Coverage Metrics

Payel Sardar

Setting the home working directory.

```
setwd("D:/KCL2024/Courses/7BBG1002_Cloud_computing/Project")
```

Reading the csv file that contains the total number of reads(sequences) per sample. This file was made from the fastqc reports of raw reads before any pre-processing

```
library_size_info <- read.csv("../metadata/sample_library_size.csv",
                              sep = ",", header = TRUE)
library_size_info
```

```
   Sample_id Total_Sequences
1 SRR1166442        40585197
2 SRR1166443        44937781
3 SRR1166444        34778661
4 SRR1166445        45543917
5 SRR1166446        38223389
6 SRR1166447        34542268
```

Computing the Average library size of the samples

```
avg_lib_size <- mean(library_size_info$Total_Sequences, na.rm = TRUE)
print(avg_lib_size)
```

```
[1] 39768536
```

From the FastQC report, the average raw read length across all samples was determined to be 50 bp. After pre-processing with Trimmomatic, the average read length was reduced to 35 bp.

```r
raw_avg_read_len <- 50
read_len_post_trim <- 35
```

Getting the length of the reference genome using the Biostrings package from BiocManager

```r
library(Biostrings)
```

```
Loading required package: BiocGenerics


Attaching package: 'BiocGenerics'

The following objects are masked from 'package:stats':

    IQR, mad, sd, var, xtabs

The following objects are masked from 'package:base':

    anyDuplicated, aperm, append, as.data.frame, basename, cbind,
    colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
    get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
    match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
    Position, rank, rbind, Reduce, rownames, sapply, setdiff, table,
    tapply, union, unique, unsplit, which.max, which.min

Loading required package: S4Vectors

Loading required package: stats4


Attaching package: 'S4Vectors'

The following object is masked from 'package:utils':

    findMatches

The following objects are masked from 'package:base':

    expand.grid, I, unname
```

```
Loading required package: IRanges


Attaching package: 'IRanges'

The following object is masked from 'package:grDevices':

    windows

Loading required package: XVector

Loading required package: GenomeInfoDb


Attaching package: 'Biostrings'

The following object is masked from 'package:base':

    strsplit
```

```
# Loading the reference genome FASTA
genome <- readDNAStringSet("../data/raw/ncbi_dataset/data/GCF_000146045.2/genome.fa")

# Calculating the total genome length
ref_genome_length <- sum(width(genome))

cat("Total genome length:", ref_genome_length, "bp\n")
```

```
Total genome length: 12157105 bp
```

Computing the coverage before and after pre-processing

```
raw_coverage <- (avg_lib_size * raw_avg_read_len)/ref_genome_length
coverage_post_trim <- (avg_lib_size * read_len_post_trim)/ref_genome_length

cat("Raw coverage:", raw_coverage, "\n")
```

```
Raw coverage: 163.5609
```

```r
cat("Coverage post-trim:", coverage_post_trim, "\n")
```

```
Coverage post-trim: 114.4926
```

However, the original study reported the coverage to be 156-fold. Since no-preprocessing on the sample datasets were performed, we assume the reported value to be raw coverage.

```r
reported_coverage <- 156

# Calculating the absolute difference
abs_difference <- abs(raw_coverage - reported_coverage)

# Calculating the percentage difference
percentage_difference <- (abs_difference / reported_coverage) * 100

# Printing the results
cat("Absolute difference:", abs_difference, "fold\n")
```

```
Absolute difference: 7.560879 fold
```

```r
cat("Percentage difference:", percentage_difference, "%\n")
```

```
Percentage difference: 4.846717 %
```