Day 3, AM Session

PLINK

whole genome association analysis toolkit

Richard Flamio Jr., Ph.D.
Madison Zimmerman
Kristina M. Ramstad, Ph.D.
University of South Carolina Aiken

What is PLINK?

- An open-source tool for analyzing genotype and phenotype data from large omics datasets

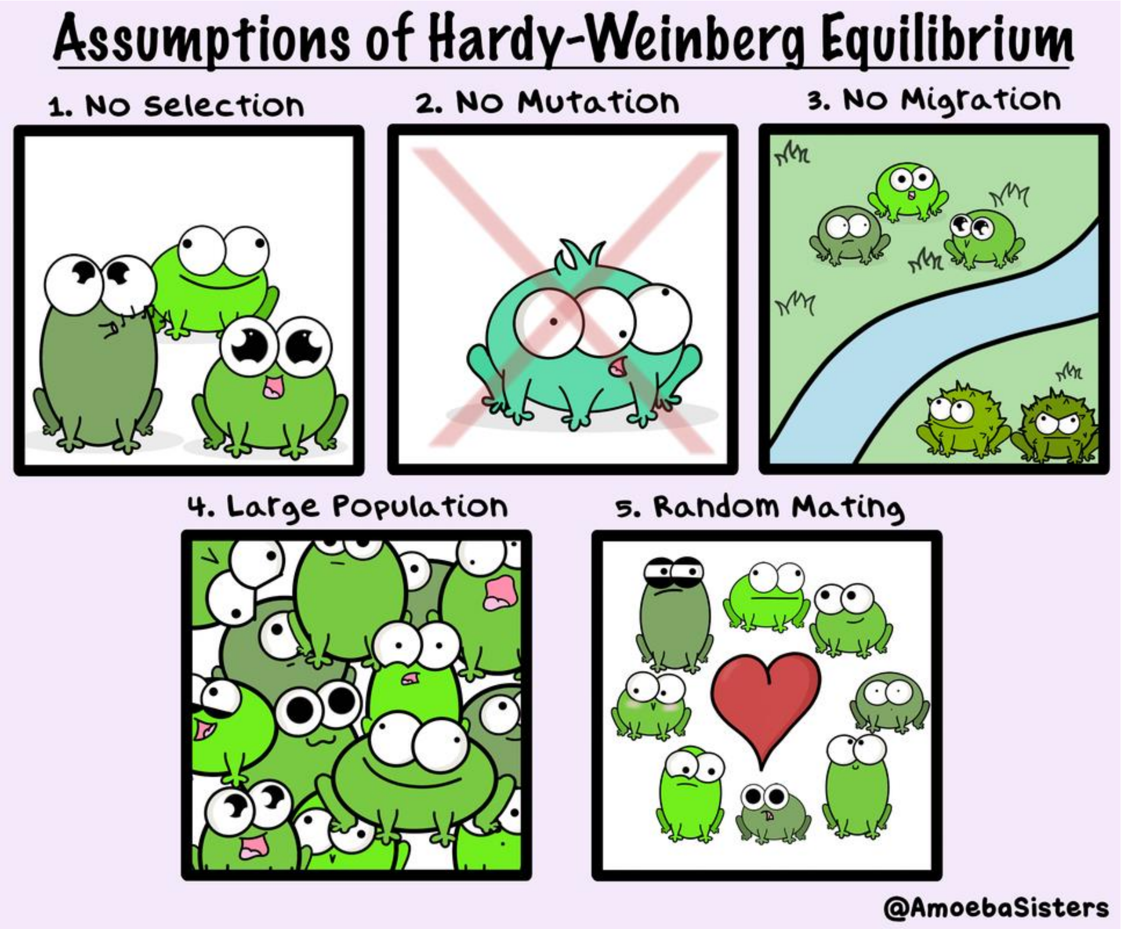
https://zzz.bwh.harvard.edu/plink/index.shtml

- Variant calling must be done beforehand

- Visualization of results has limited support --> we will visualize results using R and associated packages

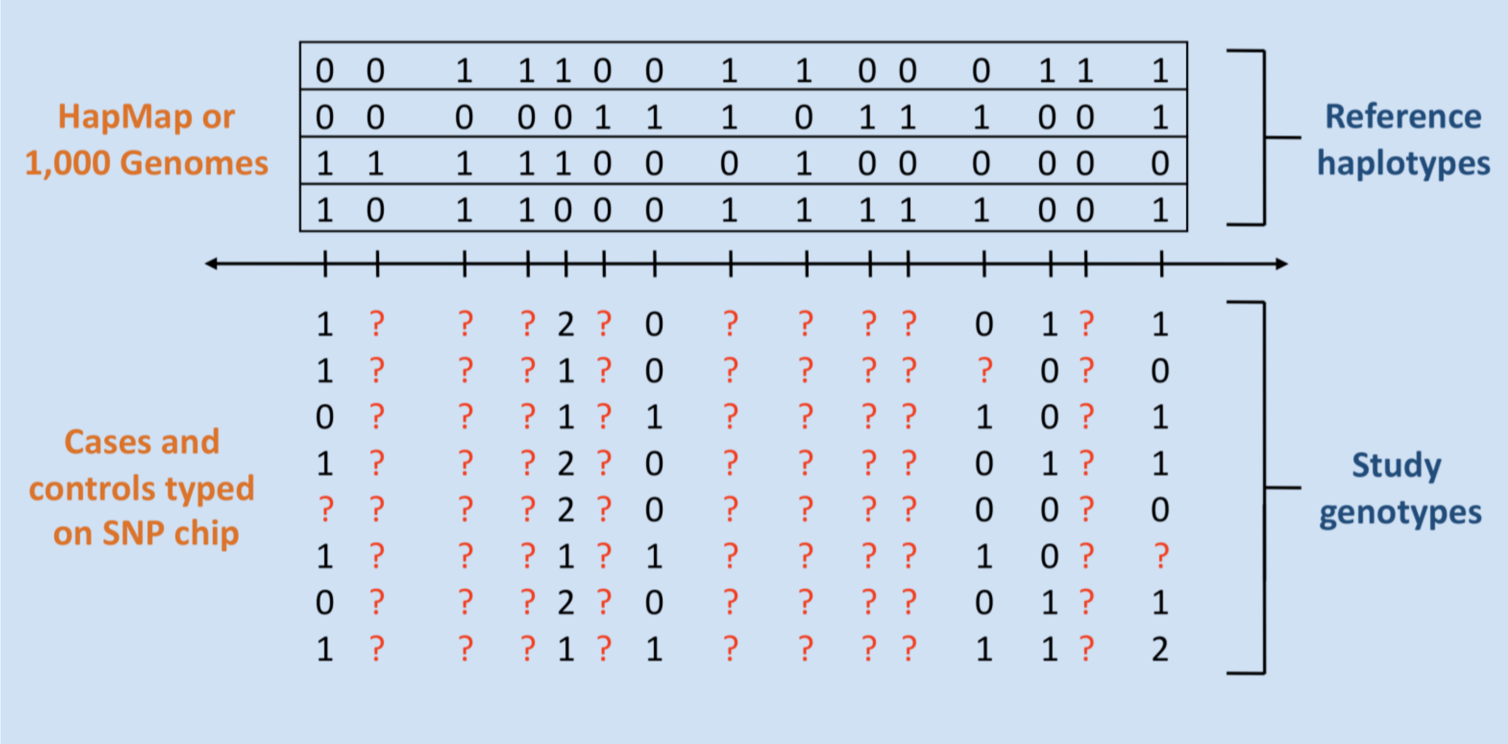
Some PLINK Capabilities

- Data management
  - Read in data in various formats
  - Merge files
  - Extract SNP subset
- Summary statistics
  - Allele frequencies
  - Hardy-Weinberg Equilibrium
  - Missing genotype rates
  - Kinship



More PLINK Capabilities

- Population stratification
  - Significance tests for whether two individuals belong to same population
- Association testing
  - Case/control
    - Fisher's exact test
    - Logistic regression
  - Quantitative traits
    - Linear regression
- Imputation



HapMap or 1,000 Genomes

Reference haplotypes

0	0	1	1	1	0	0	1	1	0	0	0	1	1	1
0	0	0	0	0	1	1	1	0	1	1	1	0	0	1
1	1	1	1	1	0	0	0	1	0	0	0	0	0	0
1	0	1	1	0	0	0	1	1	1	1	1	0	0	1

Cases and controls typed on SNP chip

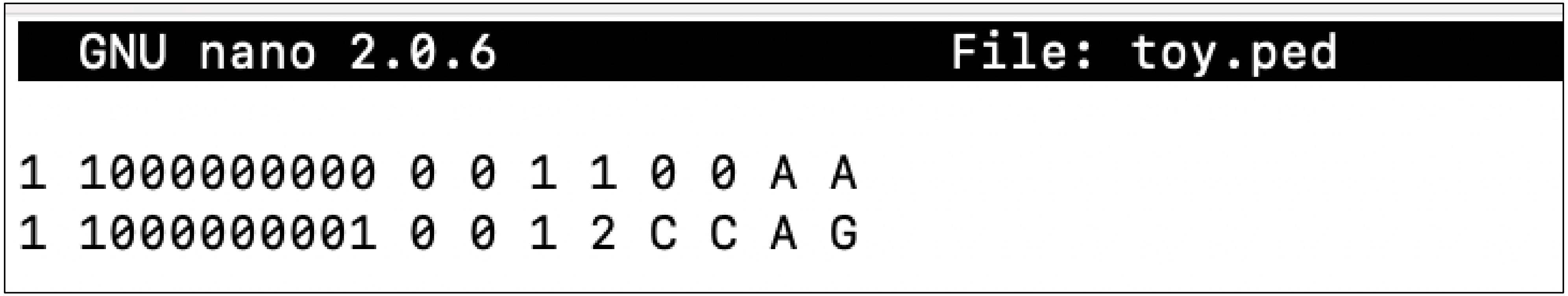
Study genotypes

1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	0	?	?	?	?	?	0	?	0
0	?	?	?	1	?	1	?	?	?	?	1	0	?	1
1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
?	?	?	?	2	?	0	?	?	?	?	0	0	?	0
1	?	?	?	1	?	1	?	?	?	?	1	0	?	?
0	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	1	?	?	?	?	1	1	?	2

https://mathgen.stats.ox.ac.uk/impute/impute_v2.html

PED (.ped) File

- Pedigree, genotype, and phenotype information
- No header line allowed



```
  GNU nano 2.0.6                File: toy.ped

1 1000000000 0 0 1 1 0 0 A A
1 1000000001 0 0 1 2 C C A G
```

https://www.cog-genomics.org/plink/1.9/general_usage

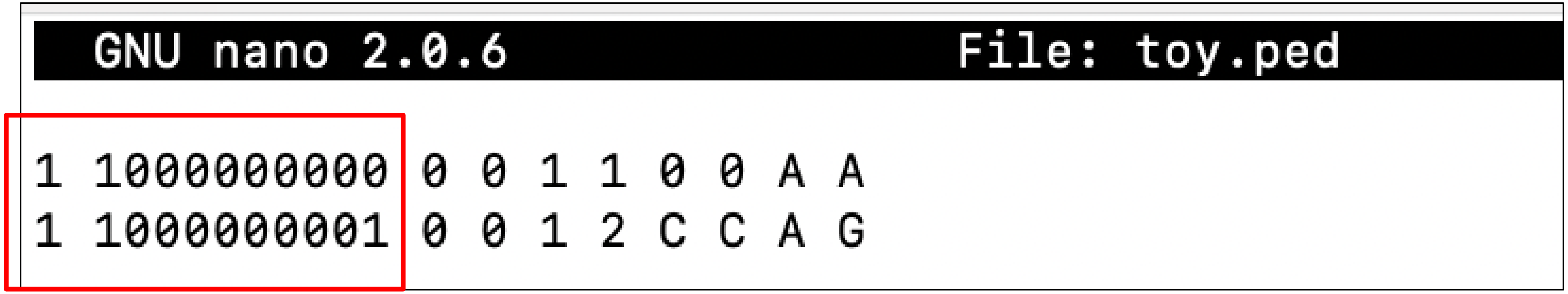
PED: Columns 1 and 2 (FID and IID)

Column 1

- Family ID (FID)
- Can be same as individual ID

Column 2

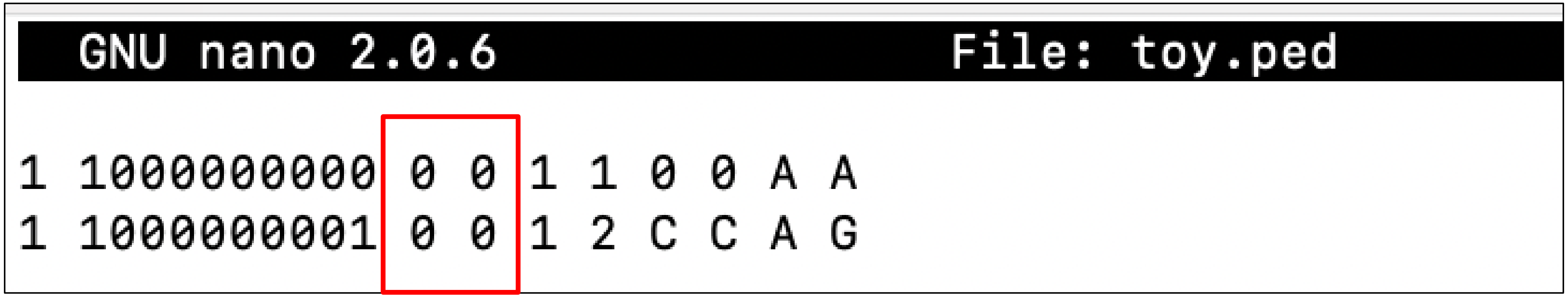
- Individual ID (IID)
- This must be unique if FIDs are the same!



```
  GNU nano 2.0.6                File: toy.ped

1 1000000000 0 0 1 1 0 0 A A
1 1000000001 0 0 1 2 C C A G
```

PED: Columns 3 (PATID) and 4 (MATID)

- Required even if pedigree is unknown
- Column 3 = paternal ID (PATID or FATID)
- Column 4 = maternal ID (MATID)
- If pedigree is known, fill in the cell with the appropriate ID
- If the pedigree is not known, fill in the cell with –9 or 0, which code for missing data

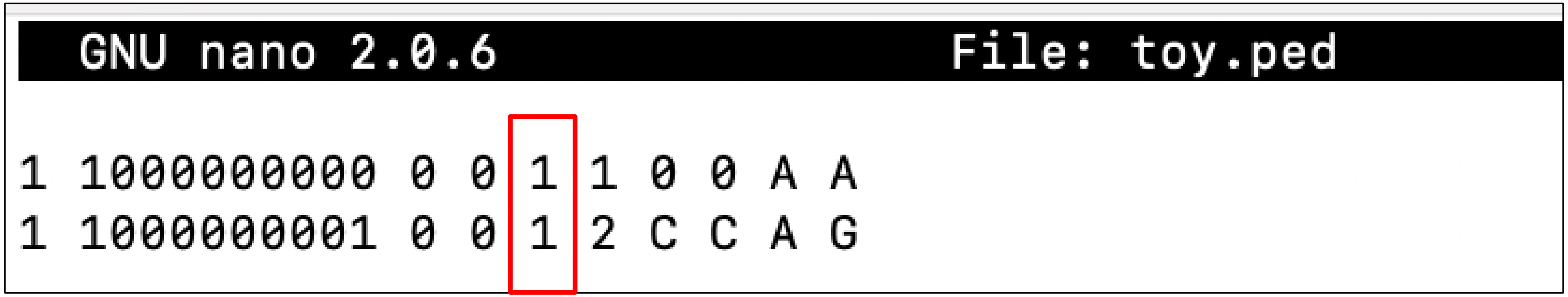


```
  GNU nano 2.0.6                   File: toy.ped

1 1000000000 0 0 1 1 0 0 A A
1 1000000001 0 0 1 2 C C A G
```

PED: Column 5 (Sex)

Column 5 = Sex

- If male --> 1
- If female --> 2
- If unknown --> 0 or -9

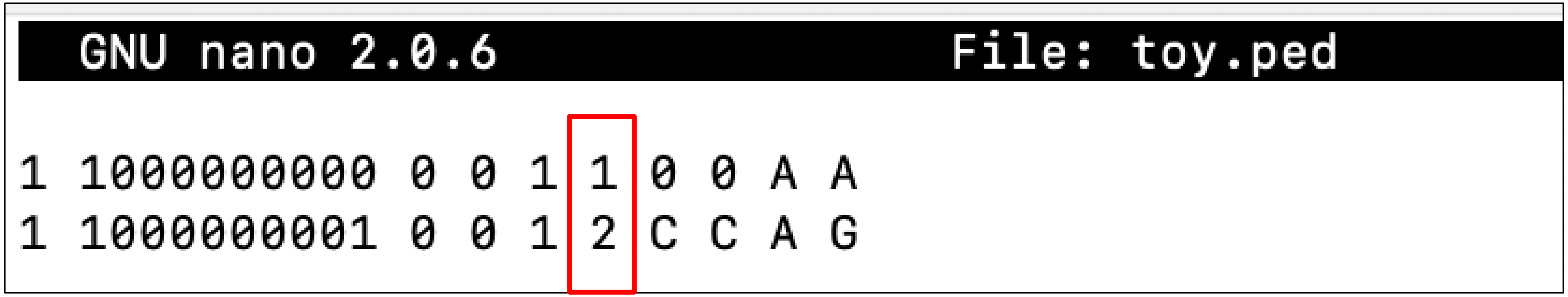


```
  GNU nano 2.0.6                      File: toy.ped

1 1000000000 0 0 1 1 0 0 A A
1 1000000001 0 0 1 2 C C A G
```

PED: Column 6 (Trait of interest)

Column 6 = Phenotype

- Quantitative
  - Plant height: 1.12
  - No commas
- Case/control
  - Unaffected = 1
  - Affected = 2
  - Missing = -9 or 0



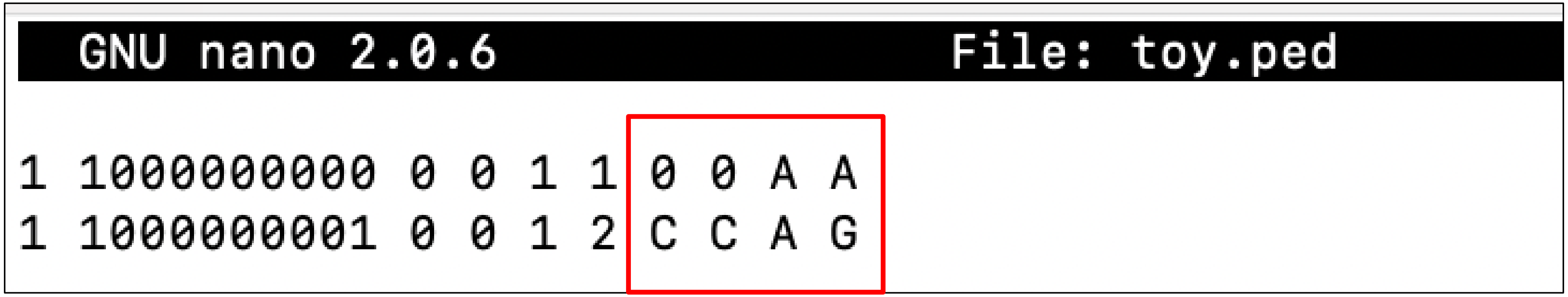
```
  GNU nano 2.0.6                    File: toy.ped

1 1000000000 0 0 1 1 0 0 A A
1 1000000001 0 0 1 2 C C A G
```

PED: Columns 7 and Beyond (Genotypes)

Genotype data

- Must be biallelic
- Can code allele as numbers 1,2,3, etc. or letters (A,B,C, etc. or A,T,G,C)

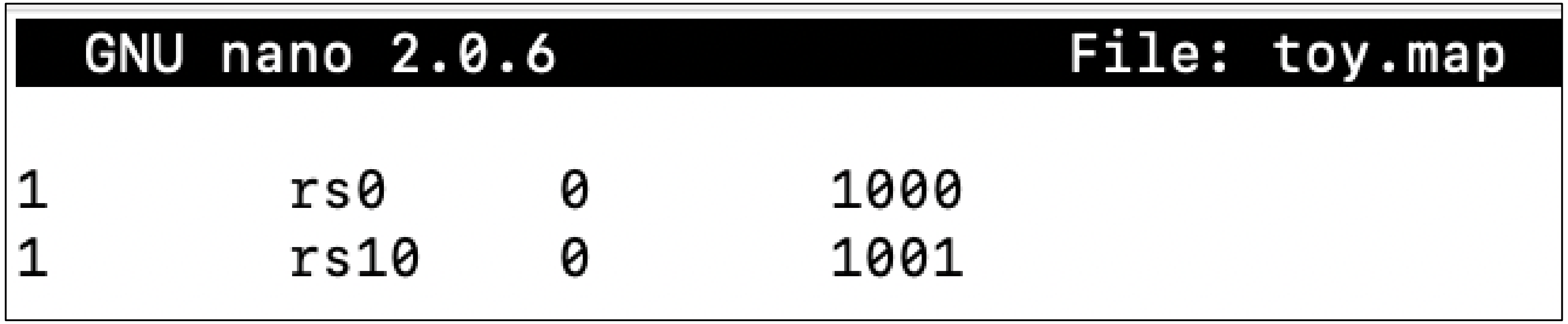
- Missing data
  - Coded as 0 or -9
  - Cannot have one allele present and one allele missing



```
  GNU nano 2.0.6                     File: toy.ped

1 1000000000 0 0 1 1 0 0 A A
1 1000000001 0 0 1 2 C C A G
```

MAP (.map) File

- Four columns
- Marker information
- Same order as PED file
- Do not need to be in genomic order



```
  GNU nano 2.0.6                      File: toy.map


1       rs0     0       1000
1       rs10    0       1001
```

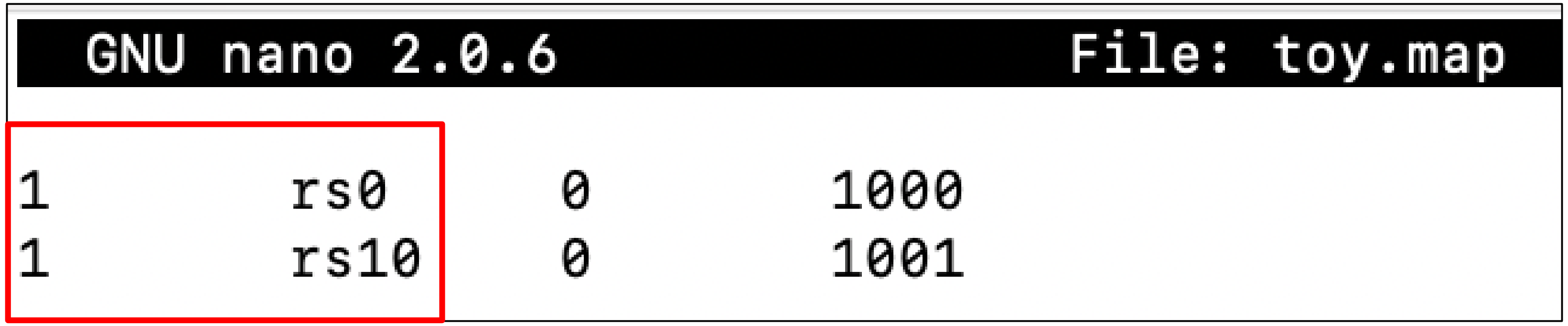
MAP: Columns 1 and 2

Column 1 = Chromosome

- Number (1, 2, 3, etc.)
- If unplaced --> 0
- X = X chromosome
- Y = Y chromosome
- XY = pseudo-autosomal region
- MT = mitochondrion

Column 2 = SNP identifier

- rs# for human SNP ID
- chromosome:position for most other organisms



```
  GNU nano 2.0.6                  File: toy.map

1        rs0      0        1000
1        rs10     0        1001
```

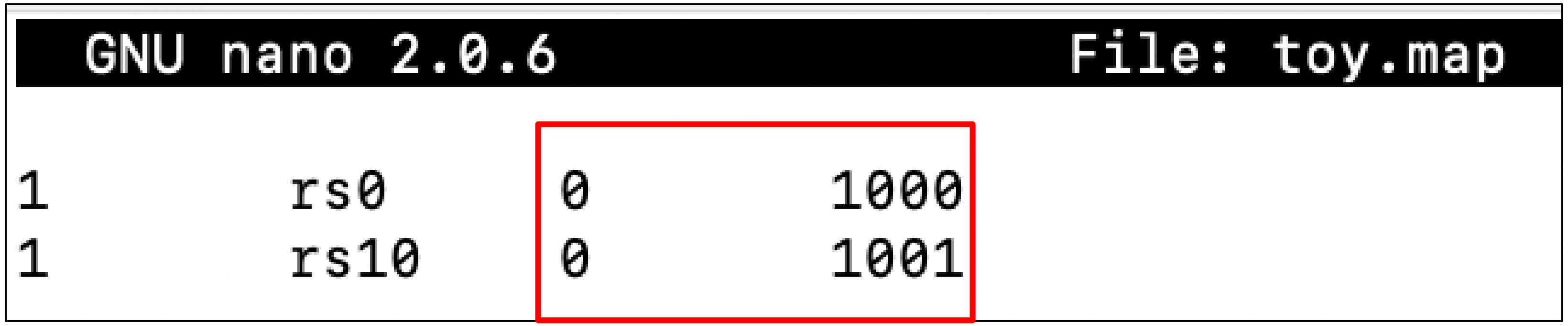
MAP: Columns 3 and 4

Column 3 = Genetic distance (cM)

- Set to 0 for association testing

Column 4 = Physical base-pair position

- If you preface this with a "-", you exclude the SNP from analysis
  - Example: -1562



```
1       rs0     0       1000
1       rs10    0       1001
```

PLINK Exercise

1. Check PLINK is installed on the system.

```
plink
```

2. Download hapmap1.zip data and unzip. Move this folder to Desktop).

   (Do this manually)

3. Open and observe each file.

```
nano hapmap1.ped
nano hapmap1.map
```

4. Identify how many individuals and SNPs are in the initial dataset.

```
plink --file hapmap1 --out hapmap1
```

Note: this implies there are the files hapmap1.ped and hapmap1.map in the directory

Answer

# 89 individuals and 83534 SNPs

Binary PED file

- Compact form that saves time and space
- Produced from command plink --file hapmap1
  - .bed = raw genotype data
    - this is the binary file
    - Cannot be viewed easily
  - .bim = raw genotype data + allele names
    - Can be viewed (nano)
  - .fam = first six columns of .ped file
    - Can be viewed (nano)

Exercise:

```
ls
nano hapmap1.bim
nano hapmap1.fam
```

Summary Statistics

- --missing --> missing data per individual (.imiss) and marker (.lmiss)
- --freq --> minor allele frequencies (.frq)
- --hardy --> Hardy-Weinberg Equilibrium (.hwe)

Adding the --chr modifier (example: --chr 1) provides the statistics for only the chromosome of interest

Adding the --snp modifier (with a SNP ID) provides the statistics for only the SNP of interest

Missing Data Exercise

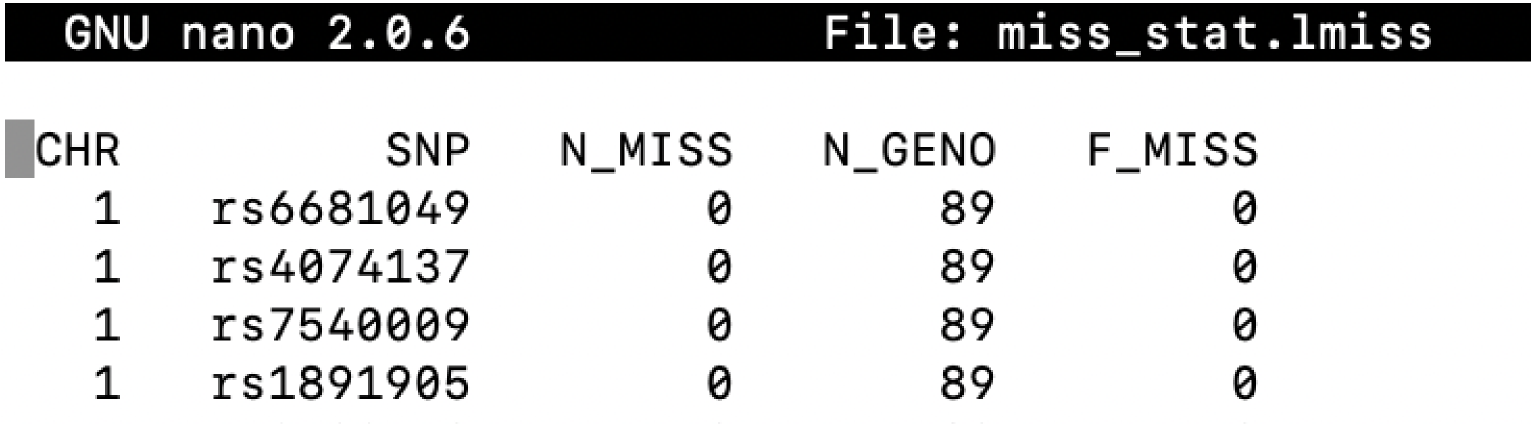
Determine amount of missing data

1. `plink --file hapmap1 --missing --out miss_stat`
2. `nano miss_stat.lmiss` Gives missing data per marker/locus

   `nano miss_stat.imiss` Gives missing data per individual

Missing data shown as counts and frequencies.

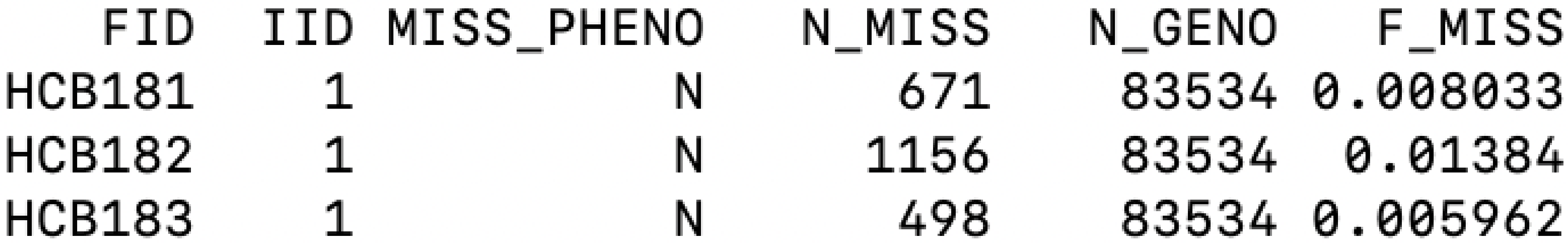
.lmiss



```
  GNU nano 2.0.6                File: miss_stat.lmiss

CHR          SNP   N_MISS   N_GENO   F_MISS
  1    rs6681049        0       89        0
  1    rs4074137        0       89        0
  1    rs7540009        0       89        0
  1    rs1891905        0       89        0
```

N_MISS = number of missing genotypes/locus
F_MISS = percent of missing data/locus

.imiss



```
   FID  IID MISS_PHENO   N_MISS   N_GENO   F_MISS
HCB181    1          N      671    83534 0.008033
HCB182    1          N     1156    83534  0.01384
HCB183    1          N      498    83534 0.005962
```

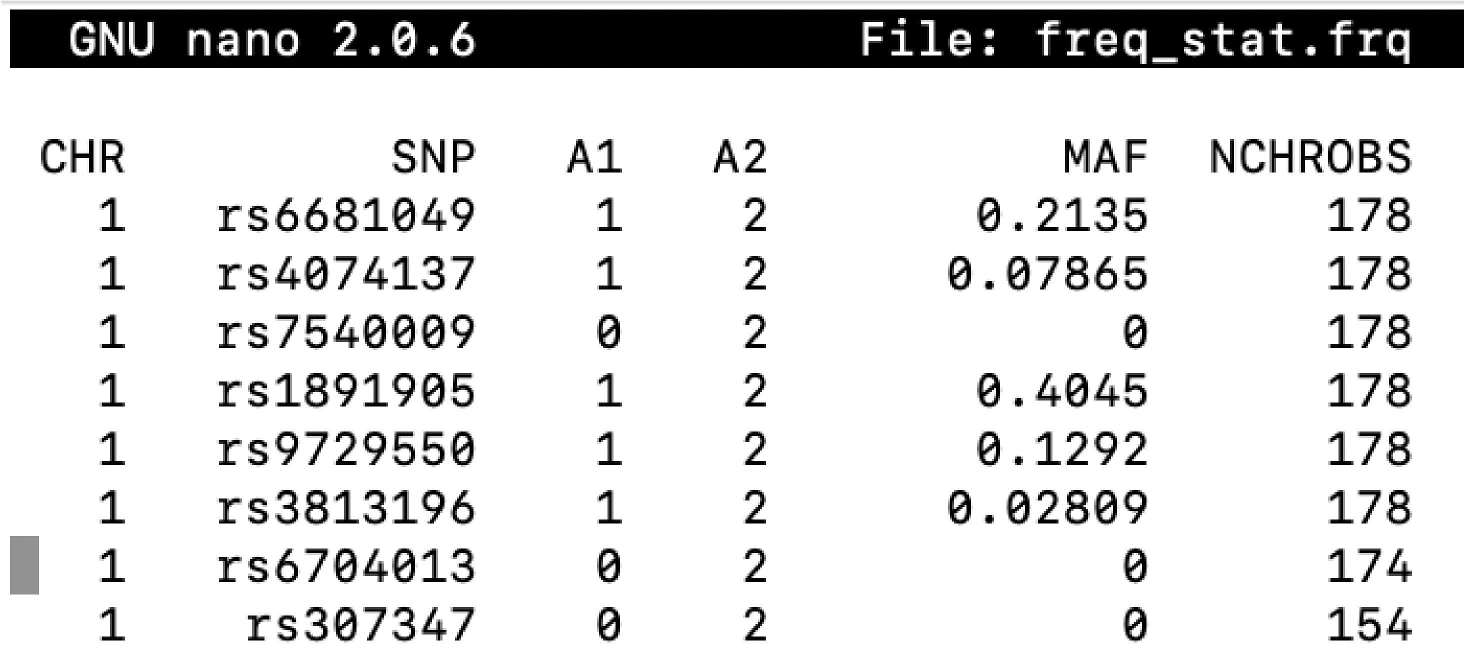
N_MISS = number of missing genotypes/individual
F_MISS = percent of missing data/individual

Minor Allele Frequency Exercise

Using the complete dataset, no population subdivision

```
plink --file hapmap1 --freq --out freq_stat
nano freq_stat.frq
```

.frq



```
  GNU nano 2.0.6                     File: freq_stat.frq

 CHR          SNP   A1   A2          MAF  NCHROBS
   1    rs6681049    1    2       0.2135      178
   1    rs4074137    1    2      0.07865      178
   1    rs7540009    0    2            0      178
   1    rs1891905    1    2       0.4045      178
   1    rs9729550    1    2       0.1292      178
   1    rs3813196    1    2      0.02809      178
   1    rs6704013    0    2            0      174
   1     rs307347    0    2            0      154
```

A1 = minor allele

A2 = major allele

MAF = minor allele frequency

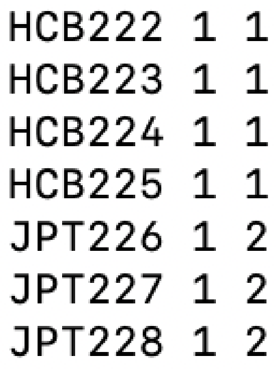
NCHROBS = number of alleles without missing data (actually genotyped)

MAF (population subdivision)

1. Look at the pop.phe file

```
nano pop.phe
```

FID, IID, Population

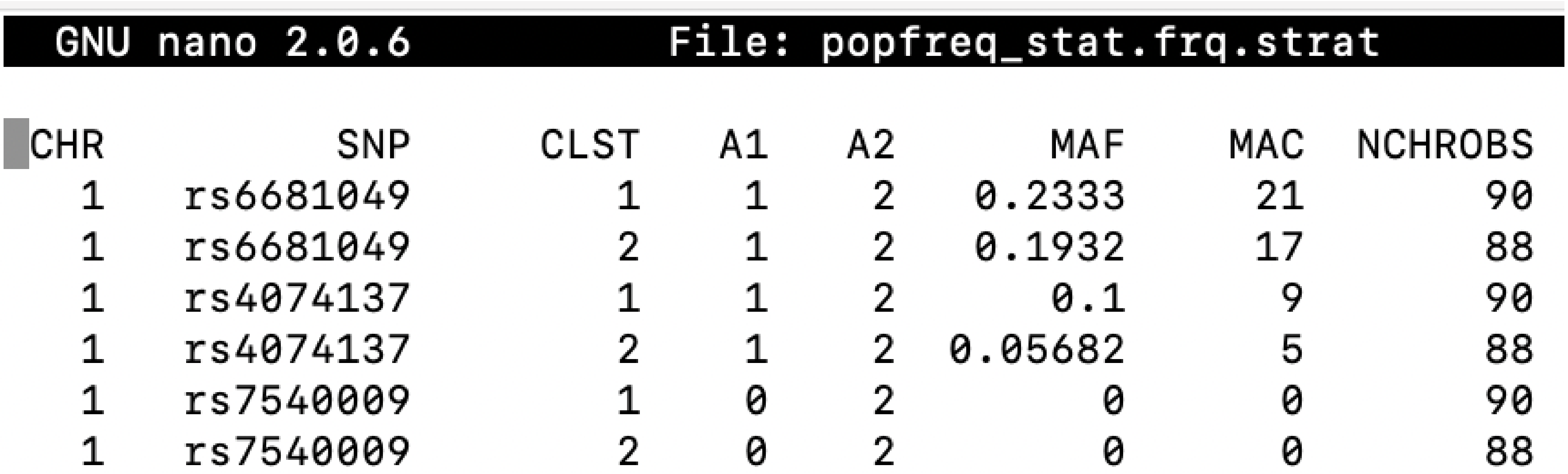


```
HCB222 1 1
HCB223 1 1
HCB224 1 1
HCB225 1 1
JPT226 1 2
JPT227 1 2
JPT228 1 2
```

2. Calculating MAF per subpopulation

```
plink --file hapmap1 --freq --within pop.phe --out popfreq_stat
nano popfreq_stat.frq.strat
```

.frq.strat



```
  GNU nano 2.0.6          File: popfreq_stat.frq.strat

 CHR          SNP     CLST   A1   A2      MAF    MAC  NCHROBS
   1    rs6681049        1    1    2   0.2333     21       90
   1    rs6681049        2    1    2   0.1932     17       88
   1    rs4074137        1    1    2      0.1      9       90
   1    rs4074137        2    1    2  0.05682      5       88
   1    rs7540009        1    0    2        0      0       90
   1    rs7540009        2    0    2        0      0       88
```

CLST = cluster
MAC = minor allele count

Filtering Options

- --mind --> allowed missing data/individual to retain individual
  - Example: --mind 0.05 for retaining individuals with 95% data present

- --geno --> allowed missing data/marker to retain marker
  - Example: --geno 0.05 for retaining markers with genotyping call rate of 95%

- --maf --> allow markers with this Minor Allele Frequency and above
  - Example: --maf 0.05 for retaining markers with MAF 0.05 and above

- --hwe --> performs Hardy-Weinberg Exact test at specified threshold
  - Be careful when using this filter
  - Make sure you do this for each population separately

- Syntax example: plink --file hapmap1 --make-bed --mind 0.05 –out filter1

Filtering Options Exercise

Retain markers with a genotype call rate of 85%.

a. How many markers were removed?

b. How many markers were retained?

Filtering Options Answer

```
plink --file hapmap1 --geno 0.15 --make-bed --out filter1
```

```
# 136 markers removed
# 83398 markers retained
```

Filtering Options

- To keep or remove individuals from an analysis
  - Two column text file (family ID, individual ID), no header
  - --keep *textfile* = keep these individuals
  - --remove *textfile* = remove these individuals

- To keep or remove certain SNP_IDs from an analysis
  - One column with SNP_IDs, no header
  - --extract *textfile* = keeps these variants
  - --exclude *textfile* = remove these variants

What type of data is the phenotype? Exercise

1. Open the hapmap1.ped file.

```
nano hapmap1.ped
```

2. Is phenotype continuous, binary, or categorical?

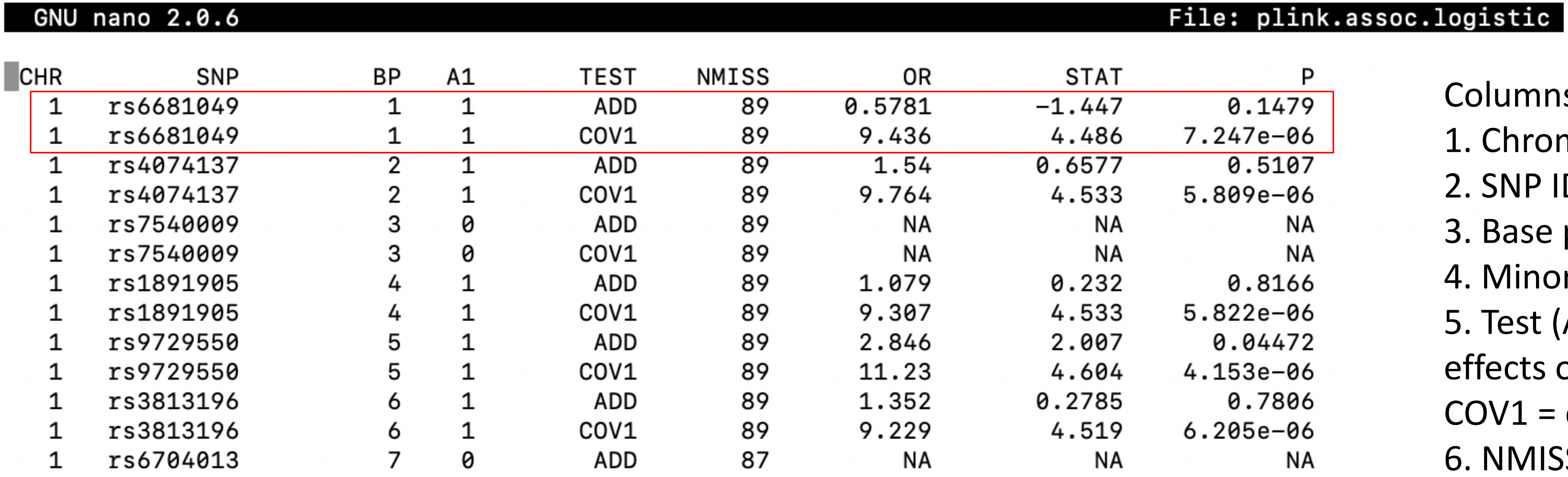
3. What type of regression should we perform?

(Options: linear, binary logistic, multinomial logistic, ordered logistic)

Logistic Regression Exercise

```
plink --file hapmap1 --make-bed --logistic --covar pop.phe --out withpopA
nano withpopA.assoc.logistic
```

Output File



```
 GNU nano 2.0.6                                                File: plink.assoc.logistic

CHR          SNP         BP   A1         TEST    NMISS         OR         STAT            P
  1    rs6681049          1    1          ADD       89     0.5781       -1.447       0.1479
  1    rs6681049          1    1         COV1       89      9.436        4.486    7.247e-06
  1    rs4074137          2    1          ADD       89       1.54       0.6577       0.5107
  1    rs4074137          2    1         COV1       89      9.764        4.533    5.809e-06
  1    rs7540009          3    0          ADD       89         NA           NA           NA
  1    rs7540009          3    0         COV1       89         NA           NA           NA
  1    rs1891905          4    1          ADD       89      1.079        0.232       0.8166
  1    rs1891905          4    1         COV1       89      9.307        4.533    5.822e-06
  1    rs9729550          5    1          ADD       89      2.846        2.007      0.04472
  1    rs9729550          5    1         COV1       89      11.23        4.604    4.153e-06
  1    rs3813196          6    1          ADD       89      1.352       0.2785       0.7806
  1    rs3813196          6    1         COV1       89      9.229        4.519    6.205e-06
  1    rs6704013          7    0          ADD       87         NA           NA           NA
```

Columns:

1. Chromosome
2. SNP ID
3. Base position
4. Minor Allele
5. Test (ADD = additive effects of allele dosage, COV1 = covariate)
6. NMISS = # of nonmissing genotypes
7. Odds ratio
8. T-statistic
9. P-value

NA is present when the SNP was monomorphic

Test Column

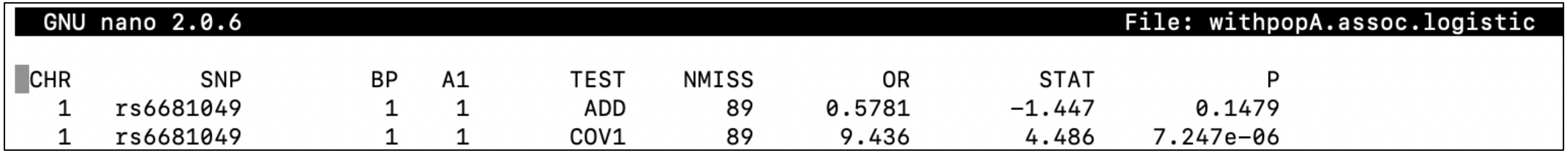
- We are interested in the additive effects of allele dosage line.
  - Additive effects of allele dosage = more copies of an allele are more or less strongly correlated with the phenotype.
- This is the significance of the SNP when controlling for the covariate (population).
- We can hide the covariate including the argument --hide-covar

Exercise

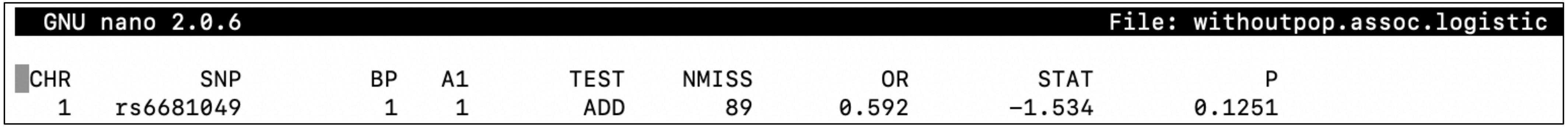
```
plink --file hapmap1 --make-bed --logistic --covar pop.phe --hide-covar
         --out withpopB
nano withpopB.assoc.logistic
```

Excluding Covariate Exercise

```
plink --file hapmap1 --make-bed --logistic --out withoutpop
nano withoutpop.assoc.logistic
```



```
  GNU nano 2.0.6                                       File: withpopA.assoc.logistic

 CHR         SNP         BP   A1      TEST    NMISS        OR      STAT          P
   1   rs6681049          1    1       ADD       89    0.5781    -1.447     0.1479
   1   rs6681049          1    1      COV1       89     9.436     4.486  7.247e-06
```



```
  GNU nano 2.0.6                                      File: withoutpop.assoc.logistic

 CHR         SNP         BP   A1      TEST    NMISS        OR      STAT          P
   1   rs6681049          1    1       ADD       89     0.592    -1.534     0.1251
```

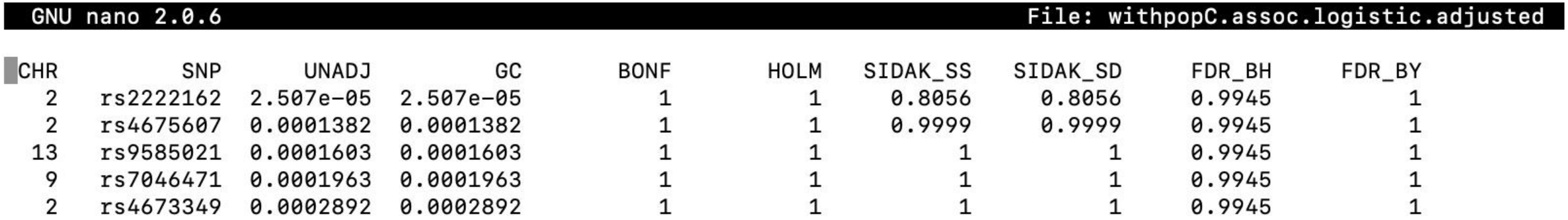
Adding the covariate, makes the SNP effect less significant.

Correcting for Multiple Testing Exercise

```
plink --file hapmap1 --make-bed --logistic --covar pop.phe --hide-covar --adjust --out withpopC
```

```
nano withpopC.assoc.logistic.adjusted
```

Correcting for Multiple Testing



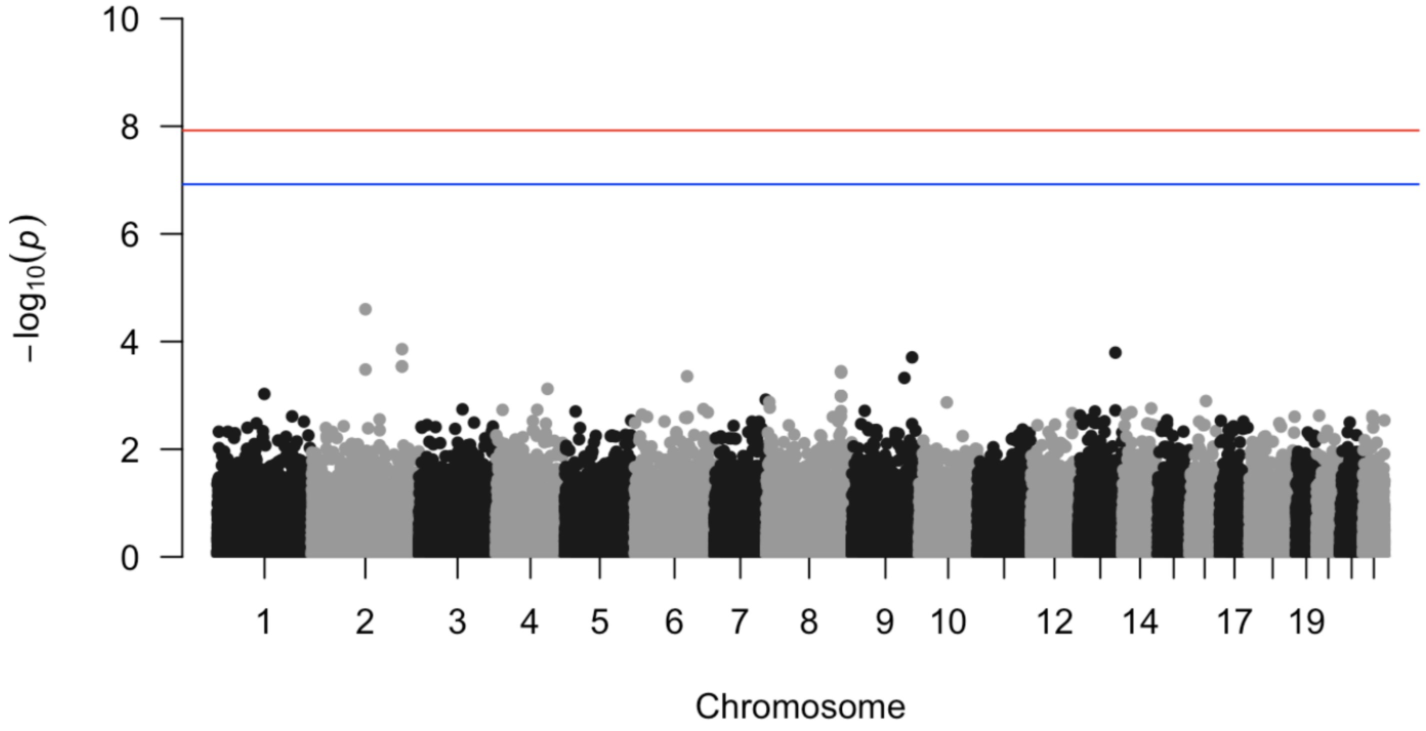
```
  GNU nano 2.0.6                                    File: withpopC.assoc.logistic.adjusted

 CHR         SNP      UNADJ         GC       BONF       HOLM   SIDAK_SS   SIDAK_SD     FDR_BH     FDR_BY
   2   rs2222162  2.507e-05  2.507e-05          1          1     0.8056     0.8056     0.9945          1
   2   rs4675607  0.0001382  0.0001382          1          1     0.9999     0.9999     0.9945          1
  13   rs9585021  0.0001603  0.0001603          1          1          1          1     0.9945          1
   9   rs7046471  0.0001963  0.0001963          1          1          1          1     0.9945          1
   2   rs4673349  0.0002892  0.0002892          1          1          1          1     0.9945          1
```

Manhattan Plot Exercise

- If you remember, we produced a file “withpopB.assoc.logistic” today.
- We are going to use this file, to produce a Manhattan plot.

- Please run the R Markdown named ‘ManhattanPlot.Rmd’ in RStudio.

Manhattan Plot Answer



References

- grunwaldlab.github
- https://bookdown.org/kdonovan125/ibis_data_analysis_r4/documenting-your-results-with-r-markdown.html