

Class 11: Genome Informatics

```
knitr::opts_chunk$set(echo = TRUE)
```

Section 2 Proportion of G/G in a population

Downloaded CSV file from Ensemble <

Here we read this CSV file

```
mxl <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378 (1).csv")
head(mxl)
```

	Sample..	Male..	Female..	Unknown..	Genotype..	forward..	strand..	Population..	s..	Father
1				NA19648	(F)		A A	ALL, AMR, MXL		-
2				NA19649	(M)		G G	ALL, AMR, MXL		-
3				NA19651	(F)		A A	ALL, AMR, MXL		-
4				NA19652	(M)		G G	ALL, AMR, MXL		-
5				NA19654	(F)		G G	ALL, AMR, MXL		-
6				NA19655	(M)		A G	ALL, AMR, MXL		-
	Mother									
1		-								
2		-								
3		-								
4		-								
5		-								
6		-								

```
table(mx1$Genotype..forward.strand.)
```

A A	A G	G A	G G
22	21	12	9

```
table(mx1$Genotype..forward.strand.) / nrow(mx1) * 100
```

A A	A G	G A	G G
34.3750	32.8125	18.7500	14.0625

Section 4: Population Scale Analysis

One sample is obviously not enough to know what is happening in a population. You are interested in assessing genetic differences on a population scale. So, you processed about ~230 samples and did the normalization on a genome level. Now, you want to find whether there is any association of the 4 asthma-associated SNPs (rs8067378...) on ORMDL3 expression.

```
expr <- read.table("https://bioboot.github.io/bimm143_S23/class-material/rs8067378_ENSG000
head(expr)
```

	sample	geno	exp
1	HG00367	A/G	28.96038
2	NA20768	A/G	20.24449
3	HG00361	A/A	31.32628
4	HG00135	A/A	34.11169
5	NA18870	G/G	18.25141
6	NA11993	A/A	32.89721

```
nrow(expr)
```

```
[1] 462
```

Q13: Read this file into R and determine the sample size for each genotype and their corresponding median expression levels for each of these genotypes.

```
table(expr$geno)
```

A/A A/G G/G
108 233 121

```
expr$geno == "A/A"
```

```
[1] FALSE FALSE TRUE TRUE FALSE TRUE FALSE TRUE FALSE FALSE FALSE FALSE
[13] TRUE FALSE TRUE TRUE FALSE TRUE FALSE FALSE FALSE FALSE FALSE TRUE
[25] FALSE FALSE TRUE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE
[37] TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE
[49] FALSE FALSE FALSE FALSE TRUE FALSE TRUE FALSE FALSE FALSE FALSE FALSE
[61] FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE
[73] FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE
[85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE
[97] TRUE FALSE TRUE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE
[109] FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[121] TRUE TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE
[133] TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE TRUE FALSE
[157] FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE
[169] TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE TRUE FALSE TRUE TRUE
[181] FALSE TRUE FALSE FALSE FALSE FALSE TRUE TRUE FALSE FALSE FALSE FALSE
[193] FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE TRUE FALSE FALSE
[205] TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE TRUE FALSE
[217] TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE TRUE TRUE
[229] TRUE TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE TRUE
[241] FALSE TRUE TRUE FALSE TRUE TRUE FALSE TRUE FALSE FALSE FALSE FALSE
[253] FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE TRUE TRUE
[265] TRUE TRUE FALSE FALSE TRUE TRUE FALSE FALSE FALSE FALSE TRUE FALSE
[277] FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[289] FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE
[301] FALSE FALSE FALSE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE
[313] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE
[325] FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE
[337] TRUE FALSE FALSE FALSE TRUE FALSE TRUE FALSE FALSE TRUE FALSE TRUE
[349] FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE TRUE TRUE
[361] FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE TRUE TRUE FALSE
[373] FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE TRUE
[385] FALSE TRUE FALSE FALSE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE
[397] FALSE FALSE FALSE TRUE TRUE FALSE FALSE FALSE TRUE FALSE FALSE TRUE
[409] TRUE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE
```

```
[421] FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
[433]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE
[445] FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE
[457] FALSE FALSE  TRUE FALSE  TRUE FALSE
```

```
median(expr$exp)
```

```
[1] 25.11561
```

```
library(ggplot2)
```

Let's make a boxplot

```
ggplot(expr) +  
  aes(x=geno, y=exp, fill=geno) +  
  geom_boxplot(notch = TRUE)
```

