

Polymorphism and Variant Analysis Lab

Alexander E. Lipka

PowerPoint by Casey Hanson
Edited by Gio Madrigal & Roberto Cucalón Tamayo

Exercise

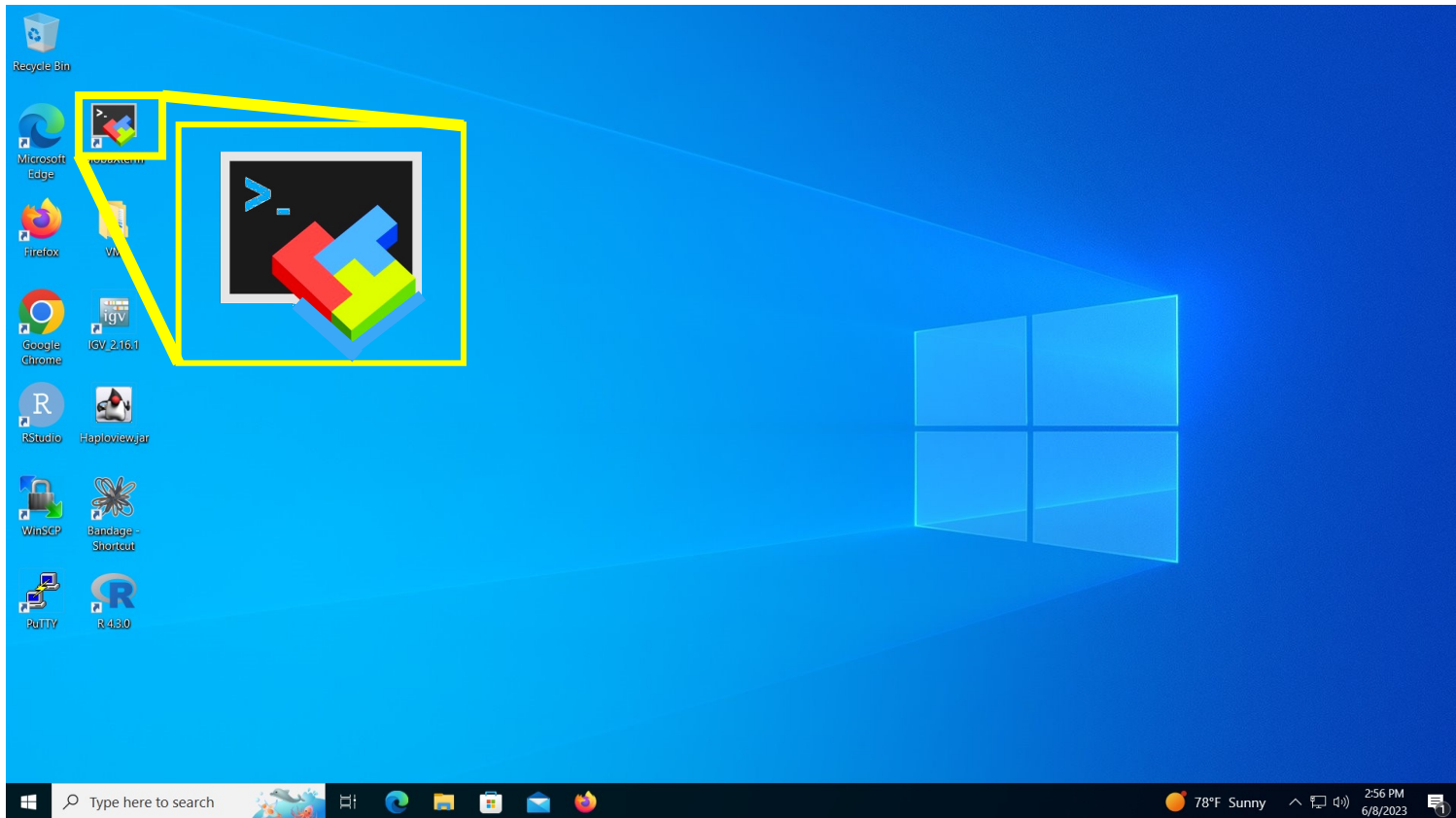
In this exercise, we will do the following:.

1. Gain familiarity with the software **PLINK**
2. Run a Quality Control (QC) analysis on genotype data of 90 individuals of two ethnic groups (Han Chinese and Japanese) genotyped for ~230,000 SNPs.
3. Use our QC data to perform a genome-wide association test (GWAS) across two phenotypes: case and control. We will compare the results of our GWAS with and without multiple hypothesis correction.

Start the VM

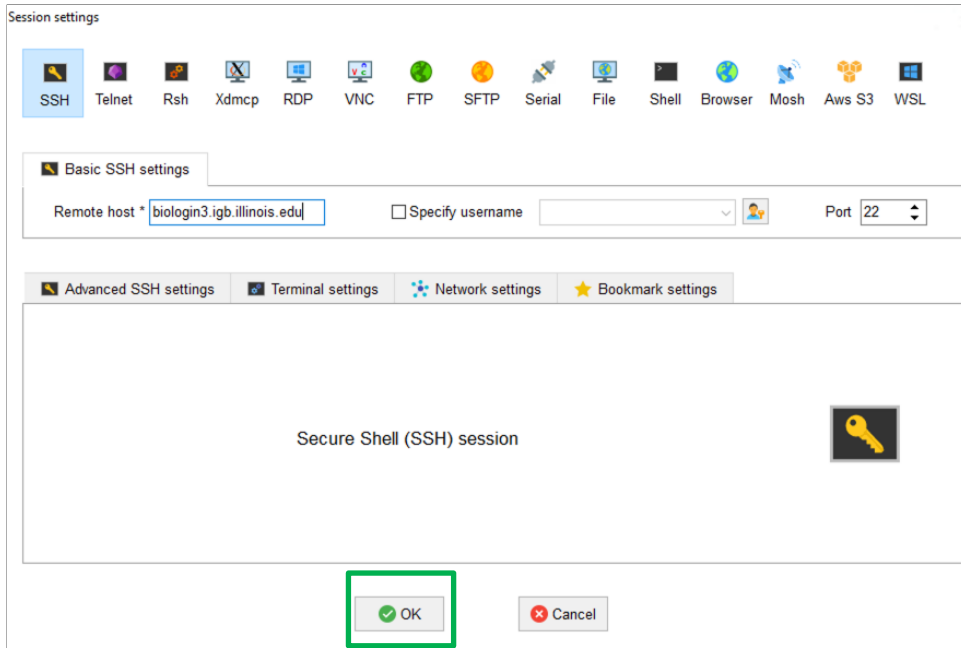
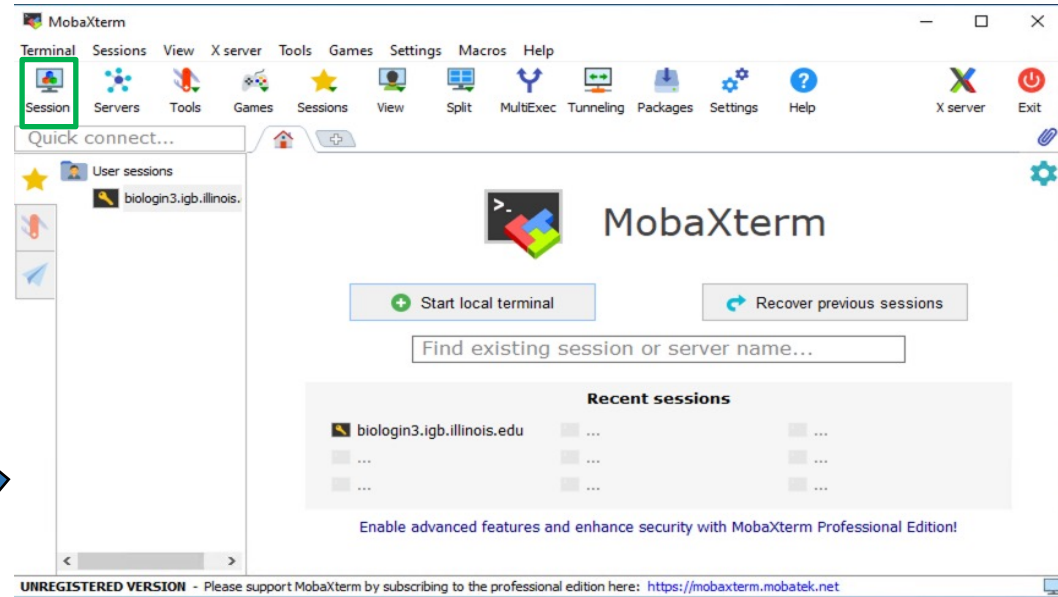
- Follow instructions for starting VM (This is the Remote Desktop software).
- The instructions are different for UIUC and Mayo participants.
- Find the instructions for this on the course website under Lab set-up:
<https://publish.illinois.edu/compgenomicscourse/2023-schedule/>

Step 0: Open MobaXterm from VM



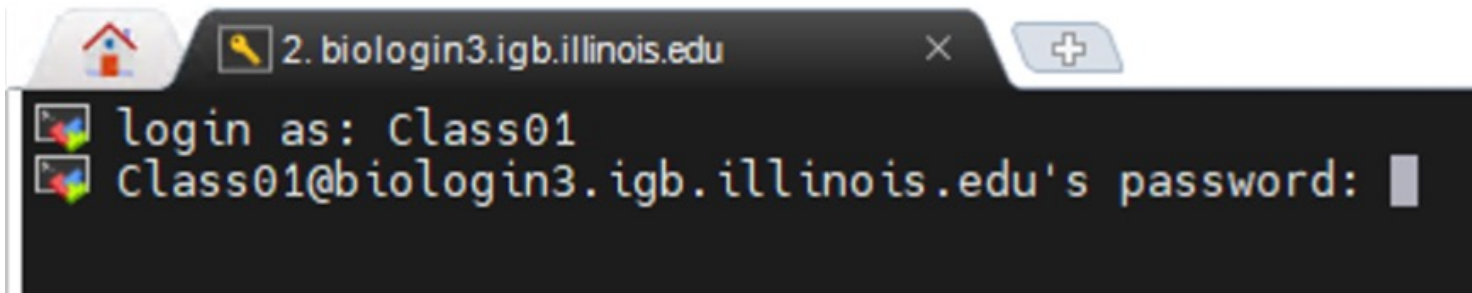
Step 0A: Accessing the IGB Biocluster for First Time

- Open **MobaXterm** from the VM
- In a new session, select **SSH** and type the following host name:
`biologin3.igb.illinois.edu`
- Click **OK**



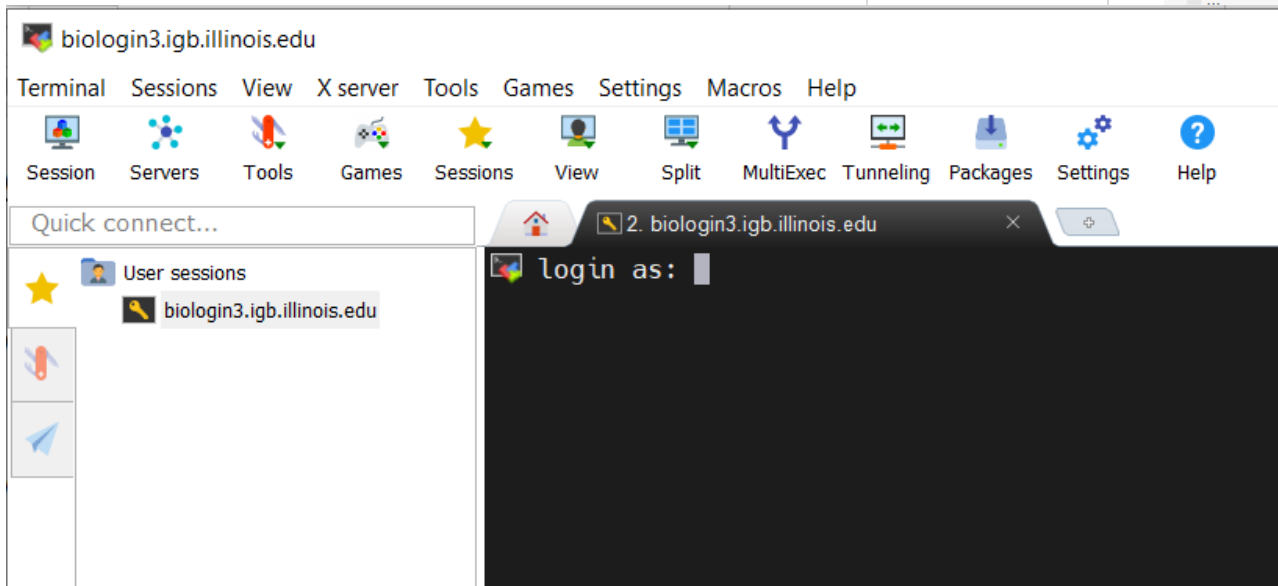
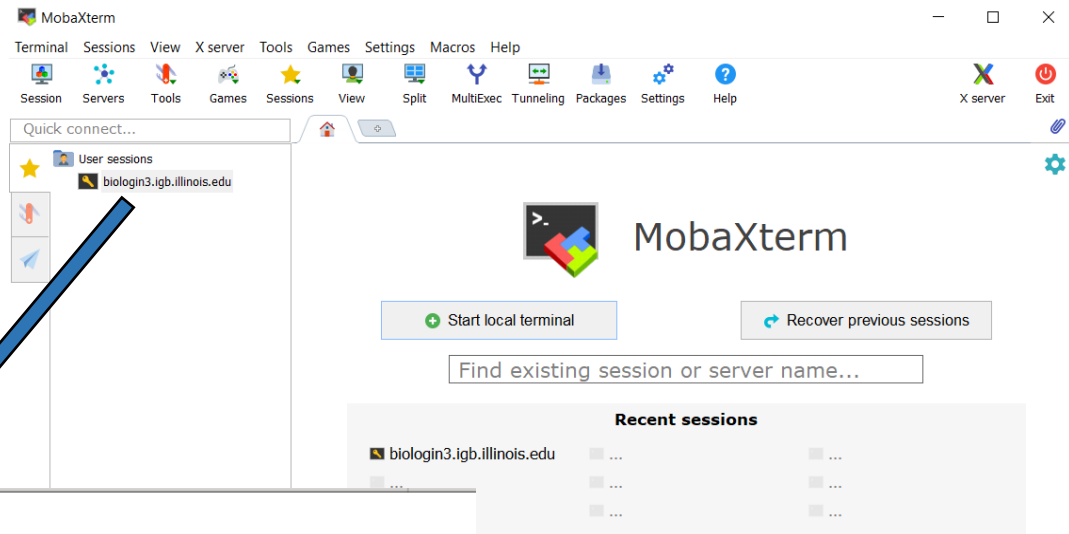
Step 0A: Accessing the IGB Biocluster

- Enter login credentials assigned to you.
- Example username: **Class01**
- You will not see any characters on screen when typing in password. Just type it.



Step 0A: Accessing the IGB Biocluster

If you have done this before, just double-click on the session you created once and type username and password.



• MobaXterm Personal Edition v23.1 •
(SSH client, X server and network tools)

- ▶ SSH session to `Class02@biologin3.igb.illinois.edu`
 - Direct SSH : ✓
 - SSH compression : ✓
 - SSH-browser : ✓
 - X11-forwarding : ✓ (remote display is forwarded through SSH)
- ▶ For more [info](#), ctrl+click on [help](#) or visit our [website](#).

```
#####  
#                                                                 #  
#           Institute for Genomic Biology                         #  
#   University of Illinois Urbana-Champaign                       #  
#   http://biocluster.igb.illinois.edu                               #  
#                                                                 #  
#####  
*Please follow the guide at http://help.igb.illinois.edu/Biocluster  
*All data on this cluster is NOT backed up. It costs $8.75 per terabyte  
per month  
*HIPAA data is not allowed on the biocluster  
*Please email help@igb.illinois.edu with any questions  
  
Last login: Wed May 31 11:26:09 2023 from 128.174.127.200  
ln: failed to create symbolic link './dropbox': File exists  
Created dropbox symbolic link  
[Class02@biologin-2 ~]$ █
```


The PED File Format

The PED File Format specifies for each individual their genotype for each SNP and their phenotype.

Family ID is either CH (Chinese) or JP (Japanese)

Paternal and Maternal IDs of 0 indicate missing.

Sex is either Male=1, Female=2, Other=Unknown

Phenotype is either 0 = missing, 1 = affected, 2 = unaffected.

Genotype 0 is used for missing genotype

Family ID	Individual ID	Paternal ID	Maternal ID	Sex	Phenotype	Genotype...
CH18526	NA18526	0	0	2	1	A A 0 G ..

The MAP File Format

The MAP File Format specifies the location of each SNP.

Note: Morgans (M) are a special kind of genetic distance derived from chromosomal recombination studies. Morgans can be used to reconstruct chromosomal maps.

chr	SNP ID	cM	Base Pair Position
8	rs17121574	12.8	12799052

Working with PLINK

In this exercise, we will analyze our data using PLINK on the command line

Additionally, we will perform a format conversion to speed up our QC analysis.

Finally, we will validate our conversion and see what individuals and SNPs would be filtered out with default filters for QC analysis.

Step 1A: Load plink

```
$ srun -p classroom -c 2 --mem 8000 --pty bash  
# Open interactive session on biocluster with 2 cpus and 8G memory.  
$ module load plink/1.07 # Load plink version 1.07 to the environment
```

Step 1B: Setting up the Directory

Copy and submit the `prep_directory.sh` job script to create a project directory with the input files.

```
# move to home directory
$ cd ~/
# copy prep_directory.sh to current directory
$ cp /home/classroom/mayo/2020/09_Variant_Analysis/prep_directory.sh ./
# submit job script
$ sbatch prep_directory.sh
```

Step 1C: Dataset Characteristics

Copy and submit the `prep_directory.sh` job script to create a project directory with the input files.

```
# move to project directory
$ cd ~/09_Variant_Analysis
# list out directory contents
$ ls
# gwas.map gwas.ped pop.cov
```

filename	meaning
gwas1.ped	Genotype data for 228,694 SNPS on 90 people.
gwas1.map	Map file for the snps in gwas1.ped.
pop.cov	Population membership of the 90 people. (1 = Han Chinese, 2 = Japanese)

Step 2A: Creating a bed file

Type in the following command to call the **PLINK** software to create a bed file

```
$ plink --file gwas1 --make-bed --out gwas2 --noweb
# --file → INPUT name
# --make-bed → operation to perform
# --out → OUTPUT name
# --noweb → tell plink not to connect to the internet
```

Step 2A: Creating a bed file

Your screen should look similar to this

```
[Class01@compute-0-1 09_Variant_Analysis]$ plink --file gwas1 --make-bed --out gwas2 --noweb
```

```
@-----@
|          PLINK!          |          v1.07          |          10/Aug/2009          |
|-----|-----|-----|
| (C) 2009 Shaun Purcell, GNU General Public License, v2 |
|-----|-----|-----|
| For documentation, citation & bug-report instructions: |
|          http://pngu.mgh.harvard.edu/purcell/plink/          |
|-----|-----|-----|
@-----@
```

```
Skipping web check... [ --noweb ]
Writing this text to log file [ gwas2.log ]
Analysis started: Sun Jun  4 14:27:25 2023
```

```
Options in effect:
  --file gwas1
  --make-bed
  --out gwas2
  --noweb
```

```
228694 (of 228694) markers to be included from [ gwas1.map ]
90 individuals read from [ gwas1.ped ]
90 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
49 cases, 41 controls and 0 missing
45 males, 45 females, and 0 of unspecified sex
Before frequency and genotyping pruning, there are 228694 SNPs
90 founders and 0 non-founders found
Total genotyping rate in remaining individuals is 0.993346
0 SNPs failed missingness test ( GENO > 1 )
0 SNPs failed frequency test ( MAF < 0 )
After frequency and genotyping pruning, there are 228694 SNPs
After filtering, 49 cases, 41 controls and 0 missing
After filtering, 45 males, 45 females, and 0 of unspecified sex
Writing pedigree information to [ gwas2.fam ]
Writing map (extended format) information to [ gwas2.bim ]
Writing genotype bitfile to [ gwas2.bed ]
Using (default) SNP-major mode
```

```
Analysis finished: Sun Jun  4 14:27:35 2023
```


Step 2B: Creating a bed file

Verify in your `09_Variant_Analysis` folder that the `gwas2` files were created

```
$ ls -lth
# -l will list out files in long format to get more information
# -t will list files in order by time (most recent files at the top of the list)
# -h will make the list human readable (e.g., 1.6K instead of 1620 for file size)
```

```
[Class01@compute-0-1 09_Variant_Analysis]$ ls -lth
total 196M
-rw-rw-r-- 1 Class01 Class01 1.7K Jun  4 14:27 gwas2.log
-rw-rw-r-- 1 Class01 Class01 5.1M Jun  4 14:27 gwas2.bed
-rw-rw-r-- 1 Class01 Class01 7.4M Jun  4 14:27 gwas2.bim
-rw-rw-r-- 1 Class01 Class01 2.2K Jun  4 14:27 gwas2.fam
-rw-rw-r-- 1 Class01 Class01 1.6K Jun  4 14:08 pop.cov
-rw-rw-r-- 1 Class01 Class01  79M Jun  4 14:08 gwas1.ped
-rw-rw-r-- 1 Class01 Class01 6.8M Jun  4 14:08 gwas1.map
[Class01@compute-0-1 09_Variant_Analysis]$
```

Step 3A: Validating the Conversion

Type in the following command to call the **PLINK** software to validate your initial output

```
$ plink --maf 0.01 --geno 0.05 --mind 0.05 --bfile gwas2 --out validate --noweb
# --maf → minor allele frequency to 0.01 (1%)
# --geno → Maximum SNP Missingness rate to 0.05 (5%)
# --mind → Maximum individual missingness rate to 0.05 (5%)
# --bfile → binary file name
# --out → output name
# --noweb → tell plink not to connect to the internet
```

Step 3A: Validating the Conversion

Your screen should look similar to this

```
[Class01@compute-0-1 09_Variant_Analysis]$ plink --maf 0.01 --geno 0.05 --mind 0.05 --bfile gwas2 --out validate --noweb
```

```
@-----@
|          PLINK!          |          v1.07          |          10/Aug/2009          |
|-----|-----|-----|
| (C) 2009 Shaun Purcell, GNU General Public License, v2 |
|-----|-----|-----|
| For documentation, citation & bug-report instructions: |
| http://pngu.mgh.harvard.edu/purcell/plink/ |
|-----|-----|-----|
@-----@
```

```
Skipping web check... [ --noweb ]
Writing this text to log file [ validate.log ]
Analysis started: Sun Jun  4 14:34:09 2023
```

```
Options in effect:
  --maf 0.01
  --geno 0.05
  --mind 0.05
  --bfile gwas2
  --out validate
  --noweb
```

```
Reading map (extended format) from [ gwas2.bim ]
228694 markers to be included from [ gwas2.bim ]
Reading pedigree information from [ gwas2.fam ]
90 individuals read from [ gwas2.fam ]
90 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
49 cases, 41 controls and 0 missing
45 males, 45 females, and 0 of unspecified sex
Reading genotype bitfile from [ gwas2.bed ]
Detected that binary PED file is v1.00 SNP-major mode
Before frequency and genotyping pruning, there are 228694 SNPs
90 founders and 0 non-founders found
Writing list of removed individuals to [ validate.irem ]
1 of 90 individuals removed for low genotyping ( MIND > 0.05 )
Total genotyping rate in remaining individuals is 0.995473
2728 SNPs failed missingness test ( GENO > 0.05 )
46834 SNPs failed frequency test ( MAF < 0.01 )
After frequency and genotyping pruning, there are 179562 SNPs
After filtering, 48 cases, 41 controls and 0 missing
After filtering, 44 males, 45 females, and 0 of unspecified sex
```

```
Analysis finished: Sun Jun  4 14:34:12 2023   Polymorphism and Variant Analysis | Saba Ghaffari | 2023
```

Step 3B: Validating the Conversion

Verify in your **09_Variant_Analysis** folder that the **validate** files were created

```
$ ls -lth
```

```
[Class01@compute-0-1 09_Variant_Analysis]$ ls -lth
total 196M
-rw-rw-r-- 1 Class01 Class01 1.8K Jun  4 14:34 validate.log
-rw-rw-r-- 1 Class01 Class01  16 Jun  4 14:34 validate.irem
-rw-rw-r-- 1 Class01 Class01 1.7K Jun  4 14:27 gwas2.log
-rw-rw-r-- 1 Class01 Class01 5.1M Jun  4 14:27 gwas2.bed
-rw-rw-r-- 1 Class01 Class01 7.4M Jun  4 14:27 gwas2.bim
-rw-rw-r-- 1 Class01 Class01 2.2K Jun  4 14:27 gwas2.fam
-rw-rw-r-- 1 Class01 Class01 1.6K Jun  4 14:08 pop.cov
-rw-rw-r-- 1 Class01 Class01 79M Jun  4 14:08 gwas1.ped
-rw-rw-r-- 1 Class01 Class01 6.8M Jun  4 14:08 gwas1.map
[Class01@compute-0-1 09_Variant_Analysis]$ █
```

Step 3C: Viewing Validation

```
[Class01@compute-0-1 09_Variant_Analysis]$ plink --maf 0.01 --geno 0.05 --mind 0.05 --bfile gwas2 --out validate --noweb
```

```
@-----@
|          PLINK!          |          v1.07          |          10/Aug/2009          |
|-----|-----|-----|
| (C) 2009 Shaun Purcell, GNU General Public License, v2 |
|-----|-----|-----|
| For documentation, citation & bug-report instructions: |
| http://pngu.mgh.harvard.edu/purcell/plink/ |
|-----|-----|-----|
@-----@
```

```
Skipping web check... [ --noweb ]
Writing this text to log file [ validate.log ]
Analysis started: Sun Jun  4 14:34:09 2023
```

```
Options in effect:
  --maf 0.01
  --geno 0.05
  --mind 0.05
  --bfile gwas2
  --out validate
  --noweb
```

```
Reading map (extended format) from [ gwas2.bim ]
228694 markers to be included from [ gwas2.bim ]
Reading pedigree information from [ gwas2.fam ]
90 individuals read from [ gwas2.fam ]
90 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
49 cases, 41 controls and 0 missing
45 males, 45 females, and 0 of unspecified sex
Reading genotype bitfile from [ gwas2.bed ]
Detected that binary PED file is v1.00 SNP-major mode
Before frequency and genotyping pruning, there are 228694 SNPs
90 founders and 0 non-founders found
Writing list of removed individuals to [ validate.irem ]
1 of 90 individuals removed for low genotyping ( MIND > 0.05 )
Total genotyping rate in remaining individuals is 0.995473
2728 SNPs failed missingness test ( GENO > 0.05 )
46834 SNPs failed frequency test ( MAF < 0.01 )
After frequency and genotyping pruning, there are 179562 SNPs
After filtering, 48 cases, 41 controls and 0 missing
After filtering, 44 males, 45 females, and 0 of unspecified sex
```

```
Analysis finished: Sun Jun  4 14:34:12 2023   Polymorphism and Variant Analysis | Saba Ghaffari | 2023
```

46834 out of ~ 230,000 SNPs were removed because they failed the MAF.

2728 SNPs were removed because they were not genotyped in enough individuals (minimum, 95%).

1 of 90 individuals removed for low genotyping (MIND > 0.05)

Step 3D: Validating the Conversion

The **validate.irem** file is small enough to print to the console, so we can use **cat** to view it

```
$ cat validate.irem  
# JA19012 NA19012
```

The family id is JA19012 (Japanese) and the individual ID is NA19012. This individual was removed because of low genotyping quality.

Quality Control Analysis

In this exercise, we will perform Quality Control Analysis (QC) to filter our data according to a set of criteria.

Quality Control Filters

The validation tool will impose the following criteria on our data.

filter	meaning	threshold
Minor Allele Frequency (MAF)	The proportion of the minor allele to the major allele of a SNP in the population must exceed this threshold for the SNP to be included in the analysis	1%
Individual Genotyping rate	The number of SNPs probed for an individual must exceed this threshold for the person to be analyzed.	95%
SNP genotyping rate	The SNP must be probed for at least this many individuals.	95%

Step 4A: Quality Control Analysis

Type in the following command to call the **PLINK** software to perform the Quality Control (QC) analysis

```
$ plink --maf 0.01 --geno 0.05 --mind 0.05 --bfile gwas2 --make-bed --out gwas3 --noweb
# --maf → minor allele frequency to 0.01 (1%)
# --geno → Maximum SNP Missingness rate to 0.05 (5%)
# --mind → Maximum individual missingness rate to 0.05 (5%)
# --bfile → binary file name
# --make-bed → operation to perform
# --out → output name
# --noweb → tell plink not to connect to the internet
```

Step 4A: Quality Control Analysis

Your screen should look similar to this

```
[Class01@compute-0-1 09_Variant_Analysis]$ plink --maf 0.01 --geno 0.05 --mind 0.05 --bfile gwas2 --make-bed --out gwas3 --noweb
```

```
@-----@
|          PLINK!          |          v1.07          |          10/Aug/2009          |
|-----|-----|-----|
| (C) 2009 Shaun Purcell, GNU General Public License, v2 |
|-----|-----|-----|
| For documentation, citation & bug-report instructions: |
| http://pngu.mgh.harvard.edu/purcell/plink/ |
|-----|-----|-----|
@-----@
```

```
Skipping web check... [ --noweb ]
Writing this text to log file [ gwas3.log ]
Analysis started: Sun Jun  4 14:40:32 2023
```

```
Options in effect:
  --maf 0.01
  --geno 0.05
  --mind 0.05
  --bfile gwas2
  --make-bed
  --out gwas3
  --noweb
```

```
Reading map (extended format) from [ gwas2.bim ]
228694 markers to be included from [ gwas2.bim ]
Reading pedigree information from [ gwas2.fam ]
90 individuals read from [ gwas2.fam ]
90 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
49 cases, 41 controls and 0 missing
45 males, 45 females, and 0 of unspecified sex
Reading genotype bitfile from [ gwas2.bed ]
Detected that binary PED file is v1.00 SNP-major mode
Before frequency and genotyping pruning, there are 228694 SNPs
90 founders and 0 non-founders found
Writing list of removed individuals to [ gwas3.irem ]
1 of 90 individuals removed for low genotyping ( MIND > 0.05 )
Total genotyping rate in remaining individuals is 0.995473
2728 SNPs failed missingness test ( GENO > 0.05 )
46834 SNPs failed frequency test ( MAF < 0.01 )
After frequency and genotyping pruning, there are 179562 SNPs
After filtering, 48 cases, 41 controls and 0 missing
After filtering, 44 males, 45 females, and 0 of unspecified sex
Writing pedigree information to [ gwas3.fam ]
Writing map (extended format) information to [ gwas3.bim ]
Writing genotype bitfile to [ gwas3.bed ]
Using (default) SNP-major mode
```

```
Analysis finished: Sun Jun  4 14:40:35 2023
```

Step 4B: Quality Control Analysis

Verify in your `09_Variant_Analysis` folder that the `gwas3` files were created

```
$ ls -lth
```

```
[Class01@compute-0-1 09_Variant_Analysis]$ ls -lth
total 215M
-rw-rw-r-- 1 Class01 Class01 2.0K Jun  4 14:40 gwas3.log
-rw-rw-r-- 1 Class01 Class01 4.0M Jun  4 14:40 gwas3.bed
-rw-rw-r-- 1 Class01 Class01 5.8M Jun  4 14:40 gwas3.bim
-rw-rw-r-- 1 Class01 Class01 2.1K Jun  4 14:40 gwas3.fam
-rw-rw-r-- 1 Class01 Class01  16 Jun  4 14:40 gwas3.irem
-rw-rw-r-- 1 Class01 Class01 1.8K Jun  4 14:34 validate.log
-rw-rw-r-- 1 Class01 Class01  16 Jun  4 14:34 validate.irem
-rw-rw-r-- 1 Class01 Class01 1.7K Jun  4 14:27 gwas2.log
-rw-rw-r-- 1 Class01 Class01 5.1M Jun  4 14:27 gwas2.bed
-rw-rw-r-- 1 Class01 Class01 7.4M Jun  4 14:27 gwas2.bim
-rw-rw-r-- 1 Class01 Class01 2.2K Jun  4 14:27 gwas2.fam
-rw-rw-r-- 1 Class01 Class01 1.6K Jun  4 14:08 pop.cov
-rw-rw-r-- 1 Class01 Class01 79M Jun  4 14:08 gwas1.ped
-rw-rw-r-- 1 Class01 Class01 6.8M Jun  4 14:08 gwas1.map
```

Genome-Wide Association Test (GWAS)

In this exercise, we will perform a GWAS on our filtered data across two phenotypes: a case study and control. We will then compare the results between unadjusted p-values and multiple hypothesis corrected p-values.

Step 5A: GWAS

Type in the following command to call the **PLINK** software to test for associations and adjust for multiple testing

```
$ plink --bfile gwas3 --assoc --adjust --out assoc1 --noweb
# --bfile → binary file name
# --assoc → operation to perform, here association testing)
# --adjust → operation to perform, here adjust p-values due to multiple
testing
# --out → output name
# --noweb → tell plink not to connect to the internet
```

Step 5A: GWAS

Your screen should look similar to this

```
[Class01@compute-0-1 09_Variant_Analysis]$ plink --bfile gwas3 --assoc --adjust --out assoc1 --noweb
```

```
@-----@
|          PLINK!          |          v1.07          |          10/Aug/2009          |
|-----|-----|-----|
| (C) 2009 Shaun Purcell, GNU General Public License, v2 |
|-----|-----|-----|
| For documentation, citation & bug-report instructions: |
| http://pngu.mgh.harvard.edu/purcell/plink/ |
|-----|-----|-----|
@-----@
```

```
Skipping web check... [ --noweb ]
Writing this text to log file [ assoc1.log ]
Analysis started: Sun Jun  4 14:50:58 2023
```

```
Options in effect:
  --bfile gwas3
  --assoc
  --adjust
  --out assoc1
  --noweb
```

```
Reading map (extended format) from [ gwas3.bim ]
179562 markers to be included from [ gwas3.bim ]
Reading pedigree information from [ gwas3.fam ]
89 individuals read from [ gwas3.fam ]
89 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
48 cases, 41 controls and 0 missing
44 males, 45 females, and 0 of unspecified sex
Reading genotype bitfile from [ gwas3.bed ]
Detected that binary PED file is v1.00 SNP-major mode
Before frequency and genotyping pruning, there are 179562 SNPs
89 founders and 0 non-founders found
Total genotyping rate in remaining individuals is 0.996307
0 SNPs failed missingness test ( GENO > 1 )
0 SNPs failed frequency test ( MAF < 0 )
After frequency and genotyping pruning, there are 179562 SNPs
After filtering, 48 cases, 41 controls and 0 missing
After filtering, 44 males, 45 females, and 0 of unspecified sex
Writing main association results to [ assoc1.assoc ]
Computing corrected significance values (FDR, Sidak, etc)
Genomic inflation factor (based on median chi-squared) is 1.25937
Mean chi-squared statistic is 1.2297
Correcting for 179562 tests
Writing multiple-test corrected significance values to [ assoc1.assoc.adjusted ]
```

```
Analysis finished: Sun Jun  4 14:51:01 2023
```

Step 5B: GWAS

Verify in your `09_Variant_Analysis` folder that the `assoc1` files were created

```
$ ls -lth
```

```
[Class01@compute-0-1 09_Variant_Analysis]$ ls -lth
total 284M
-rw-rw-r-- 1 Class01 Class01 2.0K Jun  4 14:51 assoc1.log
-rw-rw-r-- 1 Class01 Class01  19M Jun  4 14:51 assoc1.assoc.adjusted
-rw-rw-r-- 1 Class01 Class01  17M Jun  4 14:51 assoc1.assoc
-rw-rw-r-- 1 Class01 Class01 2.0K Jun  4 14:40 gwas3.log
-rw-rw-r-- 1 Class01 Class01 4.0M Jun  4 14:40 gwas3.bed
-rw-rw-r-- 1 Class01 Class01 5.8M Jun  4 14:40 gwas3.bim
-rw-rw-r-- 1 Class01 Class01 2.1K Jun  4 14:40 gwas3.fam
-rw-rw-r-- 1 Class01 Class01  16 Jun  4 14:40 gwas3.irem
-rw-rw-r-- 1 Class01 Class01 1.8K Jun  4 14:34 validate.log
-rw-rw-r-- 1 Class01 Class01  16 Jun  4 14:34 validate.irem
-rw-rw-r-- 1 Class01 Class01 1.7K Jun  4 14:27 gwas2.log
-rw-rw-r-- 1 Class01 Class01 5.1M Jun  4 14:27 gwas2.bed
-rw-rw-r-- 1 Class01 Class01 7.4M Jun  4 14:27 gwas2.bim
-rw-rw-r-- 1 Class01 Class01 2.2K Jun  4 14:27 gwas2.fam
-rw-rw-r-- 1 Class01 Class01 1.6K Jun  4 14:08 pop.cov
-rw-rw-r-- 1 Class01 Class01  79M Jun  4 14:08 gwas1.ped
-rw-rw-r-- 1 Class01 Class01 6.8M Jun  4 14:08 gwas1.map
[Class01@compute-0-1 09_Variant_Analysis]$
```

Step 6: GWAS Without Multiple Hypothesis Correction

The SNP p-values from our GWAS with no multiple hypothesis correction are located in the 9th column of **assoc1.assoc**.

Overall, 13,294 SNPs survive at p-value of 0.05 WITHOUT Multiple Hypothesis Correction

Here we will use the UNIX tool **awk** to count the number of lines/rows where the 9th column is equal or less than 0.05. At the end of the file, we then print our variable *count*, which was automatically created for us

```
$ awk '$9 <= 0.05 {count+=1} END {print count}' assoc1.assoc  
# 13294
```


Step 6: GWAS Without Multiple Hypothesis Correction

The top few SNPs are shown below after using redirecting the result of `sort` by using the "|" character to the `head` command

```
$ sort -g -k9,9 assoc1.assoc | head
```

```
# -g tells the sort command to sort using generic numeric sorting
```

```
# -k tells the sort command to set the 9th column as the key column to sort on
```

```
# head will print the first 10 lines by default
```

```
[Class01@compute-0-1 09_Variant_Analysis]$ sort -g -k9,9 assoc1.assoc | head
```

CHR	SNP	BP	A1	F_A	F_U	A2	CHISQ	P	OR
11	rs2513514	75922141	A	0.5208	0.1585	G	25.39	4.693e-07	5.769
20	rs6110115	13911728	C	0.3085	0.6829	A	24.59	7.103e-07	0.2071
11	rs2508756	75921549	A	0.5417	0.1951	G	22.5	2.105e-06	4.875
15	rs16976702	54120691	G	0.5833	0.2317	C	22.43	2.183e-06	4.642
8	rs11204005	12895576	A	0.3229	0.6585	G	19.97	7.882e-06	0.2473
9	rs16910850	94478347	T	0.09375	0.3659	C	19.14	1.216e-05	0.1793
12	rs1195747	129970575	A	0.3085	0.6375	G	18.83	1.427e-05	0.2537
17	rs7207095	77933018	G	0.5208	0.2073	A	18.52	1.682e-05	4.156
15	rs16971118	77672467	C	0.3936	0.1098	T	18.28	1.907e-05	5.265

Step 7: GWAS With Multiple Hypothesis Correction

The SNP p values from our GWAS with multiple hypothesis correction are located in the 9th column of `assoc1.assoc.adjusted`.

We will use `awk` to print every line (`$0`) where the p-value is less than or equal to 0.1

Overall, only 4 SNPS!!! show a FDR Correction of less than 0.1

```
$ awk '$9 <= 0.1 {print $0}' assoc1.assoc.adjusted
```

```
[Class01@biologin-2 09_Variant_Analysis]$ awk '$9 <= 0.1 {print $0}' assoc1.assoc.adjusted
 11  rs2513514 4.693e-07 7.131e-06 0.08427 0.08427 0.08081 0.08081 0.06378 0.8084
 20  rs6110115 7.103e-07 9.938e-06 0.1276 0.1275 0.1198 0.1198 0.06378 0.8084
 11  rs2508756 2.105e-06 2.373e-05 0.378 0.3779 0.3147 0.3147 0.098 1
 15  rs16976702 2.183e-06 2.443e-05 0.392 0.392 0.3243 0.3243 0.098 1
```

Step 8: GWAS Without Multiple Hypothesis Correction

Exit MobaXterm by either of the following:

- Close the window
- Type **exit** in the command line twice and then press <return>

```
$ exit # first to exit from compute node
```

```
$ exit # again to exit from login node
```

```
<RETURN>
```

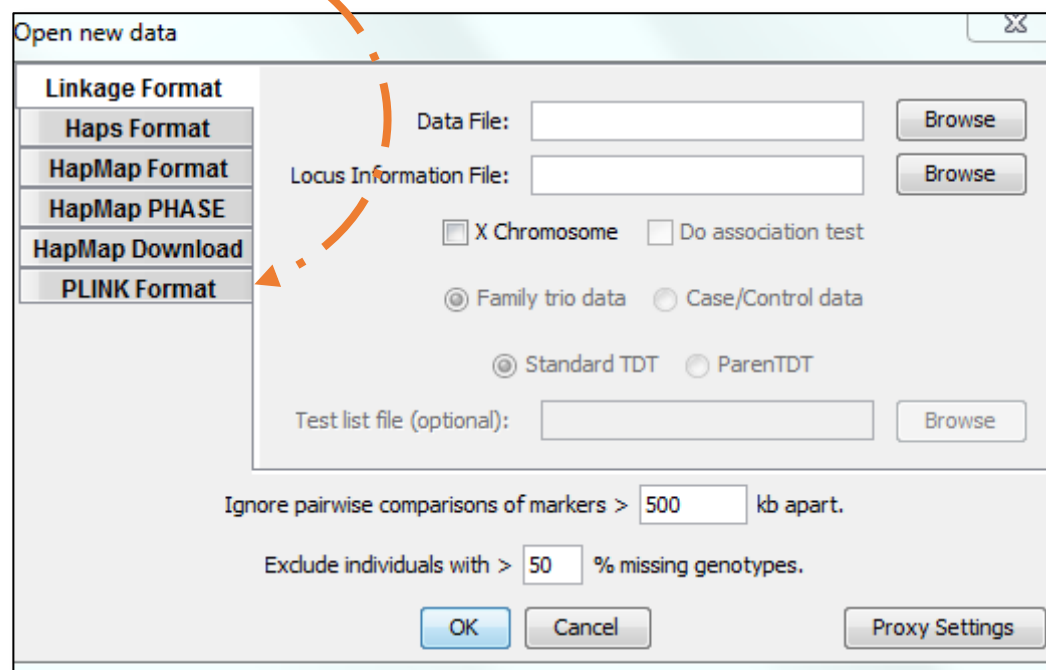
Visualization

In this exercise, we will generate a Manhattan Plot of our association results using **Haploview** from the **Broad Institute**.

Step 9A: Configuring Haploview

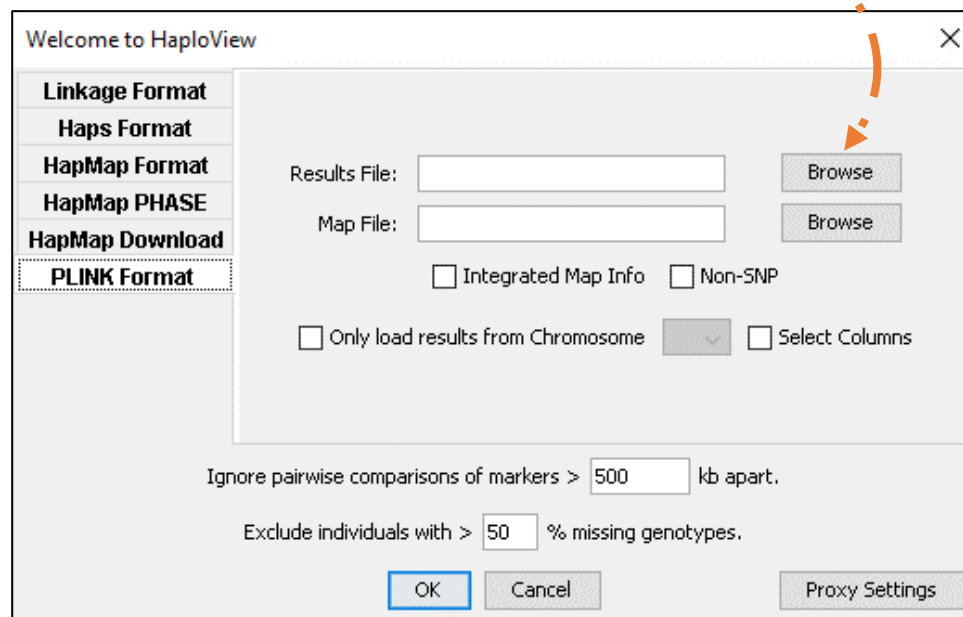
Open **Haploview** from VM.

Click **PLINK Format**



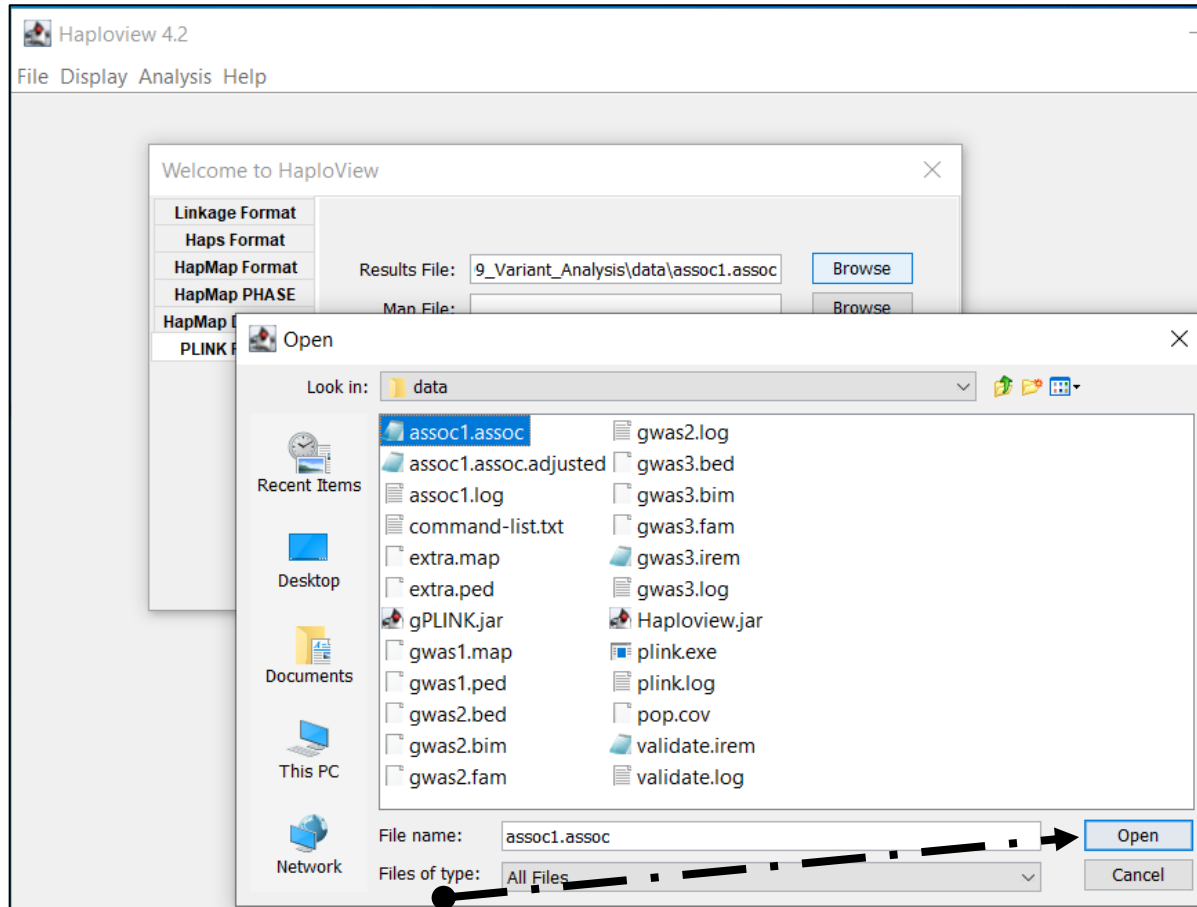
Step 9B: Configuring Haploview

Click on **Browse** next to **Results File**:



Step 9C: Configuring Haploview

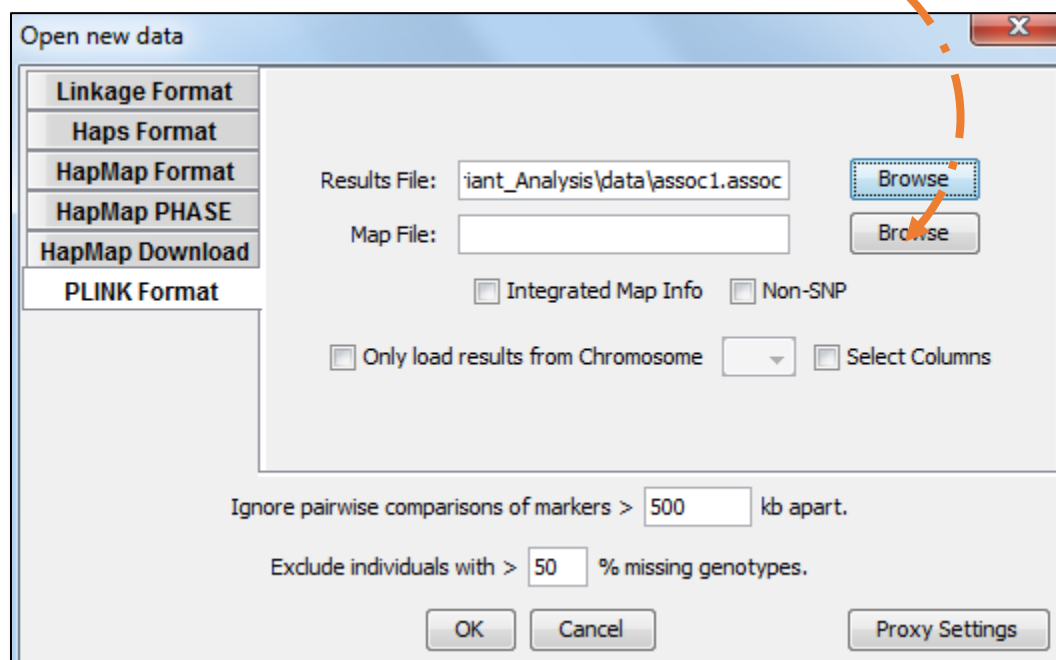
Navigate to the directory **PLINK** saved the file **assoc1.assoc**. It should be saved in the **data** sub folder in the **09_Variant_Analysis** folder



Select **assoc1.assoc** and click **Open**.

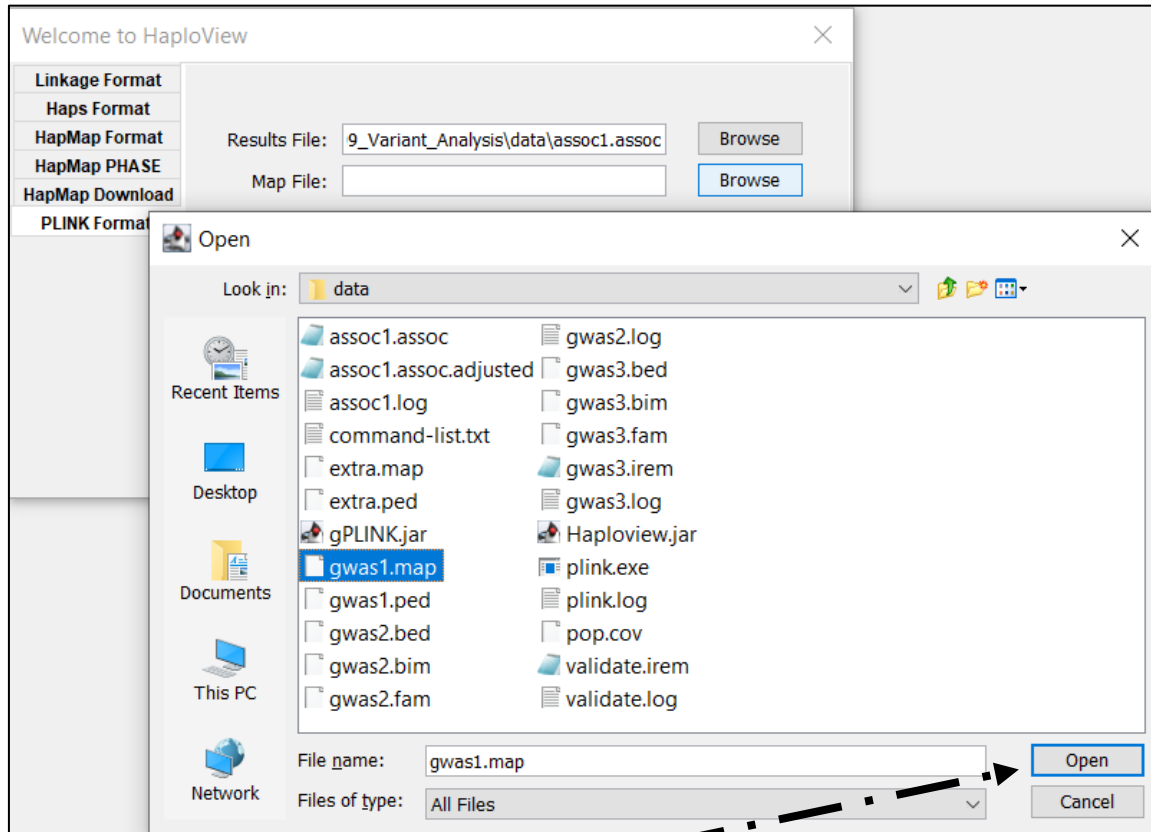
Step 9D: Configuring Haploview

Click on **Browse** next to **Map File**:



Step 9E: Configuring Haploview

Navigate to the data directory containing **gwas1.map**



Select **gwas1.map** and click **Open**.

Step 9F: Configuring Haploview

Click on OK.

Welcome to HaploView

Linkage Format
Haps Format
HapMap Format
HapMap PHASE
HapMap Download
PLINK Format

Results File: 9_Variant_Analysis\data\assoc1.assoc Browse

Map File: \09_Variant_Analysis\data\gwas1.map Browse

Integrated Map Info Non-SNP

Only load results from Chromosome Select Columns

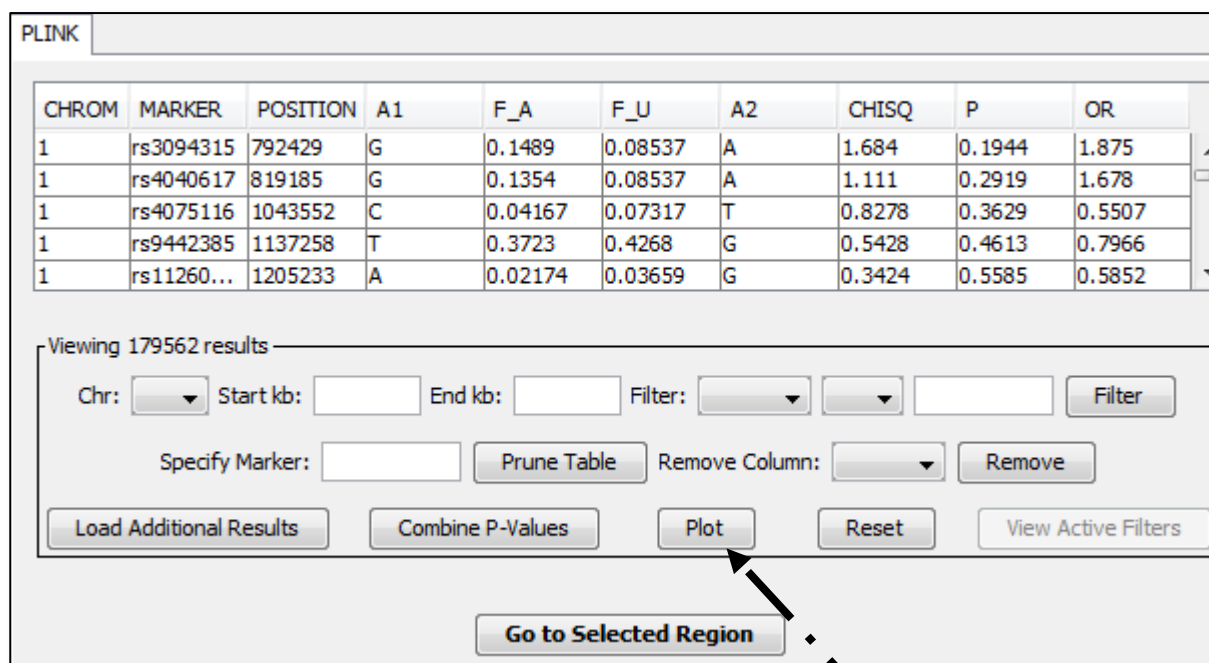
Ignore pairwise comparisons of markers > 500 kb apart.

Exclude individuals with > 50 % missing genotypes.

OK Cancel Proxy Settings

Step 9G: Configuring Haploview

Your **assoc1** should be shown in **Haploview** in tabular format.



The screenshot displays the PLINK web interface. At the top, there is a tab labeled "PLINK". Below it is a table with the following columns: CHROM, MARKER, POSITION, A1, F_A, F_U, A2, CHISQ, P, and OR. The table contains five rows of data for chromosome 1. Below the table, there is a control panel with the text "Viewing 179562 results". The control panel includes several input fields and buttons: "Chr:" with a dropdown menu, "Start kb:" and "End kb:" with text input fields, "Filter:" with a dropdown menu and a "Filter" button, "Specify Marker:" with a text input field, "Prune Table" button, "Remove Column:" with a dropdown menu and a "Remove" button, "Load Additional Results" button, "Combine P-Values" button, "Plot" button (highlighted with a black arrow), "Reset" button, and "View Active Filters" button. At the bottom of the control panel is a "Go to Selected Region" button.

CHROM	MARKER	POSITION	A1	F_A	F_U	A2	CHISQ	P	OR
1	rs3094315	792429	G	0.1489	0.08537	A	1.684	0.1944	1.875
1	rs4040617	819185	G	0.1354	0.08537	A	1.111	0.2919	1.678
1	rs4075116	1043552	C	0.04167	0.07317	T	0.8278	0.3629	0.5507
1	rs9442385	1137258	T	0.3723	0.4268	G	0.5428	0.4613	0.7966
1	rs11260...	1205233	A	0.02174	0.03659	G	0.3424	0.5585	0.5852

To create a **Manhattan Plot**, click **Plot**

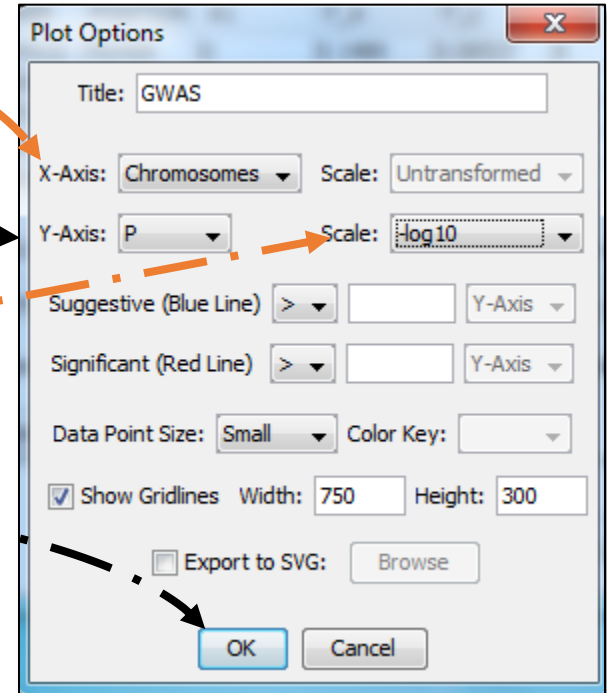
Step 9H: Configuring Haploview

Select **Chromosomes** for X-Axis

Select **P** for Y-Axis

Select **$-\log_{10}$** for Y-Axis Scale

Click **OK**



Step 10: Manhattan Plot

Haploview then should generate the following **Manhattan Plot**

