

**UNIVERSIDADE FEDERAL DE VIÇOSA  
CENTRO DE CIÊNCIAS AGRÁRIAS  
DEPARTAMENTO DE ZOOTECNIA**

## **TUTORIAL**

# **Análise de corridas de homozigose e assinaturas de seleção no programa PLINK**

Renata de Fátima Bretanha Rocha

Pamela Itajara Otto

Arielly Oliveira Garcia

Mateus Guimarães dos Santos

Marcos Vinicius Gualberto Barbosa da Silva

Marta Fonseca Martins

Marco Antonio Machado

João Cláudio do Carmo Panetto

Simone Eliza Facioni Guimarães

**Viçosa**

**2023**

# TUTORIAL

## Análise de corridas de homozigose e assinaturas de seleção no programa PLINK

Renata de Fátima Bretanha Rocha<sup>1</sup>, Pamela Itajara Otto<sup>2</sup>, Arielly Oliveira Garcia<sup>1</sup>, Mateus Guimarães dos Santos<sup>1</sup>, Marcos Vinicius Gualberto Barbosa da Silva<sup>3</sup>, Marta Fonseca Martins<sup>3</sup>, Marco Antonio Machado<sup>3</sup>, João Cláudio do Carmo Panetto<sup>3</sup>, Simone Eliza Facioni Guimarães<sup>1</sup>

<sup>1</sup>Departamento de Zootecnia, Universidade Federal de Viçosa, Viçosa, MG 36570-900, Brasil.

<sup>2</sup>Departamento de Zootecnia, Universidade Federal de Santa Maria, Santa Maria, RS 97105-900, Brasil.

<sup>3</sup>EMBRAPA – Gado de Leite, Juiz de Fora, MG 36038-330, Brasil

**ISBN:** 978-65-5668-138-2

**DOI:** <http://dx.doi.org/10.26626/9786556681382.2023B0001>

### Agradecimentos

Agradecemos às fazendas e à Empresa Brasileira de Pesquisa Agropecuária (Embrapa) – Gado de Leite, Juiz de Fora – MG, que forneceram os dados para este estudo.

### Financiamento

Este estudo recebeu o apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil (CNPQ) - Processos 402935/2021-7, 142600/2019-9 e 200147/2022-6, da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)/PROEX 88887.844747/2023-00 e do Instituto Nacional de Ciência e Tecnologia de Ciência Animal (INCT-CA).

**Este tutorial é para ajudar a fazer uma análise de corridas de homozigose (ROH) e assinaturas de seleção pela metodologia FST usando o software PLINK, por exemplo, para quem e não sabe por onde começar.**

**Este tutorial não é definitivo! Estude sobre o assunto para saber qual o tipo de análise se adequa melhor à sua pesquisa e como ela pode ser realizada!**

## SUMÁRIO

1. Arquivos necessários para a análise de ROH.....	3
1.1 O arquivo de mapa .....	3
1.2 O arquivo de pedigree .....	3
1.3 O programa PLINK.....	4
2. Como montar os arquivos .ped e .map a partir de um arquivo de genótipos usados nos programas do Misztal .....	6
2.1 Transformando o arquivo de dados .csv para .txt.....	6
2.2 Formatando arquivo de pedigree/genótipo.....	7
2.3 Formatando arquivo de mapa no software R.....	13
3. Análise de ROH no PLINK.....	15
3.1 Rodando a análise .....	15
3.2 Resultados da análise .....	20
3.3 Como cada análise é descrita no artigo .....	24
3.4 Número e Tamanho de ROHs por cromossomo e classe .....	24
3.5 Genome Coverage por cromossomo .....	27
4. Calculando a endogamia baseada em ROH (FROH) .....	28
4.1 FROH por cromossomo e por classe .....	28
4.2 Variação de uma característica com a FROH .....	29
Referências.....	31
5. Análise de FST – assinatura de seleção.....	33
Resultados .....	33

## 1. Arquivos necessários para a análise de ROH

- Arquivo de mapa
- Arquivo de pedigree

Os arquivos de mapa e fenótipos/genótipos são diferentes dos usados em GWAS. Este tutorial mostra a formatação dos arquivos necessários para a análise de ROH a partir de arquivos pré-existentes para uma análise nos programas do Misztal. As formatações são feitas no software R e no LINUX.

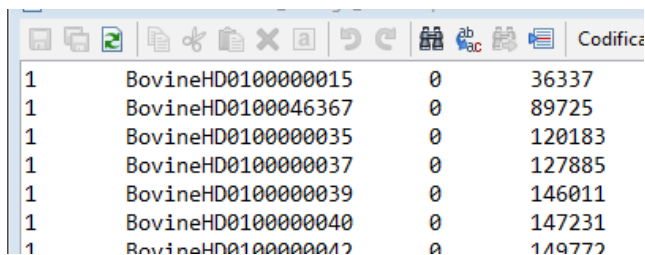
Em seguida tem o passo a passo para realizar a análise de ROH, o cálculo da endogamia baseada em ROH (FROH) e, também, para realizar a análise de FST.

### 1.1 O arquivo de mapa

As colunas do arquivo de mapa tem que ser todas separadas por espaço ou todas separadas por TAB. Por padrão, cada linha do arquivo de mapa descreve um único marcador e deve conter exatamente quatro colunas:

- Cromossomo
- rs# ou identificador/nome do SNP
- Distância genética em morgans
- Posição em pares de bases (unidades de pb)

Neste caso a coluna de distância genética em morgans está zerada, pois não preciso dessa informação para esta análise.



1	BovineHD0100000015	0	36337
1	BovineHD0100046367	0	89725
1	BovineHD0100000035	0	120183
1	BovineHD0100000037	0	127885
1	BovineHD0100000039	0	146011
1	BovineHD0100000040	0	147231
1	BovineHD0100000042	0	149772

Os arquivos de mapa e de pedigree devem ter o mesmo nome, mas o de mapa deve ter a extensão .map e o de pedigree .ped. Exemplo:

- arquivo123.map
- arquivo123.ped

### 1.2 O arquivo de pedigree

As colunas do arquivo de pedigree tem que ser todas separadas por espaço ou todas separadas por TAB. As primeiras seis colunas são obrigatórias:

- Family ID (identificação da família – se não tiver, deixar 0)
- Individual ID (identificação do indivíduo)
- Paternal ID (identificação do pai – se não tiver, deixar 0)
- Maternal ID (identificação da mãe – se não tiver, deixar 0)
- Sexo (1=macho; 2=fêmea; outro número=desconhecido)
- Fenótipo (para essa análise não precisa, então deixar é 0)
- Em seguida vem as colunas com as letras dos genótipos (vai ter uma coluna para cada SNP, então é um arquivo muito grande)

Exemplo:

0	AB52502	Tgir-824	Gir-2697	2	0	A	B	A	B	A	B	A	A	B
0	AB52503	Tgir-879	RRP6537	2	0	A	B	A	A	A	B	A	A	B
0	AB52505	Tgir-1037	Gir-4774	2	0	A	B	A	B	A	B	A	A	B
0	AB52508	Tgir-1399	Gir-2705	2	0	A	B	A	A	B	B	A	B	B
0	AB52511	Tgir-1009	Gir-716	2	0	A	B	B	B	A	B	A	A	A
0	AB52512	Tgir-1009	Ne21RRP6419	2	0	A	B	A	B	A	B	A	A	B
0	AB52514	Tgir-1295	Gir-716	2	0	A	A	A	B	A	A	B	A	B
0	AB52516	Tgir-1437	RRP6395	2	0	A	B	A	B	A	B	A	A	A
0	AB52517	Tgir-879	Gir-2805	2	0	A	B	A	B	A	B	A	A	B

### 1.3 O programa PLINK

Com a versão 1.07 do PLINK, pode ser feita a análise de ROH. Com a versão 1.09 do PLINK, além da análise de ROH também pode ser feita a análise de FST

- Mais informações em: <https://zzz.bwh.harvard.edu/plink/data.shtml>
- Ou só pesquisar: “PLINK: Whole genome data analysis toolset”

### Referências

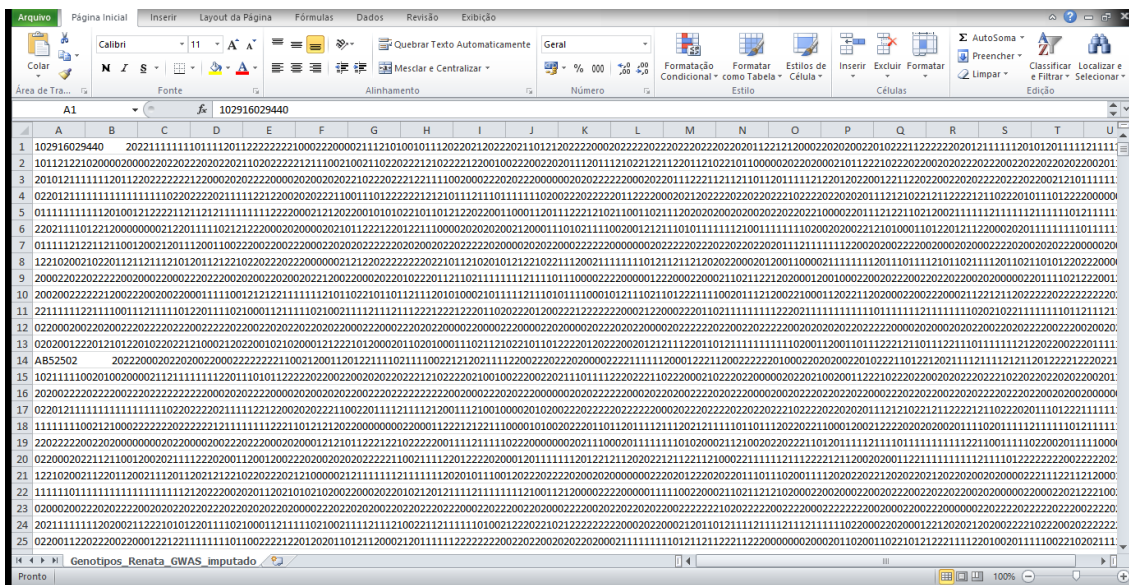
- Para a versão 1.07 do PLINK:  
Purcell S, Neale B, Todd-Brown K, et al (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 81:559–575. <https://doi.org/10.1086/519795>
- Para a versão 1.09 do PLINK:

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4:s13742–015–0047–8. <https://doi.org/10.1186/s13742-015-0047-8>

## 2. Como montar os arquivos .ped e .map a partir de um arquivo de genótipos usados nos programas do Misztal

Como estes arquivos são grandes, pode ser necessário usar um servidor LINUX. No servidor pode-se instalar o software R e usá-lo para montar os arquivos ou usar o próprio sistema LINUX para montar os arquivos.

*Arquivo de genótipos no formato para programas do Misztal: genotipos.csv*



O arquivo de genótipos acima contém duas colunas: na primeira tem a identificação dos animais e na segunda tem os genótipos (formato para serem usados em programas do Misztal para análise de GWAS e seleção genômica, por exemplo).

O arquivo deste exemplo contém 420.718 marcadores SNPs nas colunas e 2.093 animais nas linhas. Como o arquivo é grande, não é viável manipular os dados pela planilha no Excel. Arquivos maiores que este pode ser que nem abram neste formato.

### 2.1 Transformando o arquivo de dados .csv para .txt

#### a. Se o arquivo de genótipos for pequeno

Alguns arquivos de genótipos para exercícios em sala de aula, por exemplo, podem conter pouca quantidade de animais e marcadores, apenas para fins didáticos. É possível copiar o conteúdo para o Notepad++ e salvar o arquivo.

#### b. No software R



```

geno<-read.csv(genotipos.csv) # importando os dados
library(gdata) # abrindo a biblioteca para rodar o comando abaixo
write.fwf(geno, file = './genotipos.txt', rownames=F, colnames=F, quote=F, sep = ““,
justify='left') # salvando o arquivo no formato para os programas do Msztal

```

*c. No LINUX*

O comando abaixo pega o arquivos em .csv e salva em .txt

```
tr ',' '\n' < genotipos.csv > genotipos.txt
```

## 2.2 Formatando arquivo de pedigree/genótipo

O arquivo de genótipos do PLINK é uma junção do pedigree e dos genótipos nos animais, além de mais algumas informações, como visto no tópico 1.2. Então a partir do arquivo de pedigree e de genótipos no formato dos programas do Msztal, podemos formar o arquivo para ser usado no PLINK.

A seguir são apresentadas duas formas (a. No software R) e (b. Parte no software R e outra parte no LINUX) para obter o arquivo de pedigree/genótipo para PLINK.

*a. No software R*

Neste caso, o servidor já tem o software R instalado. A forma de abrir o programa no servidor do exemplo é apenas digitando:

R

Clicar na tecla 'ENTER'

```

renata.rocha@galloway: ~
renata.rocha@galloway:~$ R

R version 3.6.3 (2020-02-29) -- "Holding the Windssock"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> █

```

Como mostrado na imagem acima o servidor vai direcionar para dentro do software R.

Em seguida, usar os comandos do R para montar o arquivo:

```

rm(list=ls()) #Limpando o environment (conteúdo do R)
options(stringsAsFactors=F) # Este é um argumento para a função 'data.frame()' no R.
É uma lógica que indica se as strings em um quadro de dados devem ser tratadas como
variáveis de fator ou apenas como strings simples.
setwd("/home/pasta_com_arquivos") # Direcionando para a pasta
ped<-read.table("ped.txt",h=F,sep="\t") # Importando arquivo de pedigree
dim(ped) # Verificando a dimensão do arquivo: número de linhas e de colunas
colnames(ped)<-c("ID","PAI","MAE","SEX") # Atribuindo nomes nas colunas

```

A função `read.table()` demoraria muito pra importar o arquivo de genótipos, por isso, usar o seguinte:

```

library('data.table')
genotipos <- fread("genotipos.txt",h=F) # Importando arquivo de genótipos
colnames(genotipos)<-c("ID","GEN") # Atribuindo nomes nas colunas do arquivo
dim(genotipos) # Verificando a dimensão do arquivo: número de linhas e de colunas

```

```
ped1<-merge(ped, genotipos[,1], by = intersect("ID", "ID")) # Pegando o pedigree só
dos animais que tem genótipo
```

```
head(ped1) #Visualizando as primeiras 6 linhas do arquivo
```

```
### Recodificando os genótipos – nesta parte estamos transformando os números em
letras e separando todos por espaço simples
```

```
# Se estiver rodando a análise no R do Windows, pode deixar as letras dentro do
comando gsub separadas por TAB.
```

```
# Assim: gsub("0", "A      A", x)
```

```
# Se estiver rodando a análise no R no servidor, ao copiar e colar a linha de comando lá
o espaçamento do TAB não aparecer, então tem que deixar separado por espaço
simples.
```

```
geno_total<-as.data.frame(genotipos[,2])
```

```
total1<-data.frame(lapply(geno_total, function(x) {gsub("0", "A A", x)}))
```

```
total2<-data.frame(lapply(total1, function(x) {gsub("1", "A B", x)}))
```

```
total3<-data.frame(lapply(total2, function(x) {gsub("2", "B B", x)}))
```

```
total4<-data.frame(lapply(total3, function(x) {gsub("AA", "A A", x)}))
```

```
total5<-data.frame(lapply(total4, function(x) {gsub("BB", "B B", x)}))
```

```
total6<-data.frame(lapply(total5, function(x) {gsub("AB", "A B", x)}))
```

```
total7<-data.frame(lapply(total6, function(x) {gsub("BA", "B A", x)}))
```

```
dim(total7) # As linhas representam o número de animais, mas vai ter apenas uma
coluna.
```

```
### Formando o arquivo para o PLINK
```

```
seq0<-rep(0,nrow(geno_total)) # Essa sequência de zeros ficará no lugar das colunas de
família e de ano de nascimento
```

```
arquivo123_esp<-cbind(seq0, ped1, seq0, total7) # Juntando as colunas
```

```
# No arquivo formado terão as seguintes colunas nesta sequência:
```

```
# Coluna zerada que seria a identificação da família
```

```
# Identificação do indivíduo
```

```
# Identificação do pai
```

```
# Identificação da mãe
```

```
# Sexo
```

```

# Coluna zerada (que seria o fenótipo)
# As colunas seguintes são as letras dos genótipos

#### PARA SALVAR ####

### Se estiver rodando a análise no R do Windows:
write.table(arquivo123_esp," arquivo123_esp.ped", quote = F, row.names = F,
col.names = F, sep = "\t")
# Este arquivo será salvo com as colunas separadas por TAB
# Assim temos o arquivo de pedigree+genótipos para o LINUX!!! (com apenas as
colunas de genótipos separadas por espaço)

### Se estiver rodando a análise no R no servidor:
write.table(arquivo123_esp," arquivo123_esp.ped",quote=F,row.names=F,col.names=F)
# Este arquivo será salvo com todas as colunas separadas por espaço simples
# Agora para pegar esse arquivo e separar as colunas por TAB:
# 1ª Opção -> Recomendada (mais rápida): sair do R no servidor
q() #comando para sair do R
Save workspace image? [y/n/c]:
n

# No LINUX, podemos separar as colunas do arquivo por TAB com o seguinte
comando:
sed -e 's/ /\t/g' arquivo123_esp.ped > arquivo123.ped
# Assim temos o arquivo de pedigree+genótipos para o LINUX!!!

# 2ª Opção -> Pode ser usada para arquivos pequenos, por exemplo, porque quanto
maior o arquivo mais pode demorar.
# No próprio R, importar o arquivo novamente (Ao importar novamente, os genótipos já
vêm divididos em colunas)
# Não é necessário fazer nenhuma edição no arquivo
setwd("/home/pasta_com_arquivos")

```

```
library('data.table')
arquivo123<-fread("arquivo123_esp.ped",h=F) # Importando o arquivo
# O comando acima demora um pouco (~ 5 a 10min) neste exemplo.
# No comando para salvar, usamos sep="\t" que separa todas as colunas por TAB
write.table(arquivo123," arquivo123.ped",quote=F,row.names=F,col.names=F,sep="\t")
# Assim temos o arquivo de pedigree/genótipos para o LINUX!!!
```

*b. Parte no software R e outra parte no LINUX*

### Essa primeira parte feita no software R é usada apenas para separar o pedigree dos animais que tem genótipo.

## Se já tiver o pedigree desses animais separados, pode pular a parte do script do R.

## Precisamos ter os arquivos:

# ped.txt (com animal, pai, mae e sexo) -> Se não tiver info de sexo e não precisar para a análise basta criar uma coluna de zeros

# ids.txt -> Arquivo que tem uma coluna com as identificações dos animais do arquivo de genótipos

Novamente, no servidor, digitar:

R

```
rm(list=ls()) #limpa todo o conteúdo/arquivos que existem no R
```

```
options(stringsAsFactors=F)
```

```
setwd("/home/pasta_com_arquivos") # Direcionando para a pasta onde estão os arquivos
```

```
ped<-read.table("ped.txt",h=F,sep="\t") # Importando arquivo de pedigree
```

```
colnames(ped)<-c("ID","PAI","MAE","SEX") # atribuindo os nomes de cada coluna do arquivo de pedigree
```

```
ids<-read.table("ids.txt",h=F) # Importando as identificações dos animais do arquivo de genótipos
```

```
colnames(ids)<-c("ID") # atribuindo o nome da coluna do arquivo de identificação
```

```
ped_roh<-merge(ids, ped, by=intersect("ID","ID")) # pegando os animais em comum do arquivo de identificações (genótipos) e do pedigree
```

```

dim(ped_roh) # mostra o número de linhas (número de animais) e 4 colunas (id, pai,
mae e sexo)
write.table(ped_roh, "ped_roh.ped", quote=F, row.names=F, col.names=F, sep="\t")
# Criando arquivo com sequência de zeros para usar no LINUX
seq0<-as.data.frame(rep(0,nrow(ped_roh)))
write.table(seq0,"seq0.txt",quote=F, row.names=F, col.names=F) # salvando o arquivo
q() #sair do R

## No LINUX ##
# Cuidado ao copiar comandos com TAB e colar no servidor - ele não cola o
espaçamento do TAB
# Fazer cada comando por vez, porque cada comando é um pouco lento.
awk '{ print $2 }' genotipos.txt > snps.ped #pegando a coluna só com os genótipos
sed -i 's/0/A A/g' snps.ped
sed -i 's/1/A B/g' snps.ped
sed -i 's/2/B B/g' snps.ped
sed -i 's/AA/A A/g' snps.ped
sed -i 's/BB/B B/g' snps.ped
sed -i 's/AB/A B/g' snps.ped
sed -i 's/BA/B A/g' snps.ped
## Formando o arquivo de genótipos para PLINK ##
#Juntando seq0, ped_roh, zeros, snps
# Esse arquivo seq0 foi criado no script do R (acima)
# Tem que ter certeza que os animais no arquivo de ped estão na mesma ordem do
arquivo de genótipos!
paste seq0.txt ped_roh.ped seq0.txt snps.ped > arquivo123_esp.ped

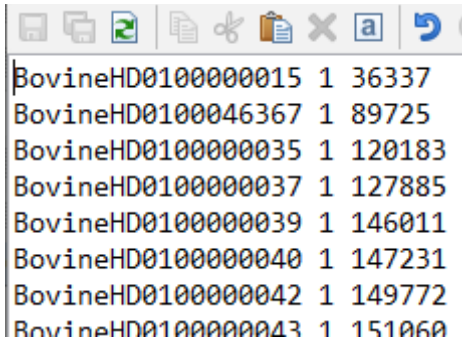
# Por padrão, a função paste() separa as linhas de cada coluna do arquivo com TAB:
# https://www.vivaolinux.com.br/dica/O-comando-paste
## Atenção: uma parte dos genótipos continua sendo uma coluna com as letras
separadas por espaço.
# Então, para substituir os espaços por TAB:
sed -e 's/ /\t/g' arquivo123_esp.ped > arquivo123.ped

```

### 2.3 Formatando arquivo de mapa no software R

Como visto no tópico 1.1, o arquivo de mapa deve ter um formato específico para análises no PLINK.

Supondo que temos um arquivo de mapa prévio com o seguinte formato: três colunas → a primeira tem o nome do SNP, a segunda tem o cromossomo e a terceira tem a posição:



```
BovineHD0100000015 1 36337
BovineHD0100046367 1 89725
BovineHD0100000035 1 120183
BovineHD0100000037 1 127885
BovineHD0100000039 1 146011
BovineHD0100000040 1 147231
BovineHD0100000042 1 149772
BovineHD0100000043 1 151060
```

Este arquivo pode ser transformado em um arquivo de mapa específico para análise no PLINK com o seguinte script para o software R (pode ser no Windows ou no LINUX, porque esse arquivo é menor).

- No software R:

```
rm(list=ls())
options(stringsAsFactors=F) # tem que ter esse comando, se não os nomes dos SNPs
viram números na hora de formar o mapa novo
setwd("/home/pasta_com_arquivos") # Direcionando para a pasta
mapa<-read.table("snmpmap.txt",h=F) # Importando arquivo de mapa
dim(mapa) # Verificando a dimensão do arquivo de mapa - número de linhas (SNPs) e
colunas (três neste caso: nome do SNP, cromossomo e posição)
head(mapa) # Visualizando as seis primeiras linhas do arquivo de mapa
coluna0<-rep(0,nrow(mapa)) # Criando uma coluna só com valores zero, com o mesmo
número de linhas do arquivo de mapa.
length(coluna0) # Comprimento da coluna de zeros
snppnames<-mapa[,1] # Criando um vetor que contém apenas o nome dos SNPs
# Em seguida, criamos um arquivo com as colunas do arquivo de mapa para o PLINK:
1ª col = cromossomo, 2º col = nomes dos SNPs, 3º col = coluna de zeros (esse seria
preenchida pela posição dos marcadores de Morgans, mas essa informação não é
```

necessária nesta análise, portanto zeramos a coluna toda) e 4ª = posição dos marcadores em pares de bases.

```
mapa_novo<-cbind(mapa[,2],snpnames, coluna0,mapa[,3])
```

```
head(mapa_novo) # Visualizando as primeiras linhas do arquivo de mapa criado para o PLINK
```

```
#Em seguida, salvar o arquivo criado:
```

```
write.table(mapa_novo,"arquivo123.map", quote=F, row.names=F, col.names=F, sep="\t") # Salvando o arquivo de mapa para o PLINK
```



### 3. Análise de ROH no PLINK

Para a análise de ROH, usamos o programa PLINK para formar arquivos binários.

Podemos rodar a análise sem transformar para arquivo binário, neste caso, é só tirar o "b" de "bfile" no comando, mas tendo os arquivos binários a análise é mais rápida.

Em uma pasta no servidor, deixar os seguintes arquivos:

- Arquivo de mapa (Ex.: arquivo123.map)
- Arquivo de pedigree (Ex.: arquivo123.ped)
- Programa PLINK

**Os arquivos .map e .ped tem que ter o mesmo nome**, pois na linha de comando da análise só escrevemos “arquivo123” sem identificar o “.map” ou “.ped”. O programa lê os dois de uma vez.

#### 3.1 Rodando a análise

- No servidor, direcionar para a pasta com os arquivos:

```
cd meusdados/pasta_arquivos
```

Dá ENTER

- Usar o comando seguinte para abrir espaço no servidor:

```
ulimit -s unlimited
```

Dá ENTER

- Usar o comando seguinte no servidor para criar o arquivo binário:

```
./plink --noweb --cow --file arquivo123 --make-bed --out plinkbfile_arquivo123
```

Dá ENTER

```
renata.rocha@galloway:~/roh/3_total/Tutorial$ ./plink --noweb --cow --file arquivo123 --make-bed --out plinkbfile_arquivo123
PLINK v1.90b6.24 64-bit (6 Jun 2021)      www.cog-genomics.org/plink/1.9/
(C) 2005-2021 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to plinkbfile_arquivo123.log.
Options in effect:
  --cow
  --file arquivo123
  --make-bed
  --noweb
  --out plinkbfile_arquivo123

Note: --noweb has no effect since no web check is implemented yet.
128830 MB RAM detected; reserving 64415 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (420718 variants, 2093 cattle).
--file: plinkbfile_arquivo123-temporary.bed +
plinkbfile_arquivo123-temporary.bim + plinkbfile_arquivo123-temporary.fam
written.
420718 variants loaded from .bim file.
2093 cattle (170 males, 1923 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 1 founder and 2092 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is exactly 1.
420718 variants and 2093 cattle pass filters and QC.
Note: No phenotypes present.
--make-bed to plinkbfile_arquivo123.bed + plinkbfile_arquivo123.bim +
plinkbfile_arquivo123.fam ... done.
renata.rocha@galloway:~/roh/3_total/Tutorial$
```

- Depois que a análise acima termina, vai soltar as seguintes saídas na pasta:
  - plinkbfile\_arquivo123.beb
  - plinkbfile\_arquivo123.bim
  - plinkbfile\_arquivo123.fam
  - plinkbfile\_arquivo123.log

Nome	Tamanho	Data de modificação	Direitos	Proprie...
..		16/04/2022 17:37:43	rw-rwxr-x	renata...
arquivo123.map	13.357 KB	25/06/2021 19:04:05	rw-rw-r--	renata...
arquivo123.ped	3.439.76...	28/06/2021 19:06:46	rw-rw-r--	renata...
plink	40.680 KB	06/06/2021 17:54:48	rw-rwxr-x	renata...
plinkbfile_arquivo123.bed	215.290 KB	16/04/2022 17:52:11	rw-rw-r--	renata...
plinkbfile_arquivo123.bim	15.000 KB	16/04/2022 17:52:11	rw-rw-r--	renata...
plinkbfile_arquivo123.fam	69 KB	16/04/2022 17:52:11	rw-rw-r--	renata...
plinkbfile_arquivo123.log	2 KB	16/04/2022 17:52:11	rw-rw-r--	renata...

- Em seguida, usar a linha de comando abaixo para rodar a análise de ROH:

```
./plink --bfile plinkbfile_arquivo123 --cow --noweb --homozyg-density 50 --homozyg-gap 1000 --homozyg-kb 1000 --homozyg-snp 50 --homozyg-window-het 1 --homozyg-window-missing 5 --homozyg-window-snp 50 --homozyg-window-threshold 0.05 --nonfounders --geno 0.02 --maf 0.05 --mind 0.1 --out roh_out_arquivo123
```

Dá ENTER

- **Detalhes importantes:**
- **Linha de comando:** aqui neste arquivo a linha de comando fica “quebrada” por causa do limite de espaço, mas a linha de comando não pode estar “quebrada” quando é colada no Linux, pois dá erro e a análise não é finalizada como deveria.
- **Sugestão:** Colar a linha de comando no Notepad++ e daí copiar ela inteira para colar no servidor.

```

380 Obtendo os arquivos.bfile:
381 ./plink --noweb --cow --file gir_total --make-bed --out plinkbfile_gir_total
382
383 Linha de comando:
384 ### Análise 1 ###
385 ./plink --bfile plinkbfile_gir_total --cow --noweb --homozyg-density 50 --homozyg-gap 1000 --homozyg-kb 1000 --homozyg-snp 50 --homozyg-window-het 1 --homozyg-wi
386
387
388

```

Normal text file | length: 86.569 | lines: 1.763 | Ln: 385 | Col: 1 | Sel: 305 | 2 | Windows (CR LF) | UTF-8 | INS

```

PLINK v1.90b6.24 64-bit (6 Jun 2021)          www.cog-genomics.org/plink/1.9/
(C) 2005-2021 Shaun Purcell, Christopher Chang  GNU General Public License v3
Logging to roh_out_arquivol23.log.
Options in effect:
  --bfile plinkbfile_arquivol23
  --cow
  --geno 0.02
  --homozyg-density 50
  --homozyg-gap 1000
  --homozyg-kb 1000
  --homozyg-snp 50
  --homozyg-window-het 1
  --homozyg-window-missing 5
  --homozyg-window-snp 50
  --homozyg-window-threshold 0.05
  --maf 0.05
  --mind 0.1
  --nonfounders
  --noweb
  --out roh_out_arquivol23

Note: --noweb has no effect since no web check is implemented yet.
128830 MB RAM detected; reserving 64415 MB for main workspace.
420718 variants loaded from .bim file.
2093 cattle (170 males, 1923 females) loaded from .fam.
0 cattle removed due to missing genotype data (--mind).
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 1 founder and 2092 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is exactly 1.
0 variants removed due to missing genotype data (--geno).
24850 variants removed due to minor allele threshold(s)
(--maf/--max-maf/--mac/--max-mac).
395868 variants and 2093 cattle pass filters and QC.
Note: No phenotypes present.
--homozyg: Scan complete, found 104609 ROH.
Results saved to roh_out_arquivol23.hom + roh_out_arquivol23.hom.indiv +
roh_out_arquivol23.hom.summary .
renata.rocha@galloway:~/roh/3_total/Tutorial$ █

```

- Depois que a análise termina, solta as seguintes saídas:
- roh\_out\_arquivo123.hom
- roh\_out\_arquivo123.hom.indiv
- roh\_out\_arquivo123.hom.summary
- roh\_out\_arquivo123.log

Nome	Tamanho	Data de modificação	Direitos	Proprie...
..		16/04/2022 17:37:43	rw-rw-r-x	renata...
arquivo123.map	13.357 KB	25/06/2021 19:04:05	rw-rw-r--	renata...
arquivo123.ped	3.439.76...	28/06/2021 19:06:46	rw-rw-r--	renata...
plink	40.680 KB	06/06/2021 17:54:48	rw-rw-r-x	renata...
plinkbfile_arquivo123.bed	215.290 KB	16/04/2022 17:52:11	rw-rw-r--	renata...
plinkbfile_arquivo123.bim	15.000 KB	16/04/2022 17:52:11	rw-rw-r--	renata...
plinkbfile_arquivo123.fam	69 KB	16/04/2022 17:52:11	rw-rw-r--	renata...
plinkbfile_arquivo123.log	2 KB	16/04/2022 17:52:11	rw-rw-r--	renata...
roh_out_arquivo123.hom	19.104 KB	16/04/2022 17:56:43	rw-rw-r--	renata...
roh_out_arquivo123.hom.indiv	109 KB	16/04/2022 17:56:43	rw-rw-r--	renata...
roh_out_arquivo123.hom.summary	29.381 KB	16/04/2022 17:56:43	rw-rw-r--	renata...
roh_out_arquivo123.log	2 KB	16/04/2022 17:56:43	rw-rw-r--	renata...

### **FLAGS relacionadas à ROH:**

- `--homozyg-density` **##** No default: uma ROH deve ter pelo menos um SNP por 50 kb em média.
- `--homozyg-gap` **##** No default: se dois SNPs consecutivos estiverem separados por mais de 1000 kb, eles não podem estar no mesmo ROH.
- `--homozyg-snp` **##** No default: apenas ROH contendo pelo menos 100 SNPs e de comprimento total  $\geq 1000$  kb são anotadas.
- `--homozyg-kb` **##** Podemos alterar esses mínimos com `--homozyg-snp` e `--homozyg-kb`, respectivamente.
- `--homozyg-window-het` **##** No default: uma ocorrência de janela de varredura pode conter no máximo 1 chamada heterozigótica.
- `--homozyg-window-missing` **##** No default: uma ocorrência de janela de varredura pode conter no máximo 5 chamadas perdidas.
- `--homozyg-window-snp` **##** No default: a janela de varredura contém 50 SNPs.
- `--homozyg-window-threshold` **##** No default: para que um SNP seja elegível para inclusão em um ROH, a taxa de acerto de todas as janelas de varredura contendo o SNP deve ser de pelo menos 0,05.
- `--nonfounders` **##** Somente fundadores são normalmente considerados por esses filtros (use a flag para mudar isso).

FLAGS relacionadas ao controle de qualidade (CQ) do arquivo de genótipos.

- `--geno` **##** No default: filtra todos os marcadores com taxas de chamadas ausentes excedendo o valor fornecido (padrão 0,1) para serem removidas.

- --maf ## No default: filtra todas marcadores com frequência de alelo menor abaixo do limite fornecido (padrão 0,01).
- --mind ## Idem --geno só que para amostras.
- --hwe ## Exclui amostras (individuos) e/ou marcadores com base no critério definido para o equilíbrio de Hardy-Weinberg
- --hardy## Faz uma estatística resumida das taxas de HWE

### 3.2 Resultados da análise

Mais detalhes em: <https://www.cog-genomics.org/plink/1.9/formats#hom>

- roh\_out\_arquivo123.log

Este arquivo tem um resumo dos parâmetros usados na análise:

```
PLINK v1.90b6.24 64-bit (6 Jun 2021)
Options in effect:
  --bfile plinkbfile_arquivo123
  --cow
  --geno 0.02
  --homozyg-density 50
  --homozyg-gap 1000
  --homozyg-kb 1000
  --homozyg-snp 50
  --homozyg-window-het 1
  --homozyg-window-missing 5
  --homozyg-window-snp 50
  --homozyg-window-threshold 0.05
  --maf 0.05
  --mind 0.1
  --nonfounders
  --noweb
  --out roh_out_arquivo123

Hostname: galloway
Working directory: /home/renata.rocha/roh/3_total/Tutorial
Start time: Sat Apr 16 20:56:36 2022
```

E também resumo dos resultados da análise:

- ...
- 420718 é o número de marcadores/SNPs no arquivo original
- 2093 é o número de indivíduos (170 machos e 1923 fêmeas)
- 0 indivíduos removidos devido a genótipo ausente.
- ...
- 0 marcadores removidos devido a genótipo ausente. O arquivo imputado original neste exemplo realmente não tinha genótipos imputados.
- 24850 marcadores removidos devido à MAF
- 395868 marcadores e 2093 indivíduos passaram no filtro do controle de qualidade
- --homozyg: varredura completa, encontrou 104609 ROHs

```

Note: --noweb has no effect since no web check is implemented yet.
Random number seed: 1650142596
128830 MB RAM detected; reserving 64415 MB for main workspace.
420718 variants loaded from .bim file.
2093 cattle (170 males, 1923 females) loaded from .fam.
0 cattle removed due to missing genotype data (--mind).
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 1 founder and 2092 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is exactly 1.
0 variants removed due to missing genotype data (--geno).
24850 variants removed due to minor allele threshold(s)
(--maf/--max-maf/--mac/--max-mac).
395868 variants and 2093 cattle pass filters and QC.
Note: No phenotypes present.
--homozyg: Scan complete, found 104609 ROH.
Results saved to roh_out_arquivo123.hom + roh_out_arquivo123.hom.indiv +
roh_out_arquivo123.hom.summary .

End time: Sat Apr 16 20:56:43 2022

```

- roh\_out\_arquivo123.hom
  - Produzido quando uma bandeira da família --homozyg está presente. Acompanhado por pelo menos um arquivo .hom.indiv e um arquivo .hom.summary
  - Cada linha do arquivo .hom representa uma ROH encontrada! Então o total de linhas (excluindo o cabeçalho) indica o total de ROHs encontradas. O número de ROHs encontradas pode alterar dependendo do número de animais e dos parâmetros usados na análise!
  - O arquivo .hom tem as seguintes colunas:
    - FID # Identificação da família -> Estará zerada neste caso, porque não era uma informação necessária.
    - IID # Identificação do indivíduo. Um mesmo indivíduo pode aparecer várias vezes, pois ele tem várias ROHs.
    - PHE # Valor fenotípico -> Neste caso não trabalhamos com o valor fenotípico, por isso está zerada.
    - CHR # Cromossomo em que está aquela ROH
    - SNP1 # Identificação do primeiro SNPs na ROH
    - SNP2 # Identificação do último SNPs na ROH
    - POS1 # Posição do primeiro SNP em pares de bases

- POS2 # Posição do último SNP em pares de bases
- KB # Comprimento da ROH em kilobases
- NSNP # Número de SNPs nessa ROH
- DENSITY # Densidade de SNP inverso em Kb/SNP
- PHOM# Proporção de chamadas (marcadores?) homozigóticas
- PHET # Proporção de chamadas (marcadores?) heterozigóticas

Dica: Se o arquivo de genótipos tiver animais com genótipos imputados, seria interessante comparar o PHET resultante de uma análise com todos os genótipos com o PHET resultante de uma análise que tem apenas animais originalmente genotipados em HD, porque o processo de imputação “adiciona” heterozigotos. Assim, pode-se verificar se existe alguma influência do processo de imputação na análise de ROH.

FID	IID	PHE	CHR	SNP1	SNP2	POS1	POS2	KB	NSNP	DENSITY	PHOM	PHET
0	102916029440	-9.000	1	BovineHD0100006914	BovineHD0100012022	23581349	42095937	18714.589	2949	6.346	1.000	0.000
0	102916029440	-9.000	1	BovineHD0100047134	BovineHD0100029888	102458103	105162097	2783.995	466	5.803	0.994	0.006
0	102916029440	-9.000	1	BovineHD0100041142	BovineHD0100043043	143141556	148784184	5642.629	1056	5.343	0.998	0.002
0	102916029440	-9.000	1	BovineHD0100043081	BovineHD0100044179	148857503	152002893	3145.391	410	7.672	1.000	0.000
0	102916029440	-9.000	1	BovineHD0100044217	BovineHD0100045574	152126033	156040374	3914.342	708	5.529	0.999	0.001
0	102916029440	-9.000	2	BovineHD0200019851	BovineHD0200020565	68732542	71627667	2895.126	199	14.548	1.000	0.000
0	102916029440	-9.000	2	BovineHD0200022016	BovineHD0200025272	76593223	89069133	12475.911	1465	8.516	0.999	0.001
0	102916029440	-9.000	2	BovineHD0200031011	BovineHD0200032436	107756558	112727237	4970.680	600	8.175	0.998	0.002
0	102916029440	-9.000	2	BovineHD0200033469	BovineHD0200033810	116160522	117331836	1171.315	155	7.557	1.000	0.000
0	102916029440	-9.000	3	BovineHD0300008694	BovineHD0300010884	27353121	32151804	4798.684	964	4.978	0.999	0.001
0	102916029440	-9.000	3	BovineHD0300030376	BovineHD0300030777	105981690	107048733	1067.044	181	5.895	0.994	0.006
0	102916029440	-9.000	4	BovineHD0400003894	BovineHD0400003894	13098702	15575489	2476.788	439	5.642	0.998	0.002
0	102916029440	-9.000	4	BovineHD0400005756	BovineHD0400006069	19227607	20249307	1021.701	180	5.676	1.000	0.000
0	102916029440	-9.000	4	BovineHD0400009568	BovineHD0400010275	33749851	36697411	2947.561	459	6.422	1.000	0.000
0	102916029440	-9.000	4	BovineHD0400018018	BovineHD0400018470	65727477	67206589	1479.113	162	9.130	0.994	0.006
0	102916029440	-9.000	4	BovineHD0400034641	BovineHD0400034641	118313905	120555019	2241.115	249	9.000	0.988	0.012
0	102916029440	-9.000	5	BovineHD0500010967	BovineHD0500011316	38339336	39502190	1162.855	147	7.911	0.993	0.007
0	102916029440	-9.000	5	BovineHD0500013540	BovineHD0500014334	47015593	49821847	2806.255	261	10.752	0.992	0.008
0	102916029440	-9.000	5	BovineHD0500017708	BovineHD0500018068	63361467	64604457	1242.991	198	6.278	1.000	0.000
0	102916029440	-9.000	5	BovineHD0500030954	BovineHD0500035330	107529728	120963715	13433.988	1618	8.303	0.999	0.001
0	102916029440	-9.000	6	BovineHD0600010245	BovineHD0600010487	36759019	37885679	1126.661	144	7.824	0.986	0.014
0	102916029440	-9.000	7	BovineHD0700015341	BovineHD0700015761	53361738	54507678	1145.941	173	6.624	1.000	0.000
0	102916029440	-9.000	8	BovineHD0800000328	BovineHD08000003181	975443	10479876	9504.434	1975	4.812	1.000	0.000
0	102916029440	-9.000	9	BovineHD0900004946	BovineHD0900005513	18250661	20272219	2021.559	399	5.067	0.997	0.003
0	102916029440	-9.000	9	BovineHD0900018186	BovineHD0900018563	66064158	67172083	1107.926	216	5.129	1.000	0.000
0	102916029440	-9.000	9	BovineHD0900029583	BovineHD0900030799	101651200	104924541	3273.342	573	5.713	1.000	0.000
0	102916029440	-9.000	12	BovineHD1200007319	BovineHD1200007831	24396971	25955944	1558.974	189	8.249	0.995	0.005
0	102916029440	-9.000	12	BovineHD1200008340	BovineHD1200008948	28035606	29685420	1649.815	74	22.295	0.973	0.027
0	102916029440	-9.000	12	BovineHD1200015688	BovineHD1200015688	56835613	57900635	1065.023	118	9.026	0.983	0.017
0	102916029440	-9.000	13	BovineHD1300023645	BovineHD1300024087	81578587	82896251	1317.665	175	7.530	0.994	0.006
0	102916029440	-9.000	15	BovineHD1500006023	BovineHD1500006528	23467684	24799473	1331.790	296	4.499	1.000	0.000
0	102916029440	-9.000	15	BovineHD1500007739	BovineHD1500008073	28783351	33106808	4323.458	645	6.703	0.997	0.003
0	102916029440	-9.000	15	BovineHD1500022754	BovineHD1500024224	78218321	82853210	4634.890	688	6.737	1.000	0.000

- roh\_out\_arquivo123.hom.indiv
  - Produzido quando uma bandeira da família --homozyg está presente.
  - Cada linha do arquivo .hom.indiv representa um indivíduo. O total de linhas (exceto o cabeçalho) representa o número de indivíduos.
  - O arquivo .hom.indiv tem as seguintes colunas:
    - FID # Identificação da família
    - IID # Identificação do indivíduo
    - PHE # Valor fenotípico
    - NSEG # Número de ROHs encontradas no indivíduo
    - KB # Comprimento total de ROHs (kb) → Soma de todos os comprimentos de ROHs do indivíduo



- **KBAVG** # Comprimento médio de ROHs (kb) → Média de todos os comprimentos de ROHs do indivíduo

FID	IID	PHE	NSEG	KB	KBAVG
0	102916029440	-9	51	198256	3887.37
0	AB52502	-9	52	213103	4098.13
0	AB52503	-9	42	102783	2447.22
0	AB52505	-9	57	202055	3544.82
0	AB52508	-9	57	262986	4613.79
0	AB52511	-9	64	285553	4461.76
0	AB52512	-9	61	346225	5675.82
0	AB52514	-9	64	375508	5867.32
0	AB52516	-9	47	108046	2298.85
0	AB52517	-9	47	975503	2075.54

- roh\_out\_arquivo123.hom.summary

- Produzido quando uma bandeira da família --homozyg está presente.

- Cada linha do arquivo .hom.summary representa um marcador. O total de linhas (exceto o cabeçalho) representa o número de marcadores.

- O arquivo .hom.summary tem as seguintes colunas:

- **CHR** # Cromossomo
- **SNP** # Identificação do SNP
- **BP** # Posição do SNP/marcador em pares de bases
- **AFF** # Número de casos com ROHs incluindo esse SNP/marcador
- **UNAFF** # Número de não-casos com ROHs incluindo esse SNP/marcador

Observe que as amostras com fenótipos ausentes são contadas na coluna 'UNAFF'. Se o fenótipo for quantitativo, todos serão contados em 'UNAFF'.

CHR	SNP	BP	AFF	UNAFF
1	BovineHD0100046367	89725	0	46
1	BovineHD0100000035	120183	0	46
1	BovineHD0100000039	146011	0	46
1	BovineHD0100000040	147231	0	46
1	BovineHD0100000042	149772	0	46
1	BovineHD0100000043	151060	0	46
1	BovineHD0100000044	152374	0	46
1	BovineHD0100000048	158820	0	46
1	BovineHD0100000049	160007	0	46
1	BovineHD0100000052	164683	0	46
1	BovineHD0100000054	168997	0	46
1	BovineHD0100000057	183000	0	46

### 3.3 Como cada análise é descrita no artigo

Tomando a seguinte análise como exemplo:

```
--homozyg-density 100 --homozyg-gap 1000 --homozyg-kb 1000 --homozyg-snp 50 --
homozyg-window-het 1 --homozyg-window-missing 5 --homozyg-window-snp 50 --
homozyg-window-threshold 0.05 --geno 0.02 --mind 0.1 --maf 0.05 --hwe 0.15 --
nonfounders --out roh_out_arquivo123
```

A descrição abaixo segue **respectivamente** cada uma das FLAGS do exemplo acima.

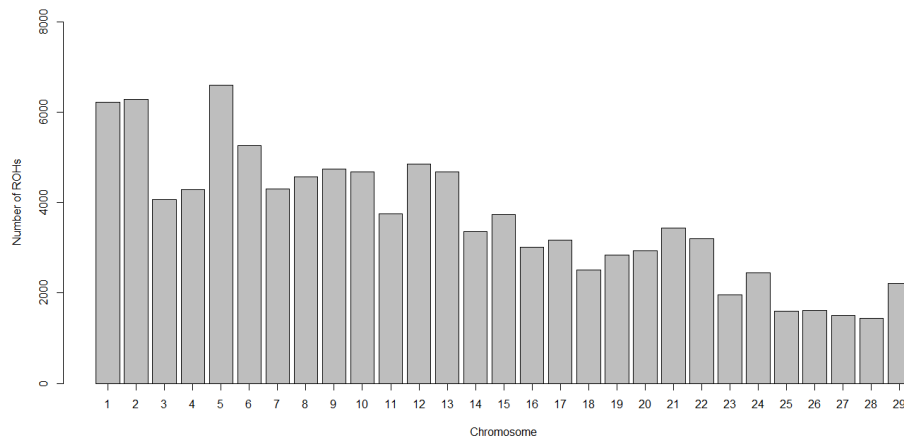
(Esta é apenas uma sugestão!):

*“A density of one SNP per 100 kb was used. The maximum gap between consecutive homozygous SNPs was 1000 kb. The minimum length of a ROH was set to 1 Mb. The minimum number of consecutive SNPs included in a ROH was 50. Up to one heterozygous genotype were allowed in a ROH. A maximum of 5 SNPs with missing genotypes and a sliding window of 50 SNPs across the genome were used. The proportion of homozygous overlapping windows was 0.05. For the quality control, the analysis considered call rate of 0.98 for genotype and 0.90 for samples, minor allele frequency (MAF) of 0,05 and Hardy Weinberg equilibrium of 0,15.”*

### 3.4 Número e Tamanho de ROHs por cromossomo e classe

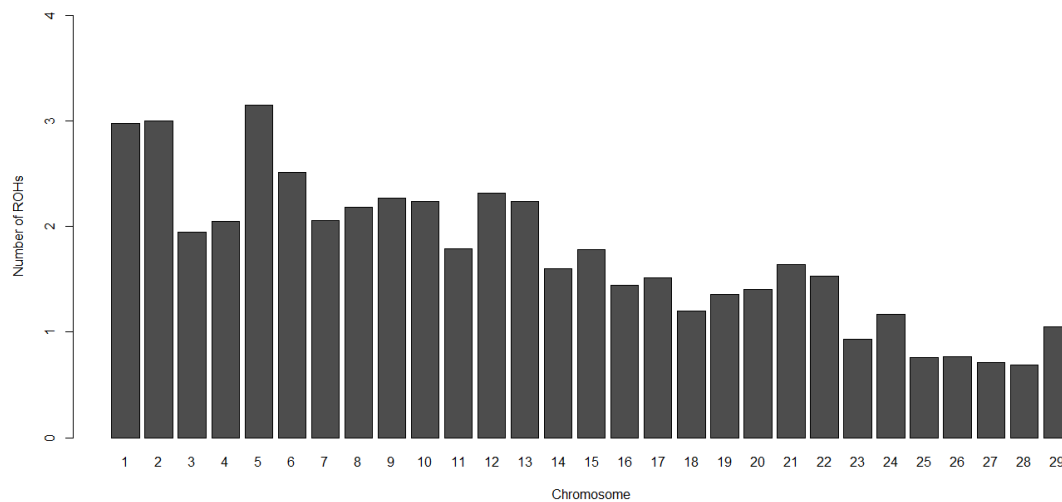
Com o arquivo .hom e o script “1\_ROH\_graficos\_classes.R” (anexa no final) podem ser feitos alguns exemplos práticos das análises, como a montagem de gráficos com o número e tamanho de ROHs por animal para incluir na publicação do artigo ou montar uma tabela com as informações de ROH por classe de tamanho. Os resultados gráficos podem ser apresentados como o da Figura 1:

**Figura 1. Exemplo do número total de ROHs por cromossomo**



Alguns artigos publicam o número total de ROHs por cromossomo (Figura 1). Outros artigos publicam o gráfico com as médias calculadas por animal e cromossomo (Figura 2). Para montar esses gráficos, verificar o script do R (1\_ROH\_graficos\_classes.R).

**Figura 2. Exemplo do número médio de ROHs por indivíduo e cromossomo**

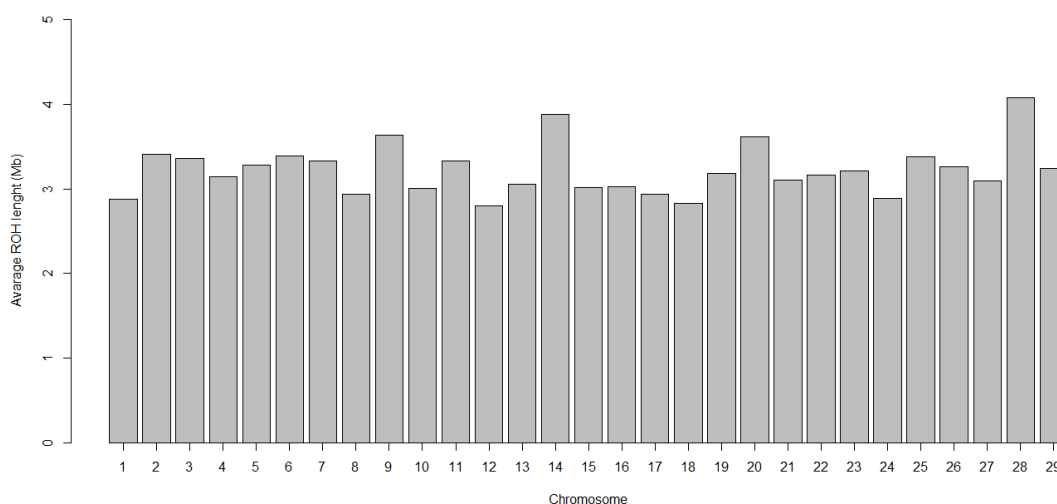


Os gráficos das Figuras 1 e 2 não mudam “em estrutura”, só mudam em escala (eixo Y).

O autor pode escolher qual acha ser mais adequado.

Com o script do R (1\_ROH\_graficos\_classes.R) também podemos montar o gráfico do tamanho médio de ROHs por cromossomo.

**Figura 3. Exemplo do tamanho médio de ROHs por cromossomo.**



Script do R (1\_ROH\_graficos\_classes.R) → **Seção:** Proporção de ROHs por tamanho  
 Muitos artigos exploram as ROHs com relação ao seu tamanho, pois as ROHs maiores (mais longas) indicam endogamia mais recente. O que alguns pesquisadores fazem é dividir as ROHs em cinco classes de tamanho: de 0 ou 1 a 2 Mb, 2 a 4 Mb, 4 a 8 Mb, 8 a 16 Mb e acima de 16 Mb.

Assim, o arquivo .hom e o script do R (1\_ROH\_graficos\_classes.R) também pode ser usado para montar resultados para essas classes de ROH.

**Tabela 1. Exemplo de parâmetros avaliados por classe de ROH.**

Class	NROH	Percent (%)	LROH (Mb)	Number of animals	SROH	Genome coverage
ROH <sub>1-2 Mb</sub>	61,269	58.17	1.35	2,093	29.27	0.05%
ROH <sub>2-4 Mb</sub>	23,949	22.74	2.78	2,093	11.44	0.11%
ROH <sub>4-8 Mb</sub>	11,942	11.34	5.53	2,081	5.74	0.22%
ROH <sub>8-16 Mb</sub>	5,805	5.51	11.05	1,847	3.14	0.44%
ROH <sub>&gt;16 Mb</sub>	2,362	2.24	24.66	1,132	2.09	0.99%

<sup>1</sup>NROH = número de ROHs; LROH = comprimento médio de ROHs; SROH = número médio de ROHs por animal.

A coluna Genome Coverage mostra a proporção do genoma que é coberta por essa classe do ROH. Considerando que o tamanho total do genoma bovino (espécie deste exemplo) é de 2489,37 Mb, a forma de se calcular a cobertura do genoma é dividir o

LROH pelo comprimento total do genoma, assim,  $(1,35 / 2489,35) * 100 = 0,05\%$  na classe  $ROH_{1-2\text{ Mb}}$  e assim por diante.

### 3.5 Genome Coverage por cromossomo

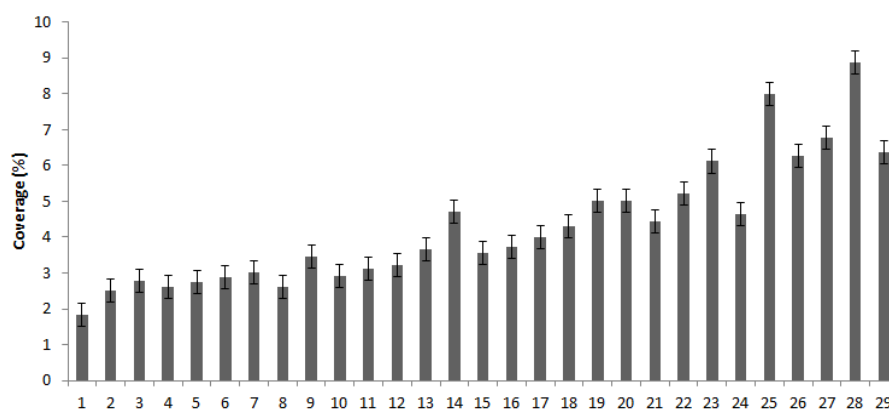
A informação de tamanho médio das ROHs por cromossomo, usada para construir a Figura 3, também pode ser usada juntamente com a informação do tamanho total em Megabases (Mb) de cada cromossomo para montar um gráfico com a proporção do genoma que é coberto por ROHs.

#### *Como saber o tamanho de cada cromossomo?*

Caso não tenha a informação do tamanho de cada cromossomo da espécie com a qual está trabalhando, isso pode ser pesquisado no site NCBI. O processo é relativamente rápido. Verificar tutorial “Como saber o tamanho de cada cromossomo.PDF”.

Para construir um gráfico de Genome Coverage por cromossomo como na Figura 4 foi utilizado uma planilha do Excel. Ao final deste tutorial podem ser encontrados os arquivos em Rmarkdown para os scripts do R e também o esquema para montar a figura 4 no Excel (2\_Genome\_Coverage\_xlsx).

**Figura 4. Porcentagem média de cobertura dos cromossomos que é coberta por corridas de homozigose. As barras de erro indicam erro padrão**



#### 4. Calculando a endogamia baseada em ROH (FROH)

Para calcular a endogamia com base em ROH ( $F_{ROH}$ ), neste caso, foi usado o pacote detectRUNS (Biscarini et al., 2019) no software R (R version 4.0.2; R Foundation for Statistical Computing, Vienna, Austria). O pacote detectRUNS é aplicado para genomas diploides.

A forma de se calcular a  $F_{ROH}$  é pela seguinte fórmula (McQuillan et al., 2008):

$$F_{ROH} = \frac{\sum_{j=1}^n L_{ROH}}{L_{aut}}$$

onde  $L_{ROH}$  é a soma de ROH por animal acima de um certo critério de comprimento e  $L_{aut}$  é o tamanho total dos cromossomos autossômicos, coberto por marcadores.  $L_{aut}$  para o genoma amplo aqui neste exemplo foi tomado como 2489.37Mb em comprimento, baseado nas posições do mapa ARS-UCD1.2 da montagem do genoma bovino (Rosen et al., 2020).

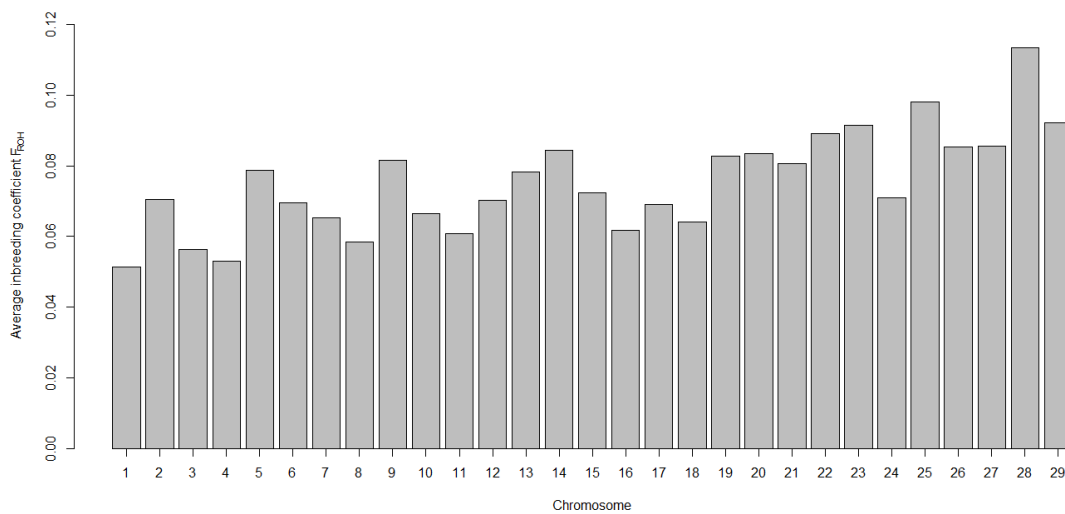
A endogamia pode ser obtida para cada cromossomo, para cada animal e, também, por animal e classe de ROH (de 1 a 2 Mb, 2 a 4 Mb, 4 a 8 Mb, 8 a 16 Mb e acima de 16 Mb).

O script do R (3\_Calculo\_FROH.R) tem a forma de se calcular a endogamia baseada em ROH com o pacote detectRUNS. Para calcular a endogamia será necessário o arquivo de mapa (arquivo123.map) e o arquivo de saída com os resultados (.hom) usado na análise de ROH do PLINK.

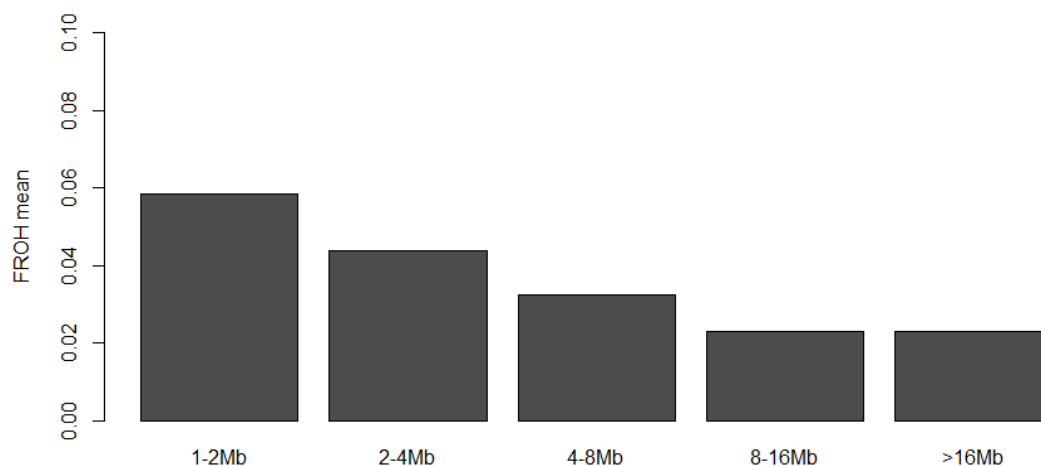
##### 4.1 FROH por cromossomo e por classe

Para montar os gráficos do coeficiente de endogamia (FROH) por cromossomo (Figura 5) e por classe de ROH (Figura 6), usar o script do R (4\_FROH\_graficos.R).

**Figura 5. Coeficiente de endogamia ( $F_{ROH}$ ) por cromossomo**



**Figura 6. Coeficiente de endogamia ( $F_{ROH}$ ) por classe de ROH**



#### 4.2 Variação de uma característica com a $F_{ROH}$

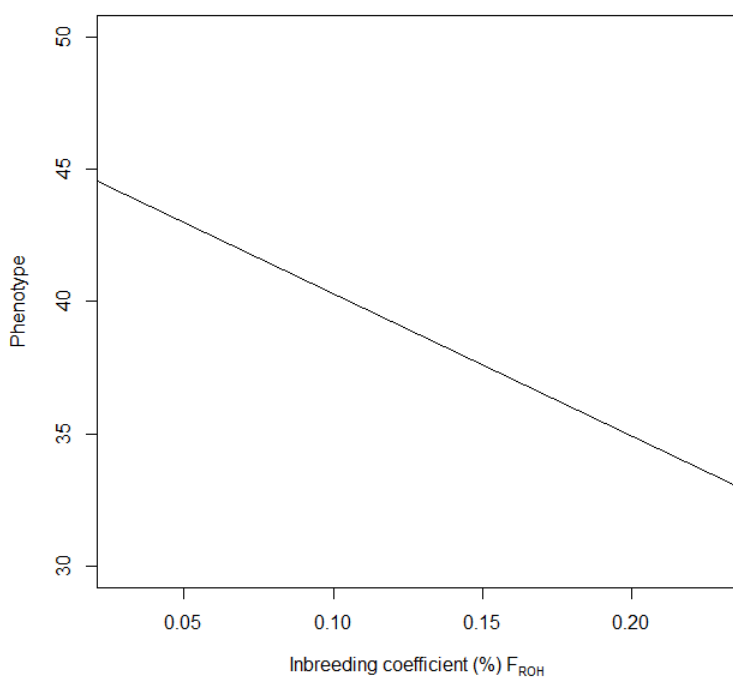
Caso tenha um fenótipo e queira verificar como a endogamia está interferindo nesta característica, verificar o script do R: 5\_Trait\_FROH.R

Para montar um gráfico com a relação entre o fenótipo e a endogamia, será preciso o arquivo de fenótipos com pelo menos uma coluna com a identificação do animal e uma coluna com os valores da característica. Também precisa do arquivo Froh\_GW.txt que

contém uma coluna de identificação do animal e uma coluna com o valor da  $F_{ROH}$  por animal, gerado anteriormente com o script do R (3\_Calculo\_FROH.R).

Montando o gráfico da variação da característica em relação ao coeficiente de endogamia ( $F_{ROH}$ ), podemos observar se a  $F_{ROH}$  afeta a característica de forma positiva ou negativa. Com o exemplo hipotético (dados2.txt) podemos observar na Figura 7 que com o aumento da endogamia temos uma queda no fenótipo (depressão endogâmica).

**Figura 7. Exemplo hipotético de variação de uma característica com o coeficiente de endogamia ( $F_{ROH}$ )**



Para afirmar que a queda ou aumento das características é significativo, precisamos verificar o coeficiente linear e nível de significância. O script do R (5\_Trait\_FROH.R) também pode ser usado para salvar vários resultados da regressão da característica em função do coeficiente de endogamia ( $F_{ROH}$ ).

**Tabela 2. Exemplo de resultados da regressão da característica em função da FROH**

Item	Resultado
Intercepto	45.6974910287186
Coefficiente_linear	-53.9241104581322
Desvio_padrao_residual	17.4429714950595
Graus_de_liberdade	48
$R^2$	0.0122640135781325
$R^2_{ajustado}$	-0.00831381947232313



---

Estadística\_F  
p-value

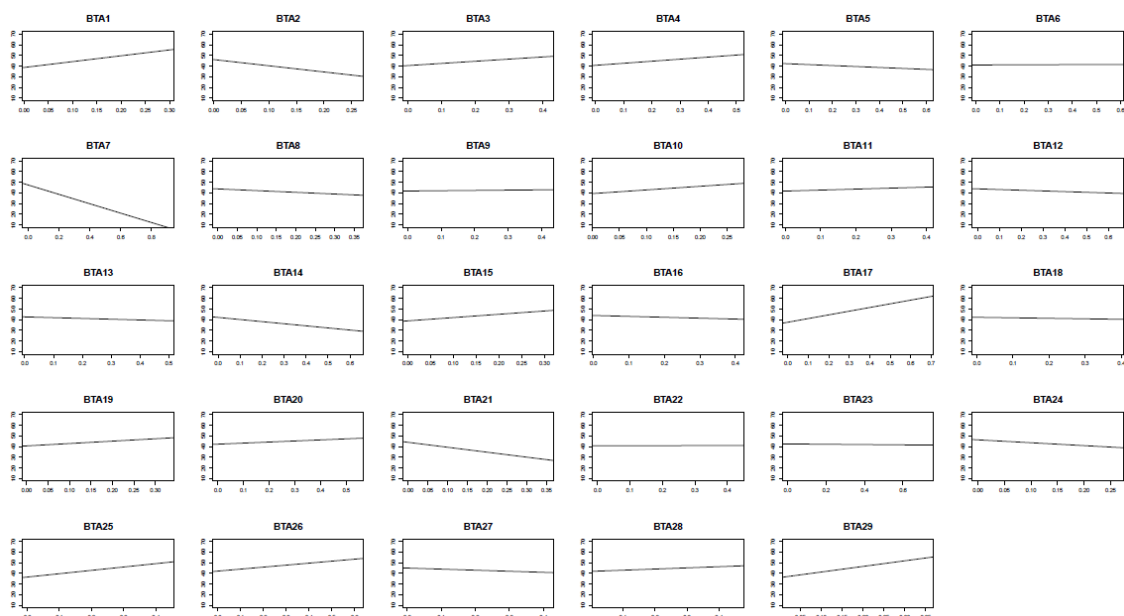
---

0.595981780397474  
0.443899685018646

---

Também é possível verificar como as características variam com a  $F_{ROH}$  em cada cromossomo (5\_Trait\_FROH.R → Seção ‘Variação da característica com a  $F_{ROH}$  e cada cromossomo’). Com a endogamia por cromossomo podemos observar que a característica varia de forma positiva, negativa ou não sofre variação dependendo do cromossomo (Figura 8).

**Figura 8. Exemplo hipotético de variação de uma característica com o coeficiente de endogamia ( $F_{ROH}$ ) por cromossomo**



Os cromossomos autossômicos da espécie *Bos taurus* são identificados pela sigla BTA: *Bos taurus autosome*.

O script do R (5\_Trait\_FROH.R) também pode ser usado para salvar os resultados da regressão da característica em função do coeficiente de endogamia ( $F_{ROH}$ ) por cromossomo.

### Referências

Biscarini, F., P. Cozzi, G. Gaspa, and G. Marras. 2019. detectRUNS: an R package to detect runs of homozygosity and heterozygosity in diploid genomes. Univ. Guelph.

<https://CRAN.R-project.org/package=detectRUNS>.

McQuillan, R., A.L. Leutenegger, R. Abdel-Rahman, C.S. Franklin, M. Pericic, L.

Barac-Lauc, N. Smolej-Narancic, B. Janicijevic, O. Polasek, A. Tenesa, A.K. MacLeod, S.M. Farrington, P. Rudan, C. Hayward, V. Vitart, I. Rudan, S.H. Wild, M.G. Dunlop, A.F. Wright, H. Campbell, and J.F. Wilson. 2008. Runs of Homozygosity in European Populations. *Am. J. Hum. Genet.* 83.

doi:10.1016/j.ajhg.2008.08.007.

R Core Team. 2022. 'R: A language and environment for statistical computing' R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>.

Rosen, B.D., D.M. Bickhart, R.D. Schnabel, S. Koren, C.G. Elsik, E. Tseng, T.N. Rowan, W.Y. Low, A. Zimin, C. Couldrey, R. Hall, W. Li, A. Rhie, J. Ghurye, S.D. McKay, F. Thibaud-Nissen, J. Hoffman, B.M. Murdoch, W.M. Snelling, T.G. McDanel, J.A. Hammond, J.C. Schwartz, W. Nandolo, D.E. Hagen, C. Dreischer, S.J. Schultheiss, S.G. Schroeder, A.M. Phillippy, J.B. Cole, C.P. Van Tassell, G. Liu, T.P.L. Smith, and J.F. Medrano. 2020. De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience* 9.  
doi:10.1093/gigascience/giaa021.

## 5. Análise de FST – assinatura de seleção

Para encontrar assinaturas de seleção por meio do índice de fixação de Wright (FST), a análise pode ser feita na mesma linha de comando quando fazemos a busca por corridas de homozigose no PLINK.

- Para a análise de assinatura de seleção por FST usando o PLINK, precisamos de:
  - Arquivo adicional com as colunas: FID (família), ID (identificação do animal), cluster (grupo). Exemplo:

FID	ID	cluster
0	1	1
0	2	1
0	3	1
0	4	2
0	5	2
0	6	2
0	7	2
0	8	3
0	9	3
0	10	3

A coluna de família está zerada, pois não será usada nesta análise. A coluna de Id identifica o indivíduo e a coluna cluster indica a qual grupo esse animal pertence. O grupo pode ser referente a raças (1, 2 e 3), a indivíduos com alto e baixo valor genético para uma característica (1: alto valor genético e 2: baixo valor genético) ou outro tipo de classificação – isso vai depender da metodologia da pesquisa.

- Parâmetros adicionados na linha de comando: `--within`, `--fst`

```
./plink --bfile plinkbfile_arquivo123 --cow --noweb --homozyg-density 50 --homozyg-gap 1000 --homozyg-kb 1000 --homozyg-snp 50 --homozyg-window-het 1 --homozyg-window-missing 5 --homozyg-window-snp 50 --homozyg-window-threshold 0.05 --nonfounders --geno 0.02 --maf 0.05 --mind 0.1 --within arquivo_com_grupos.ped --fst -out roh_out_arquivo123
```

## Resultados

Além dos resultados da análise de ROH, um arquivo `.fst` será resultante dessa análise.

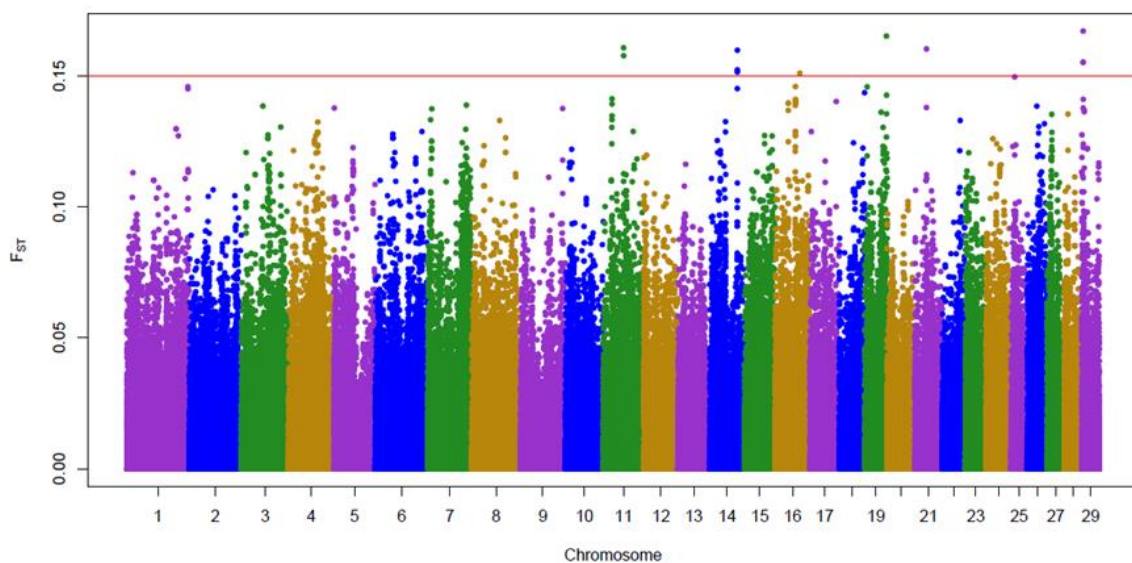
Este arquivo contém as colunas:

- CHR # Cromossomo
- SNP # Identificação do marcador

- POS # Posição do marcador em pares de bases
- NMISS # Número total de indivíduos usados na análise
- FST # Valor de FST para cada marcador

O script do R (6\_ManhattanPlot\_FST.R) pode ser usado para montar um Manhattan Plot com os resultados do arquivo .fst, como no exemplo abaixo:

**Figura 9. Exemplo de Manhattan Plot que pode ser construído para identificar assinaturas de seleção pela metodologia FST**



### Como o threshold (nível de significância) foi definido?

- Fst é uma medida de diferenciação entre duas populações.
- Valores variando de 0 (nenhuma diferença entre as populações) a 1 (diferenças fixas entre as populações).
- Valores de Fst entre 0 e 0,05 indicam pouca diferenciação genética
- Valores de Fst entre 0,05 e 0,15 indicam diferenciação genética moderada
- Valores de Fst entre 0,15 e 0,25 indicam grande diferenciação genética
- Valores de Fst acima de 0,25 indicam um grau muito grande de diferenciação genética

Fonte: Wright, S. 1978. Evolution and the Genetics of Populations. Vol. 4. Variability Within and Among Natural Populations. University Chicago Press, Chicago, USA.

- A forma de definir o threshold pode depender do pesquisador.

- O threshold estabelecido neste caso foi de 0,15, porque não havia valores acima de 0,25.

Com o script **6\_ManhattanPlot\_FST.R** também podemos verificar e salvar as regiões (posições dos marcadores e cromossomos) acima do threshold estabelecido, no caso,  $FST > 0,15$ .

O próximo passo seria uma análise de Ontologia Gênica:

- Busca de genes nestas regiões genômicas
- Pesquisa por processos biológicos relacionados a estes genes

# 1\_ROH\_graficos\_classes.R

Particular

2023-10-04

```
#####  
#### TUTORIAL - GRAFICOS DE ROHs ####  
#####
```

```
rm(list=ls())  
options(stringsAsFactors=F)
```

```
# direcionar para o diretorio onde estao os arquivos  
setwd("D:\\PessoaID\\Doutorado\\Cap 2 ROH\\Tutorial\\1_ROH_FROH")
```

```
saida_hom<-read.table("roh_out_gir_total.hom",h=T) # Lendo o arquivo .hom  
dim(saida_hom) # N?mero de Linhas e colunas (dimens?o) do arquivo
```

```
## [1] 105327 13
```

```
head(saida_hom) # Visualizando as primeiras 6 linhas do arquivo
```

```
## FID IID PHE CHR SNP1 SNP2 POS1 POS2 KB NSNP  
DENSITY PHOM PHET  
## 1 0 102916029440 -9 1 BovineHD0100006914 BovineHD0100012022 23381349 42095937 18714.589 2816  
6.646 1.000 0.000  
## 2 0 102916029440 -9 1 BovineHD0100047134 BovineHD0100029888 102458103 105162097 2703.995 377  
7.172 0.995 0.005  
## 3 0 102916029440 -9 1 BovineHD0100041142 BovineHD0100043043 143141556 148784184 5642.629 975  
5.787 0.998 0.002  
## 4 0 102916029440 -9 1 BovineHD0100043081 BovineHD0100044197 148857503 152070453 3212.951 397  
8.093 0.997 0.003  
## 5 0 102916029440 -9 1 BovineHD0100044217 BovineHD0100045574 152126033 156040374 3914.342 685  
5.714 0.999 0.001  
## 6 0 102916029440 -9 2 BovineHD0200019851 BovineHD0200020565 68732542 71627667 2895.126 181  
15.995 1.000 0.000
```

```
saida_hom$MB<-saida_hom$KB/1000 # Criando uma coluna com o tamanho das ROHs em Mbases
```

```
#### Numero de ROHs por individuo ####
```

```
roh_animais<-as.data.frame(table(saida_hom$IID)) # Numero de ROHs por individuo  
# Esse numero de ROHs por individuo é o mesmo que a coluna NSEG do arquivo .hom.indiv  
head(roh_animais)
```

```
## Var1 Freq  
## 1 102916029440 51  
## 2 AB52502 52  
## 3 AB52503 41  
## 4 AB52505 55  
## 5 AB52508 58  
## 6 AB52511 64
```

```
mean(roh_animais$Freq) # Numero medio de ROHs por individuo
```

```
## [1] 50.32346
```

```
sd(roh_animais$Freq) # Desvio padrao do numero medio de ROHs por individuo
```

```
## [1] 8.591984
```

```
min(roh_animais$Freq) # Numero minimo de ROHs por individuo
```

```
## [1] 26
```

```
max(roh_animais$Freq) # Numero maximo de ROHs por individuo
```

```
## [1] 95
```

```
#### Numero de ROHs por cromossomo ####
```

```
N_ROH_CHR<-as.data.frame(table(saida_hom$CHR))
```

```
head(N_ROH_CHR) # Numero total de ROHs por cromossomo
```

```
##   Var1 Freq  
## 1    1 6227  
## 2    2 6289  
## 3    3 4071  
## 4    4 4284  
## 5    5 6600  
## 6    6 5260
```

```
# Se quiser, pode fazer o grafico com esse arquivo
```

```
# Se for usar e quiser salvar esse arquivo, tirar o # da linha de comando abaixo.
```

```
#write.table(N_ROH_CHR, "Resultado_numero_ROH_por_CHR.txt",quote=F,row.names=F,colnames=T)
```

```
#### Grafico do numero total de ROHs por cromossomo ####
```

```
# Exemplo do numero total de ROHs por cromossomo
```

```
max(N_ROH_CHR$Freq) # Verificando qual o valor maximo para estabelecer o ylim no comando abaixo
```

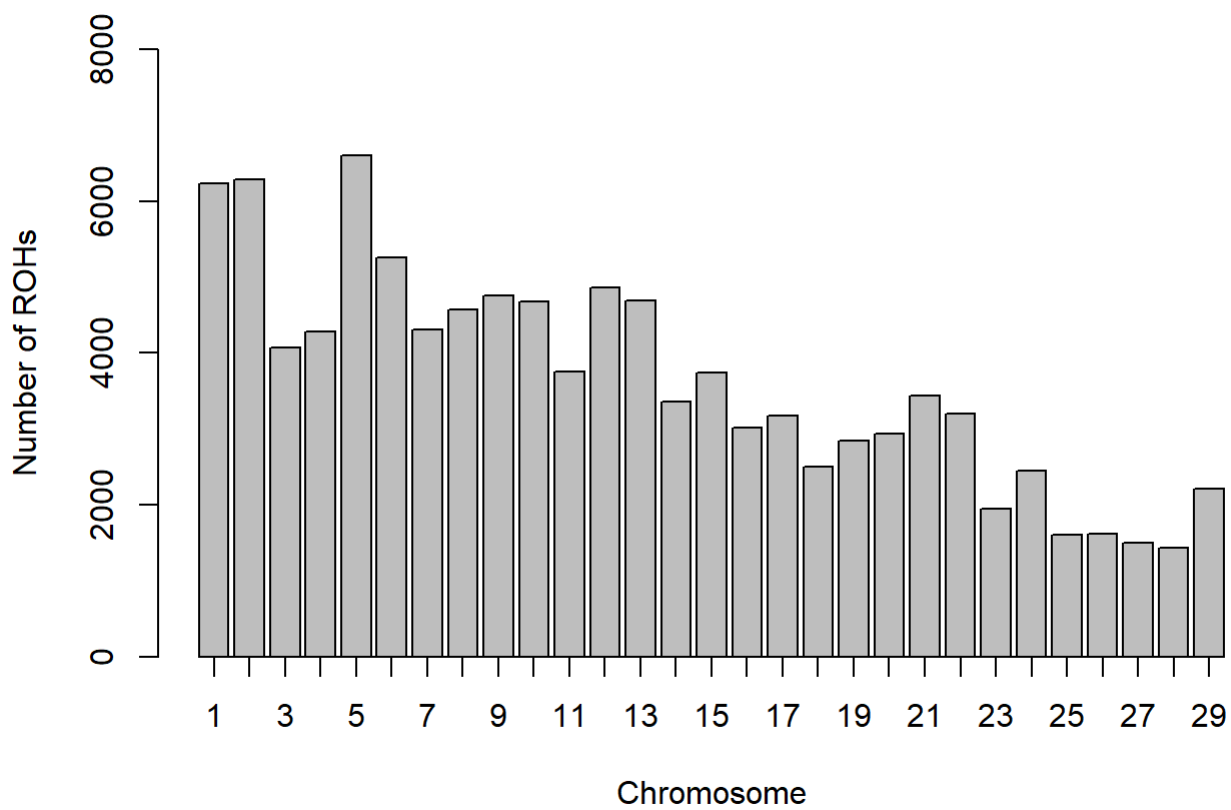
```
## [1] 6600
```

```
w<-barplot(N_ROH_CHR$Freq,ylim=c(0,8000),ylab="Number of ROHs",xlab="Chromosome")
```

```
w
```

```
##      [,1]
## [1,] 0.7
## [2,] 1.9
## [3,] 3.1
## [4,] 4.3
## [5,] 5.5
## [6,] 6.7
## [7,] 7.9
## [8,] 9.1
## [9,] 10.3
## [10,] 11.5
## [11,] 12.7
## [12,] 13.9
## [13,] 15.1
## [14,] 16.3
## [15,] 17.5
## [16,] 18.7
## [17,] 19.9
## [18,] 21.1
## [19,] 22.3
## [20,] 23.5
## [21,] 24.7
## [22,] 25.9
## [23,] 27.1
## [24,] 28.3
## [25,] 29.5
## [26,] 30.7
## [27,] 31.9
## [28,] 33.1
## [29,] 34.3
```

```
axis(1, at=w, labels=1:29)
```





# Se for usar esse grafico, SALVAR! Export -> Save as Image ... ou ... Arquivo -> salvar como...

```
#### Numero de ROHs por individuo e cromossomo ####
```

```
N_ROH_IID_CHR<-table(saida_hom$IID, saida_hom$CHR)
N_ROH_IID_CHR1<-as.data.frame(colMeans(N_ROH_IID_CHR))
N_ROH_IID_CHR1
```

```
##      colMeans(N_ROH_IID_CHR)
## 1          2.9751553
## 2          3.0047778
## 3          1.9450549
## 4          2.0468227
## 5          3.1533684
## 6          2.5131390
## 7          2.0592451
## 8          2.1839465
## 9          2.2704252
## 10         2.2365026
## 11         1.7940755
## 12         2.3210702
## 13         2.2398471
## 14         1.6053512
## 15         1.7849976
## 16         1.4409938
## 17         1.5150502
## 18         1.1978022
## 19         1.3602484
## 20         1.4056378
## 21         1.6445294
## 22         1.5284281
## 23         0.9331104
## 24         1.1720019
## 25         0.7634974
## 26         0.7716197
## 27         0.7152413
## 28         0.6865743
## 29         1.0549451
```

# Se for usar e quiser salvar esse arquivo, tirar o # da linha de comando abaixo.

```
#write.table(N_ROH_IID_CHR1, "Resultado_numero_ROH_por_indiv_CHR.txt",quote=F,row.names=F,colnames=T)
```

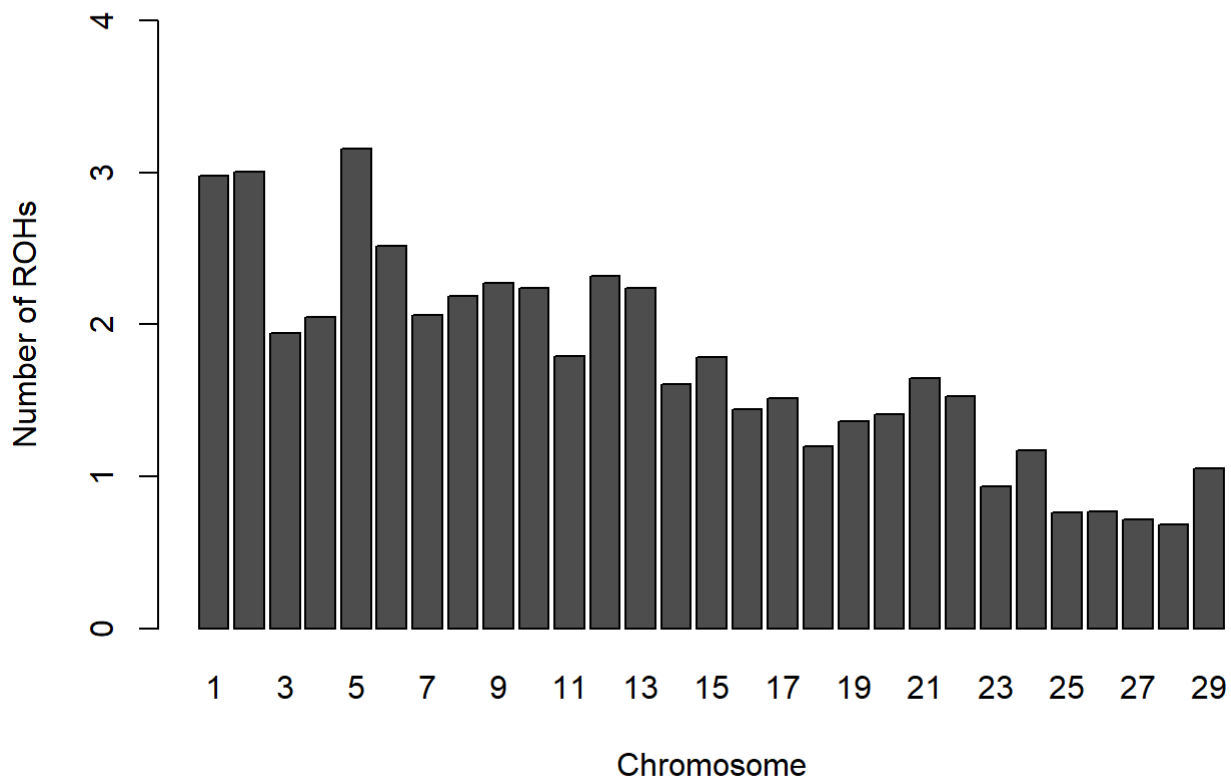
```
#### Grafico do numero de ROHs por individuo e cromossomo ####
```

```
# Exemplo do numero medio de ROHs por individuo e cromossomo
```

```
max(N_ROH_IID_CHR1)
```

```
## [1] 3.153368
```

```
barplot(t(N_ROH_IID_CHR1),ylim=c(0,4),ylab="Number of ROHs",xlab="Chromosome")
```



*# Se for usar esse grafico, SALVAR!*

*#### Comprimento medio de ROHs por cromossomo (Mb) ####*

```
size_CHR<-aggregate(MB ~ CHR, saida_hom, mean)
```

```
size_CHR
```