Brain Imaging Genetics Software (BIGS)

Version 0.0.1 (work in progress)

1. Table of contents

2. Introduction

BIGS aims to provide a hands-on & reproducible pipeline for brain imaging genomics, including downloading data from publicly available databases, e.g., UKBB and ADNI, standard genetic quality check pipeline, and common genetic analysis, e.g., GWAS.

2.1. Motivation

Brain imaging genomics is an emerging scientific field, but interdisciplinary crosstalk is not easy for newcomers, e.g., a neuroscientist or machine learning practitioner wants to include genetics into their studies. As open science and reproducible research have drawn increasing attention in the neuroimaging community, BIGS aims to provide an end-to-end solution to go through all these steps. BIGS covers only the genetic part, assuming that users are familiar with imaging analysis. If not, please refer to our previously-proposed reproducible software: Clinica (http://www.clinica.run/).

2.2. Third-party Software

As BIGS performs all analyses with third-party software, we list here all necessary software needed to be installed.

Plink: https://plink.readthedocs.io/en/latest/
King: https://www.kingrelatedness.com/
GCTA: https://yanglab.westlake.edu.cn/software/gcta/#Overview
LDSC: https://github.com/bulik/ldsc
MAGMA: https://ctg.cncr.nl/software/magma
PRsice: https://www.prsice.info/
FlashPCA: https://github.com/gabraham/flashpca
liftOver: http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/liftOver
McCarthy Group Tools: https://www.well.ox.ac.uk/~wrayner/tools/
Annovar: https://annovar.openbioinformatics.org/en/latest/user-guide/startup/
Michigan Imputation Server: https://imputationserver.sph.umich.edu/index.html#!pages/home

3. Genetic quality check protocol

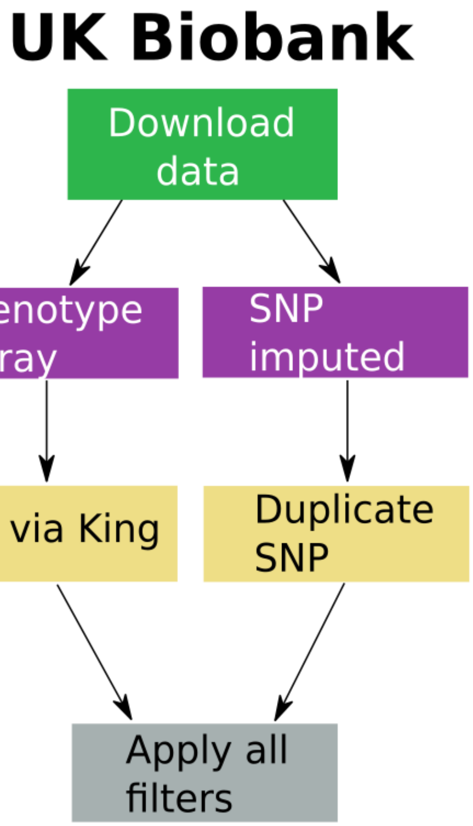

Figure 1. A schematic diagram of the UKBB and ADNI genetic quality check pipeline.

3.1. UKBB

We are going through a standard genetic QC protocol to process the UK Biobank (UKBB) genetic data. Our genetic data were downloaded on 14/07/2021. **Fig. 1** is an overview of our QC protocol.

3.1.1. Download data

UKBB provides single-nucleotide polymorphisms (SNPs) from *i*) genotype arrays (SNP array thereafter) and *ii*) its imputed genetic data (SNP imputed thereafter) of Version 3. Links for more details are:
i) https://www.ukbiobank.ac.uk/enable-your-research/about-our-data/genetic-data
ii) https://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=263

You need to download the UKBB genetic data with your username and key, which in total takes 13T, but we will use only the SNP arrays and SNP imputed data.

3.1.2. Relationship inference via King

This step is to remove related individuals that might account for population stratification. The King software and the SNP arrays are used to infer this relationship.

A. First, we process the SNP calls wit the following command line:

```
#!/bin/bash
for i in 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X XY
do
```

```
    plink --threads 8 --bed ${bed} --fam ${fam} --bim ${bim} --maf 0.01 --geno 0.03 --make-bed --out
${output_file}
done
```

B. Merge all chromosomes into one

```
plink \
  --bed chr1_calls_geno_0.03_maf_0.01.bed \
  --bim chr1_calls_geno_0.03_maf_0.01.bim \
  --fam chr1_calls_geno_0.03_maf_0.01.fam \
  --merge-list list_beds.txt \
  --make-bed --out chr_all
```

C. Run King

```
"king -b ${bed} --unrelated  --degree 2 --cpus 32 --prefix ${prefix}"
```

This will end up with a file that contains all unrelated individuals from UKBB: "*chr_all_degree2_unrelated.txt*". This contains the unrelated people for your analysis.

3.1.3. Duplicated SNPs

To find the duplicate SNPs, we use the SNP imputed data (which we will use for future GWAS).

```
for loop over the 1-22, X and Y chromosomes:
        plink2 --freq --threads 16 --bgen ${bgen} ref-first --sample ${sample} --keep ${keep} --maf 0.01 --geno
0.03 --hwe 1e-10 --rm-dup --make-bed --export bgen-1.2 'bits=8' --out ${output_file}
```

This will generate a list of duplicated SNPs for each chromosome.

3.1.4. Apply all filters to SNPs

We first run all QC filters to each chromosome, excluding duplicate SNPs, excluding related individuals, and other standard filters. See the exact parameters below in the command line.

```
for loop over the 1-22, X and Y chromosomes:
        plink2 --freq --threads 16 --bgen ${bgen} ref-first --sample ${sample} --keep ${keep} --maf 0.01 --geno
0.03 --hwe 1e-10 --mind 0.03 --rm-dup --make-bed --export bgen-1.2 'bits=8' --out ${output_file} --exclude
${snp_mismatch_removed}
```

We then merge the preprocessed data of each chromosome to a final binary plink file:

```
plink \
  --bed chr1_UKBB_unrelated_all_ancestry_geno_0.03_maf_0.01_mind_0.03_hwe_1e-10.bed \
  --bim chr1_UKBB_unrelated_all_ancestry_geno_0.03_maf_0.01_mind_0.03_hwe_1e-10.bim \
  --fam chr1_UKBB_unrelated_all_ancestry_geno_0.03_maf_0.01_mind_0.03_hwe_1e-10.fam \
  --merge-list list_beds.txt \
  --make-bed --out chr_all
```

3.1.5. PCA

We need genetic PCs to control population stratification. UKBB provided pre-computed PCs that can be readily used. Instead, you can extract these PCs by yourself tailored to your study population. Here, we demonstrate how to do so with SNP array data.

First, we need to do a standard QC for the SNP array data. We first run all QC filters to each of the chromosomes, excluding duplicate SNPs, excluding related individuals, and other standard filters. See the exact parameters below in the command line.

```
for loop over the 1-22, X and Y chromosomes:
        plink --threads 16 --bed ${bed} --fam ${fam} --bim ${bim} --maf 0.01 --geno 0.1 --hwe 1e-15 --mind 0.1
--keep ${keep} --list-duplicate-vars suppress-first --make-bed --out ${output_file}
```

We then merge the preprocessed data of each chromosome to a final plinke binary file:

```
plink \
  --bed chr1_UKBB_genotyped_unrelated_all_ancestry_geno_0.1_maf_0.01_mind_0.1_hwe_1e-15.bed \
  --bim chr1_UKBB_genotyped_unrelated_all_ancestry_geno_0.1_maf_0.01_mind_0.1_hwe_1e-15.bim \
  --fam chr1_UKBB_genotyped_unrelated_all_ancestry_geno_0.1_maf_0.01_mind_0.1_hwe_1e-15.fam \
  --merge-list list_beds.txt \
  --make-bed --out chr_all
```

Next, we do pruning using Plink for the merged SNP arrays.

```
plink --bfile /cbica/home/wenju/Dataset/UKBB/UKBB_genetic_preprocess/S3_apply_all_Calls/chr_all
--indep-pairwise 1000 50 0.05
plink --bfile /cbica/home/wenju/Dataset/UKBB/UKBB_genetic_preprocess/S3_apply_all_Calls/chr_all --extract
plink.prune.in --make-bed --out chr_all_calls_pruning
```

Finally, we use FlashPCA software to compute the first 50th PCs. For this, please refer to the R package of FlashPCA: https://github.com/gabraham/flashpca