Capstone Project Report

Emma Chandler

2025-05-30

Project Goal

The goal of this project was to map sequences of *E. coli* samples from a long-term genome evolution experiment (Tenaillon et al. 2017) to the *E. coli* Genome and analyze the number of variant sites that differ between the samples and the reference genome.

Link to GitHub Repository

The GitHub Repository includes all scripts described in the methods below, the output for the quality control analyses, and the csv file of the summary statistic from the genomic pipeline.

Methods

Three complete genomes from a long-term *E. coli* study were downloaded from the NCBI database (https://www.ncbi.nlm.nih.gov/bioproject/?term=414462) for Bio Project 414462 (Tenaillon et al. 2017). Quality control was run on all samples using the program FastQC (version 0.11.9-Java-11) and results from all samples were compiled using MultiQC (version 1.14-foss-2022a). The percentage of sequences that were adapters ranged from around 10% to 20% for the three sequences. To remove the adapters, the sequences were trimmed using Trimmomatic (version 0.39-Java-13).

The sequences were then aligned to the *E. coli* genome downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/017/985/GCA_000017985.1_ASM1798v1/GCA_000017985.1_ASM1798v1_genomic.fna.gz). The genome was indexed and aligned using the Burrows-Wheeler Aligner (BWA; version 0.7.18-GCCcore-13.3.0). SAMtools (version 1.18-GCC-12.3.0) was used to sort the aligned sequences and BCFtools (version 1.18-GCC-12.3.0) was used to call variants.

Read counts for the raw, trimmed, and aligned sequences were extracted. Since the sequences are paired-end reads, the aligned read counts were determined using the SAMtools view command with the -F 0x4 flag, which pulls out only reads that were mapped to the genome. However, those reads may include reads that mapped to multiple parts in the genome. To only include primary reads, the unique reads were isolated for the read count. The number of variants were also extracted for analysis. The read counts and variants were plotted in R 4.4.1 (R Core Team, 2024).

Analysis of Summary Statistics

Load Data and Packages

```r
# Load Packages
library(tidyverse)
library(ggpubr)
```

```r
# Load data
stats <- read.csv("results/summary_stats.csv", header = TRUE)
```

Read counts for each sample

```r
# Factor to order read type and samples in desired order
stats$read_type <- factor(stats$read_type, levels = c("Raw", "Trimmed", "Aligned", "Variants"))
stats$sample <- factor(stats$sample, levels = c("SRR2584866", "SRR2584863", "SRR2589044"))

# Convert counts from individual to every million counts
stats <- stats %>%
  mutate(count_mill = count / 1e+6)

# Plot including the raw, trimmed, and aligned read counts
plot1 <- stats %>%
  filter(read_type != "Variants")  %>%
  ggplot() +
    aes(x = read_type, y = count_mill, group = sample, color = sample) +
    geom_line(size = 1.2) +
    theme_classic() +
    labs(x = "", y = "Read Counts (million)", color = "Sample") +
    scale_color_viridis_d() +
    theme(legend.position = "none")

plot2 <- stats %>%
  filter(read_type != "Variants")  %>%
  ggplot() +
    aes(x = read_type, y = count_mill, group = sample, color = sample) +
    geom_line(size = 1.2) +
    theme_classic() +
    labs(x = "", y = "Read Counts (million)", color = "Sample") +
    scale_color_viridis_d() +
    facet_wrap(~ sample, ncol = 1, scales = "free")

ggarrange(plot1, plot2, nrow = 1, labels = c("A", "B"))
```

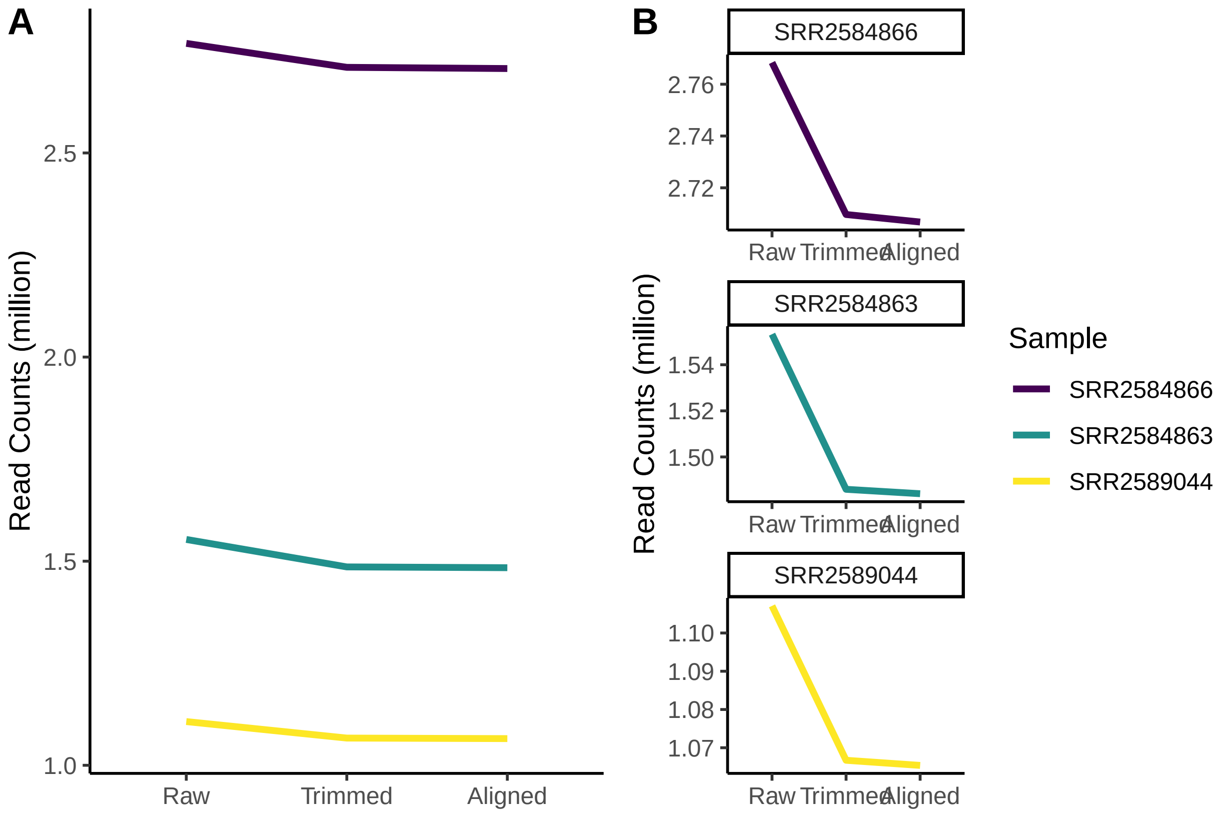

Figure 1. Read counts in million reads for the raw, trimmed, and aligned sequences of each of the three samples plotted together (A) and separately (B). Note the differing y-axes in plot B.

```
# Compare average read counts
stats %>%
  select(!count_mill) %>%
  pivot_wider(names_from = read_type, values_from = count) %>%
  summarise_all(mean) %>%
  summarise(Raw_Trimmed_diff = Raw - Trimmed,
            Percent_Aligned = Aligned / Trimmed)
```

```
## # A tibble: 1 x 2
##   Raw_Trimmed_diff Percent_Aligned
##              <dbl>           <dbl>
## 1           55465.           0.999
```

Variants in each sample

```
# Plot a bar chart of variants
stats %>%
  filter(read_type == "Variants") %>%
  ggplot() +
    aes(x = sample, y = count, fill = sample) +
    geom_col() +
    theme_classic() +
    scale_fill_viridis_d() +
    labs(x = "Sample", y = "Variants") +
    theme(legend.position = "none")
```

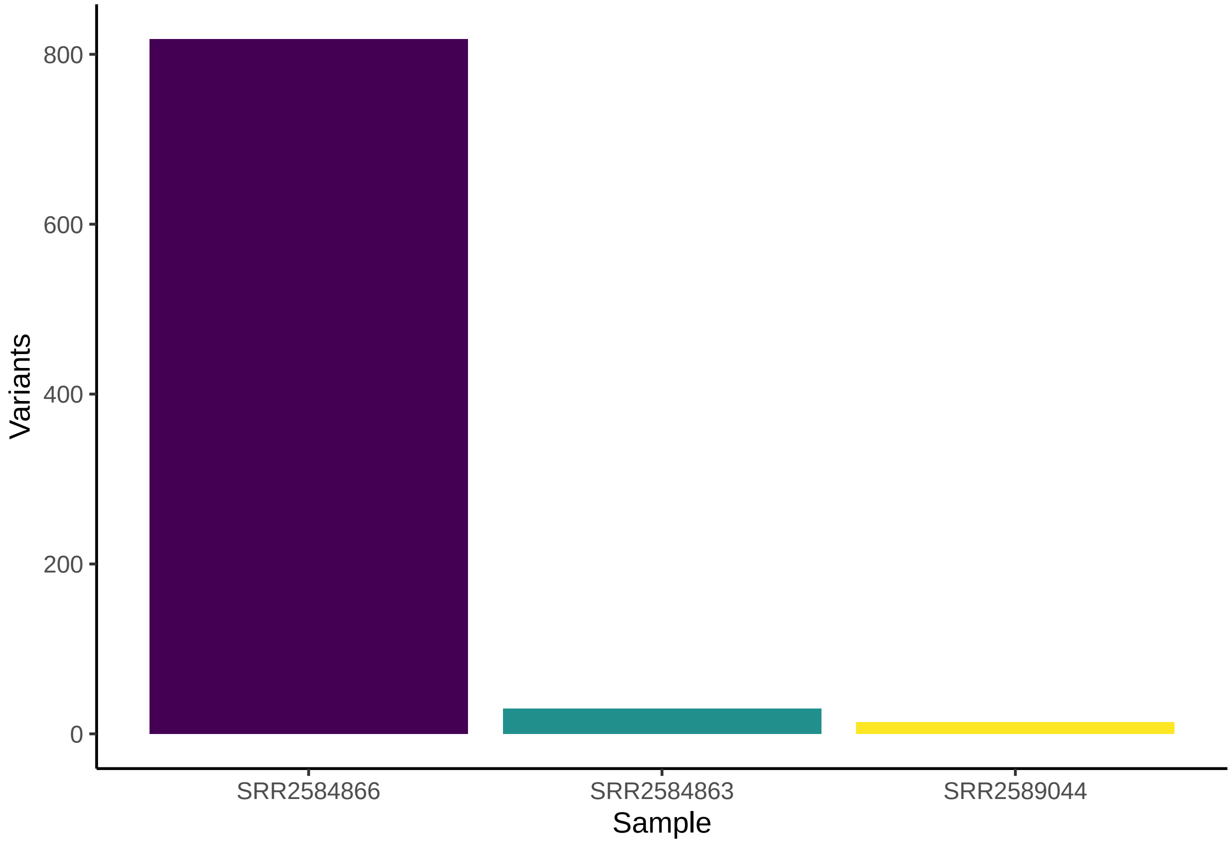

Figure 2. The number of variant sites that differ between the *E. coli* reference genome and sequences in each of the three samples.

Interpretation

The raw read counts differed substantially among the three samples. SRR2584866 had a much higher number of reads (2.8 mill) than the other two sequences (1.6 mill and 1.1 mill; Figure 1A). Trimming the sequences dropped the read lengths by about 55,000 on average (Figure 1B). Almost all (99.9%) of the reads aligned to the reference genome (Figure 1B).

Similar to the number of reads, the number of variants was much higher for the SRR2584866 sample (818) compared to the other two samples (30 and 14; Figure 2). However, relative to the number of reads, the SRR2584866 sample had a much higher number of variants than the other two samples. This suggests that the SRR2584866 sample is more genetically different from the reference genome than the other two samples.

References

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Tenaillon O, Barrick JE, Ribeck N, Deatherage DE, Blanchard JL, Dasgupta A, Wu GC, Wielgoss S, Cruveiller S, Médigue C, Schneider D, Lenski RE. Tempo and mode of genome evolution in a 50,000-generation experiment. Nature. 2016 Aug 11;536(7615):165-70. doi: 10.1038/nature18959.