Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

small RNAseq data analysis

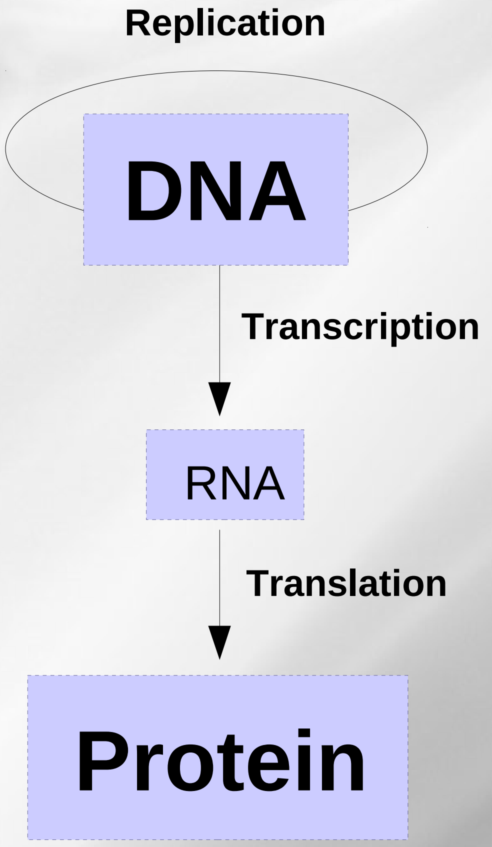
Philippe Bardou, Christine Gaspin,
Jérôme Mariette & Olivier Rué

Introduction to miRNA world and sRNAseq

Central dogma of molecular biology

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

  DNA structure descovery.

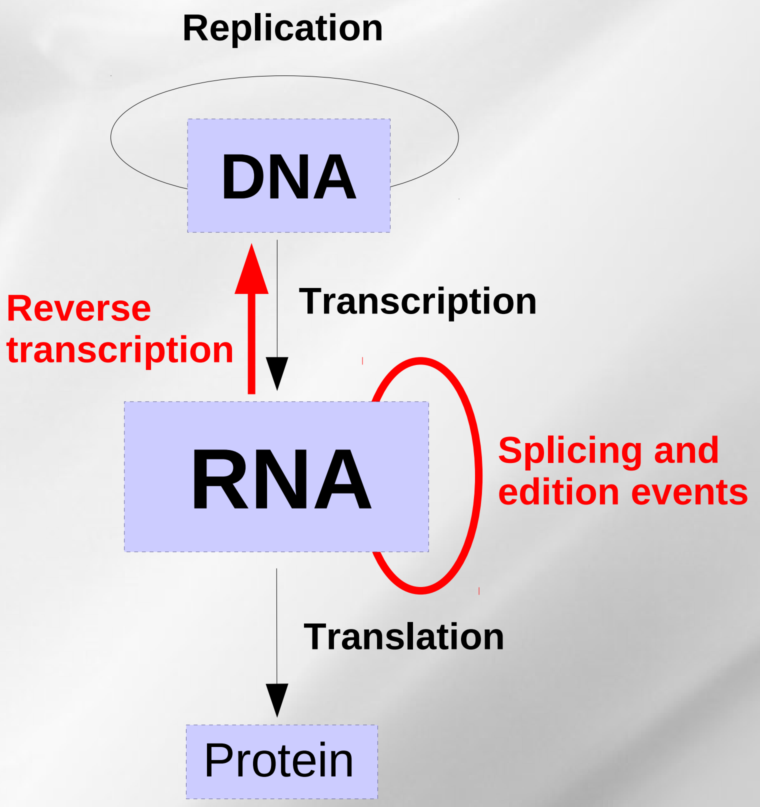


One gene = one function

Central dogma of molecular biology

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

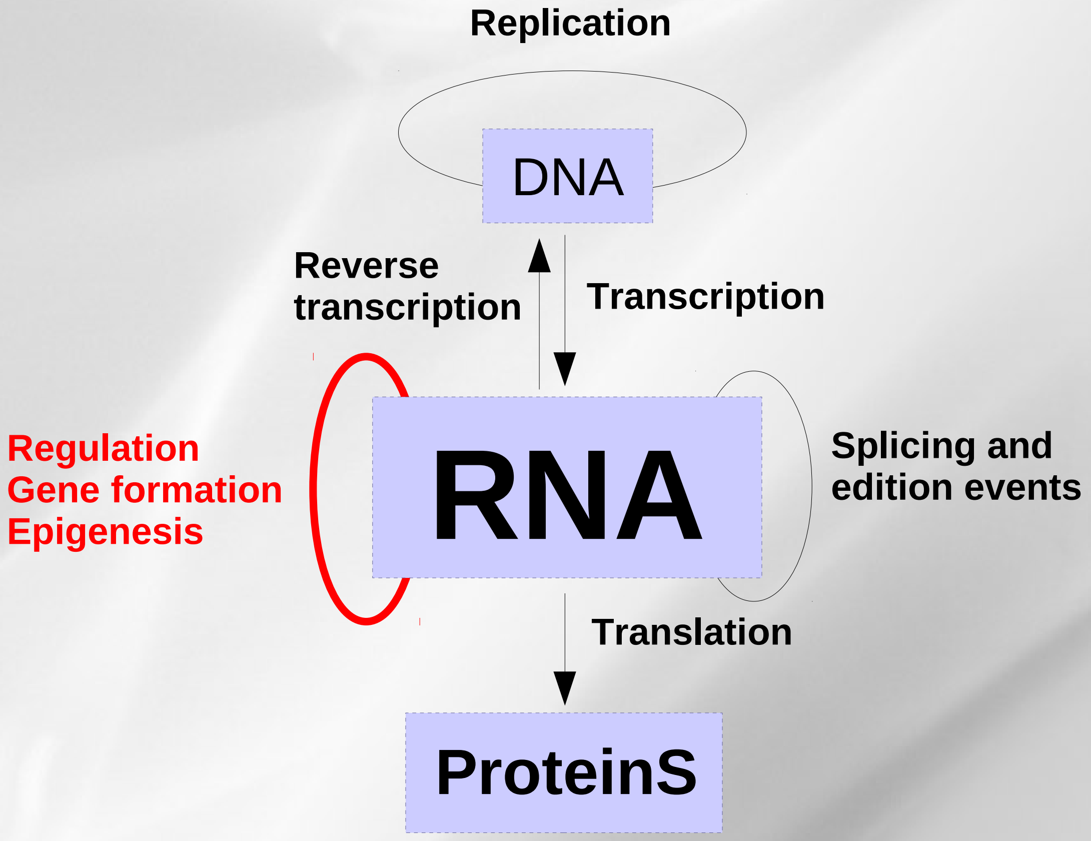
Genome analysis



Central dogma of molecular biology

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

Genome analysis + Sequencing

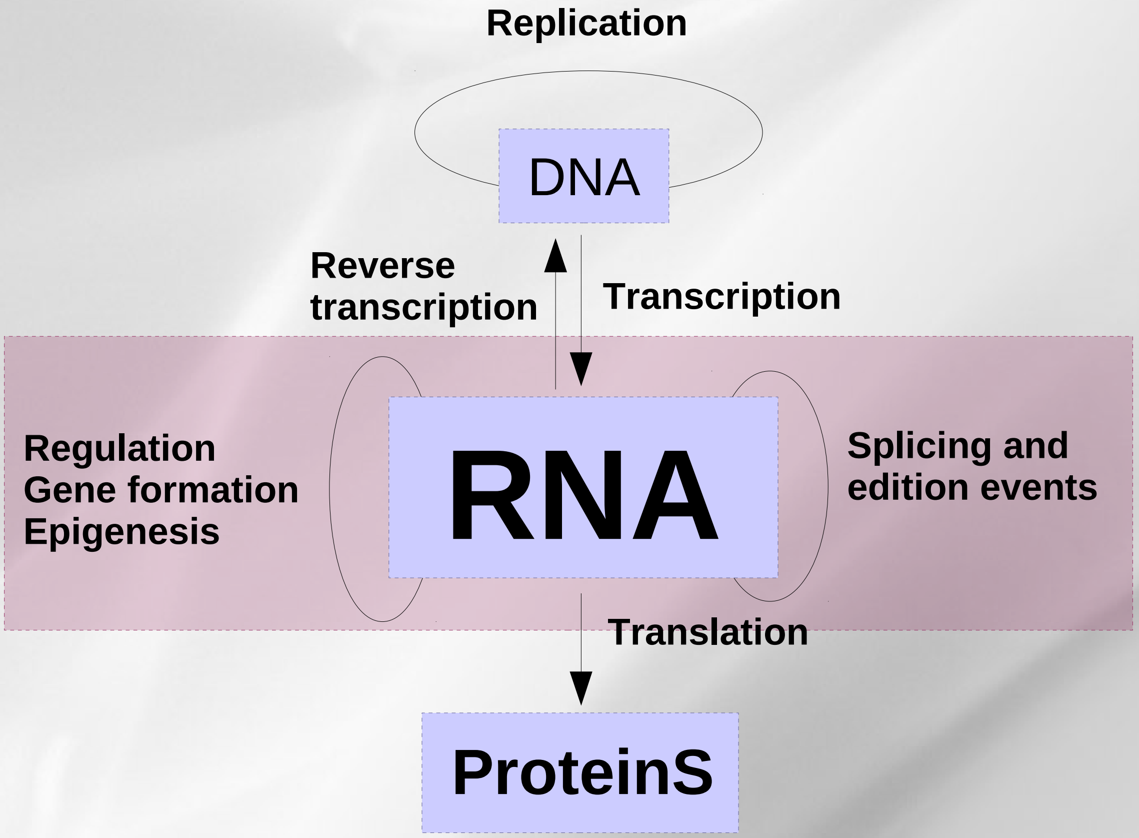


Many genes = one functionnel complex

Central dogma of molecular biology

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

Genome analysis + Sequencing

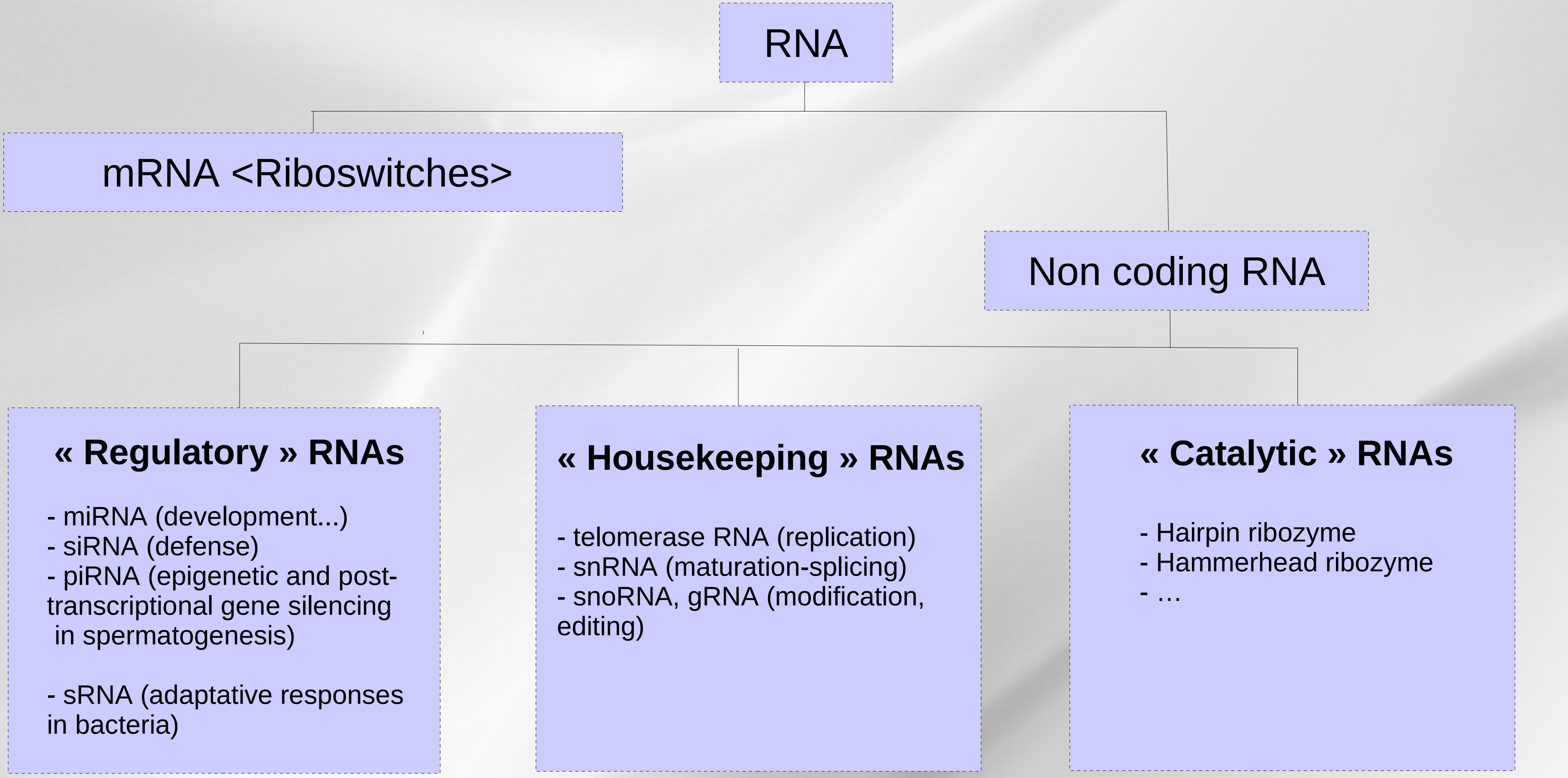


Many genes = one functionnel complex

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)

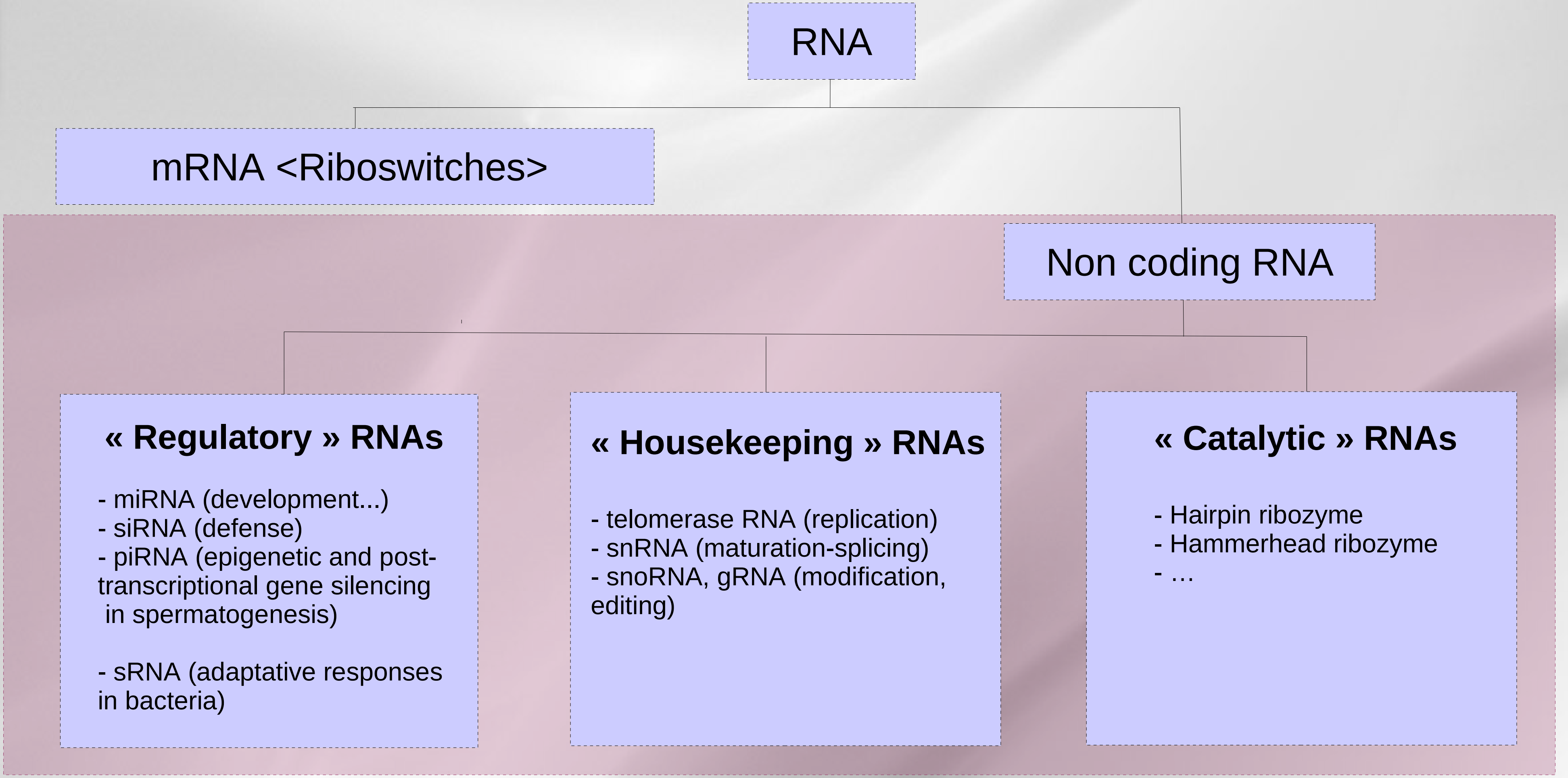
« Catalytic » RNAs

- 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)

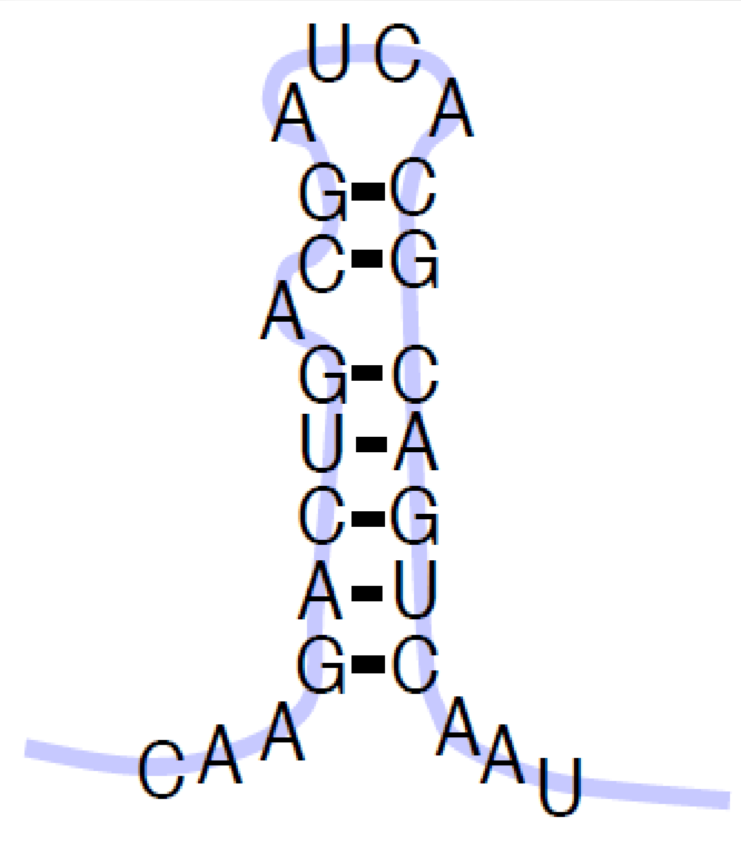
« Catalytic » RNAs

- Hairpin ribozyme
- Hammerhead ribozyme
- …

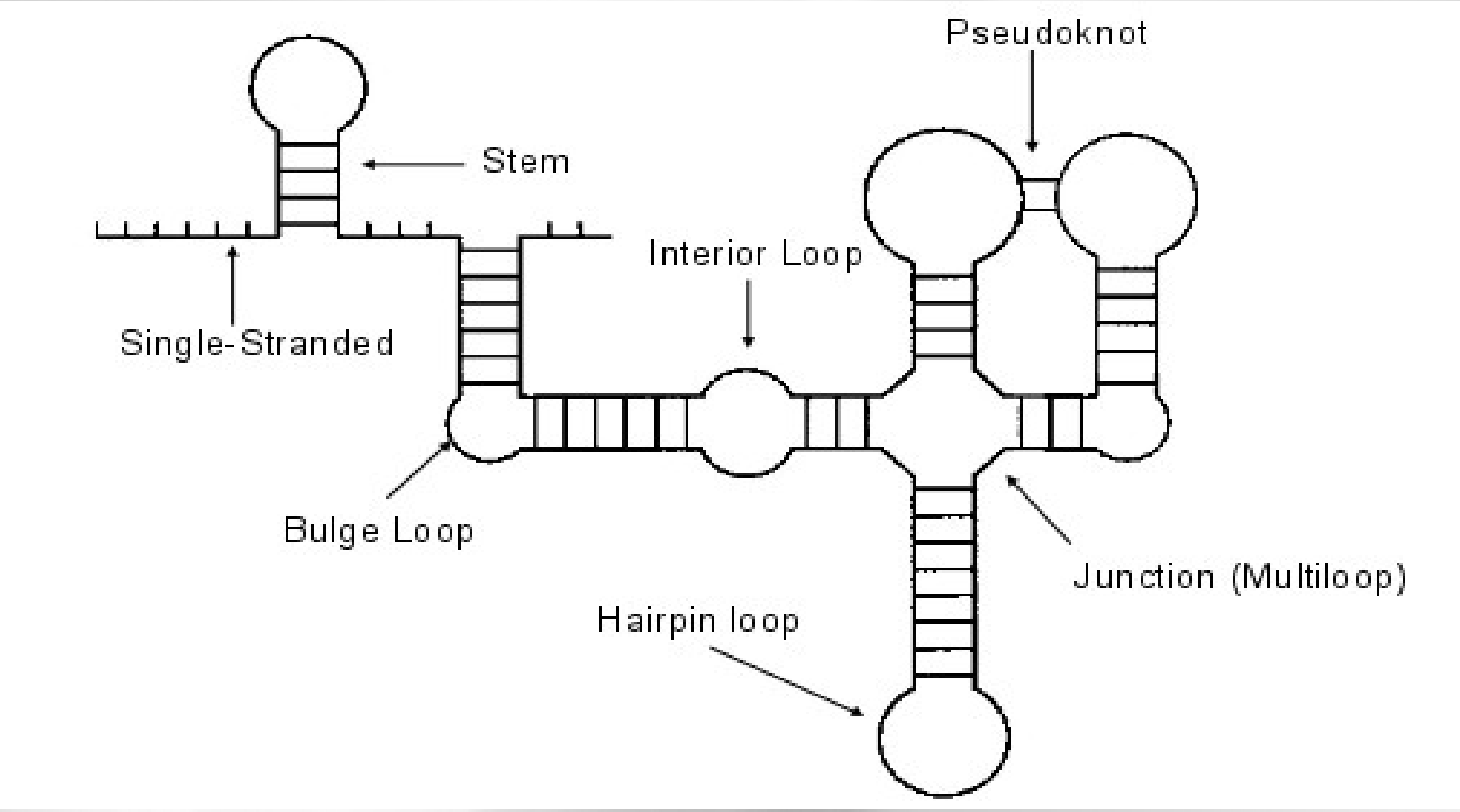
→ **Multiple roles of RNA in genes regulation**

RNA background

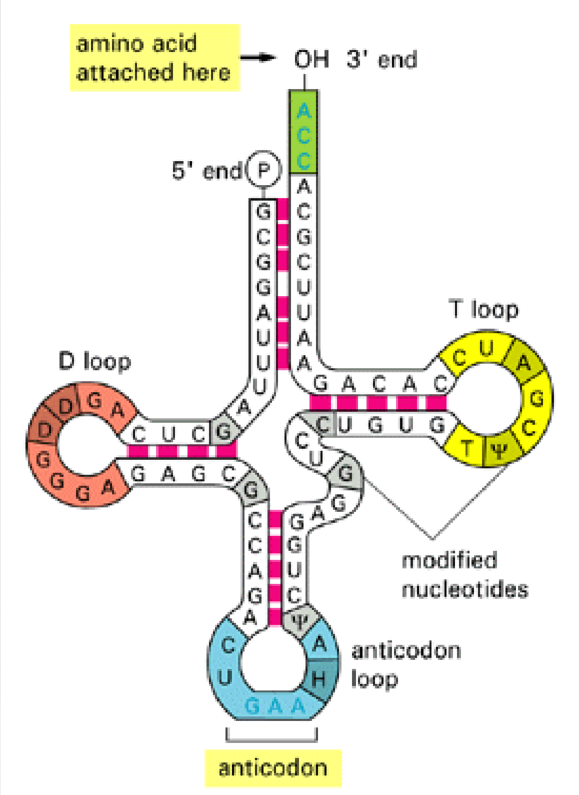
- RNA folds on itself by base pairing :
  - A with U : A-U, U-A
  - C with G : G-C, C-G
  - Sometimes G with U : U-G, G-U
- Folding = Secondary structure
- Structure related to function : ncRNA of the same family have a conserved structure
- Sequence less conserved

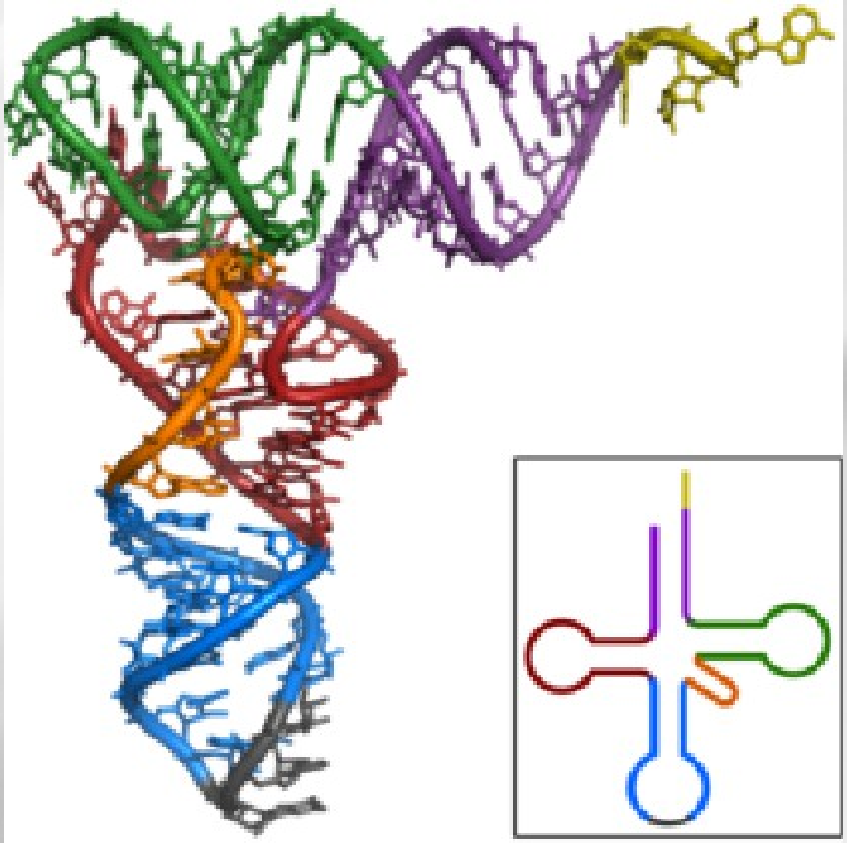


RNA background
Different elementary motifs



RNA background
Example: tRNA structure





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

The non coding protein RNA world

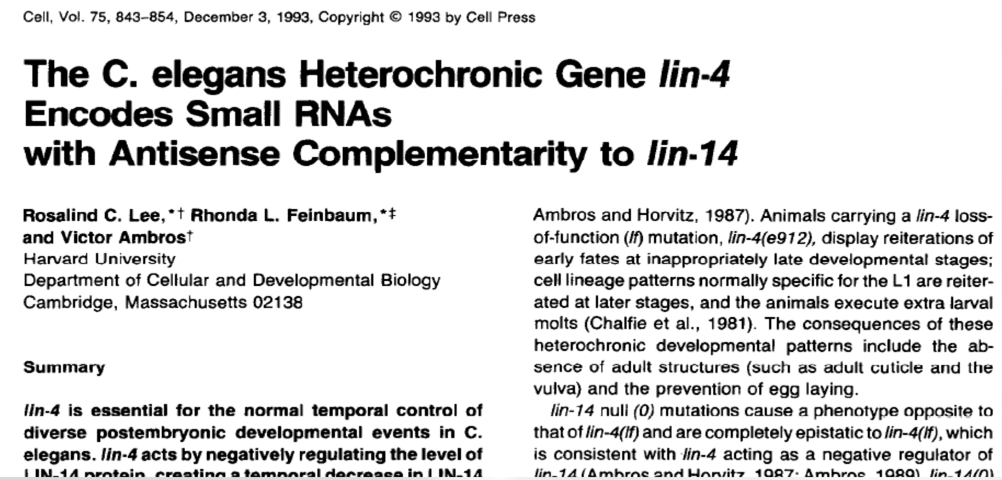
- **Large non coding protein RNA**
  - >300 nt
  - rRNA, tRNA, Xist, H19, ...
  - Genome structure & expression
- **Small non coding protein RNA**
  - >30 nt
  - snoRNA, snRNA...
  - mRNA maturation, translation
- **Micro non coding protein RNA**
  - 18-30 nt
  - miRNA, hc-siRNA, ta-siRNA, nat-siRNA,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

The miRNA world

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



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*

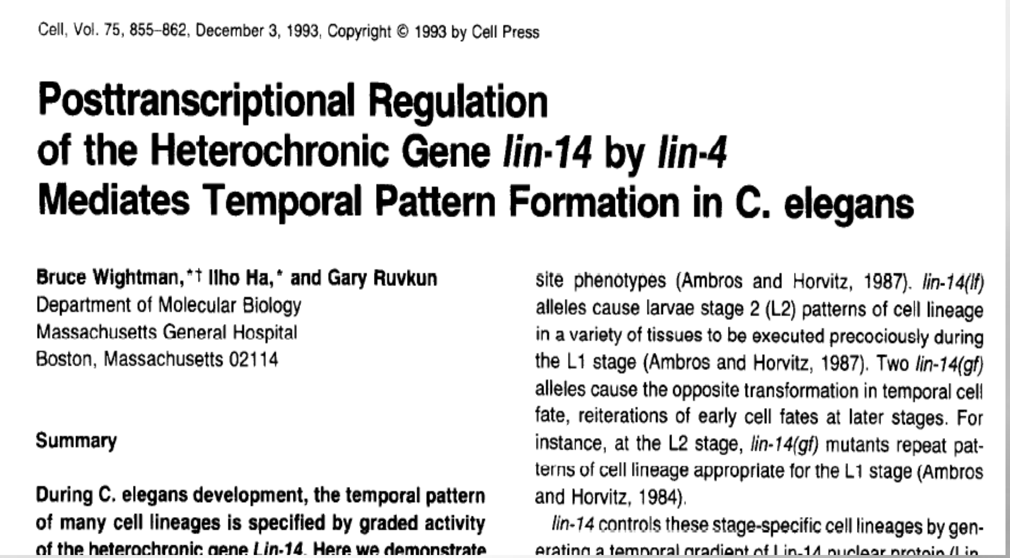
Rosalind C. Lee,*† Rhonda L. Feinbaum,*‡ and Victor Ambros†
Harvard University
Department of Cellular and Developmental Biology
Cambridge, Massachusetts 02138

Summary

***lin-4* is essential for the normal temporal control of diverse postembryonic developmental events in C. elegans. *lin-4* acts by negatively regulating the level of LIN-14 protein, creating a temporal decrease in LIN-14**

Ambros and Horvitz, 1987). Animals carrying a *lin-4* loss-of-function (*lf*) mutation, *lin-4(e912)*, display reiterations of early fates at inappropriately late developmental stages; cell lineage patterns normally specific for the L1 are reiterated at later stages, and the animals execute extra larval molts (Chalfie et al., 1981). The consequences of these heterochronic developmental patterns include the absence of adult structures (such as adult cuticle and the vulva) and the prevention of egg laying.

lin-14 null *(0)* mutations cause a phenotype opposite to that of *lin-4(lf)* and are completely epistatic to *lin-4(lf)*, which is consistent with *lin-4* acting as a negative regulator of *lin-14* (Ambros and Horvitz, 1987; Ambros, 1989). *lin-14(0)*



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

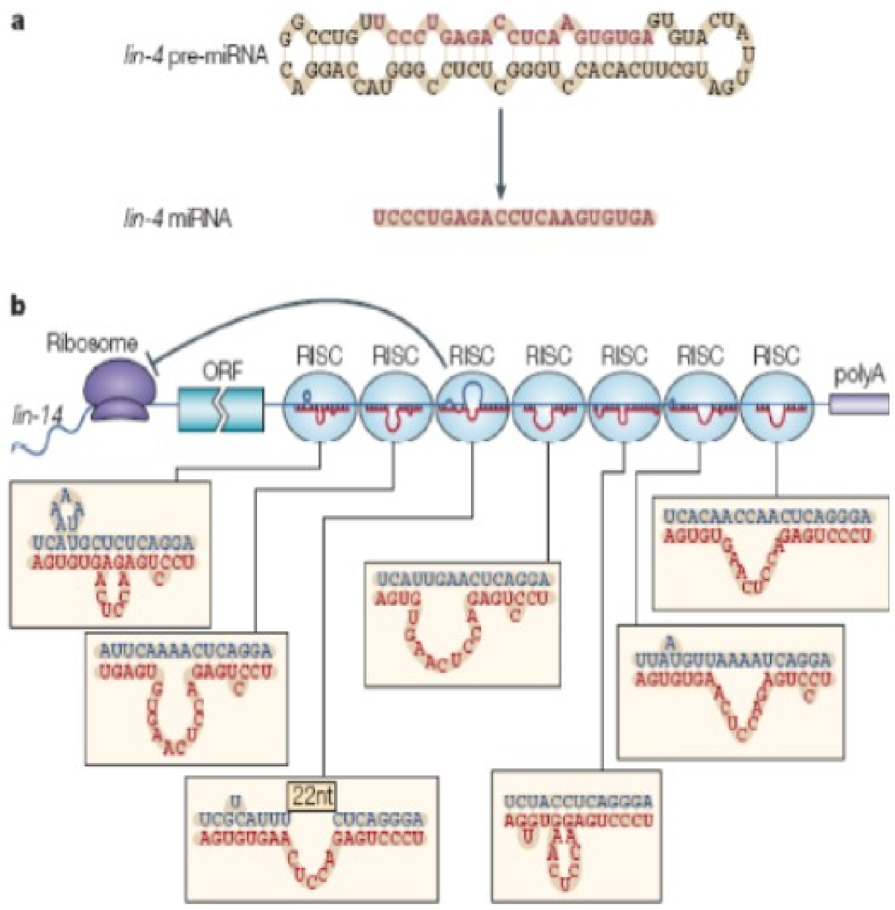
Bruce Wightman,*† Ilho Ha,* and Gary Ruvkun
Department of Molecular Biology
Massachusetts General Hospital
Boston, Massachusetts 02114

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).

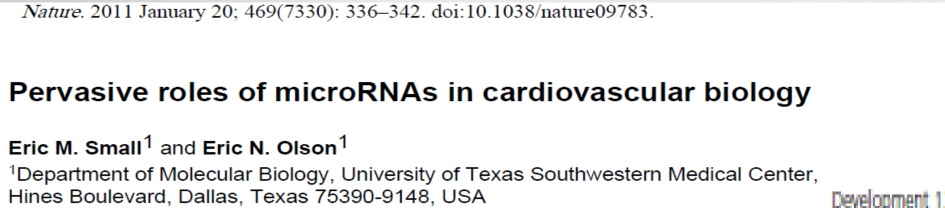
lin-14 controls these stage-specific cell lineages by generating a temporal gradient of Lin-14 nuclear protein (Lin-



(He & Hannon, Nature reviews, 2004)

The miRNA world

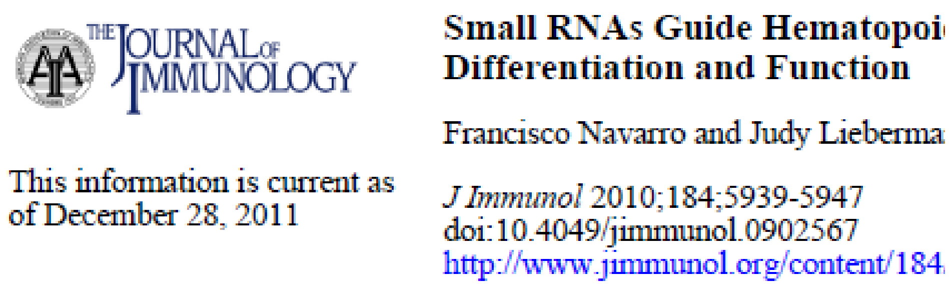
- A key regulation function



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



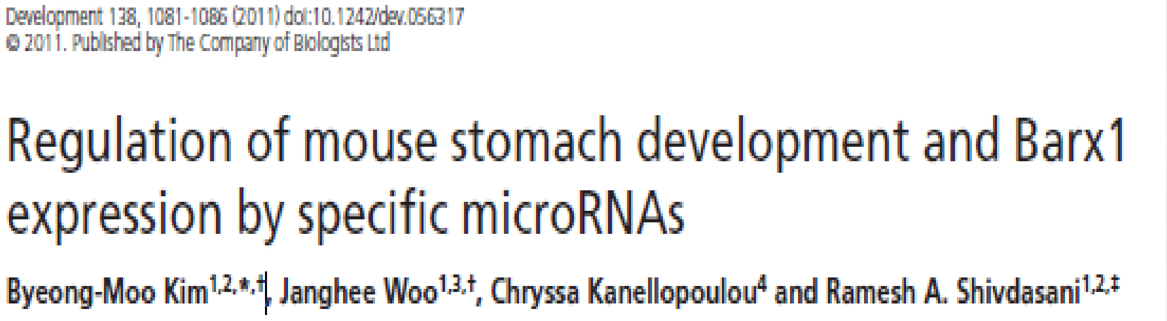
THE JOURNAL OF IMMUNOLOGY

Small RNAs Guide Hematopoi
Differentiation and Function

Francisco Navarro and Judy Lieberma

This information is current as of December 28, 2011

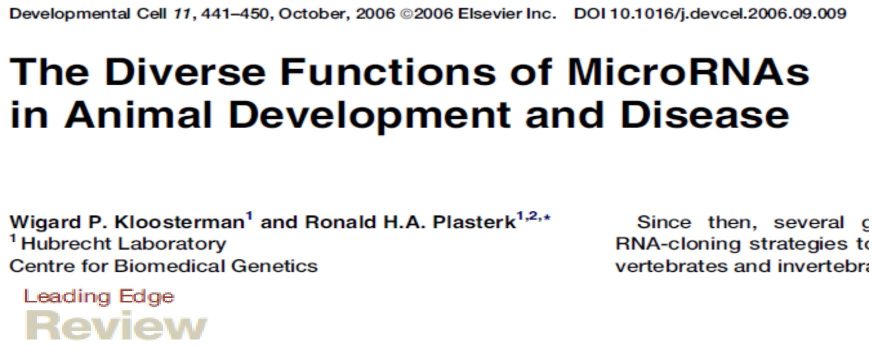
J Immunol 2010;184;5939-5947
doi:10.4049/jimmunol.0902567
http://www.jimmunol.org/content/184



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,‡]



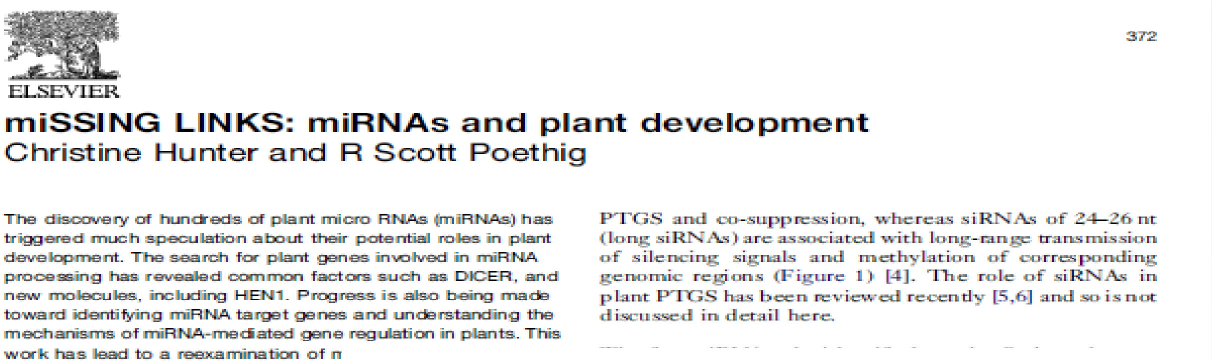
Developmental Cell *11*, **441–450, October, 2006 ©2006 Elsevier Inc. DOI 10.1016/j.devcel.2006.09.009**

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 t
vertebrates and invertebra

Leading Edge
Review



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

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.

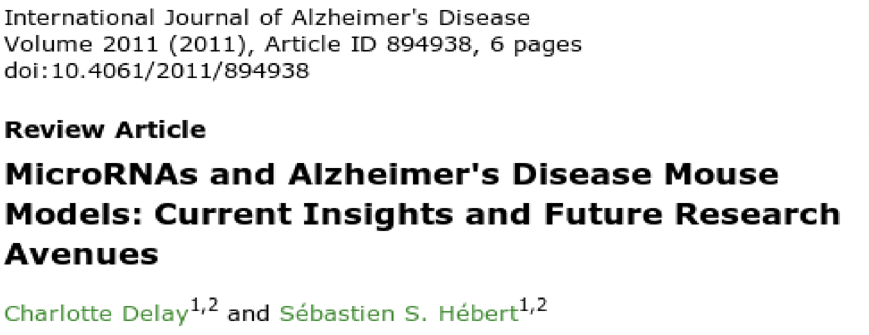
Addresses
Plant Science Institute, Department of Biol
Pennsylvania, Philadelphia, Pennsylvania 1

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

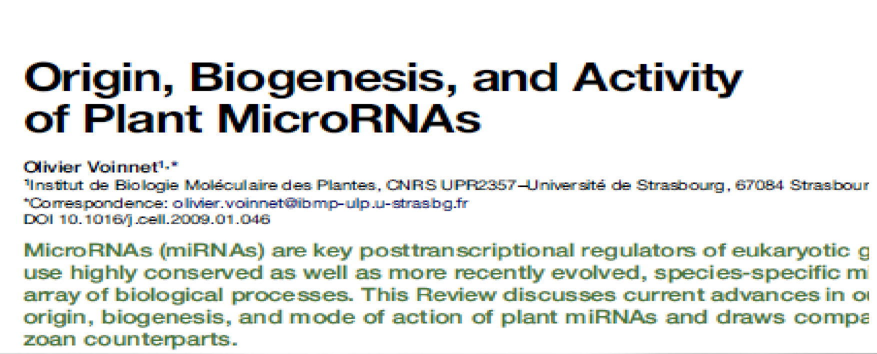


International Journal of Alzheimer's Disease
Volume 2011 (2011), Article ID 894938, 6 pages
doi:10.4061/2011/894938

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]



Origin, Biogenesis, and Activity of Plant MicroRNAs

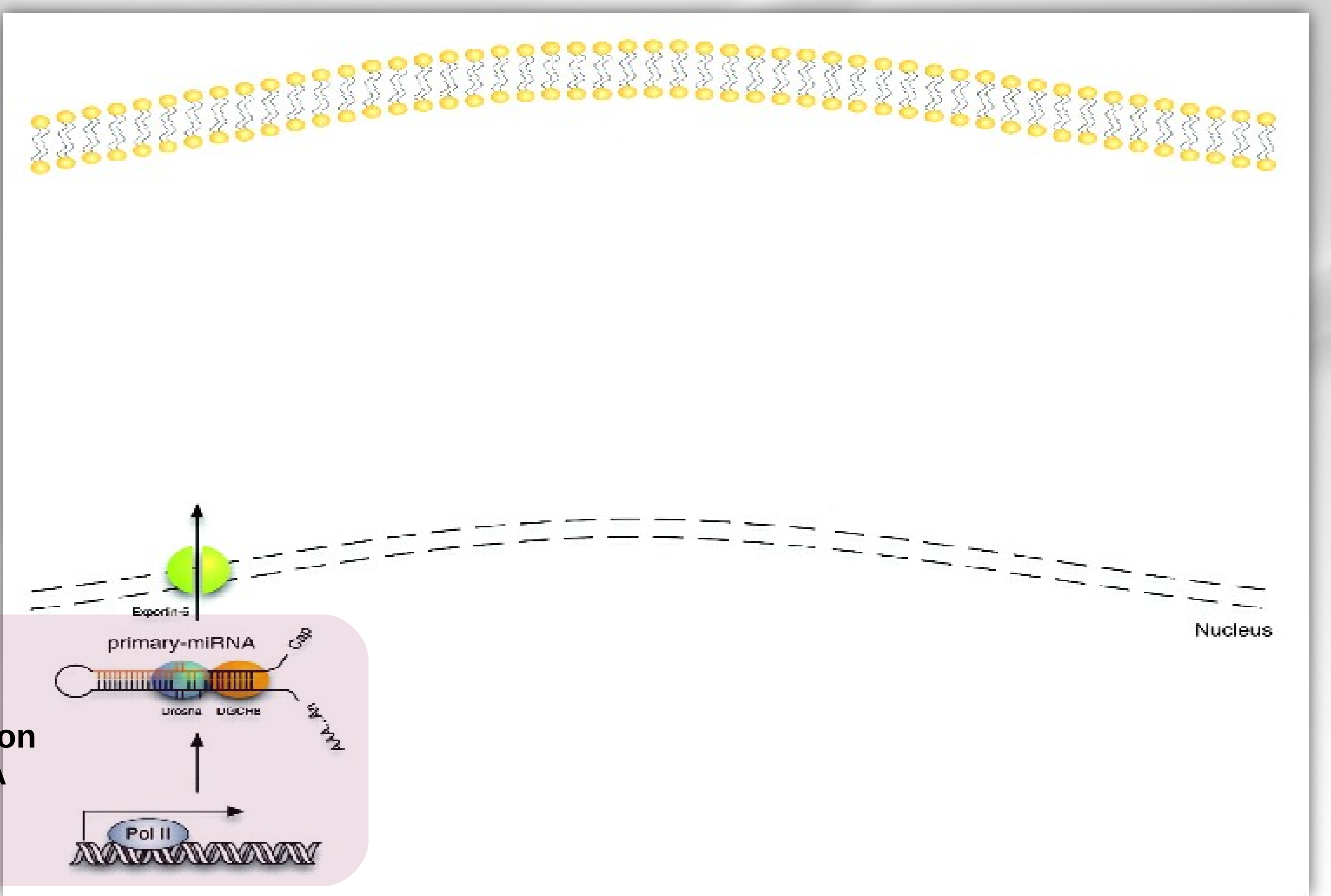
Olivier Voinnet[1,*]
[1]Institut de Biologie Moléculaire des Plantes, CNRS UPR2357–Université de Strasbourg, 67084 Strasbour
*Correspondence: olivier.voinnet@ibmp-ulp.u-strasbg.fr
DOI 10.1016/j.cell.2009.01.046

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.

The miRNA world

- **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**
  - **…**
- **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**
  - **...**

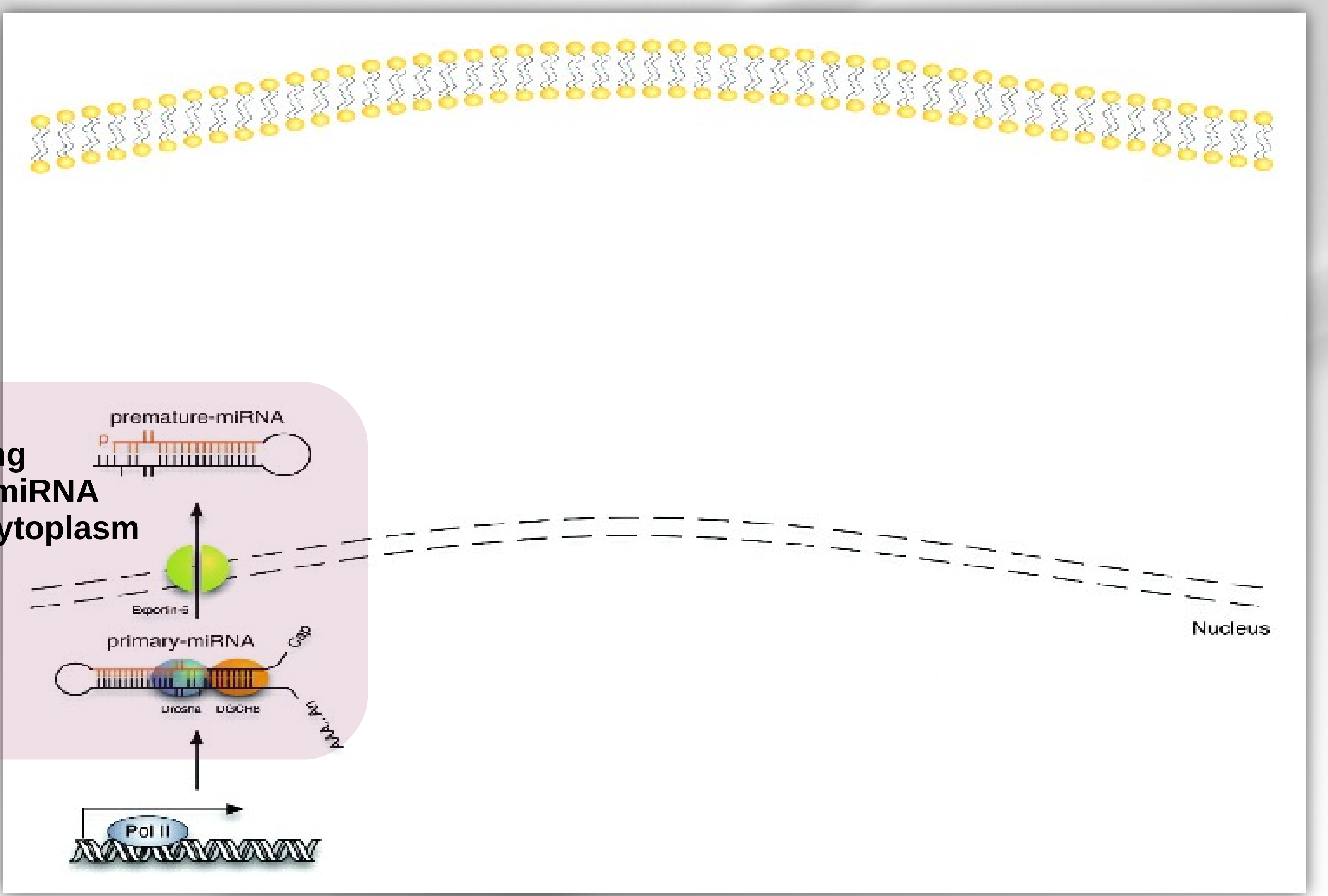
The miRNA biogenesis



Pol II transcription
Into a pri-miRNA

Plateforme Bioinformatique Midi-Pyrénées

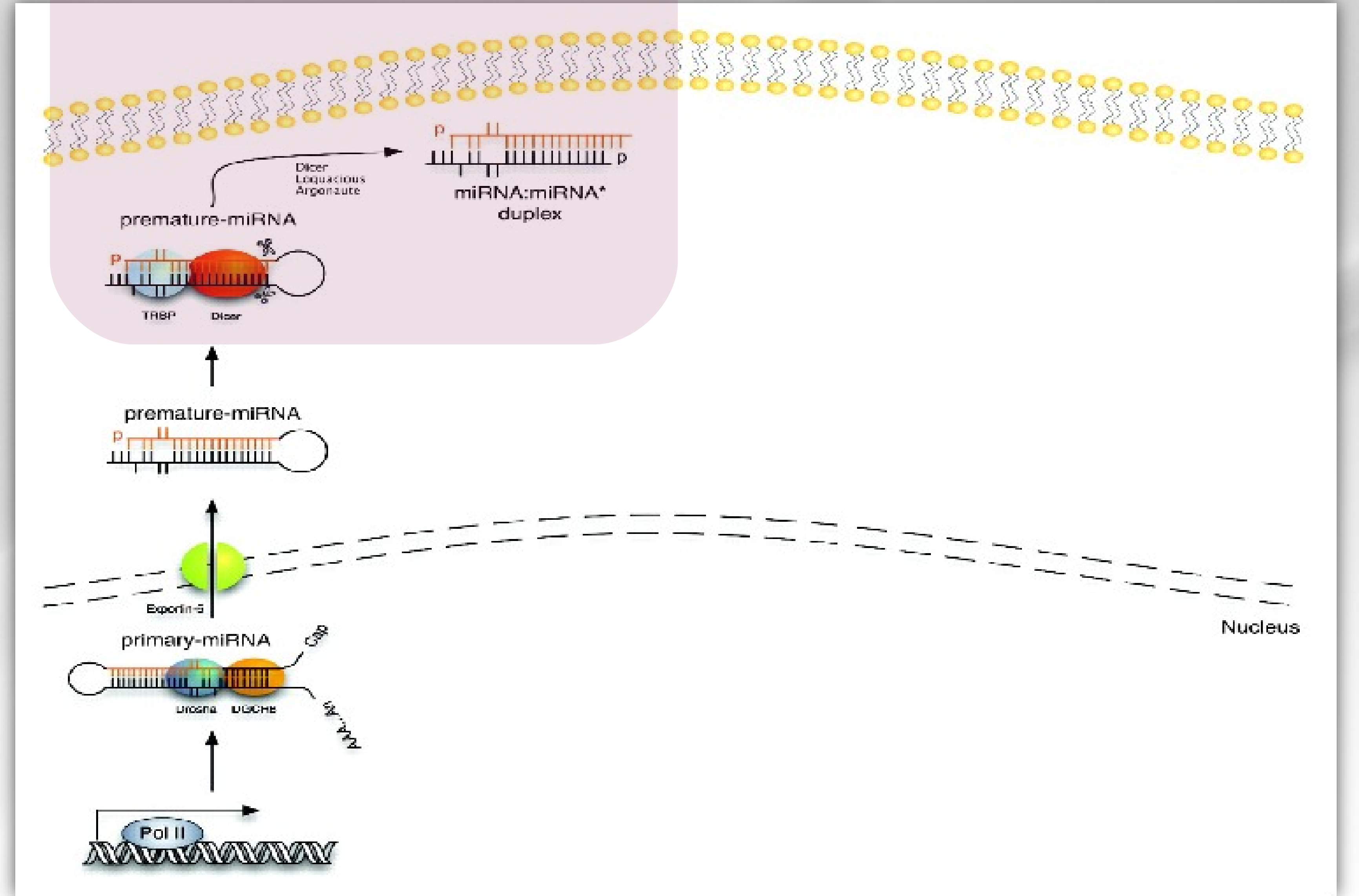
The miRNA biogenesis



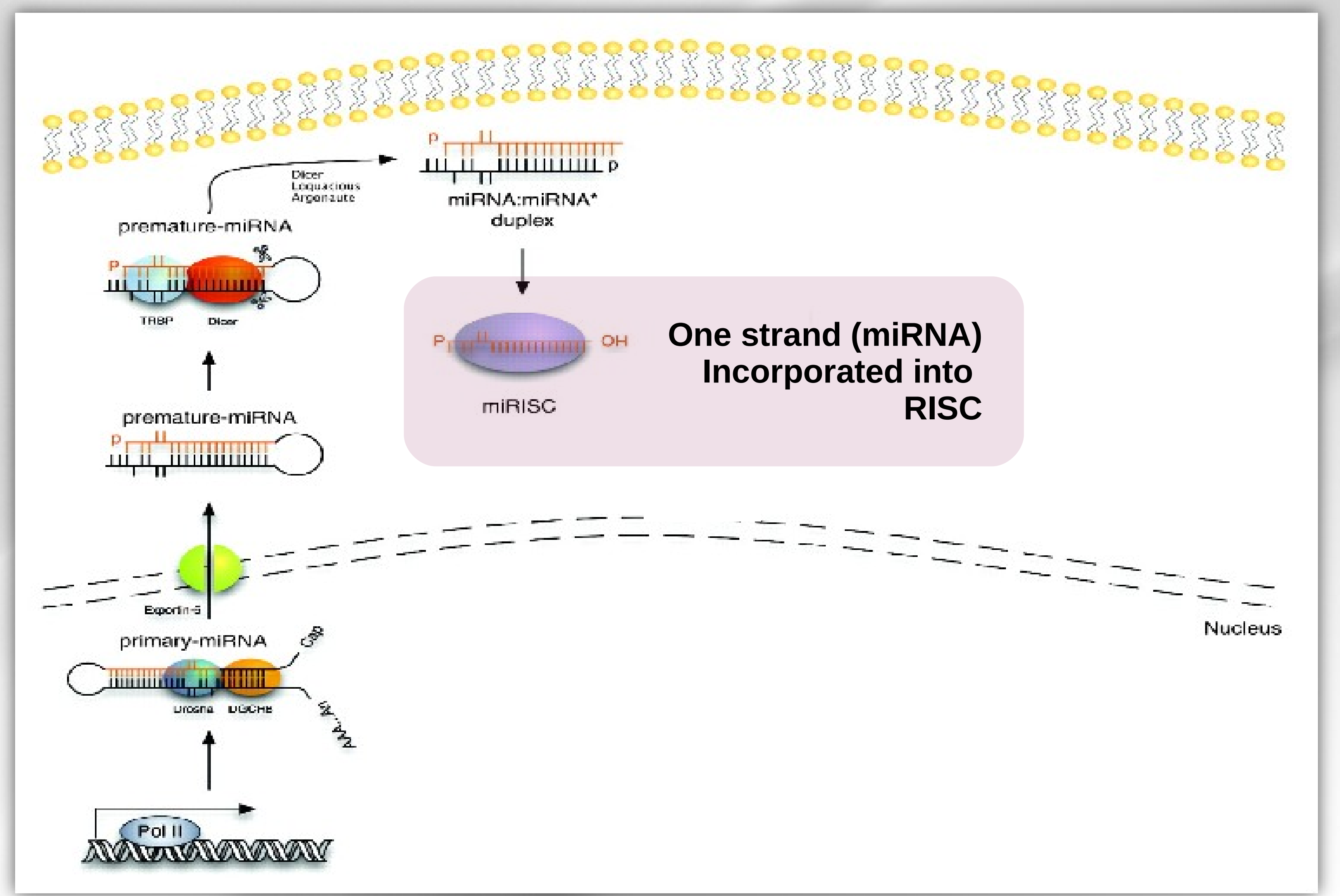
Drosha processing
one or more pre-miRNA
Exported in the cytoplasm

The miRNA biogenesis

Dicer processing
Into a duplex miRNA
Structure



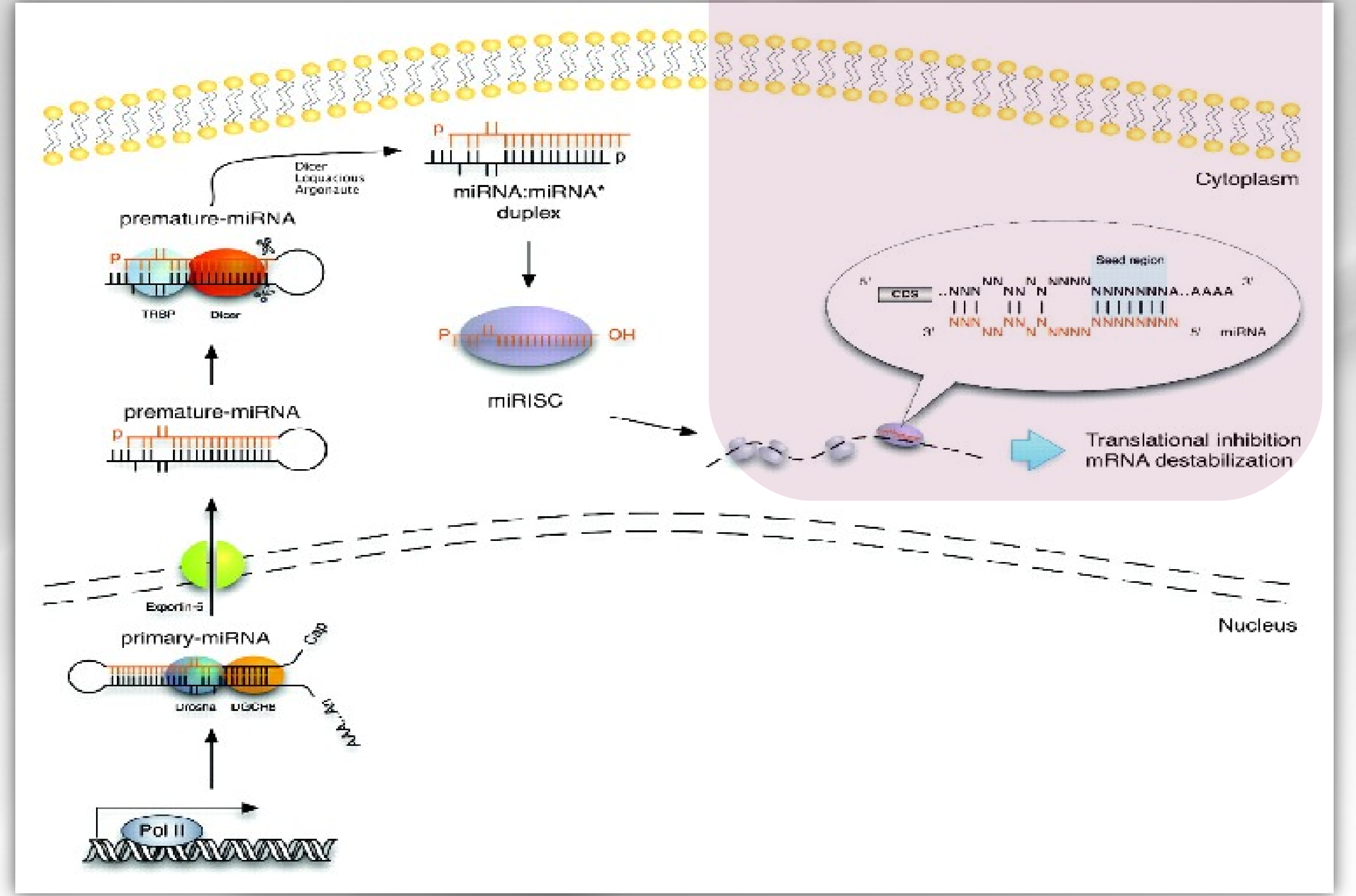
The miRNA biogenesis



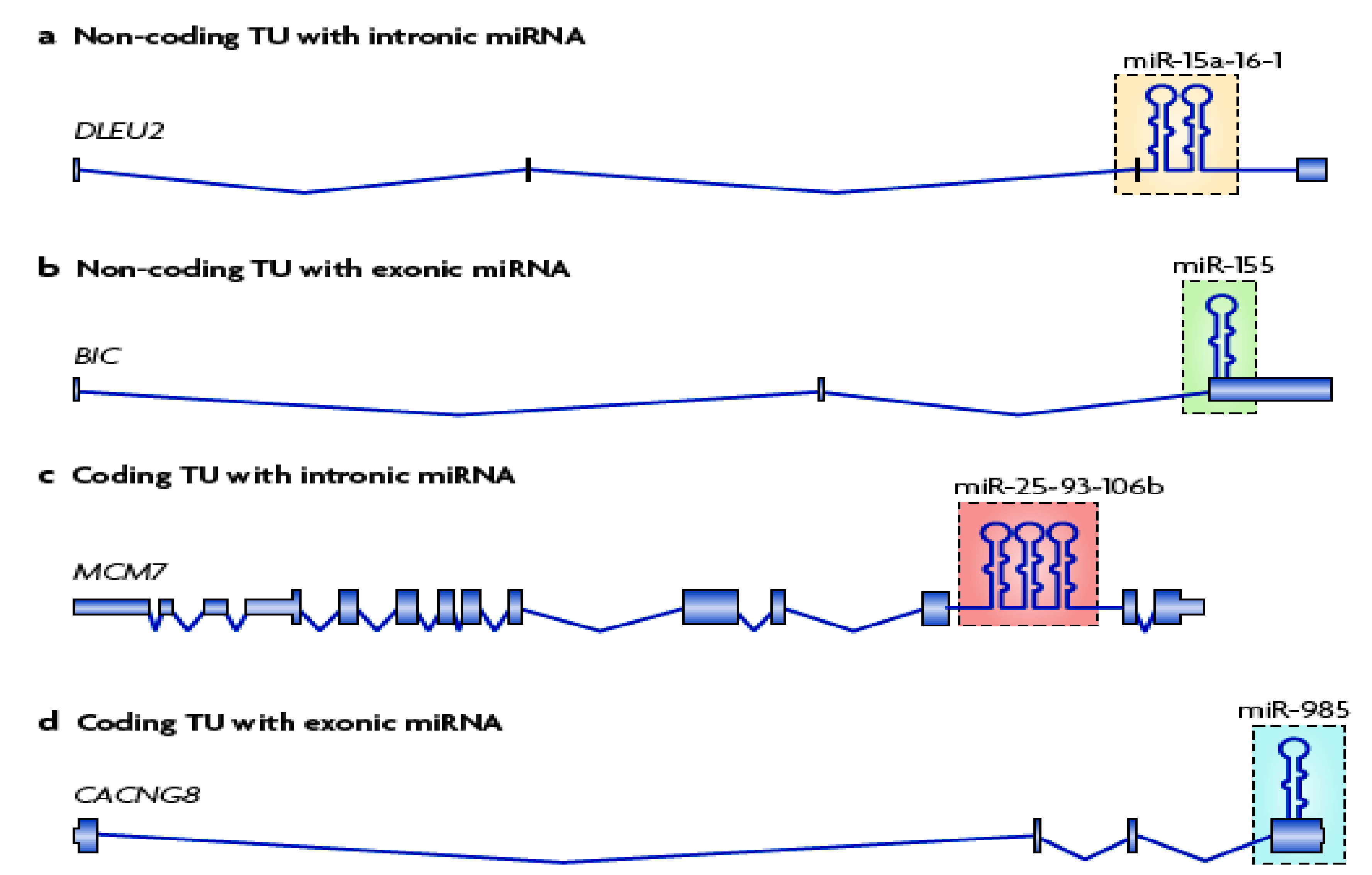
One strand (miRNA)
Incorporated into
RISC

The miRNA biogenesis

target mRNA translationally repressed

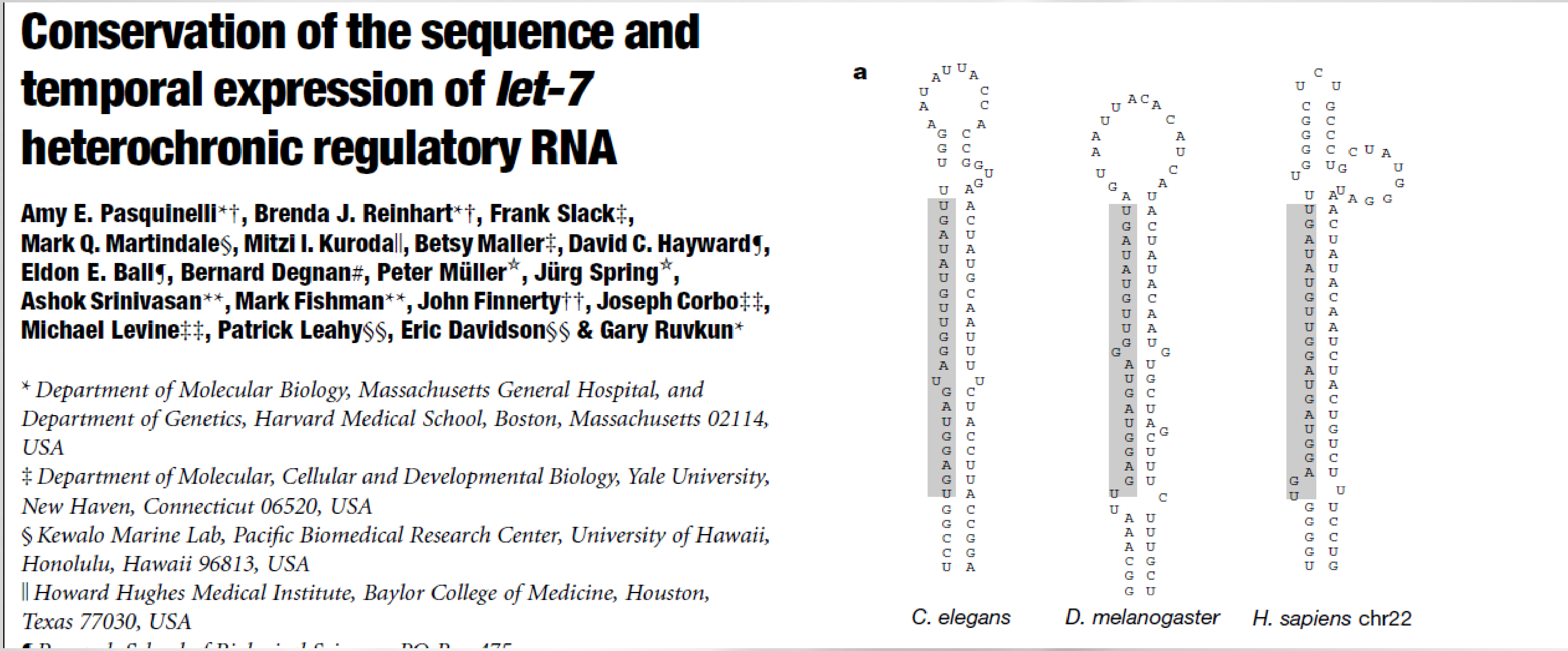


The miRNA location


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**

The 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

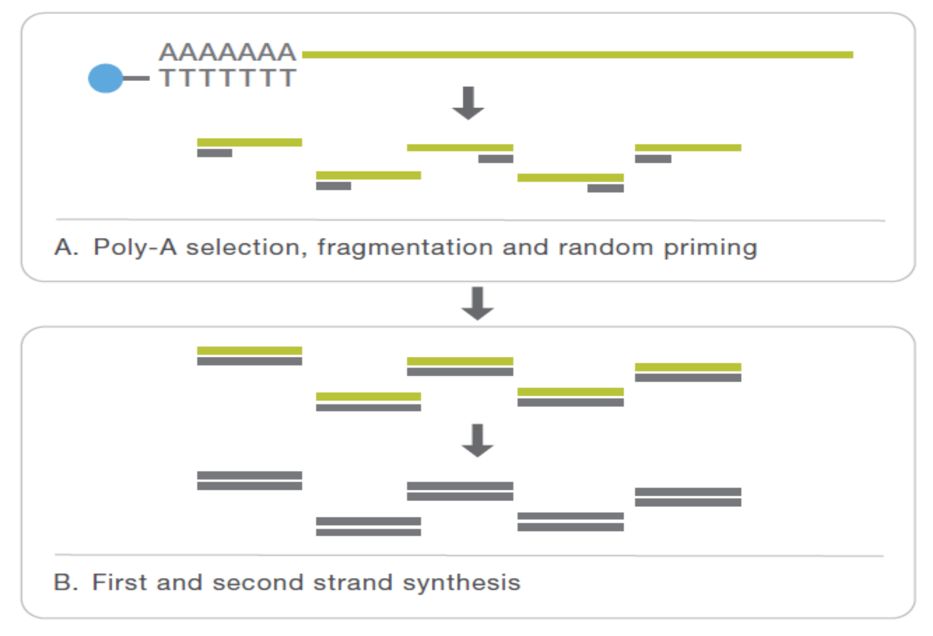
‖ Howard Hughes Medical Institute, Baylor College of Medicine, Houston, Texas 77030, USA

¶ Research School of Biological Sciences, PO Box 475

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

How can we study miRNA ?

- RNAseq not suited for miRNA (protocol and size)


AAAAAAA
TTTTTTT
A. Poly-A selection, fragmentation and random priming
B. First and second strand synthesis

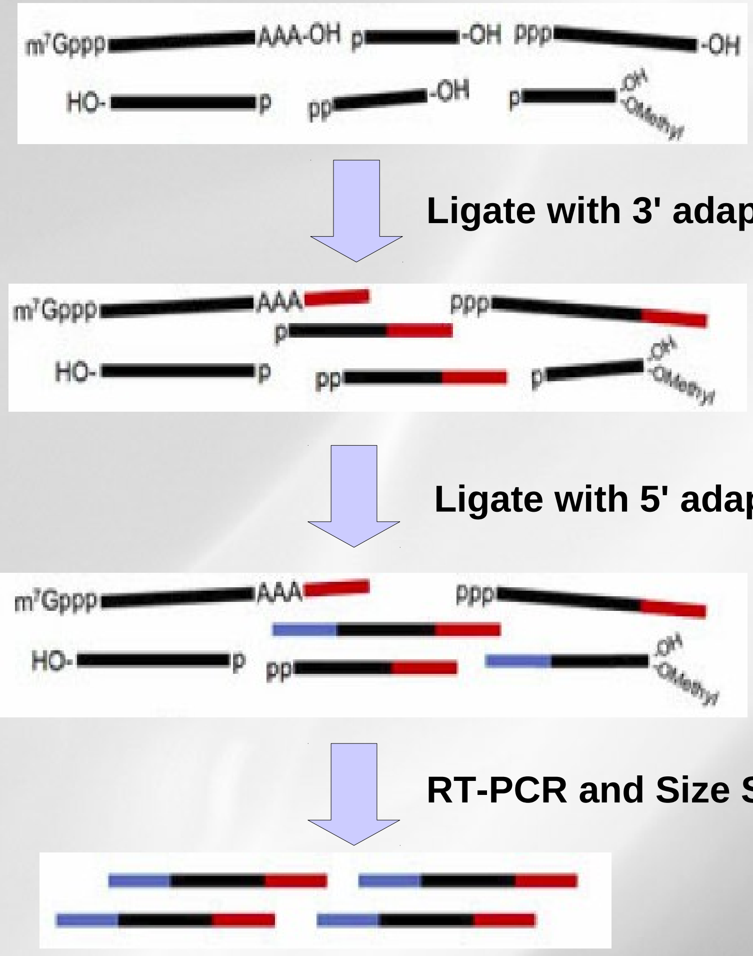
- small RNAseq: ability of high throughput sequencing to
  - Interrogate known and new small RNAs
  - Quantify them
  - Profile them on a large number of samples
  - Cost-effective

small RNAseq platforms comparisons

Platform	**454** **Roche Titanium**	**HiSeq2000** **Illumina**	**Solid 3+** **Life Technologies**
Caracteristics	-Titanium chemistry -Pyrosequencing -PCR amplification	- Polymerase-based sequence-by-synthesis -PCR amplification -Multiplexing	-ligation-base-sequencing -PCR amplification
Applications	-De novo sequencing -Small genomes -Transcriptome	-Resequencing -Transcriptome -Epigenomic -Small RNA -Allele specific sequencing	-De novo sequencing -Resequencing -Transcriptome -Epigenomic -Small RNA
Paired end separation	Not used	200bp	200bp
Mb / run	800Mb	600Gb	60Gb
Read length	800 bp	100bp	50bp
Known Biases	- Long homopolymer - makes signal saturation - read duplication	- Rich GC or AT regions: under-representation during amplification - Most error in end of cycle	- read duplication ?

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

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

MicroRNA sequencing library

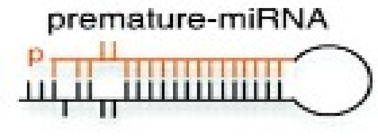
What are we looking for ?

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

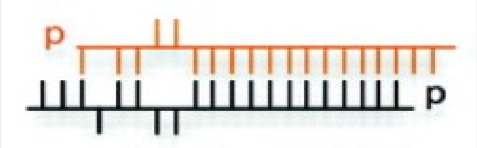
small RNAseq data analysis

What should we retain for data analysis ?

- Pre-miRNA information:



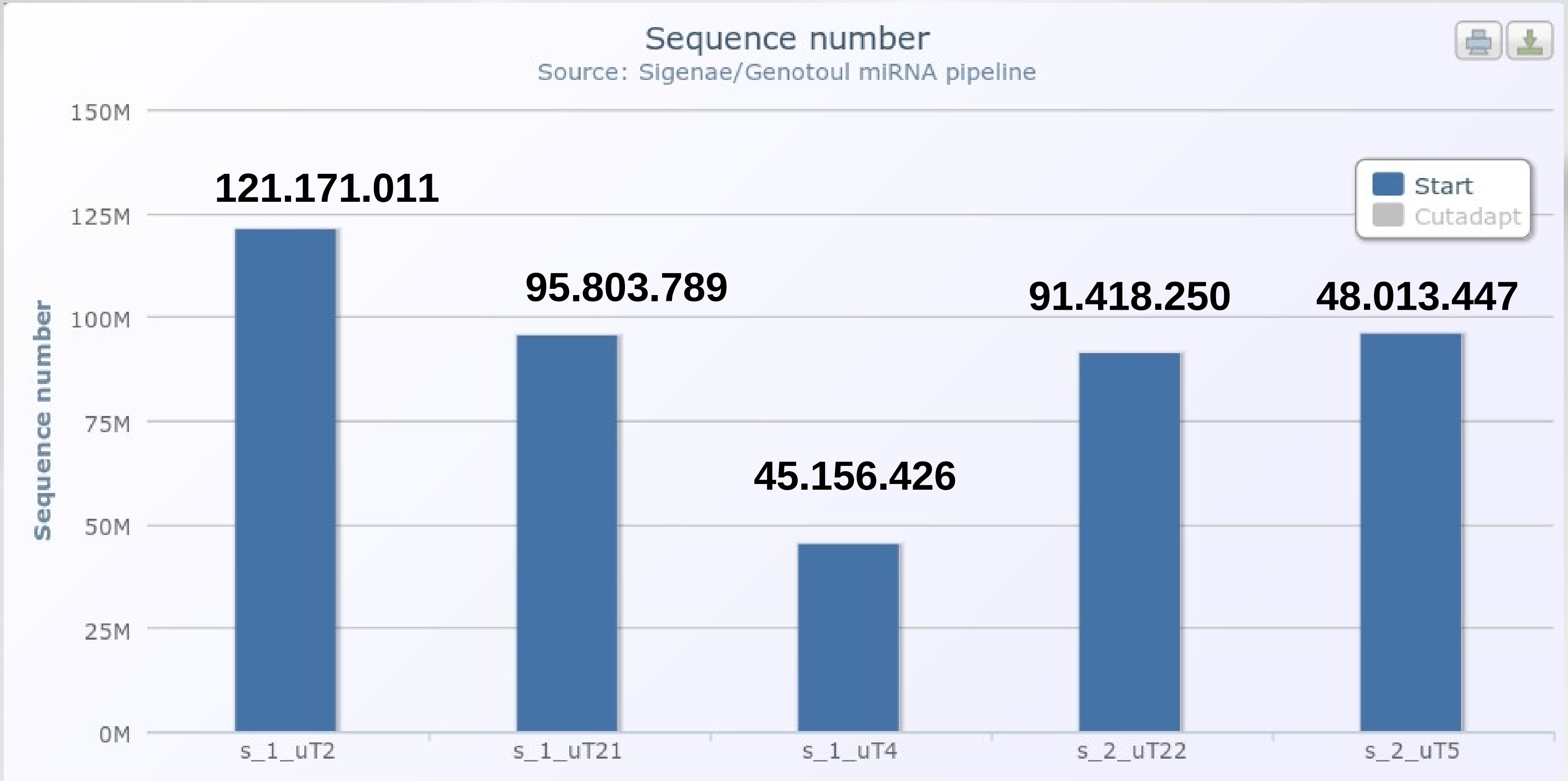
  - Hairpin structure of the pre-miRNA
  - Pre-miRNA localisation (coding/non coding TU intronic/exonic)
  - Presence of cluster
  - Size of the pre-miRNA
- miRNA and miRNA* information:



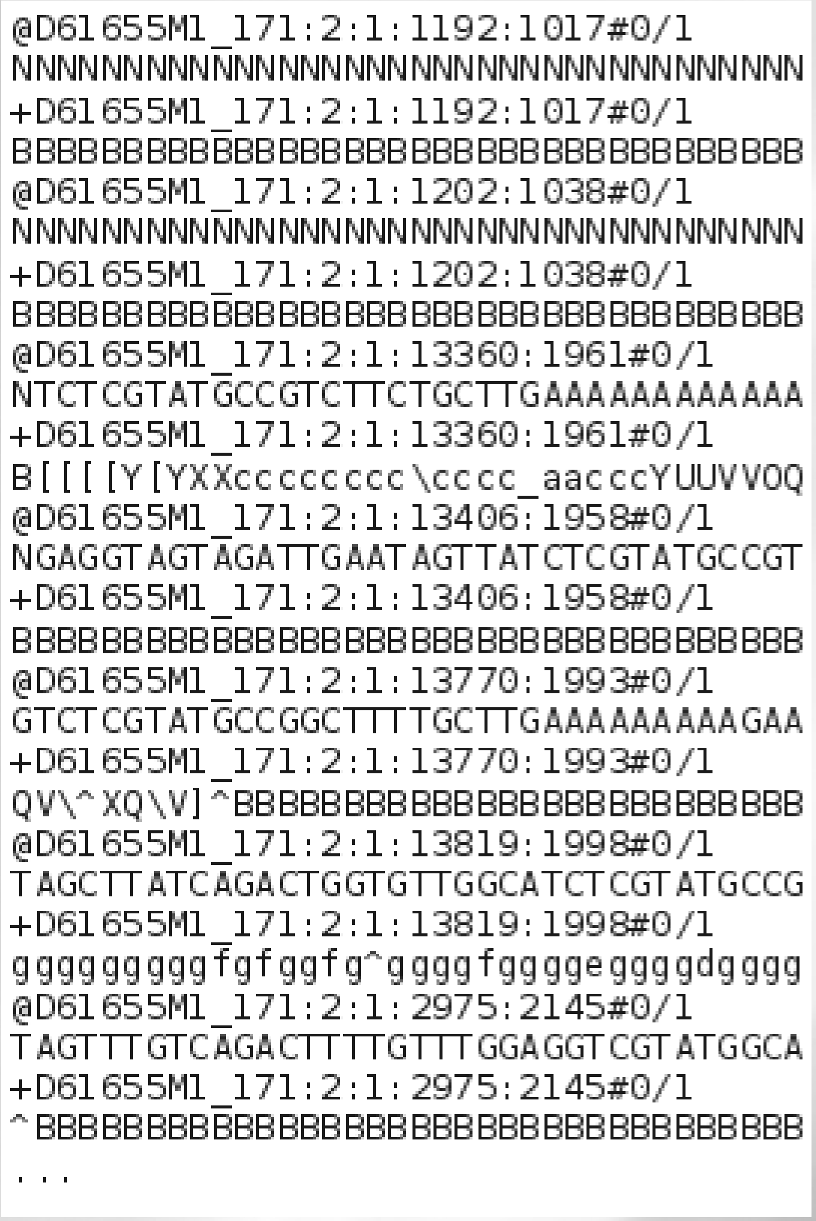
  - Existence of both miRNA and miRNA*
  - Sequence conservation
  - Overhang (around 2 nt) related to drosha and Dicer cuts
  - Size of miRNA and miRNA*
  - Overexpression of the miRNA compared to the miRNA*
- Existence of other products in sRNAseq data

Description of the dataset

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

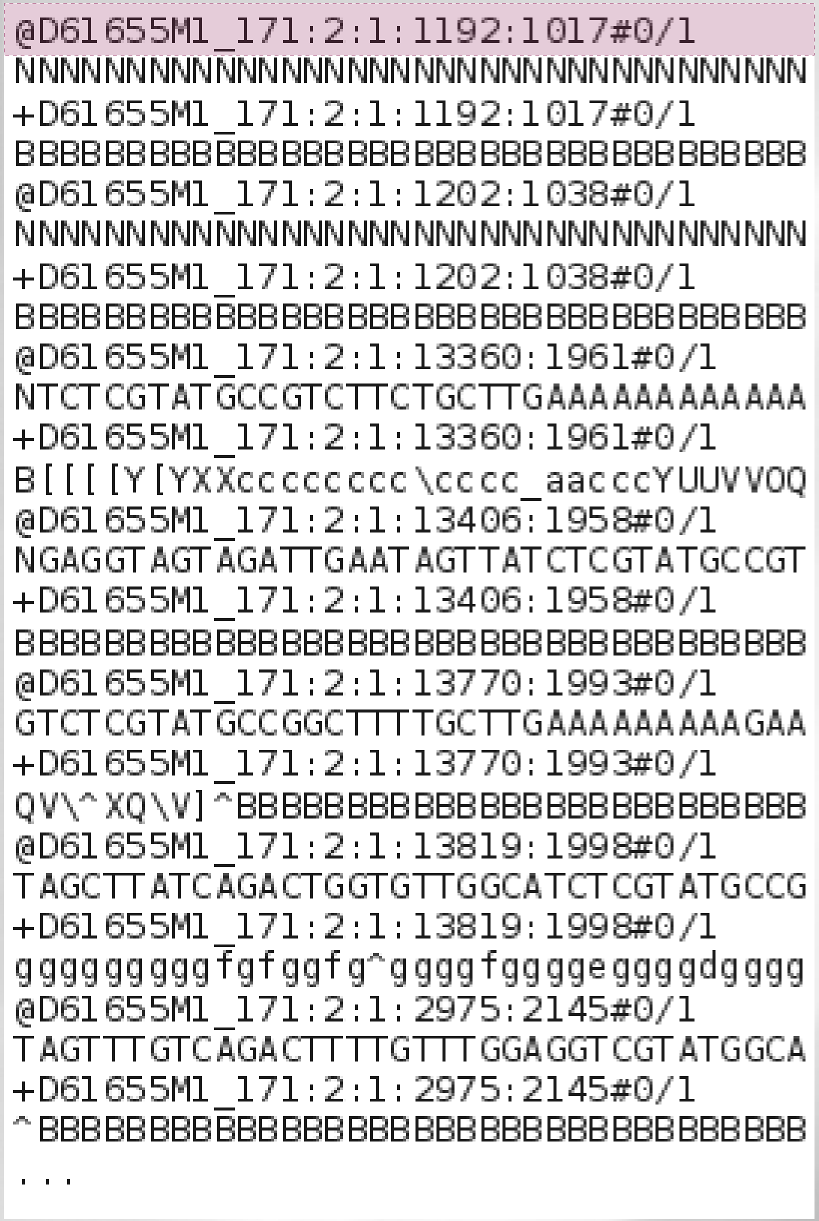


Fastq format



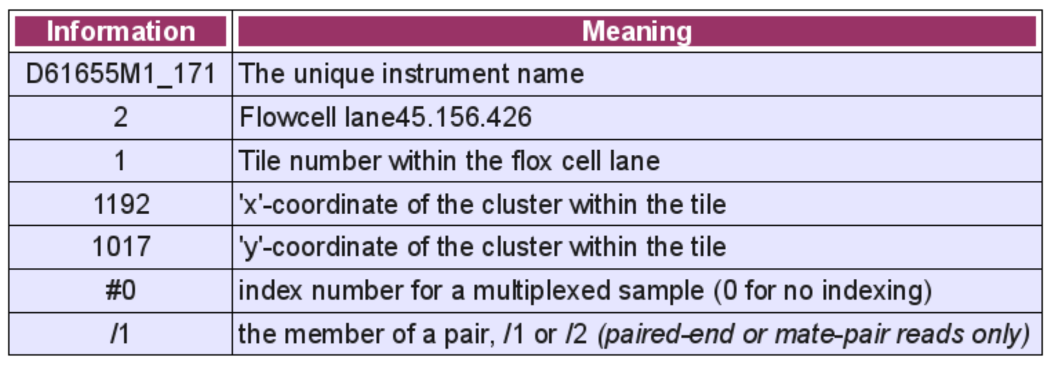
```
@D61655M1_171:2:1:1192:1017#0/1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+D61655M1_171:2:1:1192:1017#0/1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:1202:1038#0/1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+D61655M1_171:2:1:1202:1038#0/1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:13360:1961#0/1
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



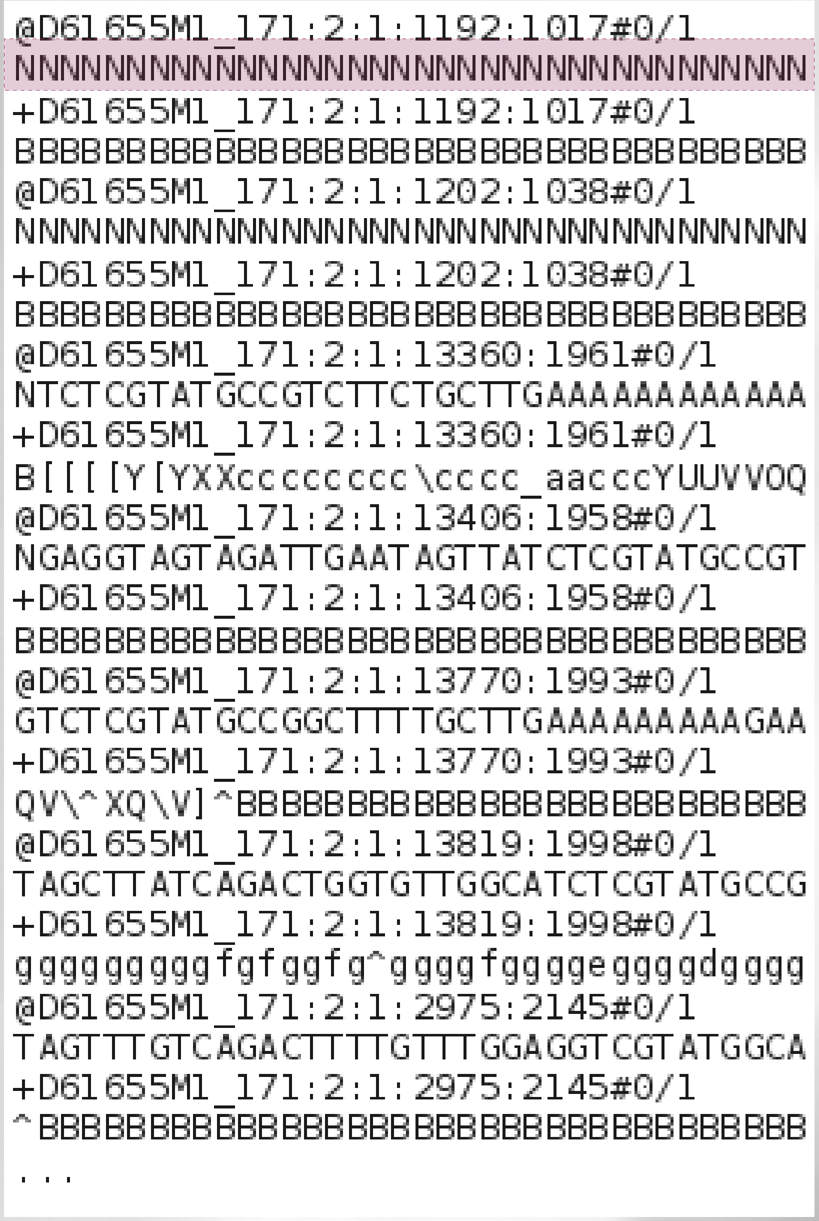
```
@D61655M1_171:2:1:1192:1017#0/1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+D61655M1_171:2:1:1192:1017#0/1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:1202:1038#0/1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+D61655M1_171:2:1:1202:1038#0/1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@D61655M1_171:2:1:13360:1961#0/1
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 @



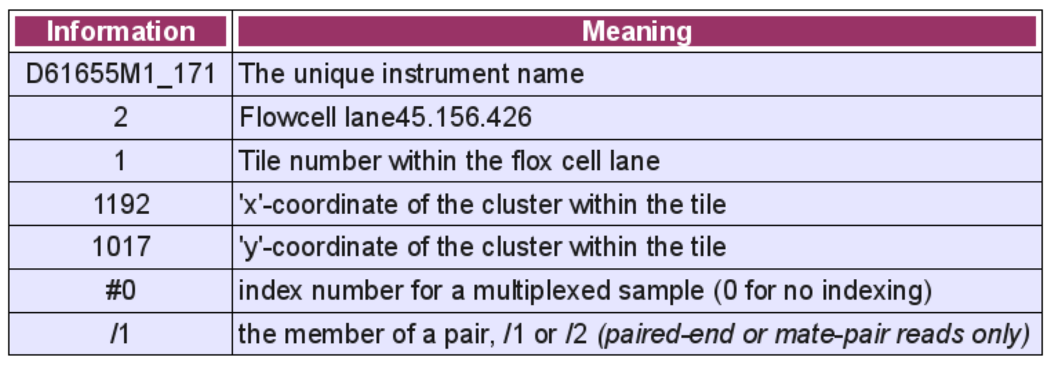
Information	Meaning
D61655M1_171	The unique instrument name
2	Flowcell lane45.156.426
1	Tile number within the flox cell lane
1192	'x'-coordinate of the cluster within the tile
1017	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member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

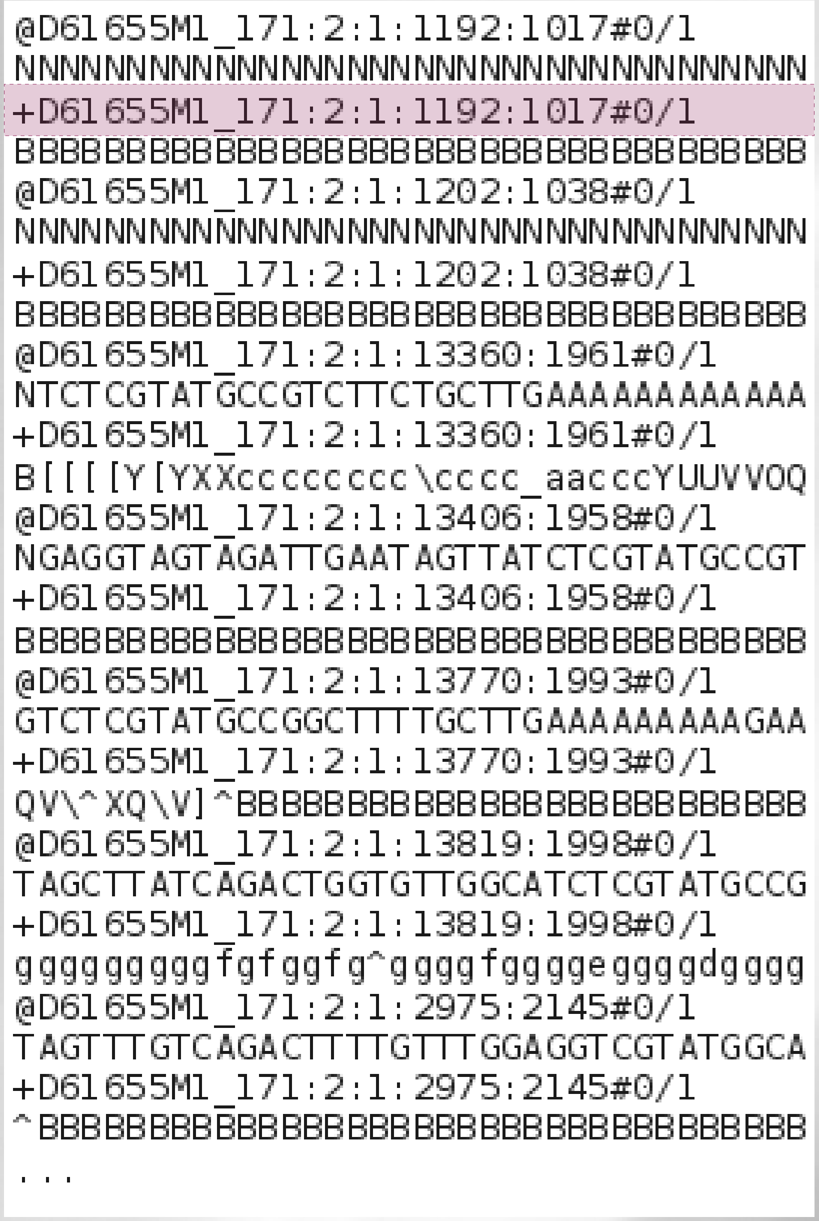
Line 1 starts with @



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)*

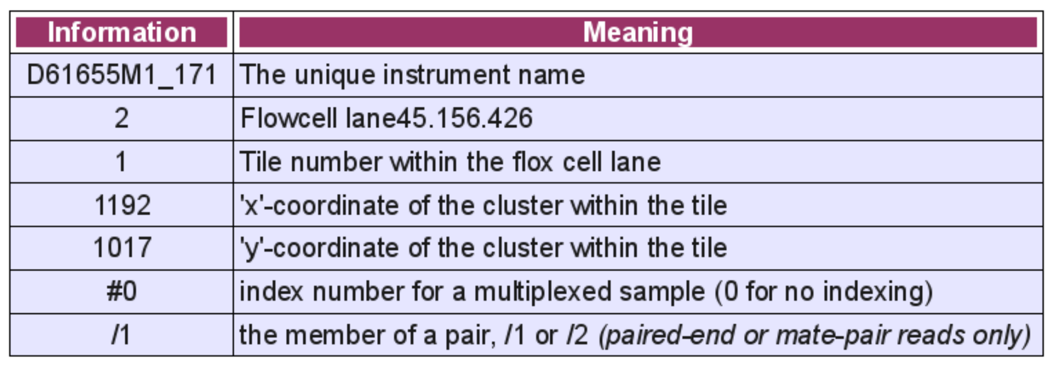
Line 2 Raw sequence of 36 nt (36 cycles in sequencing)

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 @

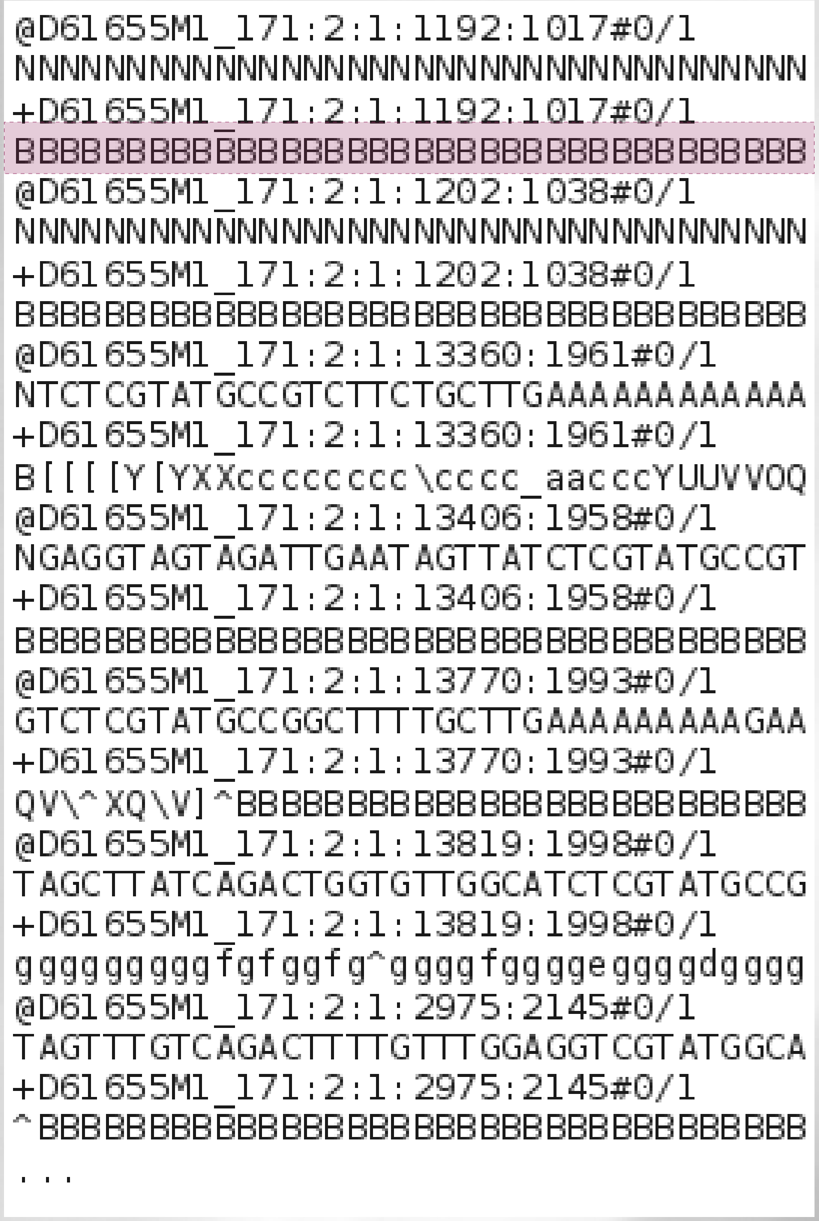


Information	Meaning
D61655M1_171	The unique instrument name
2	Flowcell lane45.156.426
1	Tile number within the flox cell lane
1192	'x'-coordinate of the cluster within the tile
1017	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member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

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

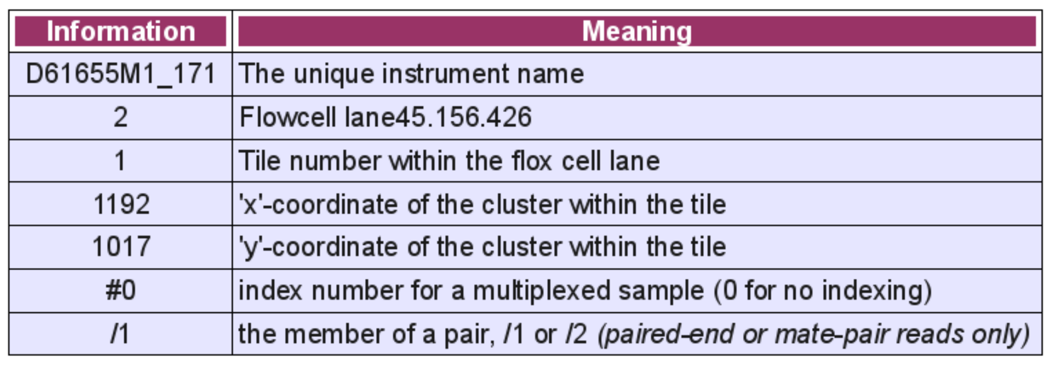
Line 3 starts with a '+' character and is optionally followed by the same sequence identifier (and any description) again.

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 @



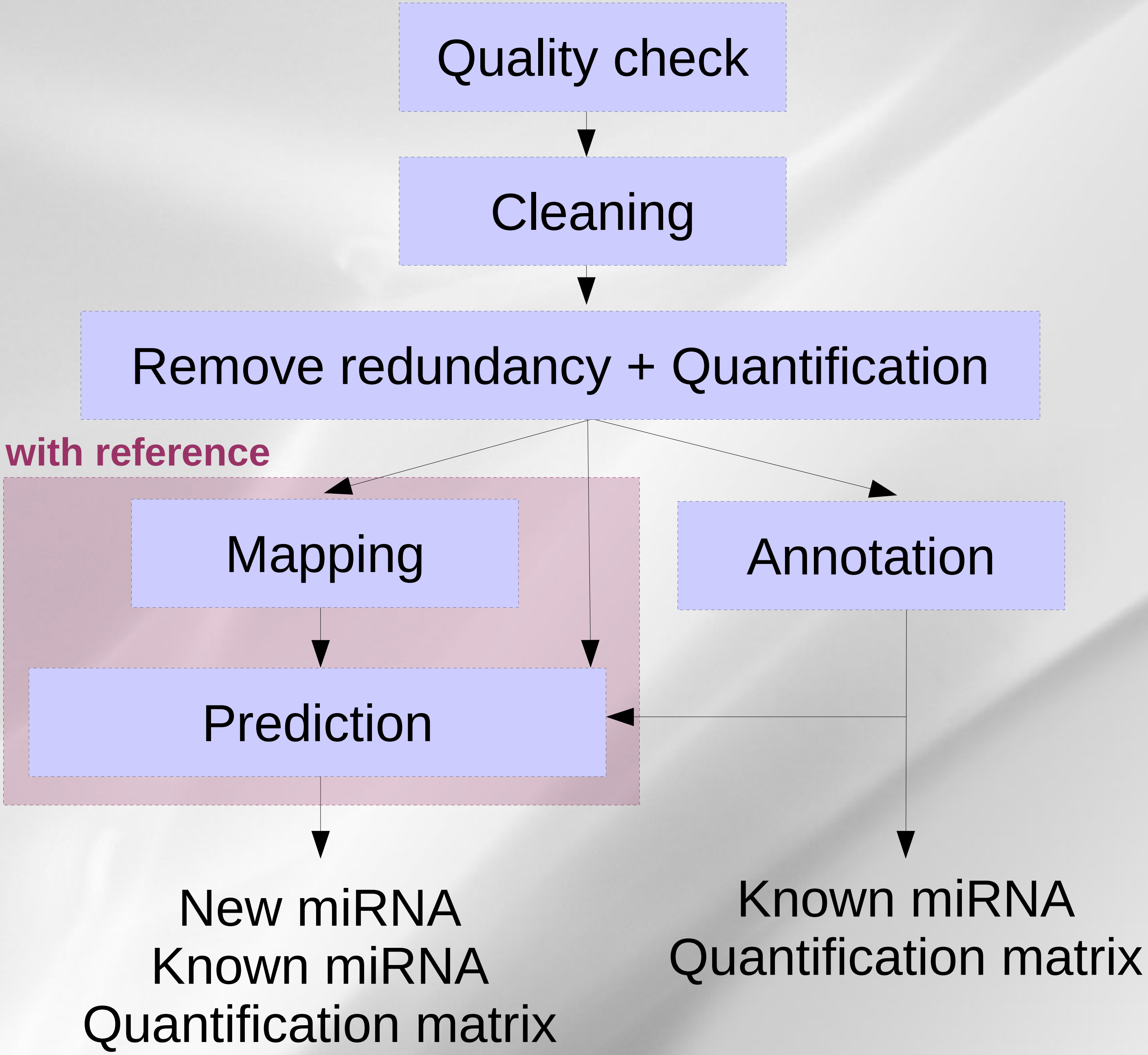
Information	Meaning
D61655M1_171	The unique instrument name
2	Flowcell lane45.156.426
1	Tile number within the flox cell lane
1192	'x'-coordinate of the cluster within the tile
1017	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member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

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

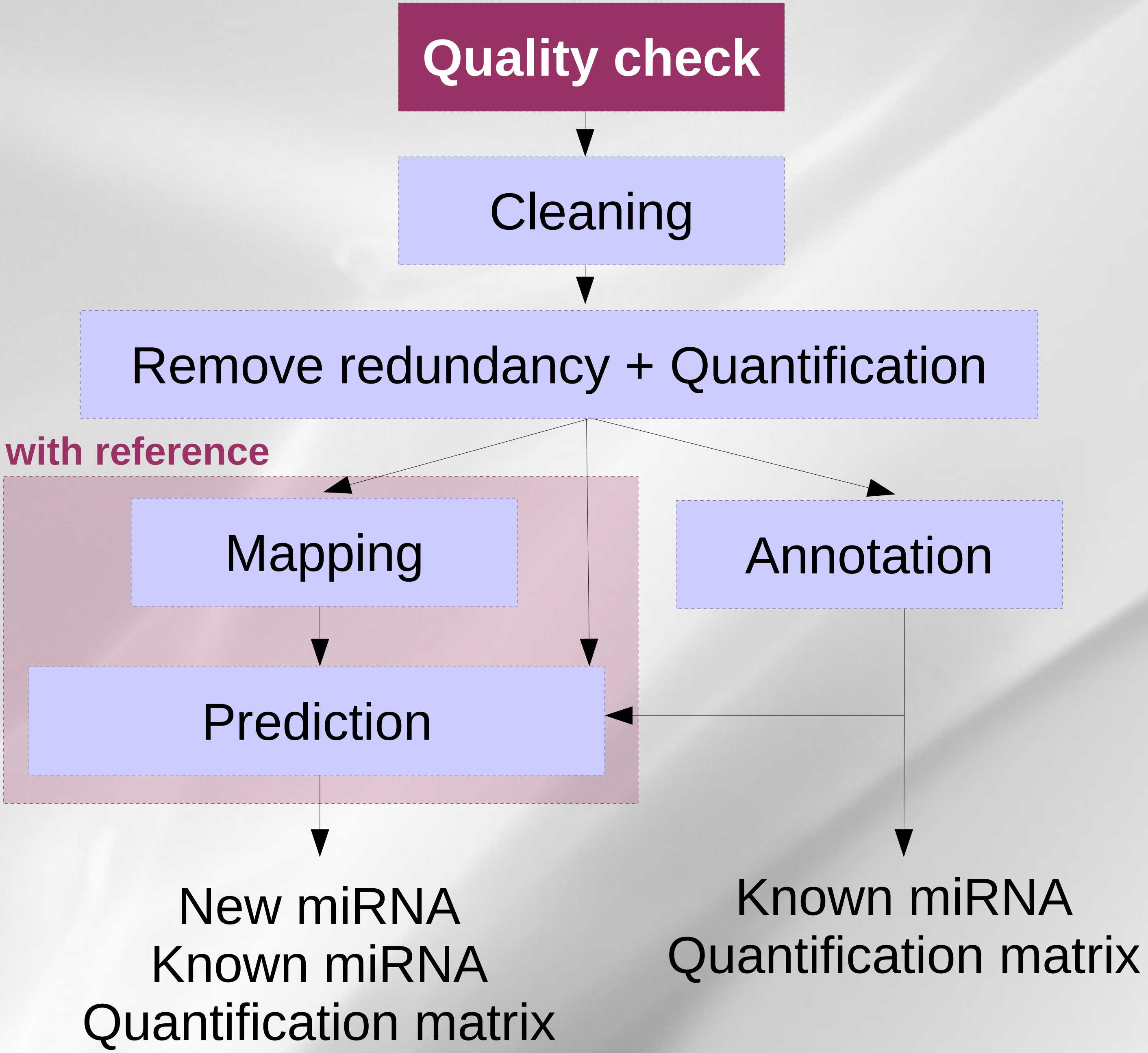
Line 3 starts with a '+' character and is optionally followed by the same sequence identifier (and any description) again.

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.

small RNAseq pipeline



small RNAseq pipeline



Quality check

Cleaning

Remove redundancy + Quantification

with reference

Mapping

Prediction

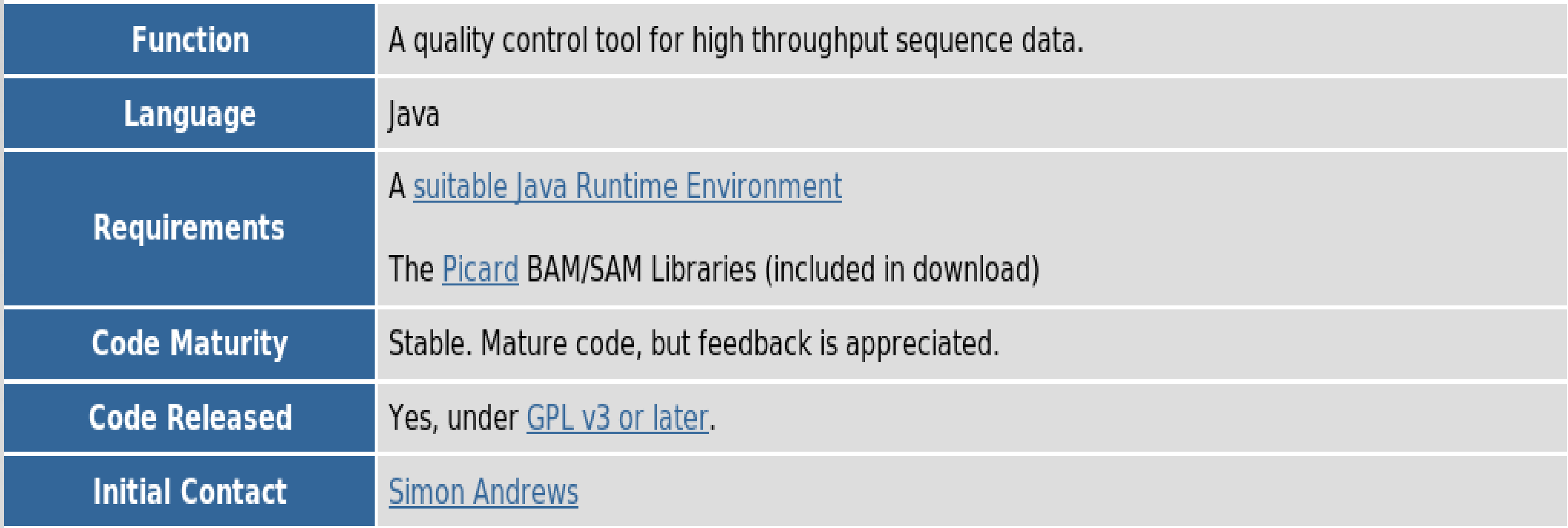
Annotation

New miRNA
Known miRNA
Quantification matrix

Known miRNA
Quantification matrix

1. Quality control

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



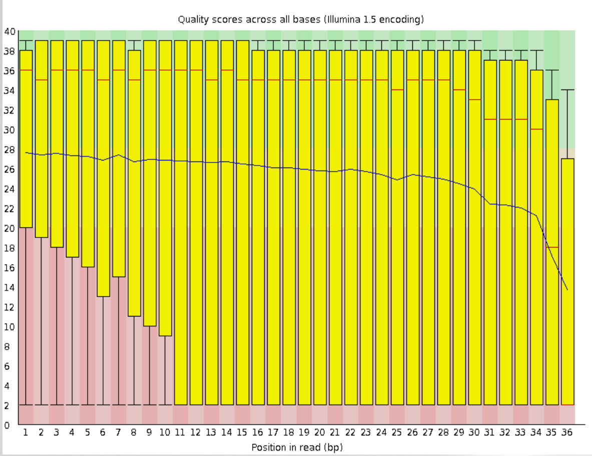
Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later.
Initial Contact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

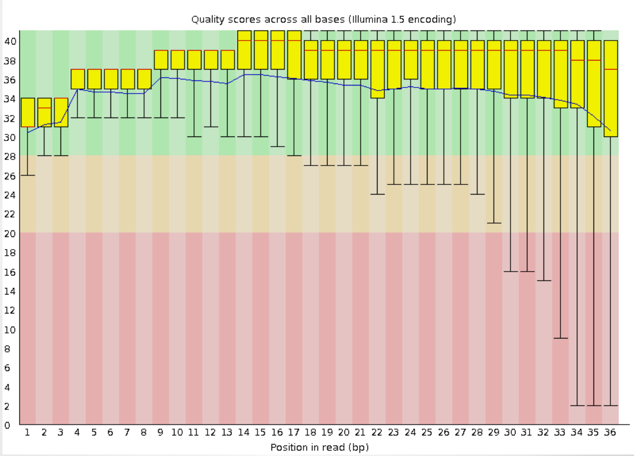
A simple way to do quality control. It provides a modular set of analyses to give a quick impression of whether data has any problems of which you should be aware before doing any further analysis. The main functions of FastQC are:
- Import of data from BAM, SAM or FastQ files (any variant)
- Provide 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

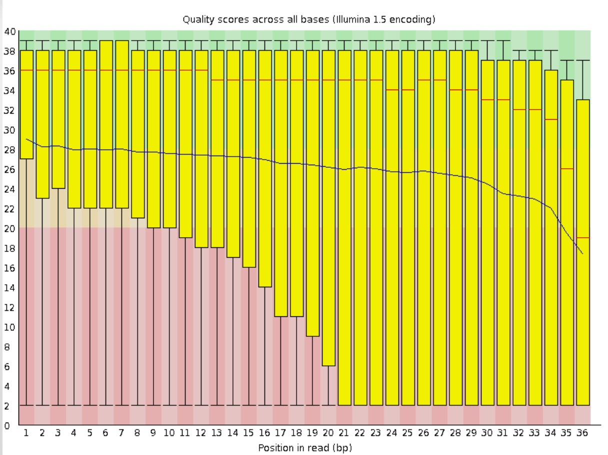
```
Fastqc -o nf.out nf_in.fastq
```

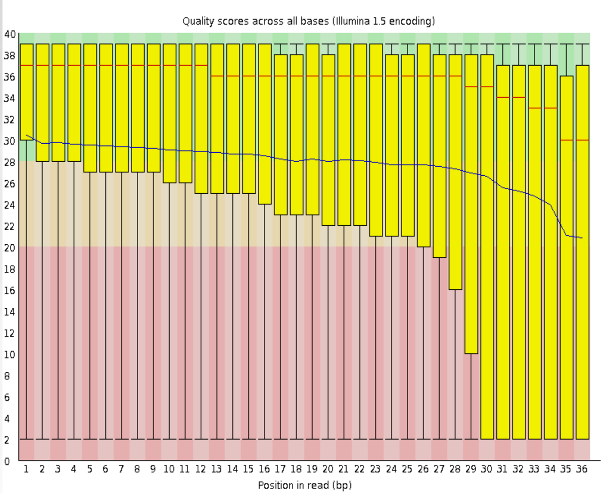
1. Quality control

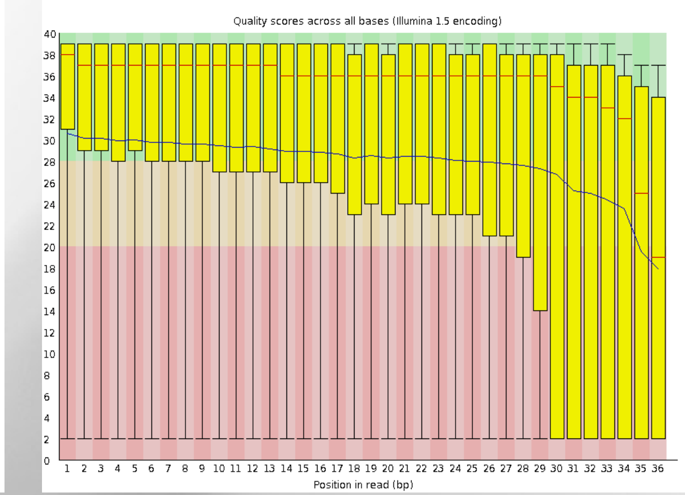
- **Per base quality**





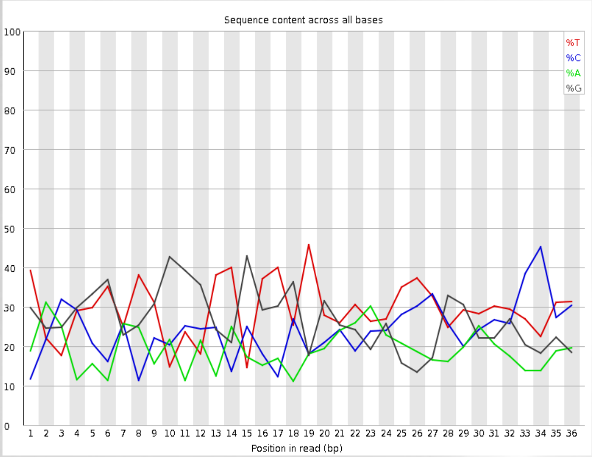


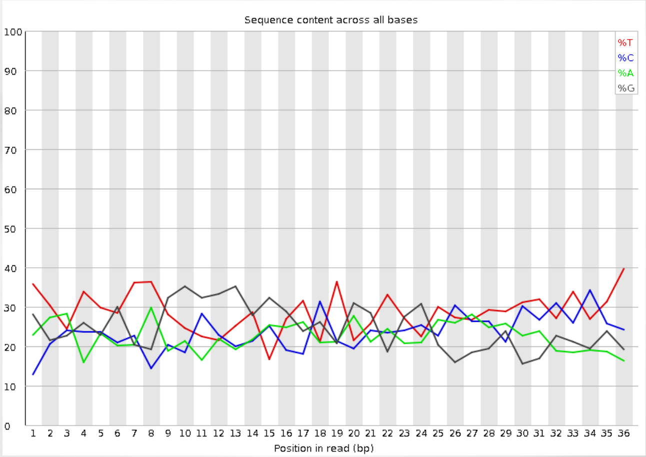


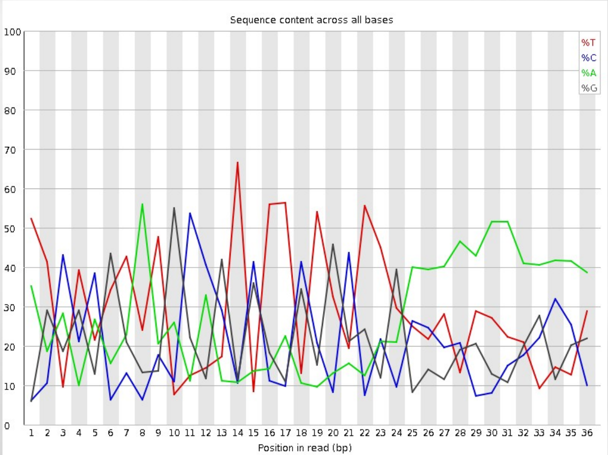


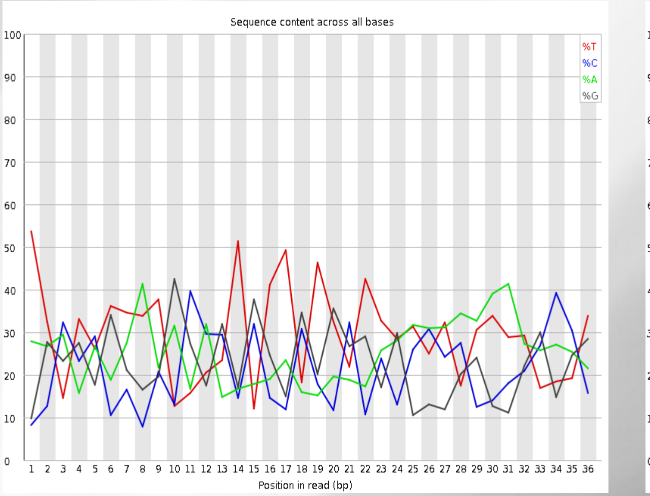
1. Quality control

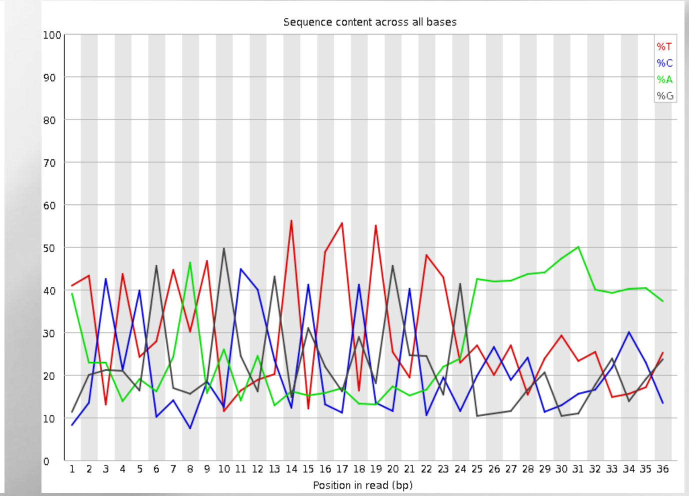
- **Sequences content in nucleotides**



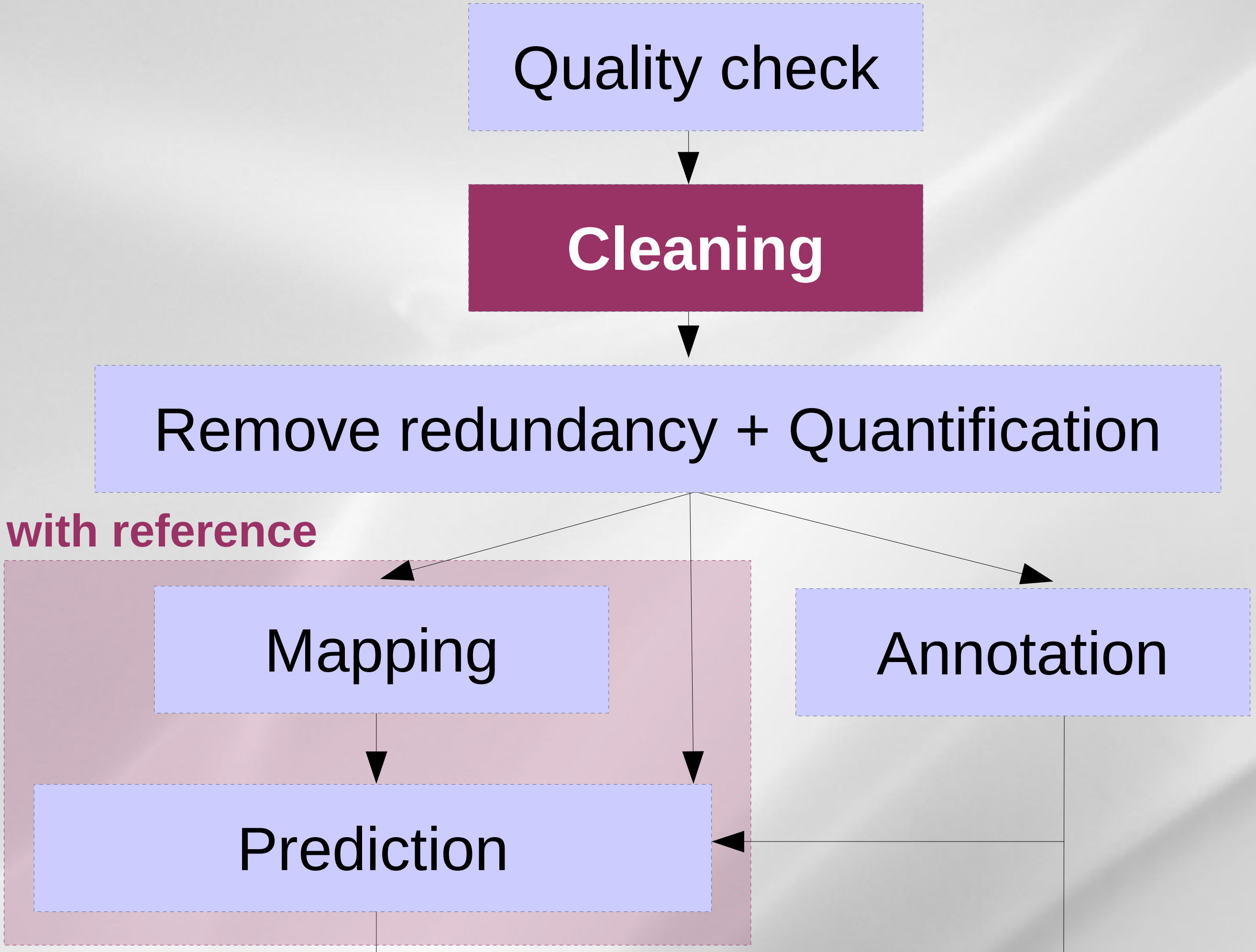








small RNAseq pipeline

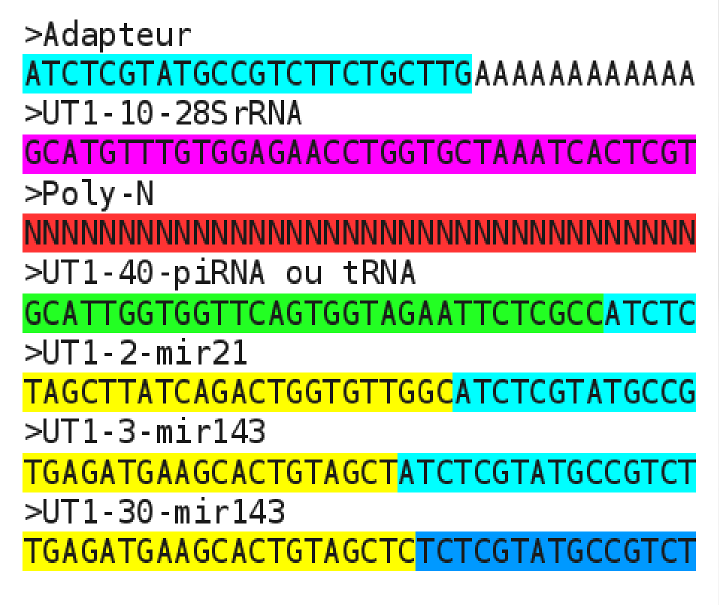


New miRNA
Known miRNA
Quantification matrix

Known miRNA
Quantification matrix

2. Why cleaning ?

Outputed reads

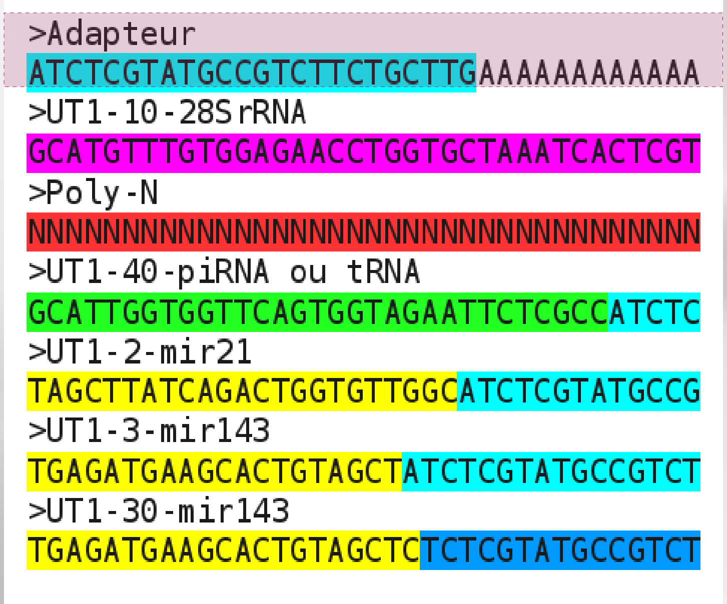


```
>Adapteur
ATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAA
>UT1-10-28S rRNA
GCATGTTTGTGGAGAACCTGGTGCTAAATCACTCGT
>Poly-N
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
>UT1-40-piRNA ou tRNA
GCATTGGTGGTTCAGTGGTAGAATTCTCGCCATCTC
>UT1-2-mir21
TAGCTTATCAGACTGGTGTTGGCATCTCGTATGCCG
>UT1-3-mir143
TGAGATGAAGCACTGTAGCTATCTCGTATGCCGTCT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

2. Why cleaning ?

Outputed reads

- Some sequences contain only adapters



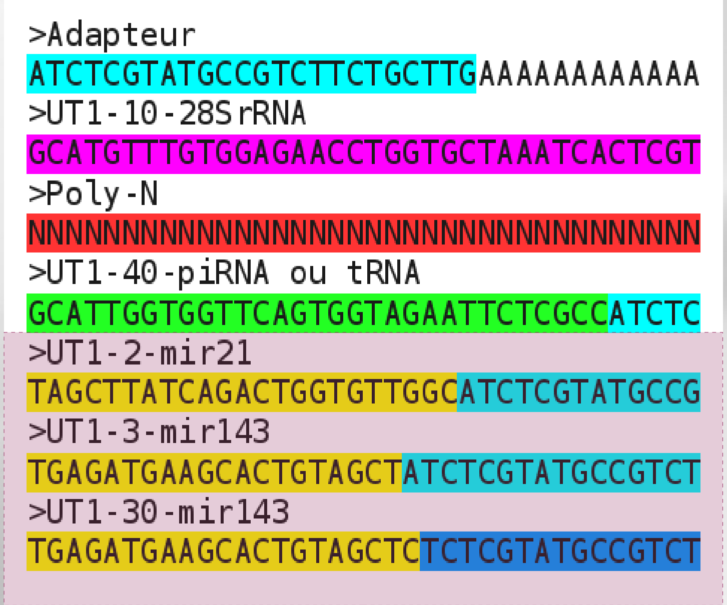
```
>Adapteur
ATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAA
>UT1-10-28S rRNA
GCATGTTTGTGGAGAACCTGGTGCTAAATCACTCGT
>Poly-N
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
>UT1-40-piRNA ou tRNA
GCATTGGTGGTTCAGTGGTAGAATTCTCGCCATCTC
>UT1-2-mir21
TAGCTTATCAGACTGGTGTTGGCATCTCGTATGCCG
>UT1-3-mir143
TGAGATGAAGCACTGTAGCTATCTCGTATGCCGTCT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

Plateforme Bioinformatique Midi-Pyrénées

2. Why cleaning ?

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

2. Why cleaning ?

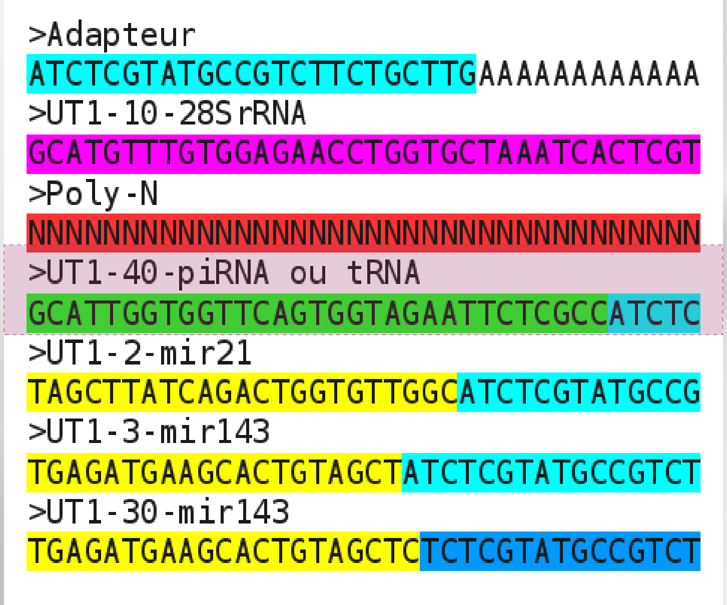
Outputed reads

- 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 RNAs (green).

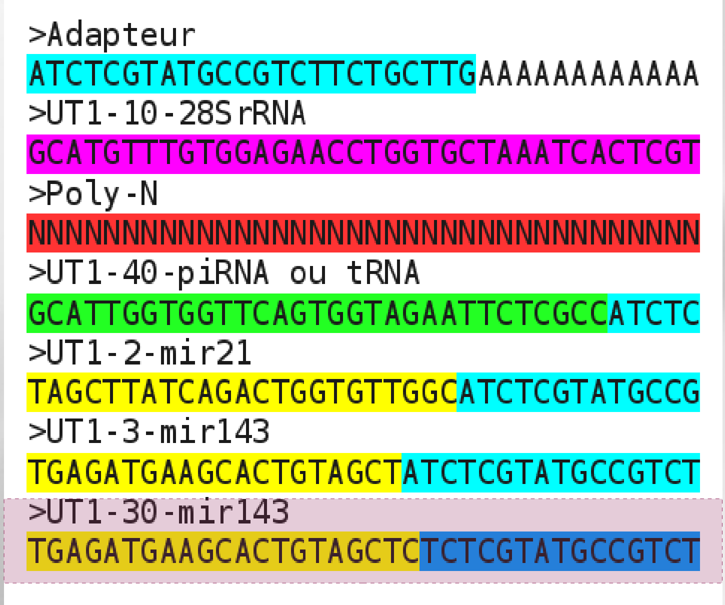


```
>Adapteur
ATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAA
>UT1-10-28SrRNA
GCATGTTTGTGGAGAACCTGGTGCTAAATCACTCGT
>Poly-N
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
>UT1-40-piRNA ou tRNA
GCATTGGTGGTTCAGTGGTAGAATTCTCGCCATCTC
>UT1-2-mir21
TAGCTTATCAGACTGGTGTTGGCATCTCGTATGCCG
>UT1-3-mir143
TGAGATGAAGCACTGTAGCTATCTCGTATGCCGTCT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

2. Why cleaning ?

Outputed 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 RNAs (green).
    - Some adapters contain errors (blue).

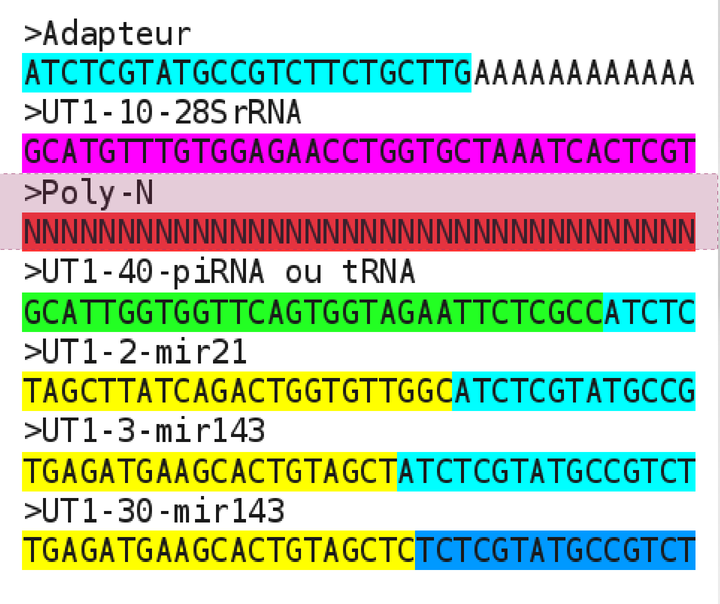


```
>Adapteur
ATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAA
>UT1-10-28S rRNA
GCATGTTTGTGGAGAACCTGGTGCTAAATCACTCGT
>Poly-N
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
>UT1-40-piRNA ou tRNA
GCATTGGTGGTTCAGTGGTAGAATTCTCGCCATCTC
>UT1-2-mir21
TAGCTTATCAGACTGGTGTTGGCATCTCGTATGCCG
>UT1-3-mir143
TGAGATGAAGCACTGTAGCTATCTCGTATGCCGTCT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

2. Why cleaning ?

Outputed 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 RNAs (green).
    - Some adapters contain errors (blue).
- Some sequences contain polyN (red)



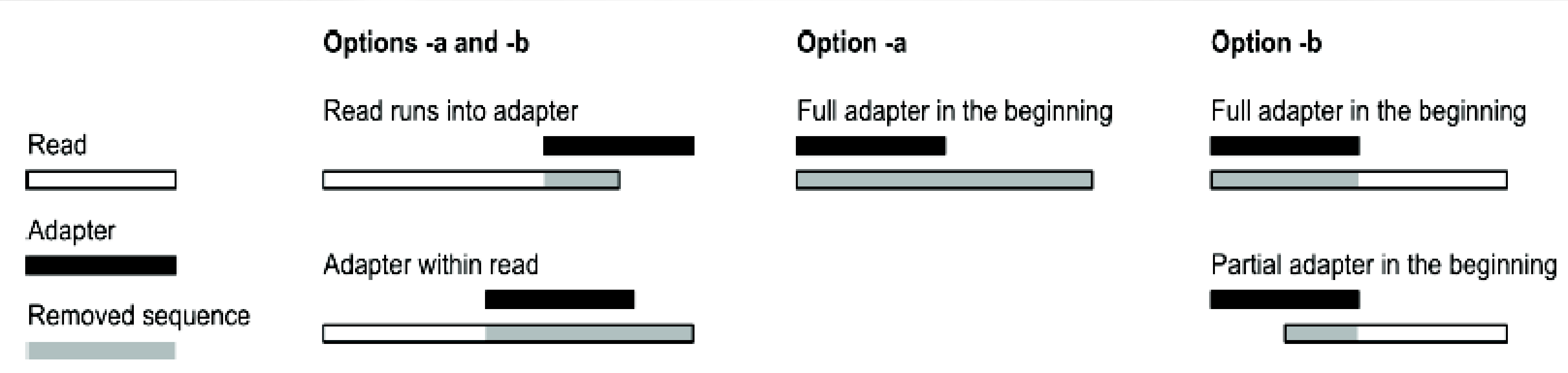
```
>Adapteur
ATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAA
>UT1-10-28S rRNA
GCATGTTTGTGGAGAACCTGGTGCTAAATCACTCGT
>Poly-N
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
>UT1-40-piRNA ou tRNA
GCATTGGTGGTTCAGTGGTAGAATTCTCGCCATCTC
>UT1-2-mir21
TAGCTTATCAGACTGGTGTTGGCATCTCGTATGCCG
>UT1-3-mir143
TGAGATGAAGCACTGTAGCTATCTCGTATGCCGTCT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

2. Cleaning

• Adapters removing and length filtering

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

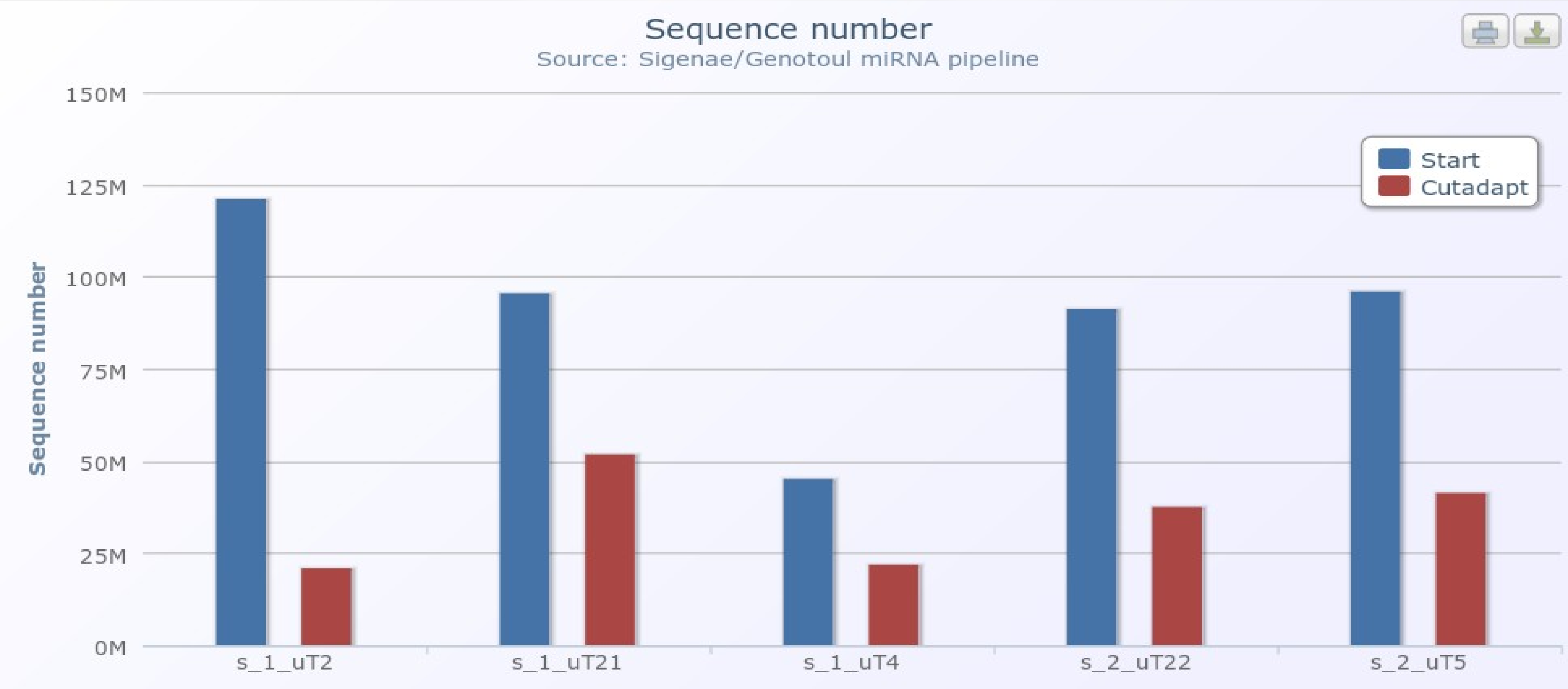
Cutadapt removes adapter sequences from high-throughput sequencing data. Indeed, reads are usually longer than the RNA, and therefore contain parts of the 3’ adapter. It also allows to keep only sequences of desired length (15<length<29).



```
cutadapt -a ATCTCGTATGCCGTCTTCTGCTTG -m 15 -M 29 -o nf_out.fg nf_in.fq
```

2. Cleaning

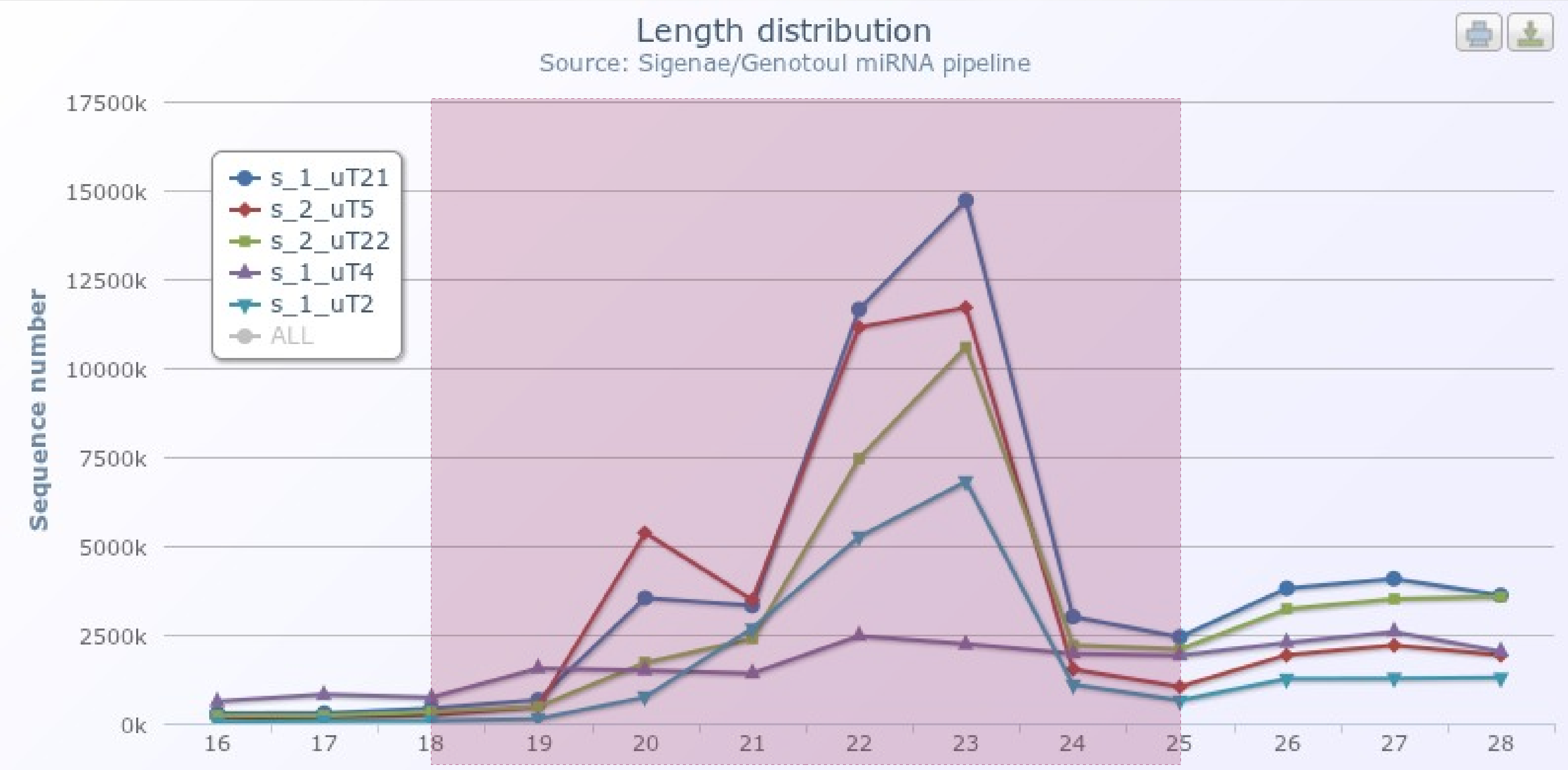
- **56 % of reads discarded**



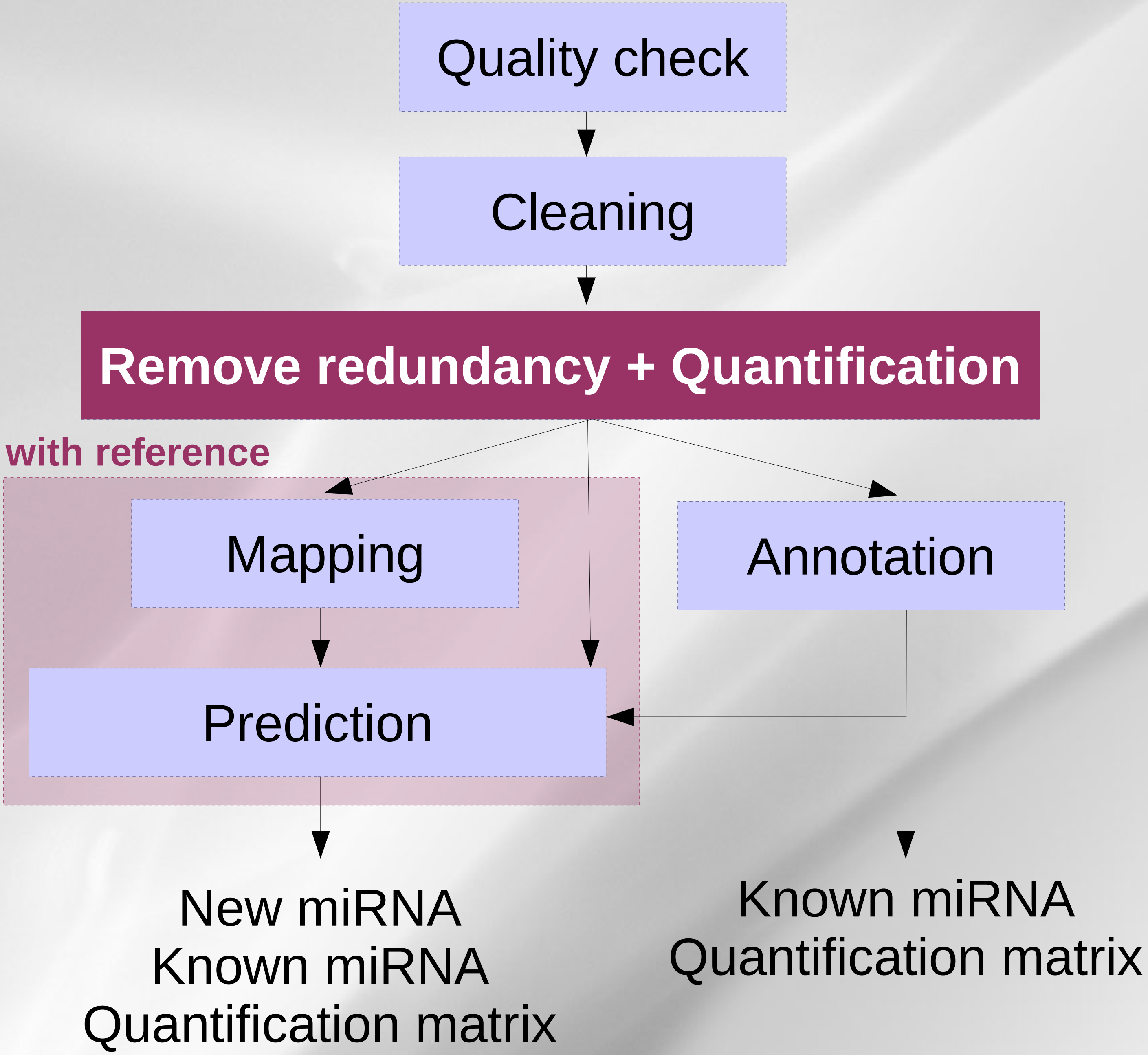
2. Cleaning

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

  → miRNA ?



small RNAseq pipeline



3. Remove redundancy

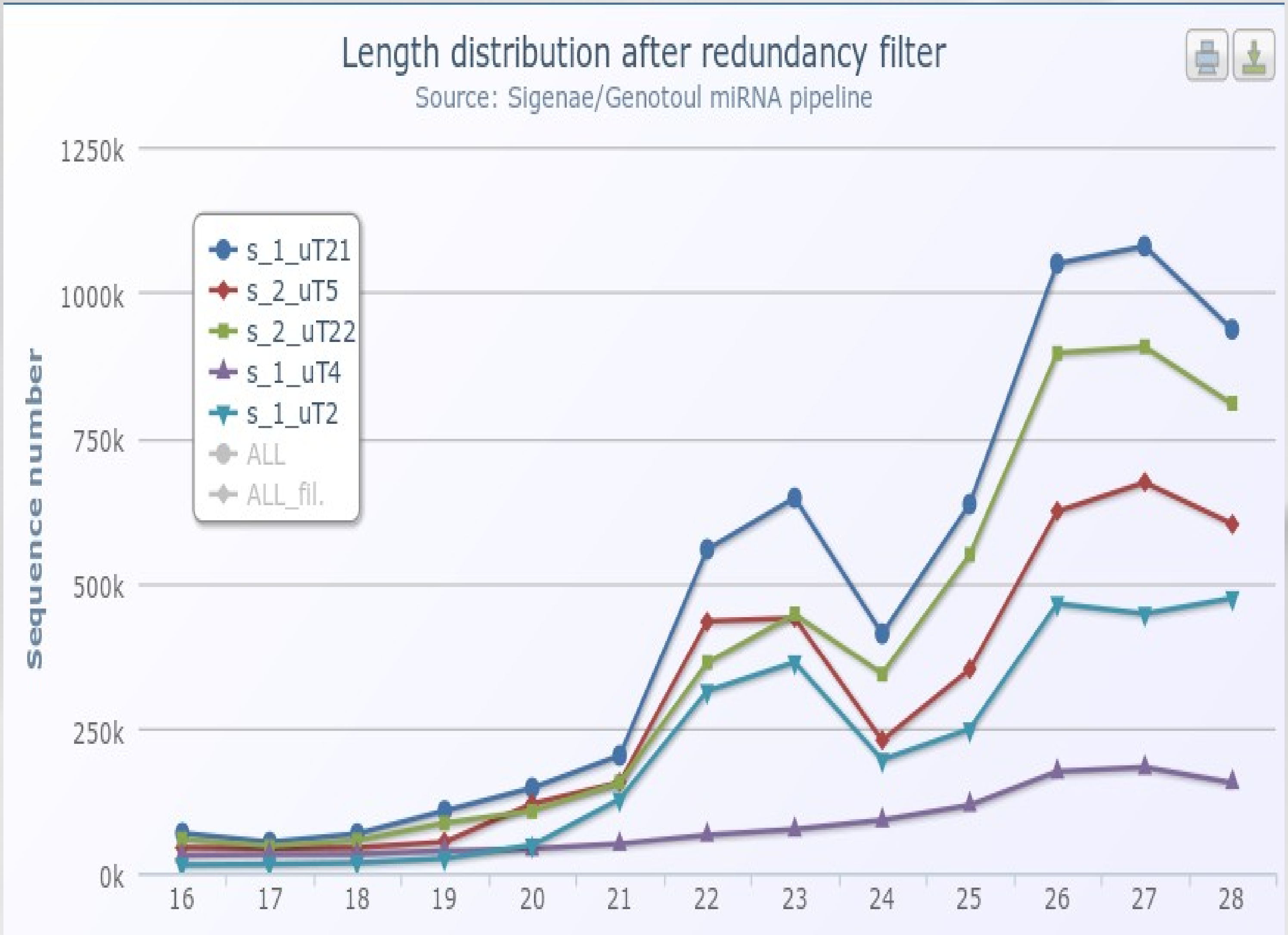
- **Removing identical reads**
  - **save computational time**
  - **useless to keep all the read**
  - **Keep the number of occurrence for each reads**

```
...
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
...
```

```
fastqnr.pl sample.fq | sort -k1,1 >  sample.matrix
```

Plateforme Bioinformatique Midi-Pyrénées

3. Remove redundancy

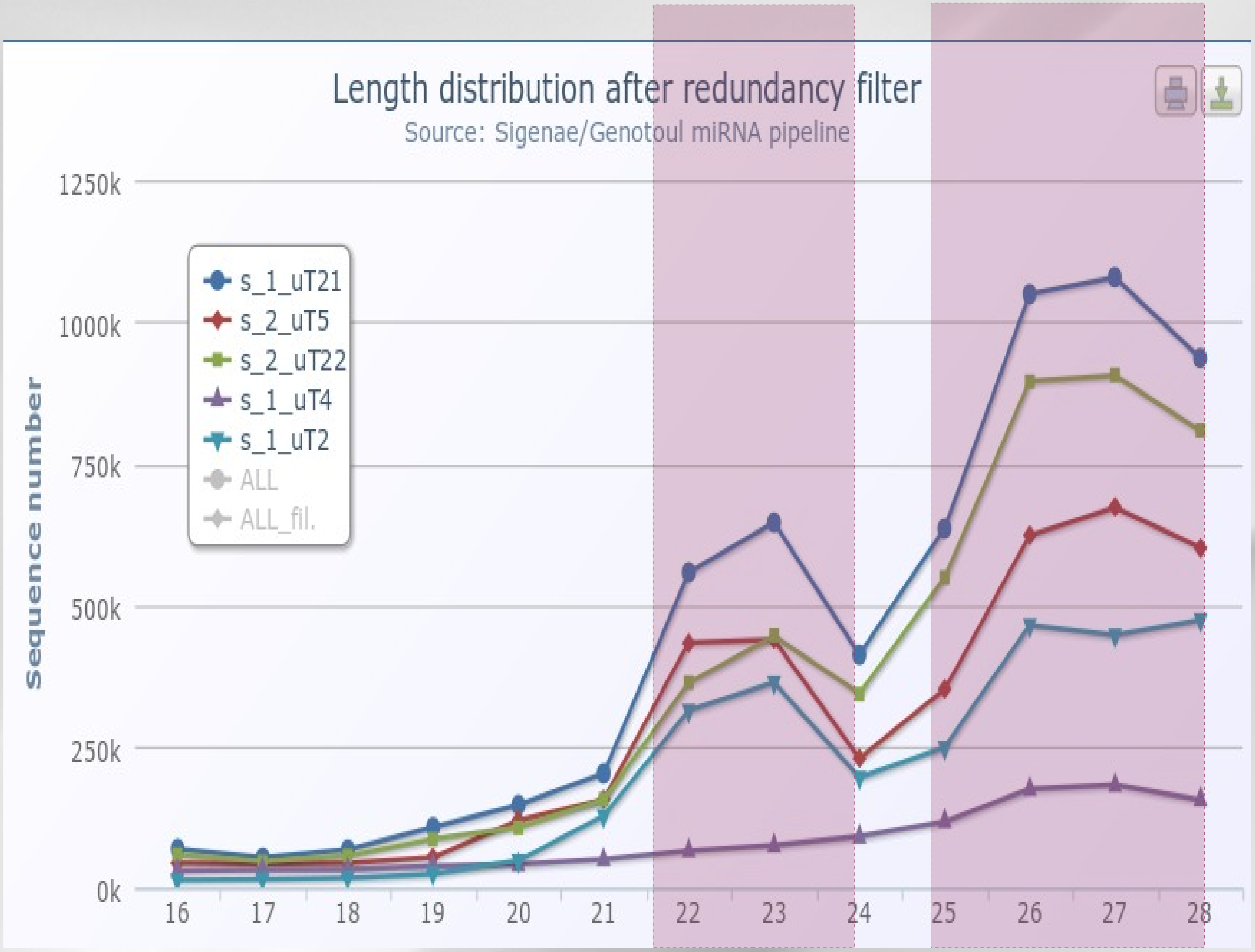


Plateforme Bioinformatique Midi-Pyrénées

3. Remove redundancy

miRNA ?

piRNA ?

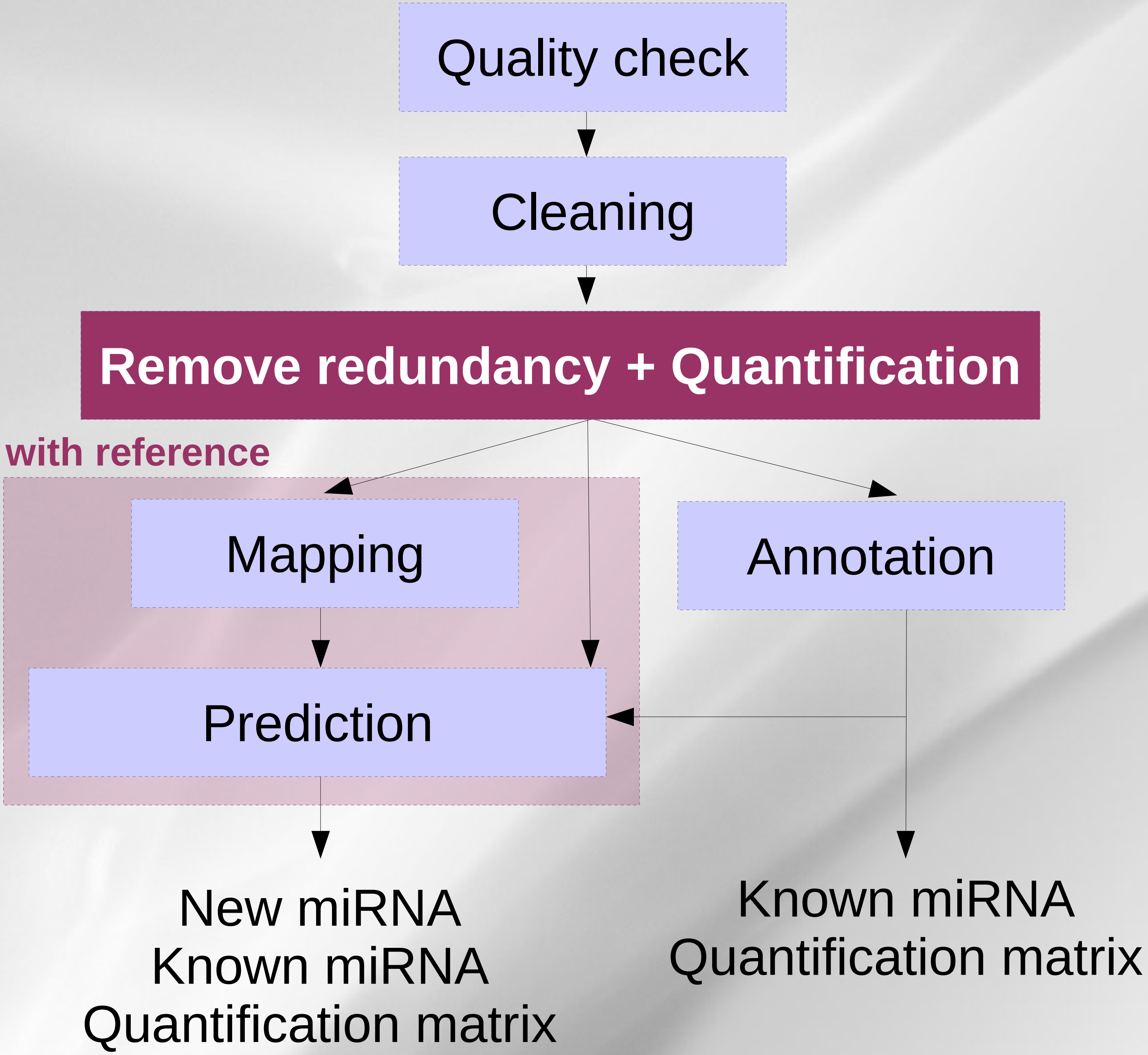


- **More differencies between piRNAs than with miRNAs ?**

Exercice 1:

- Quality control
- Cleaning
- Remove redundancy

small RNAseq pipeline



Quality check

Cleaning

Remove redundancy + Quantification

with reference

Mapping

Annotation

Prediction

New miRNA
Known miRNA
Quantification matrix

Known miRNA
Quantification matrix

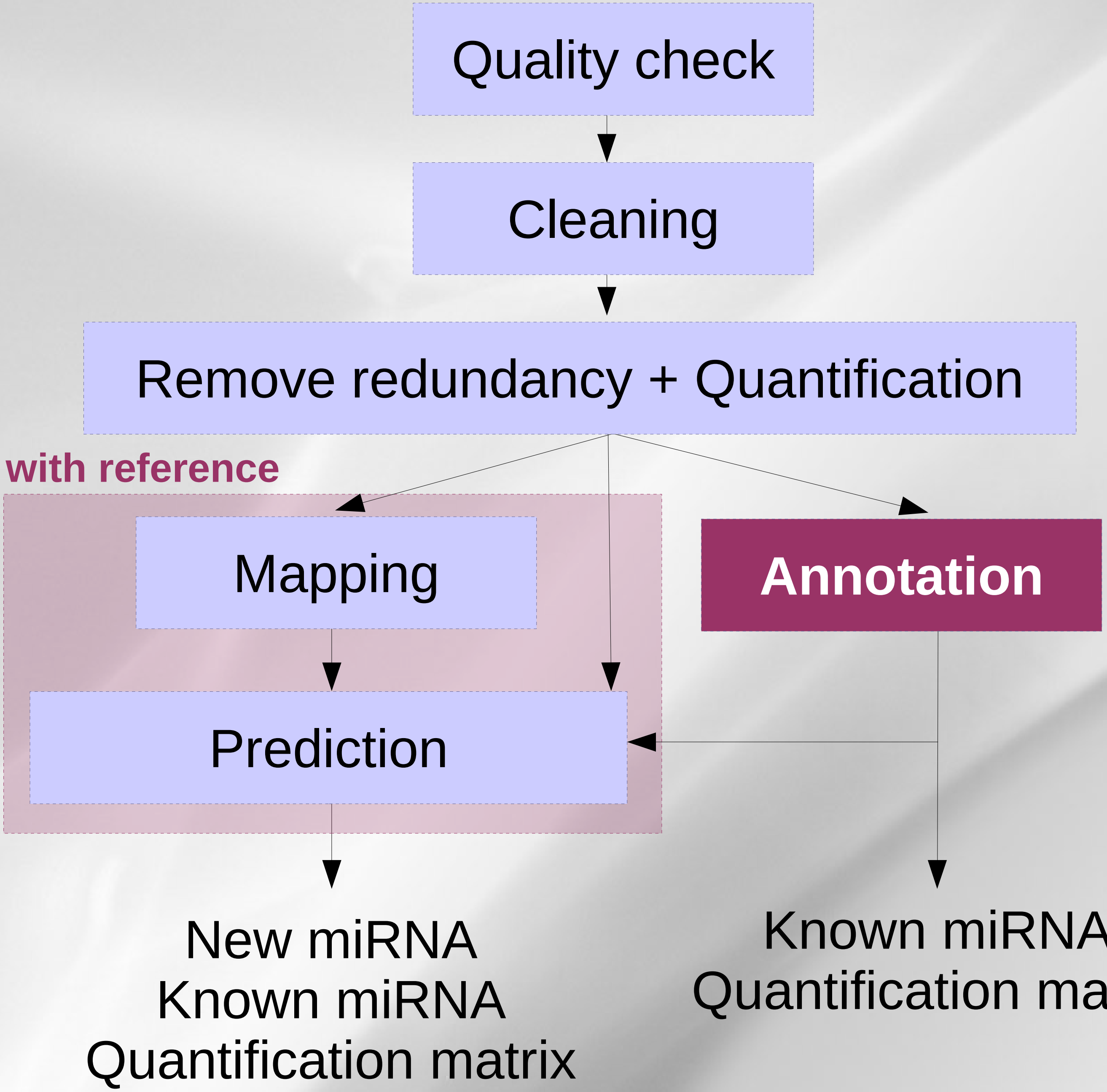
3. Quantification

- **Computes an expression matrix**
  - **Read must be at least in 2 samples if present less than 5 times**

```
#seq                            s_1_uT21   s_1_uT2   s_1_uT4   s_2_uT22   s_2_uT5
...
AAAAGGGCTGTTTGTGCAGGCAG         87         14        0         85         5
AAAAGGGCTGTTTGTGCAGGCAGA        1          0         0         1          0
AAAAGGGCTGTTTGTGCAGGCAGG        1          0         0         2          0
AAAAGGGCTGTTTGTGCAGGCAGT        1          0         0         3          0
AAAAGGGCTGTTTGTGCAGGCAGTTT      0          0         0         0          1
AAAAGGGCTGTTTGTGCAGGCAT         1          2         0         3          0
AAAAGGGCTGTTTGTGCAGGCTA         0          0         0         1          0
AAAAGGGCTGTTTGTGCAGGCTG         1          0         0         1          0
AAAAGGGCTGTTTGTGCAGGCTT         1          0         0         0          0
AAAAGGGCTGTTTGTGCAGGG           6          1         0         4          2
AAAAGGGCTGTTTGTGCAGGGA          11         1         0         3          4
AAAAGGGCTGTTTGTGCAGGGAG         88         9         0         62         11
AAAAGGGCTGTTTGTGCAGGGAGC        1          0         0         0          0
AAAAGGGCTGTTTGTGCAGGGAGCTGA     0          0         0         1          0
AAAAGGGCTGTTTGTGCAGGGAGT        0          1         0         0          0
AAAAGGGCTGTTTGTGCAGGGAGTT       0          0         0         1          0
AAAAGGGCTGTTTGTGCAGGGAT         2          0         0         0          1
AAAAGGGCTGTTTGTGCAGGGATT        1          0         0         0          0
...
```

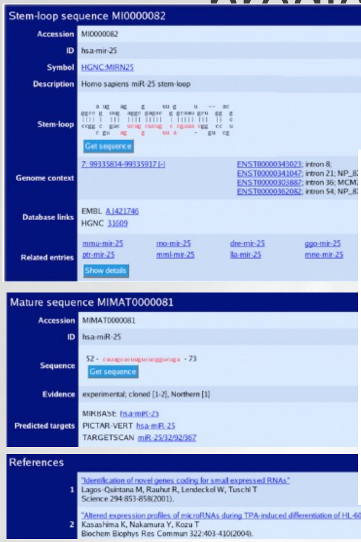
```
quatification.pl  -i 2 -a 5 sample1.matrix sample2.matrix ... > quantification.matrix
```

small RNAseq pipeline



4. Annotation

- Useful databases:
  - – miRbase (http://microrna.sanger.ac.uk/)
    - 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	(structure shown below)
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

```
    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 cguaa cgg  cc  u
    c gu   ag    g     uu a     -   gu  cg
```

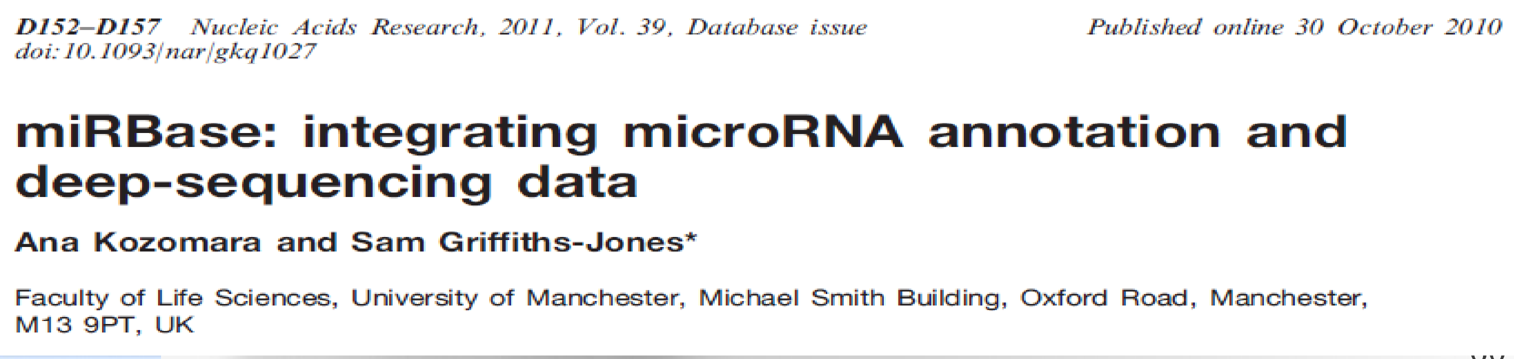
Mature sequence MIMAT0000081

Accession	MIMAT0000081
ID	hsa-miR-25
Sequence	52 - cauugcacuugucucggucuga - 73
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

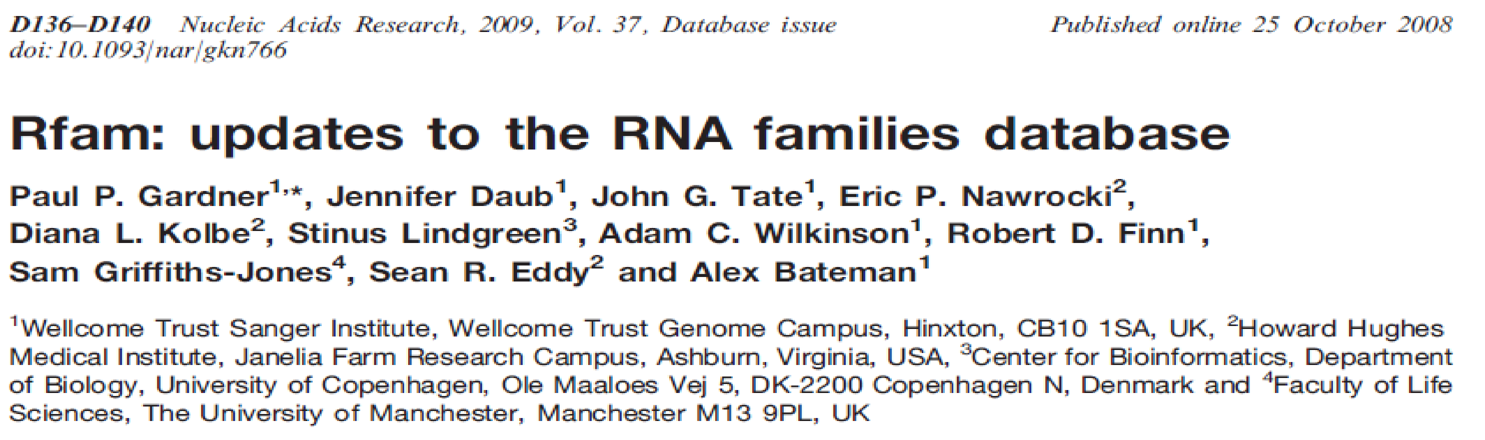
miRBase: integrating microRNA annotation and deep-sequencing data

Ana Kozomara and Sam Griffiths-Jones*

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

4. 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

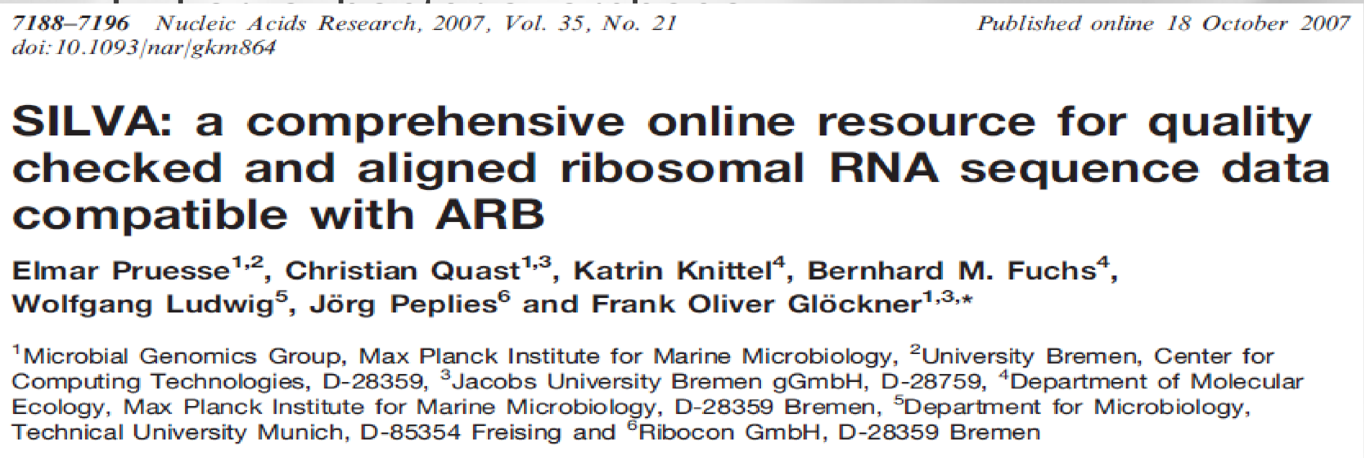
Rfam: updates to the RNA families database

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

4. 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)
      - Eukarya, bacteria, archaea



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

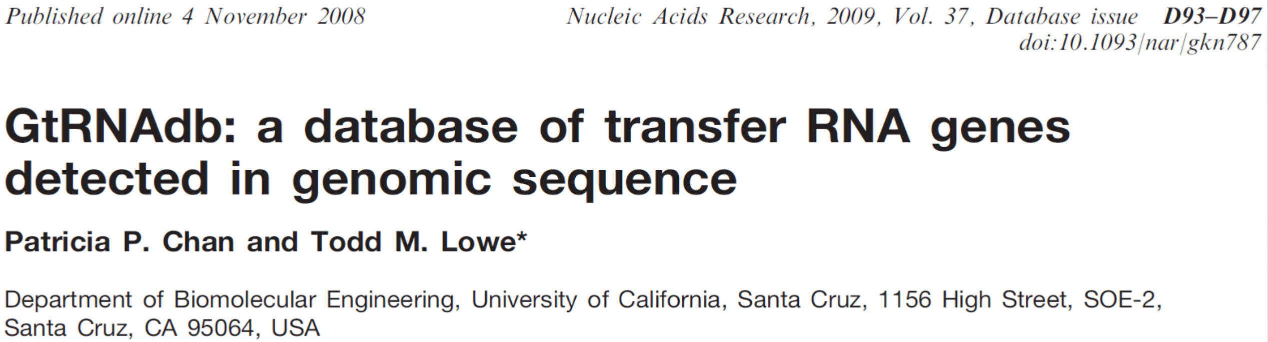
SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB

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

4. Annotation

- Useful databases:
  - – miRbase (http://microrna.sanger.ac.uk/)
  - – Rfam (http://rfam.sanger.ac.uk/)
  - – Silva (http://www.arb-silva.de/)
  - – GtRNAdb(http://gtrnadb.ucsc.edu/)
    - 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

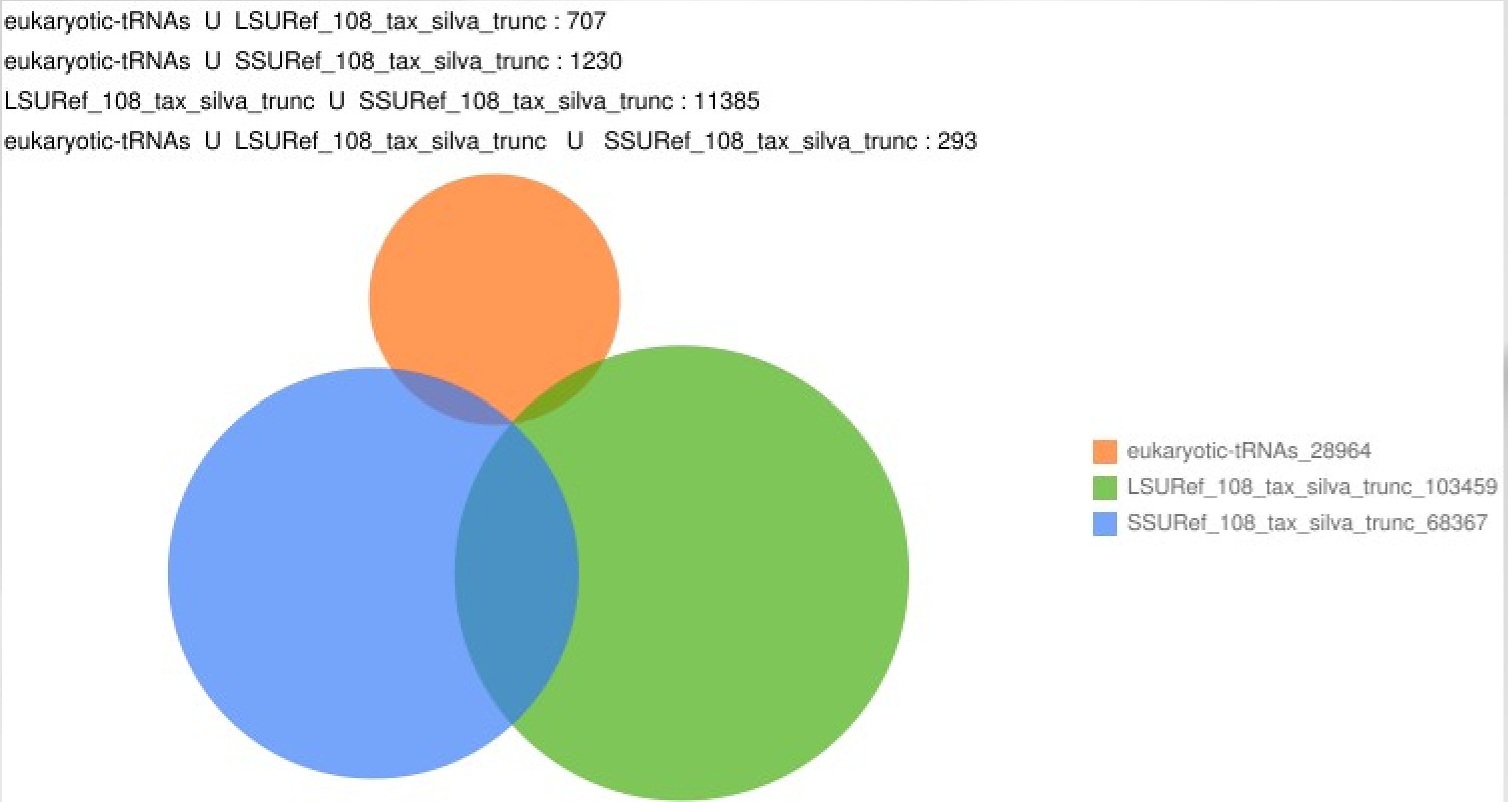
GtRNAdb: a database of transfer RNA genes detected in genomic sequence

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

4. Annotation

- **Reads with multiple annotation**



eukaryotic-tRNAs U LSURef_108_tax_silva_trunc : 707

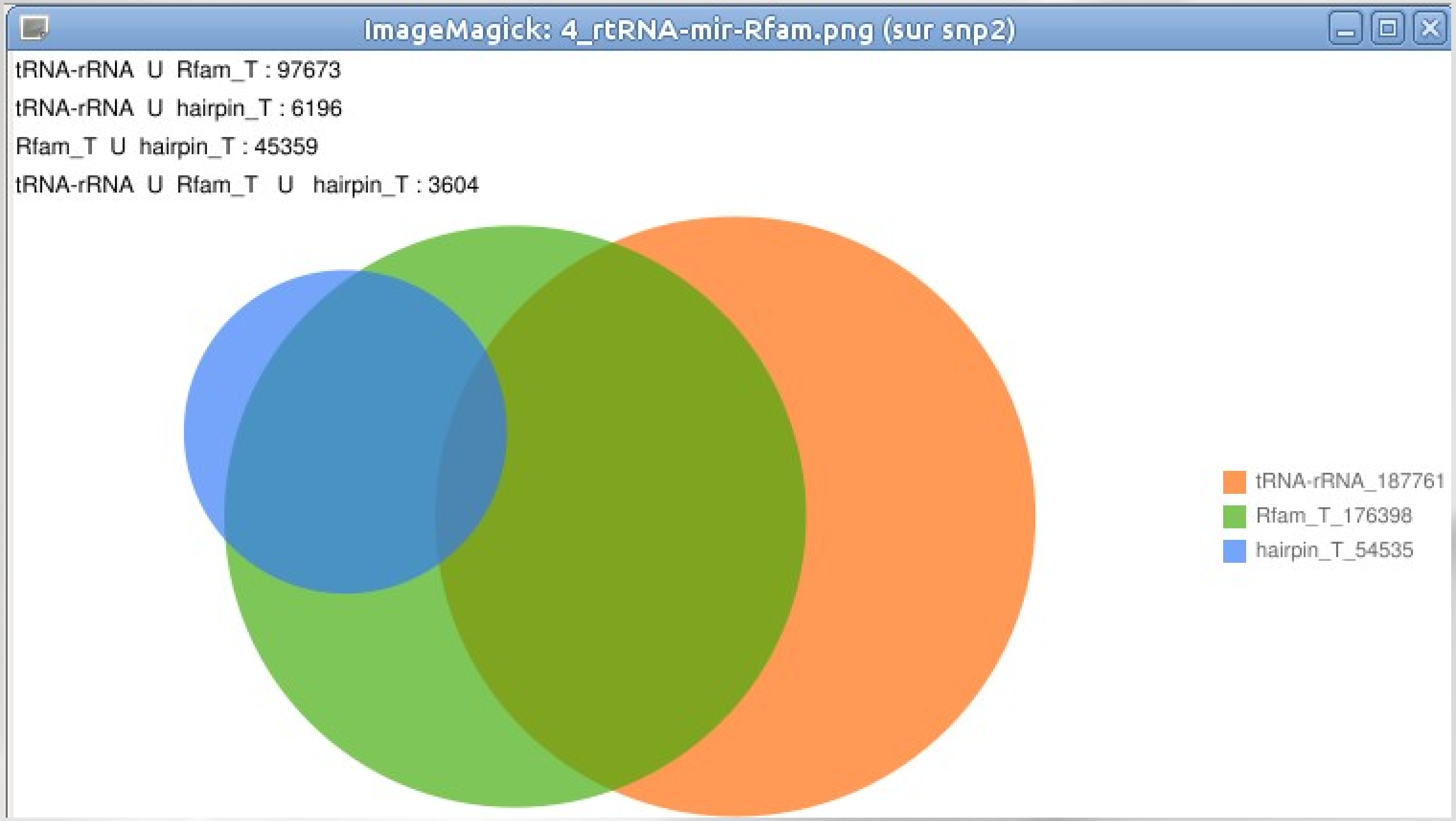
eukaryotic-tRNAs U SSURef_108_tax_silva_trunc : 1230

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

4. Annotation

- **Reads with multiple annotation**



tRNA-rRNA U Rfam_T : 97673

tRNA-rRNA U hairpin_T : 6196

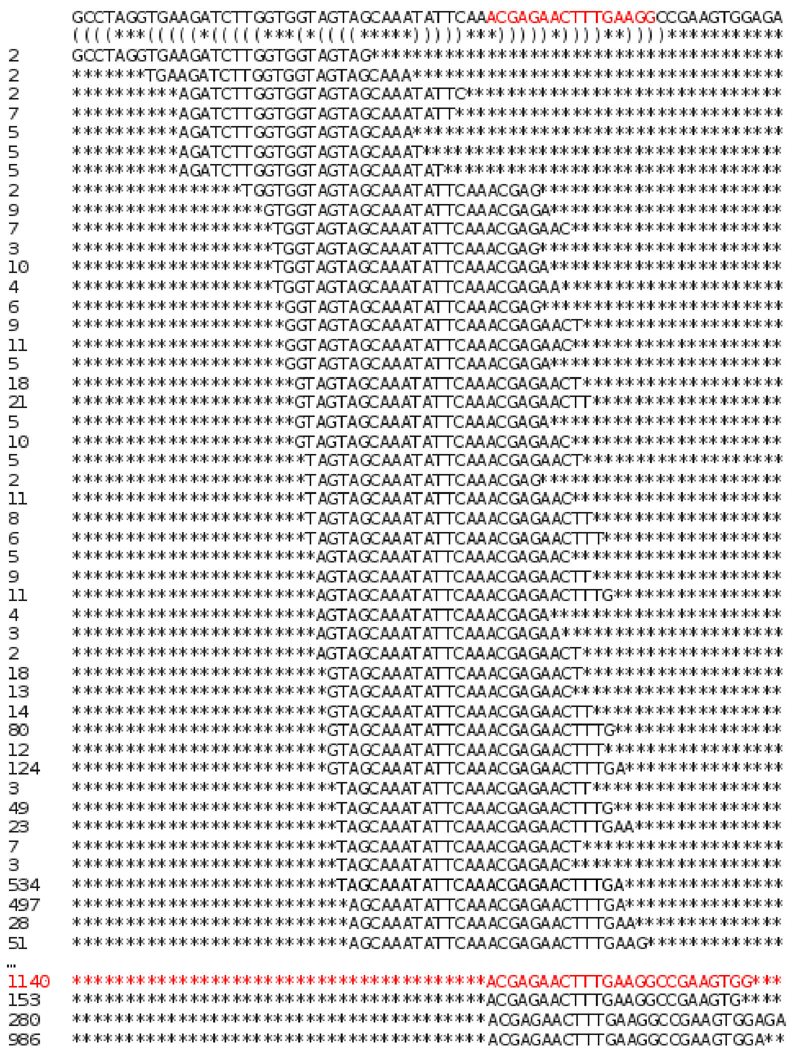
Rfam_T U hairpin_T : 45359

tRNA-rRNA U Rfam_T U hairpin_T : 3604

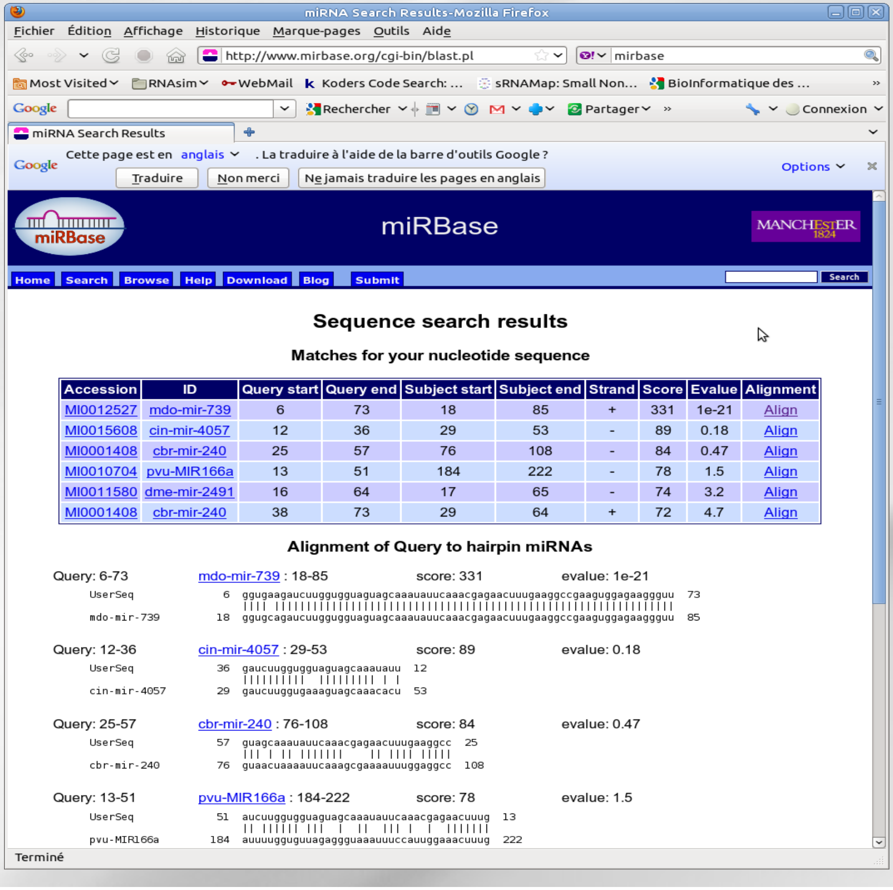
→ A lot of reads annotated with mirBase but also with tRNA and rRNA database

- **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

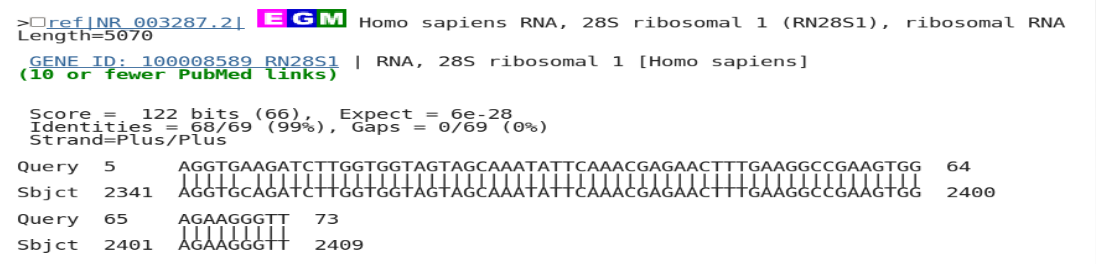
```
UserSeq         36  gaucuuggugguaguagcaaauauu  12
                    ||||||||||  ||||||||| | |
cin-mir-4057    29  gaucuuggugaaaguagcaaacacu  53
```

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  auuuuggugguuagaggguaaauuuccauuggaaacuuug  222
```

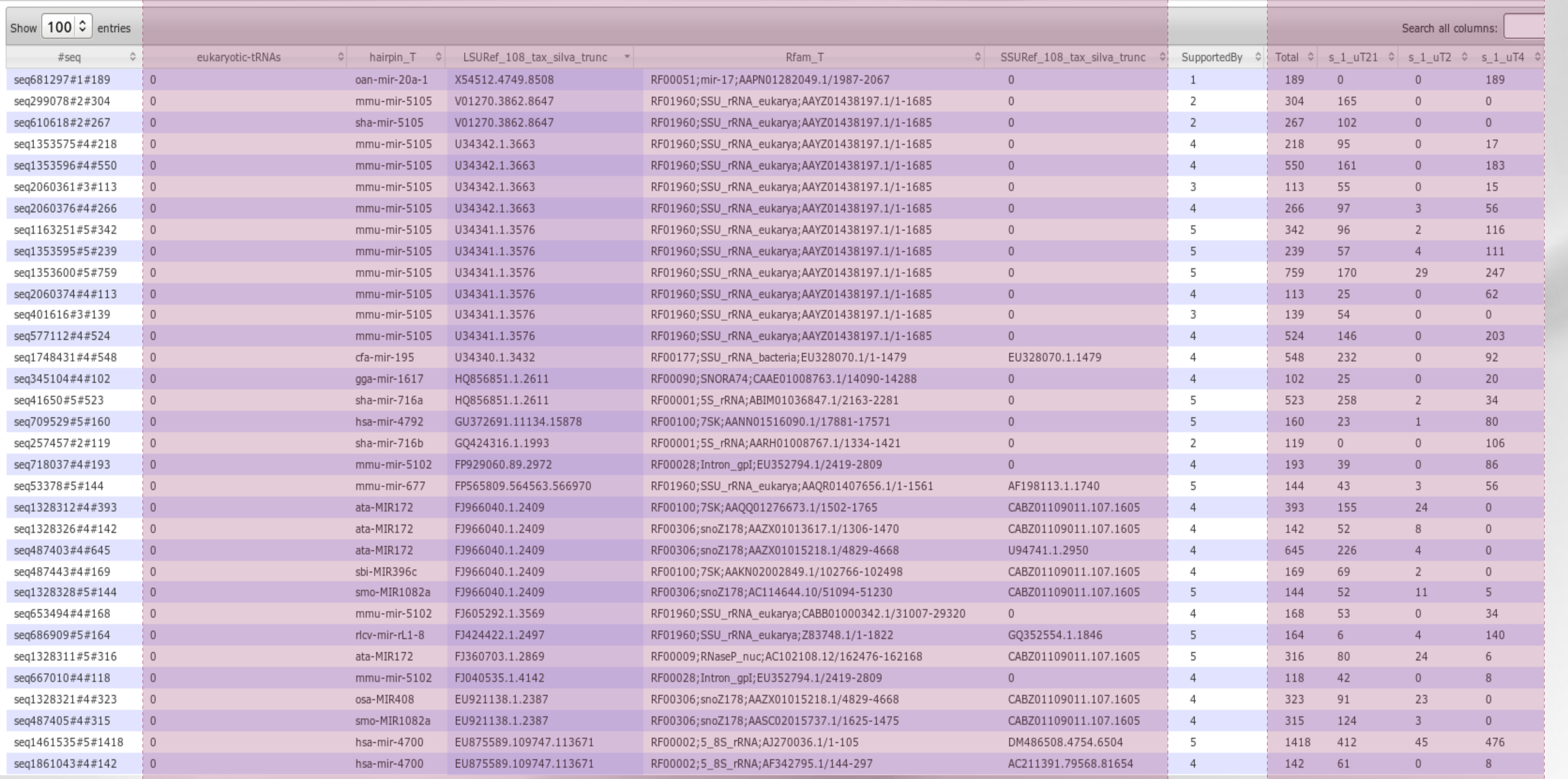


```
>□ref|NR_003287.2|  Homo sapiens RNA, 28S ribosomal 1 (RN28S1), ribosomal RNA
Length=5070

 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
```

4. Annotation

Annotation

occurences



Show 100 entries

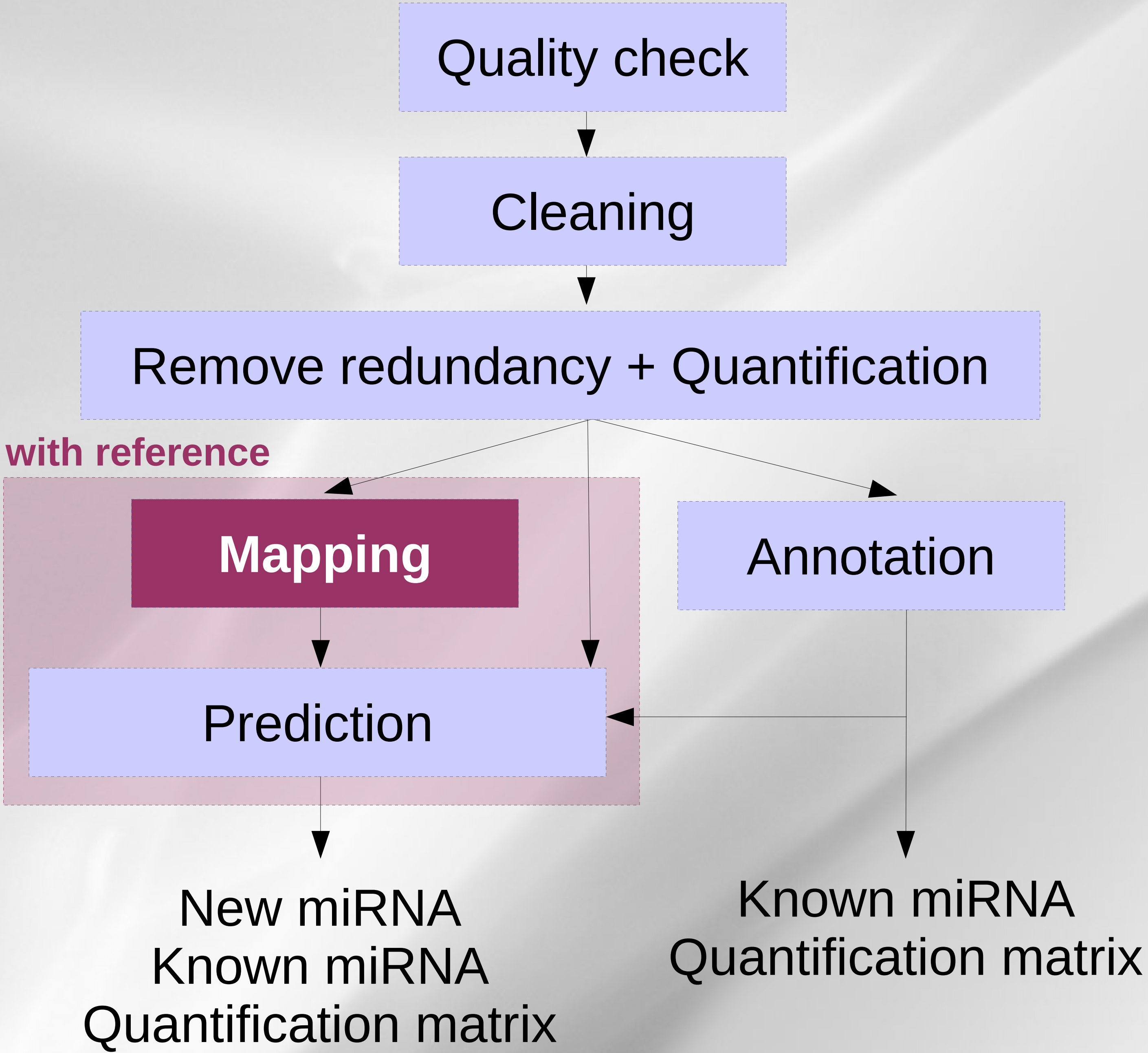
Search all columns:

#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 2:

- Annotation

small RNAseq pipeline

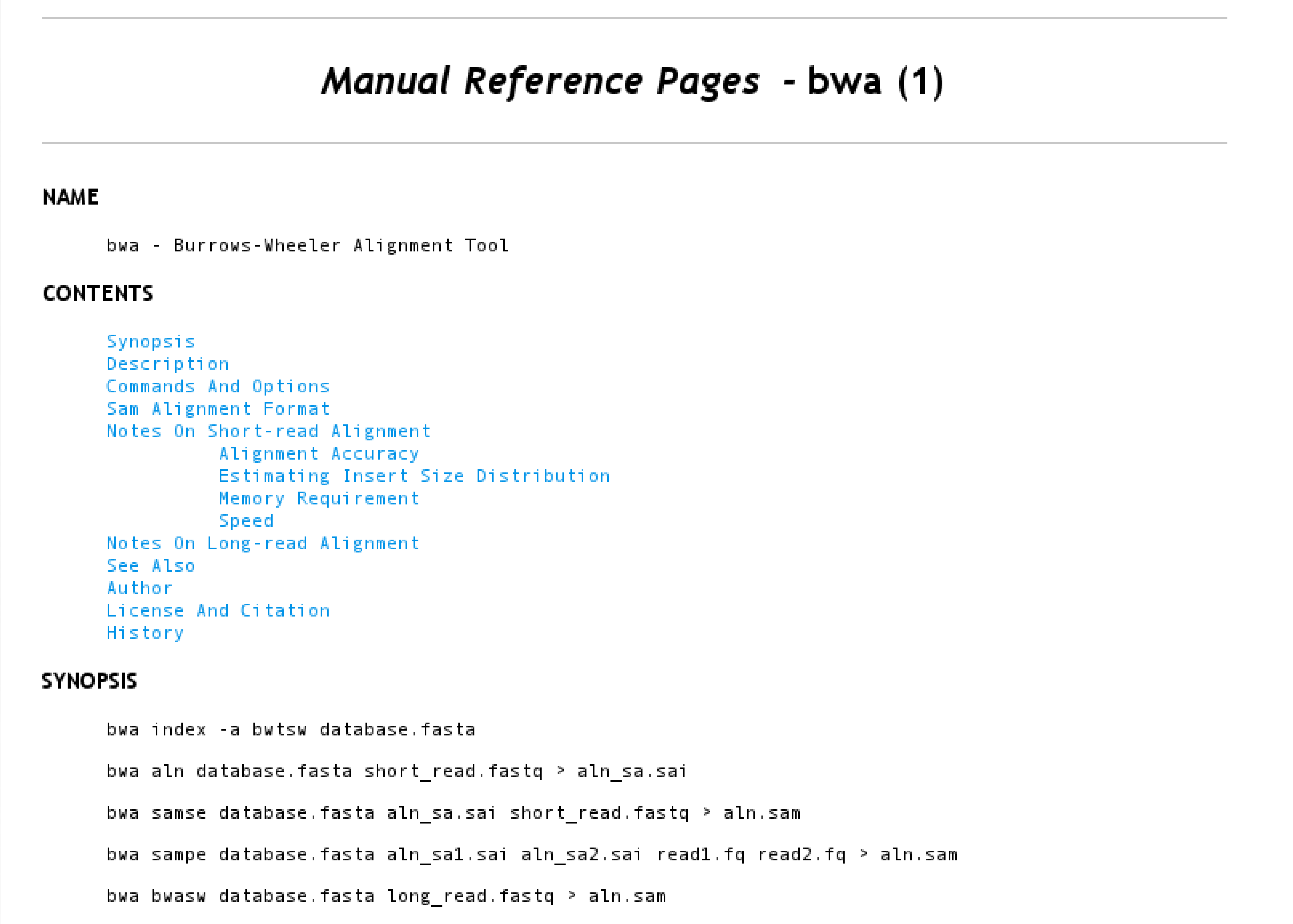


5. Mapping the reads

- 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**
- ...

Plateforme Bioinformatique Midi-Pyrénées

5. Mapping the reads with bwa



Manual Reference Pages - bwa (1)

NAME

```
bwa - Burrows-Wheeler Alignment Tool
```

CONTENTS

SYNOPSIS

```
bwa index -a bwtsw database.fasta

bwa aln database.fasta short_read.fastq > aln_sa.sai

bwa samse database.fasta aln_sa.sai short_read.fastq > aln.sam

bwa sampe database.fasta aln_sa1.sai aln_sa2.sai read1.fq read2.fq > aln.sam

bwa bwasw database.fasta long_read.fastq > aln.sam
```

5. Mapping the reads with bwa

- Reference sequence indexing:

```
bwa index -a bwtsw db.fasta
```

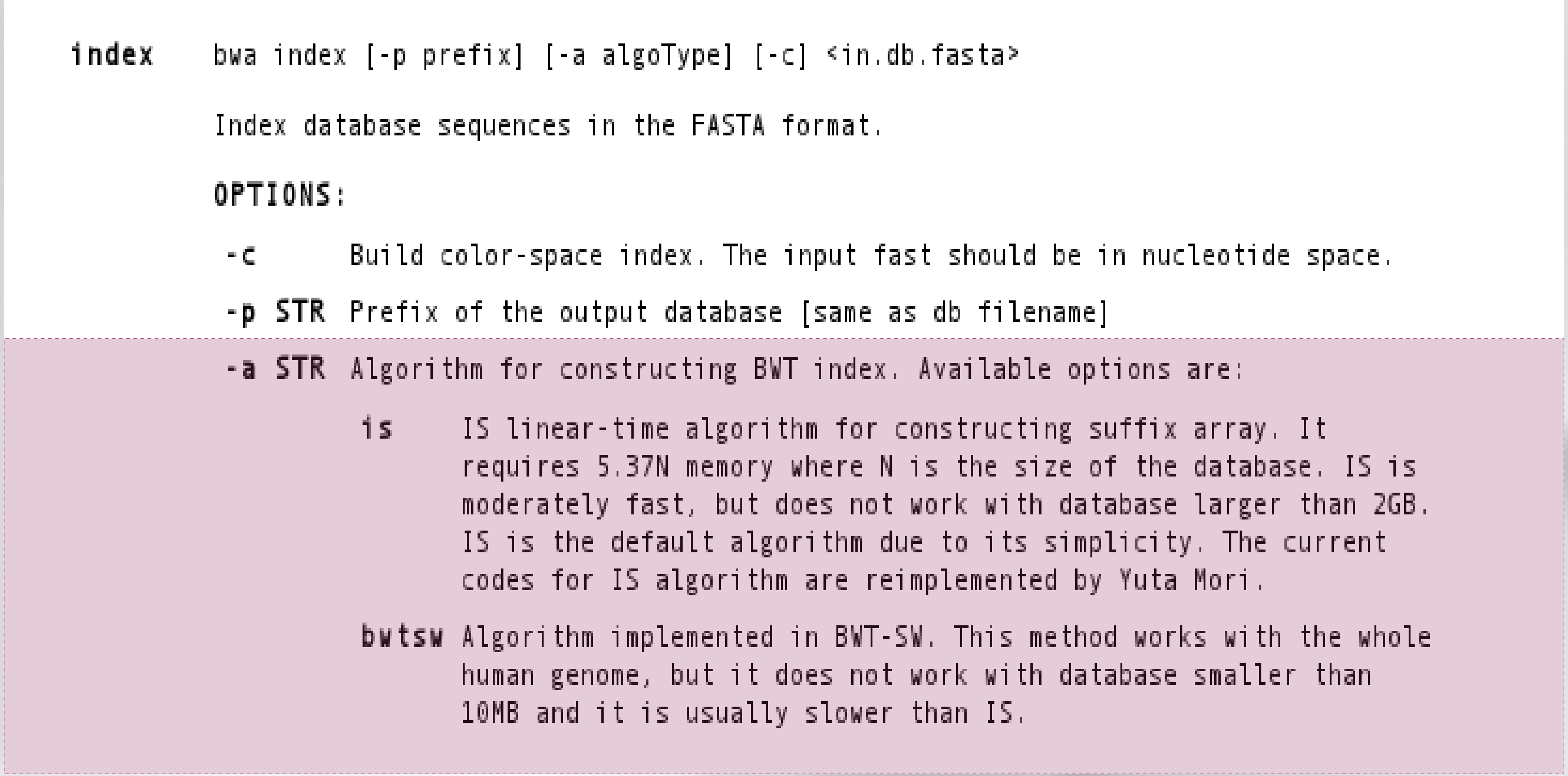
- Read alignment:

```
bwa aln db.fasta short_read.fastq > short_read.sai
```

- Formatting reads:

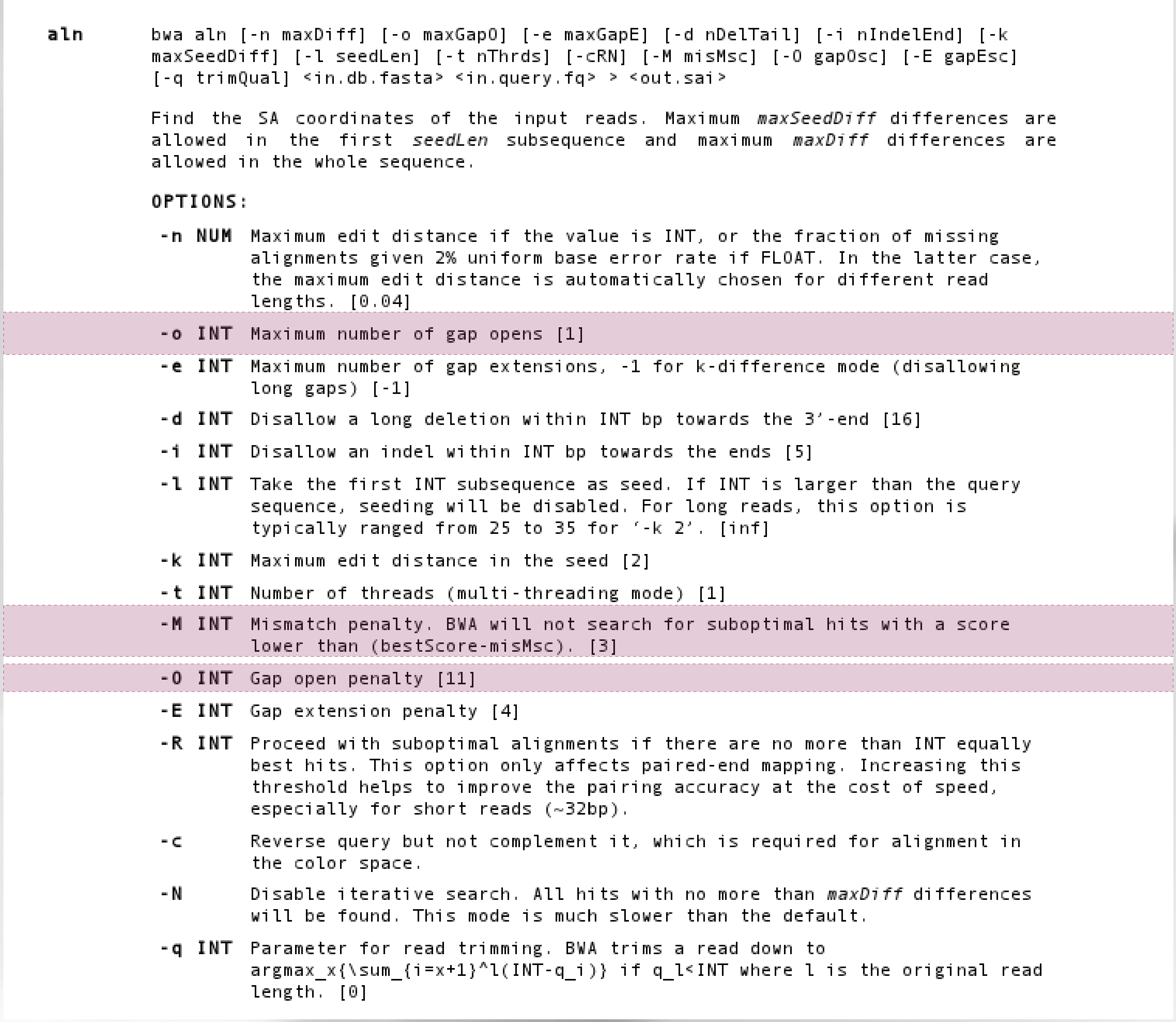
```
bwa samse db.fasta short_read.sai short_read.fastq > short_read.sam
```

5. Mapping the reads with bwa



```
index    bwa index [-p prefix] [-a algoType] [-c] <in.db.fasta>

         Index database sequences in the FASTA format.

         OPTIONS:

          -c     Build color-space index. The input fast should be in nucleotide space.

          -p STR Prefix of the output database [same as db filename]

          -a STR Algorithm for constructing BWT index. Available options are:

                 is    IS linear-time algorithm for constructing suffix array. It
                       requires 5.37N memory where N is the size of the database. IS is
                       moderately fast, but does not work with database larger than 2GB.
                       IS is the default algorithm due to its simplicity. The current
                       codes for IS algorithm are reimplemented by Yuta Mori.

                 bwtsw Algorithm implemented in BWT-SW. This method works with the whole
                       human genome, but it does not work with database smaller than
                       10MB and it is usually slower than IS.
```

5. Mapping the reads with bwa



aln

```
bwa aln [-n maxDiff] [-o maxGapO] [-e maxGapE] [-d nDelTail] [-i nIndelEnd] [-k
maxSeedDiff] [-l seedLen] [-t nThrds] [-cRN] [-M misMsc] [-O gapOsc] [-E gapEsc]
[-q trimQual] <in.db.fasta> <in.query.fq> > <out.sai>
```

Find the SA coordinates of the input reads. Maximum *maxSeedDiff* differences are allowed in the first *seedLen* subsequence and maximum *maxDiff* differences are allowed in the whole sequence.

OPTIONS:

-n NUM Maximum edit distance if the value is INT, or the fraction of missing alignments given 2% uniform base error rate if FLOAT. In the latter case, the maximum edit distance is automatically chosen for different read lengths. [0.04]

-o INT Maximum number of gap opens [1]

-e INT Maximum number of gap extensions, -1 for k-difference mode (disallowing long gaps) [-1]

-d INT Disallow a long deletion within INT bp towards the 3'-end [16]

-i INT Disallow an indel within INT bp towards the ends [5]

-l INT Take the first INT subsequence as seed. If INT is larger than the query sequence, seeding will be disabled. For long reads, this option is typically ranged from 25 to 35 for '-k 2'. [inf]

-k INT Maximum edit distance in the seed [2]

-t INT Number of threads (multi-threading mode) [1]

-M INT Mismatch penalty. BWA will not search for suboptimal hits with a score lower than (bestScore-misMsc). [3]

-O INT Gap open penalty [11]

-E INT Gap extension penalty [4]

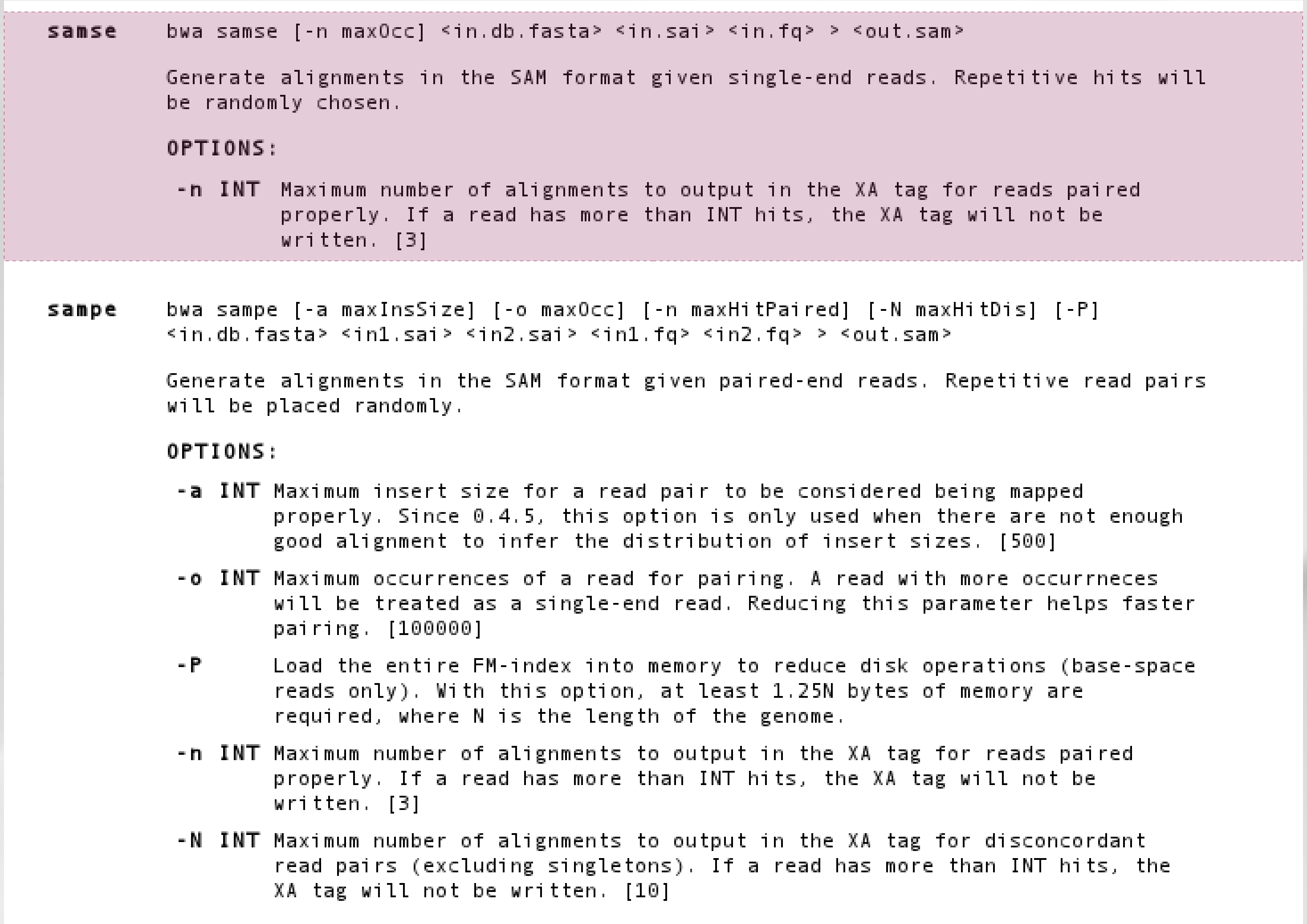
-R INT Proceed with suboptimal alignments if there are no more than INT equally best hits. This option only affects paired-end mapping. Increasing this threshold helps to improve the pairing accuracy at the cost of speed, especially for short reads (~32bp).

-c Reverse query but not complement it, which is required for alignment in the color space.

-N Disable iterative search. All hits with no more than *maxDiff* differences will be found. This mode is much slower than the default.

-q INT Parameter for read trimming. BWA trims a read down to argmax_x{\sum_{i=x+1}^l(INT-q_i)} if q_l<INT where l is the original read length. [0]

5. Mapping the reads with bwa

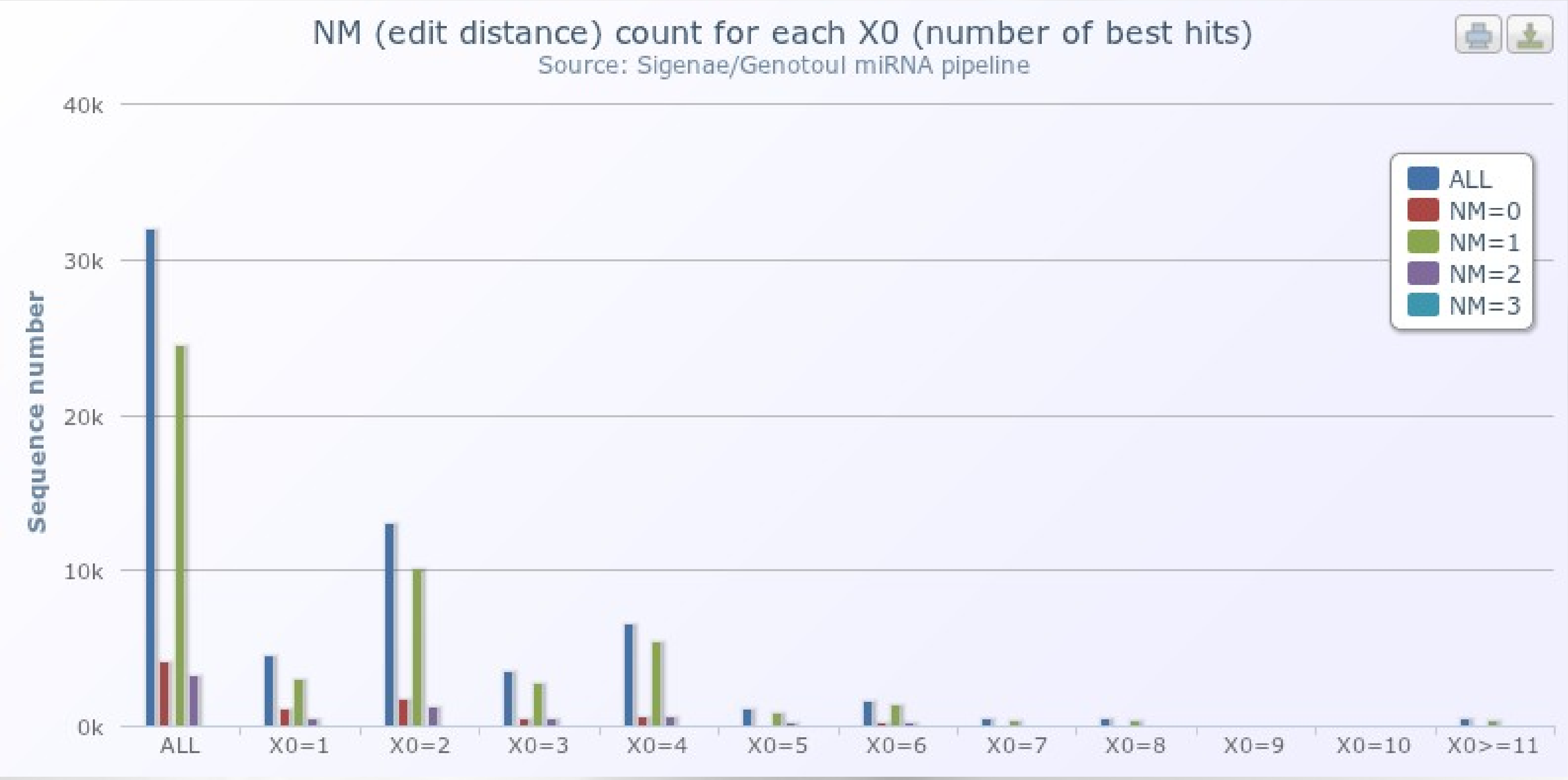


```
samse    bwa samse [-n maxOcc] <in.db.fasta> <in.sai> <in.fq> > <out.sam>

         Generate alignments in the SAM format given single-end reads. Repetitive hits will
         be randomly chosen.

         OPTIONS:

          -n INT Maximum number of alignments to output in the XA tag for reads paired
                 properly. If a read has more than INT hits, the XA tag will not be
                 written. [3]

sampe    bwa sampe [-a maxInsSize] [-o maxOcc] [-n maxHitPaired] [-N maxHitDis] [-P]
         <in.db.fasta> <in1.sai> <in2.sai> <in1.fq> <in2.fq> > <out.sam>

         Generate alignments in the SAM format given paired-end reads. Repetitive read pairs
         will be placed randomly.

         OPTIONS:

          -a INT Maximum insert size for a read pair to be considered being mapped
                 properly. Since 0.4.5, this option is only used when there are not enough
                 good alignment to infer the distribution of insert sizes. [500]

          -o INT Maximum occurrences of a read for pairing. A read with more occurrneces
                 will be treated as a single-end read. Reducing this parameter helps faster
                 pairing. [100000]

          -P     Load the entire FM-index into memory to reduce disk operations (base-space
                 reads only). With this option, at least 1.25N bytes of memory are
                 required, where N is the length of the genome.

          -n INT Maximum number of alignments to output in the XA tag for reads paired
                 properly. If a read has more than INT hits, the XA tag will not be
                 written. [3]

          -N INT Maximum number of alignments to output in the XA tag for disconcordant
                 read pairs (excluding singletons). If a read has more than INT hits, the
                 XA tag will not be written. [10]
```

Plateforme Bioinformatique Midi-Pyrénées

5. Mapping the reads with bwa

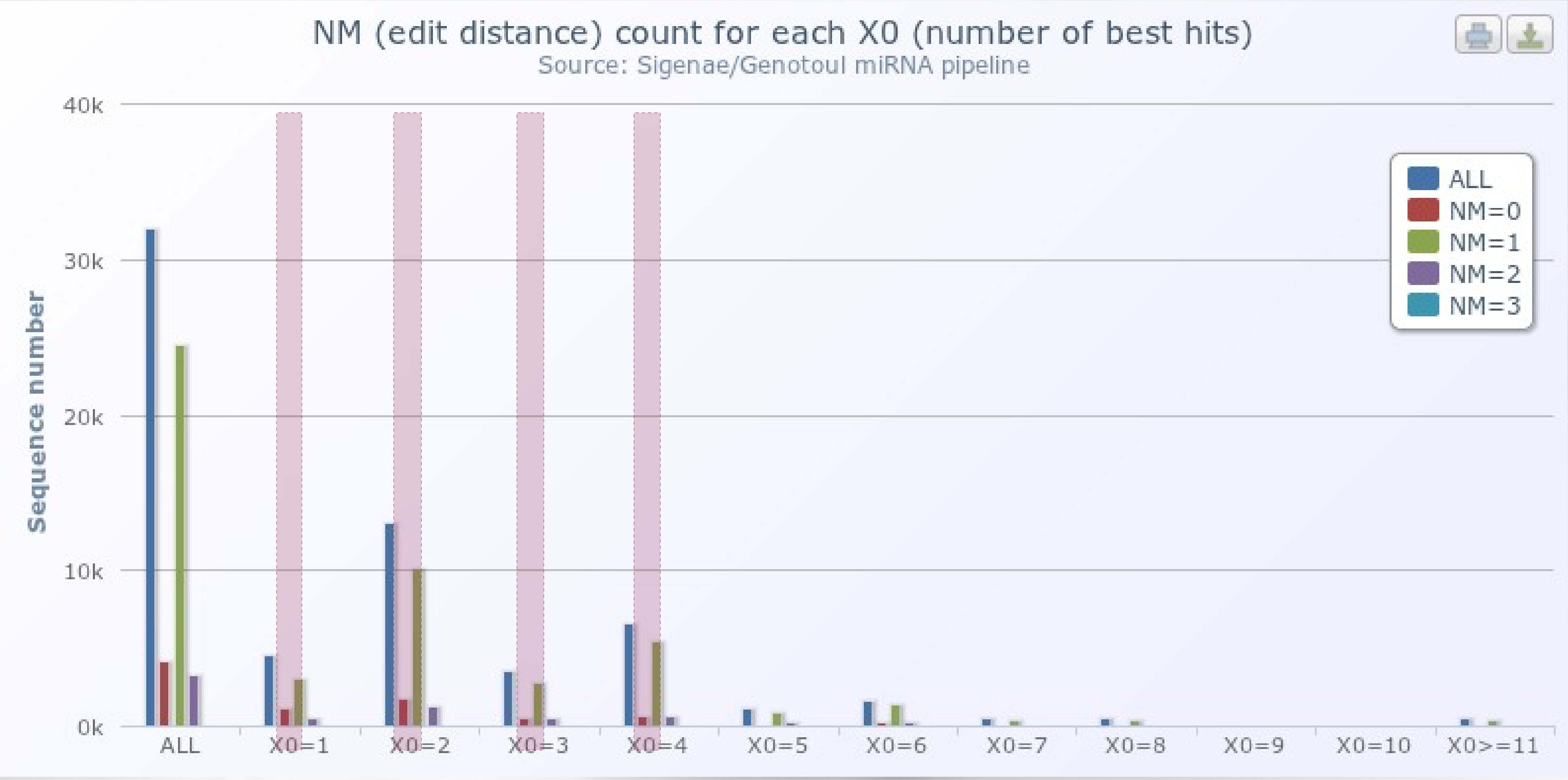
- **Alignement of annotated reads**


NM (edit distance) count for each X0 (number of best hits)

Source: Sigenae/Genotoul miRNA pipeline

5. Mapping the reads with bwa

- **Alignement of annotated reads**


NM (edit distance) count for each X0 (number of best hits)

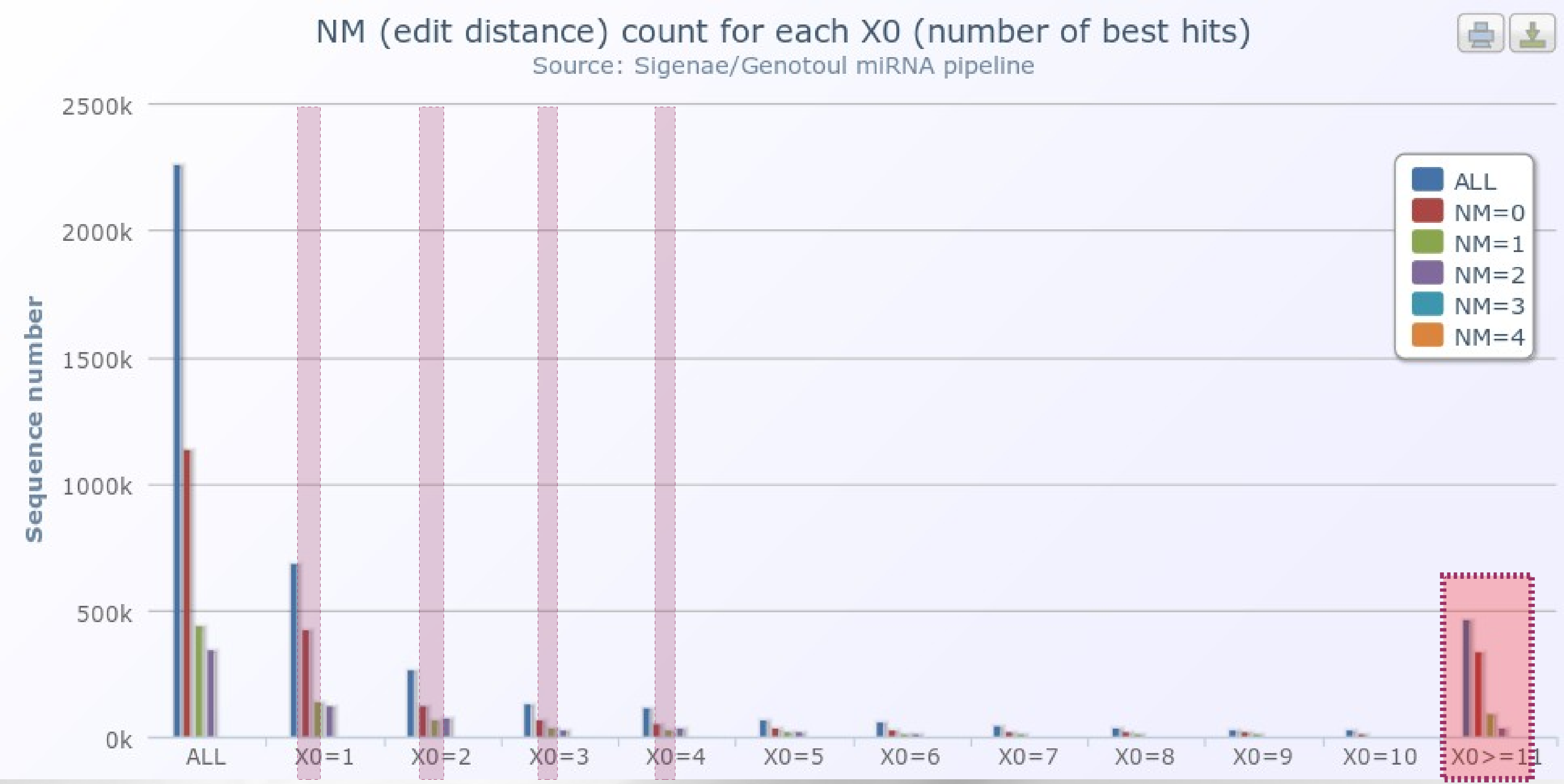
Source: Sigenae/Genotoul miRNA pipeline

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

Plateforme Bioinformatique Midi-Pyrénées

5. Mapping the reads with bwa

- **Alignement of all reads**



NM (edit distance) count for each X0 (number of best hits)

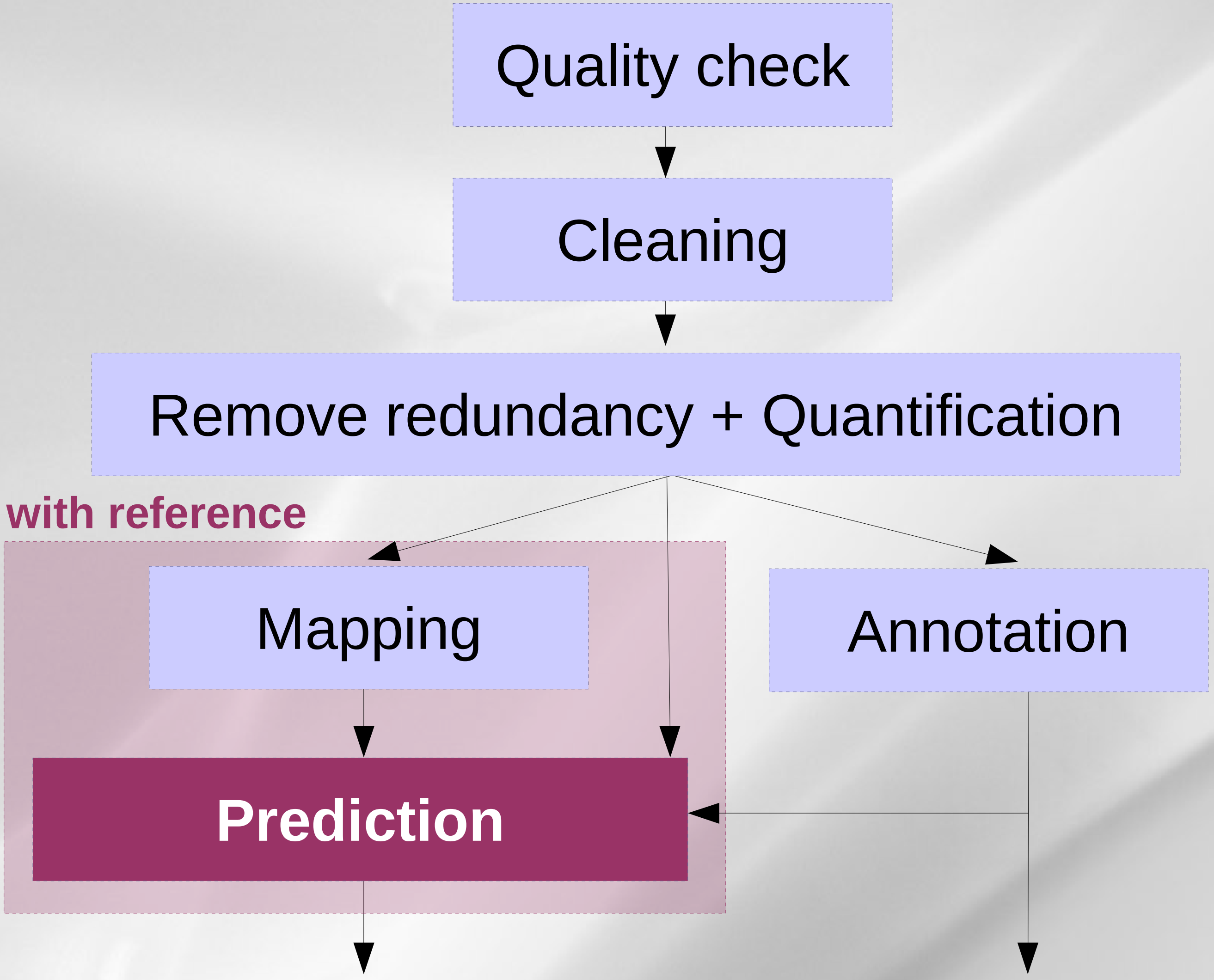
Source: Sigenae/Genotoul miRNA pipeline

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

Exercice 3:

- Mapping the reads

small RNAseq pipeline

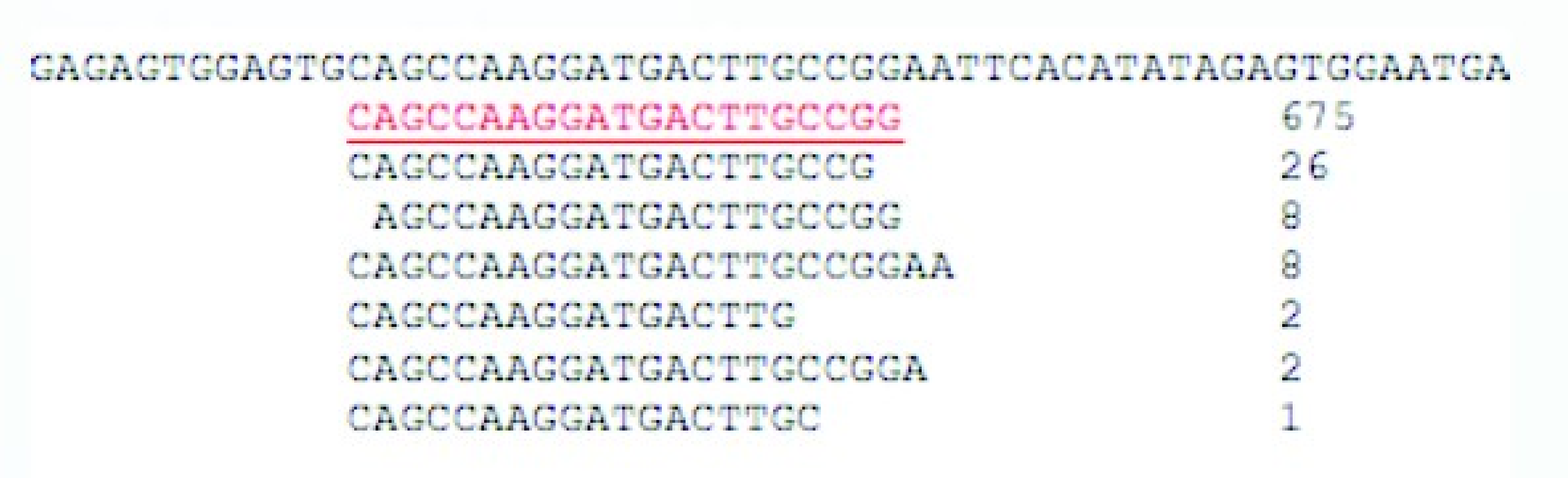


New miRNA
Known miRNA
Quantification matrix

Known miRNA
Quantification matrix

6. Prediction

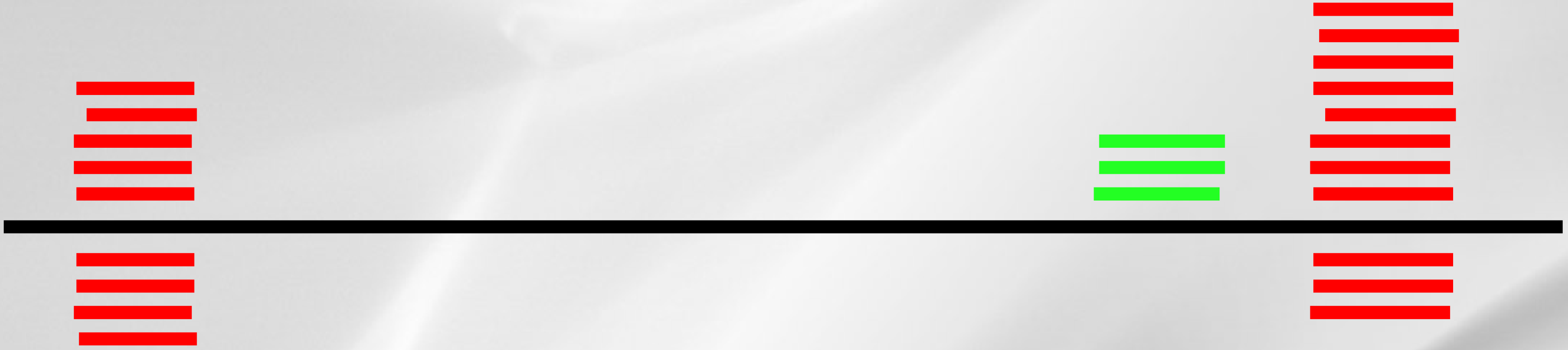
- Precise excision of a 21-22mer is typical of microRNA
  - less represented reads are products of Dicer errors and sequencing/sample preparation artifacts



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

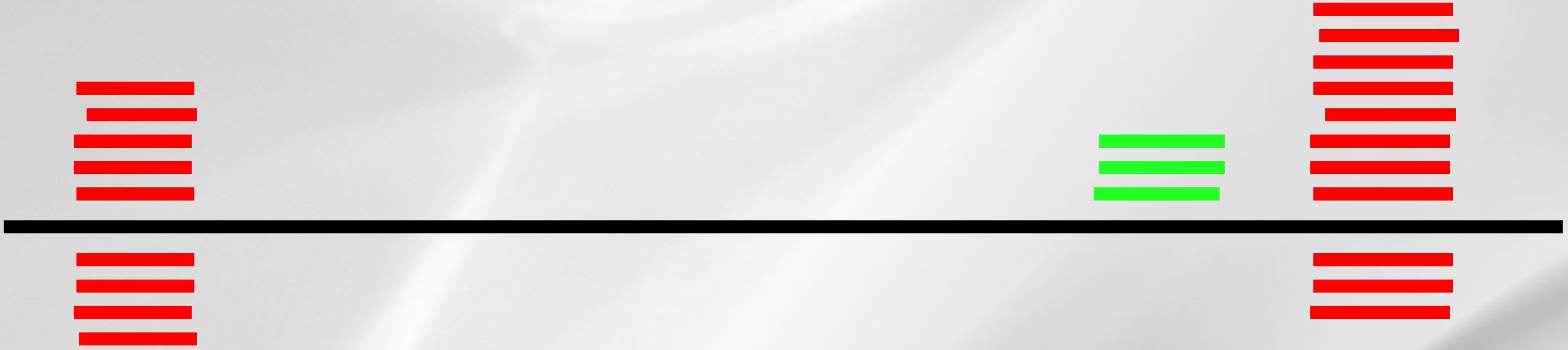
6. Prediction

- Once the reads mapped



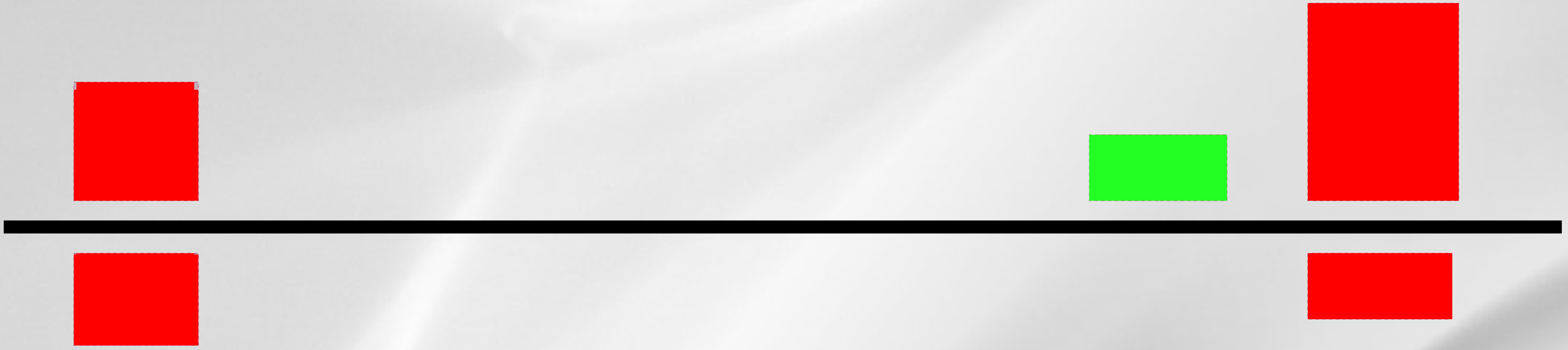
6. Prediction

- Identify all contiguous read regions



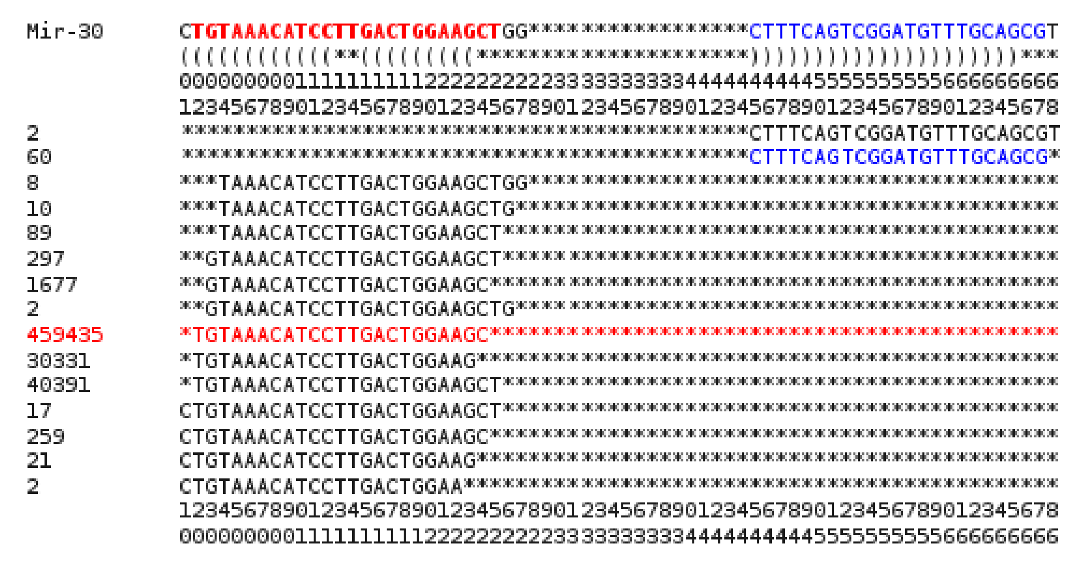
6. Prediction

- Identify all contiguous read regions

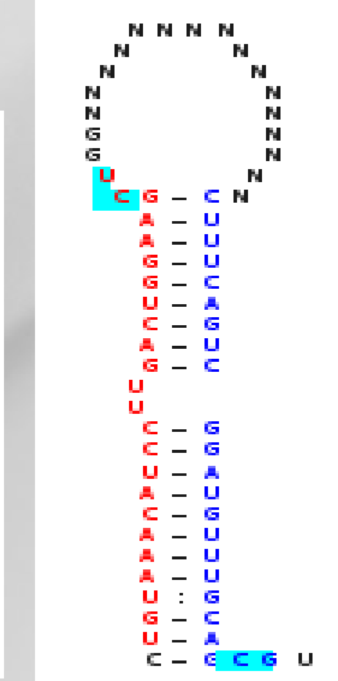


6. Prediction

- 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.



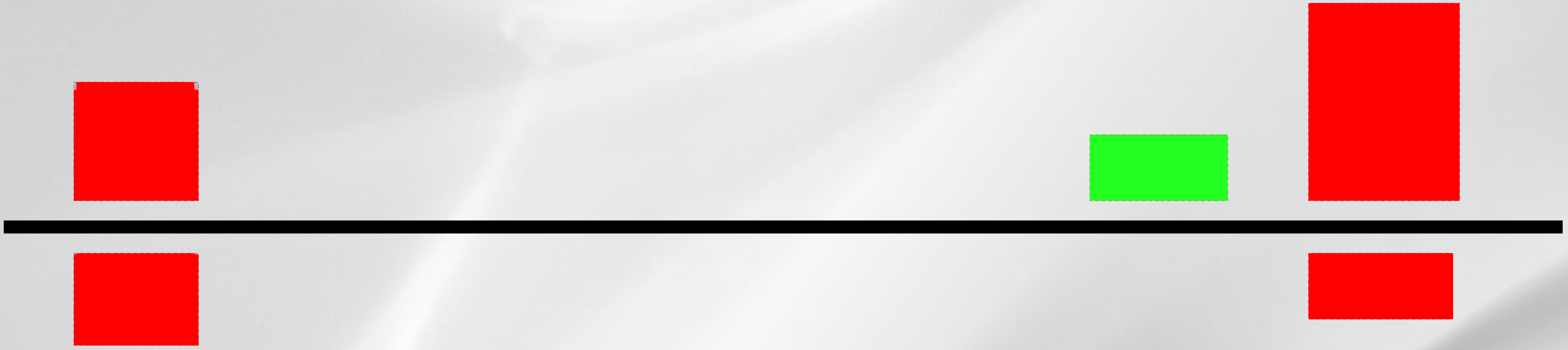
```
Mir-30      CTGTAAACATCCTTGACTGGAAGCTGG*****************CTTTCAGTCGGATGTTTGCAGCGT
            ((((((((((((**(((((((((*********************)))))))))))))))))))))***
            00000000011111111112222222222333333333344444444445555555555666666666
            12345678901234567890123456789012345678901234567890123456789012345678
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
```



Plateforme Bioinformatique Midi-Pyrénées

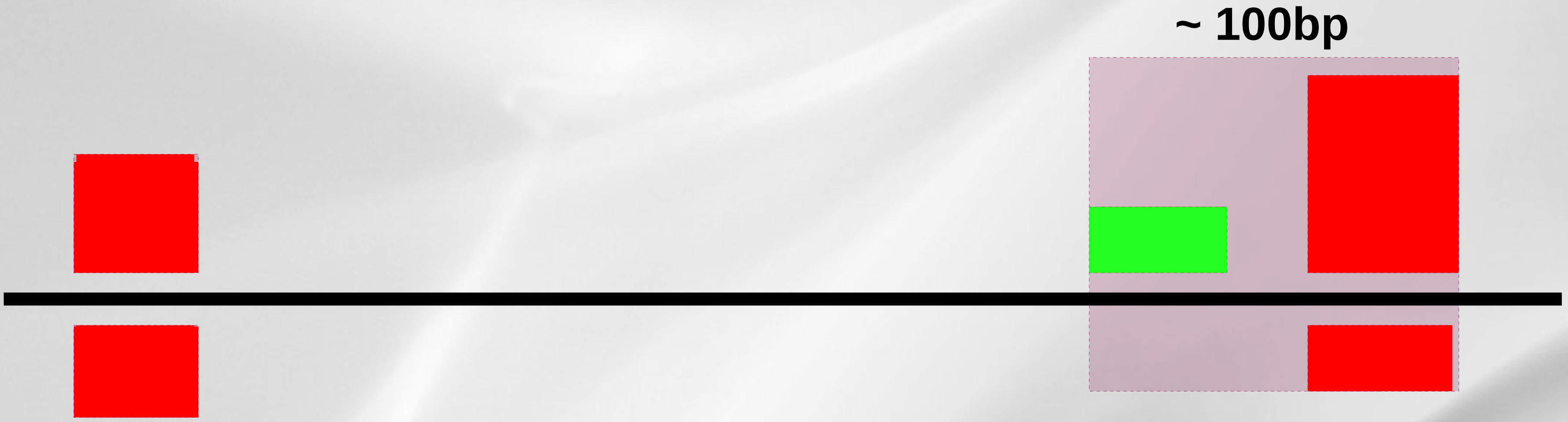
6. Prediction

- Extend and fold read regions



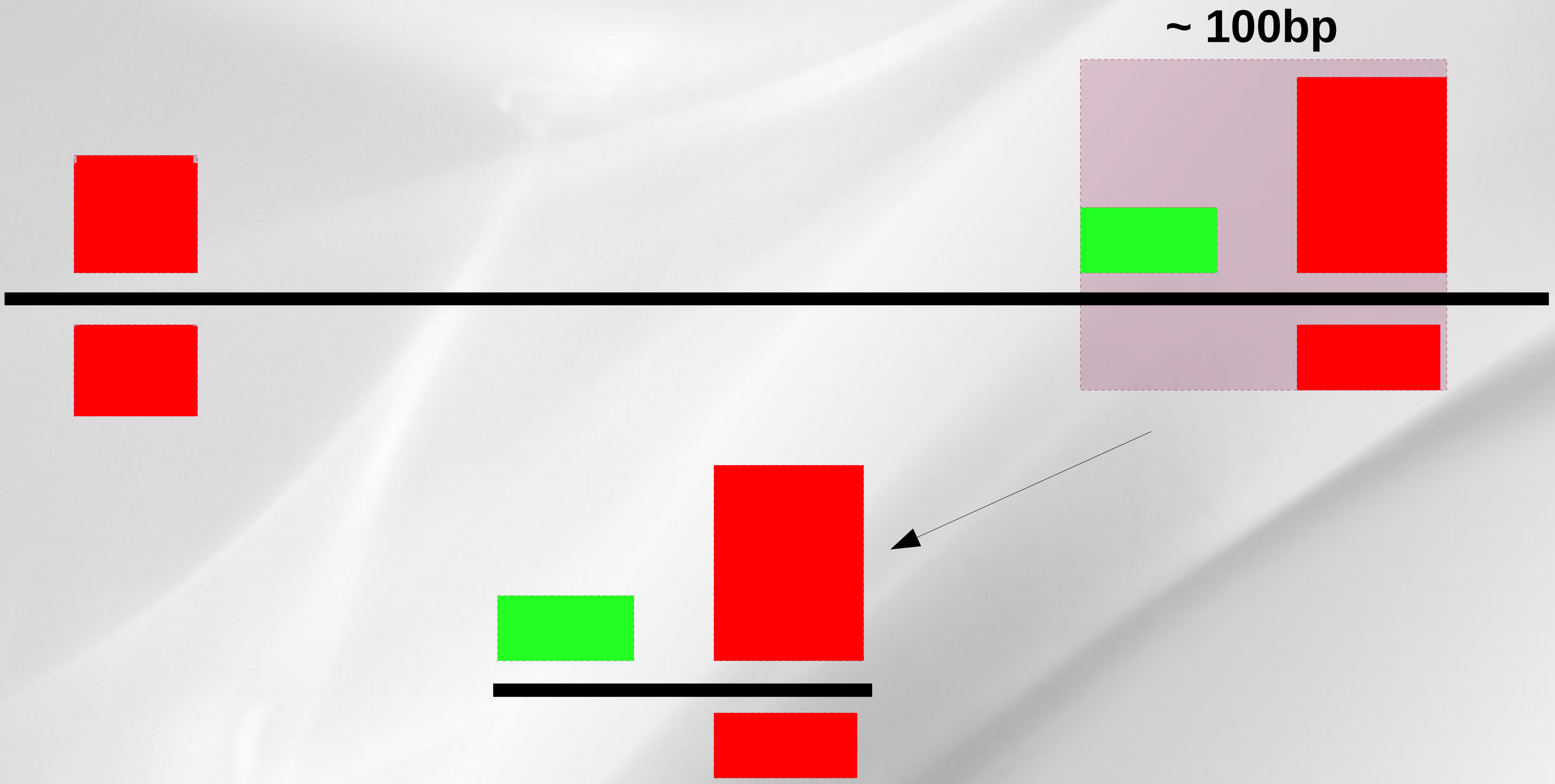
6. Prediction

- Extend and fold read regions



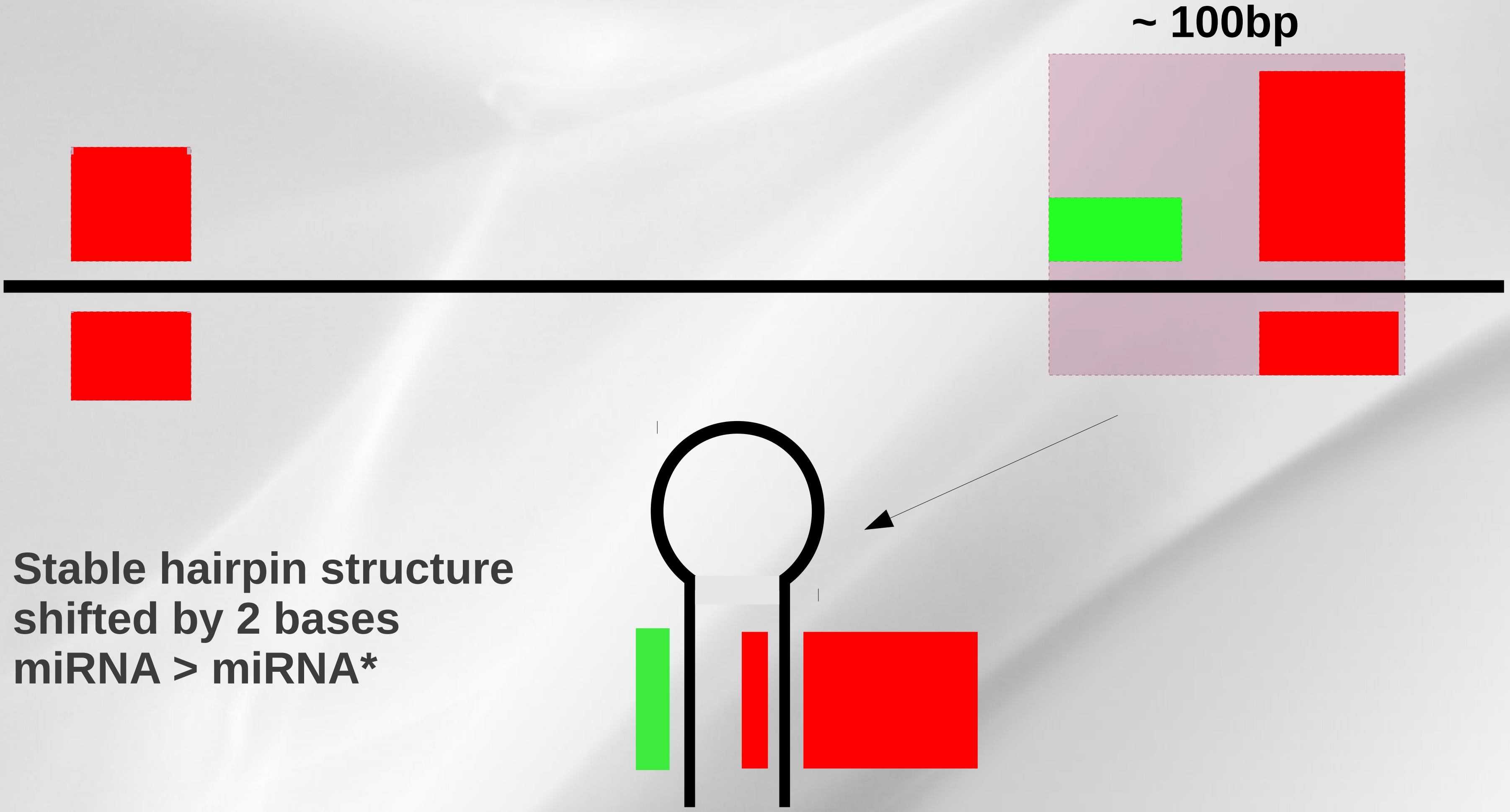
6. Prediction

- Extend and fold read regions



6. Prediction

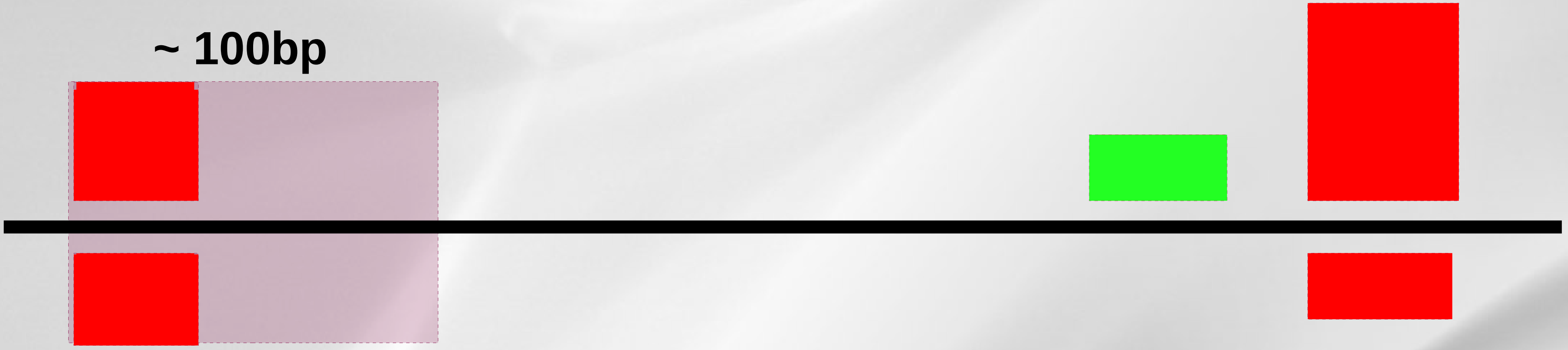
- Extend and fold read regions



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

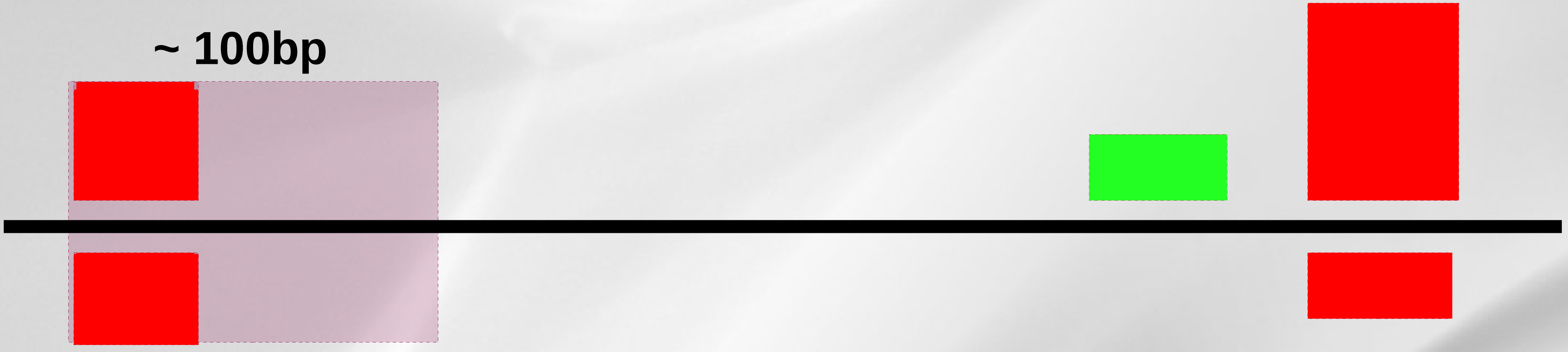
6. Prediction

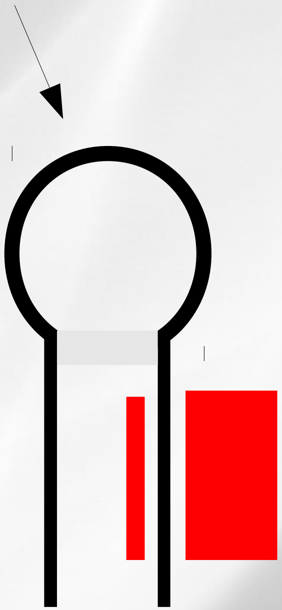
- Extend and fold read regions



6. Prediction

- Extend and fold read regions

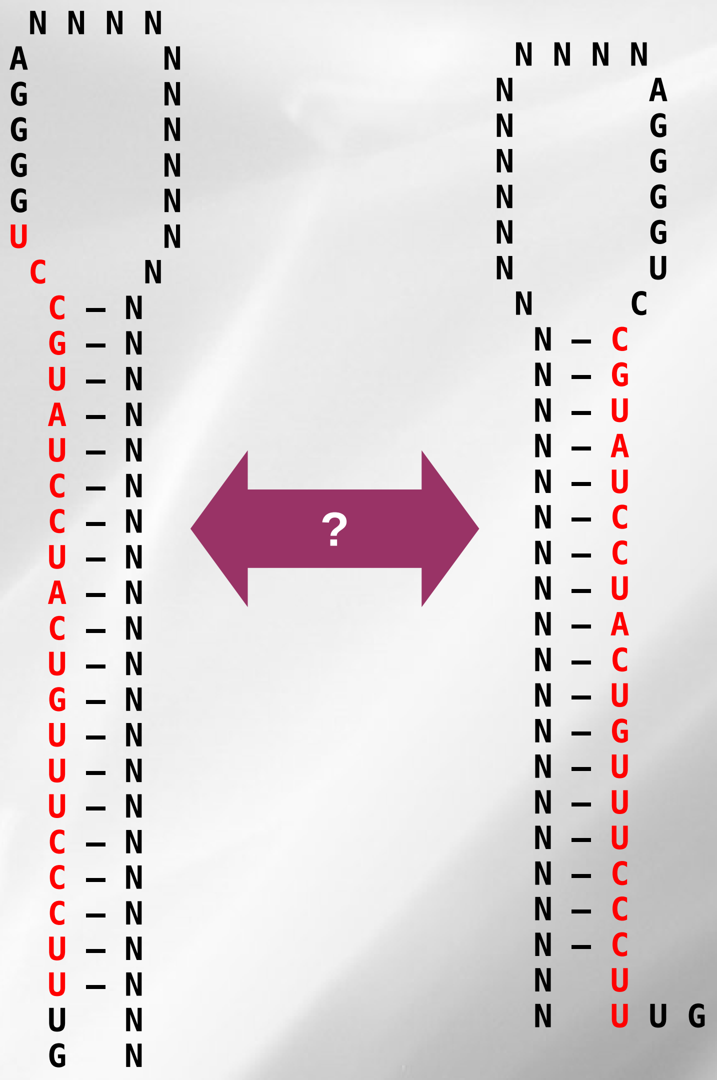




- **In the absence of reads corresponding to an expected miRNA*, additional checks on the structure are:**
  - Degree of pairing in the miRNA region
  - Hairpin: around 70nt in length
  - The secondary structure is significantly more stable than randomly shuffled versions of the same sequence
  - miRNA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

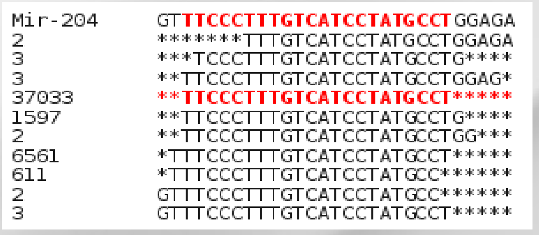
6. Prediction

- Which one should be used ?



```
   N N N N
 A         N
 G         N
 G         N
 G         N
 G         N
 U         N
   C     N
     C – N
     G – N
     U – N
     A – N
     U – N
     C – N
     C – N
     U – N
     A – N
     C – N
     U – N
     G – N
     U – N
     U – N
     U – N
     C – N
     C – N
     C – N
     U – N
     U – N
     U   N
     G   N
```

```
   N N N N
 N         A
 N         G
 N         G
 N         G
 N         G
 N         U
   N     C
     N – C
     N – G
     N – U
     N – A
     N – U
     N – C
     N – C
     N – U
     N – A
     N – C
     N – U
     N – G
     N – U
     N – U
     N – U
     N – C
     N – C
     N – C
     N   U
     N   U U G
```



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

Exercice 4:

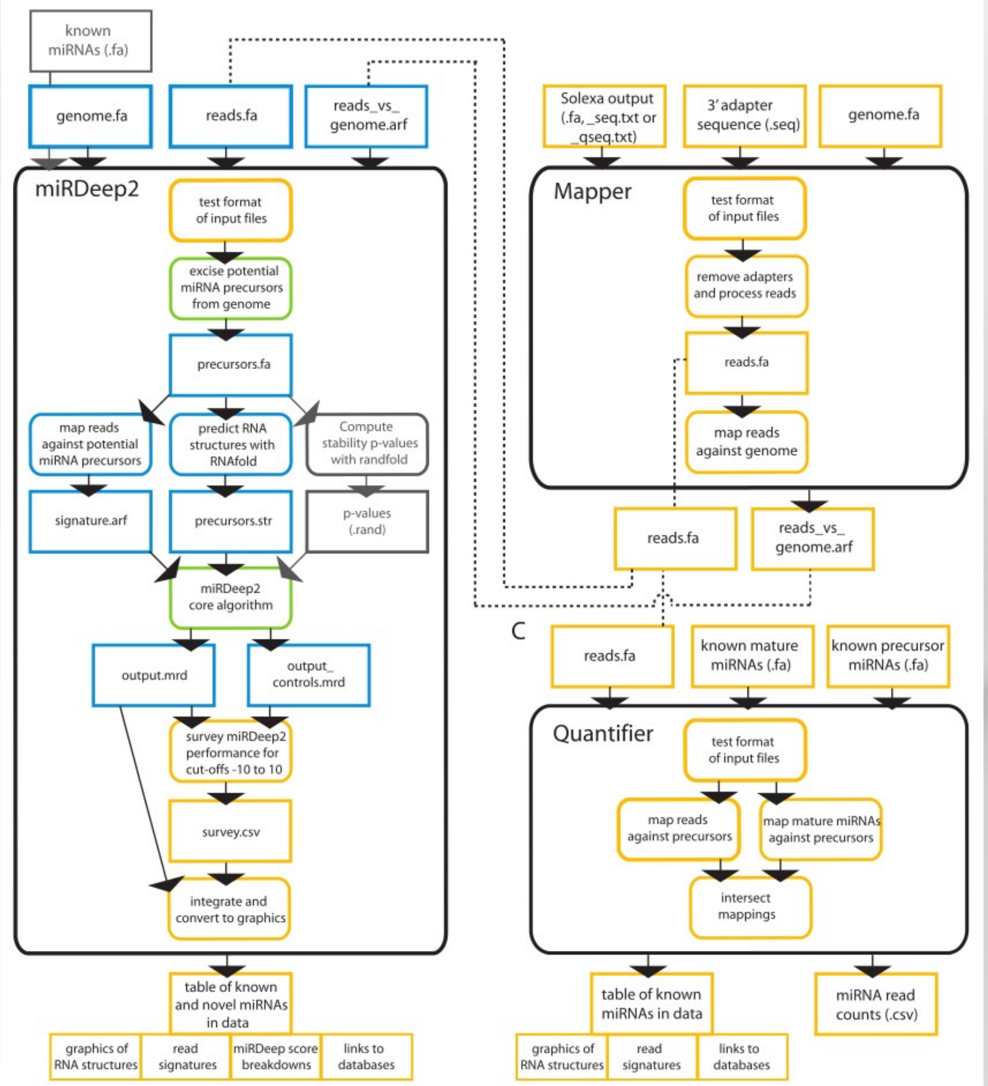
- Locus identification

miRDeep

- Tool for identification of known and novel miRNA
- Animals
  - *Friedländer MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N. (2008) Discovering microRNAs from deep sequencing data using* ***miRDeep****. Nat Biotechnol 26(4), 407-15.*
  - *Friedländer, M.R., Mackowiak, S.D., Li, N., Chen, W., and Rajewsky, N. 2011.* ***miRDeep2*** *accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res.*
- Tool for plants but nothing to do with miRDeep !
  - Plants : *Xiaozeng Yang, Lei Li. 2011* ***miRDeep-P****: a computational tool for analyzing the microRNA transcriptome in plant. Bioinformatics, doi: 10.1093*

miRDeep2

- Complex pipeline (3 main steps)


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
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
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

miRDeep2

1 : Mapper

Mapping of the SGS data on the reference genome

Pipeline :

* Filter reads (not [ACGTN])

* Clip adapters

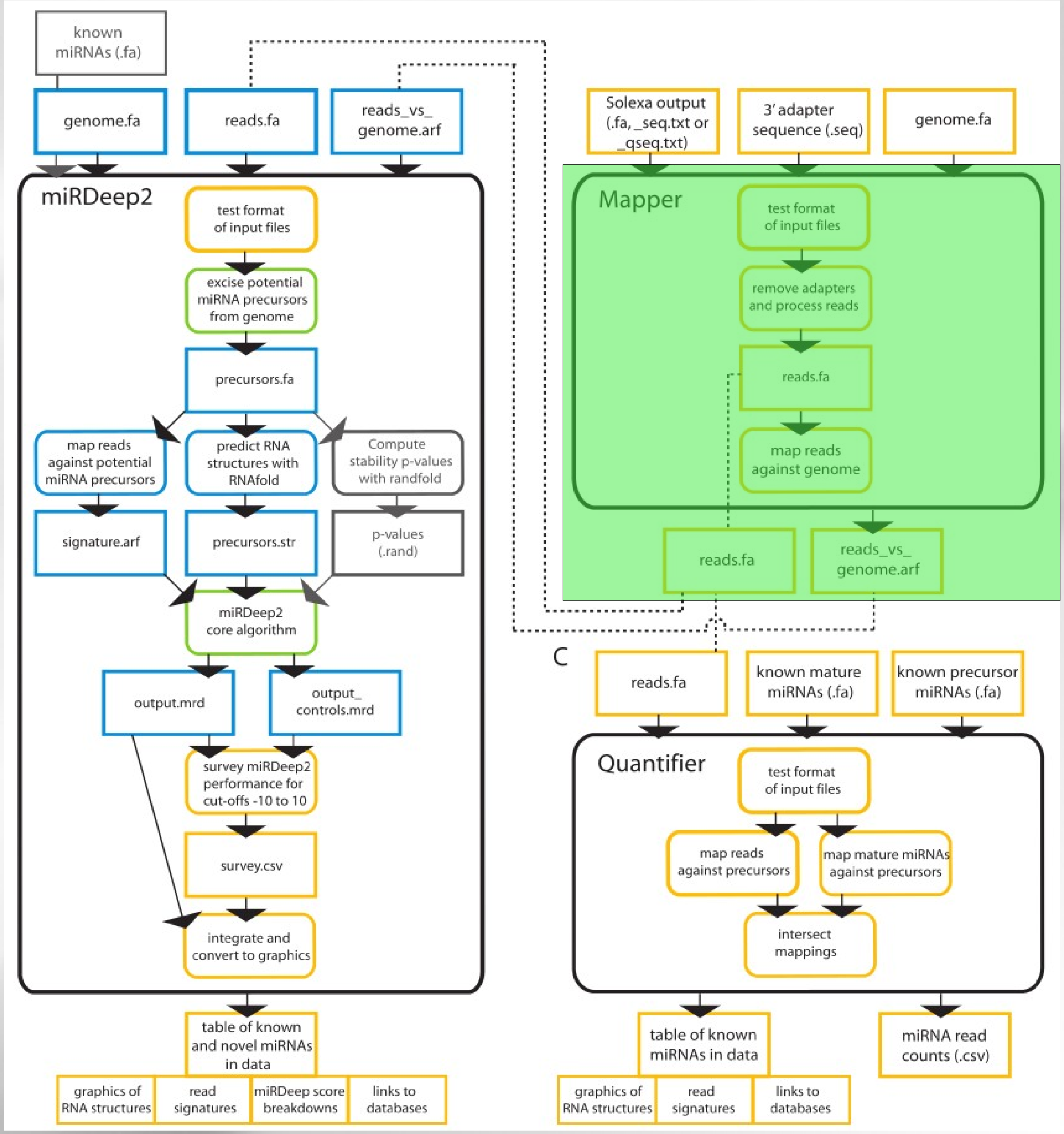
* Filter reads on size (<18 nt)

* Collapse reads

* Align with bowtie

* Transform bowtie output to specific miRDeep2 .arf format

* Filter the .arf file (soft clip)


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
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
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

miRDeep2

2 : Quantifier

Annotation of sequences on miRBase database

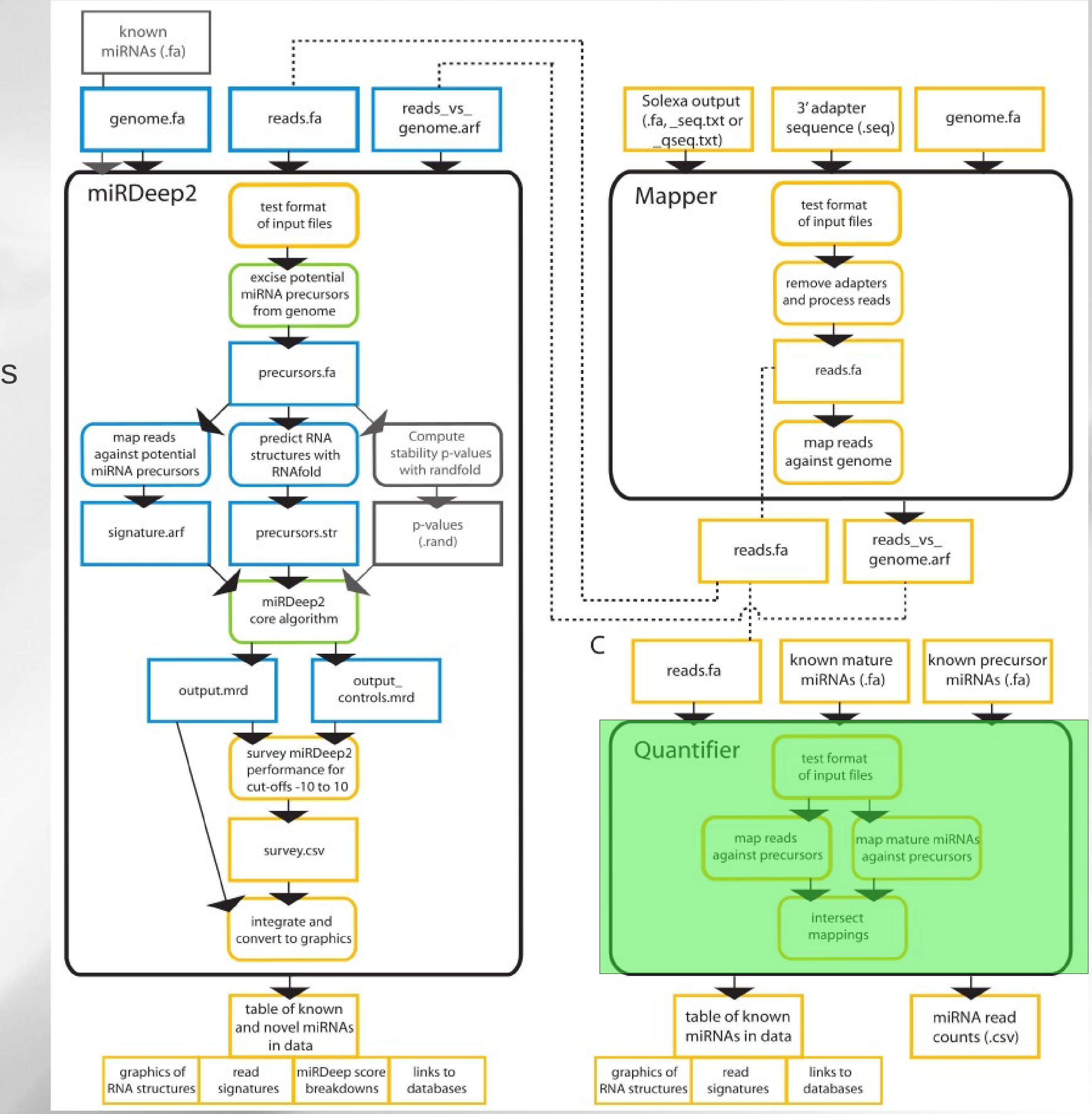
Pipeline :

* Map mature miRNAs on precursors

* Map reads on precursors

* Intersect the 2 mappings

* Output signature and structure of annotated miRNAs


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
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
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

Plateforme Bioinformatique Midi-Pyrénées

miRDeep2

3 : miRDeep2

Prediction of novel miRNAs

Pipeline :

* Test input files

* Keep only perfect mappings of at least 18 nt

* Excise potential precursors within 20 & 70 nt up and down

* Map reads and known miRNAs on potential precursors

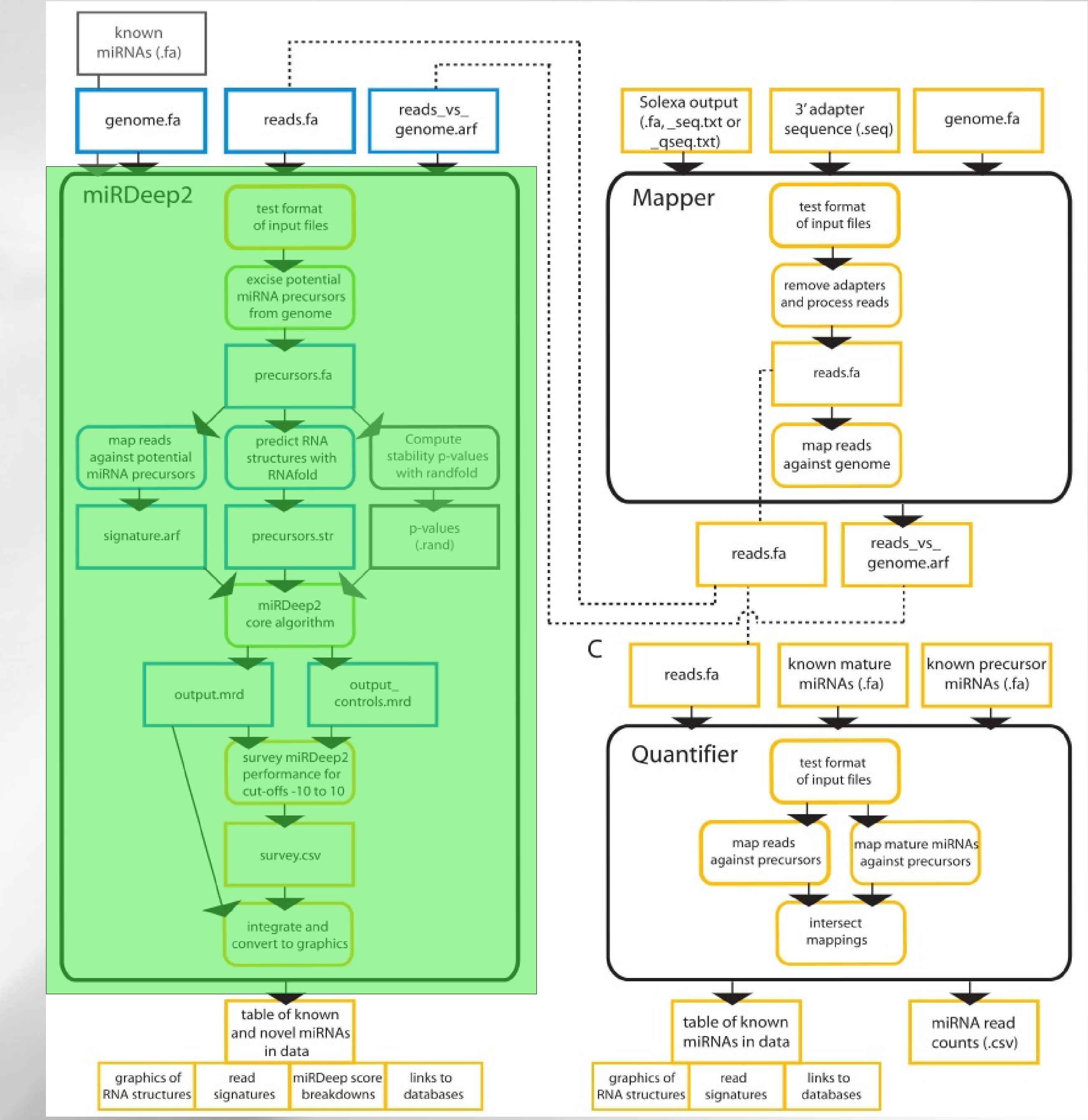
* Merge alignments

* RNAfold + randfold

* Run permuted controls

* Filter potential precursors

* Output novel miRNAs

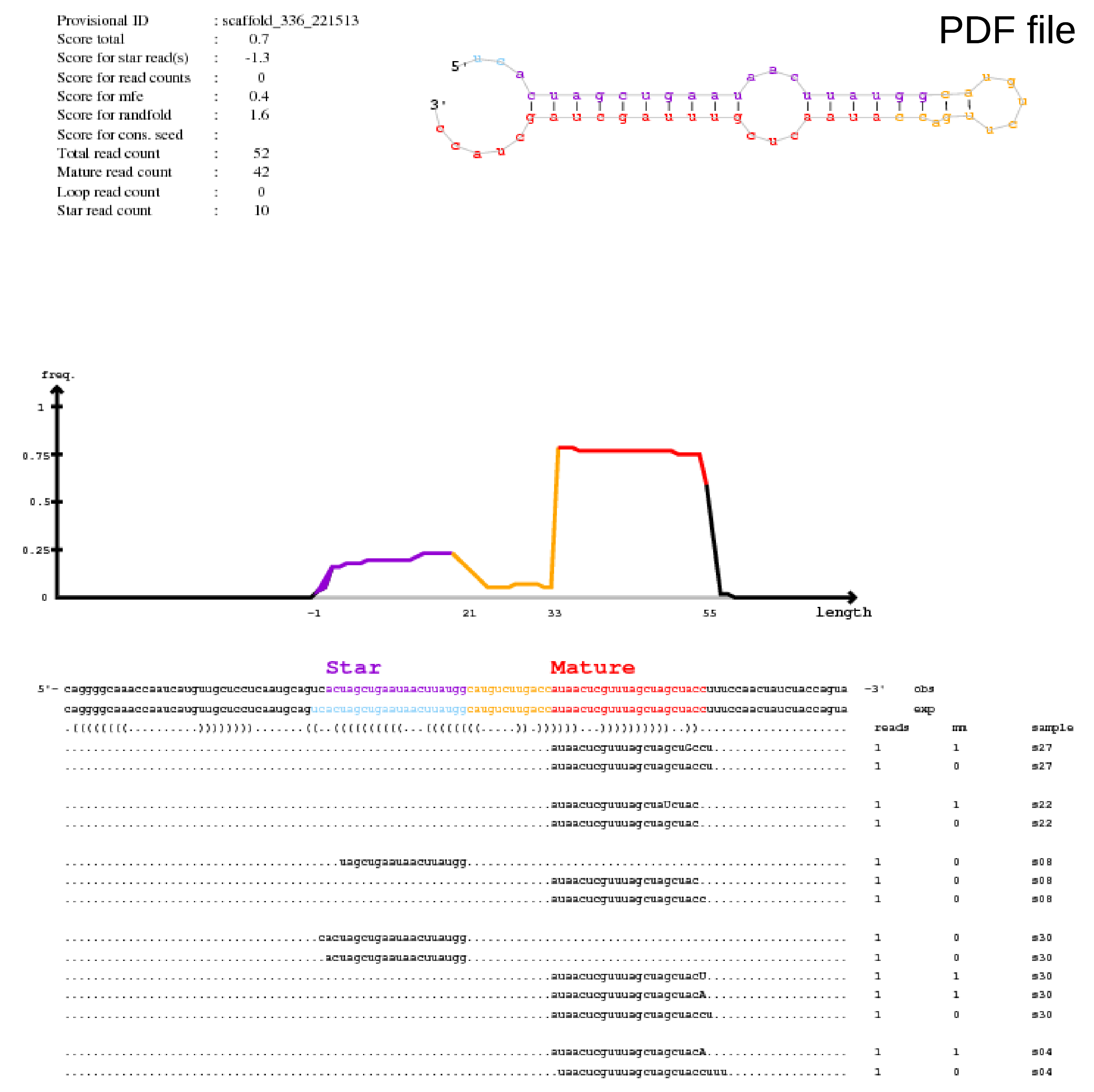


known miRNAs (.fa)
genome.fa
reads.fa
reads_vs_ genome.arf
miRDeep2
test format of input files
excise potential miRNA precursors from genome
precursors.fa
map reads against potential miRNA precursors
predict RNA structures with RNAfold
Compute stability p-values with randfold
signature.arf
precursors.str
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
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
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

miRDeep2 output



PDF file

Provisional ID	: scaffold_336_221513
Score total	: 0.7
Score for star read(s)	: -1.3
Score for read counts	: 0
Score for mfe	: 0.4
Score for randfold	: 1.6
Score for cons. seed	:
Total read count	: 52
Mature read count	: 42
Loop read count	: 0
Star read count	: 10

```
5'- caggggcaaaccaaucauguuugcuccucaaugcagucacuagcugaauaacuuaugzcaugucuugaccauaacucguuuagcuagcuaccuuuccaacuaucuaccagua   -3'   obs
    caggggcaaaccaaucauguuugcuccucaaugcagucacuagcugaauaacuuauggcaugucuugaccauaacucguuuagcuagcuaccuuuccaacuaucuaccagua         exp
    .((((((((..........))))))))........((..((((((((((...((((((((.....)).))))))...))))))))))..)).....................   reads   mm   sample
    ......................................................................auaacucguuuagcuagcuGccu...................   1       1    s27
    ......................................................................auaacucguuuagcuagcuaccu...................   1       0    s27

    ......................................................................auaacucguuuagcuaUcuac.....................   1       1    s22
    ......................................................................auaacucguuuagcuagcuac.....................   1       0    s22

    ........................................uagcugaauaacuuaugg......................................................   1       0    s08
    ......................................................................auaacucguuuagcuagcuac.....................   1       0    s08
    ......................................................................auaacucguuuagcuagcuacc....................   1       0    s08

    .....................................cacuagcugaauaacuuaugg......................................................   1       0    s30
    ......................................acuagcugaauaacuuaugg......................................................   1       0    s30
    ......................................................................auaacucguuuagcuagcuacU....................   1       1    s30
    ......................................................................auaacucguuuagcuagcuacA....................   1       1    s30
    ......................................................................auaacucguuuagcuagcuaccu...................   1       0    s30

    ......................................................................auaacucguuuagcuagcuacA....................   1       1    s04
    .......................................................................uaacucguuuagcuagcuaccuuu.................   1       0    s04
```

text file

```
scaffold_336 280330 280384
novel:scaffold_336_221513
0.7 -   280330   280384
0,0,255
```

Why develop a new tool ?

* MiRDeep2 pipeline is not optimized :

- A lot of redundant steps (mapped reads filtering, inter-fastq redundant reads kept)
- A lot of temporary files :
  - Input : 166 Go
  - Output : 1,5 To

  x 10 !
- A lot of time-processing :
  - Mapper : 37 h
  - MiRDeep2 : 390 h

* Bugs :

- Bad algorithm of 3' adapters clipping
- Quantification step not used for prediction
- Options not available
- ...

* Not enough user-defined parameters (bowtie, RNAfold ...)

* Not adapted for discovering other small RNAs (tRNA...)

Why develop a new tool ?

* MiRDeep2 pipeline is not optimized :

- A lot of redundant steps (mapped reads filtering, inter-fastq redundant reads kept)
- A lot of temporary files :
  - Input : 166 Go
  - Output : 1,5 To

x 10 !

- A lot of time-processing :
  - Mapper : 37 h
  - MiRDeep2 : 390 h

* Bugs :

- Bad algorithm of 3' adapters clipping
- Quantification step not used for prediction
- Options not available
- ...

Keep only 6 first nuc of **ADAPTE**R

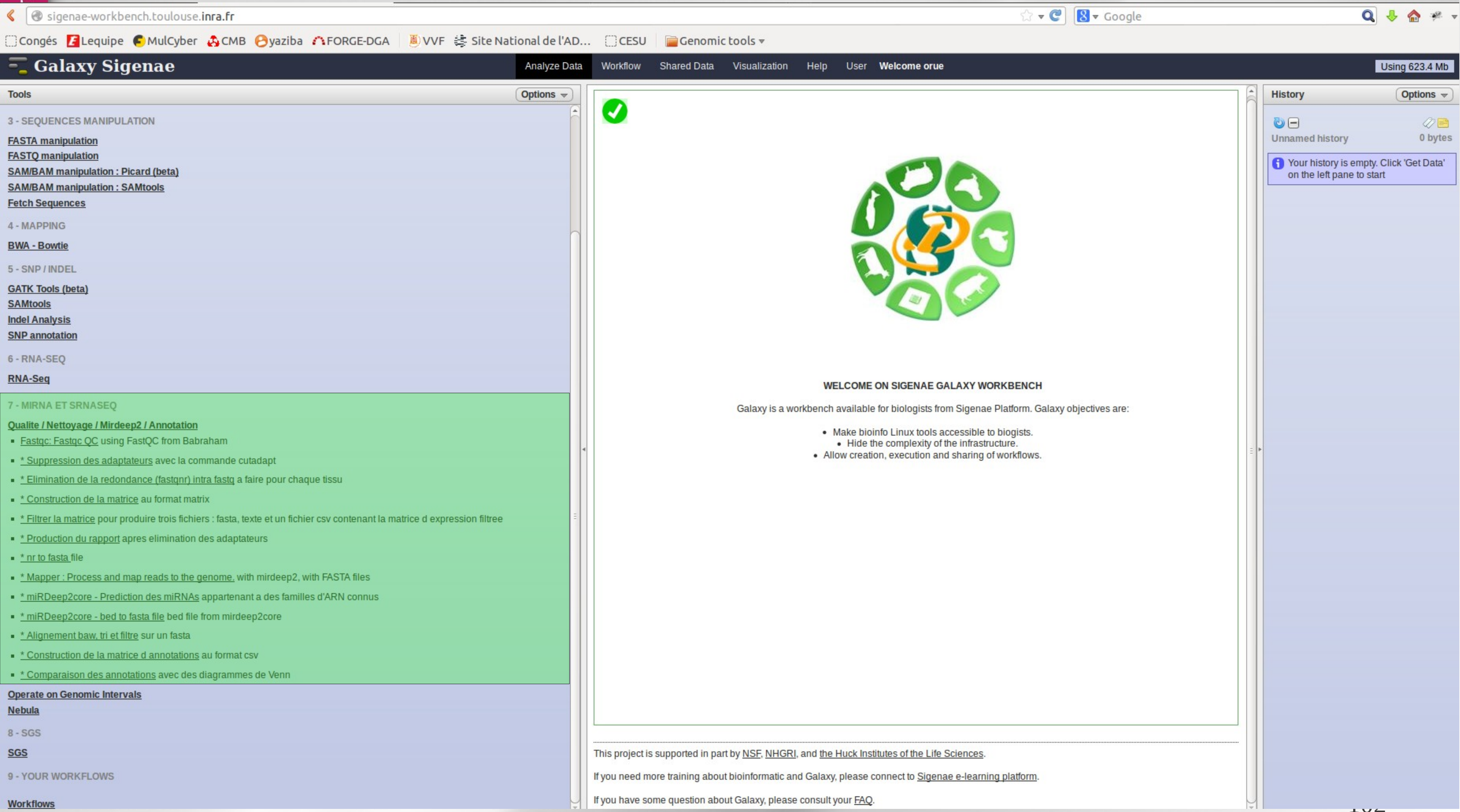
>1
ADAPTEBLABLA → ADAPTE
>2
MYSEQUENCE**A** → MYSEQUENCE

* Not enough user-defined parameters (bowtie, RNAfold ...)

* Not adapted for discovering other small RNAs (tRNA...)

sRNAseq & GALAXY

http://sigenae-workbench.toulouse.inra.fr/



3 - SEQUENCES MANIPULATION

- FASTA manipulation
- FASTQ manipulation
- SAM/BAM manipulation : Picard (beta)
- SAM/BAM manipulation : SAMtools
- Fetch Sequences

4 - MAPPING

- BWA - Bowtie

5 - SNP / INDEL

- GATK Tools (beta)
- SAMtools
- Indel Analysis
- SNP annotation

6 - RNA-SEQ

- RNA-Seq

7 - MIRNA ET SRNASEQ

Qualite / Nettoyage / Mirdeep2 / Annotation

- Fastqc: Fastqc QC using FastQC from Babraham
- * Suppression des adaptateurs avec la commande cutadapt
- * Elimination de la redondance (fastqnr) intra fastq a faire pour chaque tissu
- * Construction de la matrice au format matrix
- * Filtrer la matrice pour produire trois fichiers : fasta, texte et un fichier csv contenant la matrice d expression filtree
- * Production du rapport apres elimination des adaptateurs
- * nr to fasta file
- * Mapper : Process and map reads to the genome, with mirdeep2, with FASTA files
- * miRDeep2core - Prediction des miRNAs appartenant a des familles d'ARN connus
- * miRDeep2core - bed to fasta file bed file from mirdeep2core
- * Alignement baw, tri et filtre sur un fasta
- * Construction de la matrice d annotations au format csv
- * Comparaison des annotations avec des diagrammes de Venn

Operate on Genomic Intervals

Nebula

8 - SGS

- SGS

9 - YOUR WORKFLOWS

- Workflows

WELCOME ON SIGENAE GALAXY WORKBENCH

Galaxy is a workbench available for biologists from Sigenae Platform. Galaxy objectives are:

- Make bioinfo Linux tools accessible to biogists.
- Hide the complexity of the infrastructure.
- Allow creation, execution and sharing of workflows.

This project is supported in part by NSF, NHGRI, and the Huck Institutes of the Life Sciences.

If you need more training about bioinformatic and Galaxy, please connect to Sigenae e-learning platform.

If you have some question about Galaxy, please consult your FAQ.

Your history is empty. Click 'Get Data' on the left pane to start