Testing the Heterozygosity per Island

Grace Saville

01/02/2022

Preamble

```
library(stringr)
library(dplyr)
library(ggplot2)
library(car)
getwd()
setwd("C:/Users/airhe/OneDrive/Documents/Masters/Project 3/kiore-project")
```

Dataset loading

```
data <- read.csv("./data/RStudio/ratsSNPs_clean.csv")

copy <- data # making a copy of the data
colnames(copy)
copy <- copy[,-c(2:16)] # removing unnecessary columns for this analysis
```

Replaces SNP's with symbols for heterozygous and homozygous

```
dim(copy) # 370 rows 283 columns
str(copy)
unique(unlist(copy[,17:283])) # checking what SNP combinations are present
# "?"   "A:A" "G:G" "A:G" "T:T" "A:T" "C:C" "C:T" "T:G" "C:G" "A:C"

het <- c("A:G", "A:T", "C:T", "T:G", "C:G", "A:C") # vector of heterozygous combinations
hom <- c("G:G", "A:A", "C:C", "T:T") # vector of homozygous combinations

# Replacing specific combos
for (i in 1:nrow(copy)){
  copy[i,][copy[i,] %in% het] <- "E"
  copy[i,][copy[i,] %in% hom] <- "O"
}

rm(i)
```

```
str(copy)
copy[copy == "?"] <- NA # replacing ? with NA's
```

Calculating heterozygous and homozygous totals per specimen

```
o.freq <- vector()
e.freq <- vector()
na.freq <- vector()
for (i in 1:370) {
  x <- sum(grepl("O", copy[i,], fixed = TRUE)) # counting row O's
  o.freq <- append(o.freq, x) # adding sum to vector
  x <- sum(grepl("E", copy[i,], fixed = TRUE)) # counting row E's
  e.freq <- append(e.freq, x)
  x <- sum(is.na(copy[i,])) # counting row NA's
  na.freq <- append(na.freq, x)
  rm(x)
}

rm(i)

per.specimen <- data.frame(copy$island, o.freq, e.freq, na.freq) # making a df with the freq totals
perc.o <- round(as.vector((per.specimen$o.freq / (per.specimen$o.freq + per.specimen$e.freq)) * 100), di
per.specimen$perc.o <- perc.o
perc.missing <- round(as.vector((per.specimen$na.freq / (per.specimen$o.freq + per.specimen$e.freq + pe
per.specimen$perc.missing <- perc.missing

rm(o.freq, e.freq, na.freq, perc.o, perc.missing)

str(per.specimen)
```

Calculating heterozygous and homozygous totals per island

```
data <- data[order(data$island, decreasing = FALSE),] # ordering df alphabetically by island
names(data)
data[c(grep("Mainland", data$island.1)),c(1, 8)] # checking which populations fall in the mainland cate
unique(data$island)
shortpopnames <- as.character(
  c(
    "aotea",
    "borneo",
    "cambodia",
    "grtmercury",
    "halmaher",
    "hatutaa",
    "honuea",
    "kaikura",
    "kamaka",
```

```
    "kayangel",
    "laos",
    "late",
    "luzon",
    "malenge",
    "mohotani",
    "motukawa",
    "newbrita",
    "newguine",
    "normanby",
    "rakiura",
    "reiono",
    "rimatuu",
    "slipper",
    "southland",
    "sulawesi",
    "tahanea",
    "thailand",
    "wake"
  )
) # writing shortened names as is in the data df so I can use the character strings with grep()

# splitting the data df into df objects by island:
for (i in 1:length(shortpopnames)) {
  y <- as.vector(grep(shortpopnames[i], copy[,1], ignore.case = TRUE, value = FALSE))
  assign(paste(shortpopnames[i]), copy[y,])
}
rm(y, i)

# making a "mainland" df:
mainland <- rbind(cambodia, thailand, laos)
rm(cambodia, thailand, laos)
shortpopnames
shortpopnames <- shortpopnames[-c(3, 11, 27)] # removing the names now in mainland category
shortpopnames <- append(shortpopnames, "mainland")

# double checking the dfs have the correct specimens per island:
length(shortpopnames) #26
for (i in 1:26) {
  print(shortpopnames[i])
  y <- get(shortpopnames[i])
  print(y[,1])
}
rm(y)
```

```
# counting values:
per.island <- data.frame()
for (i in 1:length(shortpopnames)) {
  o <- sum(grepl("O", unlist(get(shortpopnames[i])), fixed = TRUE)) # counting row O's
  e <- sum(grepl("E", unlist(get(shortpopnames[i])), fixed = TRUE)) # counting row E's
  na <- sum(is.na(get(shortpopnames[i]))) # counting row NA's
  vec <- as.vector(c(shortpopnames[i], nrow(get(shortpopnames[i])), o, e, na))
  per.island <- rbind(per.island, vec)
```

```
}
rm(o, e, na, vec, i)

colnames(per.island) <- c("island", "specimens.based.on", "o.freq", "e.freq", "na.freq") # renaming col
str(per.island)
per.island$o.freq <- as.integer(per.island$o.freq) # changing the numbers from characters to integers
per.island$e.freq <- as.integer(per.island$e.freq)
per.island$na.freq <- as.integer(per.island$na.freq)

# calculating percentages:
perc.o <- round(as.vector((per.island$o.freq / (per.island$o.freq + per.island$e.freq)) * 100), digits =
per.island$perc.o <- perc.o
perc.missing <- round(as.vector((per.island$na.freq / (per.island$o.freq + per.island$e.freq + per.islan
per.island$perc.missing <- perc.missing

rm(perc.o, perc.missing, het, hom)
rm(list = shortpopnames)
```

```
getwd()
write.csv(per.island, "./results/heterozygosity_testing_results_table.csv", row.names = FALSE)
```

Considering island size and distance from mainland

```
island.km <- read.csv("./data/island_size_data.csv", header = TRUE)
head(island.km)
island.km <- island.km[,c(2, 4)]
island.km <- island.km[order(island.km$ISLAND, decreasing = FALSE),] # sorting alphabetically
```

```
distance.from.ML <- data[,c(1,8,11,12)]
distance.from.ML <- distance.from.ML[!duplicated(distance.from.ML$island.1),] # keeping only 1 coordina
distance.from.ML <- distance.from.ML[order(distance.from.ML$island.1, decreasing = FALSE),] # sorting a
row.names(distance.from.ML) <- seq(nrow(distance.from.ML)) # renaming row numbers to be sequential
distance.from.ML # checking

#shortening some of the long island names to match what is already in the island.km df:
distance.from.ML[1,2] <- "Aotea"
distance.from.ML[20,2] <- "Rakiura"
distance.from.ML[22,2] <- "Rimatuu"

#decided to keep Cambodia as the mainland coordinates, will use this as the base for distance from main
distance.from.ML <- distance.from.ML[,-1] # removing specimen ID column since no longer necessary

library(geosphere)
km.from.ML <- vector()
for (i in 1:nrow(distance.from.ML)) {
  x <- distGeo(as.vector(distance.from.ML[13,c(3,2)]), as.vector(distance.from.ML[i,c(3,2)]))
  km.from.ML <- append(km.from.ML, x)
}

km.from.ML <- km.from.ML / 1000 # converting from metres to kilometres
```

```
distance.from.ML$km.from.ML <- km.from.ML # adding the kms to the distance df

rm(i, x, km.from.ML)
```

```
names(island.km)
names(distance.from.ML)
island.km <- merge(island.km, distance.from.ML, by.x = "ISLAND", by.y = "island.1", all = TRUE)

rm(distance.from.ML)

island.km$ISLAND
per.island$island
per.island$island <- c(
  "Aotea",
  "Borneo",
  "Great Mercury Island",
  "Halmahera",
  "Hatutaa",
  "Honuea",
  "Kaikura Island",
  "Kamaka",
  "Kayangel",
  "Late Island",
  "Luzon",
  "Malenge",
  "Mohotani",
  "Motukawanui",
  "New Britain",
  "New Guinea",
  "Normanby Island",
  "Rakiura",
  "Reiono",
  "Rimatuu",
  "Slipper Island",
  "Doubtful Sound",
  "Sulawesi",
  "Tahanea",
  "Wake Island",
  "Mainland"
) # editing the names to match those in the other df so I can merge them

per.island <- merge(per.island, island.km, by.x = "island", by.y = "ISLAND", all = TRUE)
```

```
getwd()
write.csv(per.island, "./results/heterozygosity_testing_results_table.csv", row.names = FALSE)
```

Linear regression and plots

```
per.island <- read.csv("./results/RStudio_Heterozygosity/heterozygosity_testing_results_table.csv", hea
```

Issues to consider before/during the statistical analyses:

- Some islands may be outliers because they are based on too few specimens (e.g. due to removal because of too many missing SNPs)
- Some islands may have specimens that are closely related to each other if they were sampled from the same site therefore may not accurately represent the population
- Sample size is small therefore diagnostics could be overinterpreted and results might not represent reality well
- Distance from the mainland may not be the best measure because it doesn't indicate the difficultly of reaching the island in all cases, from example Normanby isl. is right next to New Guinea which is very large and likely has a diverse population that can easily move to Normanby and back.

a. Distance

i. Model Construction and Diagnostics

```
testLM <- lm(perc.o ~ km.from.ML, data = per.island) # model

plot(perc.o ~ km.from.ML, data = per.island)
abline(coef = coef(testLM), col = 4, lwd = 2)

summary(testLM) # model results
```

```
par(mfrow = c(2, 2)) # changes the number of plots visible at once
plot(testLM) # diagnostic plots

par(mfrow = c(1, 1))
residualPlots(testLM) # curved therefore some non-linearity possible, although Tukey test p-value is no

qqPlot(testLM$residuals, line = "quartiles") # non-normal dist
shapiro.test(testLM$residuals) # W = 0.91825, p-value = 0.04092, indicates non-normality of resid
hist(testLM$residuals, breaks = 10) #

ncvTest(testLM) # homoscedasticity test: Chisquare = 5.719721, Df = 1, p = 0.016775, H0 of constant var

influenceIndexPlot(testLM) # outliers
# Cooks distances: none larger than 0.5,
# Studentised residuals: point 18 less than -3
# Bonferroni p-value: point 18 smaller then 0.05,
# Hat-values: point 13 influential, higher than 1
outlierTest(testLM) # 18: rstudent = -3.479726, unadjusted p-value = 0.0020257, Bonferroni p = 0.052668

boxCox(testLM) # suggests sqrt transformation
```

Notes on diagnostic plots

- Residuals vs. Fitted plot shows a line not quite horizontal, but close enough to flat consider the relationship linear. Cone shape indicates heteroscedasticity. Questionable points: 10 (Kayangel), 18 (New Guinea), 19 (Normanby Island).

- Normal QQ plot shows residuals generally in line, indicating normality, however the tails are fat ant bottom tail prominent, indicating fat tails and/or slight left-skewed distribution (towards the right). The Shapiro-Wilk test (only just) rejects H0 that residuals are normal. Questionable points are again: 10 (Kayangel), 18 (New Guinea), 19 (Normanby Island).
- Scale-Location plot is used to indicate constant residual variance with a line that does not trend up or down overall. Here it may be flat or trend down however it is not entirely clear due to the questionable points (10 (Kayangel), 18 (New Guinea), 19 (Normanby Island)) pulling a section upwards. The NCV test indicates non-constant variance.
- Residuals vs Leverage (Cook's distance) plot assesses outliers. None of the points are over 1 (Cook's distance line which would make them statistical outliers) nor are any over 0.5 (which would make them questionable). Point 13 (Mainland) has high leverage, however this makes sense because the distance to the Mainland from the Mainland is 0, which is an unusually low number in this model. Point 18 (New Guinea) is identified as a potential issue in the outlier tests.

ii. Adjusted Model and Diagnostics

I have decided to remove the Mainland point for several reasons; it is an outlier with a high leverage and a value of 0, also not valuable since the Mainland is the reference point, not a new value. Testing removing 18 and 19 (New Guinea and Normanby island) saw an improvement in normality, which makes the model more reliable, however I don't think they are true outliers outside of the dataset, so the results from this model will come with this caveat. Additionally, New Guinea and Normanby island are based off of 3 and 1 specimen respectively, which mean the points may not represent their populations accurately.

```
z <- per.island[-c(13, 18, 19),] # removing points mentioned
LM <- lm(perc.o ~ km.from.ML, data = z) # model

plot(perc.o ~ km.from.ML, data = z)
abline(coef = coef(LM), col = 4, lwd = 2)

summary(LM) # model results
```

```
par(mfrow = c(2, 2))
plot(LM) # diagnostic plots

par(mfrow = c(1, 1))
residualPlots(LM) # curved therefore some non-linearity possible, although Tukey test p-value 0.5348 (H

qqPlot(LM$residuals, line = "quartiles") # normal residual distribution
shapiro.test(LM$residuals) # W = 0.95512, p-value = 0.3723, indicates normality of residuals
hist(LM$residuals, breaks = 10)

ncvTest(LM) # homoscedasticity test: Chisquare = 2.469013, Df = 1, p = 0.11611, H0 of constant variance

influenceIndexPlot(LM) # nothing overly concerning
outlierTest(LM) # 10: rstudent = 2.472633, unadjusted p-value = 0.022509, Bonferroni p = 0.5177
# not significant
```

The distribution of residuals is now normal, although there appears to be some level of heteroscedasticity visible in the plots (residuals vs. fitted, scale-location, residualPlot) even though the ncv test does not reject HO of constant variance of the residuals. I am not too concerned with this because the sample size is small and I do not want to over-analyse the diagnostics when my goal is merely to check for a correlation.

iii. Results

Residuals:

Min	1Q	Median	3Q	Max
-4.4315	-2.2204	-0.2978	2.3458	5.9795

Coefficients:

	Estimate	Std. Error	t value	p-value
(Intercept)	86.95	1.379	63.035	< 2e-16 ***
km.from.ML	0.0005221	0.0001509	3.459	0.00235 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.858 on 21 degrees of freedom
Multiple R-squared: 0.363
Adjusted R-squared: 0.3326
F-statistic: 11.96 on 1 and 21 DF, p-value: 0.002349

- Higher t-value is generally more significant, related to p-value.
- Multiple R-squared indicates ~36% of the variance is explained by the model
- Significant p-value(s), can reject H0 where x and y are not correlated.

vi. Model Plot

```
ggplot(data = per.island, aes(x = km.from.ML, y = perc.o, label = island)) +
  geom_point(colour = "red", size = 3) +
  geom_smooth(method = 'lm', se = TRUE, level = 0.95) +
  geom_text(data = subset(per.island, km.from.ML < 6500 & (perc.o < 80 | perc.o > 90)), hjust = 0.1, vj
  ggtitle("Effect of Island distance from Mainland on Homozygosity") +
  xlab("Distance from Mainland (km)") + ylab("Homozygosity (%)") +
  theme_light()
```

b. Area

i. Model Construction and Diagnostics

```
testLM2 <- lm(perc.o ~ area_km2, data = per.island) # model

plot(perc.o ~ area_km2, data = per.island) # points clustered in small area_km2
abline(coef = coef(testLM2), col = 4, lwd = 2)

summary(testLM2) # model results
```

```r
par(mfrow = c(2, 2)) # changes the number of plots visible at once
plot(testLM2) # diagnostic plots

par(mfrow = c(1, 1))
residualPlots(testLM2) # curved therefore non-linearity possible, however Tukey test p-value 0.5124. Sp

qqPlot(testLM2$residuals, line = "quartiles") # normal dist, outlier point 19
shapiro.test(testLM2$residuals) # W = 0.89572, p-value = 0.01482, indicates non-normality of residuals
hist(testLM2$residuals, breaks = 10)

ncvTest(testLM2) # homoscedasticity test: Chisquare = 0.4593862, Df = 1, p = 0.49791, H0 of constant va

influenceIndexPlot(testLM2) # outliers
# Cooks distances: points 2 and 18 larger than 0.5,
# Studentised residuals: point 19 less than -3
# Bonferroni p-value: point 19 smaller then 0.05,
# Hat-values: points 2 and 18 visibly different, perhaps influential but not higher than 1
outlierTest(testLM2) # 19: rstudent = -3.900653, unadjusted p-value = 0.00076823, Bonferroni p = 0.0192

boxCox(testLM2) # suggests log transformation
```

Notes on diagnostic plots

- Residuals vs. Fitted plot shows a line not horizontal, therefore relationship may not be linear here. Points 18 and 19 are more extreme.
- Normal QQ plot shows residuals generally in line, indicating normality although there is a fat lower tail. The Shapiro-Wilk test indicates non-normality of residuals, although not if the alpha is set to 0.01. Questionable point is once again 18 (New Guinea).
- Scale-Location plot is used to indicate constant residual variance with a line that does not trend up or down overall. Line trends downwards which means non-constant variance, although strangely the ncvTest indicates H0 of constant variance should not be rejected.
- Residuals vs Leverage (Cook's distance) plot assesses outliers. Points 2 (Borneo) and 18 (New Guinea) is over the 0.5 Cook's line and high leverage and therefore is noteworthy. This could be explained by them being the largest island areas in the dataset/model. The outlier tests also bring attention to point 19.
- The Box-Cox plot indicates a transformation to perform and here lambda lines up almost perfectly with 0 == log transformation.

ii. Adjusted Model and Diagnostics

Trying log10 of the area since there's a cluster of small islands. Bigger islands will be "squeezed" more than smaller islands so that ratios/distances between points are preserved, abline is estimated over them (but if "unsqueezed" the abline would become curved). A log10 relationship might make sense because the effect of islands size decreases as the islands get bigger.

I'm hesitant to remove any outliers here, but I'm seeing how it goes removing New Guinea and Normanby Island again. Removing them doesn't mean I will not consider them again, however due to having so few data points I think they strongly influence the results and it may be easier to see the general picture without them. Additionally, Normanby Island is missing 57% and New Guinea 56% of the SNPs, making them less reliable.

```r
z <- per.island[-c(18, 19),] # removing New Guinea and Normanby Island again

LM2 <- lm(perc.o ~ log10(area_km2), data = z)

plot(perc.o ~ log10(area_km2), data = z)
abline(coef = coef(LM2), col = 4, lwd = 2)

summary(LM2)
```

```r
par(mfrow = c(2, 2)) # changing the number of plots visible at once
plot(LM2) # diagnostic plots; relationship now linear and outliers are now less influential

par(mfrow = c(1, 1))
residualPlots(LM2) # curved line but Tukey test p-value 0.6773 which means linear relationship

qqPlot(LM2$residuals, line = "quartiles") # normal
shapiro.test(LM2$residuals) # W = 0.98906, p-value = 0.9945, indicates normality of residuals
hist(LM2$residuals, breaks = 10, xlim = c(-10,10))

ncvTest(LM2) # homoscedasticity test: Chisquare = 0.8664813, Df = 1, p = 0.35193, H0 of constant varian

influenceIndexPlot(LM2) # outliers
# Cooks distances: none larger than 0.5,
# Studentised residuals: none less than -3 or more than 3
# Bonferroni p-value: none smaller then 0.05,
# Hat-values: points 2 influential, higher than 1
outlierTest(LM2) # 14: rstudent = -2.493984, unadjusted p-value = 0.021508, Bonferroni p = 0.49468
```

iii. Results

Call: lm(formula = perc.o ~ log10(area_km2), data = per.island)

Residuals:

Min	1Q	Median	3Q	Max
-6.491	-1.622	-0.317	1.754	5.747

Coefficients:

	Estimate	Std. Error	t value	p-value
(Intercept)	92.9729	0.8395	110.750	< 2e-16 ***
log10(area_km2)	-0.9171	0.3012	-3.045	0.00616 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.982 on 21 degrees of freedom (1 observation deleted due to missingness)
Multiple R-squared: 0.3062

Adjusted R-squared: 0.2732
F-statistic: 9.27 on 1 and 21 DF, p-value: 0.006159

vi. Model Plot

```
ggplot(data = per.island, aes(x = log10(area_km2), y = perc.o, label = island)) +
  geom_point(colour = "red", size = 3) +
  geom_smooth(method = 'lm', se = TRUE, level = 0.95) +
  geom_text(data = subset(per.island,
    (log10(area_km2) < 2 & perc.o < 90) | (log10(area_km2) > 2.9 & (perc.o > 92 | perc.o < 80))
    ),
    hjust = 0.8, vjust = -0.8, size = 3) +
  ggtitle("Effect of Island Area on Homozygosity") +
  xlab("log Area of Island (log10 km^2)") + ylab("Homozygosity (%)") +
  theme_light()
```

c. Multiple Regression: both Distance and Area

i. Model Construction and Diagnostics

Taking a moment to look at if there's a relationship between area and distance from the mainland:

```
par(mfrow = c(1, 1))
adLM <- lm(log(area_km2) ~ km.from.ML, data = per.island)
summary(adLM)
plot(adLM)

plot(log(area_km2) ~ km.from.ML, data = per.island)
abline(coef = coef(adLM), col = 4, lwd = 2)

shapiro.test(adLM$residuals) # p-value = 0.8809, normality
cor.test(per.island$km.from.ML, log(per.island$area_km2), method = "pearson") # p-value = 0.0002738, co
```

The above plot shows that generally the area of the islands decreases with distance (e.g. Borneo and New Guinea are large, Reiono and Tahanea are small) which I believe is a combination of sampling issues and how the islands are naturally on the Pacific. The correlation between these two variables is an issue with only a basic linear models because I cannot test the affect of each variable compared with each other because their effect is similar (collinearity in this case). I will go through diagnostics to check if the effect is great enough to warrant a different method e.g. GLM.

```
z <- per.island[-c(18, 19),] # starting by removing New Guinea and Normanby Island again (after brief c

testLM3 <- lm(perc.o ~ km.from.ML + log(area_km2), data = z) # not sure if I should keep log transforma
summary(testLM3)
```

```
par(mfrow = c(2, 2))
plot(testLM3)

par(mfrow = c(1,1))
residualPlots(testLM3)
```

```r
hist(testLM3$residuals, breaks = 10, xlim = c(-10,10))

qqPlot(testLM3$residuals, line = "quartiles") # normal
shapiro.test(testLM3$residuals) # W = 0.96668, p-value = 0.6101, indicates normality of residuals

ncvTest(testLM3) # homoscedasticity test: Chisquare = 0.8398007, Df = 1, p = 0.35945, H0 of constant va

influenceIndexPlot(testLM3) # outliers
# Cooks distances: no. 2 larger than 0.5,
# Studentised residuals: none less than -3 or more than 3
# Bonferroni p-value: none smaller then 0.05,
# Hat-values: none influential, higher than 1
outlierTest(testLM3) # 14: rstudent = 2.238249, unadjusted p-value = 0.037375, Bonferroni p = 0.85962

mctest::mctest(testLM3, type = "b") # tests for collinearity, most tests did not detect e.g. condition
```

iii. Results

Call: lm(formula = perc.o ~ km.from.ML + log10(area_km2), data = per.island)

Residuals:

Min	1Q	Median	3Q	Max
-4.8047	-2.0464	-0.1323	2.0577	4.7434

Coefficients:

	Estimate	Std. Error	t value	p-value
(Intercept)	89.1097394	2.2543620	39.528	<2e-16 ***
km.from.ML	0.0003642	0.0001989	1.832	0.0819 .
log10(area_km2)	-0.1985810	0.1651672	-1.202	0.2433

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.828 on 20 degrees of freedom (1 observation deleted due to missingness)
Multiple R-squared: 0.4059
Adjusted R-squared: 0.3465
F-statistic: 6.832 on 2 and 20 DF, p-value: 0.005478

- Residuals appear well distributed (approximately 5 on eaither side of 0, which the median is close to)
- Here the explanatory variables are no longer significant, which may mean that both contribute to the overall model significance but the exact effect of each variable is unclear. This matches up with my findings that the two explanatory variables are at least somewhat colinear.
- 23 observations - 3 parameters = 20 degrees of freedom
- The R^2 value indicates that 40% of the variance is explained (35% when considering the extra parameter added)
- The f-test is significant, therefore the variance that is explained is significant (?)

vi. Model Plot

```
ggplot(data = per.island, aes(x = km.from.ML, y = perc.o, label = island)) +
  geom_point(mapping = aes(colour = log10(area_km2)), size = 3) +
  geom_smooth(method = 'lm', se = TRUE, level = 0.95) +
  scale_color_gradient(low = "red", high = "blue") +
  geom_text(data = subset(per.island,
    (perc.o < 80) | (km.from.ML < 5000 & perc.o > 90)),
    hjust = -0.1, vjust = -0.8, size = 3) +
  ggtitle("Effect of Island distance from Mainland on Homozygosity, including island area") +
  xlab("Distance from Mainland (km)") + ylab("Homozygosity (%)") +
  theme_light()
```