Workflow notebook

Jay Gillenwater

29 August, 2021

Overview

This document will be my journal of sorts for the development of the workflow for the analysis of the NC Raleigh x Soja mapping population. I'll try to both document the functions I make and also give results of the analysis as I go along.

Phenotype file overview

The phenotype input file is **Pheno_five_locations.xlsx**. This file has measurements for the genotypes split into sheets for each of the five locations they were grown in. The three phenotypes that were measured were percent nitrogen, carbon, and sulfur content. Each sheet has additional information relevant to the measurement process like run date and time, and various standards and quantities involved in the calculation of each elemental content.

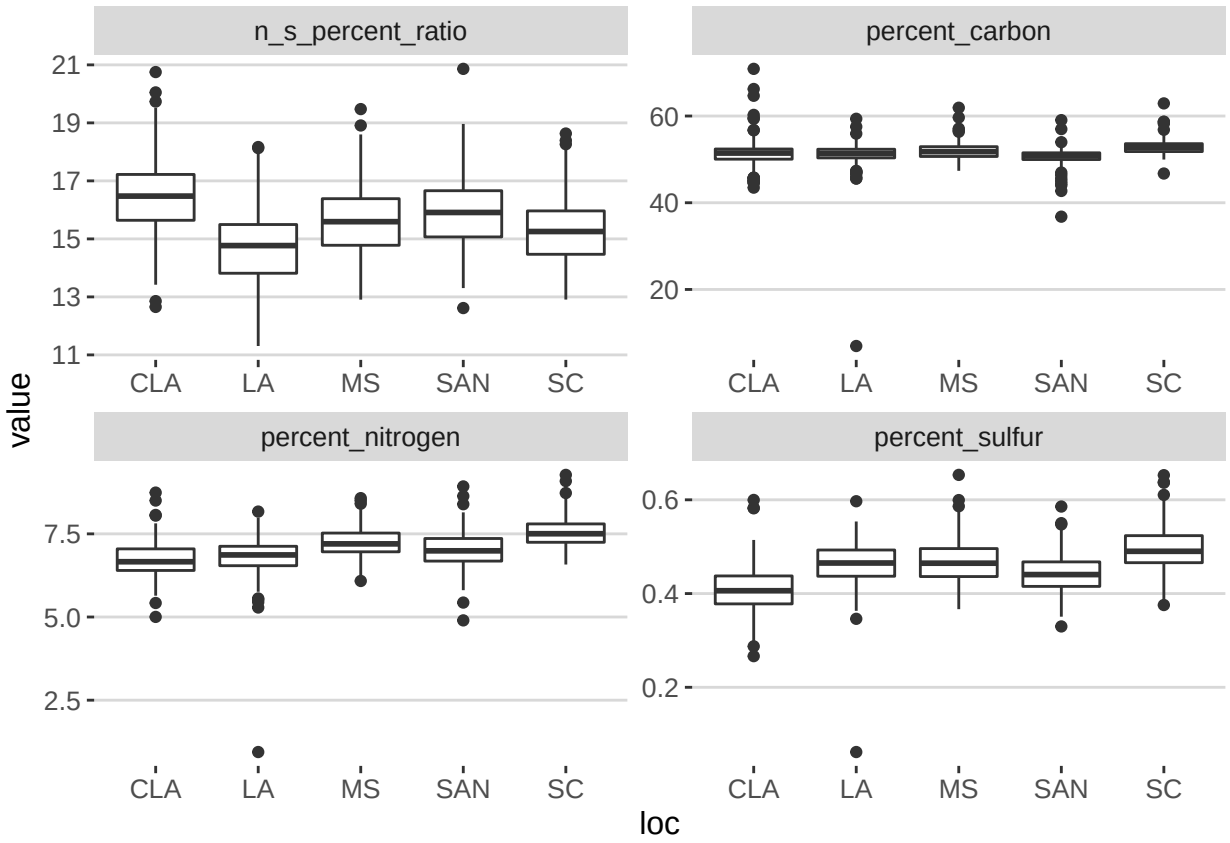
Phenotype data cleaning

To make analysis easier, I want to get all the phenotype data into one dataframe first. This is done with the **read_clean** function. Briefly, this function reads in each of the location sheets, adds the location to the data, and then selects a subset of columns (the sample name, location, and phenotypes). The plot of each sample is then derived from the sample name, and the rep number in turn from the plot number. A fourth phenotype is then calculated by dividing the percent nitrogen by the percent sulfur (n_s_percent_ratio). Any sample that does not have a code is filtered out at this point. It looks like these samples without a code name were the standards and checks. The code, location, and rep columns are converted to factors before the final dataframe is returned. Here's what this data looks like.

```
## Rows: 1,343
## Columns: 10
## $ sample_name       <chr> "SC PLOT: 263", "SC PLOT: 119", "SC PLOT: 457", "SC ~
## $ code              <fct> 1901, 1902, 1902, 1903, 1903, 1904, 1905, 1906, 1906~
## $ wt_mg             <dbl> 4.18, 3.38, 5.05, 5.11, 5.57, 5.08, 5.39, 4.12, 6.26~
## $ percent_nitrogen  <dbl> 7.705441, 8.362654, 7.288193, 7.260160, 7.299695, 7.~
## $ percent_carbon    <dbl> 55.13361, 54.67350, 51.75912, 52.86848, 55.54053, 53~
## $ percent_sulfur    <dbl> 0.5214009, 0.5693756, 0.4781052, 0.4444401, 0.475756~
## $ loc               <fct> SC, SC, SC, SC, SC, SC, SC, SC, SC, SC, SC, SC, SC, ~
## $ plot              <dbl> 263, 119, 457, 159, 283, 27, 61, 3, 393, 71, 437, 21~
## $ rep               <fct> 1, 1, 2, 1, 2, 1, 1, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1~
## $ n_s_percent_ratio <dbl> 14.77834, 14.68741, 15.24391, 16.33552, 15.34335, 15~
```

EDA

I'll check out the distributions of the phenotypes, boxplot first.



While there's a few samples that fall outside the 1.5 IQR, it looks like there's one sample in carbon, nitrogen, and sulfur in the LA environment that is way lower than the other samples.

```
## # A tibble: 4 x 8
## # Groups:   name [4]
##   sample_name  code  wt_mg loc    plot rep   name                value
##   <chr>        <fct> <dbl> <fct> <dbl> <fct> <chr>               <dbl>
## 1 LA PLOT: 255 2087  46.3  LA      255 1     percent_sulfur     0.0619
## 2 LA PLOT: 255 2087  46.3  LA      255 1     percent_nitrogen   0.941
## 3 LA PLOT: 255 2087  46.3  LA      255 1     percent_carbon     6.96
## 4 LA PLOT: 133 2046   3.06 LA      133 1     n_s_percent_ratio 11.3
```

It looks like these three low observations all come from the same plot. Beyond the three phenotypes, it looks like this plot also has an abnormally high weight. I'll look at the other measurements for this genotype to see if it is only this plot that has very low phenotypes and very high weight.

```
## # A tibble: 10 x 10
##    sample_name  code  wt_mg percent_nitrogen percent_carbon percent_sulfur loc
##    <chr>        <fct> <dbl>            <dbl>          <dbl>          <dbl> <fct>
##  1 CLA PLOT: 1~ 2087   6.79            6.61           51.4          0.402  CLA
##  2 CLA PLOT: 4~ 2087   5.09            7.14           52.7          0.456  CLA
##  3 LA PLOT: 255 2087  46.3             0.941           6.96         0.0619 LA
##  4 LA PLOT: 517 2087   3.24            6.76           51.9          0.465  LA
##  5 MS PLOT: 29  2087   4.06            7.66           51.4          0.473  MS
##  6 MS PLOT: 415 2087   3.34            7.79           54.5          0.535  MS
##  7 SAN PLOT: 1~ 2087   4.27            7.24           51.9          0.507  SAN
##  8 SAN PLOT: 4~ 2087   5.17            7.40           50.7          0.455  SAN
##  9 SC PLOT: 97  2087   4.64            6.90           52.3          0.468  SC
```

```
## 10 SC PLOT: 479 2087   5.29            7.64           51.9          0.482  SC
## # ... with 3 more variables: plot <dbl>, rep <fct>, n_s_percent_ratio <dbl>
```

It looks like it's only this one plot with very low values, I'll remove this plot from the analysis it seems like it was probably due to a measurement error.

Reproducibility

Reproducibility receipt

```
## [1] "2021-08-29 12:52:07 EDT"
```

```
## Local:    main C:/Users/Jay/Desktop/Documents/R/NCRaleigh_Soja_QTL_Mapping
## Remote:   main @ origin (https://github.com/jhgille2/NCRaleigh_Soja_QTL_Mapping.git)
## Head:     [22bb69b] 2021-08-25: Update README.md
```

```
## R version 4.1.0 (2021-05-18)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19043)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] future.callr_0.6.1 future_1.22.1      job_0.3.0          ASMap_1.0-4
##  [5] RColorBrewer_1.1-2 fields_12.5        viridis_0.6.1      viridisLite_0.4.0
##  [9] spam_2.7-0         dotCall64_1.0-1    gtools_3.9.2       vcfR_1.12.0
## [13] vroom_1.5.3        emmeans_1.6.2-1    reactable_0.2.3    lattice_0.20-44
## [17] ggthemes_4.2.4     magrittr_2.0.1     janitor_2.1.0      readxl_1.3.1
## [21] here_1.0.1         qtl_1.48-1         rmarkdown_2.9      forcats_0.5.1
## [25] stringr_1.4.0      dplyr_1.0.7        purrr_0.3.4        readr_1.4.0
## [29] tidyr_1.1.3        tibble_3.1.2       ggplot2_3.3.5      tidyverse_1.3.1
## [33] tarchetypes_0.2.1  targets_0.4.2      dotenv_1.0.3       conflicted_1.0.4
## [37] pacman_0.5.1
##
## loaded via a namespace (and not attached):
##  [1] colorspace_2.0-2  ellipsis_0.3.2    rprojroot_2.0.2   estimability_1.3
##  [5] snakecase_0.11.0  fs_1.5.0          rstudioapi_0.13   farver_2.1.0
##  [9] listenv_0.8.0     bit64_4.0.5       fansi_0.5.0       mvtnorm_1.1-2
## [13] lubridate_1.7.10  xml2_1.3.2        codetools_0.2-18  splines_4.1.0
## [17] cachem_1.0.5      knitr_1.33        jsonlite_1.7.2    broom_0.7.9
## [21] cluster_2.1.2     dbplyr_2.1.1      compiler_4.1.0    httr_1.4.2
## [25] backports_1.2.1   assertthat_0.2.1  Matrix_1.3-3      fastmap_1.1.0
## [29] cli_3.0.0         htmltools_0.5.1.1 tools_4.1.0       igraph_1.2.6
## [33] coda_0.19-4       gtable_0.3.0      glue_1.4.2        maps_3.3.0
## [37] Rcpp_1.0.7        cellranger_1.1.0  vctrs_0.3.8       ape_5.5
```

```
## [41] nlme_3.1-152      pinfsc50_1.2.0    xfun_0.24         globals_0.14.0
## [45] ps_1.6.0          rvest_1.0.1       lifecycle_1.0.0   MASS_7.3-54
## [49] scales_1.1.1      hms_1.1.0         parallel_4.1.0    yaml_2.2.1
## [53] gridExtra_2.3     stringi_1.6.2     highr_0.9         permute_0.9-5
## [57] rlang_0.4.11      pkgconfig_2.0.3   evaluate_0.14     labeling_0.4.2
## [61] htmlwidgets_1.5.3 bit_4.0.4         processx_3.5.2    tidyselect_1.1.1
## [65] parallelly_1.27.0 R6_2.5.0          generics_0.1.0    DBI_1.1.1
## [69] pillar_1.6.2      haven_2.4.1       withr_2.4.2       mgcv_1.8-35
## [73] modelr_0.1.8      crayon_1.4.1      utf8_1.2.1        tzdb_0.1.2
## [77] data.table_1.14.0 git2r_0.28.0      callr_3.7.0       vegan_2.5-7
## [81] reprex_2.0.0      digest_0.6.27     xtable_1.8-4      munsell_0.5.0
```