CHL 5224 – Project 2 Readme
Running a GWAS via PLINK using the 1000 Genome Project Data

Learning GWAS

- Revisit the lecture notes and the WTCCC GWAS study
- Additional background reading: from NIH: https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet
- A excellent tutorial on how to run GWAS:
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6001694/pdf/MPR-27-e1608.pdf
  https://github.com/MareesAT/GWA_tutorial/

Learning PLINK

- Homepage: http://zzz.bwh.harvard.edu/plink/; updated version is here: https://www.cog-genomics.org/plink2
- Good to start with the tutorial: http://zzz.bwh.harvard.edu/plink/tutorial.shtml

Learning the 1000 Genome Project (1kGP)

- The 1000 Genome Project (1kGP) background: https://www.internationalgenome.org/
- A cleaned set of 1000 genome project data: http://tcag.ca/tools/1000genomes.html
  - Important to read: the report (http://tcag.ca/documents/tools/omni25_qcReport.pdf) on the quality control (QC) steps performed
  - Good to repeat the QC steps after learning PLINK

Project 2 - Conduct and report a GWAS as a practice

- The cleaned dataset has 1736 unrelated/indepedent individuals and ~2M high-qualtiy SNPs.
- Randomly assign an individual to case or control, then conduct the apporopriate GWAS, (The random assignment is important: there are clear population stratifications based on PCA analysis; see page 10 of the report (http://tcag.ca/documents/tools/omni25_qcReport.pdf) on the QC steps.)
- Results should include the Manhanttan plot, the QQ-plot, and the histogram.
- Additional considerations (to achieve A+): the PCA analysis, the effect of minor allele frequency, and reproducing the QC analysis.