ProyectoUnidad5

Bárbara Cruz Salazar

2/24/2020

## R Markdown

## Install Bioconductor

if (!requireNamespace("BiocManager", quietly = TRUE))  
 install.packages("BiocManager")  
BiocManager::install(version = "3.10")

## Bioconductor version 3.10 (BiocManager 1.30.10), R 3.6.2 (2019-12-12)

## Old packages: 'gsl'

BiocManager::install(c("gdsfmt","SNPRelate"))

## Bioconductor version 3.10 (BiocManager 1.30.10), R 3.6.2 (2019-12-12)

## Installing package(s) 'gdsfmt', 'SNPRelate'

##   
## The downloaded binary packages are in  
## /var/folders/\_y/ccz7ddn963dg6ngl\_dnz1pm40000gn/T//RtmpLGfdnE/downloaded\_packages

## Old packages: 'gsl'

## Call libreries

library(SNPRelate)

## Loading required package: gdsfmt

## SNPRelate -- supported by Streaming SIMD Extensions 2 (SSE2)

library(ape)  
library(ggplot2)

## Change session

## Convert .bed, .fam, .bim

### It is converted with plink, from .ped file; to have the .ped file you must convert the .vcf file to .ped with vcftools

## Load dates in gds format from plink   
snpgdsBED2GDS("../data/maicesArtegaetal2015.bed",   
 "../data/maicesArtegaetal2015.fam",   
 "../data/maicesArtegaetal2015.bim",   
 out.gdsfn="../data/maices.gds", # here indicate name of gds file  
 option = snpgdsOption(Z=10)) # 10 cromosomas, this information is know

## Start file conversion from PLINK BED to SNP GDS ...  
## BED file: '../data/maicesArtegaetal2015.bed'  
## SNP-major mode (Sample X SNP), 1.5M  
## FAM file: '../data/maicesArtegaetal2015.fam'  
## BIM file: '../data/maicesArtegaetal2015.bim'  
## Mon Feb 24 17:07:55 2020 (store sample id, snp id, position, and chromosome)  
## start writing: 165 samples, 36931 SNPs ...  
## [..................................................] 0%, ETC: --- [==================================================] 100%, completed, 0s  
## Mon Feb 24 17:07:55 2020 Done.  
## Optimize the access efficiency ...  
## Clean up the fragments of GDS file:  
## open the file '../data/maices.gds' (1.6M)  
## # of fragments: 39  
## save to '../data/maices.gds.tmp'  
## rename '../data/maices.gds.tmp' (1.6M, reduced: 252B)  
## # of fragments: 18

# See summary   
snpgdsSummary("../data/maices.gds")

## Some values of 'snp.chromosome' are invalid (should be finite and >= 0)!

## Hint: specifying 'autosome.only=FALSE' in the analysis could avoid detecting chromosome coding.

## Some of 'snp.allele' are not standard (e.g., 0/G).

## The file name: /Users/barbaracruz-salazar/Desktop/BioinfinvRepro\_viejo/Unidad5/Prac\_Uni5/maices/data/maices.gds   
## The total number of samples: 165   
## The total number of SNPs: 36931   
## SNP genotypes are stored in SNP-major mode (Sample X SNP).  
## The number of valid samples: 165   
## The number of biallelic unique SNPs: 30804

# Load gds file created from .bed, .fam and .bim files  
gdsfile <- snpgdsOpen("../data/maices.gds")  
  
# Check snp.ids  
head(read.gdsn(index.gdsn(gdsfile, "snp.id")))

## [1] "abph1.15" "ae1.8" "an1.3" "ba1.5" "ba1.7" "csu1138.4"

# Check sample.ids  
head(read.gdsn(index.gdsn(gdsfile, "sample.id")))

## [1] "maiz\_3" "maiz\_68" "maiz\_91" "maiz\_39" "maiz\_12" "maiz\_41"

# Create an object to names the samples of gdsn  
sample.id <- read.gdsn(index.gdsn(gdsfile, "sample.id"))  
sample.id

## [1] "maiz\_3" "maiz\_68" "maiz\_91" "maiz\_39" "maiz\_12" "maiz\_41"   
## [7] "maiz\_35" "maiz\_58" "maiz\_51" "maiz\_82" "maiz\_67" "maiz\_93"   
## [13] "maiz\_21" "maiz\_6" "maiz\_101" "maiz\_27" "maiz\_43" "maiz\_1"   
## [19] "maiz\_33" "maiz\_100" "maiz\_24" "maiz\_103" "maiz\_72" "maiz\_10"   
## [25] "maiz\_28" "maiz\_49" "maiz\_56" "maiz\_66" "maiz\_52" "maiz\_47"   
## [31] "maiz\_80" "maiz\_65" "maiz\_94" "maiz\_36" "maiz\_26" "maiz\_105"   
## [37] "maiz\_30" "maiz\_16" "maiz\_42" "maiz\_4" "maiz\_31" "maiz\_17"   
## [43] "maiz\_46" "maiz\_5" "maiz\_32" "maiz\_19" "maiz\_50" "maiz\_8"   
## [49] "maiz\_34" "maiz\_23" "maiz\_54" "maiz\_14" "maiz\_37" "maiz\_25"   
## [55] "maiz\_55" "maiz\_40" "maiz\_29" "maiz\_60" "maiz\_44" "maiz\_74"   
## [61] "maiz\_89" "maiz\_64" "maiz\_83" "maiz\_75" "maiz\_92" "maiz\_69"   
## [67] "maiz\_84" "maiz\_76" "maiz\_97" "maiz\_70" "maiz\_85" "maiz\_77"   
## [73] "maiz\_71" "maiz\_86" "maiz\_78" "maiz\_102" "maiz\_73" "maiz\_88"   
## [79] "maiz\_79" "maiz\_106" "maiz\_119" "maiz\_148" "maiz\_108" "maiz\_120"   
## [85] "maiz\_151" "maiz\_134" "maiz\_123" "maiz\_153" "maiz\_135" "maiz\_124"   
## [91] "maiz\_184" "maiz\_140" "maiz\_125" "maiz\_141" "maiz\_131" "maiz\_189"   
## [97] "maiz\_57" "maiz\_126" "maiz\_113" "maiz\_138" "maiz\_63" "maiz\_127"   
## [103] "maiz\_114" "maiz\_139" "maiz\_99" "maiz\_129" "maiz\_116" "maiz\_142"   
## [109] "maiz\_110" "maiz\_132" "maiz\_118" "maiz\_144" "maiz\_133" "maiz\_121"   
## [115] "maiz\_111" "maiz\_137" "maiz\_109" "maiz\_146" "maiz\_164" "maiz\_157"   
## [121] "maiz\_171" "maiz\_149" "maiz\_165" "maiz\_159" "maiz\_172" "maiz\_150"   
## [127] "maiz\_166" "maiz\_160" "maiz\_173" "maiz\_152" "maiz\_167" "maiz\_161"   
## [133] "maiz\_174" "maiz\_154" "maiz\_169" "maiz\_162" "maiz\_176" "maiz\_156"   
## [139] "maiz\_170" "maiz\_163" "maiz\_177" "maiz\_178" "maiz\_192" "maiz\_185"   
## [145] "maiz\_201" "maiz\_179" "maiz\_193" "maiz\_186" "maiz\_202" "maiz\_180"   
## [151] "maiz\_195" "maiz\_187" "maiz\_181" "maiz\_197" "maiz\_188" "maiz\_182"   
## [157] "maiz\_198" "maiz\_190" "maiz\_200" "maiz\_183" "maiz\_191" "teos\_96"   
## [163] "teos\_203" "teos\_911" "teos\_9107"

# Metadates  
# load  
metadata<- read.delim(file= "../meta/maizteocintle\_SNP50k\_meta\_extended.txt")  
  
# check  
head(metadata)

## OrderColecta NSiembra Origen Raza Estado  
## 1 2 3 INIFAP-2009-16 Apachito Chihuahua  
## 2 1 68 2009-72 Chalque̱o Tlaxcala  
## 3 1 91 2007-33 Dulcillo del Noroeste Sonora  
## 4 2 39 Chis-2009-18 Dzit-Bacal Chiapas  
## 5 2 12 Celaya-2009-114 Celaya Guanajuato  
## 6 2 41 Celaya-2009-2 Elotes Occidentales Guanajuato  
## Num\_Colecta Nombre\_comun Raza\_Primaria Raza\_Secundaria  
## 1 16 Ocho carreras Apachito NA  
## 2 22 Chalque̱o Criollo Chalque̱o NA  
## 3 SON2007-033 Ma\xcc\_z Dulce Dulcillo del Noroeste NA  
## 4 Repetida 18 Olotillo crema Dzit-Bacal NA  
## 5 2009-REPO-114 Ma\xcc\_z Criollo Celaya NA  
## 6 2009-REPO-002 Ma\xcc\_z Criollo Elotes Occidentales NA  
## A.o.\_de\_colecta Localidad Municipio Estado.1 Longitud  
## 1 2009 Santo Tom\xcc\xc1s Guerrero Chihuahua -107.58300  
## 2 2008 Ignacio Zaragoza Cuapiaxtla Tlaxcala -97.92944  
## 3 2007 Agua Blanca Y̩cora Sonora -108.92467  
## 4 2009 Nuevo Vicente Guerrero Villacorzo Chiapas -92.97972  
## 5 2009 El Ahuacate Uriangato Guanajuato -101.11889  
## 6 2009 Comonfort Comonfort Guanajuato -100.76583  
## Latitud Altitud Ruizetal2008\_grupo  
## 1 28.68578 1975 1A\_templado540-640mm  
## 2 19.29083 2497   
## 3 28.53703 1435 2A\_semicalido500-870mm  
## 4 16.03225 618 3A\_muycalido990-1360mm  
## 5 20.09111 1877 2A\_semicalido500-870mm y 1B\_templado>650mm  
## 6 20.74417 1800 2C\_semicalido740-855mm y 1B\_templado>650mm  
## Sanchezetal\_grupo Categ.Altitud Rbiogeo  
## 1 Sierra de Chihuahua mid Sierra Madre Occidental  
## 2 C\xcc\_nico high Eje Volcanico  
## 3 Chapalote mid Sierra Madre Occidental  
## 4 Maduraci\xcc\_n tard\xcc\_a low Costa del Pacifico  
## 5 Dentados tropicales mid Eje Volcanico  
## 6 Ocho Hileras mid Eje Volcanico  
## DivFloristic PeralesBiog  
## 1 Sierra Madre Occidental Ca\xe5\_ones Chi  
## 2 Serranias Meridionales Mesa Centra  
## 3 Sierra Madre Occidental Sierras del  
## 4 Serranias Transismicas Chiapas  
## 5 Altiplanicie Baj\xe5\xc1o  
## 6 Altiplanicie Baj\xe5\xc1o

nrow(metadata)

## [1] 165

head(metadata$NSiembra) # ID number of the samples

## [1] 3 68 91 39 12 41

head(sample.id)

## [1] "maiz\_3" "maiz\_68" "maiz\_91" "maiz\_39" "maiz\_12" "maiz\_41"

# Do PCA

pca <- snpgdsPCA(gdsfile, num.thread=2)

## Principal Component Analysis (PCA) on genotypes:  
## Excluding 5,290 SNPs on non-autosomes  
## Excluding 837 SNPs (monomorphic: TRUE, MAF: NaN, missing rate: NaN)  
## Working space: 165 samples, 30,804 SNPs  
## using 2 (CPU) cores  
## PCA: the sum of all selected genotypes (0,1,2) = 2543593  
## CPU capabilities: Double-Precision SSE2  
## Mon Feb 24 17:07:56 2020 (internal increment: 2156)  
## [..................................................] 0%, ETC: --- [==================================================] 100%, completed, 0s  
## Mon Feb 24 17:07:56 2020 Begin (eigenvalues and eigenvectors)  
## Mon Feb 24 17:07:56 2020 Done.

# Calculate the variation percentaje for first components   
pc.percent <- pca$varprop\*100 # multiply by 100 to see the value in percentage  
head(round(pc.percent, 2)) # see the values with two digits

## [1] 6.44 2.13 1.42 1.19 1.10 1.08

# To round the values, use round function  
x<-round(pc.percent, 2) # 2 digits  
  
# See sum of comonents  
sum(x[1:4])

## [1] 11.18

sum(x[1:10])

## [1] 17.08

sum(x[1:30])

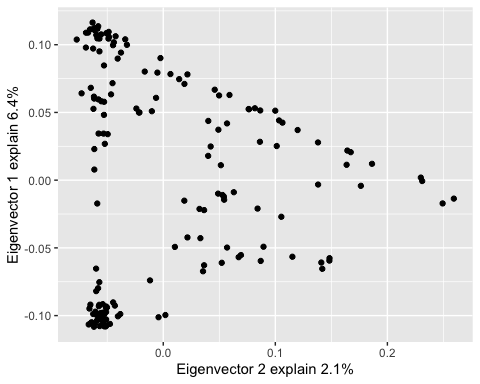
## [1] 31.92

# For do a plot I have to do a data frame  
table\_pca <- data.frame(sample.id = pca$sample.id, # sample.id of PCA in a column  
 EV1 = pca$eigenvect[,1], # the first eigenvector  
 EV2 = pca$eigenvect[,2], # the second eigenvector  
 stringsAsFactors = FALSE)  
head(table\_pca)

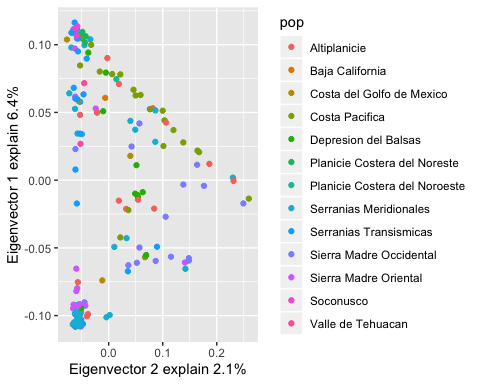
## sample.id EV1 EV2  
## 1 maiz\_3 -0.05654163 0.11522741  
## 2 maiz\_68 -0.09909519 -0.05460776  
## 3 maiz\_91 -0.01718140 0.24928583  
## 4 maiz\_39 0.10705139 -0.05925765  
## 5 maiz\_12 -0.01515252 0.01879029  
## 6 maiz\_41 -0.02124046 0.03231120

## Do a plot

ggplot(data = table\_pca, aes(x=EV2, y=EV1)) + geom\_point() + # plot will be of points  
 ylab(paste0("Eigenvector 1 explain ", round(pc.percent, 1)[1], "%")) + # round will put the rounding you did up  
 xlab(paste0("Eigenvector 2 explain ", round(pc.percent, 1)[2], "%")) # 1 indicate one digit; the number in brackets is the component

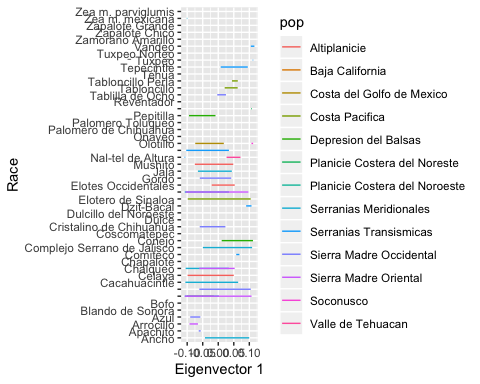


### paste0 paste  
  
# Load metadates for PCA  
group <- as.vector(metadata$DivFloristic) # Do a vector with the information of geographic zone to group the genetic data with that factor  
  
# Do the data frame again including the factor (DivFloristic)  
table\_pca\_group <- data.frame(sample.id = pca$sample.id,  
 pop = factor(group)[match(pca$sample.id, sample.id)], # pop will be the factor (group),   
 # match function makes it fit the data frame considering the new column included,   
 # and that the data is not messy.   
 EV1 = pca$eigenvect[,1], # the first eigenvector  
 EV2 = pca$eigenvect[,2], # the second eigenvector  
 stringsAsFactors = FALSE)  
  
# Do plot of PCA  
  
ggplot(data = table\_pca\_group, aes(x=EV2, y=EV1, colour = pop)) + geom\_point() + #Indicate that color according to the group that was created (group=DivFloristic)  
 ylab(paste0("Eigenvector 1 explain ", round(pc.percent, 1)[1], "%")) +  
 xlab(paste0("Eigenvector 2 explain ", round(pc.percent, 1)[2], "%"))



# Do an others plots

# Do an other object  
race\_by\_state <- as.vector(metadata$Raza)  
  
# Do an other data frame that include the race  
table\_two <- data.frame(sample.id=pca$sample.id, pop = factor(group)[match(pca$sample.id, sample.id)], group=factor(race\_by\_state)[match(pca$sample.id, sample.id)],  
 EV1 = pca$eigenvect[,1])  
  
  
ggplot (data =table\_two, aes(x=EV1, y=group,colour=pop)) + geom\_line() +  
 labs (x= "Eigenvector 1", y = "Race")



ggplot (data = metadata, aes(x=Raza, y=Altitud, col=DivFloristic)) + geom\_line() +  
 labs (x="Raza", y="Estado")

