Reviewing the half-clean data with a Hardy-Weinberg and Structure Analysis

Grace Saville

19/04/2022

Preamble

```
library(reshape)
library()
getwd()
setwd("C:/Users/airhe/OneDrive/Documents/Masters/Project 3/kiore-project")
```

Prepping df for HWE Analysis

I would like to remove the SNPs not in Hardy-Weinberg equilibrium, therefore I need to reformat the data for input into HWE program.

```
# load(".RData") # if necessary
data <- read.csv("./data/RStudio/ratsSNPs_halfclean.csv")
```

```
copy <- data # making a copy
copy[1,1:20] # checking column names
copy <- copy[,-c(2:16)] # removing all but specimen names and SNPs
copy[1,1:20] # checking

copy[copy == "?"] <- "?:?" # replacing single ? with double ? so alleles can be split

x <- data.frame(island = copy$island) # setting up new df for for loop
coln <- as.vector(colnames(copy)) # prepping to paste the column names into the for loop
dim(copy) # 379 rows 283 columns
for (i in 2:283) {
  y <- colsplit(copy[,i], split = ":", names = c(coln[i], paste("blank", i, sep = "."))) # splitting ea
  x <- cbind(x, y) # combining output with current df
  rm(i, y) # removing temp objects
}

# Checking:
# dim(x3) # 379 rows 565 columns
# x2[1:5,1:5]
# x3[1:5,1:5] # comparing the 2 dfs to check the column naming worked correctly
```

```
copy <- x
rm(x, coln) # removing excess objects
```

Producing the file necessary for PGDSpider program

```
copy <- copy[order(copy$island, decreasing = FALSE), ] # ordering df alphabetically by island
print(as.matrix(copy[, 1])) # printing the island names and row numbers

# A=1, T=2, G=3, C=4
copy[copy == "A"] <- "1"
copy[copy == "T"] <- "2"
copy[copy == "G"] <- "3"
copy[copy == "C"] <- "4"

# row numbers in dataset df listed below for each popn.
popnames <- as.character(
  c(
    "pop = Aotea", # 1:10
    "pop = Borneo", # 11:28
    "pop = Doubtful_Sound", # 315
    "pop = Great_Mercury_Island", # 30
    "pop = Halmahera", # 31:42
    "pop = Hatutaa", # 43:63
    "pop = Honuea", # 64:83
    "pop = Kaikura_Island", # 84:103
    "pop = Kamaka", # 104:124
    "pop = Kayangel", # 125:145
    "pop = Late_Island", # 148:168
    "pop = Mainland", # 29, 146, 147, 169, 358, 359 (including Luzon here)
    "pop = Malenge", # 170:181
    "pop = Mohotani", # 182:195
    "pop = Motukawanui", # 196:216
    "pop = New_Britain", # 217:226
    "pop = New_Guinea", # 227:229
    "pop = Normanby_Island", # 230
    "pop = Rakiura", # 231:251
    "pop = Reiono", # 252:272
    "pop = Rimatuu", # 273:293
    "pop = Slipper_Island", # 294:314
    "pop = Sulawesi", # 316:337
    "pop = Tahanea", # 338:357
    "pop = Wake_Island" # 360:379
  )
)

# Creating population dfs
a <- as.data.frame(copy[1:10,]) # Aotea
b <- as.data.frame(copy[11:28,]) # Borneo
c <- as.data.frame(copy[315,]) # Doubtful_Sound
d <- as.data.frame(copy[30,]) # Great_Mercury_Island
e <- as.data.frame(copy[31:42,]) # Halmahera
f <- as.data.frame(copy[43:63,]) # Hatutaa
```

```
g <- as.data.frame(copy[64:83,]) # Honuea
h <- as.data.frame(copy[84:103,]) # Kaikura_Island
i <- as.data.frame(copy[104:124,]) # Kamaka
j <- as.data.frame(copy[125:145,])  # Kayangel
k <- as.data.frame(copy[148:168,]) # Late_Island
l <- as.data.frame(copy[c(29, 146, 147, 169, 358, 359),]) # Mainland
m <- as.data.frame(copy[170:181,]) # Malenge
n <- as.data.frame(copy[182:195,]) # Mohotani
o <- as.data.frame(copy[196:216,]) #  Motukawanui
p <- as.data.frame(copy[217:226,]) #  New_Britain
q <- as.data.frame(copy[227:229,]) # New_Guinea
r <- as.data.frame(copy[230,]) #  Normanby_Island
s <- as.data.frame(copy[231:251,]) # Rakiura
t <- as.data.frame(copy[252:272,]) # Reiono
u <- as.data.frame(copy[273:293,]) # Rimatuu
v <- as.data.frame(copy[294:314,]) # Slipper_Island
w <- as.data.frame(copy[316:337,]) #  Sulawesi
x <- as.data.frame(copy[338:357,]) # Tahanea
y <- as.data.frame(copy[360:379,]) # Wake_Island

pops <- as.character(c(letters[seq(from = 1, to = 25)])) # list of popn object names
```

```
ncol(copy) #565
getwd()

sink("./data/ratsSNPs_PGDSpider_input.txt") # create empty file
cat("rats_SNPS", "npops = 25", "nloci = 282", fill = 1)
cat("\t", fill = FALSE)
cat(colnames(copy[,c(FALSE,TRUE)]), "\n", sep = "\t\t", fill = FALSE) # column/SNP names (even columns
for (i1 in 1:25) {
  cat(popnames[i1], fill = 1) # island name
  foo <- get(pops[i1]) # calling the island object based on the pops vector
  for (i2 in 1:nrow(foo)) {
    cat(as.character(foo[i2, ]), "\n", fill = FALSE, sep = "\t") # printing the SNP rows
  } # inner loop close
} # outer loop close
sink() # closing the sink connection (do not forget!)

rm(i1, i2, foo, popnames, pops)
rm(list = c(letters[seq(from = 1, to = 25)])) # removing excess objects
```

HWE Analysis and removal

Creating loop for reading the HWE files

```
getwd()
setwd("./results/Arlequin_HardyWeinberg/hwe_results_by_island_14032022")
filenames <- as.vector(list.files())

for (i in 1:length(filenames)) {
```

```
  df <- read.delim(filenames[i])
  m <- as.vector(grep("This locus is monomorphic", df[,1], value = FALSE, fixed = TRUE)) # making list
  df <- as.data.frame(df[-c(1,m),]) # removing the rows listed above, plus the dashed line

  df <- as.data.frame(gsub("  ", " ", df[,1], fixed = TRUE)) # removing spaces
  df <- as.data.frame(gsub("  ", " ", df[,1], fixed = TRUE)) # removing spaces
  df <- as.data.frame(gsub("  ", " ", df[,1], fixed = TRUE)) # removing spaces

  colnames(df) <- "Var1"
  df <- tidyr::separate(df, sep = " ", col = Var1, into = c("foo", "Locus", "Genot", "Obs.Het", "Exp.He
  df <- df[,-1] # removing extra row
  for (ii in 1:ncol(df)) {df[,ii] <- as.numeric(df[,ii])} # converting to numeric rather than character
  assign(paste(filenames[i]), df) # renaming object
  # write.table(df, paste("df", filenames[i], sep = "_"), row.names = FALSE, sep = "\t") # save to file
}

rm(i, ii, m, df, filenames)
setwd("C:/Users/airhe/OneDrive/Documents/Masters/Project 3/kiore-project")
```

Checking the HWE P-values

```
objectnames <- as.vector(ls()) # should be islands only, otherwise remove extras from vector
objectnames
# If necessary:
objectnames <- objectnames[-c(3, 27, 28)] # removing non-island objects

# making df of all hwe results
hwe.all <- data.frame()
for (i in 1:length(objectnames)) {
  foo <- get(objectnames[i])
  islandpop <- c(rep(paste(objectnames[i]), paste(nrow(foo)))) # making a vector of the popn. name
  foo$islandpop <- islandpop # adding the column to the results df to identify popn.
  hwe.all <- rbind(hwe.all, foo) # adding the popn. df to the combined hwe results df
}

rm(islandpop, foo, i)
rm(list = objectnames) # removes all the island objects
```

Running Holm's Sequential Bonferroni test to adjust p-values

```
nrow(hwe.all) # 2622

p.value.adjusted <- c(p.adjust(hwe.all$P.value, method = "holm")) # adjusting p-values
hwe.all$p.value.adjusted <- p.value.adjusted # making new column

rm(p.value.adjusted)
getwd()
write.csv(hwe.all, "./results/Arlequin_HardyWeinberg/HWEanalysis_allresults.csv", row.names = FALSE)
```

Examining significant hwe p-values

```
hwe.all <- read.csv("./results/Arlequin_HardyWeinberg/HWEanalysis_allresults.csv")
hwe.signif <- hwe.all[which(hwe.all$p.value.adjusted <= 0.05),]
hwe.signif[,c(1,8,9)]
plyr::count(hwe.signif$Locus) # 2 at locus 41 (Kaikura and Reiono), rest are singles
plyr::count(hwe.signif$islandpop) # concerning that 17 of 23 are Kayangel
```

Adjusted p-values per Locus

Locus	islandpop	p.value.adjusted
127	honuea	0.02613
182	honuea	0.00000
41	kaikura_island	0.02613
16	kayangel	0.02613
37	kayangel	0.02613
50	kayangel	0.02613
54	kayangel	0.02613
61	kayangel	0.02613
62	kayangel	0.00000
67	kayangel	0.00000
84	kayangel	0.00000
93	kayangel	0.02613
122	kayangel	0.00000
170	kayangel	0.02613
171	kayangel	0.02613
177	kayangel	0.00000
211	kayangel	0.02613
252	kayangel	0.02613
266	kayangel	0.00000
278	kayangel	0.00000
107	rakiura	0.02613
128	rakiura	0.02613
41	reiono	0.00000

Number of islands with a significant adjusted p-value at a particular locus

loci column	freq
16	1
37	1
41	2
50	1
54	1
61	1
62	1
67	1
84	1
93	1
107	1

loci column	freq
122	1
127	1
128	1
170	1
171	1
177	1
182	1
211	1
252	1
266	1
278	1

Number of loci per island population with significant adjusted p-values

island population	no. of signif. loci
honuea	2
kaikura_island	1
kayangel	17
rakiura	2
reiono	1

Removing samples/loci with issues identified in HWE and Structure analyses

```
getwd()
halfclean <- read.csv("./data/RStudio/ratsSNPs_halfclean.csv")

# need to remove Kamaka_009, and Rimatuu_19 and Rimatuu_20 due to position in Structure.
# both the (pre-cleanup) NeighborNet and Structure identify Kayangel17 as concerning, as well as Kayang

remove <- c(
    "Kamaka_009",
    "Rimatuu_19",
    "Rimatuu_20",
    "Kayangel11",
    "Kayangel13",
    "Kayangel15",
    "Kayangel17",
    "Kayangel19",
    "Kayangel21"
  ) # names of specimens to remove (each name should have exactly 10 characters)
x <- sapply(remove, function(i) grep(i, x = halfclean$island, value = FALSE)) # finding the row numbers
halfclean[c(x),1] # checking the names match
clean <- halfclean[-c(x),] # removing rows described above

getwd()
write.csv(clean, "./data/RStudio/ratsSNPs_clean.csv", row.names = FALSE)
```

```
rm(x, remove)
```

Double checking for monomorphic columns (SNP loci)

```
ncol(clean) #298
monocols <- integer() # empty vector for the for loop
for (i in 17:298) {
  z <- length(unique(clean[,i])) # no. of unique values in the row (looking for 1, or 2 if there's "?")
    if (z <= 3)
      {monocols <- append(monocols, i) # if z is as so, add the column number to the vector
    }
  rm(z)
}
# tried with z <= 2 but no result, therefore tried z <= 3 and checked the results manually below.

monocols # 29  30  43  52  54  76  84  86  89 105 113 115 123 127 136 149 154 155 156 170 179 182 183 2
for (i in monocols) {
  print(count(clean[,i]))
}
# none with only 1 unique SNP in each column. It's possible since the SNP loci were selected for their
```

```
# x <- x[,-c(monocols)] # for removal of monomorphic columns, but none found

rm(i, monocols)
```

Specimens per Island after full data clean-up

```
plyr::count(clean$island.1)
```

Island	freq before cleanup	freq after cleanup	difference
Aotea (Great Barrier I)	10	10	0
Borneo	25	18	7
Doubtful Sound	1	1	0
Great Mercury Island	1	1	0
Halmahera	25	12	13
Hatutaa	21	21	0
Honuea	21	20	1
Kaikura Island	20	20	0
Kamaka	21	20	1
Kayangel	21	15	6
Late Island	21	21	0
Luzon	1	1	0
Mainland	5	5	0
Malenge	25	12	13
Mohotani	14	14	0
Motukawanui	21	21	0

Island	freq before cleanup	freq after cleanup	difference
New Britain	26	10	16
New Guinea	25	3	22
Normanby Island	25	1	24
Rakiura (Stewart Isl)	21	21	0
Reiono	21	21	0
Rimatuu (Tetiaroa)	21	19	2
Slipper Island	21	21	0
Sulawesi	25	22	3
Tahanea	20	20	0
Wake Island	20	20	0

Islands represented by very few specimens (<= 3) are Doubtful Sound, Great Mercury Island, Luzon, New Guinea, and Normanby Island.