y-genes

Kristina Nesporova

2025-01-09

Data merge and filtering

Metadata from EnteroBase are merged with the results of y-genes screen and genomes with low coverage are filtered out

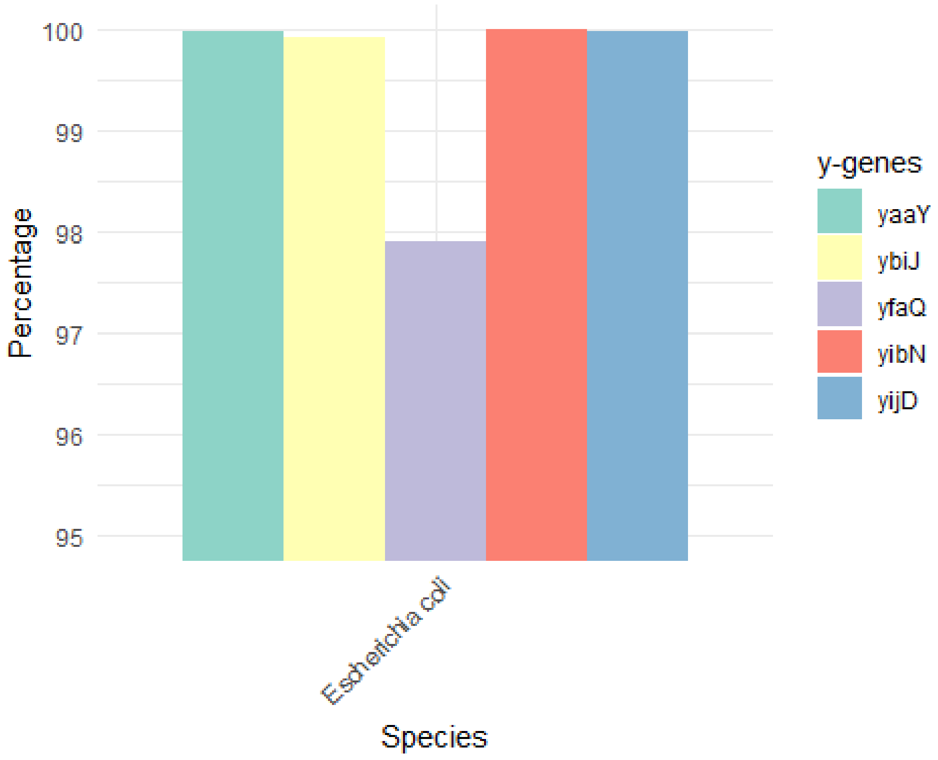
```
EB_PG_v.7_20240605 <- read.csv("~/y_genes/EB_PG_v.7_20240605.csv", sep=";",
stringsAsFactors=TRUE)

ygenes_05_2024 <- read.csv("~/y_genes/ygenes_05_2024.csv", sep=";",
stringsAsFactors=TRUE)

EB_PG_ygenes <- merge(EB_PG_v.7_20240605, ygenes_05_2024, by = "Barcode", all
= TRUE)

EB_PG_ygenes <- subset(EB_PG_ygenes, !is.na(Coverage)) #Filter out the
genomes without metadata (the metadata sheet has been already filtered for
low coverage and other issues)

write.table(EB_PG_ygenes, file = "EB_PG_ygenes.txt", sep = "\t", quote =
FALSE, row.names = F)
```

Prevalence of y-genes in E. coli collection

```
summary_data_ecoli.2 <- EB_PG_ygenes %>%
     group_by(Species) %>%
     summarize(
         yaaY = mean(yaaY == 1) * 100,
         yfaQ = mean(yfaQ == 1) * 100,
         yibN = mean(yibN == 1) * 100,
         yijD = mean(yijD == 1) * 100,
         ybiJ = mean(ybiJ == 1) * 100,
     ) %>%
     pivot_longer(cols = -Species, names_to = "y", values_to = "Percentage")
```

Figure-conservation B-1

```
ggplot(summary_data_ecoli.2, aes(x = Species, y = Percentage, fill = y)) +
          geom_bar(stat = "identity", position = "dodge") +
          labs(fill = "y-genes", y = "Percentage") +
          scale_fill_brewer(palette = "Set3") +  # Set a custom color palette
          theme_minimal() + coord_cartesian(ylim = c(95, 100)) +
          theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

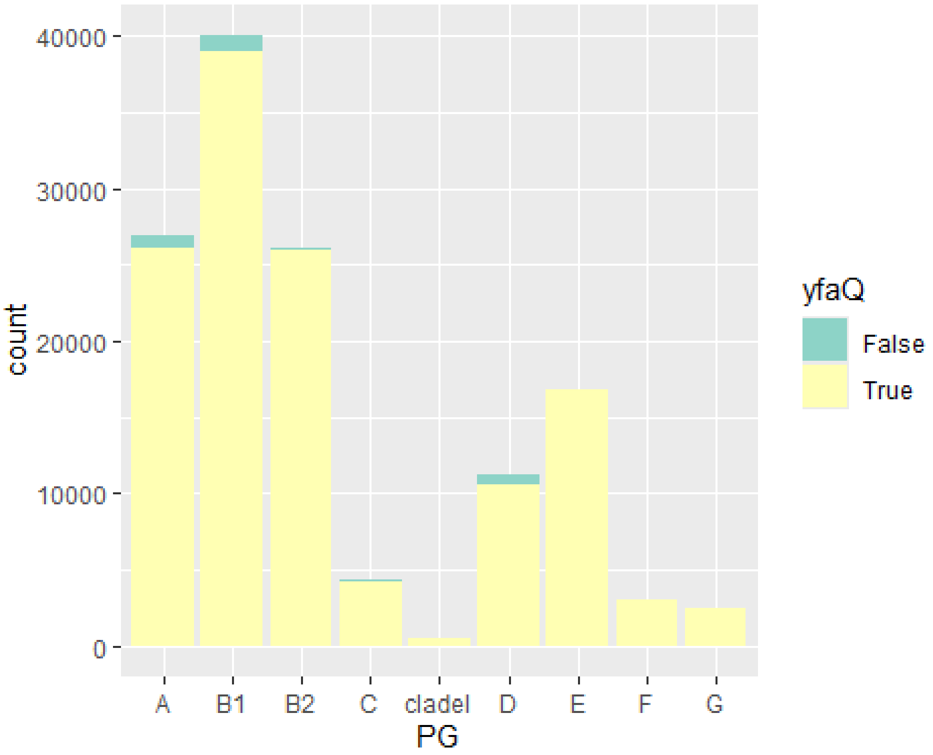


Prevalence of yfaQ in phylogenetic groups

```
EB_PG_ygenes.tf <- EB_PG_ygenes %>%
  mutate(across(c(yaaY, yfaQ, yibN, yijD, ybiJ),
                ~ ifelse(. == 1, "True", "False")))
```

Figure-conservation B-2

```
ggplot(EB_PG_ygenes.tf,aes(x=PG,fill=yfaQ)) + geom_bar() +
scale_fill_brewer(palette = "Set3")
```



Absence of yfaQ in specific sequence types

```
yfaQ_0 <- subset(EB_PG_ygenes, yfaQ == "0")

yfaQ_0_STs <- yfaQ_0 %>% count(ST)

write.table(yfaQ_0_STs, file = "yfaQ_0_ST.txt", sep = "\t", quote = FALSE,
row.names = F)
```

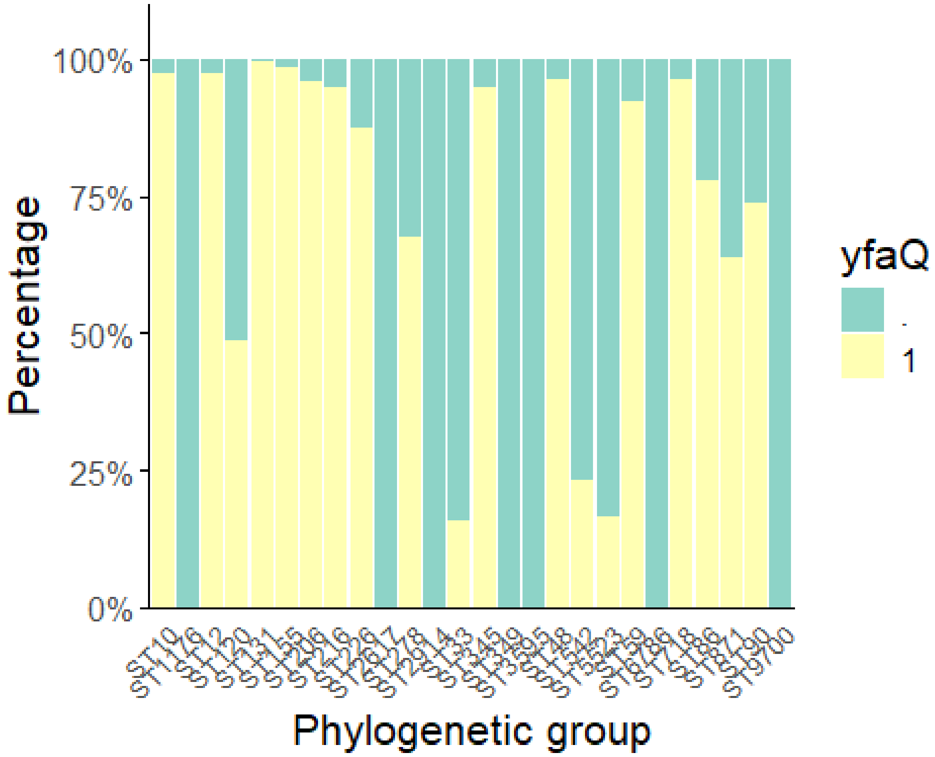
Main 26 STs wiht missign yfaQ

```
yfaQ_0_26ST <- subset(EB_PG_ygenes, ST %in% c("ST33", "ST349", "ST10",
"ST542", "ST278", "ST90", "ST3595", "ST131", "ST1176", "ST48", "ST86",
"ST155", "ST59", "ST12", "ST2914", "ST206", "ST120", "ST226", "ST871",
"ST345", "ST216", "ST6786", "ST2617", "ST9700", "ST5523", "ST718"))

proportions_26ST_yfaQ <- yfaQ_0_26ST %>%
         group_by(ST, yfaQ) %>%
          summarise(Count = n()) %>%
          mutate(Percentage = (Count / sum(Count)) * 100)

## `summarise()` has grouped output by 'ST'. You can override using the
`.groups`
## argument.
```

Figure-conservation B-2

```
ggplot(proportions_26ST_yfaQ, aes(x = ST, y = Percentage, fill = yfaQ)) +
     geom_bar(stat = "identity", position = "stack") +
     labs(
```

```r
        x = "Phylogenetic group",
        y = "Percentage",
        fill = "yfaQ"
    ) +
    scale_fill_brewer(palette = "Set3") +
    scale_y_continuous(labels = scales::percent_format(scale = 1), expand =
expansion(mult = c(0, 0.1))) +  # Ensure space at the top
    theme_minimal() +
    theme(
        panel.background = element_blank(),  # Remove grey background
        panel.grid = element_blank(),        # Remove gridlines
        axis.line = element_line(color = "black"),  # Highlight x and y axes
        axis.text.x = element_text(size = 10, angle = 45, hjust = 1),
        axis.text.y = element_text(size = 12),
        axis.title.x = element_text(size = 16),
        axis.title.y = element_text(size = 16),
        axis.ticks.y = element_line(color = "black"),  # Add y-axis ticks
        legend.title = element_text(size = 16),
        legend.text = element_text(size = 12)
    )
```


Percentage
ST10
ST9700
Phylogenetic group
yfaQ