Bin Quality Report

Tip

Table of Contents

Below is the list of sections covered in this document.
Click on any section to quickly navigate to it.

Library and Data Import

```
# install libraries if they are not availible
# required libraries
libraries <- c("ggplot2", "hrbrthemes", "dplyr", "tidyr", "viridis",
               "readr", "magick", "scales", "randomNames", "ggrepel",
               "stringr", "gridExtra", "gt")

# Install missing packages
missing_packages <- libraries[!(libraries %in% installed.packages()[,"Package"])]
if(length(missing_packages)) install.packages(missing_packages, dependencies = TRUE)

# Load required libraries
library(ggplot2)      # Data visualization
library(hrbrthemes)   # Themes for ggplot2
library(dplyr)        # Data manipulation
library(tidyr)        # Data tidying
library(viridis)      # Color palettes
library(readr)        # Data import
library(magick)       # Image processing
library(scales)       # Scaling functions
library(randomNames)  # Generate random names
library(ggrepel)      # Repel overlapping text labels
library(stringr)      # String manipulation
library(gridExtra)    # Arrange multiple plots
library(gt)           # Create tables
```

import two csv containing bins quality statistics

```
path_50_10_bins_stats_92 <- "/project/asteen_1130/deep_vs_surface/manual_results/07_bin_r
data_50_10_bins_stats_92 <- read_tsv(path_50_10_bins_stats_92) |> mutate(location = "deep

path_50_10_bins_stats_93 <- "/project/asteen_1130/deep_vs_surface/manual_results/07_bin_r
data_50_10_bins_stats_93 <- read_tsv(path_50_10_bins_stats_93) |> mutate(location = "surf

combined_data <- bind_rows(data_50_10_bins_stats_92, data_50_10_bins_stats_93)
```

Overview

Note

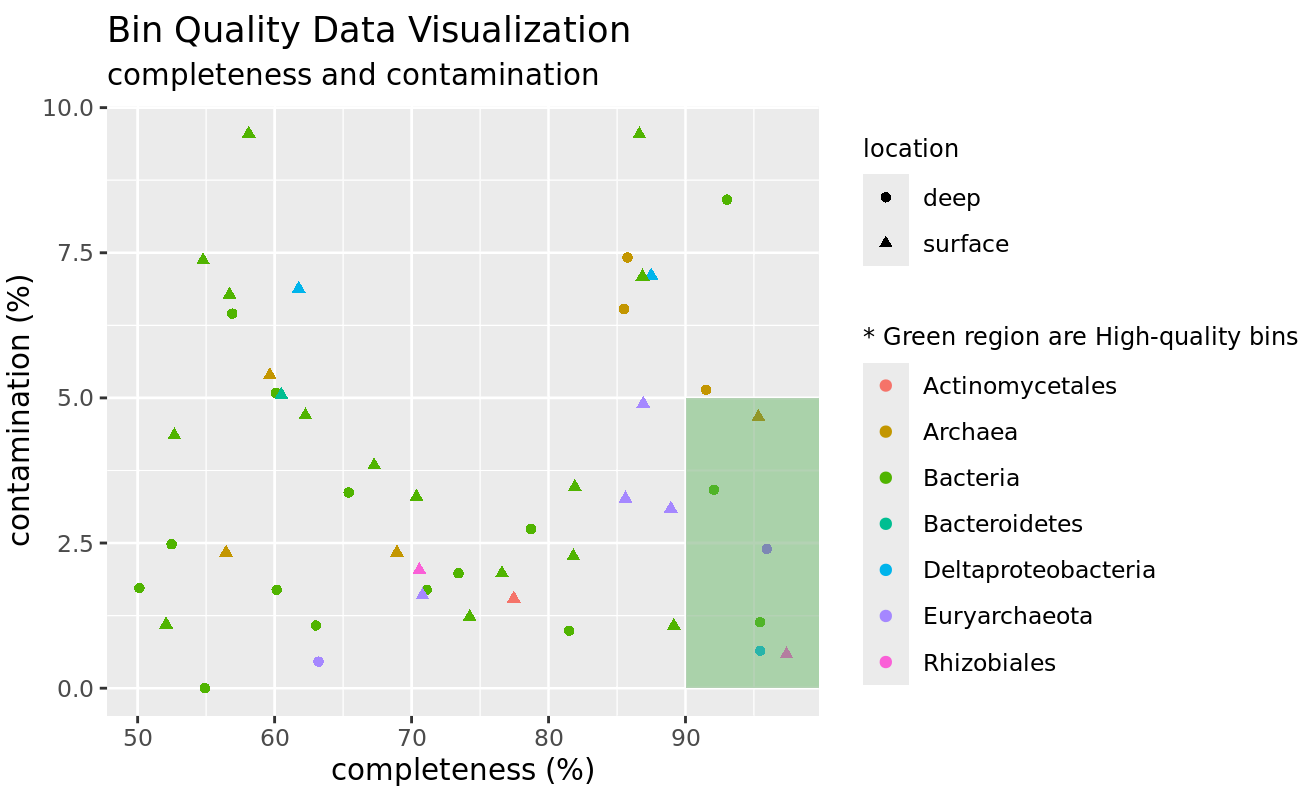
We identified 29 metagenome-assembled genome (MAG) bins in the surface sample (SRR7066493) and 21 bins in the deep sample (SRR7066492).

Bins with at least 50% completeness and no more than 10% contamination are classified as medium-quality MAGs. In the visualization, completeness and contamination levels of each bin are represented, with the green region highlighting high-quality bins (completeness > 90% and contamination < 5%). All other bins fall into the medium-quality category.

The bin origins are also depicted: circles represent bins from the deep sample, while triangles correspond to bins from the surface sample. Additionally, bin lineage information is conveyed through color coding, allowing for taxonomic differentiation.

Reference

```
scatter_plot <- ggplot(combined_data, aes(x = completeness, y = contamination)) +
  geom_point(aes(color = lineage, shape = location),size = 1.5) +
  # Add axes labels, title, and subtitle
  labs(
    title = "Bin Quality Data Visualization",
    subtitle = "completeness and contamination",
    x = "completeness (%)",
    y = "contamination (%)") +
  geom_rect(aes(xmin = 90, xmax = Inf, ymin = 0, ymax = 5), fill = "light green", alpha = 0.01) +
  labs(color = "* Green region are High-quality bins ") +
  theme(legend.title = element_text(size = 9))


combined_data$bin_quality <- "Medium-quality bins"
combined_data$bin_quality[combined_data$completeness>90 & combined_data$contamination<5] <- "High
freq_table <- table(combined_data$location, combined_data$bin_quality)
table_grob <- tableGrob(freq_table)


# show both
grid.arrange(scatter_plot, table_grob, nrow=2, heights=c(5, 1))
```

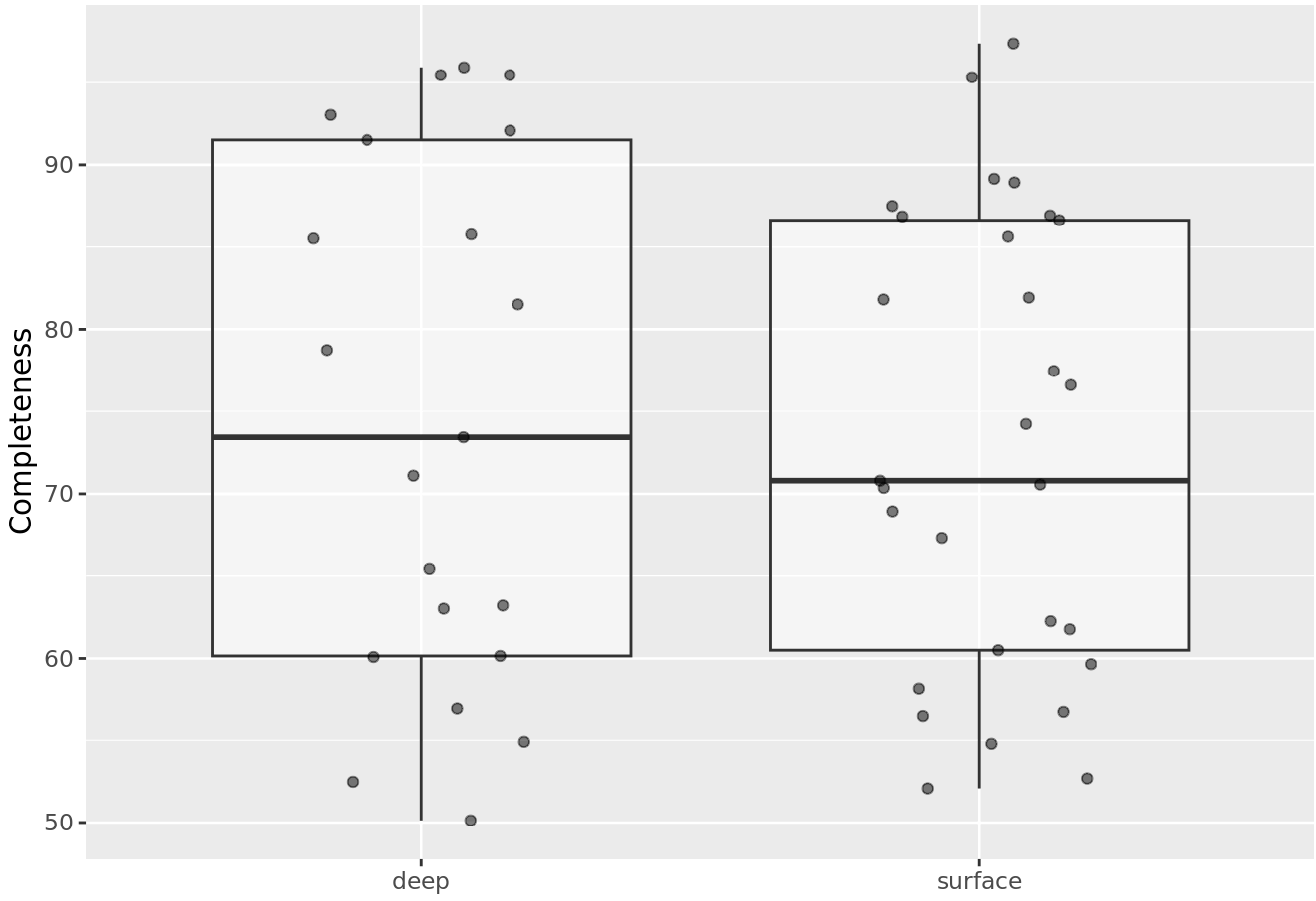


	High-quality bins	Medium-quality bins
deep	4	17
surface	2	27

completeness

Completeness in metagenome assembly refers to the extent to which the assembled contigs or scaffolds represent the total genomic content of the sampled microbial community. here we first try to compare distribution of completeness in both samples.

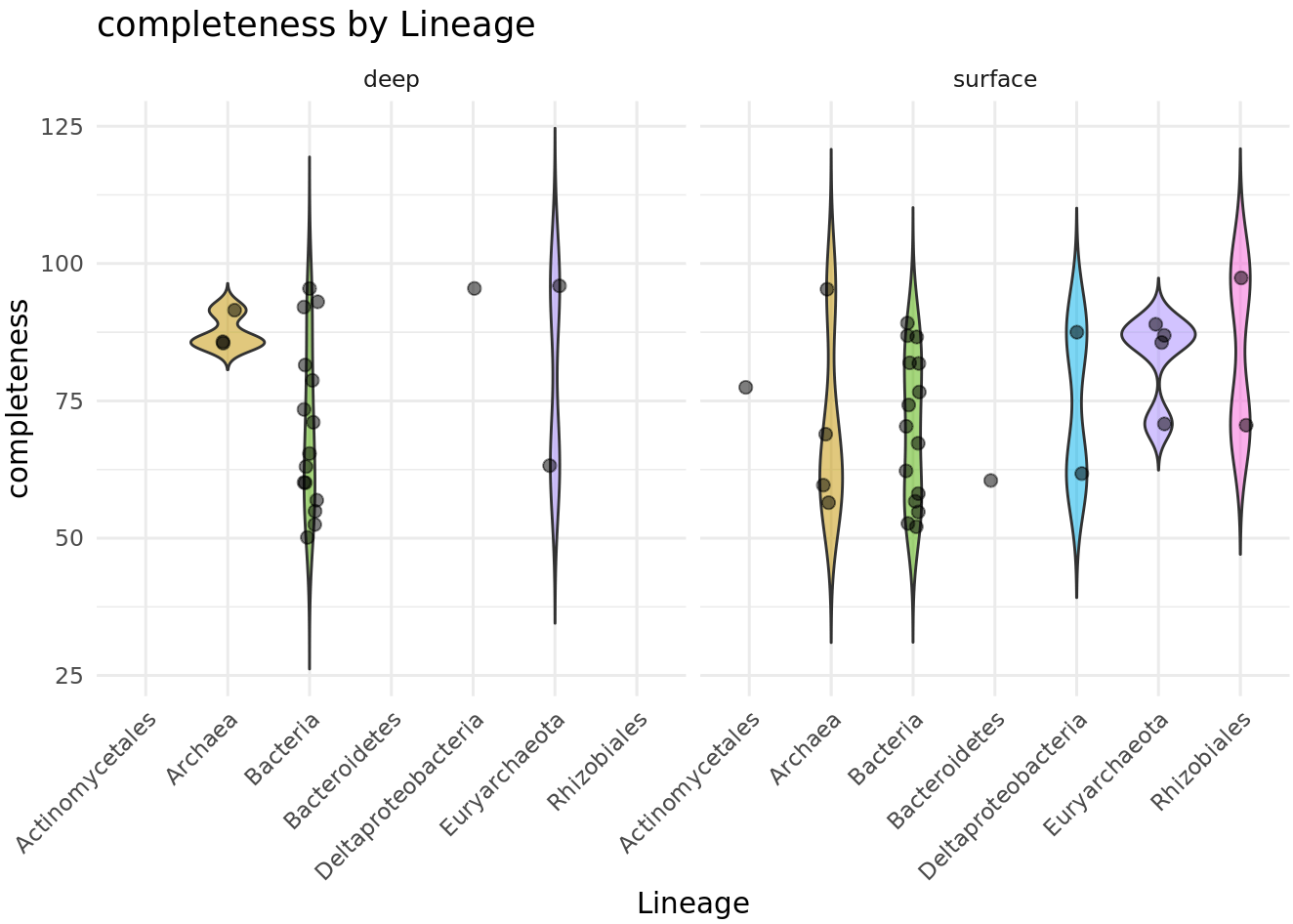
```
completeness_plot <- ggplot(data = combined_data, aes(x = location, y = completeness)) +
    geom_boxplot(alpha = 0.6) +  # Adjust transparency if needed
    geom_jitter(width = 0.2, alpha = 0.5) +
    labs(x = "", y = "Completeness")  # Removing x-axis label for clarity

completeness_plot
```



bins completeness scores based on lineage

Here, we present the distribution of completeness scores across different lineages, highlighting variations between surface and deep samples. Notably, the deep samples exhibit a substantial number of archaeal lineages with high completeness, indicating their strong representation in these environments. In contrast, within the surface samples, the Euryarchaeota lineage demonstrates particularly high completeness, suggesting its dominance or prevalence in these conditions.

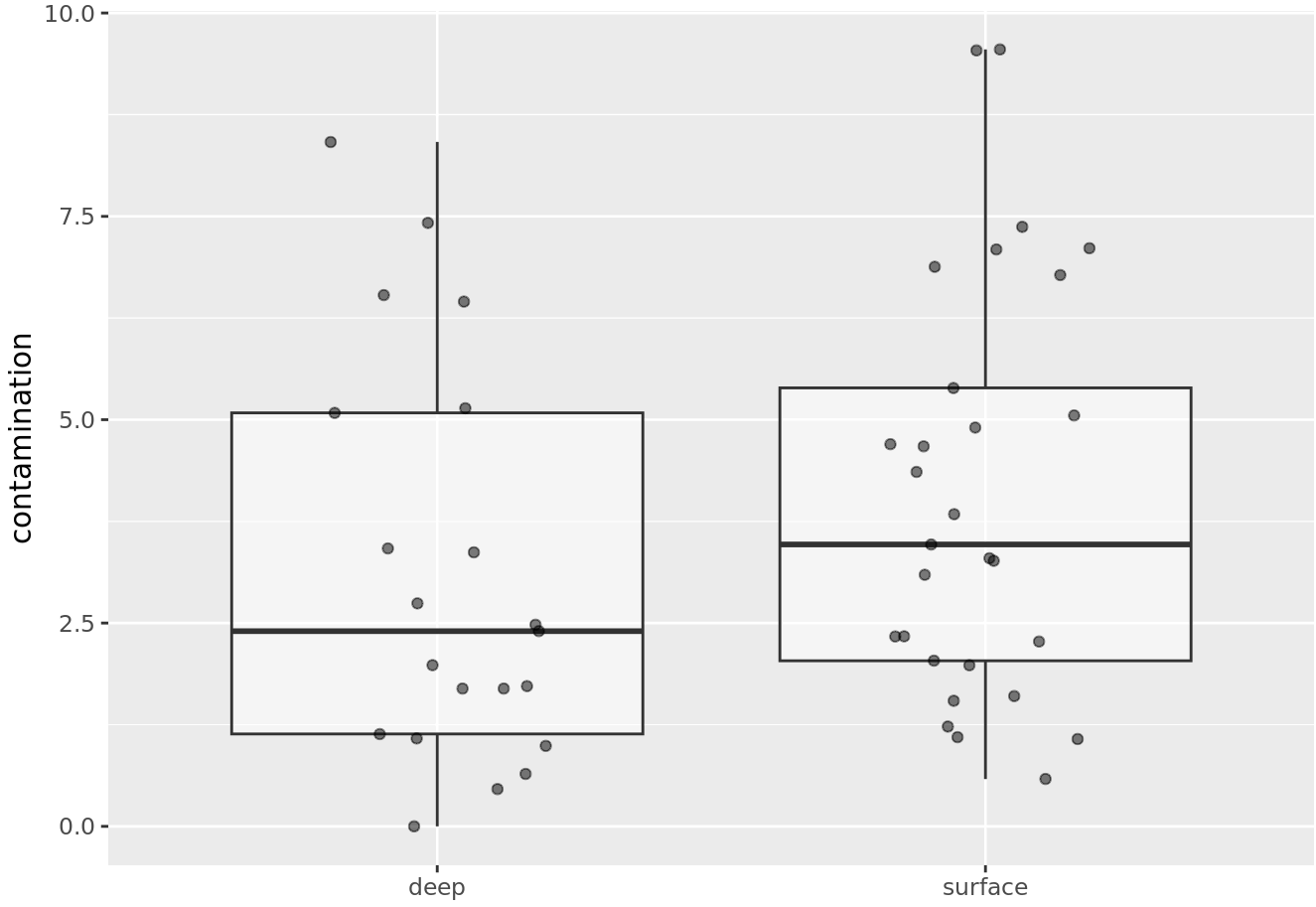
```
completeness_lineage_plot <- ggplot(data=combined_data, aes(x=lineage, y=completeness, fi
    geom_violin(alpha=0.5, trim=FALSE) +
    geom_jitter(width=0.1, alpha=0.5, size=2) +  # Adds individual data points
    facet_wrap(~location) +
    theme_minimal() +
    theme(
        legend.position = "none",
        axis.text.x = element_text(angle = 45, hjust = 1)  # Rotates x-axis text
    ) +
    labs(y = "completeness", x = "Lineage", title = "completeness by Lineage")
completeness_lineage_plot
```



contamination

the unwanted sequences in the bins that do not originate from the target microbial community.Let's have a overall look at contamination in two samples.

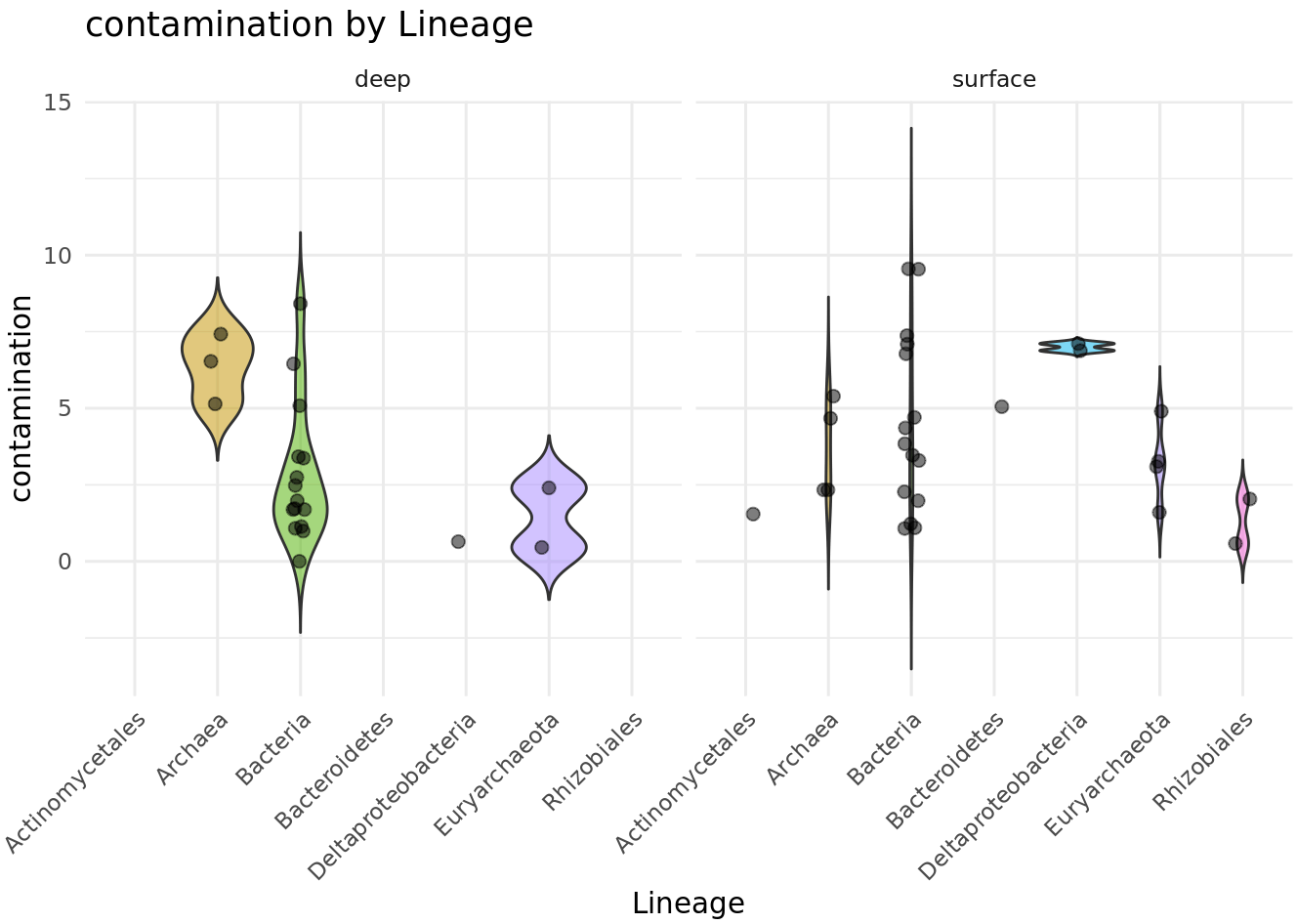
```
contamination_plot <- ggplot(data = combined_data, aes(x = location, y = contamination))
    geom_boxplot(alpha = 0.6) +  # Adjust transparency if needed
    geom_jitter(width = 0.2, alpha = 0.5) +
    labs(x = "", y = "contamination")  # Removing x-axis label for clarity

contamination_plot
```



bins contamination scores based on lineage

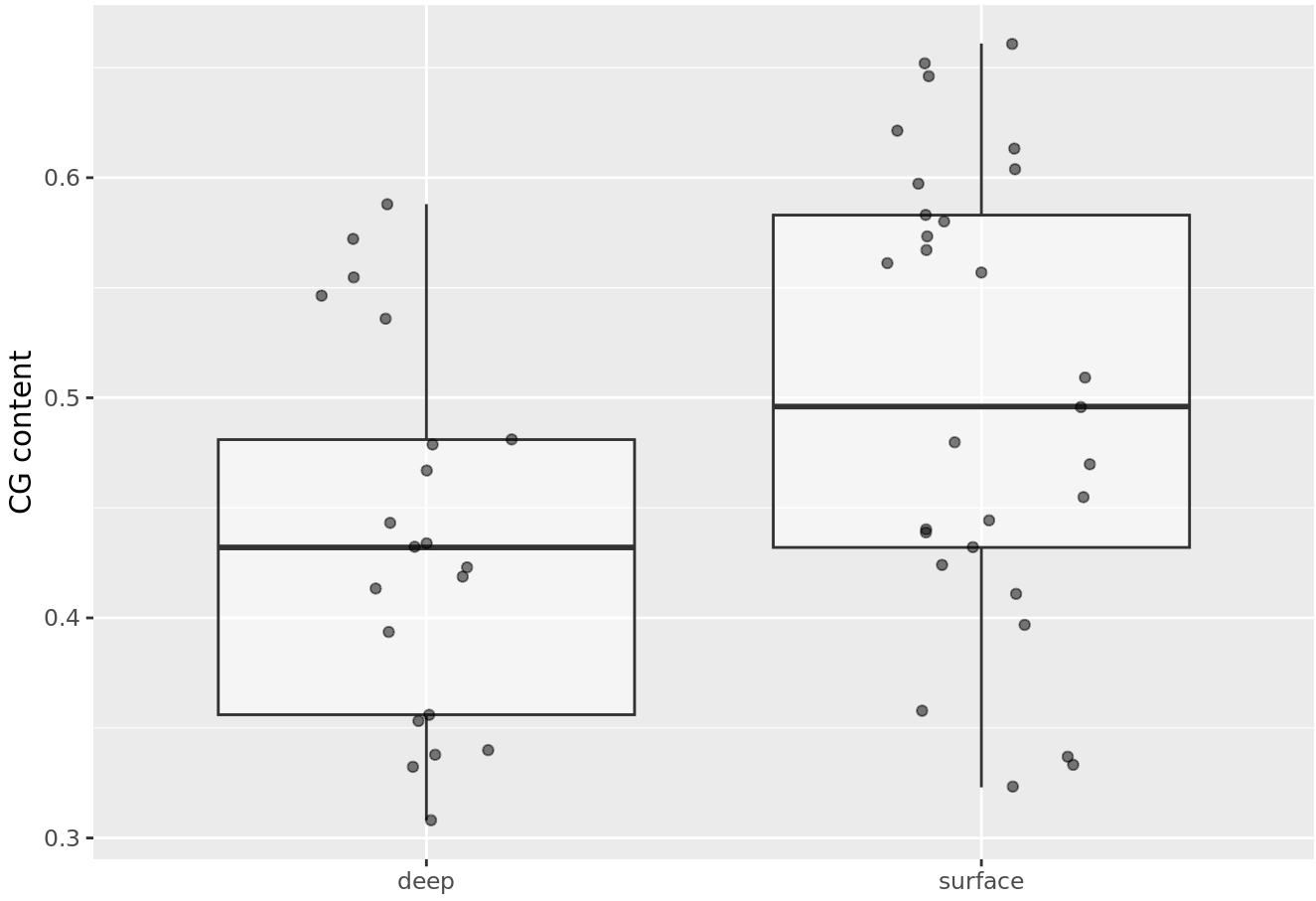
Here, we present the distribution of contamination scores across different lineages The first thing that caught my attention was the level of contamination in Deltaproteobacteria. My initial thought is that the sequence may have high genomic diversity and complexity, or it might share many genomic features with other taxa. Alternatively, it could be due to low abundance and assembly artifacts.

```
contamination_lineage_plot <- ggplot(data=combined_data, aes(x=lineage, y=contamination,
    geom_violin(alpha=0.5, trim=FALSE) +
    geom_jitter(width=0.1, alpha=0.5, size=2) +  # Adds individual data points
    facet_wrap(~location) +
    theme_minimal() +
    theme(
        legend.position = "none",
        axis.text.x = element_text(angle = 45, hjust = 1)  # Rotates x-axis text
    ) +
    labs(y = "contamination", x = "Lineage", title = "contamination by Lineage")
contamination_lineage_plot
```

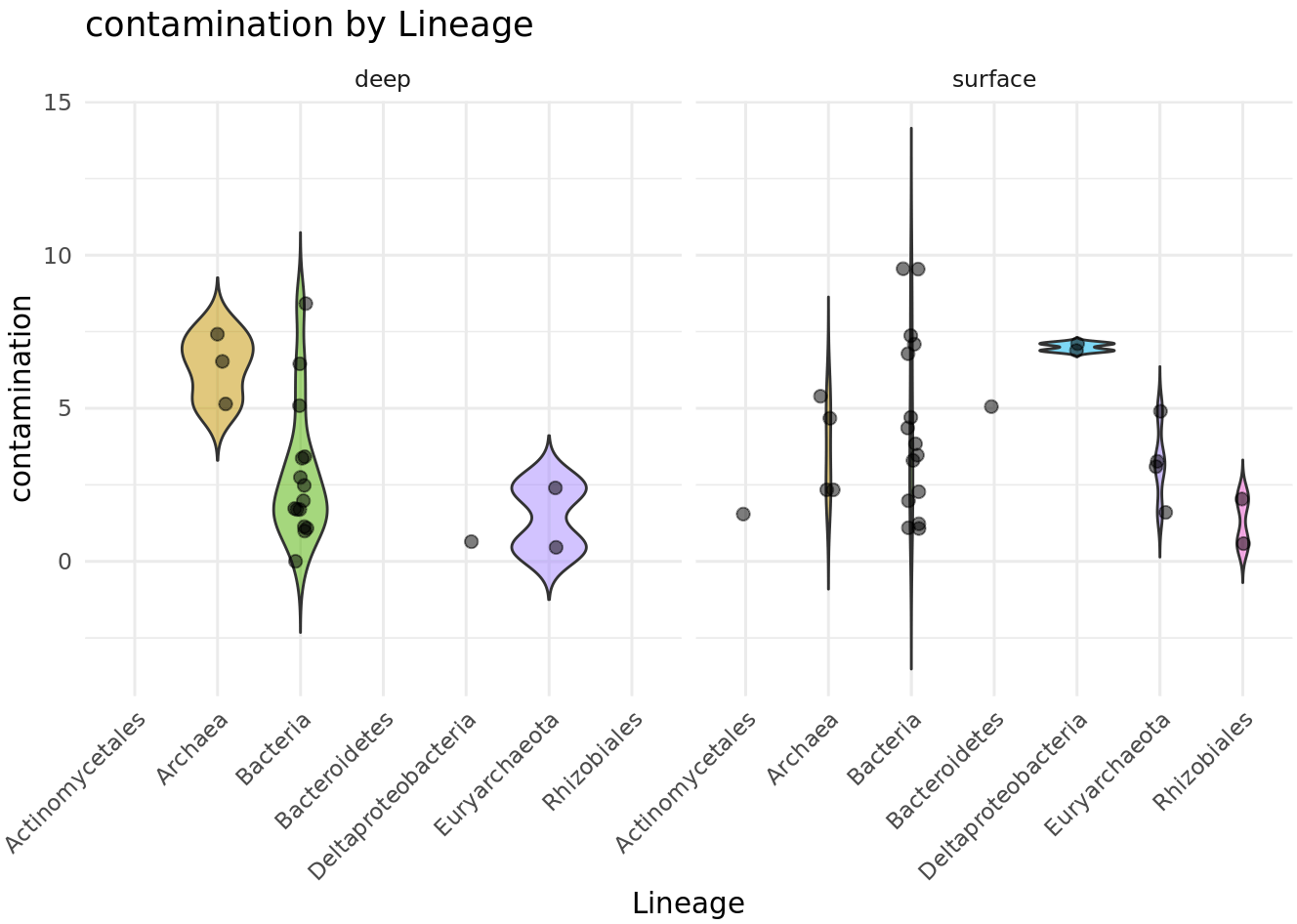


bins CG content based on lineage

```
CG_content_plot <- ggplot(data = combined_data, aes(x = location, y = GC)) +
    geom_boxplot(alpha = 0.6) +  # Adjust transparency if needed
    geom_jitter(width = 0.2, alpha = 0.5) +
    labs(x = "", y = "CG content")  # Removing x-axis label for clarity

CG_content_plot
```



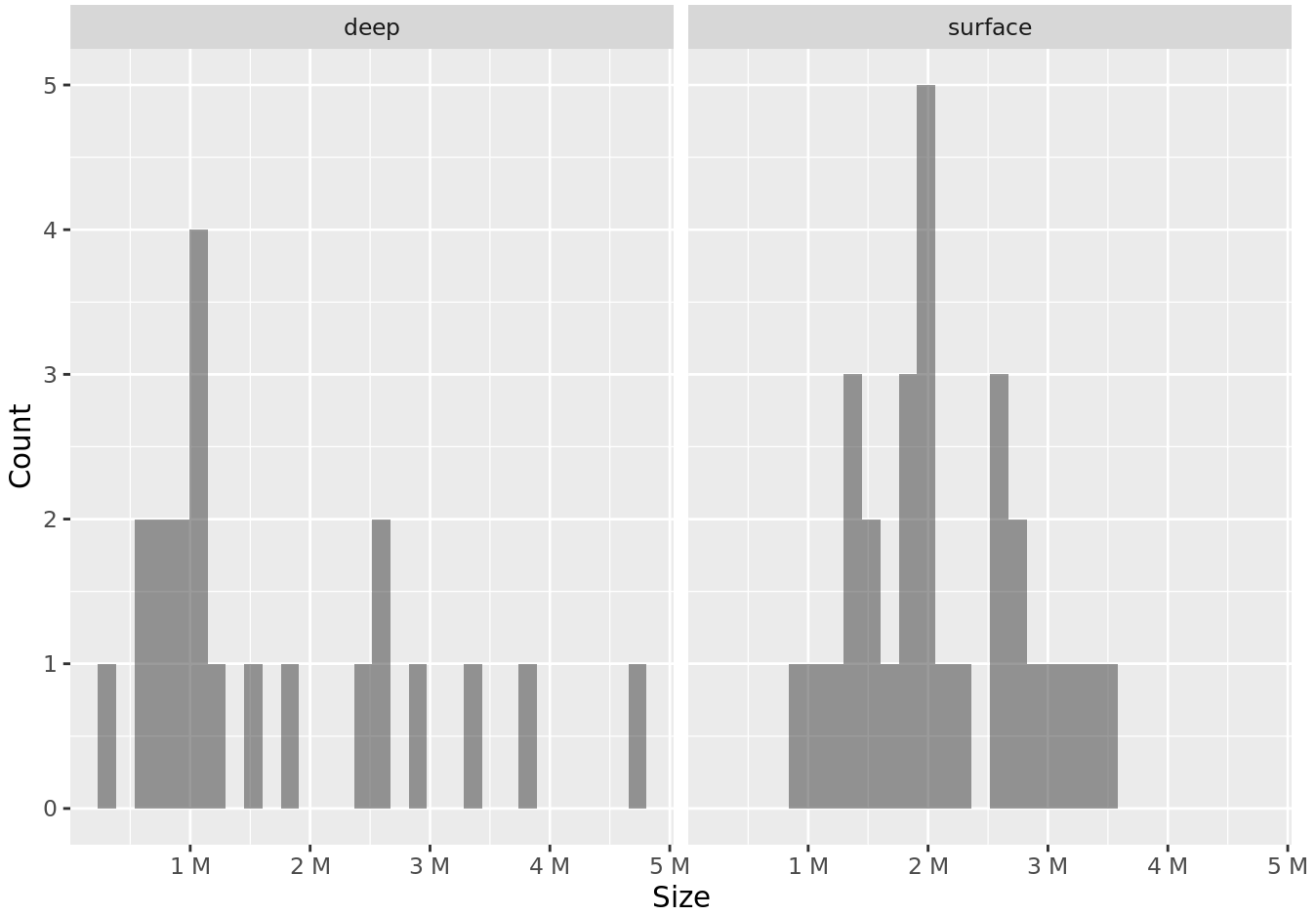

```
contamination_lineage_plot <- ggplot(data=combined_data, aes(x=lineage, y=contamination,
    geom_violin(alpha=0.5, trim=FALSE) +
    geom_jitter(width=0.1, alpha=0.5, size=2) +  # Adds individual data points
    facet_wrap(~location) +
    theme_minimal() +
    theme(
        legend.position = "none",
        axis.text.x = element_text(angle = 45, hjust = 1)  # Rotates x-axis text
    ) +
    labs(y = "contamination", x = "Lineage", title = "contamination by Lineage")
contamination_lineage_plot
```



Bin size

numbers are in million. Here we see the size of bins in million base and the frequency of each, we have some bigger bins in deep sample but the frequency of overall big bins is higher in surface sample.

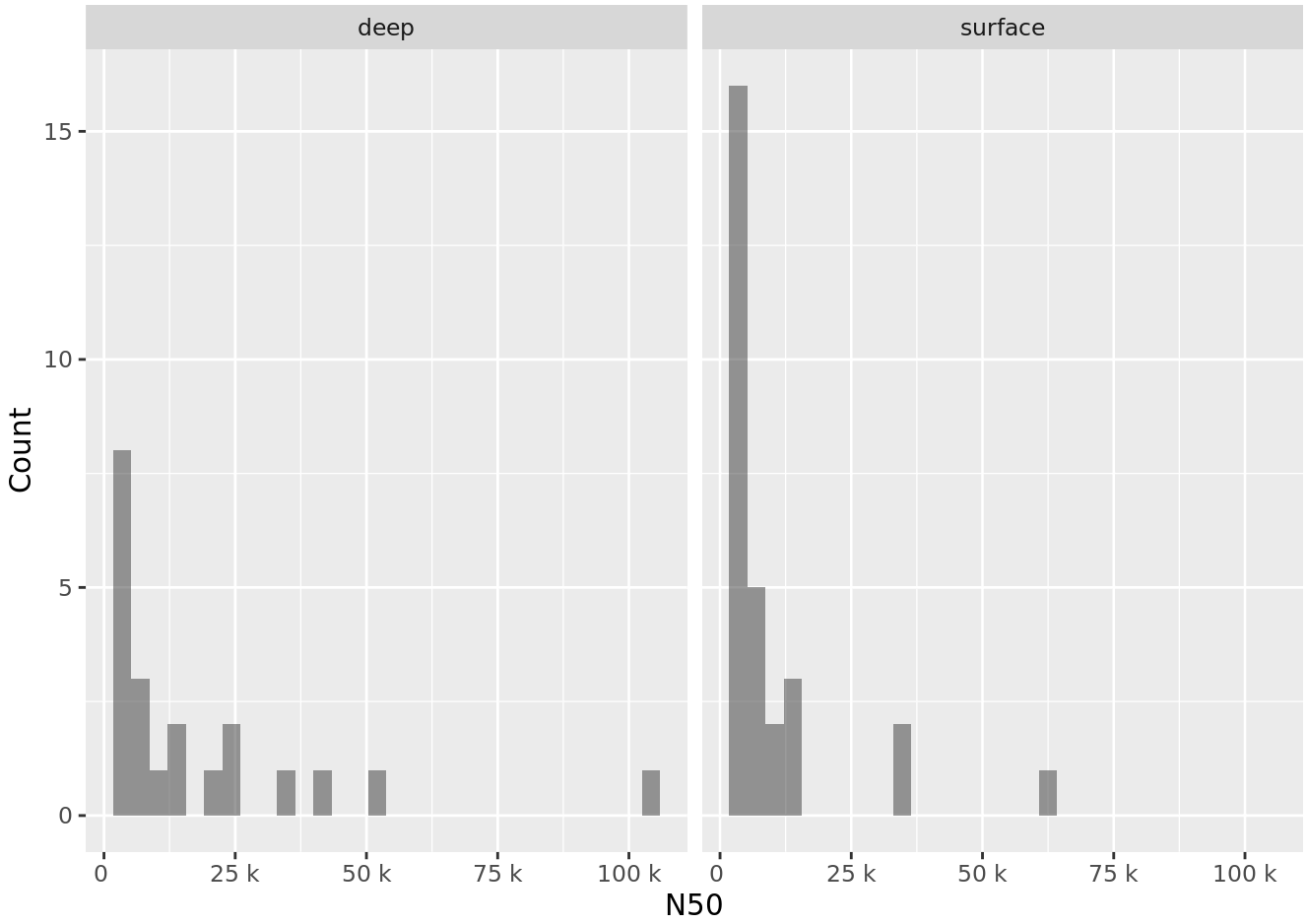
```
size_plot <- ggplot(data=combined_data, aes(x=size)) +
    geom_histogram( alpha=0.6, position = 'identity') +
    facet_wrap(~location) +
    scale_x_continuous(
        labels = scales::label_number(scale_cut = scales::cut_si(""))
    ) +
    labs(x = "Size", y = "Count", fill = "")
size_plot
```



N50

N50 describes the quality of assembled genomes or contigs. It refers to the length at which 50% of the assembled bases are contained in sequences at or above that length. describe the quality of assembled genomes or contigs. It refers to the length at which 50% of the assembled bases are contained in sequences at or above that length.

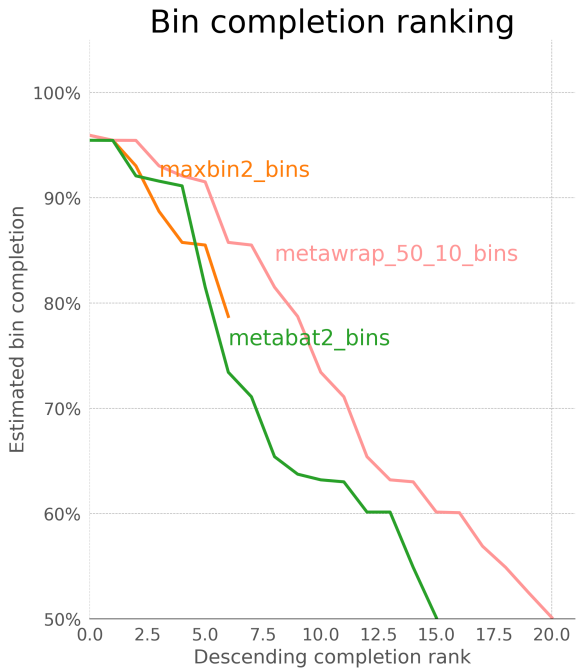
```
N50_plot <- ggplot(data=combined_data, aes(x=N50)) +
    geom_histogram( alpha=0.6, position = 'identity') +
    facet_wrap(~location) +
    scale_x_continuous(
        labels = scales::label_number(scale_cut = scales::cut_si(""))
    ) +
    labs(x = "N50", y = "Count", fill = "")
N50_plot
```

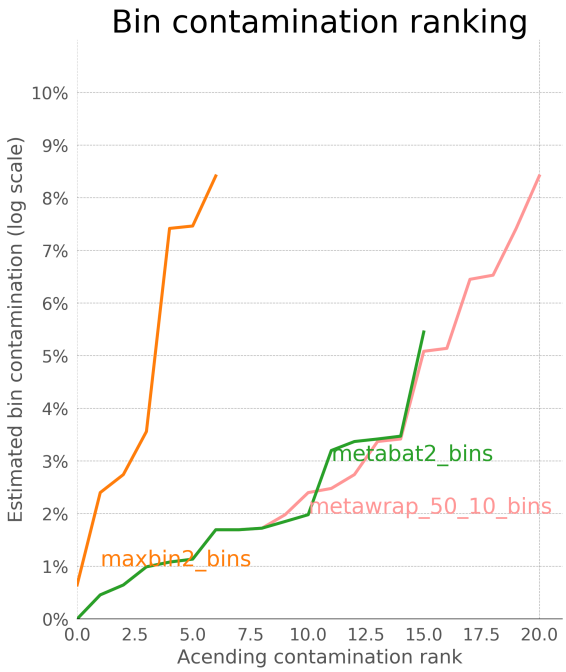


Bin Comparision

```
binning_results_compare <- image_read("/project/asteen_1130/deep_vs_surface/manual_result
print(binning_results_compare)
```

```
# A tibble: 1 × 7
  format width height colorspace matte filesize density
  <chr>  <int>  <int> <chr>      <lgl>    <int> <chr>
1 PNG     4800   2400 sRGB       TRUE    501959 +118x+118
```






```
intermediate_binning_results_compare <- image_read("/project/asteen_1130/deep_vs_surface/
print(intermediate_binning_results_compare)
```

```
# A tibble: 1 × 7
  format width height colorspace matte filesize density
  <chr>  <int>  <int> <chr>      <lgl>    <int> <chr>
1 PNG     4800   2400 sRGB       TRUE    541789 +118x+118
```