

Introducción a la Genómica

UNAL nov 2017

Alejandro Caceres
ISGlobal, Barcelona

October 13, 2017

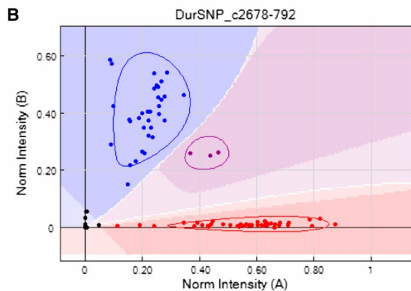
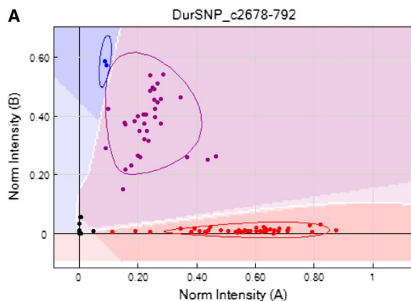
Control de Calidad

Los datos de microarrays son sometidos un control de calidad

- ▶ SNPs: Calidad del genotipado o que biológicamente no correspondan a lo esperado
- ▶ Sujetos: Sujetos que son "outliers de la muestra: ancestría o parentesco

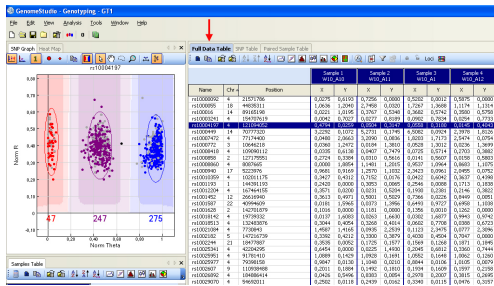
Genotipado

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)



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

Control de Calidad por Sujetos

- ▶ que el sexo es el reportado o detección de aneuploidas como XXY, XYY, etc
- ▶ cosanginidad mas alta de la esperado en el diseo del estudio

Identidad por descendencia

La consanguinidad (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 Equilibrium: menos de 3 o 4 desviaciones standard
- ▶ Mendelian errors: si hay trios que no tengan errores de transmisión

Equilibrio de Hardy Weinberg

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

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

que se deducen por la asociación aleatoria entre los cromosomas paternos y maternos

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 (Alta desviación-Probable).
- ▶ Fuerzas evolutivas sobre el SNP (Baja desviación-Menos probable)).

Si la desviación es biológica como en el segundo caso, SNPs cercanos 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 alternativo
- ▶ 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 errores de transmisión mendelianos e indican error en genotipación

Software

- ▶ el control de calidad para MAF, HWE y call-rate se puede hacer con la mayoría de paquetes
- ▶ veamos como se hace 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(snpSNPstats)
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

obtenemos 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

obtenemos los SNPs con frecuencia 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
```

```
snppnames <- colnames(snppsNPstats)
```

```
length(snppnames)
```

```
## [1] 1863
```

```
head(snppnames)
```

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

```
## [6] "rs188856175"
```

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

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

```
## 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 errores de transmisión mendelianos e indican error en genotipación

ejercicio

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