REPORTE COMPLETO PRÁCTICA 7: Introducción a los estudios de asociación genómica

Jacqueline Vásquez

08-07-2021

ACTIVIDAD 1. Conexión con RStudio.cloud

Si tienes cuenta en la nube de RStudio puedes conectarte usando el siguiente link

Si no tienes cuenta, te recomiendo descargar RStudio en el siguiente link e instalarla en tu computadora.

ACTIVIDAD 2. Cargar librerías.

library(utils)

library(ggplot2)

library(rrBLUP)

library(dplyr)

ACTIVIDAD 3. Importar y explorar archivos de genotipos y fenotipos.

a) Importar los datos geno y pheno Importe el archivo de genotipos geno.txt y fenotipos Pheno.txt usando la función read.delim.

```
# Importar los archivos de genotipos
geno <- read.delim("geno.txt", sep = "\t", dec = ",", header = T)
head(geno[1:6,1:6])
```

```
##   marker chrom pos animal_1 animal_2 animal_3
## 1   snp1     1   1        1        0        1
## 2   snp2     1   2        0        0       -1
## 3   snp3     1   3        0        0       -1
## 4   snp4     1   4       -1        0        0
## 5   snp5     1   5        0        0        1
## 6   snp6     1   6        1        0       -1
```

```
# Importar los archivos de fenotipos
pheno <- read.delim("pheno.txt", sep = "\t", dec = ",", header = T)
head(pheno)
```

```
##     animal          y
## 1 animal_1 2.52813203
## 2 animal_2 1.44348297
## 3 animal_3 1.11555299
## 4 animal_4 2.01623154
## 5 animal_5 0.08267191
## 6 animal_6 1.80765254
```

b) Explorar datos. Luego realice un análisis exploratorio de ambos set de datos con las funciones head(), dim(). También realice un histograma de la variable cuantitativa y del archivo pheno, use la función hist().

```
# Explorar los archivos de genotipos y fenotipos
dim(geno)
```

```
## [1] 1000  203
```

```
head(geno[1:6,1:6])
```

```
##   marker chrom pos animal_1 animal_2 animal_3
## 1   snp1     1   1        1        0        1
## 2   snp2     1   2        0        0       -1
## 3   snp3     1   3        0        0       -1
## 4   snp4     1   4       -1        0        0
## 5   snp5     1   5        0        0        1
## 6   snp6     1   6        1        0       -1
```

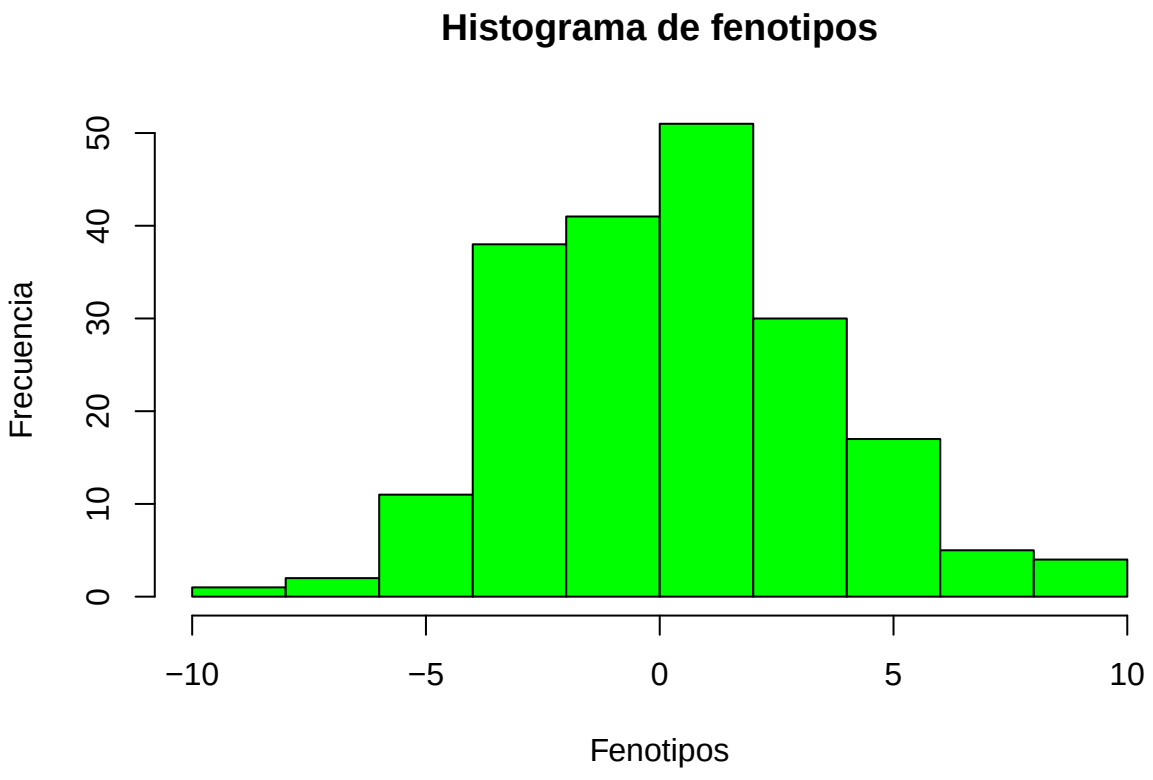
```
dim(pheno)
```

```
## [1] 200   2
```

```
head(pheno)
```

```
##     animal          y
## 1 animal_1 2.52813203
## 2 animal_2 1.44348297
## 3 animal_3 1.11555299
## 4 animal_4 2.01623154
## 5 animal_5 0.08267191
## 6 animal_6 1.80765254
```

```
hist(pheno$y, main="Histograma de fenotipos",xlab="Fenotipos",ylab="Frecuencia",col="green")
```



1. ¿Cómo están codificados los genotipos de cada polimorfismo de nucleótido único (SNP)?
Respuesta: Los genotipos estan codificados como un número de dosis de un solo alelo (-1,0 y 1) donde:

-1 homocigoto recesivo

0 heterocigoto

1 homocigoto dominante

2. ¿Observa heterocigotos? Respuesta: Sí, y estan representados con el valor 0

ACTIVIDAD 4. Estudio de asociación de genoma completo (GWAS)

```
# Investigar el codigo A.mat
help("A.mat")
# Este codigo Calcula la matriz de relación aditiva realizada.
```

```
# Cálculo y gráfica de la matriz de parentesco genómico según método de Van Raden para los 200 animales
A <- A.mat(geno[4:203])
dim(A)
```

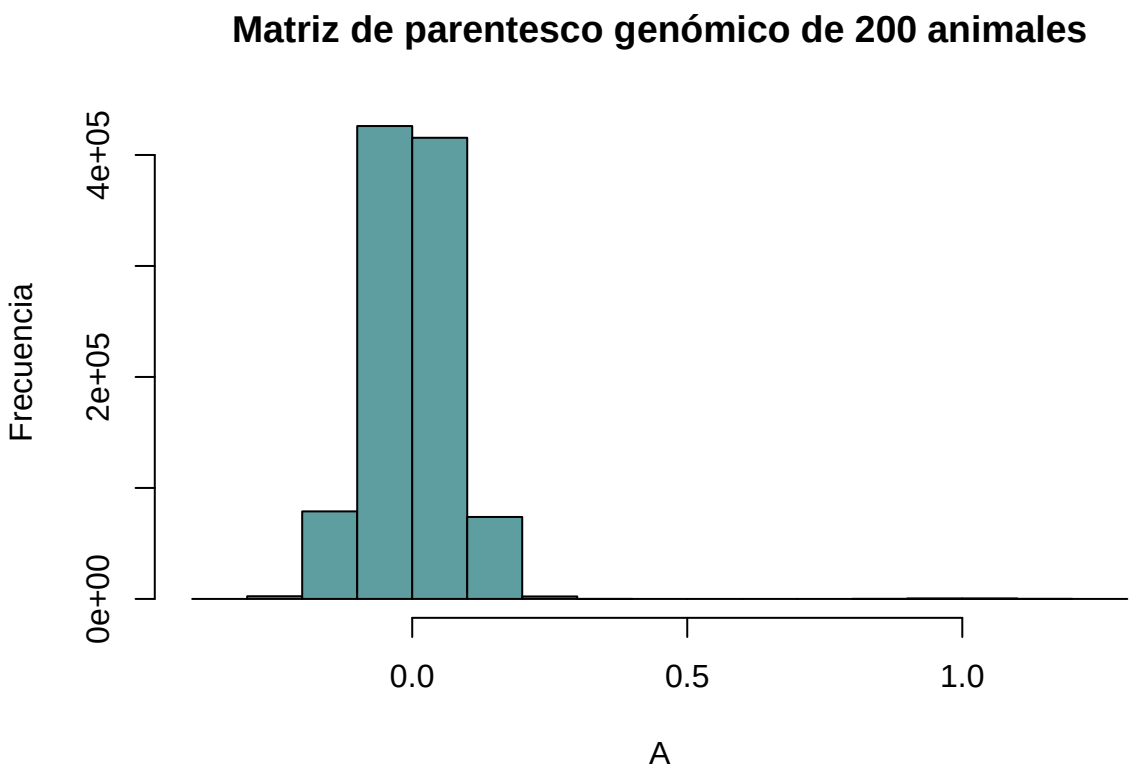
a) Investigue el uso de la función A.mat de la librería rrBLUP y calcule la matriz de parentesco genómico para el set de datos geno. Explore la matriz con las funciones dim(), head() y hist().

```
## [1] 1000 1000
```

```
head(A[1:6,1:6])
```

```
##             [,1]        [,2]        [,3]        [,4]        [,5]        [,6]
## [1,]  1.09783609  0.03753355 -0.03269311 -0.04195663  0.06084240 -0.02389977
## [2,]  0.03753355  0.98799439 -0.10261966 -0.07186798 -0.08911453 -0.02379974
## [3,] -0.03269311 -0.10261966  0.98759424 -0.18210984  0.05073856  0.06603437
## [4,] -0.04195663 -0.07186798 -0.18210984  1.04909759  0.03147125  0.07677845
## [5,]  0.06084240 -0.08911453  0.05073856  0.03147125  1.05461969 -0.07051747
## [6,] -0.02389977 -0.02379974  0.06603437  0.07677845 -0.07051747  1.05519991
```

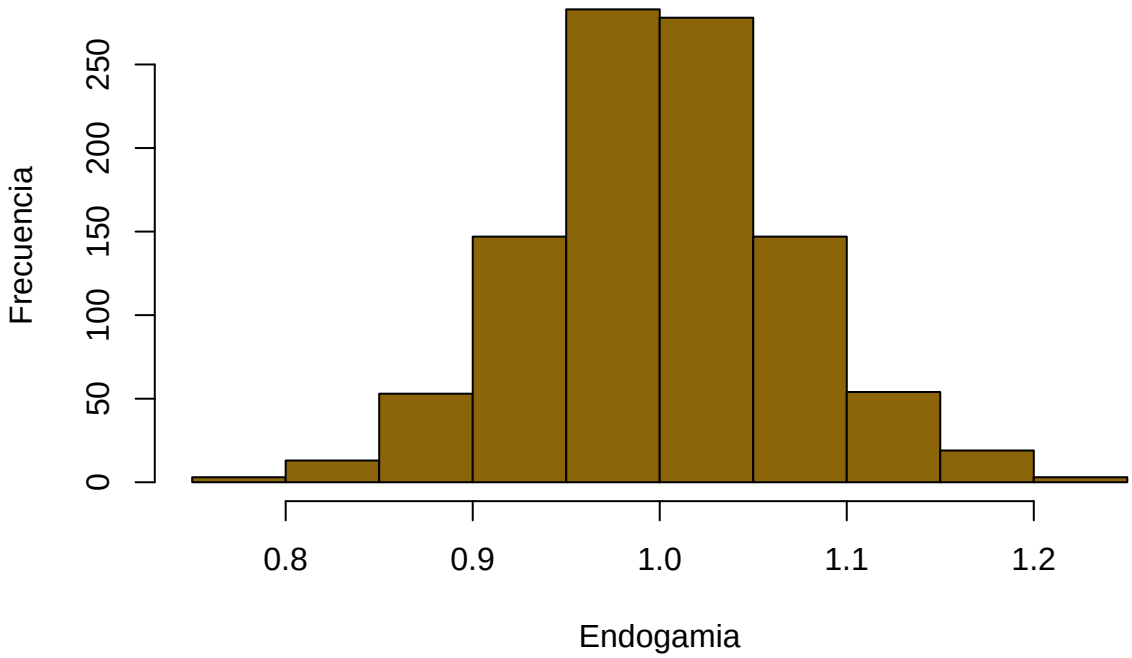
```
hist(A, main="Matriz de parentesco genómico de 200 animales",xlab="A",ylab="Frecuencia",col="cadetblue"
```



b) Ahora, cree un objeto llamado endogamia con la diagonal de la matriz y grafique con hist().

- La diagonal contiene el parentesco del individuo consigo mismo (E(diagonal)= 1+f), y permite estimar el coeficiente de endogamia de la población. (Endelman and Jannink, 2012)

```
endogamia <- diag(A)
hist(endogamia, main = "Histograma de endogamia", xlab="Endogamia",ylab="Frecuencia",col="darkgoldenrod4
```




```
summary(endogamia)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7687  0.9565  1.0001  1.0003  1.0439  1.2441
```

```
sd(endogamia)
```

```
## [1] 0.06850077
```

1. ¿Cuál es el nivel de endogamia promedio de esta población?. Respuesta: El valor de endogamia promedio de la población es de 1.0003 con una desviación estándar de 0.069

2. ¿Qué significa un valor de endogamia de 1.1? Respuesta: Un valor de endogamia sobre 1, como en el caso de 1.1, representa un coef de consanguinidad positivo en los individuos de la población analizada.

3. ¿Qué representa un valor de endogamia de 0.90? Respuesta: Un valor de endogamia bajo 1, como en el caso de 0.90, representa un coef de consanguinidad negativo en los individuos de la población analizada.

```
# Investigar el código GWAS
help("GWAS")
# Este código realiza en análisis de asociación de todo el genoma basado en el modelo mixto de [Yu et a
```

```
# Análisis GWAS

score <- GWAS(pheno,geno,plot=TRUE)
```

c) Investigue el uso de la función GWAS() de la librería rrBLUP y realice un análisis de asociación genómica GWAS. Incluya el argumento plot=TRUE. Almacene el resultado del

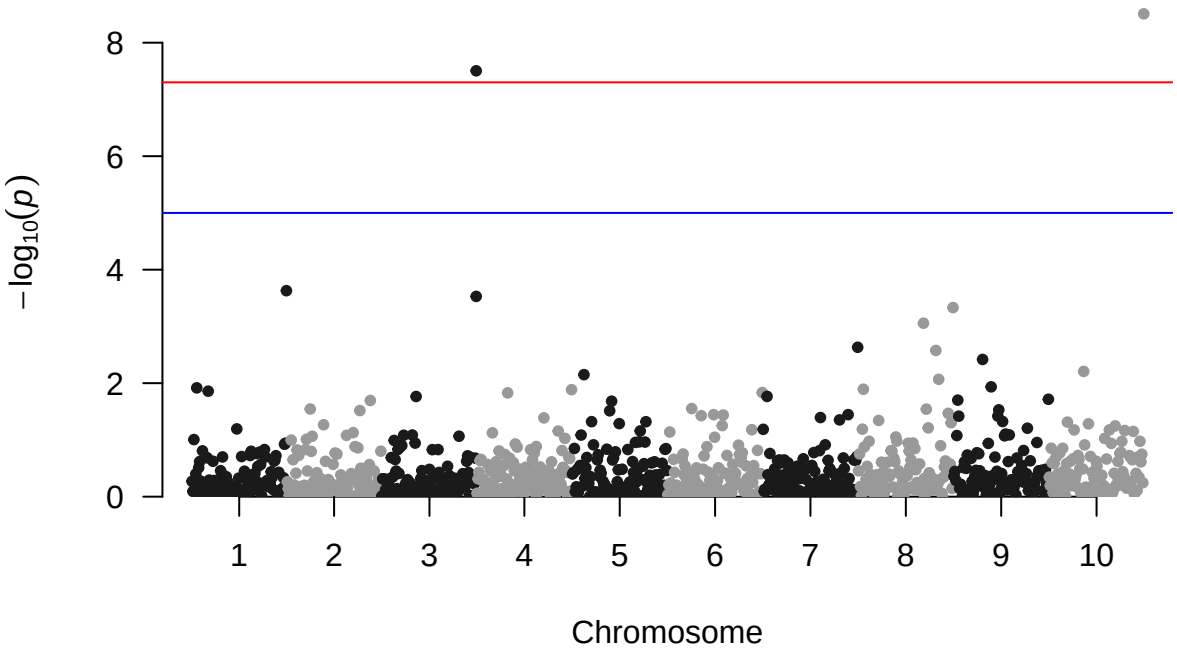
GWAS en un objeto denominado score y responda:

```
## [1] "GWAS for trait: y"
## [1] "Variance components estimated. Testing markers."
```

```
### Calcula la probabilidad de significancia a partir del inverso de y=-log10(p).

score$p <- 1/(10^score$y)
```

```
# Grafica manhattan plot
manhattan(score, chr="chrom", bp="pos", snp="marker", p="p" )
```




```
# Determinar que tipo de objeto es scrore
class(score)
```

```
## [1] "data.frame"
```

1. Visualmente, ¿Cuantos QTLs fueron detectados por el análisis GWAS? ¿En qué cromosomas se encuentran? Respuesta: La gráfica GWAS muestra 2 QTLs significativos en el análisis (sobre la línea roja) y corresponden al cromosoma 3 y 10 respectivamente.

ACTIVIDAD 5. locus que afecta un rasgo cuantitativo (QTLs)

Ahora exploraremos el efecto de los QTLs detectados por el GWAS

```
# View(scores)
head(score)
```

a) Explore el objeto score con el comando head() y View().

```
##   marker chrom pos          y          p
## 1   snp1     1   1 0.27068411 0.53618652
## 2   snp2     1   2 0.09321965 0.80682686
## 3   snp3     1   3 1.00621907 0.09857821
## 4   snp4     1   4 0.22399674 0.59703977
```

```
## 5   snp5     1   5 0.40095387 0.39723374
## 6   snp6     1   6 1.91710405 0.01210308
```

Sugerencia Extraiga el score de los SNP significativos (score>5) usando la función filter() de la librería **dplyr*

```
dplyr::filter(score, y > 5)
```

```
##    marker chrom pos        y            p
## 1  snp300     3 100 7.504724 3.128069e-08
## 2 snp1000    10 100 8.508100 3.103847e-09
```

Sugerencia note que el score corresponde a -log(p), donde p es la probabilidad o significancia. Transforme el score a p usando exp()

```
# snp300
exp(-7.5047236)
```

```
## [1] 0.000550478
```

```
# snp1000
exp(-8.5080997)
```

```
## [1] 0.000201827
```

```
# pruebe calcular log(-0.000550478)
-log(0.000550478)
```

```
## [1] 7.504724
```

```
-log10(0.0000008)
```

```
## [1] 6.09691
```

```
1/(10^6.09691)
```

```
## [1] 8e-07
```

1. ¿Qué SNP fueron significativos? Respuesta: Los polimorfismos de nucleótido único que fueron representativos son snp300 y snp1000 correspondientes al cromosoma 3 y 10 respectivamente.

2. ¿Con que nivel de significancia se concluye que fueron significativos?. Respuesta: SNP300 tiene una significancia alta p-value=0.00055 y SNP1000 tambien tiene una significancia alta p-value= 0.0002 siendo SNP1000 más significativo que el SNP300

b) Realice un gráfico de regresión lineal de fenotipo en función de los genotipo para cada SNP significativo. Sugerencia: Transponga la matriz geno y cree un nuevo data.frame solo con los snp signiticativos, luego una al data.frame el rasgo cuantitativo.

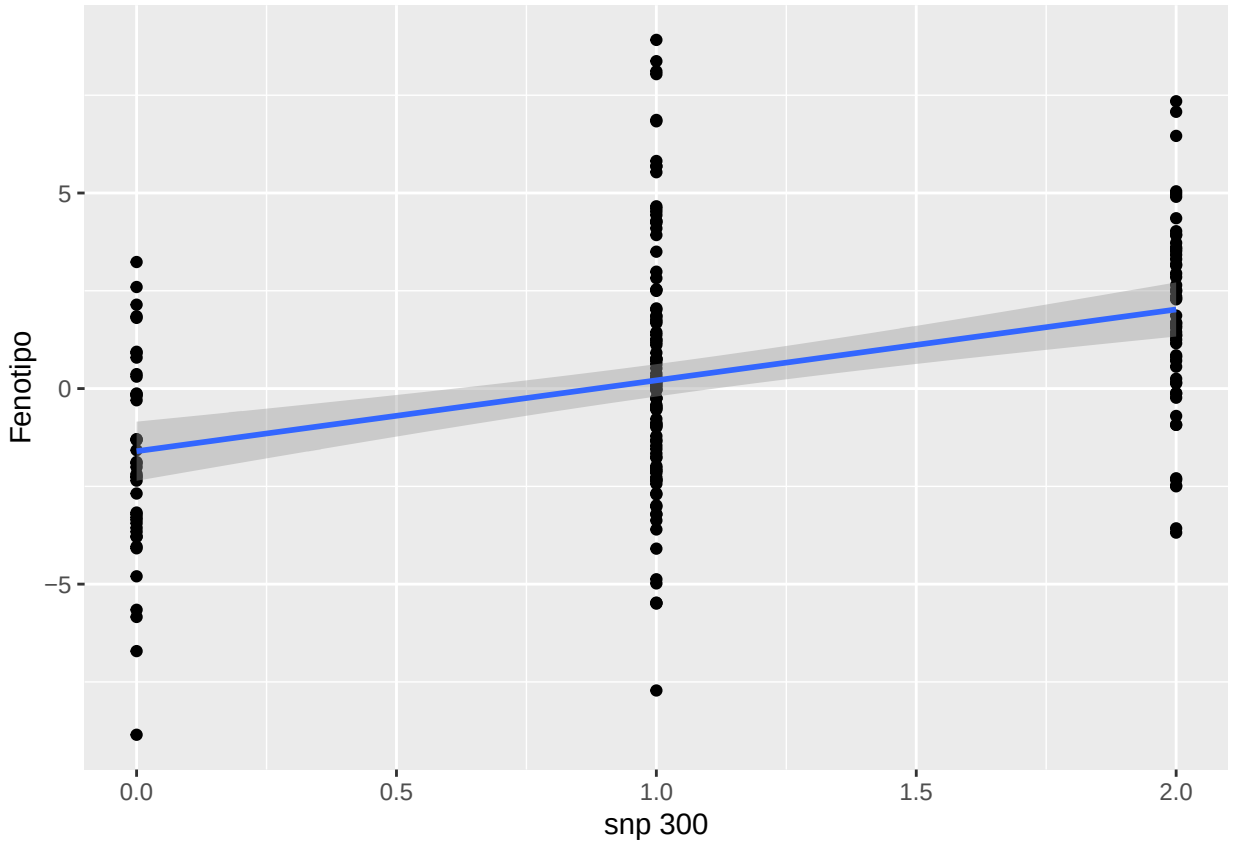
```
# SNP significativos
t_geno_300 <- t(geno[300,4:203])+1
t_geno_1000 <- t(geno[1000,4:203])+1
qtl <- data.frame(t_geno_300,t_geno_1000,pheno$y)
head(qtl)
```

```
##          X300 X1000    pheno.y
## animal_1    1     1 2.52813203
## animal_2    2     0 1.44348297
## animal_3    1     1 1.11555299
## animal_4    1     1 2.01623154
```

```
## animal_5    1     0 0.08267191
## animal_6    0     2 1.80765254
```

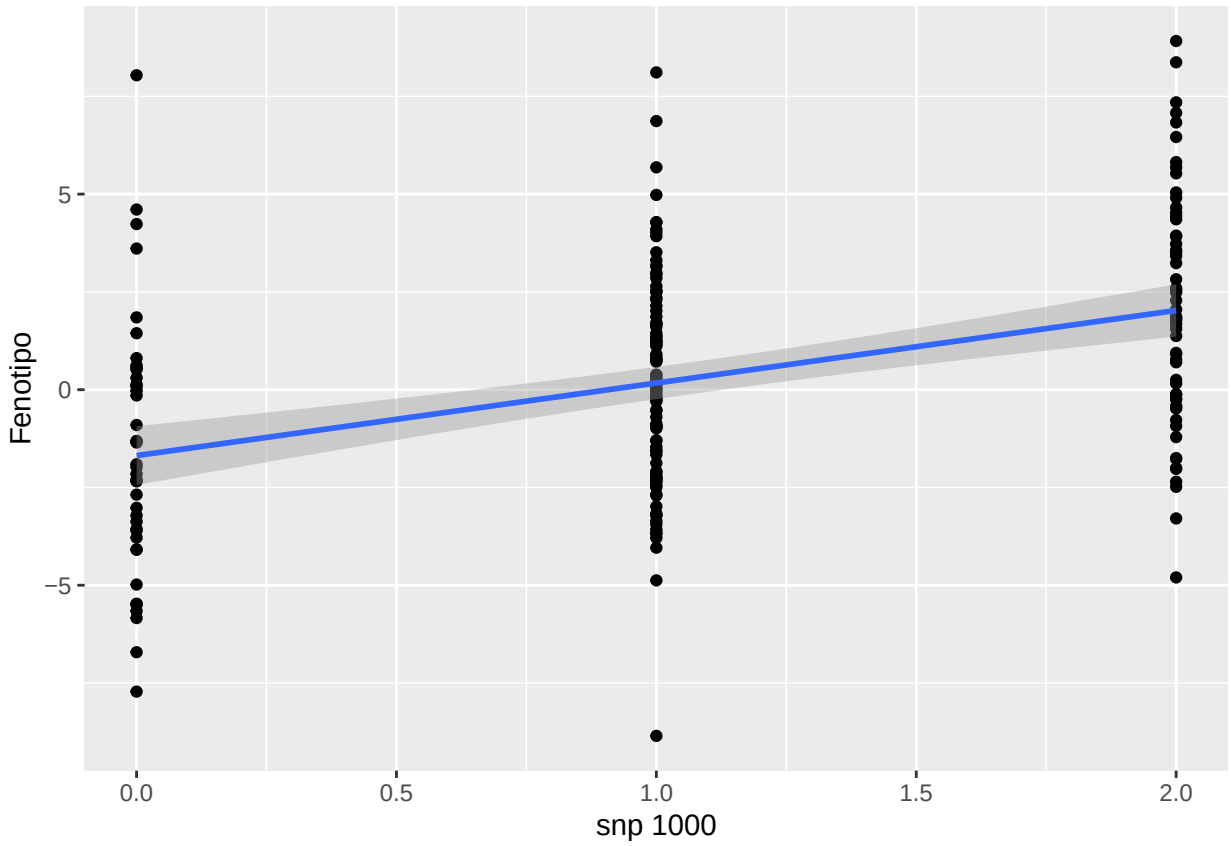
```
qtl.1 <- ggplot(qtl, aes(x = X300, y = pheno.y))
qtl.1 + geom_point() + xlab("snp 300") +  ylab("Fenotipo")+ geom_smooth(method=lm)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
qtl.2 <- ggplot(qtl, aes(x = X1000, y = pheno.y))
qtl.2 + geom_point() + xlab("snp 1000") +  ylab("Fenotipo")+ geom_smooth(method=lm)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
#SNP300
lm.qtl.300 <- lm(pheno.y ~ X300, data = qtl)
summary(lm.qtl.300)
```

c) Estime el efecto (beta o pendiente) de los QTLs con mayor score usando un modelo lineal lm().

```
##
## Call:
## lm(formula = pheno.y ~ X300, data = qtl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9292 -1.9076 -0.2454  1.5984  8.7040
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.6019     0.3831  -4.181 4.35e-05 ***
## X300          1.8121     0.3015   6.011 8.72e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.956 on 198 degrees of freedom
## Multiple R-squared:  0.1543, Adjusted R-squared:  0.1501
## F-statistic: 36.13 on 1 and 198 DF,  p-value: 8.717e-09
```

```r
cat("El efecto del snp300 sobre el rasgo cuantitativo 1.8121")
```

```
## El efecto del snp300 sobre el rasgo cuantitativo 1.8121
```

```r
#SNP1000
lm.qtl.1000 <- lm(pheno.y ~ X1000, data = qtl)
summary(lm.qtl.1000)
```

```
##
## Call:
## lm(formula = pheno.y ~ X1000, data = qtl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0213 -2.2599 -0.1655  1.9001  9.7233
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.6845     0.3805  -4.428 1.57e-05 ***
## X1000         1.8549     0.2940   6.310 1.79e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.933 on 198 degrees of freedom
## Multiple R-squared:  0.1674, Adjusted R-squared:  0.1632
## F-statistic: 39.81 on 1 and 198 DF,  p-value: 1.792e-09
```

```r
cat("El efecto del snp1000 sobre el rasgo cuantitativo 1.8549")
```

```
## El efecto del snp1000 sobre el rasgo cuantitativo 1.8549
```