**PLINK PRACTICAL**

Tuesday 8am

|  |
| --- |
| In this practical, we will:* Learn about genetic relatedness / relationship matrices by estimating one.
* Learn about some concepts and methods from molecular genetic methods.
 |

This practical will introduce you to Genome Wide Association Study (GWAS) datasets and direct estimations of genetic relatedness using a program called PLINK. PLINK is an open-source GWAS analysis toolset that allows manipulating and analyzing large-scale datasets in an efficient way.

PLINK includes functions for: data organization, formatting, quality control, association testing, population stratification, and more. Details about PLINK and its documentation are available in the developers’ website: http://zzz.bwh.harvard.edu/plink/

For this tutorial, we simulated a dataset based on real data made available by Myers et al. The original dataset included only unrelated individuals. We have modified that data to include MZ and DZ twin pairs:

*Myers, A. J., Gibbs, J. R., Webster, J. A., Rohrer, K., Zhao, A., Marlowe, L., ... & Zismann, V. L. (2007). A survey of genetic human cortical gene expression. Nature genetics, 39(12), 1494.*

We will use PLINK via the Unix console available in your laptops. You can copy and paste specific commands from this document to your UNIX terminal.

|  |
| --- |
| UNIX terminal commands are shown in a monospaced font inside a box. |

**1) Change working directory**

First, open Unix terminal. Then, we need to make sure we are working in the right folder:

|  |
| --- |
| cd /home*/*jose/2020/Tuesday |

**2) Check file format: PLINK-friendly formats (.bed and .ped)**

PLINK has preferred file formats. These are PED and BED. PED is the original standard text format for PEDigree information. By default, it is a human-readable text format (i.e., it makes sense when you open it with a text editor). On the other hand, BED (i.e., a *binary* PED) is a compressed version of PED, which saves space and speeds up analysis (crucial when working GWAS datasets) but cannot be visualized directly.

|  |
| --- |
| plink --file gwas\_plinkdata --make-bed --out gwas\_plinkdata |

**3) Clean the data (quality control)**

PLINK includes several options to *clean* genetic data. This means filtering out low quality data or outliers. We are going to run a very basic quality control. In practice, this process would involve looking at more descriptive statistics, graphs, etc. This topic is addressed in depth at the IBG workshop for analyzing GWAS and genome sequence data.

|  |
| --- |
| plink --bfile gwas\_plinkdata --geno 0.05 --mind 0.05 --hwe 1e-6 --maf 0.1 --make-bed --out gwas\_plinkdata\_clean |



**4) Estimate the genetic relatedness matrix (GRM)**

We can estimate the genetic relatedness across individuals in our dataset based on their shared genetic variants.

* **Output style A: genetic relatedness matrix**

|  |
| --- |
| plink --bfile gwas\_plinkdata\_clean --make-rel triangle |

Take a look to the results in Unix…

|  |
| --- |
| zless -S plink.rel |

Note: Press q to exit view

This can a bit a tad overwhelming, especially with real life large-datasets. However, our dataset is small enough to take a look to the actual GRM. For the sake of explanation, open the file ‘GRM\_highlighted.pdf’. A GRM is symmetric, here we present only the lower triangle. Zoom in to find which individuals are likely to be MZ twins (**in green**), DZ twins (**in red**), and genetically unrelated (not highlighted). NOTE: genetic relatedness of each individual with themselves appears on the diagonal. Values are not 1 because this is an empirical estimation of relatedness. More on this in Rob’s session on Thursday.

* **Output style B: relatedness pair by pair**

Another way to look at genetic relatedness is pair by pair. We can ask PLINK to give us the degree of relatedness and number of overlapping genetic variants for each pair in our dataset.

|  |
| --- |
| plink --bfile gwas\_plinkdata\_clean -make-grm-gz no-gz |

Take a look to the results in Unix…

|  |
| --- |
| zless –S plink.grm |

The file is structured like this (output from PLINK doesn’t have headers):

|  |  |  |  |
| --- | --- | --- | --- |
| ID1 | ID2 | common SNPs | gen relatedness |
| 1 | 1 | 251751 | 0.9945 |
| 2 | 1 | 247262 | -0.0233 |
| 2 | 2 | 247777 | 1.0011 |
| 3 | 1 | 250987 | -0.0227 |
| 3 | 2 | 247011 | -0.0209 |

Again, our small dataset comes handy and we can easily look at the output. Open the file ‘grel\_highlighted.xls’. Pairs likely to be MZ twins **in green**, DZ twins **in red**, and unrelated not highlighted.

**In summary,**

* Genotype data allow us to directly estimate the degree of genetic relatedness across individuals.
* A genetic relatedness (/relationship) matrix contains degree of genetic similarity between individuals.

This concludes this tutorial!