AMR_WGS_001 Assembly overview

2023-11-14

Load packages

```
library(tidyverse)
library(readxl)
library(hrbrthemes)
library(viridis)
library(writexl)
```

The purpose of this Rmarkdown is to analyse the first isolate sequencing batch

Load metadata

```
metadata<-read_csv("/home/projects/cu_00014/data/sepseq_WGS/metadata/metadata_AMR_WGS_001.csv")
```

Load checkm2 data in

```
read_checkm<-function(file){
    sample_id = str_split(file, pattern = "/")[[1]][length(str_split(file, pattern = "/")[[1]])-1]
    checkm<-read_tsv(file = file, col_names = T, comment = "#", show_col_types = FALSE) %>%
      mutate(sample_id=sample_id)

    return(checkm)
}

combine_checkm<-function(file_list){
  combined_checkm<-data.frame()
  for (i in file_list) {
    if (file.info(i)$size > 0) {
      checkm_batch<-read_checkm(i)
      combined_checkm<-combined_checkm %>%
        bind_rows(
          checkm_batch
        )
    }
  }
  return(combined_checkm)

}

checkm2_paths<-paste0("data/qc/20_checkm/", list.files("data/qc/20_checkm/", pattern = "quality_report.
```

```
df_checkm2<-combine_checkm(checkm2_paths) %>%
  select(-Name, -Completeness_Model_Used, -Translation_Table_Used)
head(df_checkm2)
```

```
##   Completeness Contamination Coding_Density Contig_N50 Average_Gene_Length
## 1          100          1.13          0.873    3149626            314.4639
## 2          100          0.29          0.869    3635892            300.2399
## 3          100          0.65          0.873    4279368            309.6323
## 4          100          0.49          0.875    5045369            311.5563
## 5          100          0.76          0.875    4817838            309.2612
## 6          100          0.16          0.873    4010379            307.2231
##   Genome_Size GC_Content Total_Coding_Sequences Additional_Notes sample_id
## 1     5290295       0.51                  4902             None    822_A8
## 2     5711871       0.51                  5524             None    822_B8
## 3     5217408       0.51                  4914             None    822_B9
## 4     5228452       0.51                  4902             None    822_C8
## 5     5192987       0.51                  4908             None    822_H9
## 6     5227258       0.51                  4963             None    822_I6
```

Load gtdb-tk data in

```
read_gtdbtk<-function(file){
    sample_id = str_split(file, pattern = "/")[[1]][length(str_split(file, pattern = "/")[[1]])-1]
    gtdbtk<-read_tsv(file = file, col_names = T, comment = "#", show_col_types = FALSE) %>%
      mutate(sample_id=sample_id,
             closest_placement_radius=as.character(closest_placement_radius),
             closest_placement_ani=as.character(closest_placement_ani),
             closest_placement_af=as.character(closest_placement_af))

    return(gtdbtk)
}

combine_gtdbtk<-function(file_list){
  combined_gtdbtk<-data.frame()
  for (i in file_list) {
    if (file.info(i)$size > 0) {
      gtdbtk_batch<-read_gtdbtk(i)
      combined_gtdbtk<-combined_gtdbtk %>%
        bind_rows(
          gtdbtk_batch
        )
    }
  }
  return(combined_gtdbtk)

}

gtdbtk_paths<-paste0("data/qc/30_gtdbtk/", list.files("data/qc/30_gtdbtk/", pattern = "gtdbtk.bac120.su
gtdbtk_paths<-gtdbtk_paths[!grepl("classify", gtdbtk_paths)]

df_gtdbtk<-combine_gtdbtk(gtdbtk_paths) %>%
  select(-user_genome)
```

Combine data

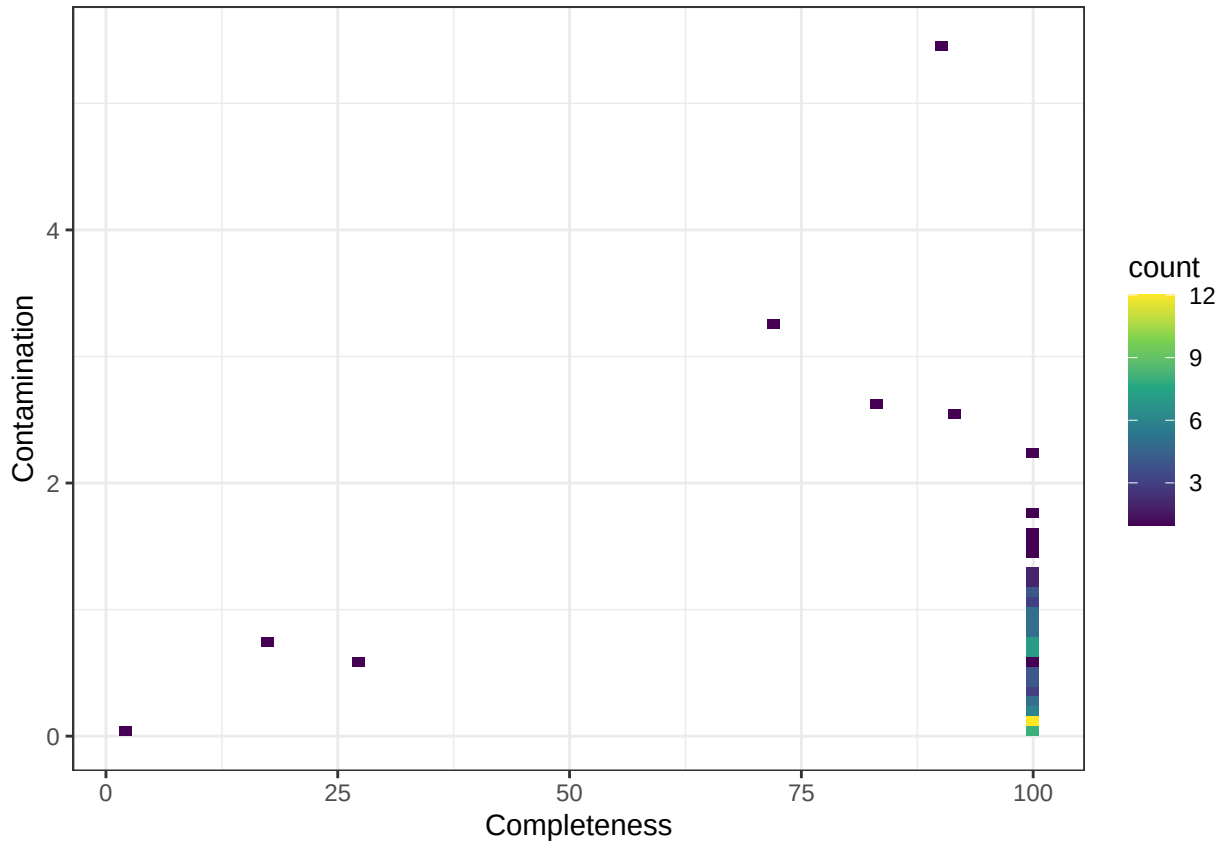
```
df_merged<-metadata %>%
  left_join(df_checkm2, by = "sample_id") %>%
  left_join(df_gtdbtk, by = "sample_id") %>%
  mutate(type = if_else(grepl("NC", sample_id), "NC",
           if_else(grepl("PC", sample_id), "PC",
             if_else(grepl("Pantoea", sample_id), "Pantoea", "Sample"))
         ),
         label = if_else(Completeness < 95, sample_id, ""))

head(df_merged)
```

```
## # A tibble: 6 x 59
##   Indate       Lar  Labnr   Box Pit              sample_id final_pos scrape_date
##   <date>     <dbl>  <dbl> <dbl> <chr>            <chr>     <chr>     <chr>
## 1 2022-02-19  2022 101672   827 "0\u001e^827\u0~ 827_A5    A1        30-11-2023
## 2 2022-02-17  2022  97592   827 "0\u001e^827\u0~ 827_I2    B1        30-11-2023
## 3 2022-01-28  2022  57596   824 "0\u001e^824\u0~ 824_B8    C1        30-11-2023
## 4 2022-01-24  2022  46874   824 "0\u001e^824\u0~ 824_B5    D1        30-11-2023
## 5 2022-01-23  2022  46096   824 "0\u001e^824\u0~ 824_H4    E1        30-11-2023
## 6 2022-01-17  2022  31727   823 "0\u001e^823\u0~ 823_I8    F1        30-11-2023
## # i 51 more variables: scrape_person <chr>, scrape_comments <lgl>,
## #   ext_date <chr>, ext_person <chr>, ext_kit <chr>, ext_lot <lgl>,
## #   ext_plate_id <chr>, ext_pos <chr>, ext_conc <dbl>, ext_conc_2 <lgl>,
## #   ext_upconc <lgl>, lib_date <date>, lib_person <chr>, lib_kit <chr>,
## #   lib_lot <lgl>, lib_plate_id <chr>, lib_pos <chr>, lib_input_sample <dbl>,
## #   lib_input_nf <dbl>, lib_barcode <chr>, lib_flowcell_id <chr>,
## #   Completeness <dbl>, Contamination <dbl>, Coding_Density <dbl>, ...
```

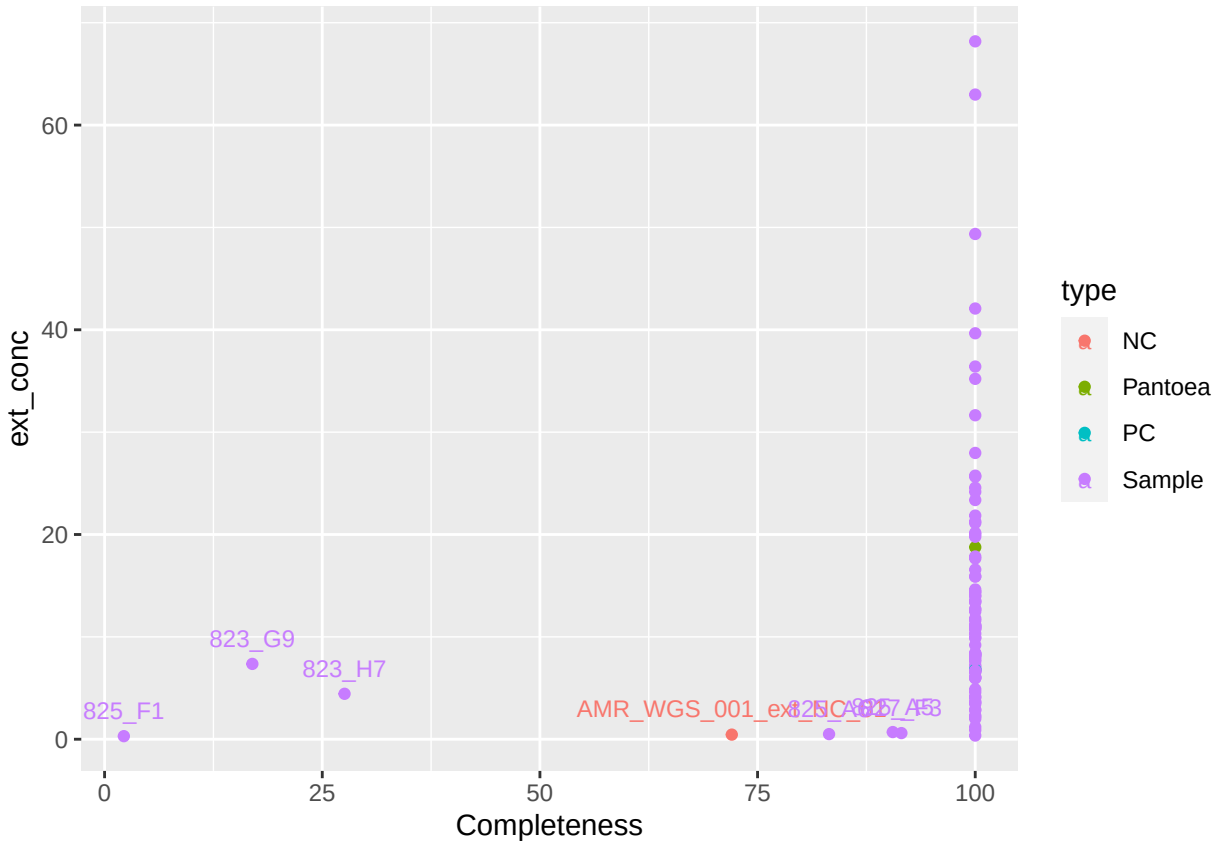
Plot the data

```
df_merged %>%
  ggplot(aes(x=Completeness, y=Contamination)) +
  geom_bin2d(bins = 70) +
  scale_fill_continuous(type = "viridis") +
  theme_bw()
```



Ext_conc against completeness

```
df_merged %>%
  ggplot(aes(x=Completeness, y=ext_conc, color = type, label = label)) +
  geom_point() +
  geom_text(vjust=-1, size = 3)
```


ext_conc
60
40
20
0
0
25
50
75
100
Completeness
823_G9
823_H7
825_F1
type
NC
Pantoea
PC
Sample