Lab Notebook

Joyce Wang

9/8/2021

Contents

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 indels and ambiguous variants.

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/`.

To get the IDs of WB, I ran `ukbconv ukb45020.enc_ukb txt -s34 -oYOB` and extracted the IDs. the extracted IDs were stored in the file `wb_id.txt` under `data/ukb_meta/`. A list of non-WB individuals were stored in `nwb_id.txt`.

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 ambigupus indel SNPs.

01b_remove_ambiguous_indel_snps.py

I haven't used this script yet.

01c_find_duplicates.py

I haven't used this script yet.

01d_import_1KG.sh

I downloaded the 1000 Genome VCF dataset for each chromosome from https://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/

02_PCA_plink.sh

I'm skipping this file for now because I don't need to project the UKB individuals onto the 1000 Genome dataset.

`03_predict_populations.sh`

I'm skipping this file for now because I don't need to project the UKB individuals onto the 1000 Genome dataset.