GWAS, Quality Control and Family based analysis workshop February 2016 – Matthew McQueen

TUTORIAL 2 - Data Cleaning, Relatedness Measures, Genetic Ancestry

NOTE: These files are located in the *plink* directory under either the *raw* or *clean* directories. We will be using PLINK to generate the information that we need. We will first be using a dataset that has not been cleaned to illustrate the quality control steps related to sample and marker cleaning. Then, we will use a different dataset that has been cleaned to explore genetic ancestry as well as subsequent analysis approaches later in the workshop.

Create a workspace (*~/plinkout*)

1) First, let's create a directory for our PLINK output. Make sure you are in your home directory:

```
student@courses:~$ cd
```

2) Next, we will create a new directory (please name this directory "plinkout"):

```
student@courses:~$ mkdir plinkout
```

3) Finally, let's change directory to **plinkout** and then we will run PLINK from there:

```
student@courses:~$ cd plinkout
```

Inspecting genome-wide data *(~/cbio2016/plink/raw)*

1) We can generate some missingness and call rate information for the raw dataset. The command `[--missing]` will provide information by sample as well as by marker.

```
plink --bfile /student_data/cbio2016/plink/raw/raw --missing --out raw
```

2) Next, let's estimate heterozygosity for each sample

```
plink --bfile /student_data/cbio2016/plink/raw/raw --het --out raw
```

3) We can also run tests for Hardy-Weinberg Equilibrium (HWE) for the markers using the `[--hardy]` command.

```
plink --bfile /student_data/cbio2016/plink/raw/raw --hardy --out raw
```

NOTE: We will come back to the files generated by these commands at a later step.

TUTORIAL 2 - Data Cleaning, Relatedness Measures, Genetic Ancestry

Genetic Ancestry *(~/cbio2016/plink/clean)*

We will use a set of data that has been cleaned to estimate genome-wide IBD matrices for all pairs of subjects in the data. This can take a long time to run so ***we will skip this step*** in the interest of time (we have run this previously and it will be available for viewing). If we were to run this analysis, we would use the `[--genome]` command. Note that we are also using a file that has been pruned of markers (`bmi_ldp` where `ldp` = linkage disequilibrium pruned). We prune markers to create a set of independent markers that results in a vast improvement in computational efficiency to estimate genome-wide IBD.

DO NOT RUN THIS: `[plink --bfile bmi_ldp --genome]`

4) Once the pairwise IBD information is generated as we have done, it is rather easy to run multidimensional scaling (MDS) analysis to set up visualizing genetic ancestry using the `[--read-genome]`, `[--cluster]` and `[--mds-plot]` commands.

```
plink --bfile /student_data/cbio2016/plink/clean/bmi_ldp --read-genome
/student_data/cbio2016/plink/clean/bmi.genome --cluster --mds-plot 10 --out bmi
```

5) We can now use these files to evaluate ancestry as well as relationship status (note that the HapMap data was also preprocessed so that we could visualize that data as well).

We can now inspect the files generated for both the BMI file as well as the HapMap sample.