A resource-efficient tool for mixed model association analysis of large-scale data

Longda Jiang[1,$], Zhili Zheng[1,2,$], Ting Qi[1], Kathryn E. Kemper[1], Naomi R. Wray[1,3], Peter M. Visscher[1,3], Jian Yang[1,2,3,*]

[1] Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland 4072, Australia

[2] Institute for Advanced Research, Wenzhou Medical University, Wenzhou, Zhejiang 325027, China

[3] Queensland Brain Institute, The University of Queensland, Brisbane, Queensland 4072, Australia

[$] These authors contributed equally to this work.

[*] Correspondence: Jian Yang (jian.yang@uq.edu.au)

ABSTRACT

The genome-wide association study (GWAS) has been widely used as an experimental design to detect associations between genetic variants and a phenotype. Two major confounding factors, population stratification and relatedness, could potentially lead to inflated GWAS test-statistics and thereby spurious associations. Mixed linear model (MLM)-based approaches can be used to account for sample structure. However, genome-wide association (GWA) analyses in biobank samples such as the UK Biobank (UKB) often exceed the capability of most existing MLM-based tools especially if the number of traits is large. Here, we developed an MLM-based tool (called fastGWA) that controls for population stratification by principal components and relatedness by a sparse genetic relationship matrix for GWA analyses of biobank-scale data. We demonstrated by extensive simulations that fastGWA is reliable, robust and highly resource-efficient. We then applied fastGWA to 3,613 traits on 456,422 array-genotyped and imputed individuals and 2,090 traits on 46,191 whole-exome-sequenced (WES) individuals in the UKB.

INTRODUCTION

The genome-wide association study (GWAS) is a powerful experimental design to detect genetic variants associated with a phenotype of interest. Over the past decade, a number of statistical methods have been developed for GWAS, facilitating the discovery of thousands of genetic loci associated with complex traits and diseases [1,2]. In the early GWAS era, the most commonly used approach was linear or logistic regression [3-11], which is also the basis of most GWAS software tools [12-14]. The statistical power of a GWAS depends on the proportion of phenotypic variance explained by a variant and the sample size [15]. In other words, for the detection of variants with small effects, a large sample size is required. This can be achieved by a meta-analysis of a large number of cohorts even if the sample size of each individual cohort is limited (e.g. GIANT [16] and PGC [17]). Due to the substantial decrease in genotyping costs in recent years, sample sizes of GWAS have dramatically increased to 100,000s in single cohorts, such as the UK Biobank (UKB) [18] and the Biobank Japan Project [19]. These large cohorts not only provide new opportunities to make novel discoveries but also bring challenges in computing especially for methods based on multivariate models. New software tools based on linear regression (LR) have also been developed to accommodate the increasing scale of data, including PLINK1.9 [20] and BGENIE [18]. Population stratification [21,22] and relatedness [23,24] are the two major confounders in GWAS, which could potentially lead to spurious associations if not well-controlled for. In an LR analysis, the effect of population stratification is usually accounted for by fitting the first few eigenvectors (also called PCs) from a principal component analysis (PCA) of the SNP genotypes [25]; the confounding due to relatedness can be avoided by excluding one member of each pair of related individuals based on pedigree or SNP-derived relatedness [14,26], which, however, results in a loss of power, especially because the proportion of individuals with close relatives in the sample is expected to increase as biobanks get larger [18].

The mixed linear model (MLM) approach has been widely used in GWAS to control for population stratification and relatedness [27-40]. The basic principle is to test for association between each genetic variant and the phenotype, conditioning on the sample structure inferred from all the genome-wide SNPs [39]. However, the runtime of most existing MLM-based methods ranges from $O(MN^2)$ to $O(M^2N)$ [29,32,34,37-39], where M is the number of variants and N is the sample size. Several recent studies have focused on the application of MLM-based methods in biobank-scale data [41-43]. Yet it is still resource demanding to run MLM-based GWAS analyses with millions of genetic variants especially when the number of phenotypes to be analysed is large.

In this study, we propose an extremely resource-efficient approach to perform an MLM-based genome-wide association (GWA) analysis (called fastGWA), implemented in the software GCTA

package [26]. We show by extensive simulations that fastGWA is robust in controlling for false positive associations in the presence of population stratification and relatedness, and that fastGWA is ~8 times faster and only requires ~3% of RAM compared to the most efficient existing MLM-based GWAS tool in a very large sample (n = 400,000). We then demonstrate the utility of fastGWA by analysing GWAS data of 456,422 array-genotyped and imputed individuals (including close relatives) of European ancestry for 3,613 traits and a subset of 49,960 whole-exome sequence-based individuals for 2,090 traits in the UKB. All the summary statistics are publicly available at our data portal (**URLs**).

RESULTS

Overview of the methods

The fastGWA model can be written as

$$\boldsymbol{y} = x_{snp}\beta_{snp} + \boldsymbol{X}_c\boldsymbol{\beta}_c + \boldsymbol{g} + \boldsymbol{e} \quad [1]$$

where $\boldsymbol{y}$ is a vector of mean centred phenotypic records; x_{snp} is the mean-centred genotype variable of a SNP of interest across the individuals with its effect β_{snp}; $\boldsymbol{X}_c$ is the incidence matrix of fixed covariates (e.g., sex, age,and the first few PCs) with their corresponding coefficients $\boldsymbol{\beta}_c$; $\boldsymbol{g}$ is a vector of the total genetic effects captured by pedigree relatedness with $\boldsymbol{g} \sim N\ (0, \boldsymbol{\pi}\sigma_g^2)$; $\boldsymbol{\pi}$ is the family relatedness matrix (FAM) based on pedigree structure [44], e.g., 0.5 for a full-sib or parent-offspring pair; $\boldsymbol{e}$ is a vector of residuals with $\boldsymbol{e} \sim N(0, \mathbf{I}\sigma_e^2)$. The variance-covariance matrix of $\boldsymbol{y}$ is $\boldsymbol{V} = \boldsymbol{\pi}\sigma_g^2 + \boldsymbol{I}\sigma_e^2$. In practice, if pedigree information is missing or largely incomplete, $\boldsymbol{\pi}$ can be replaced by an SNP-derived genetic relationship matrix (GRM) with all the small off-diagonal elements (e.g., those < 0.05) set to zero. This is because it has been shown in a previous study [45] that the sparse GRM captures approximately the same proportion of phenotypic variance as the FAM does (confirmed by our simulation below). Here we present two closely related versions of our method, the fastGWA, based on sparse GRM computed from SNP genotype data, and the fastGWA-Ped, based on FAM constructed from pedigree information.

The fastGWA model imposes control over relatedness by pedigree information or realised sparse GRM with the effect of population stratification captured by the SNP-derived PCs. The variance components σ_g^2 and σ_e^2 are unknown but can be estimated by REML [46] or Haseman-Elston regression (HE regression) [47]. HE regression only requires solving a one-dimensional equation ($y_i y_j = \pi_{ij}\sigma_g^2 + e$) with runtime of $O(N)$ [47]. Recent work has suggested that HE regression provides unbiased estimation of variance components in an MLM with a slightly larger standard error compared to that from REML [48], and this would become less of a problem as samples get larger. In the presence of moderate to strong common environmental effects shared among

relatives, the genetic variance estimated from closely related individuals (e.g., pairs of individuals with relatedness coefficients > 0.05, see **Methods**) may be a better quantity than that estimated based on genetic relatedness between all pairwise individuals in the sample as in most existing MLM-based methods [27-40]. This is because the variance component estimated from close relatives captures the variation due to the overall additive genetic effect as well as that due to common environmental effects (see **Discussion** for details).

Once the estimates of variance components are obtained, the variance-covariance matrix $\boldsymbol{V}$ and its inverse can be computed efficiently using the sparse matrix algorithms implemented in Eigen C++ library (**URLs**). Therefore, β_{snp} can be estimated using the generalised least squares approach:

$$\hat{\beta}_{snp} = \frac{x_{snp}^T \boldsymbol{V}^{-1} \boldsymbol{y}}{x_{snp}^T \boldsymbol{V}^{-1} x_{snp}} \text{ with } \mathrm{var}\left(\hat{\beta}_{snp}\right) = \frac{1}{x_{snp}^T \boldsymbol{V}^{-1} x_{snp}} \qquad [2]$$

where the parameters used to compute $\boldsymbol{V}$ are unknown but can be replicated by the estimates from REML or HE regression under the null that $\beta_{snp} = 0$ as in most existing methods [27,29-31,37,39,40]. We have implemented fastGWA in the GCTA software [26] with a user-friendly command-line interface (**URLs**).

Runtime and resource requirements

Given that fastGWA is specifically designed for large-scale data, we chose to evaluate its computational performance (e.g., runtime and resource requirements) using the UKB data [18]. We randomly sampled subsets of individuals from the UKB (456,422 individuals and 18,138,215 imputed SNPs; **Methods**) with sample sizes ranging from 50,000 to 400,000, and compared GCTA-fastGWA with BOLT-LMM v2.3 [40,42] (a highly efficient MLM-based association tool) on a computing platform with 96 GB memory and 16 CPU cores with the runtime capped at 7 days (168 hours). The tests were performed using a real trait, body mass index (BMI), with an estimated SNP-based heritability of ~0.27 [49,50]. The SNP genotype data were stored in PLINK binary PED format [14,20]. Each test was repeated 10 times to obtain an average. The results showed that GCTA-fastGWA completed the analysis in ~5 hours for the 400K sample – approximately one-eighth of that of BOLT-LMM (~42 hours) (**Table 1**). While BOLT-LMM requires a running time of $O(MN^{1.5})$, the computational complexity of fastGWA is approximately $O(N^{*3}) + O(MN)$, where $O(N^{*3})$ is for the computation of the inverse of $\boldsymbol{V}$ (using Cholesky decomposition) and $O(MN)$ is for the computation of association statistics. Considering that $\boldsymbol{V}$ is very sparse in population-based biobank data such as the UKB, the actual computational complexity of inverting $\boldsymbol{V}$ is usually very low. A detailed speed comparison between the two methods can be found in **Table 1**. BOLT-LMM is optimized for genotype data in BGEN v1.2 format [51] (Po-Ru Loh, personal communication)

133 and our benchmark testing confirmed that BOLT-LMM with data in BGEN format reduced the
134 runtime of the "association" procedure (**Table 1**) by approximately ½ but did not have significant
135 improvement on the other procedures. While all the tests of the two methods were conducted on
136 the same computing platform (i.e., 96 GB memory and 16 CPU cores), the actual memory usage
137 differed substantially. GCTA-fastGWA used much less resource than BOLT-LMM (**Table 2**). For a
138 data set with sample size of 400K, GCTA-fastGWA required less than 2 GB of memory to complete
139 the whole computation, only ~3% of the usage of BOLT-LMM. Note that if the pedigree
140 information is not available, we will need to take the computing cost of the GRM into
141 consideration for GCTA-fastGWA (**Supplementary Note 3**). Nevertheless, the GRM computation
142 is often part of the quality control process and only needs to be done once for the analyses of all
143 traits, meaning that the additional computational cost for GCTA-fastGWA due to GRM
144 computation is not very expensive.
145
146 **False positive rate and statistical power**
147 We used extensive simulations to quantify the genomic inflation factor, the false positive rate
148 (FPR, number of false positive associations divided by the total number of tests) and the statistical
149 power of fastGWA in comparison with linear regression (implemented in PLINK1.9) and BOLT-
150 LMM (**Methods**). A sample of 100,000 individuals was generated by random sampling of
151 chromosome segments from a subset of the UKB data to mimic a cohort with substantial
152 population stratification and relatedness (**Supplementary Note 1**, **Supplementary Figure 1-4**).
153 One of the main aims of this simulation study was to investigate the influences of common
154 environmental effects on different association test methods. We generated the phenotypes from
155 a number of causal SNPs randomly sampled from all SNPs on the odd chromosomes, leaving those
156 on the even chromosomes as the null SNPs to quantify the genomic inflation/deflation in test-
157 statistics and the FPR. We mimicked the effect of population stratification by generating mean
158 phenotype difference between two populations and the effect of relatedness by specifying
159 different degrees of common environmental effects among close relatives (**Methods**). The
160 genomic inflation factor, or λ_{GC}, is defined as the median chi-squared statistic divided by its
161 expected value at the null SNPs [21,52].
162
163 Our simulation results showed that there was inflation in the test-statistics of null SNPs from LR-
164 All (i.e., LR analysis of all the individuals including close relatives) even in the absence of common
165 environmental effects (**Figure 1a**). This is because of the inter-chromosome correlations
166 between the causal and null SNPs induced by the relatedness in the sample. The test-statistics of
167 null SNPs from BOLT-LMM-Mix (a Bayesian mixture model) were inflated, and the inflation was
168 higher than that of LR-All. BOLT-LMM-Inf (BOLT-LMM infinitesimal model) is a special case of

169 BOLT-LMM-Mix, which assumes only a single distribution of the SNP effects (similar but
170 computationally more efficient than the MLM leave-one-chromosome-out (LOCO) [39] approach
171 implemented in GCTA). We also observed inflation in the test-statistics of null SNPs from BOLT-
172 LMM-Inf and the inflation increased as the variance explained by common environmental effects
173 increased (**Figure 1a**), likely because BOLT-LMM-Inf failed to capture the variance due to
174 common environmental effects by fitting all common SNPs on the other chromosomes as random
175 effects (**Supplementary Figure 5**; see below for more discussion).
176
177 In contrast, there was almost no inflation for both LR-unRel (i.e., LR analysis restricted to
178 unrelated individuals) and fastGWA. As the level of common environmental effects increased, the
179 genomic inflation factors of LR-All, BOLT-LMM-Mix, and BOLT-LMM-Inf all increased slightly, but
180 not for LR-unRel and fastGWA, demonstrating the robustness of fastGWA in accounting for
181 common environmental effects among relatives. We also found that the results from fastGWA-
182 Ped were very similar to those from fastGWA (**Supplementary Figure 6,** recognising that
183 pedigree relationships were known without error in simulation). Additionally, we quantified the
184 FPR using the null SNPs (i.e., all the SNPs on the even chromosomes), where FPR is defined as the
185 proportion of null SNPs with p-values < 0.05 in each simulation replicate. The FPRs of the
186 methods were in line with their observed genomic inflation factors (**Supplementary Figure 7**).
187
188 Next, we extended the simulation with larger numbers of causal SNPs. We kept the proportions
189 of variance explained by common environmental effects and population stratification the same
190 as that in one of the scenarios above (i.e., common environmental effects explained 10% of
191 phenotypic variance (V_p) among all relatives and population stratification explained 5% of V_p).
192 The results were similar to those presented above, i.e., the test-statistics were inflated for BOLT-
193 LMM and LR-All but not for fastGWA and LR-unRel (**Figure 1b**). The inflation in test-statistics
194 from BOLT-LMM-Mix decreased as the number of causal variants increased (**Figure 1b**). It is of
195 note that, in all the simulation scenarios where there were shared environmental effects, the test-
196 statistics at the null SNPs from BOLT-LMM-Mix appeared to be even more inflated than those
197 from LR-All (i.e., linear regression without correcting for relatedness).
198
199 To quantify the statistical power of each method, we used the mean χ^2 statistic at the causal SNPs.
200 Because the test-statistics from some of the methods were inflated at the null SNPs, we re-
201 calibrated the mean χ^2 statistic at the causal SNPs by dividing it by the genomic inflation factor at
202 the null SNPs described above to compare the power of different methods given the same level of
203 FPR (**Methods**), similar to the idea of computing the area under the power-FPR ROC curve. We
204 found that BOLT-LMM-Mix showed the highest power among all the methods (**Figure 2a**). When

205 the number of causal SNPs was relatively small, there was a relatively large gap in power between
206 BOLT-LMM-Mix and all other methods including BOLT-LMM-Inf (**Figure 2a**). BOLT-LMM-Inf
207 model showed the second highest power, in line with the theory that MLM leaving out the target
208 chromosome from the polygenic component gains power [39]. fastGWA showed a similar level of
209 power to LR-All, and LR-unRel showed the lowest power among all the methods owing to its
210 smaller sample size. We also observed that the power of all the methods were almost independent
211 of the variance explained by common environmental effects. Increasing the number of causal
212 SNPs led to smaller differences in power between methods (**Figure 2b**), suggesting that the
213 difference in power increased as the per-SNP variance explained increased (**Methods**).
214
215 **Application of fastGWA to 3,613 traits in the UKB**
216 We used fastGWA to conduct a genome-wide association analysis of 13,778,261 genotype or
217 imputed variants (minor allele frequency, MAF ≥ 0.0001) in all the UKB individuals of European
218 ancestry (n = 456,422) for 3,613 real phenotypes, and compared the results to those produced by
219 the Neale Lab using LR-unRel (361,194 unrelated individuals, see **URLs**). As noted above,
220 fastGWA is expected to be more powerful than LR-unRel because of the larger sample size. We
221 applied the same QC criteria and association analysis strategies as used in the Neale Lab's analysis
222 for consistency (**Methods**). We confirmed that the test statistics were highly correlated between
223 the two sets of results (mean correlation of z-statistics of 0.86 for 1,962 overlapping phenotypes).
224 We chose 24 representative traits (**Supplementary Table 1**) to compare our results with the
225 Neale Lab's results.
226
227 We first attempted to quantify the genomic inflation in the two sets of results. LD score regression
228 (LDSR) is an approach developed to partition the inflation in GWAS test-statistics into
229 components due to polygenic variation and population structure [53], but recent studies suggest
230 that LDSR intercept is a function of heritability, sample size and population genetic differentiation
231 in the sample, and is expected to deviate from 1 in a genetically stratified sample even if the
232 phenotype is not stratified [42,53,54]. We therefore sought to quantify the inflation due to sample
233 structure by the attenuation ratio, i.e., (LDSC intercept – 1) / (mean χ^2 – 1), which is independent
234 of sample size, as suggested in a recent study [42]. On average across the 24 traits, the attenuation
235 ratio of fastGWA was only 1.03-fold larger than that of LR-unRel (**Supplementary Table 2**),
236 consistent with what we observed in simulations (**Figure 1**) that the inflation due to relatedness
237 can be reasonably well corrected for by fastGWA.
238
239 We then compared the discovery power between the two sets by the clumping analysis in
240 PLINK1.9 [20] (P-value threshold = 5×10^{-9}, window size = 5 Mb, and LD r^2 threshold = 0.01). Of all

241 the 24 traits, the number of clumped genome-wide significant SNPs was 8,077 in our results,
242 substantially higher than that (5,799) in the Neale Lab's results (see **Table 3** for the comparison
243 of each trait), suggesting a nearly 40% of improvement in the number of GWAS discoveries for
244 fastGWA over LR-unRel. In addition, Canela-Xandri et al. [55] have also applied an MLM-LOCO [39]
245 association analysis to 778 UKB traits by DISSECT [41] using parallel computing in a supercomputer
246 and released all the GWAS summary data in a public database, GeneATLAS (**URLs**). We compared
247 the results from GeneATLAS to those from the Neale Lab and our fastGWA analysis for 10 traits
248 available in all the three sets (**Table 3** and **Supplementary Table 2**). There was no significant
249 difference in attenuation ratio between the three sets but GeneATLAS had more genome-wide
250 significant discoveries than the other two sets, likely because of the MLM-LOCO scheme used in
251 GeneATLAS [55], consistent with the simulation results from this study (**Figure 2**) and previous
252 studies [39,40] that MLM-LOCO gains power.
253
254 During the revision of this manuscript, whole-exome sequence (WES) data of 49,960 participants
255 became available in the UKB. We therefore applied fastGWA to the WES data (151,497 variants
256 with MAF ≥ 0.01 and 46,191 individuals of European descent) for 2,090 traits, following the same
257 analysis pipeline described above (**Methods**). We identified 158 near-independent associations
258 at an exome-wide significance level (P-value $< 0.05/m$ with m being the number of variants tested
259 for a trait) for the 24 traits described above (**Supplementary Table 3**). For each of the exome-
260 wide significant associations, we repeated the fastGWA analysis conditioning on the GWAS
261 signals (within 10Mb of the WES variant in either direction) identified from the imputed data of
262 the whole UKB sample described above. Conditioning on the imputed GWAS signals, only 7
263 associations remained exome-wide significant (**Supplementary Table 4**), suggesting that most
264 common variants in the WES data have been well tagged by SNP array-based genotyping and
265 imputation. Full summary statistics of 13,778,261 array-genotyped or imputed variants for 3,613
266 traits and 151,497 WES variants for 2,090 traits are publicly available at our data portal without
267 restricted access (**URLs**).
268
269 **DISCUSSION**
270 In this study, we developed a reliable, robust and resource-efficient association analysis tool,
271 fastGWA. This tool requires much smaller system resources (i.e., runtime and memory usage)
272 than existing tools, which makes it feasible to conduct GWA analyses of thousands of traits in
273 large cohorts like the UKB without the need to remove related individuals. The tool is also
274 applicable to family-structured data with a very large number of omic phenotypes.
275

Apart from computational efficiency, fastGWA also shows greater robustness than existing MLM-based methods in the presence of confounding factors. It has long been known the existence of relatedness in the data would lead to inflated association test statistics [23,24,56], confirmed by our simulation (**Figure 1a**). MLMs can be used to account for relatedness because the fixed effect is tested conditional on the phenotypic covariance structure among all individuals (**Equation 1**) [44]. It should be noted that the estimate of the "genetic variance" component based on close relatives (as in fastGWA) is a compound of σ_g^2 (the true genetic variance) and σ_C^2 (the amount of phenotypic variance attributable to shared environmental effect). More specifically, in our simulated data, the phenotypic covariance between two close relatives is $cov(y_i, y_j) = \pi_{ij}\sigma_g^2 + \sigma_C^2$ with π_{ij} being the family relatedness. In the fastGWA analysis, however, we do not seek to explicitly partition σ_g^2 and σ_C^2 but to use a single variance component to model the phenotypic covariance among close relatives so that the estimated phenotypic variance-covariance matrix ($\widehat{\boldsymbol{V}}$) of the single-component model is similar to that of the two-component model (**Supplementary Figure 8**). In the fastGWA analysis, the deviation of HE regression coefficient from σ_g^2 is a function of $\frac{\sigma_C^2}{\pi_{ij}}$ and the proportion of related pairs in the sample. In contrast, the estimate of the "genetic variance" component from an MLM analysis based on a dense GRM (SNP-derived GRM between all pairwise individuals, as in BOLT-LMM) is a weighted average of the SNP-based genetic variance (equals to σ_g^2 in our simulation but often smaller than σ_g^2 in reality because of imperfect tagging) and the pedigree-based genetic variance (σ_{ped}^2, higher than the σ_g^2 because of the confounding of σ_C^2), leading to an under-estimation of the covariance between closely related individuals, especially when the proportion of relative pairs is very small compared to the unrelated pairs. It was shown in our simulations that when $\sigma_C^2 = 0$, the estimate of the "genetic variance" component from either fastGWA or BOLT-LMM was unbiased (**Supplementary Figure 5**). As σ_C^2 increased, the estimate of the "genetic variance" component from fastGWA increased but the estimate from BOLT-LMM-Inf was almost unchanged (**Supplementary Figure 5**). This may explain the inflation of test-statistics in BOLT-LMM-Inf at the null SNPs observed in our simulations with non-zero σ_C^2 (Figure 1). These observations are in line with the simulation results from a recent study [57] and caution the use of BOLT-LMM for traits with a large component of σ_C^2 in samples with high degree of relatedness. To further demonstrate the issue above, we selected 24 real phenotypes from the UKB to estimate the genetic variance using different methods (the same 24 traits as used in the UKB real data analyses except for education attainment which was reconstructed following the method in Ref. [58]). Our result shows that the BOLT-REML [59] (the method used in BOLT-LMM to estimate the variance components) estimate of "genetic variance" is equivalent to the estimate of genetic variance corresponding to the full dense GRM (**Supplementary Figure 9a**), leaving the covariance between close relatives

311 due to common environmental effect (and/or rare variants) uncaptured (**Supplementary Figure**
312 **9b**). Two particular examples were educational attainment (EA) and birth weight (BW), which
313 have been shown in previous studies with strong common environmental effects (e.g., shared
314 maternal effect among sibs) [60-62]. The estimate of the "genetic variance" component for EA and
315 BW from fastGWA were much higher than those from BOLT-REML (**Supplementary Figure 9a**),
316 consistent with a substantial estimate of σ_C^2 in a two-component model (**Supplementary Figure**
317 **9b**).
318
319 The increased power of fastGWA compared to LR-unRel is mainly because more individuals are
320 included in the association test. Large population-based cohorts such as the UKB tend to
321 oversample relatives as participants when an assessment-centre based recruitment strategy is
322 implemented [18]. Taking the UKB cohort as an example, to generate a set of unrelated individuals,
323 107,864 out of 456,422 European participants need to be excluded given a relatedness threshold
324 of 0.05 in GCTA. Excluding these participants would significantly compromise the statistical
325 power, which can be avoided by implementing MLM-based methods such as fastGWA. In addition,
326 the higher power of BOLT-LMM compared to fastGWA or LR is mainly driven by its LOCO scheme.
327 Previous studies have showed that leaving the target chromosome out of the polygenic
328 component gains power because the effects of other SNPs are conditioned out from the model
329 and proximal contamination (i.e., the focal SNP being fitted twice in the model, once as a fixed
330 effect and again as a random effect) is avoided [32,39,40]. We did not observe any increase in power
331 when applying the LOCO scheme to fastGWA (**Supplementary Figure 10**) because fastGWA
332 estimates pedigree relatedness by a sparse GRM to model phenotypic covariance between close
333 relatives due to genetic and/or common environmental effects and the pedigree relatedness
334 estimated using all autosomes are similar to those using 21 chromosomes under the LOCO
335 scheme. It is of note that to save computational time, BOLT-LMM-Inf estimates the genetic
336 variance only once using all "model SNPs" and applies it to the association tests of all SNPs
337 without re-estimating the genetic variance when a chromosome is left out, assuming that genetic
338 variance attributable to a single chromosome is relatively small. This approximation could
339 potentially induce inflation [40] and might explain the small inflation for BOLT-LMM-Inf in our
340 simulation when $\sigma_C^2 = 0$ (**Figure 1**).
341
342 There are a few caveats of applying fastGWA in practice. First, if the pedigree information is
343 unavailable or incomplete (as is the case for UKB; shown in **Supplementary Figure 11** and
344 further discussed in **Supplementary Note 3**), it is necessary to compute the GRM from SNP data.
345 We have implemented in GCTA a very efficient tool to compute the SNP-based GRM along with a
346 function that can subdivide the GRM computation into a large number of components for

347 parallelized computing (**Supplementary Note 3**). These GRM components can be finally
348 assembled to a full GRM using a simple but efficient Linux/Unix command. Second, fastGWA uses
349 SNP-derived PCs to correct for the effect due to population stratification. PCs are often provided
350 as part of the QC package in the downloaded data [18]. If PCs are not available, we would
351 recommend efficient PCA tools such as fastPCA or FlashPCA [63,64]. Another more efficient approach
352 is to compute PCs in a subset of the sample, and project the PCs to the rest of the sample. This
353 approach has been implemented in GCTA (**URLs**). It is likely that PCs are also required for other
354 MLM-based methods including BOLT-LMM because although in theory MLM-based methods
355 accounts for population stratification by fitting all (or a subset of selected) SNPs as random effects
356 [29,30,39], MLM-based association analyses in large samples suggested that fitting PCs as covariates
357 improves robustness [42,43]. It is also noteworthy that calculating PCs from all the SNPs might be
358 suboptimal, as the test-statistics would tend to be deflated under the null (**Supplementary Note**
359 **4** and **Supplementary Figure 12**). It is therefore recommended to compute PCs from a set of LD-
360 pruned SNPs. Third, as discussed above, in the presence of common environmental effects, the
361 estimate of "genetic variance" component is a function of $\hat{\sigma}_g^2$ and $\hat{\sigma}_C^2$. We did not attempt to
362 differentiate common environmental components among different degrees of relatedness (e.g.,
363 siblings might share stronger common environmental effects than cousins). Nevertheless, our
364 simulation showed that this simple modelling did not lead to inflated test-statistics at the null
365 SNPs in the scenario where common environmental effects decreased as relatedness decreased
366 (**Figure 1**). Fourth, the fastGWA analysis involves the computation of an $n \times n$ sparse matrix.
367 GRM density, defined as the proportion of non-zero entries in a matrix, is essential for the
368 computational efficiency under certain conditions. We have demonstrated the runtime of
369 fastGWA with a wide range of GRM density levels (**Supplementary Figure 13**). In general, with
370 a GRM density less than 5×10^{-6} (equivalent to approximately 200,000 individuals with at least
371 one relative among 400,000 individuals), fastGWA shows consistently high computational
372 efficiency. For GWAS cohorts like the UKB, with nearly half of the sample (213,620) having at least
373 one close relatives, the number of unique related pairs (estimated pairwise genetic relatedness >
374 0.05) is 178,075, corresponding to GRM density of around 3.9×10^{-6} (note that the UKB cohort
375 has over-sampled relatives due to its recruitment strategy [18]). Last but not least, the fastGWA
376 program will switch to use LR for analysis (allowing for covariates) if the estimate of the genetic
377 variance component is not significant at a nominal significance level, cautioning the use of
378 fastGWA in a sample with a small number of related pairs.
379
380 Despite these caveats, fastGWA is an MLM-based association analysis method that is orders of
381 magnitude more resource-efficient and has more robust control over relatedness than existing
382 methods. The computational efficiency of fastGWA has been manifested by its successful

application to the genome-wide association analyses of 3,613 traits on 456,422 array-genotyped in the UKB. The summary statistics released from this study are useful resources for post-GWAS analyses (e.g., functional enrichment, genetic correlation, polygenic risk score, and causal inference) and phenome-wide association studies (PhWAS). Additionally, fastGWA can be modified for omic-data-based QTL (xQTL) analyses in biobank samples in the future.

METHODS

UK Biobank data

The UK Biobank (UKB) is a large cohort study consists of approximately half a million participants aged between 40 and 69 at recruitment, with extensive phenotypic records [18]. In this study, we selected 456,422 individuals of European ancestry from the UKB cohort for simulation and real data analyses. Genetic data were genotyped by two different arrays, the Applied Biosystems™ UK Biobank Axiom™ Array and the Applied Biosystems™ UK BiLEVE Axiom™ Array [18], of which 556,269 genotyped SNPs were used to generate genotypes for simulation and 13,778,261 SNPs imputed by UKB consortium (imputed Version 3) were used for real data analyses [18]. Genotyped/imputed SNPs data were filtered with standard quality control (QC) criteria in PLINK1.9 [20], e.g., MAF ≥ 0.01 in simulations or ≥ 0.0001 in real data analyses, Hardy-Weinberg Equilibrium test $P \geq 10^{-6}$, genotyping rate ≥ 0.95, and imputation info score ≥ 0.8 in real data analyses. In addition, the UKB released its first tranche of whole-exome sequence (WES) data of 49,960 participants in March 2019 [65]. The WES variants had been called and cleaned by two different pipelines, Regeneron's Seal Point Balinese (SPB) [65] and Functionally Equivalent (FE) [66]. We used the SPB data for analysis and excluded from the analysis variants with MAF < 0.01 (151,497 variants remained) and individuals with non-European ancestry (46,191 individuals remained).

Simulating genotypes

In order to test the performance of fastGWA in the presence of relatedness and substantial population stratification, we simulated a total of 100,000 artificial individuals from two different ancestry backgrounds, with a moderate proportion of relatives (10% of all samples) using a "mosaic-chromosome" scheme modified from [40]. We first selected all individuals with self-reported "British" and "Irish" ancestry from the UKB as the founders. We then filtered the samples based on their genetic ancestry inferred from SNP data to assure that the two groups were genetically distinct (**Supplementary Note 1** and **Supplementary Figure 1-2**). Next, we divided the genome into consecutive segments of 2,000 SNPs and generated unrelated individuals by selecting each segment from two of the founders chosen at random, and simulated related individuals by selecting the segments from a limited number of founders according to the

419 relatedness (**Supplementary Note 1**). Finally, we obtained 45,000 independent and 5,000
420 related "British individuals", and 45,000 independent and 5,000 related "Irish individuals".
421 Detailed description of the parameters and procedures have been described in the
422 **Supplementary Note 1**. We used GCTA to compute LD scores from the simulated genotype data
423 (**Supplementary Note 5**).
424
425 **Simulating phenotypes**
426 The phenotypes were simulated based on the following model
427 $$\boldsymbol{y} = \boldsymbol{g} + \boldsymbol{z}b_p + \boldsymbol{e_C} + \boldsymbol{e} \qquad [3]$$
428 where $\boldsymbol{g} = \sum_{i=1}^{m} \boldsymbol{x}_i b_i$, is the sum of the genetic effect of m causal variants with $\boldsymbol{x}_i$ a vector of SNP
429 genotypes and $b_i \sim N(0,1)$; $\boldsymbol{z}$ is a vector consisting of 0 (British) and 1 (Irish) to indicate ancestry
430 with b_p being the mean difference in phenotype between the two groups; $\boldsymbol{e_C}$ is a vector of shared
431 environmental effects with the individual(s) in each family assigned by the same value generated
432 from $N(0,1)$; and $\boldsymbol{e}$ is a vector of individual environmental effects (i.e., the residuals) with
433 $e_j \sim N(0,1)$. We considered different levels of relatedness in different simulation scenarios
434 including:
435 a) no common environmental effects, denoted by (0,0);
436 b) common environmental effects explaining 10% or 20% of the total phenotypic variance
437 (V_p) among the 1st degree relatives, denoted by (1st, 0.1V_p) and (1st, 0.2V_p), respectively;
438 c) common environmental effects explaining 10% or 20% of V_p among the 1st and 2nd degree
439 relatives, denoted by (≥2nd, 0.1V_p) and (≥2nd, 0.2V_p), respectively;
440 d) common environmental effects explaining 20% of V_p among the 1st degree relatives and
441 10% of V_p among the 2nd degree relatives, denoted by (≥2nd, Gradient).
442 Each simulation was repeated 100 times. Detailed description of the parameter settings can be
443 found in the **Supplementary Note 2.**
444
445 **Assessing false positive rate and statistical power**
446 Genome-wide association analyses were conducted on the simulated data with 6 different
447 methods. The simulated phenotypes were pre-adjusted by the top 10 PCs computed from a set of
448 LD-pruned SNPs using flashPCA2 [67] (**Supplementary Note 4** and **Supplementary Figure 14,**
449 **15)**. Since the number of SNPs was not large (556,269 SNPs after QC), we used all the SNPs to
450 compute the sparse GRM for fastGWA and included all the SNPs as the
451 "model SNPs" in the polygenic component for BOLT-LMM. After performing GWAS, we quantified
452 the false positive rate, genomic inflation, and statistical power of each association method under
453 each simulation setting.
454

455 **Real data analyses**

456 We used fastGWA to perform a genome-wide association analysis for 3,613 traits in the UKB. We
457 applied exactly the same QC criteria as used in the Neale Lab's UKB GWAS (**URLs**). In brief, only
458 the participants with imputed genotype data and labelled as European ancestry (see the UKB
459 Data-field 1001) were included in the association analyses (n = 456,422). Quantitative traits
460 with >20% participants having the same phenotypic value and categorical traits were
461 transformed into binary (TRUE/FALSE) or ordered-categorical variables, and the other
462 quantitative traits were converted to z-scores by rank-based inverse normal transformation
463 (INT). For association analyses, we fitted age, age^2, sex, age×sex, age^2×sex, and the top 20 PCs
464 provided by the UKB as covariates. The sex-specific traits were adjusted for age, age^2, and the top
465 20 PCs. Clumping analysis was performed in a subset of 24 traits (listed in **Supplementary Table**
466 **1**) using PLINK1.9 with stringent criteria (bi-allelic variants only, P-value threshold = 5×10^{-9},
467 window size = 5 Mb, and LD r^2 threshold = 0.01), and a random subset of 10,000 unrelated
468 individuals in the UKB was used as a LD reference panel.

469 We applied the same analysis pipeline used above to the WES data (151,497 variants with MAF ≥
470 0.01 and 46,191 participants) for all the available traits in the UKB (excluding traits with n < 5000
471 or case/control traits with sample prevalence < 1%). We performed clumping analysis (bi-allelic
472 variants only, P-value threshold = 0.05 / m with m being the number of variants tested for each
473 trait, window size = 5 Mb, and LD r^2 threshold = 0.01) of the fastGWA result for each of the 24
474 selected traits based on LD estimated from WES data of 42,974 unrelated individuals (estimated
475 pairwise genetic relatedness < 0.05).

476

477 **URLs**

478 GCTA-fastGWA: http://cnsgenomics.com/software/gcta/#fastGWA
479 GWAS summary statistics from fastGWA analysis of 3,613 UKB traits:
480 http://cnsgenomics.com/software/gcta/#DataResource
481 R-script to generate FAM: http://cnsgenomics.com/software/gcta/#fastGWA
482 Computing GRM in biobank data: http://cnsgenomics.com/software/gcta/#MakingaGRM
483 UK Biobank: http://www.ukbiobank.ac.uk
484 PLINK1.9: https://www.cog-genomics.org/plink2
485 BOLT-LMM: https://data.broadinstitute.org/alkesgroup/BOLT-LMM/
486 LD score regression: https://github.com/bulik/ldsc
487 FlashPCA2: https://github.com/gabraham/flashpca
488 UKB Phenotype Processing scripts: https://github.com/Nealelab/UK_Biobank_GWAS
489 GeneATLAS: http://geneatlas.roslin.ed.ac.uk

UKB GWAS Results from the Neale Lab: http://www.nealelab.is/uk-biobank

Eigen C++ library: http://eigen.tuxfamily.org/index.php?title=Main_Page

Data availability

See the URLs.

Code availability

See the URLs.

Acknowledgements

We thank Huanwei Wang and Julia Sidorenko for assistance in data preparation, Allan McRae for organising computing resource, and Po-Ru Loh for constructive comments on the manuscript. We also thank the Neale Lab for making all the data processing pipelines publicly available. This research was supported by the Australian Research Council (DP160101343, DP160101056 and FT180100186), the Australian National Health and Medical Research Council (1078037, 1078901, 1113400 and 1107258), and the Sylvia & Charles Viertel Charitable Foundation. This study makes use of data from the UK Biobank (project ID: 12514). A full list of acknowledgments of this data set can be found in **Supplementary Note 6**.

Author contributions

J.Y. conceived the study. J.Y., L.J. and Z.Z. designed the experiment. Z.Z. developed the software tool. L.J. and Z.Z. performed the simulations and data analyses under the assistance and guidance from J.Y., P.M.V., T.Q., N.R.W. and K.E.K.. P.M.V., N.R.W. and J.Y. contributed resources and funding. L.J. and J.Y. wrote the manuscript with the participation of all authors. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

References

1. Visscher, P.M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics* **101**, 5-22 (2017).
2. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research* **47**, D1005-D1012 (2018).
3. Klein, R.J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385-389 (2005).
4. DeWan, A. *et al.* HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science* **314**, 989-992 (2006).

5. Burton, P.R. *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678 (2007).
6. Frayling, T.M. *et al.* A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889-894 (2007).
7. Scott, L.J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *science* **316**, 1341-1345 (2007).
8. Sanna, S. *et al.* Common variants in the GDF5-UQCC region are associated with variation in human height. *Nature genetics* **40**, 198-203 (2008).
9. Unoki, H. *et al.* SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nature genetics* **40**, 1098-1102 (2008).
10. Yasuda, K. *et al.* Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. *Nature genetics* **40**, 1092-1097 (2008).
11. Hunter, D.J. *et al.* A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature genetics* **39**, 870-874 (2007).
12. Aulchenko, Y.S., Ripke, S., Isaacs, A. & Van Duijn, C.M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294-1296 (2007).
13. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics* **39**, 906-913 (2007).
14. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**, 559-575 (2007).
15. Purcell, S., Cherny, S.S. & Sham, P.C. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19**, 149-150 (2003).
16. Wood, A.R., et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* (2014).
17. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet* **381**, 1371-1379 (2013).
18. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
19. Nagai, A. *et al.* Overview of the BioBank Japan Project: study design and profile. *Journal of epidemiology* **27**, S2-S8 (2017).
20. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
21. Cardon, L.R. & Palmer, L.J. Population stratification and spurious allelic association. *The Lancet* **361**, 598-604 (2003).
22. Freedman, M.L. *et al.* Assessing the impact of population stratification on genetic association studies. *Nature genetics* **36**, 388-393 (2004).
23. Voight, B.F. & Pritchard, J.K. Confounding from cryptic relatedness in case-control association studies. *PLoS genetics* **1**, e32 (2005).
24. Astle, W. & Balding, D.J. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 451-471 (2009).
25. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).
26. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
27. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics* **38**, 203-208 (2006).
28. Aulchenko, Y.S., de Koning, D.J. & Haley, C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177**, 577-85 (2007).
29. Kang, H.M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709-1723 (2008).

582 30. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-
583 wide association studies. *Nature genetics* **42**, 348-354 (2010).
584 31. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies.
585 *Nature genetics* **42**, 355-360 (2010).
586 32. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nature*
587 *methods* **8**, 833 (2011).
588 33. Korte, A. *et al.* A mixed-model approach for genome-wide association studies of correlated
589 traits in structured populations. *Nature genetics* **44**, 1066-1071 (2012).
590 34. Listgarten, J. *et al.* Improved linear mixed models for genome-wide association studies. *Nat*
591 *Methods* **9**, 525-6 (2012).
592 35. Segura, V. *et al.* An efficient multi-locus mixed-model approach for genome-wide
593 association studies in structured populations. *Nature genetics* **44**, 825-830 (2012).
594 36. Svishcheva, G.R., Axenovich, T.I., Belonogova, N.M., van Duijn, C.M. & Aulchenko, Y.S. Rapid
595 variance components-based method for whole-genome association analysis. *Nature*
596 *genetics* **44**, 1166-1170 (2012).
597 37. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association
598 studies. *Nature genetics* **44**, 821-824 (2012).
599 38. Jakobsdottir, J. & McPeek, M.S. MASTOR: mixed-model association mapping of quantitative
600 traits in samples with related individuals. *Am J Hum Genet* **92**, 652-66 (2013).
601 39. Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M. & Price, A.L. Advantages and pitfalls in
602 the application of mixed-model association methods. *Nature genetics* **46**, 100-106 (2014).
603 40. Loh, P.-R., et al. Efficient Bayesian mixed-model analysis increases association power in
604 large cohorts. *Nat Genet* (2015).
605 41. Canela-Xandri, O., Law, A., Gray, A., Woolliams, J.A. & Tenesa, A. A new tool called DISSECT
606 for analysing large genomic data sets using a Big Data approach. *Nature communications* **6**,
607 10162 (2015).
608 42. Loh, P.R., Kichaev, G., Gazal, S., Schoech, A.P. & Price, A.L. Mixed-model association for
609 biobank-scale datasets. *Nat Genet* **50**, 906-908 (2018).
610 43. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in
611 large-scale genetic association studies. *Nature Genetics* **50**, 1335-1341 (2018).
612 44. Eu-Ahsunthornwattana, J. *et al.* Comparison of methods to account for relatedness in
613 genome-wide association studies with family-based data. *PLoS Genet* **10**, e1004445 (2014).
614 45. Zaitlen, N. *et al.* Using extended genealogy to estimate components of heritability for 23
615 quantitative and dichotomous traits. *PLoS Genet* **9**, e1003520 (2013).
616 46. Patterson, H.D. & Thompson, R. Recovery of inter-block information when block sizes are
617 unequal. *Biometrika* **58**, 545-554 (1971).
618 47. Haseman, J. & Elston, R. The investigation of linkage between a quantitative trait and a
619 marker locus. *Behavior genetics* **2**, 3-19 (1972).
620 48. Yang, J., Zeng, J., Goddard, M.E., Wray, N.R. & Visscher, P.M. Concepts, estimation and
621 interpretation of SNP-based heritability. *Nat Genet* **49**, 1304-1310 (2017).
622 49. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing
623 heritability for human height and body mass index. *Nat Genet* **47**, 1114-20 (2015).
624 50. Ge, T., Chen, C.-Y., Neale, B.M., Sabuncu, M.R. & Smoller, J.W. Phenome-wide heritability
625 analysis of the UK Biobank. *PLoS genetics* **13**, e1006711 (2017).
626 51. Band, G. & Marchini, J. BGEN: a binary file format for imputed genotype and haplotype data.
627 *BioRxiv*, 308296 (2018).
628 52. Devlin, B., Roeder, K. & Wasserman, L. Genomic control, a new approach to genetic-based
629 association studies. *Theoretical population biology* **60**, 155-166 (2001).
630 53. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity
631 in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
632 54. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass
633 index in approximately 700000 individuals of European ancestry. *Hum Mol Genet* **27**, 3641-
634 3649 (2018).

55. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nat Genet* **50**, 1593-1599 (2018).
56. Kennedy, B., Quinton, M. & Van Arendonk, J. Estimation of effects of single genes on quantitative traits. *Journal of Animal Science* **70**, 2000-2012 (1992).
57. Mefford, J.A. *et al.* Efficient estimation and applications of cross-validated genetic predictions. *bioRxiv* (2019).
58. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539 (2016).
59. Loh, P.R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat Genet* **47**, 1385-92 (2015).
60. Rowe, D.C., Vesterdal, W.J. & Rodgers, J.L. Herrnstein's syllogism: Genetic and shared environmental influences on IQ, education, and income. *Intelligence* **26**, 405-423 (1998).
61. Kong, A. *et al.* The nature of nurture: Effects of parental genotypes. *Science* **359**, 424-428 (2018).
62. Lunde, A., Melve, K.K., Gjessing, H.K., Skjærven, R. & Irgens, L.M. Genetic and environmental influences on birth weight, birth length, head circumference, and gestational age by use of population-based parent-offspring data. *American journal of epidemiology* **165**, 734-741 (2007).
63. Abraham, G. & Inouye, M. Fast principal component analysis of large-scale genome-wide data. *PloS one* **9**, e93766 (2014).
64. Galinsky, K.J. *et al.* Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *The American Journal of Human Genetics* **98**, 456-472 (2016).
65. Van Hout, C.V. *et al.* Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. *bioRxiv*, 572347 (2019).
66. Regier, A.A. *et al.* Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nature communications* **9**, 4038 (2018).
67. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776-2778 (2017).

666 **Table 1.** Comparison of runtime between fastGWA and BOLT-LMM

Sample Size	GCTA-fastGWA			BOLT-LMM v2.3		
	Data input and parameter estimation (h)	Association (h)	Total (h)	Data input and parameter estimation (h)	Association (h)	Total (h)
50,000	0.01	0.23	0.24	1.05	1.36	2.41
100,000	0.04	0.47	0.52	2.66	2.68	5.34
200,000	0.21	1.43	1.64	7.61	5.33	12.94
300,000	0.55	2.72	3.26	15.76	7.86	23.62
400,000	1.16	4.04	5.2	28.30	13.87	42.16

667 Shown are the runtimes of GCTA-fastGWA (assuming that the GRM is available) and BOLT-LMM
668 v2.3 for different sample sizes (i.e., 50k, 100k, 200k, 300k, and 400k) in the simulation (**Methods**).
669 The genotype data consisted of 18,138,215 SNPs (note that 565,631 LD-pruned SNPs were used
670 as "model SNPs" by the BOLT-LMM; see **Supplementary Note 3**). The runtime of both methods
671 can be divided into two steps: a) the estimation of parameters for mixed linear model and
672 Bayesian mixture model ("parameter estimation"), and b) the association test ("association"). All
673 tests were performed under the same computing environment: 96 GB of memory and 16 CPU
674 cores in 1 computer node.

675 **Table 2.** Comparison of memory usage between fastGWA and BOLT-LMM

Sample size	GCTA-fastGWA		BOLT-LMM v2.3		$\frac{Mem_{fastGWA}}{Mem_{BOLT-LMM}}$	$\frac{VMem_{fastGWA}}{VMem_{BOLT-LMM}}$
	Mem (GB)	VMem (GB)	Mem (GB)	VMem (GB)		
50,000	1.2	1.5	8.8	9.0	13.6%	16.7%
100,000	1.5	1.6	16.9	17.2	8.9%	9.3%
200,000	1.6	1.7	32.2	32.5	5.0%	5.2%
300,000	1.7	1.8	47.0	47.2	3.6%	3.8%
400,000	2.0	2.0	63.2	64.0	3.2%	3.1%

676 Shown are the actual memory / virtual memory usage of association analysis using GCTA-
677 fastGWA and BOLT-LMM v2.3 with different sample sizes (i.e., 50k, 100k, 200k, 300k, and 400k)
678 in the simulation (**Methods**). "Mem" represents the actual memory usage and "VMem" represents
679 virtual memory usage, both in gigabyte (GB) units. All tests were performed under the same
680 computing environment: 96 GB of memory and 16 CPU cores in 1 computer node.
681

682 **Table 3.** Number of near-independent genome-wide significant associations from the fastGWA
683 analysis of the imputed data for 24 quantitative traits in the UKB

Trait	No. signals (fastGWA)	*n* (fastGWA)	No. signals (Neale Lab)	*n* (Neale Lab)	No. signals (GeneATLAS)
WC	402	455,545	247	360,564	435
HC	469	455,495	332	360,521	517
HT	1,742	455,332	1,377	360,388	2,450
WT	631	455,010	454	360,116	747
BMI	537	454,841	368	359,983	599
HGSR	164	454,473	96	359,729	199
HGSL	144	454,417	86	359,704	172
MTCIM	52	453,043	36	358,695	38
BMR	816	448,348	614	354,825	936
BFP	453	448,114	307	354,628	520
DBP	256	430,029	176	340,162	\
SBP	287	430,025	217	340,159	\
FVC	463	415,931	333	329,404	\
FEV	393	415,931	272	329,404	\
PEF	137	415,931	90	329,404	\
NTS	64	369,407	46	293,006	\
EA	62	304,998	20	240,547	\
hBMD	483	262,294	377	206,589	\
BW	122	258,857	77	205,475	\
AMena	181	240,378	126	188,644	\
AFLB	20	168,097	10	131,987	\
PR	72	149,082	51	118,850	\
FIS	39	146,808	23	117,131	\
AMeno	88	141,926	64	111,593	\

684 Shown are the numbers of near-independent GWAS signals (clumping criteria: *P*-value threshold
685 = 5×10^{-9}, window size = 5 Mb, and LD r^2 threshold = 0.01) for 24 selected UKB traits identified by
686 three association methods, fastGWA, LR-unRel (results from the Neale Lab) and DISSECT-LOCO
687 (results from GeneATLAS). Phenotypes are ordered by descending sample size (*n*). The
688 abbreviated and full names of the traits are listed in **Supplementary Table 1**.

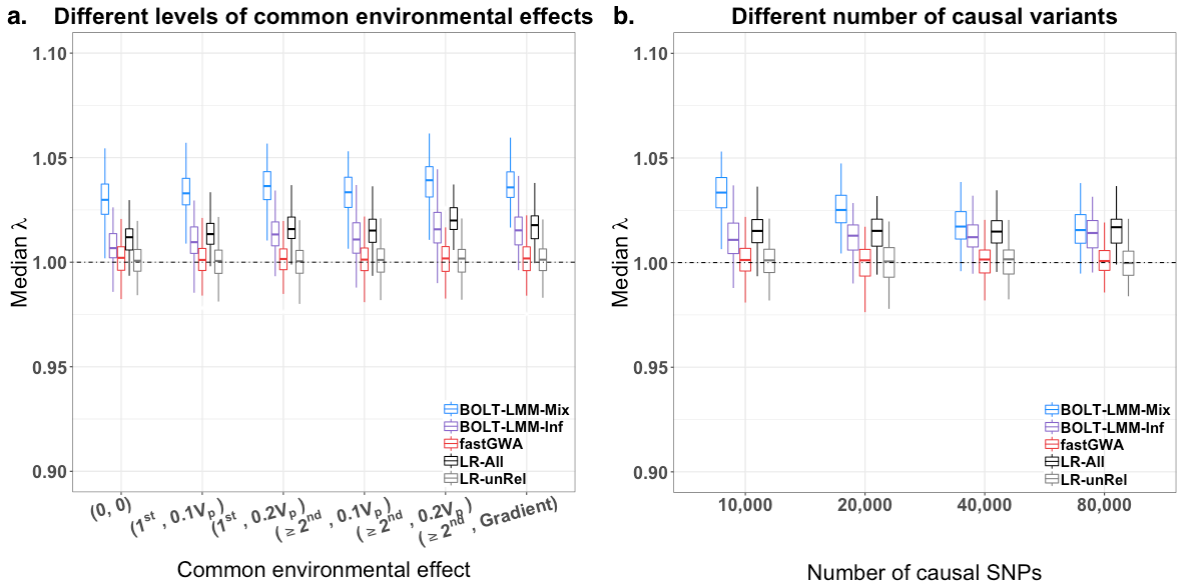

Figure 1. Median λ of null SNPs under different simulation scenarios. a) Median λ of null SNPs with different levels of common environmental effects in the simulations (**Methods**). The y axis represents the median λ of the null SNPs (i.e., all the SNPs on the even chromosomes), and the x axis represents different levels of common environment effects with (0, 0) representing no common environmental effects; (1st, $0.1V_p$) and (1st, $0.2V_p$) representing common environmental effects explaining 10% and 20% of the phenotypic variance (V_p) among the 1st degree relatives, respectively; ($\geq 2^{nd}$, $0.1V_p$) and ($\geq 2^{nd}$, $0.2V_p$) representing common environmental effects explaining 10% and 20% of V_p among the 1st and 2nd degree relatives (note that this actually includes all the relatives simulated), respectively; ($\geq 2^{nd}$, Gradient) representing 20% of V_p among the 1st degree relatives and 10% of V_p among the 2nd degree relatives. Each boxplot represents the distribution of median λ across 100 simulation replicates. b) Median λ of null SNPs with different number of causal variants (i.e., 10k, 40k, and 80k)) in the simulations (**Methods**). The y-axis represents the median λ of the null SNPs and the x-axis represents the different number of causal variants simulated.

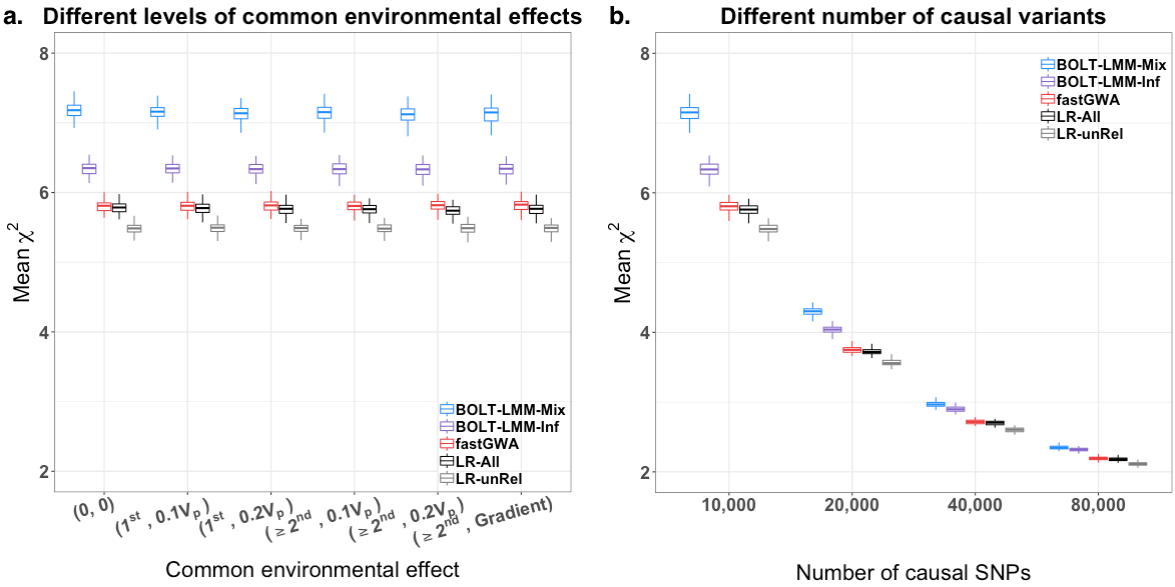

Figure 2. Mean χ^2 of causal SNPs under different simulation scenarios. a) Mean χ^2 of causal SNPs with different levels of common environmental effects in the simulations (**Methods**). The y-axis represents the mean χ^2 values for all the 10,000 causal variants on the odd chromosomes and the x-axis represents the different levels of common environmental effects as described in **Figure 1a**. The mean χ^2 has been adjusted by the median λ of null SNPs on the even chromosomes. Each boxplot represents the distribution of mean χ^2 across 100 simulation replicates. b) Mean χ^2 at causal SNPs with different number of causal variants in the simulations (**Methods**). The y-axis represents the average of mean χ^2 of the causal variants and the x-axis represents the different number of causal variants (i.e., 10k, 20k, 40k, 80k). The mean χ^2 has been adjusted by the median λ of null SNPs on the even chromosomes. Each boxplot represents the distribution of mean χ^2 across 100 simulation replicates.