$--make-pheno$: How does it work

Zhang Leyi

Supervisor: Doug Speed

2023-11

Contents

1 Example Commands

```
${dir_LDAK} \
  --make-phenos ${dir_RA}/data/makepheno/Trait_1 \
  --bfile ${dir_data}/geno \
  --weights ${dir_RA}/data/geno_weighting_thin.thin \
  --power -0.25 \
  --her 0.9 \
  --num-phenos 5 \
  --num-causals 50000 \
  --extract ${dir_RA}/data/snps_1_to_12_geno.txt
```

2 Required options

$--bfile$: the genotype files, with .bed/.bim/.fam

$--weights$: the predictor weightings, serves as w_j in the following equation.

$$E[h_j^2\ ] = w_j I_j [f_j(1-f_j)]^(1+\ ) \tag{1}$$

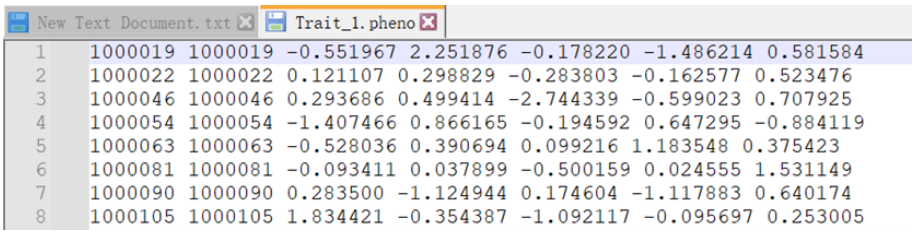
–power: to specify how the predictors are scaled, as in the above equation.

–her: to specify the heritability for the simulated phenotype, serving as $E[h_j^2]$.

–number-phenos: the number of phenotypes to generate at once.

–num-causals: to specify the number of SNP predictors contributing to the phenotype.

–extract: a file with a list of SNP ID, and removes all unlisted variants from the current analysis. In this analysis, the SNPs in the first half chromosomes are genetically related to the phenotype, and the second half are non-genetically related.



New Text Document.txt | Trait_1.pheno

```
1000019 1000019 -0.551967 2.251876 -0.178220 -1.486214 0.581584
1000022 1000022 0.121107 0.298829 -0.283803 -0.162577 0.523476
1000046 1000046 0.293686 0.499414 -2.744339 -0.599023 0.707925
1000054 1000054 -1.407466 0.866165 -0.194592 0.647295 -0.884119
1000063 1000063 -0.528036 0.390694 0.099216 1.183548 0.375423
1000081 1000081 -0.093411 0.037899 -0.500159 0.024555 1.531149
1000090 1000090 0.283500 -1.124944 0.174604 -1.117883 0.640174
1000105 1000105 1.834421 -0.354387 -1.092117 -0.095697 0.253005
```

(a)

Figure 1: **Example of a phenotype file.**

As shown in Fig.1, a phenotype file contains FID, IID and Phenotype values. For a simulated

phenotype, the values are standardized.

$$Y \ N(0, \delta^2 \ ) = \sum_j X_j \beta_j + e \tag{2}$$

Heritability explains how much percentage of phenotype is explained by SNPs, not the environment. So Y can also be written as:

$$Y = \sqrt{h^2} X\beta + \sqrt{1 - h^2} \times e \tag{3}$$

There are various heritability models to assume the heritability h^2, of which the simplest is called GCTA model. The command –her here provides the number of τ_1, and GCTA takes that:

$$E[h_j^2] = \tau_1 \tag{4}$$

While there is another complex model called LDAK-thin, which computes $E[h_j^2]$:

$$E[h_j^2] = \tau_1 I_j (f_j(1 - f_j))^2 \tag{5}$$

Every parameter on the right is scalar.

3 Implementation by python

Shown in the following link: Link to –make-pheno Code

1. About input data:

n_inds: number of individuals, 5 as default.

n_snps: number of snps in one individual, 10 as default.

tau: $E[h^2] = \tau$ in GCTA model.

2. Some parameters and functions:

genotype: randomly chosen among 0, 1, 2, making a matrix with $n\_inds \times n\_snps$.

effects: genetic effects, normally distributed, a vector of size n_snps.

genetics: genetic contribution to the phenotype, np.dot(genotype,effects).

environments: environmental contribution to the phenotype, normal distribution. Both *genetics* and *environments* are of size n_inds.

her: heritability, measuring how much percentage of phenotype is contributed by genetics, generally is written as h^2. In different models, h^2 is measured differently. In the GCTA model, the simplest

one, is $h^2 = \tau$.

LDAK-thin: There is another model to measure h^2, where it is necessary to calculate the LD and MAF. The formula is shown in Equ. (5).

Phenotype: The output phenotype, measuring as:

$$Phenotype = h^2 \times G + (1 - h^2) \times E \tag{6}$$

3. Details about LDAK-thin

W: LDAK weights, representing the LD. In LDAK-thin, the values are simplified to be binary as 0 or 1. I am constructing the function: Calc_LD now, so I arbitrarily make W binomial distributed.

MAF: shown as f in Equ. (5).

$$MAF = \frac{number \quad of \quad minor \quad alleles}{total \quad number \quad of \quad alleles} \tag{7}$$

4. Output: phenotype_file: A data frame contains 3 columns, which are FID, IID, and Phenotype.

References