Final Exam

Alexander Frieden

April 27, 2016

1. What is PCA? Why would it be used?

2. Use the following to do a Bonferoni Correction on the data **(Hint: Use p.adjust())**

```
Input = (

"Factor  Raw.p
 A        .001
 B        .01
 C        .025
 D        .05
 E        .1
")
Data = read.table(textConnection(Input),header=TRUE)
```

How many pass a threshold of $\alpha = 0.01$ before and after?

3. For this problem we are going to use the gene expression data set found here: http://www-bcf.usc.edu/~gareth/ISL/Ch10Ex11.csv

This data is gene expression from 40 tissue samples with measurements on 1000 genes. The first 20 samples are from healthy patients while the second 20 are from a diseased group.

A) Load data into R. Remember to set **header=F**.

B) Apply hierarchial clustering to the samples using correlation-based distance, and plot the dendrogram. Do the genes seperate the samples into the two groups? Do your results depend on the type of linkage used?

C) Your collaborator wants to know which genes differ the most across the two groups. Suggest a way to answer the question. For a bonus, apply it here.

4. Using the genetics package, run:

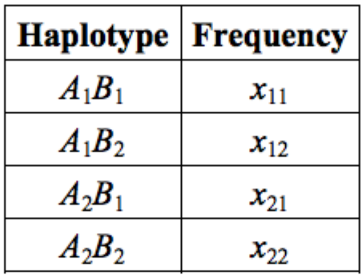
```
install.packages("genetics", repos="http://cran.rstudio.com/")
library(genetics)
```

Then use the **genotype()** method and the **LD()** method to compute the r^2 pairwise linkage disequilibrium on the following arrays. What does this tell us?

```
v1<- c('A/A','A/C','C/C','C/A',NA,'A/A','A/C','A/C')
v2<- c('A/A','C/C','C/A','C/A',NA,'A/A','A/C','A/C')
```

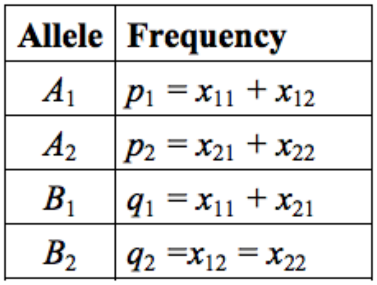
Bonus : For the following haplotype frequencies

```
##
## The downloaded binary packages are in
##  /var/folders/tj/8dxhxfns3fb0fx5kswwdvjbr0000gp/T//Rtmpxto0r9/downloaded_packages
```



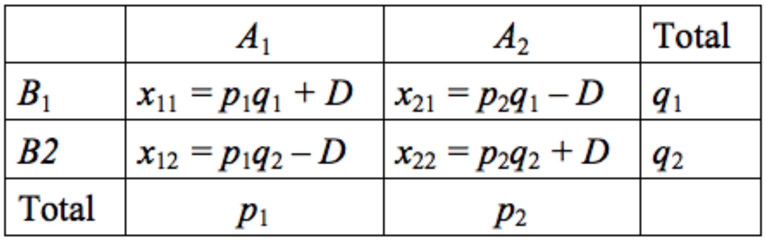
Haplotype	Frequency
A_1B_1	x_{11}
A_1B_2	x_{12}
A_2B_1	x_{21}
A_2B_2	x_{22}

And for the following allele frequencies



Allele	Frequency
A_1	$p_1 = x_{11} + x_{12}$
A_2	$p_2 = x_{21} + x_{22}$
B_1	$q_1 = x_{11} + x_{21}$
B_2	$q_2 = x_{12} = x_{22}$

Which can be rewritten



	A_1	A_2	Total
B_1	$x_{11} = p_1q_1 + D$	$x_{21} = p_2q_1 - D$	q_1
$B2$	$x_{12} = p_1q_2 - D$	$x_{22} = p_2q_2 + D$	q_2
Total	p_1	p_2	

Prove that the Linkage Disequilibrium D is $D = (x_{11})(x_{22})\breve{\ }(x_{12})(x_{21})$