microARNs - Prédiction à partir des données NGS

Mohamed Amine Remita

9 mars 2016

1. Introduction

2. Étapes de l'analyse des données NGS
   - 1. Prétraitement
   - 2. Mappage
   - 3. Élimination des reads par filtres
   - 4. Extraction et repliement
   - 5. Prédictions
   - 6. Analyse fonctionnelle

3. Travail pratique

Séquençage à haut débit (NGS)

Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000
Illumina/ Solexa's GA_{II}	Frag, MP/ solid-phase	RTs	75 or 100	4‡, 9§	18‡, 35§	540,000
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7‡, 14§	30‡, 50§	595,000

*Average read-lengths. ‡Fragment run. §Mate-pair run. Frag, fragment; GA, Genome Analyzer; GS, Genome Sequencer; MP, mate-pair; N/A, not available; NGS, next-generation sequencing; PS, pyrosequencing; RT, reversible terminator; SBL, sequencing by ligation; SOLiD, support oligonucleotide ligation detection.

Sequencing technologies - the next generation (Metzker 2009)

Formats des données

```
@HWI-ST748:152:C0Y17ACXX:4:1101:1572:2125 1:N:0:ATCACG
CCTTGAACTCGCCAATCTGCTCCTCGCTGAGCTGGTCCGCCATGGCAAACCGAAACCCCCAAGGGGCGACGAGGGCCTGCAAATGTTACGCGAGCTGCGC
+
@@@;A8BDHHADHGGGBDFF>F<FD<AHIDCB4DE9B9?6FE(=BA@G>DEEHH/?B(9A:?=;?/'))0905999???A33:(44::@@>>5@B9>999
@HWI-ST748:152:C0Y17ACXX:4:1101:1730:2146 1:N:0:ATCACG
GCCCGCTTCAGGTCGCTCGACTTCGGCGAGCGCAACGGGTACCTCAAGGGCGTCGTCACCGATGTCATCCACGACCCGGGGCGTGGTGCGCCGCTGGCCA
+
@??DDDADH>FFCD:)1AG::?F>GHIHIF5AA5:?ECB33;;ACCCCCB1;79@BB@?3255<8@@@4::92950<99@0)55<9(+49>599>BB?19
@HWI-ST748:152:C0Y17ACXX:4:1101:1697:2160 1:N:0:ATCACG
TATGCAACCCCTGGTAGTGTCCCTGATAGTGTGAACTACTGTGACCAAATTGAAATTCATCACCCAATCAGCATTCCTGGTATGGAAGGCATTACATTAC
+
@CCFDEDFHAHBHICFGEBEHHJIG;AFHFHCFG@FDG>DGBD@FHGC=FE<DGHIFGGABBBAHG;=@EDCA@HAEHEH;BDEBC>>BABCCDDDEDCC
@HWI-ST748:152:C0Y17ACXX:4:1101:1508:2163 1:N:0:ATCACG
```

Format fastq

Formats des données

```
>1_103_386_F3
T123121010223103000010300001000000000
>1_131_554_F3
T302011022321131010230300011300000010
>1_134_560_F3
T002023010020000122010330201000303 01
>1_141_466_F3
T233020113031011230200002032010300 10
```

```
>1_103_386_F3
4 14 2 4 11 11 27 3 19 17 17 26 5 9 11 10 26 14 23 11 13 23 23 20 4 5 22 5 12 9 5 10 23 25 3
>1_131_554_F3
20 9 4 20 20 6 19 13 14 17 7 4 19 21 21 4 4 9 14 11 7 11 24 6 25 5 2 16 22 6 6 6 6 5 8
>1_134_560_F3
27 24 11 8 28 23 24 8 20 27 23 26 24 11 27 19 16 20 11 15 8 22 5 17 13 26 20 15 21 26 5 7 6 18 23
>1_141_466_F3
16 19 10 11 8 20 20 4 21 15 24 6 8 5 8 16 13 6 8 13 23 18 9 5 20 8 3 8 6 17 2 9 11 6 17
```

Format color space (csfata)

Formats des données

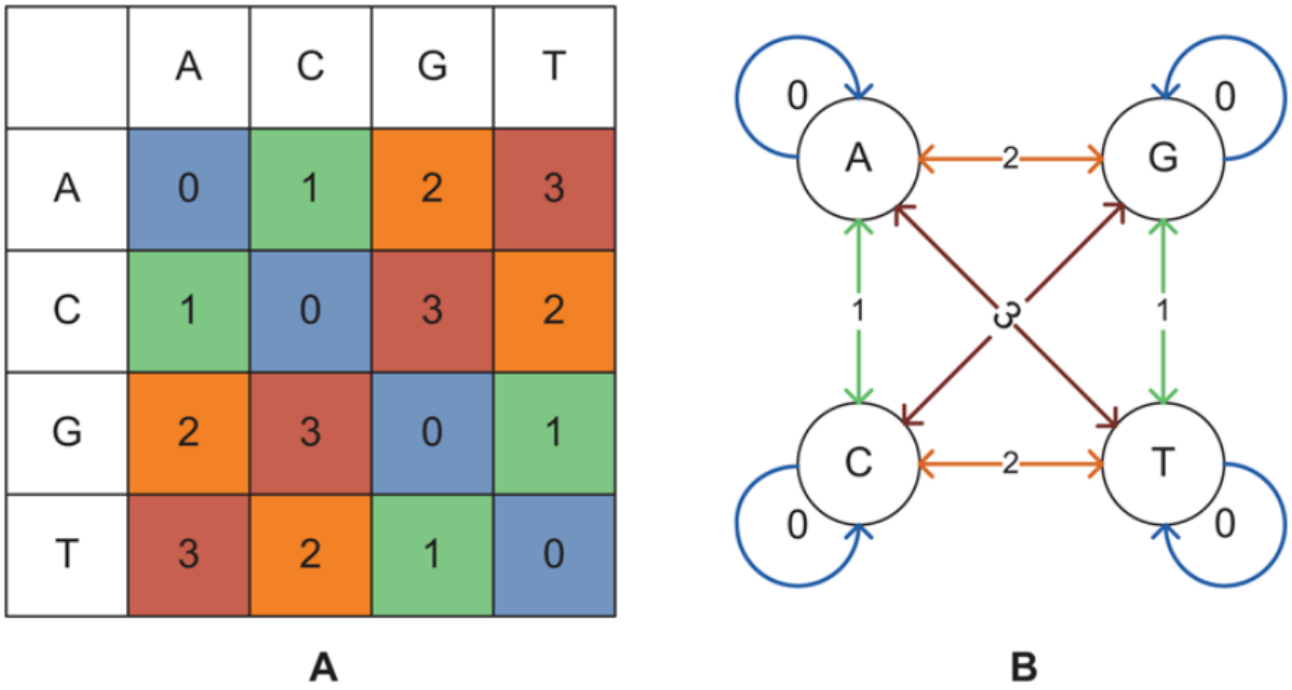
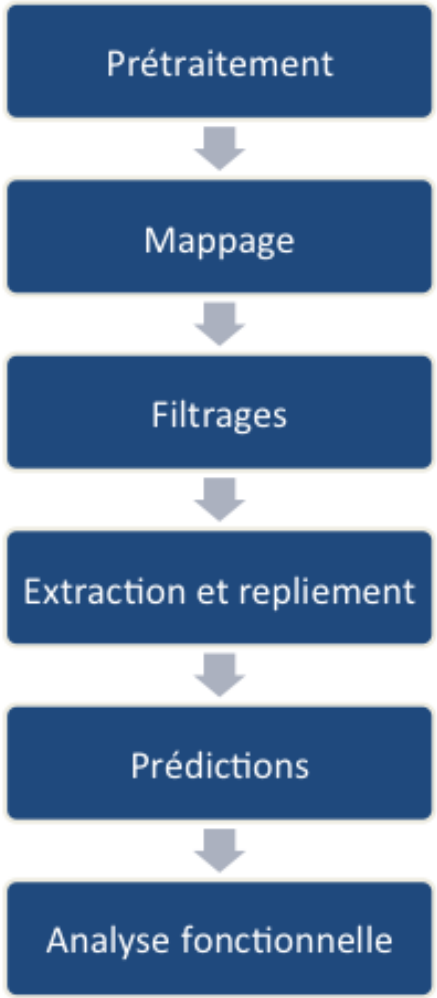


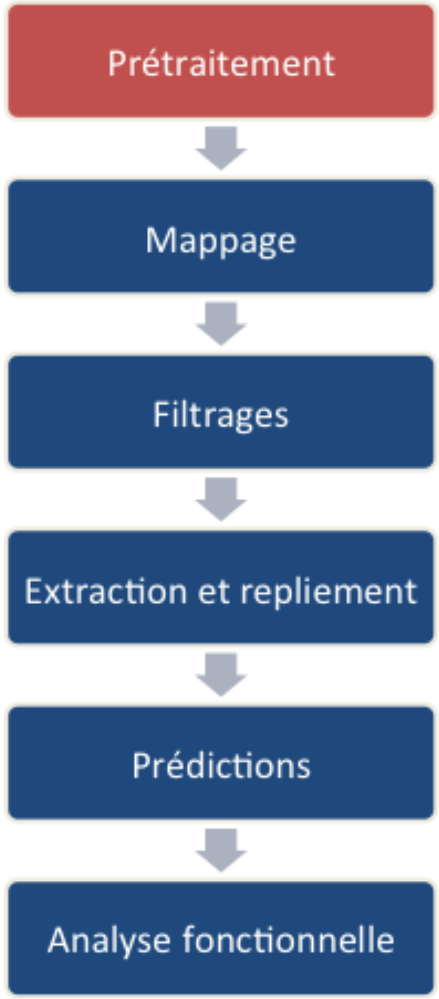
Figure 2. Two representations of the color-space (dibase) encoding used by the AB SOLiD sequencing system. A: The standard representation, with the first and second letter of the queried pair along the horizontal and vertical axes, respectively. **B:** The equivalent Finite State Automaton representation, with edges labelled with the readouts and nodes corresponding to the basepairs of the underlying genome.
doi:10.1371/journal.pcbi.1000386.g002

Étapes de l'analyse



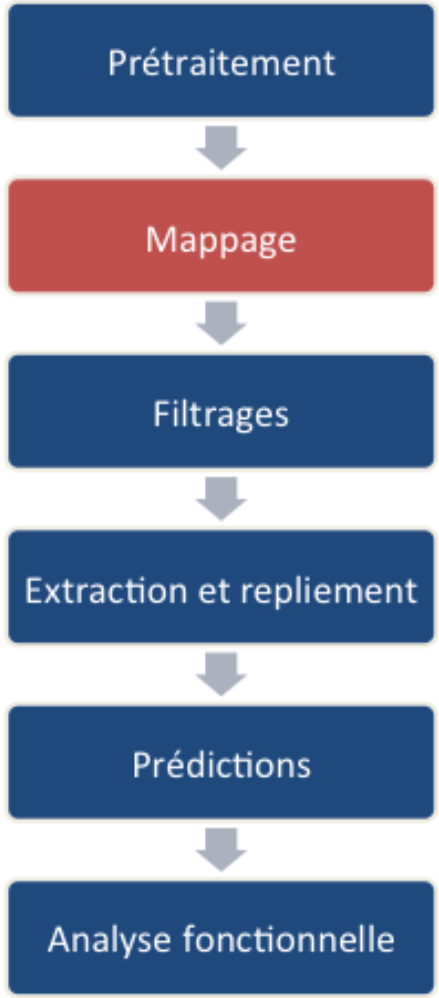
Prétraitement

- Élimination des reads avec mauvaise qualité
- Suppression des adaptateurs (cutadapt)

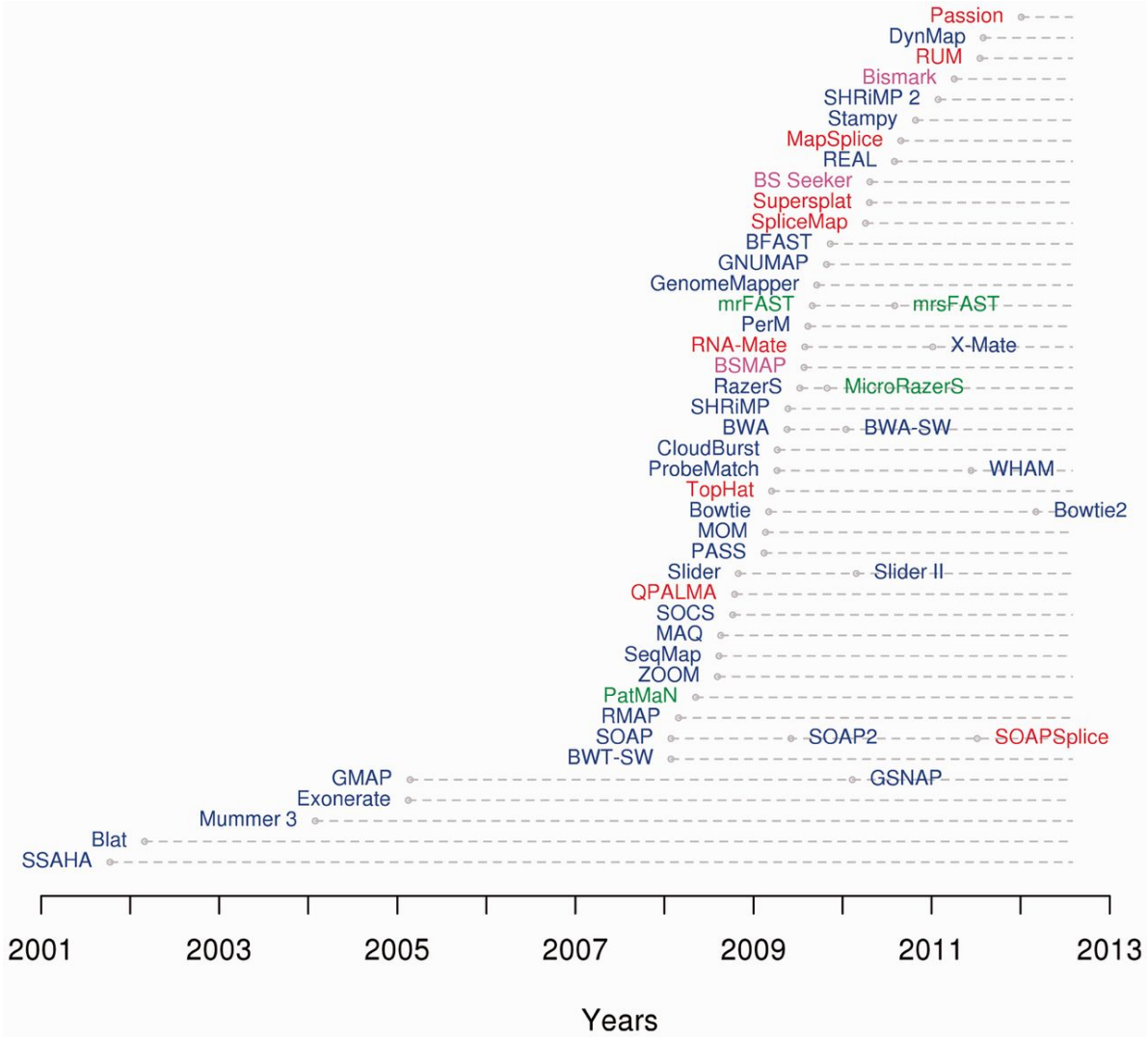


Mappage

- Alignement des reads contre :
  - Génome de reference
  - Ensemble de séquence EST et/ou GSS
  - Ensemble de microARNs valides (mirBase)
- Plusieurs programmes de mappage
  - BLAST, BLAT
  - MAQ, BOWTIE, SHRIMP
  - etc.

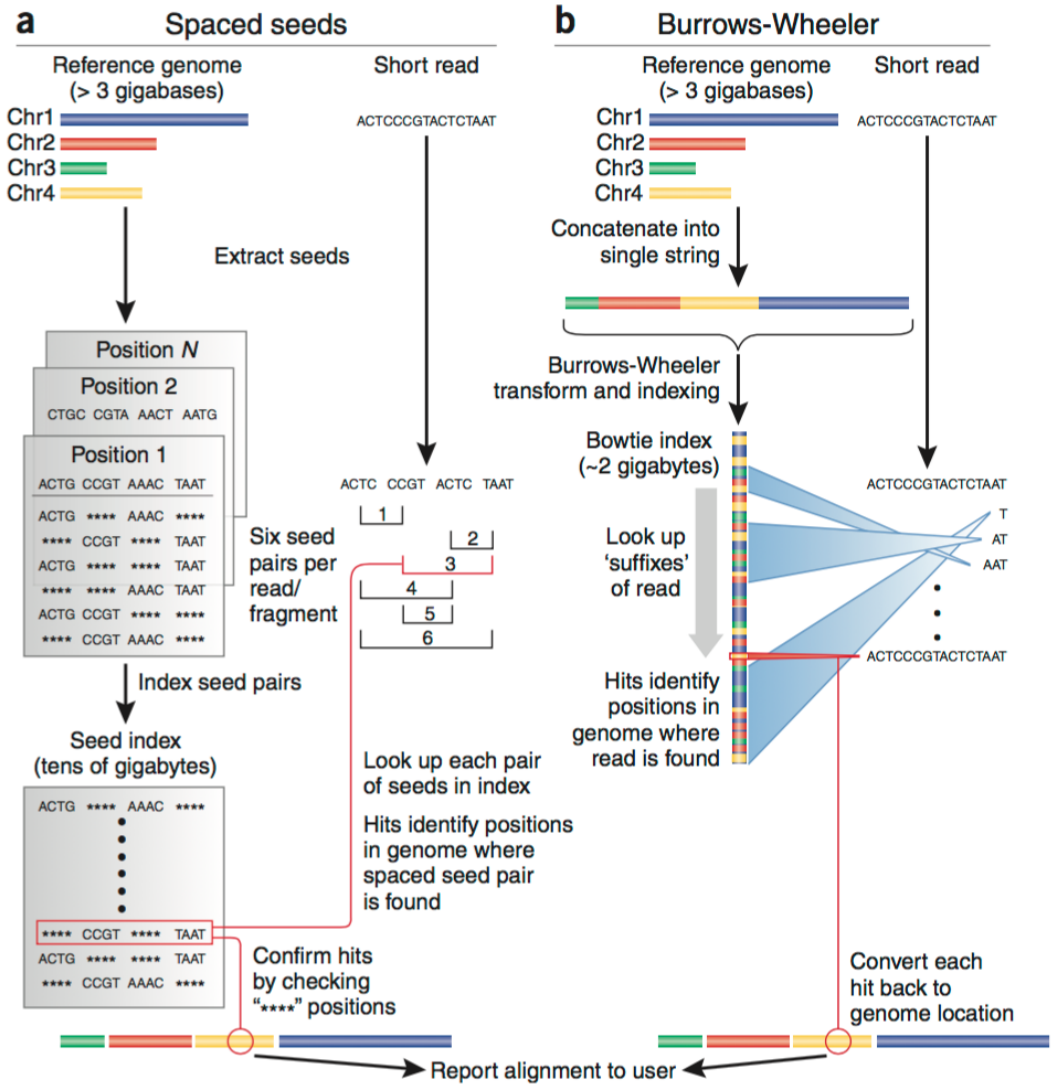


Mappage



Tools for mapping high-throughput sequencing data (Fonseca et al 2012)

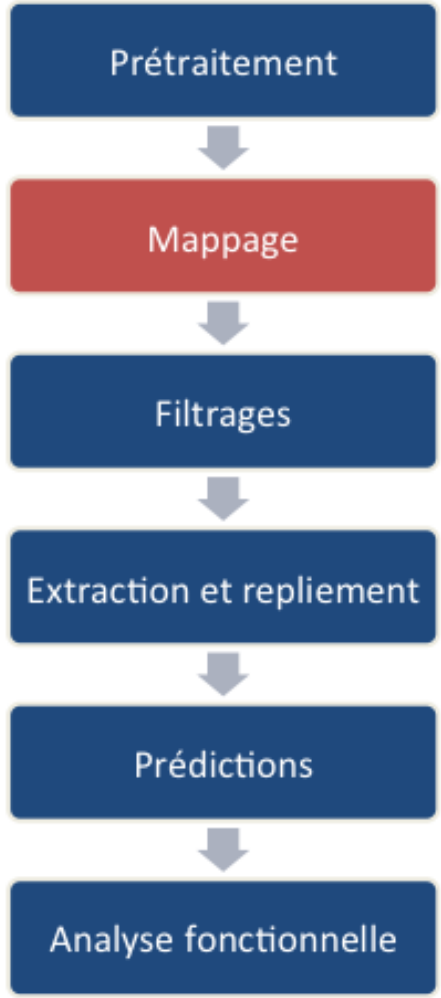
Mappage



How to map billions of short reads onto genomes (Trapnell and Salzberg 2009)

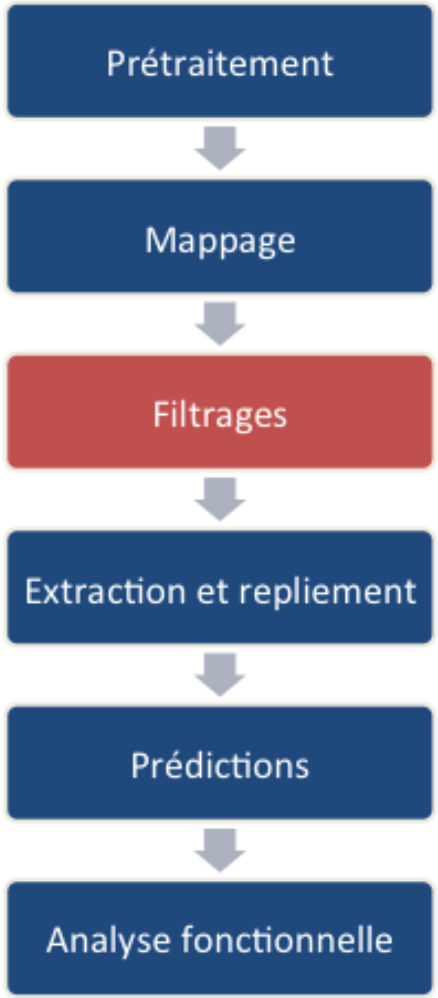
Mappage

- Calcul de l'abondance (expression) des reads dans chaque librairie
- Normalisation des abondances



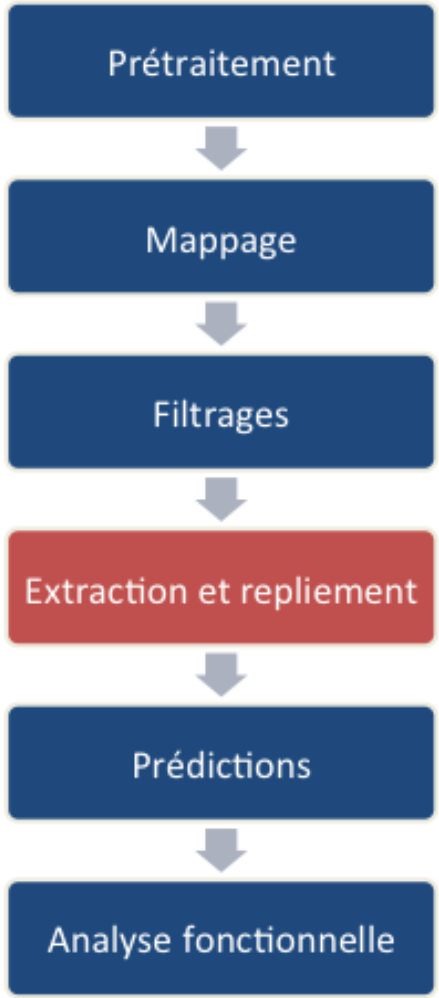
Filtrages

- Élimination des reads faiblement exprimés
- Élimination des reads qui s'alignent contre
  - Les ARNs non codants sauf les miARNs (ARNr, ARNt, snoARN, etc.)
  - Les ARNs messagers (proteines)
- Élimination des reads de faible complexité (RepeatMasker)



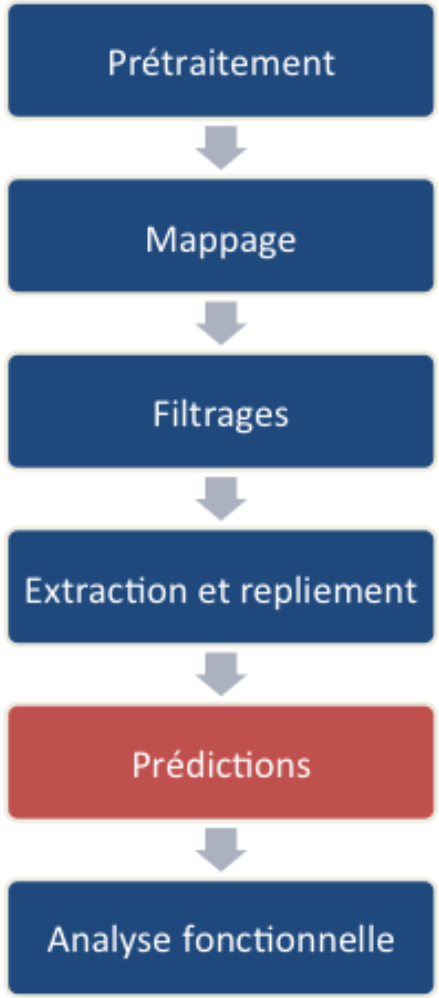
Extraction et repliement

- Extraction du précurseur (pre-miARN)
  - Un miARN peut être 5p ou 3p
  - Deux précurseurs possibles pour chaque reads
  - -160 nd +20 nd
  - -20 nd +160 nd
- Repliement des séquence extraites
  - Structures secondaires et MFE
  - Mfold, RNAFold



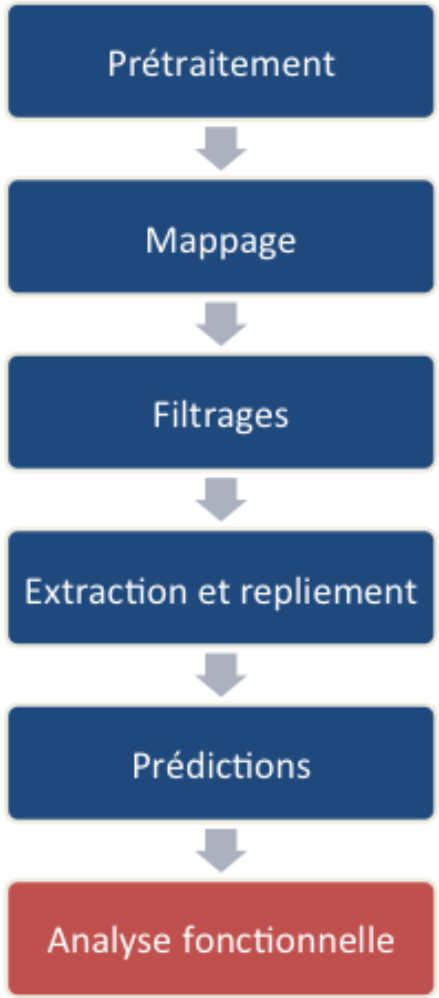
Prédictions

- Prédiction des précurseurs des miARNs
  - Triplet-SVM, Mipred
- Prédiction des miARNs
  - miRcheck, miRdup
- Prédiction des gènes ciblés par les miARNs
  - Tapir, DIANA-microT, psRNATarget



Analyse fonctionnelle

- Expression differentielle
- Enrichissement des fonctions des gènes cibles (Gene ontology, KEGG)
- Étude évolutive des miARNs prédits (inférence phylogénétique)



Travail pratique no 2

- Présentation de l'énoncé
- Réponse aux questions