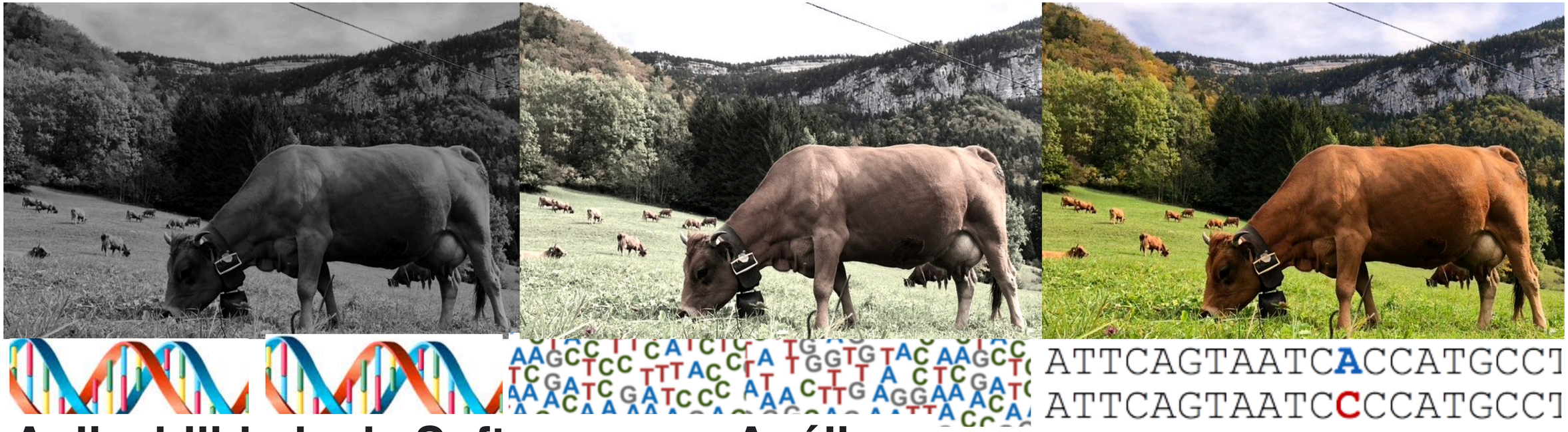

Aplicabilidade de Softwares em Análises Genômicas - EXERCÍCIOS

Limpeza de Dados Utilizando Sequência de Genoma Completo

03.02.2021

Isis Hermisdorff
Agropartners Consulting
https://agropartners.com.br
isis.dorff@agropartners.com.br

Mãos a obra!

Tamanho aproximado do genoma bovino?

- File: chr_size.txt
  - read.table()
  - head=T
  - sep=“”
  - colClasses=c(“character”,”numeric”)

Comandos úteis:

head()

tail()

max()

min()

mean()

plot()

Mãos a obra!

PERGUNTAS:

- Qual é o maior cromossomo?

- Qual é o menor cromossomo?

- Média do tamanho dos autossomos?

Mãos a obra!

```
file <- read.table("Genotipos_Pratica/Passar_alunos/CHR_size.txt",
                   header=TRUE,
                   sep='',
                   colClasses=c("character", "numeric"))

str(file)


##Três primeiras linhas do arquivo
head(file, n=3)

#Três ultimas linhas do arquivo
tail(file, n=3)

#Maior cromossomo:
head(file[order(file[ ,2], decreasing=TRUE), ], n=1)
```

Localização do Arquivo

Mãos a obra!

```r
##Tamanho medio dos autossomos:
mean(file$length)

## Melhorando a apresentacao do tamanho medio dos autossomos:
round(mean(file$length) / 1000000, 2)

## Tamanho do genoma - sem contar os cromossomos sexuais em milhos de bp
round(sum(file[ ,2]) / 1000000, 2)


### Outra forma para visualizar e conhecer melhors os seus dados
##Plotando
par(mar=c(5,5,4,2))
colour <- rep("seagreen", nrow(file))
colour[file[,1]=="1"] <- "orange"
plot(file[,2]/1000000, type="h", lwd=12, lend=1, col=colour, ylim=c(0,160), ylab="Length (Mb)",
     xlab="Chromosome", cex.lab=1.05, font.lab=2, axes=FALSE)
axis(2, las=1, cex.axis=1.2)
axis(1, at=c(1, 5, 10, 15, 20, 25,29), labels=c(1, 5, 10, 15, 20, 25,29), cex.axis=1.05)


legend("top", pch=c(15,15), col=c("seagreen", "orange"), c("Autosomes", "Maior BTA"), cex=1.1,
       pt.cex=1.2, ncol=2)
```

Mãos a obra!

Frequência Alélica

- Tarefa:
  - Calcule p e q:

Genótipos	Número de Animais	Número de Alelos A	Número de Alelos C	Total
AA	453	453 X 2	0	906
AC	1129	1129	1129	2258
CC	776	0	776 X 2	1552
		A= 906 + 1129	C= 1129 + 1552	2N=4716

Freqüência do alelo A
f(A) = nA/2N

Freqüência do alelo A2
f(C) = nC/2N

Freqüência do alelo A f(A) = (906 + 1129)/4716 = 0,43 = 43%
Freqüência do alelo C f(C) = (1129 + 1552)/4716 = 0,57 = 57%

Calculo das frequências Alélica e Genotipica

Frequência Genotípica - Observada (Gobs) e Esperada (Gexp)

Função no R - Método Simples

Função no R:

```
AA <- 453
AB <- 1129
BB <- 776

Gobs <- c(AA, AB, BB)
n <- sum(Gobs)          # Numero de animais

cat(paste("Distribuicao observada dos genotipos: "), Gobs)

cat(paste("\nNumero de amostras genotipadas (n): "), n)
```

Saída/ Resultado:

```
Console   Terminal   Jobs
~/
> AA <- 453
> AB <- 1129
> BB <- 776
> Gobs <- c(AA, AB, BB)
> n <- sum(Gobs)
> cat(paste("Distribuicao observada dos genotipos: "), Gobs)
Distribuicao observada dos genotipos:  453 1129 776
> cat(paste("\nNumero de amostras genotipadas (n): "), n)

Numero de amostras genotipadas (n):  2358
>
```

Frequências genotípicas esperada e observada

```
p <- (AA*2 + AB) / (n*2)
q <- (BB*2 + AB) / (n*2)
cat(paste("Frequency of p: "), p)
cat(paste("\nFrequency of q: "), q)

# Vamos calcular a frequencia genotipica esperada
AA_exp <- (p^2) * n
AB_exp <- (2*p*q) * n
BB_exp <- (q^2) * n

Gexp <- c(AA_exp, AB_exp, BB_exp)
cat(paste("Expected genotype distribution: "), Gexp)
```

Saída/ Resultado:

Console | Terminal | Jobs

~/

```
> p <- (AA*2 + AB) / (n*2)
> q <- (BB*2 + AB) / (n*2)
> cat(paste("Frequency of p: "), p)
Frequency of p:  0.4315098
> cat(paste("\nFrequency of q: "), q)

Frequency of q:  0.5684902
>
> # Vamos calcular a frequencia genotipica esperada
> AA_exp <- (p^2) * n
> AB_exp <- (2*p*q) * n
> BB_exp <- (q^2) * n
> Gexp <- c(AA_exp, AB_exp, BB_exp)
> cat(paste("Expected genotype distribution: "), Gexp)
Expected genotype distribution:  439.0612 1156.878 762.0612
>
```

Frequências genotípicas esperada e observada

```
### Como saber se esta ou nao, em equilibrio? Teste qui-quadrado

chisq <- ((Gobs[1] - Gexp[1])^2 / Gexp[1]) + ((Gobs[2] - Gexp[2])^2 / Gexp[2]) + ((Gobs[3] - Gexp[3])^2 / Gexp[3])
cat(paste("Chisquare statistic (sum): "), chisq)


## O que esse valor implica?
x <- rchisq(10000, 1)
hist(x, prob=TRUE, col='grey85', breaks=40, main=expression(chi^2-"distribution (df=1)"), las=1, cex.main=.8)
curve( dchisq(x, df=1), col='red2', add=TRUE)


# 95% dos valores do teste qui-quadrado sao menores que 3.841
round(qchisq(0.95, 1), 3)


# 99% dos valores do teste qui-quadrado sao menores que 6.635
round(qchisq(0.99, 1), 3)


x <- rchisq(10000, 1)
hist(x, prob=TRUE, col='grey85', breaks=40, main=expression(chi^2-"distribution (df=1)"), las=1, cex.main=.8)
curve( dchisq(x, df=1), col='red2', add=TRUE)
abline(v=round(qchisq(0.95, 1), 3), col='blue', lwd=2)
abline(v=round(qchisq(0.99, 1), 3), col='seagreen', lwd=2)
legend("right", lwd=c(2,2), col=c('blue', 'seagreen'), c(expression(alpha==0.05), expression(alpha==0.01)),
       bty="n", cex=.9, ncol=2)
```

Frequências genotípicas esperada e observada

```
#### Para cada teste qui-quadrado, nos podemos calcular a propabilidade (P) de achar um valor na distribuicao,
#como por exemplo e.g., para chisq=3.841, a probabilidade de achar um valor maior e de 0.05001.

chisq <- 0.9658303
1-pchisq(chisq, df=1)

##O que isso significa??
##A probabilidade de achar um valor maior que 0.9658303 na distribuicao qui-quadrado e de 0.32.
## A populacao esta em equilibrio
```

Hardy-Weinberg Equilibrium - Função no R

Aqui temos a função R HWE ()
A função HWE () espera o número de animais para cada genótipo AA, AB e BB
A saída de HWE () é o resultado do teste qui-quadrado

```
HWE <- function(AA, AB, BB, alpha=0.05){
    Gobs <- c(AA, AB, BB)
    n <- sum(Gobs)
    p <- (AA*2 + AB) / (n*2)
    Gexp <- c((p^2)*n, (2*p*(1-p))*n, ((1-p)^2)*n)
    chisq_gt <- (Gobs-Gexp)^2 / (Gexp)
    chisq <- sum(chisq_gt)
    HWE_pval <<- signif(pchisq(chisq, df=1, lower.tail=FALSE), 5)
    cat(paste(" Observed genotype distribution: "), Gobs)
    cat(paste("\nMinor allele frequency: "), min(p, 1-p), "\n\n")
    cat(paste("\n Expected genotype distribution: "), signif(Gexp, 3))
    cat(paste("\n Chi-square (GTs): "), signif(chisq_gt, 3))
    cat(paste("\n Chi-square (sum): ", signif(chisq, 3), " (P=", HWE_pval, ")", sep=''))
    if (HWE_pval > alpha) {
        cat(paste("\n\n >>> RESULT...This population is in HWE <<<"))
    }
    else {
        cat(paste("\n\n >>> RESULT...This population is NOT in HWE <<<"))
    }
}
```

Hardy-Weinberg Equilibrium - Função no R

- Usando a função:
- HWE(453, 1129, 776)

```
> HWE(453, 1129, 776)
 Observed genotype distribution:  453 1129 776
Minor allele frequency:  0.4315098


 Expected genotype distribution:  439 1160 762
 Chi-square (GTs):  0.443 0.672 0.255
 Chi-square (sum): 1.37 (P=0.24194)

 >>> RESULT...This population is in HWE <<<
```

Minor Allele Frequency - MAF

função no R

```r
## MAF
file<- read.table("Genotipos_Pratica/Passar_alunos/geno_20210302.txt", head=T)

## Quantos SNPs tem no arquivo?
length(unique(file[,2]))
## Quantos individuos foram genotipados??
length(unique(file[,1]))

##Distribuicao genotipica para o SNP "ARS-BFGL-NGS-36479"
table(file[file[,2] == "ARS-BFGL-NGS-36479",3])

## Este SNP esta em equilibrio??
HWE(140,51,8)

##Call Rate por individuo #175
ind<- 175
nrow(file[file$Animal == ind & file$Genotipo == "NA" , ]) / nrow(file[file$Animal == ind ,])


##E pra todos os individuos?? Como fariamos??
## Loop

unique_ids<- cbind(unique(file[,1]), rep(0, length(unique(file[,1]))))

for (i in 1:nrow(unique_ids)){
  individual <- unique_ids[i, 1]
  unique_ids[i,2] <- nrow(file[file[,1] == individual & file[,3] == 'NA', ])/ nrow(file[file[,1] == individual, ])
}
```

Localização do Arquivo

Minor Allele Frequency - MAF

função no R

```
QC_snp <- function(snp, alpha=0.05){
  sub <- file[file[,2]==snp, ]
  alleles <- c(substr(sub$GT, 1, 1), substr(sub$GT, 2, 2))
  alleles <- alleles[alleles != '-']
  allele_A <- unique(alleles)[1]
  allele_B <- unique(alleles)[2]
  AA_char <- paste(allele_A, allele_A, sep="")
  AB_char <- paste(allele_A, allele_B, sep="")
  BB_char <- paste(allele_B, allele_B, sep="")
  missing <- nrow(sub[sub$GT == '--', ])
  gt_AA <- nrow(sub[sub$GT == AA_char, ])
  gt_BB <- nrow(sub[sub$GT == BB_char, ])
  gt_AB <- nrow(sub) - gt_AA - gt_BB - missing
  cols <- rep("orange", nrow(sub))    # heterozigotos
  cols[sub$GT =='--'] <- "grey"     # missing
  cols[sub$GT == AA_char] <- "red2"   # AA
  cols[sub$GT == BB_char] <- "seagreen" # BB
  freq_A <- length(alleles[alleles==allele_A]) / length(alleles[alleles != '-'])
  freq_B <- length(alleles[alleles==allele_B]) / length(alleles[alleles != '-'])
  freqs <- c(freq_A, freq_B)
  maf <- min(freqs)
  call_rate <- 1 - (missing / nrow(sub))
  cat(paste(" Observed genotypes: "), c(AA_char, AB_char, BB_char), "\n")
  cat(paste(" Call rate: "), call_rate, "\n")
  cat(paste(" Minor allele frequency: "), maf, "\n\n")
  cat(paste("---- Testando para HWE ----\n"))
  HWE(gt_AA, gt_AB, gt_BB, alpha)
  plot(sub$X, sub$Y, main=paste(snp, "\nHWE P=", HWE_pval), xlim=c(0, max(c(sub$X, sub$Y)) + .5*sd(c(sub$X, sub$Y))),
       ylim=c(0, max(c(sub$X, sub$Y)) + .5*sd(c(sub$X, sub$Y))), bg=cols, col="grey25", lwd=.3, pch=23, xlab="Channel 1",
       ylab="Channel 2", font.lab=2)
  legend("topright", pch=c(23, 23, 23, 23), pt.bg=c("grey65", "red2", "orange", "seagreen"), col="grey25",
         pt.cex=1.3, pt.lwd=.3, c("--", AA_char, AB_char, BB_char))
}
```

Limpeza dos Dados utilizando o software PLINK

- Critérios a serem escolhidos de acordo com o objetivo do estudo

```
./plink2 \ ## Comando para chamar o software PLINK -
    --bfile ~/Documents/PLINK2/genoIZ \ ## Arquivo com genotipos a serem limpos
     --cow \ ## Especie em estudo
     --hwe 1e-6 \ ## Criterio de limpeza para Hardy-Weinberg Equilibrium
     --maf 0.05 \ ## Criterio de limpeza para MINOR ALLELE FREQUENCY
     --mind 0.05 \ ## Criterio de limpeza para Animais/Amostras com baixa taxa de chamada
     --geno 0.05 \ ## Criterio de limpeza para variantes com baixa taxa de chamada
     --make-bed \ ## Output em formato bfile - .fam/.bim/.bed
     --out qc_hwe ## Nome do Arquivo de saida
```

Maiores informações no Manual: https://www.cog-genomics.org/plink/

Qualquer dúvida:

icdorff@gmail.com