HE

**Real-check**

Real-check helps to identify the same individuals in terms of their genotypic similarity. The individuals in two files have similarity score within the range of the upper bounder and the lower bounder will be lined up and output. Only overlapping markers, which have the same marker names and the same reference alleles, are used for further analysis.

Caution:

Strand issues for ambiguous SNPs which have A/T and G/C alleles could not be distinguished in the analysis.

**Options**

--realcheck

This is the general command for invoking realcheck option.

--bfile2

Specifying the second file in real-check operation.

--realcheck-threshold-upper

The upper bounder of similarity for realcheck, the default value is 1.

--realcheck-threshold-lower

The lower bounder of similarity for realcheck, the default value is 0.

--realcheck-snps

It specifies the SNPs that are used for generate similarity scores. Only available SNPs will be used for generating similarity scores.

--realcheck-marker-number

The number of markers, which is randomly sampled from the genotype files, is used to construct the similarity scores. By default, 100 markers will be randomly picked. By selecting another set of markers, the user can choose another seed for random number --seed.

Example:

HE --realcheck --bfile cd0 --bfile2 uc0 --realcheck-threshold-lower 0.9 --realcheck-marker-number 100 --out cd-uc

HE --realcheck --bfile cd0 --bfile2 uc0 --realcheck-threshold-lower 0.9 --realcheck-marker-number 100 --seed 2000 --out cd-uc

If only one pedigree file is specified, the genetic relationship can be checked without the second one.

HE --realcheck --bfile cd0 --realcheck-threshold-lower 0.9 --realcheck-marker-number 100 –out cd0

**Merge two files**

This option merges two files together.

--merge

this is the general option to invoke the procedure.

--bfile2

the second file for merging

--merge-maf-cutoff

the option to specify the maf cutoff when combining ambiguous alleles together. The maf exceeds the cutoff will be eliminated.

--merge-p-cutoff

the option to specify the p-value cutoff when combining alleles together. The p-value exceeds the cutoff will be eliminated.

--keep-atgc

this option will keep biallelic loci that have A/T or G/C.

--remove-flip

this option will remove biallelic loci that have flipped coding SNPs, such as coded as A/G in the first file but to T/C the second.

Example

HE --realcheck --bfile cd0 --realcheck-threshold-lower 0.9 --realcheck-marker-number 100 --out cd0

output

\*.mergesnp

\*.mergebadsnp

One of the obstacles in merging binary genotypes together is strand issue that causes coding ambiguous SNPs, such as A/T, G/C pairs of SNPs at a locus. Assuming that A1, with frequency denoted as ref1, and A2 are the reference and the second alleles in the first binary genotype file and B1, with frequency denoted as ref2 and B2 for the second binary file, their matching schemes can be summarized as below

|  |  |  |
| --- | --- | --- |
| SNP Alleles at the first panel | |  |
|  | SNP alleles at the second panel | | | | | |
|  | AC | AG | AT | CG | CT | GT |
|  |  |  |  |  |  |  |
| Biallelic | AC |  |  |  |  |  |  |  |
| AG |  |  |  |  |  |  |  |
| AT |  |  |  |  |  |  |  |
| CG |  |  |  |  |  |  |  |
| CT |  |  |  |  |  |  |  |
| GT |  |  |  |  |  |  |  |

The match up schemes can be illustrated as the table above, and the ambiguity of merging SNPs lays in the cells in red. The diagonal matchups are summarized in sheme1 and sheme2 below. The green cells along the anti-diagonal are summarized in scheme3 and scheme4.

|  |  |  |  |
| --- | --- | --- | --- |
| Schemes | Matched allele frequency | Subtype | Flip B |
| Scheme1  A1 matches B1 directly. | If the polymorphism of A is neither A/T nor G/C. |  | No conversion for B1 to B2, accuracy 1. |
|  | If the polymorphism of A is A/T or G/C.  Ref1<0.5 and ref2<0.5, | If (ref1 < p-cutoff & ref2 < p-cutoff) accept this SNP | No conversion, and its accuracy p1\*p2 |
|  | If the polymorphism of A is A/T or G/C.  Ref1<0.5 and ref2>0.5, | If (ref1 < p-cutoff & ref2 > 1 - p-cutoff) accept this SNP, | Convert B1 to B2, and its accuracy p1\*p2;  ref2=1-ref2. |
|  | If the polymorphism of A is A/T or G/C.  Ref1>0.5 and ref2<0.5, | If (ref1 > 1 - p-cutoff and ref2 < p-cutoff) accept this SNP, | Convert B1 to B2, and its accuracy p1\*p2;  ref2=1-ref2. |
|  | If the polymorphism of A is A/T or G/C.  Ref1>0.5 and ref2>0.5, | If (ref1 > 1 - p-cutoff and ref2 > 1 - p-cutoff) accept this SNP, | No conversion, and its accuracy p1\*p2 |
|  |  |  |  |
| Scheme2  A1 matches B2 directly. | If the polymorphism of A is neither A/T nor G/C. |  | Conversion for B1 to B2, accuracy 1. |
|  | If the polymorphism of A is A/T or G/C.  Ref1<0.5 and (1-ref2)<0.5, | If (ref1 < p-cutoff & (1-ref2) < p-cutoff) accept this SNP | Flip, and its accuracy p1\*p2;  ref2=1-ref2. |
|  | If the polymorphism of A is A/T or G/C.  Ref1<0.5 and (1-ref2)>0.5, | If (ref1 < p-cutoff & (1-ref2) > 1 - p-cutoff) accept this SNP, | Without flip, and its accuracy p1\*p2 |
|  | If the polymorphism of A is A/T or G/C.  Ref1>0.5 and (1-ref2)<0.5, | If (ref1 > 1 - p-cutoff and (1-ref2) < p-cutoff) accept this SNP, | Without flip, and its accuracy p1\*p2 |
|  | If the polymorphism of A is A/T or G/C.  Ref1>0.5 and (1-ref2)>0.5, | If (ref1 > 1 - p-cutoff and (1-ref2) > 1 - p-cutoff) accept this SNP, | Flip, and its accuracy p1\*p2;  ref2=1-ref2. |
| Scheme3  A1 matches flipped B1. |  |  | No conversion, and its accuracy p1\*p2 |
| Scheme4  A1 matches flipped B2. |  |  | Conversion.  ref2=1-ref2. |

It should be noted that the matching scheme is from scheme 1 to scheme 4, otherwise mismatch will occur. Pi is defined as p(refi<0.5|refi<0.5) or p(refi>0.5|refi>0.5).

Furthermore, for further quality control, if the statistical difference between ref1 and ref2 exceeds predefined p-cutoff, the SNP will be eliminated too.

The the test statistic is defines as

, se=, .

Make predictor panel

--build-predictor

--build-predictor2

--predictor-file

The first two columns are SNP and the reference allele, if the predictor; if build-predictor2 is used, the first five columns are SNP, A1, A2, MAF, number of individual.

--predictor-idx

default 1

--keep-atgc

The default option is to remove AT/GC loci.

--remove-flip

The default option is to keep flip SNPs.

--logit

--linear

The default option is linear.

Example

--build-predictor --predictor-file pgc.txt --predictor-idx 1 --logit --out

**Haseman-Elston Regression**

This session estimated the additive variance components based on the generic relationship matrix.

--he

This is the general option to invoke Haseman-Elston Regression.

--grm

This option looks for a pair of files: file.grm.gz, file.grm.id. The grm files can be generated with GCTA.

--pheno

It specifies the phenotype file. The format of the file is family id, individual id, and phenotypes.

--mpheno

It specifies the default phenotype index that is used in Haseman-Elston regression. By default, the first phenotype listed in file specified in --pheno will be used. If more than one phenotypes are specified—in reversed Haseman-Elston regression, comma delimits indexes. By default, mpheno=1.

--reverse

If want to make grm regress on the phenotypes, this option should be pronounced.

Various kinds of Haseman-Elston regression

There are three kinds of Haseman-Elston regressions.

--he-sd

It models the phenotype as

--he-ss

It models the phenotype as

--he-cp

It models the phenotype as

Example

--he --grm test --pheno test.phe --mpheno 1 --out he-test

--he --grm test --pheno test.phe --mpheno 1,2 --reverse --out he-test

--he-cp --grm test --pheno test.phe --mpheno 1 --out he-test

--he-ss --grm test --pheno test.phe --mpheno 1 --out he-test

**Simulation**

--simu-poly-cc

this option invokes polygenic simulation for generating case-control sample.

--poly-loci

Specify the number of causal loci, which is 1000 by default.

--poly-ld

Specify the fraction for LD, which is 1 implicating the simulated loci are causal.

--poly-U

If want to the effects to be uniformly distributed, turn this option on; otherwise, the additive effects follow a normal distribution N(0,), in which is the heritability and N is the number of loci.

--simu-cc

Specify the number of cases, which by default is 500.

--simu-k

The prevalence of the cases in the population. By default k=0.05.

--simu-hsq

Specify the heritability of the trait. By default the heritability is 0.5 under the liability scale.

Example

--simu-poly-cc --poly-loci 100 --simu-cc 500 500 --simu-k 0.01 --simu-hsq 0.8 --out poly

Output

\*.cov

\*.ped

\*.map

\*.phe

\*.plog

\*.rnd

Prediction for dosage scores

It reads MaCH dosage scores to predict risk profiles.

**Single file format**

HE --score discover\_panel.txt --mach-dosage mach.mldosage.gz --mach-infor mach.mlinfo

The score file discover\_panel.txt it has three fields

This file has the format of one or more lines, each with exactly three fields

SNP ID Reference allele Score (numeric)

for example

SNPA A 1.95

SNPB C 2.04

SNPC C -0.98

SNPD C -0.24

The dosage file mach.mldosage.gz has the format below

Fam->IID mode dosage\_RefAllele\_locus1 dosage\_RefAllele\_locus2 dosage\_RefAllele\_locus3

FID->IID1 ML\_Dose 2.00 0.25 1.3

FID->IID2 ML\_Dose 1.50 1.3 1.6

The information file has the format below

SNP Al1 Al2 Freq1 MAF Quality Rsq

RS001 A C 0.2 0.2 0.9 0.94

Rs002 T C 0.7 0.3 0.83 0.87

**Batch format**

HE --score discover\_panel.txt --mach-dosage-batch mach\_dose.txt --mach-infor-batch mach\_infor.txt

The mach-dosage-batch file mach\_dose.txt has the list of dosage files:

Mach\_chr1.mldosage.gz

Mach\_chr2.mldosage.gz

The mach-infor-batch file mach\_infor.txt has the list of information files:

Mach\_chr1.mlinfo

Mach\_chr2.mlinfo

**Multiple scores from SNP subsets**

To calculate multiple scores from subsets of SNPs in a single --score file, it is possible to use the two commands, each followed by a filename, e.g.

--q-score-file snpval.dat --q-score-range q.ranges

in addition to --score, where snpval.dat is a file that contains for each SNP a number (e.g. that might be the p-value from some test)

rs00001 0.234

rs00002 0.046

rs00003 0.887

...

and q.ranges is a file in which each row corresponds to a different score, containing a label, then a lower and upper bound for the values as given in the other file, e.g.

S1 0.00 0.01

S2 0.00 0.20

S3 0.10 0.50

**For single file the output file is out.profile**

**FID IID score**

**0 1 0.009**

**1 1 -0.008**

**For multiple scores from SNP substes, the output file**

FID IID score.S1 score.S2 score.s3 …

0 1 0.009 0.003 0.002 …

1 2 -0.008 0.001 -0.002 …