CIAO-GWAS-1.2 user guide

Matthew Traylor
matthew.traylor@kcl.ac.uk

Prerequisities

CIAO-GWAS-1.2 will not function correctly without the following:

1. You must have R installed and R must be called by typing “R”.
2. The “calculate_liabilities.R” and “liability_assoc.R” or “liability_assoc-mach.R” files must be in your current working directory.
3. The “R_liability-assoc” or “R_liability-assoc-mach” and “R_liability_calc” files must be in your path
   (to find your path type “echo $PATH”. Copy the files to any of the given directories, or add a directory to your path with PATH=$PATH:/yourdirectory/)

IMPUTE format data

For analysis of SNPTEST / IMPUTE format data (.gen and .sample files), some manipulation of the .sample file is first required. The sample file should include ONLY the following columns:
ID_1 ID_2 sex phenotype liability_value(age) covariates

IMPORTANT: the 2nd line of the sample file (the 0 0 0 D P line) **MUST** be removed (e.g with awk 'NR!=2{print}' file.sample > newfile.sample)

You will need to create a new sample file for each phenotype analysed. As in a SNPTEST analysis, the order of the rows must not be changed as they must correspond to specific columns in the gen file.

Examples of the file formats used are given in the example.gen and example.sample and files.

Examples of how to perform basic analyses are given below, and can be tested using the files provided in the software tarball:

1. Calculating liabilities

--liability-calc mode

basic analysis

```
ciao-gwas-1.2 --liability-calc --samplefile example1.sample --
parameter stroke-subtype-men.out --out example1.liabilities
```

analysis separated by gender (or any binary variable)

```
ciao-gwas-1.2 --liability-calc --samplefile example1.sample --
gender --parameter-female stroke-subtype-women.out --parameter-
male stroke-subtype-men.out --out example1.liabilities
```

2. Association analysis

--assoc mode

options:

--chunk *N* controls number of SNPs read into R at one time. A larger value speeds up analysis at the expense of extra memory usage. Default value of 50,000 should be suitable for most datasets.

--covariates determines whether covariates included in the sample file should be included in the analysis

--line-count N Gives total number of SNPs in file. For gzipped data, this flag MUST be included

basic analysis

```
ciao-gwas-1.2 --assoc --genfile example1.gen --samplefile
example1.sample --liabilities example1.liabilities --out
example1.out --chunk 101
```

analysis including covariates

```
ciao-gwas-1.2 --assoc --genfile example1.gen --samplefile
example1.sample --liabilities example1.liabilities --out
example1.out --chunk 101 --covariates
```

basic analysis (gzipped data)

```
ciao-gwas-1.2 --assoc --genfile example1.gen.gz --samplefile
example1.sample --liabilities example1.liabilities --out
example1.out --chunk 101 --line-count 101
```

MACH format data

For analysis of MaCH format data .info, .dose and .pheno files are used.The pheno file should include ONLY the following columns:
ID_1 ID_2 sex phenotype liability_value(age) covariates

Examples of the file formats used are given in the example.pheno, example.mldose and example.mlinfo files.

You will need a different phenotype file for each analysis. As in a ProbABEL analysis, the order of the rows in the pheno file must not be changed as they must correspond to specific rows in the dose file.

Examples of how to perform basic analyses are given below and can be tested using the files provided in the software tarball:

1. Calculating liabilities

--liability-calc mode

basic analysis

```
ciao-gwas-1.2 --liability-calc --samplefile example.pheno --
parameter stroke-subtype-men.out --out example.liabilities
```

analysis separated by gender (or any binary variable)

```
ciao-gwas-1.2 --liability-calc --samplefile example.pheno --gender
--parameter-female stroke-subtype-women.out --parameter-male
stroke-subtype-men.out --out example.liabilities
```

2. Association analysis

--assoc mode

options:

--chunk *N* controls number of SNPs read into *R* at one time. A larger value speeds up analysis at the expense of extra memory usage. Default value of 50,000 should be suitable for most datasets.

--covariates determines whether covariates included in the sample file should be included in the analysis

basic analysis

```
ciao-gwas-1.2 --assoc --mldose example.mldose --mlinfo
example.mlinfo --mlpheno example.pheno --liabilities
example.liabilities --out example.out --chunk 101
```

analysis including covariates

```
ciao-gwas-1.2 --assoc --mldose example.mldose --mlinfo
example.mlinfo --mlpheno example.pheno --liabilities
example.liabilities --out example.out --chunk 101 --covariates
```

Example SGE cluster script – stroke subtype analysis

If *R* is installed to `/apps/R/2.15.1/bin/` and the "ciao-gwas-1.2-gwas-1.0", "R_liability-assoc" or "R_liability-assoc-mach" and "R_liability_calc" are in a directory named `/home/traylorm/apps` then the following script can be used to submit the analysis to an SGE cluster:

```
#!/bin/sh
#$ -S /bin/sh
#$ -cwd

PATH=$PATH:/home/traylorm/apps:/apps/R/2.15.1/bin/
export PATH

ciao-gwas-1.2 --liability-calc --samplefile example.pheno --gender
--parameter-female stroke-subtype-women.out --parameter-male
stroke-subtype-men.out --out example.liabilities

ciao-gwas-1.2 --assoc --mldose example.mldose --mlinfo
example.mlinfo --mlpheno example.pheno --liabilities
example.liabilities --out example.out --chunk 101 --covariates
```