nature genetics

TECHNICAL REPORT

https://doi.org/10.1038/s41588-021-00870-7

Check for updates

Computationally efficient whole-genome regression for quantitative and binary traits

Joelle Mbatchou, Leland Barnard, Joshua Backman, Anthony Marcketta, Jack A. Kosmicki, Andrey Ziyatdinov, Christian Benner, Colm O'Dushlaine, Mathew Barber, Boris Boutkov, Lukas Habegger, Manuel Ferreira, Aris Baras, Jeffrey Reid, Goncalo Abecasis, Evan Maxwell and Jonathan Marchini ✉

Genome-wide association analysis of cohorts with thousands of phenotypes is computationally expensive, particularly when accounting for sample relatedness or population structure. Here we present a novel machine-learning method called REGENIE for fitting a whole-genome regression model for quantitative and binary phenotypes that is substantially faster than alternatives in multi-trait analyses while maintaining statistical efficiency. The method naturally accommodates parallel analysis of multiple phenotypes and requires only local segments of the genotype matrix to be loaded in memory, in contrast to existing alternatives, which must load genome-wide matrices into memory. This results in substantial savings in compute time and memory usage. We introduce a fast, approximate Firth logistic regression test for unbalanced case–control phenotypes. The method is ideally suited to take advantage of distributed computing frameworks. We demonstrate the accuracy and computational benefits of this approach using the UK Biobank dataset with up to 407,746 individuals.

Since the first large genome-wide association studies[1] were carried out in 2007, there has been a steady increase in sample size—now reaching hundreds of thousands of individuals—which is enabled by a parallel stream of methods with ever-increasing computational efficiency. Initial methods used simple linear or logistic regression using programs such as SNPTEST[1] and PLINK[2], but these have largely been replaced by the use of linear mixed models (LMMs) and the closely related whole-genome regression models. These approaches have been shown to account for population structure and relatedness, and offer advantages in power by conditioning on associated markers from across the whole genome[3–7].

The initial methods were focused on quantitative traits[3] for studies with a few thousand samples and assumed a Gaussian distribution on SNP effect sizes. These approaches were extended to datasets including tens of thousands of individuals by computational strategies that avoided repeated matrix inversions when testing each SNP[7,8]. Building on work from the plant and animal breeding literature[9,10], even more efficient whole-genome regression approaches were developed that allowed for more flexible (non-Gaussian) prior distributions of SNP effect sizes[11,12]. BOLT-LMM and LEMMA are implementations of this approach[13–15]. The fastGWA LMM approach reduces the computational time by using a sparse representation of the genetic correlations present in the sample[16]. For simple linear regression of quantitative traits, the BGENIE method (https://jmarchini.org/BGENIE/) introduced the idea of the simultaneous analysis of multiple quantitative traits, which required only a single pass through the genetic data and provided substantial speed-ups over PLINK[17].

BOLT-LMM and fastGWA have also been applied to binary (case–control) traits when the case–control ratio is reasonably balanced and relatively common variants are tested for association. However, these approaches break down when applied to unbalanced case–control studies tested with rarer variants, such as those found in exome sequencing studies. The SAIGE method implements a logistic mixed-model approach and a saddle-point approximation (SPA) to the null distribution of the test statistic, which is effective at controlling Type 1 errors[18].

The BOLT-LMM, fastGWA and SAIGE methods all proceed in two main steps that are applied one trait at a time. In Step 1, a model is fit to a set of SNPs from across the whole genome, such as all of the SNPs on a genotyping array. The resulting model fit is then used to create either a prediction of individual trait values based on the genetic data (in BOLT-LMM and SAIGE) or an estimate of the trait variance–covariance matrix (in fastGWA). In Step 2, a larger set of imputed or sequenced variants on the same set of samples are tested for association, conditional on the predictions or variance–covariance matrix in Step 1. This is usually carried out using the so-called leave-one-chromosome-out (LOCO) scheme, where each imputed SNP on a chromosome is tested conditional on the Step 1 predictions ignoring that chromosome. This approach avoids proximal contamination, which can reduce the power of association tests[8,19].

In this paper we propose a new machine-learning method within this two-step paradigm, called REGENIE (https://rgcgithub.github.io/regenie/), that is substantially faster than existing approaches. Extended Data Fig. 1 provides an overview of the REGENIE method. In Step 1, array SNPs are partitioned into consecutive blocks of B SNPs and a small set of J ridge regression predictions are generated from each block (this is referred to as Level 0). Within each block, the ridge regression predictors each use a slightly different set of shrinkage parameters. The idea of using a range of shrinkage values is to capture the unknown number and size of truly associated genetic markers in each window. This approach is equivalent to placing a Gaussian prior on the effect sizes of the SNPs in the block and finding the maximum a posteriori estimate of the effect sizes and the resulting prediction. One can think of these predictions as local polygenic scores that account for local linkage disequilibrium (LD) within blocks. Combining the predictions from across

Regeneron Genetics Center, Tarrytown, NY, USA. ✉e-mail: jonathan.marchini@regeneron.com

Table 1 | Computational performance of REGENIE, fastGWA and BOLT-LMM when analyzing 50 quantitative traits with UK Biobank data

Method	Step	Benchmark		
		CPU time (h)	Elapsed time (h)	Memory usage (GB)
REGENIE	1	111	12	12.9
REGENIE–LOOCV	1	192	19	23.6
REGENIE	2	2,916	197	6.0
fastGWA	0	454	201	–
fastGWA	1–2	9,191	624	2.0
BOLT-LMM	1	60,735	1,815	49.6
BOLT-LMM	2		2,271	

For REGENIE and BOLT-LMM, 469,336 LD-pruned SNPs were used as model SNPs when fitting the null model (Step 1). For fastGWA, these SNPs were used to compute the sparse genetic-relatedness matrix with a default relatedness threshold of 0.05 (Step 0). Tests were performed on 11.4 million imputed variants and the timings were projected to 30 million variants (Step 2). For Step 1, REGENIE was run in multi-trait mode analyzing all traits together at once using fivefold CV (REGENIE) as well as LOOCV (REGENIE–LOOCV). All of the runs were performed on the same computing environment (16 virtual CPU cores of a 2.1-GHz AMD EPYC 7571 processor, 64 GB of memory and a 600-GB solid-state disk), except for the genetic-relatedness-matrix calculation required for fastGWA, where we used 250 partitions in a computing environment with four virtual CPU cores and 8 GB of memory. The BGEN file input needed for Step 2 was split by chromosome, so fastGWA had to be run separately for each chromosome being tested. The sample sizes for the 50 traits ranged from 332,739 to 407,662 individuals (see Supplementary Table 2).

Table 2 | Computational performance of REGENIE–Firth, REGENIE–SPA and SAIGE when analyzing 50 binary traits with UK Biobank data

Method	Step	Benchmark		
		CPU time (h)	Elapsed time (h)	Memory usage (GB)
REGENIE	1	1,590	117	11.8
REGENIE–LOOCV	1	777	108	19.5
REGENIE–Firth	2	115,492	8,237	7.7
REGENIE–SPA	2	79,363	5,090	9.1
SAIGE	1	275,070	21,428	48.7
SAIGE	2	239,865	173,992	2.1

When fitting the null model (Step 1), 469,336 LD-pruned SNPs were used as model SNPs. Tests were performed on 11.8 million imputed SNPs and the timings were projected to 30 million variants (Step 2). For Step 1, REGENIE was run in multi-trait mode analyzing all traits together at once using fivefold CV (REGENIE) as well as LOOCV (REGENIE–LOOCV). For Step 2, REGENIE was run using Firth correction (REGENIE–Firth) and SPA (REGENIE–SPA). Step 1 of SAIGE did not finish for 2/50 traits as it exceeded the four-week limit; the reported timings of SAIGE Step 2 are thus projections based on the timings of the completed runs. All of the runs were performed on the same computing environment (16 virtual CPU cores of a 2.1-GHz AMD EPYC 7571 processor, 64 GB of memory and a 600-GB solid-state disk). The sample sizes for the 50 traits ranged from 381,591 to 407,746 individuals (see Supplementary Table 6).

the genome results in a large reduction in the size of the genetic dataset. In this paper we use $B=1{,}000$ and $J=5$, and this reduces a set of $M=500{,}000$ SNPs to $M=2{,}500$ predictors. The method then uses a second ridge regression (referred to as Level 1) to combine the M predictors into a single predictor, which is then decomposed into 23 chromosome predictions for a LOCO approach. Linear or logistic regression is used at Level 1, depending on the phenotype. The resulting LOCO predictions are then used as a covariate in Step 2 when each imputed SNP is tested. This approach completely decouples Step 1 and Step 2 so that the Step 1 predictions can be re-used when running Step 2 on distinct sets of markers (for example, imputed and exome markers) or even when distinct statistical tests are needed at Step 2. All of the predictions at Level 0 and Level 1 are obtained within a cross-validation (CV) scheme (either K-fold CV or leave-one-out CV (LOOCV)) to prevent over-fitting.

This approach exhibits a number of desirable properties. First, many of the calculations in Steps 1 and 2 can be carried out for multiple traits in parallel. This leads to substantial gains in speed as the files containing the variants in Steps 1 and 2 are read only once, rather than repeatedly for each trait. In practice, we find that for Step 1, REGENIE can be over 150× faster than BOLT-LMM and 300× faster than SAIGE when analyzing 50 UK Biobank quantitative and binary traits with up to 407,746 samples (Tables 1 and 2). In Step 2, each variant is tested for association and the overall computational burden will depend on the number of variants tested. For example, analyzing imputed variants across the genome will result in a higher computational burden for Step 2 relative to Step 1, but this will be less so when Step 2 involves testing coding variants from exome sequencing. The computational differences between methods in Step 2 are less extreme and mostly depend on the type of trait, test statistic and implementations of file format reading and parallelization schemes. However, REGENIE analyzes multiple traits in parallel and this can result in substantial computational savings, especially for quantitative traits. On an imputed dataset with 30 million tested variants and 50 traits, we find that over both Steps 1 and 2, REGENIE is 19.5× and 4.4× faster than BOLT-LMM and SAIGE, respectively. In the Supplementary Note and Supplementary Table 1 we provide an analysis of the computational complexity of REGENIE.

Second, in Step 1 of REGENIE, only B SNPs need to be stored in memory at once, which leads to a low memory footprint, which can reduce the costs on cloud-based platforms. Third, the method is applicable to both quantitative and binary traits and we have implemented a new, fast Firth logistic regression test as well as a SPA test for binary traits. Finally, our algorithm is ideally suited to implementation on distributed computing frameworks, such as Apache Spark, where both the dataset and application of the method and computation can be parallelized across a large number of machines. The main implementation of REGENIE is a standalone C++ program (https://rgcgithub.github.io/regenie/) but these methods have also been implemented for quantitative traits in the Apache Spark-based Glow project (http://projectglow.io; see Supplementary Note). All of the main experiments and results in the paper were obtained using the C++ program.

Results

Quantitative traits. Figure 1 shows the results of applying REGENIE, BOLT-LMM and fastGWA to three quantitative phenotypes measured on white British participants of the UK Biobank (low-density lipoprotein cholesterol, $n=389{,}189$; body mass index, $n=407{,}609$; and bilirubin, $n=388{,}303$), where Step 2 testing was performed on 9.8 million imputed SNPs (see Supplementary Note). The Manhattan plots for all three phenotypes show good agreement between the methods (see also Extended Data Fig. 2) with both REGENIE and BOLT-LMM showing increased power gains relative to fastGWA at known peaks of association.

To demonstrate the advantages of analyzing multiple traits in parallel using REGENIE, we compared it to BOLT-LMM and fastGWA on a set of 50 quantitative traits from the UK Biobank, each with a distinct missing data pattern (Supplementary Table 2). Whereas REGENIE can analyze all traits at once within a single run of the software, the BOLT-LMM and fastGWA software must be run once for each of the 50 traits. Across all 50 traits, we found that the P values for REGENIE and BOLT-LMM were in very close

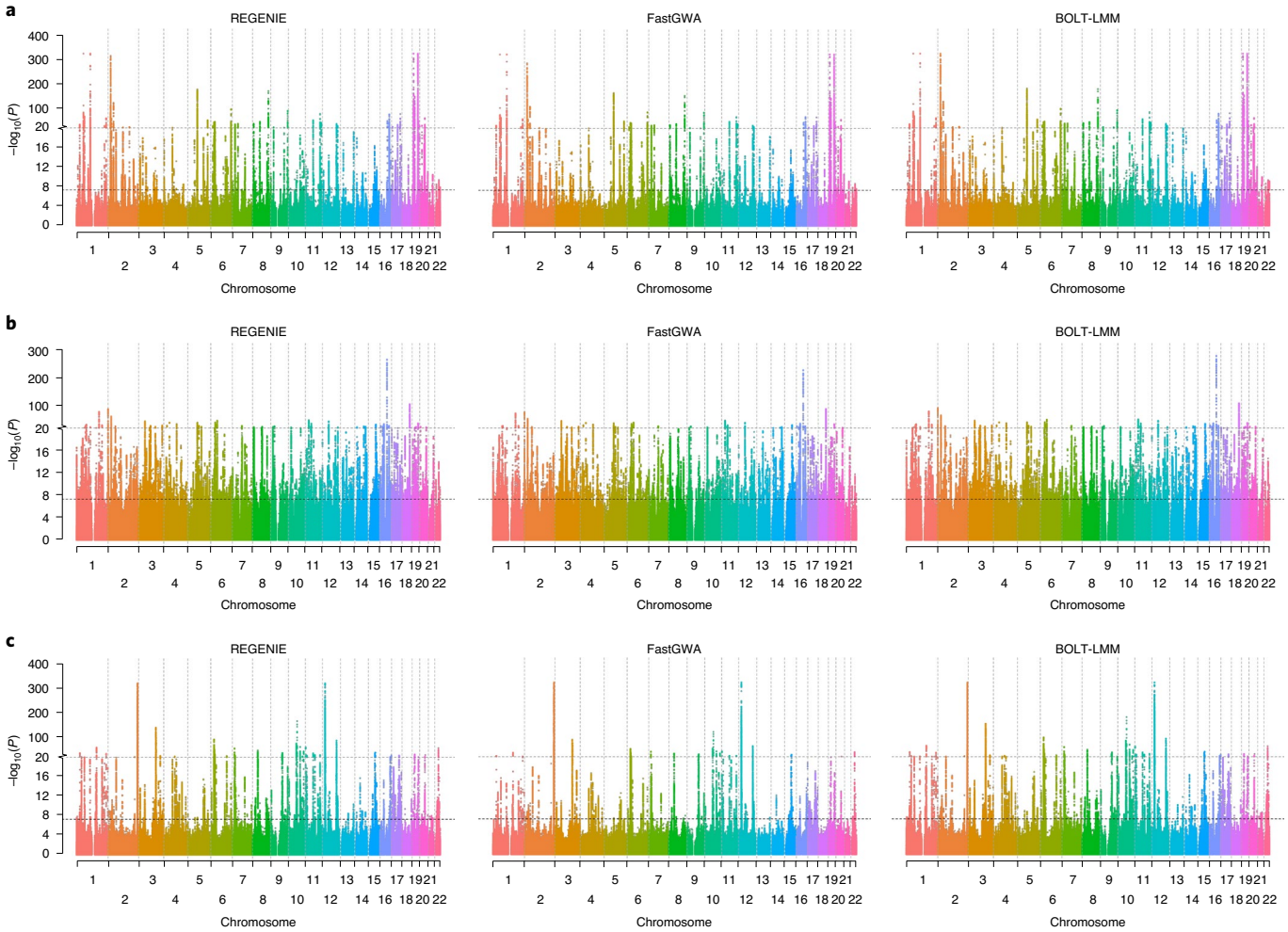


Fig. 1 | Comparison of methods on three quantitative traits from the UK Biobank. a–c, Results from REGENIE, fastGWA and BOLT-LMM for low-density lipoprotein cholesterol (**a**; $n = 389{,}189$), body mass index (**b**; $n = 407{,}609$) and bilirubin (**c**; $n = 388{,}303$) of samples from white British individuals. Tests were performed on 9.8 million imputed SNPs with a minor allele frequency greater than 1%. The bottom dashed horizontal line represents the genome-wide significance ($P = 5 \times 10^{-8}$) and the top dashed horizontal line represents the breakpoint for the different scaling of the *y* axis. The dashed vertical lines separate the 22 chromosomes.

agreement on the majority of traits tested, with some evidence that REGENIE is slightly less powerful for a few traits (Supplementary Fig. 1), whereas the fastGWA *P* values were noticeably deflated compared with REGENIE and BOLT-LMM. The compute time and memory usage of the three methods is given in Table 1. The table shows that in this 50-trait scenario, REGENIE is 151× faster than BOLT-LMM in elapsed time for Step 1 and 11.5× faster for Step 2, and this translates into an overall speed-up in terms of elapsed time of approximately 20× when projecting to 30 million tested variants obtained using imputation information score > 0.8. Similar to BOLT-LMM, Step 2 of REGENIE has been optimized for input genotype data in BGEN v1.2 format, which highly helped reduce the runtime. In addition, REGENIE has a maximum memory usage of 12.9 GB, which is mostly due to REGENIE reading onlya small portion of the genotype data at a time, whereas BOLT-LMM required 50 GB. To keep memory usage low when analyzing the 50 traits, within-block predictions are stored on disk and read separately for each trait working across blocks. The added input/output operations incur a small cost on the overall runtime but substantially decrease the amount of memory needed by REGENIE (Supplementary Table 3). When running analyses on cloud-based services such as Amazon Web Services, these time and memory reductions both contribute to large reductions in cost as cheaper Amazon Web Services instance types can be used and for less time. In the same 50-traits scenario, we find that REGENIE is about 3× faster than fastGWA but fastGWA is very memory efficient and uses a maximum of only 2 GB.

Binary traits. In addition to analyzing quantitative traits, REGENIE was also designed for the analysis of binary traits, including those with unbalanced case–control ratios. REGENIE includes implementations of both Firth and SPA corrections to handle this scenario (see Methods). Figure 2 (see also Extended Data Fig. 3) shows the results of applying REGENIE, BOLT-LMM and SAIGE to four binary phenotypes measured on white British participants of the UK Biobank (coronary artery disease, $N = 352{,}063$; glaucoma, $N = 406{,}927$; colorectal cancer, $N = 407{,}746$; and thyroid cancer, $N = 407{,}746$) where Step 2 testing was performed on 11.6 million imputed SNPs (Supplementary Note). All four approaches demonstrated very good agreement for the most balanced trait (coronary artery disease; case–control ratio = 1:11), but as the fraction of cases decreased, BOLT-LMM tended to give inflated test statistics. However, both REGENIE with Firth and SPA corrections as well as SAIGE are robust to this inflation and show similar agreement for the associations detected.

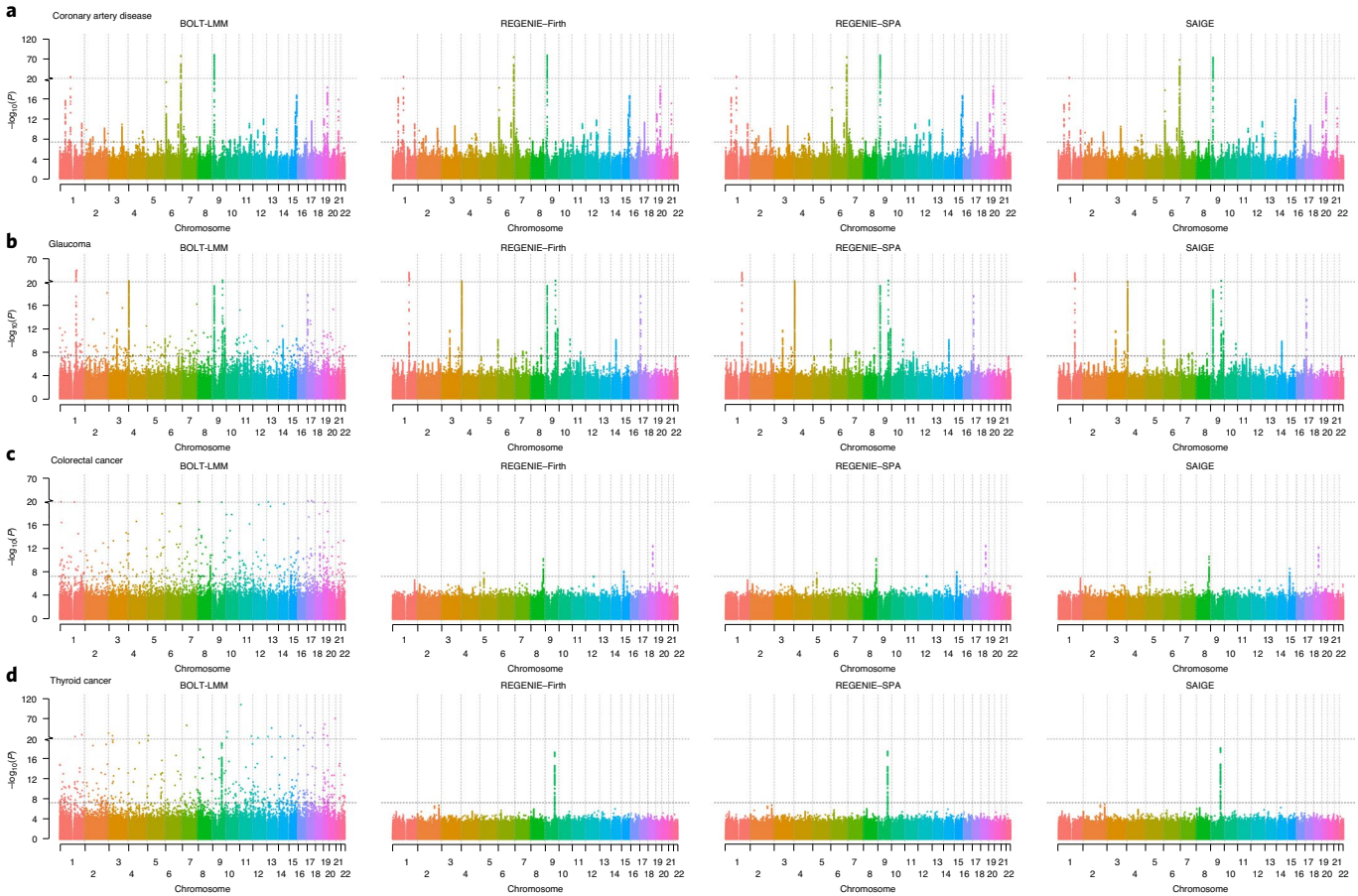


Fig. 2 | Comparison of methods on four binary traits from the UK Biobank. a–**d**, Results from REGENIE using Firth and SPA correction, BOLT-LMM and SAIGE on samples from white British individuals for coronary artery disease (**a**; case–control ratio = 1:11, $N = 352{,}063$), glaucoma (**b**; case–control ratio = 1:52, $N = 406{,}927$), colorectal cancer (**c**; case–control ratio = 1:97, $N = 407{,}746$) and thyroid cancer (**d**; case–control ratio = 1:660, $N = 407{,}746$). Tests were performed on 11 million imputed SNPs. The bottom dashed horizontal line represents the genome-wide significance ($P = 5 \times 10^{-8}$) and the top dashed horizontal line represents the breakpoint for the different scaling of the *y* axis. The dashed vertical lines separate the 22 chromosomes.

The SPA approach calculates a standard score test statistic and approximates the null distribution, whereas the Firth correction uses a penalized likelihood approach to estimate the SNP-effect-size parameters in an asymptotic likelihood-ratio test. Although both provide good control of Type 1 error for rare binary traits, we found that the SPA approach implemented in SAIGE can result in very inflated effect-size estimates (Supplementary Fig. 2). However, the Firth correction used in REGENIE provides reasonable effect-size estimates and standard errors when the minor allele count is low (Supplementary Table 4). The fast Firth correction that we developed agrees well with the exact Firth correction (Supplementary Figs. 3 and 4) but is approximately 60 times faster (Supplementary Table 5).

To assess the computational resources needed to analyze a larger number of traits, we again ran REGENIE using Firth/SPA correction and SAIGE on a set of 50 binary traits from the UK Biobank with a range of different case–control ratios and distinct missing data patterns (Supplementary Table 6). The compute time and memory usage details are given in Table 2.

For Step 1, we found that REGENIE (using the LOOCV scheme) was about 350 times faster (CPU time of 777 versus 275,070 h) and required only 40% of the memory used by SAIGE (19.5 versus 49 GB). In Step 2, REGENIE–Firth and REGENIE–SPA were 2× and 3× faster than SAIGE in CPU time, respectively, but 21× and 34× faster than SAIGE in elapsed time, respectively, which suggests that REGENIE makes better use of parallelization in this step. Overall, in this 50-trait setting, REGENIE–Firth was 4.4× faster than SAIGE in terms of CPU time and 23× faster in elapsed time when projected to 30 million tested variants obtained using INFO > 0.8%. REGENIE reduces the CO_2 footprint by more than 85% compared with SAIGE (Supplementary Table 7). Supplementary Figs. 5–10 compare the accuracy of REGENIE and SAIGE across all 50 traits and show good agreement.

A large portion of the compute time in SAIGE is used to implement the LOCO, but it has been suggested that for binary traits, the effect of proximal contamination is not as substantial for less prevalent traits[18]. We ran SAIGE without LOCO on the same 50 binary traits and observed that the impact of using LOCO was indeed more apparent with low case–control imbalance, where it can be highly beneficial (Extended Data Fig. 4 and Supplementary Figs. 11–13). These results would caution against the perfunctory use of SAIGE without LOCO for the analysis of all traits in a study such as the UK Biobank.

CV scheme. We implemented both a *K*-fold CV and a LOOCV scheme in Step 1 for both quantitative and binary traits (see Methods). Both approaches provided almost identical accuracy (Supplementary Figs. 14 and 15). For the dataset of 50 quantitative traits, LOOCV required 192 h of CPU time and 23.6 GB of memory, whereas the *K*-fold CV required 111 h of CPU time and 12.9 GB of

memory (Table 1). For the dataset of 50 binary traits, the LOOCV approach required approximately 50% of the CPU time used by the *K*-fold CV approach (CPU time of 777 versus 1,590 h) and 65% more memory, but the elapsed time of the two methods was similar (108 versus 117 h). The LOOCV approach requires fewer relatively expensive logistic regression calls compared with the *K*-fold CV, but the extra calls needed are easily parallelized across multiple cores.

Missing phenotype data. When analyzing multiple traits together with different missing data patterns, we use mean imputation of missing phenotype values in Step 1 but keep only samples with non-missing phenotypes in Step 2. This approach gave almost identical results to an exact approach that uses only samples with non-missing phenotypes in both Step 1 and Step 2 (Supplementary Figs. 16–18).

Simulation studies. Through simulations, we investigated the Type 1 error and power of the tests in REGENIE, BOLT-LMM, fastGWA, SAIGE and simple linear/logistic regression with the top principal component as a covariate (PCA). We also assessed the LOCO scheme used in REGENIE and the accuracy of the effect-size estimates from REGENIE–Firth and SAIGE. We used real genetic array data from the UK Biobank to simulate realistic genetic LD patterns and population structure. We sampled 100,000 individuals from either the white British or the full European ancestry set of the UK Biobank and selected one of the following: (1) only unrelated individuals, (2) individuals at random or (3) half of the samples from the related individuals and the remaining half from the unrelated individuals. To consider scenarios of more extreme relatedness, we sampled only first-degree relatives ($N = 22{,}990$), first- and second-degree relatives ($N = 30{,}775$) or first- to third-degree relatives ($N = 70{,}684$) in the set of white British participants. More details can be found in the 'Data simulation' section of Methods.

We used genomic inflation (λ_{GC}) and empirical Type 1-error rate (defined as the proportion of null tests with a *P* value less than a nominal level α) to assess the calibration of the tests across 100 simulation replicates. For the quantitative traits, the PCA method was well calibrated when the sample consisted only of unrelated individuals but became inflated with increasing levels of relatedness (Extended Data Fig. 5 and Supplementary Tables 8,9). However, REGENIE, BOLT-LMM with a mixture of Gaussian's model (BOLT-LMM-MoG), the BOLT-LMM infinitesimal model (BOLT-LMM-Inf) and fastGWA retained good Type 1-error control in all of the settings considered. REGENIE had slightly deflated type I error rates when half of the samples were related. This was also observed in more extreme relatedness scenarios, where the use of the Step 1 predictions in REGENIE led to good calibration of the test unlike the PCA method, which was inflated (Supplementary Table 10).

For binary traits with low case–control imbalance, REGENIE–Firth, REGENIE–SPA, BOLT-LMM-MoG, BOLT-LMM-Inf and fastGWA had good control of the Type 1 error with the more common variants tested (Extended Data Fig. 6, Supplementary Fig. 19d and Supplementary Tables 11,12), and SAIGE had slightly deflated Type 1 error rates. BOLT-LMM-MoG, BOLT-LMM-Inf and fastGWA were inflated for more unbalanced traits and this was worse for rarer variants (Supplementary Fig. 19). However, REGENIE–Firth, REGENIE–SPA and SAIGE were robust against this inflation and for extremely unbalanced traits; REGENIE–SPA was conservative with rarer variants. In more extreme relatedness scenarios, the PCA method became inflated with higher heritability levels but REGENIE–Firth and REGENIE–SPA retained good control of the Type 1 error, although they became conservative for more heritable traits (Supplementary Table 13).

To quantify power, we used the mean χ^2 test statistic at causal SNPs. For quantitative traits, REGENIE and BOLT-LMM-Inf had similar power performance, which was higher than for fastGWA across all settings (Supplementary Fig. 20 and Supplementary Table 14). BOLT-LMM-MoG had the highest power performance with fewer causal SNPs, and the power difference with REGENIE decreased as the number of causal SNPs increased. BOLT-LMM-MoG uses a more flexible mixture of Gaussian's prior, which may model traits with highly non-infinitesimal genetic architectures better. For binary traits, we compared REGENIE–Firth, REGENIE–SPA and SAIGE, which were all well calibrated. With low case–control imbalance, REGENIE–Firth and REGENIE–SPA had slightly higher power than SAIGE and for more unbalanced traits, REGENIE–Firth and SAIGE had similar power performance and REGENIE–SPA had slightly lower performance at rarer variants (Supplementary Fig. 21 and Supplementary Table 15).

We further investigated the accuracy of the effect sizes from REGENIE–Firth and SAIGE. They were similar for moderately unbalanced traits but as the case–control imbalance increased, the estimates from SAIGE became highly inflated (Extended Data Fig. 7). Finally, when comparing the approximate LOCO scheme in REGENIE to an exact LOCO scheme, we found that they gave similar results, both for the genetic predictions in Step 1 (squared sample correlation coefficient (R^2) = 0.99932 for quantitative traits and $R^2 = 0.98710$ for binary traits) and the *P* values in Step 2 (Supplementary Fig. 22).

We assessed the single-trait performance of REGENIE, BOLT-LMM and SAIGE and found that REGENIE takes approximately 3× less CPU time than BOLT-LMM for quantitative traits and >8× less CPU time than SAIGE for binary traits (Supplementary Table 16). With five traits, the computational efficiency of REGENIE improved—it took approximately 10× less CPU time than BOLT-LMM and >22× less CPU time than SAIGE. More generally, we observed that Step 1 of REGENIE scales sub-linearly with the number of traits and the scaling gets closer to linear when the number of traits become large (Extended Data Fig. 8).

Inter-chromosomal LD in the UK Biobank. While developing REGENIE, we and others[20] identified an anomaly in the UK Biobank array genotypes that leads to reduced performance of some of the LMMs being tested. We observed a sizeable number of SNP pairs that exhibited inter-chromosome LD, which breaks the assumptions of the LOCO scheme and can result in loss of power when any one of the SNPs in a pair is associated with a trait (see Supplementary Note).

Discussion

In this study we present a machine-learning method that implements simultaneous whole-genome-wide regression of multiple quantitative or binary traits. The method uses a strategy that splits computation into blocks of consecutive SNPs and does not require loading of a genome-wide set of SNPs into memory. This approach also facilitates the analysis of multiple traits in parallel. Overall, this results in substantial computational savings in terms of both CPU time and memory usage compared with existing methods such as BOLT-LMM, fastGWA and SAIGE. As the number of large-scale cohorts with deep phenotyping grows, this approach will probably become even more relevant. The parallel nature of the approach is ideally suited to distributed environments such as Apache Spark. We have developed a first version of REGENIE for quantitative traits within the Glow project as well as the full version of the method for quantitative and binary traits in a standalone C++ program with source code that is openly available.

Analysis of large cohorts for which phenotypes are derived from electronic health records often results in many binary traits with substantial case–control imbalance. REGENIE is applicable to binary traits and we have proposed an approximate Firth regression approach, which we show is almost identical to an exact Firth

regression implementation, and much faster. This approach has the added benefit that it avoids the parameter estimate inflation that occurs when SAIGE is used to analyze ultra-rare variants.

Like many existing mixed-model-based approaches, REGENIE is well able to handle relatedness in the sample, although it can become conservative in more extreme cases and hence, we recommend that it not be used for smaller cohorts with high levels of relatedness—such as founder populations, where exact mixed-model methods can be used. As previous methods have proposed to address this issue[13,18,21], we plan to explore extending REGENIE to compute and incorporate a calibration factor in its association testing step.

The approach used in REGENIE is inspired by, but not the same as, the machine-learning approach of stacked regressions[22]. REGENIE uses ridge regression to combine a set of correlated predictors, whereas Breiman's stacking approach used non-negative least squares to combine a set of highly correlated predictions in an ensemble learning approach. We have not yet investigated whether non-negative least squares might have advantages here. Furthermore, our simulations with traits that have a sparser genetic architecture also highlight the potential improvement of the REGENIE method by using more flexible priors on the effect size of predictors, as is done in BOLT-LMM with a mixture of Gaussian's prior.

There are many other potential avenues for development of this approach. It will be easy to expand the functionality to include tests such as SNP × covariate interactions[16], variance tests[23] and a whole range of gene-based tests[24–26]. Multivariate probit regression for binary traits[27], multivariate linear regression for quantitative traits[28] and multi-trait burden tests[29] will all be straightforward to implement.

We also plan to investigate whether REGENIE can be extended to handle time-to-event data[30] and multinomial regression in a mixed-model framework[31,32]. We suspect it may also be possible to leverage the REGENIE output to estimate SNP heritability, polygenic scores and multi-trait missing data imputation using mixed models on a scale that is not possible using the existing approaches[33].

One novel application would be to use REGENIE to analyze cohorts that have undergone both RNA-sequencing and either whole-genome SNP genotyping or sequencing. In this setting, the expression levels of up to 20,000 genes would represent the multiple traits of interest, and running a whole-genome regression analysis would allow for joint inference of *cis* and *trans* expression quantitative trait loci in a single analysis. This would be equivalent to an LMM analysis of an RNA-sequencing study, which has been performed in previous studies[34,35].

Cohorts will continue to grow in terms of sample size, the number of phenotypes and the number of variants available for testing, either via imputation from whole-genome-sequenced reference panels or via direct whole-genome sequencing of the study samples. It seems clear to us that Step 1 of the whole-genome regression paradigm is now highly computationally tractable using the REGENIE approach. However, further advances will be needed to reduce the compute time in Step 2, as whole-genome sequencing produces ever-increasing numbers of rare variants. Efficient utilization of the sparsity of such variants will help to improve memory efficiency and substantially reduce the cost of computation.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00870-7.

Received: 21 June 2020; Accepted: 13 April 2021;
Published online: 20 May 2021

References

1. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
2. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
3. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
4. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).
5. Yu, J. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).
6. Zhang, Z. et al. Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).
7. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
8. Listgarten, J. et al. Improved linear mixed models for genome-wide association studies. *Nat. Methods* **9**, 525–526 (2012).
9. Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829 (2001).
10. Campos, G. d. L., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D. & Calus, M. P. L. Whole genome regression and prediction methods applied to plant and animal breeding. *Genetics* **193**, 327–345 (2012).
11. Logsdon, B. A., Hoffman, G. E. & Mezey, J. G. A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinform.* **11**, 58 (2010).
12. Carbonetto, P. & Stephens, M. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal.* **7**, 73–108 (2012).
13. Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
14. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).
15. Kerin, M. & Marchini, J. Inferring gene-by-environment interactions with a Bayesian whole-genome regression model. *Am. J. Hum. Genet.* **107**, 698–713 (2020).
16. Jiang, L. et al. A resource-efficient tool for mixed model association analysis of large-scale data. *Nat. Genet.* **51**, 1749–1755 (2019).
17. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
18. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
19. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).
20. Kunert-Graf, J., Sakhanenko, N. & Galas, D. Allele frequency mismatches and apparent mismappings in UK Biobank SNP data. Preprint at *bioRxiv* https://doi.org/10.1101/2020.08.03.235150 (2020).
21. Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M. & Aulchenko, Y. S. Rapid variance components-based method for whole-genome association analysis. *Nat. Genet.* **44**, 1166–1170 (2012).
22. Breiman, L. Stacked regressions. *Mach. Learn.* **24**, 49–64 (1996).
23. Young, A. I., Wauthier, F. L. & Donnelly, P. Identifying loci affecting trait variability and detecting interactions in genome-wide association studies. *Nat. Genet.* **50**, 1608–1614 (2018).
24. Wu, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
25. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–775 (2012).
26. Zhou, W. et al. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nat. Genet.* **52**, 634–639 (2020).
27. Chib, S. & Greenberg, E. Analysis of multivariate probit models. *Biometrika* **85**, 347–361 (1998).
28. Korte, A. et al. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.* **44**, 1066–1071 (2012).
29. Dutta, D., Scott, L., Boehnke, M. & Lee, S. Multi-SKAT: general framework to test for rare-variant association with multiple phenotypes. *Genet. Epidemiol.* **43**, 4–23 (2018).
30. Rizvi, A. A. et al. gwasurvivr: an R package for genome wide survival analysis. *Bioinformatics* **35**, 1968–1970 (2018).
31. Morris, A. P. et al. A powerful approach to sub-phenotype analysis in population-based genetic association studies. *Genet. Epidemiol.* **34**, 335–343 (2010).

32. Jostins, L. & McVean, G. Trinculo: Bayesian and frequentist multinomial logistic regression for genome-wide association studies of multi-category phenotypes. *Bioinformatics* **32**, 1898–1900 (2016).
33. Dahl, A. et al. A multiple-phenotype imputation method for genetic studies. *Nat. Genet.* **48**, 466–472 (2016).
34. Kang, H. M., Ye, C. & Eskin, E. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* **180**, 1909–1925 (2008).
35. Shang, L. et al. Genetic architecture of gene expression in European and African Americans: an eQTL mapping study in GENOA. *Am. J. Hum. Genet.* **106**, 496–512 (2020).

Methods

Whole-genome linear regression. In a sample of N individuals, $\mathbf{y}$ denotes the N-element phenotype vector, G represents the $N \times M$ genotype matrix, where $G_{ij} \in \{0, 1, 2\}$ is the allele count for individual i at the jth marker and $\boldsymbol{X}$ represents the $N \times C$ matrix of covariates (including an intercept), which is assumed to be full rank. We consider a whole-genome regression model

$$\mathbf{y} = X\boldsymbol{\alpha} + G_S\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1)$$

where $\boldsymbol{\alpha}$ are the fixed covariate effects, G_S is a standardized version of G, the genotypes have been transformed to have a mean of zero and variance of one, $\boldsymbol{\beta} \sim \text{MVN}(0, \sigma_g^2 \text{I}_M)$ and $\boldsymbol{\varepsilon} \sim \text{MVN}(0, \sigma_e^2 \text{I}_N)$, where MVN denotes the multivariate normal distribution. This is the standard infinitesimal model, which can also be re-written as

$$\mathbf{y} = X\boldsymbol{\alpha} + \mathbf{g} + \boldsymbol{\varepsilon} \quad (2)$$

with $\mathbf{g} \sim \text{MVN}(0, \sigma_a^2 K)$, where $K = G_S G_S^T / M$ is usually referred to as a genetic-relatedness matrix or empirical kinship matrix and $\sigma_a^2 = M\sigma_g^2$ is the additive polygenic variance.

Covariate effects are removed from both the trait and the genotypes in equation (1) by first computing an orthonormal basis for the covariates, projecting the genotypes and the trait onto that basis and then subtracting out the resulting vectors to obtain the residuals. This is equivalent to using a projection matrix $P_X = I_N - X(X^TX)^{-1}X^T$ with

$$\tilde{\mathbf{y}} = P_X\mathbf{y} \quad (3)$$

$$\tilde{G} = P_X G_S \quad (4)$$

Both the genotype and phenotype residuals are then scaled to have a variance of one.

Stacked block ridge regression. Fitting equation (1) is computationally intensive given that G typically has many hundreds of thousands of columns. Instead, for Step 1, we transform the model to

$$\tilde{\mathbf{y}} = W\boldsymbol{\eta} + \boldsymbol{\varepsilon} \quad (5)$$

where W is a matrix derived from G with substantially fewer columns. Specifically, we divide G into blocks of B consecutive and non-overlapping SNPs, and from each block we derive a small set of predictors using ridge regression across a range of J shrinkage parameters (see Supplementary Note). The idea behind using a range of shrinkage values is to capture the unknown number and size of truly associated genetic markers within each window. This approach is equivalent to placing a Gaussian prior on the effect sizes of the SNPs in the block and finding the maximum a posteriori estimate of the effect sizes and the resulting prediction. Another approach would be to integrate out the effect sizes over the Gaussian prior to obtain the best linear unbiased prediction[36] but we have not investigated that approach in this paper.

The ridge predictors are re-scaled to have unit variance and are stored in place of the genetic markers in matrix W, providing a large reduction in data size. If $M = 500{,}000$, $B = 1{,}000$ and $J = 5$ are used, then the reduced dataset will have $JM/B = 2{,}500$ predictors. We refer to this part of the method as the Level 0 ridge regression.

To keep the memory usage low when analyzing multiple traits, the within-block predictions are stored on disk and read separately for each trait when fitting models at Level 1 (see below). The added input/output operations incur a small cost on the overall runtime and substantially decrease the amount of memory needed.

The ridge regression takes account of the LD within each block but not between blocks. One option that we have considered, but not implemented yet, is to condition the ridge regression on the estimates from the previous block, which may better account for LD across block boundaries.

The predictors in W will all be positively correlated with the phenotype. Thus, it is important to account for that correlation when building a whole-genome-wide regression model. The predictors will also be correlated with each other, especially within each block, but also between blocks that are close together due to LD. We use a second level of ridge regression on W for a range of shrinkage parameters and choose a single best value using the K-fold CV scheme[22]. This assesses the predictive performance of the model using held-out sets of data and aims to control any over-fitting induced by using the first level of ridge regression to derive the predictors (see Supplementary Note). We refer to this part of the method as the Level 1 ridge regression.

The result of this model fit is a single $N \times 1$ predicted phenotype $\hat{\mathbf{y}}^*$, and this can be partitioned into 22 LOCO predictions (denoted $\hat{\mathbf{y}}^*_{\text{LOCO}}$), which are used when testing SNPs for association in Step 2 to avoid proximal contamination (see Supplementary Note).

Association testing. When testing for association of the phenotype with a variant $\mathbf{g}$ in Step 2, we consider a simple linear model

$$\hat{\mathbf{y}}^*_{\text{resid,LOCO}} = \tilde{\mathbf{g}}\beta + \tilde{\boldsymbol{\varepsilon}} \quad (6)$$

where $\hat{\mathbf{y}}^*_{\text{resid,LOCO}} = \tilde{\mathbf{y}} - \hat{\mathbf{y}}^*_{\text{LOCO}}$ refers to the phenotype residuals where the polygenic effects estimated from the null model with LOCO have been removed, $\tilde{\mathbf{g}} = P_X\mathbf{g}$ are residuals obtained from removing the covariate effects from the tested variant and $\tilde{\boldsymbol{\varepsilon}} = P_X\boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \text{MVN}(0, \sigma_e{}^2 I_N)$.

A score test statistic for H_0: $\beta = 0$ is

$$T_{\text{linear}} = \frac{\tilde{\mathbf{g}}^T\hat{\mathbf{y}}^*_{\text{resid,LOCO}}}{[\hat{\sigma}_e^2 \cdot \tilde{\mathbf{g}}^T\tilde{\mathbf{g}}]^{1/2}} \quad (7)$$

where we use $\hat{\sigma}_e^2 = ||\hat{\mathbf{y}}^*_{\text{resid,LOCO}}||_2^2/(N - C)$. In equation (7), when estimating the variance of the numerator, we assume that the polygenic effects are given, which leads to the denominator involving only $O(N)$ computation. While other methods make use of a calibration factor in the denominator to account for the variance of the polygenic effects[13,18,21], we found in applications that the results obtained using this simple form match up closely to those using a calibration factor. Finally, we use a normal approximation, $T^2_{\text{linear}} \sim \chi^2_1$, to estimate the P value. As with Step 1 above, the REGENIE software reads the genetic data file in blocks of B SNPs and these are processed together, taking advantage of parallel linear algebra routines in the Eigen library.

Multiple traits. Both Step 1 and Step 2 above are easily extended so that multiple phenotypes can be processed in parallel. The genetic data files in both steps can be read once, in blocks of B SNPs, which means the method uses a small amount of memory. In addition, the linear algebra operations for the covariate residualization, ridge regression and association testing can be shared across traits. This is similar to the approach implemented in the BGENIE software for single SNP linear regression analysis[17]. The fine details of the multiple phenotype approach are given in the Supplementary Note.

Binary traits. For binary traits, we use exactly the same Level 0 ridge regression approach, which effectively treats the trait as if it were quantitative. However, at Level 1, instead of a linear regression in equation (5) we use logistic regression

$$\text{logit}\,(p_i) = \mathbf{X}_i^T\boldsymbol{\alpha} + \mathbf{W}_i^T\boldsymbol{\eta} \quad (8)$$

where $p_i = \text{E}(y_i) = \text{P}(y_i = 1)$ with y_i indicating the case status of the ith individual, $\mathbf{X}_i$ is the covariate vector for the ith individual, $\boldsymbol{\alpha}$ are the fixed covariate effects, $\mathbf{W}_i$ are the within-block (BR) predictions for the ith individual and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{\text{BR}})^T$, with $\eta_i \sim \text{N}(0, 1/\omega)$. This model corresponds to logistic regression with ridge penalty applied to the effects of within-block predictions in W.We approximate the model in equation (8) by first fitting a null model for each trait that only has

$$\text{logit}\,(p_i) = \mathbf{X}_i^T\boldsymbol{\alpha} \quad (9)$$

covariate effects and then using the resulting estimated effects as an offset in the model in equation (8),

$$\text{logit}\,(p_i) = \mathbf{X}_i^T\hat{\boldsymbol{\alpha}} + \mathbf{W}_i^T\boldsymbol{\eta} \quad (10)$$

where $\hat{\boldsymbol{\alpha}}$ represents the effects estimated in equation (9). As the covariate effects are not expected to change substantially (unless correlation between covariates and block predictions are very large), this approximation is expected to work well in most analyses.

As with quantitative traits, we used K-fold CV to choose the Level 1 ridge regression parameter. However, for extremely unbalanced traits, it may happen that one of the folds contains no cases. To avoid this situation, we also implemented an efficient version of LOOCV. Although at first sight it may seem that LOOCV is more computationally intensive than K-fold CV given that the model has to be fitted N times (rather than K times) on data with $N - 1$ samples, the leave-one-out estimates can actually be obtained (approximately for binary traits) from rank 1 updates to the results from fitting the model once to the full data (see Supplementary Note). In practice we have found that LOOCV gives similar association results to K-fold CV (Supplementary Figs. 14 and 15) and can be computationally faster in some cases (see Tables 1 and 2). A LOCO scheme is applied to the polygenic effect estimates and the resulting predictions $\hat{\mathbf{w}}_{\text{LOCO}} = W_{\text{LOCO}}\hat{\boldsymbol{\eta}}_{\text{LOCO}}$ are then stored.

In Step 2, we use a logistic regression model score test to test for association between each marker and binary trait. Covariate effect sizes are estimated along with genetic marker effect sizes but we include the LOCO predictions from Step 1 as a fixed offset (see Supplementary Note).

When rare variants are tested for association with a highly unbalanced trait (that is, a trait that has low sample prevalence), the use of asymptotic test statistic distributions does not work well and results in elevated Type 1 error rates. REGENIE implements several methods to handle this situation. First, it includes the SPA test[37], which is also included in SAIGE[18]. This approach better

approximates the null distribution of the test statistic but we have found that it can sometimes fail to produce good estimates of SNP effect sizes and standard errors, which are highly desirable for meta-analysis applications (Supplementary Table 4 and Supplementary Fig. 2).

Second, we use Firth logistic regression, which uses a penalized likelihood to remove much of the bias from the maximum-likelihood estimates in the logistic regression model. This approach results in well-calibrated Type 1 errors and usable SNP effect sizes and standard errors. Given that the use of Firth regression can be relatively computationally intensive, we have developed an approximate Firth regression approach that is much faster (Supplementary Table 5), which involves estimating the covariate effects in a null Firth regression model and then including covariate effects along with the LOCO genetic predictor as offset terms in a Firth logistic regression test (see Supplementary Note). In practice, we have found this approximation to give very similar results to when the exact Firth test is used (Supplementary Fig. 3). This approach has been used to analyze COVID-19 outcomes across four studies and four ancestries[38], and proved vital to provide accurate effect-size estimates for the meta-analysis.

Handling missing data. As a key goal of our approach is to analyze multiple traits all at once, one issue that remains to be addressed is the presence of 'missingness' in the data, which could differ among the traits. We consider different approaches based on the nature of the trait as well as whether the null model is being fitted or whether association testing is being performed.

For quantitative traits, when fitting the null model missing data is addressed by replacing the missing values by the sample averages for each trait and in the association testing step, individuals with missing phenotype observations are removed from the analysis for each trait. The latter is done by ensuring that when taking sums over individuals, those with a missing phenotype have a zero contribution to the sum. This is similar to the approach implemented in the BGENIE software (https://jmarchini.org/BGENIE/) for single SNP linear regression analysis[17]. We assume that covariates are fairly well-balanced in the sample and project them out of the phenotypes using all of the samples (that is, ignoring the missingness within each trait). In the case where phenotypes have the same or very similar patterns of missingness, or if only a single phenotype is being analyzed, it may be more logical to discard the missing observations rather than impute them with the sample averages per trait. Hence, we implement an alternative approach where, in both the null-fitting and the association testing steps, all samples with missingness at any of the P phenotypes are dropped. An approach we have not yet implemented, but may produce better results for quantitative traits, would involve using a multivariate normal model to jointly model correlation between the set of traits and impute missing data, either before or conditional on the output of Step 1.

For binary traits, we use the mean-imputed phenotypes to fit the Level 0 linear ridge regression models within blocks but discard missing observations when fitting Level 1 logistic ridge regressions. As the logistic ridge regressions are fitted separately for each trait, this makes it straightforward to account for the missingness patterns separately for each trait. Similarly, in the testing step, we discard missing observations when fitting logistic regression for each trait as well as when using Firth or SPA corrections.

UK Biobank dataset. The UK Biobank[17] (http://www.ukbiobank.ac.uk) is a large prospective study of about 500,000 individuals who are 40–69 years old and for whom extensive phenotype information is being recorded. Genotyping was performed using the Affymetrix UK BiLEVE Axiom array on an initial set of 50,000 participants and the Affymetrix UK Biobank Axiom array was used for the remaining participants. Up to 11,914,699 variants imputed by the Haplotype Reference Consortium panel that either have a minor allele frequency above 0.5% or a minor allele count above five and are annotated as functional in 462,428 samples of European ancestry were used in the data analyses. We selected up to 407,746 individuals of white British ancestry for whom genotype and imputed data were available and applied quality-control filters on the genotype data using PLINK2 (ref. [39]; version v2.00aLM, https://www.cog-genomics.org/plink2) that included: a minor allele frequency of ≥1%, a Hardy–Weinberg equilibrium test not exceeding $P = 1 \times 10^{-15}$, a genotyping rate above 99%, not present in low-complexity regions, not involved in inter-chromosomal LD and LD pruning using a R^2 threshold of 0.9 with a window size of 1,000 markers and a step size of 100 markers. This resulted in up to 471,762 genotyped SNPs that were kept in the analyses.

Data simulation. We performed simulations to assess the performance of the tests in REGENIE under various population-structure configurations for both quantitative and binary traits. To mimic realistic scenarios, we used genotype array data from the UK Biobank European samples (679,209 array SNPs with a minor allele count > 5). We considered scenarios with 100,000 samples obtained from the set of white British participants or from the full European set so as to incorporate various amounts of population structure. In addition, we varied the proportion of related individuals selected from 0 to 50% of the sample, where we defined a pair of individuals as related if their estimated kinship coefficient, provided by UK Biobank using KING[40], was above 0.044. This is to assess how REGENIE would perform in samples with higher amounts of relatedness. We also considered randomly selected samples from the white British or European set, irrespective of the relatedness information, where about 30% of samples in these sets are related up to the third degree. Finally, to consider scenarios of more extreme relatedness, we considered scenarios with samples consisting of only first-degree white British relatives ($N = 22{,}990$), first- and second-degree white British relatives ($N = 30{,}775$), and first- to third-degree white British relatives ($N = 70{,}684$).

We generated quantitative traits as

$$Y_i = \sum_{j=1}^{M} G_{ij}\beta_j + A_\gamma + \varepsilon_i$$

where the M causal SNPs were randomly selected only from odd chromosomes with a minor allele count above 100 and not involved in inter-chromosomal LD, and G_{ij} represents the standardized genotype for individual i at the jth causal SNP, A_i represents the score of the individual for the top principal component from a genotype relatedness matrix using SNPs on odd chromosomes and ε_i represents the environmental effects. The effect sizes for the causal SNPs were sampled from a normal distribution with a mean of zero, where the variance was determined based on the desired proportion of trait variance explained by the causal SNPs h_g^2. The effect from population structure γ was set so that the proportion of the trait variance explained by the top principal component was 5%. The environmental effects were sampled from a normal distribution with a mean of zero and the variance was set to correspond to a trait variance of one.

For binary traits, we used the model described above to obtain a quantitative phenotype and then applied a threshold based on a target sample prevalence value K to dichotomize the phenotype and obtain a binary trait. We also considered simulations to assess the effect-size estimates using a logistic model where

$$\text{logit}\,(p_i) = \beta_0 + \sum_{j=1}^{M} G_{ij}\beta_j$$

and $Y_i|p_i \sim \text{Bernoulli}(p_i)$, independently, where β_0 was chosen to achieve the desired prevalence level, and the effect sizes of the causal SNPs were sampled from a normal distribution with a mean of zero and the variance parameter was chosen so that they explain 20% of the variance on the logistic scale.

We simulated up to 100 phenotypic replicates for each simulation setting and selected 10,000, 25,000 or 50,000 SNPs to be causal. For binary traits, we varied the sample prevalence K between 0.1, 0.01 or 0.001, corresponding to a case–control ratio of 1:9, 1:99 or 1:999, respectively, and fixed the number of causal SNPs to 10,000. SNPs on even chromosomes ($M_{\text{null}} = 324{,}838$ variants) were used to assess the Type 1-error performance and the power was estimated using the set of causal SNPs for each trait. REGENIE was compared with BOLT-LMM with a mixture of Gaussian's model (BOLT-LMM-MoG), BOLT-LMM with infinitesimal model (BOLT-LMM-inf), fastGWA, SAIGE (only for binary traits and run with LOCO scheme) and PCA (using only the top principal component as a covariate in Step 2 of REGENIE without the LOCO predictions from Step 1). The top principal component was included as a covariate for all methods. For REGENIE–Firth, REGENIE–SPA and SAIGE, the P-value fallback threshold for Firth/SPA correction was set to 0.05.

Statistical analyses. We used REGENIE to perform genome-wide association analyses on up to approximately 11 million imputed variants for 50 quantitative traits and 54 binary traits of up to 407,746 white British participants in the UK Biobank. Quantitative phenotypes were converted to z-scores using rank-inverse-based normal transformation. In the statistical models used, the covariates included age, age^2, sex, age × sex and the top-10 principal components provided by the UK Biobank to appropriately correct for population stratification. To assess the performance of REGENIE in genome-wide association studies, we compared the results from REGENIE with those of existing approaches for large-scale analysis, which included BOLT-LMM (version 2.3; https://data.broadinstitute.org/alkesgroup/BOLT-LMM/) and fastGWA (GCTA version 1.93.0beta; https://cnsgenomics.com/software/gcta/#Overview) for quantitative phenotypes and SAIGE (version 0.36.5.1; https://github.com/weizhouUMICH/SAIGE) with the LOCO option for binary traits. For all methods, Step 1 was run on a set of array SNPs stored in bed/bim/fam format and Step 2 was run on imputed data stored in BGEN format. All association analyses used a $\chi^2_{\text{df}=1}$ statistic to test a variant for association with a trait (that is H_0:$\beta_{\text{SNP}} = 0$). All programs were called within R[41], where we used the function system.time to track the CPU and wall-clock timings.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The individual-level genotype and phenotype data are available through formal application to the UK Biobank (http://www.ukbiobank.ac.uk). Results from the genome-wide association study analyses in this paper have been deposited in the GWAS Catalog under the accession numbers GCST90013862–GCST90014022.

Code availability

The C++ source code for REGENIE is available from https://rgcgithub.github.io/regenie/ under an MIT License. Analysis code for the main results in the paper can be found at https://github.com/rgcgithub/regenie/tree/master/scripts.

References

36. Robinson, G. K. That BLUP is a good thing: the estimation of random effects. *Stat. Sci.* **6**, 15–32 (1991).
37. Dey, R., Schmidt, E. M., Abecasis, G. R. & Lee, S. A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *Am. J. Hum. Genet.* **101**, 37–49 (2017).
38. Horowitz, J. E. et al. Common genetic variants identify therapeutic targets for COVID-19 and individuals at high risk of severe disease. Preprint at *medRxiv* https://doi.org/10.1101/2020.12.14.20248176 (2020).
39. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
40. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
41. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2013).

Acknowledgements

We thank F. A. Nothaft, H. Davidge, K. Kianfar, K. Feng and Y. Huang for their ongoing advice on developing the REGENIE code in the Databricks environment.

Author contributions

J. Marchini conceived and supervised the study. J. Marchini, J. Mbatchou, L.B. and E.M. developed the method for quantitative traits. J. Mbatchou and J. Marchini developed the method for binary traits. J. Mbatchou and J. Marchini coded the C++ implementation of the method. J. Mbatchou carried out all of the testing and real data analysis of the C++ method. L.B. and E.M. developed the Apache Spark implementation of the method. C.B. provided advice and code for the LD calculations. J.B., A.M. and J.A.K. tested and provided comments on the C++ version. J. Marchini and J. Mbatchou wrote the manuscript. A.M., J.A.K., A.Z., C.O., M.B., B.B., L.H., J.R., M.F., A.B. and G.A. provided helpful comments at various stages of the project.

Competing interests

The authors declare no competing interests.

Additional information

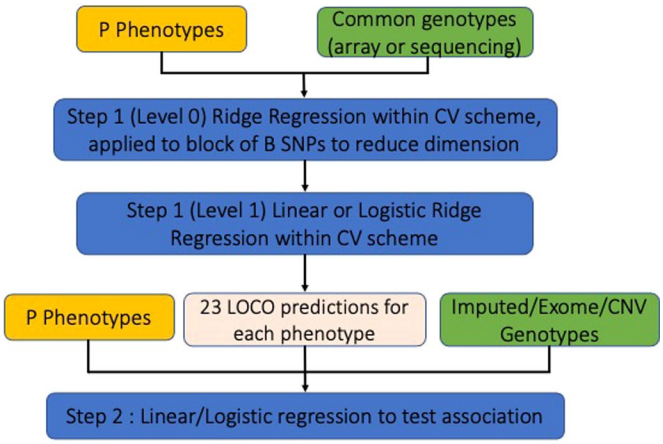
Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00870-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00870-7.

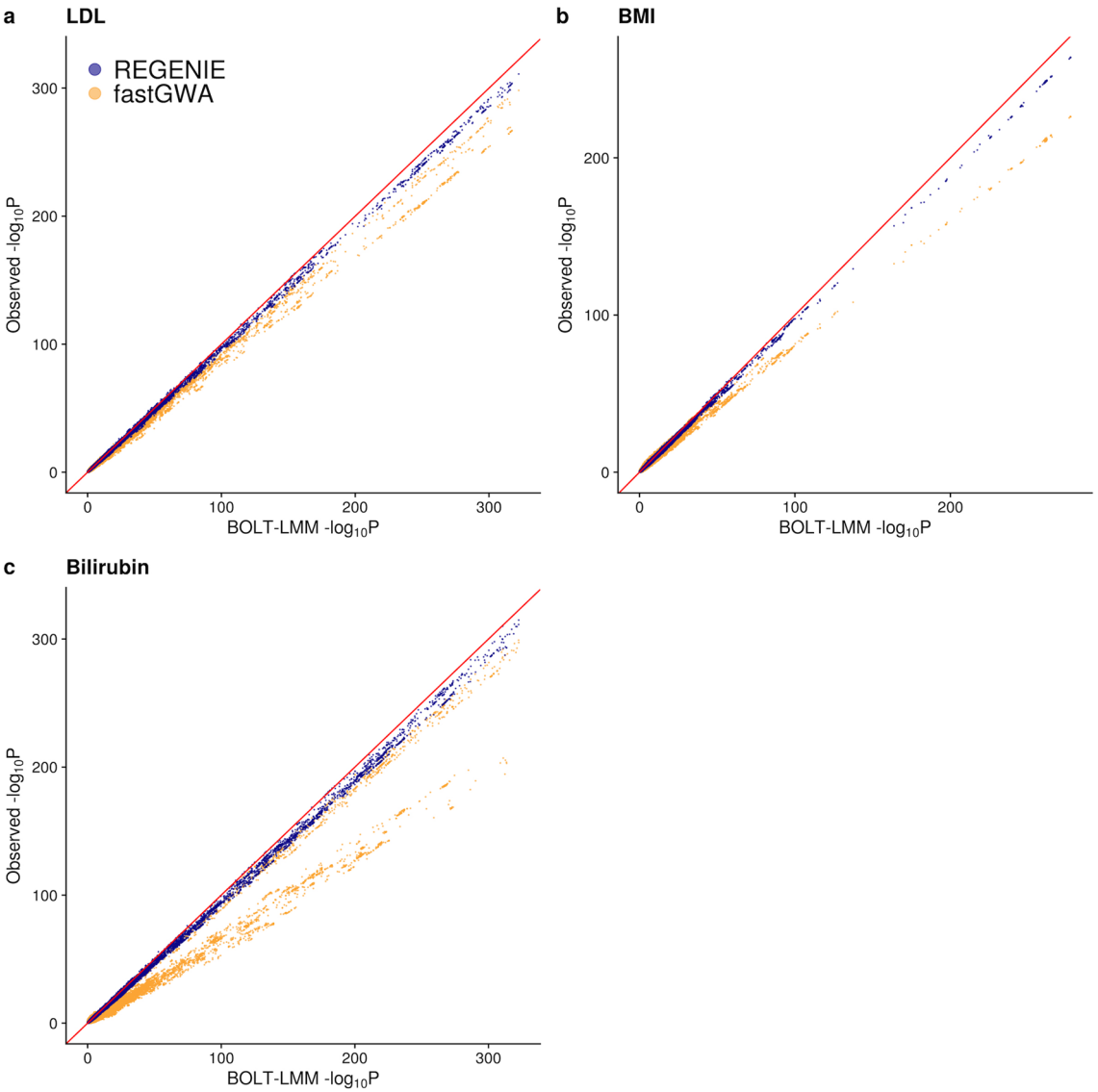
Correspondence and requests for materials should be addressed to J.M.

Peer review information *Nature Genetics* thanks Xia Shen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

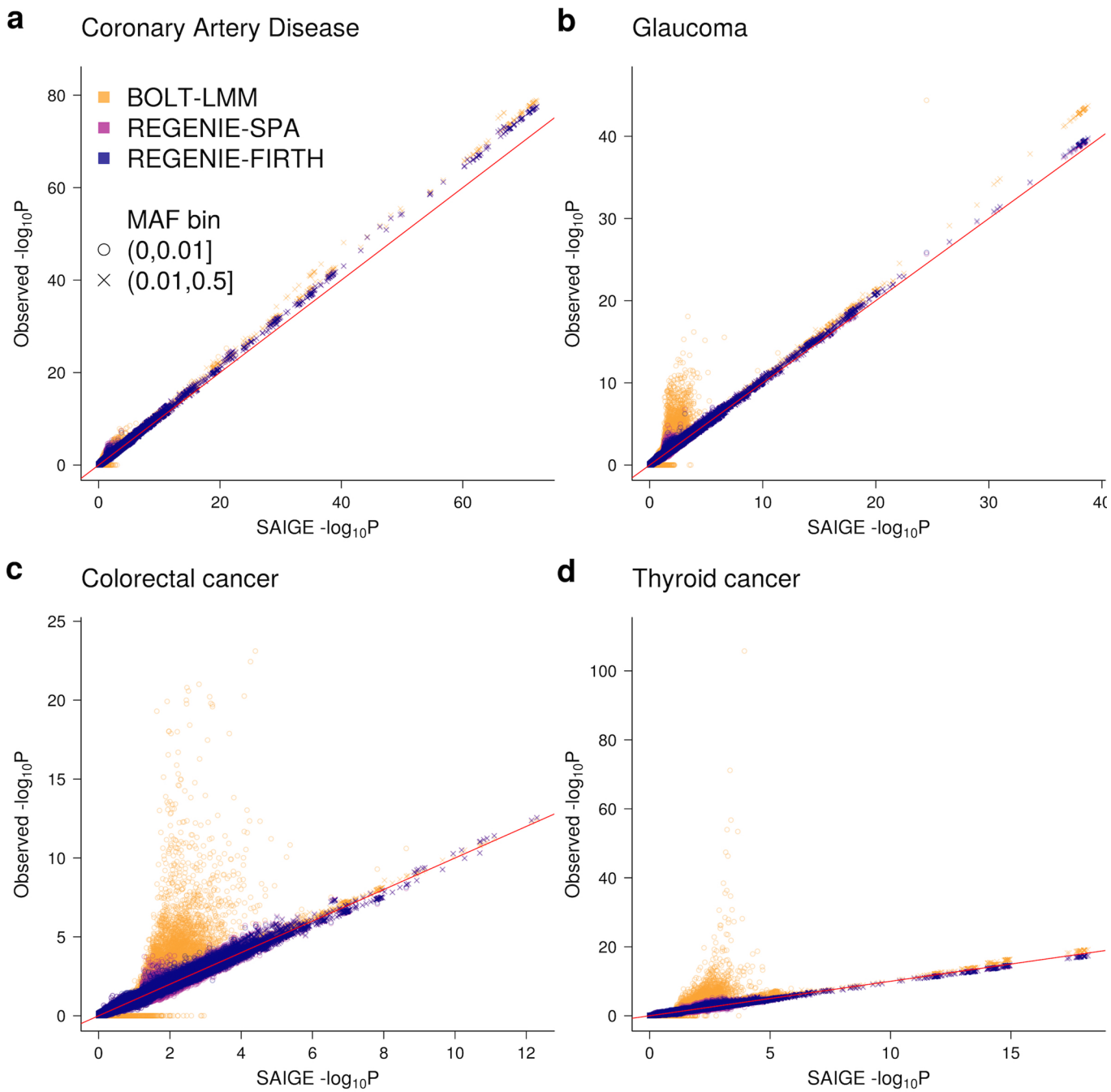
Reprints and permissions information is available at www.nature.com/reprints.



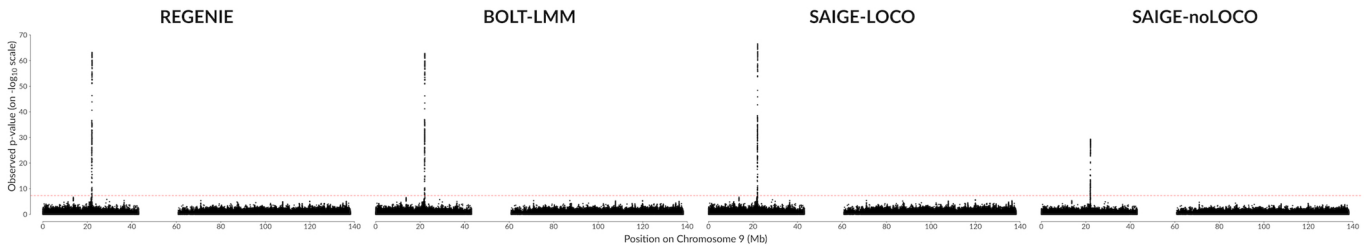
Extended Data Fig. 1 | Overview of the REGENIE method. REGENIE consists of two steps: (1) In Step 1, the dimension of the genetic data is reduced using ridge regression applied to blocks of SNPs, and then the resulting predictors are combined using a second round of linear or logistic ridge regression to produce an overall prediction for each trait, split into 23 LOCO predictors. (2) In Step 2, these LOCO predictors are used when testing each phenotype against a set of either imputed, exome or CNV markers.



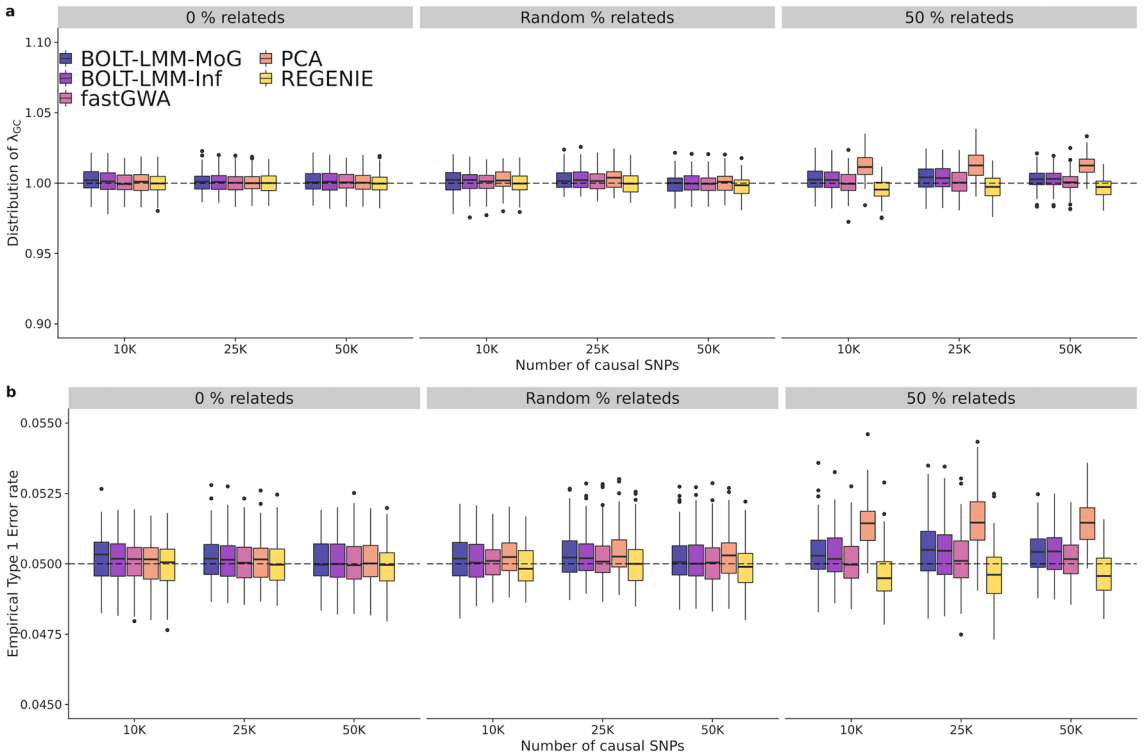
Extended Data Fig. 2 | Scatterplots comparing three LMM methods for three quantitative traits using UK Biobank white British samples. Results from REGENIE, fastGWA and BOLT-LMM are compared for (a) LDL (N = 389, 189), (b) BMI (N = 407, 609) and (c) Bilirubin (N = 388, 303). 9.8 million imputed SNPs with minor allele frequency above 1% are tested for association with each trait. For each trait, the p-value for each variant was obtained using a $\chi^2_{df=1}$ test statistic.



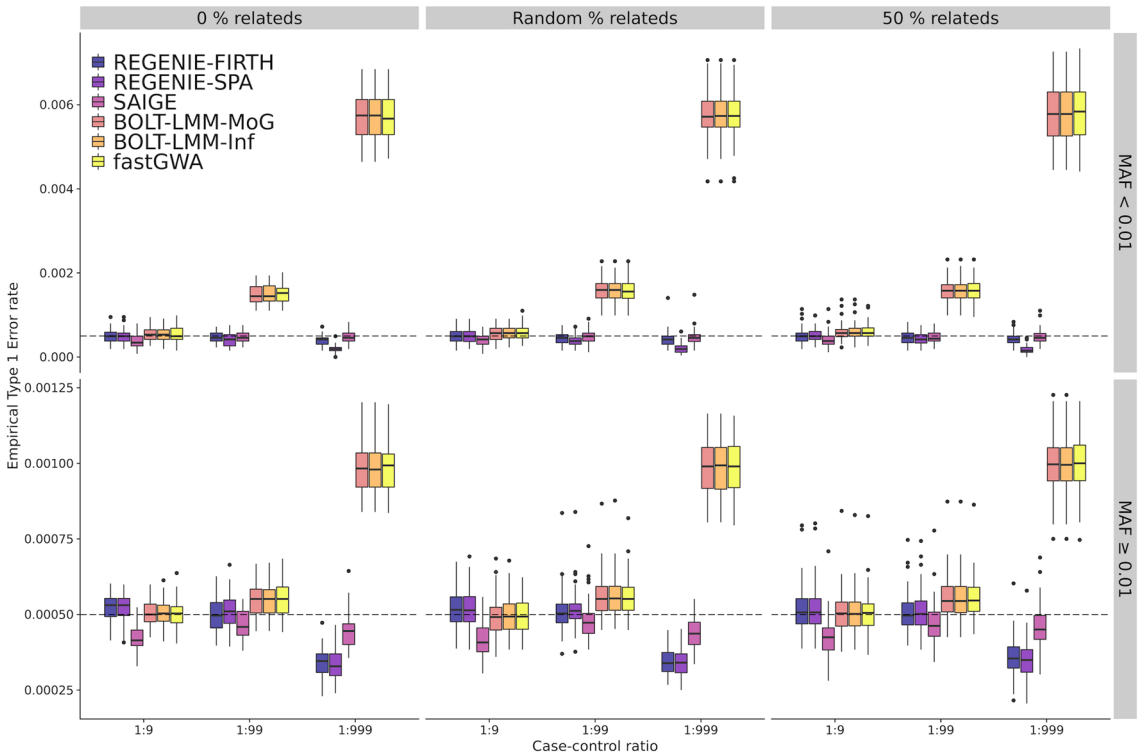
Extended Data Fig. 3 | Scatterplots comparing results from different mixed model methods for 4 binary traits using UK Biobank white British samples. Results from REGENIE using Firth and SPA correction, BOLT-LMM and SAIGE are compared for (a) coronary artery disease (case-control ratio=1:11, N = 352,063), (b) glaucoma (case-control ratio=1:52, N = 406,927), (c) colorectal cancer (case- control ratio=1:97, N = 407,746), and (d) thyroid cancer (case-control ratio=1:660, N = 407,746). Tests were performed on 11.6 million imputed SNPs, and the plotting symbols represents variant categories based on using a minor allele frequency (MAF) threshold of 1%. For each trait, the p-value for each variant was obtained using a $\chi^2_{df=1}$ test statistic.



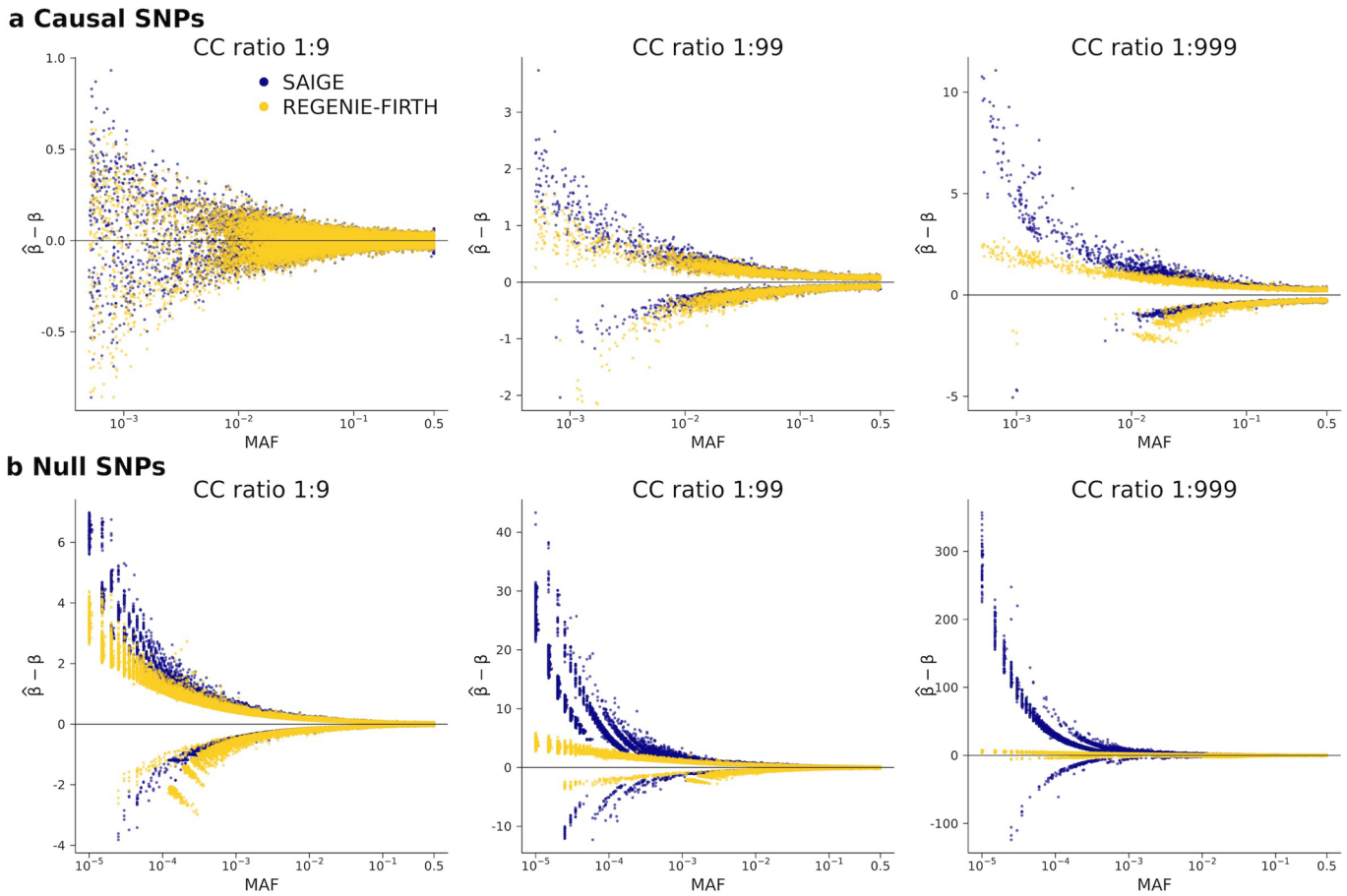
Extended Data Fig. 4 | Manhattan plots comparing association results for coronary artery disease using 337,484 unrelated white British participants from UK Biobank. For REGENIE, BOLT-LMM and SAIGE-noLOCO, 329,641 genotyped SNPs from chromosomes 1-22 are included as model SNPs in step 1, and for SAIGE-LOCO all SNPs from chromosome 9 are excluded which results in 314,309 SNPs. In step 2, 482,884 imputed SNPs on chromosome 9 are tested for association. The red dashed horizontal line represents the genome-wide significance level of 5×10^{-8}.



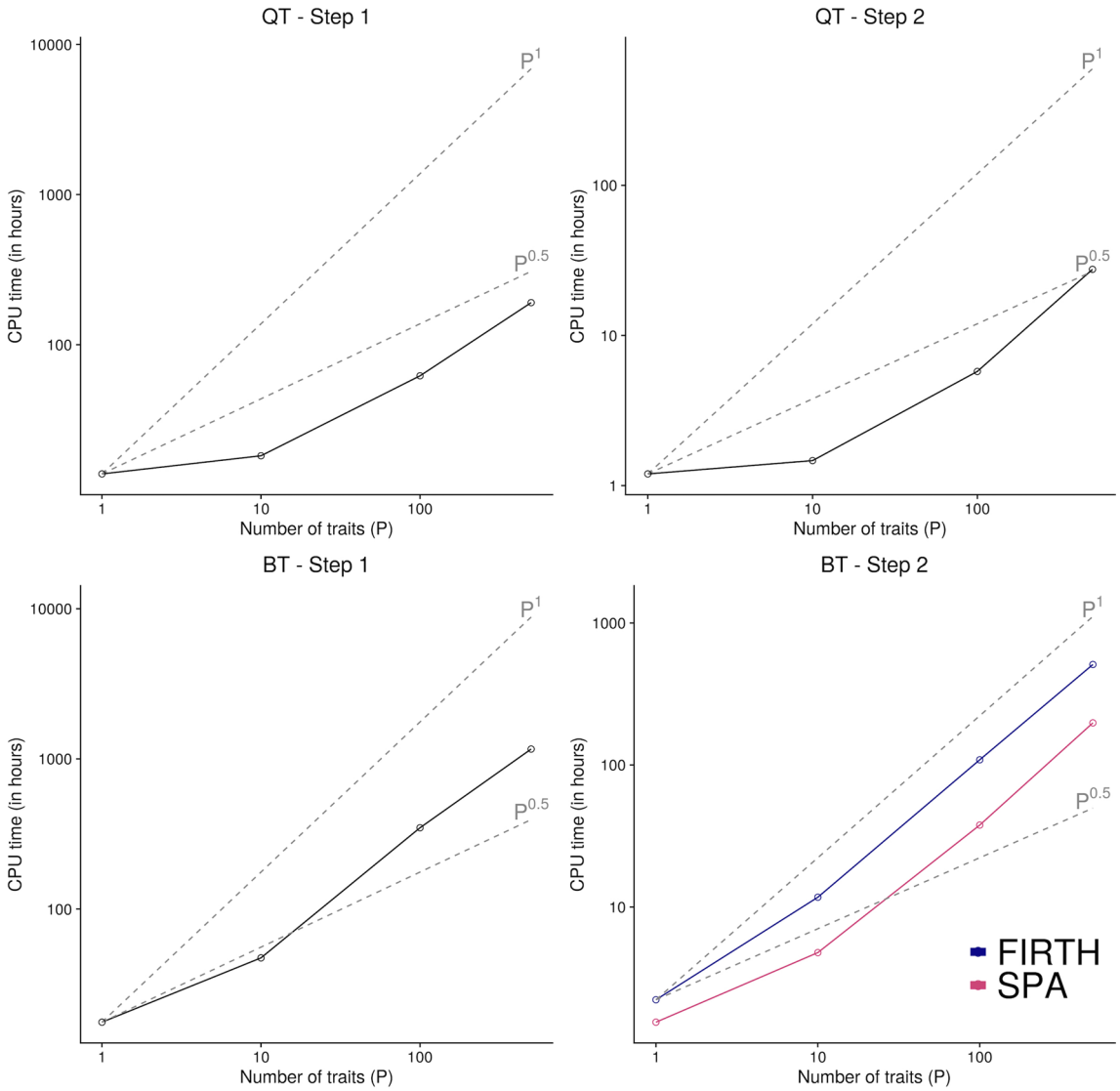
Extended Data Fig. 5 | Type 1 error performance on simulated quantitative traits with UK Biobank white British samples. (a) Distribution of λ_{GC} computed at null SNPs. (b) Distribution of empirical type 1 error rates at nominal level 0.05 computed at null SNPs. Each boxplot represents the distribution of the estimated quantity across 100 simulation replicates. Quantitative traits were simulated fixing h_g^2 (the proportion of trait variance explained by causal SNPs) to 0.2 and the number of causal SNPs was varied from 10,000 to 50,000. The proportion of related individuals in the sample of size 100,000 was varied from 0% to 50% (including randomly selecting individuals from the white British set which includes about 30% related individuals). Each box indicates the interquartile range (IQR) and the line inside each box is the median value, the whiskers indicate data up to 1.5 times the IQR, and outliers are indicated by individual dots.



Extended Data Fig. 6 | Type 1 error performance on simulated binary traits with UK Biobank white British samples. Each boxplot represents the distribution of empirical type 1 error rates at nominal level 5×10^{-4} across 100 simulation replicates. Each type 1 error rate was evaluated at 324,838 null SNPs using a minor allele frequency filter of 1%. Binary traits were simulated fixing h_g^2 (the proportion of variance on liability scale explained by 10,000 causal SNPs) to 0.2 and the case–control ratio was varied from 1:999 to 1:9. With a total sample size of 100,000, the proportion of related individuals was also varied from 0% to 50% (including randomly selecting individuals from the white British set which includes about 30% related individuals). Each box indicates the interquartile range (IQR) and the line inside each box is the median value, the whiskers indicate data up to 1.5 times the IQR, and outliers are indicated by individual dots.



Extended Data Fig. 7 | Effect size estimates for REGENIE-FIRTH and SAIGE on simulated binary traits with 100,000 UK Biobank white British samples. REGENIE-FIRTH and SAIGE were run on 10 simulated binary trait replicates with case–control ratio varied between 1:9 and 1:999. A logistic model was used to simulate the traits randomly selecting 10,000 SNPs on odd chromosomes with minor allele count (MAC) above 100 to be causal. The effect size estimates $\hat{\beta}$ are compared to the true effect sizes β for (a) causal SNPs or (b) the null SNPs on even chromosomes. Summary statistics were obtained for variants with minor allele count greater than 5, p-values in REGENIE-FIRTH and SAIGE below 0.05 (fallback p-value threshold for Firth/SPA correction, respectively), and not involved in inter-chromosomal LD.



Extended Data Fig. 8 | Computation time of REGENIE as the number of quantitative traits analyzed increases. 100,000 samples were used in a single run of REGENIE with 1, 10, 100 or 500 simulated quantitative trait (QT) or binary trait (BT) replicates. All axes are on $\log_{10}$ scale. The slope of the dotted lines represents the power law scaling with the number of traits *P*.

nature research

Corresponding author(s): Jonathan Marchini

Last updated by author(s): Jan 6, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection
Data analysis	REGENIE (version 1.0) https://rgcgithub.github.io/regenie/ PLINK (version 2.0)https://www.cog-genomics.org/plink2 BOLT-LMM (version 2.3) https://data.broadinstitute.org/alkesgroup/BOLT-LMM SAIGE (version 0.36.5.1) https://github.com/weizhouUMICH/SAIGE fastGWA (GCTA version 1.93.0beta) https://cnsgenomics.com/software/gcta/#Overview LDstore (version 2.0) http://www.christianbenner.com/ KING (version 2.2.5) https://people.virginia.edu/~wc9c/KING/

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The individual-level genotype and phenotype data are available through formal application to the UK Biobank http://www.ukbiobank.ac.uk

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We used genetic data on 462,428 samples of European ancestry who were participants of the UK Biobank. This dataset is one of the largest genetic datasets available and provides a sufficient test dataset for the new method described in the paper. We ran analysis on several different subsets of the datasets to evaluate different aspects of the method.
Data exclusions	We used up to 11,914,699 variants imputed by the Haplotype Reference Consortium (HRC) panel that either have minor allele frequency above 0.5% or have minor allele count above 5 and are annotated as functional in 462,428 samples of European ancestry were used in the data analyses. We selected up to 407,746 individuals of white British ancestry who had genotype and imputed data available and applied quality control filters on the genotype data using PLINK2, which included MAF >1%, Hardy-Weinberg equilibrium test not exceeding 10^-15 significance, genotyping rate above 99%, and LD pruning using a R^2 threshold of 0.9 with a window size of 1000 markers and a step size of 100 markers. This resulted in up to 471,762 genotyped SNPs that were kept in the analyses.
Replication	The purpose of this study is to describe the performance of the new method compared to other methods. We make no claim on new discoveries from the analysis, so there is no requirement for replication.
Randomization	The purpose of this study is to describe the performance of the new method compared to other methods. We use real phenotypes omn the UKB participants, so there is no requirement for randomization of subjects.
Blinding	The purpose of this study is to describe the performance of the new method compared to other methods, so blinding is not relevant to this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging