Nouvelles approches optimisées pour le traitement de données sRNA-Seq

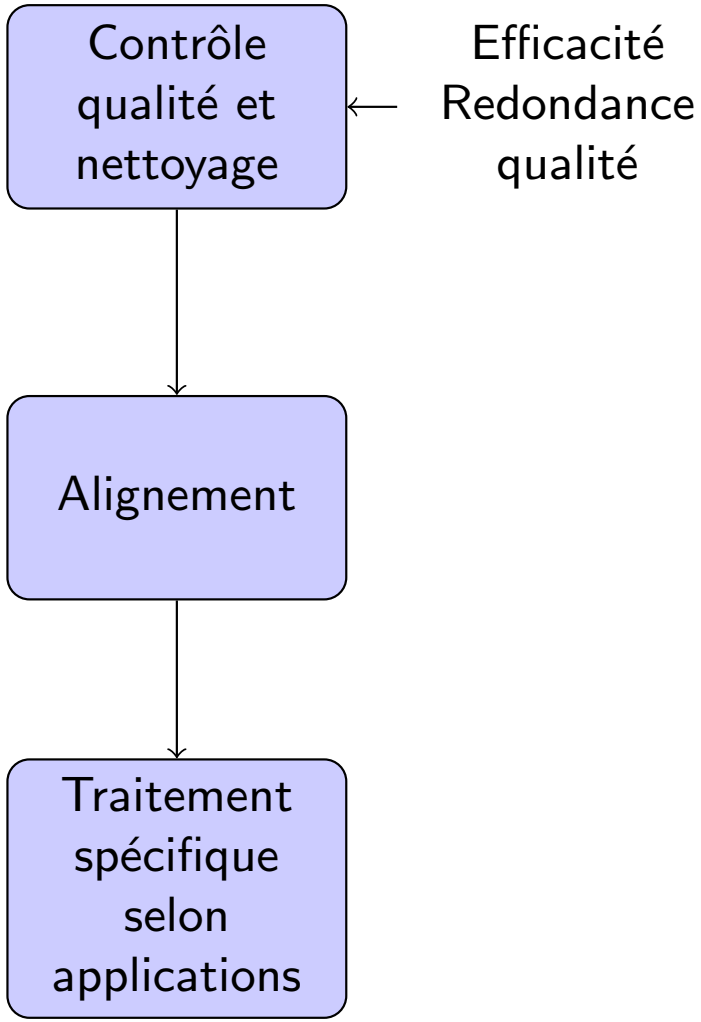
Walid BEN SAOUD BENJERRI

Sommaire

Génération de données

- Le génome = ensemble de séquences sur l'alphabet {A,T,C,G}, composé de plusieurs gènes, transcrits ou non.
- Le séquençage haut-débit
  - Assemblage de génomes
  - Analyse de variants (SNPs,...)
  - RNA-Seq : Accès aux gènes transcrits et leur niveau d'expression. En particulier on s'intéresse au sRNA-Seq : séquençage de petits ARNs.
- Verrou : Gros volume de données et redondance des lectures. Possibilité de synthétiser.

Pipeline générique d'un traitement de données

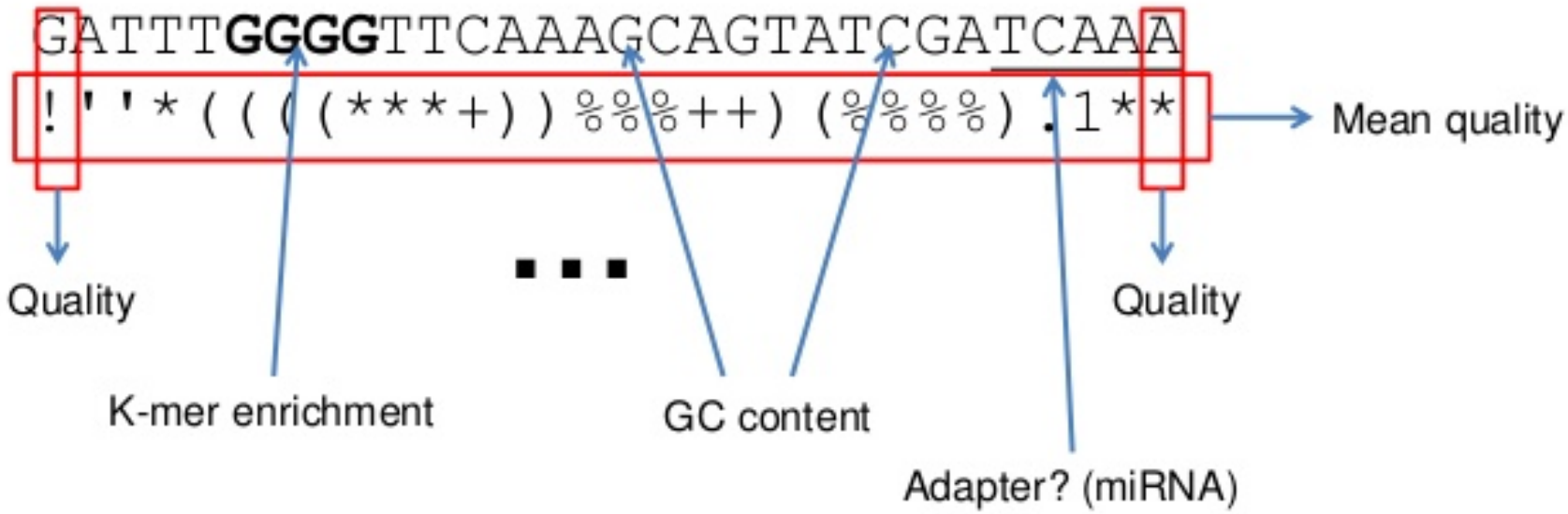


Exemple de fichier produit

Mock_A_TATAGCGA-GACACCGT_L001_R1_001.fastq

```
@HWI-M01141:63:A4NDL:1:1101:16668:1377 1:N:0:TATAGCGAGACACCGT
NACAGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCGCGTAGGTGGTTTGTTA
+
#>>AA>CAABBBGGGGGGGFFGHFGEGGGFHHHGHFEGCEHHFEGGGGG@EEHHGGGHGHHC
@HWI-M01141:63:A4NDL:1:1101:14849:1418 1:N:0:TATAGCGAGACACCGT
NACGAAGGGTGCAAGCGTTACTCGGAATTACTGGGCGTAAAGCGTGCGTAGGTGGTGGTTTA
+
#>>>>A??AFAA1BGGEGGAAFGGCA0BFF1D2BCF/EEG/DBEE/E?GAEEFGAEFAEFG1
@HWI-M01141:63:A4NDL:1:1101:13802:1421 1:N:0:TATAGCGAGACACCGT
NACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTA
+
#>>AAABBBABBGGGGGGG?FGHGGGGGHHHHHHHHHGGGGHHHGGGGGGEGGGGGGEGG?FD
@HWI-M01141:63:A4NDL:1:1101:15928:1426 1:N:0:TATAGCGAGACACCGT
NACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGCGCAGGCGGTTGTGCA
+
#>>AABFB@FBBGGGGGGGGGGHGGGGFHHHHHHHHHGGGGHHHGGGGGGGGGGGGGGEE?G4
@HWI-M01141:63:A4NDL:1:1101:14861:1431 1:N:0:TATAGCGAGACACCGT
NACGAAGGGTGCAAGCGTTACTCGGAATTACTGGGCGTAAAGCGTGCGTAGGTGGTGGTTTA
+
#>>AAAABBFABGGGGGGECEGHGGEFFHHHHHHHGHGGGGHHHGGGGGGEFGHHGGGGEGHE
@HWI-M01141:63:A4NDL:1:1101:15264:1465 1:N:0:TATAGCGAGACACCGT
NACGTAGGGTGCGAGCGTTGTCCGGAATTACTGGGCGTAAAGAGCTCGTAGGTGGTTTGTCC
+
```

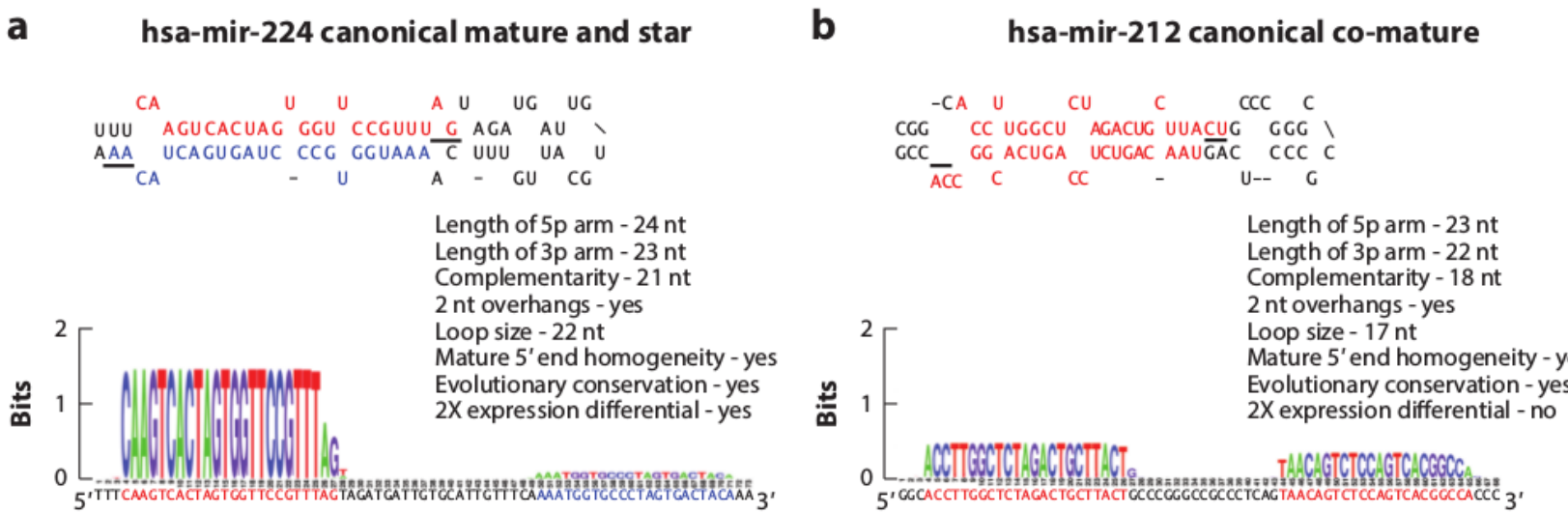
Une paire séquence/qualité



MicroARNs

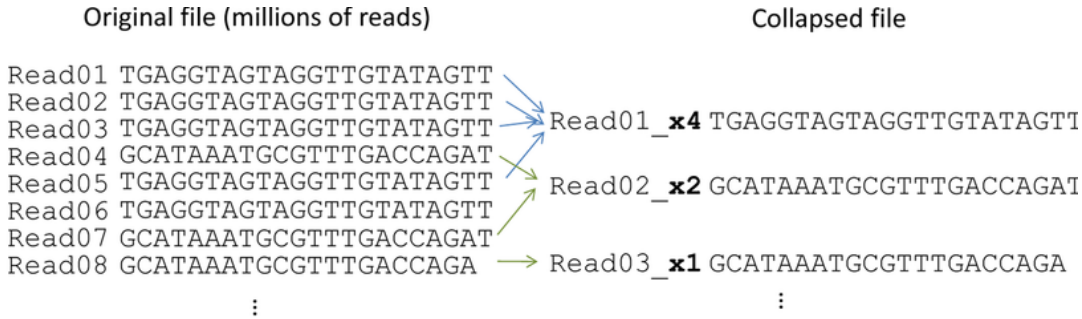
Les microARNs ont un profil d'expression abondant

- Critères biologiques pour déterminer si une lecture est ou non un bras d'un mir



sRNAcollapser

- Multiples séquences identiques, ces séquencent mappent au même endroit, on économise du temps en mappant une seule séquence



- Mais les outils de mapping utilisent la qualité, on perd de l'information.
- Problème : Après avoir collapsé N séquences, que faire des N qualités : $(Q_1, Q_2, ..., Q_n) => Q_{synthetic} =?$

sRNAcollapser

- Hypothèse : N lectures identiques viennent d'un même locus, correspondant à une séquence (inconnue) notée $s_1, ..., s_n$. On peut alors calculer une qualité à jour car pour une base donné une qualité supérieure discard toute qualité inférieur
- Ici la 1ère séquence nous dit que s_3 est C avec 30% de probabilités, la 2ème dit la même chose avec 40% de probabilités, et la troisième également avec 50% de probabilités. Synthèse : $P(s_3 = C) >= 50\%$

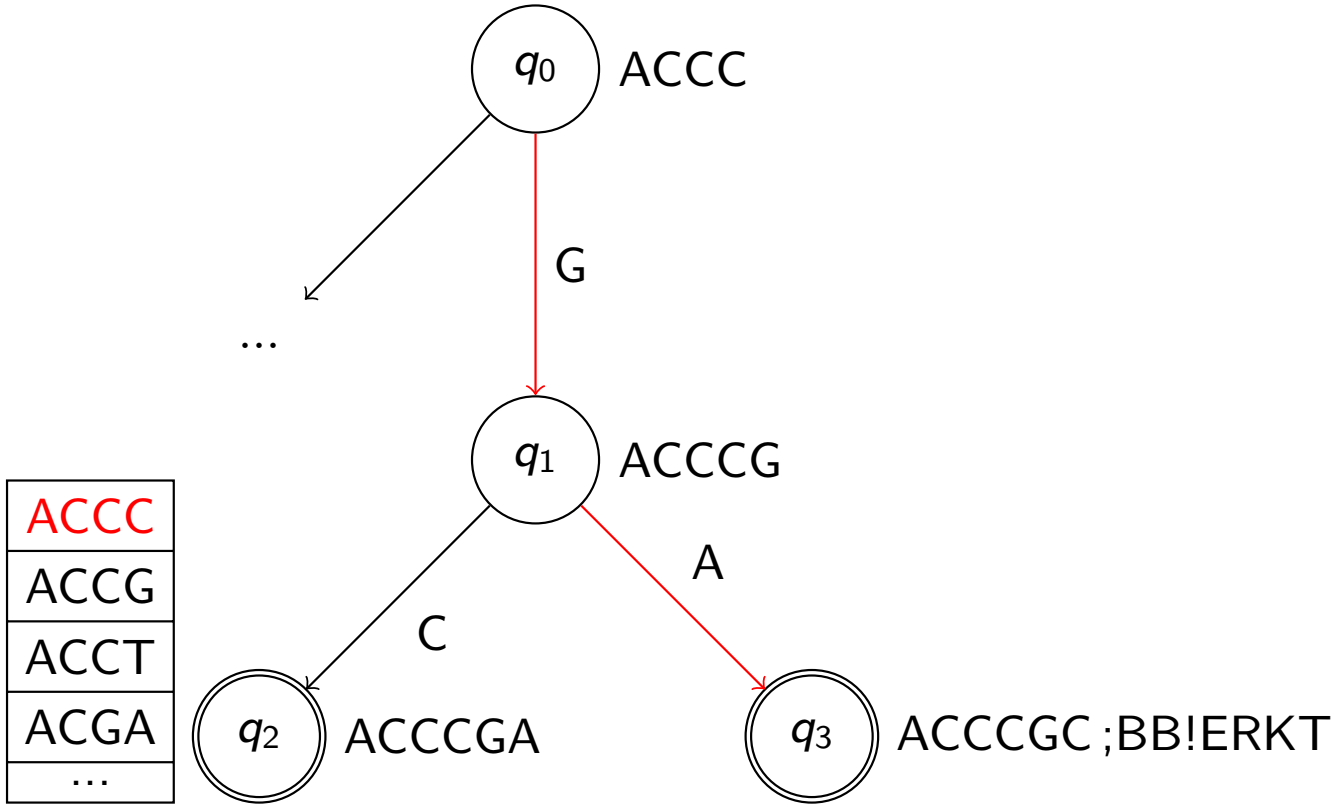
ATCG ——30–
ATCG ——40–
ATCG ——50–

- Comptage des lectures par fichier.

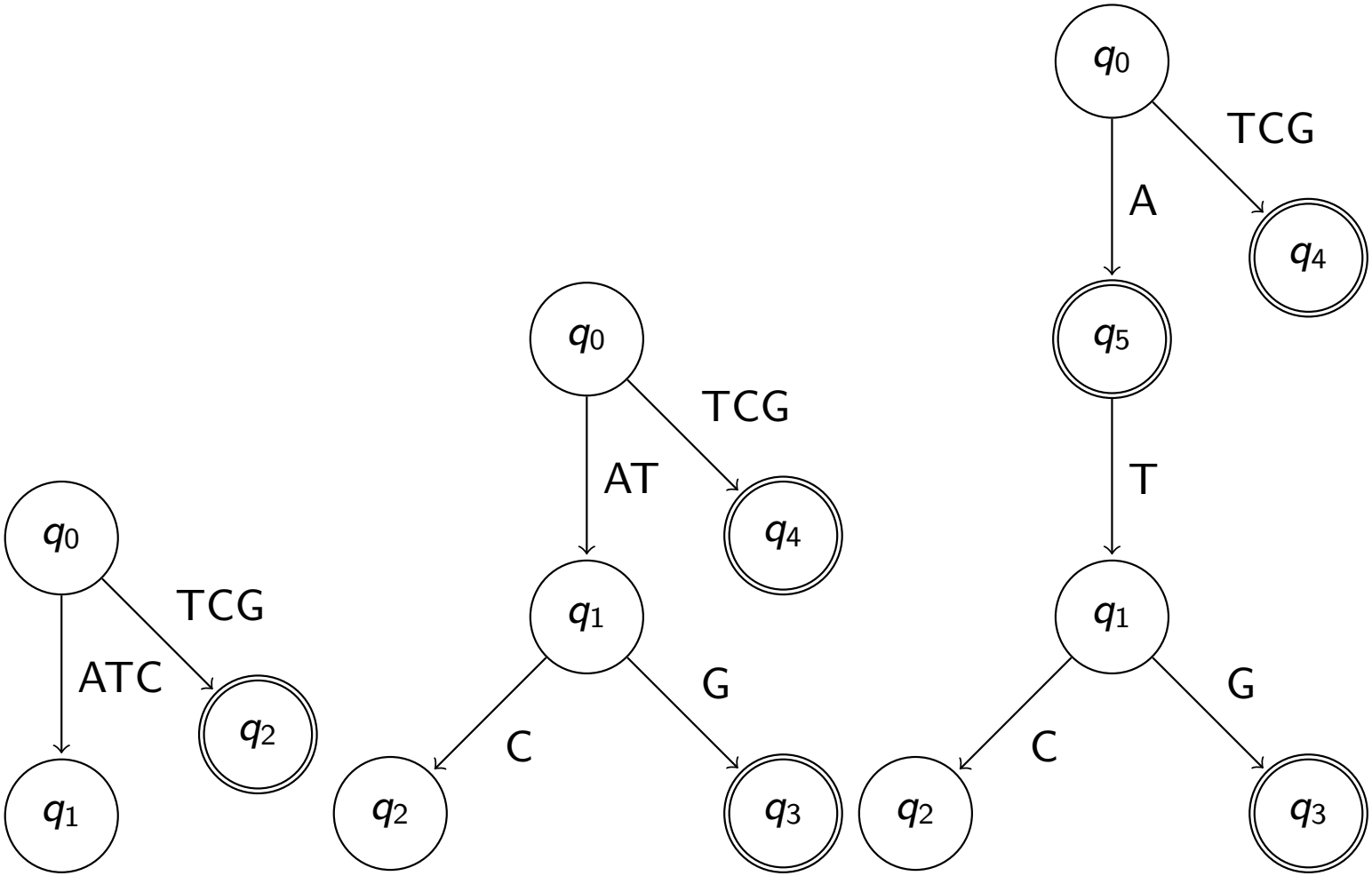
Implémentation

- Méthode : utilisation de plusieurs tries (un pour chaque préfixe de séquence) : celà augmente la localité de la référence car les noeuds d'un trie sont stoqués et lus ensemble.
- encodage binaire de l'ADN (2 bits overhead par nucléotide).
- les tries sont compacts, la recherche d'un LCP pour l'insertion se réduit à ctz (count trailing zeros) qui est une instruction directement implémentée dans les CPUs.
- implémenté en C++ pour un temps de calcul raisonnable même pour plusieurs gros fichiers. Optimizations bas niveau et multithreading pour optimiser l'usage de ressources.

sRNAcollapser tries



Exemples d'insertions



Résultats

	sRNAcollapser	Naïve
A.Thaliana 1.7GB	16s	$\geq$2 minutes
+ copy	+13s	$\geq$2 minutes (and incrementing)

Perspectives

- Correction d'erreurs
- Analyse différentielle
- Tableaux de hachage creux

Perspectives : autres outils en cours de développement

- Mettre ensemble des mirs ayant supposèment même fonction, comptabiliser erreurs de séquençage, séquences multimappés, comptage par famille au lieu de par locus pour éviter le rejet/biais
- Méthode : relier deux mirs si une certaine relation binaire est vraie