Prepare Phenotype and Covariate Data

Izel Fourie Sørensen, Pernille Merete Sarup, Palle Duun Rohde

April 3, 2017

In this script we prepare a phenotype and covariate data frame to be used in downstream genomic analyses. As an example of phenotype data we are using the phenotype "resistance to starvation" from the *Drosophila melanogaster* Genetic Reference Panel (DGRP). The data is available at http://dgrp2.gnets.ncsu.edu/data.html under "Phenotype files", "Mackay, et al., Nature, 2012". Data for both males and females are used. Inversion status (chromosomal inversions) and Wolbachia (*Wolbachia* infection status) can be found under the heading "Other useful files" at the bottom of the page.

We use the plyr, dplyr and tidyr packages for editing data. The readxl package is used for reading .xlsx files. Install these packages as follows:

```
#install.packages("plyr")
#install.packages("dplyr")
#install.packages("tidyr")
#install.packages("readxl")
```

```
library(plyr)
library(dplyr)
library(tidyr)
library(readxl)
```

Download phenotype and covariate data

```
# Female
download.file("http://dgrp2.gnets.ncsu.edu/data/website/starvation.female.csv",
              destfile = "C:/Users/Izel/Dropbox/qgg-usersguide/data/starvation.female.csv")

# Male
download.file("http://dgrp2.gnets.ncsu.edu/data/website/starvation.male.csv",
              destfile = "C:/Users/Izel/Dropbox/qgg-usersguide/data/starvation.male.csv")

# Inverstion status
download.file("http://dgrp2.gnets.ncsu.edu/data/website/inversion.xlsx", mode = "wb",
              destfile = "C:/Users/Izel/Dropbox/qgg-usersguide/data/inversion.xlsx")

# Wolbachia
download.file("http://dgrp2.gnets.ncsu.edu/data/website/wolbachia.xlsx", mode = "wb",
              destfile = "C:/Users/Izel/Dropbox/qgg-usersguide/data/wolbachia.xlsx")
```

Read and edit phenotype data

Read female data.

```
starF <- read.csv(file="./data/starvation.female_2017.csv", header = FALSE)
head(starF)
```

```
##         V1       V2
## 1 line_100 77.92000
## 2 line_101 57.76000
```

```
## 3 line_105  73.12000
## 4 line_109  53.44000
## 5 line_129  42.77551
## 6 line_136 104.32000
```

```
dim(starF)
```

```
## [1] 203   2
```

Read male data.

```
starM <- read.csv(file="./data/starvation.male_2017.csv", header = FALSE)
head(starM)
```

```
##         V1       V2
## 1 line_100 49.28000
## 2 line_101 47.20000
## 3 line_105 51.04000
## 4 line_109 44.96000
## 5 line_129 33.08475
## 6 line_136 63.04000
```

```
dim(starM)
```

```
## [1] 203   2
```

Give column names. "L" = lines, "F" = female, "M" = male.

```
colnames(starF) <- c("L", "F")
colnames(starM) <- c("L", "M")
```

In dplyr a data frame has to be converted to a tibble (`tbl`). Convert `starF` and `starM` to tibbles.

```
starF <- tbl_df(starF)
starM <- tbl_df(starM)
```

Look at the tibbles

```
starF
```

```
## # A tibble: 203 x 2
##           L         F
##      <fctr>     <dbl>
##  1 line_100  77.92000
##  2 line_101  57.76000
##  3 line_105  73.12000
##  4 line_109  53.44000
##  5 line_129  42.77551
##  6 line_136 104.32000
##  7 line_138  59.52000
##  8 line_142  59.26531
##  9 line_149  47.00000
## 10 line_153  59.04000
## # ... with 193 more rows
```

```
starM
```

```
## # A tibble: 203 x 2
##           L        M
```

```
##      <fctr>    <dbl>
##  1 line_100 49.28000
##  2 line_101 47.20000
##  3 line_105 51.04000
##  4 line_109 44.96000
##  5 line_129 33.08475
##  6 line_136 63.04000
##  7 line_138 47.83673
##  8 line_142 38.40000
##  9 line_149 35.84000
## 10 line_153 40.32000
## # ... with 193 more rows
```

Join the tibbles for males and females.

```
starMF <- left_join(starM, starF, by= "L")
starMF
```

```
## # A tibble: 203 x 3
##           L        M         F
##      <fctr>    <dbl>     <dbl>
##  1 line_100 49.28000  77.92000
##  2 line_101 47.20000  57.76000
##  3 line_105 51.04000  73.12000
##  4 line_109 44.96000  53.44000
##  5 line_129 33.08475  42.77551
##  6 line_136 63.04000 104.32000
##  7 line_138 47.83673  59.52000
##  8 line_142 38.40000  59.26531
##  9 line_149 35.84000  47.00000
## 10 line_153 40.32000  59.04000
## # ... with 193 more rows
```

Create a column for sex information and a column for the phenotype (y), in this case resistance to starvation.

```
starv <- gather(starMF, sex, y, -L)
head(starv)
```

```
## # A tibble: 6 x 3
##          L   sex        y
##     <fctr> <chr>    <dbl>
## 1 line_100     M 49.28000
## 2 line_101     M 47.20000
## 3 line_105     M 51.04000
## 4 line_109     M 44.96000
## 5 line_129     M 33.08475
## 6 line_136     M 63.04000
```

```
head(starv, 3)
```

```
## # A tibble: 3 x 3
##          L   sex     y
##     <fctr> <chr> <dbl>
## 1 line_100     M 49.28
## 2 line_101     M 47.20
```

```
## 3 line_105     M 51.04
```

Remove prefix "line_" from the contents of the "L" column.

```
starv$L <- gsub("line_", "", starv$L, fixed = TRUE)
head(starv$L)
```

```
## [1] "100" "101" "105" "109" "129" "136"
```

Read and edit Inversion status

Abbreviations used are: INV = inversion karyotype (homozygous), INV / ST = heterozygote for the inversion and ST = standard configuration in a homozygous form.

```
inv <- read_excel("C:/Users/Izel/Dropbox/qgg-usersguide/data/inversion.xlsx",
                  sheet = 1, col_names = TRUE)

head(inv)
```

```
## # A tibble: 6 x 17
##   `DGRP Line` `In(2L)t` `In(2R)NS` `In(2R)Y1` `In(2R)Y2` `In(2R)Y3`
##         <chr>     <chr>      <chr>      <chr>      <chr>      <chr>
## 1     DGRP_21        ST         ST         ST         ST         ST
## 2     DGRP_26       INV         ST         ST         ST         ST
## 3     DGRP_28        ST        INV         ST         ST         ST
## 4     DGRP_31        ST         ST         ST         ST         ST
## 5     DGRP_32       INV         ST         ST         ST         ST
## 6     DGRP_38        ST         ST         ST         ST         ST
## # ... with 11 more variables: `In(2R)Y4` <chr>, `In(2R)Y5` <chr>,
## #   `In(2R)Y6` <chr>, `In(2R)Y7` <chr>, `In(3L)P` <chr>, `In(3L)M` <chr>,
## #   `In(3L)Y` <chr>, `In(3R)P` <chr>, `In(3R)K` <chr>, `In(3R)Mo` <chr>,
## #   `In(3R)C` <chr>
```

```
dim(inv)
```

```
## [1] 205  17
```

Remove the "DGRP_" prefix from the contents of the "DGRP Line" column. Save column names of `Inv` as a vector `Invcols`. Then edit the column names in the `Invcols` vector: "DGRP Line" becomes "L" and the brackets "()" in the inversion names are changed to underscores.

```
inv$`DGRP Line` <- gsub("DGRP_", "", inv$`DGRP Line`, fixed=TRUE)
head(inv, 3)
```

```
## # A tibble: 3 x 17
##   `DGRP Line` `In(2L)t` `In(2R)NS` `In(2R)Y1` `In(2R)Y2` `In(2R)Y3`
##         <chr>     <chr>      <chr>      <chr>      <chr>      <chr>
## 1          21        ST         ST         ST         ST         ST
## 2          26       INV         ST         ST         ST         ST
## 3          28        ST        INV         ST         ST         ST
## # ... with 11 more variables: `In(2R)Y4` <chr>, `In(2R)Y5` <chr>,
## #   `In(2R)Y6` <chr>, `In(2R)Y7` <chr>, `In(3L)P` <chr>, `In(3L)M` <chr>,
## #   `In(3L)Y` <chr>, `In(3R)P` <chr>, `In(3R)K` <chr>, `In(3R)Mo` <chr>,
## #   `In(3R)C` <chr>
```

```
invcols <- colnames(inv)
invcols[1] <- "L"
```

```
invcols[2:17] <- gsub("(", "_", invcols[2:17], fixed = TRUE)
invcols[2:17] <- gsub(")", "_", invcols[2:17], fixed = TRUE)
colnames(inv) <- invcols
inv
```

```
## # A tibble: 205 x 17
##        L In_2L_t In_2R_NS In_2R_Y1 In_2R_Y2 In_2R_Y3 In_2R_Y4 In_2R_Y5
##    <chr>   <chr>    <chr>    <chr>    <chr>    <chr>    <chr>    <chr>
##  1    21      ST       ST       ST       ST       ST       ST       ST
##  2    26     INV       ST       ST       ST       ST       ST       ST
##  3    28      ST      INV       ST       ST       ST       ST       ST
##  4    31      ST       ST       ST       ST       ST       ST       ST
##  5    32     INV       ST       ST       ST       ST       ST       ST
##  6    38      ST       ST       ST       ST       ST       ST       ST
##  7    40      ST       ST       ST       ST       ST       ST       ST
##  8    41      ST       ST       ST       ST       ST       ST       ST
##  9    42      ST       ST       ST       ST       ST       ST       ST
## 10    45      ST       ST       ST       ST       ST       ST       ST
## # ... with 195 more rows, and 9 more variables: In_2R_Y6 <chr>,
## #   In_2R_Y7 <chr>, In_3L_P <chr>, In_3L_M <chr>, In_3L_Y <chr>,
## #   In_3R_P <chr>, In_3R_K <chr>, In_3R_Mo <chr>, In_3R_C <chr>
```

Read and edit *Wolbachia* status

```
wo <- read_excel("C:/Users/Izel/Dropbox/qgg-usersguide/data/wolbachia.xlsx",
                  sheet = 1, col_names = TRUE)
wo
```

```
## # A tibble: 205 x 2
##    `DGRP Line` `Infection Status`
##          <chr>              <chr>
##  1    DGRP__21                  y
##  2    DGRP__26                  n
##  3    DGRP__28                  n
##  4    DGRP__31                  n
##  5    DGRP__32                  n
##  6    DGRP__38                  n
##  7    DGRP__40                  y
##  8    DGRP__41                  n
##  9    DGRP__42                  n
## 10    DGRP__45                  n
## # ... with 195 more rows
```

```
dim(wo)
```

```
## [1] 205   2
```

Change column names of `wo`

```
colnames(wo) <- c("L", "wo")
wo$L <- gsub("DGRP__", "", wo$L, fixed=TRUE)
wo
```

```
## # A tibble: 205 x 2
##        L    wo
##    <chr> <chr>
```

```
##  1    21     y
##  2    26     n
##  3    28     n
##  4    31     n
##  5    32     n
##  6    38     n
##  7    40     y
##  8    41     n
##  9    42     n
## 10    45     n
## # ... with 195 more rows
```

Create final data frame

Merge phenotype data with inversion status and *Wolbachia* infection status.

```
starvInv <- left_join(starv, inv, by= "L")
starvInv
```

```
## # A tibble: 406 x 19
##        L   sex        y In_2L_t In_2R_NS In_2R_Y1 In_2R_Y2 In_2R_Y3
##    <chr> <chr>    <dbl>   <chr>    <chr>    <chr>    <chr>    <chr>
##  1   100     M 49.28000  INV/ST       ST       ST       ST       ST
##  2   101     M 47.20000  INV/ST       ST       ST       ST       ST
##  3   105     M 51.04000      ST       ST       ST       ST       ST
##  4   109     M 44.96000  INV/ST       ST       ST       ST       ST
##  5   129     M 33.08475      ST       ST       ST       ST       ST
##  6   136     M 63.04000      ST       ST       ST       ST       ST
##  7   138     M 47.83673      ST       ST       ST       ST       ST
##  8   142     M 38.40000      ST       ST       ST       ST       ST
##  9   149     M 35.84000      ST       ST       ST       ST       ST
## 10   153     M 40.32000      ST       ST       ST       ST       ST
## # ... with 396 more rows, and 11 more variables: In_2R_Y4 <chr>,
## #   In_2R_Y5 <chr>, In_2R_Y6 <chr>, In_2R_Y7 <chr>, In_3L_P <chr>,
## #   In_3L_M <chr>, In_3L_Y <chr>, In_3R_P <chr>, In_3R_K <chr>,
## #   In_3R_Mo <chr>, In_3R_C <chr>
```

```
starvIW <- left_join(starvInv, wo, by="L")
starvIW[1:5,15:20]
```

```
## # A tibble: 5 x 6
##   In_3L_Y In_3R_P In_3R_K In_3R_Mo In_3R_C    wo
##     <chr>   <chr>   <chr>    <chr>   <chr> <chr>
## 1      ST      ST     INV       ST      ST     y
## 2      ST      ST      ST       ST      ST     n
## 3      ST      ST     INV       ST      ST     n
## 4      ST      ST      ST       ST      ST     n
## 5      ST      ST      ST       ST      ST     n
```

An example of how one can look at the data. Here we show a summary (in table form) of the first three inversions.

```
apply(starvIW[,4:19], 2, table)[1:3]
```

```
## $In_2L_t
##
##    INV INV/ST     ST
```

```
##     38     50    318
##
## $In_2R_NS
##
##    INV INV/ST     ST
##     14     20    372
##
## $In_2R_Y1
##
## INV/ST     ST
##      2    404
```

Convert the tibble to a data frame and save the edited phenotype data.

```
starv <- as.data.frame(starvIW)
save(starv, file="./phenotypes/starv_inv_wo.Rdata")
```