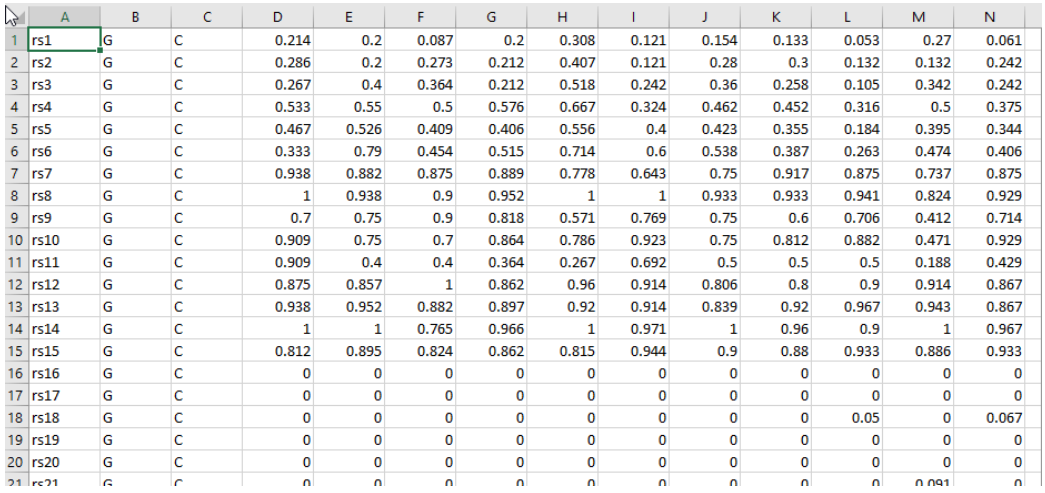
WINDOWS OS

Programs downloaded and used: PuTTY, WinSCP, RStudio

* * = fill in with your information

1. **Setting Up Hoffman**
    - Download PuTTY, hostname: hoffman2.idre.ucla.edu
    - On the page where it says do not compute on node-ssh into your hoffman login using:
        - ssh *username*@hoffman2.idre.ucla.edu
        - Prompted to type in password
    - Request an interactive shell to program
        - qrsh -l h_rt=3:00:00,h_data=3g
            - Can change h_rt to be less time for the interactive shell
2. **Setting Up Gemma**
    - Download latest file here
    - Use WinSCP to move gemma download from computer into the hoffman terminal
        - Changed name of gemma download to gemma.gz
    - gunzip gemma.gz (unzips the file name gemma)
    - chmod 700 gemma (to turn it into an executable)
3. **Formatting Genotype Matrices**
    - Transpose the genotype matrix so that the chromosome position and number is on the first column, this means that each individual sample has its own column as opposed to its own row.
    - Duplicate the excel file and delete the name of the sample, chromosome and position (make sure to duplicate because you still need the chromosome and position for the SNP annotation file)
    - Add a rs1, rs2, rs3,..., # of rows you have for the data as the first column
    - 2nd column add a column of all G
    - 3rd column add a column of all C
    - It should look something like this
        - 

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	rs1	G	C	0.214	0.2	0.087	0.2	0.308	0.121	0.154	0.133	0.053	0.27	0.061
2	rs2	G	C	0.286	0.2	0.273	0.212	0.407	0.121	0.28	0.3	0.132	0.132	0.242
3	rs3	G	C	0.267	0.4	0.364	0.212	0.518	0.242	0.36	0.258	0.105	0.342	0.242
4	rs4	G	C	0.533	0.55	0.5	0.576	0.667	0.324	0.462	0.452	0.316	0.5	0.375
5	rs5	G	C	0.467	0.526	0.409	0.406	0.556	0.4	0.423	0.355	0.184	0.395	0.344
6	rs6	G	C	0.333	0.79	0.454	0.515	0.714	0.6	0.538	0.387	0.263	0.474	0.406
7	rs7	G	C	0.938	0.882	0.875	0.889	0.778	0.643	0.75	0.917	0.875	0.737	0.875
8	rs8	G	C	1	0.938	0.9	0.952	1	1	0.933	0.933	0.941	0.824	0.929
9	rs9	G	C	0.7	0.75	0.9	0.818	0.571	0.769	0.75	0.6	0.706	0.412	0.714
10	rs10	G	C	0.909	0.75	0.7	0.864	0.786	0.923	0.75	0.812	0.882	0.471	0.929
11	rs11	G	C	0.909	0.4	0.4	0.364	0.267	0.692	0.5	0.5	0.5	0.188	0.429
12	rs12	G	C	0.875	0.857	1	0.862	0.96	0.914	0.806	0.8	0.9	0.914	0.867
13	rs13	G	C	0.938	0.952	0.882	0.897	0.92	0.914	0.839	0.92	0.967	0.943	0.867
14	rs14	G	C	1	1	0.765	0.966	1	0.971	1	0.96	0.9	1	0.967
15	rs15	G	C	0.812	0.895	0.824	0.862	0.815	0.944	0.9	0.88	0.933	0.886	0.933
16	rs16	G	C	0	0	0	0	0	0	0	0	0	0	0
17	rs17	G	C	0	0	0	0	0	0	0	0	0	0	0
18	rs18	G	C	0	0	0	0	0	0	0	0	0.05	0	0.067
19	rs19	G	C	0	0	0	0	0	0	0	0	0	0	0
20	rs20	G	C	0	0	0	0	0	0	0	0	0	0	0
21	rs21	G	C	0	0	0	0	0	0	0	0	0	0.091	0

4. **Format Phenotype Data**
    - Select column of phenotype data and paste into a separate csv file. Get rid of any other data so it's just number values in the csv
        - Make sure there is a phenotype for every genotype
            - Example: if there are 61 individual genotype data make sure there is also 61 phenotypes

Important Note: If the number of genotype and phenotype data do not match up in terms of individuals then the program will not be able to compute a kinship matrix.

If there is more genotype data than phenotype, delete the missing individuals from the genotype data file so that the number of phenotype and genotype samples match and vice versa otherwise.

5. **Creating SNP Annotation File**
    - 1st column rs 1, rs2, rs3,....,# of rows in genotype matrices
    - 2nd column chromosome position
    - 3rd column chromosome number
        - Should look something like this
            -

rs1	101847018	1
rs2	101847024	1
rs3	101847035	1
rs4	101847042	1
rs5	101847053	1
rs6	101847055	1
rs7	10496238	1
rs8	10688062	1
rs9	10688071	1
rs10	10688076	1

Excel Lesson: If chromosome number and position looks like this: “chr1_101847018”, can use this command, =LEFT(D1,FIND("_",D1)-1), to get the chr1 isolated in its own cell and use this command,=RIGHT(D1,LEN(D1)-FIND("_",D1)), to get 101847018 isolated in its own cell. Assuming that “chr1_101847018” is on the cell D1.

6. **Using GEMMA to get Kinship Matrix and p values**
    - ./gemma -h (to open gemma program)
    - ./gemma -g *matrices csv file name*.csv -p *phenotype csv file name*.csv -gk -o *the kinship matrix file name* (compute kinship matrix)
    - ./gemma -g *matrices csv file name*.csv -p *phenotype csv file name*.csv -a *snp file name*.csv -k *kinship matrix name*.cXX.txt -lmm 4 -o *name of association file* (compute 4 different p values using linear mixed model)
    -
7. **Manhattan Plot in RStudio**
    - install.packages(‘qqman’) (installs packages, type into command line)
    - library(‘qqman’) (loads packages,usually kept as first line of code)

```
civicage <- read.csv("C:\\Users\\Angel\\Desktop\\pellegrini research\\gemma project\\manhattan data\\manhattanage.csv")
civicbmi <- read.csv("C:\\Users\\Angel\\Desktop\\pellegrini research\\gemma project\\manhattan data\\manhattanbmi.csv")
civicgender <- read.csv("C:\\Users\\Angel\\Desktop\\pellegrini research\\gemma project\\manhattan data\\manhattangender.csv")
civicheight <- read.csv("C:\\Users\\Angel\\Desktop\\pellegrini research\\gemma project\\manhattan data\\manhattanheight.csv")
civicweight <- read.csv("C:\\Users\\Angel\\Desktop\\pellegrini research\\gemma project\\manhattan data\\manhattanweight.csv")

manhattan(civicweight, chr= "CHR", bp= "BP", snp= "ï..SNP", p= "P",
          col = c("gray10", "gray60"),
          chrlabs = NULL,
          suggestiveline = -log10(1e-5),
          genomewideline = -log10(6.7e-06),
          logp = TRUE,
          annotatePval = NULL,
          annotateTop = TRUE, main="CIVIC Data Weight Association")
```

- Can change suggestiveline and genomewideline to whatever threshold the project requires
- main=name of the project
- Change first variable in manhattan() function to get a manhattan plot for a different phenotype
- Here is a link for the description of each variable in the manhattan function