isGWAS: ultra-high-throughput, scalable and equitable inference of genetic associations with disease

Christopher N Foley[1,2*], Zhana Kuncheva[1,2], Riccardo E. Marioni[3], Heiko Runz[4], Benjamin B Sun[4*].

[1]Optima Partners, Edinburgh, UK.

[2]Bayes Centre, University of Edinburgh, Edinburgh, UK.

[3]Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK.

[4]Translational Sciences, Biogen Inc., Cambridge, MA, US.

*Correspondence:

Dr Christopher N Foley

Email: chris.foley@optimapartners.co.uk or chris.neal.foley@gmail.com

Dr Benjamin B Sun

Email: benjamin.sun@biogen.com or bbsun92@outlook.com

Abstract

Genome-wide association studies (GWAS) have proven a powerful tool for human geneticists to generate biological insights or hypotheses for drug discovery. Nevertheless, a dependency on sensitive individual-level data together with ever-increasing cohort sample sizes, numbers of variants and phenotypes studied put a strain on existing algorithms, limiting the GWAS approach from maximising potential. Here we present in-silico GWAS (isGWAS), a uniquely scalable algorithm to infer regression parameters in case-control GWAS from cohort-level summary data. For any sample size, isGWAS computes a variant-disease association parameter in ~1 millisecond, or ~11m variants in UK-Biobank within ~4 minutes (~1500-fold faster than state-of-the-art). Extensive simulations and empirical tests demonstrate that isGWAS results are highly comparable to traditional regression-based approaches. We further introduce a heuristic re-sampling algorithm, leapfrog re-sampler (LRS), to extrapolate association results to semi-virtually enlarged cohorts. Owing to significant computational gains we anticipate a broad use of isGWAS and LRS which are customizable on a web interface.

Main

Genome wide association studies (GWAS) have been immensely successful in unravelling the genetic contribution to human disease. Cost-effective genotyping and large biobank cohorts now make it possible to routinely conduct GWAS for tens of millions of variants in hundreds of thousands of individuals across thousands of phenotypes[1]. With the advent of population-scale whole genome sequencing and expansion of GWAS to research participants of non-European ancestries, these numbers can be expected to increase by another magnitude over the next few years[2], [3].

Current GWAS approaches that compute variant-disease associations in a regression framework, such as PLINK[4], fastGWA[5], BOLT-LMM[6], SAIGE[7] and REGENIE[8], require access to and input from ever increasing individual-level data (ILD). The efforts of individual-level GWAS sample collection, genotyping and data analysis tend to grow as a polynomial function of sample size[7], [8]. Moreover, the exchange of ILD between researchers is non-trivial due partly to data size but increasingly to strict – but essential - data protection regulations, which can limit the scope of collaborative analyses and biological insights gained[9]–[12]. Finally, the substantial computational and financial burden of running massive-scale GWAS, especially for binary disease outcomes, is exacerbating inequity between researchers, typically favouring already well-equipped institutions. There is therefore a pressing need for innovative approaches that help attenuate the increasing resource and financial inequities for conducting contemporary GWAS and to help decide where limited resources should best be allocated.

Here we present in-silico GWAS (isGWAS), a biobank-scalable and computationally highly efficient algorithm to infer genetic regression parameters in case-control GWAS from just four broadly ascertained cohort-level summary parameters: the counts of cases and controls within a cohort, as well as case and control minor allele frequencies (MAFs). isGWAS is highly parallelisable, exceeding efficiencies of current GWAS analysis tools by several orders of magnitude. Furthermore, we demonstrate that isGWAS yield association summary statistics highly comparable to traditional ILD regression-based approaches through extensive simulations and empirical tests in UK Biobank[13], Biobank Japan[14] and the Psychiatric Genomics Consortium cohort[15]. Owing to the sizeable computational gains, we introduce a heuristic re-sampling algorithm, called the leapfrog re-sampler (LRS), which can confidently extrapolate GWAS results to larger sample sizes, both at a locus or genome-wide scale. Our underlying methodology also leads to several desirable high-utility properties. We release a web tool available to the wider public to conduct customized isGWAS at www.optima-isgwas.com.

Results

Genome-wide association testing from sufficient statistics

isGWAS assumes disease-variant associations can be evaluated via a logistic-link function and, similar to widely used methods[7], [8], uses a Firth adjusted maximum likelihood procedure and Newton-Raphson solver to estimate genetic effects, standard errors and association *p*-value[7], [8], [16]. isGWAS' notable advance is based on the insight that the Newton-Raphson procedure can be simplified so that: (a) elements of the Fisher information matrix and score function vector are collapsed by taking expectation over the empirical or *a priori* distribution of a genetic variant; and (b) sufficient statistics – a specific type of summary data - are used as input variables in the score function (see **Methods** for details). We provide several options to initialise the Newton-Raphson algorithm[17], [18] that improve computational performance and reduce analysis time (**Supplementary Information**). In brief, let y_i denote disease status for the i-th individual and $g_{ij,M}$ denote the j-th genetic variant under model M (e.g., additive, recessive or dominant). The sufficient statistic triple used by isGWAS is:

$$\left\{T_{1j} = \sum_{i=1}^{N_j} y_i = N_j^*, \qquad T_{2j} = \sum_{i=1}^{N_j} y_i\, g_{ij,M}, \qquad T_{3j} = \sum_{i=1}^{N_j} g_{ij,M}\right\},$$

where T_{1j} is the total number of cases for variant j, T_{2j} is the covariance between the outcome y and genotype g for variant j under model M, and T_{3j} is the minor allele count for variant j under model M. For each variant, data can be provided as either: the sufficient statistic triple $\{T_{1j}, T_{2j}, T_{3j}\}$ plus sample size N_j (necessary for computing standard errors) or separately, on assuming Hardy-Weinberg equilibrium (HWE), as $\{N_j, N_j^*, MAF_{j,M}, MAF_{j,M}^*\}$. Default GWAS analyses assume HWE, making input data widely available[13]–[15], [19]–[21] for researchers to perform isGWAS, replicate or further expand on classical GWAS analyses (**Methods**). If MAFs for cases and controls are supplied, isGWAS will automatically convert to the pair

$\{MAF_{j,M}, MAF^*_{j,M}\}$ (**Methods**). After convergence, which is guaranteed for most scenarios by a re-initialization approach (empirically all scenarios converged using isGWAS-Firth), the estimated genetic effect parameter $\hat{\beta}_{j,M}$ and standard error $SE(\hat{\beta}_{j,M})$ are used to construct Wald *p*-values (**Methods**). Additional options include a sample-level likelihood ratio-test or *p*-values computed using sandwich-robust standard errors (**Supplementary Methods**). A simplified illustration highlighting differences and computational advantages of isGWAS against ILD-based genetic association analyses is summarised in **Figure 1**.

isGWAS reliably identifies genetic associations across cohorts and diseases

We benchmarked isGWAS in real-data settings and performed simulation studies to compare isGWAS performance and results relative to several existing individual-level data (ILD)-based approaches. Our assessments broadly fall into two categories: (1) methods which require ILD, i.e., REGENIE[8] , logistic and Firth corrected regression[16], and (2) approaches which do not require ILD directly, i.e., the logistic ad-hoc estimator[17] and Fisher's Exact Test (FET)[22]. We note FET was successfully leveraged for efficient large-scale GWAS analyses recently[23]. Using data from UK Biobank (UKB), we first assessed isGWAS performance against the popular ILD based regression approach REGENIE[8] by deploying both methods for analyses of seven diseases some of which were previously used for establishing GWAS methodology[5], [7], [8]. Second, we evaluated isGWAS's ability to replicate 309 significantly associated variants from the Biobank Japan (BBJ) meta-analysis of 30 diseases[24] using only the published sample-summaries, i.e., without access to ILD, per variant and disease pair. Lastly, by considering multiple iterations of nested schizophrenia meta-analysis GWAS[25]–[27] we assessed isGWAS's ability to accurately predict genomic regions and significant novel associations. The isGWAS sample size of the meta-analysis from 2014 was virtually expanded to match the numbers from the larger 2022 GWAS whilst holding constant the allele frequency and prevalence information in 2014 cohort.

Performance against ILD based regression GWAS in UK Biobank

We compared isGWAS results and computational performance against the state-of-the-art ILD-based method REGENIE for seven diseases in UKB: asthma (IC10:J45); atherosclerosis (IC10:I25); colon cancer (IC10:C18); hypertension (IC10:I10); glaucoma (IC10:H40); stroke (IC10:I63); and thyroid gland cancer (IC10:C73). Case-control ratios varied from 1:2 in hypertension to 1:669 in thyroid gland cancer across the diseases (**Supplementary Table 1a**), allowing for the review of performance in near balanced to highly imbalanced case-control settings. To help attenuate the possible influence of confounders when deploying isGWAS, particularly sample relatedness and population structure, we describe and apply additional data quality control (QC) steps before computing the required sample-level sufficient statistics (**Methods**). After additional QC, the total sample size analysed was ~335,000 individuals per disease (**Supplementary Table 1**). This sample was used to perform and contrast analyses in both isGWAS and REGENIE. We review approaches which leverage additional insight from any removed samples in the **Discussion**. For each of the seven diseases, we applied isGWAS (no covariates) and two-step REGENIE with Firth correction (including the covariates age, sex and ethnicity principal components) to ~11 million autosomal variants. Results are presented in **Table 1**, **Figures 2-3, Supplementary Figures 1-15, Supplementary Tables 1-9 and Supplementary Files 1-2**.

Across all seven traits tested we observed close to perfect consistency between REGENIE and isGWAS association results, as illustrated in the mirrored Manhattan and *p-p* plots for asthma (**Figure 2a**) and other diseases (**Supplementary Figures 1-6**). Concordance between

REGENIE and isGWAS is further validated by benchmarking accuracy (**Table 1**) and Pearson correlations between estimated p-values $(cor(p_{isG}, p_R)) > 0.94$ at log_{10} scale (**Supplementary Tables 2-3**). The results are consistent for varying prevalence levels (**Supplementary Figures 1-6**) and are not affected by covariate adjustment (**Supplementary Tables 2-3**). The consistency translated to the regional locus level. This is exemplified by a locus zoom plot of the *FLG2* gene region for asthma (**Figure 2b**) where isGWAS not only nominated the identical GWAS lead variants but also largely recapitulated the overall association pattern identified through REGENIE. This observation is consistent across the lead independent loci from the asthma GWAS (**Supplementary Figure 7**) and translates to all other diseases studied (**Supplementary Figure 8**). For a comprehensive numerical comparison of association results, we took REGENIE derived p-values as the ground truth, retaining all SNPs with $p < 0.01$ and setting the true positive threshold as $p < 5 \times 10^{-8}$ (excluding stroke (**Supplementary Figure 4**) which did not yield any significant associations). We computed the accuracy, false positive (FPR), true positive (TPR) and false discovery rates (FDR) of isGWAS (**Figure 3a** and **Table 1**). Accuracy of isGWAS was $\geq 99.98\%$ for each disease, highlighting excellent overall correspondence between methods. The FPR was low, i.e., $FPR \lesssim 10^{-5}$, and TPR was generally good at $> 88\%$ - excluding hypertension which had a $TPR = 0.63$. The FDR was below $\leq 5\%$ for each disease, revealing that the positive predictive value of isGWAS was greater than 95%.

Importantly, isGWAS and REGENIE results differed for two broad categories: (i) the estimation of genetic effect sizes; and (ii) computational performance. When non-confounding covariates are excluded (β_{no_cov}) or included (β_{cov}) in a model, previous and extensive investigations of effect size estimates in logistic regression deduce that $|\beta_{no\_cov}| \leq |\beta_{cov}|$, i.e., regression estimates are smaller in magnitude when excluding covariates but the null-hypothesis of no association is maintained[28], [29]. Overall, we replicate these results in our analyses. isGWAS computed effect sizes are smaller in absolute value, but largely concordant with covariate-adjusted REGENIE. Moreover, we fail to reject the null hypothesis for the same variants almost always between methods - suggesting that the isGWAS QC helps attenuate issues of population confounding (**Figures 2d-e, Supplementary Figures 1d-e – 6d-e, 9, 10 and 11**). An investigation of the performance of isGWAS without removing related individuals highlights potential expansion of isGWAS beyond the recommended QC (**Methods, Supplementary Information**), but further investigation – possibly leveraging the re-sampling potential of isGWAS - is required on the reliability of isGWAS in family-based cohorts and ethnically diverse populations (**Supplementary Figures 12-13, Supplementary Tables 5-7)**. In our full-QC analyses, all estimated effects between isGWAS and REGENIE were observed to be in the same direction, and the correlation between estimates was on average $cor(\beta_{isG}, \beta_{REGENIE}) \approx 0.7$ (**Supplementary Tables 1-3**). The relative drop in the correlation between effect estimates (≈ 0.7) and p-values ($\gtrsim 0.94$) is anticipated[28] and can be explained on noting that, across all diseases more precise effect estimates (i.e., those with smaller standard errors) have stronger concordance between approaches (**Figure 2e and Supplementary Figures 1d-6d**). Overall, we found that at least 98% of isGWAS and REGENIE confidence intervals (CI) overlap, (**Supplementary Table 4**). When effect estimates are viewed as a function of MAF, the absolute value of REGENIE-derived estimates seemingly increase (along with standard errors) as MAF decreases across all scenarios. This contrasts with isGWAS where the relationship between MAF and effect size is less clear: fewer variants with low MAF are associated with relatively larger effect sizes. However, the correspondingly narrower standard errors guarantee the same significance p-values as REGENIE. The isGWAS derived distribution of effect sizes is consistent with the hypothesis of a flattened heritability distribution under negative selection[30]. Genomic inflation computed from isGWAS results across all analyses was on average ≈ 1.07 and ranged between (0.94, 1.26) which was similar to REGENIE with average ≈ 1.1 and range (1.01,1.3) (**Supplementary Figure 14** and **Supplementary Table 8**). We deploy isGWAS with genotype imputation in our primary analyses and, as secondary sensitivity analyses, without imputation. Our investigation reveals some surprising results. Imputation occasionally led to changes in MAF between cases and controls such that estimated genetic effects switched sign (i.e., effect direction) relative to results computed from non-imputed data (**Supplementary Figure 15, Supplementary Table 3, Supplementary Files 2-3**). The approach might be used to efficiently flag ambiguous significant results in analyses that are the result of the missing values imputation strategy (mean-imputed in the case of REGENIE).

Finally, the computational gains of isGWAS relative to REGENIE Step 2 are striking: a full genome-wide association assessment for each disease took approximately 4 minutes using isGWAS and, on average over different prevalence, this is around 1,300 times faster than a like-for-like assessment using REGENIE Step-2 (**Figure 7, Supplementary Table 9, and Supplementary File 4**).

Replicating significant associations in Biobank Japan analyses

Using only publicly available summary information from Biobank Japan (BBJ), i.e., without access to ILD, we looked to compare and replicate BBJ GWAS results across 42 diseases[24]. We considered 309 variants that were identified in [24] as genome-wide significant ($p < 5 \times 10^{-8}$) across 30 of the 42 diseases. Our results reveal very close alignment between isGWAS computed associations and those of [24] - correlation between p-values at log_{10} scale was $cor(p_{isG}, p_{BBJ}) = 0.98$ with 92.2% of isGWAS computed genetic effects within the 95% CI of the original study (**Supplementary Figure 16).** Using the published study-level results[24] as the ground truth, isGWAS demonstrated good sensitivity and specificity (**Supplementary Figure 17**). We alternatively assessed performance when setting more stringent significance thresholds - returning near identical conclusions when classifying variants at $p < 9.58 \times 10^{-9}$ (used in the original publication). Results for X-chromosome variants in males and females were similarly concordant (results not presented).

isGWAS model validation using simulations

We generated simulated datasets to assess performance of isGWAS - with and without Firth correction - against a variety of classical methods which either: (a) do not require ILD, the logistic ad-hoc estimator[17] and Fisher's Exact Test[22]; or (b) require ILD, logistic and Firth corrected regression[7], [8], [16]. We perform two simulation studies (**Figure 4**, **Supplementary Figure 18, Supplementary Information)**. isGWAS-Firth outperformed all other approaches in terms of either computational cost or robustness of results over the range of scenarios considered. It is well documented that computational performance is reduced when using Firth's bias correction in ILD regression analyses[7], [8], [16], we discover, however, that no-ILD isGWAS-Firth regression has significantly improved performance relative to uncorrected isGWAS (**Supplementary Table 21**). As anticipated[23], when disease prevalence is rare (i.e., $\pi \leq 0.01$) parameter estimates computed using non-Firth corrected ILD regression were unreliable. The MSE and distribution of parameters estimated via ILD logistic regression were often orders of magnitude poorer than other methods (**Figure 4a-c**). **Figure 4f-h** highlights the chronological evolution of no-ILD *p*-value estimates, from Sasieni's logistic ad-hoc estimator (1997)[17], Fisher's Exact Test (1922)[22] to isGWAS-Firth, illustrating improvements in estimation via successive approaches. See **Supplementary Information** for detailed results review.

Leapfrog re-sampling: using isGWAS to extrapolate variant association results to larger sample sizes

When ILD are available, the computational benefits of isGWAS make it possible to deploy resampling approaches to estimate empirical effect sizes, *p*-values and corresponding confidence intervals, previously considered computationally daunting in GWAS[31], [32]. We extend the idea by introducing a heuristic leapfrog re-sampling (LRS) algorithm to help forecast future results in larger hypothetical GWAS sample sizes (**Methods**). The LRS is summarised in three key steps: (1) specify a target sample size along with the number and size of sub-samples to be generated; (2) (leapfrog-step) compute sufficient statistics in the sub-samples and re-scale the estimated number of cases and controls to match the larger target sample size; and (3) deploy isGWAS in each leapfrog sample to recover a distribution of association *p*-values over the collection of sub-samples. In our testing of the LRS, we use the median *p*-value as a generally robust estimate of a target p-value (weighted or distribution-based summaries can alternatively be considered). Thus, the LRS leverages variation in both genotype and disease status between individuals in the current sample to help predict updates of parameters after adding new samples. Despite perceived similarities, traditional GWA power calculators[33] and the isGWAS-LRS are different. isGWAS-LRS does not require input of case-control ratios, heritability (i.e., beta estimates) or type-I error rates. Instead, multiple regression analyses are combined to forecast and test parameter estimates in expanding sample sizes.

We run the leapfrog re-sampler in both simulation and real-data settings, informed by the seven studied diseases in UKB (**Methods and Supplementary Information**). We evaluate performance over a range of initialisations, starting from a 10% increase to a maximum of 100% (i.e., 2-fold) increase in GWAS sample size relative to the current actual UKB sample size. Results are presented in **Figure 5**, **Supplementary Tables 10-11**. As is standard, we assume a true positive association of $p<5\times10-8$ in the target sample. Our results in simulated scenarios (**Figure 5f**) reveal that: when doubling sample size from $N_{current} = 276{,}204$ to a maximum $N_{target} = 552{,}408$, the accuracy and TPR progressively dropped for subsequent increases in the target sample size, but values for each measurement were typically ≥80% across the range. Our real world LRS analyses of UK Biobank data replicate and further elucidate performance across the six of the seven diseases (**Figure 5a-e**). Using a subsample size of $N_{current} \approx 135{,}000$ we increased target sample size up to $N_{target} \approx 270{,}000$, taken as the maximum observed sample size we could benchmark against. For all choices of target sample size, and across each disease, we observe high accuracy rates (≥95%). However, the TPR was sensitive to disease prevalence, reducing monotonically as the target sample size increased. Broadly, TPR remained reasonable ($\gtrsim 60\%$) up to a 2-fold increase in sample size, except for the very rare (case-control ratio of 1:669) thyroid gland cancer. This is due to fewer significant variants being included in the assessment as a result of lower percentage of heritability explained, which can artificially reduce the TPR for each new locus with relatively high odds ratios. Naturally, TPR reduces as a function of decreasing disease prevalence, as re-sampling from fewer cases can increase the variability in MAFs and thus isGWAS forecasting. We note that our theoretical sub-sampling approach had better predictive capabilities, owing to the prevalence preserving sampling strategy taken (**Methods and Supplementary Figure 19**).

We also assessed isGWAS's ability to extrapolate results when ILD were not available, using a highly constrained version of the leapfrog re-sampler (**Supplementary Information**). In this scenario, MAF and disease prevalence per variant are fixed, computed from the maximum current sample (i.e., without sub-sampling), and the number of cases and controls are proportionately increased to match the target sample size. We did this for two GWAS of schizophrenia: (a) 2014 analyses with up to $N = 77{,}096\ (\text{cases} = 33{,}640, \text{controls} = 43456)$ European ancestry individuals[25]; and (b) the larger (and future) 2022 analyses with up to $N = 130{,}644$ $(\text{cases} = 53386, \text{controls} = 77258)$ European ancestry individuals[27]. We treat the 2014 study as the current sample size and the 2022 sample size as the target future state, which we benchmark predictive performance against. The studies were selected because, for each variant j, necessary data to run isGWAS, i.e., $\{N_j, N_j^*, MAF_j, MAF_j^*\}$, were made publicly available. Note these data are pooled estimates, computed across all European cohorts. Despite not accessing ILD, our results reveal reasonable concordance between isGWAS 2014 extrapolated results and the published analyses of 2022 (**Figure 6**, **Supplementary Figures 20-21, Supplementary Tables 13-16**). Like our Biobank Japan analyses, we also used a more stringent significance threshold $(p < 10^{-10})$ to help attenuate false positives, observing improved overall performance by recovering a good TPR $\geq 70\%$ (**Supplementary Table 13**). We do not report FDR as these cannot be accurately computed when filtering results based on a p-value inclusion/exclusion threshold. Of the overlapping 608 clumped variants considered, isGWAS-LRS identified 136 associations that were not yet deemed GWAS significant (i.e,. $p > 5e - 8$) in the 2014 study but later identified as significant in the 2022 study. Moreover, of the 436 significant associations predicted by isGWAS, 75% overlap with observed significant associations in 2022. isGWAS predicted an additional 74 associations as significant that were not significant in 2022 – of those 52 were near the significance threshold with $p < 9e - 07$. There were 121 variants not correctly predicted by the 2014 cohort. This could be due to increased ethnical and relatedness heterogeneity in the 2022 cohort that was not present in the 2014 analysis.

Computational performance and convergence details

isGWAS is an iterative algorithm whose convergence (i.e., ability to estimate model parameters) depends on several tuning parameters (**Methods**). Using default parameter settings, isGWAS-Firth converged in all real-data and simulated scenarios tested (**Figure 4e and**

Supplementary Tables 17, 19-20). Convergence was achieved in around 0.001 seconds per variant (**Supplementary Table 21**) on a 2.4 GHz 8-Core Intel Core i9 processor. The non-Firth corrected isGWAS algorithm may require more iterations, particularly for diseases with lower prevalence (e.g., case:control ratio of 1:94 and lower) which included scenarios where convergence was not achieved (**Figure 4e, Supplementary Tables 17-20, Supplementary Information**).

When distributed over 32 CPU cores on a high-performance cluster, Firth-corrected isGWAS analysed a single disease from UK Biobank across ~11 million SNPs and for ~335,000 individuals in ~4 minutes (**Figure 7, Supplementary Tables 9 and 18**). This means that isGWAS-Firth can perform around 1,500 disease GWAS for every one GWAS performed using an alternative methodology. The same analysis with a small number of CPU cores was completed in tens of minutes using isGWAS-Firth (**Figure 7**). Further computational gains at larger sample sizes will likely be achieved as ILD methods can scale poorly with sample size, whereas isGWAS has near fixed computational cost at any size. As isGWAS currently computes associations for each variant independently, additional improvements such as parallelisation are possible. Full details are available in **Supplementary File 4.**

Discussion

In this study, we developed isGWAS, an efficient, biobank-scalable method for genetic association testing which can: (a) compute regression parameters and test for a variant-disease association in real-time (i.e., approximately one millisecond) for any sample size; (b) bypass the need to run large-scale GWAS using high-performance computing facilities owing to ultra-low system resource demands (i.e., runtime and memory); and (c) infer GWAS results from virtually enlarged sample sizes using a novel re-sampling procedure. The isGWAS algorithm design allows analyses to be run without the need to hold or access individual-level data (ILD) directly, thereby providing a single methodological framework to utilise a wide range of data sources such as published summary-level data from biobanks and repositories.

isGWAS draws inspiration from classical methodologies to overcome significant computational bottlenecks associated with massive-scale analyses. The practical simplicity and quick runtime of classical approaches have seen them deployed in a recent large-scale analysis[23]. Rather than using ILD, as contemporary GWAS regression analyses do, isGWAS distils the required input data down to sufficient statistics – a low-dimensional summary of ILD that captures all necessary information required to compute a genetic-disease association model parameter. In combination with modifications to the Newton-Raphson procedure, used to estimate model parameters in a logistic regression, our use of sufficient statistics dramatically reduces the computational time for disease association testing relative to existing methods. Achieving up to a 1,500-fold improvement in computational runtime, when benchmarked against a state-of-the-art GWAS tool, isGWAS reduced time to genome-wide insight from several days down to ~4 minutes. Thereby unlocking potential for massive scale exploration of genetic-disease associations in real-time and making feasible the routine assessment of thousands of disease endpoints and studies. Computational bottlenecks associated with existing GWAS methodologies are fast approaching. Analyses of resources such as UK Biobank WGS data, the emerging massive cohorts of the Global Biobank Initiative[1], and Our Future Health[34] are expected to push current GWAS tools to their system resource limits with significant associated time-to-insight penalties. Conversely, with no computational sensitivity to sample size, as the number of variants assessed and sample sizes continue to increase, the relative savings and benefits of isGWAS can be expected to grow.

To attenuate possible issues of confounding and population stratification, we propose that additional QC-steps are performed before computing sufficient statistics for isGWAS. In our analyses, these steps reduced our UK Biobank sample size from ~408k individuals (used in the original testing of REGENIE[8]) down to a more homogeneous sample of ~335k individuals used to generate and compare results from isGWAS and REGENIE. Our results reveal an often-striking concordance between approaches genome-wide as well as at the regional locus level. The reduction in sample size was compensated by the isGWAS leapfrog re-sampler (LRS), which we demonstrate efficiently helped extrapolate GWA results onto larger sample sizes (up to 2-times). While we note sensitivity of an LRS extrapolation to disease prevalence, across the range considered the TPR and FPR were well calibrated to at least a 1.5-fold increase in sample size. The LRS might therefore be leveraged to aid GWAS cohort design, for example to quantify the potential benefit of sampling more participants with a disease of interest against cost. In our analyses of the PGC Schizophrenia cohort, we deployed a highly restricted (i.e., no ILD) version of the LRS: forecasting results from a smaller 2014 cohort[25] onto a sample size that matched a future 2022 study[27]. Despite no guarantees of sufficiency, isGWAS LRS identified 75% of significant variants that were later identified in the larger 2022 cohort (almost double the size) while maintaining a low FDR. Unlike extrapolation via the ILD leapfrog re-sampler, this naïve extrapolation does not account for differences in the MAF of cases and controls between 2014 and 2022 data. Regardless, the above findings highlight potential for isGWAS to furnish reasonable forecasts of future results without accessing ILD directly. Our recent predictions from FinnGen consortium data[18], [35] provide confidence that the isGWAS algorithm is applicable also to multi-ethnic GWAS through analyzing each ethnicity separately and combining results in a meta-analysis, as it is common practice[36].

Beyond, isGWAS can be applied to help address routinely asked questions about future scenarios and evaluate enrichment contribution of biobanks to disease-specific associations[35] or to protein-specific variant associations[18], particularly in the rare spectrum. isGWAS-Firth provides a timely, rapid regression-based analysis of common, rare and ultra-rare variants. Unlike ILD-based analyses, where Firth's correction significantly increases computational time[7], [8], there is no computational penalty when using Firth's correction in the isGWAS framework - in fact, we observe improved computational performance. The advantage of considerable improvements in computational runtime is that it allows for the introduction of forecasting, re-sampling and other non-parametric techniques - the LRS being one example. These might widen robust association testing strategies as well as provide new avenues to tackle confounding or population sub-structure. For now, we envisage the possibility that the wider human genetics community routinely compute and make available the sufficient statistics, i.e. MAF in the cases and the cohort, and the corresponding sample sizes per variant, toward a publicly available, privacy compliant, data asset. In addition to avoiding the need for expensive high-performance computing facilities and memory intensive data storage, the data asset might enhance meta-analyses and biological insight, improve equitable access, and enable faster collaborations between teams and help bridge financial and resource gaps between institutions and research groups internationally.

Methods

Disease SNP association model

Let $S_M \in \{1,2\}$ be the maximum number of copies of the effect allele for an individual under model $M \in \{A, D, R\}$, where A denotes an additive model, R a recessive model and D a dominant model, i.e.,

$$S_M = \begin{cases} 2, & M = A, \\ 1, & M = D \cap R. \end{cases}$$

Furthermore let,

$$MAF_{j,M} = MAF_j \mid M$$

$$MAF_{j,M}^* = MAF_j^* \mid M$$

where, for a given model M, $MAF_{j,M}$ is the minor allele frequency for variant j in the sample, ancestry, or population and $MAF_{j,M}^*$ the minor allele frequency in the cases. We let Y denote disease status and G_j the jth genotype in the sample. For convenience we write $G_{j,M} = G_j | M$. It is assumed that the outcome model for Y, conditional on G_j, is given by:

$$\mathbb{E}[Y \mid G_j, M] = h^{-1}\left(\alpha_{j,M} + \beta_{j,M} G_{j,M}\right), \qquad j = 1,2, \dots, Q,$$

where, conditional on model M, the pair $\{\alpha_{j,M}\ \beta_{j,M}\}$ denote the intercept and genotype effect and h is a function linking the outcome to genotype $G_{j,M}$ for all $j = 1,2, \dots, Q$ genetic variants considered. In deriving the isGWAS estimation procedure we assume that h is the logit function, i.e.,

$$\pi_{Y|G_{j,M}} = P\left(Y = 1 \mid G_j, M\right) = \frac{e^{\left(\alpha_{j,M} + \beta_{j,M} G_{j,M}\right)}}{1 + e^{\left(\alpha_{j,M} + \beta_{j,M} G_{j,M}\right)}}.$$

The isGWAS methodology can, however, be broadened to other link functions and outcome types. We take $\left\{\hat{\alpha}_{j,M}, \hat{\beta}_{j,M}, \hat{\sigma}_{\alpha_{j,M}}, \hat{\sigma}_{\beta_{j,M}}\right\}$ to be sample based estimates of the intercept $\alpha_{j,M}$ and coefficient $\beta_{j,M}$, and their associated standard errors $\left\{\hat{\sigma}_{\alpha_{j,M}}, \hat{\sigma}_{\beta_{j,M}}\right\}$. We allow the genetic effect

$\beta_{j,M}$ and genotype G_j (or an observation thereof g_j) to be analyzed on either (i) the standardized scale or (ii) non standardized scale. To accommodate this, we introduce the variable s_m^*, so that:

$$\{\beta_{j,s_m^*}, g_{j,s_m^*}\} = \left\{(1+(\sigma_{g_i}-1)s_m^*)\beta_j, {}^{g_j}\!\Big/\!\left(1+(\sigma_{g_i}-1)s_m^*\right)\right\}.$$

Hence, when $s_m^* = 0$ analyses are performed on the non-standardized scale and $s_m^* = 1$ on the standardized scale. Default analyses assume the genetic effect is assessed on the non-standardized scale, i.e., $s_m^* = 0$. Note that, while p-values are generally invariant to the choice of effect scale s_m^*, betas and standard errors are dependent on the specification of s_m^*.

Sample-Level Newton-Raphson (SaLN-R) algorithm

Here we detail the isGWAS procedure for computing summary statistics $\left\{\hat{\alpha}_{j,M}, \hat{\beta}_{j,M,s_m^*}, \hat{\sigma}_{\alpha_{j,M}}, \hat{\sigma}_{\beta_{j,M,s_m^*}}\right\}$ using only four data points,

$$\left\{N_j, \sum_{i=1}^{N_j} y_i = N_j^*, \sum_{i=1}^{N_j} y_i g_{ij,M}, \sum_{i=1}^{N_j} g_{ij,M} = n_{j1} + 2n_{j2} I(M = A)\right\}$$

or, as we show, the quadruple

$$\{N_j, N_j^*, MAF_{j,M}, MAF_{j,M}^*\},$$

where N_j denotes the study or population sample size and N_j^* the number of cases in the sample, see ad-hoc estimator (Supplementary Information) for definitions of $n_{j.}$. Note that we have allowed the sample size N and number of cases N^* to vary by genotype j, this is useful when emulating results from GWAS. This is because the number of individuals analyzed in GWAS can vary by genotype owing to (e.g.,) quality of imputation or available data per variant and participant in a study. Ideally the sample size and number of cases would not vary by genotype and when using isGWAS to forecast GWAS results, users do not need not vary N_j, i.e,.

$$N_j = N \quad \text{and} \quad N_j^* = N^*, \qquad j = 1,2,\dots,Q.$$

Given a vector of observed data $\{\boldsymbol{y}, \boldsymbol{g}_{j,s_m^*}\}$, where $\boldsymbol{y} = \{y_1, y_2, \dots, y_{N_j}\}$ and $\boldsymbol{g}_{j,s_m^*} = \{g_{1j,}, g_{2j}, \dots, g_{N_j}\}$, estimates of model parameters are typically derived by maximizing the log-likelihood function

$$L\left(\boldsymbol{\beta}_{j,M,s_m^*}\right) = \sum_{i=1}^{N_j} \log P\left(y_i \,|\boldsymbol{\beta}_{j,M,s_m^*}, g_{ij,M,s_m^*}\right),$$

which is equivalent to identifying parameter values $\boldsymbol{\beta}_{j,M,s_m^*} = \{\alpha_{j,M}, \beta_{j,M,s_m^*}\}$ which satisfy:

$$\frac{\partial L\left(\boldsymbol{\beta}_{j,M,s_m^*}\right)}{\partial \boldsymbol{\beta}_{j,M,s_m^*}} = V\left(\boldsymbol{\beta}_{j,M,s_m^*}, I_F\right) = 0,$$

where $V\left(\boldsymbol{\beta}_{j,M,s_m^*}, I_F\right)$ denotes the logistic score function, i.e.,

$$V\left(\boldsymbol{\beta}_{j,M,s_m^*}, I_F\right) = \tilde{\boldsymbol{g}}^T_{j,M,s_m^*}\left(\boldsymbol{y} - \pi_{\boldsymbol{y}|\tilde{\boldsymbol{g}}_{j,M,s_m^*}}\right) + I_F \mathrm{K}\left(\boldsymbol{\beta}_{j,M,s_m^*}\right) = 0,$$

with I_F denoting an indicator function used to highlight that a Firth modified version of the score function has been used. For ease of mathematical presentation initially, we detail the Firth adjusted SaLN-R algorithm later, i.e., we set $I_F = 0$ in this section. Additionally, to improve succinctness of notation, we drop the use of the parameter s_m^* - reintroducing where necessary - and set $\boldsymbol{\beta}_{j,M} = \{\alpha_{j,M}, \beta_{j,M}\}$ and $\tilde{\boldsymbol{g}}_{j,M} = (1, g_{j,M})$ above, so that $\boldsymbol{\beta}_{j,M}\tilde{\boldsymbol{g}}^T_{j,M} = \alpha_{j,M} + \beta_{j,M} g_{j,M}$. We compute candidate solutions to by expanding $V\left(\boldsymbol{\beta}_{j,M}\right)$ as a Taylor series about a value $\boldsymbol{\beta}_{j0,M}$ and up to second order, i.e., using the Newton-Raphson (N-R) method:

$$V\left(\boldsymbol{\beta}_{j,M}\right) = V\left(\boldsymbol{\beta}_{j0,M}\right) + \left.\frac{\partial V\left(\boldsymbol{\beta}_{j,M}\right)}{\partial \boldsymbol{\beta}_{j,M}}\right|_{\boldsymbol{\beta}_{j0,M}} \left(\boldsymbol{\beta}_{j,M} - \boldsymbol{\beta}_{j0,M}\right) + \mathcal{O}\left(\left(\boldsymbol{\beta}_{j,M} - \boldsymbol{\beta}_{j0,M}\right)^2\right),$$

Which is re-written as

$$\boldsymbol{\beta}_{j,M} = \boldsymbol{\beta}_{j0,M} + \mathcal{I}^{-1}\left(\boldsymbol{\beta}_{j0,M}\right) V\left(\boldsymbol{\beta}_{j0,M}\right) + \mathcal{O}\left(\left(\boldsymbol{\beta}_{j,M} - \boldsymbol{\beta}_{j0,M}\right)^2\right)$$

and generalized into an N-R iterative algorithm:

$$\boldsymbol{\beta}_{j(k+1),M,s_m^*} = \boldsymbol{\beta}_{jk,M,s_m^*} + \mathcal{I}^{-1}\left(\boldsymbol{\beta}_{jk,M,s_m^*}\right) V\left(\boldsymbol{\beta}_{jk,M,s_m^*}\right) + \mathcal{O}\left(\left(\boldsymbol{\beta}_{j(k+1),M,s_m^*} - \boldsymbol{\beta}_{jk,M,s_m^*}\right)^2\right), \quad k = 0,1, \dots, K,$$

where we have re-introduced s_m^* to highlight that the algorithm is dependent on the choice of effect scale. The variable $\mathcal{I}^{-1}$ denotes the inverse Fisher Information matrix, where

$$\mathcal{I}(\boldsymbol{\beta}_{jk,M}) = -\frac{\partial V(\boldsymbol{\beta}_{j,M})}{\partial \boldsymbol{\beta}_{j,M}}\bigg|_{\boldsymbol{\beta}_{j,M}=\boldsymbol{\beta}_{jk,M}}$$

$$= \tilde{\boldsymbol{g}}_{j,M}{}^{T}\operatorname{diag}\left(\pi_{y_i|g_{j,M}}\left(1-\pi_{y_i|g_{jk,M}}\right)\right)\tilde{\boldsymbol{g}}_{j,M}\Big|_{\boldsymbol{\beta}_{j,M}=\boldsymbol{\beta}_{jk,M}}$$

$$= \begin{pmatrix} \sum_{i=1}^{N_j} \pi_{y_i|g_{ij,M}}\left(1-\pi_{y_i|g_{ij,M}}\right) & \sum_{i=1}^{N_j} \pi_{y_i|g_{ij,M}}\left(1-\pi_{y_i|g_{ij,M}}\right) g_{ij,M} \\ \sum_{i=1}^{N_j} \pi_{y_i|g_{ij,M}}\left(1-\pi_{y_i|g_{ij,M}}\right) g_{ij,M} & \sum_{i=1}^{N_j} \pi_{y_i|g_{ij,M}}\left(1-\pi_{y_i|g_{ij,M}}\right) g_{ij,M}^2 \end{pmatrix}\Bigg|_{\boldsymbol{\beta}_{j,M}=\boldsymbol{\beta}_{jk,M}}$$

and the score function is given by

$$V(\boldsymbol{\beta}_{jk,M}) = \tilde{\boldsymbol{g}}_{j,M}{}^{T}\left(\boldsymbol{y}-\pi_{\boldsymbol{y}|\boldsymbol{g}_{j,M}}\right)\Big|_{\boldsymbol{\beta}_{,Mj}=\boldsymbol{\beta}_{jk,M}}$$

$$= \begin{pmatrix} \sum_{i=1}^{N_j}\left(y_i-\pi_{y_i|g_{ij,M}}\right) \\ \sum_{i=1}^{N_j}\left(y_i-\pi_{y_i|g_{ij,M}}\right) g_{ij,M} \end{pmatrix}\Bigg|_{\boldsymbol{\beta}_{j,M}=\boldsymbol{\beta}_{jk,M}}$$

Both $\mathcal{I}(\boldsymbol{\beta}_{jk,M})$ and $V(\boldsymbol{\beta}_{jk,M})$ above require individual-level data to compute their values. isGWAS aims to estimate values for these variables using sample-level information only, thereby avoiding the immediate need for individual data. To achieve this, we approximate both the Fisher Information matrix and the Score function via the pair $\{\mathcal{I}_{\mathbb{e}}(\boldsymbol{\beta}_{jk,M}), V_{\mathbb{e}}(\boldsymbol{\beta}_{jk,M})\}$, where:

$$\mathcal{I}_{\mathbb{e}}(\boldsymbol{\beta}_{jk,M}) = N_j\begin{pmatrix} \mathbb{E}_{g_{j,M}}\left[\pi_{y|g_{j,M}}\left(1-\pi_{y|g_{j,M}}\right);\ \boldsymbol{\beta}_{jk,M}\right] & \mathbb{E}_{g_{j,M}}\left[\pi_{y|g_{j,M}}\left(1-\pi_{y|g_{j,M}}\right)g_{j,M};\ \boldsymbol{\beta}_{jk,M}\right] \\ \mathbb{E}_{g_{j,M}}\left[\pi_{y|g_{j,M}}\left(1-\pi_{y|g_{j,M}}\right)g_{j,M};\ \boldsymbol{\beta}_{jk,M}\right] & \mathbb{E}_{g_{j,M}}\left[\pi_{y|g_{j,M}}\left(1-\pi_{y|g_{j,M}}\right)g_{j,M}^2;\ \boldsymbol{\beta}_{jk,M}\right] \end{pmatrix}$$

and

$$V_{\mathbb{e}}(\boldsymbol{\beta}_{jk,M}) = \begin{pmatrix} \sum_{i=1}^{N_j} y_i - N_j\,\mathbb{E}_{g_{j,M}}\left[\pi_{y|g_{j,M}};\ \boldsymbol{\beta}_{jk,M}\right] \\ \sum_{i=1}^{N_j} y_i g_{ij,M} - N_j\mathbb{E}_{g_{j,M}}\left[\pi_{y|g_{j,M}}g_{j,M}\ ;\ \boldsymbol{\beta}_{jk,M}\right] \end{pmatrix}$$

$$\approx \begin{pmatrix} N_j^* - N_j\,\mathbb{E}_{g_{j,M}}\left[\pi_{y|g_{j,M}};\,\boldsymbol{\beta}_{jk,M}\right] \\ N_j^*\mathbb{E}_{g_{j,M}}\left[g_{j,M}|\,y=1\right] - N_j\mathbb{E}_{g_{j,M}}\left[\pi_{y|g_{j,M}}g_{j,M}\,;\,\boldsymbol{\beta}_{jk,M}\right]\end{pmatrix}, \quad N_j^* \gg 1.$$

with $\mathbb{E}_{g_{j,M}}[\,\cdot\,;\boldsymbol{\beta}_{jk,M}]$ denoting that expectation is taken with respect to $g_{j,M}$ and evaluated at $\boldsymbol{\beta}_{j,M} = \boldsymbol{\beta}_{jk,M}$. Note that $\{\mathcal{J}_{\mathbb{e}}(\boldsymbol{\beta}_{jk,M}), V_{\mathbb{e}}(\boldsymbol{\beta}_{jk,M})\}$ are motivated by switching from empirical, i.e., sample-based, estimates in $\{\mathcal{J}(\boldsymbol{\beta}_{jk,M}), V(\boldsymbol{\beta}_{jk,M})\}$ to their expected value analogues, which reverses the usual mode of estimation. Sample size N_j is presumed large and thus switching from sample-based to expected values in the N-R algorithm is well motivated. However, when the number of cases N_j^* is 'small', an approximation of $\sum_{i=1}^{N_j} y_i g_{j,M} \approx N_j^*\mathbb{E}_{g_{j,M}}[g_{j,M}|\,y=1]$ becomes weaker and we recommend using the statistic $\sum_{i=1}^{N_j} y_i g_{j,M}$. Values for the elements in $\{\mathcal{J}_{\mathbb{e}}(\boldsymbol{\beta}_{jk,M}), V_{\mathbb{e}}(\boldsymbol{\beta}_{jk,M})\}$ are computed via:

$$\mathbb{E}_{g_{j,M,s_m^*}}\left[\pi_{y|g_{j,M,s_m^*}}\left(1-\pi_{y|g_{j,M,s_m^*}}\right)g_{j,M,s_m^*}^{c}\,;\,\boldsymbol{\beta}_{jk,M,s_m^*}\right] =$$

$$\sum_{l=0}^{s_M}\pi_{y|g_{j,M}={}^{l}/_{w_{s_m^*}};\boldsymbol{\beta}_{jk,M,s_m^*}}\left(1-\pi_{y|g_{j,M}={}^{l}/_{w_{s_m^*}};\boldsymbol{\beta}_{jk,M,s_m^*}}\right)\left(\frac{l}{w_{s_m^*}}\right)^{c}p\left(g_{j,M}={}^{l}/_{w_{s_m^*}}\right)$$

$$= e_{j,M}^{(c,k)}$$

and

$$\mathbb{E}_{g_{j,M,s_m^*}}\left[\pi_{y|g_{j,M,s_m^*}}g_{j,M,s_m^*}^{c}\,|\,\boldsymbol{\beta}_{jk,M,s_m^*}\right] =$$

$$\sum_{l=0}^{s_M}\pi_{y|g_{j,M}={}^{l}/_{w_{s_m^*}};\boldsymbol{\beta}_{jk,M,s_m^*}}\left(\frac{l}{w_{s_m^*}}\right)^{c}p\left(g_{j,M}={}^{l}/_{w_{s_m^*}}\right)$$

$$= \tilde{e}_{j,M}^{(c,k)},$$

where $w_{s_m^*} = \left(1+(\sigma_{g_i}-1)s_m^*\right)$ and the superscript and subscript in $e_M^{(c,k)}$, $\tilde{e}_M^{(c,k)}$ are used to highlight that expectation has been taken conditional on k-th iteration $\beta_{jk,M}$ and under modelling assumption M (and implicitly effect scale s_m^*). Probability mass $p\left(g_{j,M} = {}^{l}/{w_{s_m^*}}\right)$ is either defined a-priori or can be approximated empirically, which we detail later. In combination, therefore, it follows that:

$$\mathcal{J}_{\mathbb{e}}\left(\boldsymbol{\beta}_{jk,M,s_m^*}\right) = N_j \begin{pmatrix} e_{j,M}^{(0,k)} & e_{j,M}^{(1,k)} \\ e_{j,M}^{(1,k)} & e_{j,M}^{(2,k)} \end{pmatrix}$$

and

$$\mathcal{J}_{\mathbb{e}}^{-1}\left(\boldsymbol{\beta}_{jk,M,s_m^*}\right) = \frac{1}{N_j\left(e_{j,M}^{(0,k)} e_{M}^{(2,k)} - \left(e_{j,M}^{(1,k)}\right)^2\right)} \begin{pmatrix} e_{j,M}^{(2,k)} & -e_{j,M}^{(1,k)} \\ -e_{j,M}^{(1,k)} & e_{j,M}^{(0,k)} \end{pmatrix}.$$

Following the same process that led to the above, we re-write the sample-level Score function $V_{\mathbb{e}}\left(\boldsymbol{\beta}_{jk,M,s_m^*}\right)$ as:

$$V_{\mathbb{e}}\left(\boldsymbol{\beta}_{jk,M,s_m^*}\right) = \begin{pmatrix} N_j^* - N_j\, \tilde{e}_{j,M}^{(0,0)} \\ \sum_{i\,:\,y_i=1} g_{ij,M,s_m^*} - N_j \tilde{e}_{j,M}^{(1,0)} \end{pmatrix}$$

$$\approx \begin{pmatrix} N_j^* - N_j\, \tilde{e}_{j,M}^{(0,0)} \\ \dfrac{S_M N_j^* MAF_{j,M}^*}{w_{s_m^*}} - N_j \tilde{e}_{j,M}^{(1,0)} \end{pmatrix}, \quad N_j^* \gg 1,$$

where we have used the following approximation:

$$\sum_{i\,:\,y_i=1} g_{ij,M,s_m^*} \approx N_j^* \mathbb{E}_{g_{j,M}}\left[g_{j,M,s_m^*} \mid y = 1\right] = \frac{S_M N_j^* MAF_{j,M}^*}{w_{s_m^*}}.$$

Hence, replacing the pair $\left\{\mathcal{J}\left(\boldsymbol{\beta}_{jk,M,s_m^*}\right), V\left(\boldsymbol{\beta}_{jk,M,s_m^*}\right)\right\}$ with the sample-level approximations $\left\{\mathcal{J}_{\mathbb{e}}\left(\boldsymbol{\beta}_{jk,M,s_m^*}\right), V_{\mathbb{e}}\left(\boldsymbol{\beta}_{jk,M,s_m^*}\right)\right\}$ we furnish the SaLN-R algorithm:

$$\boldsymbol{\beta}_{j(k+1),M,s_m^*} = \boldsymbol{\beta}_{jk,M,s_m^*} + \mathcal{J}_{\mathbb{e}}^{-1}\left(\boldsymbol{\beta}_{jk,M,s_m^*}\right) V_{\mathbb{e}}\left(\boldsymbol{\beta}_{jk,M,s_m^*}\right)$$

$$= \boldsymbol{\beta}_{jk,M,s_m^*} + \frac{1}{N_j\left(e_{j,M}^{(0,k)} e_{j,M}^{(2,k)} - \left(e_{j,M}^{(1,k)}\right)^2\right)} \begin{pmatrix} e_{j,M}^{(2,k)} & -e_{j,M}^{(1,k)} \\ -e_{j,M}^{(1,k)} & e_{j,M}^{(0,k)} \end{pmatrix} \begin{pmatrix} N_j^* - N_j\, \tilde{e}_{j,M}^{(0,k)} \\ \dfrac{S_M N_j^* MAF_{j,M}^*}{w_{s_m^*}} - N_j \tilde{e}_{j,M}^{(1,k)} \end{pmatrix}.$$

The standard error of the updates, $\hat{\boldsymbol{\sigma}}_{\boldsymbol{\beta}_{j(k+1),M,s_m^*}} = \left\{\hat{\sigma}_{\alpha_{j(k+1),M,s_m^*,}}, \hat{\sigma}_{\beta_{j(k+1),M,s_m^*}}\right\}$, are given by the diagonal of the inverse Fisher information matrix, i.e.,

$$\hat{\sigma}_{\alpha_{j(k+1),M,s_m^*}} = \sqrt{\frac{e_{j,M}^{(2,k)}}{N_j\left(e_{j,M}^{(0,k)}e_{j,M}^{(2,k)} - \left(e_{j,M}^{(1,k)}\right)^2\right)}},$$

$$\hat{\sigma}_{\beta_{j(k+1)},M,s_m^*} = \sqrt{\frac{e_{j,M}^{(0,k)}}{N_j\left(e_{j,M}^{(0,k)}e_{j,M}^{(2,k)} - \left(e_{j,M}^{(1,k)}\right)^2\right)}}.$$

It can be shown from the above that:

$$\boldsymbol{\beta}_{jk,M,s_m^*=1} = \sigma_{g_i}\boldsymbol{\beta}_{jk,M,s_m^*=0},$$

$$\hat{\sigma}_{\beta_{j(k+1)},M,s_m^*=1} = \sigma_{g_i}\hat{\sigma}_{\beta_{j(k+1)},M,s_m^*=0}.$$

We set $s_m^* = 0$ to compute values for the pair $\left\{\hat{\boldsymbol{\beta}}_{jk,M,0}, \hat{\sigma}_{\beta_{j(k+1)},M,0}\right\}$ and use the above identities to return parameter estimates on the standardized scale $s_m^* = 1$. The data required to run the SaLN-R algorithm are:

$$\left\{N_j, N_j^*, \sum_{i=1}^{N_j} y_i g_{ij,M}\,, p\left(g_{j,M} = 1\right)\right\} \bigcup \begin{Bmatrix} \emptyset, & S_M = 1, \\ p\left(g_{j,M} = 2\right), & S_M = 2. \end{Bmatrix}$$

We use the approximations

$$p\left(g_{j,M} = 1\right) \approx \frac{n_{j1}}{N_j} \underset{HWE}{\xrightarrow{N_j \gg 1}} \begin{cases} MAF_{j,M}, & S_M = 1, \\ 2MAF_{j,M}\left(1 - MAF_{j,M}\right), & S_M = 2, \end{cases}$$

$$p\left(g_{j,M} = 2\right) \approx \frac{n_{j2}}{N_j} \underset{HWE}{\xrightarrow{N_j \gg 1}} MAF_{j,M}^2,$$

where $\underset{HWE}{\longrightarrow}$ is used to denote under Hardy-Weinberg equilibrium.

The SaLN-R algorithm is extended to include Firth's penalty function (see **Supplementary Information for more details**):

$$\boldsymbol{\beta}_{j(k+1),M,s_m^*} = \boldsymbol{\beta}_{jk,M,s_m^*} + \frac{1}{\left(\left(e_M^{(0,k)} - \frac{\partial_{\alpha_{j,M}}\mathrm{K}_{\mathbb{e}}^{(\alpha_{j,M})}}{N_j}\right)\left(e_M^{(2,k)} - \frac{\partial_{\beta_{j,M}}\mathrm{K}_{\mathbb{e}}^{(\beta_{j,M})}}{N_j}\right) - \left(e_M^{(1,k)} - \frac{\partial_{\beta_{j,M}}\mathrm{K}_{\mathbb{e}}^{(\alpha_{j,M})}}{N_j}\right)^2\right)}$$

$$\times \begin{pmatrix} e_M^{(2,k)} - \dfrac{\partial_{\alpha_{j,M}} \mathrm{K}_{\mathbb{e}}^{(\alpha_{j,M})}}{N_j} & -e_M^{(1,k)} + \dfrac{\partial_{\beta_{j,M}} \mathrm{K}_{\mathbb{e}}^{(\alpha_{j,M})}}{N_j} \\ -e_M^{(1,k)} + \dfrac{\partial_{\beta_{j,M}} \mathrm{K}_{\mathbb{e}}^{(\alpha_{j,M})}}{N_j} & e_M^{(0,k)} - \dfrac{\partial_{\beta_{j,M}} \mathrm{K}_{\mathbb{e}}^{(\beta_{j,M})}}{N_j} \end{pmatrix} \begin{pmatrix} \pi_j^* - \tilde{e}_M^{(0,k)} + \mathrm{K}_{\mathbb{e}}^{(\alpha_{j,M})} \Big/ N_j \\ \dfrac{S_M \pi_j^* MAF_{j,M}^*}{w_{s_m^*}} - \tilde{e}_M^{(1,k)} + \mathrm{K}_{\mathbb{e}}^{(\beta_{j,M})} \Big/ N_j \end{pmatrix},$$

where

$$\mathrm{K}_{\mathbb{e}}(\boldsymbol{\beta}_{j,M}) = \frac{1}{2\left(e_{j,M}^{(0,k)} e_{j,M}^{(2,k)} - \left(e_{j,M}^{(1,k)}\right)^2\right)} \begin{pmatrix} d_{\alpha,j,M}^{(0,k)} e_{j,M}^{(2,k)} + d_{\alpha,j,M}^{(2,k)} e_{j,M}^{(0,k)} - 2 d_{\alpha,j,M}^{(1,k)} e_{j,M}^{(1,k)} \\ d_{\beta,j,M}^{(0,k)} e_{j,M}^{(2,k)} + d_{\beta,j,M}^{(2,k)} e_{j,M}^{(0,k)} - 2 d_{\beta,j,M}^{(1,k)} e_{j,M}^{(1,k)} \end{pmatrix}$$

and

$$d_{\alpha,j,M}^{(c+1,k)} = d_{\beta,j,M}^{(c,k)} = \frac{\partial e_{j,M}^{(c,k)}}{\partial \beta_{j,M}}.$$

isGWAS is computed using sufficient statistics

Under Hardy-Weinberg equilibrium, the quadruple $\{N_j, N_j^*, MAF_{j,M}, MAF_{j,M}^*\}$ are combined to form the global and local (under a wide radius of convergence) sufficient statistics from the logistic model. Consequently, they hold all necessary information to compute regression parameter estimates $\left\{\hat{\alpha}_{j,M}, \hat{\beta}_{j,M,s_m^*}, \hat{\sigma}_{\alpha_{j,M}}, \hat{\sigma}_{\beta_{j,M,s_m^*}}\right\}$ over a broad range of scenarios. Regardless of Hardy-Weinberg being valid or not, we show that the triple $\{T_{1j}, T_{2j}, T_{3j}\}$,

$$\left\{ T_{1j} = \sum_{i=1}^{N_j} y_i = N_j^*, \qquad T_{2j} = \sum_{i=1}^{N_j} y_i\, g_{ij,M,s_m^*}, \qquad T_{3j} = \sum_{i=1}^{N_j} g_{ij,M,s_m^*} \right\},$$

are the two global and one local sufficient statistics and these can alternatively be used as input variables in isGWAS. To show this, we write:

$$L(\boldsymbol{\beta}_{j,M,s_m^*}) = \sum_{i=1}^{N_j} \log P(y_i \,|\boldsymbol{\beta}_{j,M,s_m^*}, g_{ij,M,s_m^*})$$

$$= \alpha_{j,M,s_m^*} \sum_{i=1}^{N_j} y_i + \beta_{j,M,s_m^*} \sum_{i=1}^{N_j} y_i \, g_{ij,M,s_m^*}$$

$$+ \sum_{i=1}^{N_j} \log\left(1 - P\left(y_i = 1 \mid \boldsymbol{\beta}_{j,M,s_m^*}, g_{ij,M,s_m^*}\right)\right)$$

$$= \alpha_{j,M,s_m^*} T_{1j} + \beta_{j,M,s_m^*} T_{2j} - \sum_{i=1}^{N_j} \log\left(1 + \exp \alpha_{j,M,s_m^*}\right)$$

$$- \sum_{i=1}^{N_j} \log\left(1 + \frac{\exp \alpha_{j,M,s_m^*}}{1 + \exp \alpha_{j,M,s_m^*}} \left(\left(\exp \beta_{j,M,s_m^*} \, g_{ij,M,s_m^*}\right) - 1\right)\right)$$

$$= \alpha_{j,M,s_m^*} T_{1j} + \beta_{j,M,s_m^*} \left(T_{2j} - \frac{\exp \alpha_{j,M,s_m^*}}{1 + \exp \alpha_{j,M,s_m^*}} T_{3j}\right) - N_j \log\left(1 - \exp \alpha_{j,M,s_m^*}\right)$$

$$+ \mathcal{O}\left(\frac{\exp \alpha_{j,M,s_m^*}}{1 + \exp \alpha_{j,M,s_m^*}} \left(\beta_{j,M,s_m^*} g_{\cdot j,M,s_m^*}\right)^2\right)$$

$$= f\left(T_{1j}, T_{2j}, T_{3j}; \alpha_{j,M,s_m^*}, \beta_{j,M,s_m^*}\right) + \mathcal{O}\left(\frac{\exp \alpha_{j,M,s_m^*}}{1 + \exp \alpha_{j,M,s_m^*}} \left(\beta_{j,M,s_m^*} g_{\cdot j,M,s_m^*}\right)^2\right)$$

and valid when

$$\frac{\exp \alpha_{j,M,s_m^*}}{1 + \exp \alpha_{j,M,s_m^*}} \left|\left(\exp \beta_{j,M,s_m^*} \, g_{\cdot j,M,s_m^*}\right) - 1\right| < 1.$$

Hence, the global sufficient statistics are $\{T_{1j}, T_{2j}\}$ and (on assuming random g_{ij,M,s_m^*} as in the SaLN-R algorithm) the locally sufficient statistic is $\{T_{3j}\}$, where:

$$T_{1j} = \sum_{i=1}^{N_j} y_i = N_j^*, \qquad T_{2j} = \sum_{i=1}^{N_j} y_i \, g_{ij,M,s_m^*}$$

and

$$T_{3j} = \sum_{i=1}^{N_j} g_{ij,M,s_m^*} = \begin{cases} n_{j1}, & S_M = 1, \\ n_{j1} + 2n_{j2}, & S_M = 2. \end{cases}$$

Under Hardy-Weinberg equilibrium, we can write

$$T_{2j} = s_m^* N_j^* MAF_j^* \quad \text{and} \quad T_{3j} = s_m^* N_j MAF_j.$$

Leapfrog re-sampler: forecasting results in target sample sizes

To estimate regression parameters $\{\alpha_{j,M,s^*_m}, \beta_{j,M,s^*_m}\}$ in larger target sample sizes, i.e., $\widehat{N}_j > N_j$, we propose the following strategy:

1. **Specify number** K, **sub-sample** γ_1 **and target sample** γ_2 **parameters**, where $K \geq 1$, $0 < \gamma_1 < 1$ and $\gamma_2 > 1$.
2. **Generate random sub-samples of individuals of size** $\widetilde{N}_j = \gamma_1 N_j < N_j$. For each of $k = 1,2,\dots,K,$ generate a random sub-sample $D_{k,\gamma_1} \subset D$, where $|D_{k,\gamma_1}| = \widetilde{N}_j = \gamma_1 N_j$.
3. **(Leapfrog-step) Compute subsample quadruple and project to target sample size** $\widehat{N}_j$. For each subsample D_{k,γ_1}, compute values $\{\widetilde{N}^*_{ij}, \widetilde{MAF}_{kj,M}, \widetilde{MAF^*}_{kj,M}\}$ and project these on to the target sample size, i.e., $d_{k,\gamma_{1,2}} = \left\{\left(\frac{\gamma_2}{\gamma_1}\right)\widetilde{N}_{kj}, \left(\frac{\gamma_2}{\gamma_1}\right)\widetilde{N}^*_{ij}, \widetilde{MAF}_{kj,M}, \widetilde{MAF^*}_{kj,M}\right\}$ for sample D_{k,γ_1}
   - Note that $\left(\frac{\gamma_2}{\gamma_1}\right)\widetilde{N}_{kj} = \widehat{N}_j$, which is the target ‘future’ sample size.
4. **Deploy isGWAS across all *K* (projected) quadruples** $d_{k,\gamma_{1,2}}$ and record each estimate of the genetic effects, standard error and p-value $\left\{\hat{\beta}_{k,j,M,s^*_m},\ \hat{\sigma}_{\beta_{k,j,M,s^*_m}}, p_{k,j,M,s^*_m}\right\}_{k=1:K}$.
5. **Estimate p-value in target sample size** as a summary point estimate (e.g., median) or range across all K sub-samples,

$$p_{target_{j,M,s^*_m}} = \text{median}\{p_{k,j,M,s^*_m}\}_{k=1:K}.$$

Data Quality Control: preparation of sufficient statistics for isGWAS

In order to deploy isGWAS successfully, the sufficient statistics are required to be prepared in a sample where only a single individual (preferably case) from pairs or n-tuples of 3rd, 2nd and 1st degree relatives is retained. Additionally, ethnical outliers must also be removed. In summary, to deploy isGWAS successfully we require either: (a) access to the sufficient statistics computed after duplications of related n-tuples and ethnical outliers are removed; or (b) access to the individual level data, whereupon the sufficient statistics can be prepared as described in (a). We provide a detailed outline of recommended Quality Control for genetic Individual Level Data (ILD) to running successfully isGWAS in **Supplementary Information**.

Application to Biobank data

The GWAS results used in the assessment of isGWAS were taken from large-scale analyses of UK Biobank[13], Biobank Japan[14] and the Psychiatric Genomics Consortium[15].

The UK Biobank[13] is a large-scale biomedical database and research resource containing in-depth genetic and health information from half a million UK participants. From the full available UK Biobank cohort, we obtain phenotypes for seven different diseases with varying levels of prevalence. These are Hypertension (IC10:I10), Asthma (IC10:J45), Atherosclerosis (IC10:I25), Glaucoma (IC10:H40), Stroke (IC10:I63), Colon Cancer (IC10:C18) and Thyroid Gland Cancer (IC10:C73) patients. From a total cohort of 502,422 participants, we used the following inclusion criteria: white British (Field 22006), non-related (>3rd degree), no patients with difference in reported (Field 31) and genetic (Field 22001) sex, no patients with aneuploidy (Field 22019), no patients with unusual heterozygosity and high missing rates

(Field 22027). The ethnicity component is obtained from samples who self-identified as 'White British' according to Field 21000 and have very similar genetic ancestry based on a principal components analysis of the genotypes. Retaining one related individual (where we favour the retention of cases) we obtain a working sample size of ~335,000 individuals; the approximate value is owing to small differences in the number of cases between disease phenotypes (**Supplementary Information**). Comparative analysis for these varying populations is reported in the main text. The prevalence ratios and exact number of cases and controls are provided in **Supplementary Table 1**. The variant based statistics needed for isGWAS were obtained from the imputed UK Biobank dataset. A quality info score>0.9 is applied to the data, and the number of cases and controls per variant and the MAF for variant in cases and controls is based on patients with non-missing genotypes for the variant using software PLINK[4]. Sample-level MAF>0.001 is used as inclusion criteria for the variants to analyse. For each disease, we run isGWAS analysis using default settings under the 'additive' genetic model. In addition, we also perform GWAS analysis using two-step REGENIE[8] applied to all variants with a MAF>0.001 and Genotype Score>0.99. Firth correction was enabled and performed on variants with p-value<0.1. REGENIE was also adjusted for covariate information (age, sex, ancestry). For each disease we provide the following diagnostic plots: 1) mirrored Manhattan plot comparing directly p-values for isGWAS and REGENIE, 2) p-value – p-value plots comparing REGENIE and isGWAS, 3) $\beta - \beta$ plots comparing REGENIE and isGWAS where we have colored the values by a) MAF and b) ratios of computed standard error (SE) between methods, i.e., $log_2(\frac{SE(isGWAS)}{SE(REGENIE)})$. Across all diseases and variants considered, we compare performance of isGWAS and isGWAS-Firth to REGENIE-Firth.

Schizophrenia data from the Psychiatric Genomics Consortium[15] was used to conduct two different large-scale GWAS analysis. The first GWAS analysis was executed with data from

77,096 European individuals (33,640 cases, 43,456 controls)[25]. The second GWAS analysis was executed with data from the larger 130,644 European individuals (53,386 cases, 77,258 controls)[27]. We used the 2014 dataset to infer the 2022 results. To do this, we refine the significant results from both imputed 2014 and 2022 summary statistics using clumping with PLINK. The European 1000 Genomes Project v3[19] dataset was used as a reference population for the clumping procedure. Twelve strategies for clumping were explored: three were LD R^2-based only, the other nine were a combination of clumping by LD block information and p-value thresholding. The refined variants are used to assess the inference capabilities of isGWAS both within each of the two datasets and the enrichment capabilities of isGWAS to infer p-values of the 2022 dataset using the 2014 dataset. For the 2014 dataset, 225 variants were remaining after the clumping. For the 2022 dataset, 451 variants were remaining after the clumping. From those, 54 are overlapping and 608 is the unique set between the two datasets.

The Biobank Japan data was used to conduct a large-scale GWAS with 212,453 Japanese individuals across 42 different diseases[24]. We obtained the published significantly associated loci ($P < 5e-08$) in autosomes from the GWAS findings which amounted to 309 variants across 30 different diseases. Similarly, we used the significantly associated X chromosome findings for males and females that amounted to a total of nine significantly associated loci across five diseases, although results are omitted from text. We applied isGWAS to the three different sets of variants using default parameters to assess the performance of isGWAS. To aid association interpretation, we use the following additional statistical tests to assess the accuracy and sensitivity of the isGWAS calculator for the Biobank Japan data. First, a classical ROC curve was produced where the true/false actual value was determined by various p-value thresholds (benchmarked against published Biobank Japan results). The isGWAS calculator is an inferential tool thus this usage of the ROC curve is unconventional, however, it provides us with the opportunity to assess the sensitivity to the choice of thresholds used to correct for multiple testing. These are 10^{-10}, 10^{-8}, 5×10^{-8}, 10^{-7}, where we have also used 9.58×10^{-9} for Biobank Japan as recommended by the authors[24]. AUC values were not obtained as this is not a standard classification problem and they are not interpretable in this context. Second, an adapted ROC curve was produced which accounts for two different thresholds – one more stringent one to determine the true positive rate and one less stringent one to determine the true negative rate. **Supplementary Figure 22** showcases this scenario and highlights the importance of a threshold choice and its impact on a sensitivity analysis. The main aim of isGWAS calculator is to be used as an inferential tool for truly significant or truly non-significant genetic signals. Thus, using two thresholds – one for truly significant and one for truly non-significant – provides us the assess the sensitivity of isGWAS to this scientific question. Third, the obtained β values were compared to the true ones by obtaining the percentage of 1) predicted β values in the 95% C.I.s of the true β values and 2) 95% C.I.s of the predicted β values in the 95% C.I.s of the true β values.

Simulation scenarios

In the first scenario, for each individual i and iteration index k, we randomly generate disease status via $y_{ik} \sim Ber(\pi_{ik}; \alpha_k, \beta)$ with probability of disease $\pi_{ik} = expit(\alpha_k + \beta g_{ik})$ and $g_{ik} \sim Bin(2, MAF_k)$. Minor allele frequency is randomly selected from the set $MAF_k \in \{10^{-4}, 5 \times 10^{-4}, 0.01\} \cup \{0.025, 0.05, \dots, 0.5\}$ and the genetic effect on disease risk is fixed as $\beta = 0.5$. In the second study, we allow the genetic effect to vary, i.e., $\beta \equiv \beta_k$, by fixing disease status per individual and generating genotype data in controls $g_{ik}|y_{ik} = 0 \sim Bin(2, MAF_k)$ or cases $g_{ik}|y_{ik} = 1 \sim Bin(2, MAF_k^*)$, where minor allele frequency in cases is taken as the outer product with the sample minor allele frequency, with a random increase or decrease in frequency (which controls the magnitude and direction of genetic effect), i.e., we introduce the set $MAF_k^* \in MAF_k \otimes (1 \pm MAF_k)$. The parameter β_k is then estimated via each of the 5 estimators using the vector of simulated data $\{\boldsymbol{y}_k, \boldsymbol{g}_k\}$.

We compare isGWAS and isGWAS-Firth against classical logistic and Firth corrected regression[16], [37], [38]. Details for the second scenario are provided alongside full description of the simulation protocol in the **Supplementary Information**.

Leapfrog re-sampler: simulation and real-data analyses

The parameters $\{K, \gamma_1, \gamma_2\}$ in the leapfrog re-ampler are assessed over a variety of values. To attenuate the computational burden of a 3-dimensional grid search, we considered scenarios in which: $K = 100$, $\gamma_1 = {}^{1}/_{\gamma_2}$ and a γ_2-fold increase in sample size of $\gamma_2 \in \{1.1, 1.25, 1.5, \ldots, 2.5\}$, i.e., a 10% to 150% increase in sample size. Furthermore, we take our working sample size to be 276,204 individuals, which matches the number of all unrelated individuals in our UKB sample (i.e., on not retaining any member of a related pair – which is therefore fixed between diseases). We used our simulation protocol (**Supplementary Information**) to generate synthetic samples and additionally assessed performance across all seven disease datasets in UK Biobank. Variants for assessment were selected after pruning in PLINK[4] was applied to the ~11 million variants with the following parameters: genotype quality>0.99, MAF>0.01, HWE $p < 10e - 15$, 1000 bp windows, 100 variant increments, $R^2 > 0.9$. From the pruned variants, 5% were selected uniformly from variants with $p > 10^{-6}$ and all variants with $p \leq 10^{-6}$ were retained. Final number of variants progressed for LRS for the seven diseases are provided in **Supplementary Table 10**. For simulated data, data for smaller sub-samples were simulated using full cohort and empirical distributions for MAF and disease prevalence. In our tests of the LRS, we assess the predictive properties of isGWAS on real-life data where the ground truth is either computed from the entire sample or provided in the literature. Predictions from X-fold increases in sample size are compared using standard accuracy, FDR, FPR and TPR measures based on a putative true significance threshold of $5e - 08$.

Computational resources

Real-life analyses were performed using up to 48 virtual CPU cores of a 2.5 GHz Intel Xeon Gold 6240R processor with 64 GB of memory. Simulation analyses were performed using up to 8 virtual CPU cores of a 2.4 GHz Intel Core i9 processor.

Computational comparison protocol

We contrast the computational performance of isGWAS and REGENIE (Step-2 only). For clarity, REGENIE Step-1 simplifies the outcome and model by projecting out covariate information, before variant-disease association analyses are performed in Step-2. To directly compare both methods, we performed individual GWA analyses of each of the seven diseases considered in UK Biobank across ~11m variants for ~335,000 individuals. Owing to computational cost of the ILD method, we summarise results from a single GWA analysis per trait. Performance of isGWAS across repeated runs, for varying numbers of SNPs and available CPUs, up to a maximum of 10m variants, is also performed.

Data availability

The genotype data, phenotype status and allele counts were extracted from UK Biobank[13] to support the findings of this study. The genome-wide association summary data with available allele frequencies and cohort counts that was used to support the findings of this study are available from: Psychiatric Genomics Consortium[15] and Biobank Japan[14].

Code availability

The tool is available for use on the webportal www.optima-isgwas.com. The isGWAS algorithm is also available on github (https://github.com/cnfoley/isgwas/).

References

[1] W. Zhou *et al.*, ‘Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease’, *Cell Genomics*, vol. 2, no. 10, Oct. 2022, doi: 10.1016/J.XGEN.2022.100192.

[2] A. Abdellaoui, L. Yengo, K. J. H. Verweij, and P. M. Visscher, ‘15 years of GWAS discovery: Realizing the promise’, *Am. J. Hum. Genet.*, vol. 110, no. 2, pp. 179–194, Feb. 2023, doi: 10.1016/J.AJHG.2022.12.011.

[3] M. Claussnitzer *et al.*, ‘A brief history of human disease genetics’, *Nature*, vol. 577, no. 7789, pp. 179–189, Jan. 2020, doi: 10.1038/S41586-019-1879-7.

[4] C. C. Chang, C. C. Chow, L. C. A. M. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee, ‘Second-generation PLINK: Rising to the challenge of larger and richer datasets’, *GigaScience*, vol. 4, no. 1, Feb. 2015, doi: 10.1186/s13742-015-0047-8.

[5] L. Jiang *et al.*, ‘A resource-efficient tool for mixed model association analysis of large-scale data’, *Nat. Genet. 2019 5112*, vol. 51, no. 12, pp. 1749–1755, Nov. 2019, doi: 10.1038/s41588-019-0530-8.

[6] P. R. Loh *et al.*, ‘Efficient Bayesian mixed-model analysis increases association power in large cohorts’, *Nat. Genet. 2015 473*, vol. 47, no. 3, pp. 284–290, Feb. 2015, doi: 10.1038/ng.3190.

[7] W. Zhou *et al.*, ‘Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies’, *Nat. Genet. 2018 509*, vol. 50, no. 9, pp. 1335–1341, Aug. 2018, doi: 10.1038/s41588-018-0184-y.

[8] J. Mbatchou *et al.*, ‘Computationally efficient whole-genome regression for quantitative and binary traits’, *Nat. Genet.*, vol. 53, no. 7, pp. 1097–1103, Jul. 2021, doi: 10.1038/s41588-021-00870-7.

[9] N. Homer *et al.*, ‘Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays’, *PLOS Genet.*, vol. 4, no. 8, p. e1000167, Aug. 2008, doi: 10.1371/JOURNAL.PGEN.1000167.

[10] X. Shi and X. Wu, ‘An overview of human genetic privacy’, *Ann. N. Y. Acad. Sci.*, vol. 1387, no. 1, p. 61, Jan. 2017, doi: 10.1111/NYAS.13211.

[11] E. W. Clayton *et al.*, ‘The law of genetic privacy: applications, implications, and limitations’, *J. Law Biosci.*, vol. 6, no. 1, pp. 1–36, Oct. 2019, doi: 10.1093/JLB/LSZ007.

[12] Z. Wan, J. W. Hazel, E. W. Clayton, Y. Vorobeychik, M. Kantarcioglu, and B. A. Malin, ‘Sociotechnical safeguards for genomic data privacy’, *Nat. Rev. Genet. 2022 237*, vol. 23, no. 7, pp. 429–445, Mar. 2022, doi: 10.1038/s41576-022-00455-y.

[13] ‘UK Biobank’. https://www.ukbiobank.ac.uk/ (accessed Jul. 27, 2023).

[14] ‘BioBank Japan’. https://biobankjp.org/en/ (accessed Jul. 27, 2023).

[15] ‘PGC – Psychiatric Genomics Consortium’. https://pgc.unc.edu/ (accessed Jul. 27, 2023).

[16] D. Firth, ‘Bias Reduction of Maximum Likelihood Estimates’, *Source Biom.*, vol. 80, no. 1, pp. 27–38, 1993.

[17] P. D. Sasieni, ‘From Genotypes to Genes: Doubling the Sample Size’, *Biometrics*, vol. 53, no. 4, p. 1253, Dec. 1997, doi: 10.2307/2533494.

[18] B. B. Sun *et al.*, ‘Genetic associations of protein-coding variants in human disease’, *Nature*, vol. 603, no. 7899, pp. 95–102, Mar. 2022, doi: 10.1038/S41586-022-04394-W.

[19] ‘1000 Genomes | A Deep Catalog of Human Genetic Variation’. https://www.internationalgenome.org/home (accessed Jul. 27, 2023).

[20] ‘Estonian Biobank’, *Tartu Ülikool*, Dec. 13, 2021. https://genomics.ut.ee/en/content/estonian-biobank (accessed Jul. 27, 2023).

[21] 'FinnGen research project is an expedition to the frontier of genomics and medicine | FinnGen'. https://www.finngen.fi/en (accessed Jul. 27, 2023).

[22] R. A. Fisher, 'On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P', *J. R. Stat. Soc.*, vol. 85, no. 1, p. 87, Jan. 1922, doi: 10.2307/2340521.

[23] Q. Wang *et al.*, 'Rare variant contribution to human disease in 281,104 UK Biobank exomes', *Nat. 2021 5977877*, vol. 597, no. 7877, pp. 527–532, Aug. 2021, doi: 10.1038/s41586-021-03855-y.

[24] K. Ishigaki *et al.*, 'Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases', *Nat. Genet.*, vol. 52, no. 7, pp. 669–679, Jul. 2020, doi: 10.1038/s41588-020-0640-3.

[25] S. Ripke *et al.*, 'Biological insights from 108 schizophrenia-associated genetic loci', *Nature*, vol. 511, no. 7510, pp. 421–427, Jul. 2014, doi: 10.1038/nature13595.

[26] M. Lam *et al.*, 'Comparative genetic architectures of schizophrenia in East Asian and European populations', *Nat. Genet.*, vol. 51, no. 12, pp. 1670–1678, Dec. 2019, doi: 10.1038/s41588-019-0512-x.

[27] V. Trubetskoy *et al.*, 'Mapping genomic loci implicates genes and synaptic biology in schizophrenia', *Nature*, vol. 604, no. 7906, pp. 502–508, Apr. 2022, doi: 10.1038/s41586-022-04434-5.

[28] M. Pirinen, P. Donnelly, and C. C. A. Spencer, 'Including known covariates can reduce power to detect genetic effects in case-control studies', *Nat. Genet. 2012 448*, vol. 44, no. 8, pp. 848–851, Jul. 2012, doi: 10.1038/ng.2346.

[29] J. Mefford and J. S. Witte, 'The Covariate's Dilemma', *PLOS Genet.*, vol. 8, no. 11, p. e1003096, Nov. 2012, doi: 10.1371/JOURNAL.PGEN.1003096.

[30] L. J. O'Connor, A. P. Schoech, F. Hormozdiari, S. Gazal, N. Patterson, and A. L. Price, 'Extreme Polygenicity of Complex Traits Is Explained by Negative Selection', *Am. J. Hum. Genet.*, vol. 105, no. 3, p. 456, Sep. 2019, doi: 10.1016/J.AJHG.2019.07.003.

[31] N. Chatterjee, B. Wheeler, J. Sampson, P. Hartge, S. J. Chanock, and J. H. Park, 'Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies', *Nat. Genet. 2013 454*, vol. 45, no. 4, pp. 400–405, Mar. 2013, doi: 10.1038/ng.2579.

[32] G. Kimmel and R. Shamir, 'A fast method for computing high-significance disease association in large population-based studies', *Am. J. Hum. Genet.*, vol. 79, no. 3, pp. 481–492, 2006, doi: 10.1086/507317.

[33] P. M. Visscher *et al.*, 'Statistical Power to Detect Genetic (Co)Variance of Complex Traits Using SNP Data in Unrelated Samples', *PLOS Genet.*, vol. 10, no. 4, p. e1004269, 2014, doi: 10.1371/JOURNAL.PGEN.1004269.

[34] 'Our Future Health', *Our Future Health*. https://ourfuturehealth.org.uk/ (accessed Jul. 27, 2023).

[35] M. I. Kurki *et al.*, 'FinnGen provides genetic insights from a well-phenotyped isolated population', *Nat. 2023 6137944*, vol. 613, no. 7944, pp. 508–518, Jan. 2023, doi: 10.1038/s41586-022-05473-8.

[36] J. K. Pritchard and P. Donnelly, 'Case-control studies of association in structured or admixed populations', *Theor. Popul. Biol.*, vol. 60, no. 3, pp. 227–237, Nov. 2001, doi: 10.1006/tpbi.2001.1543.

[37] R. Puhr, G. Heinze, M. Nold, L. Lusa, and A. Geroldinger, 'Firth's logistic regression with rare events: accurate effect estimates and predictions?', *Stat. Med.*, vol. 36, no. 14, pp. 2302–2317, Jun. 2017, doi: 10.1002/SIM.7273.

[38] G. Heinze, M. Ploner, D. Dunkler, H. Southworth, L. Jiricka, and G. Steiner, ‘logistf: Firth’s Bias-Reduced Logistic Regression’. May 03, 2023. Accessed: Jul. 27, 2023. [Online]. Available: https://cran.r-project.org/web/packages/logistf/index.html

Acknowledgements

The authors would like to thank Prof Peter Visscher for his insightful comments and suggestions on the methodology. The authors are additionally grateful to Manju Dissanayake, Chris Marshall, Michael Mulholland, Andrew Donald and the Optima Partners Data Science and Engineering software development team.

Author contributions

CNF developed the mathematical and statistical methodologies, developed the statistical software and conceptualized the design. ZK developed the statistical software and webtool, designed the methodological analysis pipeline and conducted the real-life analyses. REM contributed to the interpretation of results. HR conceptualized and supervised the study. BBS conceptualized, designed the study, contributed to the method, application, and contextualization of the study. All authors contributed to the writing of the manuscript.

Competing interests

BBS and HR are employed by Biogen. CNF and ZK are employed by Optima Partners. REM is an advisor to the Epigenetic Clock Development Foundation and Optima Partners.

Tables

Table 1. Accuracy, true positive rate, false positive rate and false discovery rate of isGWAS using REGENIE results as gold-standard and a threshold of $p = 5e - 08$ as classification rule. Results are obtained on all 11,079,229 variants used for the analysis of seven diseases in UK Biobank without clumping/finemapping.

Disease (ICD code)	Case:control ratio	TPR	FPR	Acc	FDR
Hypertension (I10)	1:2	0.625	0.000026	0.99961	0.041
Asthma (J45)	1:6	0.982	0.000028	0.99994	0.016
Atherosclerosis (I25)	1:9	0.886	0.000011	0.99995	0.043
Glaucoma (H40)	1:26	0.891	0	0.99997	0.036
Stroke (I63)	1:56	NA	0	1	NA
Colon Cancer (C18)	1:94	0.944	0	0.99999	0.037
Thyroid Gland Cancer (C73)	1:669	1	0	0.99999	0.051

Figures

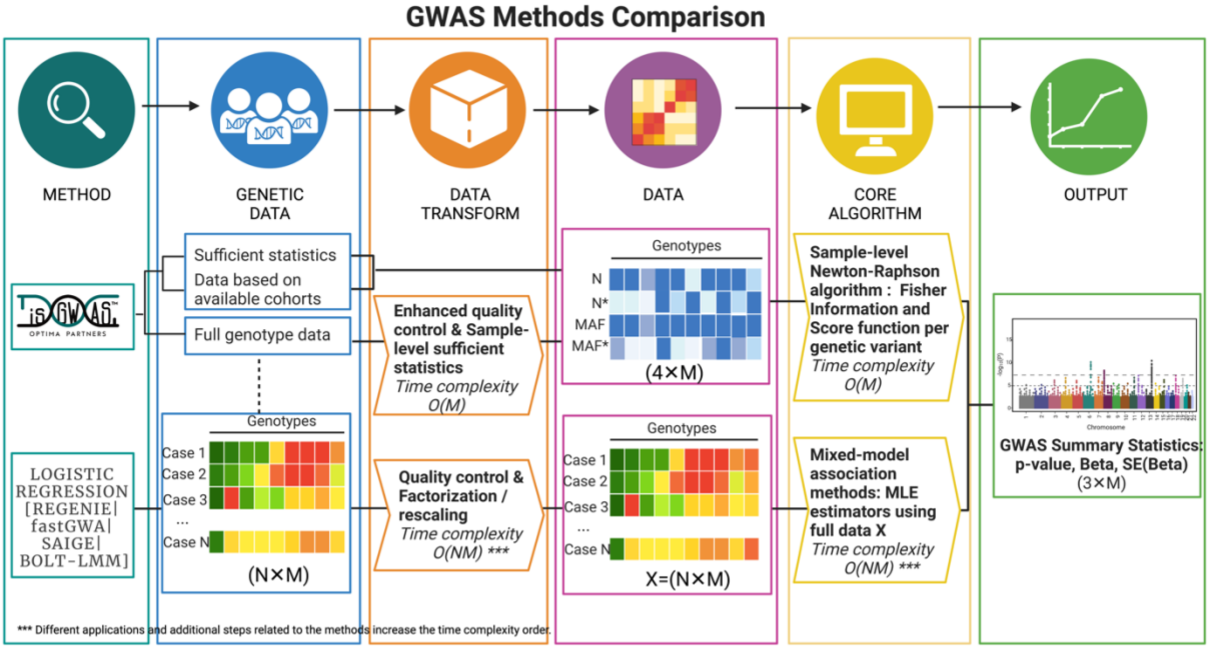

GWAS Methods Comparison
METHOD
GENETIC DATA
DATA TRANSFORM
DATA
CORE ALGORITHM
OUTPUT
isGWAS
OPTIMA PARTNERS
Sufficient statistics
Data based on available cohorts
Full genotype data
Genotypes
Case 1
Case 2
Case 3
...
Case N
(N×M)
LOGISTIC REGRESSION [REGENIE| fastGWA| SAIGE| BOLT-LMM]
Enhanced quality control & Sample-level sufficient statistics
Time complexity O(M)
Quality control & Factorization / rescaling
Time complexity O(NM) ***
Genotypes
N
N*
MAF
MAF*
(4×M)
Genotypes
Case 1
Case 2
Case 3
...
Case N
X=(N×M)
Sample-level Newton-Raphson algorithm : Fisher Information and Score function per genetic variant
Time complexity O(M)
Mixed-model association methods: MLE estimators using full data X
Time complexity O(NM) ***
Chromosome
GWAS Summary Statistics: p-value, Beta, SE(Beta)
(3×M)
*** Different applications and additional steps related to the methods increase the time complexity order.

Figure 1. Diagram highlighting main differences between isGWAS and other GWAS approaches.

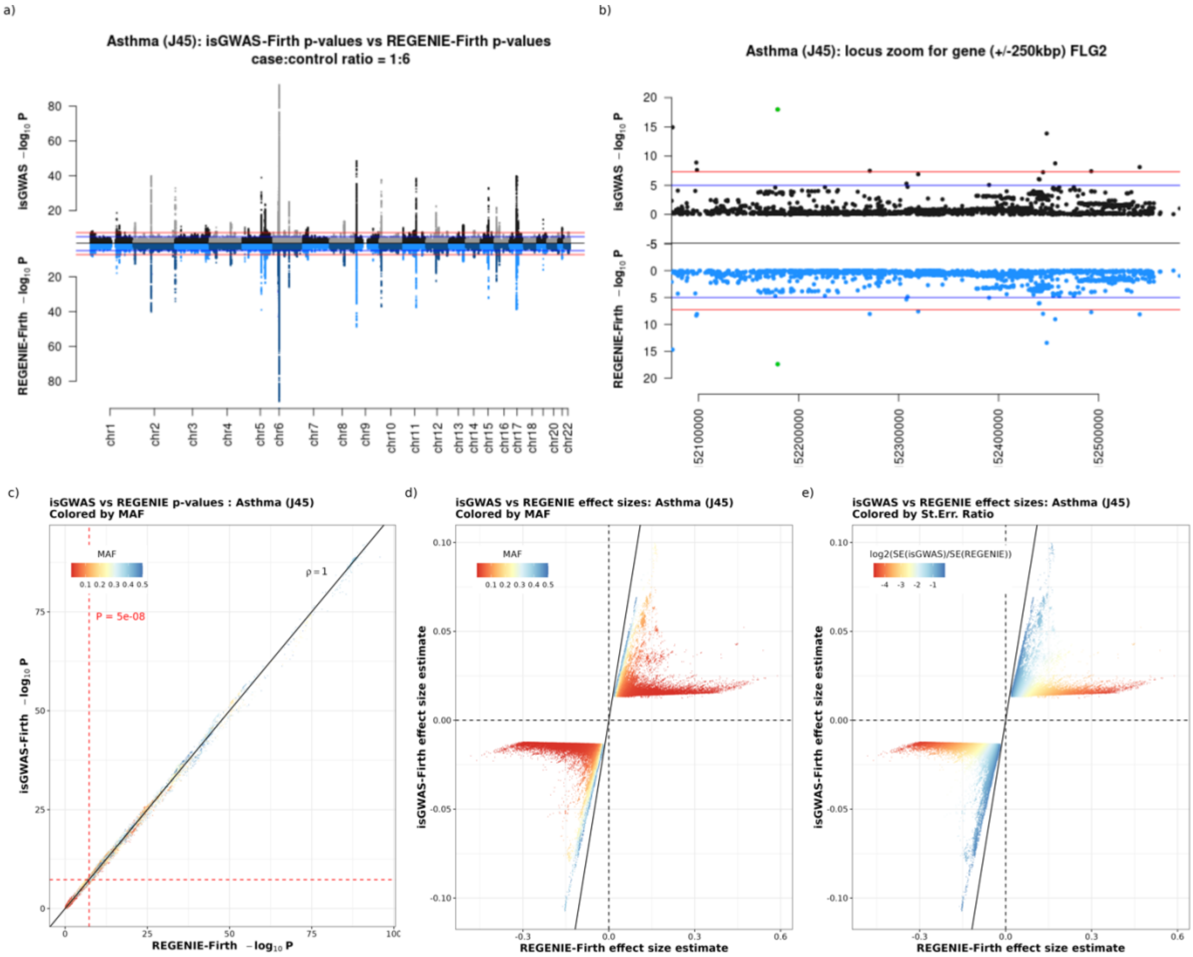
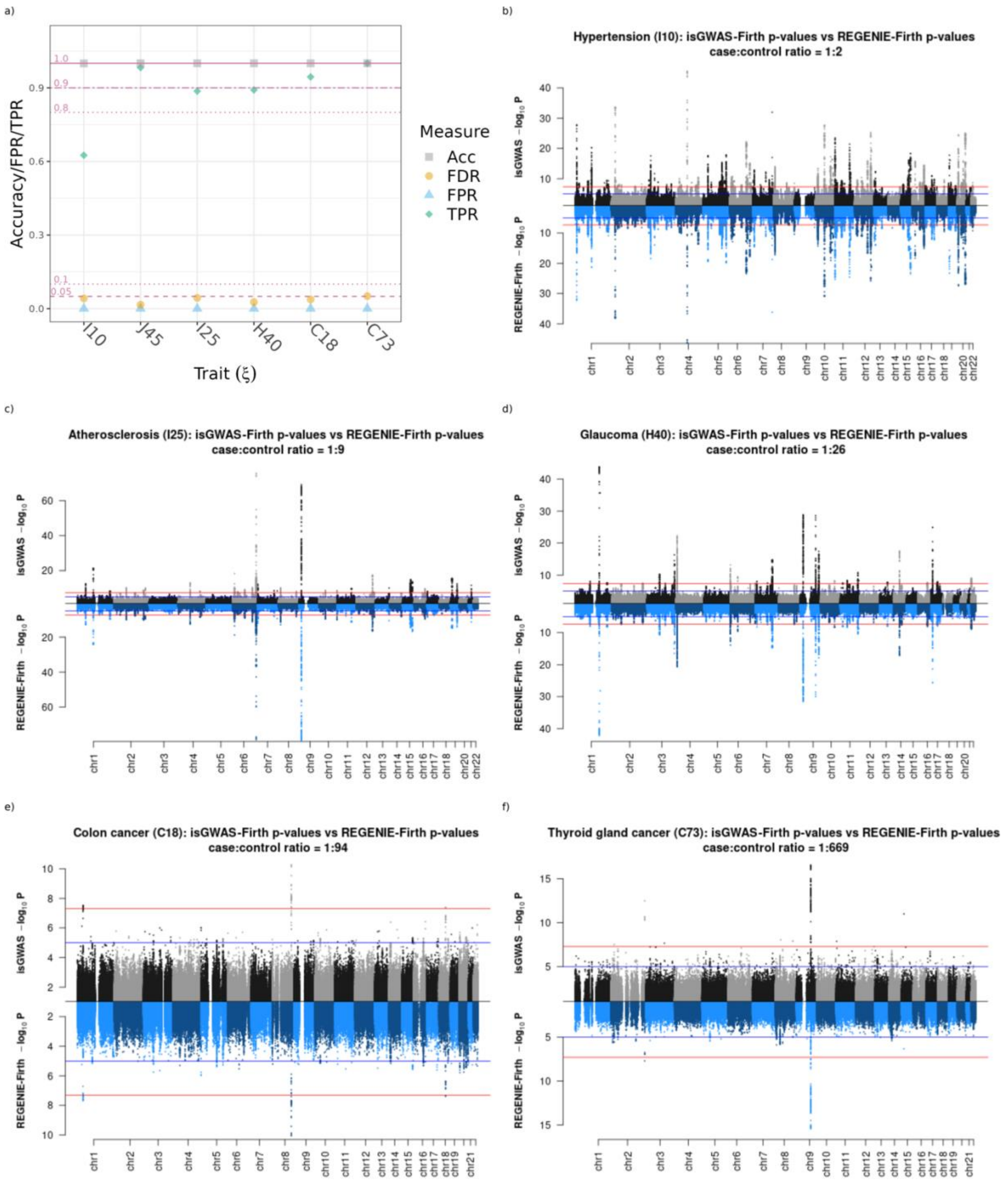


Figure 2. Comparative results for Asthma (IC10:J45) from UK Biobank. Subplot (a) is a mirror Manhattan plot comparing $-\log_{10} P$ values for isGWAS and REGENIE-Firth and subplot (b) is a locus zoom of the gene FLG2 region +/-250kbp on chromosome 7. Subplot (c) plots $-\log_{10} P$ values for isGWAS and REGNIE-Firth with the standard threshold P-value indicated colored by population-level MAF. Subplots (d) and (e) showcase $\beta - \beta$ effect size estimates for variants with p-values<0.05 and are coloured by population-level MAF and $\log_2(\frac{SE(isGWAS)}{SE(REGENIE)})$.



a)

b)

Hypertension (I10): isGWAS-Firth p-values vs REGENIE-Firth p-values
case:control ratio = 1:2

c)

Atherosclerosis (I25): isGWAS-Firth p-values vs REGENIE-Firth p-values
case:control ratio = 1:9

d)

Glaucoma (H40): isGWAS-Firth p-values vs REGENIE-Firth p-values
case:control ratio = 1:26

e)

Colon cancer (C18): isGWAS-Firth p-values vs REGENIE-Firth p-values
case:control ratio = 1:94

f)

Thyroid gland cancer (C73): isGWAS-Firth p-values vs REGENIE-Firth p-values
case:control ratio = 1:669

Figure 3. a) Accuracy/TPR/FPR comparing REGENIE-Firth and isGWAS results, where $p = 5e - 08$ threshold has been used as indicator for correct classification accuracy. Results are obtained on all 11,079,229 variants used for the analysis without clumping/finemapping. See Supplementary Table 1 for full results. Manhattan plots b), c), d), e) and f): Comparative results for five diseases from UK Biobank. Mirror Manhattan plots comparing $-\,log_{10}\,P$ values for isGWAS and REGENIE-Firth for six different diseases obtained from UK Biobank. Stroke was excluded from the analysis due to no variants passing significance threshold.

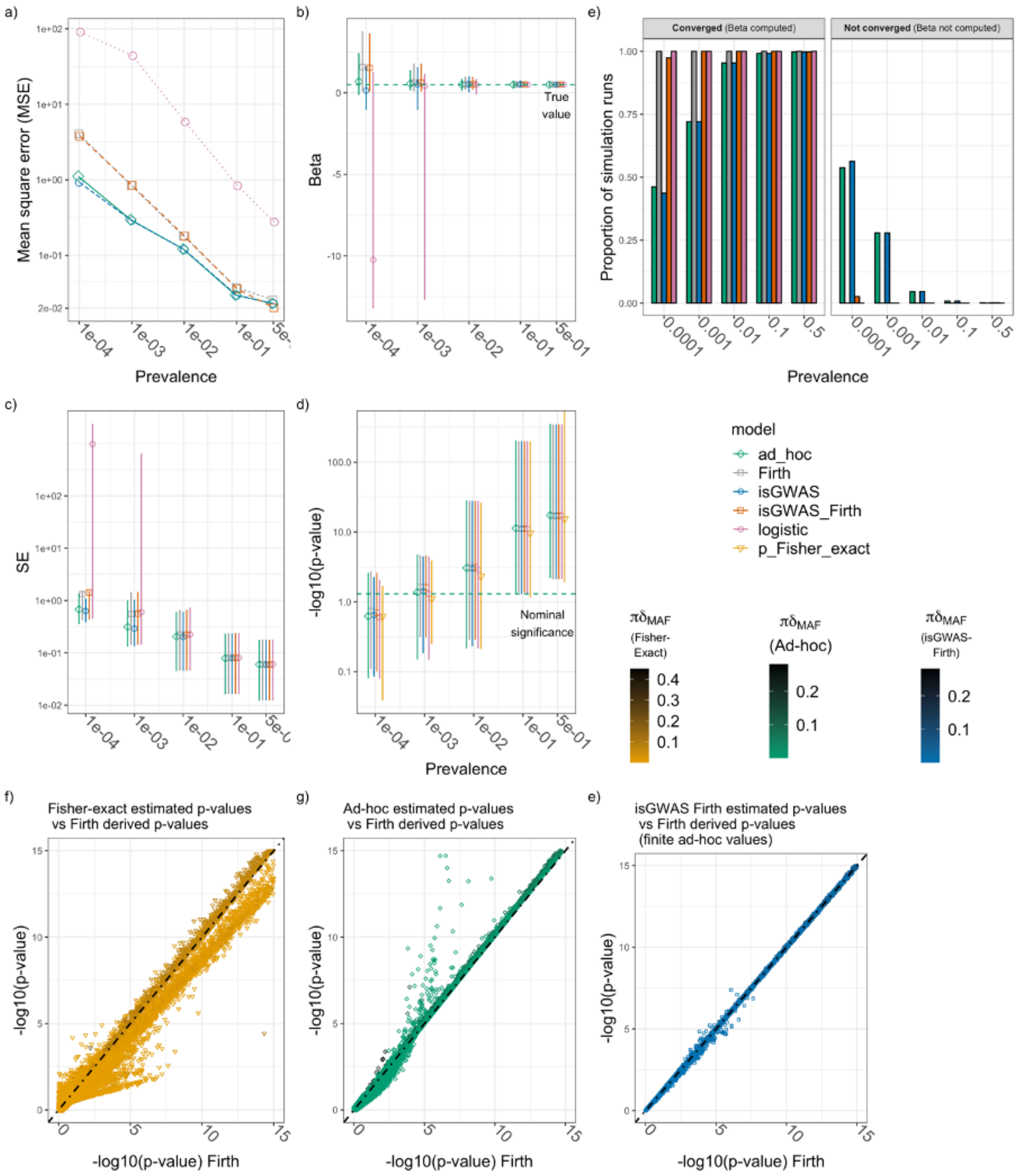


Figure 4. Simulation I results. Clockwise from top: a) Mean square error; b) distribution of estimated beta values; c) distribution of associated standard errors and d) distribution of -log_{10}(p-value), for each model - logistic regression, firth regression, 'isGWAS', 'isGWAS_Firth' and, for p-values only, Fisher's Exact Test - and specification of disease prevalence. Panel e) presents the relationship between Firth regression derived -log_{10}(p-value), along the horizontal axis, and the corresponding 'isGWAS_Firth' and Fisher's Exact Test (FET) computed values, on the vertical axis. Results are presented in the range $[0,15]$ *as FET regularly failed to converge for very small p-values and coloured according to value of* $\pi\delta_{MAF}$, *where*

π denotes prevalence and $\delta_{MAF} = {(MAF^{*} - MAF)}\big/{(MAF(1 - MAF))}$. In panels b)-d) a point denotes the median value and error-bars the first and ninth deciles of the range.

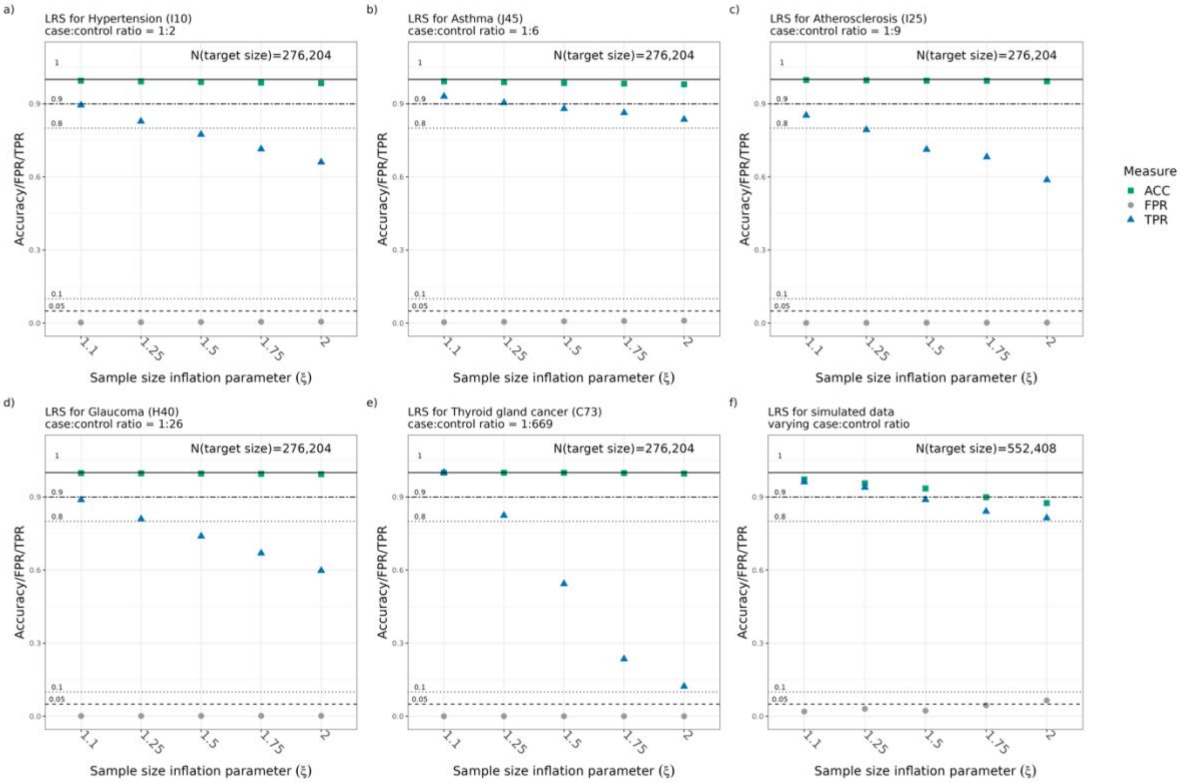


Figure 5. Performance of Leapfrog re-sampler (LRS) benchmarked against results derived from a)-e) the sample of ~276,204 individuals from UK Biobank with five different diseases and b) a simulated sample of 552,408 individuals (i.e., double UK Biobank sample size). For each value of ξ, we subset the target sample size down to ${N_{target}}\big/{\xi}$ individuals and deploy the LRS to compute predictions for the target sample N_{target}. As the maximum number of UK Biobank samples was 276,204, this was taken as the target. For example, when $\xi = 2$, we subset the full sample to 138,102 individuals and run the LRS to compute predictions of the larger 276,204 sample. We use results from the disease analysis, benchmarking LRS predictions against those computed on the pruned genome sampling uniformly across significance associations, resulting in ~3500 variants per studied disease. Colon cancer and Stroke are excluded from this figure as they don't have significant variants or a very low number of such after pruning. In the right panel we generated 1000 simulated datasets under the null of no genetic association or the alternative (see simulation protocol for details).

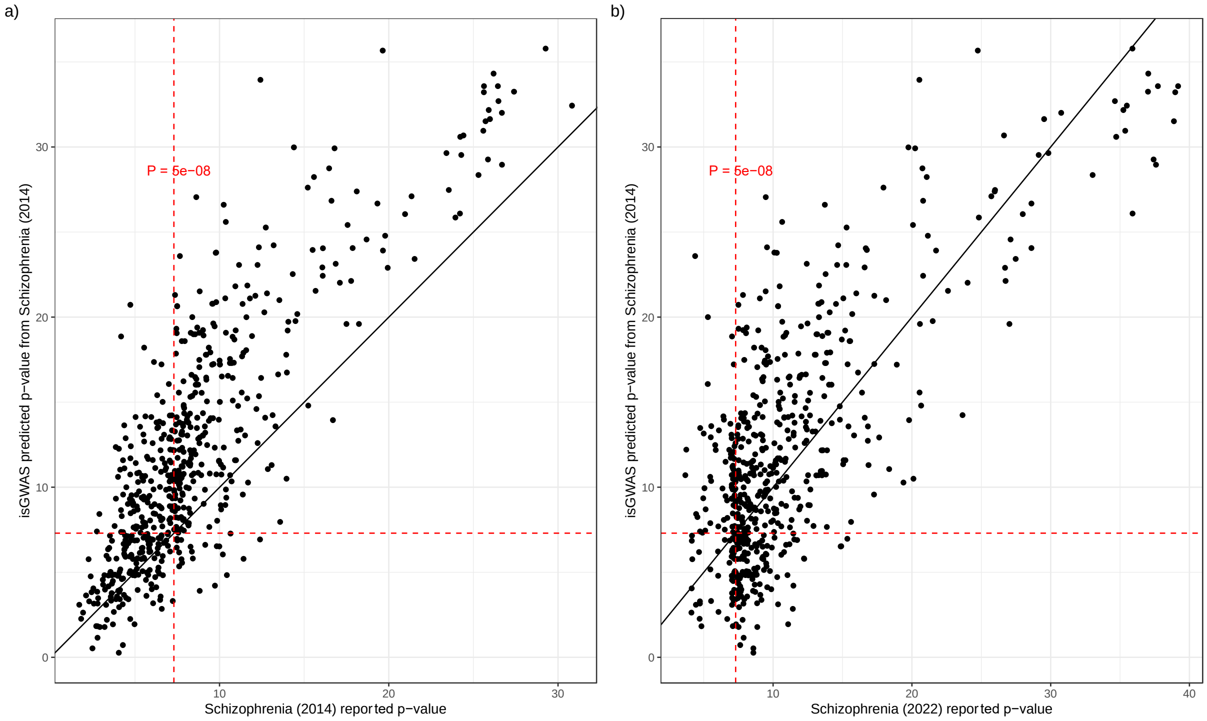


Figure 6. Prediction results for Schizophrenia (2022) using data from Schizophrenia (2014): population-level information for 608 significantly associated loci (P<1e-07) obtained from clumping with parameters ($R^2 = 0.2, p_1 = 1e-7, p_2 = 1e-7$) has been used to infer p-values. a) The figure compares reported GWAS Schizophrenia (2014) p-values and isGWAS predicted p-values using population-level information from Schizophrenia (2014) matching for the larger 2022 cohort size. b) The figure compares reported GWAS Schizophrenia (2022) p-values and isGWAS predicted p-values using population-level information from Schizophrenia (2014) matching for the larger 2022 cohort size. The dashed red line represents the threshold P=5e-08.

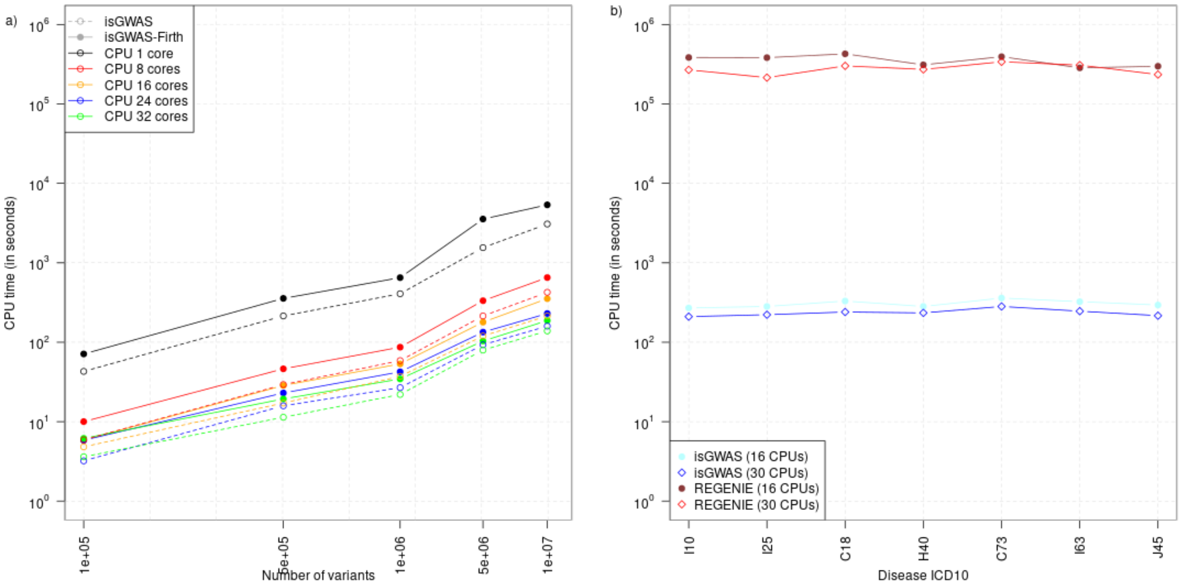


Figure 7. a) Computational CPU time (in seconds) for an increasing number of variants. The results compare performance of isGWAS with and without Firth distributed over different number of CPU cores. The data was obtained from UK Biobank ICD10:C73 disease with low disease prevalence (case-control ratio = 1:669). The x- and y-axis are on log_{10} scale. b) Computational CPU time (in seconds) for seven UK Biobank diseases on 11,079,229 variants for ~335,000 individuals. We compare the performance of isGWAS running on 16 and 30 CPUs vs the performance of REGENIE Step 2 running on 16 CPUs (with 16 threads) and 30 CPUs (with 30 threads). The computation is performed on the same machine. The y-axis is on log_{10} scale.