# Introducción a la Genómica
# UNAL nov 2017

Alejandro Cáceres
ISGlobal, Barcelona

November 9, 2017

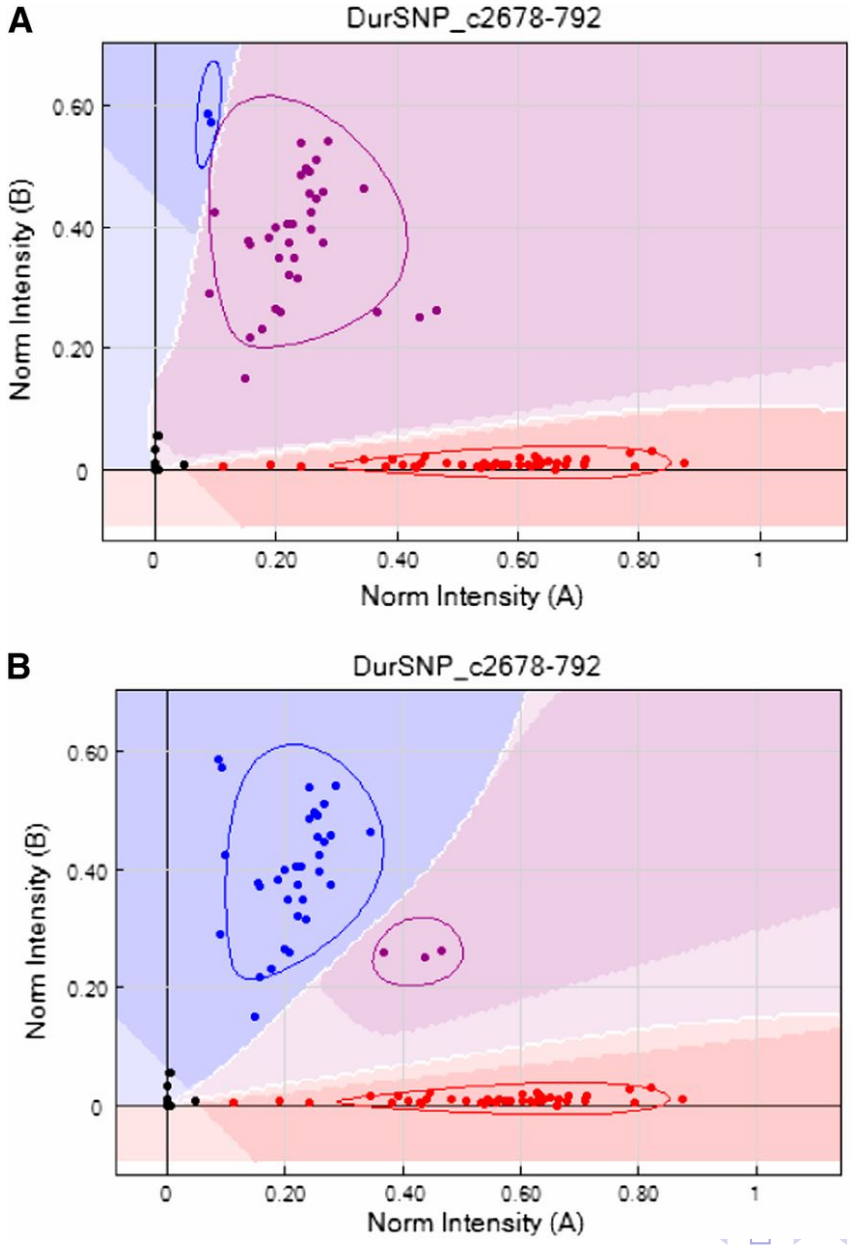
# Control de Calidad

Los datos de microarrays son sometidos a control de calidad

- `SNPs`: Se revisa la calidad del genotipado y que los genotipos biológimcamente correspondan a lo esperado
- `Sujetos`: Se identifican sujetos que son "outliers" de la muestra por diferencias en ancestría o parentesco cercano (si es muestra poblacional)

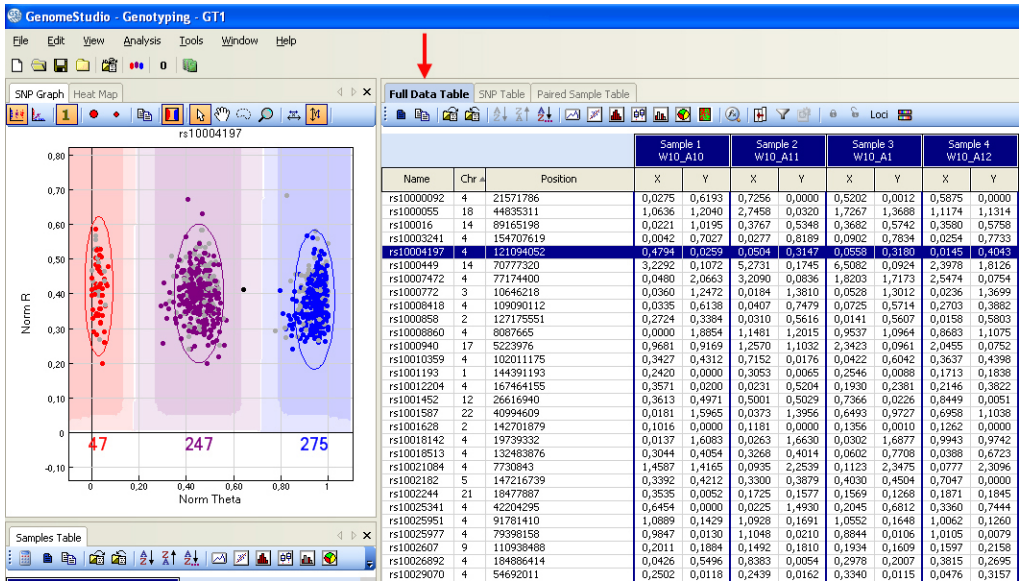
# Genotipado

En illumina los datos crudos son las intensidades de cada alelo



# GenomeStudio

Es el software de Illumina para hacer el genotipado de los SNPs (clustering)




| Name | Chr | Position | Sample 1 W10_A10 X | Sample 1 W10_A10 Y | Sample 2 W10_A11 X | Sample 2 W10_A11 Y | Sample 3 W10_A1 X | Sample 3 W10_A1 Y | Sample 4 W10_A12 X | Sample 4 W10_A12 Y |
|---|---|---|---|---|---|---|---|---|---|---|
| rs10000092 | 4 | 21571786 | 0,0275 | 0,6193 | 0,7256 | 0,0000 | 0,5202 | 0,0012 | 0,5875 | 0,0000 |
| rs1000055 | 18 | 44835311 | 1,0636 | 1,2040 | 2,7458 | 0,0320 | 1,7267 | 1,3688 | 1,1174 | 1,1314 |
| rs100016 | 14 | 89165198 | 0,0221 | 1,0195 | 0,3767 | 0,5348 | 0,3682 | 0,5742 | 0,3580 | 0,5758 |
| rs10003241 | 4 | 154707619 | 0,0042 | 0,7027 | 0,0277 | 0,8189 | 0,0902 | 0,7834 | 0,0254 | 0,7733 |
| rs10004197 | 4 | 121094052 | 0,4794 | 0,0259 | 0,0504 | 0,3147 | 0,0558 | 0,3180 | 0,0145 | 0,4043 |
| rs1000449 | 14 | 70777320 | 3,2292 | 0,1072 | 5,2731 | 0,1745 | 6,5082 | 0,0924 | 2,3978 | 1,8126 |
| rs10007472 | 4 | 77174400 | 0,0480 | 2,0663 | 3,2090 | 0,0836 | 1,8203 | 1,7173 | 2,5474 | 0,0754 |
| rs1000772 | 3 | 10646218 | 0,0360 | 1,2472 | 0,0184 | 1,3810 | 0,0528 | 1,3012 | 0,0236 | 1,3699 |
| rs10008418 | 4 | 109090112 | 0,0335 | 0,6138 | 0,0407 | 0,7479 | 0,0725 | 0,5714 | 0,2703 | 0,3882 |
| rs1000858 | 2 | 127175551 | 0,2724 | 0,3384 | 0,0310 | 0,5616 | 0,0141 | 0,5607 | 0,0158 | 0,5803 |
| rs10008860 | 4 | 8087665 | 0,0000 | 1,8854 | 1,1481 | 1,2015 | 0,9537 | 1,0964 | 0,8683 | 1,1075 |
| rs1000940 | 17 | 5223976 | 0,9681 | 0,9169 | 1,2570 | 1,1032 | 2,3423 | 0,0961 | 2,0455 | 0,0752 |
| rs10010359 | 4 | 102011175 | 0,3427 | 0,4312 | 0,7152 | 0,0176 | 0,0422 | 0,6042 | 0,3637 | 0,4398 |
| rs1001193 | 1 | 144391193 | 0,2420 | 0,0000 | 0,3053 | 0,0065 | 0,2546 | 0,0088 | 0,1713 | 0,1838 |
| rs10012204 | 4 | 167464155 | 0,3571 | 0,0200 | 0,0231 | 0,5204 | 0,1930 | 0,2381 | 0,2146 | 0,3822 |
| rs1001452 | 12 | 26616940 | 0,3613 | 0,4971 | 0,5001 | 0,5029 | 0,7366 | 0,0226 | 0,8449 | 0,0051 |
| rs1001587 | 22 | 40994609 | 0,0181 | 1,5965 | 0,0373 | 1,3956 | 0,6493 | 0,9727 | 0,6958 | 1,1038 |
| rs1001628 | 2 | 142701879 | 0,1016 | 0,0000 | 0,1181 | 0,0000 | 0,1356 | 0,0010 | 0,1262 | 0,0000 |
| rs10018142 | 4 | 19739332 | 0,0137 | 1,6083 | 0,0263 | 1,6630 | 0,0302 | 1,6877 | 0,9943 | 0,9742 |
| rs10018513 | 4 | 132483876 | 0,3044 | 0,4054 | 0,3268 | 0,4014 | 0,0602 | 0,7708 | 0,0388 | 0,6723 |
| rs10021084 | 4 | 7730843 | 1,4587 | 1,4165 | 0,0935 | 2,2539 | 0,1123 | 2,3475 | 0,0777 | 2,3096 |
| rs1002182 | 5 | 147216739 | 0,3392 | 0,4212 | 0,3300 | 0,3879 | 0,4030 | 0,4504 | 0,7047 | 0,0000 |
| rs1002244 | 21 | 18477887 | 0,3535 | 0,0052 | 0,1725 | 0,1577 | 0,1569 | 0,1268 | 0,1871 | 0,1845 |
| rs10025341 | 4 | 42204295 | 0,6454 | 0,0000 | 0,0225 | 1,4930 | 0,2045 | 0,6812 | 0,3360 | 0,7444 |
| rs10025951 | 4 | 91781410 | 1,0889 | 0,1429 | 1,0928 | 0,1691 | 1,0552 | 0,1648 | 1,0062 | 0,1260 |
| rs10025977 | 4 | 79398158 | 0,9847 | 0,0130 | 1,1048 | 0,0210 | 0,8844 | 0,0106 | 1,0105 | 0,0079 |
| rs1002607 | 9 | 110938488 | 0,2011 | 0,1884 | 0,1492 | 0,1810 | 0,1934 | 0,1609 | 0,1597 | 0,2158 |
| rs10026892 | 4 | 184886414 | 0,0426 | 0,5496 | 0,8383 | 0,0054 | 0,2978 | 0,2007 | 0,3815 | 0,2695 |
| rs10029070 | 4 | 54692011 | 0,2502 | 0,0118 | 0,2439 | 0,0162 | 0,3340 | 0,0115 | 0,0476 | 0,3157 |

- el control de calidad en el genotipado está dado por el clustering
- reporta un callrate y un valor del número de sujetos con genotipado aceptable por SNP
- genera los genotipos en formato PLINK

# Control de Calidad por Sujetos

- Se revisa el sexo es el reportado por los individuos y se detecta aneploidías como XXY, XYY, etc
- cosanginidad mas alta de la esperado por el diseno del estudio

# Identidad por descendencia

La consanginidad (identity by descent) se puede calcular en PLINK por medio de la opción –genome

```
acaceres@IMW00680:~$ plink --bfile [filename prefix] --genome
```

y produce un reporte en el fichero `plink.genome`

# Control de calidad de SNPs

- Call-rate (dada por genomeStudio): $> 80\%$
- Minor allele frequency (MAF): $> 0.01$ o $> 0.05$
- Hardy Weinberg Equillibrium: menos de 3 o 4 desviaciones standard
- Mendelian errors: si hay trios que no tengan errores de trasmisión

# Equilibrio de Hardy Weinberg

En una población en donde no hay fuerzas que cambien la frecuencia alelica $p$ (de un SNP) debemos encontar

- $p^2$ fracción de homicigotos
- $2 * p * (p - 1)$ fracción de heretocigotos
- $(1 - p)^2$ fracción de homocigotos variantes

que se producen por la asociación aleatoria entre los cromosomas paternos y maternos en la población

# Equilibrio de Hardy Weinberg

SNPs que no están en WHE pueden presentar

- Problemas de genotipación. Por ejemplo una sonda defectuosa en la mitad de la población: se espera una alta desviación y es una situacin probable.
- Fuerzas evolutivas sobre el SNP: se espera una desviación mas moderada y es una situación menos probable.

Si la desviación es biológica como en el segundo caso, SNPs vecinos que están en LD también deben estar desviados.

# Errores Mendelianos

Para un estudio con trios los SNPs de

- dos padres homocigotos para el mismo alelo no pueden tener un hijo que no sea homocigoto
- un padre homocigoto y una madre heterocigota no pueden tener un hijo que no sea homocigoto variante
- un padre homocigoto para un alelo y una madre homocigota para el otro alelo no pueden tener un hijo homocigoto para ningún alelo

desviaciones en estas reglas son herrores de trasmisión mendelianos e indican error en genotipación

# Software

- el control de calidad para MAF, HWE y call-rate se puede hacer con la mayoría de paquertes
- veamos como se have con SNPstats

En R cargemos la librería y los datos

```
library("snpStats")

## Loading required package:  survival
## Loading required package:  Matrix

load("datos/snpsSNPstats.RData")
snpsSNPstats

## A SnpMatrix with  2504 rows and  1863 columns
## Row names:  HG00096 ... NA21144
## Col names:  rs555347111 ... rs558158882
```

# SNPStats

`col.summary` calcula call-rate, MAF, HWE para todos los SNPs en la base de datos

```
sum <- col.summary(snpsSNPstats)
dim(sum)

## [1] 1863    9

head(sum)

##             Calls Call.rate Certain.calls         RAF         MAF
## rs555347111  2504         1             1 0.073682109 0.073682109
## rs573543994  2504         1             1 0.012180511 0.012180511
## rs542617372  2504         1             1 0.076078275 0.076078275
## rs562398147  2504         1             1 0.011781150 0.011781150
## rs576107214  2504         1             1 0.021365815 0.021365815
## rs188856175  2504         1             1 0.008785942 0.008785942
##                   P.AA        P.AB        P.BB      z.HWE
## rs555347111 0.8554313 0.141773163 0.002795527   1.930779
## rs573543994 0.9796326 0.016373802 0.003993610 -15.991828
## rs542617372 0.8582268 0.131389776 0.010383387  -3.271542
## rs562398147 0.9844249 0.007587859 0.007987220 -33.733301
## rs576107214 0.9740415 0.009185304 0.016773163 -39.048892
```

# SNPStats

obtengamos los SNPs con call-rate $> 0.8$

```
Callrate <- sum$Call.rate
selectCallRate <- Callrate > 0.8
length(selectCallRate)

## [1] 1863

head(selectCallRate)

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

# SNPStats

obtengamos los SNPs con frequencia mayor a 0.01

```
MAF <- sum$MAF
selectMAF <- MAF > 0.01
length(selectMAF)

## [1] 1863

head(selectMAF)

## [1]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
```

# SNPStats

## Cuales SNPs tienen $MAF > 0.01$ y $CallRate > 0.80$

```
selectMAFCAllrete <- selectMAF & selectCallRate
head(selectMAFCAllrete)

## [1]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE

table(selectMAFCAllrete)

## selectMAFCAllrete
## FALSE  TRUE
##   420  1443
```

```
snpnames <- colnames(snpsSNPstats)
length(snpnames)

## [1] 1863

head(snpnames)

## [1] "rs555347111" "rs573543994" "rs542617372" "rs562398147" "rs57610
## [6] "rs188856175"
```

# SNPStats

teniendo los SNPs se puede seleccionar una submatriz de los genotipos con sólo estos SNPs

```
NewsnpsSNPstats<-snpsSNPstats[,selsnpnames]
NewsnpsSNPstats

## A SnpMatrix with  2504 rows and  1443 columns
## Row names:  HG00096 ... NA21144
## Col names:  rs555347111 ... rs558158882
```

# Ejercicio

- Seleccionar ahora los SNPs que tienen $abs(sum\$z.HWE) < 6$
- Crear una nueva matriz con los SNPs seleccionados

desviaciones en estas reglas son herrores de trasmisión mendelianos e indican error en genotipación

# Ejercicio-solución

```
selhw<-abs(sum$z.HWE) < 6
NewsnpsSNPstats<-snpsSNPstats[, selhw & selectMAF & selectCallRate]
save(NewsnpsSNPstats, file="NewsnpsSNPstats.RData")
```