Workshop on GWAS

Exploratory data analysis (EDA) and the concept of regression

Prof. Lei Sun

Department of Statistical Sciences, FAS

Division of Biostatistics, DLSPH

University of Toronto

21 June, 2021

At the end of this lecture

- Get to know GWAS-workshop-sample-dataset-ERAP2.txt
- Types of data and descriptive statistics
- Exploratory data analysis (EDA)
- Towards statistical inference (and learning), regression

The (small but real) dataset used throughout the lectures

extracted from a cleaned genotype dataset of the 1000 Genomes Project data, and the gene expression of *ERAP2* of the International Genome Sample Resource

```
FID     IID     PID     MID     SEX     PHENO   POP     SNP1-5618704    SNP1-57815437   SNP1-64302980   SNP1-104336159  SNP1-151435036  SNP1-158018135
SNP1-173714419  rs2782524       SNP2-23882292   SNP2-60375263   rs10192914      SNP2-112710623  rs999891        SNP2-211265378  SNP3-19002158   SNP3-
61162380        SNP3-64305918   SNP3-71108737   SNP3-125827295  SNP3-127964403  SNP3-157693188  SNP3-178358711  SNP3-184470209  rs1518872       SNP4-
48537429        SNP4-67185059   SNP4-76791598   SNP4-121158599  SNP4-129152543  rs10007083      SNP4-162915830  SNP4-167449697  SNP4-173723483  SNP5-
33837406        rs10054860      rs1056893       rs4360063       rs10044354      SNP5-106184146  rs40588 SNP6-24967500    SNP6-37032449   rs2746304
SNP6-69359962   SNP6-94678620   SNP6-108159912  rs1006932       SNP7-6655522    SNP7-8339546    SNP7-52236127   rs10282724      SNP7-126907925  SNP8-
97403295        SNP9-12237310   SNP9-27159704   rs7864801       SNP9-109943261  SNP10-15114789  rs7912144       SNP10-28792948  rs7913102
rs9415825       SNP10-80281274  rs2020163       SNP11-89516883  SNP11-95415516  SNP12-23576923  SNP12-47113506  rs10220224      rs9582475
rs7339421       SNP13-104488021 rs4772700       SNP14-27762092  rs2737721       SNP14-41252775  rs1813500       SNP14-50513718  SNP14-89810169
rs11845053      SNP15-91733944  SNP16-8669923   SNP16-20312407  SNP16-57425543  SNP16-77578033  SNP17-60824552  SNP17-67165654  rs733383
rs11870893      SNP18-549352    SNP18-46491800  SNP18-66700491  SNP19-2446724   SNP19-6387304   rs427366        SNP20-4494424   rs227134
rs4134385       SNP22-47353519  SNP23-99734210
1328    NA06984 0       0       1       11.4981 CEU     AA      AA      CC      CC      CC      TT      TT      CC      TT      CC      GG      CC
AA      TT      TT      TC      TT      TT      GG      AA      TT      CT      AA      CC      CC      AA      CC      CC      AA      GG      TT
CC      AA      CC      TT      CT      GA      TC      TA      GG      CC      AA      TT      TT      CC      CG      AG      GG      TT      TT
AA      TT      CC      AA      AA      GG      AA      TT      CA      CC      CC      TT      AG      GG      AA      TT      AA      GG      CC
GA      TT      GG      GG      TT      TT      AG      AA      CC      CC      GG      AA      CC      GG      TT      TT      TT      AG      GG
CC      TT      GG      AA      GG      TC      CC      GG      AA      GA      AA      GG
1328    NA06989 0       0       2       10.6796 CEU     AA      AA      CC      CC      CC      TT      CT      AC      TT      CC      GG      CC
GA      TT      TT      TC      TT      TT      GG      GA      TT      TT      GG      TC      CT      GA      CC      CC      AA      GG      TT
CC      AT      CC      TT      CT      GA      TC      TA      GA      CC      AA      TC      TT      TC      GG      GG      GG      TT      CC
AA      TT      CC      CC      AA      AG      GA      TT      AA      CC      CC      TT      AA      GG      AA      TT      AA      GG      CC
GA      TT      GG      GG      TT      CT      GG      AA      CC      CC      GG      AA      CC      GG      TT      TT      TT      GG      GG
CC      TT      GG      AA      AG      TC      TC      GG      GG      AA      AA      AG
1330    NA12340 0       0       1       10.5453 CEU     AA      AA      TT      CC      CC      TT      TT      AC      TT      CC      GG      CC
AA      TT      CT      TT      TT      TT      GG      AA      TT      CT      AA      TT      CT      GA      CC      CC      AA      AA      TT
CC      AT      CC      GT      CT      GA      TC      TA      GA      CC      AA      TC      TT      CC      CG      AG      GG      TT      TT
AA      TT      CC      CC      AA      GG      AA      TT      AA      AC      CC      TT      GG      GG      AA      CT      AA      GG      CT
GA      CT      GG      AG      CC      TT      AA      AA      CC      CC      GG      AA      TC      GG      TT      TT      TT      GG      TG
CC      TT      GG      AA      GG      TC      TC      GG      GA      AA      GA      GG
1330    NA12341 0       0       2       11.4555 CEU     AA      AA      TC      CC      CC      TT      CT      CC      TT      CC      GG      CC
AA      CT      TT      TT      TT      TT      AG      GA      TT      TT      GG      CC      CT      AA      CC      CC      AA      AA      TT
CC      AT      CC      GG      CC      GA      TC      TA      GA      TC      AA      TC      TT      TC      CG      GG      GG      TT      CT
AA      TT      CC      CA      AA      GG      AA      TT      AA      CC      CC      TT      AG      GG      AA      CT      GA      TG      TT
GA      TT      GG      AG      CT      CT      GG      AA      CC      CC      GG      AA      CC      GG      TT      TT      TT      GG      TG
CC      TT      GG      AA      GG      TC      CC      GG      AA      AA      AA      GG
1330    NA12342 0       0       1       7.98708 CEU     AA      AA      CC      CC      CC      TT      TT      CC      TT      TC      GG      CC
AA      CT      TT      TC      TT      TT      AA      GA      TT      CT      AG      CC      TT      AA      CC      AC      AA      GG      TT
CC      AT      CC      GG      TT      AA      CC      TA      AA      CC      AA      TC      TT      TC      CG      AG      GG      TT      CT
AA      TT      CC      CA      AA      GG      AA      TT      CC      CC      CC      TT      GG      GG      AA      TT      GA      GG      CC
AA      TT      GG      AG      TT      TT      AA      AA      CC      CC      GG      AA      TC      GG      TT      TT      TT      GG      TG
TC      TT      GG      AA      GG      CC      TC      GG      GA      GA      AA      GG
1330    NA12343 0       0       2       11.2674 CEU     AA      AA      CC      CC      CC      TT      CT      CC      TT      CC      GG      CC
AA      CT      TT      TC      TT      TT      GG      AA      TT      CT      GG      CC      CC      GA      CC      CC      AA      GG      TT
CC      TT      CC      TT      CT      GA      TC      AA      GA      CC      AA      CC      TT      TC      CG      AG      GG      TT      CC
AA      TT      CC      AA      AA      AG      AA      TT      AA      CC      CC      TT      GG      GG      AA      TT      AA      GG      CC
AA      TT      GG      AG      TT      CT      GG      AA      CC      CC      GG      AA      TC      GG      TT      GT      TT      AG      GG
```

. . . . . .

Getting the data, GWAS-workshop-sample-dataset-ERAP2.txt

```
# mydata.ERAP2=data.table::fread("http://www.utstat.toronto.edu/sun/data/GWAS-workshop-sample-dataset-ERAP2
# Locally if you have already downloaded the data to your working directory
mydata.ERAP2=read.table("GWAS-workshop-sample-dataset-ERAP2.txt",header=T)

# a sneak peak at the data; male=1 and female=2 by convention
# CEU: Utah residents with Northern and Western European ancestry from the CEPH collection
# YRI: Yoruba in Ibadan, Nigeria
mydata.ERAP2[c(1:3,103:105,193:195),1:9]
```

```
##       FID     IID PID MID SEX    PHENO POP SNP1.5618704 SNP1.57815437
## 1    1328 NA06984   0   0   1 11.49810 CEU           AA            AA
## 2    1328 NA06989   0   0   2 10.67960 CEU           AA            AA
## 3    1330 NA12340   0   0   1 10.54530 CEU           AA            AA
## 103 13291 NA07435   0   0   1 11.55000 CEU           AA            AA
## 104 13292 NA07051   0   0   1 10.68810 CEU           AA            AA
## 105  Y001 NA18486   0   0   1 10.65360 YRI           CA            GA
## 193  Y116 NA19236   0   0   1 10.84090 YRI           CA            GG
## 194  Y120 NA19247   0   0   2  8.57924 YRI           AA            AA
## 195  Y120 NA19248   0   0   1 11.95420 YRI           CC            AA
```

```
# Our data matrix is 195 by 107 in dimension; the last 100 columns are genotypes of 100 SNPs
c(length(mydata.ERAP2[,1]),length(mydata.ERAP2[1,]))
```

```
## [1] 195 107
```

Types of data

Data can take many forms, such as free-form text, images, audio recordings, and networks of relationships.

***a conventional form of data**: we have a collection of cases (observations), and we measure multiple characteristics of each case. The characteristics may also be called attributes, or variables.*

*Such data may be said to have **a rectangular form**. By convention, the rows of the rectangle correspond to the cases [individuals] and the columns correspond to the variables.*

```
mydata.ERAP2[c(1:3,103:105,193:195),1:9]
```

```
##       FID     IID PID MID SEX    PHENO POP SNP1.5618704 SNP1.57815437
## 1    1328 NA06984   0   0   1 11.49810 CEU           AA            AA
## 2    1328 NA06989   0   0   2 10.67960 CEU           AA            AA
## 3    1330 NA12340   0   0   1 10.54530 CEU           AA            AA
## 103 13291 NA07435   0   0   1 11.55000 CEU           AA            AA
## 104 13292 NA07051   0   0   1 10.68810 CEU           AA            AA
## 105  Y001 NA18486   0   0   1 10.65360 YRI           CA            GA
## 193  Y116 NA19236   0   0   1 10.84090 YRI           CA            GG
## 194  Y120 NA19247   0   0   2  8.57924 YRI           AA            AA
## 195  Y120 NA19248   0   0   1 11.95420 YRI           CC            AA
```

Types of data cont'd

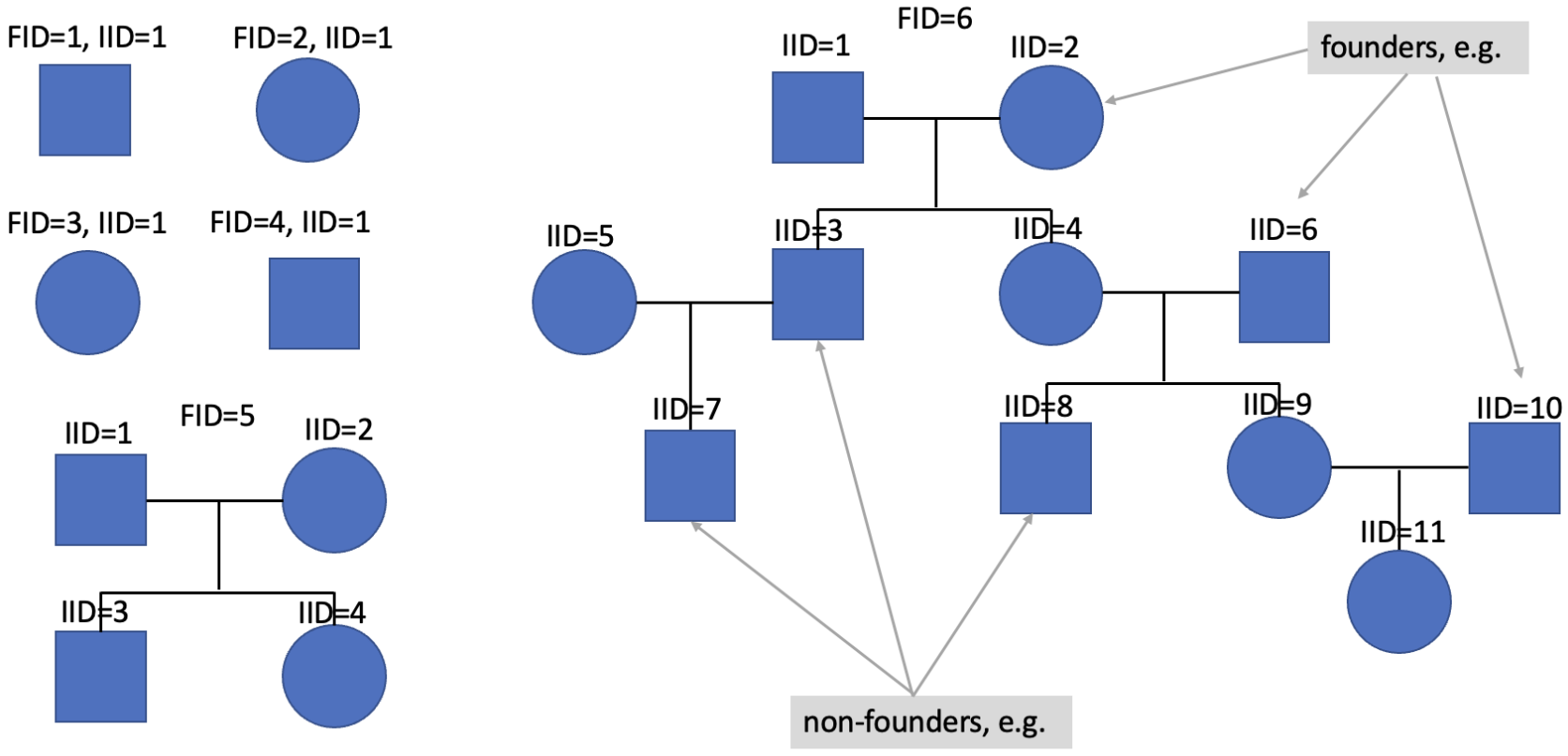
```
mydata.ERAP2[c(1:3,103:105,193:195),1:9]
```

```
##       FID     IID PID MID SEX    PHENO POP SNP1.5618704 SNP1.57815437
## 1    1328 NA06984   0   0   1 11.49810 CEU           AA            AA
## 2    1328 NA06989   0   0   2 10.67960 CEU           AA            AA
## 3    1330 NA12340   0   0   1 10.54530 CEU           AA            AA
## 103 13291 NA07435   0   0   1 11.55000 CEU           AA            AA
## 104 13292 NA07051   0   0   1 10.68810 CEU           AA            AA
## 105  Y001 NA18486   0   0   1 10.65360 YRI           CA            GA
## 193  Y116 NA19236   0   0   1 10.84090 YRI           CA            GG
## 194  Y120 NA19247   0   0   2  8.57924 YRI           AA            AA
## 195  Y120 NA19248   0   0   1 11.95420 YRI           CC            AA
```

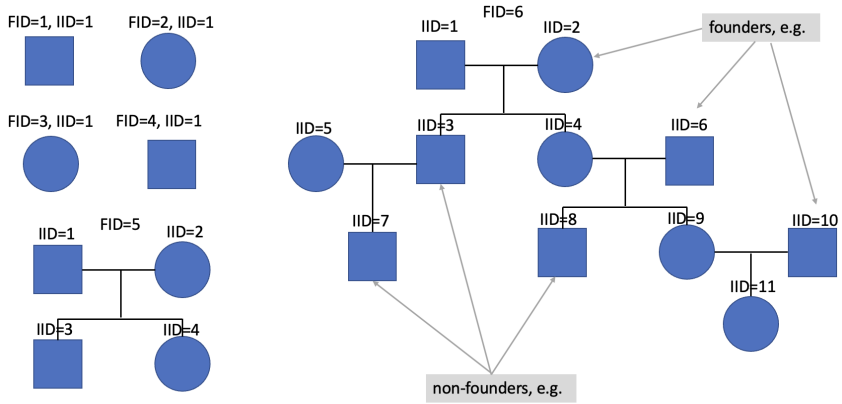
Can you identify

- *A **nominal variable** is an unordered set of labels.*
- *An **ordinal variable** is an ordered set of labels, with no implication that we can attribute meaning to the size of "gaps" between the labels.*
- *A **quantitative [discrete or continuous] variable** measures something in numeric terms to high precision, often capturing an amount or the change in an amount.*

Unrelated vs. related individuals, and pedigree drawing



Question: How to code the data/graphs?
(Much harder: how do you translate the codes to graphs?)



FID	IID	PID	MID	SEX
1	1	0	0	1
2	1	0	0	2
3	1	0	0	1
4	1	0	0	2
5	1	0	0	1
5	2	0	0	2
5	3	1	2	1
5	4	1	2	2
6	1	0	0	1
6	2	0	0	2
6	3	1	2	1
6	4	1	2	2
6	5	0	0	2
6	6	0	0	1
6	7	3	5	1
6	8	4	6	1
6	9	4	6	2
6	10	0	0	1
6	11	10	9	2

Some descriptive statistics

```
mydata.ERAP2[c(1,2,194,195),1:9]
```

```
##       FID     IID PID MID SEX    PHENO POP SNP1.5618704 SNP1.57815437
## 1    1328 NA06984   0   0   1 11.49810 CEU           AA            AA
## 2    1328 NA06989   0   0   2 10.67960 CEU           AA            AA
## 194 Y120 NA19247   0   0   2  8.57924 YRI           AA            AA
## 195 Y120 NA19248   0   0   1 11.95420 YRI           CC            AA
```

```
summary(mydata.ERAP2$POP) # number of individuals in each population
```

```
## CEU YRI
## 104  91
```

```
summary(mydata.ERAP2$SEX) # Why not summary(mydata.ERAP2$SEX)?
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.492   2.000   2.000
```

```
c(sum(mydata.ERAP2$SEX==1),sum(mydata.ERAP2$SEX==2)) # Instead; male=1 and female=2 by convention
```

```
## [1] 99 96
```

```
# Sex by Population; male=1 and female=2 by convention
matrix(c(sum(mydata.ERAP2$POP=='CEU'&mydata.ERAP2$SEX==1),sum(mydata.ERAP2$POP=='CEU'&mydata.ERAP2$SEX==2)
         sum(mydata.ERAP2$POP=='YRI'&mydata.ERAP2$SEX==1),sum(mydata.ERAP2$POP=='YRI'&mydata.ERAP2$SEX==2)
```

```
##      [,1] [,2]
## [1,]   53   51
## [2,]   46   45
```

Question: Are there any siblings or genetically related individuals in the data?

The nominal genetic data, genotype

```
mydata.ERAP2$SNP1.5618704
```

```
##   [1] AA AA AA AA AA AA AA AA AA AA AA AA AA AA AA AA AA AA CA AA AA AA AA AA AA AA
##  [26] CA AA AA AA AA AA AA CC AA AA AA AA AA AA AA CA AA AA AA AA AA AA AA AA AA
##  [51] AA CA AA AA AA AA AA AA AA AA AA AA AA AA AA AA AA AA AA AA CA AA AA AA AA CA
##  [76] AA AA AA AA AA AA AA AA AA AA AA AA AA AA AA AA AA AA AA AA AA AA CA AA AA AA
## [101] AA AA AA AA CA AA AA AA AA CA AA AA AA CC CA CA CA CA CA CA CA CA CA CA AA
## [126] AA CA CC CC CA CA AA AA CA CA AA AA AA CA AA AA AA CA CA AA AA CA AA CA CC
## [151] CC CC CC AA CA CC CA AA CC CA AA CA AA AA CA CA AA CA CA CC CA CA CA CA CA
## [176] AA AA AA AA AA AA AA CA CA CC AA CA CC CA AA CC AA CA AA CC
## Levels: AA CA CC
```

Transform the nomimal genotype data to quantitative discrete data using the additive genotype coding scheme

```
# Count the number of C's in each genotype
as.numeric(mydata.ERAP2$SNP1.5618704)-1
```

```
##   [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 2 0 0 0 0
##  [38] 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
##  [75] 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 0
## [112] 0 0 2 1 1 1 1 1 1 1 1 1 1 0 0 1 2 2 1 1 0 0 1 1 0 0 0 1 0 0 0 1 1 0 0 1 0
## [149] 1 2 2 2 2 0 1 2 1 0 2 1 0 1 0 0 1 1 0 1 1 2 1 1 1 1 1 0 0 0 0 0 0 0 1 1 2
## [186] 0 1 2 1 0 2 0 1 0 2
```

The quantitative (continous) phenotype/trait/outcome (Y) of interest

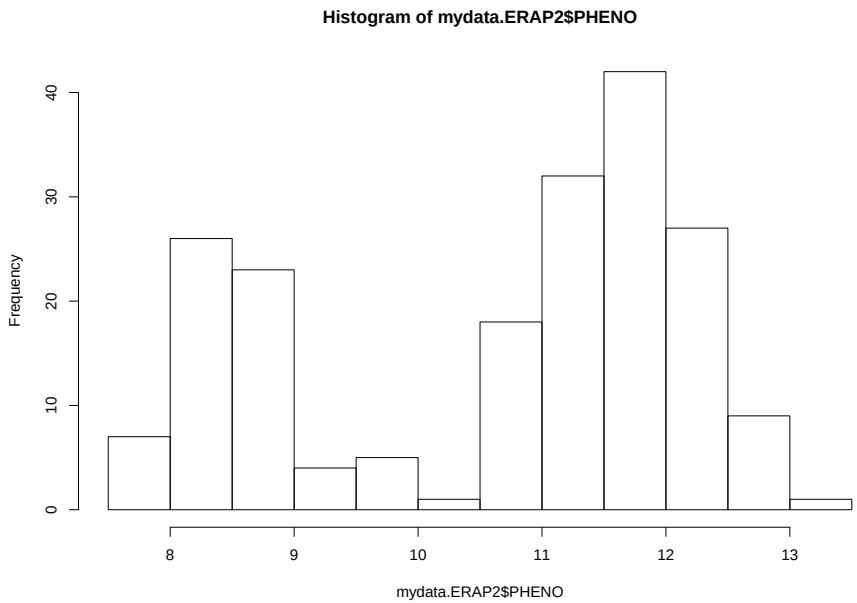
```
round(mydata.ERAP2$PHENO,1)[1:50]
```

```
##  [1] 11.5 10.7 10.5 11.5  8.0 11.3 12.1 11.8 12.7  9.6 11.9 12.8  9.0 10.8 11.8
## [16] 10.7 12.6  8.6 12.3 11.4 12.0 11.5 11.2 12.3  9.3  8.9 12.2  8.6  8.6 10.9
## [31]  8.4 10.6 10.6 12.2 11.0 12.3 11.4  8.7 11.8 11.9  8.3  8.9  8.0  9.5 11.9
## [46] 11.6 11.7 11.4 12.0 12.0
```

```
summary(mydata.ERAP2$PHENO)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.914   8.843  11.295  10.625  11.827  13.208
```

```
hist(mydata.ERAP2$PHENO)
```



Towards statistical inference (and learning)

Are the patterns/distributions of males and females (Y)? similar between the CEU and YRI populations (X)?

(Yes, what is Y and what is X depend your scientific question!)

```
# Sex by Population; male=1 and female=2 by convention
matrix(c(sum(mydata.ERAP2$POP=='CEU'&mydata.ERAP2$SEX==1),sum(mydata.ERAP2$POP=='CEU'&mydata.ERAP2$SEX==2)
         sum(mydata.ERAP2$POP=='YRI'&mydata.ERAP2$SEX==1),sum(mydata.ERAP2$POP=='YRI'&mydata.ERAP2$SEX==2)
```

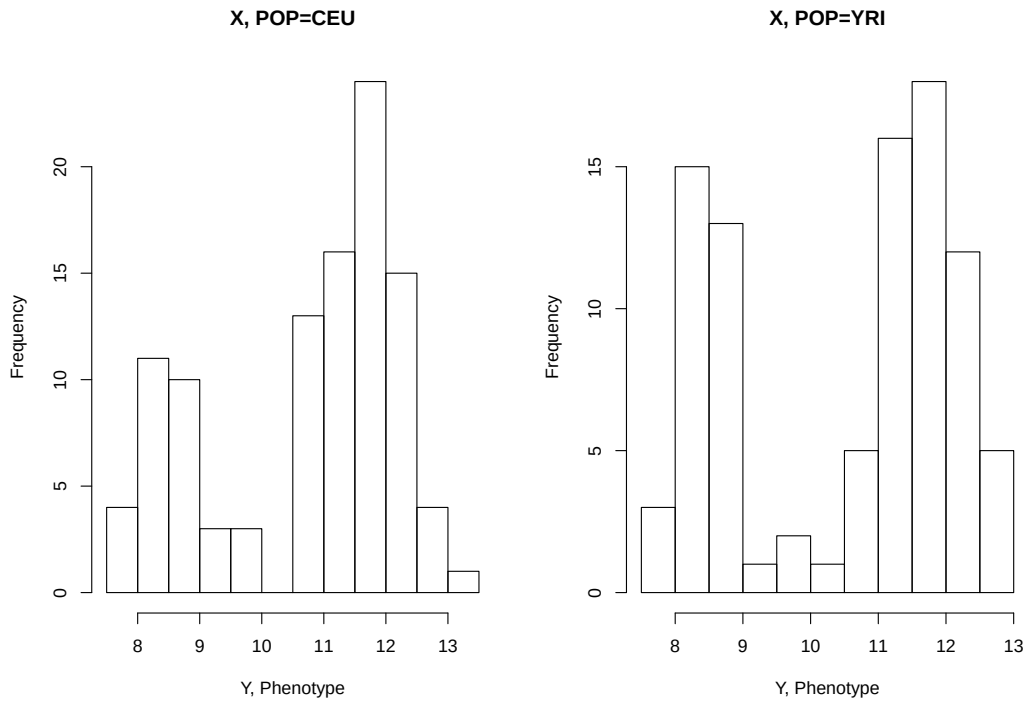
```
##      [,1] [,2]
## [1,]   53   51
## [2,]   46   45
```

- Proprotion of males in CEU: $\frac{53}{53+51} = 50.96\%$
- Proprotion of males in YRI: $\frac{46}{46+45} = 50.55\%$
- No need for statistical inference in this case.

How about this one?

Are the distributions of PHENO (Y) similar between CEU and YRI (X)?

```
par(mfrow=c(1,2))
hist(mydata.ERAP2$PHENO[mydata.ERAP2$POP=='CEU'],main="X, POP=CEU",xlab="Y, Phenotype")
hist(mydata.ERAP2$PHENO[mydata.ERAP2$POP=='YRI'],main="X, POP=YRI",xlab="Y, Phenotype")
```




```
summary(mydata.ERAP2$PHENO[mydata.ERAP2$POP=='CEU'])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.935   9.239  11.384  10.722  11.831  13.208
```

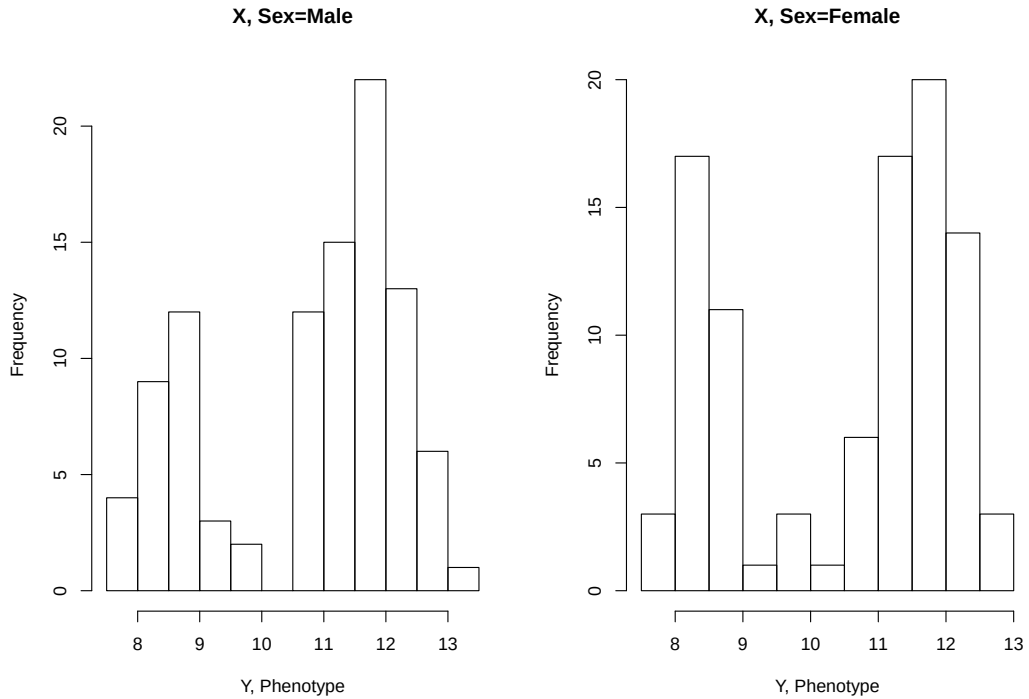
```
summary(mydata.ERAP2$PHENO[mydata.ERAP2$POP=='YRI'])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.914   8.728  11.256  10.515  11.827  12.650
```

Yet another one

PHENO (Y) similar between males and females (X)?

```
par(mfrow=c(1,2))
hist(mydata.ERAP2$PHENO[mydata.ERAP2$SEX==1],main="X, Sex=Male",xlab="Y, Phenotype")
hist(mydata.ERAP2$PHENO[mydata.ERAP2$SEX==2],main="X, Sex=Female",xlab="Y, Phenotype")
```




```
summary(mydata.ERAP2$PHENO[mydata.ERAP2$SEX==1]);
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.935   9.010  11.387  10.740  11.897  13.208
```

```
summary(mydata.ERAP2$PHENO[mydata.ERAP2$SEX==2])
```

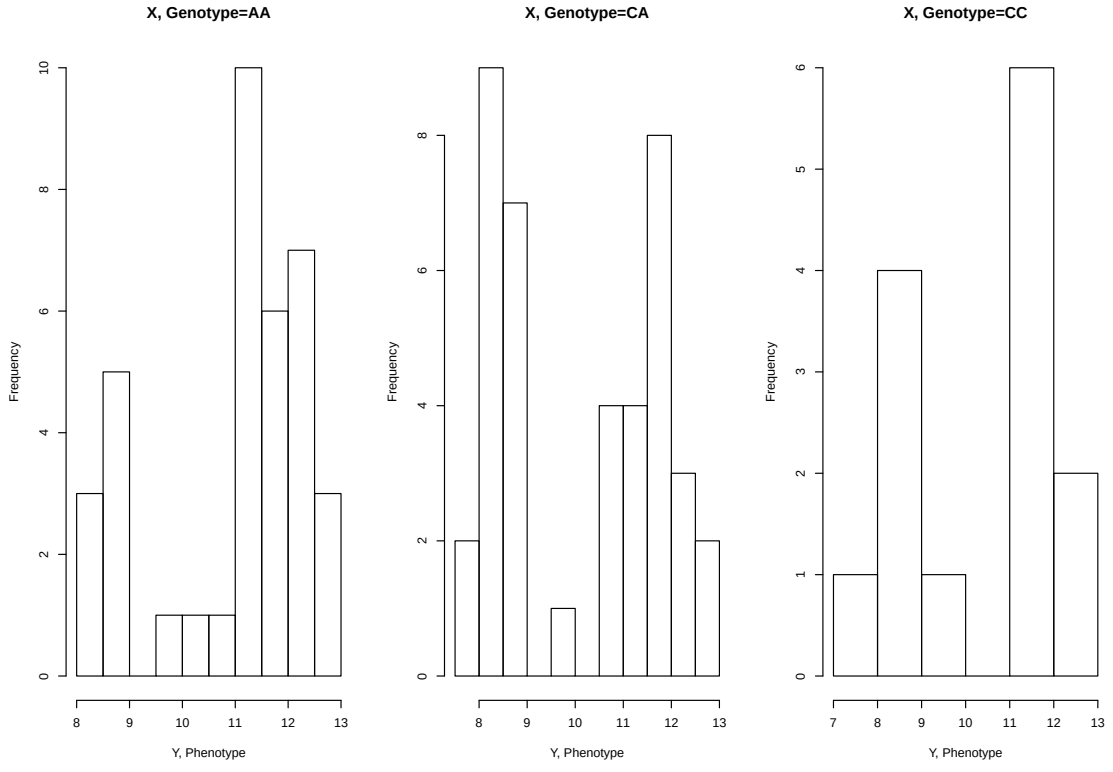
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.914   8.668  11.275  10.507  11.785  12.779
```

Finally, the phenotype-genotype association analysis

```
mydata.ERAP2$SNP1.5618704[mydata.ERAP2$POP=='YRI']
```

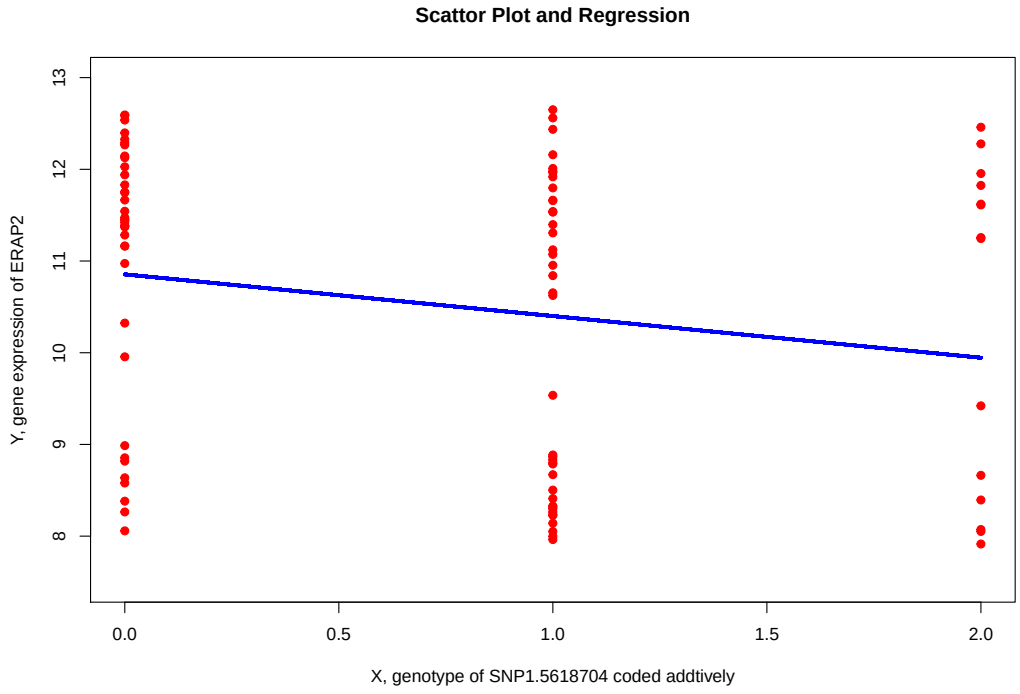
```
##  [1] CA AA AA AA AA CA AA AA AA CC CA CA CA CA CA CA CA CA CA CA AA AA CA CC CC
## [26] CA CA AA AA CA CA AA AA AA CA AA AA AA CA CA AA AA CA AA CA CC CC CC CC AA
## [51] CA CC CA AA CC CA AA CA AA AA CA CA AA CA CA CC CA CA CA CA CA AA AA AA AA
## [76] AA AA AA CA CA CC AA CA CC CA AA CC AA CA AA CC
## Levels: AA CA CC
```

```
par(mfrow=c(1,3))
hist(mydata.ERAP2$PHENO[mydata.ERAP2$SNP1.5618704=="AA"&mydata.ERAP2$POP=='YRI'],main="X, Genotype=AA",xla
hist(mydata.ERAP2$PHENO[mydata.ERAP2$SNP1.5618704=="CA"&mydata.ERAP2$POP=='YRI'],main="X, Genotype=CA",xla
hist(mydata.ERAP2$PHENO[mydata.ERAP2$SNP1.5618704=="CC"&mydata.ERAP2$POP=='YRI'],main="X, Genotype=CC",xla
```



Simple Linear Regression, Expected or average value of $Y = \beta_0 + \beta X$

```
sample.index=(mydata.ERAP2$POP=="YRI")
x=as.numeric(mydata.ERAP2[sample.index,]$SNP1.5618704)-1
y=mydata.ERAP2[sample.index,]$PHENO
plot(x,y,pch=19,col="red", ylim=c(7.5,13), main="Scattor Plot and Regression",
     xlab="X, genotype of SNP1.5618704 coded addtively",
     ylab="Y, gene expression of ERAP2")
lines(x,fitted(lm(y~x)),col="blue",lwd=3) # fitted regression line
```



An important scientific question: Does the genotype of SNP1.5618704 (X) influence the gene expression of *ERAP2* (Y)? Does Y depend on X?

To be answered statistically: Is the regression line flat, i.e. the slope $\beta = 0$?

Ordinary least squares

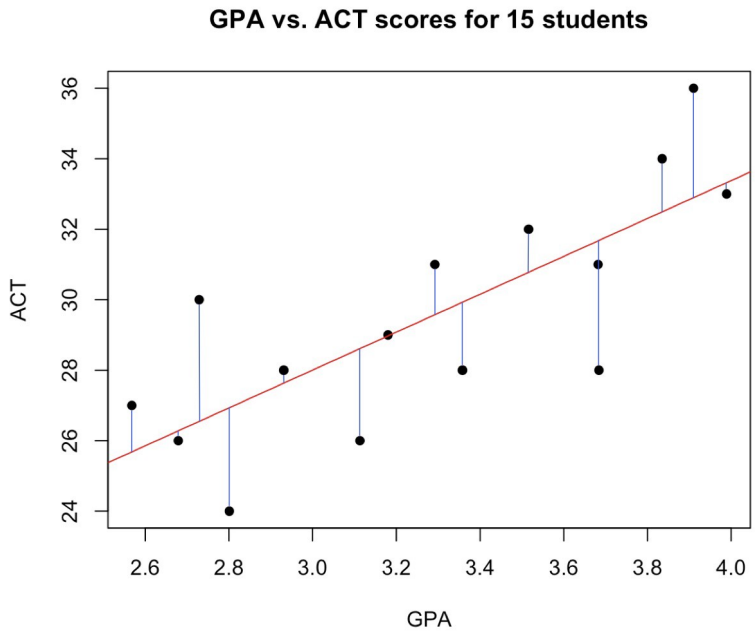
Use one X or a set of predictors, $X_1, \ldots, X_p$, to predict an outcome Y:

Expected value of Y, phenotype $= \beta_0 + \beta_1 X_1$(e.g. genotype of a SNP)$+\beta_2 X_2$(e.g. sex)

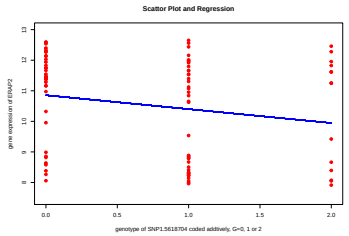
Find $\beta = (\beta_0, \beta_1, \ldots \beta_p)$ that **minimizes the sum of the squares of the differences between the observed y_i values and the predicted $\hat{y}_i$ values**, predicted based on the set of predictors, $x_1, \ldots, x_p$.

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2.$$

Visualize Least Squares (LS) estimation



Regression coefficient and p-value, a 'black-boxed' approach

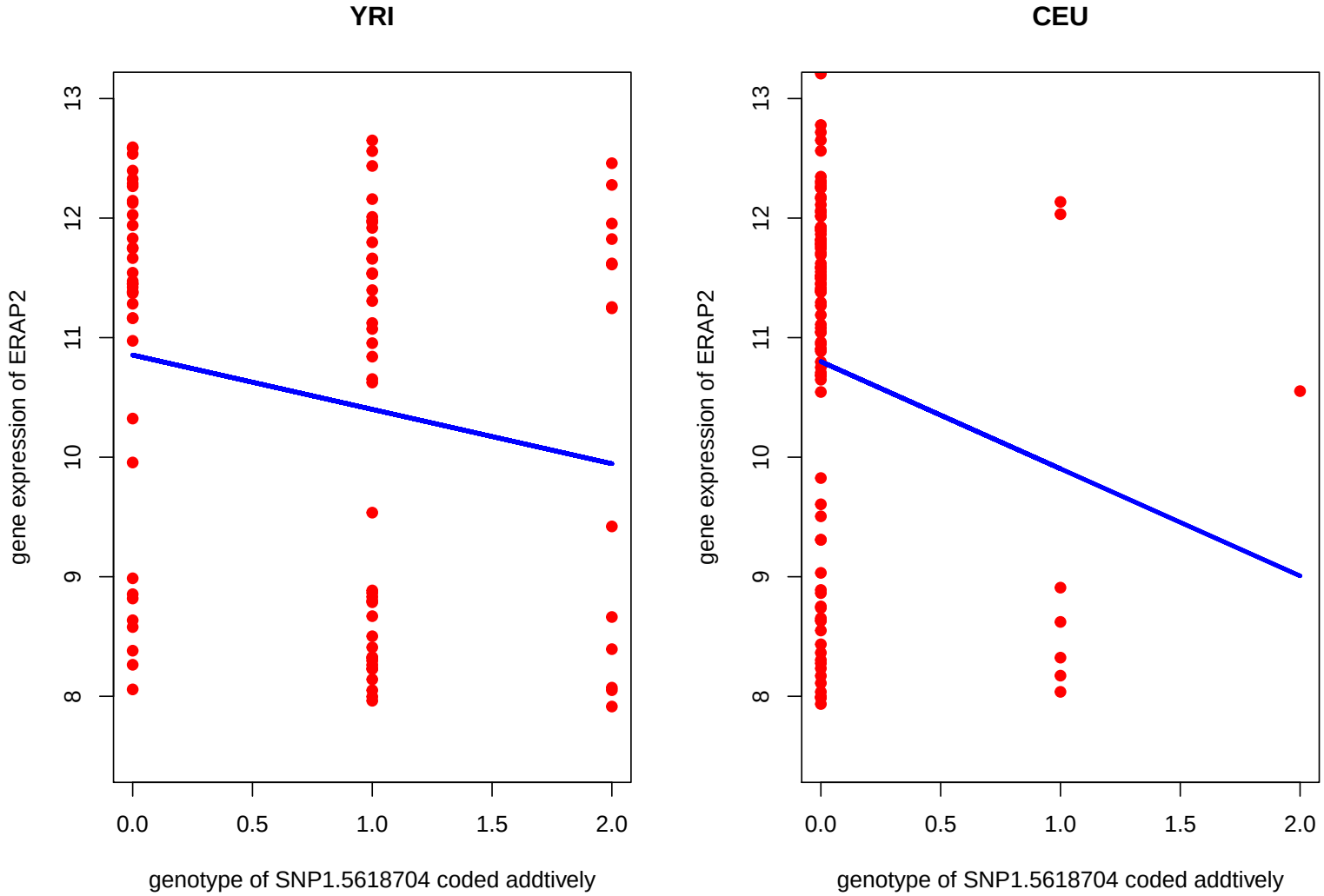


The slope (the regression coefficient) is -0.4545. The slope is not statistically different from zero: the **p-value of testing the slope = 0 is 0.0594, not statistically significant.**

```
summary(lm(y~x))
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7969 -1.7987  0.5538  1.3051  2.5135
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.8544     0.2445  44.402   <2e-16 ***
```

How about YRI (left) vs. CEU (right)?



Live Quiz 1: Which slope (statistically) departs more from zero?

A: YRI (left)

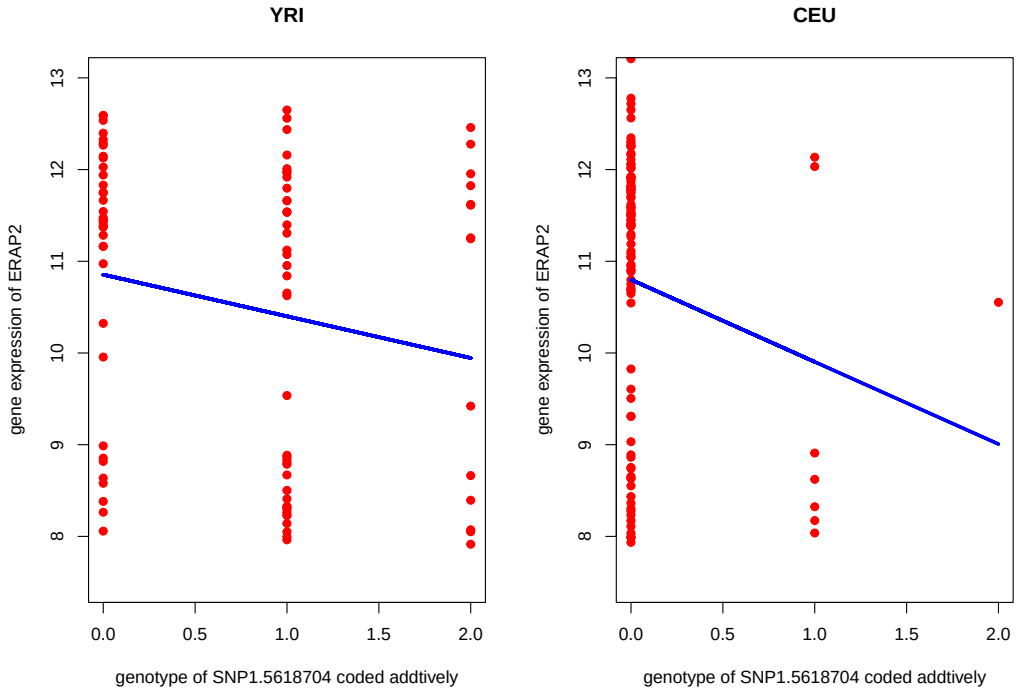
B: CEU (right)

C: Cannot determine; need more data

The 'solution'

```
## [1] "YRI"
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 10.8544007  0.2444552 44.402418 1.624690e-62
## x           -0.4544541  0.2380040 -1.909438 5.942701e-02
```

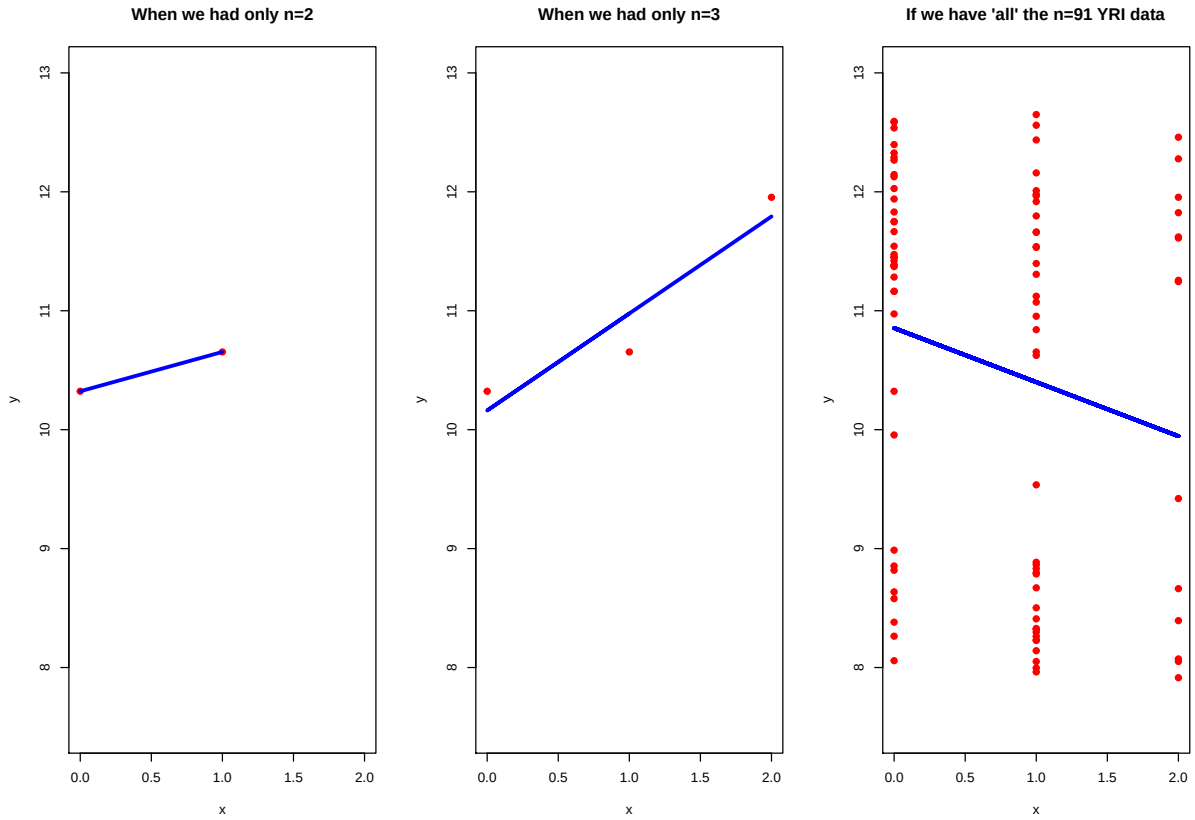



```
## [1] "CEU"
```

```
##                Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 10.7996640  0.1511875 71.43227 6.364624e-89
## x           -0.8963905  0.4648749 -1.92824 5.660573e-02
```

'Large' *n* is important!

Continue the study of association between X=SNP1.5618704 and Y=trait (gene expression of *ERAP2*) in the YRI sample:



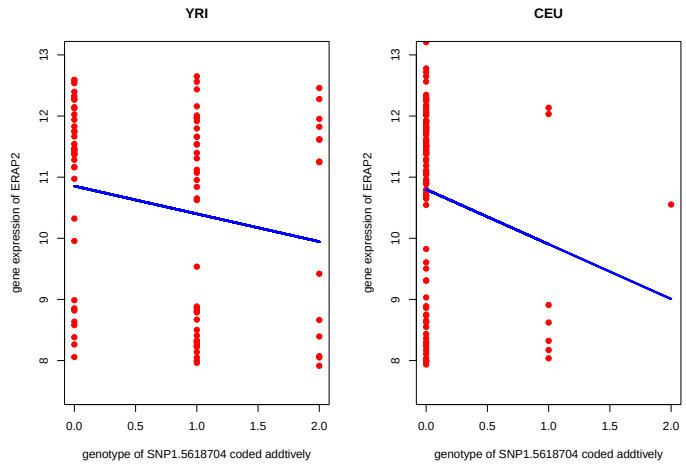
Looking ahead (i.e. if you take university level statistics courses): how to determine your *n* is large enough (i.e. sample size calculation)?

Why not use both YRI and CEU samples to increase n?

Recall the (potential) confounding issue!

```
## [1] "YRI"

##      [,1] [,2] [,3]
## [1,] "AA" "CT" "CC"
## [2,] "37" "40" "14"
```




```
## [1] "CEU"

##      [,1] [,2] [,3]
## [1,] "AA" "CT" "CC"
## [2,] "96" "7"  "1"
```

Other questions can be asked

```
## [1] "YRI"

##      [,1] [,2] [,3]
## [1,] "AA" "CT" "CC"
## [2,] "37" "40" "14"

## [1] "CEU"

##      [,1] [,2] [,3]
## [1,] "AA" "CT" "CC"
## [2,] "96" "7"  "1"
```

- Are the genotype patterns similar between the YRI and CEU samples?
- How do you estimate genotype frequencies of AA, CA and CC?
- How do you estimate allele frequencies of A and C?
- What is a good estimate? What is a estimate anyway?!
- Do genotype frequencies (or allele frequencies) differ between the YRI and CEU populations?
- **Statistical inference is needed again!**

What's next

Introduction to p-values, first using simulation of flipping a coin then regression

'homework'

Use the gene expression of *LCT* dataset GWAS-workshop-sample-dataset-LCT.txt to repeat some of the analyses and graphing shown so far, e.g. start with

```
# Don't forget to first downloaded the data to your WORKING directory
mydata.LCT=read.table("GWAS-workshop-sample-dataset-LCT.txt",header=T)
mydata.LCT[c(1:3,103:105,193:195),1:9]
```

```
##       FID     IID PID MID SEX   PHENO POP SNP1.5618704 SNP1.57815437
## 1    1328 NA06984   0   0   1 7.44421 CEU           AA            AA
## 2    1328 NA06989   0   0   2 7.27992 CEU           AA            AA
## 3    1330 NA12340   0   0   1 6.85208 CEU           AA            AA
## 103 13291 NA07435   0   0   1 8.39611 CEU           AA            AA
## 104 13292 NA07051   0   0   1 6.98187 CEU           AA            AA
## 105  Y001 NA18486   0   0   1 6.87917 YRI           CA            GA
## 193  Y116 NA19236   0   0   1 7.25701 YRI           CA            GG
## 194  Y120 NA19247   0   0   2 7.50315 YRI           AA            AA
## 195  Y120 NA19248   0   0   1 7.09252 YRI           CC            AA
```