PROJET FINAL STATISTIQUES POUR LA GÉNÉTIQUE 2019-2020

UNIVERSITÉ DE PARIS / MASTER 1 INGÉNIERIE MATHÉMATIQUE ET BIOSTATISTIQUE

James Kelson LOUIS

4/30/2020

Exercice 1

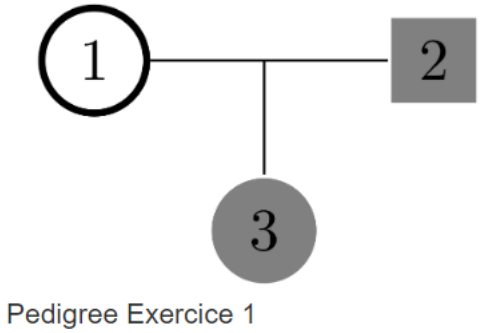


Figure 1: A caption

On se pose le problème d'identifier le modèle sous-jacent à une maladie donnée par analyse de ségrégation. Pour cela on considère le pedigree de la figure dans lequel les individus en gris sont atteints par la maladie. On note $\mathbf{G_i}$ le génotype de l'individu i et $\mathbf{P_i}$ son phénotype (1 pour malade et 0 pour sain).

On fait l'hypothèse que la maladie est une maladie génétique due à mutation dans un seul locus d'allèles S,s. On suppose que la fréquence de l'allèle de susceptibilité S est q et que il y a équilibre de Hardy-Weinberg.

On commence par considérer un modèle de maladie dominant à pénétrance complète.

1. Écrivons les fonctions de pénétrance $\mathbf{P}(P = \mathbf{1}|G = \mathbf{g})$ pour tous les trois génotypes possibles.

a) $\mathbf{P}(P = \mathbf{1}|G = \mathbf{SS}) = 1$

b) $\mathbf{P}(P = \mathbf{1}|G = \mathbf{Ss}) = 1$

c) $\mathbf{P}(P = \mathbf{1}|G = \mathbf{ss}) = 0$

2. Trouvons les génotypes possibles pour chaque individu.

Il s'agit d'un modèle de maladie dominant à pénétrance complète, et d'après le pedigree l'individu 1 n'est pas malade et les autres sont malades, forcémment $\mathbf{G_1} = \mathbf{ss}$.

Pour l'individu 2 on peut considérer deux cas.

$G2\|G1$	ss	
Ss	Ss	ss
SS	Ss	

a) Si $\mathbf{G_2} = \mathbf{Ss}$, alors $\mathbf{G_3} = \mathbf{Ss}$ ou $\mathbf{G_3} = \mathbf{ss}$, par contre on ne peut pas avoir $\mathbf{G_3} = \mathbf{ss}$ pour le modèle considéré car l'individu 3 est malade.

b) Si $\mathbf{G_2} = \mathbf{SS}$ alors $\mathbf{G_3} = \mathbf{Ss}$.

Au final les génotypes possibles sont: $\mathbf{G_1} = \mathbf{ss}$, $\mathbf{G_2} = \mathbf{Ss}$ ou $\mathbf{G_2} = \mathbf{SS}$ et $\mathbf{G_3} = \mathbf{Ss}$.

3. Pour tout génotype possible g_1 pour l'individu 1, calculons $\mathbf{P}(G_1 = g_1)$.

Pour l'individu 1 $g_1 = ss$ Sous l'hypothèse Hardy-Weinberg
$\mathbf{P}(G_1 = \mathbf{ss}) = (1 - q)^2$.

Pour l'individu 2 :

Si $g_2 = Ss$, sous l'hypothèse Hardy-Weinberg
$\mathbf{P}(G_2 = \mathbf{Ss}) = 2 \times q \times (1 - q)$.

Si $g_2 = SS$, sous l'hypothèse Hardy-Weinberg
$\mathbf{P}(G_2 = \mathbf{SS}) = q^2$.

4. Pour chaque combinaison de génotypes possibles (g_1, g_2, g_3), calculons $\mathbf{P}(G_3 = g_3|G_1 = g_1, G_2 = g_2)$
$\mathbf{P}(G_3 = \mathbf{Ss}|G_1 = \mathbf{ss}, G_2 = \mathbf{Ss}) = \frac{1}{2}$

$\mathbf{P}(G_3 = \mathbf{Ss}|G_1 = \mathbf{ss}, G_2 = \mathbf{SS}) = 1$.

5. Écrivons la vraissemblance du trio sous ce modèle de maladie, à l'aide des questions précédentes. $\mathbf{V_1} = \mathbf{P}(P_1 = \mathbf{0} | P_2 = \mathbf{1}, P_3 = \mathbf{1})$
Pour calculer cette probabilité, on utilise la règle de la chaine.

$V = \mathbf{P}(P_1 = 0) \times \mathbf{P}(P_2 = \mathbf{1} | P_1 = \mathbf{0}) \times \mathbf{P}(P_3 = \mathbf{1} | P_1 = \mathbf{0}, P_2 = \mathbf{1})$

** $\mathbf{P}(P_1 = 0) = \mathbf{P}(P_1 = 0 | G = \mathbf{SS}) \times \mathbf{P}(G = SS) + \mathbf{P}(P_1 = 0 | G = \mathbf{Ss}) \times \mathbf{P}(G = Ss) + \mathbf{P}(P_1 = 0 | G = \mathbf{ss}) \times \mathbf{P}(G = ss)$

** $\mathbf{P}(P_1 = 0) = 0 \times 1 + 0 \times 2q(1-q) + 1 \times (1-q)^2 = (1-q)^2$

** $P_2 \perp P_1$

** $\mathbf{P}(P_2 = \mathbf{1} | P_1 = \mathbf{0}) = \mathbf{P}(P_2 = \mathbf{1})$

** $\mathbf{P}(P_2 = 1) = \mathbf{P}(P_2 = 1 | G = \mathbf{SS}) \times \mathbf{P}(G = SS) + \mathbf{P}(P_2 = 1 | G = \mathbf{Ss}) \times \mathbf{P}(G = Ss) + \mathbf{P}(P_2 = 1 | G = \mathbf{ss}) \times \mathbf{P}(G = ss)$

** $\mathbf{P}(P_2 = 1) = 1 \times q^2 + 1 \times 2q(1-q) + 0 \times (1-q)^2 = q^2 + 2q(1-q) = q(2-q)$

** $\mathbf{P}(P_3 = 1 | P_1 = 0, P_2 = 1) = \mathbf{P}(P_3 = 1 | G_3 = Ss) \times \mathbf{P}(G_3 = Ss | G_1 = ss, G_2 = SS) + \mathbf{P}(P_3 = 1 | G_3 = Ss) \times \mathbf{P}(G_3 = Ss | G_1 = ss, G_2 = Ss)$

** $\mathbf{P}(P_3 = 1 | P_1 = 0, P_2 = 1) = 1 \times 1 + 1 \times \frac{1}{2} = \frac{3}{2}$

Il vient alors:
$\mathbf{V_1} = (1-q)^2 \times q(2-q) \times \frac{3}{2}$

6. Calculons la vraissemblance du trio si $q = \frac{1}{2}$

Pour $q = \frac{1}{2}$

$\mathbf{V_1} = \frac{9}{32}$

ou $\boxed{\mathbf{V_1} = 0.28125}$

Pour l'instant on va considérer un modèle de maladie récessif à pénétrance complète.

Les fonctions de pénétrances pour ce modèle sont:

a) $\mathbf{P}(P = \mathbf{1} | G = \mathbf{SS}) = 1$

b) $\mathbf{P}(P = \mathbf{1} | G = \mathbf{Ss}) = 0$

c) $\mathbf{P}(P = \mathbf{1}|G = \mathbf{ss}) = 0$

Trouvons les génotypes possibles pour ce modèle

L'individu 1 n'est pas malade, son génotype est soit $\mathbf{G_1} = \mathbf{ss}$ soit $\mathbf{G_1} = \mathbf{Ss}$. L'individu 2 est malade, forcément son génotype est $\mathbf{G_2} = \mathbf{SS}$, pareil pour l'individu 3, $\mathbf{G_3} = \mathbf{SS}$.

$G1\|G2$	SS	
ss	Ss	
Ss	SS	Ss

1^{er} **cas, si $\mathbf{G_1} = \mathbf{ss}$ et $\mathbf{G_2} = \mathbf{SS}$, alors $\mathbf{G_3} = \mathbf{Ss}$ ce qui est impossible pour le modèle considéré, car l'individu 3 est malade.**

$2^{ème}$ **cas, si $\mathbf{G_1} = \mathbf{Ss}$ et $\mathbf{G_2} = \mathbf{SS}$, alors soit $\mathbf{G_3} = \mathbf{SS}$ soit $\mathbf{G_3} = \mathbf{Ss}$, le même raisonnement montre que ce dernier cas $\mathbf{G_3} = \mathbf{Ss}$ n'est pas possible.**

Les génotypes possibles sont: $\mathbf{G_1} = \mathbf{Ss}$, $\mathbf{G_2} = \mathbf{SS}$ et $\mathbf{G_3} = \mathbf{SS}$

Calculons la vraissemblance de ce trio pour ce modèle.
$\mathbf{V_2} = \mathbf{P}(P_1 = \mathbf{0}|P_2 = \mathbf{1}, P_3 = \mathbf{1})$

$\mathbf{V_2} = \mathbf{P}(P_1 = 0) \times \mathbf{P}(P_2 = \mathbf{1}|P_1 = \mathbf{0}) \times \mathbf{P}(P_3 = \mathbf{1}|P_1 = \mathbf{0}, P_2 = \mathbf{1})$

** $\mathbf{P}(P_1 = 0) = \mathbf{P}(P_1 = 0|G = \mathbf{SS}) \times \mathbf{P}(G = SS) + \mathbf{P}(P_1 = 0|G = \mathbf{Ss}) \times \mathbf{P}(G = Ss) + \mathbf{P}(P_1 = 0|G = \mathbf{ss}) \times \mathbf{P}(G = ss)$

** $\mathbf{P}(P_1 = 0) = 0 \times q^2 + 1 \times 2q(1-q) + 1 \times (1-q)^2 = 1 - q^2$

** $\mathbf{P}(P_2 = 1|P_1 = 0) = \mathbf{P}(P_2 = \mathbf{1})$

** $\mathbf{P}(P_2 = 1) = \mathbf{P}(P_2 = 1|G = \mathbf{SS}) \times \mathbf{P}(G = SS) + \mathbf{P}(P_2 = 1|G = \mathbf{Ss}) \times \mathbf{P}(G = Ss) + \mathbf{P}(P_2 = 1|G = \mathbf{ss}) \times \mathbf{P}(G = ss)$

** $\mathbf{P}(P_2 = 1) = 1 \times q^2 + 0 \times 2q(1-q) + 0 \times (1-q)^2 = q^2$

** $\mathbf{P}(P_3 = 1|P_1 = 0, P_2 = 1) = \mathbf{P}(P_3 = 1|G_3 = SS) \times \mathbf{P}(G_3 = SS|G_1 = Ss, G_2 = SS)$

** $\mathbf{P}(P_3 = 1|P_1 = 0, P_2 = 1) = 1 \times \frac{1}{2} = \frac{1}{2}$

La vraissemblance s'écrit:

$\mathbf{V_2} = (1 - q^2) \times q^2 \times \frac{1}{2}$

$\mathbf{V_2} = (1 - (\frac{1}{2})^2) \times (\frac{1}{2})^2 \times \frac{1}{2}$

Soit $\mathbf{V_2} = \frac{3}{32}$

ou $\boxed{\mathbf{V_2} = 0.09375}$

Enfin, on suppose que la maladie n'a pas de composante génétique. On note F la fréquence de la maladie dans la population.

8. Écrivons la vraissemblance du trio observé sous cette nouvelle hypothèse.
$\mathbf{V_3} = (1 - F) \times F^2$

9. Calculons la vraissemblance du trio sous cette nouvelle hypothèse, si $F = \frac{1}{20}$.
$\mathbf{V_3} = (1 - \frac{1}{20}) \times (\frac{1}{20})^2$

$\mathbf{V_3} = \frac{19}{20} \times (\frac{1}{400})$

$\mathbf{V_3} = \frac{19}{8000}$

$\boxed{\mathbf{V_3} = 0.002375}$

10. Au vu du trio observé, nous avons $\mathbf{V_3} < \mathbf{V_2} < \mathbf{V_1}$, donc le modèle de maladie le plus probable est le modèle 1.

Exercice 2

Explorations préliminaires-covariables

1 Charegement du fichier des données

```
hgdp<- read.delim("C:/Users/james/OneDrive/Desktop/DocParisDescartes/S2/StatGen2/HGDP_AKT1.txt", header=
```

```
attach(hgdp)
```

2a Trouvons le nombre d'observations (individus)

```
nrow(hgdp)
```

```
## [1] 1064
```

2b Trouvons le nombre de SNPs disponibles, sachant que les SNPs dans le gène AKT1 sont dénotés avec le préfixe AKT1

```
SNP_AKT1 <- names(hgdp)[startsWith(names(hgdp),"AKT1")]
length(SNP_AKT1)
```

```
## [1] 4
```

3a Le nombre de femmes et d'hommes dans l'étude est:

```
table(Gender)
```

```
## Gender
##   F   M
## 380 684
```

Soit 380 Femmes et 684 Hommes.

3b Le nombre de populations dans l'étude est:

```
length(levels(Population))
```

```
## [1] 52
```

3b Le nombre de zones géographiques regroupant plusieurs pays on a des données est:

```
length(unique(Geographic.area))
```

```
## [1] 14
```

```
length(summary(Geographic.area))
```

```
## [1] 14
```

4a Les Populations les plus représentées sont:

```
Plus <- rev(tail(as.factor(sort(summary(Population))),4))
Plus1 <- as.numeric(rev(levels(Plus)))[1:4]
names(Plus1) <- attributes(Plus)$names
Plus1
```

```
## Palestinian     Bedouin       Druze         Han
##          51          49          48          45
```

4b Les Populations les moins représentées sont:

```
Moins <- (head(as.factor(sort(summary(Population))),4))
Moins1 <- as.numeric(levels(Moins))[1:4]
names(Moins1) <- attributes(Moins)$names
Moins1
```

```
##   San  Tuscan   Xibo    Dai
##     7      8      9     10
```

4c Les zones géographiques les plus représentées sont:

```
plus <- rev(tail(as.factor(sort(summary(Geographic.area))),4))
plus1 <- as.numeric(rev(levels(plus)))[1:4]
names(plus1) <- attributes(plus)$names
plus1
```

```
##        Pakistan           China          Israel Southern Europe
##             200             184             148             125
```

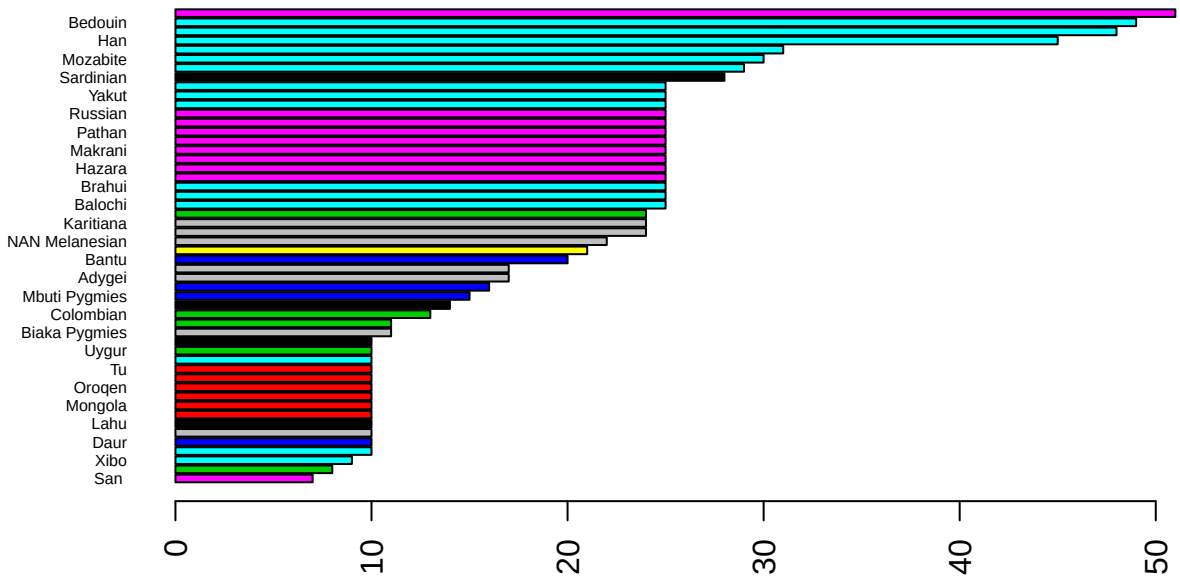
4d Les zones géographiques les moins représentées sont:

```
moins <- (head(as.factor(sort(summary(Geographic.area))),4))
moins1 <- as.numeric(levels(moins))[1:4]
names(moins1) <- attributes(moins)$names
moins1
```

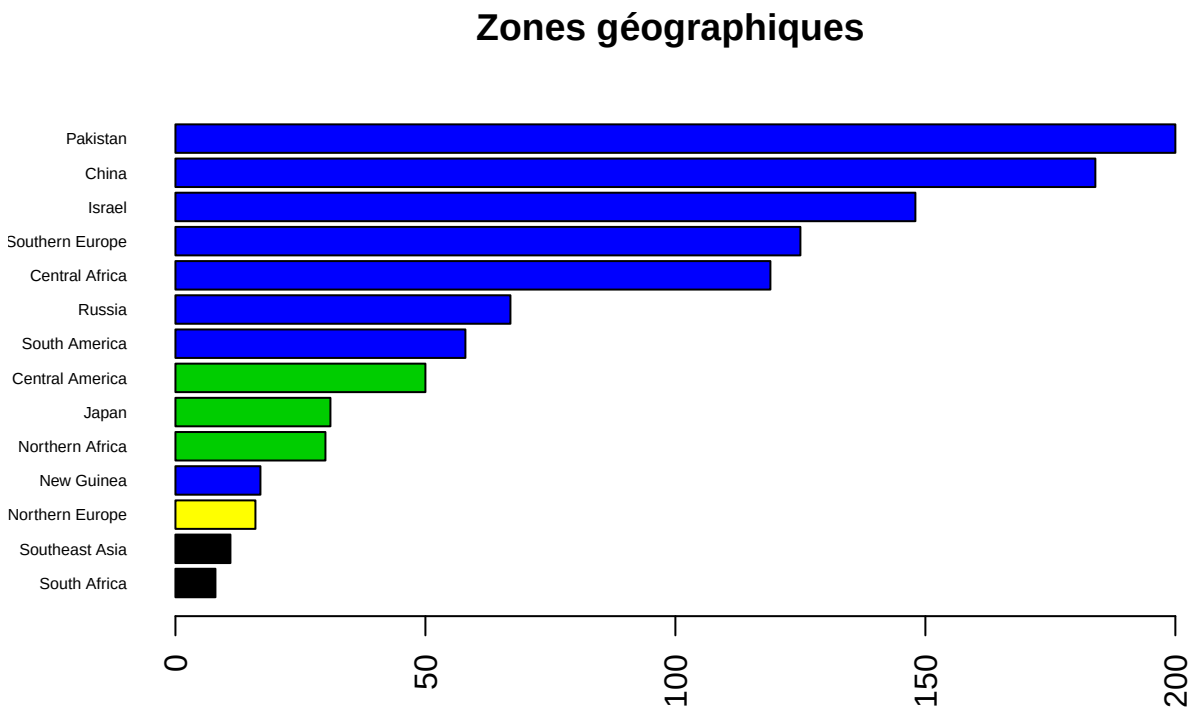
```
##    South Africa  Southeast Asia Northern Europe      New Guinea
##               8              11              16              17
```

4e Représentons graphiquement ces deux variables à l'aide d'un diagramme en batons

```
barplot(sort(table(Population)),horiz=TRUE,las=2,cex.names=0.5, col = Population,main=" Population")
```




```
barplot(sort(table(Geographic.area)),horiz=TRUE,las=2,cex.names=0.5, col = Geographic.area, main="Zones
```



Exercice 3

Explorations préliminaires-génotypes

1 Estimons les fréquences alléliques et génotypiques du SNP AKT1.C6024T, on suppose que les fréquences génotypiques ne changent pas en considérant uniquement les génotypes observés. Car, si un génotype est plus difficile à observer par voie expérimentale que les autres, sous cette hypothèse on sous-estime sa fréquence réelle.

```
geno1 <- genotype(AKT1.C6024T, sep="")
```

```
summary(geno1)
```

```
##
## Number of samples typed: 1063 (99.9%)
##
## Allele Frequency: (2 alleles)
##    Count Proportion
## C   1732       0.81
## T    394       0.19
## NA     2         NA
##
##
```

```
## Genotype Frequency:
##     Count Proportion
## C/C   719       0.68
## C/T   294       0.28
## T/T    50       0.05
## NA      1         NA
##
## Heterozygosity (Hu)  = 0.3021008
## Poly. Inf. Content   = 0.2563692
```

Affichons les proportions des génotypes de AKT1.C6024T pour chaque zone géographique à l'aide d'un mosaicplot

```
(gen <- table(Geographic.area, AKT1.C6024T))
```

```
##                  AKT1.C6024T
## Geographic.area    CC  CT  TT
##   Central Africa   55  50  14
##   Central America  39  10   1
##   China           142  39   3
##   Israel          106  36   6
##   Japan            19  12   0
##   New Guinea       16   1   0
##   Northern Africa  17  11   2
##   Northern Europe   9   7   0
##   Pakistan        140  52   8
##   Russia           47  15   5
##   South Africa      5   3   0
##   South America    46   5   7
##   Southeast Asia    9   1   0
##   Southern Europe  69  52   4
```

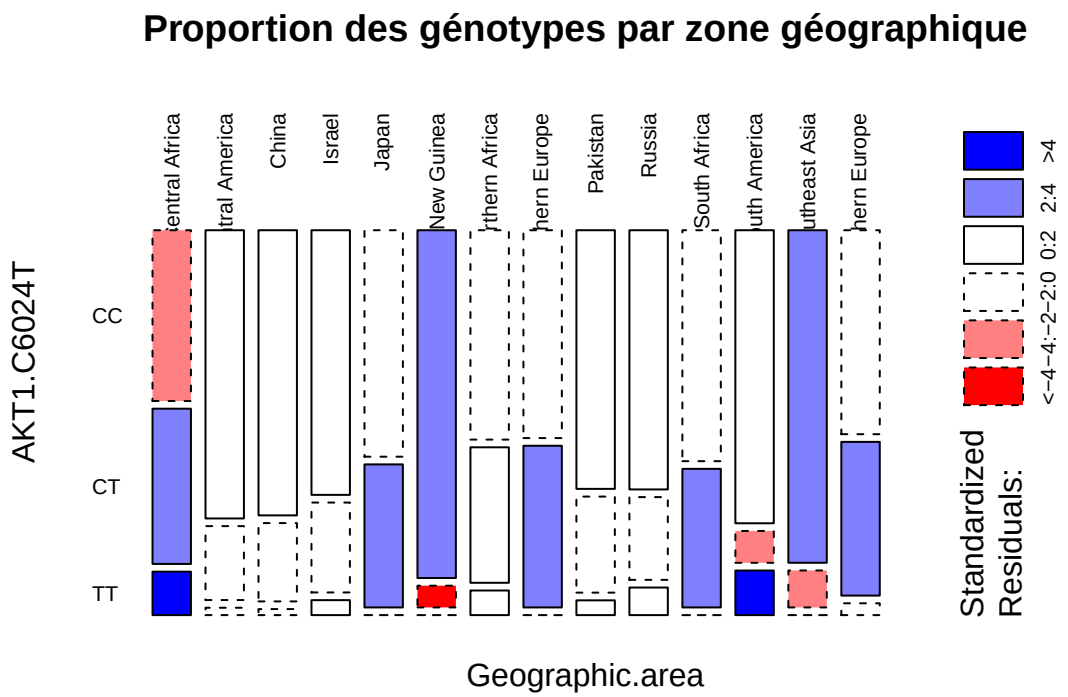
```
(prop_gen <- prop.table(gen,1)*100)
```

Proportions

```
##                  AKT1.C6024T
## Geographic.area          CC        CT        TT
##   Central Africa  46.218487 42.016807 11.764706
##   Central America 78.000000 20.000000  2.000000
##   China           77.173913 21.195652  1.630435
##   Israel          71.621622 24.324324  4.054054
##   Japan           61.290323 38.709677  0.000000
##   New Guinea      94.117647  5.882353  0.000000
##   Northern Africa 56.666667 36.666667  6.666667
##   Northern Europe 56.250000 43.750000  0.000000
##   Pakistan        70.000000 26.000000  4.000000
##   Russia          70.149254 22.388060  7.462687
##   South Africa    62.500000 37.500000  0.000000
```

```
##   South America   79.310345  8.620690 12.068966
##   Southeast Asia  90.000000 10.000000  0.000000
##   Southern Europe 55.200000 41.600000  3.200000
```

```
library("graphics")
mosaicplot(prop_gen, shade = TRUE, las=2,
           main = "Proportion des génotypes par zone géographique")
```



Au vu du mosaicplot on a envie de conclure que la proportion des génotypes CC est toujours supérieure par rapport à la proportion des génotypes CT qui ont une proportion supérieure aux génotypes TT.

Faisons un test pour vérifier la réponse précédente

```
chisq.test(prop_gen)
```

```
## Warning in chisq.test(prop_gen): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  prop_gen
## X-squared = 180.57, df = 26, p-value < 2.2e-16
```

Au vu de la p-valeur, on peut conclure que les proportions des génotypes par zones géographiques sont statistiquement liées.

Calculons la mesure D' de déséquilibre de liaison(LD) pour toute paire de SNPs dans le gène AKT1

```
snp_names <- names(hgdp)[substr(names(hgdp),start=1,stop=5)=='AKT1.']
snps <- hgdp[snp_names]
for (i in 1:length(snp_names)) snps[,i]<-genotype(snps[,i],sep='')
LD(snps)$"D'"
```

```
##             AKT1.C0756A AKT1.C6024T AKT1.G2347T AKT1.G2375A
## AKT1.C0756A          NA   0.9934369   0.9481195   0.9843420
## AKT1.C6024T          NA          NA   0.9429031   0.9842787
## AKT1.G2347T          NA          NA          NA   0.9995238
## AKT1.G2375A          NA          NA          NA          NA
```

Au vu des résultats, on constate que la mesure D' de déséquilibre de liaison entre les paires de SNPs dans le gène AKT1 est toujours très proche de 1, on conclut qu'il y a un fort déséquilibre de liaison entre eux.

3e partie: Etude Fuctional Single Nucleotide Polymorphisms Associated with Human Muscle Size and Strength (FAMuSS)

Exercice 4 (HWE)

On s'interese à la variable NDRM.CH, le changement en pourcentage de la force de la force du bras non dominant avant et après le programme d'entrainnement physique prévu dans l'étude. On se demande si NDRM.CH est associée à un ou plusieurs SNPs.

1a Chargement des données

```
fms <- read.delim("C:/Users/james/OneDrive/Desktop/DocParisDescartes/S2/StatGen2/FMS_data.txt", header=
```

1b On exclut le seul individu d'origine amérindienne car cette observation crée des problèmes quand on essaie d'automatiser l'analyse sur les strates.

```
table(fms$Race)
```

```
##
## African Am  Am Indian      Asian  Caucasian   Hispanic      Other
##         44          1         97        791         52         49
```

```
which(fms$Race=="Am Indian")
```

```
## [1] 1107
```

```r
fms <- fms[-1107,]
```

```r
table(fms$Race)
```

```
##
## African Am  Am Indian      Asian  Caucasian   Hispanic      Other
##         44          0         97        791         52         49
```

```r
geno2 <- genotype(fms$akt1_t10726c_t12868c,sep='')
summary(fms$akt1_t10726c_t12868c)
```

Testons si le SNP akt1_t10726c_t12868c est en HWE dans l'ensemble de la population

```
##   CC   TC   TT NA's
##  881  301   62  152
```

```r
n=sum(table(geno2))
GenoCount <- table(geno2)
(GenoFreq <- GenoCount/n)
```

```
## geno2
##        C/C        C/T        T/T
## 0.70819936 0.24196141 0.04983923
```

```r
FreqC <- setNames(GenoFreq[1]+0.5*GenoFreq[2],c("C")); FreqC
```

```
##         C
## 0.8291801
```

```r
FreqT <- setNames(GenoFreq[3]+0.5*GenoFreq[2],c("T")); FreqT
```

```
##         T
## 0.1708199
```

Sous l'hypothèse Hardy-Weinberg

```r
FreqCC = setNames(FreqC^2,'CC'); FreqCC
```

```
##        CC
## 0.6875396
```

```r
FreqTC = setNames(2*FreqT*FreqC,'TC'); FreqTC
```

```
##       TC
## 0.283281
```

```r
FreqTT = setNames(FreqT^2,'TT'); FreqTT
```

```
##         TT
## 0.02917945
```

```r
(ExpCount <- c(FreqCC,FreqTC,FreqTT)*n)
```

```
##        CC        TC        TT
## 855.29924 352.40153  36.29924
```

La statistique de test est:

```r
ChisqStat <- sum((GenoCount-ExpCount)^2/ExpCount); ChisqStat
```

```
## [1] 26.46652
```

```r
pchisq(ChisqStat,df=1,lower.tail = F)
```

```
## [1] 2.681452e-07
```

Donc on rejette H0, il n'est pas en HWE dans l'ensemble de la population.

3 Testons si le SNP akt1_t10726c_t12868c est en HWE dans chaque strate de la variable Race}

```r
A <- levels(fms$Race)
HWEGeoArea <- tapply(geno2,INDEX=fms$Race,HWE.chisq)
for (i in A) { print(HWEGeoArea[i])
  }
```

```
## $`African Am`
##
##  Pearson's Chi-squared test with simulated p-value (based on 10000
##  replicates)
##
## data:  tab
## X-squared = 2.0144, df = NA, p-value = 0.2226
##
##
## $`Am Indian`
## NULL
##
## $Asian
##
##  Pearson's Chi-squared test with simulated p-value (based on 10000
##  replicates)
##
```

```
## data:  tab
## X-squared = 2.3611, df = NA, p-value = 0.1795
##
##
## $Caucasian
##
##  Pearson's Chi-squared test with simulated p-value (based on 10000
##  replicates)
##
## data:  tab
## X-squared = 0.0030898, df = NA, p-value = 1
##
##
## $Hispanic
##
##  Pearson's Chi-squared test with simulated p-value (based on 10000
##  replicates)
##
## data:  tab
## X-squared = 0.70512, df = NA, p-value = 0.5725
##
##
## $Other
##
##  Pearson's Chi-squared test with simulated p-value (based on 10000
##  replicates)
##
## data:  tab
## X-squared = 0.60706, df = NA, p-value = 0.6621
```

On conclut que le SNP akt1_t10726c_t12868c est en HWE dans chaque strate de la variable Race.

Exercice 5 (Association, tests multiples)

On s'interese à la variable NDRM.CH, le changement en pourcentage de la force de la force du bras non dominant avant et après le programme d'entrainnement physique prévu dans l'étude. On se demande si NDRM.CH est associée à un ou plusieurs SNPs.

Contruisons la variable aléatoire Y qui vaut 1 si NDRM.CH>60 et 0 autrement

```
Y <- as.numeric(fms$NDRM.CH>60)
```

Testons l'association entre tous les SNPs et Y

```
W=c("A","C","G","T")
A <- (expand.grid(x=W,y=W, stringsAsFactors=T))
(W=paste(A$x,A$y, sep=""))
```

Pour cela, on va récupérer les SNPs dans le jeu de données.

```
##  [1] "AA" "CA" "GA" "TA" "AC" "CC" "GC" "TC" "AG" "CG" "GG" "TG" "AT" "CT" "GT"
## [16] "TT"
```

```
vect <- c()
for (i in 1:ncol(fms)) {
R <-lapply(list(W),match, fms[,i])[[1]]
if (sum(R,na.rm = TRUE)==0){vect[i]=FALSE}
else{vect[i]= TRUE}
}
```

```
ind=which(vect==TRUE)
length(ind)
```

Nombre de SNPs

```
## [1] 222
```

```
names(fms[,ind])[1:10]
```

```
##  [1] "acdc_rs1501299"  "actn3_r577x"     "actn3_rs540874"  "actn3_rs1815739"
##  [5] "actn3_1671064"   "ardb1_1801253"   "adrb2_1042713"   "adrb2_1042714"
##  [9] "adrb2_rs1042718" "adrb3_4994"
```

```
snps <- fms[,vect]
dim(snps)
```

```
## [1] 1396  222
```

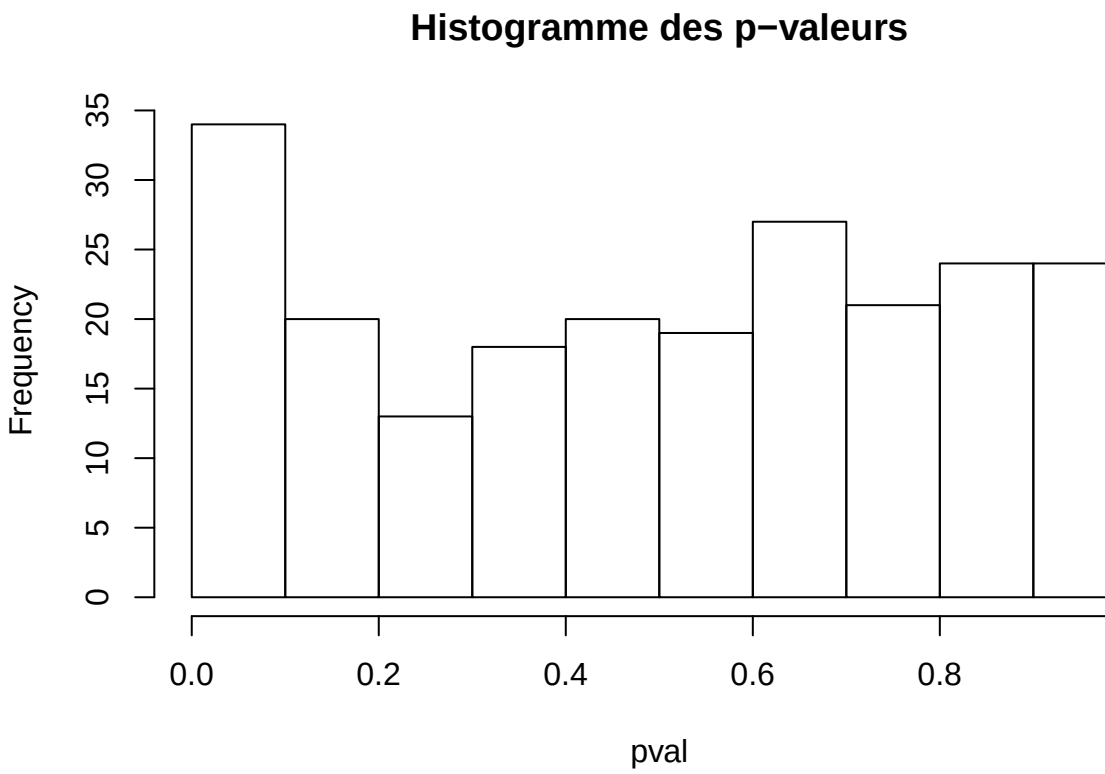
```
length(Y)
```

```
## [1] 1396
```

```
which(names(fms)=='NDRM.CH')
```

```
## [1] 236
```

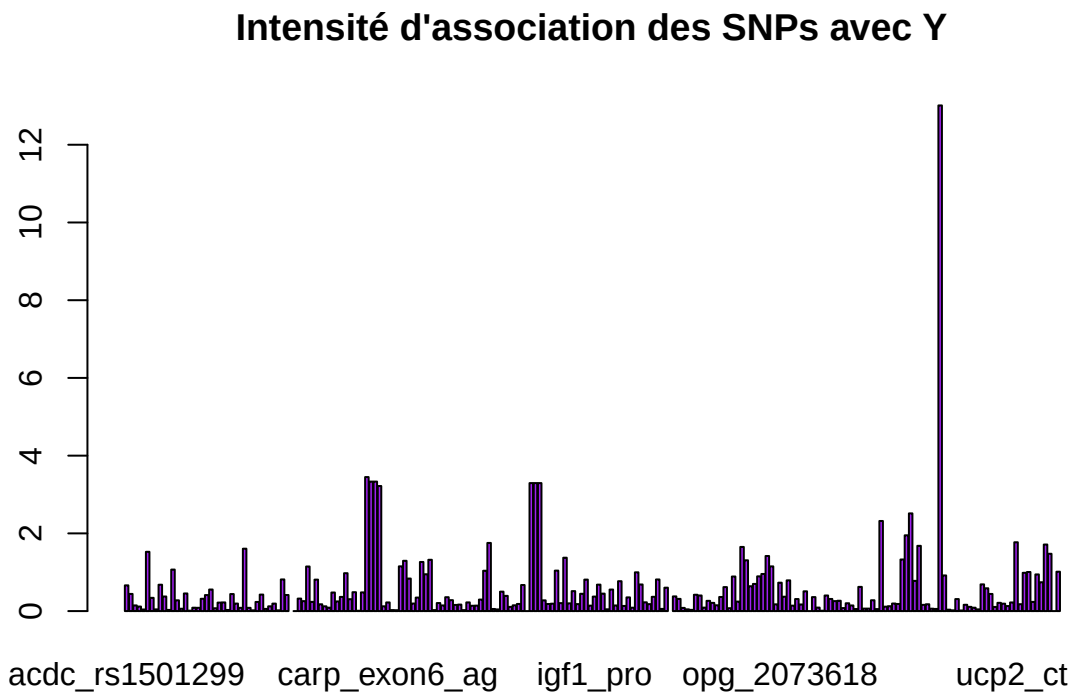
```
pval = suppressWarnings(apply(snps, 2, function(x) chisq.test(table(x,Y))$p.value))
hist(pval, main="Histogramme des p-valeurs")
```



On calcule les p-valeurs

Affichons le manhattan plot montrant l'intensité d'association de chaque SNP avec Y

```
barplot(-log10(pval),col="purple", main="Intensité d'association des SNPs avec Y")
```




```
datasnp=data.frame(SNP=colnames(snps),CHR=rep(1,ncol(snps)),BP=1:ncol(snps),P=pval,row.names = 1:ncol(sn
```

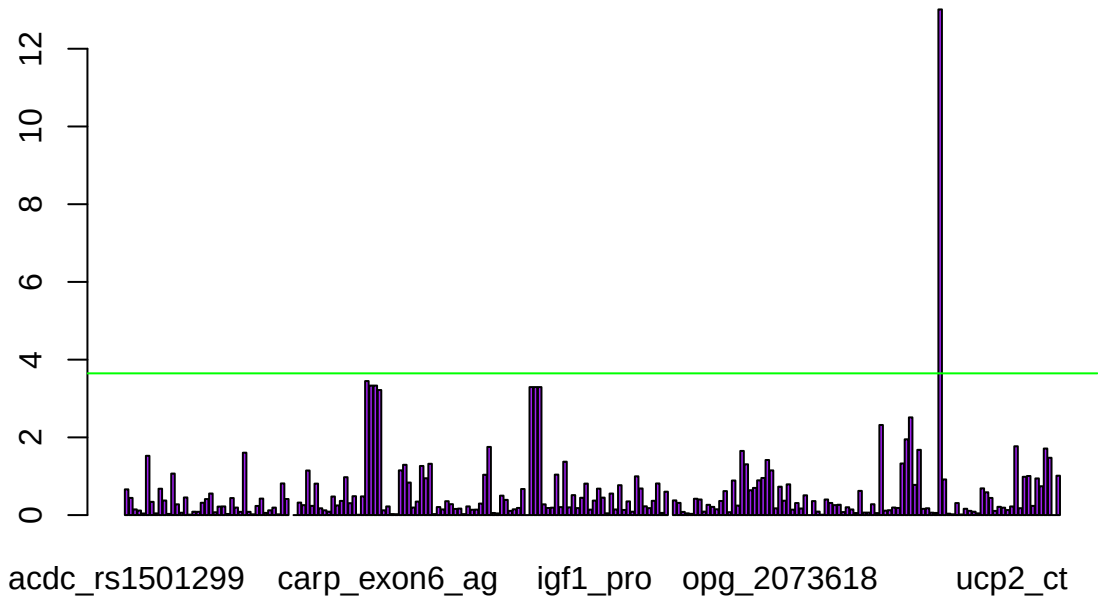
Pour que la probabilité de faire au moins un faux positif sur la totalité des tests soit inférieure à 0.05 il faut faire chaque test au seuil de:

```
alpha=0.05
(seuil=alpha/length(snps))
```

```
## [1] 0.0002252252
```

Affichons une barre horizontale à ce niveau sur le Manhattan plot

```
barplot(-log10(pval), col="purple",main="Intensité d'association des SNPs avec Y")
abline(h =-log10(seuil), col = "green")
```



Correction avec tests multiples

```
Bonfpv = p.adjust(pval,method="bonferroni")
sum(Bonfpv<=0.05,na.rm = T)
```

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

Après correction pour les tests multiples,le seul SNP qui semble avoir une association significative avec Y est

```
which(Bonfpv<=0.05)
```

```
## rs849409
##      194
```

Faisons un test pour voir si on peut conclure sur l'association entre Y et ce SNP

```
chisq.test(table(fms$rs849409,Y))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(fms$rs849409, Y)
## X-squared = 55.42, df = 1, p-value = 9.735e-14
```

Au vu de la p-valeur, on peut conclure qu'il y a une association significative entre ce SNP et Y.

4e Partie: Autres données

```
exo5 <- read.table("C:/Users/james/OneDrive/Desktop/DocParisDescartes/S2/StatGen2/exo5.txt", quote="\""
```

Chargement du jeu de données

Description

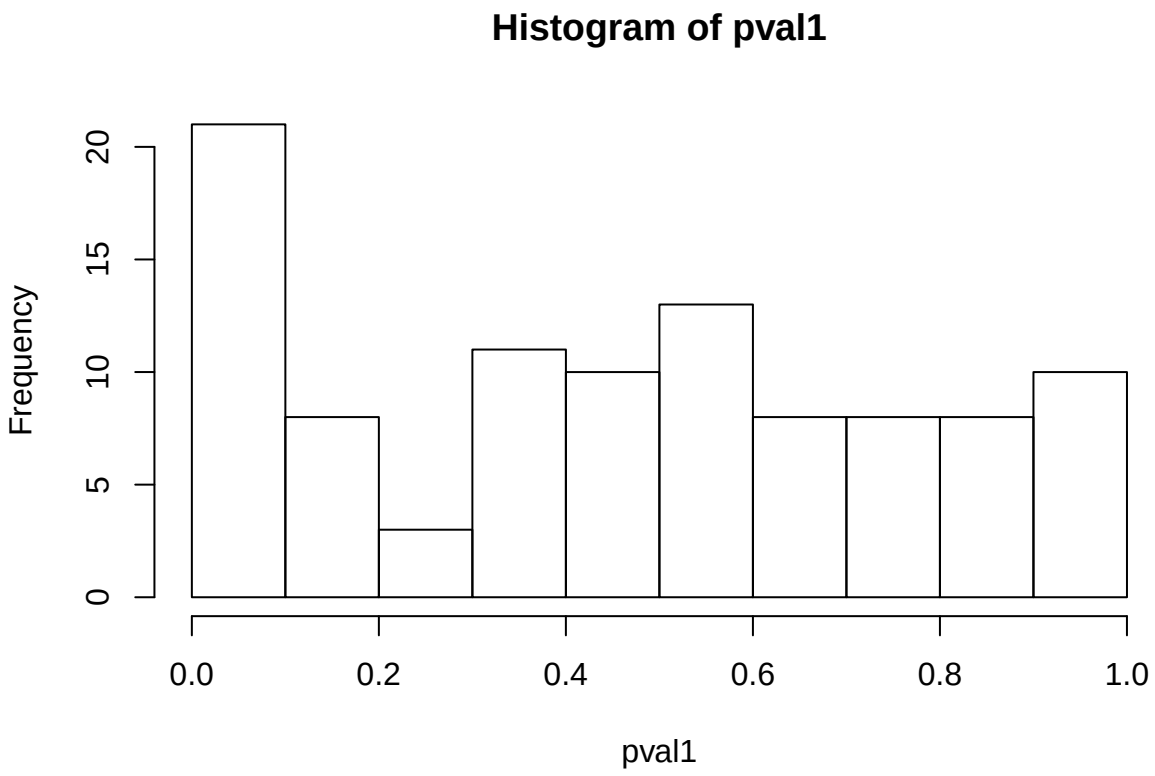
M: Phénotype, 1 pour les cas et 0 pour les témoins

E: Covariable d'exposition environnementale, 1 pour les exposés, 0 pour les non exposés.

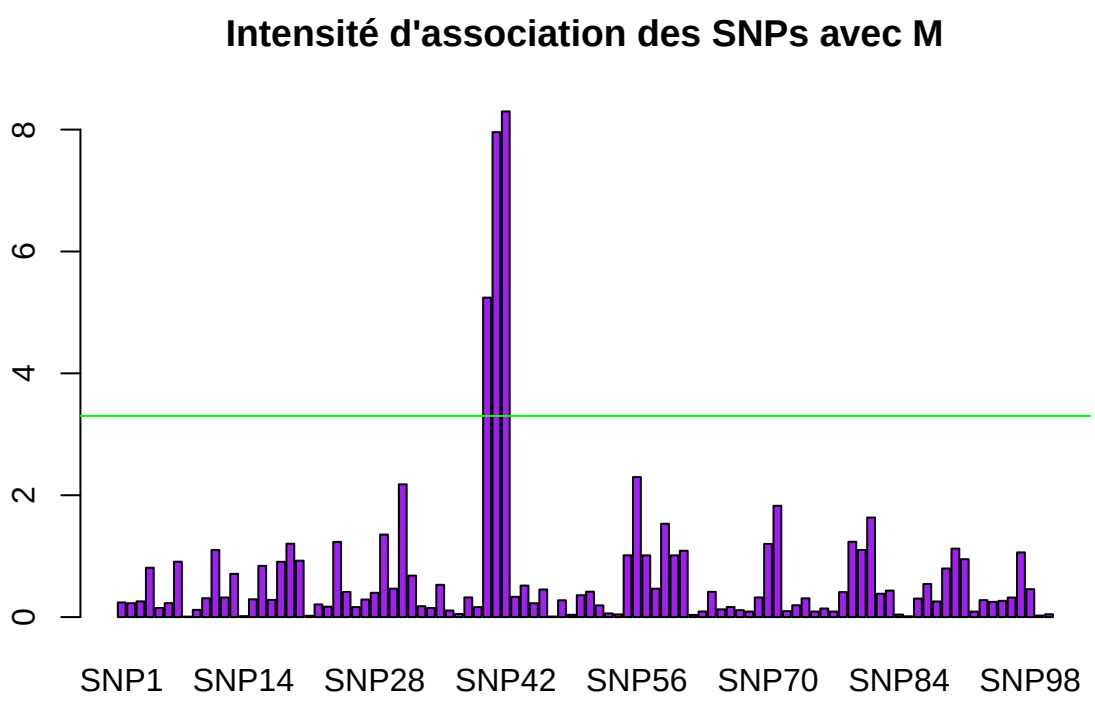
SNP1-SNP100: génotype 0,1,2 pour 100 SNPs biallélіques, la valeur du génotype indique le nombre d'allèles rares.

Testons l'association entre tous les SNPs et M

```
pval1 = suppressWarnings(apply(exo5[,-c(1,2)], 2, function(x) chisq.test(table(x,exo5$M))$p.value))
hist(pval1)
```




```
seuil2=alpha/(ncol(exo5)-2)
barplot(-log10(pval1),col="purple",main="Intensité d'association des SNPs avec M")
abline(h = -log10(seuil2), col = "green")
```




```
sum(pval1<=0.05)
```

```
## [1] 9
```

Correction avec tests multiples

```
Bonfpv1 = p.adjust(pval1,method="bonferroni")
sum(Bonfpv1<=0.05,na.rm = T)
```

```
## [1] 3
```

Les SNPs associés à la maladie sont:

```
which(Bonfpv1<=0.05)
```

```
## SNP40 SNP41 SNP42
##    40    41    42
```

3. Considérons le SNP 42, celui pour lequel le signal d'association est plus forte (c'est à dire pour lequel le test d'association a donné la plus petite p-valeur). Écrivons l'équation du modèle de régression logistique de M sur X

```
model<-glm(exo5$M~exo5$SNP42, family=binomial(link='logit'),na.action=na.exclude); summary(model)
```

```
##
## Call:
## glm(formula = exo5$M ~ exo5$SNP42, family = binomial(link = "logit"),
##     na.action = na.exclude)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.506  -1.045  -1.045   1.089   1.316
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.32044    0.07918  -4.047 5.19e-05 ***
## exo5$SNP42   0.53265    0.08981   5.931 3.01e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1651.1  on 1190  degrees of freedom
## Residual deviance: 1614.6  on 1189  degrees of freedom
## AIC: 1618.6
##
## Number of Fisher Scoring iterations: 4
```

Les coefficients sont significatifs, de plus, $\hat{\beta}_1 > 0$, une augmentation du nombre d'allèles rare de 1 augmente les chances pour que le phénotype M soit égal à 1 de $\exp(0.53265)$

```
new_model <- lm(exo5$M ~ exo5$SNP42 + exo5$E, na.action=na.exclude); summary(new_model)
```

4. On suspecte que l'effet de X sur M dépend de l'exposition E. Proposons un modèle pour vérifier cette hypothèse.

```
##
## Call:
## lm(formula = exo5$M ~ exo5$SNP42 + exo5$E, na.action = na.exclude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7716 -0.5061 -0.3949  0.4724  0.6051
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  0.39491    0.02079  18.996  < 2e-16 ***
## exo5$SNP42   0.13273    0.02127   6.239 6.11e-10 ***
## exo5$E       0.11124    0.03427   3.246   0.0012 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4908 on 1188 degrees of freedom
## Multiple R-squared:  0.03878,    Adjusted R-squared:  0.03716
## F-statistic: 23.96 on 2 and 1188 DF,  p-value: 6.264e-11
```

Au vu des résultats,la variable E est significative, on conclut que l'effet de X sur M dépend de E.