Bioinformatics Assignment

Stepwise Workflow

1) Started OnDemand and navigated to my personal directory in “tillers”
2) Uploaded my file “Rice_GRIN-Global.csv” using the upload function in the “On demand”
3) Logged into the HPC on the terminal using “*ssh username@login.palmetto.clemson.edu*”
4) My folder: *cd /pathway_to_directory*
5) Created a new directory: *mkdir Rice_project*
6) Created an R file: *touch codes.R*
7) Requested computing nodes and memory using salloc: *salloc --nodes=1 --ntasks=32 --mem=12G --time=2:00:00 --partition=floret*
8) Loaded the R module: *module load r/4.4.0* and started *R*
9) Installed required packages in R
    - *install.packages (“ggplot2”)*
    - *install.packages (“tidyverse”)*
    - *install.packages (“dplyr”)*
    - *install.packages (“corrplot”)*
10) Opened a file and wrote R code
    - *vim codes.R* and saved file using “*escape*” then “:*wq!*”
11) R code in the file

```
>   library("corrplot")
>   library("dplyr")
>   library("tidyr")
>   library(ggplot2)

>   data_rice <- read.csv("/project/skresov/tillers/djasrot/Rice_project/Rice_GRIN-Global.csv")

>   # Convert columns into numeric
>   data_rice$AMYLOSE <- as.numeric(data_rice$AMYLOSE)
>   data_rice$`KERNEL.LENGTH.WIDTH.RATIO` <- as.numeric(data_rice$`KERNEL.LENGTH.WIDTH.RATIO`)
>   data_rice$`LENGTH.OF.KERNEL` <- as.numeric(data_rice$`LENGTH.OF.KERNEL`)
>   data_rice$`SALT.TOLERANCE` <- as.numeric(data_rice$`SALT.TOLERANCE`)
>   data_rice$`WIDTH.OF.KERNEL` <- as.numeric(data_rice$`WIDTH.OF.KERNEL`)


>   traits_rice <- c(
      "AMYLOSE",
      "KERNEL.LENGTH.WIDTH.RATIO",
      "LENGTH.OF.KERNEL",
      "WIDTH.OF.KERNEL",
      "SALT.TOLERANCE"
    )
```

```
>   #plot1 - correlation matrix
>   cm <- cor(data_rice[traits_rice], use = "pairwise.complete.obs", method = "pearson")
>   corrplot(
      cm,
      method = "circle",
      type   = "lower",
      diag   = FALSE,
      tl.col = "red",
      tl.cex = 0.5,
      addCoef.col = "black",
      number.cex = 0.7,
      mar = c(0,0,2,0),
      title  = "Correlation between traits (Rice-GRIN)",
    )

>   #trying to visualize number of origins in the data file
>   length(unique(data_rice$ORIGIN))
>   unique(data_rice$ORIGIN)

>   #no of origins in usa
>   sum(grepl("United States", data_rice$ORIGIN, ignore.case = TRUE))

>   #plot2 - scatter plot between amylose and kernel length/width ratio
>   r_val <- cor(data_rice$AMYLOSE, data_rice$`KERNEL.LENGTH.WIDTH.RATIO`, use =
    "pairwise.complete.obs")

>   p <- ggplot(data_rice, aes(x = AMYLOSE, y = `KERNEL.LENGTH.WIDTH.RATIO`)) +
      geom_point(color = "blue", alpha = 0.6, size = 2) +
      geom_smooth(method = "lm", se = TRUE, color = "red") +
      annotate("text",
               x = max(data_rice$AMYLOSE, na.rm = TRUE) * 0.95,
               y = max(data_rice$`KERNEL.LENGTH.WIDTH.RATIO`, na.rm = TRUE) * 0.95,
               label = paste0("R = ", round(r_val, 2)),
               size = 5, color = "black", hjust = 1) +
      theme_minimal(base_size = 13) +
      labs(title = "Scatterplot: Amylose vs kernel length/width ratio",
           x = "Amylose Content",
           y = "Kernel Length/Width Ratio") +
      theme(panel.border = element_rect(color = "black", fill = NA, linewidth = 1),
            plot.title = element_text(hjust = 0.5, face = "bold"))

    ggsave("scatter_amylose_vs_ratio_v2.pdf", p, width = 8, height = 6, dpi = 300)
```

12) Ran the R script file using “Rscript codes.R”

13) Two plots were generated and saved

a) Correlation plot:

Correlation analysis showed that amylose content is moderately correlated with kernel length/width ratio (R = 0.48) and slightly negatively correlated with salt tolerance (r = -0.04). Whereas the length of the kernel is strongly correlated with the width (R = 0.71) of the kernel and negatively correlated with salt tolerance (R = -0.19).

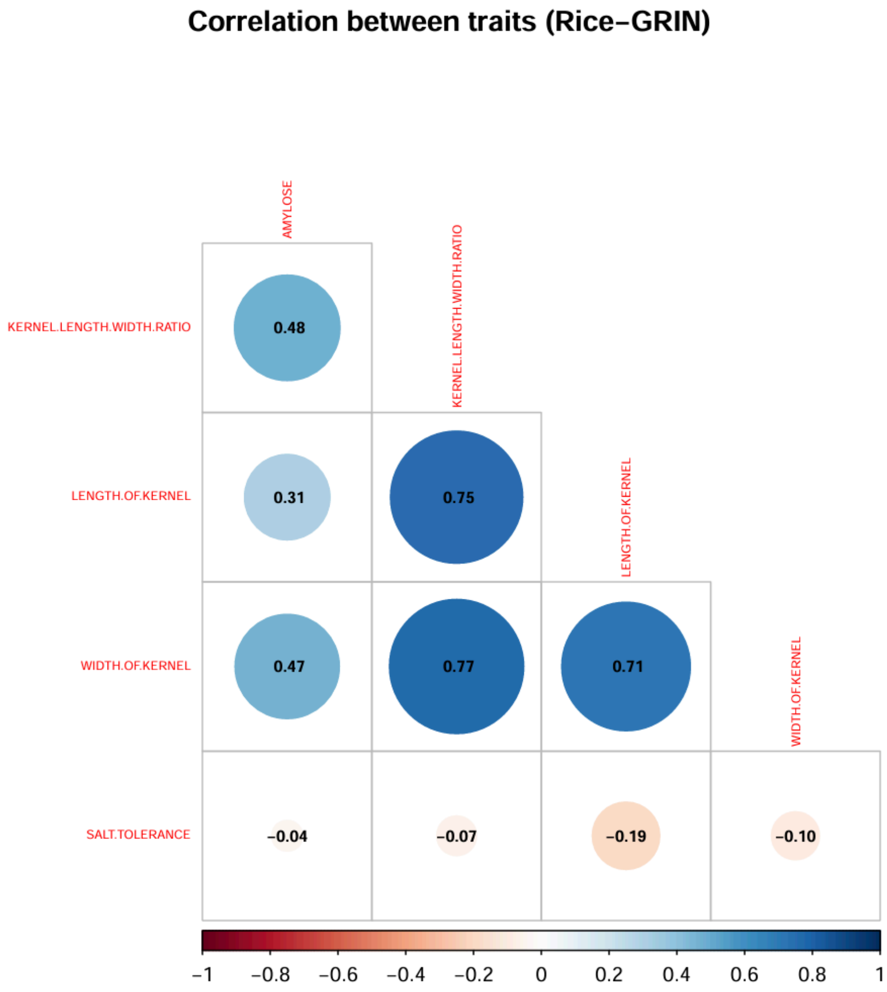

Fig.1 Correlation matrix between seed traits and salt tolerance. The color represents the direction of correlation: blue indicates positive correlation, whereas red indicates negative correlation. The size of the circle represents the magnitude of correlation between traits.

b) Scatterplot between amylose and kernel length/width ratio with regression line:
The scatter plot**)** suggests a moderate positive correlation (R = 0.48). As amylose content increases, the kernel length-to-width ratio tends to increase as well. However, the data points show some variability, meaning the relationship is not perfectly linear.

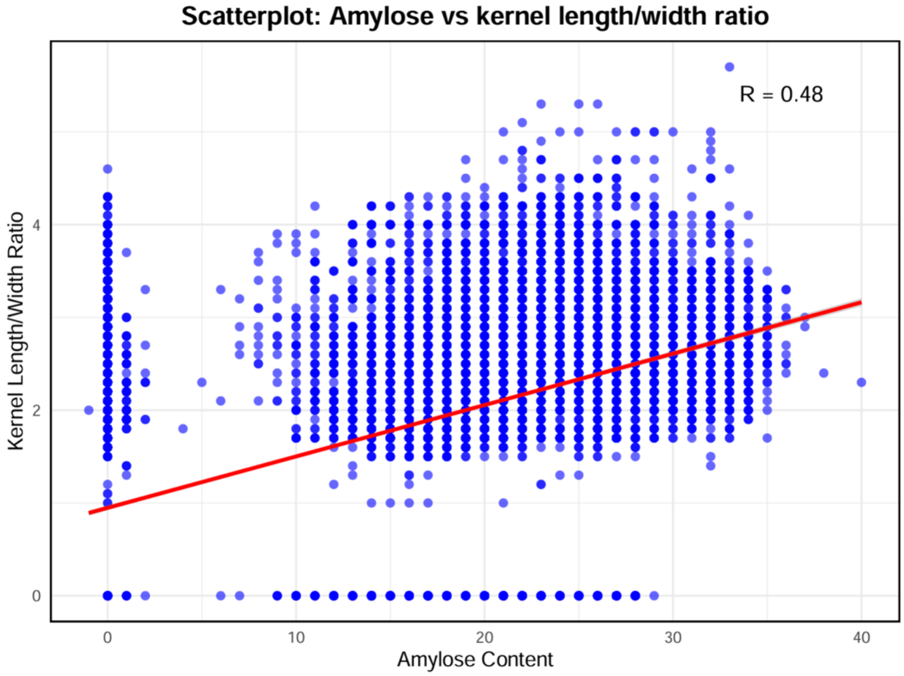

Fig. 2 Scatter plot showing the relationship between amylose content and kernel length/width ratio in rice. Each blue point represents an individual sample, while the red line shows the best-fit regression line.