Population Genomics - Assignment

Name: Dafnoudis Dimitris

Date: 11/07/2023

1. Introduction

The aim of this assignment is to explore the morphological and genetic differentiation of the 8 populations of the three-spined stickleback species (Gasterosteus aculeatus Linnaeus, 1758) as the heterozygosity. In order to achieve these goals, we'll use a vcf file and analyze the chromosome 5 (chromosome V) based on the single-nucleotide polymorphisms (SNPs). In addition we'll use a reference genome from the NCBI (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_016920845.1/).

The VCF file contains 192 samples from three-spined stickleback. There are eight populations. The L01, L02, L03, L05 are from brackish water and the L07, L09, L10, L12 are from the fresh water habitats and from each site approximately 24 individuals have been sequenced using WGS. Samples have been taken from the Belgian-Dutch lowlands. In this VCF file we have removed monomorphic SNPs to exclude all sites at which no alternative alleles are called for any of the samples and all sites at which only alternative alleles are called (all samples differ from the reference genome). Furthermore, multiallelic and low allele frequency (AF < 0.01) SNPs have also been removed.

2. Methods & Results

2.1 Inpecting the files and basic analysis

```
We will set our working directory based on our preferences and then
we'll import the R packages that we will use for the assignment.

setwd("C:/Users/dimit/Desktop/AppBio/7th_classes_Polulation_Genomics/
PopGen_Assignment")

library("vcfR")
library("DEMEtics")
library("adegenet")
library("ade4")
library("ggplot2")
library("ape")
library("StAMPP")

Importing our data
```

```
samplesInfo <- read.table("SamplesInfo.txt", sep = "\t", header =
TRUE)
sitesInfo <- read.table("SitesInfo.txt", sep = "\t", header = TRUE)
```

Basic information just to take a quick look of our data.

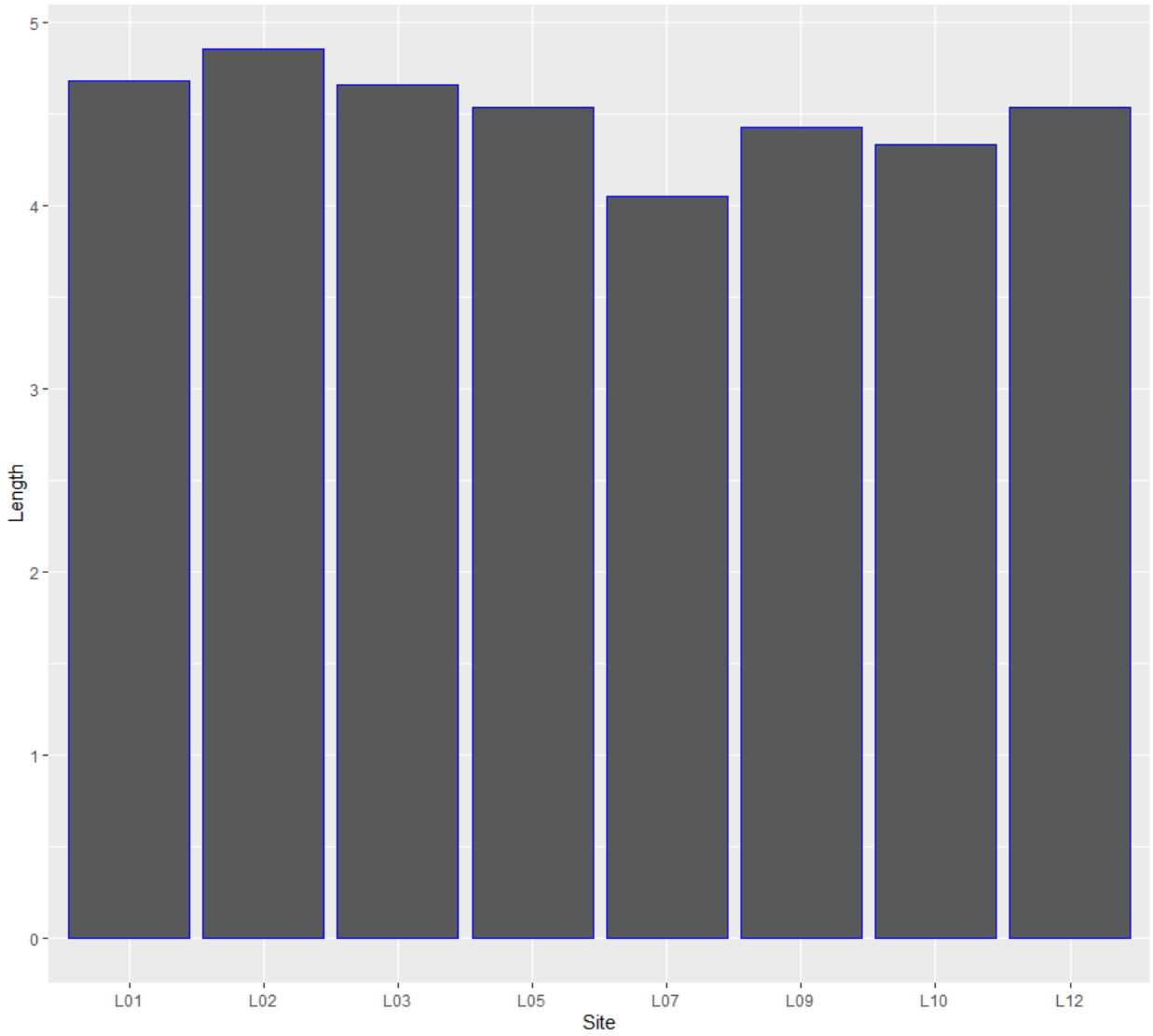
```
> head(samplesInfo)

          SeqID        ID Species Site
1 Sample_01-131 TMS_00014      3S  L12
2   Sample_01-2 TMS_00021      3S  L12
3 Sample_02-253 TMS_00338      3S  L03
4 Sample_02-350 TMS_00439      3S  L07
5   Sample_02-3 TMS_00056      3S  L12
6 Sample_03-133 TMS_00293      3S  L01
  Dissected Sex Length.cm. Body_weight.g.
1         1   M        4.9           1.96
2         1   F        5.2           2.22
3         1   F        5.8           4.40
4         1   M        4.4           0.95
5         3   F        5.4           2.14
6         1   M        4.2           0.99

> head(sitesInfo, 8)

  Site Latitude Longitude       Habitat
1  L01 51.35791  3.444077 Brackishwater
2  L02 51.37349  3.524563 Brackishwater
3  L03 51.28850  3.594585 Brackishwater
4  L05 51.24066  3.303227 Brackishwater
5  L07 51.23914  3.654950    Freshwater
6  L09 51.19932  3.666218    Freshwater
7  L10 51.18721  3.399005    Freshwater
8  L12 51.21092  3.517319    Freshwater
```

Before we dive into the VCF file we can perform some basic commands in order to provide insights about our data
Therefore, we merge the two txt files by Site, find the mean values of each and we use ggplot for visualizaton.

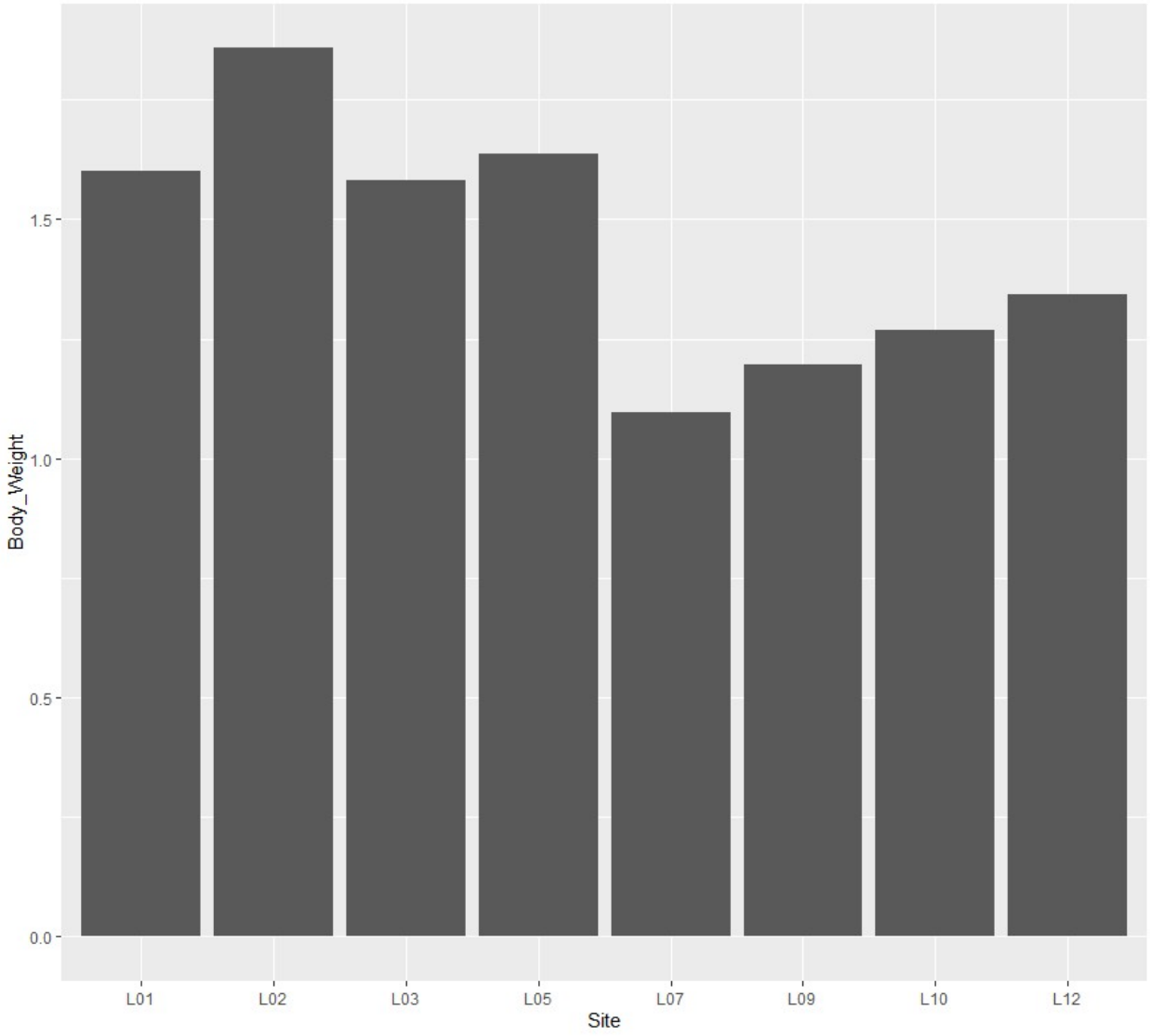
```
merge_data <- merge(samplesInfo, sitesInfo, by = "Site")

# Find the mean of Length and Body_weight based on the site of the
merged_data
site_means <- aggregate(cbind(samplesInfo$Length,
samplesInfo$Body_weight) ~ Site, merge_data, mean)
colnames(site_means) <- c("Site","Length","Body_Weight")

> site_means

  Site   Length Body_Weight
```

```
1  L01 4.683333    1.602083
2  L02 4.854545    1.858182
3  L03 4.658333    1.582083
4  L05 4.537500    1.637500
5  L07 4.054167    1.096667
6  L09 4.429167    1.196667
7  L10 4.337500    1.269167
8  L12 4.541667    1.344583
```

```r
# Visualization of Length and Body_Weight per Landscape
ggplot(site_means, aes(x=Site, y = Length)) +
  geom_bar(colour="blue", stat = "identity")

ggplot(site_means, aes(x=Site, y = Body_Weight)) +
  geom_bar(colour="red", stat = "identity")
```



We can see that based on the length of each population, the L02 has the largest number and L07 has the lowest. Nonetheless, in total, it doesn't appear to have strong differences.



The body-weight seems that have some differences. Here L07 population has also the lowest number and also L02 has the highest. Overall brackish water seems heavier than the freshwater habitats.

Subset the samplesInfo file by Site

```
L01 <- subset(samplesInfo, Site =="L01")
L02 <- subset(samplesInfo, Site =="L02")
L03 <- subset(samplesInfo, Site =="L03")
L05 <- subset(samplesInfo, Site =="L05")
L07 <- subset(samplesInfo, Site =="L07")
```

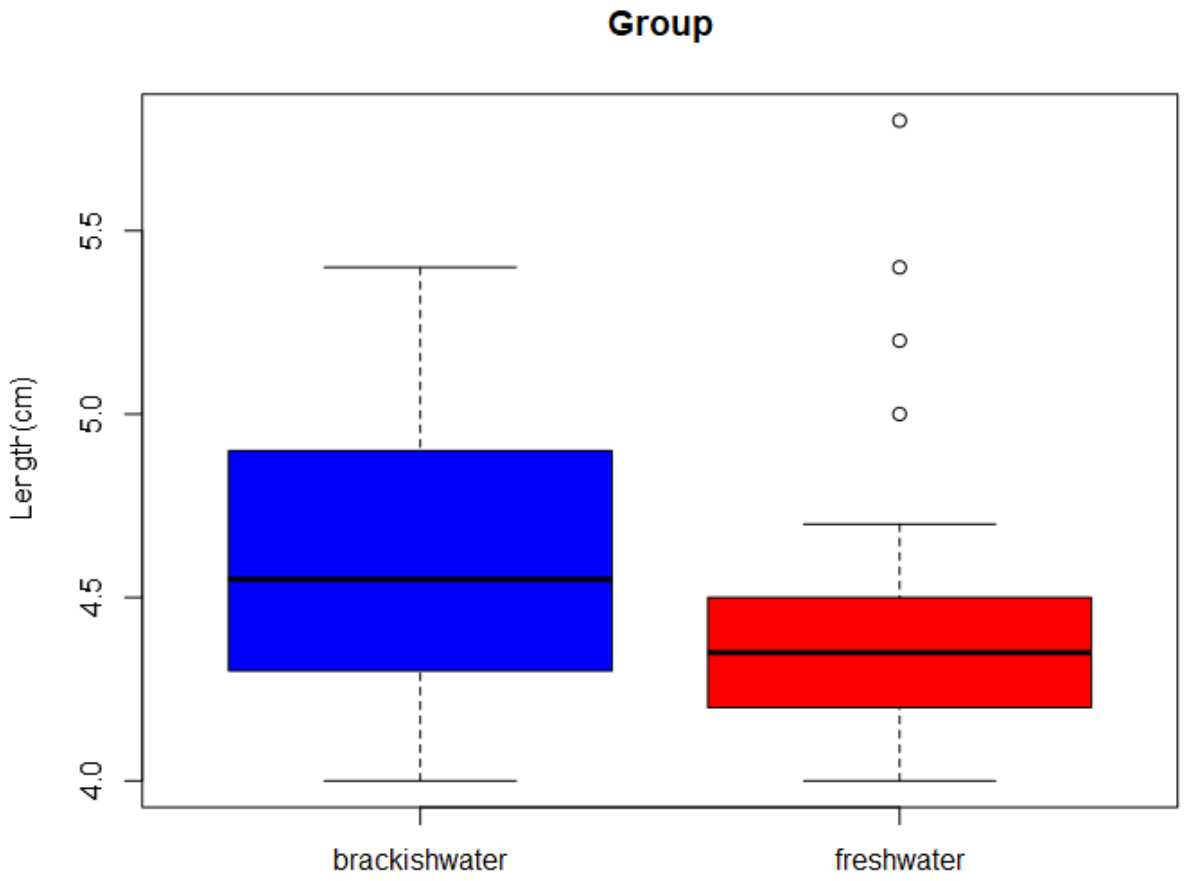
```
L09 <- subset(samplesInfo, Site =="L09")
L10 <- subset(samplesInfo, Site =="L10")
L12 <- subset(samplesInfo, Site =="L12")
```

Brackishwater and freshwater

```
brackishwater <- c(L01, L02, L03, L05)
freshwater <- c(L07, L09, L10, L12)
```

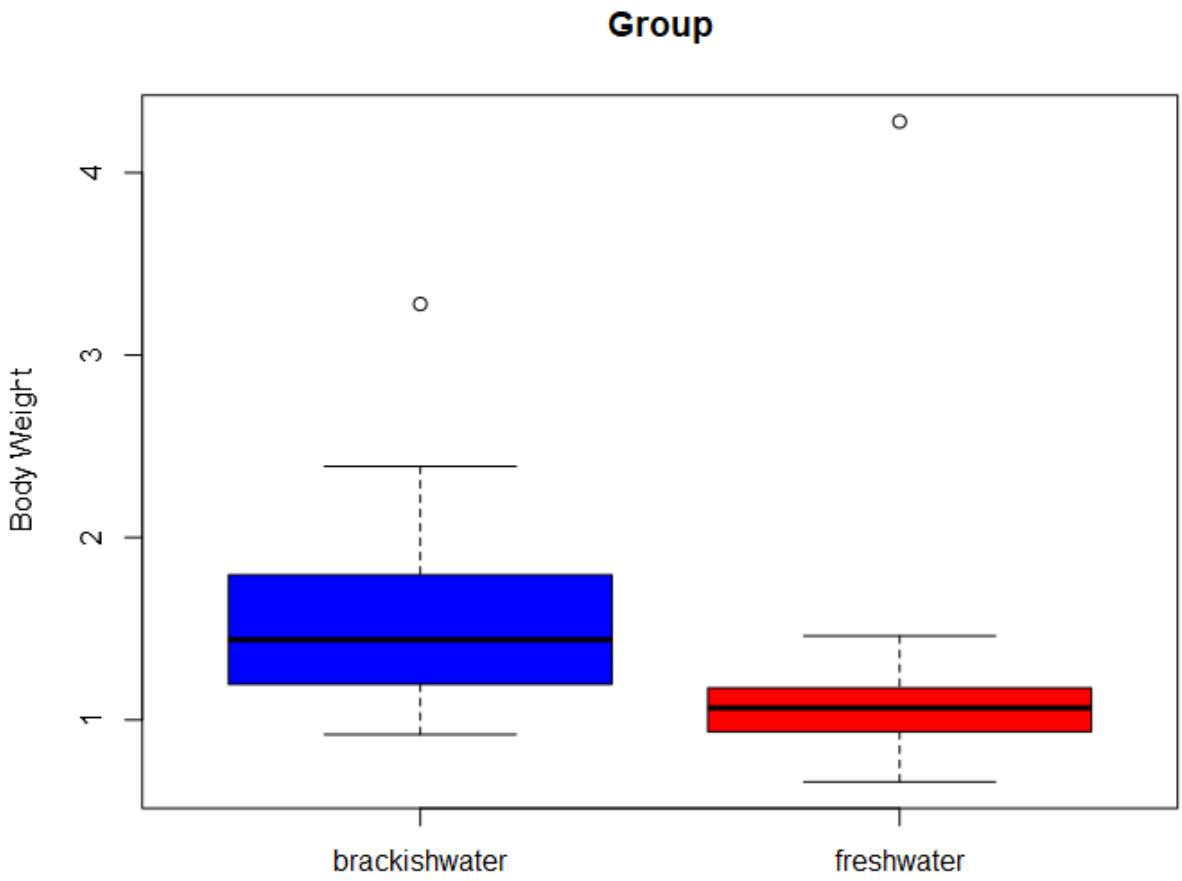
Visualization of Brackishwater and freshwater by length

```
boxplot(brackishwater$Length.cm., freshwater$Length.cm.,
        names = c("brackishwater", "freshwater"),
        col = c("blue", "red"),
        ylab = "Length(cm)",
        main = "Group")
```



Brackishwater and freshwater by Body Weight

```
boxplot(brackishwater$Body_weight.g., freshwater$Body_weight.g.,
        names = c("brackishwater", "freshwater"),
        col = c("blue", "red"),
        ylab = "Body Weight",
        main = "Group")
```



2.2 Identify genetic structure and diversity in the chromosome 5 (chromosome V) of the three-spined stickleback using SNPs

2.2.1 Subset Chromosome V

Subset the chromosome V from the ThreeSpined.vcf.gz file

```
vcf<- readLines("ThreeSpined.vcf.gz")
filtered_vcf <- vcf[grep("^##|^#CHROM|^chrV\t", vcf)]
writeLines(filtered_vcf, "chrV.vcf")
```

Information about chromosome V

```
> chrV <- read.vcfR("chrV.vcf")

Scanning file to determine attributes.
File attributes:
  meta lines: 73
  header_line: 74
  variant count: 127403
  column count: 201
Meta line 73 read in.
All meta lines processed.
gt matrix initialized.
Character matrix gt created.
  Character matrix gt rows: 127403
  Character matrix gt cols: 201
  skip: 0
  nrows: 127403
  row_num: 0
Processed variant: 127403
All variants processed
```

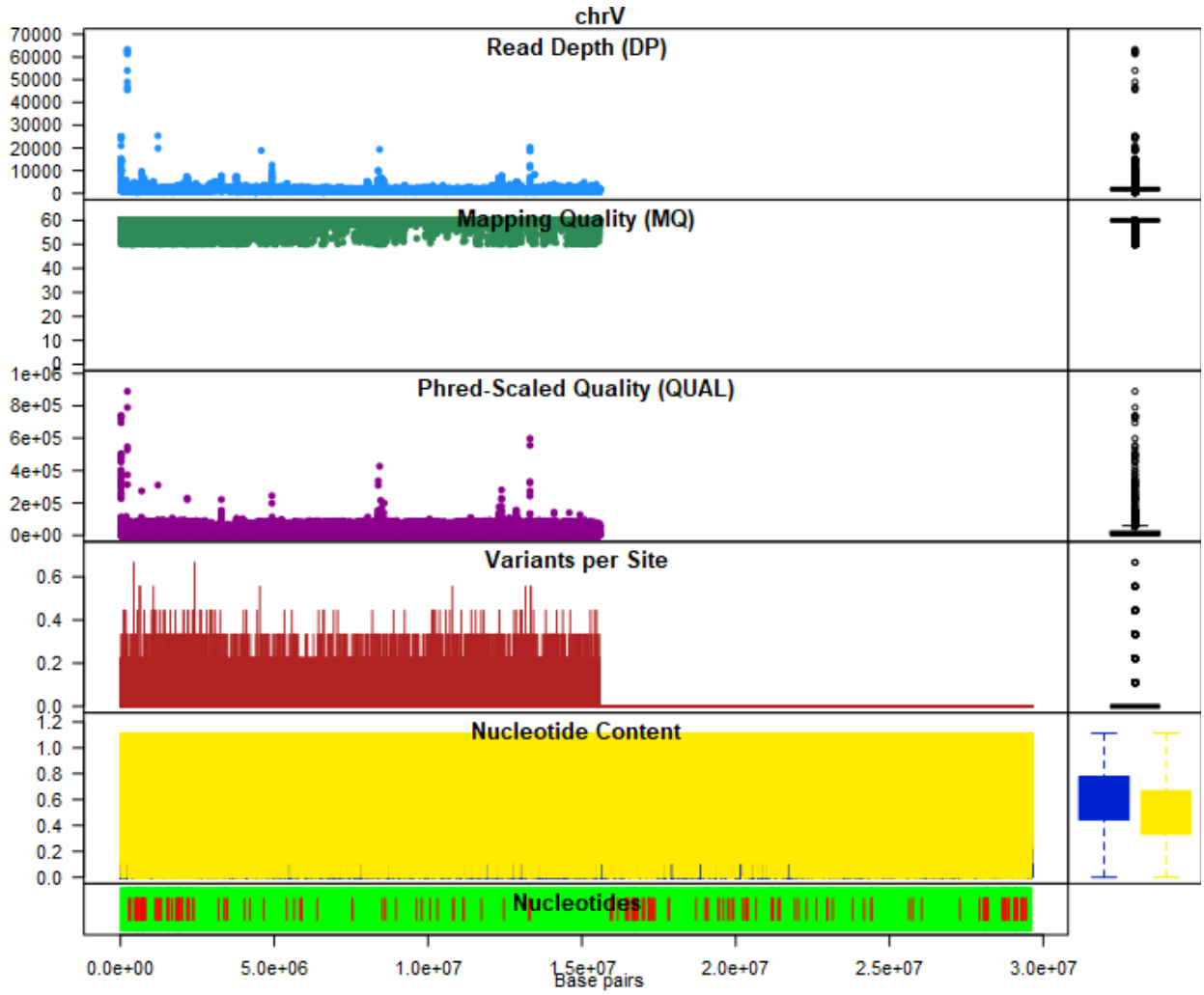
Structure of Chromosome V

```
> str(chrV)

Formal class 'vcfR' [package "vcfR"] with 3 slots
  ..@ meta: chr [1:73] "##fileformat=VCFv4.2"
"##FILTER=<ID=PASS,Description=\"All filters passed\">"
"##ALT=<ID=NON_REF,Description=\"Represents any possible alternative
allele not already represented at this loca"| __truncated__
"##FILTER=<ID=FAIL_FS60,Description=\"FS > 60.0\">" ...
  ..@ fix : chr [1:127403, 1:8] "chrV" "chrV" "chrV" "chrV" ...
  .. ..- attr(*, "dimnames")=List of 2
  .. .. ..$ : NULL
  .. .. ..$ : chr [1:8] "CHROM" "POS" "ID" "REF" ...
  ..@ gt  : chr [1:127403, 1:193] "GT:AD:DP:GQ:PGT:PID:PL:PS"
```

```
"GT:AD:DP:GQ:PL" "GT:AD:DP:GQ:PGT:PID:PL:PS"
"GT:AD:DP:GQ:PGT:PID:PL:PS" ...
  .. ..- attr(*, "dimnames")=List of 2
  .. .. ..$ : NULL
  .. .. ..$ : chr [1:193] "FORMAT" "Sample_01-131" "Sample_01-2"
"Sample_02-253" ...
```

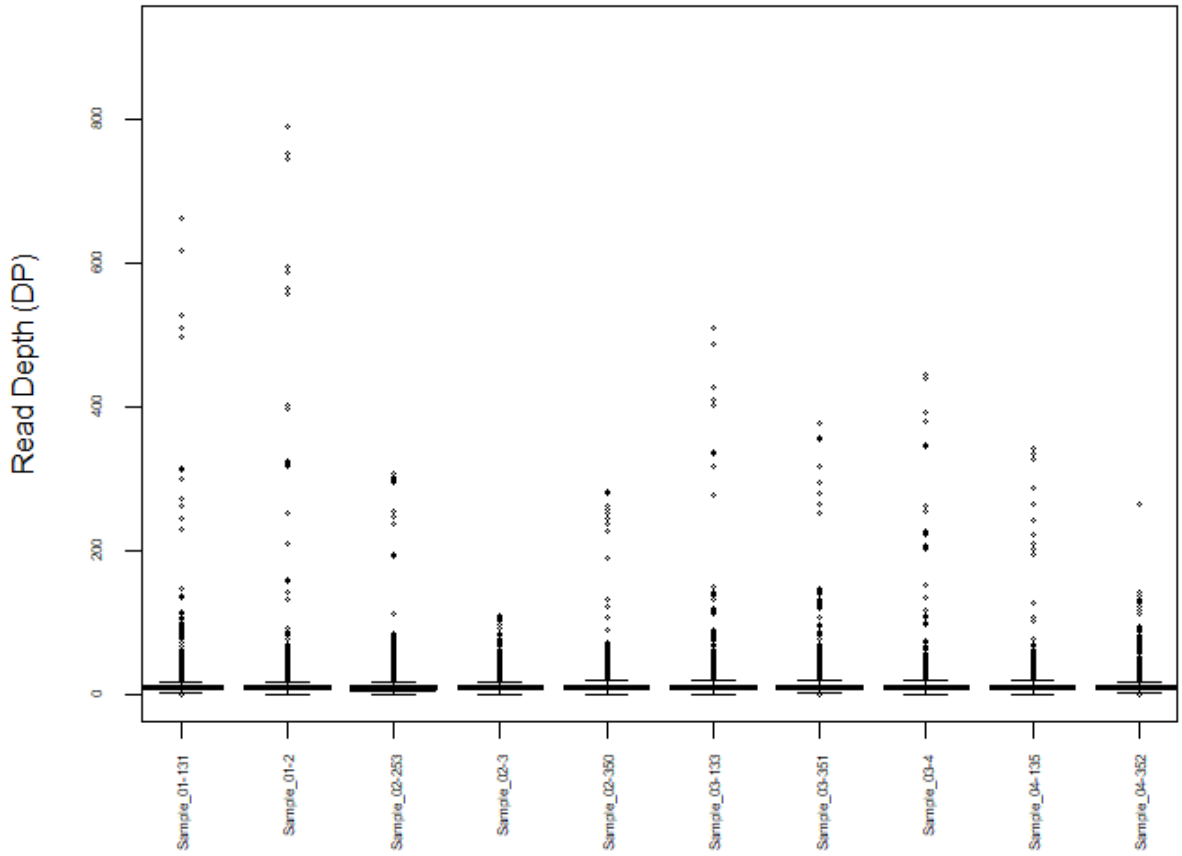
2.2.2 Visual overview of the SNP data

```
dna_file <-
read.dna("RefGenome/ncbi_dataset/data/GCF_016920845.1/GCF_016920845.1_
GAculeatus_UGA_version5_genomic.fasta",format="fasta")
chrom <- create.chromR(name = "chrV", vcf = chrV, seq = dna_file,
verbose = TRUE)
chrom <- proc.chromR(chrom, verbose = TRUE,win.size=10)

# summarize the data from our fasta and vcf files
chromoqc(chrom)
```



2.2.3 Extract the allele depths for each sample (DP field of chrV) and plot its distribution

```
dp <- extract.gt(chrV, element='DP', as.numeric=TRUE)
boxplot(dp, las=3,
        col=c("#C0C0C0", "#808080"),
        ylab="Read Depth (DP)",
        cex=0.4, cex.axis=0.5,
        xlim=c(1,10))
```



2.2.4 Genlight objects

```
genlight.data <- vcfR2genlight(chrV)

# Get the definition of a class (getClassDef)

> getClassDef("genlight") # Content of a genlight object:

Class "genlight" [package "adegenet"]
```

```
Slots:


Name:           gen        n.loc    ind.names    loc.names
loc.all   chromosome
Class:         list      integer   charOrNULL   charOrNULL
charOrNULL factorOrNULL


Name:      position       ploidy          pop       strata
hierarchy        other
Class:    intOrNULL    intOrNULL factorOrNULL     dfOrNULL
formOrNULL         list
```

Show some info of the genlight object

```
> indNames(genlight.data)

  [1] "Sample_01-131" "Sample_01-2"
  [3] "Sample_02-253" "Sample_02-3"
  [5] "Sample_02-350" "Sample_03-133"
  [7] "Sample_03-351" "Sample_03-4"
  [9] "Sample_04-135" "Sample_04-352"
                 ...
[183] "Sample_92-175" "Sample_93-127"
[185] "Sample_93-209" "Sample_93-248"
[187] "Sample_94-128" "Sample_94-249"
[189] "Sample_95-250" "Sample_95-346"
[191] "Sample_95-390" "Sample_96-251"

> nLoc(genlight.data) # number of SNPs

[1] 127403
```

In order to explore the genetic differentiation we will need the number of the SNPs (in this case the 127403).

```
# Adding information about the population membership
# and the ploidy of each sample.
pops <- as.factor(c(
    "L12", "L12", "L03", "L07", "L12", "L01",
"L07", "L12", "L01", "L07", "L12", "L01",
"L07", "L01", "L03", "L01", "L03", "L07",
"L03", "L07", "L02", "L07", "L02", "L03",
"L03", "L09", "L01", "L03", "L09", "L09",
"L10", "L01", "L03", "L09", "L12", "L09",
"L10", "L03", "L09", "L10", "L10", "L12",
"L03", "L01", "L02", "L12", "L03", "L03",
"L12", "L03", "L02", "L12", "L02", "L09",
"L09", "L12", "L03", "L10", "L12", "L01",
```

```
"L03", "L03", "L03", "L01", "L12", "L01",
"L01", "L12", "L09", "L01", "L09", "L03",
"L09", "L03", "L09", "L12", "L09", "L01",
"L01", "L10", "L12", "L05", "L01", "L12",
"L05", "L12", "L05", "L01", "L03", "L02",
"L05", "L03", "L02", "L02", "L02", "L03",
"L03", "L02", "L02", "L02", "L03", "L05",
"L02", "L02", "L05", "L10", "L05", "L02",
"L05", "L05", "L10", "L10", "L05", "L05",
"L10", "L05", "L05", "L02", "L10", "L07",
"L07", "L05", "L10", "L07", "L05", "L12",
"L05", "L01", "L05", "L10", "L02", "L07",
"L12", "L02", "L07", "L02", "L07", "L02",
"L05", "L12", "L05", "L10", "L01", "L10",
"L05", "L01", "L05", "L01", "L07", "L07",
"L05", "L07", "L01", "L01", "L01", "L02",
"L05", "L02", "L07", "L09", "L09", "L02",
"L09", "L10", "L07", "L02", "L07", "L10",
"L09", "L07", "L10", "L07", "L10", "L07",
"L10", "L09", "L07", "L10", "L10", "L10",
"L09", "L12", "L10", "L12", "L05", "L09",
"L12", "L09", "L09", "L07", "L09", "L09"
))


pop(genlight.data) <- pops
ploidyvalues <- rep(2,192)
ploidy(genlight.data) <- ploidyvalues
```

Here, we look at sample 20 to 30 and SNP 1 to 5.

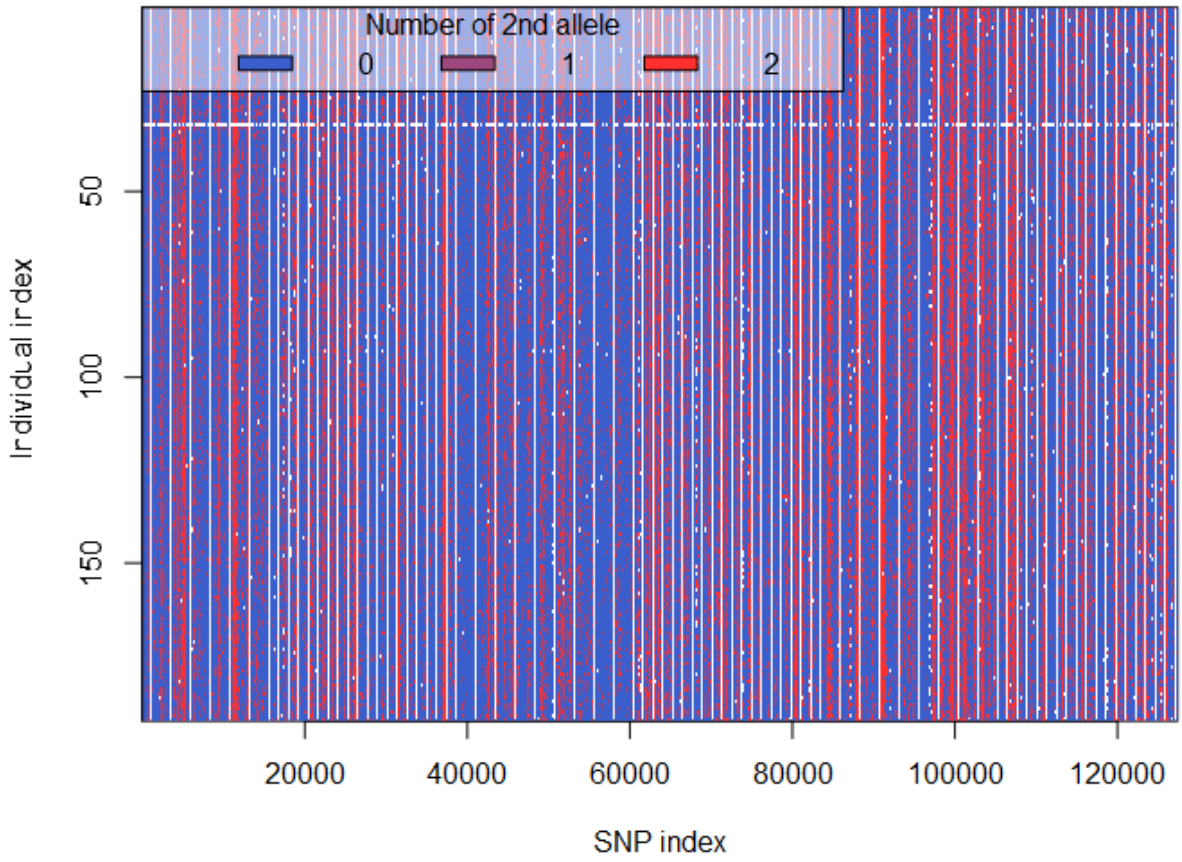
```
> as.matrix(genlight.data)[20:30,1:5]
              chrV_20518 chrV_21532
Sample_09-357          0          0
Sample_10-152          1          1
Sample_10-358          1          0
Sample_12-158          1          0
Sample_12-263          0          0
Sample_13-264          1          0
Sample_13-361          1          0
Sample_14-23           1          0
Sample_14-265          1          0
Sample_14-362          1          0
Sample_15-363          1          0
              chrV_21782 chrV_21792
Sample_09-357          0          0
Sample_10-152          0          1
Sample_10-358          0          2
Sample_12-158          0          1
Sample_12-263          0          2
```

```
Sample_13-264          0          2
Sample_13-361          0          2
Sample_14-23           0          0
Sample_14-265          0          0
Sample_14-362          0          1
Sample_15-363          0          0
              chrV_21972
Sample_09-357          0
Sample_10-152          0
Sample_10-358          0
Sample_12-158          0
Sample_12-263          2
Sample_13-264          0
Sample_13-361          0
Sample_14-23           0
Sample_14-265          0
Sample_14-362          1
Sample_15-363          0
```

```
    The original matrix of biallelic SNPs is stored in a way of one
number per individual per site that reflects the number of alternative
alleles in that site in that individual (i.e. 0, 1, or 2 in a diploid
individual).
```

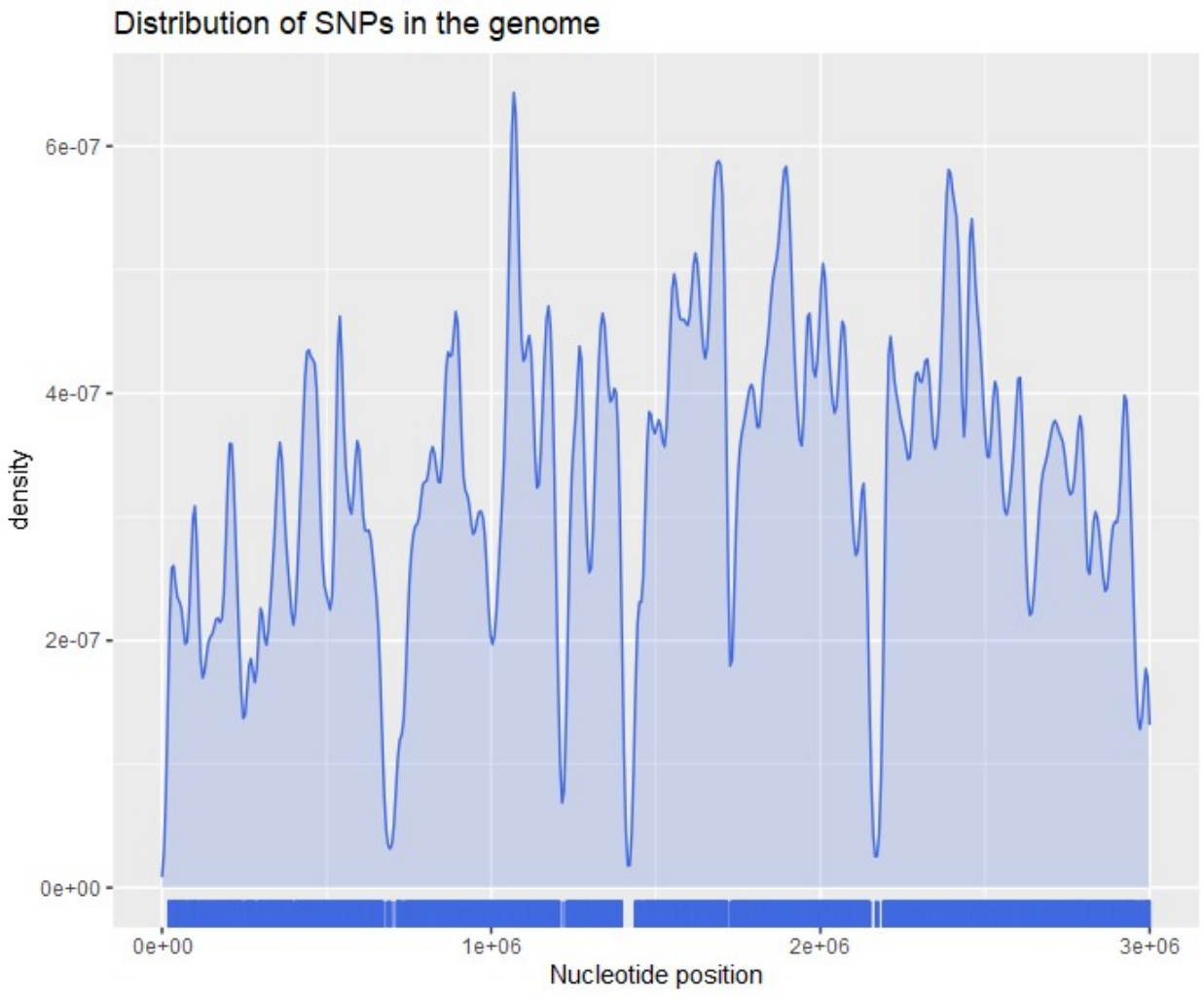
A graphical overview of alternative alleles and missing data (in white) can be obtained with the glPlot function:

```
glPlot(genlight.data, posi="topleft")
```



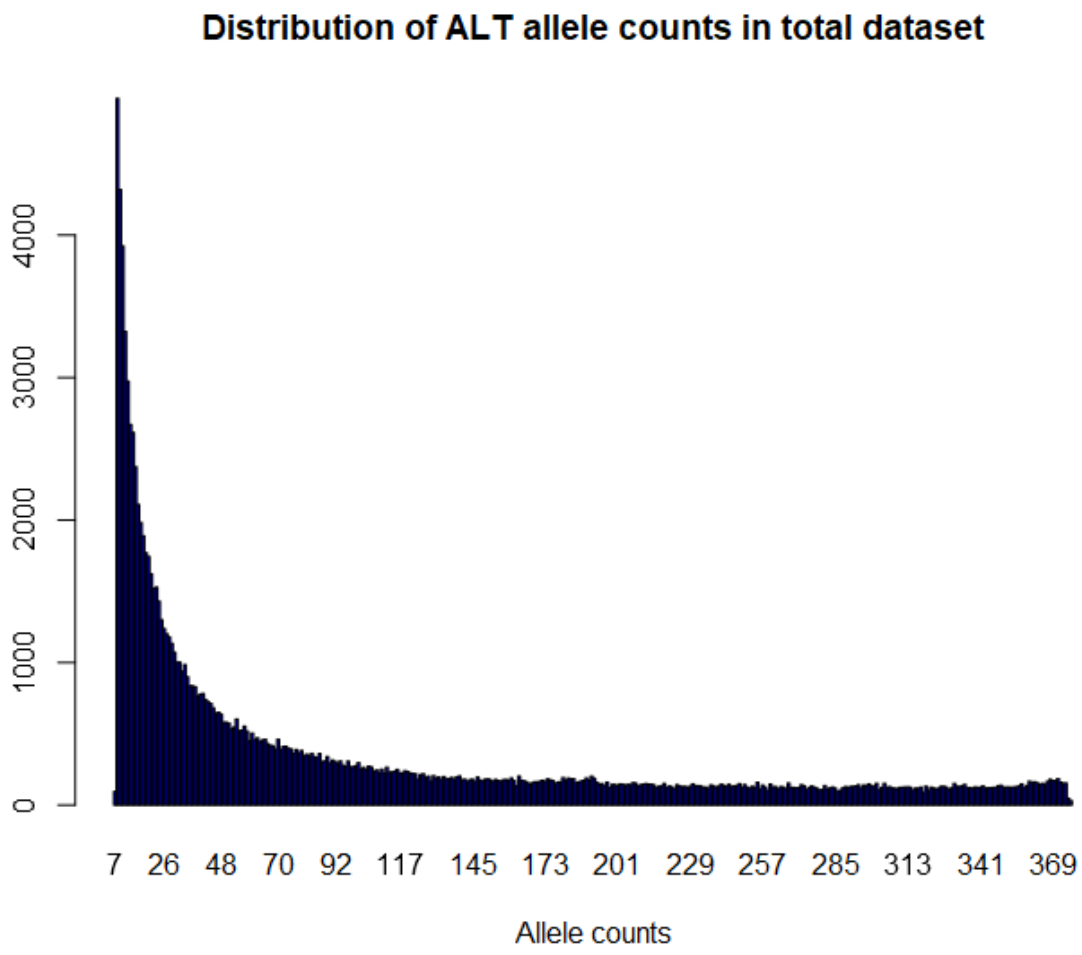
Assess the position of the polymorphic sites within the chromosome graphically

```
snpposi.plot(position(genlight.data[,genlight.data$chromosome=="chrV"]
),genome.size=3000000,codon=FALSE)
```



2.2.5 Allele frequency spectrum

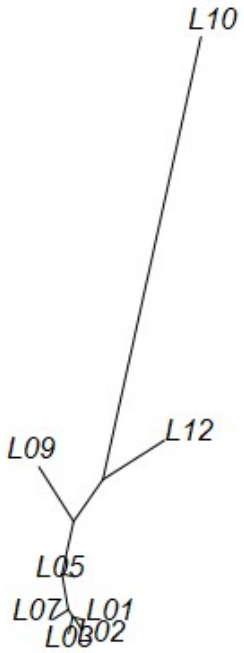
```
# plot total AFS of the dataset
mySum <- glSum(genlight.data, alleleAsUnit = TRUE) # Computes the sum
of second alleles for each SNP.
barplot(table(mySum), col="blue", space=0, xlab="Allele counts",
        main="Distribution of ALT allele counts in total dataset")
```



2.2.6 Genetic Differentiation

```
genlight.data.reduced <- genlight.data[,sample(1:127403, 50000)]

> genlight.data.reduced #checking basic information

 /// GENLIGHT OBJECT /////////

 // 192 genotypes,  50,000 binary SNPs, size: 6.7 Mb
 87528 (0.91 %) missing data

 // Basic content
   @gen: list of 192 SNPbin
   @ploidy: ploidy of each individual  (range: 2-2)

 // Optional content
   @ind.names:  192 individual labels
   @loc.names:  50000 locus labels
   @chromosome: factor storing chromosomes of the SNPs
   @position: integer storing positions of the SNPs
```

```
   @pop: population of each individual (group size range: 24-24)
   @other: a list containing: elements without names


> FstValues <- stamppFst(genlight.data.reduced, nboots = 100, percent
= 95)
> FstValues$Fsts
           L12         L03        L07         L01        L02
L09       L10 L05
L12          NA          NA         NA          NA         NA
NA        NA  NA
L03 0.07305464          NA         NA          NA         NA
NA        NA  NA
L07 0.06653382 0.013199426         NA          NA         NA
NA        NA  NA
L01 0.07015893 0.008965339 0.01263091          NA         NA
NA        NA  NA
L02 0.07124323 0.010100635 0.01235857 0.007641408         NA
NA        NA  NA
L09 0.06151285 0.054896756 0.05122086 0.052923757 0.05387723
NA        NA  NA
L10 0.15877466 0.188546516 0.18426305 0.187434783 0.18929747
0.16700181        NA  NA
L05 0.04958611 0.014059967 0.01430625 0.012256477 0.01420418
0.03219301 0.1635454  NA

# Library(ape)
# Required package to visualize the tree using the "nj" function

Tree <- nj(as.dist(FstValues$Fsts)) # conversion of the Fst values to
a tree object
plot.phylo(Tree,type="radial",show.tip.label=TRUE,edge.width=1,rotate.
tree=140)
```



2.2.7 Heterozygosity

```
# We will use the adegenet package
genind.data <- vcfR2genind(chrV)

# Specify the populations
pops <- as.factor(c(
    "L12", "L12", "L03", "L07", "L12", "L01",
"L07", "L12", "L01", "L07", "L12", "L01",
"L07", "L01", "L03", "L01", "L03", "L07",
"L03", "L07", "L02", "L07", "L02", "L03",
"L03", "L09", "L01", "L03", "L09", "L09",
"L10", "L01", "L03", "L09", "L12", "L09",
"L10", "L03", "L09", "L10", "L10", "L12",
"L03", "L01", "L02", "L12", "L03", "L03",
"L12", "L03", "L02", "L12", "L02", "L09",
"L09", "L12", "L03", "L10", "L12", "L01",
"L03", "L03", "L03", "L01", "L12", "L01",
"L01", "L12", "L09", "L01", "L09", "L03",
```

```
"L09", "L03", "L09", "L12", "L09", "L01",
"L01", "L10", "L12", "L05", "L01", "L12",
"L05", "L12", "L05", "L01", "L03", "L02",
"L05", "L03", "L02", "L02", "L02", "L03",
"L03", "L02", "L02", "L02", "L03", "L05",
"L02", "L02", "L05", "L10", "L05", "L02",
"L05", "L05", "L10", "L10", "L05", "L05",
"L10", "L05", "L05", "L02", "L10", "L07",
"L07", "L05", "L10", "L07", "L05", "L12",
"L05", "L01", "L05", "L10", "L02", "L07",
"L12", "L02", "L07", "L02", "L07", "L02",
"L05", "L12", "L05", "L10", "L01", "L10",
"L05", "L01", "L05", "L01", "L07", "L07",
"L05", "L07", "L01", "L01", "L01", "L02",
"L05", "L02", "L07", "L09", "L09", "L02",
"L09", "L10", "L07", "L02", "L07", "L10",
"L09", "L07", "L10", "L07", "L10", "L07",
"L10", "L09", "L07", "L10", "L10", "L10",
"L09", "L12", "L10", "L12", "L05", "L09",
"L12", "L09", "L09", "L07", "L09", "L09"
))


pop(genlight.data) <- pops
```

Use the Hs function to obtain the average heterozygosity for each population

```
Hs(genind.data)

      L01       L02       L03       L05       L07       L09       L10
0.2254799 0.2258295 0.2248733 0.2246247 0.2250070 0.2251745 0.2254469
      L12
0.2184511
```

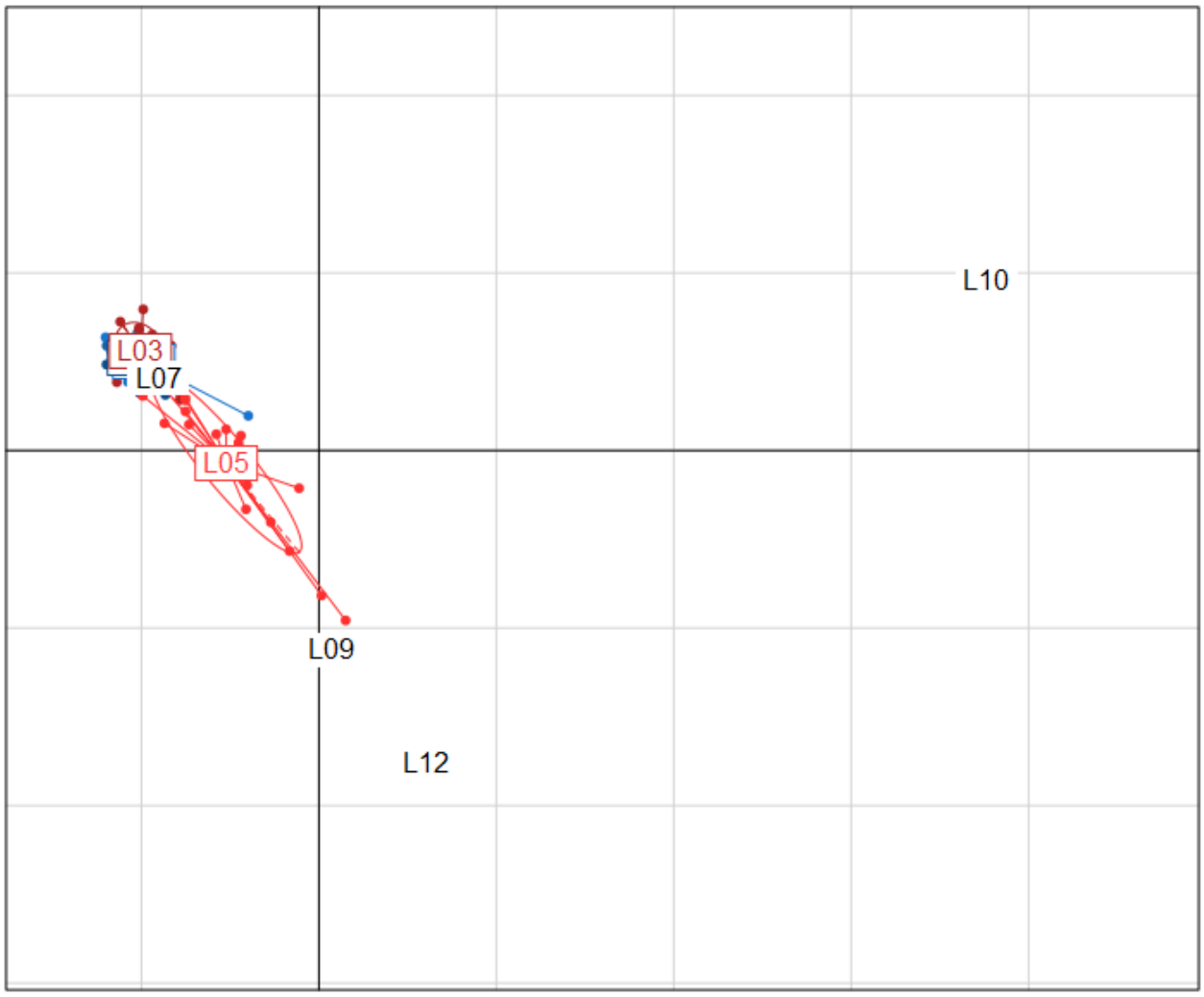
2.2.8 Principal Component Analysis (PCA)

```
# when nf=2 (number of retained factors) is not specified,
# the function displays the barplot of eigenvalues
# of the analysis and asks the user for a number of
# retained principal components.
pca.1 <- glPca(genlight.data.reduced, nf=2)

> pca.1$eig[1]/sum(pca.1$eig) # proportion of variation explained by
1st principal component

[1] 0.07517911

> pca.1$eig[2]/sum(pca.1$eig) # proportion of variation explained by
2nd principal component
```

```
[1] 0.02916904
```

Plot the samples along the first two principle components showing groups

```
s.class(pca.1$scores, pop(genlight.data.reduced), col=colors()
[c(131,132,133,134)])
```



3. Discusion

Based on our analysis, the conclusions about our findigs are:

- The Lenght between the populations of the habitats are not different. Although, the body-weight of the brackish water habitat individually and in total is heavier.
- The Fst value shows that more genetically distant are the L10 and L12 from other populations.

- The phylogenetic tree and the Principal Component Analysis (PCA) also confirms that L10 is genetically different from the rest of the populations and also L12. The L01, L02, L03, L07 are clustered while L05, L09 are distictly presented.