Exercise

In this exercise, we will do the following:

1. Gain familiarity with a graphical user interface to **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/2021-schedule/](https://publish.illinois.edu/compgenomicscourse/2021-schedule/)
Step 0: Local Files (for UIUC users)

**If you are a Mayo Clinic user, go to the next slide**

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

\[\text{course_directory}\]

We will use the files found in:

\[\text{course_directory}\]\09_Variant_Analysis\Data

**For UIUC:** \[\text{course_directory}\] = C:\Users\IGB\Desktop\VM

so the path would be:

C:\Users\IGB\Desktop\VM\09_Variant_Analysis\Data
Step 0: Local Files (for Mayo Clinic users)

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

\[\text{course_directory}\]

We will use the files found in:

\[\text{course_directory}\]\\09\_Variant\_Analysis

\textbf{Mayo Clinic:} [course_directory] = C:\\Users\\<MayoClinicLANID>\\Documents

so the path would be:

C:\\Users\\<MayoClinicLANID>\\Documents\\09\_Variant\_Analysis
## Dataset Characteristics

<table>
<thead>
<tr>
<th>filename</th>
<th>meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>plink.exe</td>
<td>An executable of the PLINK GWAS toolkit. (Preinstalled)</td>
</tr>
<tr>
<td>gPLINK.jar</td>
<td>A JAVA graphical user interface (GUI) that interfaces with plink.exe.</td>
</tr>
<tr>
<td>Haploview.jar</td>
<td>A haplotype analysis program written in JAVA. Used to view PLINK results and SNP analysis.</td>
</tr>
<tr>
<td>wgas1.ped</td>
<td>Genotype data for 228,694 SNPS on 90 people.</td>
</tr>
<tr>
<td>wgas1.map</td>
<td>Map file for the SNPs in wgas1.ped.</td>
</tr>
<tr>
<td>extra.ped</td>
<td>Genotype data for 29 SNPS on the same 90 people.</td>
</tr>
<tr>
<td>extra.map</td>
<td>Map file for the SNPs in extra.ped.</td>
</tr>
<tr>
<td>pop.cov</td>
<td>Population membership of the 90 people. (1 = Han Chinese, 2 = Japanese)</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Family ID</th>
<th>Individual ID</th>
<th>Paternal ID</th>
<th>Maternal ID</th>
<th>Sex</th>
<th>Phenotype</th>
<th>Genotype…</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH18526</td>
<td>NA18526</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>A A 0 G ..</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>chr</th>
<th>SNP ID</th>
<th>cM</th>
<th>Base Pair Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>rs17121574</td>
<td>12.8</td>
<td>12799052</td>
</tr>
</tbody>
</table>
Configuring gPLINK

In this exercise, we will configure gPLINK to work with our data.

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 gPLINK

gPLINK is a graphical user interface, written in JAVA, to the command line program PLINK.

To start gPLINK, navigate to

```
[course_directory]/09_Variant_Analysis/data/
```

Double click on gPLINK.jar
Step 1B: Starting gPLINK

A window should appear similar to the one below:
Step 2A: Configuring gPLINK

Click on the **Project** item on the **Menu Bar**.

Select **Open** from the drop down menu.

The pop-up window should look similar to the screenshot below.

Click on **Browse**.
Step 2B: Configuring gPLINK

In the file browser, navigate to the following directory:

```
/course_directory]/09_Variant_Analysis
```

Click on the **data** directory and click **Open**.

Click OK on the **Open Project** window.
Step 2C: Configuring gPLINK

You should see the files in the data folder in the Folder Viewer on the left hand side of gPLINK.
Step 3A: Creating a Binary Input File

Click the PLINK item on the Menu Bar.

Click Data Management.

Click Generate fileset.

In the next window, select Standard Input on the tab bar.

Select wgas1 under Quick Fileset.

Check Binary fileset.

Under Output File input wgas2.

Click OK.
Step 3B: Creating a Binary Input File

On the **Execute Command** window, click **OK**.

This will convert our **wgas1** files to a binary format.

Under the **Operations Viewer**, you will see **wgas2** with an **R** next to it indicating running. Wait for it to turn **GREEN**.
Step 3C: Creating a Binary Input File

In the **Folder Viewer**, you should see a bunch of new *wgas2* files created during the file creation process.
Step 4A: Validating the Conversion

Click the **PLINK** item on the **Menu Bar**.

Click **Summary Statistics**.

Click **Validate Fileset**.

In the next window, select **Binary Input** on the tab bar.

Select **wgas2** under **Quick Fileset**.

Under **Output File** input **validate**.

Click **Threshold**.
Step 4B: Validating the Conversion

On the **Threshold** window:

Set **Minor allele frequency** to 0.01.

Set **Maximum SNP missingness rate** to 0.05.

Set **Maximum individual missingness rate** to 0.05.

Click **OK**.

Click **OK**
Step 4C: Validating the Conversion

On the **Execute Command** window click **OK**.

Wait for the command to finish (validate will show the icon)

Click on the validate track:
Step 4C: Validating the Conversion

Look in the **Log viewer**

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 4D: Validating the Conversion

Click the + adjacent to the **Validate** track to expand it.

Click the + adjacent to the **Output files** track to expand it.

Right click `validate.irem` and click **Open in default viewer**.

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.

<table>
<thead>
<tr>
<th>filter</th>
<th>meaning</th>
<th>threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minor Allele Frequency (MAF)</td>
<td>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</td>
<td>1%</td>
</tr>
<tr>
<td>Individual Genotyping rate</td>
<td>The number of SNPs probed for an individual must exceed this threshold for the person to be analyzed.</td>
<td>95%</td>
</tr>
<tr>
<td>SNP genotyping rate</td>
<td>The SNP must be probed for at least this many individuals.</td>
<td>95%</td>
</tr>
</tbody>
</table>
Step 5A: Quality Control Analysis

Click the **PLINK** item on the **Menu Bar**.

Click **Data Management**.

Click **Generate Fileset**.

In the next window, select **Binary Input** on the tab bar.

Select **wgas2** under **Quick Fileset**.

Click **Binary fileset**.

Under **Output File** input **wgas3**.

Click **Threshold**.
Step 5B: Quality Control Analysis

On the **Threshold** window:

Set **Minor allele frequency** to 0.01.

Set **Maximum SNP missingness rate** to 0.05.

Set **Maximum individual missingness rate** to 0.05

Click **OK**.

Click **OK**.
Step 5C: Quality Control Analysis

On the **Execute Command** window, click **OK**.

This will create a new set of files prefixed **wgas3** that are filtered according to the thresholds on the previous slide.
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 6A: GWAS

Click the **PLINK** item on the **Menu Bar**.

Click **Association**.

Click **Allelic Association Tests**.

In the next window, select **Binary Input** on the tab bar.

Select **wgas3** under **Quick Fileset**.

Click **Adjusted p-values**.

Under **Output File** input **assoc1**.

Click **OK**.
Step 6B: GWAS

On the **Execute Command** window, click **OK**.

![Execute Command Window](image)

This will perform the **GWAS** analysis on our data and store the results under **assoc1** in the main window of **gPLINK**.
Step 7: 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 in default viewer. 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.
Step 7: 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 in default viewer. If the viewer has wrapped the text, you can go to view and under word wrap, choose no wrap.

Overall, 13,294 SNPS survive at $p$ value of 0.05 WITHOUT Multiple Hypothesis Correction.
Step 8: 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 in default viewer.

Overall, only 4 SNPs!!! show a FDR Correction of less than 1.

<table>
<thead>
<tr>
<th>CHR</th>
<th>SNP</th>
<th>UNADJ</th>
<th>GC</th>
<th>BONF</th>
<th>HOLM</th>
<th>SIDAK_SS</th>
<th>SIDAK_SD</th>
<th>FDR_BH</th>
<th>FDR_EY</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>rs2513514</td>
<td>4.693e-007</td>
<td>7.131e-006</td>
<td>0.08427</td>
<td>0.08427</td>
<td>0.08081</td>
<td>0.08081</td>
<td>0.06378</td>
<td>0.8084</td>
</tr>
<tr>
<td>20</td>
<td>rs6110115</td>
<td>7.103e-007</td>
<td>9.938e-006</td>
<td>0.1276</td>
<td>0.1276</td>
<td>0.1198</td>
<td>0.1198</td>
<td>0.06378</td>
<td>0.8084</td>
</tr>
<tr>
<td>11</td>
<td>rs2508756</td>
<td>2.105e-006</td>
<td>2.373e-005</td>
<td>0.378</td>
<td>0.3779</td>
<td>0.3147</td>
<td>0.3147</td>
<td>0.098</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>rs16976702</td>
<td>2.183e-006</td>
<td>2.443e-005</td>
<td>0.392</td>
<td>0.392</td>
<td>0.3243</td>
<td>0.3243</td>
<td>0.098</td>
<td>1</td>
</tr>
</tbody>
</table>
Visualization

In this exercise, we will generate a Manhattan Plot of our association results using Haploview from the Broad Institute.
Step 9A: Configuring Haplovview

Open **Haplovview** from **Search**.

Click **PLINK Format**
Step 9B: Configuring Haploview

Click on **Browse** next to **Results File**.
Step 9C: Configuring Haploview

Navigate to the directory \texttt{gPLINK} saved the file \texttt{assoc1.assoc}. It should be saved in the data sub folder in the 09\_Variant\_Analysis folder.

Select \texttt{assoc1.assoc} and click \texttt{Open}.
Step 9D: Configuring Haploview

Click on **Browse** next to **Map File**:
Step 9E: Configuring Haploview

Navigate to the data directory containing `wgas1.map`

Select `wgas1.map` and click Open.
Step 9F: Configuring Haploview

Click on **OK**.
Step 9G: Configuring Haploview

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

To create a **Manhattan Plot**, click **Plot**.
Step 9H: Configuring Haploview

Select **Chromosomes** for X-Axis

Select **P** for Y-Axis

Select **–log10** for Y-Axis Scale

Click **OK**
Step 10: Manhattan Plot

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

![Manhattan Plot](image)