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 (for me, these directories are under `$WORK2/pgs_portability/`), in the following structure:

- `/data/`
  - `/ambiguous_indel_snps/`: For storing the names of the indels and ambiguous variants
  - `/ukb_filtered/`: For storing the SNPs after filtering
  - `/ukb_merged/`: For storing the files after merging the data from all chromosomes together
  - `/ukb_meta/`: For storing the meta file including individuals who wish to opt out from the study
  - `/ukb_populations/`: For storing the IDs of WB, NWB, as well as the IDs of WB randomly selected to train the GWAS model for each trait
  - `/phenotypes/`: For storing the phenotype data
  - `/gwas_results/`: For storing the GWAS results
  - `/prs/`: For storing the C+T and PGS results
  - `/LDpred/`: For storing the IDs of individuals randomly selected to create the LD landscape (used for clumping)
  - `/fst/`: For storing the results of FST calculations
  - `/prs_comparisons/`: For storing the full datasets with FST groups
  - `/theory/`: For storing the IDs of each FST group
  - `/extracted_phenotypes/`: For storing the phenotypes extracted from UKB (WB, sex, age)
- `/img/`: For storing images

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, 70,000 individuals were randomly selected as the test set and the remaining is the training set. For each trait, 200,000 individuals were randomly selected from the training set for GWAS. From the WB training set, 125,000 individuals were randomly selected to represent this group's LD pattern.

The number of individuals after filtering for each group is shown below:

- WB: 337,463
  - Training: 267,463
    - For each trait, 200,000 individuals are randomly selected from this set
    - 125,000 individuals were randomly selected to represent this group's LD pattern
  - Test: 70,000
- NWB: 69,649

03d_UKBB_genotypes_EUR_train_125k.sh

This file creates the files to represent the LD pattern. These files will be useful for clumping.

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 subsets the raw phenotypes for GWAS and normalizes the values to that of WB. 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. For PCs, the first 20 are used. To save time, I used the genotype array SNPs instead of the imputed SNPs. Using the imputed SNPs will allow the discovery of more hits, but it'll take a lot more time.

`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.

`06a_clumping1.sh`

This file clumps and filters the p-values from GWAS based on the LD pattern. The 5 thresholds used are:

- `5e-8`
- `1e-5`
- `1e-4`
- `1e-3`
- `1e-2`

`06b_convert_plink2_glm_to_plink1.py`

This is a helper file for `06a_clumping1.sh` that converts a `plink2` GLM output (`.glm.linear`) into the `plink` output format (`.assoc`).

`06c_filter_snps_for_prs.py`

This is a helper file for `06a_clumping1.sh` that filters SNPs based on a given p-value.

`06d_combine_glm_threshold_4.py`

This is a helper file for `06a_clumping1.sh` that combines all significant SNPs.

`06e_after_clumping.sh`

This is a helper file for `06a_clumping1.sh` that filters for significant SNPs from other SNPs.

`07_compute_prs.sh`

This file computes the PGS using the significant SNPs and for each threshold and trait. The outputs are stored at `/data/prs/`.

`08_fst.sh`

This file calculates FST by running the helper scripts.

`08a_edit_merged_fam.R`

This file calculates edits the `.fam` files to include information on population.

`08b_fst_parallel.py`

This file creates groups of bash scripts for FST calculations.

To reduce the time for running, I'm only calculating FST using the genotype data on chromosome 1. I have also tried calculating the FST on 10% randomly chosen individuals with all 22 autosomes, the resulting FST values correlate well with the ones calculated using just chromosome 1 (`r^2 = 0.959`; `p < 1e^(-4)`), but there were some noise. See `correlation.png` under `/img/FST` for details.

`08b_submit_multiple_jobs.txt`

This file contains a simple script to submit multiple jobs on TACC in a loop. Note that TACC only allows up to 50 jobs in queue or running at the same time.

`08c_final_fst_formatting.R`

This file combines all the files from FST calculations into a single file

`09_R2.sh`

This file runs `09a_run_R2_models.R` to generate plots that correlate incremental r^2 with FST values.

`09a_run_R2_models.R`

This file creates plots under `/img/` that correlate relative incremental r^2 with FST values.

It first loads in the FST and phenotype data. It then sorts and groups individuals based on FST values, with the lowest first. For individuals with the same FST values, they will be assigned groups randomly.

For each group, it then calculates the relative incremental r^2 of PGS as follows:

1. Create a linear regression with `phenotype_value ~ covariates`. Calculated r^2 (I used the regular one instead of the adjusted one)
2. Create a linear regression with `phenotype_value ~ PGS + covariates`. Calculated r^2 (I used the regular one instead of the adjusted one)
3. Subtract r^2 in Step 2 from the one in Step 1.

For the incremental r^2 of each group, I divided it by the incremental r^2 of Group 1 (the group with the lowest FST and is the closest to the WB training set) to calculate relative incremental r^2.

I then plotted the relative incremental r^2 as a function of median FST in each group.

Because there's a lot of noise at low FST values, I have tried making 3 types of plots:

1. The regular linear regression plot
2. The plot with 2 linear regression models, separated at x = 0.01
3. The splines

`img` Folder

`Manhattan` Folder

This folder contains the Manhattan plots and QQ plots for each trait. For the Manhattan plots, I ceated an unzoomed and zoomed version.

`FST` Folder

This folder contains the scatterplots with median FST for each group on the x-axis and relative incremental r^2 on the y-axis for each trait. The most recent plots are stored directly under this folder, whereas the older versions are under the `Old` subfolder.

The FST values were calculated using chromosome 1 only.

I have tried grouping the FST values in 2 different ways:

1. Create 100 bins, make 2nd-99th equally large, and make 1st 10x larger
2. Create 100 bins, make 2nd-99th equally large, and make 1st 50x larger

For both cases, there are lot of noise near at lower FST values and I don't understand why. I tried creating 3 types of plots:

1. The regular linear regression plot
2. The plot with 2 linear regression models, separated at x = 0.01
3. The splines

I also created `correlation.png` to show the strong relationship between FST values calculated with all autosomes and the ones calculated just from chromosome 1.

`Old` Folder

This folder contains the older plots.

`1` Folder This was my first attempt at creating these plots.

`FINAL_COMB_FST1_40.png`:

- Used adjusted r^2 to calculate incremental r^2
- Used incremental r^2 instead of relative incremental r^2
- FST calculated from ~10% individuals (not randomly chosen) with all autosomes
- Negative FST values were set to 0s
- The dots were not very visible
- Used only 40 bins, making the trend not very clear

`R_h2_FST_1_40.png`:

- The motivation behind making this plot was not clear

`2` Folder This was my second attempt at creating these plots. I separated each trait into a plot to make the dots more visible. Some problems:

- Used adjusted r^2 to calculate incremental r^2
- Negative FST values were set to 0s
- Used only 40 bins, making the trend not very clear

`3-100_10` Folder This was my second attempt at creating these plots. I used the regular r^2 instead of the adjusted r^2. I increased the number of bins and made the first one 10x larger than the rest. However, negative FST values were set to 0s.

`3-100_50` Folder This was my second attempt at creating these plots. I used the regular r^2 instead of the adjusted r^2. I increased the number of bins and made the first one 50x larger than the rest. However, negative FST values were set to 0s.