GCTA: a tool for Genome-wide Complex Trait Analysis

Version 1.04, 13 Sep 2012

Overview

GCTA (Genome-wide Complex Trait Analysis) is designed to estimate the proportion of phenotypic variance explained by genome- or chromosome-wide SNPs for complex traits. GCTA was developed by Jian Yang, Hong Lee, Mike Goddard and Peter Visscher and is maintained in Peter Visscher's lab at the University of Queensland. GCTA currently supports the following functionalities:

- Estimate the genetic relationship from genome-wide SNPs;
- Estimate the inbreeding coefficient from genome-wide SNPs;
- Estimate the variance explained by all the autosomal SNPs;
- Partition the genetic variance onto individual chromosomes;
- Estimate the genetic variance associated with the X-chromosome;
- Test the effect of dosage compensation on genetic variance on the X-chromosome;
- Predict the genome-wide additive genetic effects for individual subjects and for individual SNPs;
- Estimate the LD structure encompassing a list of target SNPs;
- Simulate GWAS data based upon the observed genotype data;
- Convert Illumina raw genotype data into PLINK format;
- Conditional & joint analysis of GWAS summary statistics without individual level genotype data

Questions and Help Requests

If you have any bug reports or questions please send us an email at jian.yang@uq.edu.au

Citations

Method for estimating the variance explained by all SNPs with its applicaiton in human height:

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010 Jul 42(7): 565-9. [PubMed ID: 20562875]

Method for estimating the variance explained by all SNPs being extended for case-control design with its application to the WTCCC data:

Lee SH, Wray NR, Goddard ME and Visscher PM. Estimating Missing Heritability for Disease from Genome-wide Association Studies. Am J Hum Genet. 2011 Mar 88(3): 294-305. [PubMed ID: 21376301]

Method for partitioning the genetic variance captured by all SNPs onto chromosomes and genomic segments with its applications in height, BMI, vWF and QT interval:

Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, de Andrade M, Feenstra B, Feingold E, Hayes MG, Hill WG, Landi MT, Alonso A, Lettre G, Lin P, Ling H, Lowe W, Mathias RA, Melbye M, Pugh E, Cornelis MC, Weir BS, Goddard ME, Visscher PM: Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet. 2011 Jun 43(6): 519-525. [PubMed ID: 21552263]

Method for conditional and joint analysis using summary statistics from GWAS with its application to the GIANT meta-analysis data for height and BMI:

Yang J, Ferreira T, Morris AP, Medland SE; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Madden PA, Heath AC, Martin NG, Montgomery GW, Weedon MN, Loos RJ, Frayling TM, McCarthy MI, Hirschhorn JN, Goddard ME, Visscher PM (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies

additional variants influencing complex traits. Nat Genet. Mar 18 44(4): 369-375. [PubMed ID: 22426310]

Software tool:

Yang J, Lee SH, Goddard ME and Visscher PM. GCTA: a tool for Genome-wide Complex Trait Analysis. Am J Hum Genet. 2011 Jan 88(1): 76-82. [PubMed ID: 21167468]

OPTIONS (case sensitive)

1. Input and output

--bfile *test*

Input **PLINK** binary PED files, e.g. **test.fam**, **test.bim** and **test.bed** (see **PLINK** user manual for details).

--dosage-mach *test.mldose.gz* *test.mlinfo.gz*

Input files in **MACH** output format (compressed), e.g. **test.mldose.gz** and **test.mlinfo.gz** (see **MACH** user manual for details).

--out *test*

Specify output root filename.

2. Data management

--keep *test.indi.list*

Specify a list of individuals to be included in the analysis.

--remove *test.indi.list*

Specify a list of individuals to be excluded from the analysis.

--chr 1

Include SNPs on a specific chromosome in the analysis, e.g. chromosome 1.

--autosome-num *22*

Specify the number of autosomes for a species other than human. For example, if you specify the number of autosomes to be 19, then chromosomes 1 to 19 will be recognized as autosomes and chromosome 20 will be recognized as the X chromosome. The default number is 22 if this option not specified.

--autosome

Include SNPs on all of the autosomes in the analysis.

--extract *test.snplist*

Specify a list of SNPs to be included in the analysis.

Input file format
test.snplist
rs103645
rs175292
……

--exclude *test.snplist*

Specify a list of SNPs to be excluded from the analysis.

--extract-snp *rs123678*

Specify a SNP to be included in the analysis.

--exclude-snp *rs123678*

Specify a single SNP to be excluded from the analysis.

--maf *0.01*

Exclude SNPs with minor allele frequency (MAF) less than a specified value, e.g. 0.01.

--max-maf *0.1*

Include SNPs with MAF less than a specified value, e.g. 0.1.

--update-sex *test.indi.sex.list*

Update sex information of the individuals from a file.

Input file format

test.indi.sex.list (no header line; columns are family ID, individual ID and sex). Sex coding: "1" or "M" for male and "2" or "F" for female.

```
011        0101    1
012        0102    2
013        0103    1
……
```

--update-ref-allele *test_reference_allele.txt*

Assign a list of alleles to be the reference alleles for the SNPs included in the analysis. By default, the first allele listed in the ***.bim** file (the 5th coloumn) or ***.mlinfo.gz** file (the 2nd conlumn) is assigned to be the reference allele. **NOTE**: This option is invalid for the imputed dosage data only.

Input file format

test_reference_allele.txt (no header line; columns are SNP ID and reference allele)

```
rs103645   A
rs175292   G
……
```

--imput-rsq *0.3*

Include SNPs with imputation R^2 (squared correlation between imputed and true genotypes) larger than a specified value, e.g. 0.3.

--update-imput-rsq *test.imput.rsq*

Update imputation R^2 from a file. For the imputed dosage data, you do not have to use this option because **GCTA** can read the imputation R^2 from the ***.mlinfo.gz** file unless you want to write them. For the best guess data (usually in **PLINK** format), if you want to use a R^2 cut-off to filter SNPs, you need to use this option to read the imputation R^2 values from the specified file.

Input file format

test.imput.rsq (no header line; columns are SNP ID and imputation R^2)

```
rs103645   0.976
rs175292   1.000
```

……

--freq

Output allele frequencies of the SNPs included in the analysis (in plain text format), e.g.

Output file format

test.freq (no header line; columns are SNP ID, reference allele and its frequency)

rs103645 A 0.312

rs175292 G 0.602

……

--update-freq *test.freq*

Update allele frequencies of the SNPs from a file rather than calculating from the data. The format of the input file is the same as the output format for the option --freq.

--recode

Output the SNP genotypes in additive coding (in compressed text format), e.g. **test.xmat.gz.**

--recode-nomiss

Output the SNP genotypes in additive coding, and fill the missing genotype by its expected value i.e. 2*p* where *p* is the frequency of the reference allele.

Output file format

test.xmat.gz (The first two lines are header lines. The first line contains headers of family ID, individual ID and names of SNPs. The second line contains two nonsense words "Reference Allele" and the reference alleles of the SNPs. Any missing genotype is represented by "NA" unless the option --recode-nomiss is specified, for which the missing genotype will be assigned by 2*p*).

FID	IID	rs103645	rs175292
Reference	Allele	A	G
011	0101	1	0
012	0102	2	NA
013	0103	0	1

……

--make-bed

Save the genotype data in PLINK binary PED files (*.fam, *.bim and *.bed).

Example

 # Convert MACH (or Minimac) dosage data to PLINK binary PED format

gcta --dosage-mach test.mldose.gz test.mlinfo.gz --make-bed --out test

3. Estimation of the genetic relationships from the SNPs

--make-grm

Estimate the genetic relationship matrix (GRM) between pairs of individuals from a set of SNPs. By default, **GCTA** will save the lower triangle of the genetic relationship matrix in a compressed text file (e.g. **test.grm.gz**) and save the IDs in a plain text file (e.g. **test.grm.id**).

Output file format

test.grm.gz (no header line; columns are indices of pairs of individuals (row numbers of the ***test.grm.id***), number of non-missing SNPs and the estimate of genetic relatedness)

```
1   1   1000   1.0021
2   1    998   0.0231
2   2    999   0.9998
3   1   1000   -0.0031
……
```

test.grm.id (no header line; columns are family ID and individual ID)

```
011    0101
012    0102
013    0103
……
```

--make-grm-xchr

Estimate the GRM from SNPs on the X-chromosome. The GRM will be saved in the same format as above. Due to the speciality of the GRM for the X-chromosome, it is not recommended to manipulate the matrix by --grm-cutoff or --grm-adj, or merge it with the GRMs for autosomes (see below for the options of manipulating the GRM).

--make-grm-bin

Save the GRM in binary format with size of 8 bytes (e.g. **test.grm.bin**) and the IDs in a text file as decribed above. This option is useful when you want to read the GRM in R because it more efficient for R to read the binary file than the compressed text file.

--save-ram

Run the program in RAM-saving mode, but use more CPU time. This option is valid for estimating the GRM only.

--ibc

Estimate the inbreeding coefficient from the SNPs by 3 different methods (see the software paper for details).

Output file format

test.ibc (one header line; columns are family ID, individual ID, number of nonmissing SNPs, estimator 1, estimator 2 and estimator 3)

FID	IID	NOMISS	Fhat1	Fhat2	Fhat3
011	0101	999	0.00210	0.00198	0.00229
012	0102	1000	-0.0033	-0.0029	-0.0031
013	0103	988	0.00120	0.00118	0.00134
……					

Examples

 # Estimate the GRM from all the autosomal SNPs

gcta --bfile test --autosome --make-grm --out test

 # Estimate the GRM from the SNPs on the X-chromosome

gcta --bfile test --make-grm-xchr --out test_xchr

 # Estimate the GRM from the SNPs on chromosome 1 with MAF from 0.1 to 0.4

gcta --bfile test --chr 1 --maf 0.1 --max-maf 0.4 --make-grm --out test

 # Estimate the GRM using a subset of individuals and a subset of autosomal SNPs with MAF < 0.01

gcta --bfile test --keep test.indi.list --extract test.snp.list --autosome --maf 0.01 --make-grm --out test

 # Estimate the GRM from the imputed dosage scores for the SNPs with MAF > 0.01 and imputation R^2 > 0.3

gcta --dosage-mach test.mldose.gz test.mlinfo.gz --imput-rsq 0.3 --maf 0.01 --make-grm --out test

 # Estimate the GRM from the imputed dosage scores for a subset of individuals and a subset of SNPs

gcta --dosage-mach test.mldose.gz test.mlinfo.gz --keep test.indi.list --extract test.snp.list --make-grm --out test

 # Estimate the inbreeding coefficient from all the autosomal SNPs

gcta --bfile test --autosome --ibc --out test

4. Manipulation of the genetic relationship matrix

--grm *test*

Input the GRM. This option actually tells **GCTA** to read two files, e.g. **test.grm.gz** and **test.grm.id** (See the option --make-grm). **GCTA** automatically adds suffix "**.grm.gz**" and "**.grm.id**" to the specified root filename.

--mgrm *multi_grm.txt*

Input multiple GRMs. The root filenames of multiple GRMs are given in a file, e.g. **multi_grm.txt**

Input file format

multi_grm.txt (full paths can be specified if the GRM files are in different directories)

```
test_chr1
test_chr2
test_chr3
……
test_chr22
```

Example

This option is very useful to deal with large dataset. You can firstly run the jobs (split one job into 22 pieces)

```
gcta --bfile test --chr 1 --make-grm --out test_chr1
gcta --bfile test --chr 2 --make-grm --out test_chr2
…
gcta --bfile test --chr 22 --make-grm --out test_chr22
```

to estimate the GRMs from the SNPs on each chromosome, then merge them by the command

```
gcta --mgrm multi_grm.txt --make-grm --out test
```

--grm-cutoff *0.025*

Remove one of a pair of individuals with estimated relatedness larger than the specified cut-off value (e.g. 0.025). **GCTA** selectively removes individuals to maximize the remaining sample size rather than doing it at random. **NOTE**: When merging multiple GRMs, this option does not apply to each single GRM but to the final merged GRM.

--grm-adj *0*

When using the SNPs to predict the genetic relationship at causal loci, we have to adjust the prediction errors due to imperfect LD because of two reasons: 1) the use of only a finite number of SNPs; 2) causal loci tend to have lower MAF than the genotyped SNPs (input 0 if

you assume that the causal loci have similar distribution of allele frequencies as the genotyped SNPs) (see Yang et al. 2010 Nat Genet for details).

--dc *1*

By default, the GRM, especially for the X-chromosome, is parameterized under the assumption of equal variance for males and females, unless the option --dc is specified (1 and 0 for full and no dosage compensation, respectively). **You need to use the option --update-sex to read sex information of the individuals from a file (see the --update-sex option above).**

NOTE: you can add the option --make-grm afterwards to save the modified GRM. You can also use the option --keep and/or --remove in combination with these five commands. It is also possible to use these five commands in the REML analysis (see the section below).

Examples

```
# Prune the GRM by a cutoff of 0.025 and adjust for prediction errors assuming the causal variants have similar distribution of allele frequencies as the genotyped SNPs)
gcta  --grm test  --grm-adj  0  --grm-cutoff  0.025  --make-grm  --out test_adj
# Use --keep or --remove option
gcta  --grm test  --keep test.indi.list  --grm-cutoff  0.025  --make-grm  --out test_adj
gcta  --grm test  --remove test.indi.list  --grm-adj 0  --make-grm  --out test_adj
# Assume full and no dosage compensation for the X chromosome
gcta  --grm test_xchr  --dosage-compen 1  --update-sex test.indi.sex.list  --make-grm  --out test_xchr_fdc
gcta  --grm test_xchr  --dosage-compen 0  --update-sex test.indi.sex.list  --make-grm  --out test_xchr_ndc
```

5. Principal component analysis

--pca *20*

Input the GRM and output the first *n* (*n* = 20, by default) eigenvalues (saved as ***.eigenval**, plain text file) and eigenvectors (saved as ***.eigenvec**, plain text file), which are equivalent to those calcuated by the progrom **EIGENSTRAT**. The only purpose of this option is to calcuate the first *m* eigenvectors, and subsquently include them as covariates in the model when estimating the variance explained by all the SNPs (see below for the option of estimating the

variance explained by genome-wide SNPs). Please find the **EIGENSTRAT** software if you need more sophisticated principal component analysis of the population structure.

Output file format

test.eigenval (no header line; the first *m* eigenvalues)

```
20.436
7.1293
6.7267
……
```

test.eigenvec (no header line; the first *m* eigenvectors; columns are family ID, individual ID and the first *m* eigenvectors)

```
011    0101    0.00466824    -0.000947    0.00467529    -0.00923534
012    0102    0.00139304    -0.00686406    -0.0129945    0.00681755
013    0103    0.00457615    -0.00287646    0.00420995    -0.0169046
……
```

Examples

 # Input the GRM file and output the first 20 eigenvectors for a subset of individuals

```
gcta  --grm test --keep test.indi.list  --pca 20  --out test
```

6. Estimation of the phenotypic variance explained by the SNPs using the REML method

--reml

Perform REML (restricted maximum likelihood) analysis. This option is usually followed by the option --grm (one GRM) or --mgrm (multiple GRMs) to estimate the variance explained by the SNPs that were used to estimate the genetic relationship matrix.

--reml-priors *0.45 0.55*

Specify the starting values for REML iterations. The number of starting values specified should NOT be less than the number of variance components in the model. By default, GCTA will use equal variances of all the components as the starting values if this option is not specified.

--reml-alg *0*

Specify the algorithm to do REML iterations, 0 for average information (AI), 1 for Fisher-scoring and 2 for EM. The default option is 0, i.e. AI-REML, if this option is not specified.

--reml-no-constrain

By default, if an estimate of variance component escapes from the parameter space (i.e. negative value), it will be set to be a small positive value that is $V_P \times 10^{-6}$ with V_P being the phenotypic variance. If the estimate keeps on escaping from the parameter space, the estimate will be constrained to be $V_P \times 10^{-6}$. If the option --reml-no-constrain is specified, the program will allow an estimate of variance component to be negative, which may result in the estimate of proportion variance explained by all the SNPs > 1.

--reml-maxit *100*

Specify the maximum number of iterations. The default number is 100 if this option is not specified.

--pheno *test.phen*

Input phenotype data from a plain text file, e.g. **test.phen**. If the phenotypic value is coded as 0 or 1, then it will be recognized as a case-control study (0 for controls and 1 for cases). Missing value should be represented by "-9" or "NA".

Input file format

test.phen (no header line; columns are family ID, individual ID and phenotypes)

```
011     0101   0.98
012     0102   -0.76
013     0103   -0.06
……
```

--mpheno *2*

If the phenotype file contains more than one trait, by default, **GCTA** takes the first trait for analysis (the third column of the file) unless this option is specified. For example, --mpheno *2* tells **GCTA** to take the second trait for analysis (the fourth column of the file).

--gxe *test.gxe*

Input an environmental factor from a plain text file, e.g. **test.gxe**. Apart from estimating the genetic variance, this command tells **GCTA** to estimate the variance of genotype-environment (GE) interaction. You can fit multiple environmental factors simultaneously. The main effects of an environmental factor will be included in the model as fixed effects and the GE interaction effects will be treated as random effects. **NOTE:** the design matrix of the overall mean in the model (which is a vector of all ones) is always a linear combination of the design matrix of a discrete environmental factor so that not all the main effects (fixed effects) are estimable. GCTA will always constrain the main effect of the first level to be zero and the main effect of any other level represents its difference in effect compared to the first level. For example, if you fit sex as an environmental factor, GCTA will fit only one main effect in the model, i.e. the mean difference between males and females.

Input file format

test.gxe (no header line; columns are family ID, individual ID and environmental factors)

```
01    0101   F    smoker
02    0203   M    nonsmoker
03    0305   F    smoker
……
```

--covar *test.covar*

Input discrete covariates from a plain text file, e.g. **test.covar**. Each discrete covariate is recognized as a categorical factor with several levels. The levels of each factor can be represented by a single character, word or numerical number. **NOTE:** the design matrix of the mean in the model (which is a vector of all ones) is always a linear combination of the design matrix of a discrete covariate so that not all the effects of the levels (or classes, e.g. male and female) of a discrete covariate are estimable. GCTA will always constrain the effect of the first level to be zero and the effect of any other level represents its difference in effect compared to the first level.

Input file format

test.covar (no header line; columns are family ID, individual ID and discrete covariates)

```
01    0101   F    Adult        0
02    0203   M    Adult        0
03    0305   F    Adolescent   1
……
```

--qcovar *test.qcovar*

Input quantitative covariates from a plain text file, e.g. **test.qcovar**. Each quantitative covariate is recognized as a continuous variable.

Input file format

test.qcovar (no header line; columns are family ID, individual ID and quantitative covariates)

```
01      0101    -0.024    0.012
02      0203    0.032     0.106
03      0305    0.143     -0.056
……
```

--reml-lrt *1*

Calculate the log likelihood of a reduce model with one or multiple genetic variance components dropped from the full model and calculate the LRT and p-value. By default, GCTA will always calculate and report the LRT for the first genetic variance component, i.e. --reml-lrt *1*, unless you re-specify this option, e.g. --reml-lrt *2* assuming there are a least two genetic variance components included in the analysis. You can also test multiple components simultaneously, e.g. --reml-lrt *1 2 4*. See **FAQ #1** for more details.

--reml-no-lrt

Turn off the LRT.

--prevalence *0.01*

Specify the disease prevalence for a case-control study. Once this option is specified, GCTA will transform the estimate of variance explained, V(1)/Vp, on the observed scale to that on the underlying scale, V(1)/Vp_L. The prevalence should be estimated from a general population in literatures rather than that estimated from the sample.

NOTE:

1. You do not have to have exactly the same individuals in these files. **GTCA** will find the individuals in common in the files and sort the order of the individuals.

2. Please be aware that if the GRM is estimated from the imputed SNPs (either "best guess" or "dosage score"), the estimate of variance explained by the SNPs will depend on the imputation-R^2 cutoff used to select SNPs because the imputation-R^2 is correlated with MAF, so that selection on imputation-R^2 will affect the MAF spectrum and thus affect the estimate of variance explained by the SNPs.
3. **For a case-control study, the phenotypic values of cases and controls should be specified as 1 and 0 (or 2 and 1, compatible with PLINK), respectively.**
4. Any missing value (either phenotype or covariate) should be represented by "-9" or "NA".
5. The summary result of REML analysis will be saved in a plain text file (***.hsq**).

Output file format

test.hsq (rows are

header line;

name of genetic variance, estimate and standard error (SE);

residual variance, estimate and SE;

phenotypic variance, estimate and SE;

ratio of genetic variance to phenotypic variance, estimate and SE;

log-likelihood;

sample size). If there are multiple GRMs included in the REML analysis, there will be multiple rows for the genetic variance (as well as their ratios to phenotypic variance) with the names of V(1), V(2), … .

```
Source  Variance        SE
V(1)    0.389350        0.161719
V(e)    0.582633        0.160044
Vp      0.971984        0.031341
V(1)/Vp 0.400573        0.164937
The estimate of variance explained on the observed scale is transformed to that on the underlying scale:
(Proportion of cases in the sample = 0.5; User-specified disease prevalence = 0.1)
V(1)/Vp_L       0.657621        0.189123
logL    -945.65
logL0   -940.12
LRT     11.06
Pval    4.41e-4
n       2000
```

--reml-est-fix

Output the estimates of fixed effects on the screen.

--reml-pred-rand

Predict the random effects by the BLUP (best linear unbiased prediction) method. This option is actually to predict the total genetic effect (called "breeding value" in animal genetics) of each individual attributed by the aggregative effect of the SNPs used to estimate the GRM. The total genetic effects of all the individuals will be saved in a plain ext file ***.indi.blp.**

Output file format

test.indi.blp (no header line; columns are family ID, individual ID, an intermedia variable and the total genetic effect; if there are multiple GRMs fitted in the model, each GRM will add additional two columns to the file, i.e. the intermedia variable and the total genetic effect)

```
01     0101   -0.012   -0.014
02     0203   0.021    0.031
03     0305   0.097    0.102
……
```

--blup-snp *test.indi.blp*

Calculate the BLUP solutions for the SNP effects (you have to specify the option --bfile to read the genotype data). This option takes the output of the option --reml-pred-rand as input (***.indi.blp** file) and transforms the BLUP solutions for individuals to the BLUP solutions for the SNPs, which can subsequently be used to predict the total genetic effect of individuals in an independent sample by **PLINK** --score option.

Output file format

test.snp.blp (columns are SNP ID, reference allele and BLUP of SNP effect; if there are multiple GRMs fitted in the model, each GRM will add an additional column to the file)

```
rs103645   A    0.00312
rs175292   G    -0.00021
……
```

Examples

```
# Without GRM (fitting the model under the null hypothesis that the additive genetic variance is zero)
gcta  --reml  --pheno test.phen  --out test_null
gcta  --reml  --pheno test.phen  --keep test.indi.list  --out test_null
```

 # One GRM (quantitative traits)

gcta --reml --grm test --pheno test.phen --reml-pred-rand –qcovar test_10PCs.txt --out test

gcta --reml --grm test --pheno test.phen --grm-adj 0 --grm-cutoff 0.05 --out test

gcta --reml --grm test --pheno test.phen --keep test.indi.list --grm-adj 0 --out test

 # One GRM (case-control studies)

gcta --reml --grm test --pheno test_cc.phen --prevalence 0.01 --out test_cc

gcta --reml --grm test --pheno test_cc.phen --prevalence 0.01 --qcovar test_10PCs.txt --out test_cc

 # GxE interaction (LRT test for the significance of GxE)

gcta --reml --grm test --pheno test.phen --gxe test.gxe --reml-lrt 2 --out test

 # Multiple GRMs

gcta --reml --mgrm multi_grm.txt --pheno test.phen --reml-no-lrt --out test_mgrm

gcta --reml --mgrm multi_grm.txt --pheno test.phen --keep test.indi.list --reml-no-lrt --out test_mgrm

 # BLUP solutions for the SNP effects

gcta --bfile test --blup-snp test.indi.blp --out test

7. Estimation of the LD structure in the genomic regions specified by a list of SNPs

For each target SNP, GCTA uses simple regression to search for SNPs that are in significant LD with the target SNP.

--ld *ld.snplist*

Specify a list of SNPs.

--ld-wind *5000*

Search for SNPs in LD with a target SNP within d Kb (e.g. 5000 Kb) region in either direction by simple regression test.

--ld-sig *0.05*

Threshold p-value for regression test, e.g. 0.05.

Example

gcta --bfile test --ld ld.snplist --ld-wind 5000 --ld-sig 0.05 --out test

Output files

1) **test.rsq.ld**, summary of LD structure with each row corresponding to each target SNP. The columns are
   target SNP
   length of LD block
   two flanking SNPs of the LD block
   total number of SNPs within the LD block
   mean r^2
   median r^2
   maximum r^2
   SNP in highest LD with the target SNP
2) **test.r.ld**, the correlations (r) between the target SNP and all the SNPs in the LD block.
3) **test.snp.ld**, the names of all the SNPs in the LD with the target SNP.

Note: LD block is defined as a region where SNPs outside this region are not in significant LD with the target SNP. According to this definition, the length of LD block depends on user-specified window size and significance level.

8. Simulation: simulating a GWAS based on real genotype data

The phenotypes are simulated based on a set of real genotype data and a simple additive genetic model $y_j = \Sigma_i x_{ij} * b_i + \varepsilon_j$, where x_{ij} is defined as the number of reference alleles for the i-th causal variant of the j-th individual, b_j is the allelic effect of the i-th causal variant and ε_j is the residual effect generated from a normal distribution with mean of 0 and variance of $va(\Sigma_i x_{ij} * b_i)(1 - 1 / h^2)$. For a case-control study, under the assumption of threshold model, cases are sampled from the individuals with disease liabilities (y) exceeding the threshold of normal distribution truncating the proportion of K (disease prevalence) and controls are sampled from the remaining individuals.

--simu-qt

Simulate a quantitative trait.

--simu-cc *100* *200*

Simulate a case-control study. Specify the number of cases and the number of controls, e.g. 100 cases and 200 controls. Since the simulation is based on the actual genotype data, the maximum numbers of cases and controls are restricted to be $n * K$ and $n * (1-K)$, respectively, where n is the sample size of the genotype data.

--simu-causal-loci *causal.snplist*

Assign a list of SNPs as causal variants. If the effect sizes are not specified in the file, they will be generated from a standard normal distribution.

Input file format

causal.snplist (columns are SNP ID and effect size)

```
rs113645    0.025
rs185292    -0.021
……
```

--simu-hsq *0.8*

Specify the heritability (or heritability of liability), e.g. 0.8. The default value is 0.1 if this option is not specified.

--simu-k *0.01*

Specify the disease prevalence, e.g. 0.01. The default value is 0.1 if this option is not specified.

--simu-rep *100*

Number of simulation replicates. The default value is 1 if this option is not specified.

Examples

 # Simulate a quantitative trait with the heritability of 0.5 for a subset of individuals for 3 times

```
gcta --bfile test --simu-qt --simu-causal-loci causal.snplist --simu-hsq 0.5 --simu-rep 3 --keep test.indi.list --out test
```

 # Simulate 500 cases and 500 controls with the heritability of liability of 0.5 and disease prevalence of 0.1 for 3 times

```
gcta --bfile test --simu-cc 500 500 --simu-causal-loci causal.snplist --simu-hsq 0.5 --simu-k 0.1 --simu-rep 3 --out test
```

Output file format

test.par (one header line; columns are the name of the causal variant, reference allele, allele frequency, allelic effect and variance explained by the causal variant).

QTL	RefAllele	Frequency	Effect	Qsq
rs13626255	C	0.136	-0.0837	0.026

rs779725 G 0.204 -0.0677 0.023

……

test.phen (no header line; columns are family ID, individual ID and the simulated phenotypes). For the simulation of a case-control study, all the individuals involved in the simulation will be outputted in the file and the phenotypes for the indivdiuals neither sampled as cases nor as controls are treated as missing, i.e. -9.

011 0101 1 -9 1

012 0102 2 2 -9

013 0103 1 1 1

……

9. Converting illumina raw genotype data into PLINK PED format

We provide a function to convert the raw genotype data (text files generated by GenomeStudio software) into PLINK PED format. **NOTE: this option is under developing. Please contact to us if you have any suggestion.**

--raw-files *raw_geno_filenames.txt*

Input a file which lists the filenames of the raw genotype data files (one data file per individual).

Input file format

raw_geno_filenames.txt (full paths can be specified if the raw genotype data files are in different directories)

raw_geno_file1

raw_geno_file2

……

raw_geno_file1000

The format of the raw genotype data looks like

[Header]

GSGT Version 1.6.3

Processing Date 7/7/2010 9:35 AM

Content HumanOmni1-Quad_v1-0_B.bpm

Num SNPs 1140419

Total SNPs 1140419

Num Samples 1000

Total Samples 1000

File 62 of 1000

[Data]

SNP Name Sample ID Sample Group GC Score Allele1 - Forward Allele2 - Forward Allele1 - Top Allele2 - Top Allele1 - Design Allele2 - Design Allele1 - AB Allele2 - AB Theta R X Y X Raw Y Raw B Allele Freq Log R Ratio

200006 000001 000001 0.8203 T T A A A A A A 0.018 1.901 1.848 0.053 19622 2436 0.0000 -0.2777

200052 000002 000001 0.8789 T T T T A A B B 0.958 0.881 0.054 0.827 2667 19381 0.9767 -0.0438

200053 000003 000002 0.6387 T T A A T T A A 0.105 1.396 1.196 0.200 12889 5067 0.0000 0.0175

```
200070 000004 000002 0.9221 G C C G G C A B 0.603 0.545 0.228 0.317 2767 3402 0.5133 -0.0125
200078 000005 000002 0.6779 C C G G G G B B 0.973 2.048 0.084 1.964 3114 37363 1.0000 0.0710
……
```

'Allele1-Top' and 'Allele2-Top' are taken as the genotypes for the SNPs.

--raw-summary *SNP_summary_table.txt*

Input a file providing the summary information of the SNPs (one row per SNP). The headers are necessary but they are not keywords and will be ignored by the program. **Note: the program actually only read the first four columns of this file.**

```
Index Name Chr Position ChiTest100 Het Excess AA Freq AB Freq BB Freq Call Freq Minor Freq Aux P-C Errors P-P-C Errors Rep Errors 10% GC 50% GC SNP # Calls # no calls Plus/Minus Strand HumanOmni1-Quad_v1-0_B.bpm.Address HumanOmni1-Quad_v1-0_B.bpm.GenTrain Score HumanOmni1-Quad_v1-0_B.bpm.Orig Score HumanOmni1-Quad_v1-0_B.bpm.Edited HumanOmni1-Quad_v1-0_B.bpm.Cluster Sep HumanOmni1-Quad_v1-0_B.bpm.AA T Mean HumanOmni1-Quad_v1-0_B.bpm.AA T Dev HumanOmni1-Quad_v1-0_B.bpm.AB T Mean HumanOmni1-Quad_v1-0_B.bpm.AB T Dev HumanOmni1-Quad_v1-0_B.bpm.BB T Mean HumanOmni1-Quad_v1-0_B.bpm.BB T Dev HumanOmni1-Quad_v1-0_B.bpm.AA R Mean HumanOmni1-Quad_v1-0_B.bpm.AA R Dev HumanOmni1-Quad_v1-0_B.bpm.AB R Mean HumanOmni1-Quad_v1-0_B.bpm.AB R Dev HumanOmni1-Quad_v1-0_B.bpm.BB R Mean HumanOmni1-Quad_v1-0_B.bpm.BB R Dev HumanOmni1-Quad_v1-0_B.bpm.Address2 HumanOmni1-Quad_v1-0_B.bpm.Norm ID
1 200006 9 139046223 0.6913772 0.03969868 0.124057 0.4819782 0.3939648 1 0.3650461 0 0 0 0 0.8203169 0.8203169 [A/G] 1193 0 60702346 0.8030853 0.8030853 0 1 0.02950359 0.009121547 0.4321907 0.01578533 0.9878551 0.005570452 2.313316 0.2726709 2.638608 0.3402262 1.769039 0.1879732 0 3
2 200052 2 219783037 0.9122009 0.01102628 0.00 0.02181208 0.9781879 0.9991618 0.01090604 0 0 0 0 0.8789128 0.8789128 [T/A] 1192 1 37712495 0.8901258 0.8901258 0 0.7359893 0.02316774 0.02236068 0.4633549 0.03744823 0.9825876 0.009741872 1.041702 0.1 1.228919 0.1265495 0.8926759 0.1 35794467 201
……
```

--gencall *0.7*

Specify a cutoff value of GenCall score. The default value is 0.7 if this option is not specified.

Example

```
gcta --raw-files raw_geno_filenames.txt --raw-summary SNP_summary_table.txt --out test
```

The data will be saved in two files in PLINK PED format, i.e. test.ped and test.map.

10. Joint & conditional genome-wide association analysis

--massoc-file *test.ma*

Input the summary-level statistics from a meta-analysis GWAS (or a single GWAS).

Input file format

test.ma

```
SNP A1 A2 freq b se p N
rs1001 A G 0.8493 0.0024 0.0055 0.6653 129850
rs1002 C G 0.0306 0.0034 0.0115 0.7659 129799
```

```
rs1003    A    C    0.5128  0.0045  0.0038  0.2319    129830
……
```

Columns are SNP, the effect allele, the other allele, frequency of the effect allele, effect size, standard error, p-value and sample size. The headers are not keywords and will be omitted by the program. **Important: "A1" must be the effect allele with "A2" being the other allele and "freq" should be the frequency of "A1".**

NOTE: 1) For a case-control study, the effect size should be log(odds ratio) with its corresponding standard error. 2) Please always input the summary statistics of all the SNPs even if your analysis only focuses on a subset of SNPs because the program needs the summary data of all SNPs to calculate the phenotypic variance.

--massoc-slct

Perform a stepwise model selection procedure to select independently associated SNPs. Results will be saved in a *.jma file with additional file *.jma.ldr showing the LD correlations between the SNPs.

--massoc-joint

Fit all the included SNPs to estimate their joint effects without model selection. Results will be saved in a *.jma file with additional file *.jma.ldr showing the LD correlations between the SNPs.

--massoc-cond *cond.snplist*

Perform association analysis of the included SNPs conditional on the given list of SNPs. Results will be saved in a *.cma.

Input file format

cond.snplist

```
rs1001
rs1002
……
```

--massoc-p 5e-8

Threshold p-value to declare a genome-wide significant hit. The default value is 5e-8 if not specified. This option is only valid in conjunction with the option --massoc-slct. **NOTE: it will be extremely time-consuming if you set a very low significance level, e.g. 5e-3.**

--*massoc*-wind 10000

Specify a distance *d* (in Kb units). It is assumed that SNPs more than *d* Kb away from each other are in complete linkage equilibrium. The default value is 10000 Kb (i.e. 10 Mb) if not specified.

--massoc-collinear 0.9

During the model selection procedure, the program will check the collinearity between the SNPs that have already been selected and a SNP to be tested. The testing SNP will not be selected if its multiple regression R^2 on the selected SNPs is greater than the cutoff value. By default, the cutoff value is 0.9 if not specified.

--massoc-gc

If this option is specified, p-values will be adjusted by the genomic control method. By default, the genomic inflation factor will be calculated from the summary-level statistics of all the SNPs unless you specify a value, e.g. --massoc-gc 1.05.

--massoc-actual-geno

If the individual-level genotype data of the discovery set are available (e.g. a single-cohort GWAS), you can use the discovery set as the reference sample. In this case, the analysis will be equivalent to a multiple regression analysis with the actual genotype and phenotype data. Once this option is specified, GCTA will take all pairwise LD correlations between all SNPs into account, which overrides the –massoc-wind option. This option also allows GCTA to calculate the variance taken out from the residual variance by all the significant SNPs in the model, otherwise the residual variance will be fixed constant at the same level of the phenotypic variance.

Examples (Individual-level genotype data of the discovery set is NOT available)

 # Select multiple associated SNPs through a stepwise selection procedure

```
gcta  --bfile test  --chr 1 --maf 0.01 --massoc-file test.ma --massoc-slct --out test_chr1
```

 # Estimate the joint effects of a subset of SNPs (given in the file test.snplist) without model selection

```
gcta  --bfile test  --chr 1 --extract test.snplist  --massoc-file test.ma --massoc-joint --out test_chr1
```

 # Perform single-SNP association analyses conditional on a set of SNPs (given in the file cond.snplist) without model selection

gcta --bfile test --chr 1 --maf 0.01 --massoc-file test.ma --massoc-cond cond.snplist --out test_chr1

It should be more efficient to separate the analysis onto individual chromosomes or even some particular genomic regions. Please refer to the **Data management** section for some other options, e.g. including or excluding a list of SNPs and individuals or filtering SNPs based on the imputation quality score.

Examples (Individual-level genotype data of the discovery set is available)

 # Select multiple associated SNPs through a stepwise selection procedure

gcta --bfile test --maf 0.01 --massoc-file test.ma --massoc-slct --massoc-actual-geno --out test

In this case, it is recommended to perform the analysis using the data of all the genome-wide SNPs rather than separate the analysis onto individual chromosomes because GCTA needs to calculate the variance taken out from the residual variance by all the significant SNPs in the model, which could give you a bit more power.

 # Estimate the joint effects of a subset of SNPs (given in the file test.snplist) without model selection

gcta --bfile test --extract test.snplist --massoc-file test.ma --massoc-actual-geno --massoc-joint --out test

 # Perform single-SNP association analyses conditional on a set of SNPs (given in the file cond.snplist) without model selection

gcta --bfile test --maf 0.01 --massoc-file test.ma --massoc-actual-geno --massoc-cond cond.snplist --out test

Output file format

test.jma **(generate by the option --massoc-slct or --massoc-joint)**

Chr	SNP	bp	freq	refA	b	se	p	n	freq_geno	bJ	bJ_se	pJ	LD_r
1	rs2001	172585028	0.6105	A	0.0377	0.0042	6.38e-19	121056	0.614	0.0379	0.0042	1.74e-19	-0.345
1	rs2002	174763990	0.4294	C	0.0287	0.0041	3.65e-12	124061	0.418	0.0289	0.0041	1.58e-12	0.012
1	rs2003	196696685	0.5863	T	0.0237	0.0042	1.38e-08	116314	0.589	0.0237	0.0042	1.67e-08	0.0
…													

Columns are chromosome; SNP; physical position; frequency of the effect allele in the original data; the effect allele; effect size, standard error and p-value from the original GWAS or meta-analysis; estimated effective sample size; frequency of the effect allele in the reference sample; effect size, standard error and p-value from a joint analysis of all the selected SNPs; LD correlation between the SNP i and SNP i + 1 for the SNPs on the list.

test.jma.ldr **(generate by the option --massoc-slct or --massoc-joint)**

SNP	rs2001	rs2002	rs2003	…
rs2001	1	0.0525	-0.0672	…
rs2002	0.0525	1	0.0045	…
rs2003	-0.0672	0.0045	1	…
…				

LD correlation matrix between all pairwise SNPs listed in ***test.jma***.

***test.cma* (generate by the option --massoc-slct or --massoc-cond)**

Chr	SNP	bp	freq	refA	b	se	p	n	freq_geno	bC	bC_se	pC
1	rs2001	172585028	0.6105	A	0.0377	0.0042	6.38e-19	121056	0.614	0.0379	0.0042	1.74e-19
1	rs2002	174763990	0.4294	C	0.0287	0.0041	3.65e-12	124061	0.418	0.0289	0.0041	1.58e-12
1	rs2003	196696685	0.5863	T	0.0237	0.0042	1.38e-08	116314	0.589	0.0237	0.0042	1.67e-08
…												

Columns are chromosome; SNP; physical position; frequency of the effect allele in the original data; the effect allele; effect size, standard error and p-value from the original GWAS or meta-analysis; estimated effective sample size; frequency of the effect allele in the reference sample; effect size, standard error and p-value from conditional analyses.

11. Bivariate REML analysis

These options are designed to perform a bivariate REML analysis of **two quantitative traits (continuous)** from population based studies or **two disease traits (binary)** from case control studies, to estimate the genetic variance of each trait and that genetic covariance between two traits that can be captured by all SNPs.

--reml-bivar 1 2

By default, GCTA will take the first two traits in the phenotype file for analysis. The phenotype file is specified by the option --pheno as described in univariate REML analysis. All the options for univariate REML analysisare still working here except --mpheno, --gxe, --prevalence, --reml-lrt, --reml-no-lrt and --blup-snp. All the input files are in the same format as in univariate REML analysis.

--reml-bivar-nocove

By default, GCTA will model the residual covariance between two traits. However, if the traits were measured on different individuals (e.g. two diseases), the residual covariance will be automatically dropped from the model. You could also specify this option to exclude the residual covariance at all time.

--reml-bivar-prevalence 0.1 0.05

For a bivariate analysis of two disease traits, you can specify the prevalence rates of the two diseases in the general population so that GCTA will transform the estimate of variance explained by the SNPs from the observed 0-1 scale to that on the underlying scale for both

diseases.

Examples

 # With residual covariance

gcta --reml-bivar --grm test --pheno test.phen --out test

 # Without residual covariance

gcta --reml-bivar --reml-bivar-nocove --grm test --pheno test.phen --out test

 # Case-control data for two diseases (the residual covariance will be automatically dropped from the model if there are not too many samples affected by both diseases)

gcta --reml-bivar --grm test_CC --pheno test_CC.phen --reml-bivar-prevalence 0.1 0.05 --out test_CC

Output file format

test.hsq (rows are header line;

genetic variance for trait 1, estimate and standard error (SE);

genetic variance for trait 2, estimate and SE;

genetic covariance between traits 1 and 2, estimate and SE;

residual variance for trait 1, estimate and SE;

residual variance for trait 2, estimate and SE;

residual covariance between traits 1 and 2, estimate and SE;

proportion of variance explained by all SNPs for trait 1, estimate and SE;

proportion of variance explained by all SNPs for trait 2, estimate and SE;

genetic correlation;

sample size).

Source	Variance	SE
V(G)_tr1	0.479647	0.179078
V(G)_tr2	0.286330	0.181329
C(G)_tr12	0.230828	0.147958
V(e)_tr1	0.524264	0.176650
V(e)_tr2	0.734654	0.181146
C(e)_tr12	0.404298	0.146863
Vp_tr1	1.003911	0.033202
Vp_tr2	1.020984	0.033800
V(G)/Vp_tr1	0.477779	0.176457
V(G)/Vp_tr2	0.280445	0.176928
rG	0.622864	0.217458
n	3669	