PCA_Analysis_lyrataData

Carina Bravo-Chan

2023-04-12

Proposed work-flow for independent project:

This data set is used on the ongoing research project in Dr. Remington's lab. In our lab we study plants' evolutionary life history and the molecular basis of phenotype divergence within populations. The system studied consists of two sub-populations of the rock cress *Arabdopsis lyrata*. One group from Spiterstulen (SP), Norway, and the other from Mayodan (MA), North Carolina. These populations were chosen because they occupy environments on opposite ends of the spectrum, and display different patterns of resource allocation for reproductive and vegetative growth. The data consists of phenotype records of traits from F2 individuals of *A. lyrata*, accompanied by the genotype of QTL regions. The hypothesis is that the resulting F2 phenotype is influenced by the type of allele inherited, Mayodan or Spiterstulen. Individual observations were classified according to the number of allele copies, inherited from Mayodan parents, present in the genes of a QTL region. Individuals with **no** MA alleles, were identified with the number 0. It was expected plants with a more perennial-like phenotype, displaying a longer season of vegetative growth, and a short reproductive season with modest flowering. The heterozygous individuals were identified with the number 1 (one parental copy of the MA allele), and an intermediary phenotype was expected with resource allocation displaying displaying a mixed, or inconsistent, pattern. Finally, the individuals that were homozygous MA were identified with the number 2 (two parental copies of the MA allele). The expected phenotype was more annual-like, displaying stagnation of vegetative growth for overwintering, followed by an intense resource allocation for reproductive growth with significantly more bolting, branching, and flowering. The goal is to write a script to analyze the trends within this data set to find if there is any significant correlation between perennial traits and the underlying alleles. For example, a lower number of inflorescences suggesting a more perennial phenotype, and would allow us to examine if the genotype scoring for this, and other, traits would corroborate our hypothesis.

First step: Tidying and cleaning the data so only the relevant variables will be at hand.

```
#Load relevant libraries
library(tidyverse)
library(mice)
library(Hmisc)
library(mi)

#Read in the data
mtDataAll <- read.csv("QTLStudyTraitAnalysis_survMarkers.csv")

#Generate calculated fields
mtDataAll2 <- mtDataAll %>%
  mutate(dDiam21 = Diam2 - Diam1,
         dDiam32 = Diam3 - Diam2,
         meanBLvs = rowMeans(select(., 12:14), na.rm = TRUE),
         meanOrder = rowMeans(select(., 15, 18, 21), na.rm = TRUE),
         meanBrNode = rowMeans(select(., 16, 19, 22), na.rm = TRUE),
         meanFlNode = rowMeans(select(., 17, 20, 23), na.rm = TRUE)) %>%
```

```r
#Convert variables to factors
  mutate(Pop = as.factor(Pop),
         Fam = as.factor(Fam),
         PShootStatus0215 = as.factor(PShootStatus0215),
         LShootStatus0215 = as.factor(LShootStatus0215))

#Create a new dataframe with just the F2 plants
mtDataF2 <- filter(mtDataAll2, Pop == "F2")

#Calculate the mean of inflorescence branches
mtDataF2$meanBrNumber <- apply(mtDataF2[, c(17, 20, 23)], 1, mean, na.rm = TRUE) - apply(mtDataF2[, c(1

#Check variable names
colnames(mtDataF2)
```

```
##  [1] "Pop"                  "Fam"                  "Num"
##  [4] "X2nd_Data_colec_date" "LatRating0215"        "PShootStatus0215"
##  [7] "LShootStatus0215"     "Rhiz0215"             "Remarks1"
## [10] "Remarks2"             "Collector"            "S1BLvs"
## [13] "S2BLvs"               "S3BLvs"               "S4Order"
## [16] "S4BrNode"             "S4FlNode"             "S5Order"
## [19] "S5BrNode"             "S5FlNode"             "S6Order"
## [22] "S6BrNode"             "S6FlNode"             "Rhiz0615"
## [25] "CollDate"             "Comments"             "Diam1"
## [28] "Diam2"                "Diam3"                "Infl"
## [31] "Surv1216"             "PIN1"                 "PIN3"
## [34] "PIN_combined"         "LFY"                  "GI"
## [37] "SOC1"                 "GenoComment"          "dDiam21"
## [40] "dDiam32"              "meanBLvs"             "meanOrder"
## [43] "meanBrNode"           "meanFlNode"           "meanBrNumber"
```

```r
#Check variable classes
mtDataF2 %>%
  select(Pop, Fam, PShootStatus0215, LShootStatus0215) %>%
  map(class)
```

```
## $Pop
## [1] "factor"
##
## $Fam
## [1] "factor"
##
## $PShootStatus0215
## [1] "factor"
##
## $LShootStatus0215
## [1] "factor"
```

```r
#Check variable names
names(mtDataF2)
```

```
##  [1] "Pop"                  "Fam"                  "Num"
```

```
##  [4] "X2nd_Data_colec_date" "LatRating0215"        "PShootStatus0215"
##  [7] "LShootStatus0215"     "Rhiz0215"             "Remarks1"
## [10] "Remarks2"             "Collector"            "S1BLvs"
## [13] "S2BLvs"               "S3BLvs"               "S4Order"
## [16] "S4BrNode"             "S4FlNode"             "S5Order"
## [19] "S5BrNode"             "S5FlNode"             "S6Order"
## [22] "S6BrNode"             "S6FlNode"             "Rhiz0615"
## [25] "CollDate"             "Comments"             "Diam1"
## [28] "Diam2"                "Diam3"                "Infl"
## [31] "Surv1216"             "PIN1"                 "PIN3"
## [34] "PIN_combined"         "LFY"                  "GI"
## [37] "SOC1"                 "GenoComment"          "dDiam21"
## [40] "dDiam32"              "meanBLvs"             "meanOrder"
## [43] "meanBrNode"           "meanFlNode"           "meanBrNumber"
```

Second Step: Create data frames for specific calculations. One where the effect of a QTL region is decomposed into additive dominance components. Then sort the desired traits within this data frame.

```
#Load relevant library
library(tidyverse)
library(mice)
library(Hmisc)
library(mi)

#Decompose GI genotype into additive and dominance components
mtDataF2Dec <- mtDataF2 %>% #change df name
  mutate(GIAdd = GI - 1,
         GIDom = 1 - abs(1 - GI))

#Single-trait regressions on GI genotype
ST_models <- c("Diam3", "dDiam32", "Infl", "Diam2", "Diam1", "dDiam21",
                        "meanBLvs", "meanOrder", "LatRating0215", "Rhiz0615",
                        "meanFlNode", "meanBrNode", "meanBrNumber")

for (trait in ST_models) {
  formula <- formula(paste(trait, "~ GI"))
  model <- lm(formula, data = mtDataF2)
  print(summary(model))
  ggplot(mtDataF2, aes(x = GI, y = !!sym(trait))) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE, color = "blue") +
    labs(x = "GI", y = trait)
}
```

```
##
## Call:
## lm(formula = formula, data = mtDataF2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.571 -15.127  -0.849  13.401  59.151
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   71.127      1.382  51.453   <2e-16 ***
## GI             2.722      1.334   2.041   0.0418 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.42 on 454 degrees of freedom
##   (144 observations deleted due to missingness)
## Multiple R-squared:  0.009092,   Adjusted R-squared:  0.00691
## F-statistic: 4.166 on 1 and 454 DF,  p-value: 0.04183
##
##
## Call:
## lm(formula = formula, data = mtDataF2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.003 -18.003  -0.898  15.997  81.102
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -28.102      1.931 -14.555   <2e-16 ***
## GI             3.105      1.872   1.659   0.0979 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.39 on 420 degrees of freedom
##   (178 observations deleted due to missingness)
## Multiple R-squared:  0.00651,    Adjusted R-squared:  0.004145
## F-statistic: 2.752 on 1 and 420 DF,  p-value: 0.09787
##
##
## Call:
## lm(formula = formula, data = mtDataF2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.7283  -4.6238  -0.6238   4.3762  17.1672
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.6238     0.4892  38.068   <2e-16 ***
## GI            0.1045     0.4725   0.221    0.825
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.892 on 456 degrees of freedom
##   (142 observations deleted due to missingness)
## Multiple R-squared:  0.0001072,  Adjusted R-squared:  -0.002086
## F-statistic: 0.04891 on 1 and 456 DF,  p-value: 0.8251
##
##
## Call:
## lm(formula = formula, data = mtDataF2)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.162  -8.909   0.343  10.591  59.838
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  98.6565     1.2014  82.119   <2e-16 ***
## GI           -0.2472     1.1664  -0.212    0.832
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.48 on 425 degrees of freedom
##   (173 observations deleted due to missingness)
## Multiple R-squared:  0.0001057,  Adjusted R-squared:  -0.002247
## F-statistic: 0.04493 on 1 and 425 DF,  p-value: 0.8322
##
##
## Call:
## lm(formula = formula, data = mtDataF2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.090 -14.085   0.915  13.919  58.910
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  102.090      1.643  62.138   <2e-16 ***
## GI            -1.005      1.578  -0.637    0.525
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.08 on 380 degrees of freedom
##   (218 observations deleted due to missingness)
## Multiple R-squared:  0.001066,   Adjusted R-squared:  -0.001562
## F-statistic: 0.4056 on 1 and 380 DF,  p-value: 0.5246
##
##
## Call:
## lm(formula = formula, data = mtDataF2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.580 -12.600  -1.659  10.498  84.341
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.34141    1.75468  -1.904   0.0577 .
## GI          -0.07819    1.68585  -0.046   0.9630
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.72 on 346 degrees of freedom
##   (252 observations deleted due to missingness)
## Multiple R-squared:  6.217e-06,  Adjusted R-squared:  -0.002884
```

```
## F-statistic: 0.002151 on 1 and 346 DF,  p-value: 0.963
##
##
## Call:
## lm(formula = formula, data = mtDataF2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4372 -1.1039 -0.2238  0.8961  5.9895
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.1039     0.1183  34.687   <2e-16 ***
## GI           -0.2134     0.1148  -1.858   0.0638 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.66 on 451 degrees of freedom
##   (147 observations deleted due to missingness)
## Multiple R-squared:  0.007599,   Adjusted R-squared:  0.005398
## F-statistic: 3.453 on 1 and 451 DF,  p-value: 0.06378
##
##
## Call:
## lm(formula = formula, data = mtDataF2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3182 -0.2879  0.0151  0.3787  2.0757
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.25760    0.04167   54.17   <2e-16 ***
## GI           0.03032    0.04044    0.75    0.454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5847 on 451 degrees of freedom
##   (147 observations deleted due to missingness)
## Multiple R-squared:  0.001245,   Adjusted R-squared:  -0.0009698
## F-statistic: 0.5621 on 1 and 451 DF,  p-value: 0.4538
##
##
## Call:
## lm(formula = formula, data = mtDataF2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7488 -0.7750  0.1987  0.2250  2.2512
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.80130    0.05128  54.628   <2e-16 ***
## GI          -0.02626    0.04974  -0.528    0.598
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7242 on 459 degrees of freedom
##   (139 observations deleted due to missingness)
## Multiple R-squared:  0.0006068,  Adjusted R-squared:  -0.001571
## F-statistic: 0.2787 on 1 and 459 DF,  p-value: 0.5978
##
##
## Call:
## lm(formula = formula, data = mtDataF2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6568 -0.6441  0.3432  0.3559  0.3685
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.65676    0.03410  19.259   <2e-16 ***
## GI          -0.01263    0.03300  -0.383    0.702
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4789 on 454 degrees of freedom
##   (144 observations deleted due to missingness)
## Multiple R-squared:  0.0003225,  Adjusted R-squared:  -0.001879
## F-statistic: 0.1465 on 1 and 454 DF,  p-value: 0.7021
##
##
## Call:
## lm(formula = formula, data = mtDataF2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.75723 -0.71932 -0.05265  0.61401  2.94735
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.42390    0.07153  61.850  < 2e-16 ***
## GI           0.29542    0.06941   4.256 2.53e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.003 on 450 degrees of freedom
##   (148 observations deleted due to missingness)
## Multiple R-squared:  0.0387, Adjusted R-squared:  0.03656
## F-statistic: 18.12 on 1 and 450 DF,  p-value: 2.532e-05
##
##
## Call:
## lm(formula = formula, data = mtDataF2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -1.5732 -0.7235 -0.1647  0.4431  4.7601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.05686    0.06637  30.991  < 2e-16 ***
## GI           0.18299    0.06485   2.822  0.00499 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9239 on 440 degrees of freedom
##   (158 observations deleted due to missingness)
## Multiple R-squared:  0.01778,    Adjusted R-squared:  0.01554
## F-statistic: 7.963 on 1 and 440 DF,  p-value: 0.004991
##
##
## Call:
## lm(formula = formula, data = mtDataF2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8170 -0.7121 -0.0454  0.6212  3.6212
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.37876    0.07335  32.431   <2e-16 ***
## GI           0.10494    0.07166   1.464    0.144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.021 on 439 degrees of freedom
##   (159 observations deleted due to missingness)
## Multiple R-squared:  0.004861,   Adjusted R-squared:  0.002594
## F-statistic: 2.144 on 1 and 439 DF,  p-value: 0.1438
```

```r
# Regressions on GI additive and dominance components
MT_models <- c("Diam1", "Diam2")

for (trait in MT_models) {
  formula <- formula(paste(trait, "~ GIAdd + GIDom"))
  model <- lm(formula, data = mtDataF2Dec)
  print(summary(model))
  ggplot(mtDataF2Dec, aes(x = GIAdd, y = !!sym(trait), color = GIDom)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE) +
    labs(x = "GIAdd", y = trait, color = "GIDom")
}
```

```
##
## Call:
## lm(formula = formula, data = mtDataF2Dec)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.565 -13.992   1.167  13.899  59.899
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   99.384      1.656  60.029   <2e-16 ***
## GIAdd         -1.717      1.656  -1.037    0.300
## GIDom          3.181      2.264   1.405    0.161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.05 on 379 degrees of freedom
##   (218 observations deleted due to missingness)
## Multiple R-squared:  0.006244,   Adjusted R-squared:  0.0009999
## F-statistic: 1.191 on 2 and 379 DF,  p-value: 0.3052
##
##
## Call:
## lm(formula = formula, data = mtDataF2Dec)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.113  -9.044  -0.044   9.956  60.887
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  97.6815     1.2328  79.234   <2e-16 ***
## GIAdd        -0.5685     1.2328  -0.461    0.645
## GIDom         1.3625     1.6867   0.808    0.420
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.48 on 424 degrees of freedom
##   (173 observations deleted due to missingness)
## Multiple R-squared:  0.001642,   Adjusted R-squared:  -0.003067
## F-statistic: 0.3487 on 2 and 424 DF,  p-value: 0.7058
```

Third Step: The final step is select the traits to createa principle component analysis to comparing multiple QTL regions to detect relationships among them.

```
#Load relevant libraries
library(tidyverse)
library(mice)
library(Hmisc)
library(mi)

# PC Analysis
PCTraitsF2dExt <- mtDataF2[,c(5,27,31,32,24,41,30,39,40,37)]
PCTraitsF2dExtTrim <- na.omit(PCTraitsF2dExt)

#created a new data frame, using the impute function to force PCA despite empty rows within data frame
PCTraitsF2dExtImputed <- impute(PCTraitsF2dExt)

#replace all is.na=0
names(PCTraitsF2dExtImputed)
```

```
##  [1] "LatRating0215" "Diam1"         "Surv1216"      "PIN1"
##  [5] "Rhiz0615"      "meanBLvs"      "Infl"          "dDiam21"
##  [9] "dDiam32"       "SOC1"
```

```
PCModelF2dExt <- prcomp(PCTraitsF2dExtImputed, scale.=TRUE)
summary(PCModelF2dExt)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6    PC7
## Standard deviation     1.3051 1.2074 1.1566 1.0671 0.99030 0.93063 0.8786
## Proportion of Variance 0.1703 0.1458 0.1338 0.1139 0.09807 0.08661 0.0772
## Cumulative Proportion  0.1703 0.3161 0.4499 0.5637 0.66182 0.74843 0.8256
##                           PC8     PC9   PC10
## Standard deviation     0.82357 0.79652 0.6565
## Proportion of Variance 0.06783 0.06344 0.0431
## Cumulative Proportion  0.89346 0.95690 1.0000
```

```
PCModelF2dExt$rotation
```

```
##                       PC1        PC2         PC3         PC4         PC5
## LatRating0215  0.04532617 -0.1469728  0.06438522 -0.56807270  0.70554080
## Diam1         -0.32662792 -0.4483884  0.37719264 -0.04151956 -0.14995626
## Surv1216       0.20793032  0.3295264  0.17139143 -0.31994961 -0.36722514
## PIN1          -0.04265211 -0.1873149 -0.47838557 -0.37522329 -0.43744675
## Rhiz0615       0.22052901  0.2099681  0.27719058 -0.54999824 -0.11666677
## meanBLvs       0.29199915 -0.3510365 -0.32625155  0.10528470  0.22000081
## Infl          -0.57639675  0.1472433 -0.05705819 -0.17869401  0.16831332
## dDiam21        0.07369270  0.4968045 -0.49961872  0.02320843  0.17972280
## dDiam32        0.61133958 -0.2024829  0.14909829  0.06528298  0.02742616
## SOC1           0.03015064  0.3960611  0.37009861  0.28652101  0.17354537
##                       PC6         PC7         PC8          PC9         PC10
## LatRating0215 -0.12009423  0.12455109  0.31073554 -0.121335009  0.10242226
## Diam1         -0.10653141 -0.16134328 -0.05824618 -0.584834249 -0.37811730
## Surv1216      -0.60407061  0.43748486 -0.10143715 -0.115375553 -0.04104822
## PIN1          -0.16576778 -0.39003724  0.46066770  0.100033750 -0.01610950
## Rhiz0615       0.36726446 -0.44918577 -0.41701742  0.026161025  0.05565563
## meanBLvs      -0.48098621 -0.27226163 -0.55832537 -0.008834421  0.09113255
## Infl          -0.18044252 -0.02026750 -0.24837270  0.512089915 -0.47505188
## dDiam21        0.11861007 -0.06721491 -0.03487398 -0.525762733 -0.41117885
## dDiam32        0.07860634  0.03806851  0.18441739  0.282351799 -0.66080234
## SOC1          -0.40349777 -0.57676793  0.30822209  0.027945136  0.04228201
```

```
PCPredictF2 <- PCModelF2dExt$x
mtDataF3 <- cbind(mtDataF2, PCPredictF2)

#Checking the column names within the data frame for proper recall
names(mtDataF3)
```

```
##  [1] "Pop"                  "Fam"                  "Num"
##  [4] "X2nd_Data_colec_date" "LatRating0215"        "PShootStatus0215"
##  [7] "LShootStatus0215"     "Rhiz0215"             "Remarks1"
## [10] "Remarks2"             "Collector"            "S1BLvs"
```

```
## [13] "S2BLvs"               "S3BLvs"               "S4Order"
## [16] "S4BrNode"             "S4FlNode"             "S5Order"
## [19] "S5BrNode"             "S5FlNode"             "S6Order"
## [22] "S6BrNode"             "S6FlNode"             "Rhiz0615"
## [25] "CollDate"             "Comments"             "Diam1"
## [28] "Diam2"                "Diam3"                "Infl"
## [31] "Surv1216"             "PIN1"                 "PIN3"
## [34] "PIN_combined"         "LFY"                  "GI"
## [37] "SOC1"                 "GenoComment"          "dDiam21"
## [40] "dDiam32"              "meanBLvs"             "meanOrder"
## [43] "meanBrNode"           "meanFlNode"           "meanBrNumber"
## [46] "PC1"                  "PC2"                  "PC3"
## [49] "PC4"                  "PC5"                  "PC6"
## [52] "PC7"                  "PC8"                  "PC9"
## [55] "PC10"
```

```
PC1_PIN <- lm(PC1 ~ PIN_combined, data = mtDataF3)
summary(PC1_PIN)
```

```
##
## Call:
## lm(formula = PC1 ~ PIN_combined, data = mtDataF3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9481 -0.8695 -0.0164  0.8265  3.7561
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.45442    0.09925   4.578 5.98e-06 ***
## PIN_combined -0.47841    0.07922  -6.039 3.12e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.231 on 479 degrees of freedom
##   (119 observations deleted due to missingness)
## Multiple R-squared:  0.07075,    Adjusted R-squared:  0.06881
## F-statistic: 36.47 on 1 and 479 DF,  p-value: 3.116e-09
```

```
PC1_SOC1 <- lm(PC1 ~ SOC1, data = mtDataF3)
summary(PC1_SOC1)
```

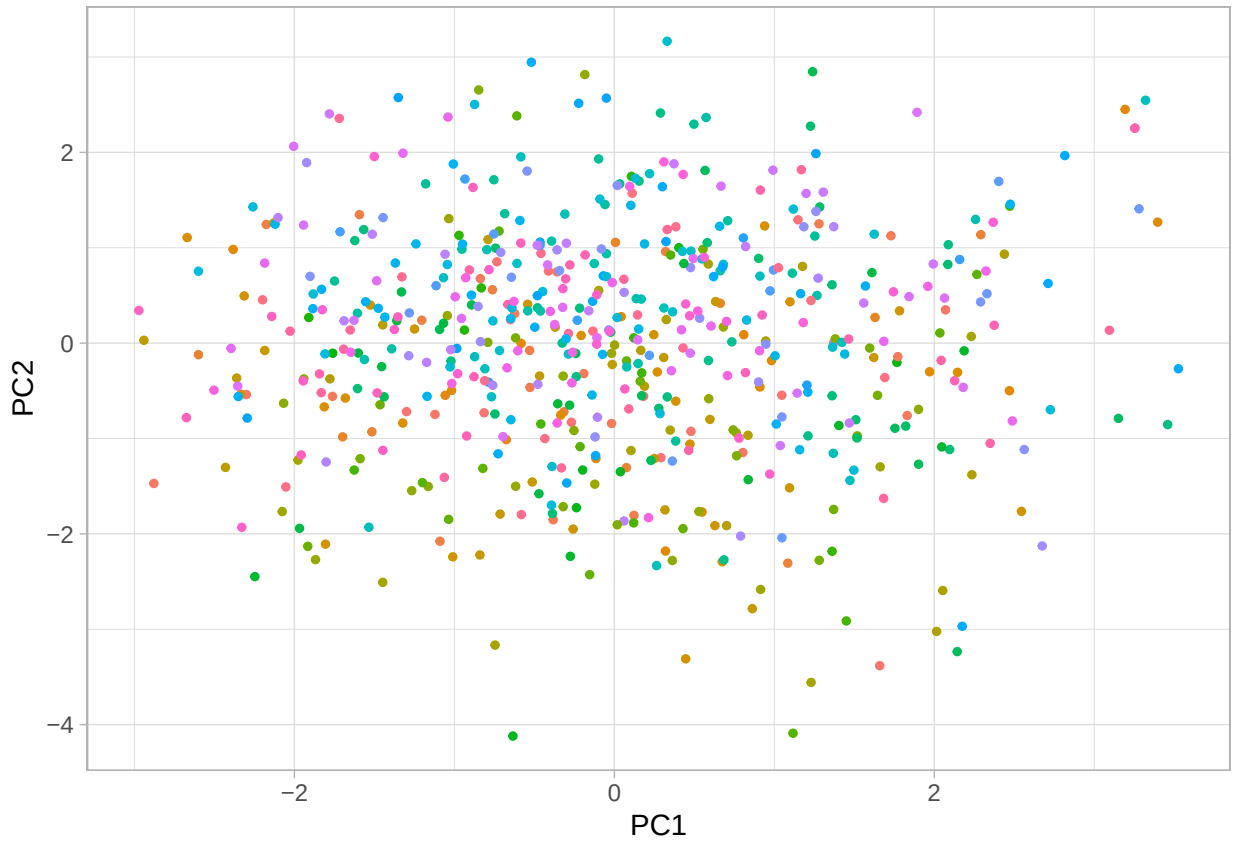
```
##
## Call:
## lm(formula = PC1 ~ SOC1, data = mtDataF3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8470 -0.9323 -0.0212  0.7710  2.6384
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.16306    0.20125  -0.810    0.419
```

```
## SOC1          0.06939    0.17621   0.394    0.694
##
## Residual standard error: 1.274 on 113 degrees of freedom
##   (485 observations deleted due to missingness)
## Multiple R-squared:  0.001371,   Adjusted R-squared:  -0.007467
## F-statistic: 0.1551 on 1 and 113 DF,  p-value: 0.6945
```

```
# Create a data frame of the principal components
PCDataF2 <- data.frame(PCModelF2dExt$x[,1:2])

# Add the sample names to the data frame
PCDataF2$Sample <- rownames(PCDataF2)

# Plot the principal components
ggplot(PCDataF2, aes(x=PC1, y=PC2, color=Sample)) +
  geom_point(size=1) +
  theme_light() +
  theme(legend.position="none")
```



Conclusion The final part of the script calculates linear models of some of the regions being genotyped. It also generates some principal component analysis among QTLs, with outputs of summary tables showing the relationships between traits and genotype. In the last block of code, a summary plot showing the distribution of the trait values in relationship to each other.