R_assignment_LF

R Markdown

Part I

Data Inspection

First, I am going to check if I am in the right working directory:

```
getwd()
```

```
## [1] "/Users/leila/Desktop/R_assignment/R_assignment"
```

Next, load the tidyverse package:

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------ tidyverse 1.2.1 --
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## -- Conflicts --------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Import data files and view them

```
library(RCurl)
```

```
## Loading required package: bitops
##
## Attaching package: 'RCurl'
## The following object is masked from 'package:tidyr':
##
##     complete
```

```
x <- getURL('https://raw.githubusercontent.com/EEOB-BioData/BCB546X-Fall2019/master/assignments/UNIX_Ass
fang_genotypes <- read.delim(text = x)
y <- getURL('https://raw.githubusercontent.com/EEOB-BioData/BCB546X-Fall2019/master/assignments/UNIX_Ass
snp_position <- read.delim(text = y)
```

Inspect files by using head() and tail() (I will not demonstrate here as it will take lots of space in the RMarkdown file). We can also determine the number of columns and rows using ncol() and nrow(), respectively. Use names() to list the headers in both files and class() to determine the class of the files.

```
ncol(fang_genotypes)
```

```
## [1] 986
```

```
ncol(snp_position)
```

```
## [1] 15
```

```
nrow(fang_genotypes)
```

```
## [1] 2782
```

```
nrow(snp_position)
```

```
## [1] 983
```

```
names(fang_genotypes)
```

```
##   [1] "Sample_ID"      "JG_OTU"         "Group"          "abph1.20"
##   [5] "abph1.22"       "ae1.3"          "ae1.4"          "ae1.5"
##   [9] "an1.4"          "ba1.6"          "ba1.9"          "bt2.5"
##  [13] "bt2.7"          "bt2.8"          "Fea2.1"         "Fea2.5"
##  [17] "id1.3"          "lg2.11"         "lg2.2"          "pbf1.1"
##  [21] "pbf1.2"         "pbf1.3"         "pbf1.5"         "pbf1.6"
##  [25] "pbf1.7"         "pbf1.8"         "PZA00003.11"    "PZA00004.2"
##  [29] "PZA00005.8"     "PZA00005.9"     "PZA00006.13"    "PZA00006.14"
##  [33] "PZA00008.1"     "PZA00010.5"     "PZA00013.10"    "PZA00013.11"
##  [37] "PZA00013.9"     "PZA00015.4"     "PZA00017.1"     "PZA00018.5"
##  [41] "PZA00029.11"    "PZA00029.12"    "PZA00030.11"    "PZA00031.5"
##  [45] "PZA00041.3"     "PZA00042.2"     "PZA00042.5"     "PZA00043.7"
##  [49] "PZA00045.1"     "PZA00047.2"     "PZA00049.12"    "PZA00050.9"
##  [53] "PZA00051.2"     "PZA00058.5"     "PZA00058.6"     "PZA00060.2"
##  [57] "PZA00061.1"     "PZA00065.2"     "PZA00069.4"     "PZA00070.5"
##  [61] "PZA00078.2"     "PZA00079.1"     "PZA00081.17"    "PZA00084.2"
##  [65] "PZA00084.3"     "PZA00086.8"     "PZA00088.3"     "PZA00090.2"
##  [69] "PZA00092.1"     "PZA00092.5"     "PZA00093.2"     "PZA00096.26"
##  [73] "PZA00097.13"    "PZA00098.14"    "PZA00100.10"    "PZA00100.12"
##  [77] "PZA00100.14"    "PZA00100.9"     "PZA00103.20"    "PZA00106.9"
##  [81] "PZA00107.18"    "PZA00108.12"    "PZA00108.14"    "PZA00108.15"
##  [85] "PZA00109.3"     "PZA00109.5"     "PZA00111.2"     "PZA00111.4"
##  [89] "PZA00111.5"     "PZA00111.6"     "PZA00111.8"     "PZA00114.3"
##  [93] "PZA00116.2"     "PZA00119.4"     "PZA00120.4"     "PZA00123.1"
##  [97] "PZA00125.2"     "PZA00131.14"    "PZA00132.17"    "PZA00132.18"
## [101] "PZA00132.3"     "PZA00135.6"     "PZA00137.2"     "PZA00139.14"
## [105] "PZA00140.10"    "PZA00140.6"     "PZA00140.9"     "PZA00142.6"
## [109] "PZA00148.2"     "PZA00153.3"     "PZA00153.6"     "PZA00163.4"
## [113] "PZA00164.1"     "PZA00164.2"     "PZA00164.3"     "PZA00166.1"
## [117] "PZA00166.3"     "PZA00170.1"     "PZA00170.3"     "PZA00170.4"
## [121] "PZA00174.1"     "PZA00174.2"     "PZA00175.2"     "PZA00176.8"
## [125] "PZA00177.4"     "PZA00178.3"     "PZA00182.3"     "PZA00182.4"
## [129] "PZA00184.1"     "PZA00184.4"     "PZA00188.1"     "PZA00188.3"
## [133] "PZA00191.5"     "PZA00192.6"     "PZA00192.7"     "PZA00193.2"
## [137] "PZA00198.39"    "PZA00200.11"    "PZA00200.17"    "PZA00200.9"
## [141] "PZA00201.2"     "PZA00204.1"     "PZA00210.1"     "PZA00210.6"
## [145] "PZA00211.7"     "PZA00212.1"     "PZA00213.19"    "PZA00214.1"
## [149] "PZA00216.9"     "PZA00218.1"     "PZA00218.6"     "PZA00219.7"
## [153] "PZA00220.11"    "PZA00220.12"    "PZA00221.7"     "PZA00225.8"
## [157] "PZA00226.7"     "PZA00227.8"     "PZA00230.5"     "PZA00232.24"
## [161] "PZA00234.21"    "PZA00235.6"     "PZA00235.8"     "PZA00237.2"
## [165] "PZA00237.7"     "PZA00237.8"     "PZA00238.3"     "PZA00240.9"
```

```
## [169] "PZA00241.6"     "PZA00243.27"    "PZA00245.14"    "PZA00245.16"
## [173] "PZA00245.17"    "PZA00245.18"    "PZA00245.19"    "PZA00249.2"
## [177] "PZA00250.1"     "PZA00251.1"     "PZA00254.3"     "PZA00255.15"
## [181] "PZA00255.17"    "PZA00256.16"    "PZA00256.21"    "PZA00256.23"
## [185] "PZA00257.11"    "PZA00257.22"    "PZA00261.6"     "PZA00263.14"
## [189] "PZA00266.5"     "PZA00270.3"     "PZA00273.1"     "PZA00274.7"
## [193] "PZA00277.17"    "PZA00277.9"     "PZA00280.14"    "PZA00287.1"
## [197] "PZA00289.11"    "PZA00294.20"    "PZA00296.6"     "PZA00297.2"
## [201] "PZA00297.3"     "PZA00297.4"     "PZA00298.4"     "PZA00298.5"
## [205] "PZA00299.2"     "PZA00300.12"    "PZA00300.13"    "PZA00300.14"
## [209] "PZA00300.16"    "PZA00301.3"     "PZA00303.19"    "PZA00303.21"
## [213] "PZA00307.12"    "PZA00307.14"    "PZA00307.17"    "PZA00309.2"
## [217] "PZA00310.5"     "PZA00314.6"     "PZA00314.8"     "PZA00315.1"
## [221] "PZA00315.6"     "PZA00318.2"     "PZA00323.3"     "PZA00323.4"
## [225] "PZA00326.16"    "PZA00326.18"    "PZA00326.19"    "PZA00332.8"
## [229] "PZA00332.9"     "PZA00334.2"     "PZA00335.12"    "PZA00337.3"
## [233] "PZA00337.4"     "PZA00337.5"     "PZA00342.9"     "PZA00344.10"
## [237] "PZA00345.15"    "PZA00346.1"     "PZA00346.2"     "PZA00346.3"
## [241] "PZA00349.3"     "PZA00349.5"     "PZA00350.2"     "PZA00352.22"
## [245] "PZA00355.1"     "PZA00355.2"     "PZA00356.9"     "PZA00364.5"
## [249] "PZA00364.6"     "PZA00367.2"     "PZA00369.1"     "PZA00370.1"
## [253] "PZA00370.5"     "PZA00380.5"     "PZA00380.7"     "PZA00381.3"
## [257] "PZA00381.4"     "PZA00381.5"     "PZA00382.17"    "PZA00385.3"
## [261] "PZA00386.3"     "PZA00390.6"     "PZA00391.2"     "PZA00392.3"
## [265] "PZA00392.4"     "PZA00393.1"     "PZA00393.4"     "PZA00394.11"
## [269] "PZA00395.1"     "PZA00395.2"     "PZA00396.12"    "PZA00401.11"
## [273] "PZA00401.6"     "PZA00406.1"     "PZA00407.9"     "PZA00408.7"
## [277] "PZA00409.3"     "PZA00411.1"     "PZA00411.4"     "PZA00411.5"
## [281] "PZA00413.17"    "PZA00413.18"    "PZA00413.21"    "PZA00417.2"
## [285] "PZA00417.3"     "PZA00419.1"     "PZA00420.4"     "PZA00422.2"
## [289] "PZA00422.5"     "PZA00422.6"     "PZA00423.16"    "PZA00423.17"
## [293] "PZA00424.1"     "PZA00425.4"     "PZA00425.9"     "PZA00429.1"
## [297] "PZA00433.5"     "PZA00436.7"     "PZA00439.6"     "PZA00440.1"
## [301] "PZA00442.3"     "PZA00442.4"     "PZA00442.5"     "PZA00442.6"
## [305] "PZA00444.1"     "PZA00444.5"     "PZA00445.18"    "PZA00449.2"
## [309] "PZA00452.4"     "PZA00458.6"     "PZA00459.5"     "PZA00460.3"
## [313] "PZA00460.5"     "PZA00460.7"     "PZA00462.2"     "PZA00463.3"
## [317] "PZA00466.1"     "PZA00468.11"    "PZA00468.7"     "PZA00470.1"
## [321] "PZA00471.2"     "PZA00471.3"     "PZA00471.4"     "PZA00472.2"
## [325] "PZA00477.10"    "PZA00477.11"    "PZA00477.5"     "PZA00477.9"
## [329] "PZA00478.10"    "PZA00478.11"    "PZA00478.7"     "PZA00478.9"
## [333] "PZA00480.10"    "PZA00481.7"     "PZA00484.5"     "PZA00485.2"
## [337] "PZA00486.2"     "PZA00487.16"    "PZA00487.24"    "PZA00487.26"
## [341] "PZA00489.1"     "PZA00493.1"     "PZA00493.2"     "PZA00493.5"
## [345] "PZA00495.3"     "PZA00495.4"     "PZA00495.6"     "PZA00496.1"
## [349] "PZA00497.1"     "PZA00497.4"     "PZA00498.4"     "PZA00499.10"
## [353] "PZA00499.12"    "PZA00499.3"     "PZA00501.12"    "PZA00501.14"
## [357] "PZA00502.5"     "PZA00503.5"     "PZA00504.1"     "PZA00504.2"
## [361] "PZA00505.4"     "PZA00505.8"     "PZA00510.2"     "PZA00510.3"
## [365] "PZA00514.1"     "PZA00514.6"     "PZA00514.7"     "PZA00515.14"
## [369] "PZA00516.3"     "PZA00517.6"     "PZA00522.12"    "PZA00523.2"
## [373] "PZA00525.16"    "PZA00525.2"     "PZA00527.6"     "PZA00527.9"
## [377] "PZA00529.3"     "PZA00531.1"     "PZA00533.3"     "PZA00533.4"
## [381] "PZA00533.5"     "PZA00533.6"     "PZA00534.2"     "PZA00536.2"
```

```
## [385] "PZA00538.12"    "PZA00538.16"    "PZA00538.8"     "PZA00543.2"
## [389] "PZA00543.4"     "PZA00543.5"     "PZA00545.21"    "PZA00545.22"
## [393] "PZA00545.4"     "PZA00547.13"    "PZA00547.18"    "PZA00552.4"
## [397] "PZA00560.1"     "PZA00560.2"     "PZA00562.4"     "PZA00565.3"
## [401] "PZA00566.5"     "PZA00568.19"    "PZA00573.3"     "PZA00578.1"
## [405] "PZA00579.6"     "PZA00582.4"     "PZA00586.1"     "PZA00587.3"
## [409] "PZA00587.6"     "PZA00588.2"     "PZA00588.4"     "PZA00589.10"
## [413] "PZA00589.8"     "PZA00589.9"     "PZA00593.2"     "PZA00595.3"
## [417] "PZA00600.11"    "PZA00603.1"     "PZA00608.1"     "PZA00608.5"
## [421] "PZA00610.18"    "PZA00610.9"     "PZA00613.22"    "PZA00614.12"
## [425] "PZA00615.3"     "PZA00615.6"     "PZA00615.8"     "PZA00617.16"
## [429] "PZA00618.22"    "PZA00620.2"     "PZA00621.2"     "PZA00622.1"
## [433] "PZA00622.2"     "PZA00623.2"     "PZA00626.3"     "PZA00626.4"
## [437] "PZA00630.9"     "PZA00636.5"     "PZA00636.6"     "PZA00637.4"
## [441] "PZA00639.12"    "PZA00639.13"    "PZA00639.15"    "PZA00641.7"
## [445] "PZA00641.8"     "PZA00644.11"    "PZA00647.9"     "PZA00650.8"
## [449] "PZA00654.10"    "PZA00654.12"    "PZA00655.1"     "PZA00656.15"
## [453] "PZA00656.16"    "PZA00656.18"    "PZA00656.4"     "PZA00658.19"
## [457] "PZA00658.23"    "PZA00662.3"     "PZA00665.6"     "PZA00667.1"
## [461] "PZA00672.6"     "PZA00672.8"     "PZA00673.2"     "PZA00674.3"
## [465] "PZA00676.2"     "PZA00680.1"     "PZA00680.3"     "PZA00682.2"
## [469] "PZA00684.12"    "PZA00686.8"     "PZA00692.5"     "PZA00693.3"
## [473] "PZA00695.1"     "PZA00698.4"     "PZA00700.3"     "PZA00704.11"
## [477] "PZA00705.5"     "PZA00706.16"    "PZA00710.1"     "PZA00710.16"
## [481] "PZA00712.4"     "PZA00715.3"     "PZA00717.14"    "PZA00719.1"
## [485] "PZA00719.2"     "PZA00719.3"     "PZA00720.2"     "PZA00720.3"
## [489] "PZA00721.4"     "PZA00721.5"     "PZA00725.4"     "PZA00726.6"
## [493] "PZA00726.7"     "PZA00726.9"     "PZA00727.11"    "PZA00727.12"
## [497] "PZA00729.18"    "PZA00729.19"    "PZA00730.2"     "PZA00731.6"
## [501] "PZA00731.7"     "PZA01104.1"     "PZA01149.1"     "PZA01149.3"
## [505] "PZA01182.1"     "PZA01240.1"     "PZA01240.2"     "PZA01420.1"
## [509] "PZA01420.2"     "PZA01420.3"     "PZA01474.2"     "PZA01637.2"
## [513] "PZA01637.3"     "PZA01637.4"     "PZA01725.1"     "PZA01725.2"
## [517] "PZA01782.2"     "PZA01782.3"     "PZA01782.4"     "PZA02789.31"
## [521] "PZA02789.36"    "PZA02791.6"     "PZA02792.16"    "PZA02792.9"
## [525] "PZA02806.4"     "PZA02806.9"     "PZA02807.5"     "PZA02808.12"
## [529] "PZA02808.16"    "PZA02819.35"    "PZA02820.6"     "PZA02822.2"
## [533] "PZA02824.1"     "PZA02824.3"     "PZA02825.8"     "PZA02831.5"
## [537] "PZA02837.5"     "PZA02844.1"     "PZA02850.18"    "PZA02850.4"
## [541] "PZA02853.10"    "PZA02853.7"     "PZA02856.1"     "PZA02862.3"
## [545] "PZA02865.11"    "PZA02869.2"     "PZA02869.8"     "PZA02872.1"
## [549] "PZA02872.3"     "PZA02878.12"    "PZA02888.3"     "PZA02890.3"
## [553] "PZA02890.4"     "PZA02890.5"     "PZA02894.1"     "PZA02897.12"
## [557] "PZA02906.12"    "PZA02906.7"     "PZA02921.9"     "PZA02923.7"
## [561] "PZA02927.1"     "PZA02938.5"     "PZA02939.6"     "PZA02940.3"
## [565] "PZA02941.3"     "PZA02941.6"     "PZA02941.8"     "PZA02947.2"
## [569] "PZA02948.19"    "PZA02948.21"    "PZA02948.22"    "PZA02949.22"
## [573] "PZA02949.26"    "PZA02952.10"    "PZA02954.2"     "PZA02955.3"
## [577] "PZA02958.17"    "PZA02959.7"     "PZA02961.1"     "PZA02962.13"
## [581] "PZA02963.5"     "PZA02966.11"    "PZA02968.4"     "PZA02969.11"
## [585] "PZA02970.9"     "PZA02972.1"     "PZA02982.5"     "PZA02982.6"
## [589] "PZA02983.38"    "PZA02984.7"     "PZA02988.2"     "PZA02993.5"
## [593] "PZA02997.16"    "PZA02997.19"    "PZA03001.15"    "PZA03001.18"
## [597] "PZA03001.9"     "PZA03009.5"     "PZA03009.6"     "PZA03009.7"
```

```
## [601] "PZA03009.8"     "PZA03011.6"     "PZA03012.10"    "PZA03013.7"
## [605] "PZA03013.8"     "PZA03014.10"    "PZA03014.21"    "PZA03014.24"
## [609] "PZA03017.10"    "PZA03017.11"    "PZA03024.16"    "PZA03024.18"
## [613] "PZA03024.7"     "PZA03028.5"     "PZA03032.16"    "PZA03034.1"
## [617] "PZA03035.5"     "PZA03037.8"     "PZA03037.9"     "PZA03041.8"
## [621] "PZA03042.1"     "PZA03042.5"     "PZA03046.2"     "PZA03046.3"
## [625] "PZA03047.12"    "PZA03047.20"    "PZA03047.22"    "PZA03048.16"
## [629] "PZA03048.17"    "PZA03049.23"    "PZA03051.1"     "PZA03051.3"
## [633] "PZA03052.15"    "PZA03054.3"     "PZA03054.5"     "PZA03058.17"
## [637] "PZA03062.15"    "PZA03062.7"     "PZA03063.17"    "PZA03063.18"
## [641] "PZA03064.6"     "PZA03067.17"    "PZA03067.20"    "PZA03068.11"
## [645] "PZA03068.13"    "PZA03069.6"     "PZA03073.23"    "PZA03073.24"
## [649] "PZA03074.24"    "PZA03078.33"    "PZA03081.1"     "PZA03081.10"
## [653] "PZA03081.11"    "PZA03081.13"    "PZA03081.6"     "PZA03083.7"
## [657] "PZA03089.12"    "PZA03090.31"    "PZA03092.7"     "PZA03094.18"
## [661] "PZA03094.6"     "PZA03095.1"     "PZA03095.2"     "PZA03095.3"
## [665] "PZA03097.4"     "PZA03097.7"     "PZA03097.9"     "PZA03102.10"
## [669] "PZA03102.2"     "PZA03102.9"     "PZA03137.1"     "PZA03172.2"
## [673] "PZA03223.3"     "PZA03258.2"     "PZA03283.2"     "PZA03284.3"
## [677] "PZA03290.1"     "PZA03290.2"     "PZA03295.4"     "PZA03296.6"
## [681] "PZA03296.7"     "PZA03298.1"     "PZA03298.2"     "PZA03301.2"
## [685] "PZA03301.4"     "PZA03302.1"     "PZA03305.6"     "PZA03305.7"
## [689] "PZA03311.2"     "PZA03311.3"     "PZA03311.4"     "PZA03311.5"
## [693] "PZA03312.1"     "PZA03312.2"     "PZA03316.2"     "PZA03317.1"
## [697] "PZA03319.3"     "PZA03319.4"     "PZA03320.3"     "PZA03320.4"
## [701] "PZA03328.5"     "PZA03329.1"     "PZA03329.2"     "PZA03333.3"
## [705] "PZA03335.2"     "PZA03335.3"     "PZA03337.1"     "PZA03338.5"
## [709] "PZA03340.2"     "PZA03342.2"     "PZA03344.4"     "PZA03344.5"
## [713] "PZA03344.6"     "PZA03345.1"     "PZA03345.2"     "PZA03345.4"
## [717] "PZA03347.1"     "PZA03348.1"     "PZA03349.1"     "PZA03349.9"
## [721] "PZA03767.1"     "PZA03767.4"     "PZA03767.5"     "PZA03773.2"
## [725] "PZA03773.3"     "PZA03774.1"     "PZA03774.10"    "PZA03774.2"
## [729] "PZA03774.4"     "PZA03774.5"     "PZA03774.6"     "PZA03774.8"
## [733] "PZA03774.9"     "PZA03775.1"     "PZA03775.11"    "PZA03775.2"
## [737] "PZA03775.3"     "PZA03775.4"     "PZA03775.6"     "PZA03775.7"
## [741] "PZA03775.8"     "PZA03775.9"     "PZA03781.1"     "PZA03781.2"
## [745] "PZA03781.3"     "PZA03781.4"     "PZA03781.5"     "PZA03781.6"
## [749] "PZA03781.7"     "PZA03781.8"     "PZA03782.1"     "PZA03782.3"
## [753] "PZA03786.1"     "PZA03786.2"     "PZA03789.1"     "PZA03789.2"
## [757] "PZA03789.4"     "PZB00011.4"     "PZB00011.5"     "PZB00041.2"
## [761] "PZB00041.4"     "PZB00049.2"     "PZB00049.4"     "PZB00049.7"
## [765] "PZB00055.1"     "PZB00060.4"     "PZB00062.6"     "PZB00062.7"
## [769] "PZB00062.8"     "PZB00067.2"     "PZB00067.3"     "PZB00067.4"
## [773] "PZB00067.5"     "PZB00078.1"     "PZB00081.2"     "PZB00081.4"
## [777] "PZB00081.5"     "PZB00081.7"     "PZB00092.1"     "PZB00092.4"
## [781] "PZB00093.3"     "PZB00093.4"     "PZB00093.6"     "PZB00096.2"
## [785] "PZB00096.3"     "PZB00136.3"     "PZB00140.1"     "PZB00145.2"
## [789] "PZB00149.2"     "PZB00149.4"     "PZB00153.1"     "PZB00153.2"
## [793] "PZB00153.3"     "PZB00153.5"     "PZB00160.1"     "PZB00160.2"
## [797] "PZB00160.4"     "PZB00165.2"     "PZB00165.6"     "PZB00169.4"
## [801] "PZB00169.6"     "PZB00175.1"     "PZB00175.2"     "PZB00175.3"
## [805] "PZB00175.4"     "PZB00175.5"     "PZB00180.1"     "PZB00180.2"
## [809] "PZB00183.3"     "PZB00188.6"     "PZB00207.3"     "PZB00221.3"
## [813] "PZB00221.8"     "PZB00229.3"     "PZB00232.1"     "PZB00232.2"
```

```
## [817] "PZB00232.4"     "PZB00232.5"     "PZB00379.3"     "PZB00379.4"
## [821] "PZB00379.5"     "PZB00393.7"     "PZB00409.3"     "PZB00416.2"
## [825] "PZB00416.5"     "PZB00454.2"     "PZB00454.3"     "PZB00454.4"
## [829] "PZB00454.5"     "PZB00498.2"     "PZB00498.4"     "PZB00598.1"
## [833] "PZB00598.2"     "PZB00603.3"     "PZB00603.4"     "PZB00603.5"
## [837] "PZB00607.2"     "PZB00761.1"     "PZB00761.2"     "PZB00849.2"
## [841] "PZB00849.3"     "PZB00849.4"     "PZB00859.1"     "PZB01109.2"
## [845] "PZB01109.3"     "PZB01110.1"     "PZB01110.2"     "PZB01110.3"
## [849] "PZB01111.6"     "PZB01111.7"     "PZB01111.8"     "PZB01112.3"
## [853] "PZB01112.4"     "PZB01112.5"     "PZB01112.6"     "PZB01113.4"
## [857] "PZB01114.1"     "PZB01114.3"     "PZB01115.1"     "PZB01115.5"
## [861] "PZB01115.6"     "PZB01116.2"     "PZB01221.1"     "PZB01222.1"
## [865] "PZB01222.3"     "PZB01223.3"     "PZB01223.4"     "PZB01223.7"
## [869] "PZB01225.1"     "PZB01225.2"     "PZB01225.4"     "PZB01228.1"
## [873] "PZB01228.3"     "PZB01228.4"     "PZB01233.2"     "PZB01233.3"
## [877] "PZB01238.5"     "PZB01238.6"     "PZB01427.1"     "PZB01427.3"
## [881] "PZB01463.2"     "PZB01463.3"     "PZB01463.4"     "PZD00003.1"
## [885] "PZD00003.3"     "PZD00007.1"     "PZD00008.3"     "PZD00011.1"
## [889] "PZD00011.3"     "PZD00011.4"     "PZD00012.1"     "PZD00012.2"
## [893] "PZD00012.3"     "PZD00012.4"     "PZD00012.5"     "PZD00013.3"
## [897] "PZD00013.4"     "PZD00014.3"     "PZD00017.1"     "PZD00019.1"
## [901] "PZD00020.2"     "PZD00020.3"     "PZD00020.4"     "PZD00020.6"
## [905] "PZD00021.2"     "PZD00021.4"     "PZD00021.5"     "PZD00022.1"
## [909] "PZD00022.3"     "PZD00022.4"     "PZD00024.2"     "PZD00025.1"
## [913] "PZD00025.2"     "PZD00030.1"     "PZD00030.4"     "PZD00030.5"
## [917] "PZD00030.6"     "PZD00034.3"     "PZD00043.1"     "PZD00043.2"
## [921] "PZD00043.3"     "PZD00043.4"     "PZD00044.2"     "PZD00044.3"
## [925] "PZD00044.4"     "PZD00045.1"     "PZD00045.2"     "PZD00045.3"
## [929] "PZD00045.4"     "PZD00049.3"     "PZD00049.4"     "PZD00049.5"
## [933] "PZD00051.1"     "PZD00052.3"     "PZD00052.4"     "PZD00062.2"
## [937] "PZD00066.1"     "PZD00067.1"     "PZD00067.2"     "PZD00067.3"
## [941] "PZD00068.1"     "PZD00069.2"     "PZD00069.3"     "PZD00069.4"
## [945] "PZD00069.5"     "PZD00073.1"     "PZD00073.2"     "PZD00073.6"
## [949] "PZD00074.1"     "PZD00075.1"     "PZD00075.2"     "PZD00076.1"
## [953] "PZD00076.2"     "PZD00076.4"     "PZD00077.10"    "PZD00077.5"
## [957] "PZD00077.7"     "PZD00077.8"     "PZD00078.2"     "Ra2_ORF.1"
## [961] "Ra2_ORF.2"      "Ra2_ORF.4"      "Ra2_promoter.1" "Ra2_promoter.2"
## [965] "Ra2_promoter.3" "sh2.5"          "sh2.6"          "sh2.7"
## [969] "sh2.9"          "su1.4"          "su1.5"          "su1.7"
## [973] "tb1.17"         "tb1.18"         "tb1.19"         "tb1.5"
## [977] "te1.3"          "te1.4"          "zagl1.1"        "zagl1.6"
## [981] "zap1.2"         "zen1.1"         "zen1.2"         "zen1.4"
## [985] "zfl2.6"         "zmm3.4"
```

```
names(snp_position)
```

```
##  [1] "SNP_ID"              "cdv_marker_id"        "Chromosome"
##  [4] "Position"            "alt_pos"              "mult_positions"
##  [7] "amplicon"            "cdv_map_feature.name" "gene"
## [10] "candidate.random"    "Genaissance_daa_id"   "Sequenom_daa_id"
## [13] "count_amplicons"     "count_cmf"            "count_gene"
```

```
class(fang_genotypes)
```

```
## [1] "data.frame"
```

```
class(snp_position)
```

```
## [1] "data.frame"
```

fang_genoytypes has 986 columns with 2782 rows (exluding the header), and snp_positions has 15 columns and 983 rows (excluding the header). Both files are data frames.

We can use summary() to provide us with statistics on the columns of the data frame. Note that I did not apply it on fang_genotypes since the number of columns is too large and the summary is not very useful in this case.

```
summary(snp_position)
```

```
##       SNP_ID    cdv_marker_id     Chromosome       Position
##  abph1.20:  1   Min.   : 3463   1      :155   unknown  : 27
##  abph1.22:  1   1st Qu.: 3978   2      :127   multiple : 11
##  ae1.3   :  1   Median : 5723   5      :122            :  6
##  ae1.4   :  1   Mean   : 5925   3      :107   100227859:  1
##  ae1.5   :  1   3rd Qu.: 6629   7      : 97   10069039 :  1
##  an1.4   :  1   Max.   :12480   4      : 91   102663486:  1
##  (Other) :977                   (Other):284   (Other)  :936
##                                                                    alt_pos
##                                                                        :952
##  Approximate: 5 prime BLAST start of amplicon plus 150 bases               :  1
##  Approximate: 5 prime BLAST start of amplicon plus 50 bases                :  6
##  Approximate: position of another SNP from the same amplicon plus 100 bases:  8
##  Approximate: position of another SNP from the same amplicon plus 150 bases:  3
##  Approximate: position of another SNP from the same amplicon plus 200 bases:  3
##  Approximate: position of another SNP from the same amplicon plus 50 bases : 10
##                                                mult_positions
##                                                       :966
##  Chr1(275076660);Chr8(71078729);                      :  1
##  Chr1(9691660);Chr2(225530627);                       :  1
##  Chr2(17581865);Chr3(143082611);Chr6(158386495);      :  1
##  Chr2(209957515;210039173);                           :  1
##  Chr4(106637739;106644768;157652727);Chr8(111504317);:  1
##  (Other)                                              : 12
##      amplicon                cdv_map_feature.name        gene
##  PZA03775:  9   NULL                     : 25       zmm28    : 11
##  PZA03774:  8   zmm28                   : 11       PZA03450 :  9
##  PZA03781:  8   AY104805                : 10       AY104805 :  8
##  pbf1    :  7   CL6994_-2               :  8       CL6994_-2:  8
##  PZA00111:  5   zag1                    :  8       PZA03455 :  8
##  PZA00245:  5   Grain Wt. QTL (McCouch):  7       ra1      :  8
##  (Other) :941   (Other)                 :914       (Other)  :931
##   candidate.random Genaissance_daa_id Sequenom_daa_id count_amplicons
##  candidate:339     Min.   : 7649      Min.   :10474   Min.   :0.0000
##  random   :644     1st Qu.: 7906      1st Qu.:10784   1st Qu.:0.0000
##                    Median : 8173      Median :11110   Median :1.0000
##                    Mean   : 8524      Mean   :11122   Mean   :0.5768
##                    3rd Qu.: 9834      3rd Qu.:11420   3rd Qu.:1.0000
##                    Max.   :10104      Max.   :11829   Max.   :1.0000
##
##    count_cmf        count_gene
```

```
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :1.0000
##  Mean   :0.5483   Mean   :0.5565
##  3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000
##
```

Data Processing

First, create a table containing only the "SNP_ID", "Chromosome", and "Position" columns:

```
snps<-snp_position[c("SNP_ID","Chromosome", "Position")]
view(snps)
```

Next, use filter() to extract the rows containing the genotypes, and dim() to check the dimensions of the resulting data frames:

```
maize <- filter(fang_genotypes, `Group` == "ZMMIL" | `Group` == "ZMMLR" | `Group` == "ZMMMR")
teosinte <- filter(fang_genotypes, `Group` == "ZMPBA" | `Group` == "ZMPIL" | `Group` == "ZMPJA")
dim(maize)
```

```
## [1] 1573  986
```

```
dim(teosinte)
```

```
## [1] 975 986
```

So, the total maize genotypes extracted are 1573, while those of teosinte are 975.

We can count the number of groups present in fang_genotypes using the following to check if the previous dimensions make sense by adding the ones corresponding to each maize or teosinte:

```
fang_genotypes %>% group_by(Group) %>% summarize(count=n())
```

```
## # A tibble: 16 x 2
##    Group count
##    <fct> <int>
##  1 TRIPS    22
##  2 ZDIPL    15
##  3 ZLUXR    17
##  4 ZMHUE    10
##  5 ZMMIL   290
##  6 ZMMLR  1256
##  7 ZMMMR    27
##  8 ZMPBA   900
##  9 ZMPIL    41
## 10 ZMPJA    34
## 11 ZMXCH    75
## 12 ZMXCP    69
## 13 ZMXIL     6
## 14 ZMXNO     7
## 15 ZMXNT     4
## 16 ZPERR     9
```

To transpose the maize and teosinte data frames, use t()

```
maize_t <- t(maize)
teosinte_t <- t(teosinte)
dim(maize_t)
```

```
## [1]  986 1573
```

```
dim(teosinte_t)
```

```
## [1] 986 975
```

By looking at the dimensions of the data frames, we can confirm that they have been transposed successfully.

Use merge() to join the snps data frame with maize/teosinte genotypes:

```
?merge()
maize_snps <- merge(snps, maize_t, by.x = 1, by.y = 0, sort = TRUE)
teosinte_snps <- merge(snps, teosinte_t, by.x = 1, by.y = 0, sort = TRUE)
```

Sorting by increasing SNP position values with missing data encoded by "?" (not changed since missing data is already encoded by "?")

```
for (i in 1:10){
maize_temp <- filter(maize_snps, Chromosome == i )
maize_increasing <- arrange(maize_temp, Position)
write.table(maize_increasing, file = file.path("./Maize/Maize_increasing/", paste0("maize_chromosome_in
}
```

```
for (i in 1:10){
teosinte_temp <- filter(teosinte_snps, Chromosome == i)
teosinte_increasing <- arrange(teosinte_temp, Position)
write.table(teosinte_increasing, file = file.path("./Teosinte/Teosinte_increasing/", paste0("teosinte_c
}
```

Sorting by decreasing SNP position values with missing data encoded by "-"

```
for (i in 1:10){
maize_tempo <- filter(maize_snps, Chromosome == i)
maize_tempor <- arrange(maize_tempo, desc(Position))
maize_decreasing <- sapply(maize_tempor, gsub, pattern = "?", replacement = "-", fixed = TRUE)
write.table(maize_decreasing, file = file.path("./Maize/Maize_decreasing/", paste0('maize_chromosome_de
}
```

```
for (i in 1:10){
teosinte_tempo <- filter(teosinte_snps, Chromosome == i)
teosinte_tempor <- arrange(teosinte_tempo, desc(Position))
teosinte_decreasing <- sapply(teosinte_tempor, gsub, pattern = "?", replacement = "-", fixed = TRUE)
write.table(teosinte_decreasing, file = file.path("./Teosinte/Teosinte_decreasing/", paste0('teosinte_c
}
```

Part II

Data Visualization

SNPs per Chromosome

First, load the required packages

```
library(ggplot2)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

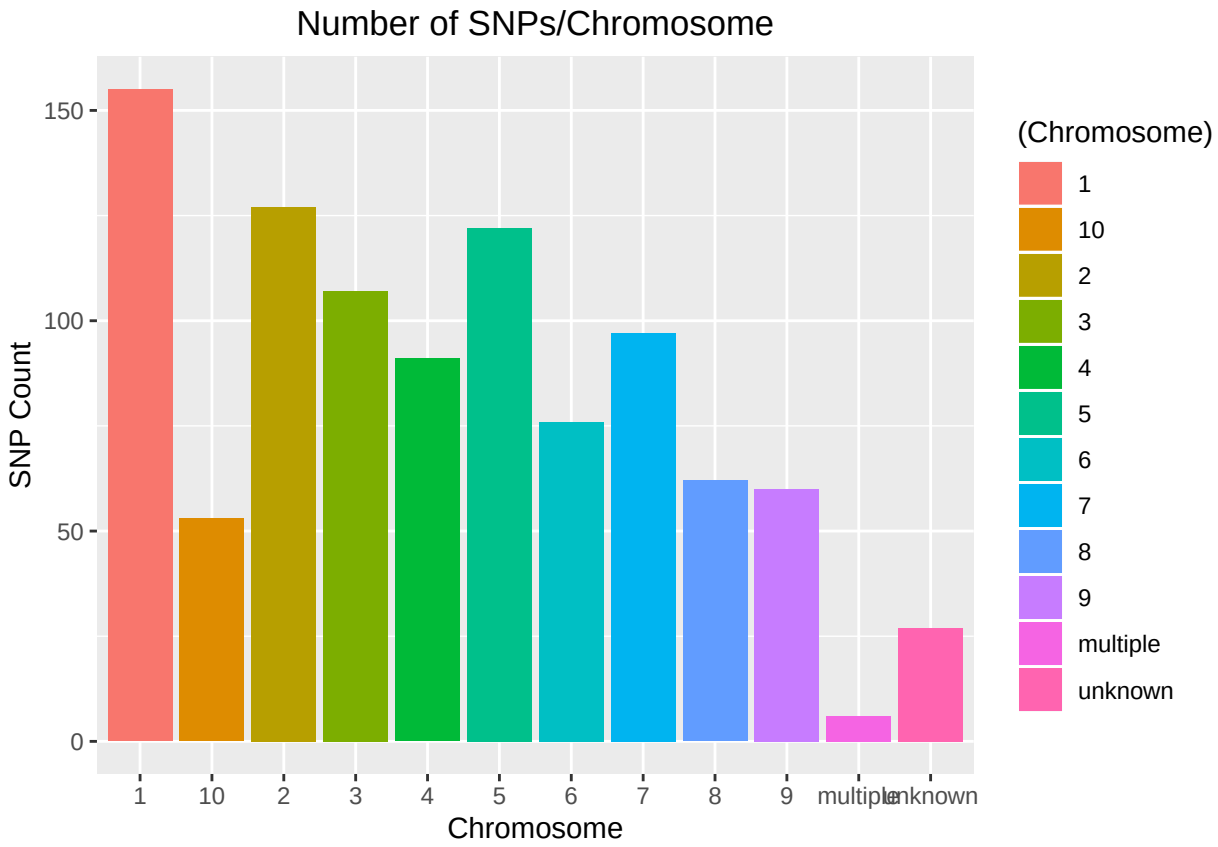
```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

Next, tidy data using pivot_longer

```
genotypes <- t(fang_genotypes)
genotypes_snps <- merge(snps, genotypes, by.x = 1, by.y = 0, sort = TRUE)
```

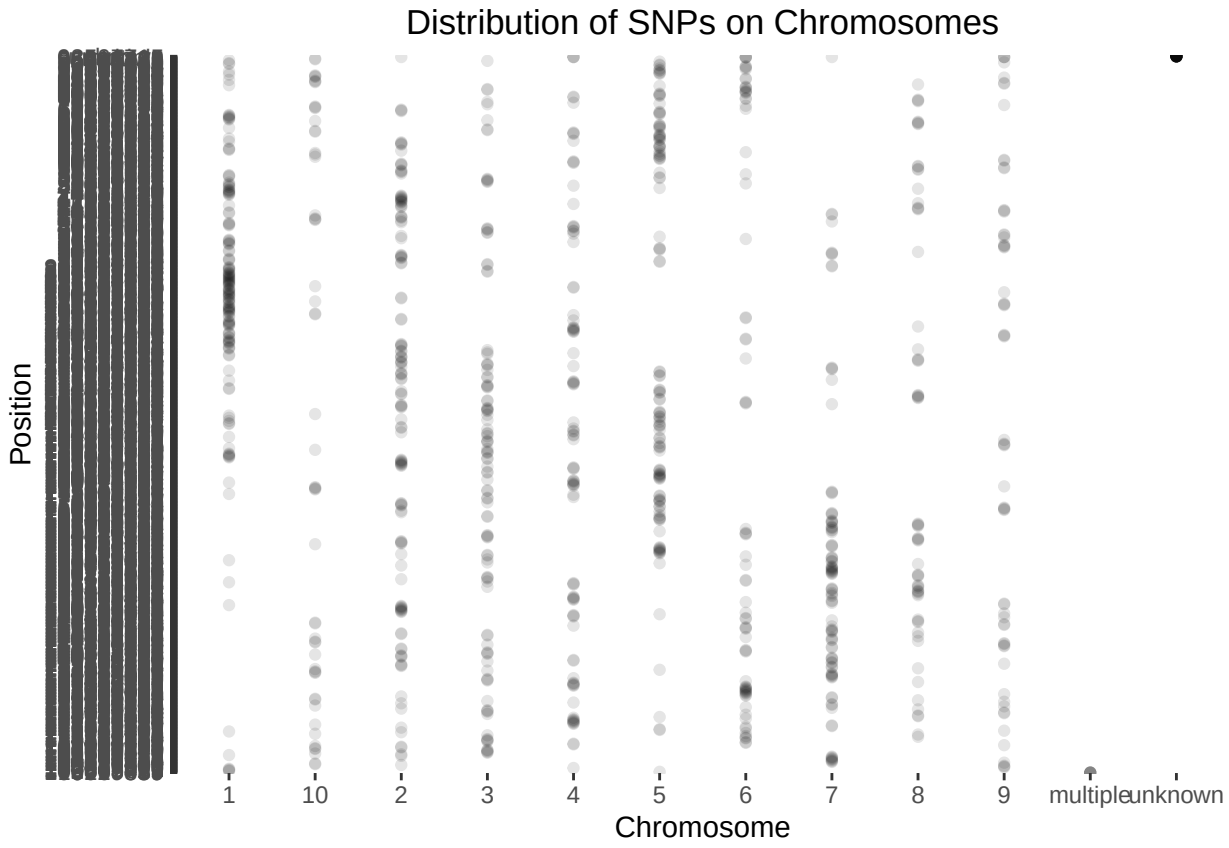
Plotting the total number of SNPs on each chromosome

```
ggplot(genotypes_snps) + geom_bar(aes(x=Chromosome, fill=(Chromosome))) + ggtitle("Number of SNPs/Chrom
```



Plotting the distribution of SNPs on chromosomes

```
ggplot(data = genotypes_snps, mapping=aes(x=Chromosome, y=Position))+
     geom_point(alpha=0.1) + ggtitle("Distribution of SNPs on Chromosomes") + theme(plot.title = elemen
```



Missing Data and Amount of Heterozygosity

```
melted_genotypes <- melt(fang_genotypes, id = c("Sample_ID", "Group"))
```

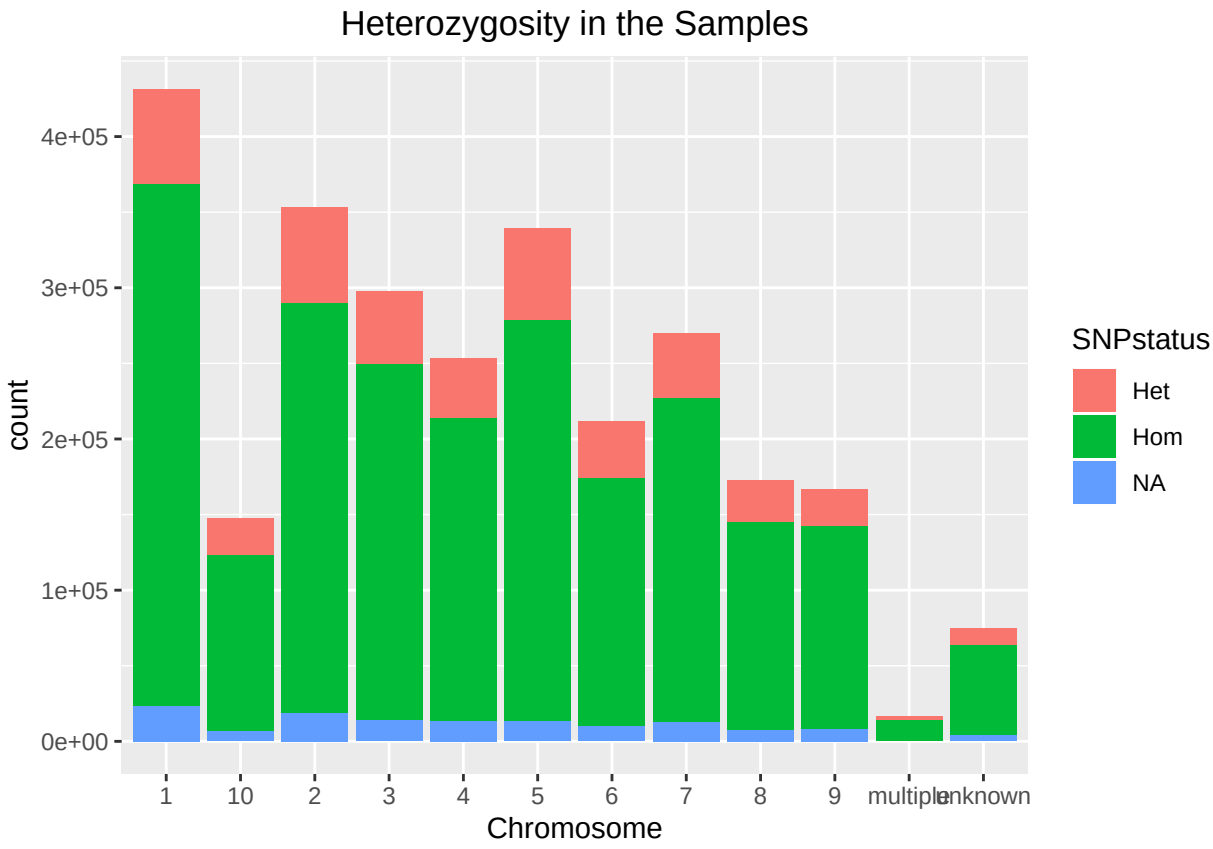
```
## Warning: attributes are not identical across measure variables; they will
## be dropped
```

```
melted_snps <- melt(snps, id = c("SNP_ID", "Chromosome"))
colnames(melted_genotypes) [3:4] <- c("SNP_ID", "SNPname")
geno_snps <- merge(melted_snps, melted_genotypes, by.x = 1, by.y = 3)
geno_snps <- geno_snps[,-3]
```

```
geno_snps$SNPstatus <- "NA"
geno_snps$SNPstatus <- geno_snps$SNPname
geno_snps$SNPstatus[geno_snps$SNPname=="?/?"] <- "NA"
geno_snps$SNPstatus[geno_snps$SNPname=="A/A" | geno_snps$SNPname=="C/C" | geno_snps$SNPname=="G/G" | ge
geno_snps$SNPstatus [geno_snps$SNPstatus!="Hom" & geno_snps$SNPstatus!= "NA"] <- "Het"
heterozygosity<-geno_snps
```

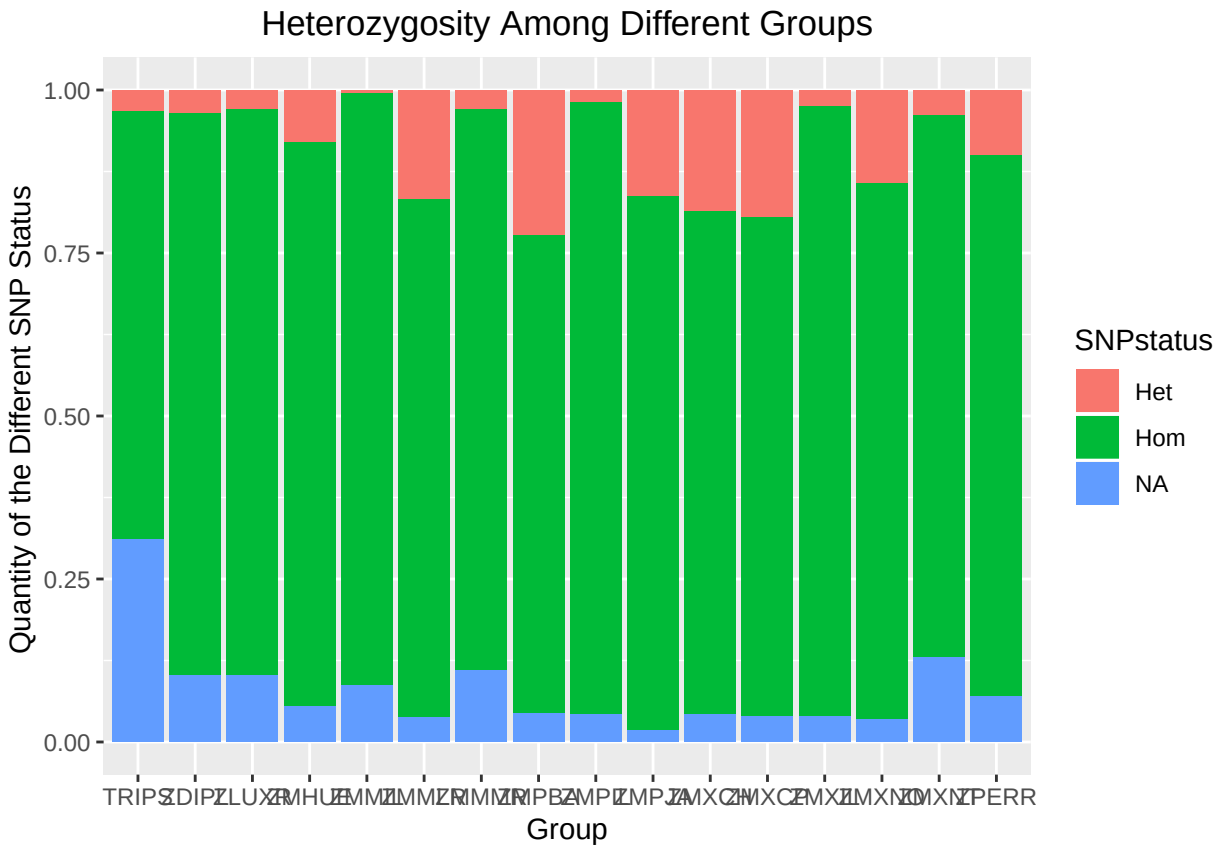
Plotting Heterozygosity in the Chromosomes

```
ggplot(data = heterozygosity) +
  geom_bar(mapping = aes(x =Chromosome, fill=SNPstatus)) + ggtitle("Heterozygosity in the Samples") + t
```



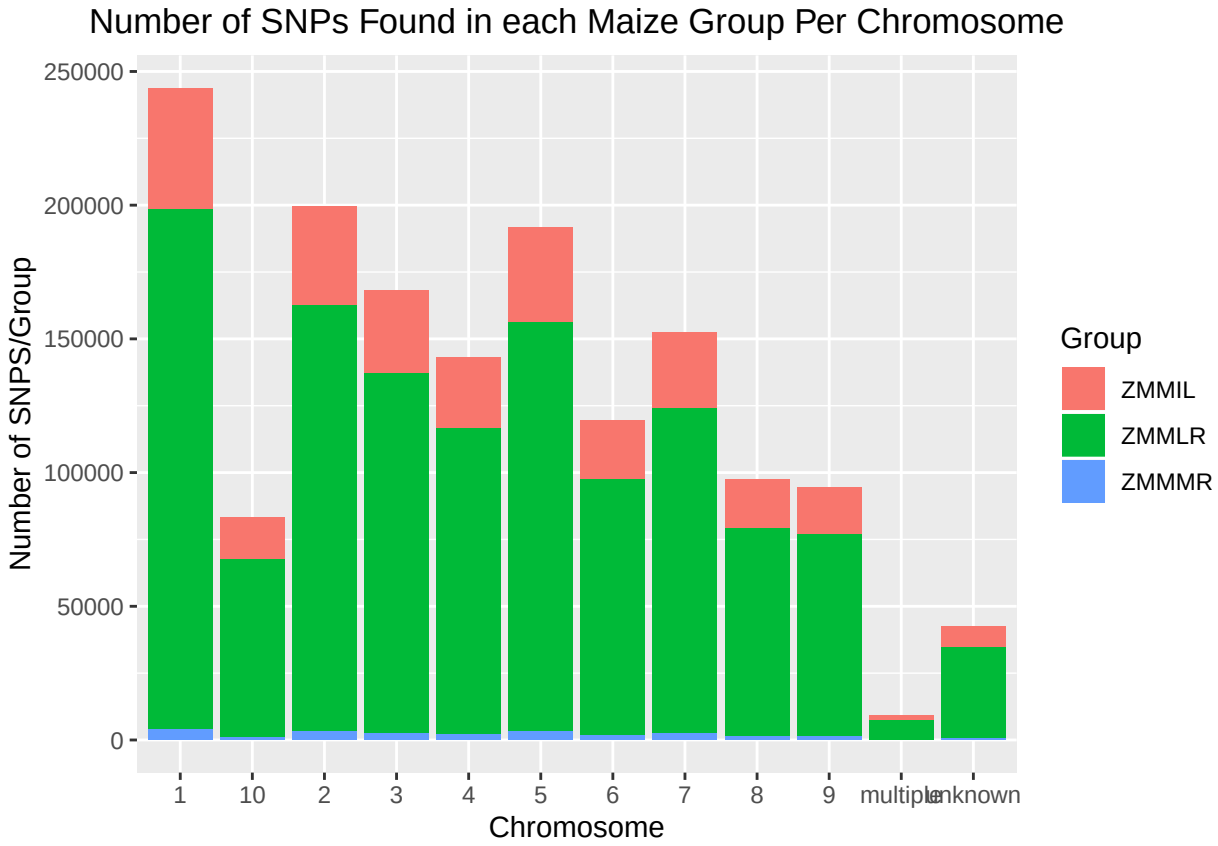
Plotting Heterozygosity Among Different Groups

```
ggplot(heterozygosity) + geom_bar(aes(x=Group, fill=SNPstatus), position = "fill") + ggtitle("Heterozyg
```



My Own Visualization

```
maize_geno_snps <- filter(geno_snps, `Group` == "ZMMIL" | `Group` == "ZMMLR" | `Group` == "ZMMMR")
ggplot(maize_geno_snps) + geom_bar(aes(x=Chromosome, fill=Group)) + ggtitle("Number of SNPs Found in ea
```


Number of SNPs Found in each Maize Group Per Chromosome
Number of SNPS/Group
250000
200000
150000
100000
50000
0
1
10
2
3
4
5
6
7
8
9
multiple
unknown
Chromosome
Group
ZMMIL
ZMMLR
ZMMMR