Polymorphism and Variant Analysis Lab

Matt Hudson

PowerPoint by Casey Hanson
Edited by Brianna Bucknor &
Giovanni Madrigal

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/2022-schedule/

Step 0: Local Files

For viewing and manipulating the files needed for this laboratory exercise, the path on the VM will be denoted as the following:

[course_directory]

We will use the files found in:

[course_directory]\09_Variant_Analysis\data

[course_directory]= Desktop\Labs **UIUC**

[course_directory]= Desktop\VM **Mayo**

Dataset Characteristics

filename	meaning
plink.exe	An executable of the PLINK GWAS toolkit. (Preinstalled)
Haploview.jar	A haplotype analysis program written in JAVA. Used to view PLINK results and SNP analysis.
wgas1.ped	Genotype data for 228,694 SNPS on 90 people.
wgas1.map	Map file for the snps in wgas1.ped.
extra.ped	Genotype data for 29 SNPS on the same 90 people.
extra.map	Map file for the SNPS in extra.ped.
pop.cov	Population membership of the 90 people. (1 = Han Chinese, 2 = Japanese)

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 prompt

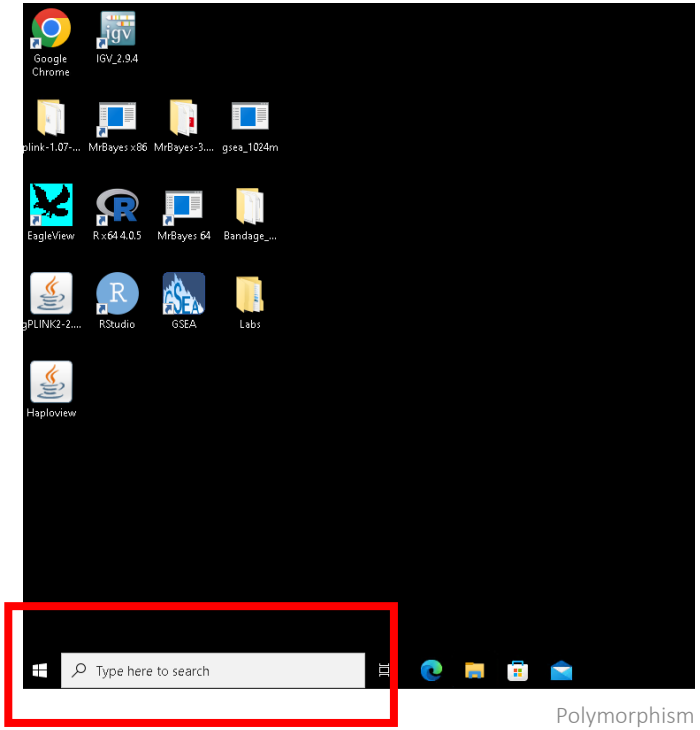
Additionally, we will perform a format conversion to speed up our QC analysis.

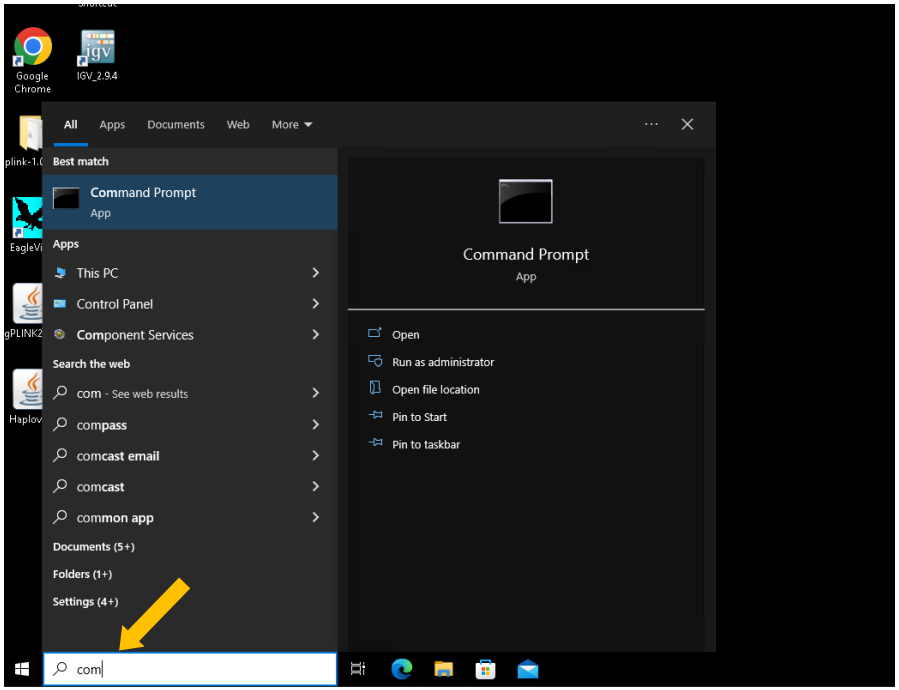
Finally, we will validate our conversion and see what individuals and SNPs would be filtered out with default filters for QC analysis.

Step 1A: Starting the Command Prompt

The **command prompt** is a program that let's us run **PLINK** directly without using additional tools

To start the **command prompt window,** navigate to the search bar at the bottom of the screen and search for the command prompt.





Step 1A: Setting up the Directory

A window should appear similar to the one below:

Step 1B: Setting up the Directory

Command prompt (do not type)

Type in the following command to head to where the data is located. Use TAB to autocomplete. Make sure to use the correct course directory

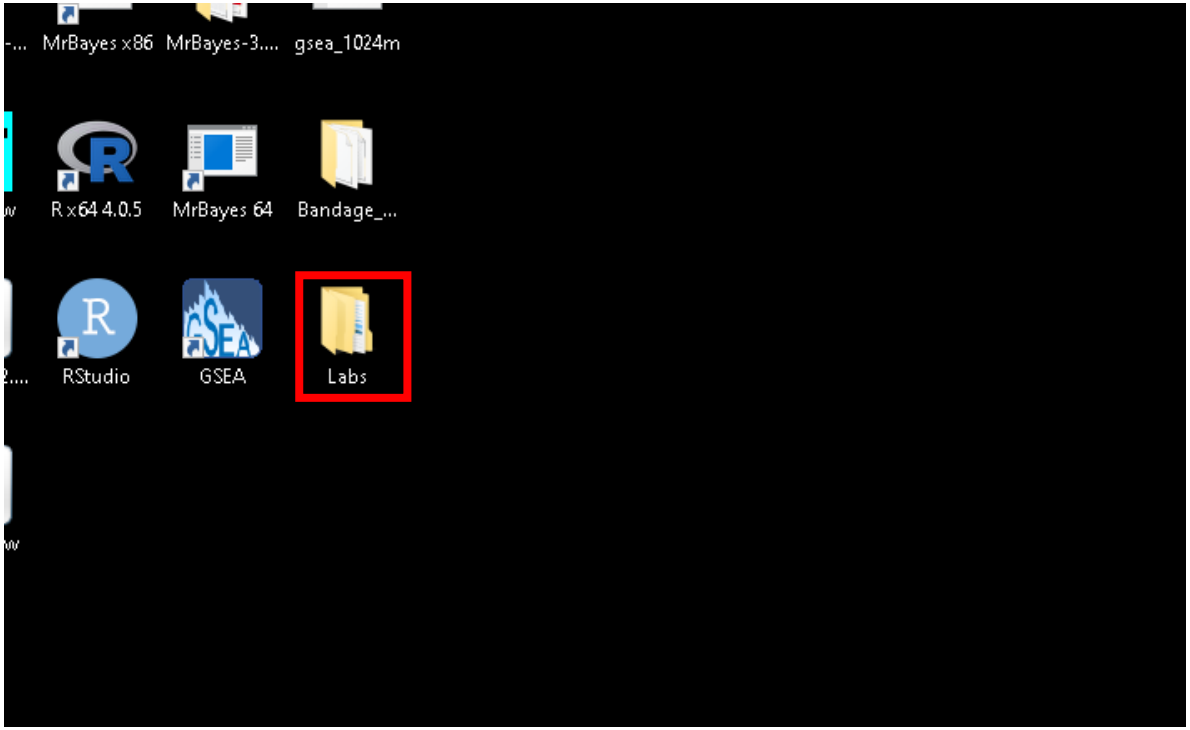
```
> cd Desktop\Labs\09_Variant_Analysis\data # use this if you are UIUC

> cd Desktop\VM\09_Variant_Analysis\data # use this if you are Mayo

# this is a comment (DO NOT TYPE)

# cd = change directory

# example shown below. Note that on windows, folders are separated by "\"
instead of "/"
```

```
C:\Users\IGB>cd Desktop\Labs\09_Variant_Analysis\data
```

Typing begins here

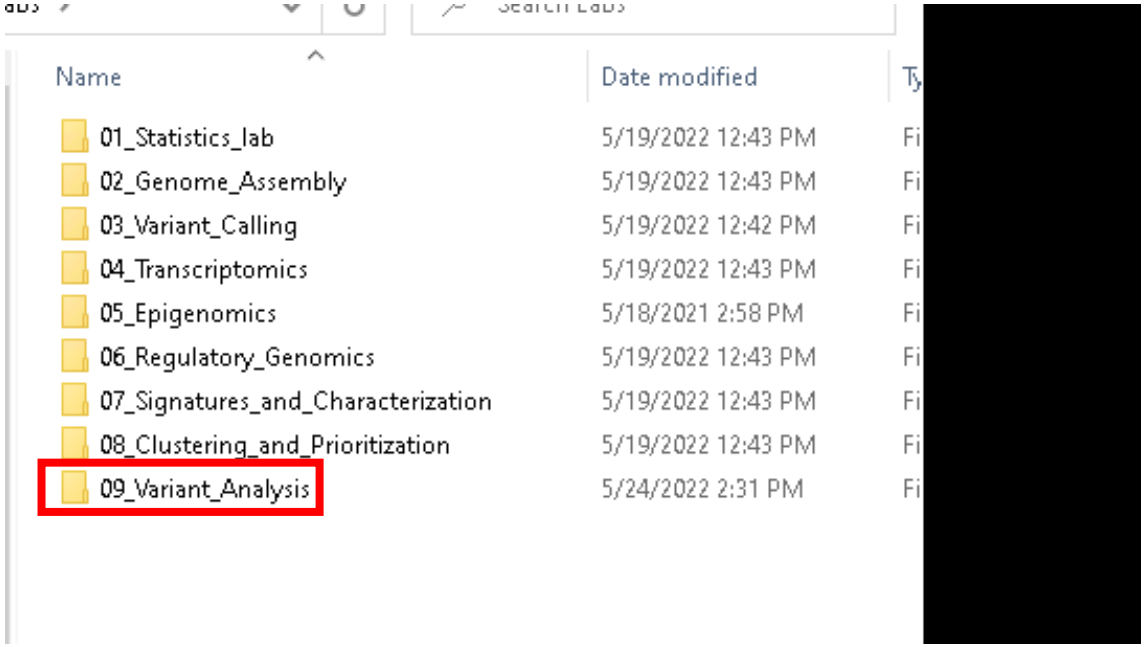
Step 1C: Setting up the Directory

To verify that you are in the **data** folder, select the **Labs** folder located in the desktop (select **VM** if you are Mayo)



Step 1D: Setting up the Directory

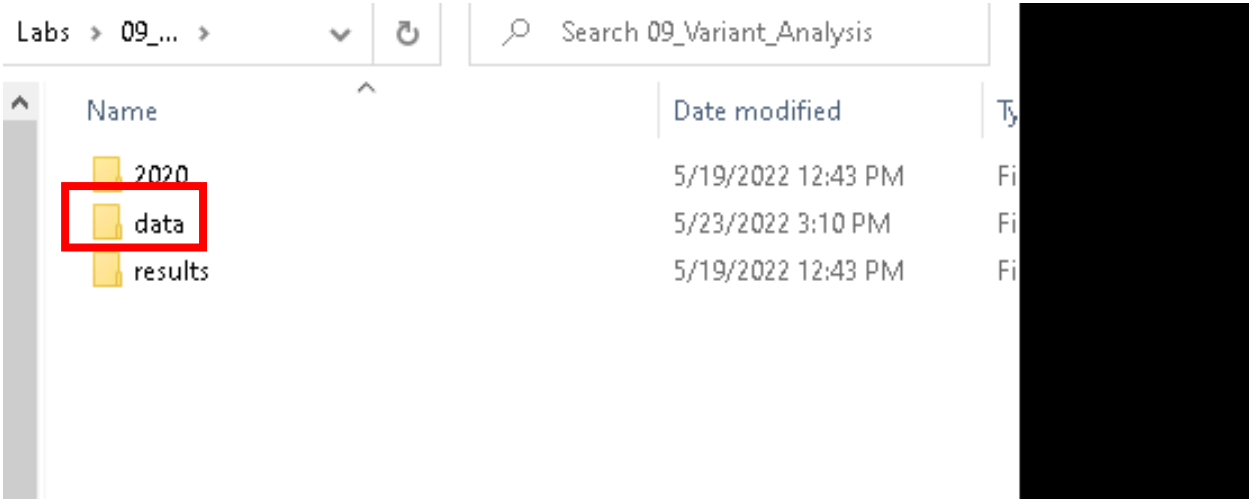
Open the **09_Variant_Analysis** folder



Name	Date modified
01_Statistics_lab	5/19/2022 12:43 PM
02_Genome_Assembly	5/19/2022 12:43 PM
03_Variant_Calling	5/19/2022 12:42 PM
04_Transcriptomics	5/19/2022 12:43 PM
05_Epigenomics	5/18/2021 2:58 PM
06_Regulatory_Genomics	5/19/2022 12:43 PM
07_Signatures_and_Characterization	5/19/2022 12:43 PM
08_Clustering_and_Prioritization	5/19/2022 12:43 PM
09_Variant_Analysis	5/24/2022 2:31 PM

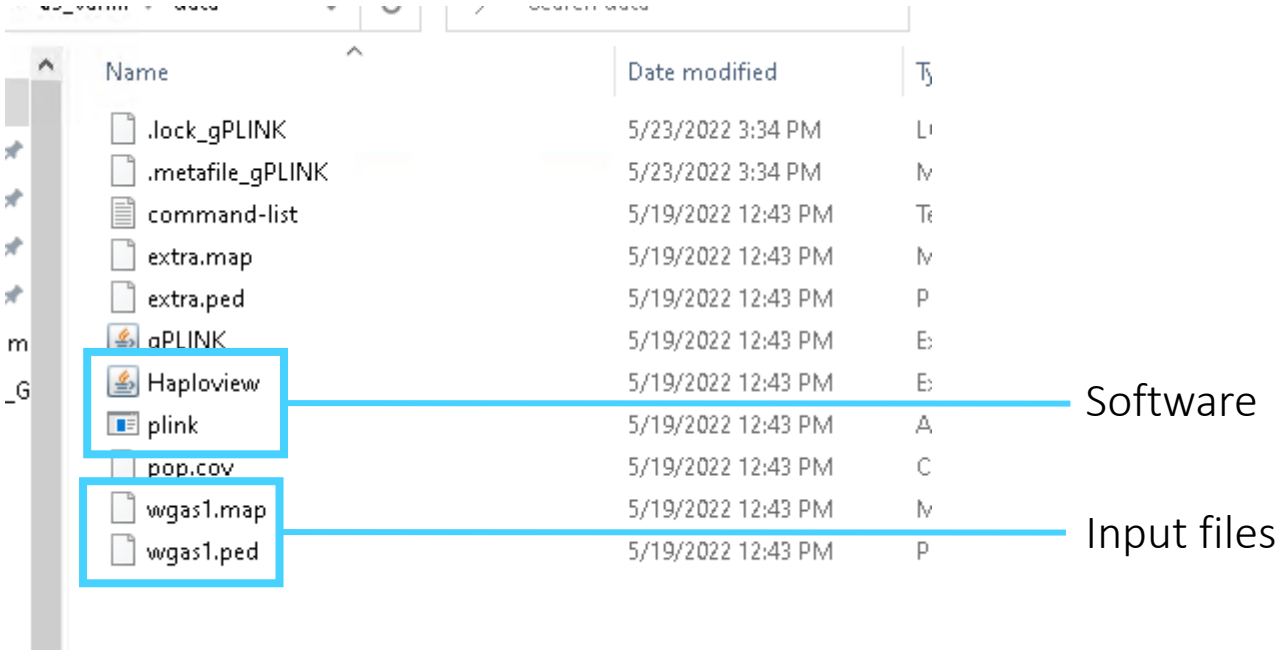
Step 1E: Setting up the Directory

Next, enter the **data** directory



Step 1F: Setting up the Directory

This directory will contain the input and output files for several analyzes in this lab. Note* you will not be using every file shown in the image below



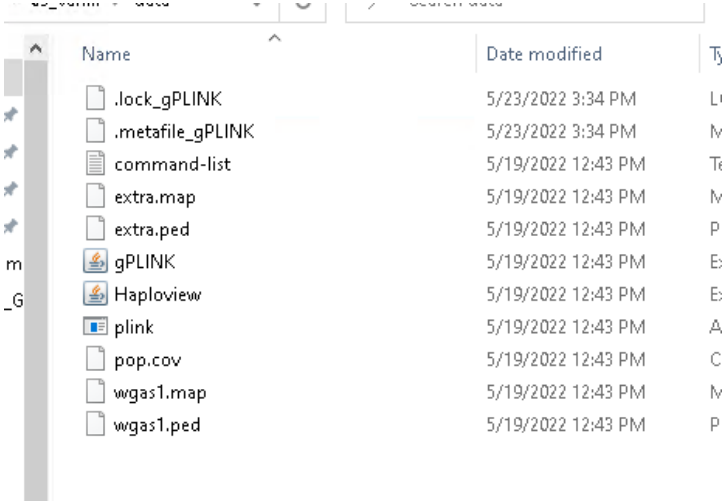
Step 1G: Setting up the Directory

For one last check, type in the following command to list out the contents of your directory. It should match with what I seen with the **data** folder open

Command prompt (do not type)

```
> dir

# this is a comment (DO NOT TYPE)

# dir is the list command in windows
```

```
C:\Users\IGB\Desktop\Labs\09_Variant_Analysis\data>dir
 Volume in drive C has no label.
 Volume Serial Number is 661F-5882

 Directory of C:\Users\IGB\Desktop\Labs\09_Variant_Analysis\data

05/23/2022  03:10 PM    <DIR>          .
05/23/2022  03:10 PM    <DIR>          ..
05/23/2022  03:34 PM                 2 .lock_gPLINK
05/23/2022  03:34 PM               141 .metafile_gPLINK
05/19/2022  12:43 PM             2,878 command-list.txt
05/19/2022  12:43 PM               509 extra.map
05/19/2022  12:43 PM             8,277 extra.ped
05/19/2022  12:43 PM         1,799,865 gPLINK.jar
05/19/2022  12:43 PM         5,300,252 Haploview.jar
05/19/2022  12:43 PM         3,965,571 plink.exe
05/19/2022  12:43 PM             1,626 pop.cov
05/19/2022  12:43 PM         7,033,003 wgas1.map
05/19/2022  12:43 PM        82,332,096 wgas1.ped
              11 File(s)    100,444,203 bytes
               2 Dir(s)  43,562,508,288 bytes free
```



Step 2A: Creating a Binary Input File

Command prompt (do not type)

Type in the following command to call the **PLINK** software to create a binary file to speed up downstream analyzes

```
> plink.exe --file wgas1 --make-bed --out wgas2

# plink.exe is the software

# --file → INPUT

# --make-bed (operation to perform)

# --out → Output name
```

Step 2A: Creating a Binary Input File

Your screen should look similar to this

```
C:\Users\IGB\Desktop\Labs\09_Variant_Analysis\data>plink.exe --file wgas1 --make-bed --out wgas2

@----------------------------------------------------------@
|        PLINK!       |     v1.02      |   25/May/2008     |
|----------------------------------------------------------|
|  (C) 2008 Shaun Purcell, GNU General Public License, v2  |
|----------------------------------------------------------|
|  For documentation, citation & bug-report instructions:  |
|        http://pngu.mgh.harvard.edu/purcell/plink/        |
@----------------------------------------------------------@

Web-check not implemented on this system...
Writing this text to log file [ wgas2.log ]
Analysis started: Tue May 24 14:51:35 2022

Options in effect:
        --file wgas1
        --make-bed
        --out wgas2

228694 (of 228694) markers to be included from [ wgas1.map ]
90 individuals read from [ wgas1.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 [ wgas2.fam ]
Writing map (extended format) information to [ wgas2.bim ]
Writing genotype bitfile to [ wgas2.bed ]
Using (default) SNP-major mode

Analysis finished: Tue May 24 14:51:52 2022
```

Step 2B: Creating a Binary Input File

Verify in your **data** folder that the **wgas2** files were created

.lock_gPLINK	5/23/2022 3:34 PM	LOCK_GPLINK File	1 KB
.metafile_gPLINK	5/23/2022 3:34 PM	METAFILE_GPLINK...	1 KB
command-list	5/19/2022 12:43 PM	Text Document	3 KB
extra.map	5/19/2022 12:43 PM	MAP File	1 KB
extra.ped	5/19/2022 12:43 PM	PED File	9 KB
gPLINK	5/19/2022 12:43 PM	Executable Jar File	1,758 KB
Haploview	5/19/2022 12:43 PM	Executable Jar File	5,177 KB
plink	5/19/2022 12:43 PM	Application	3,873 KB
pop.cov	5/19/2022 12:43 PM	COV File	2 KB
wgas1.map	5/19/2022 12:43 PM	MAP File	6,869 KB
wgas1.ped	5/19/2022 12:43 PM	PED File	80,403 KB
wgas2.bed	5/24/2022 2:51 PM	BED File	5,137 KB
wgas2.bim	5/24/2022 2:51 PM	BIM File	7,762 KB
wgas2.fam	5/24/2022 2:51 PM	FAM File	3 KB
wgas2	5/24/2022 2:51 PM	Text Document	2 KB

Step 3A: Validating the Conversion

Command prompt (do not type)

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

```
> plink.exe --maf 0.01 --geno 0.05 --mind 0.05 --bfile wgas2 --out validate

# plink.exe is the software

# --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
```

Step 3A: Validating the Conversion

Your screen should look similar to this

```
C:\Users\IGB\Desktop\Labs\09_Variant_Analysis\data>plink.exe --maf 0.01 --geno 0.05 --mind 0.05 --bfile wgas2 --out validate

@----------------------------------------------------------@
|        PLINK!       |     v1.02      |   25/May/2008     |
|----------------------------------------------------------|
|  (C) 2008 Shaun Purcell, GNU General Public License, v2  |
|----------------------------------------------------------|
|  For documentation, citation & bug-report instructions:  |
|        http://pngu.mgh.harvard.edu/purcell/plink/        |
@----------------------------------------------------------@

Web-check not implemented on this system...
Writing this text to log file [ validate.log ]
Analysis started: Tue May 24 14:56:26 2022

Options in effect:
        --maf 0.01
        --geno 0.05
        --mind 0.05
        --bfile wgas2
        --out validate

Reading map (extended format) from [ wgas2.bim ]
228694 markers to be included from [ wgas2.bim ]
Reading pedigree information from [ wgas2.fam ]
90 individuals read from [ wgas2.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 [ wgas2.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: Tue May 24 14:56:31 2022
```

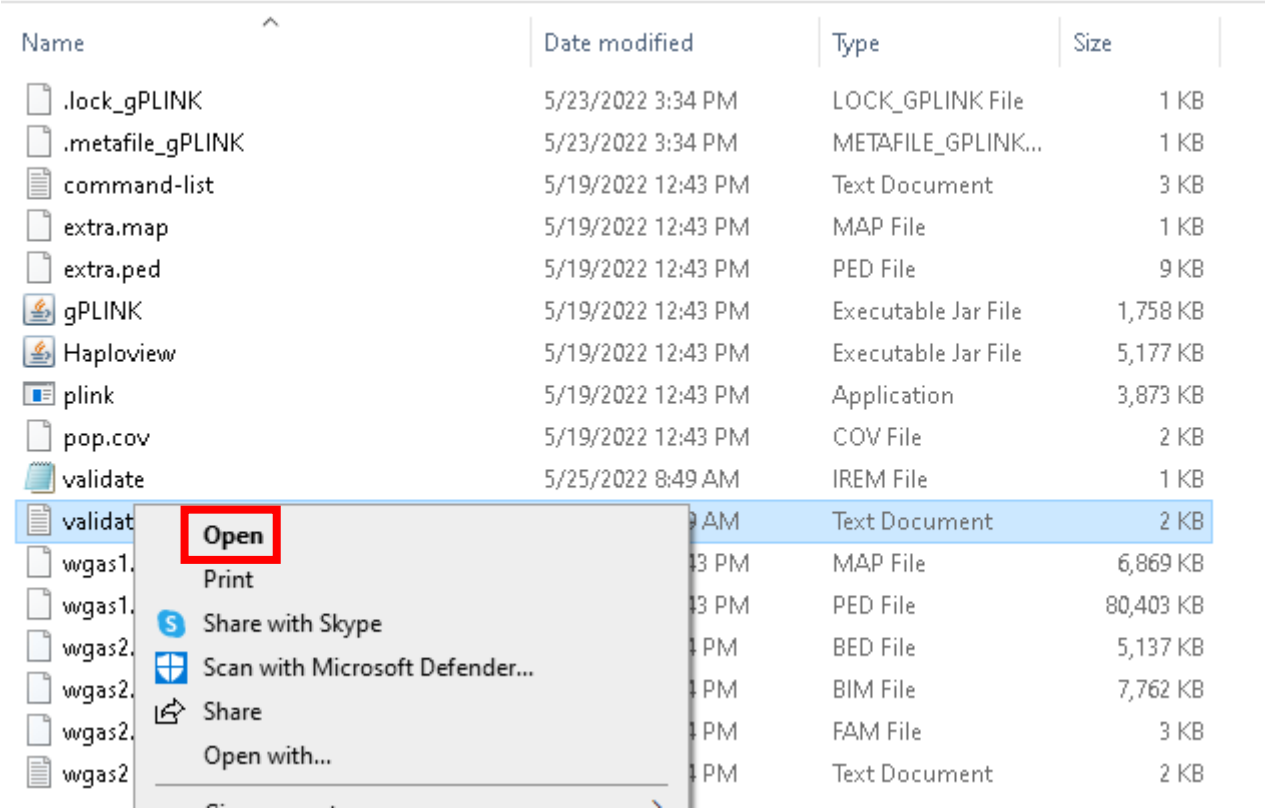
Step 3B: Validating the Conversion

Verify in your **data** folder that the **validate** files were created

Name	Date modified	Type	Size
.lock_gPLINK	5/23/2022 3:34 PM	LOCK_GPLINK File	1 KB
.metafile_gPLINK	5/23/2022 3:34 PM	METAFILE_GPLINK...	1 KB
command-list	5/19/2022 12:43 PM	Text Document	3 KB
extra.map	5/19/2022 12:43 PM	MAP File	1 KB
extra.ped	5/19/2022 12:43 PM	PED File	9 KB
gPLINK	5/19/2022 12:43 PM	Executable Jar File	1,758 KB
Haploview	5/19/2022 12:43 PM	Executable Jar File	5,177 KB
plink	5/19/2022 12:43 PM	Application	3,873 KB
pop.cov	5/19/2022 12:43 PM	COV File	2 KB
validate.irem	5/24/2022 2:56 PM	IREM File	1 KB
validate	5/24/2022 2:56 PM	Text Document	2 KB
wgas1.map	5/19/2022 12:43 PM	MAP File	6,869 KB
wgas1.ped	5/19/2022 12:43 PM	PED File	80,403 KB
wgas2.bed	5/24/2022 2:51 PM	BED File	5,137 KB
wgas2.bim	5/24/2022 2:51 PM	BIM File	7,762 KB
wgas2.fam	5/24/2022 2:51 PM	FAM File	3 KB
wqas2	5/24/2022 2:51 PM	Text Document	2 KB

Step 3C: Viewing Validation

Right click on the **validate** file and choose the **Open** option



Step 3D: Viewing Validation

validate - Notepad

File Edit Format View Help

```
@----------------------------------------------------------@
|        PLINK!       |     v1.02      |   25/May/2008     |
|----------------------------------------------------------|
|  (C) 2008 Shaun Purcell, GNU General Public License, v2  |
|----------------------------------------------------------|
|  For documentation, citation & bug-report instructions:  |
|        http://pngu.mgh.harvard.edu/purcell/plink/        |
@----------------------------------------------------------@

Web-check not implemented on this system...
Writing this text to log file [ validate.log ]
Analysis started: Tue May 24 19:24:57 2022

Options in effect:
        --maf 0.01
        --geno 0.05
        --mind 0.05
        --bfile wgas2
        --out validate

Reading map (extended format) from [ wgas2.bim ]
228694 markers to be included from [ wgas2.bim ]
Reading pedigree information from [ wgas2.fam ]
90 individuals read from [ wgas2.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 [ wgas2.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: Tue May 24 19:25:01 2022
```

46834 out of ~ 230,000 SNPs were removed because the 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)

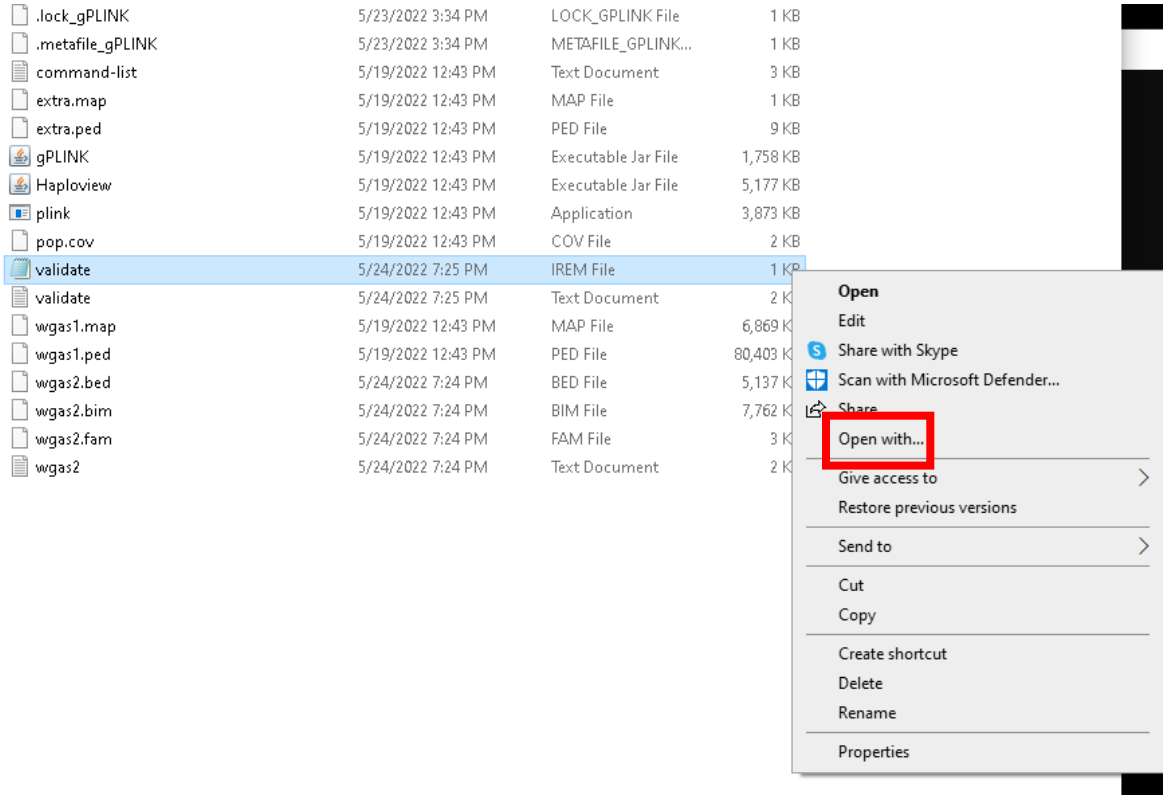
Step 3E: Validating the Conversion

Locate the **irem** file

Name	Date modified	Type	Size
.lock_gPLINK	5/23/2022 3:34 PM	LOCK_GPLINK File	1 KB
.metafile_gPLINK	5/23/2022 3:34 PM	METAFILE_GPLINK...	1 KB
command-list	5/19/2022 12:43 PM	Text Document	3 KB
extra.map	5/19/2022 12:43 PM	MAP File	1 KB
extra.ped	5/19/2022 12:43 PM	PED File	9 KB
gPLINK	5/19/2022 12:43 PM	Executable Jar File	1,758 KB
Haploview	5/19/2022 12:43 PM	Executable Jar File	5,177 KB
plink	5/19/2022 12:43 PM	Application	3,873 KB
pop.cov	5/19/2022 12:43 PM	COV File	2 KB
validate.irem	5/24/2022 2:56 PM	IREM File	1 KB
validate	5/24/2022 2:56 PM	Text Document	2 KB
wgas1.map	5/19/2022 12:43 PM	MAP File	6,869 KB
wgas1.ped	5/19/2022 12:43 PM	PED File	80,403 KB
wgas2.bed	5/24/2022 2:51 PM	BED File	5,137 KB
wgas2.bim	5/24/2022 2:51 PM	BIM File	7,762 KB
wgas2.fam	5/24/2022 2:51 PM	FAM File	3 KB
wgas2	5/24/2022 2:51 PM	Text Document	2 KB

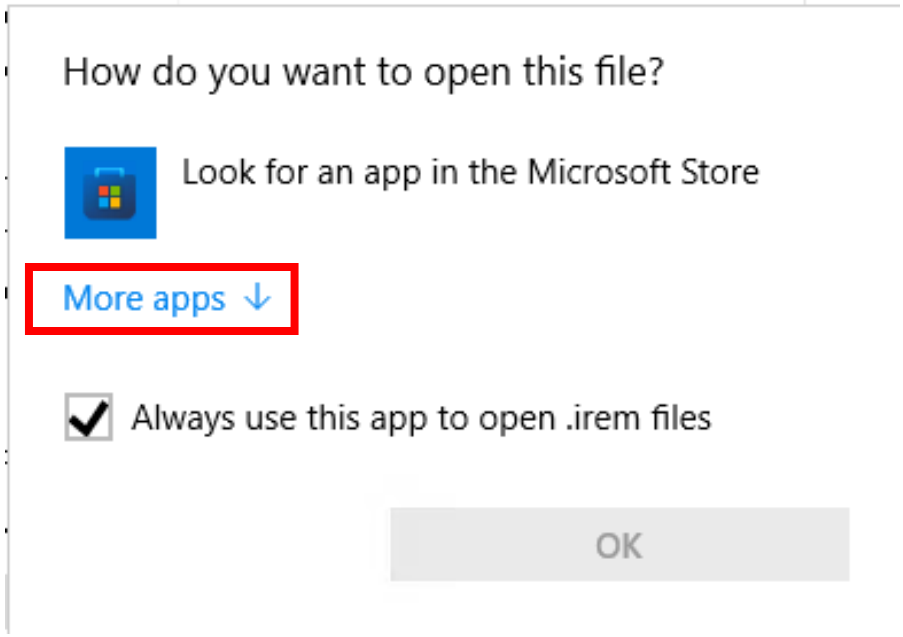
Step 3F: Validating the Conversion

Right click on **validate.irem** and choose the **Open with…** option



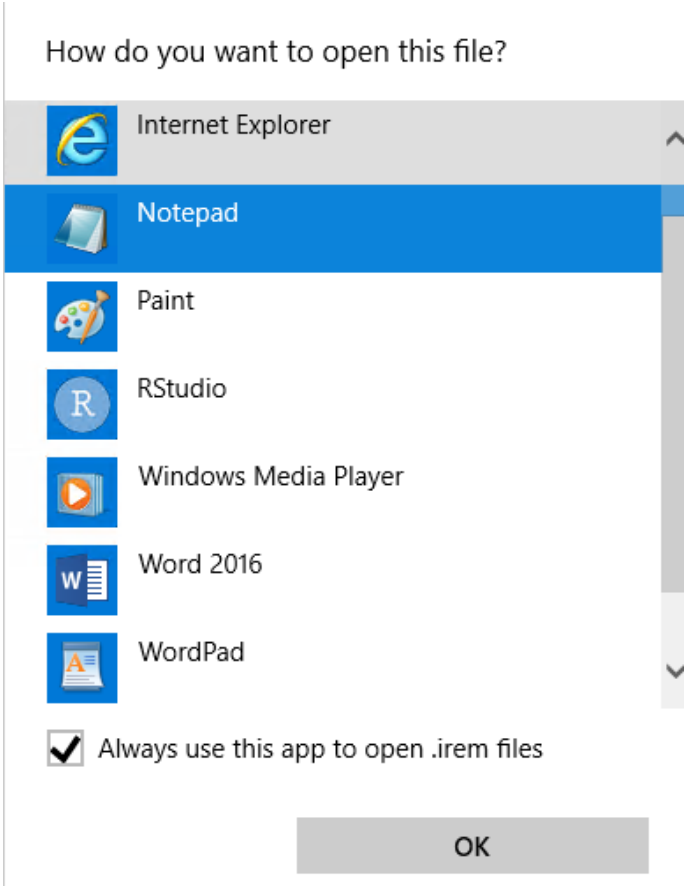
Step 3G: Validating the Conversion

Next, select **More apps** and choose the **Notepad** software



Step 3H: Validating the Conversion

Lastly, select the **Notepad** software

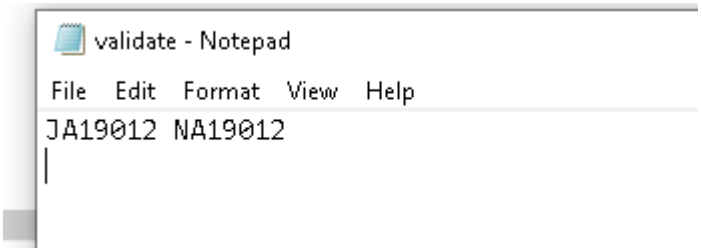


Step 3I: Validating the Conversion

You should see the following:

JA19012 NA19012

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



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

Command prompt (do not type)

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

```
> plink.exe --maf 0.01 --geno 0.05 --mind 0.05 --bfile wgas2 --make-bed –-out wgas3

# plink.exe is the software

# --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
```

Step 4A: Quality Control Analysis

Your screen should look similar to this

```
C:\Users\IGB\Desktop\Labs\09_Variant_Analysis\data>plink.exe --maf 0.01 --geno 0.05 --mind 0.05 --bfile wgas2 --make-bed --out wgas3

@----------------------------------------------------------@
|        PLINK!       |     v1.02      |   25/May/2008     |
|----------------------------------------------------------|
|  (C) 2008 Shaun Purcell, GNU General Public License, v2  |
|----------------------------------------------------------|
|  For documentation, citation & bug-report instructions:  |
|        http://pngu.mgh.harvard.edu/purcell/plink/        |
@----------------------------------------------------------@

Web-check not implemented on this system...
Writing this text to log file [ wgas3.log ]
Analysis started: Tue May 24 15:03:16 2022

Options in effect:
        --maf 0.01
        --geno 0.05
        --mind 0.05
        --bfile wgas2
        --make-bed
        --out wgas3

Reading map (extended format) from [ wgas2.bim ]
228694 markers to be included from [ wgas2.bim ]
Reading pedigree information from [ wgas2.fam ]
90 individuals read from [ wgas2.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 [ wgas2.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 [ wgas3.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 [ wgas3.fam ]
Writing map (extended format) information to [ wgas3.bim ]
Writing genotype bitfile to [ wgas3.bed ]
Using (default) SNP-major mode

Analysis finished: Tue May 24 15:03:21 2022
```

Step 4B: Quality Control Analysis

Verify in your **data** folder that the **wgas3** files were created

.lock_gPLINK	5/23/2022 3:34 PM	LOCK_GPLINK File	1 KB
.metafile_gPLINK	5/23/2022 3:34 PM	METAFILE_GPLINK...	1 KB
command	5/19/2022 12:43 PM	Text Document	3 KB
extra.map	5/19/2022 12:43 PM	MAP File	1 KB
extra.ped	5/19/2022 12:43 PM	PED File	9 KB
gPLINK	5/19/2022 12:43 PM	Executable Jar File	1,758 KB
Haploview	5/19/2022 12:43 PM	Executable Jar File	5,177 KB
plink	5/19/2022 12:43 PM	Application	3,873 KB
pop.cov	5/19/2022 12:43 PM	COV File	2 KB
validate	5/24/2022 2:56 PM	IREM File	1 KB
validate	5/24/2022 2:56 PM	Text Document	2 KB
wgas1.map	5/19/2022 12:43 PM	MAP File	6,869 KB
wgas1.ped	5/19/2022 12:43 PM	PED File	80,403 KB
wgas2.bed	5/24/2022 2:51 PM	BED File	5,137 KB
wgas2.bim	5/24/2022 2:51 PM	BIM File	7,762 KB
wgas2.fam	5/24/2022 2:51 PM	FAM File	3 KB
wgas2	5/24/2022 2:51 PM	Text Document	2 KB
wgas3.bed	5/24/2022 3:03 PM	BED File	4,034 KB
wgas3.bim	5/24/2022 3:03 PM	BIM File	6,089 KB
wgas3.fam	5/24/2022 3:03 PM	FAM File	3 KB
wgas3	5/24/2022 3:03 PM	IREM File	1 KB
wgas3	5/24/2022 3:03 PM	Text Document	3 KB

Type: METAFILE_GPLINK File
Size: 141 bytes
Date modified: 5/23/2022 3:34 PM

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

Command prompt (do not type)

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

```
> plink.exe --bfile wgas3 --assoc --adjust –-out assoc1

# plink.exe is the software

# --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
```

Step 5A: GWAS

Your screen should look similar to this

```
C:\Users\IGB\Desktop\Labs\09_Variant_Analysis\data>plink.exe --bfile wgas3 --assoc --adjust --out assoc1

@----------------------------------------------------------@
|        PLINK!       |     v1.02      |   25/May/2008     |
|----------------------------------------------------------|
|  (C) 2008 Shaun Purcell, GNU General Public License, v2  |
|----------------------------------------------------------|
|  For documentation, citation & bug-report instructions:  |
|        http://pngu.mgh.harvard.edu/purcell/plink/        |
@----------------------------------------------------------@

Web-check not implemented on this system...
Writing this text to log file [ assoc1.log ]
Analysis started: Tue May 24 15:06:17 2022

Options in effect:
        --bfile wgas3
        --assoc
        --adjust
        --out assoc1

Reading map (extended format) from [ wgas3.bim ]
179562 markers to be included from [ wgas3.bim ]
Reading pedigree information from [ wgas3.fam ]
89 individuals read from [ wgas3.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 [ wgas3.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
0 of 89 individuals removed for low genotyping ( MIND > 0.1 )
Total genotyping rate in remaining individuals is 0.996307
0 SNPs failed missingness test ( GENO > 0.1 )
0 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 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: Tue May 24 15:06:31 2022
```

Step 5B: GWAS

Verify in your **data** folder that the **assoc1** files were created

.lock_gPLINK	5/23/2022 3:34 PM	LOCK_GPLINK File	1 KB
.metafile_gPLINK	5/23/2022 3:34 PM	METAFILE_GPLINK...	1 KB
assoc1	5/24/2022 3:06 PM	ASSOC File	17,010 KB
assoc1.assoc.adjusted	5/24/2022 3:06 PM	ADJUSTED File	18,763 KB
assoc1	5/24/2022 3:06 PM	Text Document	3 KB
command-list	5/19/2022 12:43 PM	Text Document	3 KB
extra.map	5/19/2022 12:43 PM	MAP File	1 KB
extra.ped	5/19/2022 12:43 PM	PED File	9 KB
gPLINK	5/19/2022 12:43 PM	Executable Jar File	1,758 KB
Haploview	5/19/2022 12:43 PM	Executable Jar File	5,177 KB
plink	5/19/2022 12:43 PM	Application	3,873 KB
pop.cov	5/19/2022 12:43 PM	COV File	2 KB
validate	5/24/2022 2:56 PM	IREM File	1 KB
validate	5/24/2022 2:56 PM	Text Document	2 KB
wgas1.map	5/19/2022 12:43 PM	MAP File	6,869 KB
wgas1.ped	5/19/2022 12:43 PM	PED File	80,403 KB
wgas2.bed	5/24/2022 2:51 PM	BED File	5,137 KB
wgas2.bim	5/24/2022 2:51 PM	BIM File	7,762 KB
wgas2.fam	5/24/2022 2:51 PM	FAM File	3 KB
wgas2	5/24/2022 2:51 PM	Text Document	2 KB
wgas3.bed	5/24/2022 3:03 PM	BED File	4,034 KB
wgas3.bim	5/24/2022 3:03 PM	BIM File	6,089 KB
wgas3.fam	5/24/2022 3:03 PM	FAM File	3 KB
wgas3	5/24/2022 3:03 PM	IREM File	1 KB
wgas3	5/24/2022 3:03 PM	Text Document	3 KB

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**.

You can inspect this file by **Right Clicking** it and selecting **Open with…** and selecting the **Notepad** software. Open in **Excel** if you want to sort by p-value.

Overall, 13,294 SNPS survive at p value of 0.05 WITHOUT Multiple Hypothesis Correction.

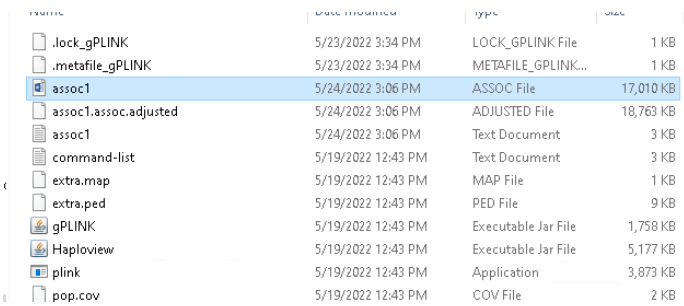
The few top SNPs are shown below, after using the unix **sort**, **awk**, and **head** commands.

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-007	5.769
20	rs6110115	13911728	C	0.3085	0.6829	A	24.59	7.103e-007	0.2071
11	rs2508756	75921549	A	0.5417	0.1951	G	22.5	2.105e-006	4.875
15	rs16976702	54120691	G	0.5833	0.2317	C	22.43	2.183e-006	4.642
8	rs11204005	12895576	A	0.3229	0.6585	G	19.97	7.882e-006	0.2473
9	rs16910850	94478347	T	0.09375	0.3659	C	19.14	1.216e-005	0.1793
12	rs1195747	129970575	A	0.3085	0.6375	G	18.83	1.427e-005	0.2537
17	rs7207095	77933018	G	0.5208	0.2073	A	18.52	1.682e-005	4.156

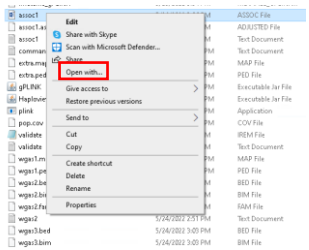
Step 6: GWAS Without Multiple Hypothesis Correction

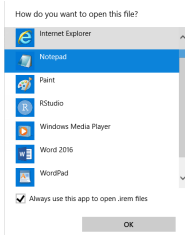
The SNP p values from our GWAS with no multiple hypothesis correction are located in the 9th column of **assoc1.assoc**.

You can inspect this file by **Right Clicking** it and selecting **Open with…** and selecting the **Notepad** software.



.lock_gPLINK	5/23/2022 3:34 PM	LOCK_GPLINK File	1 KB
.metafile_gPLINK	5/23/2022 3:34 PM	METAFILE_GPLINK...	1 KB
assoc1	5/24/2022 3:06 PM	ASSOC File	17,010 KB
assoc1.assoc.adjusted	5/24/2022 3:06 PM	ADJUSTED File	18,763 KB
assoc1	5/24/2022 3:06 PM	Text Document	3 KB
command-list	5/19/2022 12:43 PM	Text Document	3 KB
extra.map	5/19/2022 12:43 PM	MAP File	1 KB
extra.ped	5/19/2022 12:43 PM	PED File	9 KB
gPLINK	5/19/2022 12:43 PM	Executable Jar File	1,758 KB
Haploview	5/19/2022 12:43 PM	Executable Jar File	5,177 KB
plink	5/19/2022 12:43 PM	Application	3,873 KB
pop.cov	5/19/2022 12:43 PM	COV File	2 KB





assoc1 - Notepad

File Edit Format View Help

CHR	SNP	BP	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	rs11260562	1205233	A	0.02174	0.03659	G	0.3424	0.5585	0.5852
1	rs6685064	1251215	C	0.3854	0.439	T	0.5253	0.4686	0.8013
1	rs3766180	1563420	T	0.1771	0.09756	C	2.317	0.128	1.991
1	rs6603791	1586208	A	0.1771	0.08537	G	3.189	0.07413	2.306
1	rs7519837	1596068	C	0.1702	0.08537	T	2.775	0.09573	2.198
1	rs3737628	1755094	T	0.5104	0.4756	C	0.2143	0.6434	1.149
1	rs7511905	1825948	A	0.08333	0.1098	C	0.3574	0.5499	0.7374
1	rs3855951	1836464	C	0.1146	0.2125	T	3.127	0.07699	0.4796
1	rs6603803	1844850	A	0.4894	0.5122	G	0.09133	0.7625	0.9127
1	rs2803285	1920531	A	0.1354	0.08537	G	1.111	0.2919	1.678
1	rs7513222	2060063	G	0.4479	0.3415	A	2.09	0.1482	1.565
1	rs3107146	2079746	T	0.03125	0.08537	C	2.443	0.1181	0.3456
1	rs3107157	2094131	T	0.1979	0.1951	C	0.002187	0.9627	1.018
1	rs3753242	2101843	C	0.3542	0.3902	T	0.2467	0.6194	0.8569
1	rs385039	2109571	G	0.2083	0.1463	A	1.153	0.283	1.535
1	rs2292857	2138600	A	0.0625	0.06098	G	0.001773	0.9664	1.027
1	rs626479	2142422	A	0.2083	0.1585	G	0.7261	0.3941	1.397
1	rs262680	2199311	C	0.3438	0.4024	T	0.6529	0.4191	0.7778
1	rs16824948	2218382	T	0.08333	0.125	C	0.8251	0.3637	0.6364
1	rs12084736	2221742	T	0.3958	0.4146	C	0.0649	0.7989	0.9249
1	rs12045693	2237743	C	0.4167	0.4756	A	0.6225	0.4301	0.7875
1	rs2132303	2255420	T	0.2083	0.1098	C	3.151	0.07587	2.135
1	rs1496555	2266413	A	0.2292	0.122	G	3.448	0.06334	2.141
1	rs2645072	2312585	A	0.07292	0.122	C	1.231	0.2672	0.5663
1	rs7527871	2313888	C	0.4271	0.4024	A	0.1106	0.7395	1.107
1	rs2840528	2316058	G	0.4348	0.4756	A	0.2915	0.5892	0.8481

Overall, 13,294 SNPS survive at p value of 0.05 WITHOUT Multiple Hypothesis Correction.

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.**

You can inspect this file by **Right Clicking** it and selecting **Open with…** and selecting the **Notepad** software

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

CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
11	rs2513514	4.693e-007	7.131e-006	0.08427	0.08427	0.08081	0.08081	0.06378	0.8084
20	rs6110115	7.103e-007	9.938e-006	0.1276	0.1275	0.1198	0.1198	0.06378	0.8084
11	rs2508756	2.105e-006	2.373e-005	0.378	0.3779	0.3147	0.3147	0.098	1
15	rs16976702	2.183e-006	2.443e-005	0.392	0.392	0.3243	0.3243	0.098	1

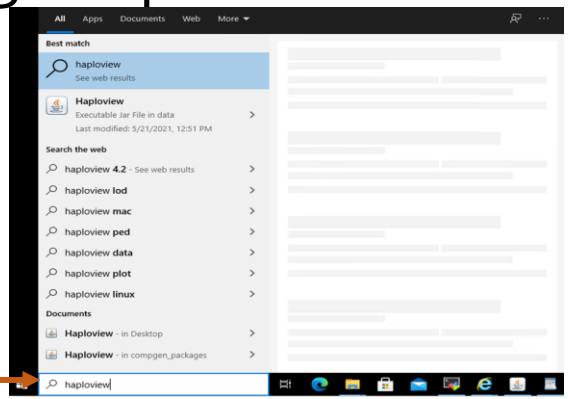
Visualization

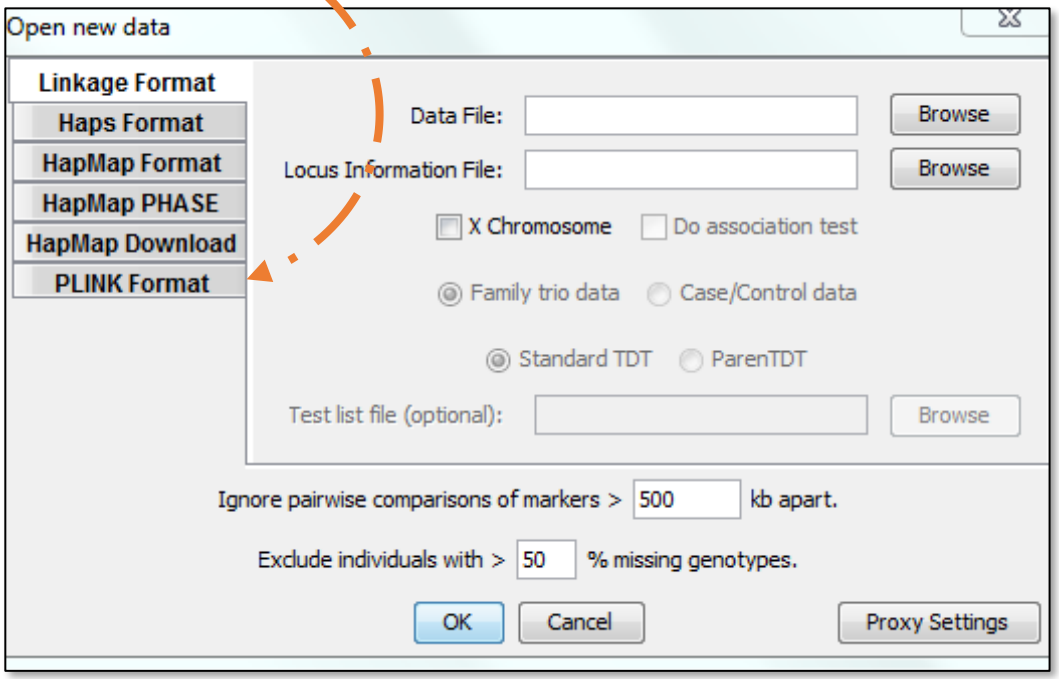
In this exercise, we will generate a Manhattan Plot of our association results using **Haploview** from the **Broad Institute.**

Step 8A: Configuring Haploview

Open Haploview from **Search.**

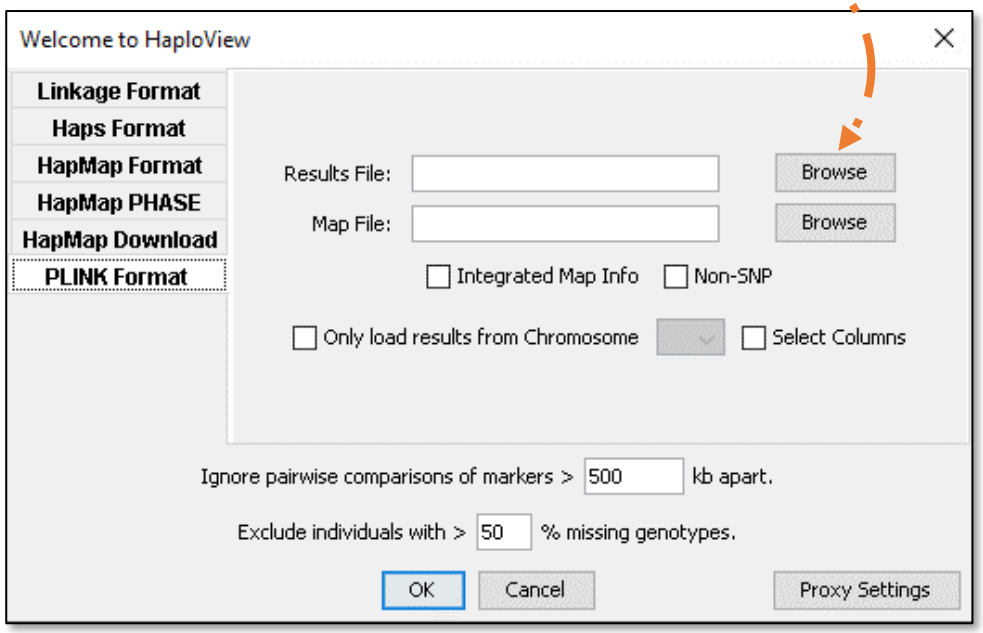
Click **PLINK Format**





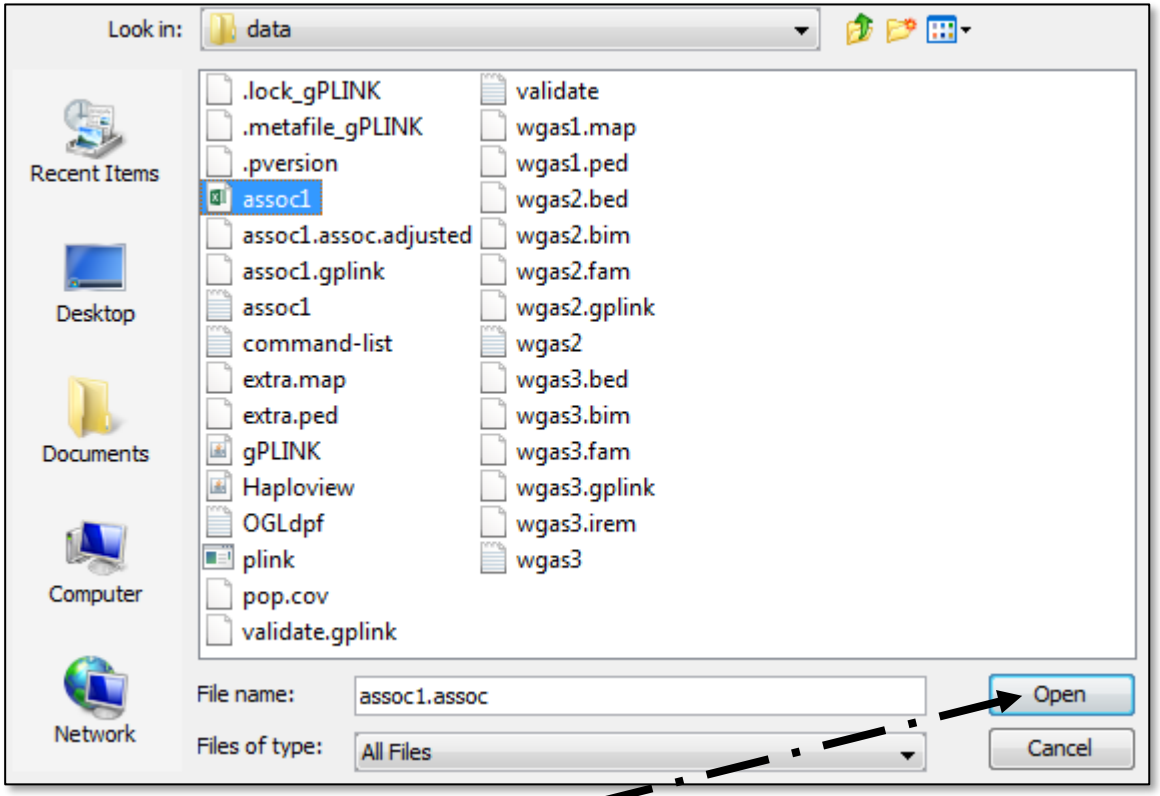
Step 8B: Configuring Haploview

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



Step 8C: Configuring Haploview

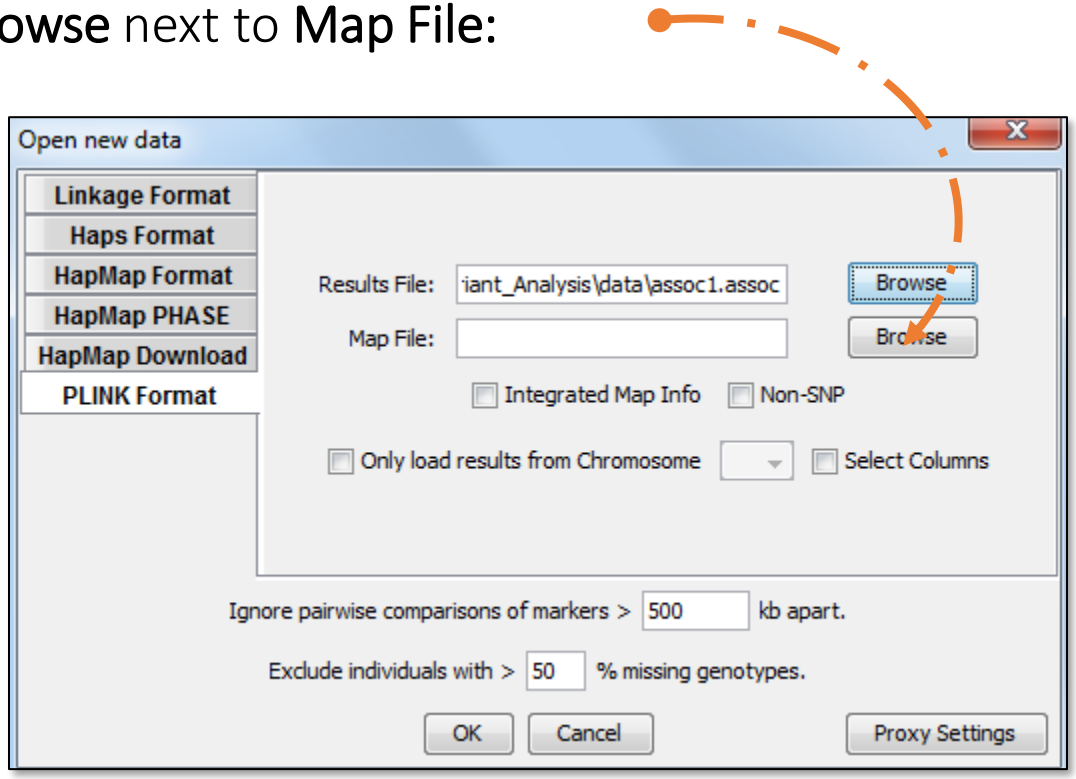
Navigate to the directory PLINK saved the file **assoc1.assoc**. It should be saved in the data sub folder in the 09_Variant_Analysis folder



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

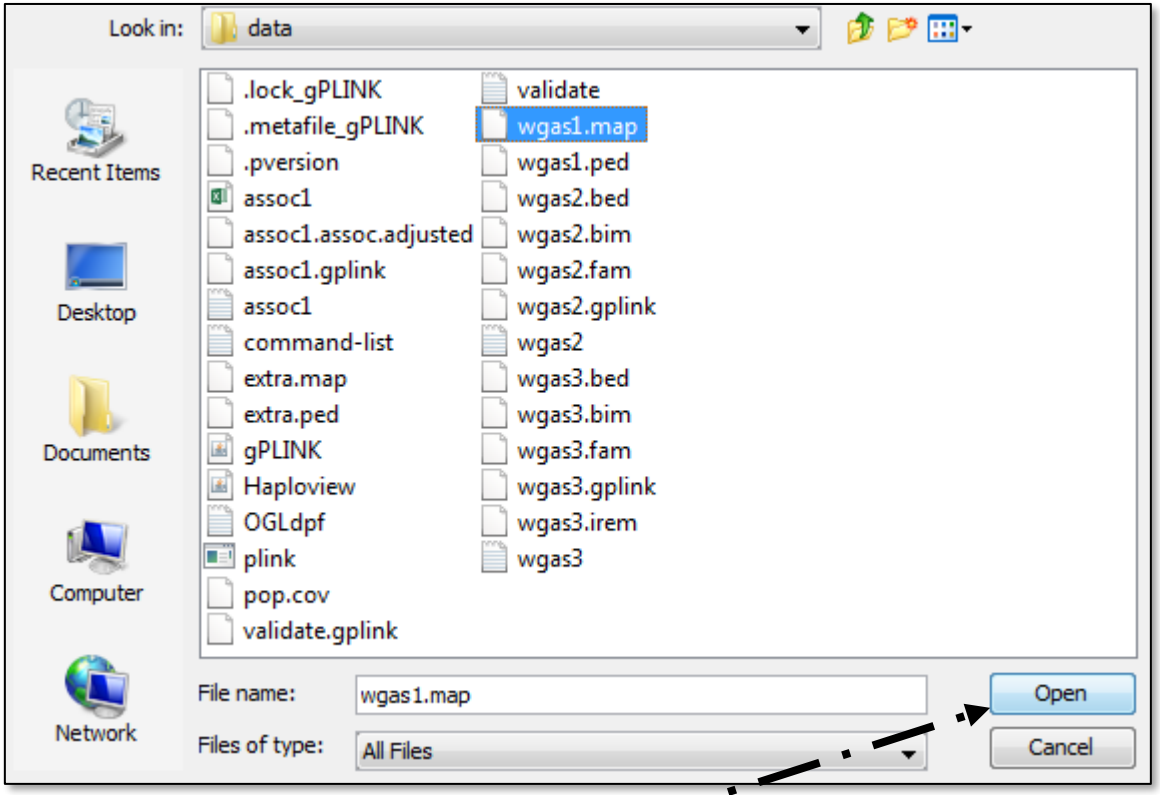
Step 8D: Configuring Haploview

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



Step 8E: Configuring Haploview

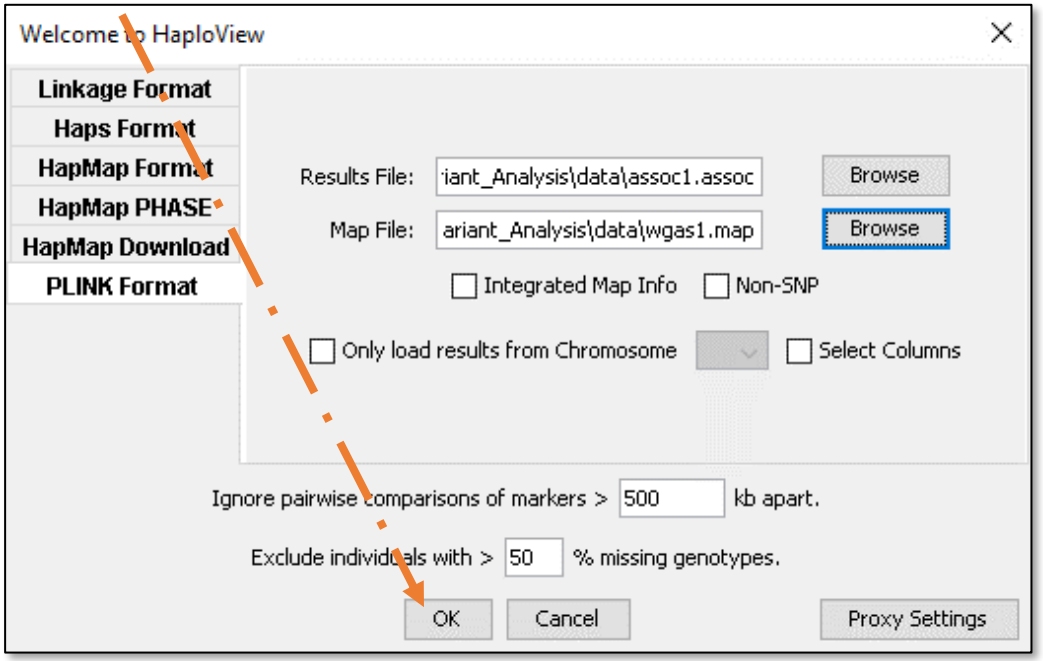
Navigate to the data directory containing **wgas1.map**



Select **wgas1.map** and click **Open**.

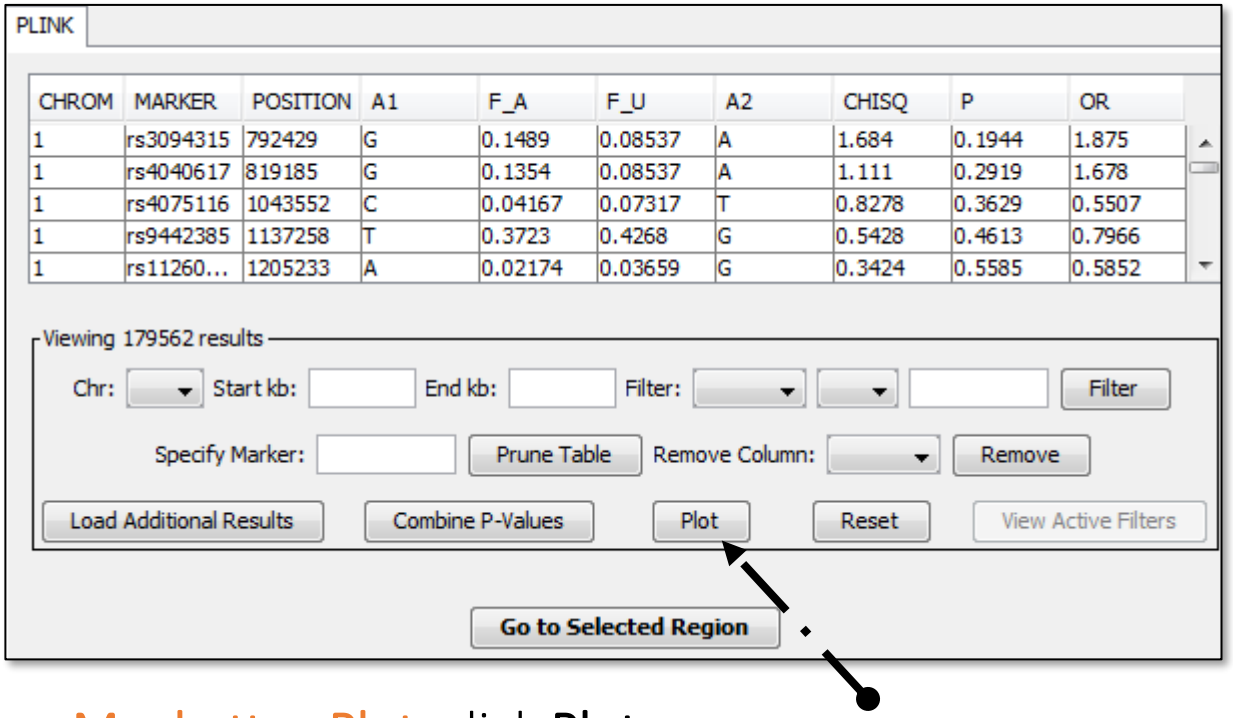
Step 8F: Configuring Haploview

Click on **OK.**



Step 8G: Configuring Haploview

Your **asssoc1** should be shown in **Haploview** in tabular format.



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**

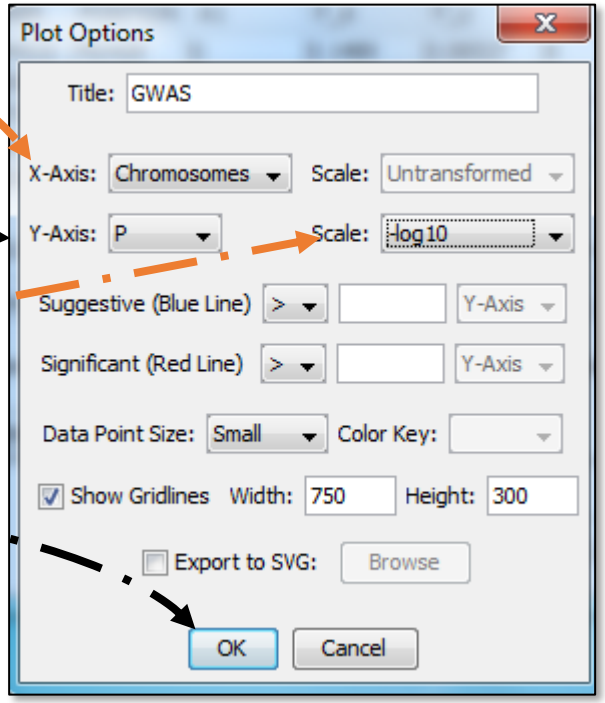
Step 8H: Configuring Haploview

Select Chromosomes for **X-Axis**

Select P for **Y-Axis**

Select –log10 for **Y-Axis Scale**

Click **OK**



Step 9: Manhattan Plot

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