Lab Notebook

Joyce Wang

9/8/2021

Contents

scripts Folder **1**

- 00_make_directories.sh 1
- 01_UKBB_genotypes_filtered.sh 2
  - 01a_get_ambiguous_indel_snps.py 2
- 03_predict_populations.sh 2
  - 03a_classify_ukb.py 2
  - 03b_separate_populations.py 3
  - 03d_UKBB_genotypes_EUR_train_125k.sh 3
- 04_produce_files_for_gwas.sh 3
  - 04a_create_covariates.R 3
  - 04b_create_phenotypes_file.R 3
- 05_gwas_plink.sh 3
  - 05b_plot_Manhattan.sh 3
  - 05c_plot_ManhattanPlots.R 3

scripts Folder

00_make_directories.sh

This file creates some subdirectories under the same directory where this file is located, in the following structure:

- data
  - ambiguous_indel_snps
  - intersecting_filtered
  - kgp_filtered
  - kgp_merged
  - kgp_meta
  - ukb_filtered
  - ukb_merged

    - `ukb_meta`
    - `ukb_populations`
    - `models`
    - `phenotypes`
    - `gwas_results`
    - `prs`
    - `kgp_populations`
    - `fst`
    - `LDpred`
        - `prs`
        - `tmp-data`
        - `val_prs`
    - `prs_comparisons`
    - `theory`
    - `theor_herit`
    - `theoretical`
- `img`

For me, these directories are under `$WORK2/pgs_portability/`.

`01_UKBB_genotypes_filtered.sh`

This file filters out the related individuals and indels and ambiguous variants.

To get the IDs of unrelated individuals, I ran `ukbconv ukb45020.enc_ukb txt -s22020 -oukb.unrelated` and extracted the IDs. The unrelated individuals are coded as `1` and otherwise as`NA`s. The extracted IDs were stored in the file `ukb.unrelated.id.txt` under `data/extracted_phenotypes/`.

I copied all the files from `/corral-repl/utexas/Recombining-sex-chro/ukb/data/genotype_calls/` into my directory `data/genotype_calls/` for this script to work, or it will throw a `FileNotFoundError`.

The list of individuals to be excluded from the study is contained in `w61666_20210809.csv` under `data/ukb_meta/`.

After running this file, the outputs are stored at `data/ukb_filtered/` and `data/ukb_merged/`.

`01a_get_ambiguous_indel_snps.py`

This is a helper script for `01_UKBB_genotypes_filtered.sh` that identifies ambiguous indel SNPs.

`03_predict_populations.sh`

This file calls `03a_classify_ukb.py` and `03b_separate_populations.py` to separate the WB and NWB.

`03a_classify_ukb.py`

This file classifies an individual either as a WB or a NWB, based on UKB field `22006`. I extracted the IDs for WB as `ukb.wb.id.txt` under `data/extracted_phenotypes/`.

03b_separate_populations.py

This file separates the WB and NWB. From the total WB population, 5,000 individuals were randomly selected as the test set and the remaining is th training set. For each trait, 200,000 individuals from the training set for GWAS.

From the WB training set, 125,000 individuals were randomly selected to represent this group's LD pattern.

03d_UKBB_genotypes_EUR_train_125k.sh

This file runs LDpred on the 125,000 individuals randomly selected from the WB training set.

04_produce_files_for_gwas.sh

This file calls 04a_create_covariates.R and 04b_create_phenotypes_file.R to create covariate and phenotype files for GWAS.

04a_create_covariates.R

This file combines sex (UKB field 31), age (UKB field 21022), and PCs (UKB field 22020) as covariates for GWAS.

04b_create_phenotypes_file.R

This file sebset the raw phwnotypes for GWAS. The extracted phenotypes include:

- BMI
- WBC
- Height
- RBC
- MCV
- MCH
- Lymphocyte
- Platelet
- Monocyte
- Eosinophil

The respective UKB field codes are stored as martin_gwas_info.txt under data/.

05_gwas_plink.sh

This file runs GWAS on the randomly selected 200,000 WB individuals from the training set for each trait.

05b_plot_Manhattan.sh

This file calls 05c_plot_ManhattanPlots.R to create Manhattan and QQ plots.

05c_plot_ManhattanPlots.R

This file creates Manhattan and QQ plots. For Manhattan plots, both zoomed and unzoomed versions will be created.