Day 2 – 01 Explore Data (Exercises – Solutions)

Seminar practice worksheet

Use this key after attempting `scripts/01_explore_data_exercises.Rmd`. The code chunks mirror the TODOs but include one possible solution for each task.

1. Load and preview the dataset

```
input_path <- file.path('..', 'data', 'dataset2_subset_long.csv')
long_df <- read.csv(input_path, stringsAsFactors = FALSE, check.names = FALSE)
cat('Rows:', nrow(long_df), '\\nColumns:', ncol(long_df), '\\n')
```

```
## Rows: 3072 \nColumns: 7 \n
```

```
head(long_df)
```

```
##                            Genome      snp_id Position     value mouse_id day
## 1 Akkermansia_muciniphila_YL44   239840-C-G   239840 0.000000     1683   0
## 2 Akkermansia_muciniphila_YL44   241793-A-G   241793 0.049587     1683   0
## 3 Akkermansia_muciniphila_YL44   355328-A-T   355328 0.138182     1683   0
## 4 Akkermansia_muciniphila_YL44   356291-C-A   356291 0.000000     1683   0
## 5 Akkermansia_muciniphila_YL44 2351445-C-T  2351445 0.000000     1683   0
## 6   Bacteroides_caecimuris_I48 1601848-T-C  1601848 0.041609     1683   0
##   treatment_group
## 1         Control
## 2         Control
## 3         Control
## 4         Control
## 5         Control
## 6         Control
```

2. Enumerate genomes and SNPs

```
unique(long_df$Genome)
```

```
## [1] "Akkermansia_muciniphila_YL44" "Bacteroides_caecimuris_I48"
## [3] "Turicimonas_muris_YL45"
```

```
table(long_df$Genome)
```

```
##
## Akkermansia_muciniphila_YL44   Bacteroides_caecimuris_I48
##                          320                         1216
##       Turicimonas_muris_YL45
##                         1536
```

```
tapply(long_df$snp_id, long_df$Genome, function(x) length(unique(x)))
```

```
## Akkermansia_muciniphila_YL44    Bacteroides_caecimuris_I48
##                             5                            19
##        Turicimonas_muris_YL45
##                            24
```

3. Summaries by genome

```
sample_values <- long_df$value[1:10]
summary(sample_values)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.03426 0.04759 0.13818
```

```
aggregate(value ~ Genome, data = long_df, function(x) summary(x))
```

```
##                           Genome value.Min. value.1st Qu. value.Median value.Mean
## 1 Akkermansia_muciniphila_YL44  0.0000000     0.0000000    0.0523090  0.1741893
## 2   Bacteroides_caecimuris_I48  0.0000000     0.0000000    0.0338305  0.1485411
## 3        Turicimonas_muris_YL45  0.0000000     0.0000000    0.2649125  0.4312851
##   value.3rd Qu. value.Max.
## 1    0.2714463  0.8712450
## 2    0.2679232  0.9912180
## 3    0.9090910  1.0000000
```

4. Focus on Turicimonas

```
turicimonas_df <- long_df[long_df$Genome == 'Turicimonas_muris_YL45', ]
cat('Rows for Turicimonas:', nrow(turicimonas_df), '\\n')
```

```
## Rows for Turicimonas: 1536 \n
```

```
head(turicimonas_df)
```

```
##                     Genome       snp_id Position    value mouse_id day
## 25 Turicimonas_muris_YL45  362534-G-A   362534 1.000000     1683   0
## 26 Turicimonas_muris_YL45 1063346-G-A  1063346 0.741935     1683   0
## 27 Turicimonas_muris_YL45 1135479-G-A  1135479 0.000000     1683   0
## 28 Turicimonas_muris_YL45 1364256-G-A  1364256 0.406250     1683   0
## 29 Turicimonas_muris_YL45 1376518-G-A  1376518 0.000000     1683   0
## 30 Turicimonas_muris_YL45 1392383-G-A  1392383 0.461538     1683   0
##    treatment_group
## 25         Control
## 26         Control
## 27         Control
## 28         Control
## 29         Control
## 30         Control
```

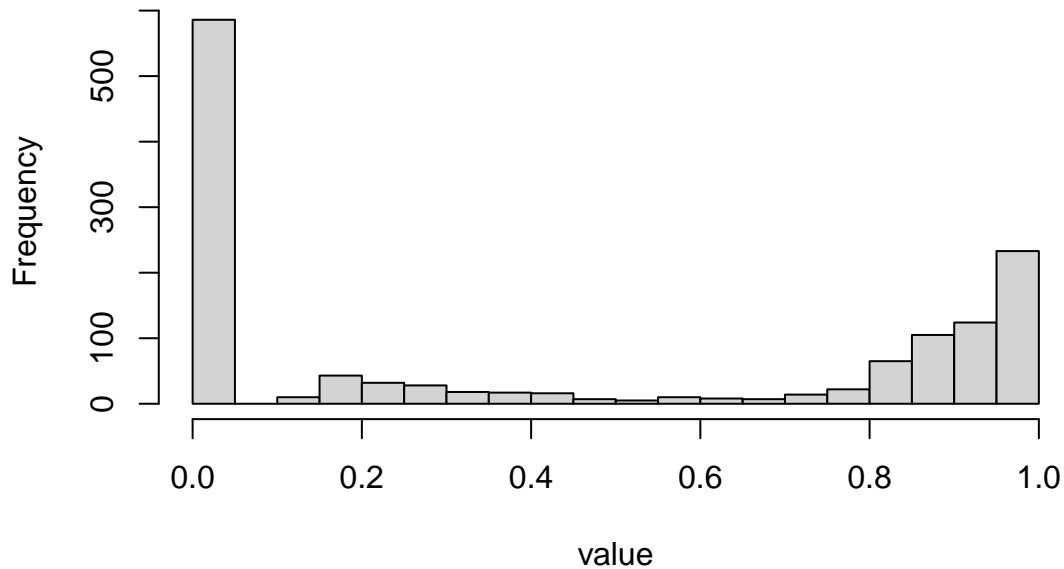```
table(turicimonas_df$mouse_id, turicimonas_df$day)
```

```
##
##          0   4   9  14  18  23  30  37  44  49  53  58  63  67  72  79
##    1683 24  24  24  24  24  24  24  24  24  24  24  24  24  24  24  24
##    1688 24  24  24  24  24  24  24  24  24  24  24  24  24  24  24  24
##    1692 24  24  24  24  24  24  24  24  24  24  24  24  24  24  24  24
##    1699 24  24  24  24  24  24  24  24  24  24  24  24  24  24  24  24
```

5. Allele-frequency focus (Turicimonas vs all genomes)

```
turicimonas_values <- long_df$value[long_df$Genome == 'Turicimonas_muris_YL45']
hist(turicimonas_values, breaks = 20, main = 'Turicimonas AF', xlab = 'value')
```

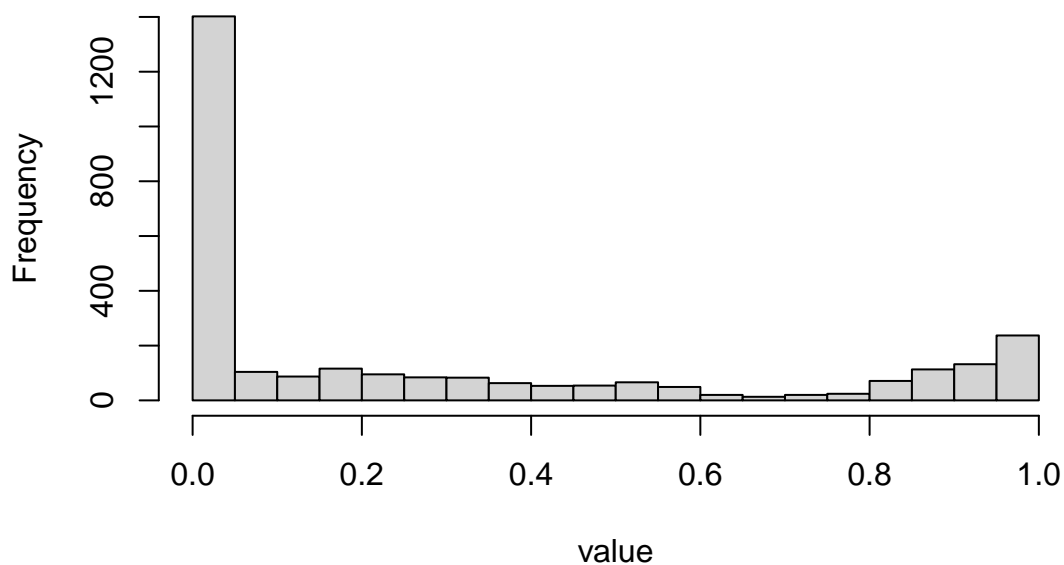**Turicimonas AF**


```
summary(turicimonas_values)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.0000  0.0000  0.2649  0.4313  0.9091  1.0000     186
```

```
hist(long_df$value, breaks = 30, main = 'All genomes AF', xlab = 'value')
```

All genomes AF

```r
summary(long_df$value)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00000 0.00000 0.06438 0.28365 0.51995 1.00000     186
```

6. Missing values

```r
na_total <- sum(is.na(long_df$value))
na_total
```

```
## [1] 186
```

```r
na_by_mouse_day <- with(long_df, tapply(value, list(mouse_id, day), function(x) sum(is.na(x))))
na_by_mouse_day
```

```
##          0  4  9 14 18 23 30 37 44 49 53 58 63 67 72 79
## 1683     0  1  0  1  3  2  1  0  0  2  0  1  0  0  0  1
## 1688    15  1  0  1  0  0  0  0  0  0  0  0  0  0  0  1
## 1692     0 24 24  0 24 20  6  0  1  8 16  0  0  6  0  0
## 1699     0  0 24  0  1  0  0  0  1  0  0  0  1  0  0  0
```

In this dataset the counts are all zero, so no genome (including Turicimonas) introduces missing allele-frequency entries.

7. Treatment groups

```r
unique(long_df$treatment_group)
```

```
## [1] "Control"       "Ciprofloxacin" "Tetracyclin"   "Vancomycin"
```

```r
table(long_df$treatment_group)
```

```
##
## Ciprofloxacin       Control   Tetracyclin    Vancomycin
##           768           768           768           768
```

```r
table(long_df$mouse_id, long_df$treatment_group)
```

```
##
##        Ciprofloxacin Control Tetracyclin Vancomycin
##    1683             0     768           0          0
##    1688           768       0           0          0
##    1692             0       0         768          0
##    1699             0       0           0        768
```

```r
long_df[long_df$day == 30, c('mouse_id','treatment_group')]
```

```
##     mouse_id treatment_group
## 193     1683         Control
## 194     1683         Control
## 195     1683         Control
## 196     1683         Control
## 197     1683         Control
## 198     1683         Control
## 199     1683         Control
## 200     1683         Control
## 201     1683         Control
```

```
## 202       1683          Control
## 203       1683          Control
## 204       1683          Control
## 205       1683          Control
## 206       1683          Control
## 207       1683          Control
## 208       1683          Control
## 209       1683          Control
## 210       1683          Control
## 211       1683          Control
## 212       1683          Control
## 213       1683          Control
## 214       1683          Control
## 215       1683          Control
## 216       1683          Control
## 217       1683          Control
## 218       1683          Control
## 219       1683          Control
## 220       1683          Control
## 221       1683          Control
## 222       1683          Control
## 223       1683          Control
## 224       1683          Control
## 225       1683          Control
## 226       1683          Control
## 227       1683          Control
## 228       1683          Control
## 229       1683          Control
## 230       1683          Control
## 231       1683          Control
## 232       1683          Control
## 233       1683          Control
## 234       1683          Control
## 235       1683          Control
## 236       1683          Control
## 237       1683          Control
## 238       1683          Control
## 239       1683          Control
## 240       1683          Control
## 961       1688     Ciprofloxacin
## 962       1688     Ciprofloxacin
## 963       1688     Ciprofloxacin
## 964       1688     Ciprofloxacin
## 965       1688     Ciprofloxacin
## 966       1688     Ciprofloxacin
## 967       1688     Ciprofloxacin
## 968       1688     Ciprofloxacin
## 969       1688     Ciprofloxacin
## 970       1688     Ciprofloxacin
## 971       1688     Ciprofloxacin
## 972       1688     Ciprofloxacin
## 973       1688     Ciprofloxacin
## 974       1688     Ciprofloxacin
## 975       1688     Ciprofloxacin
```

```
## 976    1688   Ciprofloxacin
## 977    1688   Ciprofloxacin
## 978    1688   Ciprofloxacin
## 979    1688   Ciprofloxacin
## 980    1688   Ciprofloxacin
## 981    1688   Ciprofloxacin
## 982    1688   Ciprofloxacin
## 983    1688   Ciprofloxacin
## 984    1688   Ciprofloxacin
## 985    1688   Ciprofloxacin
## 986    1688   Ciprofloxacin
## 987    1688   Ciprofloxacin
## 988    1688   Ciprofloxacin
## 989    1688   Ciprofloxacin
## 990    1688   Ciprofloxacin
## 991    1688   Ciprofloxacin
## 992    1688   Ciprofloxacin
## 993    1688   Ciprofloxacin
## 994    1688   Ciprofloxacin
## 995    1688   Ciprofloxacin
## 996    1688   Ciprofloxacin
## 997    1688   Ciprofloxacin
## 998    1688   Ciprofloxacin
## 999    1688   Ciprofloxacin
## 1000   1688   Ciprofloxacin
## 1001   1688   Ciprofloxacin
## 1002   1688   Ciprofloxacin
## 1003   1688   Ciprofloxacin
## 1004   1688   Ciprofloxacin
## 1005   1688   Ciprofloxacin
## 1006   1688   Ciprofloxacin
## 1007   1688   Ciprofloxacin
## 1008   1688   Ciprofloxacin
## 1729   1692     Tetracyclin
## 1730   1692     Tetracyclin
## 1731   1692     Tetracyclin
## 1732   1692     Tetracyclin
## 1733   1692     Tetracyclin
## 1734   1692     Tetracyclin
## 1735   1692     Tetracyclin
## 1736   1692     Tetracyclin
## 1737   1692     Tetracyclin
## 1738   1692     Tetracyclin
## 1739   1692     Tetracyclin
## 1740   1692     Tetracyclin
## 1741   1692     Tetracyclin
## 1742   1692     Tetracyclin
## 1743   1692     Tetracyclin
## 1744   1692     Tetracyclin
## 1745   1692     Tetracyclin
## 1746   1692     Tetracyclin
## 1747   1692     Tetracyclin
## 1748   1692     Tetracyclin
## 1749   1692     Tetracyclin
```

```
## 1750      1692      Tetracyclin
## 1751      1692      Tetracyclin
## 1752      1692      Tetracyclin
## 1753      1692      Tetracyclin
## 1754      1692      Tetracyclin
## 1755      1692      Tetracyclin
## 1756      1692      Tetracyclin
## 1757      1692      Tetracyclin
## 1758      1692      Tetracyclin
## 1759      1692      Tetracyclin
## 1760      1692      Tetracyclin
## 1761      1692      Tetracyclin
## 1762      1692      Tetracyclin
## 1763      1692      Tetracyclin
## 1764      1692      Tetracyclin
## 1765      1692      Tetracyclin
## 1766      1692      Tetracyclin
## 1767      1692      Tetracyclin
## 1768      1692      Tetracyclin
## 1769      1692      Tetracyclin
## 1770      1692      Tetracyclin
## 1771      1692      Tetracyclin
## 1772      1692      Tetracyclin
## 1773      1692      Tetracyclin
## 1774      1692      Tetracyclin
## 1775      1692      Tetracyclin
## 1776      1692      Tetracyclin
## 2497      1699      Vancomycin
## 2498      1699      Vancomycin
## 2499      1699      Vancomycin
## 2500      1699      Vancomycin
## 2501      1699      Vancomycin
## 2502      1699      Vancomycin
## 2503      1699      Vancomycin
## 2504      1699      Vancomycin
## 2505      1699      Vancomycin
## 2506      1699      Vancomycin
## 2507      1699      Vancomycin
## 2508      1699      Vancomycin
## 2509      1699      Vancomycin
## 2510      1699      Vancomycin
## 2511      1699      Vancomycin
## 2512      1699      Vancomycin
## 2513      1699      Vancomycin
## 2514      1699      Vancomycin
## 2515      1699      Vancomycin
## 2516      1699      Vancomycin
## 2517      1699      Vancomycin
## 2518      1699      Vancomycin
## 2519      1699      Vancomycin
## 2520      1699      Vancomycin
## 2521      1699      Vancomycin
## 2522      1699      Vancomycin
## 2523      1699      Vancomycin
```

```
## 2524     1699     Vancomycin
## 2525     1699     Vancomycin
## 2526     1699     Vancomycin
## 2527     1699     Vancomycin
## 2528     1699     Vancomycin
## 2529     1699     Vancomycin
## 2530     1699     Vancomycin
## 2531     1699     Vancomycin
## 2532     1699     Vancomycin
## 2533     1699     Vancomycin
## 2534     1699     Vancomycin
## 2535     1699     Vancomycin
## 2536     1699     Vancomycin
## 2537     1699     Vancomycin
## 2538     1699     Vancomycin
## 2539     1699     Vancomycin
## 2540     1699     Vancomycin
## 2541     1699     Vancomycin
## 2542     1699     Vancomycin
## 2543     1699     Vancomycin
## 2544     1699     Vancomycin
```

8. Stretch idea (optional)

```r
day30 <- long_df[long_df$day == 30, ]
medians_day30 <- tapply(day30$value, day30$Genome, median, na.rm = TRUE)
medians_day30
```

```
## Akkermansia_muciniphila_YL44    Bacteroides_caecimuris_I48
##                     0.124955                      0.030887
##       Turicimonas_muris_YL45
##                     0.250000
```

```r
medians_day30[which.max(medians_day30)]
```

```
## Turicimonas_muris_YL45
##                   0.25
```

Feel free to compare your answers with these outputs, then proceed to `scripts/02_simple_heatmap.Rmd`.