Normalisation Quality Control

Payel Sardar

Description

This script takes in individual normalised files and merges then into a unified dataset. A Box plot is then created to check if the samples for corresponding carbon source (glucose and cellobiose) get clustered together.

Setting the home working directory.

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

Loading necessary packages

```r
library(dplyr)
```

```

Attaching package: 'dplyr'

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

    filter, lag

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

    intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.4.2
```

```
library(tidyr)
```

Reading the file on sample information

```
sample_info <- read.csv("../metadata/sample_types.csv",
                        sep = ",", header = TRUE)
head(sample_info)
```

```
         Run    BioSample  Bases   Bytes carbon_source Experiment GEO_Accession
1 SRR1166442 SAMN02639514 2.03 G 1.29 Gb       Glucose  SRX468698    GSM1324496
2 SRR1166443 SAMN02639516 2.25 G 1.42 Gb       Glucose  SRX468699    GSM1324497
3 SRR1166444 SAMN02639513 1.74 G 1.05 Gb       Glucose  SRX468700    GSM1324498
4 SRR1166445 SAMN02639515 2.28 G 1.44 Gb    Cellobiose  SRX468701    GSM1324499
5 SRR1166446 SAMN02639512 1.91 G 1.20 Gb    Cellobiose  SRX468702    GSM1324500
6 SRR1166447 SAMN02639517 1.73 G 1.04 Gb    Cellobiose  SRX468703    GSM1324501
          create_date Sample.Name            source_name
1 2014-02-1010:25:00Z  GSM1324496    Glucose-grown cells
2 2014-02-1010:25:00Z  GSM1324497    Glucose-grown cells
3 2014-02-1010:24:00Z  GSM1324498    Glucose-grown cells
4 2014-02-1010:25:00Z  GSM1324499 Cellobiose-grown cells
5 2014-02-1010:24:00Z  GSM1324500 Cellobiose-grown cells
6 2014-02-1010:24:00Z  GSM1324501 Cellobiose-grown cells
```

Next, the sample_ids(Run) for each carbon_source is extracted

```
glucose_samples <- sample_info %>%
  filter(carbon_source == "Glucose") %>%
  pull(Run)
cellobiose_samples <- sample_info %>%
  filter(carbon_source == "Cellobiose") %>%
  pull(Run)
```

The function below reads multiple normalized RPKM files (in CSV format) for a set of sample IDs corresponding to a particular carbon_source, extracts the first column (Geneid) and the last column (normalized_rpkm) from each file, and merges the data frames by the Geneid column to create a combined dataset.

```
group_samples <- function(sample_ids, prefix = "normalized_",
                          folder = "../data/processed/Normalized_data/") {

  # Initializing an empty list to store data
  data_list <- list()

  for (sample_id in sample_ids) {

    file_name <- paste0(folder, prefix, sample_id, ".csv")

    # Reading the normalized file
    data <- read.csv(file_name)

    # Extracting the first column(Geneid) and the last column(normalized_rpkm)
    data_subset <- data[, c(1, ncol(data))]

    # Renaming the last column to the sample ID for identification
    colnames(data_subset)[2] <- sample_id

    # Appending the file data the list
    data_list[[sample_id]] <- data_subset
  }

  # Merging all data frames by "gene_id"
  merged_data <- Reduce(function(x, y) full_join(x, y, by = "Geneid"), data_list)

  return(merged_data)
}
```

Applying the function for glucose and cellobiose Samples

```
glucose_data <- group_samples(glucose_samples)
cellobiose_data <- group_samples(cellobiose_samples)
```

Veiwing the merged datasets

```
head(glucose_data)
```

```
     Geneid SRR1166442 SRR1166443 SRR1166444
1   YAL068C  8.4379100 11.1595834  10.898455
2 YAL067W-A  0.3630822  0.3290228   0.564276
3   YAL067C 55.5291618 53.1827885  40.971760
4   YAL065C  3.4225421  3.5537864   2.825754
5 YAL064W-B  3.3315888  4.0691744   3.292351
6 YAL064C-A 10.1396180 10.1729361   8.273086
```

```
head(cellobiose_data)
```

```
     Geneid SRR1166445 SRR1166446 SRR1166447
1   YAL068C 11.6956377  9.3393236 10.1604814
2 YAL067W-A  0.5115573  0.4956395  0.1336905
3   YAL067C 15.5382521 17.9623442 15.2407221
4   YAL065C  1.8685722  2.3360375  2.5991929
5 YAL064W-B  3.0000621  2.7435795  2.8801365
6 YAL064C-A 10.1022500  7.8599846 10.0004739
```

Saving the merged dataframes as CSV files in the Normalized_data folder for downstream processing

```
write.csv(glucose_data, "../data/processed/Normalized_data/glucose_merged.csv", row.names = 
write.csv(cellobiose_data, "../data/processed/Normalized_data/cellobiose_merged.csv", row.na
```

Merging the two datasets for visualisation

```
merged_norm_rpkm <- merge(glucose_data, cellobiose_data,
                          by.x = "Geneid", by.y = "Geneid")
write.csv(merged_norm_rpkm, "../data/processed/Normalized_data/CombinedNF.csv", row.names = 
```

```
# Removing gene_id column
merged_data_numeric <- merged_norm_rpkm[, -1]
# Setting gene IDs as row names
row.names(merged_data_numeric) <- merged_norm_rpkm$Geneid

head(merged_data_numeric)
```

```
      SRR1166442 SRR1166443 SRR1166444 SRR1166445 SRR1166446 SRR1166447
Q0020 0.10040356 0.21229852 0.13653495 0.50926167 0.54823925  0.3327265
Q0045 0.03438535 0.04673969 0.04007942 0.27614533 0.17602152  0.1329409
```

```
Q0050 0.11015667 0.09982329 0.07703888 0.21418017 0.28195064  0.1946919
Q0055 0.08606393 0.05849294 0.06269734 0.32739668 0.36346900  0.3446245
Q0060 0.02211077 0.04007329 0.02577222 0.05607455 0.02263738  0.0000000
Q0065 0.00000000 0.02992906 0.00000000 0.05583965 0.03381383  0.0000000
```

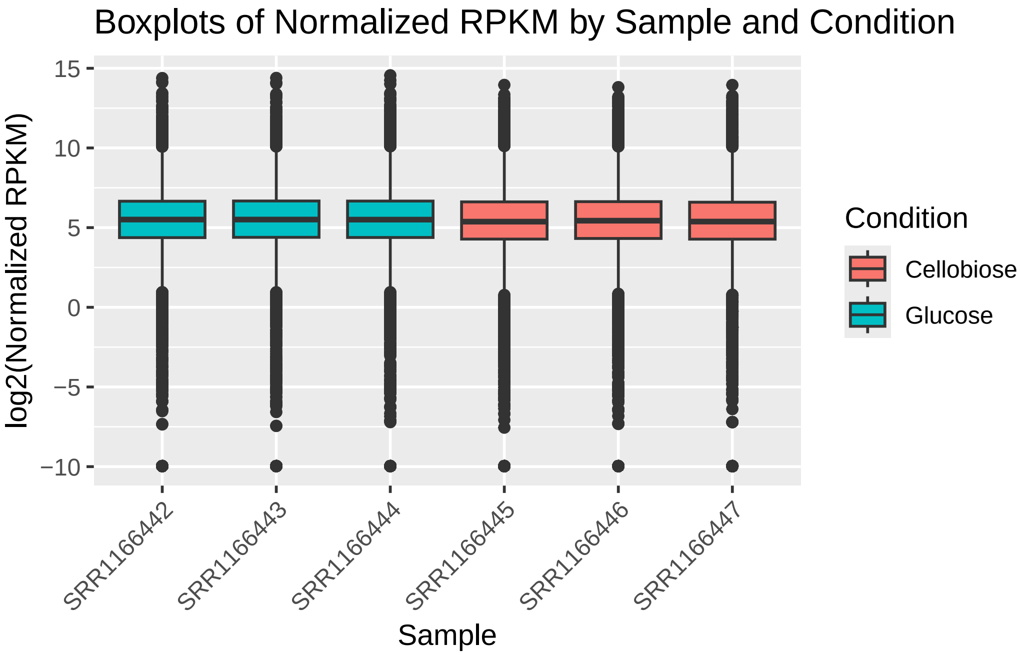
Preparing the Data for making the Box plot

```r
data_long <- pivot_longer(merged_data_numeric,
                          cols = everything(),
                          names_to = "Sample",
                          values_to = "Norm_RPKM")
data_long$log2_norm_rpkm <- log2(data_long$Norm_RPKM + 0.001) # to avoid log2(0)

sample_conditions <- c(rep("Glucose", 3), rep("Cellobiose", 3))

data_long$Condition <-
  rep(sample_conditions, times = nrow(data_long)/length(sample_conditions))

head(data_long)
```

```
# A tibble: 6 x 4
  Sample     Norm_RPKM log2_norm_rpkm Condition
  <chr>          <dbl>          <dbl> <chr>
1 SRR1166442     0.100         -3.30  Glucose
2 SRR1166443     0.212         -2.23  Glucose
3 SRR1166444     0.137         -2.86  Glucose
4 SRR1166445     0.509         -0.971 Cellobiose
5 SRR1166446     0.548         -0.864 Cellobiose
6 SRR1166447     0.333         -1.58  Cellobiose
```

Creating the box plots

```r
plot <- ggplot(data_long, aes(x = Sample, y = log2_norm_rpkm, fill = Condition)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Boxplots of Normalized RPKM by Sample and Condition",
       x = "Sample", y = "log2(Normalized RPKM)")
print(plot)
```



Saving the plot as an image

```
ggsave("../Output/plots/boxplot_rpkm_comparison.jpg",
       plot = plot, width = 10, height = 6, dpi = 300)
```