Thesis SNV Filtering

Max Barclay

2023-01-30

```
#Reads the SNP csv files for each strains snippy
#Filters these to omit NAs and hypothetical proteins
Y1_1 <- read.csv(file="R_data/Y1.1.csv", header=TRUE) %>% filter(FTYPE != "" & PRODUCT != "hypothetical
Y1_2 <- read.csv(file="R_data/Y1.2.csv", header=TRUE) %>% filter(FTYPE != "" & PRODUCT != "hypothetical
Y1_3 <- read.csv(file="R_data/Y1.3.csv", header=TRUE) %>% filter(FTYPE != "" & PRODUCT != "hypothetical

#Puts strain csv's into data frames
Y1_1Df <- data.frame(Y1_1)
Y1_2Df <- data.frame(Y1_2)
Y1_3Df <- data.frame(Y1_3)

#Filters the data frames into appropriate tables
Y1_1Df_filter <- subset(Y1_1Df, select = c("LOCUS_TAG", "PRODUCT"))
Y1_2Df_filter <- subset(Y1_2Df, select = c("LOCUS_TAG", "PRODUCT"))
Y1_3Df_filter <- subset(Y1_3Df, select = c("LOCUS_TAG", "PRODUCT"))
```

Anything hypothetical or unidentified by resequencing was filtered out here, while strains .csv files were put into data frames to allow for them to be converted to tables that only used the columns for locus tag and product to ensure no repeats and clarity for what the loci were linked to.

```
#Merged table for appendix
merged_tables <- rbind.fill(Y1_1Df_filter, Y1_2Df_filter, Y1_3Df_filter)
print(merged_tables)
```

```
##    LOCUS_TAG                                                           PRODUCT
## 1    SCO2962                              bifunctional transferase/deacetylase
## 2    SCO4127                                           ATP/GTP-binding protein
## 3    SCO4594             2-oxoglutarate ferredoxin oxidoreductase subunit beta
## 4    SCO4594             2-oxoglutarate ferredoxin oxidoreductase subunit beta
## 5    SCO4595                                                    oxidoreductase
## 6    SCO4654                          DNA-directed RNA polymerase subunit beta
## 7    SCO4659                                         30S ribosomal protein S12
## 8    SCO4659                                         30S ribosomal protein S12
## 9    SCO5065                                         transcriptional regulator
## 10   SCO6167                             proline rich protein membrane protein
## 11   SCO7350                                           membrane efflux protein
## 12   SCO2962                              bifunctional transferase/deacetylase
## 13   SCO4127                                           ATP/GTP-binding protein
## 14   SCO4594             2-oxoglutarate ferredoxin oxidoreductase subunit beta
## 15   SCO4594             2-oxoglutarate ferredoxin oxidoreductase subunit beta
## 16   SCO4595                                                    oxidoreductase
## 17   SCO4654                          DNA-directed RNA polymerase subunit beta
## 18   SCO4659                                         30S ribosomal protein S12
## 19   SCO4659                                         30S ribosomal protein S12
## 20   SCO5065                                         transcriptional regulator
## 21   SCO6167                             proline rich protein membrane protein
## 22   SCO7015                                                glycosyl hydrolase
## 23   SCO7350                                           membrane efflux protein
## 24   SCO2962                              bifunctional transferase/deacetylase
## 25   SCO4127                                           ATP/GTP-binding protein
## 26   SCO4594             2-oxoglutarate ferredoxin oxidoreductase subunit beta
## 27   SCO4594             2-oxoglutarate ferredoxin oxidoreductase subunit beta
## 28   SCO4595                                                    oxidoreductase
## 29   SCO4654                          DNA-directed RNA polymerase subunit beta
## 30   SCO4659                                         30S ribosomal protein S12
## 31   SCO4659                                         30S ribosomal protein S12
## 32   SCO5065                                         transcriptional regulator
## 33   SCO5090 actinorhodin polyketide synthase bifunctional cyclase/dehydratase
## 34   SCO6167                             proline rich protein membrane protein
## 35   SCO7350                                           membrane efflux protein
```

```
#Gives excel output of filtered tables in environment for external use
write.xlsx(Y1_1Df_filter, "Y1.1_Table.xlsx", colNames = TRUE, rowNames = TRUE)
write.xlsx(Y1_2Df_filter, "Y1.2_Table.xlsx", colNames = TRUE, rowNames = TRUE)
write.xlsx(Y1_3Df_filter, "Y1.3_Table.xlsx", colNames = TRUE, rowNames = TRUE)
```

rbind was used to give a full non-segregated table for all three strains. The individual filtered tables from before were exported to excel to allow for any modifications and use as figures in presentation etc.