Chapter 13 Population Genetics

Angelo LaCommare-Soto

2025-05-02

Genetic Diversity and Population Structure in Woodrats

Research Question 1: Does genetic differentiation increase with geographic distance among dusky-footed woodrat populations?

Section 1 - Importing Data

```
# set working directory for all chunks in this file (default working directory is wherever Rmd file is)
getwd()
```

```
## [1] "C:/Users/Angelo L/Documents/GitHub/BIOL710/RCode710/RCode/working_directory"
```

```
library(tidyverse)
```

```
## Warning: package 'purrr' was built under R version 4.4.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(adegenet)
```

```
## Warning: package 'adegenet' was built under R version 4.4.3
```

```
## Loading required package: ade4
```

```
## Warning: package 'ade4' was built under R version 4.4.3
```

```
##
##    /// adegenet 2.1.11 is loaded ////////////
##
##    > overview: '?adegenet'
##    > tutorials/doc/questions: 'adegenetWeb()'
##    > bug reports/feature requests: adegenetIssues()
```

```r
library(hierfstat)
```

```
## Warning: package 'hierfstat' was built under R version 4.4.3
```

```
##
## Attaching package: 'hierfstat'
##
## The following objects are masked from 'package:adegenet':
##
##     Hs, read.fstat
```

```r
library(vegan)
```

```
## Warning: package 'vegan' was built under R version 4.4.3
```

```
## Loading required package: permute
```

```
## Warning: package 'permute' was built under R version 4.4.3
```

```
## Loading required package: lattice
```

```r
library(vcfR)
```

```
## Warning: package 'vcfR' was built under R version 4.4.3
```

```
##
##    *****       ***   vcfR   ***       *****
##    This is vcfR 1.15.0
##      browseVignettes('vcfR') # Documentation
##      citation('vcfR') # Citation
##    *****       *****      *****       *****
```

```r
# Load metadata (locations, coordinates)
neo_fus_loc <- read.csv("Neo_fus.csv")

# Load SNP genotype data from VCF file
neo_fus_gen <- read.vcfR("Neo_fus.vcf")
```

```
## Scanning file to determine attributes.
## File attributes:
##   meta lines: 10
##   header_line: 11
##   variant count: 1000
##   column count: 77
## Meta line 10 read in.
## All meta lines processed.
## gt matrix initialized.
## Character matrix gt created.
##   Character matrix gt rows: 1000
```

```
##   Character matrix gt cols: 77
##   skip: 0
##   nrows: 1000
##   row_num: 0
## Processed variant 1000Processed variant: 1000
## All variants processed
```

Section 2 - Creating a Genind Object

```
# Convert VCF data to a genind object
genind_obj <- vcfR2genind(neo_fus_gen)

# View summary of the genetic data
#summary(genind_obj))

#converst to a dataframe for easier interpretation
test<-genind2df(genind_obj)
#View(test)
```

Question Answers

a. There are 68 woodrat individuals in the genotype dataset.
b. The dataset includes information such as number of alleles per locus, observed heterozygosity, and expected heterozygosity.

Section 3 - Calculating Hetegozygosity

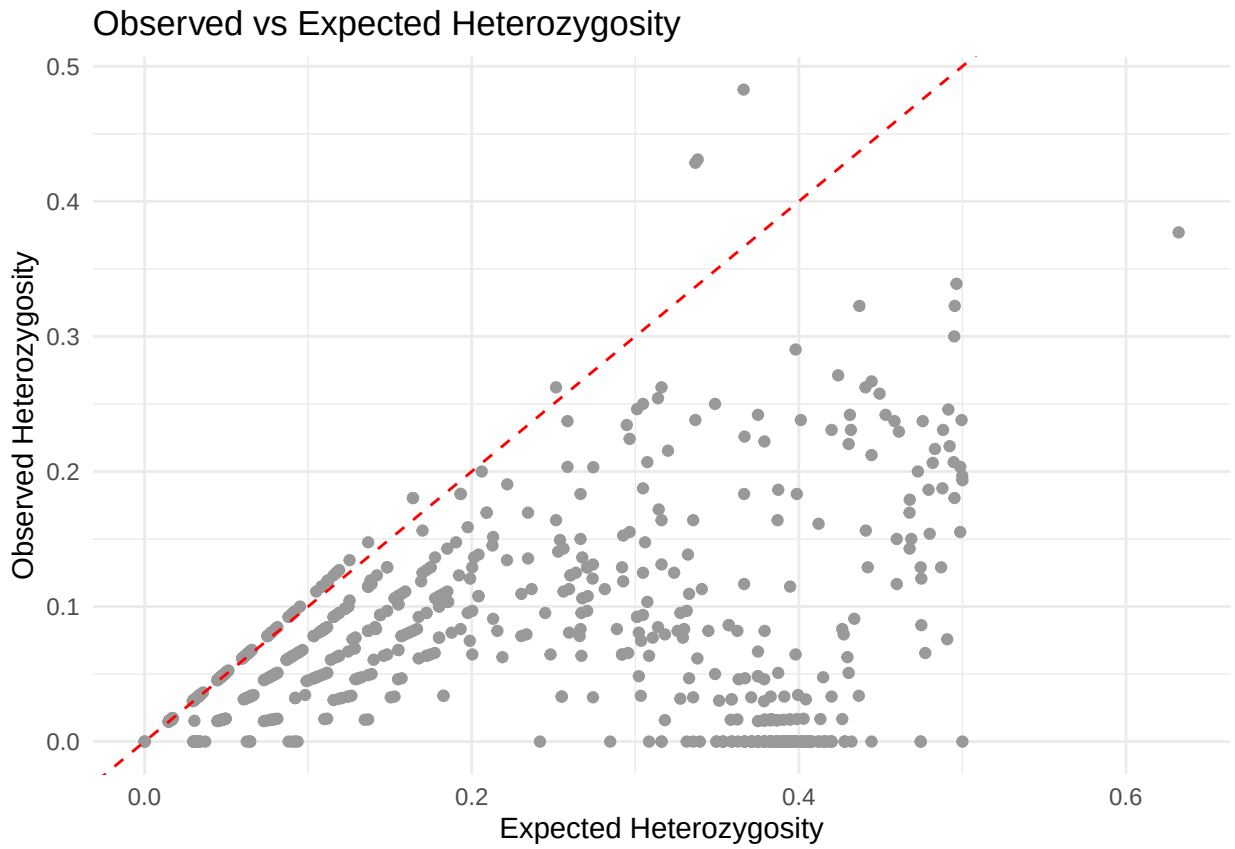
```
# Summarize the genind object to get basic statistics
genind_summary <- summary(genind_obj)

# Extract observed and expected heterozygosity per SNP locus
observed_het <- genind_summary$Hobs
expected_het <- genind_summary$Hexp

# Combine into a tidy table
het_df <- tibble(
  Locus = names(observed_het),
  Hobs = observed_het,
  Hexp = expected_het
)

# View first few rows
head(het_df)
```

```
## # A tibble: 6 x 3
##   Locus             Hobs      Hexp
##   <chr>            <dbl> <dbl[1d]>
## 1 locus_109138_73 0.0625    0.0605
## 2 locus_33786_87  0.0317    0.0312
```

```
## 3 locus_52829_25  0.0833    0.110
## 4 locus_12937_15  0.183     0.193
## 5 locus_52732_87  0         0.0312
## 6 locus_116968_44 0         0.406
```

Section 4 - Plotting Observed and Expected Heterozygosity

```r
# Plot observed vs expected heterozygosity
ggplot(het_df, aes(x = Hexp, y = Hobs)) +
  geom_point(color = "gray60") +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +
  theme_minimal() +
  labs(x = "Expected Heterozygosity",
       y = "Observed Heterozygosity",
       title = "Observed vs Expected Heterozygosity")
```



Question Answers

a. Most loci have an observed heterozygosity less than the expected heterozygosity, given by the location of a majority of points below the guiding red-colored line that shows complete congruence between expected heterozygosity and observed heterozygosity
b. Loci can deviate from the Hardy-Weinberg expectations if there is random genetic drift or selection for a trait that is associated with homozygosity within a population.

Section 5 - Calculating Fst Among Populations

```
# Assign populations based on metadata
pop(genind_obj) <- as.factor(neo_fus_loc$location)

# Convert genind object to hierfstat format
hf_data <- genind2hierfstat(genind_obj)

# Calculate basic statistics, including per-locus Fst values
basic_stats <- basic.stats(hf_data)

# View per-locus values, including Fst
basic_stats_perloc <- basic_stats$perloc

# Turn locus IDs (rownames) into a column
basic_stats_perloc$locusID <- rownames(basic_stats_perloc)

#View product
summary(basic_stats_perloc)
```

```
##        Ho                Hs                Ht                Dst
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :-0.00910
##  1st Qu.:0.01090   1st Qu.:0.0236   1st Qu.:0.02350   1st Qu.: 0.00000
##  Median :0.02380   Median :0.0890   Median :0.09205   Median : 0.00100
##  Mean   :0.05201   Mean   :0.1335   Mean   :0.15078   Mean   : 0.01731
##  3rd Qu.:0.06520   3rd Qu.:0.2360   3rd Qu.:0.28732   3rd Qu.: 0.02948
##  Max.   :0.50320   Max.   :1.0000   Max.   :1.00000   Max.   : 0.15730
##
##       Htp               Dstp               Fst                Fstp
##  Min.   :0.00000   Min.   :-0.01820   Min.   :-0.04050   Min.   :-0.0844
##  1st Qu.:0.02330   1st Qu.:-0.00010   1st Qu.:-0.00430   1st Qu.:-0.0087
##  Median :0.09445   Median : 0.00210   Median : 0.01570   Median : 0.0310
##  Mean   :0.16675   Mean   : 0.03469   Mean   : 0.05917   Mean   : 0.1014
##  3rd Qu.:0.32610   3rd Qu.: 0.05920   3rd Qu.: 0.11350   3rd Qu.: 0.2051
##  Max.   :0.65620   Max.   : 0.31460   Max.   : 0.32830   Max.   : 0.4943
##  NA's   :2         NA's   :2          NA's   :3          NA's   :5
##       Fis               Dest            locusID
##  Min.   :-0.3508   Min.   :-0.03600   Length:1000
##  1st Qu.: 0.0088   1st Qu.:-0.00010   Class :character
##  Median : 0.3825   Median : 0.00220   Mode  :character
##  Mean   : 0.4157   Mean   : 0.04716
##  3rd Qu.: 0.7952   3rd Qu.: 0.07980
##  Max.   : 1.0000   Max.   : 0.47060
##  NA's   :3         NA's   :2
```

Question Answers

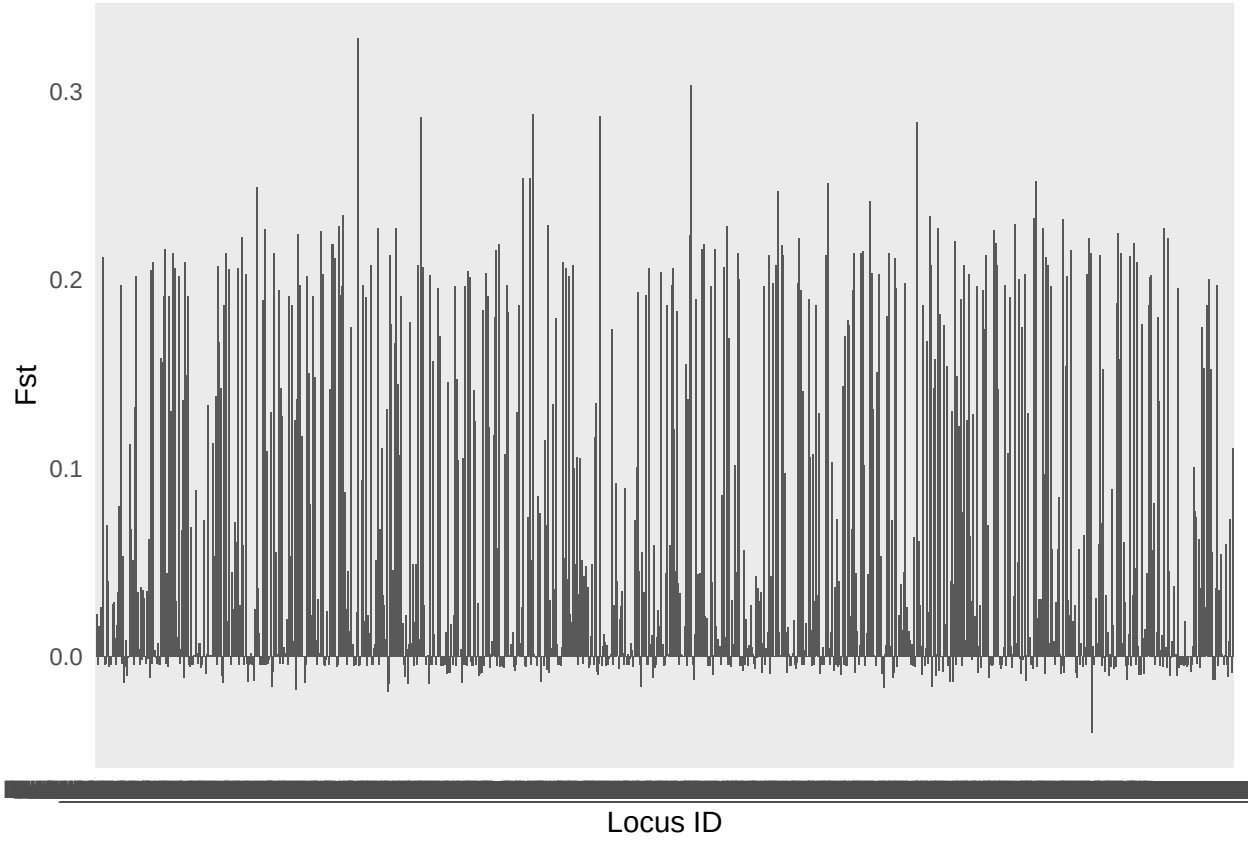
a. Loci 119007_83 and 30398_11 show the highest Fst values with values greater than 0.3. These high values suggest that there is large genetic differentiation at these loci among north and south woodrat populations.
b. There are some negative Fst values for certain loci, which are difficult to interpret, but also a good amount of positive Fst values for loci that are near zero. These low Fst values suggest that there is

no significant difference in the in genetic composition at these loci among north and south woodrat populations. These loci could be highly biologically conserved or under similar selection pressures and thus exhibit congruence.

Challenge 1: Create a plot that shows Fst values for each SNP. Hint: use ggplot2::geom_col().

```
f1<-ggplot(basic_stats_perloc, aes(x = locusID, y = Fst)) +
  geom_col() +
  labs(
    x = "Locus ID",
    y = "Fst") +
  theme_minimal()
f1
```

```
## Warning: Removed 3 rows containing missing values or values outside the scale range
## (`geom_col()`).
```



Section 6 - Spatial Patterns of Genetic Structure

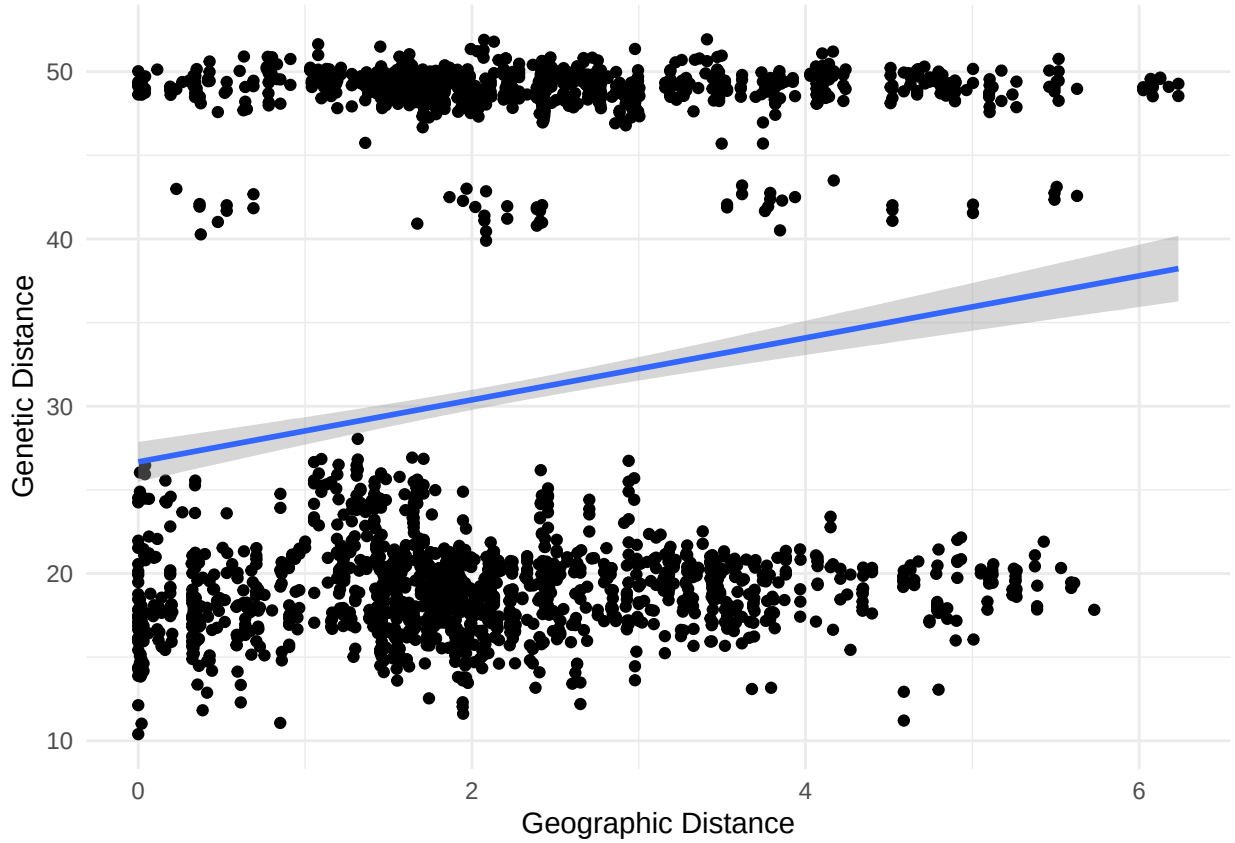
```
# Calculate pairwise geographic distances
dist_geo <- neo_fus_loc %>%
```

```
  select(longitude, latitude) %>%
  dist()

# Calculate pairwise genetic distances between individuals
dist_gen <- dist(genind_obj)


# create a tibble to combine vector-like dist data structures
GenGeoTable <- tibble(Genetic= dist_gen, Geographic= dist_geo)
```

Challenge 2: Create a plot comparing genetic and geographic distance Add a regression line to your scatterplot to help visualize the trend. Hint: use ggplot2::geom_smooth(method = "lm").

```
ggplot(GenGeoTable, aes(x=Geographic, y=Genetic)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  labs(
   x = "Geographic Distance",
   y = "Genetic Distance") +
  theme_minimal()
```

```
## Don't know how to automatically pick scale for object of type <dist>.
## Defaulting to continuous.
## Don't know how to automatically pick scale for object of type <dist>.
## Defaulting to continuous.
## `geom_smooth()` using formula = 'y ~ x'
```



Stop, Think, and Think:

If there were genetic isolation by geographic distance the relationship between the geographic distance and genetic isolation would be strongly positively correlated and residuals would be minimal and data concentrated more or less along the line of best fit.

Question Answers

a. The above plot suggests a slight positive relationship between geographic distance and genetic distance.
b. A speciation event resulting from habitat fragmentation could create a pattern where genetic distance increases with geographic distance.

Challenge 3: What kind of statistical methods could you use to formally test whether genetic distance increases with geographic distance? Justify the best method and statistically test the relationship.

The Mantel test formally compares two distance matrices to test whether they are correlated. In our case, one matrix represents genetic distances between individuals and the other represents geographic distances. A significant p-value suggests that individuals farther apart geographically are also more genetically different; a pattern expected under Isolation by Distance.

```
# Perform Mantel test between genetic and geographic distances
mantel_result <- mantel(dist_gen, dist_geo)
```

```r
# View the result
mantel_result
```

```
##
## Mantel statistic based on Pearson's product-moment correlation
##
## Call:
## mantel(xdis = dist_gen, ydis = dist_geo)
##
## Mantel statistic r: 0.1616
##       Significance: 0.003
##
## Upper quantiles of permutations (null model):
##    90%    95%  97.5%    99%
## 0.0635 0.0847 0.1012 0.1259
## Permutation: free
## Number of permutations: 999
```

With a Mantel correlation coefficient of 0.162 and a significance level of 0.003, there appears to be a statistically significant correlation between the geographic distance of the sample woodrat population and their genetic distance.

Discussion Question Answers

a. Heterozygosity is a measure of the proportion of individuals within a population that have two different alleles for a given locus. Observed heterozygosity is calculated as a proportion of the individuals in a population that are known to be heterozygous. Expected heterozygosity under Hardy-Weinberg equilibrium is calculated from genotype data by doubling the product of individual allele frequencies.
b. According to basic_stats, the overall Fst output for the comparison between north and south woodrat population genetics was 0.1148. This number under the context of a single locus with a similar individual Fst (basic_stats[["pop.freq"]][[129]]) corresponding to allele frequency differences of 0.3382, suggests that, on average, the north and south populations are genetically different from one another.
c. Fst quantifies the magnitude of genetic differences between groups using allele proportions, with larger Fst values indicating greater genetic dissimilarity between groups.
d. When using a Mantel test, a significant p-value means that there is a potential association between the two distances (in our case genetic and geographic).