**fastGWA**

**fastGWA: A fast MLM-based Genome-Wide Association tool**

fastGWA is an ultra-efficient tool for mixed linear model (MLM)-based GWAS analysis of biobank-scale data such as the UK Biobank (see Jiang et al. [Nature Genetics 2019](https://www.nature.com/articles/s41588-019-0530-8) for details of the method). Credits: [Longda Jiang](mailto:longda.jiang@uq.edu.au) (method, simulation and analysis), [Zhili Zheng](mailto:zhili.zheng@uq.edu.au) (method, software and analysis) and [Jian Yang](http://researchers.uq.edu.au/researcher/2713) (method and overseeing).

We have applied fastGWA to 2,173 traits on 456,422 array-genotyped and imputed individuals and 2,048 traits on 49,960 whole-exome-sequenced (WES) individuals in the UK Biobank. All the summary statistics are available at **[UKBiobankGWASresults](http://172.16.13.142/software/gcta/index.html" \l "DataResource)**. One can also query or visualize the summary data using the online tool: [http://fastgwa.info](http://fastgwa.info/).

**Citation**

Jiang L, Zheng Z, Qi T, Kemper KE, Wray NR, Visscher PM, Yang J (2019) A resource-efficient tool for mixed model association analysis of large-scale data. Nature Genetics, 51:1749–1755. [doi:10.1038/s41588-019-0530-8](https://www.nature.com/articles/s41588-019-0530-8).

--make-bK-sparse 0.05  
To generate a sparse genetic relationship matrix (GRM) from a full-dense GRM at a cutoff value of 0.05. Note:1) The full-dense GRM can be generated by the --make-grm or --make-grm-part option (see the [GRM page](http://172.16.13.142/software/gcta/index.html" \l "MakingaGRM)); 2) We also have an [R-script](http://172.16.13.142/software/gcta/res/pedFAM.R) to generate a family relatedness matrix (FAM, the sparse GRM constructed from expected relatedness coefficients between related individuals), without the need of calculating a full-dense GRM. It only requires the input of a [PLINK .fam file](https://www.cog-genomics.org/plink/1.9/formats" \l "fam) and a table that describes the pedigree information of the cohort.

# Partition the GRM into 100 parts

gcta64 --bfile test --make-grm-part 100 1 --thread-num 5 --out test

gcta64 --bfile test --make-grm-part 100 2 --thread-num 5 --out test

...

gcta64 --bfile test --make-grm-part 100 100 --thread-num 5 --out test

# Merge all the parts together (Linux or Mac)

cat test.part\_3\_\*.grm.id > test.grm.id

cat test.part\_3\_\*.grm.bin > test.grm.bin

cat test.part\_3\_\*.grm.N.bin > test.grm.N.bin

# Make a sparse GRM from the merged full-dense GRM

gcta64 --grm test\_grm --make-bK-sparse 0.05 --out test\_sp\_grm

Output file format  
test\_sp\_grm.grm.id (columns are family ID and individual ID)

fid1 iid1

fid2 iid2

...

test\_sp\_grm.grm.sp (columns are the indexes of a pairs of individuals and the corresponding GRM value)

0 0 0.999106

1 1 0.993465

...

Note: "0" indicates the first individual in the \*.grm.id file.

--grm-sparse test\_sp\_grm  
To input the sparse GRM. The sparse format can be generated from SNP data using the --make-bK-sparse option described above or from pedigree information using our [R-script](http://172.16.13.142/software/gcta/res/pedFAM.R).

If the --grm-sparse flag is not specified, --fastGWA will run a linear regression analysis that does not account for relatedness.

--fastGWA-mlm  
To perform an MLM-based genome-wide association analysis.

--fastGWA-mlm-exact  
To perform an exact MLM-based association analysis without the GRAMMAR-GAMMA approximation.

--fastGWA-lr  
To perform a linear regression-based association analysis.

--est-vg REML  
To specify the method used to estimate the genetic variance component (Vg). The default value is "REML", which uses the fastGWA-REML method to estimate Vg. The alternative option is "HE", which uses Haseman-Elston regression to estimate Vg.

--h2-limit 1.6  
To specify the upper limit of Vg / Vp (with Vg being the genetic variance and Vp being the phenotypic variance) used in fastGWA-REML for grid search. The default value is 1.6.

--save-fastGWA-mlm-residual  
To save fastGWA residuals in a text file (\*.fastGWA.residual). The residuals are V-1 y/gamma. The estimated gamma parameter will be saved in a text file (\*.fastGWA.gamma).

\*.fastGWA.residual

1 11 0.03135

2 22 0.05642

3 33 0.000125

...

Columns are FID, IID and residual.

--model-only  
To perform the variance component estimation step in fastGWA without the association test step and save the results in \*.fastGWA.mdl.id and \*.fastGWA.mdl.bin files.

--load-model  
To load a saved model (see the --model-only flag above) to perform association tests. This flag is useful in a scenario where the fastGWA model parameters estimated from an analysis for the autosomes can be used in that for the X chromosome (see the example below). Note that this function only works when the sample IDs in the saved model are a subset of those in genotype data. This flag also works with all the other genotype QC flags (e.g., --maf, --extract and --geno) but is incompatible with flags to input phenotype, covariate or GRM.

--dc 1  
To specify a dosage compensation model for the X chromosome. Following PLINK, GCTA labels non-PAR (chr23) and PAR (chr25) regions of chromosome X with different chromosome numbers. SNPs on chr23 are coded as 0/2 for males and as 0/1/2 for females. By default, the GRM for chromosome X 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). However, all other analyses assume a full dosage compensation model (i.e., --dc 1) by default. Individuals without gender information will be treated as females

--nofilter  
By default, fastGWA filters out low quality variants (i.e., MAF < 0.0001 or missingness rate > 0.1) if no QC flag is specified. The --nofilter flag will mute this default filtering and output the association test results of all the variants. Note that this flag is equivalent to --maf 0 and --geno 0.

--seed  
fastGWA uses a set of randomly selected variants (up to 1000) to estimate the gamma parameter used for association tests (see the Supplementary Note 2 of Jiang et al. 2019 Nat Genet for details). In a very rare scenario, a bad choice of the random seed would lead to a failure of the gamma parameter estimation. In this case, it is recommended to choose a different seed (a non-zero integer value) using this flag.

Output format  
test.fastGWA (columns are chromosome, SNP, SNP position, the effect allele, the other allele, per allele sample size, frequency of A1, SNP effect, SE and p-value)

CHR SNP POS A1 A2 N AF1 BETA SE P

1 rs3131962 756604 A G 10000 0.130613 -0.00651289 0.00665363 0.327655

1 rs12562034 768448 A G 9998 0.106691 -0.0037883 0.00724556 0.601082

1 rs4040617 779322 G A 10000 0.128422 -0.00407097 0.00670377 0.543675

1 rs79373928 801536 G T 9996 0.0147122 -0.0365657 0.0186005 0.0493158

1 rs6657440 850780 C T 9992 0.393875 -0.00596944 0.00459088 0.193504

...

Examples

# To generate a sparse GRM from SNP data

# geno\_chrs.txt is a text file containing file paths to the SNP data of each chromosome

gcta64 --mbfile geno\_chrs.txt --make-grm --thread-num 10 --out geno\_grm

gcta64 --grm geno\_grm --make-bK-sparse 0.05 --out sp\_grm

# The two steps above can be merge into one if you don't have enough disk space to store the full dense GRM

gcta64 --mbfile geno\_chrs.txt --make-grm --sparse-cutoff 0.05 --threads 10 --out sp\_grm

# To run a fastGWA analysis based on the sparse GRM generated above

gcta64 --mbfile geno\_chrs.txt --grm-sparse sp\_grm --fastGWA-mlm --pheno phenotype.txt --qcovar pc.txt --covar fixed.txt --threads 10 --out geno\_assoc

# To save the estimated fastGWA model parameters from an analysis for the autosomes and use them in a subsequent analysis for chrX

# chrX.idlist: a list of sample IDs used in the analysis for chrX

gcta64 --mbfile geno\_chrs.txt --grm-sparse sp\_grm --fastGWA-mlm --model-only --pheno phenotype.txt --qcovar pc.txt --covar fixed.txt --keep chrX.idlist --threads 10 --out geno\_assoc

# To load the saved model above to run association tests for ChrX

# chr.snplist: a list of variants on chrX to be included in this analysis

gcta64 --bfile test\_chrX --load-model geno\_assoc.fastGWA --extract chr.snplist --geno 0.1 --out test\_chrX\_assoc --threads 10

# To run a linear regression analysis using fastGWA

gcta64 --mbfile geno\_chrs.txt --fastGWA-lr --pheno phenotype.txt --qcovar pc.txt --covar fixed.txt --threads 10 --out geno\_assoc