Analyse RNA-Seq SLA : Exploration comparative STAR / CRAC via le tag NH

Mickael Coquerelle

19 juin 2025

Contents

1 Introduction

À travers ce script, je cherche à analyser un sous-ensemble représentatif de lectures issues de plusieurs runs de séquençage ciblé SLA, précisément ceux des expériences suivantes : **202304, 202312, 202402 et 202404.** Ces runs ont été sélectionnés pour couvrir différentes périodes d'acquisition, mon objectif étant d'évaluer la robustesse et la cohérence des alignements produits par les mappeurs STAR et CRAC sur un échantillon diversifié de données.

Par ailleurs, l'objectif est de comparer la distribution des lectures mappées uniques par CRAC stockées dans le fichier*merge_reads_star_crac.tsv* avec le tag NH correspondant à l'aligneur STAR, en particulier la proportion de lectures non alignées ou multi-mappées.

Le fichier TSV, généré via un script Bash, sert à analyser la répartition des lectures mappées de manière unique par CRAC dans les différentes catégories d'alignement de STAR (non mappées, multi-mappées ou uniques). En étudiant la distribution des tags NH de STAR associés à ces lectures spécifiques à CRAC, l'objectif est d'identifier les mécanismes et paramètres sous-jacents susceptibles **d'expliquer les divergences d'alignement observées entre les deux outils, tous deux configurés avec un paramétrage minimal.**

2 Méthodologie

Chargement des librairies et du dataset

```
library(readr)
library(dplyr)
library(tidyr)
library(ggplot2)
library(data.table)
library(UpSetR)
```

Chargement

```
All_reads_merge <- read_tsv("merge_reads_star_crac.tsv", col_names = c("ReadID", "CRAC_FLAG", "CRAC_
                            ReadID = col_character(),
                            CRAC_FLAG = col_integer(),
                            CRAC_CHR = col_character(),
                            CRAC_POS = col_integer(),
                            CRAC_MAPQ = col_integer(),
                            CRAC_CIGAR = col_character(),
                            STAR_NH = col_integer()
                          ))
```

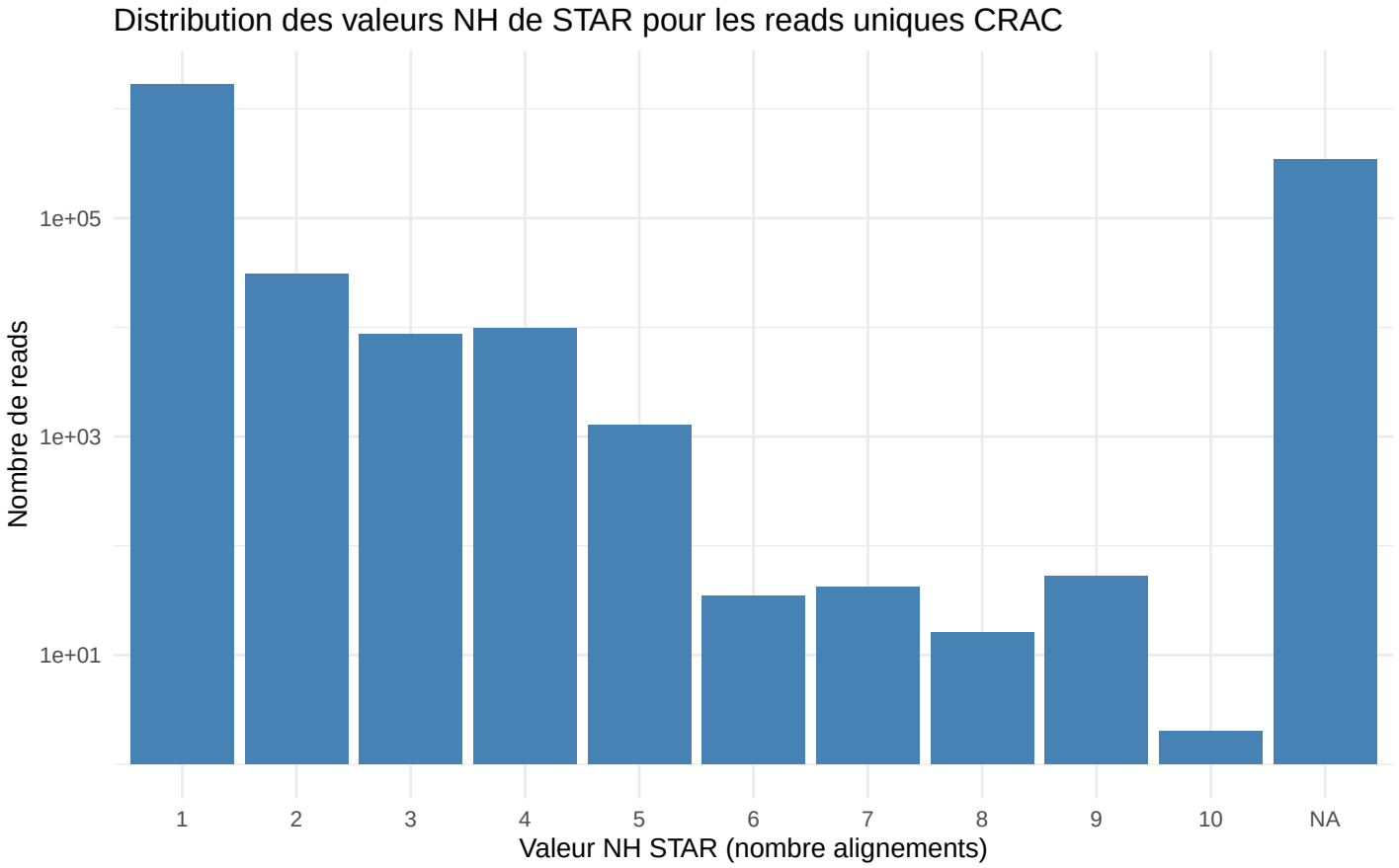
```
nh_grouped <- All_reads_merge %>%
  mutate(NH_group = case_when(
    is.na(STAR_NH) ~ "NA",
    STAR_NH == 1 ~ "Unique (NH=1)",
    STAR_NH > 1 ~ "Multiple (NH>1)"
  )) %>%
  count(NH_group, name = "n") %>%
  mutate(Proportion = round(100 * n / sum(n), 2))
```

Comptage des reads (uniques)

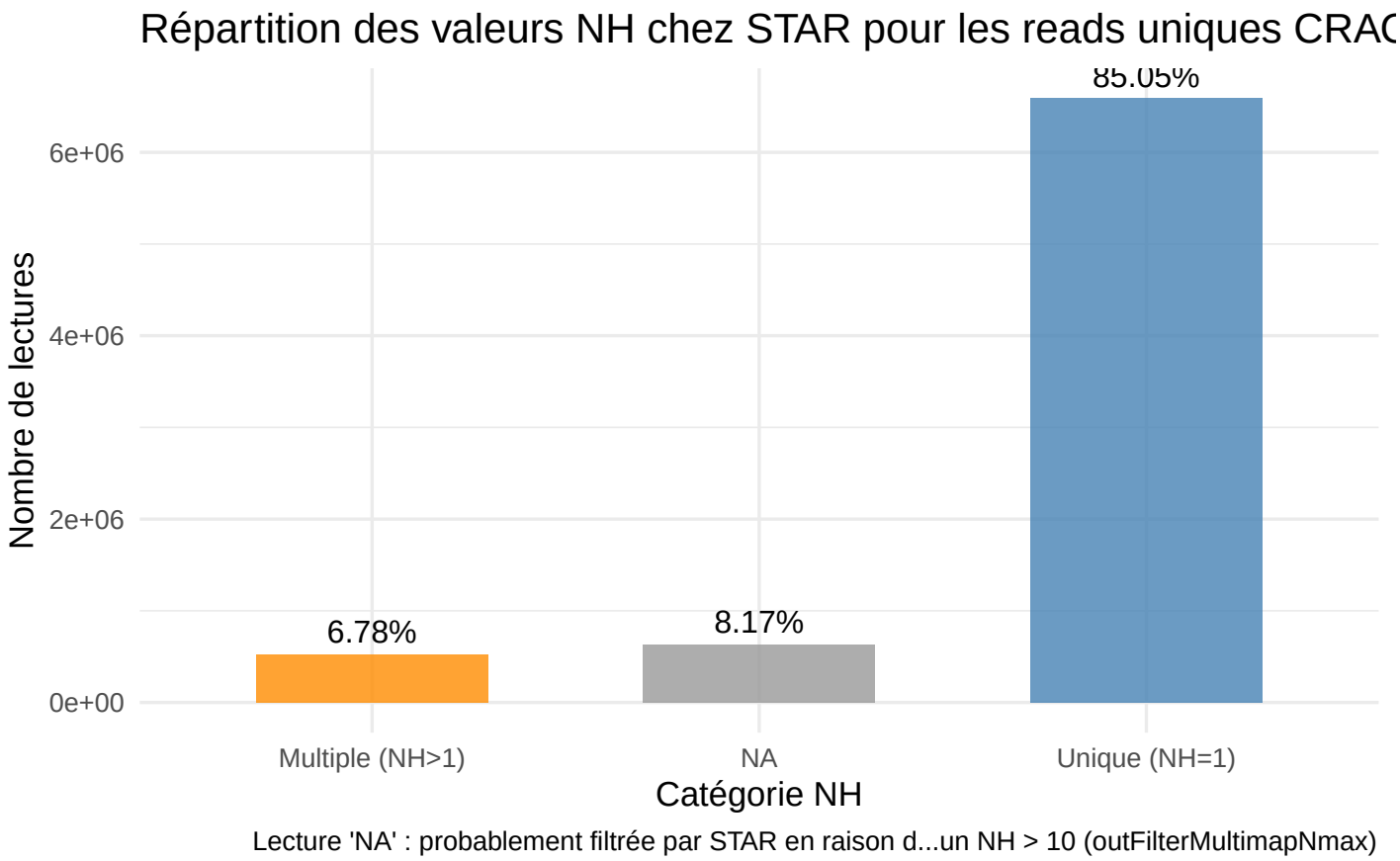
```
nh_counts <- All_reads_merge %>% distinct(ReadID, STAR_NH) %>% dplyr::count(STAR_NH, name ="n")
```

2.1 Analyse des flags NH et de la cohérence entre les flags STAR et CRAC

```
ggplot(nh_counts, aes(x = factor(STAR_NH), y = n)) +
  geom_col(fill = "steelblue") +
  scale_y_log10()+
  labs(title = "Distribution des valeurs NH de STAR pour les reads uniques CRAC",
       x = "Valeur NH STAR (nombre alignements)",
       y = "Nombre de reads") + theme_minimal()
```




```
ggplot(nh_grouped, aes(x = NH_group, y = n, fill = NH_group)) +
  geom_col(alpha = 0.8, width = 0.6) +
  geom_text(aes(label = paste0(Proportion, "%")), vjust = -0.5, size = 4.5) +
  labs(
    title = "Répartition des valeurs NH chez STAR pour les reads uniques CRAC",
    x = "Catégorie NH",
    y = "Nombre de lectures",
    caption = "Lecture 'NA' : probablement filtrée par STAR en raison d'un NH > 10 (outFilterMultima
  ) +
  scale_fill_manual(values = c("Unique (NH=1)" = "steelblue", "Multiple (NH>1)" = "darkorange", "NA"
  theme_minimal(base_size = 14) +
  theme(legend.position = "none")
```


Répartition des valeurs NH chez STAR pour les reads uniques CRAC
85.05%
6e+06
4e+06
2e+06
0e+00
Nombre de lectures
6.78%
8.17%
Multiple (NH>1)
NA
Unique (NH=1)
Catégorie NH
Lecture 'NA' : probablement filtrée par STAR en raison d...un NH > 10 (outFilterMultimapNmax)

Interprétation/ conclusion NH STAR

Le tag *NH:i*: correspond au nombre de loci auxquels un read s'aligne (nombre total d'alignements). Le tag est ajouté automatiquement lorsqu'on active l'option –outSAMattributes Standard (par défaut). Si un read s'aligne sur plusieurs loci, la ligne correspondante possède NH:i: et les alignements secondaires sont marqués par le flag 0x100 dans les FLAGS SAM.

NH:NA ou NA présents : pourquoi ? Le tag NH ne peut pas être NA dans un fichier `SAM/BAM` produit avec STAR (valeur entière >= 1). Pour autant, ici, n'ayant pas filtré les valeurs NA, on peut visualiser un nombre significatif, qui provient probablement de l'option –outSAMattributes Standard (pour NH =< 10). En effet, au-delà de 10 loci (*outFilterMultimapNmax*=10), STAR efface totalement ces alignements : il n'y a ni ligne SAM ni tag NH. Il ne s'agit pas d'un flag spécial « no map », mais d'un effet des filtres multi-mappers : les reads moins spécifiques sont traités comme non alignés.

2.2 Taux de récupération Crac versus STAR

```
total_reads <- length(unique(All_reads_merge$ReadID))
reads_star_mapped <- All_reads_merge %>% filter(STAR_NH == 1) %>% distinct(ReadID) %>% nrow()
reads_crac_mapped <- All_reads_merge %>%
  distinct(ReadID) %>% nrow()

data.frame(
  Outil = c("STAR", "CRAC"),
  Reads_mappés = c(reads_star_mapped, reads_crac_mapped),
  Pourcentage = round(c(reads_star_mapped, reads_crac_mapped) / total_reads * 100, 2)
)
```

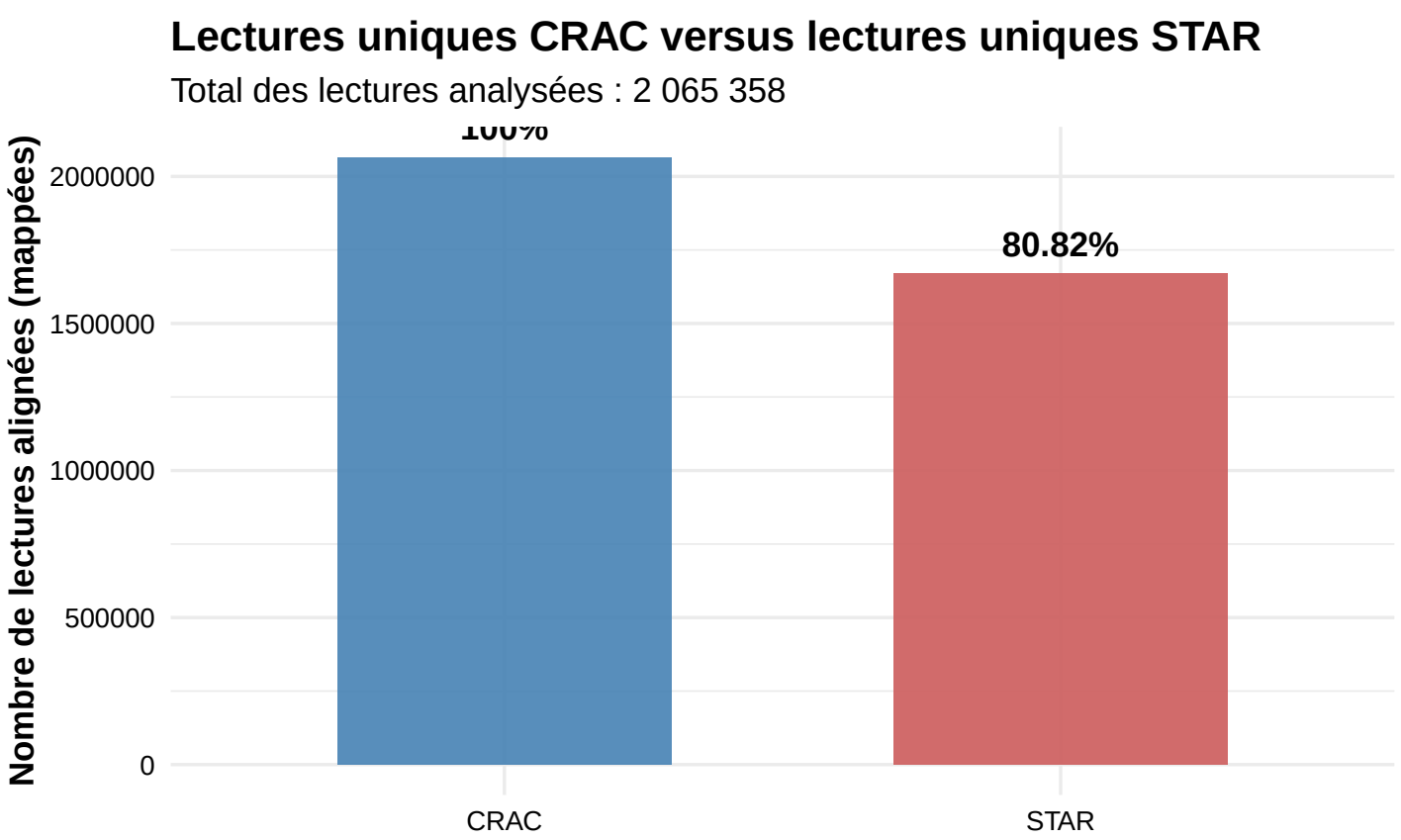
```
##   Outil Reads_mappés Pourcentage
## 1  STAR      1669253       80.82
## 2  CRAC      2065358      100.00
```

```
# Données de base
reads_stats <- data.frame(
  Outil = c("STAR", "CRAC"),
  Reads_mappés = c(reads_star_mapped, reads_crac_mapped),
  Pourcentage = round(c(reads_star_mapped, reads_crac_mapped) / total_reads * 100, 2)
)

# Graphique comparatif enrichi
ggplot(reads_stats, aes(x = Outil, y = Reads_mappés, fill = Outil)) +
  geom_col(width = 0.6, alpha = 0.9) +
  geom_text(aes(label = paste0(Pourcentage, "%")), vjust = -0.7, size = 5, fontface = "bold") +

  labs(
    title = "Lectures uniques CRAC versus lectures uniques STAR ",
    subtitle = paste("Total des lectures analysées :", format(total_reads, big.mark = " ")),
    y = "Nombre de lectures alignées (mappées)",
    x = NULL,
    caption = "Les pourcentages indiquent la proportion de lectures alignées par outil. Les lectures
  ) +
  scale_fill_manual(values = c("STAR" = "indianred", "CRAC" = "steelblue")) +
  theme_minimal(base_size = 14) +
```

```
  theme(
    legend.position = "none",
    plot.title = element_text(face = "bold"),
    axis.text = element_text(color = "black"),
    axis.title.y = element_text(face = "bold"),
    plot.caption = element_text(size = 9, color = "gray30")
  )
```



Les pourcentages indiquent la proportion de lectures alignées par outil. Les lectures CRAC incluent les alignements primaires (FLAG < 256).

Interprétation taux de récupération

```
All_reads_merge_binned <- All_reads_merge %>%
  mutate(
    NH_bin = case_when(
      is.na(STAR_NH) ~ NA_character_,
      STAR_NH >= 1 & STAR_NH <= 10 ~ as.character(STAR_NH),
      STAR_NH > 10 ~ ">10"
    )
  ) %>%
  filter(!is.na(NH_bin) & !is.na(CRAC_MAPQ)) %>%
  group_by(NH_bin, CRAC_MAPQ) %>%
  summarise(N_reads = n(), .groups = "drop")

# Graphique final
ggplot(All_reads_merge_binned, aes(x = factor(NH_bin, levels = c(as.character(1:10), ">10")),
                                   y = CRAC_MAPQ)) +
  geom_point(aes(size = N_reads), color = "steelblue3", alpha = 0.7) +
```

```r
  scale_size_continuous(
    name = "Nombre de lectures",
    range = c(1, 8),
    breaks = c(10, 100, 1000),
    labels = c("10", "100", "1000+")
  ) +
  labs(
    title = "Relation entre NH (STAR) et MAPQ (CRAC)",
    subtitle = "Chaque point représente un volume de lecture",
    x = "Nombre de positions d'alignement (NH - STAR)",
    y = "MAPQ (CRAC)",
    caption = "La ligne rouge indique MAPQ = 20 (seuil de confiance typique)."
  ) +
  geom_hline(yintercept = 20, linetype = "dashed", color = "firebrick", size = 1) +
  coord_cartesian(ylim = c(0, 256)) +
  theme_minimal(base_size = 14) +
  theme(
    panel.grid.minor = element_blank(),
    legend.position = "right"
  )
```