Genetic association testing using the GENESIS R/Bioconductor package: Supplementary Material

Contents

Supplementary Table 1

Package	Function	CPU time (s)
SNPRelate	snpgdsBED2GDS	74
SNPRelate	snpgdsVCF2GDS	595
SeqArray	seqBED2GDS	281
SeqArray	seqVCF2GDS	383

We compared run times of converting the same data from BED or VCF to GDS using the SNPRelate and SeqArray packages. Data used was 1000 Genomes phase 3, chromosome 22, with 2,504 samples and 1,103,822 variants. To get the 1000 Genomes data into BED format for testing, we used plink 1.9 to convert from VCF.

Supplementary Table 2

Package	Function	Block size	CPU time (s)	Memory (GB)
snpStats	read.plink		27	3.0
snpStats	single.snp.tests		74	3.0
GENESIS	assocTestSingle	10,000	194	1.3
GENESIS	assocTestSingle	100,000	119	7.0

We compared run times and memory usage for single-variant association testing on a set of unrelated samples with snpStats and GENESIS. Data used was the same as in Supplementary Table 1. snpStats reads the entire dataset into memory prior to running tests, so the total time for running a test is the sum of read.plink and single.snp.tests. GENESIS tests must account for the time involved in converting to GDS (Table 1), but this need be done only once for an entire project. GENESIS allows the user to balance available memory with computation speed; reading data in larger blocks requires more memory but reduces run time. We illustrate this with blocks of 10,000 and 100,000 variants.

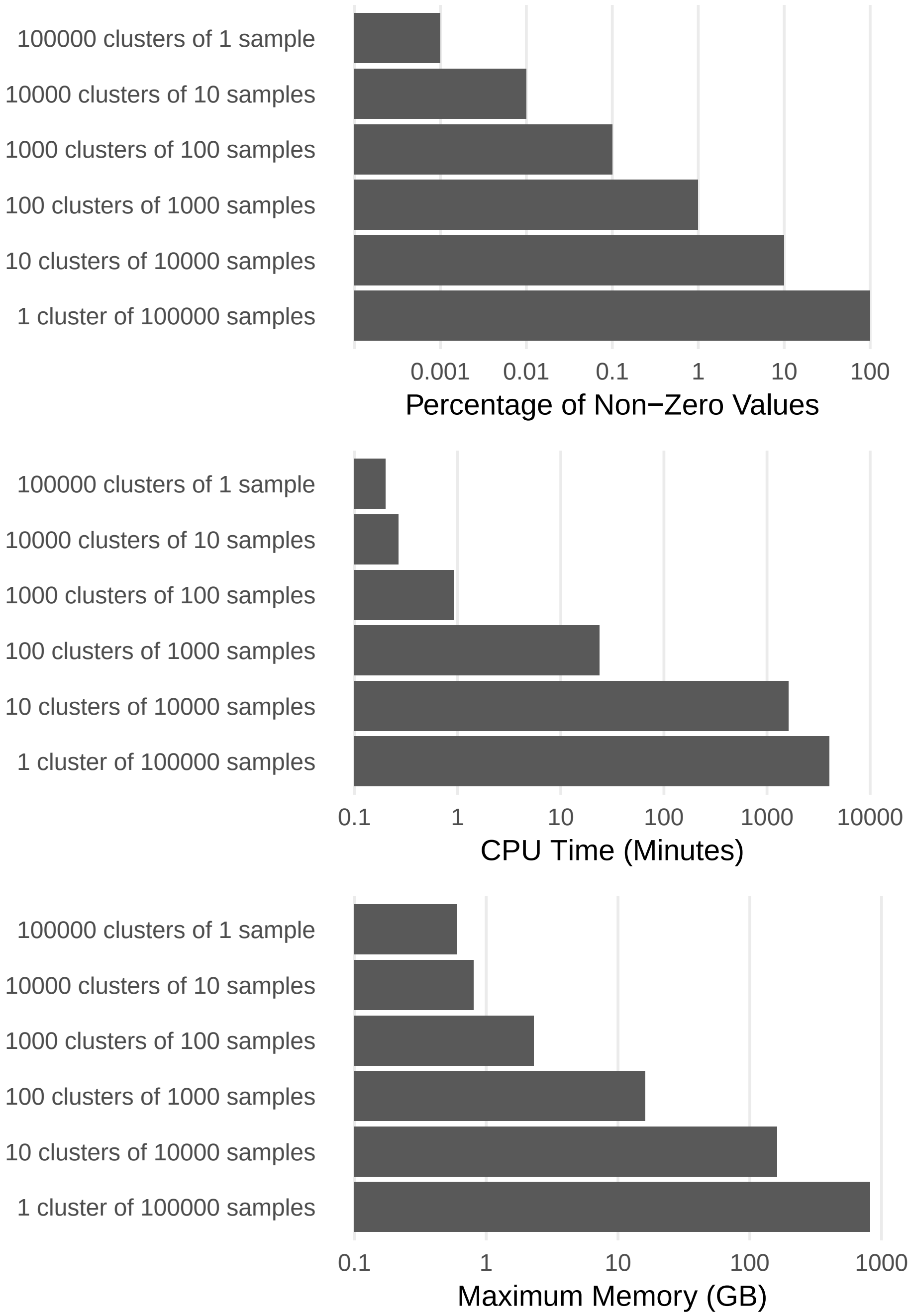
Supplementary Table 3

Approach A	Approach B	Prop Diff < 0.05	Prop Diff < 0.10	Prop Diff < 0.25	Prop Diff < 0.50	Max Diff
GRM	GRM + PCs	0.71776	0.93602	0.99862	0.99997	1.25422
GRM	Dense KM + PCs	0.62353	0.87567	0.99340	0.99988	1.40531
GRM	Sparse KM + PCs	0.58376	0.85001	0.99088	0.99983	1.50761
GRM + PCs	Dense KM + PCs	0.78487	0.94771	0.99859	0.99999	0.92791
GRM + PCs	Sparse KM + PCs	0.67647	0.90428	0.99583	0.99996	1.02340
Dense KM + PCs	Sparse KM + PCs	0.69577	0.92934	0.99904	1.00000	0.59189

Differences in association p-values at 24,456,292 SNPs for a heritable quantitative phenotype simulated on the 2,504 samples from 1000 Genomes. Mixed models with four different approaches of accounting for ancestry and relatedness are compared: (1) an empirical GRM; (2) an empirical GRM + ancestry PCs; (3) an empirical dense KM + ancestry PCs; (4) an empirical sparse KM + ancestry PCs. The empirical GRM was computed using the SNPRelate implementation of the GCTA method, the ancestry PCs were computed using PC-AiR, and the empirical KM was computed using PC-Relate. The empirical KM was made sparse at a 5^{th} degree relatedness threshold (i.e. kinship $> 2^{-13/2} \approx 0.011$) using the algorithm recommended in the manuscript; this sparse KM had 2,236 clusters, of which 2,080 were singletons and the largest had 23 members.

For each comparison of Approach A vs. Approach B, we present the proportion of SNPs with differences in $-\log_{10}(p)$ less than 0.05, 0.10, 0.25, and 0.50. We also present the maximum difference in $-\log_{10}(p)$ across all 24,456,292 SNPs tested. All four mixed model approaches provided very similar results; comparing any pair of these approaches, over 99.0% of SNPs had differences in $-\log_{10}(p)$ less than 0.25, and over 99.98% of SNPs had differences less than 0.5. Given a SNP with a true p-value of 5.0x10^{-8}, a difference in $-\log_{10}(p)$ less than 0.25 would correspond to a reported p-value in the range (8.9x10^{-8}, 2.8x10^{-8}); a difference in $-\log_{10}(p)$ less than 0.50 would correspond to a reported p-value in the range (1.6x10^{-7}, 1.6x10^{-8}).

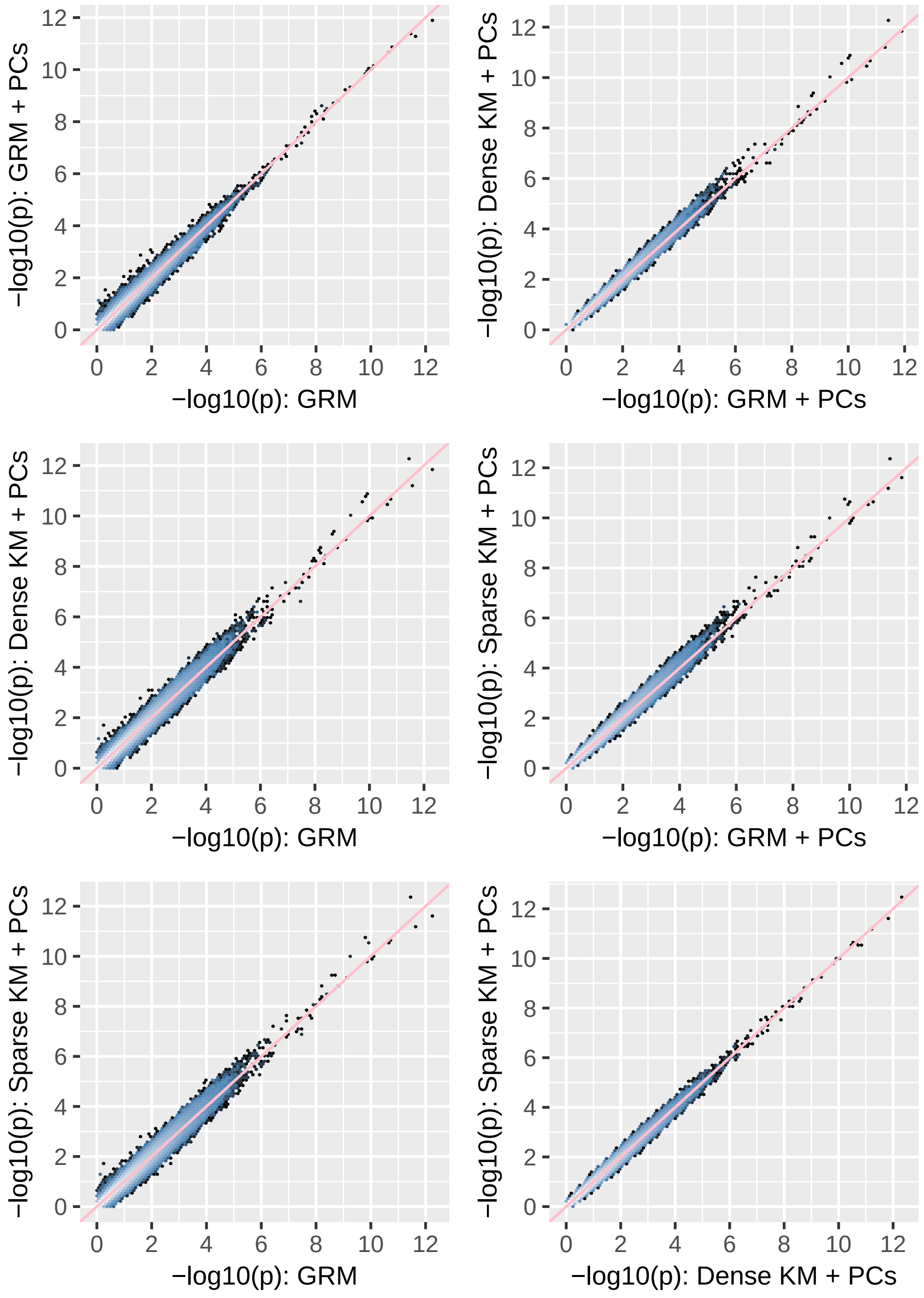
Supplementary Figure 1



Computational Comparison Using Sparse GRM/KM

The null model was fit for a simulated heritable quantitative trait measured on 100,000 samples using GRM/KMs with different sparsity. The empirical GRM/KM was made sparse block-diagonal at varying levels; cluster sizes ranged from 1 sample (diagonal matrix) to 100,000 samples (dense matrix) in 10-fold increments; values between clusters were set to 0. We display (A) the percentage of non-zero values in each GRM/KM; (B) the CPU time to fit the null model; and (C) the maximum memory required while fitting the null model. All values are shown on a log10 scale.

Supplementary Figure 2



Comparison of p-values from Mixed Models using Different GRMs, KMs, and PCs

Comparison of association p-values at 24,456,292 SNPs for a heritable quantitative phenotype simulated on the 2,504 samples from 1000 Genomes. Mixed models with four different approaches of accounting for ancestry and relatedness are compared: (1) an empirical GRM; (2) an empirical GRM + ancestry PCs; (3) an empirical dense KM + ancestry PCs; (4) an empirical sparse KM + ancestry PCs. The $-\log_{10}(p)$ for all 24,456,292 SNPs tested are shown through the use of hexbin plots.

GWAS Example Code

The GENESIS package includes vignettes illustrating how to run mixed model analyses starting from a variety of data formats. Here we provide a simplified code example showing how to perform a GWAS on a set of unrelated samples using GENESIS. The first step is to convert data from another format, such as PLINK BED, to GDS.

```
library(SNPRelate)
bedfile <- system.file("extdata", "plinkhapmap.bed.gz", package="SNPRelate")
bimfile <- system.file("extdata", "plinkhapmap.bim.gz", package="SNPRelate")
famfile <- system.file("extdata", "plinkhapmap.fam.gz", package="SNPRelate")
gdsfile <- tempfile()
snpgdsBED2GDS(bed.fn=bedfile, fam.fn=famfile, bim.fn=bimfile,
              out.gdsfn=gdsfile, verbose=FALSE)
```

Next, we load the sample annotation and simulate a phenotype and covariate. Sex values must be coded as "M"/"F" to accurately count alleles on sex chromosomes.

```
library(GWASTools)
fam <- read.table(famfile, as.is=TRUE)
names(fam) <- c("family", "scanID", "father", "mother", "sex", "phen")
set.seed(100)
fam$phen <- rnorm(nrow(fam))
set.seed(100)
fam$group <- sample(c("A", "B"), replace=TRUE, nrow(fam))
fam$sex <- c("1"="M", "2"="F", "0"=NA)[as.character(fam$sex)]
annot <- ScanAnnotationDataFrame(fam)
```

We adjust the phenotype for the covariate, creating a null model. If we were using a mixed model for related samples, doing this step only once before testing the genotypes would save substantial computation time. Since no random effects are included, in this case GENESIS does a simple linear regression.

```
library(GENESIS)
nullmod <- fitNullModel(annot, outcome="phen", covars="group", verbose=FALSE)
```

We link the sample annotation to the genotypes and set up an iterator object that controls how many SNPs are read at one time.

```
gds <- GdsGenotypeReader(gdsfile)
genoData <- GenotypeData(gds, scanAnnot=annot)
iterator <- GenotypeBlockIterator(genoData, snpBlock=10000)
```

Finally, we test the association of each SNP with the phenotype.

```
assoc <- assocTestSingle(iterator, nullmod, verbose=FALSE)
head(assoc, n=4)
```

```
##   variant.id chr     pos n.obs       freq     Score Score.SE Score.Stat
## 1          9   1 2444790    60 0.44166667  2.092129 5.693352  0.3674688
## 2         18   1 3314897    60 0.42500000 -5.492986 5.469607 -1.0042742
## 3         20   1 3644455    60 0.15833333  1.496689 4.486641  0.3335879
## 4         32   1 4221327    60 0.08333333 -1.763030 3.175936 -0.5551216
##   Score.pval
## 1  0.7132693
## 2  0.3152465
## 3  0.7386905
## 4  0.5788115
```