nature research

Peer Review Information

Journal: Nature Genetics
Manuscript Title: A generalized linear mixed model association tool for biobank-scale data
Corresponding author name(s): Professor Jian Yang

Reviewer Comments & Decisions:

Decision Letter, initial version:

11th Feb 2021

Dear Jian,

Your Technical Report, "FastGWA-GLMM: a generalized linear mixed model association tool for biobank-scale data" has now been seen by 3 referees. You will see from their comments copied below that while they find your work of considerable potential interest, they have raised some concerns that must be addressed, as well as some suggestions for improvement that we feel would greatly add to the impact of your tool.

In light of these comments, we cannot accept the manuscript for publication, but would be very interested in considering a revised version that addresses these issues.

In brief, Reviewer #1 is critical of the degree of advance presented; but does make some suggestions that would add to the study. We note that, as REGENIE has not yet been published, we would not insist on extensive benchmarking of it versus fastGWA-GLMM and SAIGE, but it would no doubt be of interest to the reader.

Reviewers #2 and #3 are both strongly supportive. They make a number of constructive comments for improvement, for instance by extended benchmarking (e.g. as per Reviewer #2, comparing the performance for PGS construction; and per Reviewer #3, larger simulations). We think that Reviewer #3's suggestion that fastGWA-GLMM could be extended to gene burden testing to be a particularly insightful and interesting one, and would be very pleased to see if you and your co-authors could achieve this in a revision; although we accept that this may be beyond the scope of the current manuscript.

We hope you will find the referees' comments useful as you decide how to proceed. If you wish to

submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to identifying key priorities that should be addressed in revision and sometimes overruling referee requests that are deemed beyond the scope of the current study. We hope that you will find the prioritised set of referee points to be useful when revising your study. Please do not hesitate to get in touch if you would like to discuss these issues further.

If you choose to revise your manuscript taking into account all reviewer and editor comments, please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

If revising your manuscript:

*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Technical Report format instructions, available here.
Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary:
https://www.nature.com/documents/nr-reporting-summary.pdf
It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.
A revised checklist is essential for re-review of the paper.

Please be aware of our guidelines on digital image standards.

You may use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

If you wish to submit a suitably revised manuscript we would hope to receive it within 6 months. If you cannot send it within this time, please let us know. We will be happy to consider your revision so long as nothing similar has been accepted for publication at Nature Genetics or published elsewhere.

Should your manuscript be substantially delayed without notifying us in advance and your article is eventually published, the received date would be that of the revised, not the original, version.

Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

Thank you for the opportunity to review your work.

Sincerely,

Michael Fletcher, PhD
Associate Editor, Nature Genetics

ORCID: 0000-0003-1589-7087

Referee expertise: statistical genetics.

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
This paper proposes fastGWA-GLMM, which extends the sparse GRM based linear mixed model, developed by the same group, to GLMM. The method basically incorporates approaches used in GMMAT (penalized quasi-likelihood (PQL) and average information) and SAIGE (saddlepoint approximation (SPA)) with sparse GRM. The authors showed that the fastGWA-GLMM can be much faster than SAIGE. The authors applied it to UK-Biobank data and showed that the method works well.

Although it is a useful method for biobank data, the overall novelty/innovation is weak, as it is a simple extension of SAIGE with sparse GRM. The main components of the method (sparse GRM+PQL+SPA) were already developed and combining them is not a difficult task. In addition, the authors do not compare it with Regenie (https://rgcgithub.github.io/regenie/), another recently developed method for biobank size data, which can also analyze binary traits. I have a few particular comments.

1. As I mentioned, the authors should compare it with Regenie.

2. It is interesting that SAIGE has missed many signals, especially for irritability. I am wondering

whether it is because of the proximal contamination, so chromosome 6 MHC region association signals are missed. How does the result look like if LOCO option is used?

3. Supplementary Table 5. It is somewhat misleading to show the number of significant SNPs only, as most of the significance is due to LD. For example, Thyroid cancer, although there are 144 significantly associated SNPs, the manhattan plot shows that all of them from 1 loci. The authors should include the number of significantly association after clumping.

Reviewer #2:
Remarks to the Author:
Jiang et al. propose a new and extremely computationally efficient implementation of a generalised linear mixed model for GWAS, which they name fastGWA-GLMM. They find that their approach is ~37 times more efficient than SAIGE, which is currently the only other similar GWAS software capable of analysing biobank data sizes, such as the UKB data. Using extensive simulations they also show that fastGWA-GLMM method does otherwise perform similarly to SAIGE in terms of power and false positive rates. These improvements allows them to apply their method to common and rare variants, and they apply it to thousands of traits in the UKB and provide the results in their nice online platform. In summary I found the paper and their analysis to be very impressive. However, given the large number of traits analyzed, I was also surprised by the relatively few number of new quasi-independent genetic associations discovered. Nevertheless, I believe this software will likely be used and inspire future software development in this direction. Several comments and suggested improvements are listed below.

Comments:

1. The authors do not compare to BOLT-LMM (Loh et al. NG 2015), arguing that it can't really be applied when case-control ratios are small (<0.1 or <0.01), as shown in the SAIGE paper. However, if we believe that the effective sample size is 4/((1/N_cases)+(1/N_controls)), then it follows that the loss in power would generally not be huge if we down-sample controls such that case-control ratios become 1/10. I would be interested in such a comparison, as I suspect BOLT-LMMs non-infinitesimal component could counter the loss in power for some traits.

2. You note that the FPR for SAIGE seems deflated. However, when looking at the most significant variants, which are the ones we care about, it seems less deflated. Also the AUC for SAIGE seems better than for fastGWA-GLMM. Hence, it's unclear to me whether SAIGE is actually worse. Maybe calculating the AUC for small FRP (the right side of the AUC plot), would be of interest, because in GWAS we usually want to apply a stringent significance threshold.

3. Another measurement for benchmarking would be to examine how useful the methods are for generating polygenic risk scores. Here you would train a polygenic predictor (e.g. using methods like SBayesR, PRScs, LDpred2, PRScs, etc.) based on the summary statistics from SAIGE vs. fastGWA-GLMM, and predict into an independent dataset (or a test subset).

4. In suppl fig 6-8, both SAIGE and fastGWA-GLMM seem to be a bit too conservative. Do you have an idea why?

Minor comments:

5. Arguing that SAIGE requires 13GB instead of 4GB memory (fastGWA-GLMM), is not really a limitation using modern computers.

6. Page 4, line 104: Maybe briefly remind the reader how P is related to \pi. (i.e. it's the inverse of ..)

7. fastGWA-GLMM is a mouthful. I get it, but maybe something simpler would be nice. (You're welcome to ignore this comment.)

Reviewer #3:
Remarks to the Author:
The manuscript describes an improvement to the fastGWA software to allow generalised linear mixed model genome-wide association analyses. The key advantages of this over fastGWA is that it provides well calibrated statistics for rare variants and extreme case/control ratios. The only other program able to do this is SAIGE, but the authors show that fastGWA is significantly more efficient and faster than SAIGE.

This a very well written paper with clear results and fastGWA is likely to become an important tool for very large-scale GWAS studies. I have only two concerns.

1. While there is a need for more efficient single variant GWAS analyses, one important addition to SAIGE is the ability to perform burden testing across rare variants for gene units. For very rare variants this is likely to be more important that single variant tests (e.g. looking at all loss of function variants in each gene). And this is becoming more important as exome sequencing becomes available in ever larger numbers - for example, UK Biobank will soon release exome sequencing data on all 500,000 individuals. Could fastGWA be easily adapted to perform gene burden testing? I think this would make it an even more attractive program.

2. It's stated in the introduction that fastGWAS-GLMM is particularly important for studies larger than UK Biobank - but no results or simulations are provided for studies in the millions. And I think it would be interesting to see these results included.

Author Rebuttal to Initial comments

Responses to the Reviewers

We thank the reviewers for their constructive comments which have helped to improve our manuscript. We have addressed all the reviewers' comments point-by-point below (in blue) in this document and made the corresponding changes in the manuscript files (highlighted in yellow). Here is a brief summary of the main changes.

1) To demonstrate the scalability of fastGWA-GLMM to GWAS data larger than the UK Biobank (UKB), we have applied fastGWA-GLMM to a pseudo GWAS data set with two million individuals generated based on the UKB data. We show that either the runtime or memory usage of fastGWA-GLMM was still linear to sample size (n) with n ranging from 400k to 2 million, demonstrating the scalability of fastGWA-GLMM to data with millions of individuals. For instance, when using 16 CPUs, fastGWA-GLMM was easily scalable to data with n = 1 million (runtime = 8.3 hours and memory usage = 14.5 GB), and it took less than a day (runtime = 17.0 hours and memory usage = 31.6 GB) to run a GWA with n = 2 million.
2) We have extended the single-variant association test in fastGWA-GLMM to a burden test (named fastGWA-BB), with an efficient implementation in GCTA. We have also implemented in GCTA a flexible set-based test (the ACAT-V test; *Liu et al. 2019 AJHG*) that only requires p-values from a single-variant test as input. Both methods showed well-calibrated test statistics in simulation.
3) We have compared fastGWA-GLMM with the latest version of REGENIE (v2.0.2, released in April 2021). When benchmarked using the UK Biobank data, fastGWA-GLMM was ~5.6-fold faster than REGENIE in the analysis of a single trait. Although REGENIE can save computing cost by analysing multiple traits jointly, the runtime per-trait of fastGWA-GLMM was still ~2.5-fold faster than that of REGENIE in the analysis of 50 traits. When benchmarked using the pseudo cohort of two million individuals, on average fastGWA-GLMM was at least 4.0-fold faster than REGENIE per trait, and REGENIE could not finish the job for 50 traits as it exceeded the 4-week time limit.
4) We have updated the runtime and memory usage of SAIGE using its latest version (v0.44.5, released on 21 Apr 2021).

Reviewer 1

Remarks to the Author:

This paper proposes fastGWA-GLMM, which extends the sparse GRM based linear mixed model, developed by the same group, to GLMM. The method basically incorporates approaches used in GMMAT (penalized quasi-likelihood (PQL) and average information) and SAIGE (saddlepoint approximation (SPA)) with sparse GRM. The authors showed that the fastGWA-GLMM can be much faster than SAIGE. The authors applied it to UK-Biobank data and showed that the method works well.

Although it is a useful method for biobank data, the overall novelty/innovation is weak, as it is a simple extension of SAIGE with sparse GRM. The main components of the method (sparse GRM+PQL+SPA) were already developed and combining them is not a difficult task. In addition, the authors do not compare it with REGENIE (https://rgcgithub.github.io/regenie/), another recently developed method for biobank size data, which can also analyze binary traits. I have a few particular comments.

Re: We thank the reviewer for the summary of our study. Regarding the comment on novelty, we believe that developing an ultra-efficient genome-wide association analysis tool that is scalable to cohorts with sample sizes of the order of 100,000s or even millions (see our response to comment #2 from Reviewer #3 for more details) is a significant technological breakthrough, not to mention the methodological innovation (e.g., the fastGWA-GLMM-REML method) that we have made to achieve this goal. Our benchmark analysis showed that fastGWA-GLMM was orders of magnitude faster than the existing tool, SAIGE (**Supplementary Table 1**). Our additional analysis showed that the relationship between the runtime (or memory usage) of fastGWA-GLMM and sample size (*n*) was still linear when *n* was larger than 400k (**Figure 2** and **Supplementary Tables 4,5**), suggesting the scalability of fastGWA-GLMM to data with millions of individuals. For instance, when using 16 CPUs, fastGWA-GLMM was easily scalable to data with *n* = 1 million (runtime = 8.3 hours and memory usage = 14.5 GB), and it took less than a day (runtime = 17.0 hours and memory usage = 31.6 GB) to run a GWA with *n* = 2 million.

1. As I mentioned, the authors should compare it with REGENIE.

Re: We thank the reviewer for this suggestion and have compared fastGWA-GLMM with the latest version of REGENIE (v2.0.2, released in April 2021). We first compared the statistical properties between fastGWA-GLMM and REGENIE using the same simulated data as used to compare between fastGWA-GLMM and SAIGE. Overall, the false-positive rate (FPR) and power of REGENIE were similar to those of fastGWA-GLMM, although the test statistics from REGENIE tended to be larger than those from fastGWA-GLMM when the disease prevalence was relatively high (e.g., > 0.05) but smaller than those from fastGWA-GLMM when the prevalence was low (**Supplementary Figures 18-20**). We next applied REGENIE to the 8 traits used in our real data analysis and found that the GWAS signals identified by REGENIE were almost identical to those by fastGWA-GLMM (**Supplementary Figures 13**, **Supplementary Table 8**).

We then benchmarked the computational performance of REGENIE on the same computing platform as used to benchmark fastGWA-GLMM and SAIGE (i.e., 80 GB memory and 8 Intel Xeon Gold 6148 CPUs). Considering that REGENIE can save computing cost by analysing multiple traits simultaneously,

we performed the benchmark analysis in three scenarios using the UKB data (n = 400k): a single trait, 8 traits (the set of traits included in our real data analysis), and 50 traits (the number used in the REGENIE preprint). We reserved 80 and 500 GB disk space for REGENIE for the analysis of 8 and 50 traits, respectively, to avoid massive memory usage. The runtime of REGENIE in the single trait analysis was 27.5 hours, ~5.6 times slower than that of fastGWA-GLMM (4.9 hours) (**Supplementary Table 12**). When analyzing 8 traits jointly, the runtime of REGENIE was 97.2 hours, corresponding to 12.2 hours per trait, which was still ~2.5 times slower than fastGWA-GLMM (**Supplementary Table 12**). The runtime of REGENIE per trait remained similar at 12.2 hours when the number of traits increased from 8 to 50 (overall runtime of 614.0 hours for 50 traits) (**Supplementary Table 12**). Hence, the runtime of fastGWA-GLMM per trait was 2.5 to 5.6 times faster than that of REGENIE, depending on the number of traits analysed jointly in REGENIE. Moreover, grouping phenotypes for a joint analysis increases the time per run, which will become more challenging when applied to larger data sets. When we applied REGENIE to a pseudo cohort of two million individuals, on average fastGWA-GLMM was at least 4.0-fold faster than REGENIE per trait, and REGENIE could not finish the job for 50 traits as it exceeded the 4-week time limit (**Supplementary Table 13**).

We have revised the manuscript accordingly based on the additional analyses described above (lines 329-349 and **Supplementary Note 13**).

2. It is interesting that SAIGE has missed many signals, especially for irritability. I am wondering whether it is because of the proximal contamination, so chromosome 6 MHC region association signals are missed. How does the result look like if LOCO option is used?

Re: We thank the reviewer for pointing this out. There are a number of signals missed by SAIGE for three traits, i.e., mood swings, irritability and asthma (**Supplementary Figures 13a-c**). To confirm whether this is due to proximal contamination, we re-ran SAIGE with the leave-one-chromosome-out (LOCO) option for all the 8 traits used in our real data analysis and found that nearly all the missing signals for mood swings, irritability and asthma could be restored by SAIGE-LOCO (**Supplementary Figures 16a-c**, **Supplementary Table 8**). For the other 5 traits with low or moderate prevalence, we did not observe an improvement of SAIGE-LOCO over SAIGE (**Supplementary Figures 16d-h**, **Supplementary Table 8**). Overall, across all the 8 traits, the signals identified by SAIGE-LOCO are highly consistent with those by fastGWA-GLMM or REGENIE (**Supplementary Figures 13** and **16**).

We have commented on this and included the new results in the revised manuscript (lines 229-231, 302-307).

3. Supplementary Table 5. It is somewhat misleading to show the number of significant SNPs only, as most of the significance is due to LD. For example, Thyroid cancer, although there are 144 significantly associated SNPs, the manhattan plot shows that all of them from 1 loci. The authors should include the number of significantly association after clumping.

Re: We have removed this table and included the number of signals after clumping for all the methods (including SAIGE-LOCO and REGENIE) in **Supplementary Table 8**. We have also modified the relevant statement accordingly (lines 227-229).

Reviewer 2

Remarks to the Author:

Jiang et al. propose a new and extremely computationally efficient implementation of a generalised linear mixed model for GWAS, which they name fastGWA-GLMM. They find that their approach is ~37 times more efficient than SAIGE, which is currently the only other similar GWAS software capable of analysing biobank data sizes, such as the UKB data. Using extensive simulations they also show that fastGWA-GLMM method does otherwise perform similarly to SAIGE in terms of power and false positive rates. These improvements allows them to apply their method to common and rare variants, and they apply it to thousands of traits in the UKB and provide the results in their nice online platform. In summary I found the paper and their analysis to be very impressive. However, given the large number of traits analyzed, I was also surprised by the relatively few number of new quasi-independent genetic associations discovered. Nevertheless, I believe this software will likely be used and inspire future software development in this direction. Several comments and suggested improvements are listed below.

Re: We thank the reviewer for the positive remarks.

Comments:

1. The authors do not compare to BOLT-LMM (Loh et al. NG 2015), arguing that it can't really be applied when case-control ratios are small (<0.1 or <0.01), as shown in the SAIGE paper. However, if we believe that the effective sample size is 4/((1/N_cases)+(1/N_controls)), then it follows that the loss in power would generally not be huge if we down-sample controls such that case-control ratios become 1/10. I would be interested in such a comparison, as I suspect BOLT-LMMs non-infinitesimal component could counter the loss in power for some traits.

Re: We thank the reviewer for this suggestion. We have applied the BOLT-LMM non-infinitesimal model (BOLT-LMM-Mix) to two traits with extremely low case-control ratios, i.e., large bowel cancer

(case-control ratio = 0.0014) and thyroid cancer (case-control ratio = 0.00087). We first confirmed that the BOLT-LMM-Mix test statistics were inflated for rare variants (MAF < 0.01) owing to the highly unbalanced case-control ratios (**Supplementary Figure 20**). We next re-ran BOLT-LMM-Mix by down-sampling the controls to achieve a case-control ratio of 1/10 (n_{case} : $n_{control}$ = 636 : 6,360 for large bowl cancer and n_{case} : $n_{control}$ = 379 : 3,790 for thyroid cancer). We found that although the strategy of down-sampling controls could reduce the inflation at minor cost of statistical power, the remaining inflation was sufficiently large to generate false-positive associations for rare variants (**Supplementary Figure 20**). We have commented on this analysis in the revised manuscript (lines 285-289).

As mentioned in our manuscript (lines 278-281), another disadvantage of applying LMM methods to binary traits is that the effect size is not directly interpretable and needs to be transformed to odds ratio by approximation (*Lloyd-Jones et al. 2018 Genetics*).

2. You note that the FPR for SAIGE seems deflated. However, when looking at the most significant variants, which are the ones we care about, it seems less deflated. Also the AUC for SAIGE seems better than for fastGWA-GLMM. Hence, it's unclear to me whether SAIGE is actually worse. Maybe calculating the AUC for small FRP (the right side of the AUC plot), would be of interest, because in GWAS we usually want to apply a stringent significance threshold.

Re: We agree with the reviewer that for GWAS discovery, a truncated AUC at a small false-positive rate (FPR) level is more informative than the overall AUC. Therefore, we calculated truncated AUCs at 4 different FPR levels, namely α = 0.05, 5×10^{-3}, 5×10^{-4}, and 5×10^{-5} (**Supplementary Note 9**). The bottom line is that the truncated AUCs of fastGWA-GLMM and SAIGE were extremely similar at all the four FPR levels (**Supplementary Figure 7**), suggesting very similar levels of power for the two methods when evaluated at the same FPR level. We have commented on this in the revised manuscript (lines 191-196).

3. Another measurement for benchmarking would be to examine how useful the methods are for generating polygenic risk scores. Here you would train a polygenic predictor (e.g., using methods like SBayesR, PRScs, LDpred2, PRScs, etc.) based on the summary statistics from SAIGE vs. fastGWA-GLMM, and predict into an independent dataset (or a test subset).

Re: We have benchmarked the performance of fastGWA-GLMM and SAIGE in disease risk prediction using two polygenic risk score (PRS) methods, the conventional Pruning + Thresholding (P + T) method and the more recently developed SBayesR method (*Luke et al. 2019 Nat Comm*). The workflow is as follows.

1) We split the UKB cohort into a training ($n_{training}$ = 380,000) and a validation set (n_{test} = 20,000). To avoid relatedness between the training and validation sets, the 20,000 participants in the validation set were randomly selected from 245,000 singletons (i.e., estimated genetic relatedness < 0.05 with any other individuals in the UKB).
2) We ran a GWAS in the training set with fastGWA-GLMM and SAIGE, respectively.
3) We carried out a P + T (also known as clumping) analysis in PLINK with a window size of 500 kb and LD r^2 threshold of 0.1 at 3 different significance levels (i.e., 0.05, 0.001, and 0.0001) based on the GWAS summary statistics from fastGWA-GLMM and SAIGE, respectively, and used the selected variants from each of the six scenarios (2 GWAS methods x 3 p-value thresholds) to compute a PRS in the validation set.
4) We performed an SBayesR analysis in GCTB v2.02 using the GWAS summary statistics from SAIGE and fastGWA-GLMM, respectively, and computed a PRS for each of the GWAS methods in the validation set. Note that we removed the variants in the MHC region for better convergence of the SBayesR model.
5) We regressed the observed disease status against the PRS along with a set of covariates (i.e., age, age2, sex, age×sex, age2×sex, and the first 20 PCs) in a generalized linear model (GLM) and calculated the AUC for both the null (fitting the covariates only) and alternative models.

We included in this analysis 5 of the 8 traits used in our real data analysis (note that we did not include the other three traits because of their low case-control ratios). For each trait, we repeated the pipeline above five times and used the mean AUC across replicates to compare methods. The result showed that there was a difference between the PRS methods, with SBayesR performing consistently better than P + T in general, in line with prior work (*Luke et al. 2019 Nat Comm; Ni et al. 2021 Biol Psychiatry*), but almost no difference between the two GWA methods (**Supplementary Table 11**).

We have included the analysis above in the revised manuscript (lines 321-327 and **Supplementary Note 12**).

4. In suppl fig 6-8, both SAIGE and fastGWA-GLMM seem to be a bit too conservative. Do you have an idea why?

Re: There are several reasons why SAIGE and fastGWA-GLMM seem very conservative in these figures. 1) In the previous version of the manuscript, the scale of the y-axis in some of the FPR plots (e.g., those in the original Supplementary Figures 6,7) was smaller than that in Figure 2, which exaggerated the differences visually. In the revised manuscript, we have harmonized the FPR plots so that the y-axis always starts from the origin (**Figure 3**, **Supplementary Figures 2-4, 8, 9, 11, 22, 23, 25, 26)**.

2) As noted in prior work (*Zhou et al. 2018 Nat Genet*) and also in this study (lines 360-365), the FPRs of both fastGWA-GLMM and SAIGE tend to be slightly lower than the expected value owing to the use of the SPA correction to correct for case-control imbalance. We have shown by an additional analysis that without the SPA correction, the FPR of fastGWA-GLMM was consistent with the expected value when the case-control rate was relatively high (e.g., ≥ 0.1) but heavily inflated when the case-control rate became extreme (e.g., 0.001) (**Supplementary Figures 25,26**). Note that the FPR of SAIGE without SPA was deflated even when the case-control ratio was not extreme (**Supplementary Figures 25,26**).

3) We also noticed in the real data analysis that the SAIGE test statistics could be deflated due to proximal contamination (see our response to comment #2 from Reviewer #1 for more details), and such deflation could be mitigated by the use of the leave-one-chromosome-out (LOCO) strategy (lines 302-307 and **Supplementary Figures 18**).

Minor comments:

5. Arguing that SAIGE requires 13GB instead of 4GB memory (fastGWA-GLMM), is not really a limitation using modern computers.

Re: We agree and have focused our discussion on the linear relationship between the memory usage of fastGWA-GLMM and sample size. Our additional analysis shows that fastGWA-GLMM can be applied to GWAS data with two million individuals with reasonable memory usage (**Figure 2** and **Supplementary Tables 4,5**).

We have adjusted statement in the revised manuscript (lines 134-135).

6. Page 4, line 104: Maybe briefly remind the reader how P is related to \pi. (i.e. it's the inverse of ..)

Re: We have been specific about the equations to compute matrices **P** and **V** and how they are related to matrix $\boldsymbol{\pi}$ in the revised manuscript (lines 101-102).

7. fastGWA-GLMM is a mouthful. I get it, but maybe something simpler would be nice. (You're welcome to ignore this comment.)

Re: We thank the reviewer for the suggestion. We wish to keep the name fastGWA-GLMM because it is straightforward to be seen as an extension of our previous tool fastGWA under the generalized linear mixed model (GLMM) framework. We also use fastGWA-B (where "B" stands for "binary") as an

alias of fastGWA-GLMM when necessary (e.g., we use fastGWA-BB to denote the fastGWA-GLMM burden test and fastGWA-B-REML to denote fastGWA-GLMM REML).

Reviewer 3

Remarks to the Author:

The manuscript describes an improvement to the fastGWA software to allow generalised linear mixed model genome-wide association analyses. The key advantages of this over fastGWA is that it provides well calibrated statistics for rare variants and extreme case/control ratios. The only other program able to do this is SAIGE, but the authors show that fastGWA is significantly more efficient and faster than SAIGE.

This a very well written paper with clear results and fastGWA is likely to become an important tool for very large-scale GWAS studies. I have only two concerns.

Re: We thank the reviewer for the positive remarks.

1. While there is a need for more efficient single variant GWAS analyses, one important addition to SAIGE is the ability to perform burden testing across rare variants for gene units. For very rare variants this is likely to be more important that single variant tests (e.g., looking at all loss of function variants in each gene). And this is becoming more important as exome sequencing becomes available in ever larger numbers - for example, UK Biobank will soon release exome sequencing data on all 500,000 individuals. Could fastGWA be easily adapted to perform gene burden testing? I think this would make it an even more attractive program.

Re: We thank the reviewer for this comment. We understand this comment as a suggestion to develop a flexible set-based association method to test the aggregate effect of multiple variants, e.g., all the rare variants in or around a gene. Developing a novel gene-based test under the fastGWA-GLMM framework requires heavy method development, which warrants further studies. Nevertheless, we have achieved to implement in GCTA two gene-based methods, the fastGWA-GLMM burden test (named fastGWA-BB) and the Aggregated Cauchy Association Test based on Variant-level p-values (ACAT-V; *Liu et al. 2019 AJHG*).

FastGWA-BB is an extension of the single-variant test in fastGWA-GLMM to assess the association of a weighted rare allele count of all the tested rare variants in or around a gene with the phenotype, conditioning on the sparse genetic relationship matrix (**Supplementary Note 4**). The variants weighting method used in fastGWA-BB follows that in SAIGE-Burden (*Zhou et al., 2020 Nat Genet*),

and also very likely in REGENIE-Burden, although we have not found the method details of REGENIE-Burden. As with the single-variant tests such as SAIGE and fastGWA-GLMM, the SPA correction is applied to adjust the gene-based test p-values for case-control imbalance. The runtime of fastGWA-BB for a genome-wide analysis of the UKB data (456,348 individuals and 11,842,647 variants) was ~3 hours, given 8 CPUs and 20 GB memory.

Compared to fastGWA-BB and other set-based methods (*Chen et al., 2019 AJHG; Zhou et al., 2020 Nat Genet*), ACTA-V is more general and flexible because it only requires summary statistics and does not rely on any specific assumption about the distribution or the directions of the effects of the rare alleles. The tail of the null distribution of the ACAT-V test statistic can be well approximated by a Cauchy distribution regardless the linkage disequilibrium (LD) structure of the variants (*Liu et al. 2019 AJHG*); hence, LD information is not needed. We have implemented an efficient version of ACAT-V in GCTA by C/C++ code, which only takes ~30 second to run a genome-wide gene-based test.

We have benchmarked fastGWA-BB and ACAT-V (both implemented in GCTA) as well as REGENIE-Burden (https://rgcgithub.github.io/regenie/options/#burden-testing) by simulations using a subset of the UKB data (n = 100,000). The result showed that all the three methods were well calibrated under the null model regardless of the disease prevalence (**Supplementary Figure 13**). To quantify the power under the alternative model, we varied the proportion of variants being causal in a gene (5%, 20% or 50%) and the directions of variant effects (random vs. consistent). When the directions of the effects of the rare alleles were simulated at random, ACAT-V was always more powerful than fastGWA-BB and REGENIE-Burden, regardless of the proportion of variants being causal (**Supplementary Figures 14a-c, 15a-c**). In the scenario where the effects of all the rare alleles in a gene were in a consistent direction, ACAT-V was still more powerful than fastGWA-BB and REGENIE-Burden when the proportion of variants being causal was relatively small (i.e., 5%), but had lower power than the two burden tests when the proportion of variants being causal was large (i.e., 20% and 50%) (**Supplementary Figures 14d-f, 15d-f**), in line with the results reported previously (*Liu et al. 2019 AJHG*). In all the simulation scenarios, we observed an extremely high Pearson correlation of >0.98 between the test statistics of fastGWA-BB and REGENIE-Burden (**Supplementary Table 6**).

We did not achieve to include SAIGE-Gene in the comparison above because of a bug in SAIGE-Gene v0.44.5. We have submitted a bug report at the SAIGE website: https://github.com/weizhouUMICH/SAIGE/issues/335.

We have applied ACAT-V and fastGWA-BB to the 8 traits used in our real data analysis. The results were in line with our observations from simulations that ACAT-V and fastGWA-BB performed differently under different patterns of genetic architecture (**Supplementary Figure 19** and **Supplementary Table 10**). We identified a total of 22 genes for 3 traits, with four genes in common between the methods, *IL4R*, *PBX2*, *NXPH4*, and *STAC3*, all associated with asthma. *IL4R* was a well-known gene related to asthma (*Chatila et al., 2004 Trends Mol Med; Wenzel et al., 2007 Am J Respir Crit Care Med*). *PBX2* was also a previously established asthma-associated gene (*Hirota et al., 2011 Nat Genet*). No prior evidence has been found for *NXPH4* and *STAC3*.

We have included a description of the fastGWA-BB method (lines 112-115, 496-505 and **Supplementary Note 4**), the ACAT-V method (lines 112-115, 505-510 and **Supplementary Note 5**), the simulations (**Supplementary Note 10**), and the real data analysis results (**Supplementary Note 11**) in the revised manuscript.

2. It's stated in the introduction that fastGWA-GLMM is particularly important for studies larger than UK Biobank - but no results or simulations are provided for studies in the millions. And I think it would be interesting to see these results included.

Re: To assess the scalability of fastGWA-GLMM to GWAS data beyond the size of the UKB, we generated a pseudo GWAS data set with two million individuals and 11,842,647 variants by duplicating the genotypes of each UKB individual multiple times (**Supplementary Note 7**). This pseudo cohort had a consistent level of relatedness as the UKB (GRM density = 3.9×10^{-6}). We used a threshold-liability model to simulate a binary disease trait with prevalence of 0.1, 10,000 causal variants (randomly sampled from all the variants), and heritability of liability of 0.2. We ran fastGWA-GLMM using the full set (n = 2 million) or a subset of the data (n = 200k, 400k, …, 1.8 million) on a computing platform with 80 GB memory and 8 or 16 CPUs. We repeated each of the analyses five times to compute the mean runtime and memory usage. It is reassuring that either the runtime or the memory usage of fastGWA-GLMM was still linear to sample size (n) when n was larger than 400k, despite whether 8 or 16 CPUs were used (**Figure 2 and Supplementary Tables 4,5**), consistent with the theory that the computational complexity of fastGWA-GLMM is close to $O(mn)$ with m being the number of variants. Note that the runtime and memory usage for n = 200k and 400k quantified in this pseudo cohort (**Supplementary Table 4**) were greater than those quantified in the UKB (**Supplementary Table 1,3**) largely because of the extra computational cost to deal with larger genotype data in the former.

Considering that fastGWA-GLMM only required 17.0 hours to run a GWAS analysis with two million individuals and ~12 million variants using 31.6 GB memory and 16 CPUs, we believe that it can be applied to GWAS data sets with millions of individuals.

We have included the additional analysis above in the revised manuscript (lines 70-72 and 140-150; **Figure 2, Supplementary Tables 4,5; Supplementary Note 7**).

Decision Letter, first revision:

7th Jul 2021

Dear Jian,

Your Technical Report, "FastGWA-GLMM: a generalized linear mixed model association tool for biobank-scale data" has now been seen by 3 referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

In brief, all three reviewers think that your manuscript has improved in revision, and appear supportive of an eventual publication.

There are two new comments from these reviews; but, in our opinion, they seem minor and, we believe, should be easily addressable.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to identifying key priorities that should be addressed in revision and sometimes overruling referee requests that are deemed beyond the scope of the current study. We hope that you will find the prioritized set of referee points to be useful when revising your study. Please do not hesitate to get in touch if you would like to discuss these issues further.

We therefore invite you to revise your manuscript taking into account all reviewer and editor comments. Please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Technical Report format instructions, available
here.
Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary:

https://www.nature.com/documents/nr-reporting-summary.pdf
It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.
A revised checklist is essential for re-review of the paper.

Please be aware of our guidelines on digital image standards.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within four to eight weeks. If you cannot send it within this time, please let us know.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,

Michael Fletcher, PhD
Associate Editor, Nature Genetics

ORCID: 0000-0003-1589-7087

Referee expertise: statistical genetics.

Reviewers' Comments:

Reviewer #1:

Remarks to the Author:
The authors mostly addressed my previous comments. I have additional comments on the computation time comparisons.

1. Currently, the evaluation was done on an 8 CPU core node, which I assume that one job was submitted with multithread. Although it is appropriate when measuring computation time for estimating parameters, when testing associations, a more realistic scenario is to submit 8 jobs (so one job for each CPU core), as researchers split the genome into many chunks and testing many phenotypes. So submitting many jobs of genotype chunk x phenotype. Based on this, the association computation time should be measured as a single thread CPU hour (not hours in an 8 CPU node). I think SAIGE was developed with this usage, so a more fair comparison can be done if the jobs were done in a single thread (i.e. single core). With this setup, we observed that the association test computation time of SAIGE is similar to (or even slightly faster than) REGENIE. Still, fastGWA-GLMM can be faster than both, but a fair comparison is needed.

2. Given the current setup (evaluation on 8 CPU cores), REGENIE is state-of-the-art in computation. The authors compared fastGWA-GLMM with REGENIE in the discussion, but I think it should be in the "Runtime and resource requirements" section, so the main comparison is with REGENIE.

Minor
Abstract, it is written "37 times faster than the state-of-the-art tool". As I mentioned previously, the comparison should be done with REGENIE (so up to 5 times faster than the stat-of art, based on 1 trait, Stable 12)

Reviewer #2:
Remarks to the Author:
I would like to thank the authors for fully addressing most of my comments. Regarding my first comment, the authors are right that "down-sampling" and applying BOLT-LMM does not seem to fully avoid biases for rare variants. However, it is still unclear to me whether "down-sampled" BOLT-LMM applied to common variants (BOLT-LMM-ds-cv) is really a bad approach. Indeed, I believe it might outcompete the proposed method for a number of traits/outcomes, because it can account for non-infinitesimal genetic architectures. I believe it would be great if you directly compared the results for BOLT-LMM-ds-cv and fastGWA-GLMM. Alternatively, you could address this in the discussion.

Reviewer #3:
Remarks to the Author:
The authors have done a very good job of responding to the reviewers concerns and I have no additional comments.

Author Rebuttal, first revision:

Response to reviewers' comments

We thank all the reviewers for their constructive comments which have helped to further improve our manuscript. We have addressed all the reviewers' comments point-by-point below (in blue) in this document and made the corresponding changes in the manuscript files (highlighted in yellow). Here is a summary of the main changes.

1) We have performed an additional analysis to quantify the association test computation time of fastGWA-GLMM, SAIGE, and REGENIE using a single CPU thread. The result shows that the association test step of fastGWA-GLMM is 7.2-fold faster than that of SAIGE and 13.0-fold faster than that of REGENIE.

2) We have demonstrated that the down-sampled BOLT-LMM analysis performed reasonably well for common variants.

Reviewer #1

Remarks to the Author: The authors mostly addressed my previous comments. I have additional comments on the computation time comparisons.

1. Currently, the evaluation was done on an 8 CPU core node, which I assume that one job was submitted with multithread. Although it is appropriate when measuring computation time for estimating parameters, when testing associations, a more realistic scenario is to submit 8 jobs (so one job for each CPU core), as researchers split the genome into many chunks and testing many phenotypes. So submitting many jobs of genotype chunk x phenotype. Based on this, the association computation time should be measured as a single thread CPU hour (not hours in an 8 CPU node). I think SAIGE was developed with this usage, so a more fair comparison can be done if the jobs were done in a single thread (i.e. single core). With this setup, we observed that the association test computation time of SAIGE is similar to (or even slightly faster than) REGENIE. Still, fastGWA-GLMM can be faster than both, but a fair comparison is needed.

Re: We thank the reviewer for the suggestion. We have performed an additional analysis to quantify the association test computation time of the three methods (i.e., fastGWA-GLMM, SAIGE, and REGENIE) using a single CPU thread. We show in **Supplementary Table 8** that under the one-thread setting, the association test step of fastGWA-GLMM is 7.2 times faster than that of SAIGE and 13.0 times faster than that of REGENIE. The observation that SAIGE is slightly faster than REGENIE under this one-trait-one-thread scenario is consistent with the referee's result. We have revised the manuscript accordingly (lines 171-174).

2. Given the current setup (evaluation on 8 CPU cores), REGENIE is state-of-the-art in computation. The authors compared fastGWA-GLMM with REGENIE in the discussion, but I think it should be in the "Runtime and resource requirements" section, so the main comparison is with REGENIE.

Re: We have moved the runtime result of REGENIE to the "Runtime and resource requirements" section (lines 152-174).

Minor

Abstract, it is written "37 times faster than the state-of-the-art tool". As I mentioned previously, the comparison should be done with REGENIE (so up to 5 times faster than the stat-of art, based on 1 trait, Stable 12)

Re: We have revised the statements accordingly in the Abstract (lines 16-18) and in the main text (lines 68-70, 277-280).

Reviewer #2

Remarks to the Author:

I would like to thank the authors for fully addressing most of my comments. Regarding my first comment, the authors are right that "down-sampling" and applying BOLT-LMM does not seem to fully avoid biases for rare variants. However, it is still unclear to me whether "down-sampled" BOLT-LMM applied to common variants (BOLT-LMM-ds-cv) is really a bad approach. Indeed, I believe it might outcompete the proposed method for a number of traits/outcomes, because it can account for non-infinitesimal genetic architectures. I believe it would be great if you directly compared the results for BOLT-LMM-ds-cv and fastGWA-GLMM. Alternatively, you could address this in the discussion.

Re: The reviewer is correct that the down-sampled BOLT-LMM analysis performed reasonably well for common variants (**Supplementary Figure 23**). We have commented on this in the revised manuscript (lines 309-311).

Reviewer #3

Remarks to the Author:

The authors have done a very good job of responding to the reviewers concerns and I have no additional comments.

Re: We thank the reviewer again for their time and effort reviewing our manuscript.

Decision Letter, second revision:

Our ref: NG-TR56536R1

21st Jul 2021

Dear Jian,

Thank you for submitting your revised manuscript "FastGWA-GLMM: a generalized linear mixed model association tool for biobank-scale data" (NG-TR56536R1).

We have considered your revision and the point-by-point response, and we believe that these have sufficiently addressed the remaining concerns and, thus, do not require another round of peer review.

Therefore, we'll be happy in principle to publish it in Nature Genetics, pending minor revisions to satisfy the referees' requests and to comply with our editorial and formatting guidelines.

If the current version of your manuscript is in a PDF format, please email us a copy of the file in an editable format (Microsoft Word or LaTex)-- we can not proceed with PDFs at this stage.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements in about a week. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Thank you again for your interest in Nature Genetics Please do not hesitate to contact me if you have any questions.

Sincerely,

Michael Fletcher, PhD
Associate Editor, Nature Genetics

ORCID: 0000-0003-1589-7087

Final Decision Letter:

In reply please quote: NG-TR56536R2 Yang

22nd Sep 2021

Dear Dr. Yang,

I am delighted to say that your manuscript "A generalized linear mixed model association tool for biobank-scale data" has been accepted for publication in an upcoming issue of Nature Genetics.

Prior to setting your manuscript, we may make minor changes to enhance the lucidity of the text and with reference to our house style. We therefore ask that you examine the proofs most carefully to ensure that we have not inadvertently altered the sense of your text in any way.

Once your manuscript is typeset and you have completed the appropriate grant of rights, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

Your paper will be published online after we receive your corrections and will appear in print in the next available issue. You can find out your date of online publication by contacting the Nature Press Office (press@nature.com) after sending your e-proof corrections. Now is the time to inform your Public Relations or Press Office about your paper, as they might be interested in promoting its publication. This will allow them time to prepare an accurate and satisfactory press release. Include your manuscript tracking number (NG-TR56536R2) and the name of the journal, which they will need when they contact our Press Office.

Before your paper is published online, we shall be distributing a press release to news organizations worldwide, which may very well include details of your work. We are happy for your institution or funding agency to prepare its own press release, but it must mention the embargo date and Nature Genetics. Our Press Office may contact you closer to the time of publication, but if you or your Press Office have any enquiries in the meantime, please contact press@nature.com.

Acceptance is conditional on the data in the manuscript not being published elsewhere, or announced in the print or electronic media, until the embargo/publication date. These restrictions are not intended to deter you from presenting your data at academic meetings and conferences, but any enquiries from the media about papers not yet scheduled for publication should be referred to us.

Please note that Nature Genetics is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

<B>Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. For submissions from January 2021, if your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route our standard licensing terms will need to be accepted, including our self-archiving policies. Those standard licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Please note that Nature Research offers an immediate open access option only for papers that were first submitted after 1 January, 2021.

You will not receive your proofs until the publishing agreement has been received through our system.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. Please let your coauthors and your institutions' public affairs office know that they are also welcome to order reprints by this method.

If you have not already done so, we invite you to upload the step-by-step protocols used in this manuscript to the Protocols Exchange, part of our on-line web resource, natureprotocols.com. If you complete the upload by the time you receive your manuscript proofs, we can insert links in your article that lead directly to the protocol details. Your protocol will be made freely available upon publication of your paper. By participating in natureprotocols.com, you are enabling researchers to more readily reproduce or adapt the methodology you use. Natureprotocols.com is fully searchable, providing your protocols and paper with increased utility and visibility. Please submit your protocol to https://protocolexchange.researchsquare.com/. After entering your nature.com username and password you will need to enter your manuscript number (NG-TR56536R2). Further information can be found at https://www.nature.com/nprot/.

Congratulations on the paper!

All the best,

Catherine

--

Catherine Potenski, PhD
Chief Editor
Nature Genetics
1 NY Plaza, 47th Fl.
New York, NY 10004
catherine.potenski@us.nature.com
https://orcid.org/0000-0002-4843-7071

on behalf of:

Michael Fletcher, PhD
Associate Editor, Nature Genetics
ORCID: 0000-0003-1589-7087