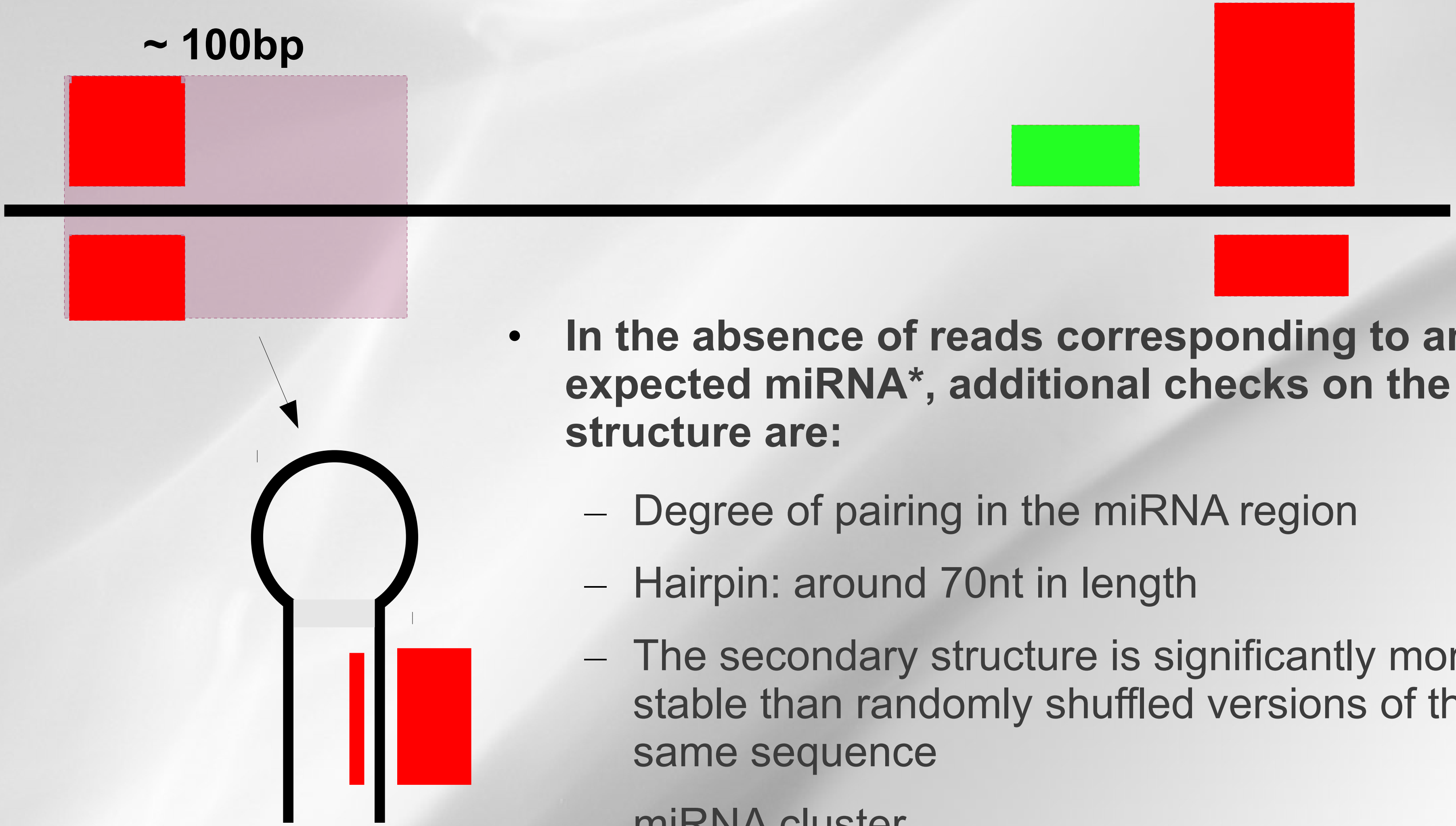

~ 100bp

OR


~ 100bp

Prediction

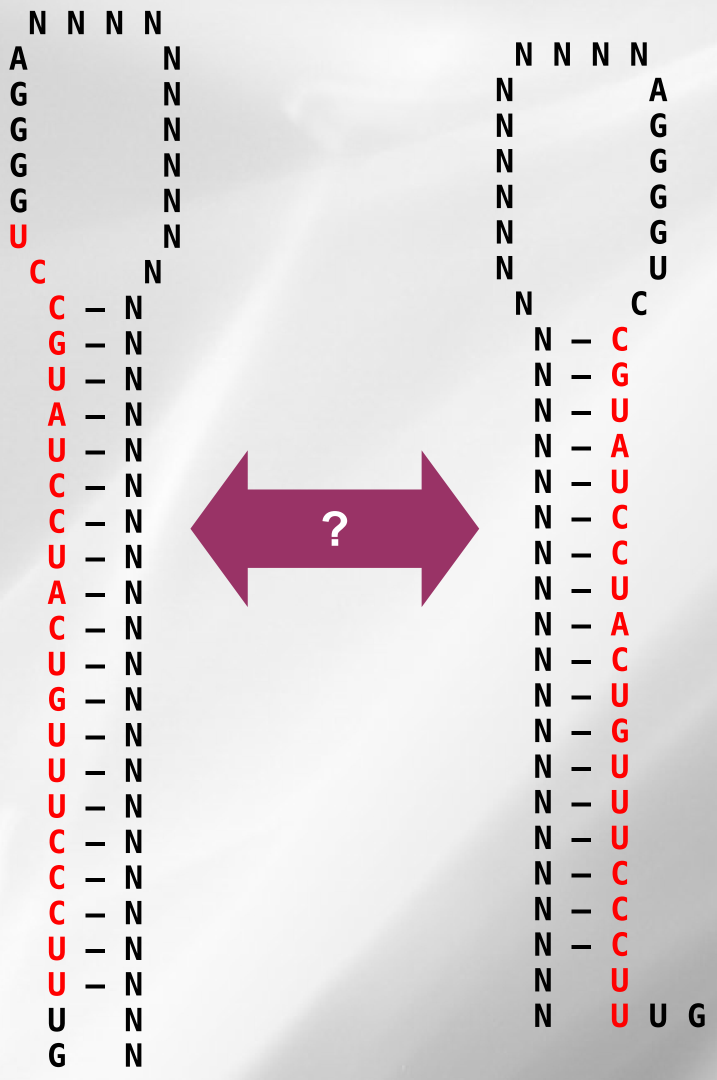
- Extend and fold read regions

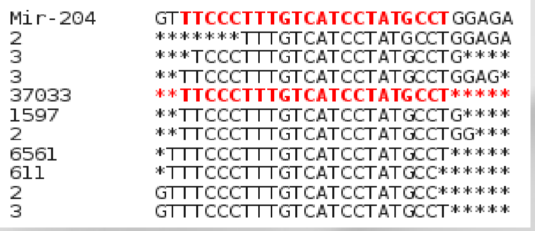


- **In the absence of reads corresponding to an expected miRNA*, additional checks on the structure are:**
  - Degree of pairing in the miRNA region
  - Hairpin: around 70nt in length
  - The secondary structure is significantly more stable than randomly shuffled versions of the same sequence
  - miRNA cluster

Prediction

- Which one should be used ?

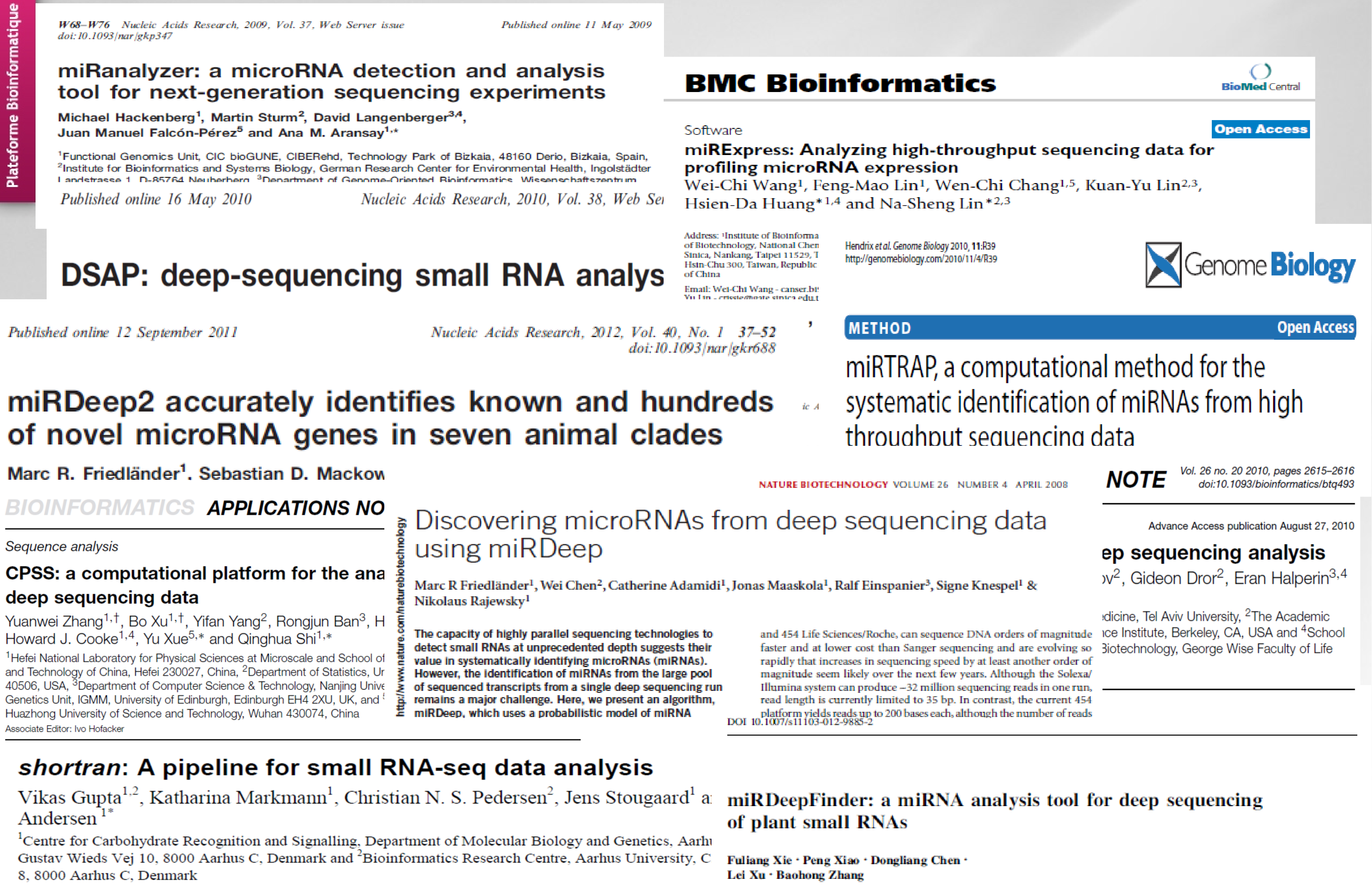




```
Mir-204       GTTTCCCTTTGTCATCCTATGCCTGGAGA
2             *******TTTGTCATCCTATGCCTGGAGA
3             ***TCCCTTTGTCATCCTATGCCTG****
3             **TTCCCTTTGTCATCCTATGCCTGGAG*
37033         **TTCCCTTTGTCATCCTATGCCT*****
1597          **TTCCCTTTGTCATCCTATGCCTG****
2             **TTCCCTTTGTCATCCTATGCCTGG***
6561          *TTTCCCTTTGTCATCCTATGCCT*****
611           *TTTCCCTTTGTCATCCTATGCC******
2             GTTTCCCTTTGTCATCCTATGCC******
3             GTTTCCCTTTGTCATCCTATGCCT*****
```

Prediction

Existing software



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

Published online 16 May 2010 Nucleic Acids Research, 2010, Vol. 38, Web Se

DSAP: deep-sequencing small RNA analys

BMC Bioinformatics BioMed Central

Software Open Access

miRExpress: Analyzing high-throughput sequencing data for profiling microRNA expression

Wei-Chi Wang[1], Feng-Mao Lin[1], Wen-Chi Chang[1,5], Kuan-Yu Lin[2,3], Hsien-Da Huang*[1,4] and Na-Sheng Lin*[2,3]

Address: [1]Institute of Bioinforma
of Biotechnology, National Chen
Sinica, Nankang, Taipei 11529, T
Hsin-Chu 300, Taiwan, Republic
of China

Email: Wei-Chi Wang - canser.bi
Yu Lin - cristie@gate.sinica.edu.t

Hendrix et al. Genome Biology 2010, 11:R39
http://genomebiology.com/2010/11/4/R39

Genome Biology

METHOD Open Access

miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data

Published online 12 September 2011 Nucleic Acids Research, 2012, Vol. 40, No. 1 37–52
doi:10.1093/nar/gkr688

miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades

Marc R. Friedländer[1], Sebastian D. Mackow

BIOINFORMATICS APPLICATIONS NO

Sequence analysis

CPSS: a computational platform for the ana deep sequencing data

Yuanwei Zhang[1,†], Bo Xu[1,†], Yifan Yang[2], Rongjun Ban[3], H Howard J. Cooke[1,4], Yu Xue[5,*] and Qinghua Shi[1,*]

[1]Hefei National Laboratory for Physical Sciences at Microscale and School of and Technology of China, Hefei 230027, China, [2]Department of Statistics, Ur 40506, USA, [3]Department of Computer Science & Technology, Nanjing Unive Genetics Unit, IGMM, University of Edinburgh, Edinburgh EH4 2XU, UK, and [5] Huazhong University of Science and Technology, Wuhan 430074, China

Associate Editor: Ivo Hofacker

NATURE BIOTECHNOLOGY VOLUME 26 NUMBER 4 APRIL 2008

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

DOI 10.1007/s11103-012-9885-2

NOTE Vol. 26 no. 20 2010, pages 2615–2616
doi:10.1093/bioinformatics/btq493

Advance Access publication August 27, 2010

ep sequencing analysis

ov[2], Gideon Dror[2], Eran Halperin[3,4]

edicine, Tel Aviv University, [2]The Academic nce Institute, Berkeley, CA, USA and [4]School Biotechnology, George Wise Faculty of Life

shortran: A pipeline for small RNA-seq data analysis

Vikas Gupta[1,2], Katharina Markmann[1], Christian N. S. Pedersen[2], Jens Stougaard[1] a Andersen[1*]

[1]Centre for Carbohydrate Recognition and Signalling, Department of Molecular Biology and Genetics, Aarhu Gustav Wieds Vej 10, 8000 Aarhus C, Denmark and [2]Bioinformatics Research Centre, Aarhus University, C 8, 8000 Aarhus C, Denmark

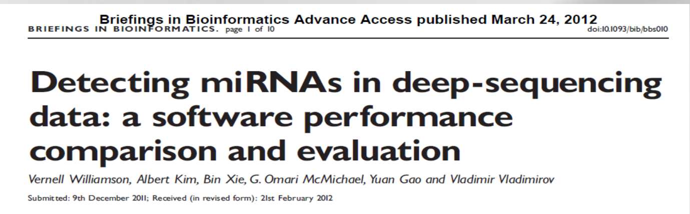
miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs

Fuliang Xie · Peng Xiao · Dongliang Chen · Lei Xu · Baohong Zhang

Prediction

Existing software

- Basic features
  - Availability (web/executable)
  - Computing resources (time, memory)
  - Reads pre-processing
  - Mapping
  - Identification



Briefings in Bioinformatics Advance Access published March 24, 2012

BRIEFINGS IN BIOINFORMATICS. page 1 of 10 doi:10.1093/bib/bbs010

Detecting miRNAs in deep-sequencing data: a software performance comparison and evaluation

Vernell Williamson, Albert Kim, Bin Xie, G. Omari McMichael, Yuan Gao and Vladimir Vladimirov

Submitted: 9th December 2011; Received (in revised form): 21st February 2012

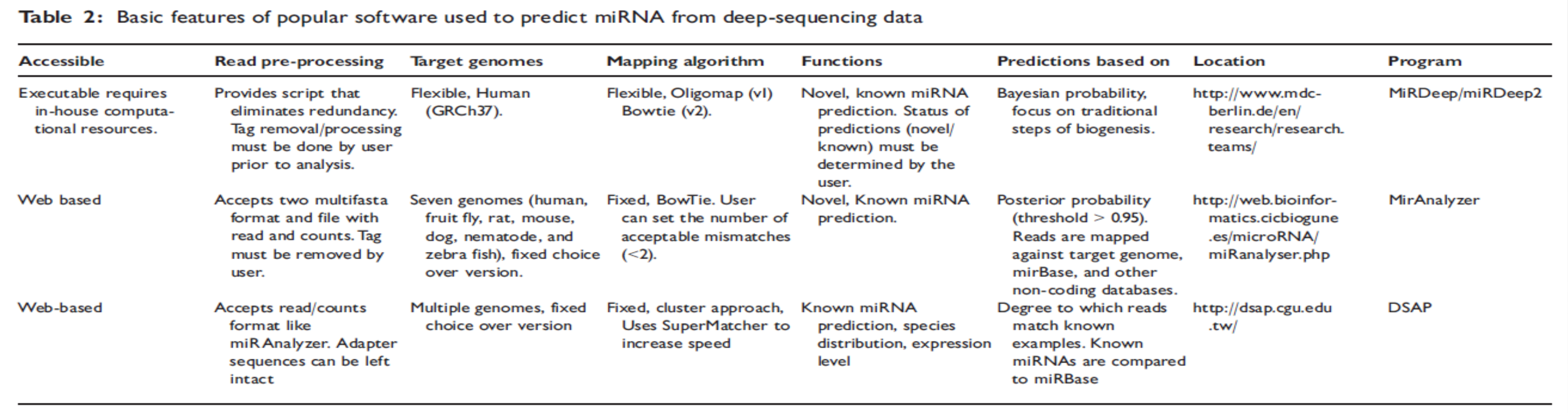


Table 2: Basic features of popular software used to predict miRNA from deep-sequencing data

Accessible	Read pre-processing	Target genomes	Mapping algorithm	Functions	Predictions based on	Location	Program
Executable requires in-house computational resources.	Provides script that eliminates redundancy. Tag removal/processing must be done by user prior to analysis.	Flexible, Human (GRCh37).	Flexible, Oligomap (vl) Bowtie (v2).	Novel, known miRNA prediction. Status of predictions (novel/known) must be determined by the user.	Bayesian probability, focus on traditional steps of biogenesis.	http://www.mdc-berlin.de/en/research/research.teams/	MiRDeep/miRDeep2
Web based	Accepts two multifasta format and file with read and counts. Tag must be removed by user.	Seven genomes (human, fruit fly, rat, mouse, dog, nematode, and zebra fish), fixed choice over version.	Fixed, BowTie. User can set the number of acceptable mismatches (<2).	Novel, Known miRNA prediction.	Posterior probability (threshold > 0.95). Reads are mapped against target genome, mirBase, and other non-coding databases.	http://web.bioinformatics.cicbiogune.es/microRNA/miRanalyser.php	MirAnalyzer
Web-based	Accepts read/counts format like miRAnalyzer. Adapter sequences can be left intact	Multiple genomes, fixed choice over version	Fixed, cluster approach, Uses SuperMatcher to increase speed	Known miRNA prediction, species distribution, expression level	Degree to which reads match known examples. Known miRNAs are compared to miRBase	http://dsap.cgu.edu.tw/	DSAP

Prediction

Existing software

- Reads pre-processing
  - Adaptators trimming
  - Redundancy
  - Repeats
  - Other ncRNA
  - Size of the mature miRNA (min/max)

Prediction

Existing software

- Mapping
  - Size and region of the read
  - Number of locations
    - Considered
    - Reported
  - Error(s) consideration in mapping
  - Quality of the read

Prediction

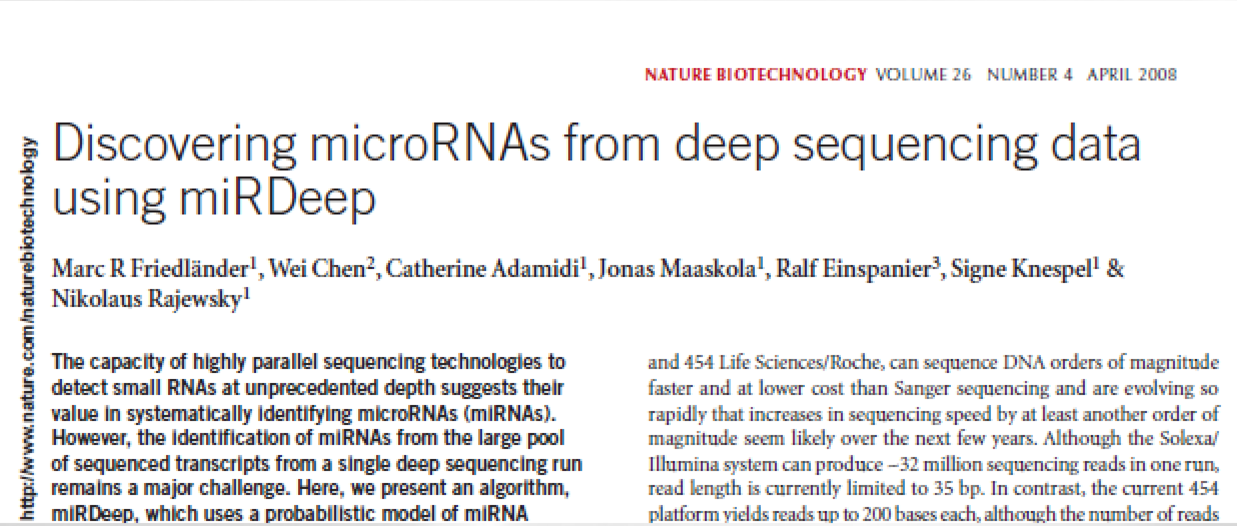
Existing software

- Precursor identification
  - Length and bounds of the theoretical sequence (folding)
  - Alignment of the read against known miRNA
- Post processing step: assessment of the potential miRNA
  - Different methods: SVM, bayesian statistics based score, combinatorial rules...
    - Location of the read on the precursor
    - 2 nt overhang of the mature miRNA/precursor
    - Accuracy of the folding (HP structure, energy, Z-score...)

Plateforme Bioinformatique Midi-Pyrénées

Existing software

miRDeep2



NATURE BIOTECHNOLOGY VOLUME 26 NUMBER 4 APRIL 2008

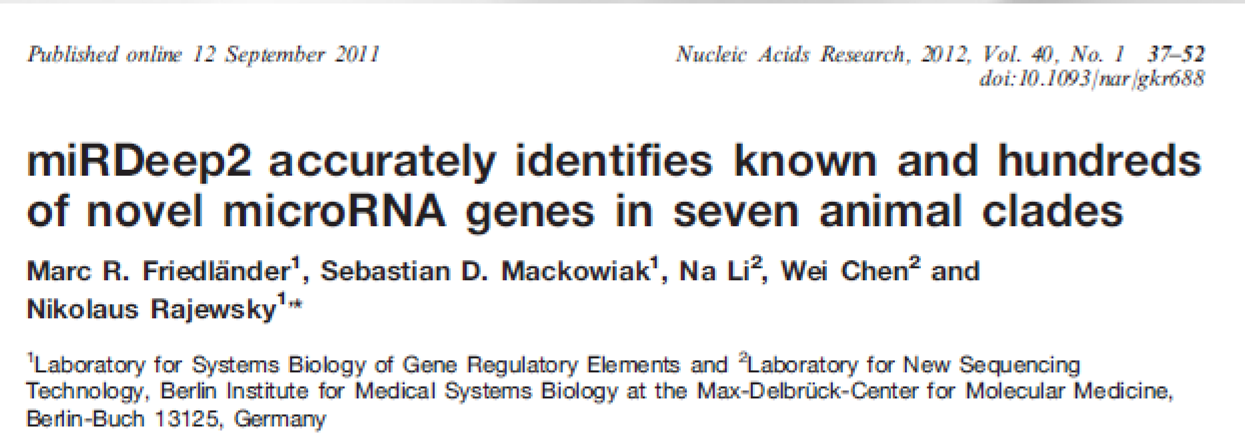
http://www.nature.com/naturebiotechnology

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



Published online 12 September 2011 *Nucleic Acids Research, 2012, Vol. 40, No. 1* ***37–52***
doi:10.1093/nar/gkr688

miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades

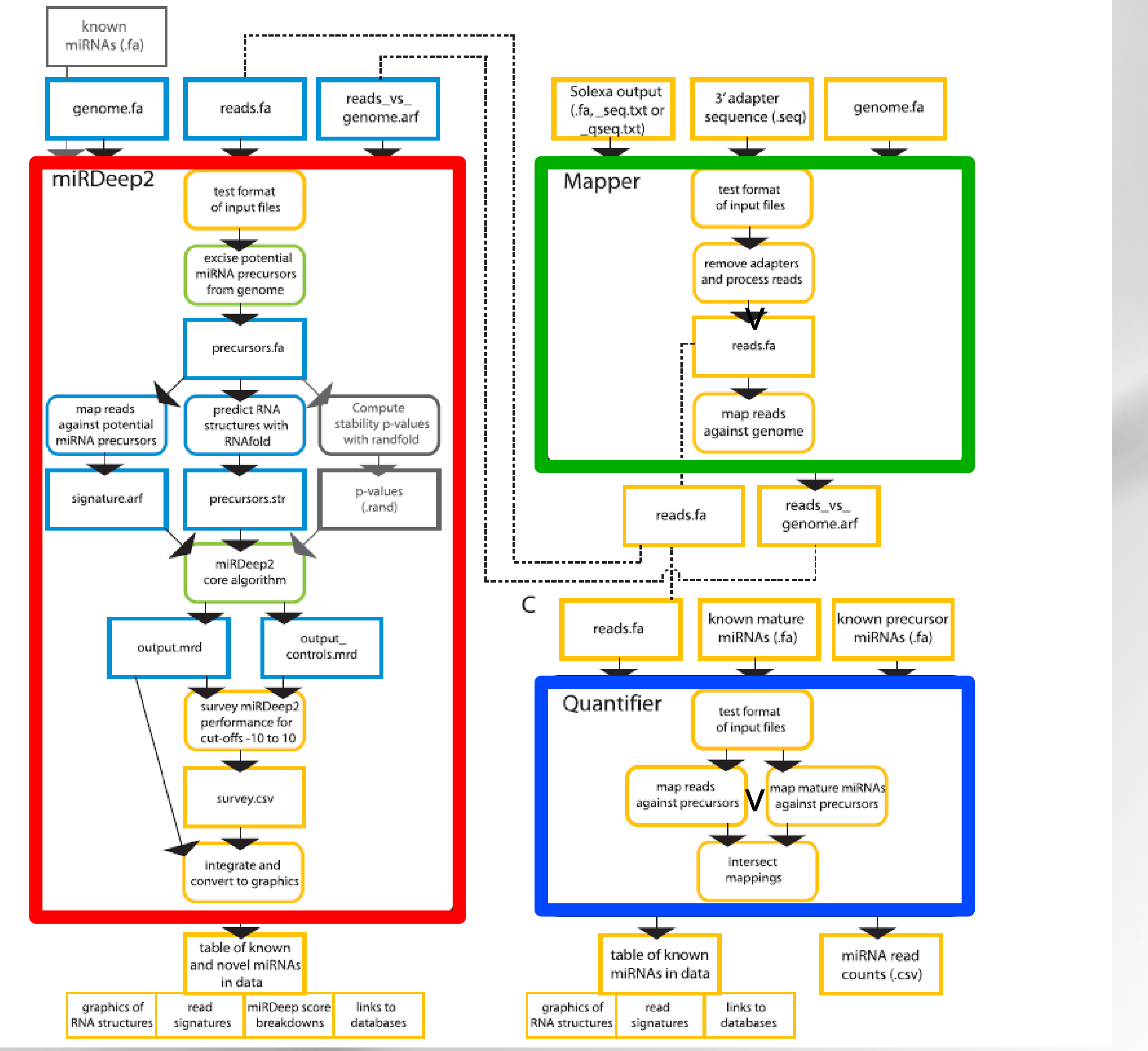
Marc R. Friedländer[1], Sebastian D. Mackowiak[1], Na Li[2], Wei Chen[2] and Nikolaus Rajewsky[1]*

[1]Laboratory for Systems Biology of Gene Regulatory Elements and [2]Laboratory for New Sequencing Technology, Berlin Institute for Medical Systems Biology at the Max-Delbrück-Center for Molecular Medicine, Berlin-Buch 13125, Germany

Existing software
miRDeep2

Three modules

- MiRDeep2
- Mapper
- Quantifier



known miRNAs (.fa)

genome.fa

reads.fa

reads_vs_genome.arf

miRDeep2

test format of input files

excise potential miRNA precursors from genome

precursors.fa

map reads against potential miRNA precursors

predict RNA structures with RNAfold

Compute stability p-values with randfold

signature.arf

precursors.str

p-values (.rand)

miRDeep2 core algorithm

output.mrd

output_controls.mrd

survey miRDeep2 performance for cut-offs -10 to 10

survey.csv

integrate and convert to graphics

table of known and novel miRNAs in data

graphics of RNA structures

read signatures

miRDeep score breakdowns

links to databases

Solexa output (.fa, _seq.txt or _qseq.txt)

3' adapter sequence (.seq)

genome.fa

Mapper

test format of input files

remove adapters and process reads

V

reads.fa

map reads against genome

reads.fa

reads_vs_genome.arf

C

reads.fa

known mature miRNAs (.fa)

known precursor miRNAs (.fa)

Quantifier

test format of input files

map reads against precursors

V

map mature miRNAs against precursors

intersect mappings

table of known miRNAs in data

miRNA read counts (.csv)

graphics of RNA structures

read signatures

links to databases

Existing software
miRDeep2

The Mapper module

- Reads processing
  - Remove redundancy and keep # occurrences

- Map reads with bowtie (or BWA)
  - Bowtie -f -n 0 -e 80 -l 18 -a -m 5 -best -strata

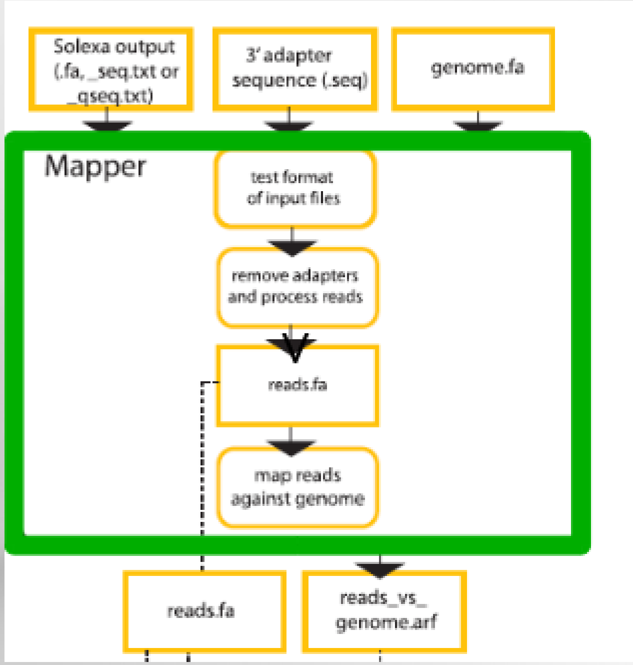
0 mismatch In the seed

2 mismatches after the seed

The size of the Seed is 18nt

Do not map more than 5

Order alignments from best to worse

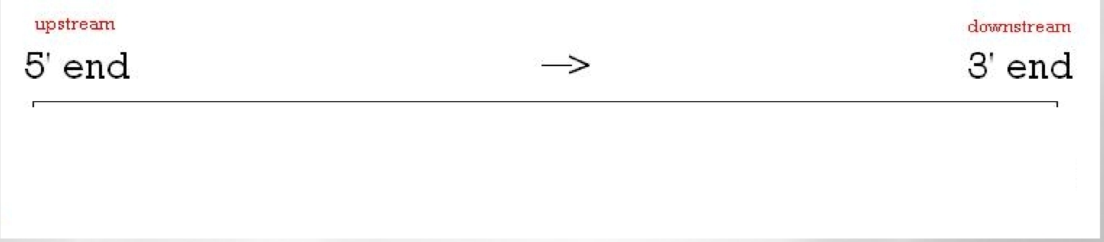

Solexa output (.fa, _seq.txt or _qseq.txt)
3' adapter sequence (.seq)
genome.fa
Mapper
test format of input files
remove adapters and process reads
reads.fa
map reads against genome
reads.fa
reads_vs_genome.arf

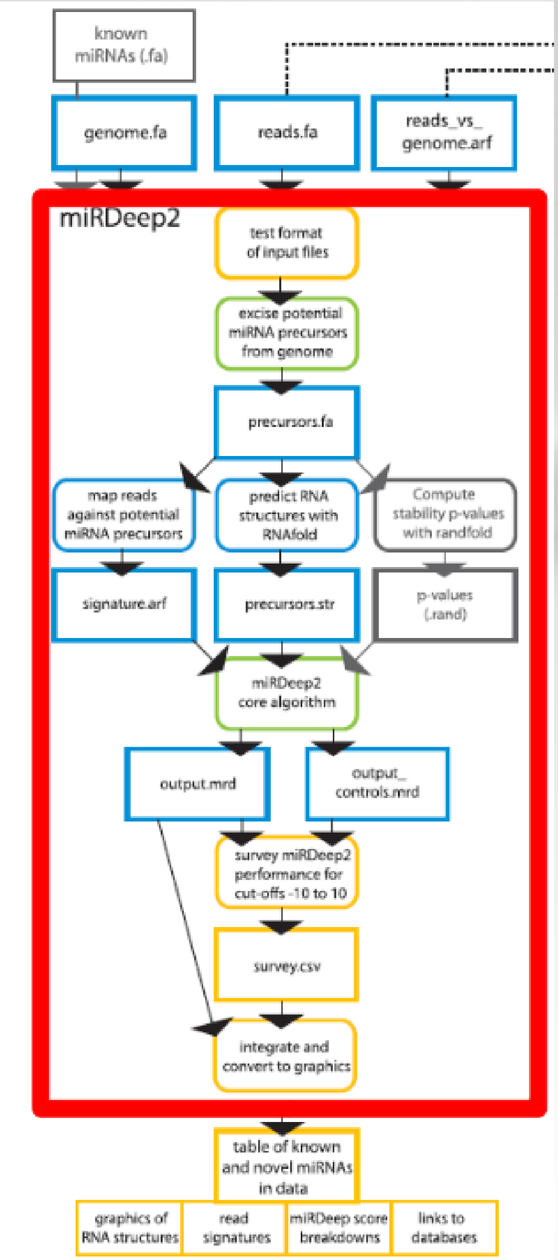
Existing software

miRDeep2

The miRDeep2 module

- Scan of both strands from 5' to 3'



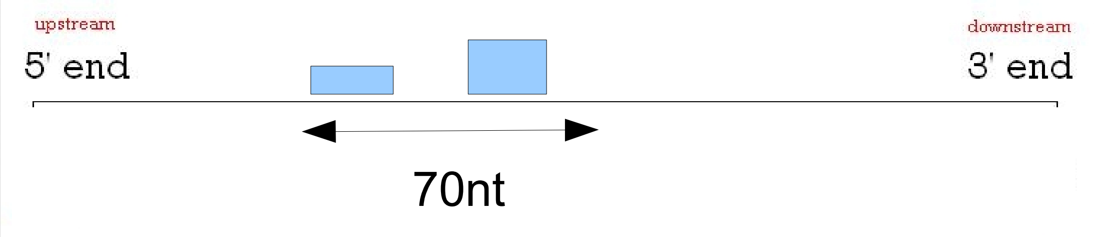

known miRNAs (.fa)
genome.fa
reads.fa
reads_vs_genome.arf
miRDeep2
test format of input files
excise potential miRNA precursors from genome
precursors.fa
map reads against potential miRNA precursors
predict RNA structures with RNAfold
Compute stability p-values with randfold
signature.arf
precursors.str
p-values (.rand)
miRDeep2 core algorithm
output.mrd
output_controls.mrd
survey miRDeep2 performance for cut-offs -10 to 10
survey.csv
integrate and convert to graphics
table of known and novel miRNAs in data
graphics of RNA structures
read signatures
miRDeep score breakdowns
links to databases

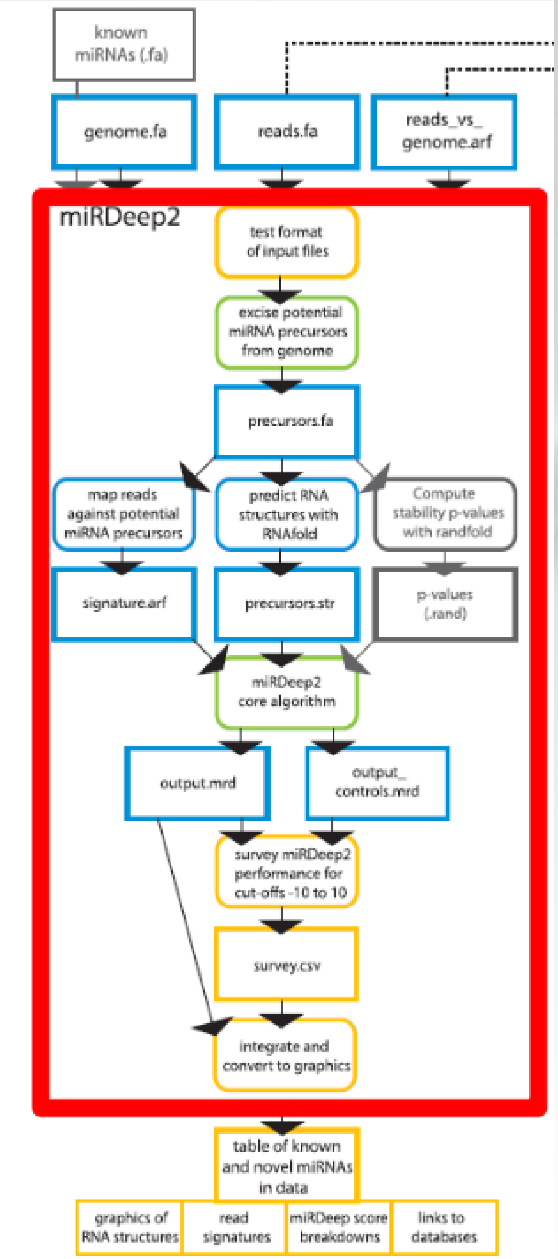
Existing software

miRDeep & miRDeep2

The miRDeep2 module

- Scan of both strands from 5' to 3
  - Search the best stack of reads (heigth 1 or more) in a distance of 70nt



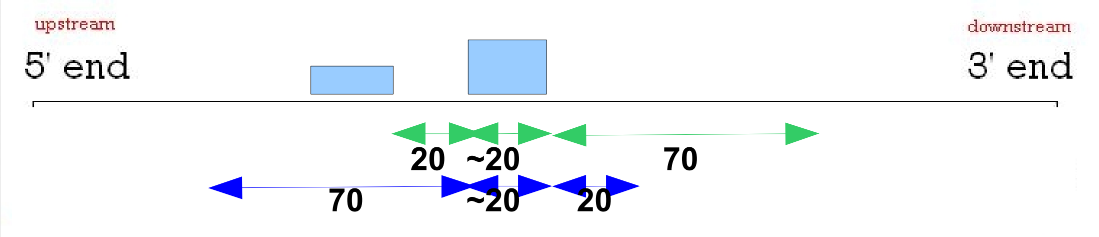

known miRNAs (.fa)
genome.fa
reads.fa
reads_vs_genome.arf
miRDeep2
test format of input files
excise potential miRNA precursors from genome
precursors.fa
map reads against potential miRNA precursors
predict RNA structures with RNAfold
Compute stability p-values with randfold
signature.arf
precursors.str
p-values (.rand)
miRDeep2 core algorithm
output.mrd
output_controls.mrd
survey miRDeep2 performance for cut-offs -10 to 10
survey.csv
integrate and convert to graphics
table of known and novel miRNAs in data
graphics of RNA structures
read signatures
miRDeep score breakdowns
links to databases

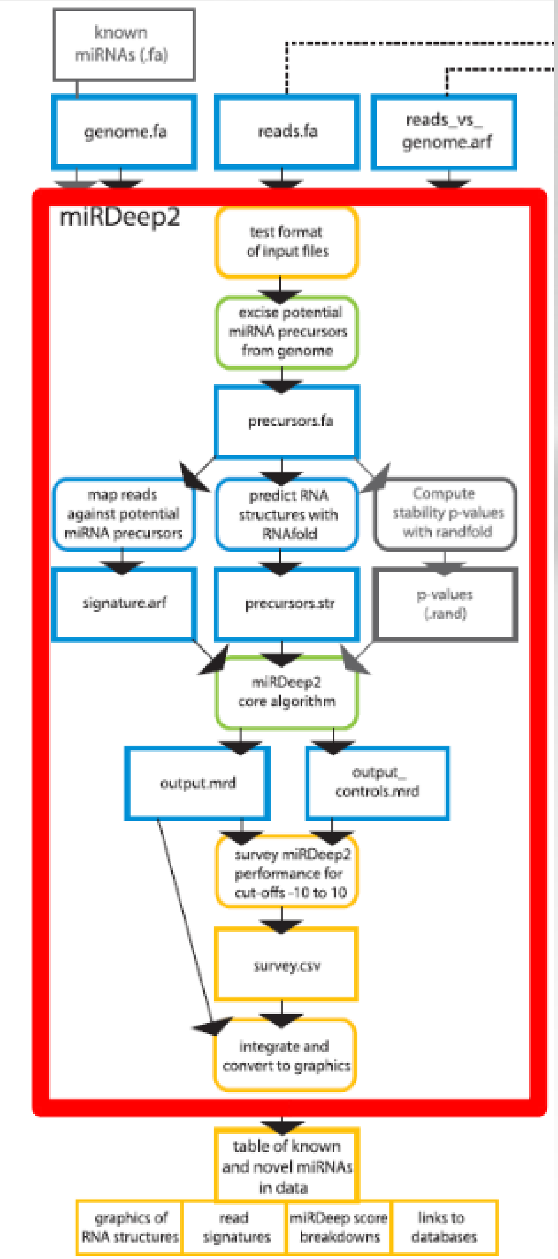
Existing software

miRDeep & miRDeep2

The miRDeep2 module

- Scan of both strands from 5' to 3
  - Search the best stack of reads (heigth 1 or more) in a distance of 70nt
  - Excise potential precursors on both sides





known miRNAs (.fa)

genome.fa

reads.fa

reads_vs_genome.arf

miRDeep2

test format of input files

excise potential miRNA precursors from genome

precursors.fa

map reads against potential miRNA precursors

predict RNA structures with RNAfold

Compute stability p-values with randfold

signature.arf

precursors.str

p-values (.rand)

miRDeep2 core algorithm

output.mrd

output_controls.mrd

survey miRDeep2 performance for cut-offs -10 to 10

survey.csv

integrate and convert to graphics

table of known and novel miRNAs in data

graphics of RNA structures

read signatures

miRDeep score breakdowns

links to databases

Existing software
miRDeep2

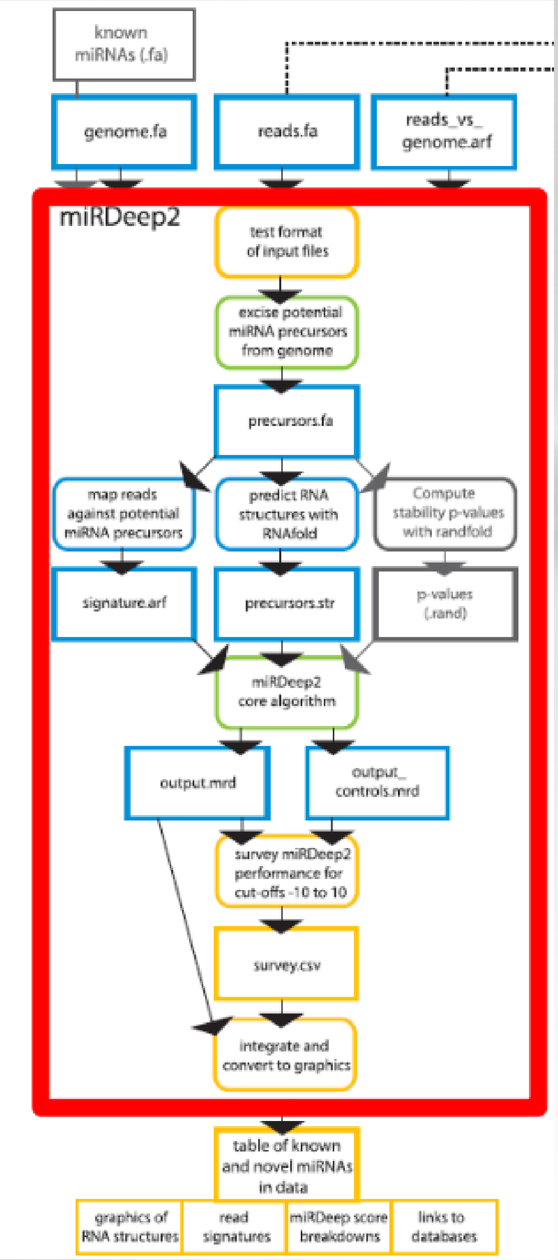
The miRDeep2 module

- Scan of both strands from 5' to 3
  - Search the best stack of reads (heigth 1 or more) in a distance of 70nt
  - Excise potential precursors on both sides
  - Go on from 1 nt after the last position excised

- If the number of candidate precursor>50.000, repeat the process (height of stack=height of stack + 1)

- Prepare the file of precursor signature
  - Align reads against precursors (1 MM allowed)
  - Align known miRNA against precursors (0 MM allowed)

- Evaluation of candidate precursors
  - Fold candidate precursors (RNAfold + Randfold)
    - Unbifurcated hairpins
  - Score the candidates
    - Valid alignment of reads on the precursor
    - 60% of nt in the mature part paired


known miRNAs (.fa)
genome.fa
reads.fa
reads_vs_ genome.arf
miRDeep2
test format of input files
excise potential miRNA precursors from genome
precursors.fa
map reads against potential miRNA precursors
predict RNA structures with RNAfold
Compute stability p-values with randfold
signature.arf
precursors.str
p-values (.rand)
miRDeep2 core algorithm
output.mrd
output_ controls.mrd
survey miRDeep2 performance for cut-offs -10 to 10
survey.csv
integrate and convert to graphics
table of known and novel miRNAs in data
graphics of RNA structures
read signatures
miRDeep score breakdowns
links to databases

Existing software

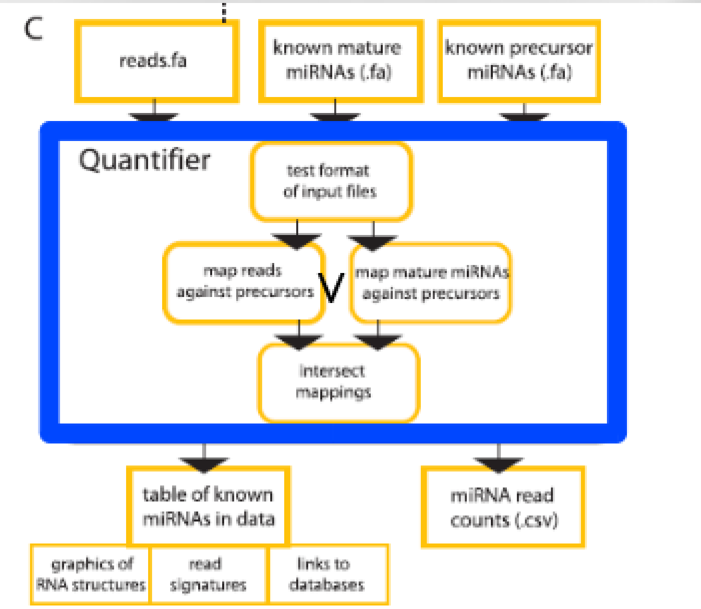
miRDeep2

The Quantifier module

Identifies and quantifies known mature miRNA given

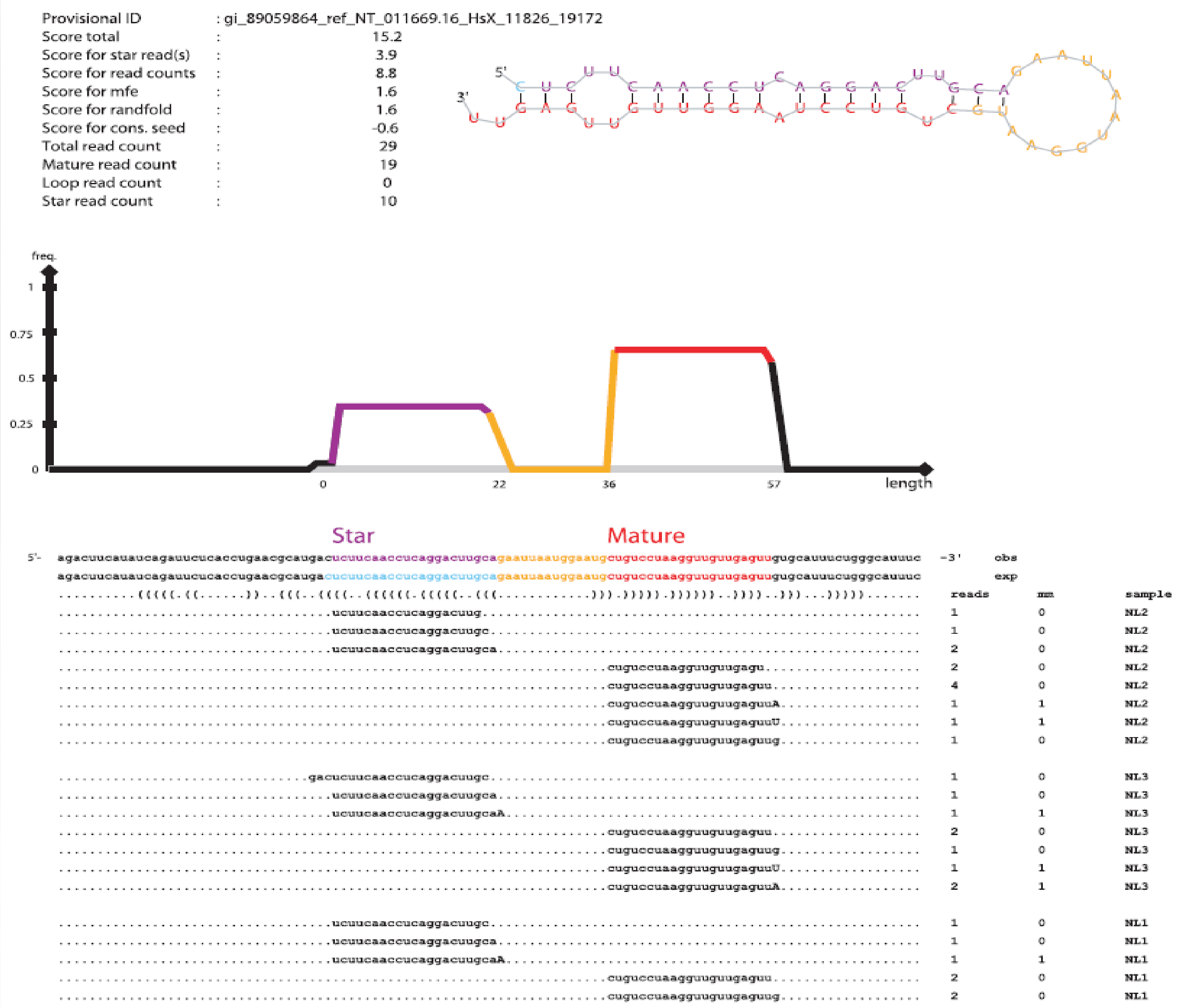
- Know mature miRNA
- Know miRNA precursors

Use Bowtie for miRNA/reads alignment


C
reads.fa
known mature miRNAs (.fa)
known precursor miRNAs (.fa)
Quantifier
test format of input files
map reads against precursors
V
map mature miRNAs against precursors
intersect mappings
table of known miRNAs in data
miRNA read counts (.csv)
graphics of RNA structures
read signatures
links to databases

Existing software

miRDeep2



```
Provisional ID             : gi_89059864_ref_NT_011669.16_HsX_11826_19172
Score total                :      15.2
Score for star read(s)     :       3.9
Score for read counts      :       8.8
Score for mfe              :       1.6
Score for randfold         :       1.6
Score for cons. seed       :      -0.6
Total read count           :       29
Mature read count          :       19
Loop read count            :        0
Star read count            :       10
```

Star   Mature

```
5'-  agacuucauaucagauucucaccugaacgcaugacucuucaaccucaggacuugcagaauuaauggaaugcuguccuaagguuguugaguugugcauuucugggcauuuc  -3'  obs
     agacuucauaucagauucucaccugaacgcaugacucuucaaccucaggacuugcagaauuaauggaaugcuguccuaagguuguugaguugugcauuucugggcauuuc       exp
     ..........(((((.((......))..(((..((((..((((((.(((((..(((............))).))))).))))))..))))..)))...))))).......  reads   mm    sample
     ...................................ucuucaaccucaggacuug........................................................  1       0     NL2
     ...................................ucuucaaccucaggacuugc.......................................................  1       0     NL2
     ...................................ucuucaaccucaggacuugca......................................................  2       0     NL2
     ......................................................................cuguccuaagguuguugagu....................  2       0     NL2
     ......................................................................cuguccuaagguuguugaguu...................  4       0     NL2
     ......................................................................cuguccuaagguuguugaguuA..................  1       1     NL2
     ......................................................................cuguccuaagguuguugaguuU..................  1       1     NL2
     ......................................................................cuguccuaagguuguugaguug..................  1       0     NL2

     ................................gacucuucaaccucaggacuugc.......................................................  1       0     NL3
     ...................................ucuucaaccucaggacuugca......................................................  1       0     NL3
     ...................................ucuucaaccucaggacuugcaA.....................................................  1       1     NL3
     ......................................................................cuguccuaagguuguugaguu...................  2       0     NL3
     ......................................................................cuguccuaagguuguugaguug..................  1       0     NL3
     ......................................................................cuguccuaagguuguugaguuU..................  1       1     NL3
     ......................................................................cuguccuaagguuguugaguuA..................  2       1     NL3

     ...................................ucuucaaccucaggacuugc.......................................................  1       0     NL1
     ...................................ucuucaaccucaggacuugca......................................................  1       0     NL1
     ...................................ucuucaaccucaggacuugcaA.....................................................  1       1     NL1
     ......................................................................cuguccuaagguuguugaguu...................  2       0     NL1
     ......................................................................cuguccuaagguuguugaguug..................  2       0     NL1
```

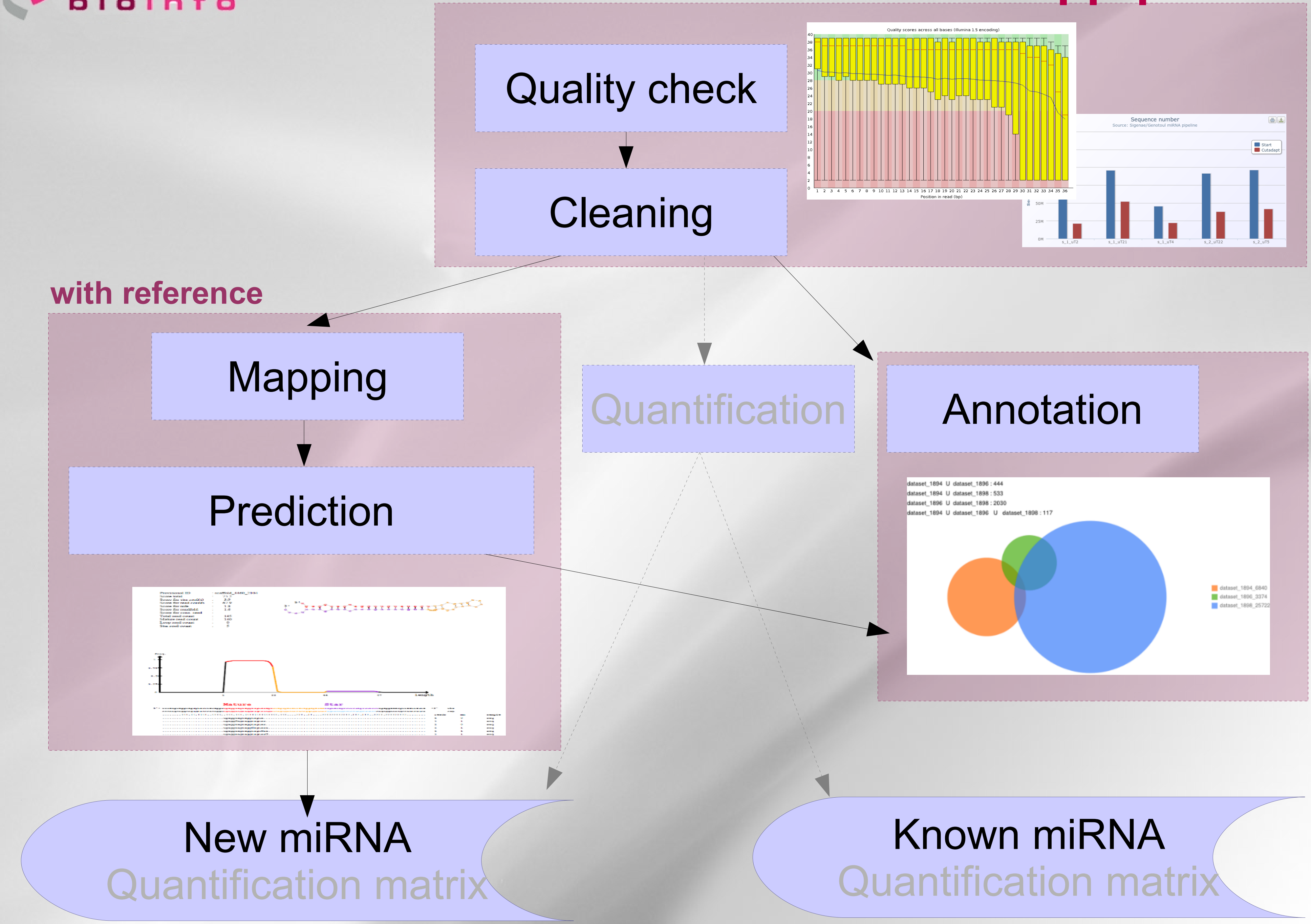
Exercice:

- Back to miRdeep2-core results

Plateforme Bioinformatique Midi-Pyrénées

genotoul bioinfo

small RNAseq pipeline



Quality check

Cleaning

with reference

Mapping

Prediction

Quantification

Annotation

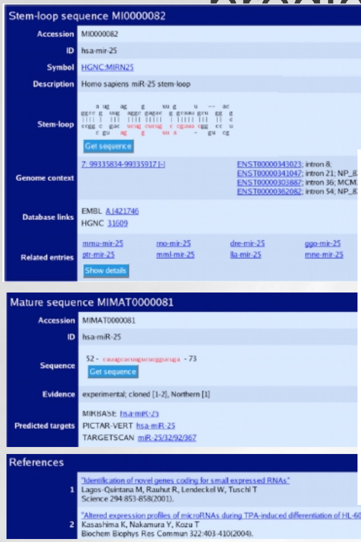
New miRNA
Quantification matrix

Known miRNA
Quantification matrix

Quality scores across all bases (Illumina 1.5 encoding)
Position in read (bp)
Sequence number
Source: Sigenae/Genotoul miRNA pipeline
Start
Cutadapt
s_1_uT2
s_1_uT21
s_1_uT4
s_2_uT22
s_2_uT5
dataset_1894 U dataset_1896 : 444
dataset_1894 U dataset_1898 : 533
dataset_1896 U dataset_1898 : 2030
dataset_1894 U dataset_1896 U dataset_1898 : 117
dataset_1894_6840
dataset_1896_3374
dataset_1898_25722
Mature
Star

Annotation

- Useful databases:
  - miRbase (http://microrna.sanger.ac.uk/) miRBase
    - miRBase::Registry provides names to novel miRNA genes prior to their publication.
    - **miRBase::Sequences provides miRNA sequence data, annotation, references and links to other resources for all published miRNAs.**
    - miRBase::Targets provides an automated pipeline for the prediction of targets for all published animal miRNAs.



Stem-loop sequence MI0000082

Accession	MI0000082
ID	hsa-mir-25
Symbol	HGNC:MIRN25
Description	Homo sapiens miR-25 stem-loop
Stem-loop	(see below) Get sequence
Genome context	7: 99335834-99335917 [-] ENST00000343023; intron 8; ENST00000341047; intron 21; NP_8 ENST00000303887; intron 36; MCM ENST00000362082; intron 54; NP_8
Database links	EMBL AJ421746 HGNC 31609
Related entries	mmu-mir-25 rno-mir-25 dre-mir-25 ggo-mir-25 ptr-mir-25 mml-mir-25 lla-mir-25 mne-mir-25 Show details

```
   a ug   ag    g     uu g    u   --  ac
ggcc g  uug  aggc gagac  g gcaau gcu  gg  g
|||| |  |||  |||| |||||  | ||||| |||  ||  c
ccgg c  gac  ucug cucug  c cguua cgg  cc  u
    c gu   ag    g     uu a    -   gu  cg
```

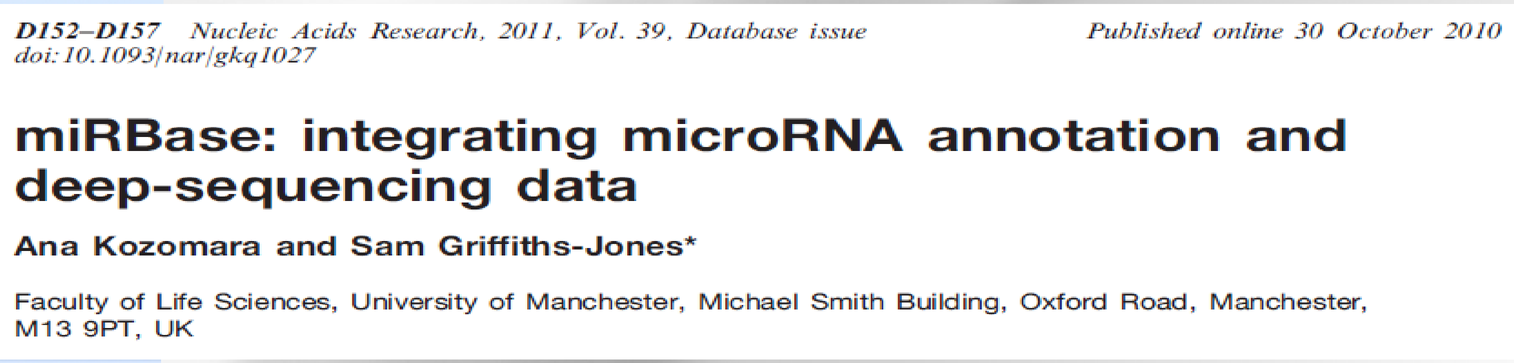
Mature sequence MIMAT0000081

Accession	MIMAT0000081
ID	hsa-miR-25
Sequence	52 - cauugcacuugucucggucuga - 73 Get sequence
Evidence	experimental; cloned [1-2], Northern [1]
Predicted targets	MIRBASE hsa-miR-25 PICTAR-VERT hsa-miR-25 TARGETSCAN miR-25/32/92/367

References

1 "Identification of novel genes coding for small expressed RNAs"
Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T
Science 294:853-858(2001).

2 "Altered expression profiles of microRNAs during TPA-induced differentiation of HL-60 cells"
Kasashima K, Nakamura Y, Kozu T
Biochem Biophys Res Commun 322:403-410(2004).



D152–D157 Nucleic Acids Research, 2011, Vol. 39, Database issue *Published online 30 October 2010*
doi:10.1093/nar/gkq1027

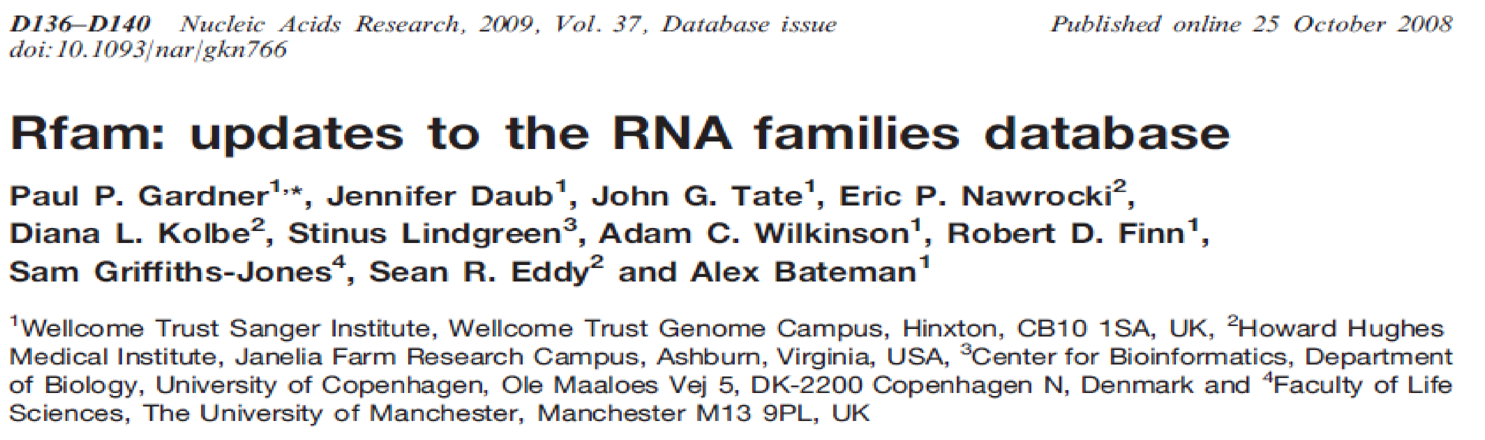
miRBase: integrating microRNA annotation and deep-sequencing data

Ana Kozomara and Sam Griffiths-Jones*

Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester, M13 9PT, UK

Annotation

- Useful databases:
  - miRbase (http://microrna.sanger.ac.uk/)
  - Rfam (http://rfam.sanger.ac.uk/)
    - A collection of RNA families
      - Rfam 10.1, June 2011, 1973 families
    - A track now included in the UCSC genome browser
    - Be careful: also contains (not all) miRNA families



D136–D140 Nucleic Acids Research, 2009, Vol. 37, Database issue *Published online 25 October 2008*
doi:10.1093/nar/gkn766

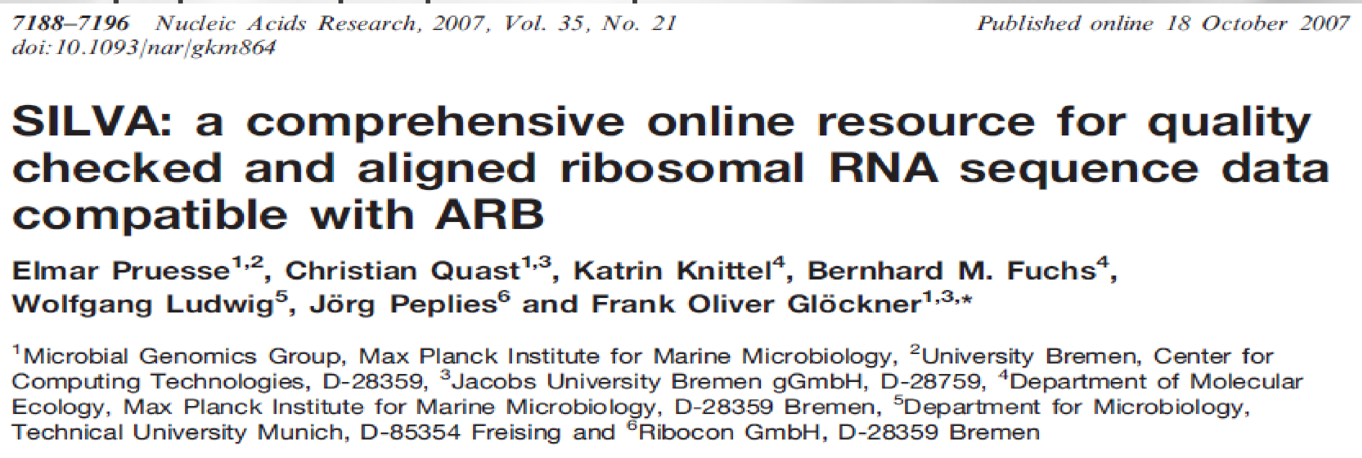
Rfam: updates to the RNA families database

Paul P. Gardner[1,*], Jennifer Daub[1], John G. Tate[1], Eric P. Nawrocki[2], Diana L. Kolbe[2], Stinus Lindgreen[3], Adam C. Wilkinson[1], Robert D. Finn[1], Sam Griffiths-Jones[4], Sean R. Eddy[2] and Alex Bateman[1]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK, [2]Howard Hughes Medical Institute, Janelia Farm Research Campus, Ashburn, Virginia, USA, [3]Center for Bioinformatics, Department of Biology, University of Copenhagen, Ole Maaloes Vej 5, DK-2200 Copenhagen N, Denmark and [4]Faculty of Life Sciences, The University of Manchester, Manchester M13 9PL, UK

Annotation

- Useful databases:
  - miRbase (http://microrna.sanger.ac.uk/)
  - Rfam (http://rfam.sanger.ac.uk/)
  - Silva (http://www.arb-silva.de/)
    - A comprehensive on-line resource for quality checked and aligned ribosomal RNA sequence data.
      - SSU (16S rRNA, 18S rRNA)
      - LSU (23S rRNA, 28S rRNA)



7188–7196 Nucleic Acids Research, 2007, Vol. 35, No. 21 *Published online 18 October 2007*
doi:10.1093/nar/gkm864

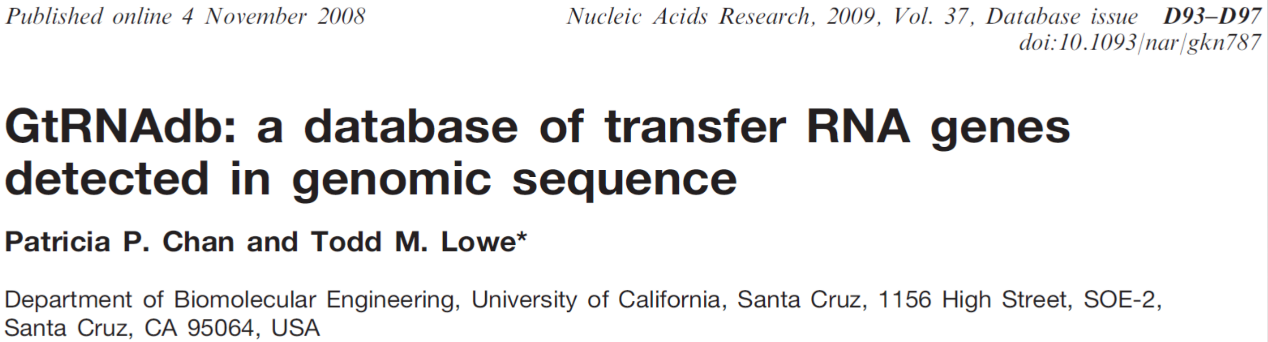
SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB

Elmar Pruesse[1,2], Christian Quast[1,3], Katrin Knittel[4], Bernhard M. Fuchs[4], Wolfgang Ludwig[5], Jörg Peplies[6] and Frank Oliver Glöckner[1,3,*]

[1]Microbial Genomics Group, Max Planck Institute for Marine Microbiology, [2]University Bremen, Center for Computing Technologies, D-28359, [3]Jacobs University Bremen gGmbH, D-28759, [4]Department of Molecular Ecology, Max Planck Institute for Marine Microbiology, D-28359 Bremen, [5]Department for Microbiology, Technical University Munich, D-85354 Freising and [6]Ribocon GmbH, D-28359 Bremen

Annotation

- Useful databases:
  - miRbase (http://microrna.sanger.ac.uk/) miRBase
  - Rfam (http://rfam.sanger.ac.uk/)
  - Silva (http://www.arb-silva.de/) silva comprehensive ribosomal RNA databases
  - GtRNAdb(http://gtrnadb.ucsc.edu/) GtRNAdb Genomic tRNA Database tRNAscan-SE analysis of complete genomes
    - Contains tRNA gene predictions made by the program tRNAscan-SE (Lowe & Eddy, Nucl Acids Res 25: 955-964, 1997) on complete or nearly complete genomes.
    - All annotation is automated and has not been inspected for agreement with published literature.



Published online 4 November 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D93–D97***
doi:10.1093/nar/gkn787

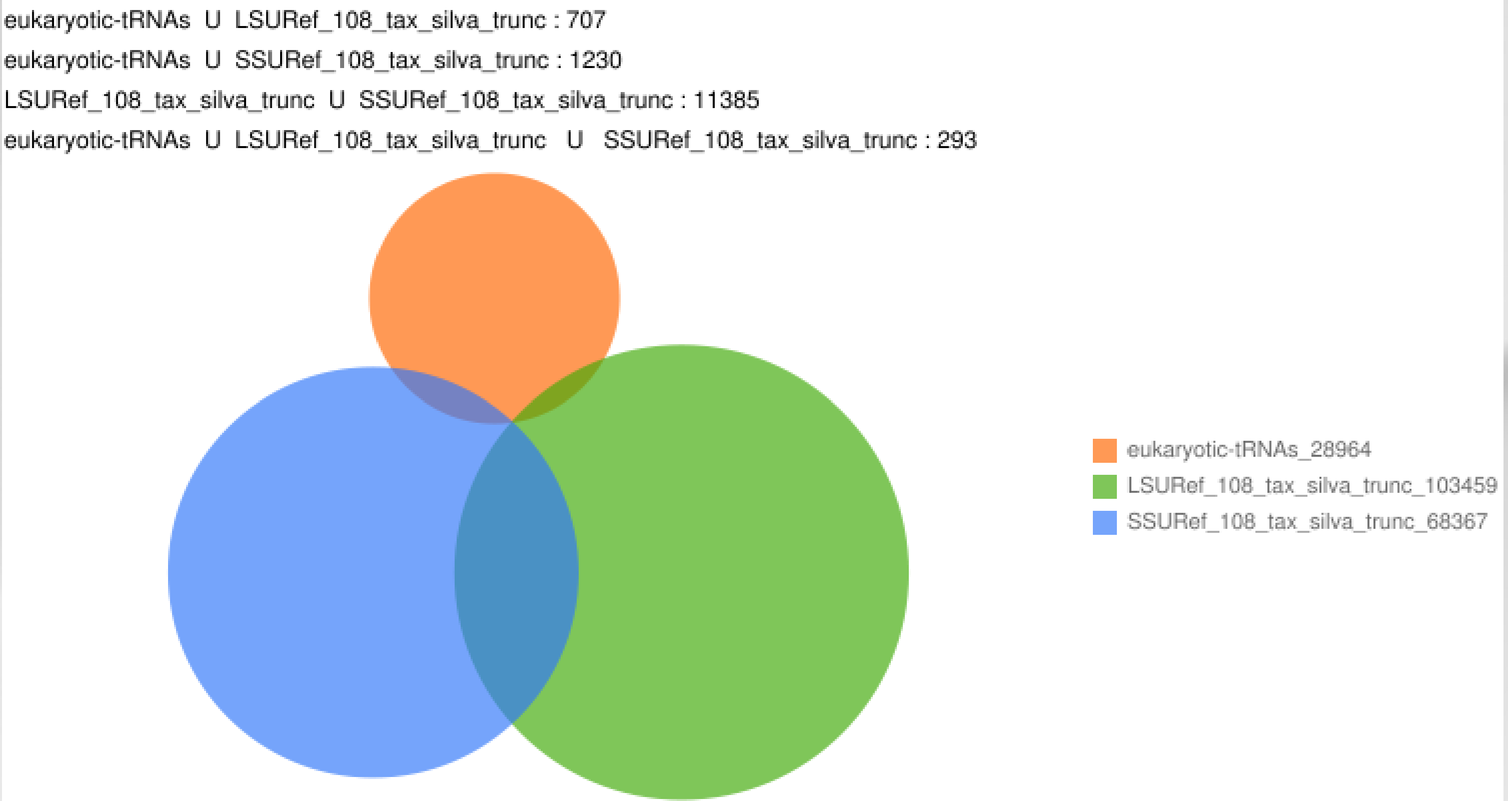
GtRNAdb: a database of transfer RNA genes detected in genomic sequence

Patricia P. Chan and Todd M. Lowe*

Department of Biomolecular Engineering, University of California, Santa Cruz, 1156 High Street, SOE-2, Santa Cruz, CA 95064, USA

Annotation

- **Reads with multiple annotation**



eukaryotic-tRNAs U LSURef_108_tax_silva_trunc : 707

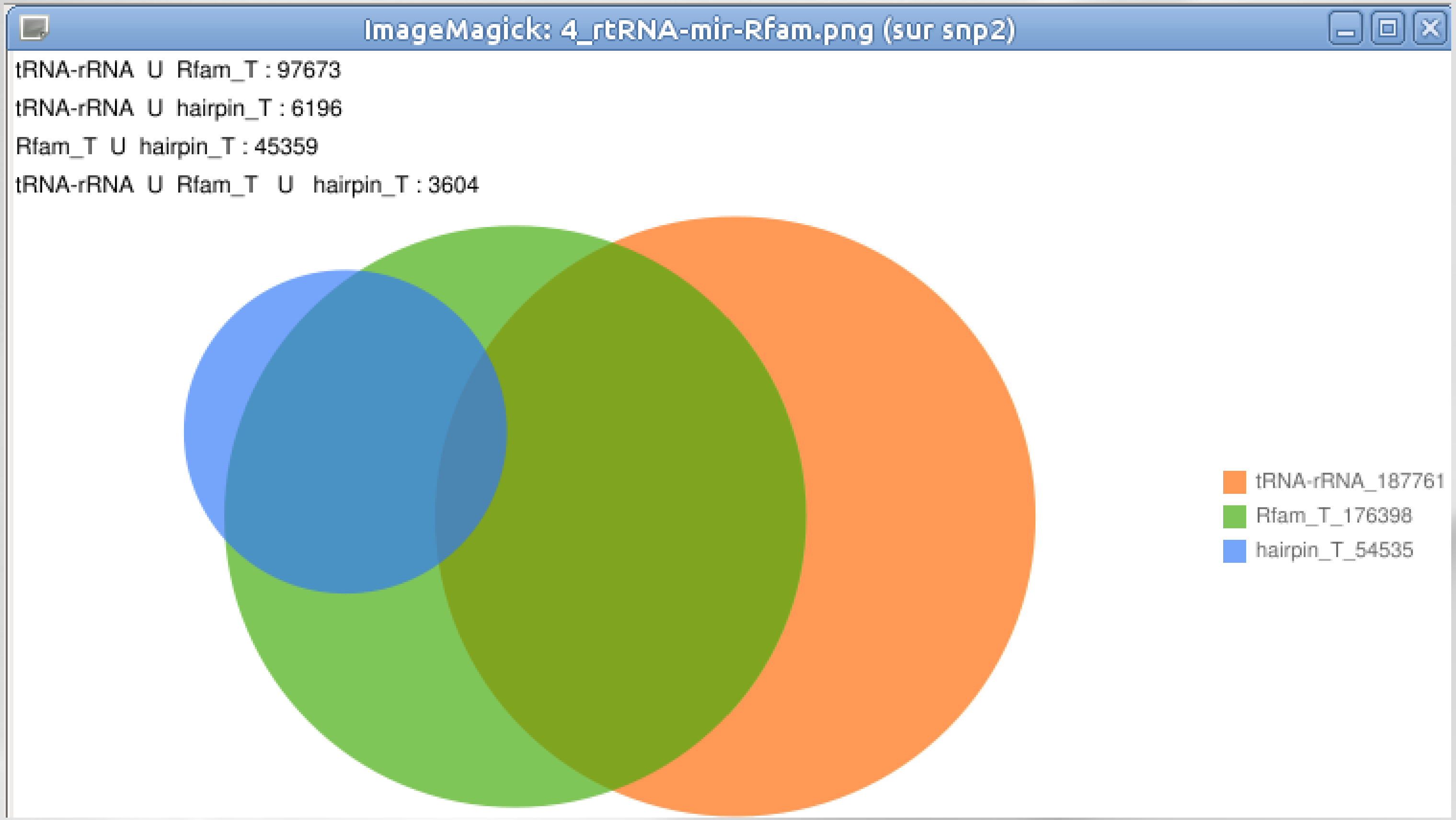
eukaryotic-tRNAs U SSURef_108_tax_silva_trunc : 1230

LSURef_108_tax_silva_trunc U SSURef_108_tax_silva_trunc : 11385

eukaryotic-tRNAs U LSURef_108_tax_silva_trunc U SSURef_108_tax_silva_trunc : 293

Annotation

- **Reads with multiple annotation**



ImageMagick: 4_rtRNA-mir-Rfam.png (sur snp2)

tRNA-rRNA U Rfam_T : 97673

tRNA-rRNA U hairpin_T : 6196

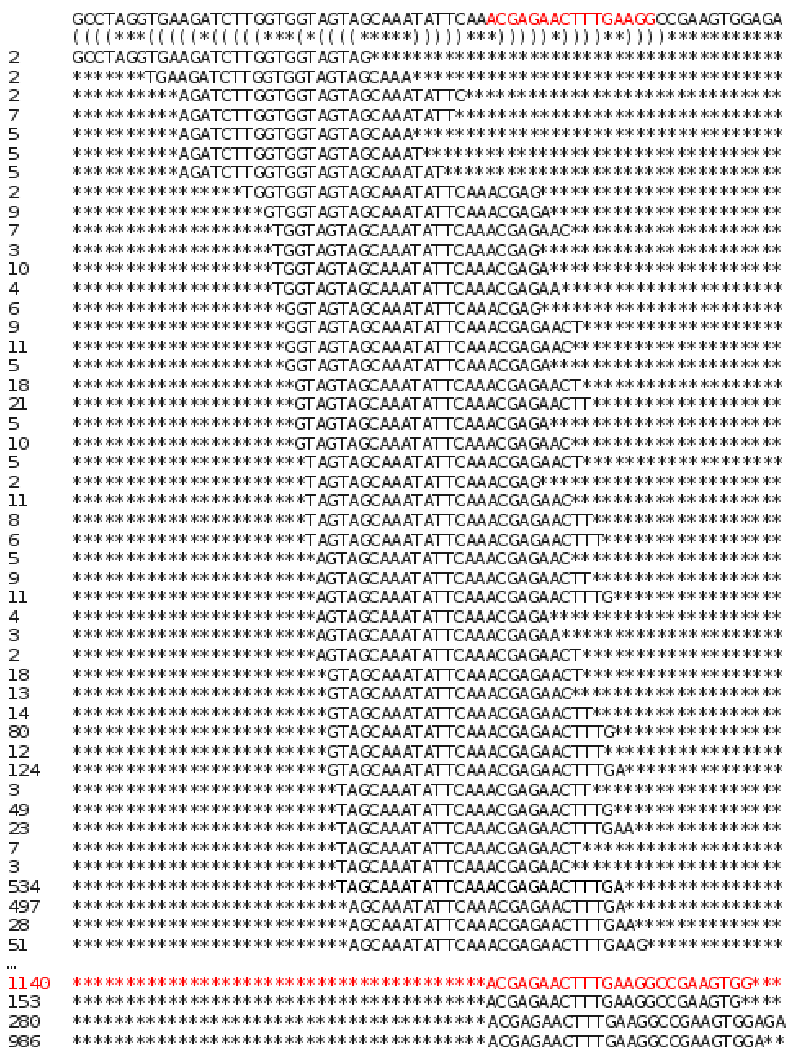
Rfam_T U hairpin_T : 45359

tRNA-rRNA U Rfam_T U hairpin_T : 3604

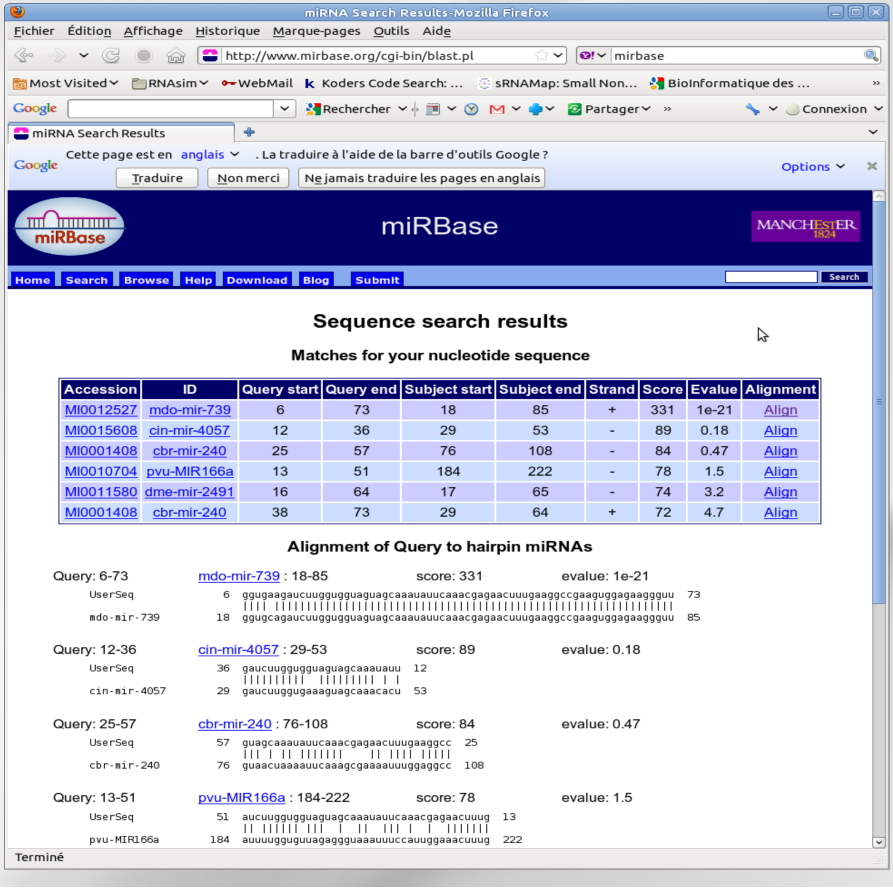
→ A lot of reads annotated with mirBase but also with tRNA and rRNA database

• rRNA present in miRBase

Mir-739 or **28S rRNA** ?



```
      GCCTAGGTGAAGATCTTGGTGGTAGTAGCAAATATTCAAACGAGAACTTTGAAGGCCGAAGTGGAGA
      ((((***(((((*(((((***(*((((*****)))))***)))))*))))**))))***********
2     GCCTAGGTGAAGATCTTGGTGGTAGTAG***************************************
2     *******TGAAGATCTTGGTGGTAGTAGCAAA***********************************
2     **********AGATCTTGGTGGTAGTAGCAAATATTC******************************
7     **********AGATCTTGGTGGTAGTAGCAAATATT*******************************
5     **********AGATCTTGGTGGTAGTAGCAAA***********************************
5     **********AGATCTTGGTGGTAGTAGCAAAT**********************************
5     **********AGATCTTGGTGGTAGTAGCAAATAT********************************
2     ****************TGGTGGTAGTAGCAAATATTCAAACGAG***********************
9     ******************GTGGTAGTAGCAAATATTCAAACGAGA**********************
7     *******************TGGTAGTAGCAAATATTCAAACGAGAAC********************
3     *******************TGGTAGTAGCAAATATTCAAACGAG***********************
10    *******************TGGTAGTAGCAAATATTCAAACGAGA**********************
4     *******************TGGTAGTAGCAAATATTCAAACGAGAA*********************
6     ********************GGTAGTAGCAAATATTCAAACGAG***********************
9     ********************GGTAGTAGCAAATATTCAAACGAGAACT*******************
11    ********************GGTAGTAGCAAATATTCAAACGAGAAC********************
5     ********************GGTAGTAGCAAATATTCAAACGAGA**********************
18    *********************GTAGTAGCAAATATTCAAACGAGAACT*******************
21    *********************GTAGTAGCAAATATTCAAACGAGAACTT******************
5     *********************GTAGTAGCAAATATTCAAACGAGA**********************
10    *********************GTAGTAGCAAATATTCAAACGAGAAC********************
5     **********************TAGTAGCAAATATTCAAACGAGAACT*******************
2     **********************TAGTAGCAAATATTCAAACGAG***********************
11    **********************TAGTAGCAAATATTCAAACGAGAAC********************
8     **********************TAGTAGCAAATATTCAAACGAGAACTT******************
6     **********************TAGTAGCAAATATTCAAACGAGAACTTT*****************
5     ***********************AGTAGCAAATATTCAAACGAGAAC********************
9     ***********************AGTAGCAAATATTCAAACGAGAACTT******************
11    ***********************AGTAGCAAATATTCAAACGAGAACTTTG****************
4     ***********************AGTAGCAAATATTCAAACGAGA**********************
3     ***********************AGTAGCAAATATTCAAACGAGAA*********************
2     ***********************AGTAGCAAATATTCAAACGAGAACT*******************
18    ************************GTAGCAAATATTCAAACGAGAACT*******************
13    ************************GTAGCAAATATTCAAACGAGAAC********************
14    ************************GTAGCAAATATTCAAACGAGAACTT******************
80    ************************GTAGCAAATATTCAAACGAGAACTTTG****************
12    ************************GTAGCAAATATTCAAACGAGAACTTT*****************
124   ************************GTAGCAAATATTCAAACGAGAACTTTGA***************
3     *************************TAGCAAATATTCAAACGAGAACTT******************
49    *************************TAGCAAATATTCAAACGAGAACTTTG****************
23    *************************TAGCAAATATTCAAACGAGAACTTTGAA**************
7     *************************TAGCAAATATTCAAACGAGAACT*******************
3     *************************TAGCAAATATTCAAACGAGAAC********************
534   *************************TAGCAAATATTCAAACGAGAACTTTGA***************
497   **************************AGCAAATATTCAAACGAGAACTTTGA***************
28    **************************AGCAAATATTCAAACGAGAACTTTGAA**************
51    **************************AGCAAATATTCAAACGAGAACTTTGAAG*************
…
1140  ***************************************ACGAGAACTTTGAAGGCCGAAGTGG***
153   ***************************************ACGAGAACTTTGAAGGCCGAAGTG****
280   ***************************************ACGAGAACTTTGAAGGCCGAAGTGGAGA
986   ***************************************ACGAGAACTTTGAAGGCCGAAGTGGA**
```



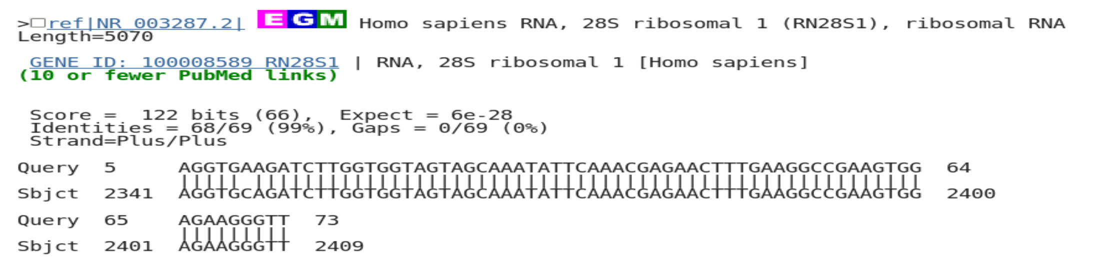
Sequence search results

Matches for your nucleotide sequence

Accession	ID	Query start	Query end	Subject start	Subject end	Strand	Score	Evalue	Alignment
MI0012527	mdo-mir-739	6	73	18	85	+	331	1e-21	Align
MI0015608	cin-mir-4057	12	36	29	53	-	89	0.18	Align
MI0001408	cbr-mir-240	25	57	76	108	-	84	0.47	Align
MI0010704	pvu-MIR166a	13	51	184	222	-	78	1.5	Align
MI0011580	dme-mir-2491	16	64	17	65	-	74	3.2	Align
MI0001408	cbr-mir-240	38	73	29	64	+	72	4.7	Align

Alignment of Query to hairpin miRNAs

```
Query: 6-73          mdo-mir-739 : 18-85          score: 331          evalue: 1e-21
     UserSeq          6  ggugaagaucuuggugguaguagcaaauauucaaacgagaacuuugaaggccgaaguggagaaggguu  73
                         |||| |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
     mdo-mir-739     18  ggugcagaucuuggugguaguagcaaauauucaaacgagaacuuugaaggccgaaguggagaaggguu  85

Query: 12-36         cin-mir-4057 : 29-53         score: 89           evalue: 0.18
     UserSeq         36  gaucuuggugguaguagcaaauauu  12
                         ||||||||||  ||||||||| | |
     cin-mir-4057    29  gaucuuggugaaaguagcaaacacu  53

Query: 25-57         cbr-mir-240 : 76-108         score: 84           evalue: 0.47
     UserSeq         57  guagcaaauauucaaacgagaacuuugaaggcc  25
                         ||| | || |||||||    || |||| |||||
     cbr-mir-240     76  guaacuaaaauucaaagcgaaaauuuggaggcc  108

Query: 13-51         pvu-MIR166a : 184-222        score: 78           evalue: 1.5
     UserSeq         51  aucuuggugguaguagcaaauauucaaacgagaacuuug  13
                         || |||||| |||  |  ||  ||| |  |  |||||||
     pvu-MIR166a    184  auuuugguguuagagggtuaaauuuccauuggaaacuuug  222
```

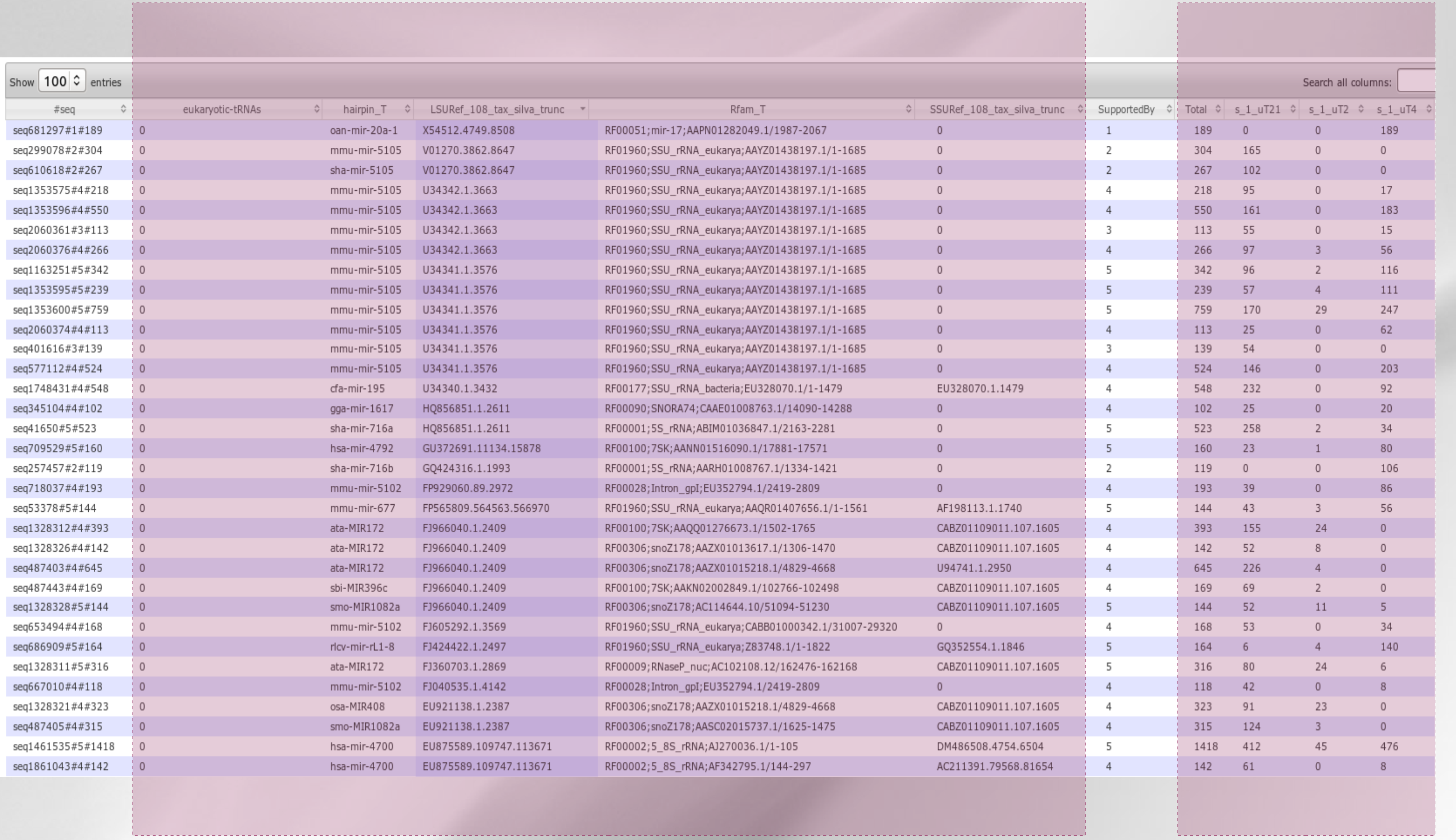


```
>□ref|NR_003287.2| EGM Homo sapiens RNA, 28S ribosomal 1 (RN28S1), ribosomal RNA
Length=5070

 GENE ID: 100008589 RN28S1 | RNA, 28S ribosomal 1 [Homo sapiens]
(10 or fewer PubMed links)

 Score =  122 bits (66),  Expect = 6e-28
 Identities = 68/69 (99%), Gaps = 0/69 (0%)
 Strand=Plus/Plus

Query  5     AGGTGAAGATCTTGGTGGTAGTAGCAAATATTCAAACGAGAACTTTGAAGGCCGAAGTGG  64
             ||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2341  AGGTGCAGATCTTGGTGGTAGTAGCAAATATTCAAACGAGAACTTTGAAGGCCGAAGTGG  2400

Query  65    AGAAGGGTT  73
             |||||||||
Sbjct  2401  AGAAGGGTT  2409
```

Annotation

Annotation

occurences



Show 100 entries

Search all columns:

#seq	eukaryotic-tRNAs	hairpin_T	LSURef_108_tax_silva_trunc	Rfam_T	SSURef_108_tax_silva_trunc	SupportedBy	Total	s_1_uT21	s_1_uT2	s_1_uT4
seq681297#1#189	0	oan-mir-20a-1	X54512.4749.8508	RF00051;mir-17;AAPN01282049.1/1987-2067	0	1	189	0	0	189
seq299078#2#304	0	mmu-mir-5105	V01270.3862.8647	RF01960;SSU_rRNA_eukarya;AAYZ01438197.1/1-1685	0	2	304	165	0	0
seq610618#2#267	0	sha-mir-5105	V01270.3862.8647	RF01960;SSU_rRNA_eukarya;AAYZ01438197.1/1-1685	0	2	267	102	0	0
seq1353575#4#218	0	mmu-mir-5105	U34342.1.3663	RF01960;SSU_rRNA_eukarya;AAYZ01438197.1/1-1685	0	4	218	95	0	17
seq1353596#4#550	0	mmu-mir-5105	U34342.1.3663	RF01960;SSU_rRNA_eukarya;AAYZ01438197.1/1-1685	0	4	550	161	0	183
seq2060361#3#113	0	mmu-mir-5105	U34342.1.3663	RF01960;SSU_rRNA_eukarya;AAYZ01438197.1/1-1685	0	3	113	55	0	15
seq2060376#4#266	0	mmu-mir-5105	U34342.1.3663	RF01960;SSU_rRNA_eukarya;AAYZ01438197.1/1-1685	0	4	266	97	3	56
seq1163251#5#342	0	mmu-mir-5105	U34341.1.3576	RF01960;SSU_rRNA_eukarya;AAYZ01438197.1/1-1685	0	5	342	96	2	116
seq1353595#5#239	0	mmu-mir-5105	U34341.1.3576	RF01960;SSU_rRNA_eukarya;AAYZ01438197.1/1-1685	0	5	239	57	4	111
seq1353600#5#759	0	mmu-mir-5105	U34341.1.3576	RF01960;SSU_rRNA_eukarya;AAYZ01438197.1/1-1685	0	5	759	170	29	247
seq2060374#4#113	0	mmu-mir-5105	U34341.1.3576	RF01960;SSU_rRNA_eukarya;AAYZ01438197.1/1-1685	0	4	113	25	0	62
seq401616#3#139	0	mmu-mir-5105	U34341.1.3576	RF01960;SSU_rRNA_eukarya;AAYZ01438197.1/1-1685	0	3	139	54	0	0
seq577112#4#524	0	mmu-mir-5105	U34341.1.3576	RF01960;SSU_rRNA_eukarya;AAYZ01438197.1/1-1685	0	4	524	146	0	203
seq1748431#4#548	0	cfa-mir-195	U34340.1.3432	RF00177;SSU_rRNA_bacteria;EU328070.1/1-1479	EU328070.1.1479	4	548	232	0	92
seq345104#4#102	0	gga-mir-1617	HQ856851.1.2611	RF00090;SNORA74;CAAE01008763.1/14090-14288	0	4	102	25	0	20
seq41650#5#523	0	sha-mir-716a	HQ856851.1.2611	RF00001;5S_rRNA;ABIM01036847.1/2163-2281	0	5	523	258	2	34
seq709529#5#160	0	hsa-mir-4792	GU372691.11134.15878	RF00100;7SK;AANN01516090.1/17881-17571	0	5	160	23	1	80
seq257457#2#119	0	sha-mir-716b	GQ424316.1.1993	RF00001;5S_rRNA;AARH01008767.1/1334-1421	0	2	119	0	0	106
seq718037#4#193	0	mmu-mir-5102	FP929060.89.2972	RF00028;Intron_gpI;EU352794.1/2419-2809	0	4	193	39	0	86
seq53378#5#144	0	mmu-mir-677	FP565809.564563.566970	RF01960;SSU_rRNA_eukarya;AAQR01407656.1/1-1561	AF198113.1.1740	5	144	43	3	56
seq1328312#4#393	0	ata-MIR172	FJ966040.1.2409	RF00100;7SK;AAQQ01276673.1/1502-1765	CABZ01109011.107.1605	4	393	155	24	0
seq1328326#4#142	0	ata-MIR172	FJ966040.1.2409	RF00306;snoZ178;AAZX01013617.1/1306-1470	CABZ01109011.107.1605	4	142	52	8	0
seq487403#4#645	0	ata-MIR172	FJ966040.1.2409	RF00306;snoZ178;AAZX01015218.1/4829-4668	U94741.1.2950	4	645	226	4	0
seq487443#4#169	0	sbi-MIR396c	FJ966040.1.2409	RF00100;7SK;AAKN02002849.1/102766-102498	CABZ01109011.107.1605	4	169	69	2	0
seq1328328#5#144	0	smo-MIR1082a	FJ966040.1.2409	RF00306;snoZ178;AC114644.10/51094-51230	CABZ01109011.107.1605	5	144	52	11	5
seq653494#4#168	0	mmu-mir-5102	FJ605292.1.3569	RF01960;SSU_rRNA_eukarya;CABB01000342.1/31007-29320	0	4	168	53	0	34
seq686909#5#164	0	rlcv-mir-rL1-8	FJ424422.1.2497	RF01960;SSU_rRNA_eukarya;Z83748.1/1-1822	GQ352554.1.1846	5	164	6	4	140
seq1328311#5#316	0	ata-MIR172	FJ360703.1.2869	RF00009;RNaseP_nuc;AC102108.12/162476-162168	CABZ01109011.107.1605	5	316	80	24	6
seq667010#4#118	0	mmu-mir-5102	FJ040535.1.4142	RF00028;Intron_gpI;EU352794.1/2419-2809	0	4	118	42	0	8
seq1328321#4#323	0	osa-MIR408	EU921138.1.2387	RF00306;snoZ178;AAZX01015218.1/4829-4668	CABZ01109011.107.1605	4	323	91	23	0
seq487405#4#315	0	smo-MIR1082a	EU921138.1.2387	RF00306;snoZ178;AASC02015737.1/1625-1475	CABZ01109011.107.1605	4	315	124	3	0
seq1461535#5#1418	0	hsa-mir-4700	EU875589.109747.113671	RF00002;5_8S_rRNA;AJ270036.1/1-105	DM486508.4754.6504	5	1418	412	45	476
seq1861043#4#142	0	hsa-mir-4700	EU875589.109747.113671	RF00002;5_8S_rRNA;AF342795.1/144-297	AC211391.79568.81654	4	142	61	0	8

Exercice:

– Annotation