Analyse Comparative des Alignements RNA-Seq : STAR vs CRAC

Mickael Coquerelle

11 juin 2025

1 Introduction

Ce script à vocation d'effectuer une analyse comparative des performances d'alignement des outils STAR et CRAC pour tenter de démontrer si oui ou non, la capacité de l'outil d'alignement intervient significativement dans la proportion de lecture mappées de manière significative.

Ce rapport détaille la méthodologie appliquée au dataset RNA-Seq ciblé pour la SLA couvrant l'extraction, la préparation, la visualisation et l'analyse statistique des données d'alignement.

2 Méthodes

2.1 Chargement des librairies et nettoyage de l'environnement:

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(readr)
library(cowplot)
rm(list = ls())
```

2.2 Chargement du jeu de données à analyser :

Stats_Log_merge.csv, contient des métriques d'alignement issues des deux outils segmentées par patients et par expérience.

```
data <- read_csv("~/Stats_Log_merge.csv", show_col_types = FALSE)
```

2.3 Préparation des données STAR par patient:

```
data_star <- data %>%
  group_by(Patient) %>%
  summarise(
    Unique = sum(STAR_Unique_reads),
    Multiple = sum(STAR_Multi_reads),
    No_map = sum(STAR_No_map_reads + STAR_Chimeric_reads),
    .groups = "drop"
  ) %>%
  mutate(Outil = "STAR") %>%
  pivot_longer(cols = Unique:No_map, names_to = "Category", values_to = "Reads")
```

2.4 Préparation des données CRAC par patient:

```r
data_crac <- data %>%
  group_by(Patient) %>%
  summarise(
    Unique = sum(CRAC_Unique_reads / 2),
    Multiple = sum(CRAC_Multi_reads / 2) + sum(CRAC_Dup_reads / 2),
    No_map = sum(CRAC_No_map_reads / 2),
    .groups = "drop"
  ) %>%
  mutate(Outil = "CRAC") %>%
  pivot_longer(cols = Unique:No_map, names_to = "Category", values_to = "Reads")
```

2.5 Fusion et calcul des pourcentages par patient

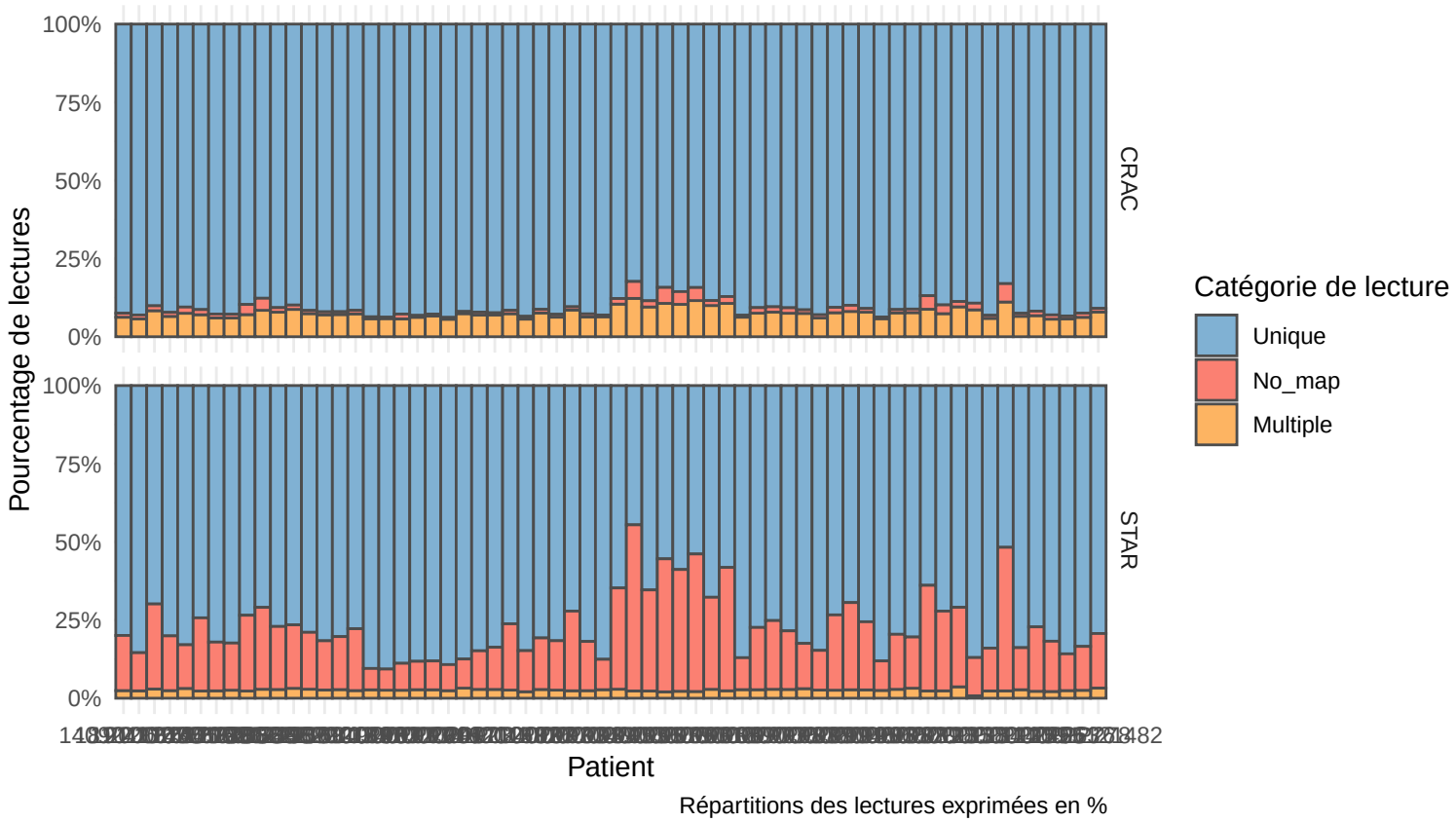
```r
data_plot <- bind_rows(data_star, data_crac) %>%
  group_by(Patient, Outil) %>%
  mutate(Percent = 100 * Reads / sum(Reads)) %>%
  ungroup()

data_plot <- data_plot %>%
  mutate(
    Patient = factor(Patient, levels = unique(data$Patient)),
    Category = factor(Category, levels = c("Unique", "No_map", "Multiple"))
  )
```

2.6 Sortie graphique /résultats:

```r
couleurs_categories <- c(
  "Unique" = "#80b1d3",
  "Multiple" = "#fdb462",
  "No_map" = "#fb8072"
)

ggplot(data_plot, aes(x = Patient, y = Percent, fill = Category)) +
  geom_bar(stat = "identity", position = "stack", color = "gray30", width = 1) +
  scale_fill_manual(values = couleurs_categories) +
  scale_y_continuous(
    labels = scales::label_percent(scale = 1),
    limits = c(0, 101)
  ) +
  facet_grid(Outil ~ ., scales = "free_y") +
  labs(
    title = "Étude comparative des alignements STAR et CRAC par patient",
    subtitle = "Index : GRCh37 -- STAR v2.7.8 & CRAC v2.5.2",
    x = "Patient",
    y = "Pourcentage de lectures",
    fill = "Catégorie de lecture",
    caption = "Répartitions des lectures exprimées en %"
  ) +
  theme_minimal(base_size = 11, base_family = "Helvetica")
```



2.7 Sauvegarde

```
ggsave("Figure_STAR_CRAC_P.pdf", width = 8, height = 5, dpi = 600)
```