PLINK 2.0 GWAS Pipeline User Guide

RH (2025-04-05)

INTRODUCTION

This pipeline makes it easy to process genotype data using PLINK by running pre-prepared scripts. The pipeline also compiles the output files and provides custom R-scripts in a pre-prepared working directory for rapid analysis and reporting.

PLINK is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner.

ABBREVIATIONS

- **VCF:** Variant Call Format
- **PGEN (.pgen):** Plink2 binary GENotype table
- **PSAM (.psam):** Plink2 SAMple information
- **PVAR (.pvar):** Plink2 VARiant information file
- **ZST (.zst):** Lossles data compressed file with Zstd (`$ sudo apt install zstd`)

RAW/INPUT DATA NEEDED
.vcf file

- - -

I. PROCEDURE

Ia. CLONE GITHUB REPOSITORY AND RENAME

- `$ git clone` https://github.com/ramTT
- `$ mv standard_folder_name my_project_folder_name`

Note: *The enfolded executable is for Linux AVX2 Intel. If a different operating system is being used make sure to download the correct plink2-executable from https://www.cog-genomics.org/plink/2.0/ and do:*

- `$ unzip downloaded_file.zip`
- `$ mv plink2* ~/my_project_folder_name/Main`

Ib. TRANSFER .vcf FILE INTO REPOSITORY

- `$ mv ~/my_vcf_file.vcf ~/my_project_folder_name/Main/Data/Raw_data`

Ic. DATA CONVERSION (.vcf file -> .pgen, .psam & .pvar files):

- `$ cd ~/my_project_folder_name/Main`
- `$ bash Script1_DataConversion.sh`
- **Output** (*stored in Raw_data folder*):
  - `output_file_name.pgen.zst` (*binary format, zst-compressed*)
  - `output_file_name.psam` (*human readable format*)
  - `output_file_name.pvar.zst` (*binary format, zst-compressed*)

Id. DATA PRE-PROCESSING

- Remain in `~/my_project_folder_name/Main`
- `$ bash Script2_DataPreProcessing.sh`

Important settings to consider:

Ie. GWAS

- Remain in `~/my_project_folder_name/Main`
- `$ bash Script3_Analysis.sh`

Important settings to consider:

- - -

II. DATA VISUALIZATION & ANALYSIS in R

1. Open Rstudio.
2. Create a new project and select `~/my_project_folder_name/Main/R_project_folder` as the project folder (i.e. the working directory). All relevant output files are stored here in a .csv format and are ready for import into R.
3. Run the R-scripts and produce the baseline quality control and analysis PDFs using the pre-prepared R-scripts.
4. Transfer any additional epidemiological data into the working directory.
5. Continue to analyze and interpret the data manually and integrate and correlate clinical and genetic data.
6. Apply machine- and deep learning algorithms if possible (eg through R (Caret), Python (PyTorch) and/or Python (JAX)).

III. CITATION

IIIa. SOFTWARE

- Package: PLINK 2.0
- Authors: Shaun Purcell, Christopher Chang
- URL: www.cog-genomics.org/plink/2.0/

IIIb. PUBLICATION

- Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, 4.