CoverM_Test

2023-08-05

Background

This script is for understanding the “relative_abundance” calculation method of CoverM more clearly. For various calculation methods of CoverM, refer to https://github.com/wwood/CoverM#calculation-methods.

What I want to calculate is the relative abundance of genome A.
Specifically, I’d like to know what would happen to the relative abundance of genome A, if I include other genomes with variable sizes that recruit variable number of metagenome reads.

```
library(tidyverse)
```

Analysis

Metagenome

Set the total number of reads = 1000.
Set the length of each read = 100 bp. To make it simple, assume that all reads have the same length.

```
tot_reads <- 1000
read_len <- 100
```

Genomes

Genome A (my focus) Genome size = 1000 bp
The number of mapped reads = 10 reads
To make it simple, assume that all matches are perfect. In other words, all mapped reads are aligned to genomes across their whole length with 100% sequence identity.
“mean” is calculated according to the CoverM formula. For simplicity, end problem is not considered.

```
size_A <- 1000
mapped_reads_A <- 10
mapped_bases_A <- mapped_reads_A * read_len
mean_A <- mapped_bases_A / size_A
```

Genomes B, C, and D (other genomes) Set the size of other genomes.
Three genomes with different sizes will be used here to test the effects of the size of other genomes.

- B is smaller than A.
- C is equal to A.
- D is larger than A.

The number of metagenome reads mapped to the other genomes is set to change from 0 to 50.
This number will be used as a variable for x-axis in the plotting.
The same range is used for the three genomes.

```
size_B <- 500
size_C <- 1000
size_D <- 2000

mapped_reads_other <- 0:50
```

Calculate the "relative_abundance" of genome A

Three genome sets will be created, each including Genome A and one other genome (B, C, or D).
"relative_abundance" of genome A will be calculated according to the CoverM formula.

```
coverm <- data.frame(mapped_reads_other) %>%
  mutate(mapped_bases_other = mapped_reads_other * read_len) %>%
  mutate(mean_B = mapped_bases_other / size_B) %>%
  mutate(mean_C = mapped_bases_other / size_C) %>%
  mutate(mean_D = mapped_bases_other / size_D) %>%
  mutate(relcov_A_with_B = mean_A / (mean_A + mean_B)) %>%
  mutate(relcov_A_with_C = mean_A / (mean_A + mean_C)) %>%
  mutate(relcov_A_with_D = mean_A / (mean_A + mean_D)) %>%
  mutate(tot_mapped_reads = mapped_reads_A + mapped_reads_other) %>%
  mutate(relabu_A_with_B = relcov_A_with_B * tot_mapped_reads / tot_reads) %>%
  mutate(relabu_A_with_C = relcov_A_with_C * tot_mapped_reads / tot_reads) %>%
  mutate(relabu_A_with_D = relcov_A_with_D * tot_mapped_reads / tot_reads)
```

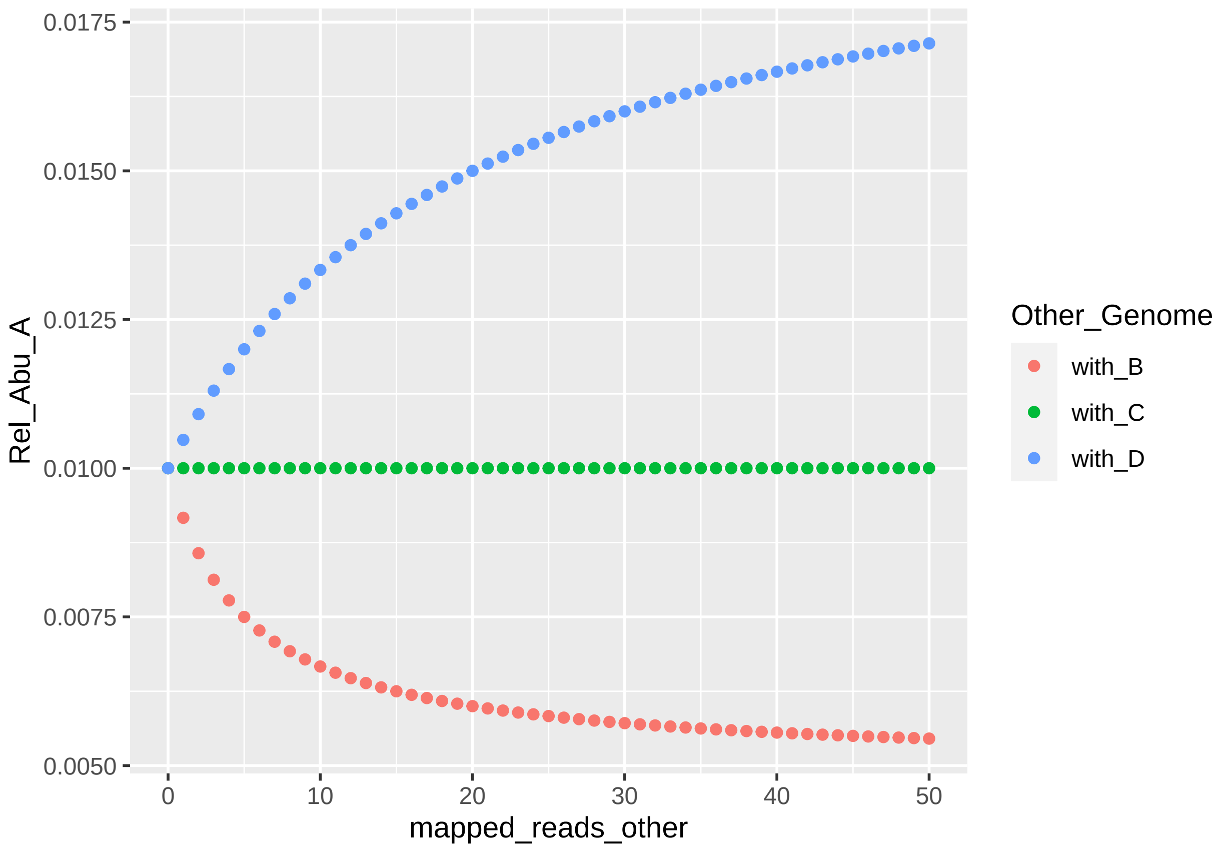
Create a data frame for plotting based on the above calculation.

```
coverm_plot <- coverm %>%
  select(mapped_reads_other, starts_with("relabu")) %>%
  rename_with(str_replace,
              pattern = "relabu_A_", replacement = "") %>%
  pivot_longer(-mapped_reads_other, names_to = "Other_Genome", values_to = "Rel_Abu_A")
```

Plotting

x-axis shows the number of reads mapped to the other genomes.
Therefore, "x=0" means that no other genomes are included.
y-axis shows the "relative_abundance" of genome A with the other three genomes of variable sizes. Color indicates the other genome.

```
ggplot(coverm_plot, aes(x = mapped_reads_other, y = Rel_Abu_A)) +
  geom_point(aes(color = Other_Genome))
```



Conclusion

1. "relative_abundance" of a genome (focal genome; A) is affected by the inclusion of another genome in the analysis
2. If the other genome is smaller than the focal genome ("with_B"), "relative_abundance" of the focal genome decreases.
3. If the other genome is larger than the focal genome ("with_C"), "relative_abundance" of the focal genome increases.
4. The degree of increase/decrease is dependent on the number of reads mapped to the other genome.