The UK Crop Microbiome Cryobank - mapping the files for the fastq checklist

Payton Yau

2024-03-11

The UK Crop Microbiome Cryobank

The UK Crop Microbiome Cryobank integrates genomic (DNA) data with a cryobank collection of samples for the soil microbiomes of the UK major crop plant systems. For this project, the microbiomes are from the rhizosphere (the soil surrounding the crop plant roots) and from bulk soil (soil outside the rhizosphere). The Cryobank provides a facility for researchers to source data and samples, including cryo-preserved microbial material and genomic and metagenomic sequences from different soil microbiome environments.

The script below was used for mapping the information for fastq checklist

Files required:

1. The receipt after the plant checklist uploaded to the server
2. Plant checklist ERC000020 (The plant checklist was preiously prepared)
3. MD5Checksum information

Step 1

Load the table from the receipt `Webin` after the plant Checklist uploaded to the server

```
Webin <- read.table("Webin-accessions-2023-12-07T15_42_52.222Z_OR.txt", header = T)

# Check the last 5 columns
tail(Webin)
```

```
##          TYPE   ACCESSION                                                ALIAS
## 37     SAMPLE ERS27620606                                             ms00351
## 38     SAMPLE ERS27620607                                             ms00352
## 39     SAMPLE ERS27620608                                             ms00355
## 40     SAMPLE ERS27620609                                             ms00356
## 41     SAMPLE ERS27620610                                             ms00357
## 42 SUBMISSION ERA27692675 ena-SUBMISSION-TAB-07-12-2023-15:42:48:241-1362
```

```r
# Remove the last column of the dataframe - `Webin`
Webin <- head(Webin, -1)

# Check the last 5 columns
tail(Webin)
```

```
##      TYPE   ACCESSION   ALIAS
## 36 SAMPLE ERS27620605 ms00350
## 37 SAMPLE ERS27620606 ms00351
## 38 SAMPLE ERS27620607 ms00352
## 39 SAMPLE ERS27620608 ms00355
## 40 SAMPLE ERS27620609 ms00356
## 41 SAMPLE ERS27620610 ms00357
```

Step 2

Load the data from the Crop Check-list

```r
Check.list <- read.delim('Checklist_GSC-MIxS_16Samplicons_OR_TESTv1.tsv', header = F, sep = "\t")

# Print the data (first 6 columns) to check if it has been read correctly
head(Check.list)
```

```
##          V1              V2                          V3           V4
## 1 Checklist       ERC000020 GSC MIxS plant associated
## 2    tax_id scientific_name              sample_alias sample_title
## 3    #units
## 4    410658 soil metagenome                  ms00315  OR-CL-BO-01
## 5    410658 soil metagenome                  ms00316  OR-CL-BO-02
## 6    410658 soil metagenome                  ms00317  OR-CL-BO-03
##                                            V5                          V6
## 1
## 2                          sample_description               project name
## 3
## 4 Oilseed rape grown in clay loam from Borders UK Crop Microbiome Cryobank
## 5 Oilseed rape grown in clay loam from Borders UK Crop Microbiome Cryobank
## 6 Oilseed rape grown in clay loam from Borders UK Crop Microbiome Cryobank
##                         V7                          V8
## 1
## 2      experimental factor   reference for biomaterial
## 3
## 4 Oilseed rape Rhizosphere Rothamsted Research Station
## 5 Oilseed rape Rhizosphere Rothamsted Research Station
## 6 Oilseed rape Rhizosphere Rothamsted Research Station
##                                          V9
## 1
## 2 sample volume or weight for DNA extraction
## 3                                         g
## 4                                       0.5
## 5                                       0.5
## 6                                       0.5
##                                         V10
```

```
## 1
## 2                   nucleic acid extraction
## 3
## 4 Qiagen(TM) Power soil gDNA extraction kit
## 5 Qiagen(TM) Power soil gDNA extraction kit
## 6 Qiagen(TM) Power soil gDNA extraction kit
##                                                                                                  V11
## 1
## 2                                                                          nucleic acid amplification
## 3
## 4 https://emea.support.illumina.com/downloads/16s_metagenomic_sequencing_library_preparation.html
## 5 https://emea.support.illumina.com/downloads/16s_metagenomic_sequencing_library_preparation.html
## 6 https://emea.support.illumina.com/downloads/16s_metagenomic_sequencing_library_preparation.html
##          V12                                                                    V13
## 1
## 2 target gene                                                          pcr primers
## 3
## 4         16S Forward: 5' CCTACGGGNGGCWGCAG; Reverse: 5' GACTACHVGGGTATCTAATCC
## 5         16S Forward: 5' CCTACGGGNGGCWGCAG; Reverse: 5' GACTACHVGGGTATCTAATCC
## 6         16S Forward: 5' CCTACGGGNGGCWGCAG; Reverse: 5' GACTACHVGGGTATCTAATCC
##                   V14
## 1
## 2 multiplex identifiers
## 3
## 4                   N/A
## 5                   N/A
## 6                   N/A
##
## 1
## 2
## 3
## 4 Forward overhang: 5' TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG; Reverse overhang: 5' GTCTCGTGGGCTCGGAGATGT
## 5 Forward overhang: 5' TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG; Reverse overhang: 5' GTCTCGTGGGCTCGGAGATGT
## 6 Forward overhang: 5' TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG; Reverse overhang: 5' GTCTCGTGGGCTCGGAGATGT
##
## 1
## 2
## 3
## 4 95 C for 3 minutes; 25 cycles of 95 C for 30 seconds, 55 C for 30 seconds, 72 C for 30 seconds; 72
## 5 95 C for 3 minutes; 25 cycles of 95 C for 30 seconds, 55 C for 30 seconds, 72 C for 30 seconds; 72
## 6 95 C for 3 minutes; 25 cycles of 95 C for 30 seconds, 55 C for 30 seconds, 72 C for 30 seconds; 72
##                V17                   V18                   V19
## 1
## 2 sequencing method sequence quality check chimera check software
## 3
## 4             MiSeq               software                   N/A
## 5             MiSeq               software                   N/A
## 6             MiSeq               software                   N/A
##                             V20                   V21
## 1
## 2 relevant electronic resources   negative control type
## 3
## 4          agmicrobiomebase.org no-template PCR control
## 5          agmicrobiomebase.org no-template PCR control
```

```
## 6          agmicrobiomebase.org no-template PCR control
##                                                    V22             V23
## 1
## 2                                 positive control type collection date
## 3
## 4 synthetic community of three known bacterial species      2022-02-20
## 5 synthetic community of three known bacterial species      2022-02-20
## 6 synthetic community of three known bacterial species      2022-02-20
##                                       V24                           V25
## 1
## 2 geographic location (country and/or sea) geographic location (latitude)
## 3                                                                     DD
## 4                            United Kingdom                       51.8094
## 5                            United Kingdom                       51.8094
## 6                            United Kingdom                       51.8094
##                               V26
## 1
## 2 geographic location (longitude)
## 3                                DD
## 4                            0.3561
## 5                            0.3561
## 6                            0.3561
##                                             V27                          V28
## 1
## 2              broad-scale environmental context  local environmental context
## 3
## 4 environmental system determined by an organism plant-associated environment
## 5 environmental system determined by an organism plant-associated environment
## 6 environmental system determined by an organism plant-associated environment
##                       V29           V30              V31      V32        V33
## 1
## 2    environmental medium plant product host common name host age host taxid
## 3                                                                    months
## 4 rhizosphere environment           N/A     Oilseed rape        3       3708
## 5 rhizosphere environment           N/A     Oilseed rape        3       3708
## 6 rhizosphere environment           N/A     Oilseed rape        3       3708
##               V34             V35                               V36
## 1
## 2 host life stage plant body site host subspecific genetic lineage
## 3
## 4       flowering            root                            Campus
## 5       flowering            root                            Campus
## 6       flowering            root                            Campus
##                      V37
## 1
## 2   climate environment
## 3
## 4 controlled glasshouse
## 5 controlled glasshouse
## 6 controlled glasshouse
```

Format the crop checklist to fit the fastq checklist requirement and the mapping/merging process.

```r
# Set the column names of 'data' to be the second row of 'data'
colnames(Check.list) <- Check.list[2,]

# Remove the first three rows of 'data'
Check.list <- Check.list[-c(1:3),]

# Create a new dataframe 'Check-list2' that only includes the 3rd and 4th columns of 'Check-list'
Check.list_2 <- Check.list[,c(3:4)]

# Print the data (first 6 columns) to check if it has been correctly
head(Check.list_2)
```

```
##   sample_alias sample_title
## 4      ms00315  OR-CL-BO-01
## 5      ms00316  OR-CL-BO-02
## 6      ms00317  OR-CL-BO-03
## 7      ms00318  OR-CL-BO-04
## 8      ms00319  OR-CL-BO-05
## 9      ms00320  OR-CL-YO-01
```

Merge 'Webin' and 'Check.list_2' on the columns "ALIAS" and "sample_alias", respectively, to create 'merged_df'

```r
merged_df <- merge(Webin, Check.list_2, by.x = "ALIAS", by.y = "sample_alias")

# Add a leading zero to all the numbers in the 'sample_title' column
merged_df$sample_title <- sub("-([0-9])([^0-9]|$)", "-0\\1\\2", merged_df$sample_title)

# Print the 'sample_title' column of 'merged_df' to check if it has been read correctly
print(merged_df$sample_title)
```

```
##  [1] "OR-CL-BO-01" "OR-CL-BO-02" "OR-CL-BO-03" "OR-CL-BO-04" "OR-CL-BO-05"
##  [6] "OR-CL-YO-01" "OR-CL-YO-02" "OR-CL-YO-03" "OR-CL-YO-04" "OR-CY-BU-01"
## [11] "OR-CY-BU-02" "OR-CY-BU-03" "OR-CY-BU-04" "OR-CY-BU-05" "OR-CY-YO-01"
## [16] "OR-CY-YO-02" "OR-CY-YO-03" "OR-CY-YO-04" "OR-CY-YO-05" "OR-SC-HE-01"
## [21] "OR-SC-HE-02" "OR-SC-HE-03" "OR-SC-HE-04" "OR-SC-HE-05" "OR-SC-SH-01"
## [26] "OR-SC-SH-02" "OR-SC-SH-03" "OR-SC-SH-04" "OR-SC-SH-05" "OR-SL-AN-01"
## [31] "OR-SL-AN-02" "OR-SL-AN-03" "OR-SL-AN-04" "OR-SL-AN-05" "OR-SL-BE-01"
## [36] "OR-SL-BE-02" "OR-SL-BE-03" "OR-SL-BE-04" "OR-SL-SH-03" "OR-SL-SH-04"
## [41] "OR-SL-SH-05"
```

Step 3

Load the data from the "md5.txt" file, which contains the MD5 checksums for each file in the directory. The MD5 checksum is a 32-digit hexadecimal number that represents the unique fingerprint of a file. It can be used to verify the integrity and authenticity of a file, especially after a file transfer.

```r
md5 <- read.table("md5.txt", sep = "")

# Print the data (first 6 columns) to check the document structure
head(md5)
```

```
##                                 V1                                V2
## 1 4077f422e00c7080d3ff8a9bc9e06cff  OR-CL-BO-1_S1_L001_R1_001.fastq.gz
## 2 2e6a47129313fdc0220d587b9d13fe40  OR-CL-BO-1_S1_L001_R2_001.fastq.gz
## 3 0c6e7f68e61d776099e9cf04ffc8e686 OR-CL-BO-2_S13_L001_R1_001.fastq.gz
## 4 051e19ade76bc57cfdfc16ae08546d33 OR-CL-BO-2_S13_L001_R2_001.fastq.gz
## 5 182745acd2ae4ff42c5deeec8c12d744 OR-CL-BO-3_S25_L001_R1_001.fastq.gz
## 6 f943fb06b03be795e9bfb328aec750d0 OR-CL-BO-3_S25_L001_R2_001.fastq.gz
```

Change the original md5 format to fit for fastq checklist

```
# Keep only the first and second columns of 'md5'
md5 <- md5[, c("V2", "V1")]

# Create a new dataframe 'md5.R1' that only includes the rows of 'md5'
# where the second column contains "_R1_001.fastq.gz"
md5.R1 <- md5[grep("_R1_001.fastq.gz", md5$V2), ]

# Add a new column to 'md5.R1', which is created by splitting
# the second column on underscores and taking the first element
md5.R1$v3 <- sapply(strsplit(as.character(md5.R1$V2), "_"), `[`, 1)

# Replace single digits at the end of the strings in the new
# column with the same digit preceded by a zero
md5.R1$v3 <- sub("(\\d)$", "0\\1", md5.R1$v3)

# Replace single digits between dashes in the new column
# with the same digit preceded by a zero
md5.R1$v3 <- sub("-([1-9])-", "-0\\1-", md5.R1$v3)

# Set the column names of 'md5.R1'
colnames(md5.R1) <- c("forward_file_name","forward_file_md5", "sample_title")

# Create a new dataframe 'md5.R2' that only includes the rows of
# 'md5' where the second column contains "_R2_001.fastq.gz"
md5.R2 <- md5[grep("_R2_001.fastq.gz", md5$V2), ]

# Add a new column to 'md5.R2', which is created by splitting
# the second column on underscores and taking the first element
md5.R2$v3 <- sapply(strsplit(as.character(md5.R2$V2), "_"), `[`, 1)

# Replace single digits at the end of the strings in
# the new column with the same digit preceded by a zero
md5.R2$v3 <- sub("(\\d)$", "0\\1", md5.R2$v3)

# Replace single digits between dashes in
# the new column with the same digit preceded by a zero
md5.R2$v3 <- sub("-([1-9])-", "-0\\1-", md5.R2$v3)

# Set the column names of 'md5.R2'
colnames(md5.R2) <- c("reverse_file_name", "reverse_file_md5", "sample_title")

# Merge 'md5.R1' and 'md5.R2' on the "sample_title" column to create 'merged.md5'
merged.md5 <- merge(md5.R1, md5.R2, by="sample_title")
```

```r
# Print the data (first 6 columns) to check the dataframe structure
head(merged.md5)
```

```
##   sample_title                  forward_file_name
## 1  OR-CL-BO-01  OR-CL-BO-1_S1_L001_R1_001.fastq.gz
## 2  OR-CL-BO-02 OR-CL-BO-2_S13_L001_R1_001.fastq.gz
## 3  OR-CL-BO-03 OR-CL-BO-3_S25_L001_R1_001.fastq.gz
## 4  OR-CL-BO-04 OR-CL-BO-4_S37_L001_R1_001.fastq.gz
## 5  OR-CL-BO-05 OR-CL-BO-5_S49_L001_R1_001.fastq.gz
## 6  OR-CL-YO-01 OR-CL-YO-1_S26_L001_R1_001.fastq.gz
##                   forward_file_md5                  reverse_file_name
## 1 4077f422e00c7080d3ff8a9bc9e06cff  OR-CL-BO-1_S1_L001_R2_001.fastq.gz
## 2 0c6e7f68e61d776099e9cf04ffc8e686 OR-CL-BO-2_S13_L001_R2_001.fastq.gz
## 3 182745acd2ae4ff42c5deeec8c12d744 OR-CL-BO-3_S25_L001_R2_001.fastq.gz
## 4 099c6af2b8ae1a2fe89d211338149a2d OR-CL-BO-4_S37_L001_R2_001.fastq.gz
## 5 aacbd1d9a6fa4b48bb168eb8c50580a8 OR-CL-BO-5_S49_L001_R2_001.fastq.gz
## 6 c9f76cd2041bbdc7e4922bf2e46dfd65 OR-CL-YO-1_S26_L001_R2_001.fastq.gz
##                   reverse_file_md5
## 1 2e6a47129313fdc0220d587b9d13fe40
## 2 051e19ade76bc57cfdfc16ae08546d33
## 3 f943fb06b03be795e9bfb328aec750d0
## 4 a73950f630c6d69b21444896ceb1fbfe
## 5 07399c859ead29ce2ecf034a2a4cd25e
## 6 9fc814cc0f1a841a4b2dc9e79a9e2dcd
```

Step 4

Merge “merged_df” and “merged.md5” on the “sample_title” column to create “merged_all”

```r
merged_all <- merge(merged_df, merged.md5, by ="sample_title")

# Add several new columns to 'merged_all' with constant values
merged_all$study <- rep("PRJEB58189", nrow(merged_all))
merged_all$instrument_model <- rep("Illumina MiSeq", nrow(merged_all))
merged_all$library_name <- rep("Nextera XT v2", nrow(merged_all))
merged_all$library_source <- rep("METAGENOMIC", nrow(merged_all))
merged_all$library_selection <- rep("PCR", nrow(merged_all))
merged_all$library_strategy <- rep("AMPLICON", nrow(merged_all))
merged_all$library_layout <- rep("PAIRED", nrow(merged_all))

# Rename the "sample" column to "ACCESSION"
colnames(merged_all)[which(colnames(merged_all) == "sample")] <- "ACCESSION"

# Rename the "ACCESSION" column back to "sample"
colnames(merged_all)[colnames(merged_all) == 'ACCESSION'] <- 'sample'

# Reorder the columns of 'merged_all'
merged_all <- merged_all[, c("sample", "study", "instrument_model", "library_name",
                             "library_source", "library_selection", "library_strategy",
                             "library_layout", "forward_file_name", "forward_file_md5",
                             "reverse_file_name", "reverse_file_md5")]
```

```r
# Print the data (first 6 columns) to check the dataframe structure
head(merged_all)
```

```
##        sample      study instrument_model  library_name library_source
## 1 ERS27620570 PRJEB58189   Illumina MiSeq Nextera XT v2    METAGENOMIC
## 2 ERS27620571 PRJEB58189   Illumina MiSeq Nextera XT v2    METAGENOMIC
## 3 ERS27620572 PRJEB58189   Illumina MiSeq Nextera XT v2    METAGENOMIC
## 4 ERS27620573 PRJEB58189   Illumina MiSeq Nextera XT v2    METAGENOMIC
## 5 ERS27620574 PRJEB58189   Illumina MiSeq Nextera XT v2    METAGENOMIC
## 6 ERS27620575 PRJEB58189   Illumina MiSeq Nextera XT v2    METAGENOMIC
##   library_selection library_strategy library_layout
## 1               PCR         AMPLICON         PAIRED
## 2               PCR         AMPLICON         PAIRED
## 3               PCR         AMPLICON         PAIRED
## 4               PCR         AMPLICON         PAIRED
## 5               PCR         AMPLICON         PAIRED
## 6               PCR         AMPLICON         PAIRED
##                     forward_file_name                 forward_file_md5
## 1  OR-CL-BO-1_S1_L001_R1_001.fastq.gz 4077f422e00c7080d3ff8a9bc9e06cff
## 2 OR-CL-BO-2_S13_L001_R1_001.fastq.gz 0c6e7f68e61d776099e9cf04ffc8e686
## 3 OR-CL-BO-3_S25_L001_R1_001.fastq.gz 182745acd2ae4ff42c5deeec8c12d744
## 4 OR-CL-BO-4_S37_L001_R1_001.fastq.gz 099c6af2b8ae1a2fe89d211338149a2d
## 5 OR-CL-BO-5_S49_L001_R1_001.fastq.gz aacbd1d9a6fa4b48bb168eb8c50580a8
## 6 OR-CL-YO-1_S26_L001_R1_001.fastq.gz c9f76cd2041bbdc7e4922bf2e46dfd65
##                     reverse_file_name                 reverse_file_md5
## 1  OR-CL-BO-1_S1_L001_R2_001.fastq.gz 2e6a47129313fdc0220d587b9d13fe40
## 2 OR-CL-BO-2_S13_L001_R2_001.fastq.gz 051e19ade76bc57cfdfc16ae08546d33
## 3 OR-CL-BO-3_S25_L001_R2_001.fastq.gz f943fb06b03be795e9bfb328aec750d0
## 4 OR-CL-BO-4_S37_L001_R2_001.fastq.gz a73950f630c6d69b21444896ceb1fbfe
## 5 OR-CL-BO-5_S49_L001_R2_001.fastq.gz 07399c859ead29ce2ecf034a2a4cd25e
## 6 OR-CL-YO-1_S26_L001_R2_001.fastq.gz 9fc814cc0f1a841a4b2dc9e79a9e2dcd
```

Step 5

Final modification for the fastq checklist

```r
# Create a new row with the same number of columns as 'merged_all'
new_row <- setNames(data.frame(matrix(ncol = ncol(merged_all), nrow = 1)), colnames(merged_all))

# Assign the values to the new row
new_row[1, c("sample", "study",
             "instrument_model")] <- c("FileType",
                                       "fastq", "Read submission file type")

# Save the column names
col_names <- colnames(new_row)

# Add the column names as a new row in the second position
new_row <- rbind(new_row, col_names)

# Add the new row to the top of the dataframe
merged_all <- rbind(new_row, merged_all)
```

```
# Remove column names
colnames(merged_all) <- NULL

# Print the data (first 6 columns) to check the dataframe structure
head(merged_all)
```

```
##
## 1    FileType      fastq Read submission file type          <NA>           <NA>
## 2      sample      study          instrument_model  library_name library_source
## 3 ERS27620570 PRJEB58189            Illumina MiSeq Nextera XT v2    METAGENOMIC
## 4 ERS27620571 PRJEB58189            Illumina MiSeq Nextera XT v2    METAGENOMIC
## 5 ERS27620572 PRJEB58189            Illumina MiSeq Nextera XT v2    METAGENOMIC
## 6 ERS27620573 PRJEB58189            Illumina MiSeq Nextera XT v2    METAGENOMIC
##
## 1              <NA>             <NA>           <NA>
## 2 library_selection library_strategy library_layout
## 3               PCR         AMPLICON         PAIRED
## 4               PCR         AMPLICON         PAIRED
## 5               PCR         AMPLICON         PAIRED
## 6               PCR         AMPLICON         PAIRED
##
## 1                                 <NA>                                <NA>
## 2                    forward_file_name                   forward_file_md5
## 3  OR-CL-BO-1_S1_L001_R1_001.fastq.gz 4077f422e00c7080d3ff8a9bc9e06cff
## 4 OR-CL-BO-2_S13_L001_R1_001.fastq.gz 0c6e7f68e61d776099e9cf04ffc8e686
## 5 OR-CL-BO-3_S25_L001_R1_001.fastq.gz 182745acd2ae4ff42c5deeec8c12d744
## 6 OR-CL-BO-4_S37_L001_R1_001.fastq.gz 099c6af2b8ae1a2fe89d211338149a2d
##
## 1                                 <NA>                                <NA>
## 2                    reverse_file_name                   reverse_file_md5
## 3  OR-CL-BO-1_S1_L001_R2_001.fastq.gz 2e6a47129313fdc0220d587b9d13fe40
## 4 OR-CL-BO-2_S13_L001_R2_001.fastq.gz 051e19ade76bc57cfdfc16ae08546d33
## 5 OR-CL-BO-3_S25_L001_R2_001.fastq.gz f943fb06b03be795e9bfb328aec750d0
## 6 OR-CL-BO-4_S37_L001_R2_001.fastq.gz a73950f630c6d69b21444896ceb1fbfe
```

Step 6

Export the table and ready to upload

```
# write.table(merged_all, file = "fastq2_template_16Samplicons_OR_TEST_with_mapping.tsv",
#             sep = "\t", row.names = FALSE, quote = FALSE, na = "")
```

```
sessionInfo()
```

```
## R version 4.3.2 (2023-10-31 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 11 x64 (build 22631)
##
## Matrix products: default
##
##
```

```
## locale:
## [1] LC_COLLATE=English_United Kingdom.utf8
## [2] LC_CTYPE=English_United Kingdom.utf8
## [3] LC_MONETARY=English_United Kingdom.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United Kingdom.utf8
##
## time zone: Europe/London
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.3.2    fastmap_1.1.1     cli_3.6.1         tools_4.3.2
##  [5] htmltools_0.5.7   rstudioapi_0.15.0 yaml_2.3.7        rmarkdown_2.25
##  [9] knitr_1.45        xfun_0.40         digest_0.6.33     rlang_1.1.1
## [13] evaluate_0.23
```