

LDAK-GBAT: fast and powerful gene-based association testing using summary statistics

Authors: Takiy-Eddine Berrandou¹, David Balding² and Doug Speed¹

¹ Quantitative Genetics and Genomics (QGG), Aarhus University, Aarhus, Denmark.

² Melbourne Integrative Genomics (MIG), Melbourne University, Melbourne, Australia.

Correspondence: doug@qgg.au.dk

ABSTRACT

We present LDAK-GBAT, a novel tool for gene-based association testing using summary statistics from genome-wide association studies. We first evaluate LDAK-GBAT using ten phenotypes from the UK Biobank. We show that LDAK-GBAT is computationally efficient, taking approximately 30 minutes to analyze imputed data (2.9M common, genic SNPs), and requiring less than 10Gb memory. In total, LDAK-GBAT finds 680 genome-wide significant genes ($P \leq 2.8 \times 10^{-6}$), which is at least 25% more than each of five existing tools (MAGMA, GCTA-fastBAT, sumFREGAT-SKAT-O, sumFREGAT-PCA and sumFREGAT-ACAT), and 48% more than found by single-SNP analysis. We then analyze 99 additional phenotypes from the UK Biobank, the Million Veterans Project and the Psychiatric Genetics Consortium. In total, LDAK-GBAT finds 7957 significant genes, which is at least 24% more than the best existing tools, and 42% more than found by single-SNP analysis.

INTRODUCTION

A genome-wide association study (GWAS) analysis typically starts by testing each SNP individually for association with the phenotype ¹. However, it is now common to also perform a gene-based analysis, which instead tests for association sets of SNPs defined by gene annotations ^{2,3}. Gene-based analyses can be more powerful than single-SNP analyses because they are able to accumulate evidence of association across multiple SNPs, and because they perform fewer statistical tests (and so require less correction for multiple

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
testing) ^{4,6}. Further, gene-based analyses are biologically justified, as it is plausible that variants within a gene act jointly to affect an outcome ⁴.

We propose LDAK-GBAT, a tool for performing gene-based association analysis. LDAK-GBAT extends a previous tool, FaST-LMM-Set ⁷, in three key ways. Firstly, LDAK-GBAT can be applied using GWAS summary statistics and a reference panel, whereas FaST-LMM-Set requires individual-level data. Secondly, LDAK-GBAT accommodates alternative heritability models (assumptions regarding how heritability is distributed across the genome), whereas FaST-LMM-Set assumes that all SNPs are expected to contribute equally towards the phenotype. Thirdly, results from LDAK-GBAT can be clumped, which provides an approximate way to divide the significant genes into those directly and indirectly associated with the phenotype.

First, we analyze summary statistics for ten phenotypes from UK Biobank (UKBB) ⁸. We show that LDAK-GBAT is computationally efficient, produces well-calibrated p-values under the null hypothesis, and finds at least a quarter more significant genes than MAGMA ⁹, GCTA-fastBAT ¹⁰, sumFREGAT-SKAT-O, sumFREGAT-PCA, sumFREGAT-ACAT ^{11,12} and single-SNP analysis. Next, we analyze summary statistics for 99 traits: 72 additional phenotypes from UK Biobank defined by the International Classification of Diseases 10th Revision (ICD-10), 18 traits from Million Veterans Project (MVP) ¹³ and nine traits from the Psychiatric Genomics Consortium (PGC) ¹⁴. Again, LDAK-GBAT finds substantially more significant genes than MAGMA, sumFREGAT-PCA, sumFREGAT-ACAT and single-SNP analysis.

LDAK-GBAT is freely-available within our software package LDAK (see Web Resources for a link to the LDAK website, that contains full documentation, including test data and example scripts).

MATERIAL AND METHODS

Here we summarize how LDAK-GBAT tests genes for association with a phenotype; for full details see the Supplementary Note. Suppose we have a GWAS of n individuals, and we wish to test a gene containing m SNPs. Let the length- n vector Y contain the phenotypic values, and the $n \times m$ matrix X contain the SNP genotypes. For convenience, we assume that X_j , the j th column of X , has been standardized to have mean zero and variance one. LDAK-GBAT assumes the linear model

$$Y = X\beta + e = X_1\beta_1 + \dots + X_m\beta_m + e$$

Here, β_j denotes the effect size for the j th SNP, while the vector e denotes environmental noise. We assign the following prior distributions to β_j and e

$$\beta_j \sim N(0, q_j \sigma_g^2) \quad \text{and} \quad e_i \sim N(0, \sigma_e^2)$$

where σ_g^2 and σ_e^2 denote genetic and environmental variance components, respectively, and the q_j are pre-specified constants determined by the choice of heritability model (see below). We estimate σ_g^2 and σ_e^2 using REML (restricted maximum likelihood)¹⁵, then obtain p-values via a likelihood ratio test of $\sigma_g^2=0$. Like FaST-LMM-Set, we derive the null distribution of the test statistic via permutation⁷.

As well as p-values, LDAK-GBAT reports $X \hat{\beta}$ for each gene, an estimate of its genetic contribution (in effect, a polygenic risk score computed using only the SNPs in the gene). These estimates can be used to clump the results from LDAK-GBAT (similar to how it is standard to clump the results from single-SNP analysis¹⁶). For example, in the analyses below, we compute the squared correlation between the estimated genetic contributions of significant genes on the same chromosome; if a pair of genes has squared correlation above 0.1, we exclude the gene with highest p-value.

The above description assumes we have individual-level data (access to X and Y). However, in the Supplementary Note, we explain how all the required calculations can be written in terms of the form $Cor(X_j, Y)$, the correlation between SNP j and the phenotype, and $Cor(X_j, X_k)$, the correlation between SNPs j and k . This enables LDAK-GBAT to be run using only summary statistics from single-SNP analysis (from which it recovers $Cor(X_j, Y)$ for each SNP) and a reference panel (which it uses to estimate $Cor(X_j, X_k)$ between pairs of SNPs).

When analyzing SNP data, the heritability model describes how $E[h_j^2]$, the expected heritability contributed by a SNP, varies across the genome. With SNPs standardized, $E[h_j^2]$ is proportional to $E[\beta_j^2]$, which is in turn proportional to q_j (i.e., the values for q_j determine the heritability model). If we set $q_j=1$ for all SNPs, then LDAK-GBAT assumes the same model as FaST-LMM-Set. This choice of q_j corresponds to assuming $E[h_j^2]$ is constant, and therefore we refer to it as the “Uniform Model”. Our previous works indicate

that the Uniform Model is sub-optimal, because when analyzing real data, we observed that h_j^2 systematically varies with features such as minor allele frequency (MAF) and linkage disequilibrium (LD)¹⁷⁻²⁰.

UKBB, MVP and PGC summary statistics. We compute summary statistics for the UKBB phenotypes from the individual-level data. We first focus on ten phenotypes: body mass index, college education, forced vital capacity, height, hypertension, impedance, neuroticism score, preference for evenings, pulse rate, and systolic blood pressure. We subsequently consider 72 ICD-10 diseases, spanning 12 of the 22 chapters (see Supplementary Table 1 for details). We compute summary statistics by performing classical linear regression (for both quantitative and binary traits), restricting to unrelated, white British individuals, and including 13 covariates (age, sex, Townsend Deprivation Index and ten principal components). For the first ten phenotypes, the sample size is 50,000 individuals (the exception is when performing single-SNP analysis, where we use up to 200,000 individuals). For the ICD-10 phenotypes, there are on average 12,781 cases (range 1,559 to 70,012) and 225,130 controls (range 103,972 to 250,346). Further details of the UKBB summary statistics are provided in the Supplementary Note.

For MVP and PGC, we use summary statistics from previously published GWAS. For MVP, we consider 18 phenotypes including type 2 diabetes, cardiovascular diseases, number of cigarettes per day, lipid levels and blood pressure. The average sample size is 242,207 (range 17,014 to 1,114,458). For PGC, we consider 9 phenotypes: alcohol use, Alzheimer's Disease, attention deficit hyperactivity disorder, autism spectrum disorder, bipolar disorder, eating disorder, major depressive disorder, post-traumatic stress disorder and schizophrenia. The average sample size is 256,250 (range 46,351 to 762,917). See Supplementary Tables 2 and 3 for more details.

Default settings of LDAK-GBAT. When running LDAK-GBAT, the user must choose a reference panel, provide gene annotations and specify a heritability model. For our analyses below, our primary reference panel comprises 10,000 unrelated UKBB individuals genotyped for 7,186,768 imputed SNPs with MAF \geq 0.01 and imputation quality information (Rsq) \geq 0.8. We use RefSeq gene annotations²¹. When testing a gene, we consider only SNPs located between the transcription start and stop sites. In total, there are 17,965 genes, that contain 2,890,640 of the 7,186,768 SNPs. Unless stated otherwise, we run LDAK-GBAT assuming the

“Human Default Model”. This sets $q_j = [p_j(1-p_j)]^{0.75}$, where p_j is the MAF of SNP j , and is our default recommendation when analyzing human traits^{19,20}.

Existing tools. We compare LDAK-GBAT with five existing summary statistic tools for gene-based association analysis: MAGMA, GCTA-fastBAT, sumFREGAT-SKAT-O, sumFREGAT-PCA and sumFREGAT-ACAT (see Supplementary Note for a brief description of each tool). We use the default settings for each tool. MAGMA, GCTA-fastBAT, sumFREGAT-SKAT-O and sumFREGAT-PCA require a reference panel, for which we use the same 10,000 unrelated UKBB individuals that we use when running LDAK-GBAT.

RESULTS

Supplementary Figures 1 and 2 show that when we run LDAK-GBAT assuming the Uniform Model, its results align closely with those from FaST-LMM-Set, confirming that it is feasible to use only GWAS summary statistics and a reference panel, instead of individual-level data.

Comparing LDAK-GBAT with existing tools. Firstly, we compare tools based on their type 1 error. For this we generate permuted versions of the UKBB phenotype height. If a tool effectively controls the type 1 error, then the p-values from analyzing the permuted phenotype should be uniformly distributed. Supplementary Table 4 shows that this is the case for LDAK-GBAT, and is generally the case for the five existing tools. For example, for LDAK-GBAT, we find that 0.0099, 1.0×10^{-4} and 1.5×10^{-6} genes have p-values less than 0.01, 1×10^{-4} and 1×10^{-6} , respectively.

Tool	Time (minutes)	Memory (GB)
LDAK-GBAT	33	5
MAGMA	47	1
GCTA-fastBAT	96	22
sumFREGAT-SKAT-O	1299	23
sumFREGAT-PCA	679	18
sumFREGAT-ACAT	20	9

Table 1: Computational efficiency of different tools for gene-based association testing. Average time and memory required to analyze each of the ten UKBB phenotypes on a single CPU.

Secondly, we compare tools based on computational efficiency. Table 1 shows that sumFREGAT-ACAT, LDAK-GBAT and MAGMA are the fastest tools, taking on average 20-50 minutes to analyze each of the ten UKBB phenotypes. Meanwhile, sumFREGAT-PCA and sumFREGAT-SKAT-O are the slowest tools, taking over 10 hours to analyze each phenotype. All tools have either low or modest memory demands, requiring at most 23Gb.

Thirdly, we compare tools based on the number of significant genes. We use the Bonferroni significance threshold $0.05/17669 = 2.8 \times 10^{-6}$. Figure 1 and Supplementary Table 5 show that, in total, LDAK-GBAT finds 680 significant genes, whereas MAGMA, GCTA-fastBAT, sumFREGAT-SKAT-O, sumFREGAT-PCA and sumFREGAT-ACAT find 410, 408, 93, 542 and 534, respectively. Supplementary Tables 5 and 6 show that LDAK-GBAT continues to find the most associations if we filter significant genes based on LD, or if we restrict our analysis to 606k directly genotyped SNPs.

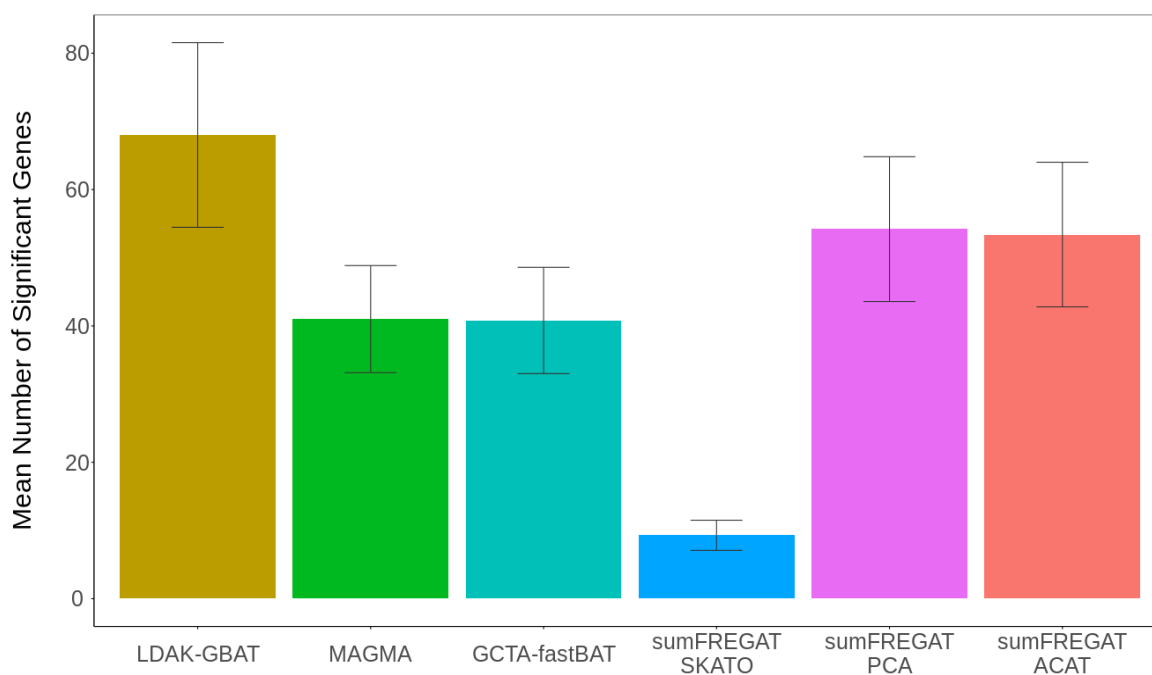


Figure 1: Comparing gene-based association testing tools for the ten UK Biobank phenotypes. Bars report the mean number of genome-wide significant genes ($P \leq 2.8 \times 10^{-6}$) for LDAK-GBAT and five existing tools. Segments mark 95% confidence intervals for the means.

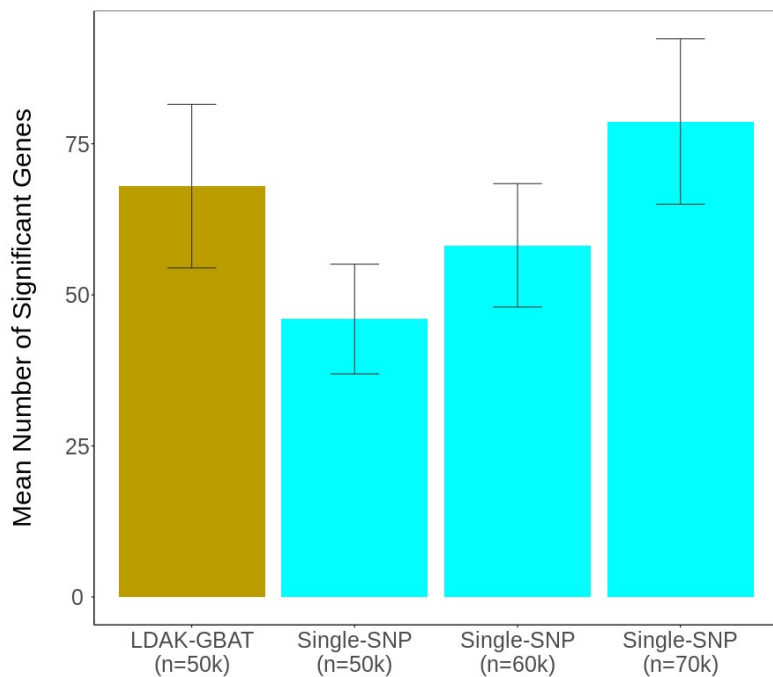


Figure 2: Comparing LDAK-GBAT and single-SNP analysis for the ten UK Biobank phenotypes. Bars report the mean number of genome-wide significant genes ($P \leq 2.8 \times 10^{-6}$) for LDAK-GBAT (using 50,000 individuals) and the number of genes containing a SNP with $P \leq 5e-8$ from single-SNP analysis (using 50,000, 60,000 or 70,000 individuals). Segments mark 95% confidence intervals for the means.

Comparing LDAK-GBAT with single-SNP analysis. When performing single-SNP analysis, we declare a gene significant if it contains at least one SNP with $P \leq 5e-8$. Figure 2 shows that when analyzing 50,000 individuals, single-SNP analysis finds in total 460 significant genes, 220 fewer than LDAK-GBAT. Further, we see that for single-SNP analysis to find the same number of significant genes as LDAK-GBAT, we must increase the sample size to approximately 65,000 individuals (i.e., by about 30%). Supplementary Table 7 shows that 249 of the 267 genes (93%) significant from LDAK-GBAT but not from single-SNP analysis, are significant from single-SNP analysis when the sample size is increased to 200,000 individuals, providing reassurance that the majority of novel associations found by LDAK-GBAT are true positives.

Choice of reference panel and heritability model. Supplementary Figure 3 shows that results from LDAK-GBAT are very similar if we replace the UKBB reference panel with 404 non-Finnish, European individuals from the 1000 Genome Project ²¹. Figure 3 and Supplementary Table 8 consider heritability models of the form $q_j = [p_j(1-p_j)]^{1+\alpha}$ for $\alpha = -1.25, -1, -0.75, -0.5, -0.25, 0$ and 0.25 . We see that changing from the Uniform Model ($\alpha = -1$) to the Human Default Model ($\alpha = -0.25$) has increased the total number of significant genes from 532 to 680 (27%).

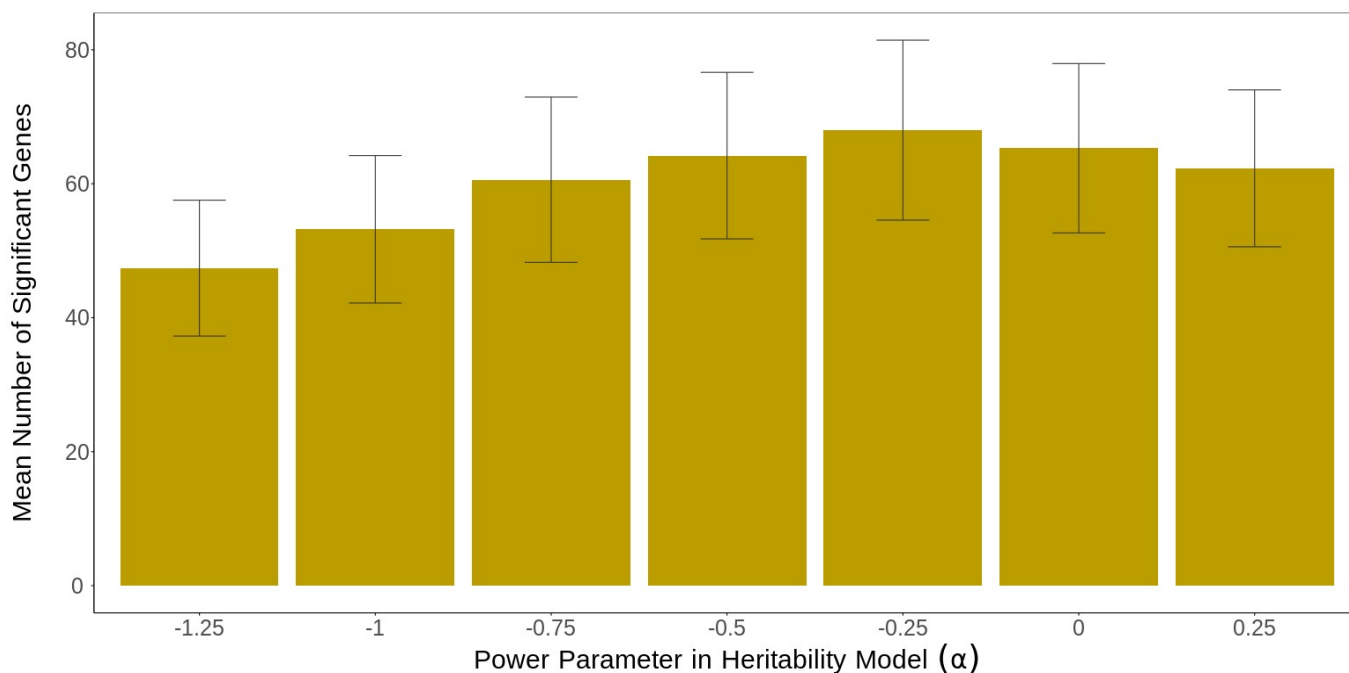


Figure 3: Varying the heritability model for the ten UK Biobank phenotypes. We run LDAK-GBAT using seven heritability models, defined by $E[h^2_j] \sim [p_j(1-p_j)]^{1+\alpha}$, where p_j is the MAF of SNP_j , and α is -1.25, -1, -0.75, -0.5, -0.25, 0 or 0.25. Bars report the mean number of genome-wide significant genes ($P \leq 2.8 \times 10^{-6}$). Segments mark 95% confidence intervals for the means.

Analysis of ICD-10, MVP and PGC phenotypes. Figure 4 and Supplementary Table 9 report results from analyzing the 99 ICD10, MVP and PGC phenotypes. Now we compare LDAK-GBAT with MAGMA, sumFREGAT-PCA and sumFREGAT-ACAT (the three best-performing existing tools when analyzing the ten UKBB phenotypes). We continue to use the Bonferroni significance threshold $2.8e-6$. Across the 72 UKBB ICD10 phenotypes, LDAK-GBAT finds 1874 significant genes, whereas MAGMA, sumFREGAT-PCA and sumFREGAT-ACAT find only 1265, 1560 and 1609, respectively. The pattern is similar for the 18 MVP traits (LDAK-GBAT finds 4681 significant genes, compared to 2964, 3456 and 3659), and also for the nine PGC traits (LDAK-GBAT finds 1402 significant genes, compared to 943, 1171 and 1130). We note that 1279 (16%) of the significant genes from LDAK-GBAT are not detected by any of the three existing tools, nor by single-SNP analysis (Supplementary Figure 4).

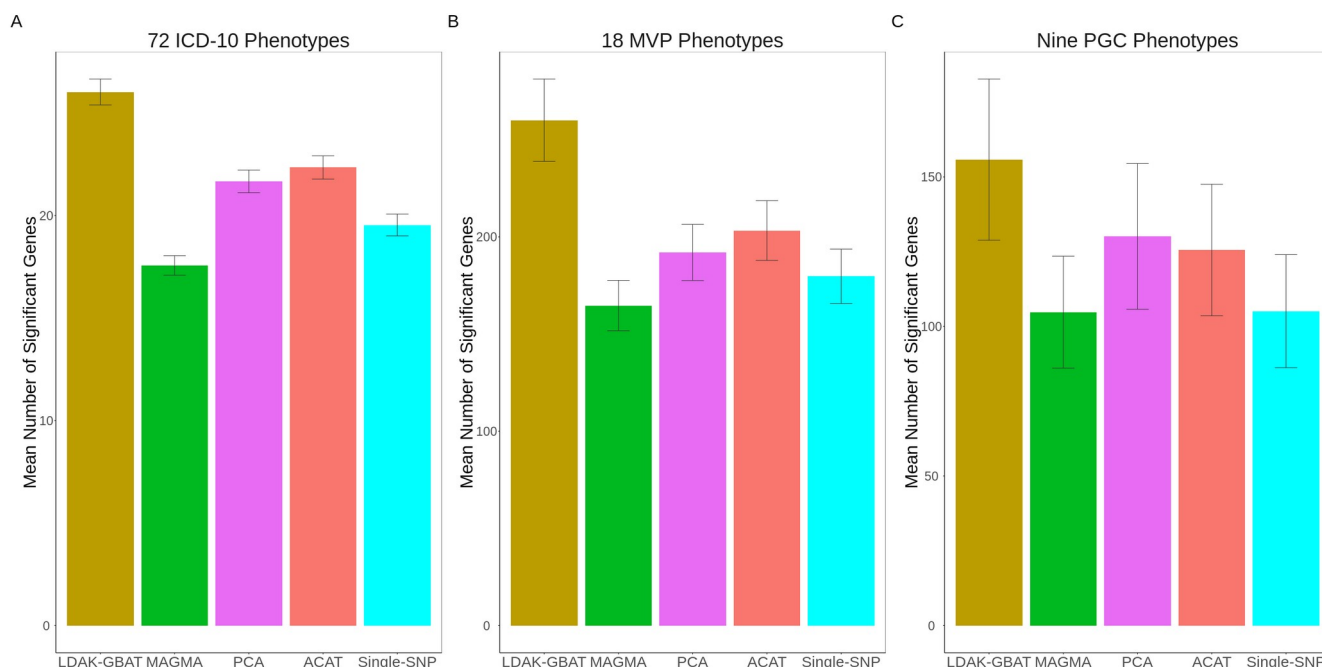


Figure 4: Comparing gene-based association testing tools for the 99 ICD10, MVP and PGC phenotypes.

Bars report the mean number of genome-wide significant genes ($P \leq 2.8 \times 10^{-6}$) for LDAK-GBAT and three existing tools, and the number of genes with a SNP with $P < 5e-8$ from single-SNP analysis, across the 72 ICD10 phenotypes (left), the 18 MVP phenotypes (middle) and nine PGC phenotypes (right). Segments mark 95% confidence intervals for the means.

For Supplementary Table 10, we clump the results from LDAK-GBAT so that no two genes on the same chromosome have estimated genetic contributions with squared correlation greater than 0.1. This reduces the 7957 significant genes to 2290 approximately independent significant genes (448, 1362 and 480 for the ICD10, MVP and PGC traits, respectively). To illustrate some of the benefits of clumping, Figure 5 compares results from LDAK-GBAT and single-SNP analysis for Type 2 Diabetes (the trait with most significant genes), for Chromosome 3 45 - 90 Mbp. In total, LDAK-GBAT finds that 50 of 261 genes in this region are significantly associated. However, clumping reduces this to ten approximately independent significant genes. For example, there are 39 significant genes in the first 10Mbp of the region, however, clumping indicates that 37 of these are associated only because they are in LD with the genes RBM6 or SFMBT1. Similarly, there are ten significant genes in the last 5Mbp of the region, however, clumping indicates that nine of these are associated only because they are in LD with the gene VGLL3. Moreover, it is interesting to note that VGLL3 is strongly significant from LDAK-GBAT ($P = 3.8 \times 10^{-8}$), but only modestly significant from single-SNP analysis (across the 43 SNPs within this gene, the smallest p -value from single-SNP analysis is 7×10^{-5}).

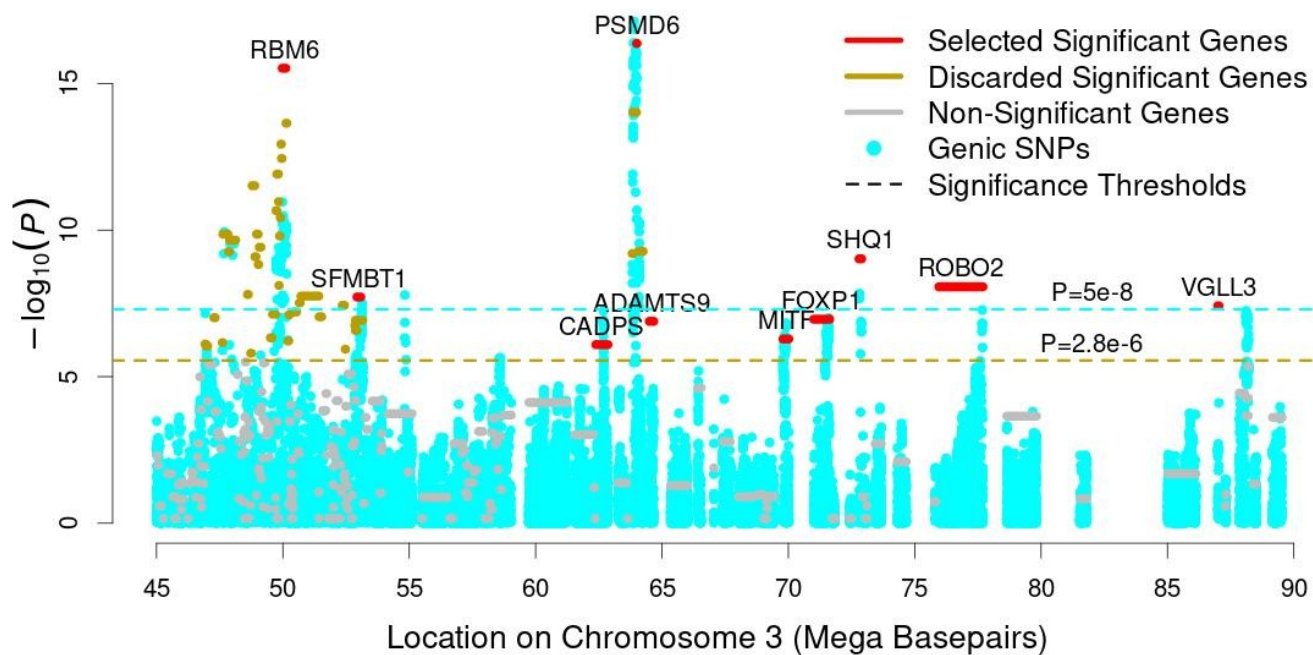


Figure 5: Clumping results from LDAK-GBAT. Segments provide p -values from LDAK-GBAT when analyzing the MVP phenotype type 2 diabetes; red segments indicate significant genes that are selected by clumping, gold segments indicate significant genes that are discarded by clumping, while gray segments indicate non-significant genes. Blue points report p -values for genic SNPs from single-SNP analysis. The gold and blue horizontal lines mark $P=2.8 \times 10^{-6}$ and $P=5 \times 10^{-8}$, respectively (the significance thresholds used with LDAK-GBAT and single-SNP analyses).

DISCUSSION

We have proposed LDAK-GBAT, a novel method for gene-based association analysis. We have shown that LDAK-GBAT is both computationally efficient and statistically more powerful than existing tools for gene-based association testing.

LDAK-GBAT improves the existing tool FaST-LMM-Set in three main ways. Firstly, LDAK-GBAT uses GWAS summary statistics and a reference panel, whereas FaST-LMM-Set requires individual-level data. This made it feasible, for example, for us to analyze phenotypes from the MVP and PGC without access to individual-level data. Secondly, LDAK-GBAT allows the user to specify the heritability model, whereas FaST-LMM-Set considers only the Uniform Model. Figure 3 showed that, for the ten UKBB phenotypes, switching from the Uniform Model to the Default Human Model resulted in 27% more significant genes (approximately equal to power difference between LDAK-GBAT and sumFREGAT-PCA, the best existing tool). We speculate that it may be possible to increase the power of LDAK-GBAT further by considering more

complex heritability models, that take into account, for example, functional annotations^{19,20}. Thirdly, results from LDAK-GBAT can be clumped. This feature is particularly useful when analyzing results from large GWAS (e.g., >100,000 samples), where it is common to find clusters of significantly associated genes, and so clumping provides a quick way to prioritize the genes in each cluster, and to estimate the number of independent significant associations.

We recognize three limitations with the LDAK-GBAT methodology. Firstly, the accuracy of LDAK-GBAT depends on the suitability of the reference panel. In this respect, it is reassuring that when analyzing the ten UKBB phenotypes, results from LDAK-GBAT were robust to replacing the UKBB reference panel with 404 European individuals from the 1000 Genome Project (i.e., both reducing the size of the reference panel, and reducing its ancestral similarity with the GWAS data). Secondly, the present version of LDAK-GBAT can only be applied to results from single-ancestry GWAS, which is problematic considering the increasing popularity of multi-ancestry GWAS²². As a possible solution, we will explore whether it is feasible to meta-analyze results from LDAK-GBAT across ancestries or develop a version of LDAK-GBAT that uses a multi-ancestry reference panel. Thirdly, although we believe that LDAK-GBAT is the first gene-based association tool to offer a clumping feature, we appreciate that this is an imperfect way to prioritize results (it assumes that the causal gene has the smallest p-value). We note that when performing single-SNP analyses, there are now a variety of fine-mapping methods that are more effective than clumping, and we believe that LDAK-GBAT will enable similar methods to be developed for gene-based analyses.

Appendices

None

Description of supplementary material

The supplementary material include three notes, six figures and ten tables.

Declaration of interests

The authors declare no competing interests.

Acknowledgments

T-E.B. and D.S. are supported by Aarhus University Research Foundation (AUFF), by the Independent Research Fund Denmark (Project no. 7025-00094B), and by a Lundbeck Foundation Experiment Grant. D.B. and D.S. are supported from the Australian Research Council (Grant DP190103188).

Author contributions

Writing and editing the manuscript: T-E.B., D.B. and D.S. Study design/conception: T-E.B., D.B. and D.S. Analyses: T-E.B. and D.S.

Web resources

LDAK-GBAT, <https://www.ldak.org>

MAGMA, <https://ctg.cncr.nl/software/magma>

GCTA-fastBAT, <https://yanglab.westlake.edu.cn/software/gcta/#fastBAT>

sumFREGAT, <https://cran.r-project.org/web/packages/sumFREGAT>

Psychiatric Genomics Consortium, <https://www.med.unc.edu/pgc/download-results>

dbGaP, <https://www.ncbi.nlm.nih.gov/gap>

TB's GitHub, <https://github.com/takiy-berrandou/LDAK-GBAT-paper-scripts>

Data and code availability

Instructions for running LDAK-GBAT are available on the LDAK website, while sample code for the analyses in this paper are on the GitHub page of TB. We applied for and downloaded individual-level UK Biobank data (via Application 21432) from the UK Biobank website. We applied for study accession phs001672.v7.p1 and downloaded summary statistics for the Million Veterans Project GWAS from the dbGaP website. We downloaded summary statistics for the Psychiatric Genomics Consortium from the consortium's website. See Web resources for links to each website.

Figure titles and legends

Figure 1: Comparing gene-based association testing tools for the ten UK Biobank phenotypes. Bars report the mean number of genome-wide significant genes ($P \leq 2.8 \times 10^{-6}$) for LDAK-GBAT and five existing tools. Segments mark 95% confidence intervals for the means.

Figure 2: Comparing LDAK-GBAT and single-SNP analysis for the ten UK Biobank phenotypes. Bars report the mean number of genome-wide significant genes ($P \leq 2.8 \times 10^{-6}$) for LDAK-GBAT (using 50,000 individuals) and the number of genes containing a SNP with $P < 5e-8$ from single-SNP analysis (using 50,000, 60,000 or 70,000 individuals). Segments mark 95% confidence intervals for the means.

Figure 3: Varying the heritability model for the ten UK Biobank phenotypes. We run LDAK-GBAT using seven heritability models, defined by $E[h^2_j] \sim [p_j(1-p_j)]^{1+\alpha}$, where p_j is the MAF of SNP_j, and α is -1.25, -1, -0.75, -0.5, -0.25, 0 or 0.25. Bars report the mean number of genome-wide significant genes ($P \leq 2.8 \times 10^{-6}$). Segments mark 95% confidence intervals for the means.

Figure 4: Comparing gene-based association testing tools for the 99 ICD10, MVP and PGC phenotypes. Bars report the mean number of genome-wide significant genes ($P \leq 2.8 \times 10^{-6}$) for LDAK-CBAT and three existing tools, and the number of genes with a SNP with $P < 5e-8$ from single-SNP analysis, across the 72 ICD10 phenotypes (left), the 18 MVP phenotypes (middle) and nine PGC phenotypes (right). Segments mark 95% confidence intervals for the means.

Figure 5: Clumping results from LDAK-GBAT. Segments provide p-values from LDAK-GBAT when analyzing the MVP phenotype type 2 diabetes; red segments indicate significant genes ($P \leq 2.8 \times 10^{-6}$) that are selected by clumping, gold segments indicate significant genes that are discarded by clumping, while gray segments indicate non-significant genes. Blue points report p-values for genic SNPs from single-SNP analysis (points above the horizontal line have $P \leq 5 \times 10^{-8}$).

Tables

Table 1: Computational efficiency of different tools for gene-based association testing. Average time and memory required to analyze each of the ten UKBB phenotypes on a single CPU.

References

- 1 Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat Rev Genet* **20**, 467-484, doi:10.1038/s41576-019-0127-1 (2019).
- 2 Neale, B. M. & Sham, P. C. The future of association studies: gene-based analysis and replication. *Am J Hum Genet* **75**, 353-362, doi:10.1086/423901 (2004).
- 3 Kang, G., Jiang, B. & Cui, Y. Gene-based Genomewide Association Analysis: A Comparison Study. *Curr Genomics* **14**, 250-255, doi:10.2174/13892029113149990001 (2013).
- 4 Wang, K., Li, M. & Bucan, M. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* **81**, 1278-1283, doi:10.1086/522374 (2007).
- 5 Ballard, D. H., Cho, J. & Zhao, H. Comparisons of multi-marker association methods to detect association between a candidate region and disease. *Genet Epidemiol* **34**, 201-212, doi:10.1002/gepi.20448 (2010).
- 6 Basu, S. *et al.* A rapid gene-based genome-wide association test with multivariate traits. *Hum Hered* **76**, 53-63, doi:10.1159/000356016 (2013).
- 7 Listgarten, J. *et al.* A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics* **29**, 1526-1533, doi:10.1093/bioinformatics/btt177 (2013).
- 8 Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779, doi:10.1371/journal.pmed.1001779 (2015).
- 9 de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* **11**, e1004219, doi:10.1371/journal.pcbi.1004219 (2015).
- 10 Bakshi, A. *et al.* Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. *Sci Rep* **6**, 32894, doi:10.1038/srep32894 (2016).
- 11 Svishcheva, G. R., Belonogova, N. M., Zorkoltseva, I. V., Kirichenko, A. V. & Axenovich, T. I. Gene-based association tests using GWAS summary statistics. *Bioinformatics* **35**, 3701-3708, doi:10.1093/bioinformatics/btz172 (2019).
- 12 Svishcheva, G. R. A generalized model for combining dependent SNP-level summary statistics and its extensions to statistics of other levels. *Sci Rep* **9**, 5461, doi:10.1038/s41598-019-41827-5 (2019).
- 13 Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* **70**, 214-223, doi:10.1016/j.jclinepi.2015.09.016 (2016).
- 14 Sullivan, P. F. *et al.* Psychiatric Genomics: An Update and an Agenda. *Am J Psychiatry* **175**, 15-27, doi:10.1176/appi.ajp.2017.17030283 (2018).
- 15 Corbeil, R. R. & Searle, S. R. Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model. *Technometrics* **18**, doi:10.2307/1267913 (1976).
- 16 Marees, A. T. *et al.* A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int J Methods Psychiatr Res* **27**, e1608, doi:10.1002/mpr.1608 (2018).
- 17 Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet* **91**, 1011-1021, doi:10.1016/j.ajhg.2012.10.010 (2012).
- 18 Speed, D. *et al.* Reevaluation of SNP heritability in complex human traits. *Nat Genet* **49**, 986-992, doi:10.1038/ng.3865 (2017).

- 19 Speed, D., Holmes, J. & Balding, D. J. Evaluating and improving heritability models using summary statistics. *Nat Genet* **52**, 458-462, doi:10.1038/s41588-020-0600-y (2020).
- 20 Speed, D., Kaphle, A. & Balding, D. J. SNP-based heritability and selection analyses: Improved models and new results. *Bioessays* **44**, e2100170, doi:10.1002/bies.202100170 (2022).
- 21 International HapMap, C. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58, doi:10.1038/nature09298 (2010).
- 22 Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* **51**, 584-591, doi:10.1038/s41588-019-0379-x (2019).