Incorporating family disease history and controlling case-control imbalance for population based genetic association studies

Yongwen Zhuang [1,2], Brooke N Wolford [3], Kisung Nam [4], Wenjian Bi [5], Wei Zhou [6], Cristen J Willer [3,7,8], Bhramar Mukherjee [2,9,10], Seunggeun Lee [1,2,4]

1. Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, Michigan, USA;

2. Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, Michigan, USA;

3. Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA;

4. Graduate School of Data Science, Seoul National University, Seoul, Korea;

5. Department of Medical Genetics, School of Basic Medical Sciences, Peking University, Beijing, China;

6. Massachusetts General Hospital, Broad Institute, Boston, Massachusetts, USA;

7. Department of Internal Medicine, Division of Cardiology, University of Michigan Medical School, Ann Arbor, Michigan, USA;

8. Department of Human Genetics, University of Michigan Medical School, Ann Arbor, Michigan, USA;

9. Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, Michigan, USA;

10. Michigan Institute of Data Science, University of Michigan, Ann Arbor, Michigan, USA

Correspondence:

Email: lee7801@snu.ac.kr

Address: Bldg 942, Graduate School of Data Science, Seoul, 08826, Republic of Korea

Abstract

In the genome-wide association analysis of population-based biobanks, most diseases have low prevalence, which results in low detection power. One approach to tackle the problem is using family disease history, yet existing methods are unable to address type I error inflation induced by increased correlation of phenotypes among closely related samples, as well as unbalanced phenotypic distribution. We propose a new method for genetic association test with family disease history, TAPE (mixed-model-based Test with Adjusted Phenotype and Empirical saddlepoint approximation), which controls for increased phenotype correlation by adopting a two-variance-component mixed model and accounts for case-control imbalance by using empirical saddlepoint approximation. We show through simulation studies and analysis of UK-Biobank data of white British samples and KoGES data of Korean samples that the proposed method is computationally efficient and gains greater power for detection of variant-phenotype associations than common GWAS with binary traits while yielding better calibration compared to existing methods.

Introduction

Genome-wide and phenome-wide studies are facilitated by the recent development of large-scale biobanks such as the UK Biobank (UKB)[1], BioBank Japan (BBJ)[2] and the Korean Genome and Epidemiology Study (KoGES)[3]. Individuals in the biobanks are samples from a target population and large numbers of phenotypes are collected for each individual, which allows phenome-wide scan. However, challenges remain to gain enough power to identify associated variants, especially for binary traits with a low prevalence.

One promising approach to improve detection power is using family disease history to infer risk of diseases of unaffected individuals. For family-based cohorts with partially-missing genotypes, association test power can be improved by using pedigree information[4-7]. The GWAX method first demonstrated that with completely-missing family genotypes, unaffected individuals with family disease history can be used as proxy-cases to find genetic associations[8]. The LT-FH method[9] further increases association power by estimating a liability of disease conditional on the observed phenotypes and family disease history, which differentiate the disease risks among the proxy cases.

Despite the progress, several important limitations remain. First, when samples are related, the increased correlation among the inferred risks (**Supplementary Figure 1**) can lead to type I error inflation. Hujoel et. al showed that since samples with close relatedness such as sibling pairs tend to have highly correlated GWAX or LT-FH phenotypes due to nearly identical family disease history, GWAX and LT-FH suffered poor calibration compared to GWAS [9]. Thus, the usage of the existing methods should be restricted to testing unrelated individuals only, which can reduce power. Second, with unbalanced case-control ratios, the distributions of inferred risks are still unbalanced, hence testing for association using linear mixed model (LMM) can yield inflated type I error rates. For example, diseases such as Parkinson's disease have low prevalence in UK-Biobank, which leads to a small number of cases and proxy cases (i.e. controls with non-zero inferred disease risk) in GWAX and a relatively low posterior liability conditioning on family history in LT-FH (**Supplementary Figures 2 and 3**). Since the gaussian approximation does not perform well in this setting, LMMs can yield inflated type I error rates. Currently no method exists to handle situations of this kind.

We propose a new method for genetic association test with family disease history, TAPE (mixed-model-based Test with Adjusted Phenotype and Empirical saddlepoint approximation), which controls for increased phenotype correlation and case-control imbalance. In standard mixed model methods, only a

dense genetic relatedness matrix is used as the variance component. TAPE uses a sparse kinship matrix as an additional variance component to further account for the increased correlation among phenotypes in closely related individuals. In addition, to adjust for case-control imbalance, TAPE uses empirical saddlepoint approximation under a linear mixed model[10-12].

Given family disease history, TAPE infers the phenotype for controls based on the proportion of affected relatives weighted by kinship coefficient and is robust to including the disease status of relatives other than parents and siblings in existing methods. TAPE is also flexible to use inferred phenotypes generated from other approaches such as the one proposed in LT-FH into the analysis.

We show through simulation studies and analysis of UK Biobank that the proposed method is computationally efficient, achieves greater power for detection of variant-phenotype associations than common GWAS with binary traits, and yields better calibration among related individuals compared to LT-FH. We applied TAPE to 10 binary traits in UK Biobank among 408,898 white British samples with imputed genotypes and parental disease information and identified 659 genome-wide significant clumped variants, among which 127 were with MAF<1%. We also analyzed two binary phenotypes in the KoGES data with 72,298 samples and identified 29 genome-wide significant clumped variants in total using TAPE.

RESULTS

Overview of methods

The TAPE method takes a three-step framework (**Figure 1**): (0) infer the disease risk for all individuals in the analysis based on the original case-control status and family disease history to be used as phenotype; (1) fit two variance components null linear mixed model to obtain parameter estimates; (2) test for genetic association using score test with empirical saddlepoint approximation.

In Step 0, the phenotypes are adjusted using inferred risk of individuals. The basic TAPE uses a weighted proportion of the affected close relatives to the control, which can be viewed as an extension of the GWAX method[8] to further differentiate disease risk of controls based on family disease history configurations. Another approach is TAPE-LTFH, which uses the liability of diseases generated from LT-FH as the adjusted phenotypes.

In Step 1, we fit the null linear mixed model to estimate model parameters. Fixed effects of the null model include covariates such as demographic information and principal components. Sample relatedness is accounted for by random effects in the model. We include two random effects, the first uses the sparse kinship matrix as covariance structure similar to that in fastGWA[13], and the second uses the dense genetic relatedness matrix (GRM). These two variance components can capture both increased correlation in phenotypes due to phenotype adjustment procedure and distance genetic relatedness among individuals. To make the method scalable, average information restricted maximum likelihood (AI-REML)[14], with preconditioned conjugate gradient (PCG) method[15] similar to that used in BOLT-LMM[16] and SAIGE, is used. The computation complexity is $\mathcal{O}(B(M_{GRM} + C_{sparse})N^{1.5})$ where B is the number of iterations until convergence, C_{sparse} is the number of non-zero elements in the sparse kinship matrix, M_{GRM} is the number of variants included in the GRM construction, N is the sample size, and the number of iterations for PCG method is assumed to be $\mathcal{O}(N^{0.5})$. We use raw genotypes as input and calculate GRM in runtime, yielding a reduced memory usage of $M_{GRM}N/4$ bytes compared to methods that facilitate a precomputed GRM, which has memory usage of fN^2 bytes where f denotes memory size for a floating number.

In Step 2, score test statistic is calculated for each genetic variant against the adjusted phenotype. Since the Gaussian approximation does not perform well at the tails of the test statistic distribution especially when the case-control ratio is unbalanced and MAF of the variant is low, we approximate the distribution by empirical saddlepoint approximation[11], which uses empirically estimated cumulant generating function

(CGF) to calculate p-value. The empirical saddlepoint approximation is utilized when the test statistic exceeds two standard deviations of the mean. Time complexity for this step is $\mathcal{O}(MN)$.

Simulation study results

Type I error and Power

Simulation results for TAPE were compared with three other methods: (1) GWAS with original binary phenotypes by SAIGE[17]; (2) BOLT-LMM with LT-FH phenotypes[9] (hereafter denoted as LT-FH), which is shown to increase association power over GWAX [8]; and (3) a hybrid method using TAPE testing procedures with LT-FH phenotypes (hereafter denoted as TAPE-LTFH). Type I error rates were evaluated at genome-wide $\alpha = 5 \times 10^{-8}$ with sample size of 10,000 and case-control ratio ranging from 1:99 to 10:90. For each case-control ratio setting, two sets of genotype data with 10^9 independent variants were generated with MAF of 0.1 and 0.01 respectively. We first simulated a population consisting of 2,500 pairs of siblings and 5,000 independent individuals (**Table 1a).** The empirical type I error rates of LT-FH were largely inflated under more unbalanced case-control ratio and lower MAF, while results from TAPE and SAIGE were well calibrated. For the hybrid method TAPE-LTFH, inflation was also observed but not as large as that of LT-FH, since the additional variance component in the mixed model further accounts for the phenotypic concordance among sibling pairs with same family disease history, and the empirical saddlepoint approximation better approximates the distribution of test statistics. Further, we evaluated type I error rates with a more complex relatedness structure, i.e., a population consisting of 625 8-member families and 5,000 independent individuals (**Table 1b).** Inflated type I error rates were observed in results from LT-FH but with lower magnitude compared to the previous setting. TAPE-LTFH had slightly inflated type I error rates. One explanation is that LT-FH phenotypes are less concordant in the latter setting since there is a smaller number of individuals sharing identical family history under a more complicated pedigree. On

the other hand, type I error rates from TAPE and SAIGE were relatively well controlled with a slight deflation.

One of the important features of TAPE is the use of a kinship matrix in addition to (dense) GRM to account for increased correlation among phenotypes. Two additional analyses were performed to investigate the influence of no kinship variance component (TAPE-nok) and mis-specified kinship matrix (TAPE-misk) on calibration of TAPE. For TAPE-nok, the sparse kinship matrix was not included as an LMM variance component and inflated empirical type I error was observed (**Supplementary Figure 4).** For TAPE-misk, the true kinship matrix of an 8-member family pedigree (**Supplementary Figure 5a)** was replaced with a slightly mis-specified one (**Supplementary Figure 5b)** in step 0 and step 1. The empirical type I error of TAPE-misk was similar to that of TAPE. The results indicated that the impact of a slightly mis-specified kinship matrix was negligible, while the inclusion of the kinship matrix as a variance component is crucial in controlling type I error rate when family information is incorporated into the analysis.

To assess empirical power, we compared the average χ^2 statistics (**Figure 2**) and the proportion of causal SNPs significant at $\alpha = 5 \times 10^{-8}$ level (**Supplementary Figure 6**) for simulated data sets with sample size 10,000 under different genetic effects and case-control ratio. For each data set, 100,000 independent variants with MAF 0.1 were simulated in which 1% were causal, and we generated 100 data sets for each setting. TAPE achieves greater detection power over SAIGE results, with a 21.0% average increase in average χ^2 statistics and a 12.1% average increase in proportion of causal SNPs detected. Although LT-FH also had increased χ^2 over SAIGE by 27.5% and had a 14.3% average increase in detection rate, it suffered from type I error inflation especially when analyzing related samples.

To investigate how more complex relatedness structures will influence simulation results, we further simulated a population in which related individuals form families with 8 members (**Supplementary Figure 5a)**. With this setting, the phenotype adjustment of TAPE method takes into account all relatedness in the

pedigree including second-degree relationships such as grandparent-grandchild, while LT-FH only integrates first-degree relationship information from parents and sibling of an individual. TAPE has higher overall power than both LT-FH and SAIGE under such relatedness structure (**Supplementary Figure 6**), with a 18.5% average increase in proportion of significant SNPs detected over SAIGE, slightly higher than LT-FH's 17.9% increase over SAIGE.

In general, the proposed TAPE method yielded well-controlled type I error rate even when case-control ratio is unbalanced, which makes the incorporation of family disease information in genetic association test feasible in the presence of sample relatedness and can achieve greater detection power than traditional GWAS analysis using unadjusted binary phenotypes.

Computation Time

Computation time was evaluated using randomly selected samples from 408,898 white British individuals in UK Biobank data for Type II diabetes (case:control=1:20) with $M = 100{,}000$ variants. Projected computation time for 21 million variants with MAF $\geq 0.01\%$ was estimated and plotted on log10 scale against sample size varying from 10,000 to 408,898 (**Supplementary Figure 7**). P-values for LT-FH method were calculated using BOLT-LMM. Computation time for TAPE-LTFH is similar to that for TAPE and is therefore omitted in the plot. A break-down of run time for null model estimation and p-value calculation is presented in **Supplementary Table 3**. Since TAPE fits the model with two variance components and uses ESPA in p-value calculation, which requires additional computation, TAPE was slower than SAIGE and LT-FH. Overall, TAPE is scalable to analyze biobank size data. For genome-wide analysis of testing 21 million variants, TAPE required 16 CPU hours with 40,000 samples and 284 CPU hours with 408,898 samples.

Analysis of binary traits in biobank data

We analyzed 10 binary disease outcomes with available parental disease status in the UK Biobank[1]. The binary traits were defined by the PheWAS codes[17] aggregated from ICD codes in the UK Biobank dataset, with case-control ratio ranging from 1:4 to 1:406 among 408,898 white British individuals (**Table 2**). We tested over 21 million variants imputed from the Haplotype Reference Consortium (HRC)[18] with minor allele frequency (MAF) $\geq 0.01\%$. Sex, age and first 10 principal components were included in the analysis model. GRM was constructed using 93,511 genotyped variants with high quality. Kinship coefficients were estimated using the KING software[19], and the sparse kinship matrix was constructed using those with estimated kinship no larger than third-degree relatedness.

Figures 3 and 4 presents Manhattan plots and Q-Q plots stratified by MAF categories for two phenotypes with different case-control ratio: Type II diabetes (case-control ratio 1:20), and Parkinson's disease (case-control ratio 1:350). Plots for all 10 diseases in the analysis are shown in **Supplementary Figures 8 and 9**. TAPE results were compared to results from SAIGE v0.44.3[17] using binary phenotype and results from BOLT-LMM using LT-FH phenotypes[9]. It is shown that TAPE yields higher power than SAIGE for variants with higher MAF, and has better calibration than LT-FH especially among lower MAF variants. **Supplementary Table 1** lists the number of significant variants and significant clumped variants at $\alpha = 5 \times 10^{-8}$ detected by TAPE, TAPE-LTFH, LT-FH and SAIGE. Significant clumped variants were further identified by clumping genome-wide significant variants with 5Mb window size and linkage disequilibrium threshold $r^2 = 0.1$ using PLINK software[20]. TAPE identified 84 more genome-wide significant clumped variants than SAIGE for Type II diabetes, and 5 more for Parkinson's disease. For all 10 diseases analyzed, a total of 659 genome-wide significant clumped variants were identified by TAPE, including 127 clumped variants with MAF < 1%; whilst a total of 344 clumped variants were identified by SAIGE, of which 71 were with MAF < 1%.

To assess the calibration of testing methods, we performed stratified LD score regression with the baselineLD model to obtain the attenuation ratios[21] (**Supplementary Table 2**). For traits with more unbalanced case-control, TAPE consistently yields relatively lower attenuation ratios than TAPE-LTFH, while LTFH generates the highest attenuation ratio, indicating poor calibration. For example, the average attenuation ratio for type II diabetes (case:control=1:20) is 0.110, 0.120 and 0.142 for TAPE, TAPE-LTFH and LT-FH respectively; for Parkinson's disease (case:control=1:360), the average attenuation ratio is 0.125, 0.222 and 0.462 for TAPE, TAPE-LTFH and LT-FH respectively. Since we used all the individuals regardless of relatedness, the observation supports the previously reported result that LT-FH suffers poor calibration in related samples due to concordance between phenotypes from closely related samples such as sibling pairs[9]. On the contrary, TAPE is able to generate better calibrated results under such situations, followed by TAPE-LTFH.

For additional analysis, we applied TAPE, TAPE-LTFH and SAIGE to two binary phenotypes for 72,298 individuals with family disease history in the KoGES data and analyzed 8 million variants. Disease prevalence among sample individuals and their relatives is shown in **Supplementary Table 4**. For diabetes (case:control=1:12), TAPE identified 14 more genome-wide significant clumped variants than SAIGE, while TAPE-LTFH identified 15 more than SAIGE. For gastric cancer (case:control=1:191), both TAPE and TAPE-LTFH identified 3 genome-wide significant clumped variants (rs760077, rs35972942, rs2978977) while no variants were genome-wide significant by SAIGE. The three clumped variants have been previously reported to be associated with gastric cancer among Chinese or Japanese population [22-24], but not among Korean samples. Manhattan plots and Q-Q plots are presented in **Supplementary Figures 10 and 11.**

DISCUSSION

We propose a robust method that incorporates family disease information for genetic association test while accounting for case-control unbalance and close relatedness in the population. Over the past decade, the construction of large biobanks linked with electronic medical record data has facilitated large-scale genome-wide studies to test association with thousands of disease-related phenotypes. Samples in biobanks usually follow cohort study design, which can have small number of cases compared to traditional case-control study design, especially for rare phenotypes. Previous studies have shown that additional information from family disease history can help improve test power in such cases, yet challenges remain (1) to control for type I error inflation induced by increased correlation of phenotypes among related individuals after incorporating family disease history; and (2) to account for unbalanced distribution of phenotypes after being adjusted by family disease information. TAPE includes both a dense genetic relatedness matrix and a sparse matrix for close relatedness as variance components in the linear mixed model framework to account for sample relatedness and family-history-induced correlation. Empirical saddlepoint approximation of the score test statistic distribution is adopted to control for type I error inflation under unbalanced phenotypic distribution. Optimization strategies such as PCG for computing components with matrix inversion, and runtime GRM calculation from raw genotypes were implemented to improve computation efficiency and reduce memory usage.

TAPE incorporates family disease history by adjusting phenotype of control samples with a product of ρ and r, where ρ indicates the increase in latent disease risk among controls given all relatives of the individual are cases, and r represents a weighted proportion of diseased relatives of the individual. We assumed a constant $\rho = 0.5$ for the analysis in this paper, and we expect more accurate estimates of this value to yield better performance in capturing potential disease risk among controls with family disease history, which might be a direction for future exploration. For example, ρ can be estimated in a similar

way as genetic nurturing effects for different phenotypes respectively under a family analysis framework where genotypes for relatives are avaiable[16], thus enabling different level of contribution from relatives to the individual's latent risk for different diseases.

For null model, both sparse estimated kinship matrix and GRM are included in TAPE as variance components to account for the potential phenotypic concordance. The use of two or more variance components in mixed model has been shown to better control for test statistics inflation and improves association power as well as prediction accuracy in standard GWAS and family studies [25, 26], yet we are not aware of existing methods that apply more than one variance components to mixed model while incorporating family disease history. From simulation studies we show that the absence of kinship matrix in variance components leads to inflated type I error rates of association test results. This result echoes previous findings from LT-FH that the inclusion of family disease history can lead to calibration issue under a single-variance-component mixed model due to similar family history for closely related individuals such as sibling pairs[9], and indicates a possible solution to control for phenotypic correlation introduced by incorporating family disease information. When estimating variance parameters, TAPE improves computation efficiency by applying PCG algorithm on top of the sparse estimated kinship matrix and the dense GRM, where sparsity of the estimated kinship matrix is ascertained by proper thresholding.

The analytical framework of TAPE allows for flexible choice of outcome variables. For example, we also investigated a hybrid method, TAPE-LTFH, which uses LT-FH phenotypes in the proposed two-variance-component mixed model. We show by simulation studies that TAPE-LTFH can partially control for type I error inflation as compared to LT-FH, but not as well-calibrated as TAPE. It remains a future work to better capture latent risk while accounting for phenotypic concordance to further improve association power using external information such as family disease history. TAPE currently supports analysis of binary

phenotypes, and has the potential to be extended to ordinal phenotypes, such as categorical diagnoses of mental disorders, for more powerful association test in a broader context.

We also note several limitations of our proposed method. First, the potential difference in the phenotype classification for genotyped individuals and their relatives is not accounted for in TAPE. For example, phenotypes of genotyped individuals in UKB dataset were defined using the PheWAS codes aggregated from ICD9 and ICD10 codes, whereas parental phenotypes were extracted from self-reported surveys. The different phenotype classification standard may induce bias in the adjusted phenotype after incorporating family disease history. The second limitation lies in the modeling assumption of infinitesimal genetic effects, i.e., the effect size of each variant follows a standard Normal distribution, which may yield less detection power when the assumption does not match the true underlying genetic architecture.

Despite the above-mentioned limitations, TAPE is the only existing approach that incorporates family disease history while handling related samples and phenotype unbalance. With the increasing accessibility to large-scale biobank data with population relatedness and family disease history information, our proposed method is expected to contribute to improving detection power for genetic association studies, especially for late-onset diseases that are underrepresented in the sample cohorts.

URLs

TAPE (version 0.3.0): https://github.com/styvon/TAPE. Link to TAPE summary association statistics for 10 phenotypes in UKB data and 2 phenotypes in KoGES data: https://github.com/styvon/TAPE/blob/main/vignettes/biobank_results.md BOLT-LMM (version 2.3.4): https://data.broadinstitute.org/alkesgroup/BOLT-LMM. LT-FH (version 2): https://alkesgroup.broadinstitute.org/UKBB/LTFH. SAIGE (version 0.44.3): https://github.com/weizhouUMICH/SAIGE. KING (version 2.2.4): https://www.kingrelatedness.com/. PLINK (version 2.00): https://www.cog-genomics.org/plink2.

ACKNOWLEDGMENTS

This research was supported by NIH grants R01-HG008773 (Y.Z.), and Brain Pool Plus (BP+, Brain Pool+) Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (2020H1D3A2A03100666, S.L). UK Biobank data were accessed under the accession number UKB: 45227.

AUTHOR CONTRIBUTIONS

Y.Z., C.J.W and S.L. designed the experiments. Y.Z., B.W. and S.L. analyzed the UK Biobank data. K.N. and S.L. analyzed the KoGES data. Y.Z implemented the program with help from W.B and W.Z. Y.Z wrote the manuscript with input and critical feedback from all authors.

REFERENCES

1. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
2. Nagai, A. *et al.* Overview of the BioBank Japan Project: study design and profile. *Journal of epidemiology* **27**, S2-S8 (2017).
3. Kim, Y., Han, B. & KoGES Group. Cohort profile: the Korean genome and epidemiology study (KoGES) consortium. *International journal of epidemiology* **46**, e20-e20 (2017).
4. Thornton, T. & McPeek, M. S. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *The American Journal of Human Genetics* **81**, 321-337 (2007).
5. Gudbjartsson, D. *et al.* Many sequence variants affecting diversity of adult human height. *Nature genetics* **40**, 609-615 (2008).
6. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868-874 (2009).
7. Zhong, S., Jiang, D. & McPeek, M. S. CERAMIC: Case-control association testing in samples with related individuals, based on retrospective mixed model analysis with adjustment for covariates. *PLoS genetics* **12**, e1006329 (2016).
8. Liu, J. Z., Erlich, Y. & Pickrell, J. K. Case-control association mapping by proxy using family history of disease. *Nature genetics* **49**, 325 (2017).
9. Hujoel, M. L., Gazal, S., Loh, P.-R., Patterson, N. & Price, A. L. Liability threshold modeling of case–control status and family history of disease increases association power. *Nature genetics* **52**, 541-547 (2020).
10. Davison, A. C. & Hinkley, D. V. Saddlepoint approximations in resampling methods. *Biometrika* **75**, 417-431 (1988).
11. Feuerverger, A. On the empirical saddlepoint approximation. *Biometrika* **76**, 457-464 (1989).

12. Bi, W., Fritsche, L. G., Mukherjee, B., Kim, S. & Lee, S. A fast and accurate method for genome-wide time-to-event data analysis and its application to UK Biobank. *The American Journal of Human Genetics* **107**, 222-233 (2020).
13. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nature genetics* **51**, 1749-1755 (2019).
14. Gilmour, A. R., Thompson, R. & Cullis, B. R. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, 1440-1450 (1995).
15. Hestenes, M. R. & Stiefel E. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards* **49**, 409-436 (1952).
16. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics* **47**, 284 (2015).
17. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature genetics* **50**, 1335-1341 (2018).
18. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics* **48**, 1279-1283 (2016).
19. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873 (2010).
20. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics* **81**, 559-575 (2007).
21. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics* **47**, 1228 (2015).
22. Du, M. *et al.* Remote modulation of lncRNA GCLET by risk variant at 16p13 underlying genetic susceptibility to gastric cancer. *Science advances* **6**, eaay5525 (2020).
23. Tanikawa, C. *et al.* Genome-wide association study identifies gastric cancer susceptibility loci at 12q24. 11-12 and 20q11. 21. *Cancer science* **109**, 4015-4024 (2018).
24. Yan, C. *et al.* Meta-analysis of genome-wide association studies and functional assays decipher susceptibility genes for gastric cancer in Chinese populations. *Gut* **69**, 641-651 (2020).
25. Speed, D. & Balding, D. J. MultiBLUP: improved SNP-based prediction for complex traits. *Genome research* **24**, 1550-1557 (2014).
26. Widmer, C. *et al.* Further improvements to linear mixed models for genome-wide association studies. *Scientific reports* **4**, 1-13 (2014).
27. Zaitlen, N. *et al.* Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet* **9**, e1003520 (2013).
28. Tucker, G. *et al.* Two-variance-component model improves genetic prediction in family datasets. *The American Journal of Human Genetics* **97**, 677-690 (2015).
29. 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. *Nature genetics* **44**, 1166-1170 (2012).
30. Daniels, H. E. Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics*, 631-650 (1954).
31. Jensen, J. L. *Saddlepoint approximations*. (Oxford University Press, 1995).
32. Kuonen, D. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* **86**, 929-935 (1999).
33. Bycroft, C. *et al.* Genome-wide genetic data on~ 500,000 UK Biobank participants. Preprint at https://doi.org/10.1101/166298 (2017).

METHODS

Phenotype adjustment

We first introduce the proposed phenotype adjustment procedure in TAPE. In a sample of N individuals where each individual has N_R relatives with phenotypic information, F is an $N \times N_R$ kinship matrix with each element F_{ij} denoting the kinship coefficient between individual i and j ($i \in \{1, \ldots, N\}; j \in \{1, \ldots, N_R\}$), D is an $N \times N_R$ matrix with each element D_{ij} denoting the phenotype of relative j of individual i, Y is an N-vector of observed binary phenotypes. The adjusted quantitative phenotype for individual i, Z_i, is expressed as:

$$Z_i = \mathbb{I}(Y_i = 1) + \mathbb{I}(Y_i = 0)\rho \cdot r_i$$

where $\mathbb{I}(\cdot)$ denotes indicator function, ρ is a pre-specified constant indicating the increase in latent disease risk, and $r_i = \frac{\sum_{j=1}^{N_R} F_{ij}\mathbb{I}(D_{ij}=1)}{\sum_{j=1}^{N_R} F_{ij}}$. If $Y_i = 0$ and all N_R relatives of the *i*th individual are cases, the latent disease risk is $Z_i = \rho$. For the analysis in this paper, we assume that latent risk of such individual is 0.5 (i.e., $\rho = 0.5$). In addition, the phenotype adjustment procedure can be adapted to include information other than family disease status that is potentially indicative of latent disease risk. See **Supplementary Note 1 and 2** for details.

Both LT-FH and TAPE-LTFH use the posterior mean genetic liability proposed by Hujoel et al. [9] as outcome in the analysis, which is computed conditioning on test samples' binary phenotypes and available disease status of parents and siblings through Monte Carlo integration.

Linear mixed model (LMM) for adjusted phenotype

We denote X_i as a (p+1)-vector of covariates with the intercept, and G_i as the allele counts for the variant to be tested. We consider the following linear model:

$$E(Z_i) = X_i\alpha + G_i\beta + b_i,$$

where α is a $(p+1)$-vector of fixed effect coefficients, β is a genetic effect coefficient, and b_i the random effect term for the *i*th individual with $b = (b_1, \dots, b_N)^T$. We assume the random effect to follow a multivariate Gaussian distribution $b \sim N(0, \tau_0 I + \sum_{k=1}^{K} \tau_k V_k)$, where τ_0 is the variance component parameter for a noise term. Parameters for other variance components are denoted as τ_k, and V_k are pre-specified $N \times N$ correlation matrices.

To better capture phenotype correlation, we use a variance component of sparse kinship in addition to the commonly used genetic relationship matrix (GRM), i.e., $K = 2$ and $\Sigma = \tau_0 I + \tau_1 V_1 + \tau_2 V_2$, where V_1 is a sparse matrix of the estimated kinship coefficients after thresholding, and V_2 is GRM computed from genetic variants. The inclusion of the sparse kinship matrix as an additional variance component can be justified by the observation that the phenotype adjustment using family disease information increases the concordance among related individuals. For example, the adjusted phenotype for a control sibling pair would be identical as they share the same parental disease status (**Supplementary Figure 1**). Such phenotypic concordance is not sufficiently captured by GRM alone and can lead to mis-calibration as pointed out by Hujoel, Margaux LA, et al.[9]. It is also shown that incorporating pedigree structure as a variance component in linear mixed models improves association outcomes[27, 28].

Parameter estimation for the null model

Under the assumption of no genetic effects, the null model can be represented as

$$Z_i = X_i\alpha + b_i$$

Treating the adjusted phenotype Z as quantitative trait, the log likelihood of (α, τ) with random effect b integrated out in REML is

$$\ell(\alpha, \beta = 0, \tau) = c - \frac{1}{2}(\log|\Sigma| + \log|X^T \Sigma^{-1} X| + Z^T P Z)$$

where c is a constant, $\Sigma = \tau_0 I + \sum_{k=1}^{2} \tau_k V_k$, $P = \Sigma^{-1} - \Sigma^{-1} X (X^T \Sigma^{-1} X)^{-1}$. Model parameters (α, β, τ) are estimated iteratively with a working model $\tilde{Z} = X\hat{\alpha}^{(l)} + \hat{b}^{(l)}$ for iteration l. Let $\hat{\Sigma} = \hat{\tau}_0 I + \sum_{k=1}^{2} \hat{\tau}_k V_k$ be the working variance matrix. The first derivatives of $\ell(\alpha, \beta = 0, \tau)$ with respect to τ are:

$$\frac{\partial \ell(\alpha, \beta = 0, \tau)}{\partial \tau_k} = \frac{1}{2}[\tilde{Z}^T P V_k P \tilde{Z} - tr(P V_k)]$$

For each iteration, variance components $\hat{\tau}$ are updated using AI-REML algorithm[14], in which the Hessian is approximated by an average information matrix AI with its entries expressed as:

$$AI_{\tau_k \tau_l} = \frac{1}{2} \tilde{Z}^T \hat{P} V_k \hat{P} V_l \hat{P} \tilde{Z},$$

where $\hat{P} = \hat{\Sigma}^{-1} - \hat{\Sigma}^{-1} X (X^T \hat{\Sigma}^{-1} X)^{-1}$. Then the variance component parameters are updated by $\tau_k^{new} = \tau_k^{old} + \{AI\}^{-1} \frac{\partial \ell(\alpha, \beta = 0, \tau^{old})}{\partial \tau_k}$.

Both the first derivative and the approximated second derivative involves matrix inverse of $\hat{\Sigma}$, which can be computationally heavy when N is large. To reduce the computational burden, the PCG method[15] with Jacobi preconditioner is adopted, which avoids directly calculating matrix inverse by finding solutions of linear systems and involves only matrix multiplication. Since $\hat{\Sigma}$ is a linear combination involving two components V_1 and V_2 in our setting, matrix multiplication with regard to the two parts can be calculated

separately. For V_1 which is a sparse matrix representing close relatedness up to the third degree, the computation cost is further lowered by scanning through the non-zero elements of V_1 only. For V_2 which represents genetic relatedness, we improve the memory usage by calculating its elements in runtime instead of using a pre-computed $N \times N$ GRM matrix. Thus, the overall time complexity for null model estimation is reduced from $\mathcal{O}(N^3)$ to $\mathcal{O}(B(M_{GRM} + C_{sparse})N^{1.5})$, where B is the number of iterations until the algorithm reaches convergence, C_{sparse} is the number of non-zero elements of sparse kinship matrix, M_{GRM} is the number of variants included in the GRM construction. Here we assume that PCG algorithm has complexity $\mathcal{O}(N^{0.5})$.[16]

To avoid double-fitting the candidate variant in the model and GRM, leave-one-chromosome-out (LOCO) scheme was implemented in both the null model parameter estimation step and the score statistic calculation step of the proposed method.

Single variant association test with empirical SPA

The score statistic for testing the null hypothesis $H_0: \beta_j = 0$ for variant j is $T = \tilde{G}_j^T(Z - \hat{\mu})$, where $\tilde{G}_j$ is an N-vector of covariate-adjusted genotypes. Under the null hypothesis, the variance of the statistic is $Var(T) = G_j^T \hat{P} G_j$. For computational efficiency, $Var(T)$ can be approximated using $Var(T)^* = \tilde{G}_j^T \tilde{G}_j$ combined with a calibration factor $r = \frac{Var(T)}{Var(T)^*}$ estimated using a subset of SNP data[16, 17, 29]. The variance-adjusted statistic after calibration is $T' = \frac{\tilde{G}_j^T(Z-\hat{\mu})}{\sqrt{\hat{r}\tilde{G}_j^T \tilde{G}_j}}$. For the proposed method, 30 SNPs were used to obtain the estimated calibration factor $\hat{r}$.

When Z is unbalanced and a variant has low minor allele count, the distribution of T' deviates from the Gaussian distribution especially at the tails, thus the usual test of the score statistic against a Gaussian distribution can result in type I error inflation. Saddlepoint approximation is shown to improve over

normal approximation in such conditions by utilizing the entire cumulant-generating function (CGF)[30, 31]. Fixing $\tilde{G}_j$, T' can be viewed as a weighted sum of residuals $Z - \hat{\mu}$, yet the adjusted phenotype Z has an intractable distribution which makes it impossible to derive the explicit cumulant-generating function (CGF).

Alternatively, we use the empirical version of saddlepoint approximation[10, 11] as a nonparametric estimator for the distribution of the test statistic. The empirical CGF approach has been utilized in methods such as SPACox[12] for an altered version of the test statistic and was shown to provide calibrated p-values. The empirical estimator for CGF of T' is $\widehat{K}(\xi) = \log(\frac{1}{N}\sum_{i=1}^{N} e^{\xi t_i})$ where t_i is the residual of the *i*th individual from Step1. The empirical approximation of the first and second derivative are $\widehat{K}'(\xi) = \frac{\sum_{i=1}^{N} e^{\xi t_i} t_i}{\sum_{i=1}^{N} e^{\xi t_i}}$ and $\widehat{K}''(\xi) = \frac{\sum_{i=1}^{N} e^{\xi t_i} t_i^2}{\sum_{i=1}^{N} e^{\xi t_i}} - \widehat{K}'(\xi)\widehat{K}^2(\xi)$ respectively. Suppose $\hat{\xi}$ is a value satisfying the equation $\widehat{K}'(\hat{\xi}) = q$, the p-value can be calculated by the following formula[32]

$$pr(T' > q) \approx 1 - \Phi\{w + \frac{1}{w}\log\frac{v}{w}\}$$

where $w = \mathrm{sign}(\hat{\xi})\sqrt{2[\hat{\xi}q - \widehat{K}(\hat{\xi})]}$, $v = \hat{\xi}\sqrt{\widehat{K}''(\hat{\xi})}$, Φ is the cumulative distribution function of the standard normal distribution.

Simulation studies

We performed simulation studies to evaluate type I error rate and power of the proposed method. To simulate a population of size N with p_r% relatedness, we set $p_r N/100$ to be related individuals with a specified relatedness structure, and the rest $1 - p_r N/100$ to be independent individuals. Given that the relatedness structure is sibling pairs, the simulation process proceeds as follows: First, sequences of M

variants for both parents of each sample individual were simulated independently with pre-specified MAFs. Genotypes for N sample individuals (offsprings) were then generated using $N_R = 2N$ parental genotypes. Binary phenotypes for sample individuals and parents were simulated from $Bernoulli(\mu_i)$ with μ_i from a logistic mixed model

$$logit(\mu_i) = \alpha_0 + X_i + G_i\beta + b_i$$

where for individual i $(i = 1, \ldots, 3N)$, X_i is a covariate randomly sampled from $Normal(0,1)$, G_i is the genotypes of the M variants, α_0 is the intercept determined by prevalence k, β is a vector of log odds ratio of genetic effects, and b_i is a random effect with underlying distribution $Normal(0, \tau K)$ depending on the true underlying kinship coefficient matrix K. Given the kinship coefficient φ_{ij} between individual *i* and individual *j*, the value for an element in *K* is $K_{ij} = 2\varphi_{ij}$.

Type I error rates were evaluated at significant level $\alpha = 5 \times 10^{-8}$ for simulated data sets with 10^9 independent null SNPs, and sample size 10,000 at case-control ratio of 1:99, 5:95 and 10:90. All SNPs were generated with minor allele frequency 0.1 while phenotypes were generated given $\tau = 1$, corresponding to liability-scale heritability 0.23[17]. We considered two types of relatedness structure for the simulated population. The first one consists of 5,000 independent individuals and 2,500 sibling pairs ($p_r = 50\%$). The second one is a mixture of independent individuals and families with 8 members in each family. The pedigree for the 8-member family was shown in **Supplementary Figure** 5. To obtain a sample size of 10,000 with 50% related individuals, 625 families were simulated with the 8-member pedigree, while the rest 5,000 are independent individuals. Four methods were compared: SAIGE, LT-FH, TAPE-LTFH and TAPE. Note that for the setting with 8-member families, TAPE's phenotype adjustment takes into account all relatedness in the pedigree, while LT-FH only integrates information from parents and sibling of an individual.

Power of the tests was assessed using simulated data sets with 10,000 individuals and 100,000 variants for each setting with 1% variants selected as causal variants. We calculated the χ^2 statistics from SAIGE, LT-FH, TAPE-LTFH and TAPE and compared between causal and non-causal variants. Genetic effect sizes ranged from 0.4 to 2.3 and three case-control ratio settings were considered, i.e., 1:99, 5:95, and 10:90. We generated 100 replications for each setting and compared the empirical power at significance level $\alpha = 5 \times 10^{-8}$ over SAIGE, LT-FH, TAPE-LTFH and TAPE.

Computation time

Computation time was evaluated with $M = 100{,}000$ variants and sample size N ranging from 10,000 to 408,898 sampled from white British individuals in UK Biobank data for Type II diabetes (case:control=1:20). Projected time for the analysis of 21 million variants with MAF $\geq 0.01\,\%$ was calculated based on the evaluation results. Two other methods were evaluated in addition to TAPE: analysis of binary phenotypes by SAIGE and analysis of LT-FH phenotypes by BOLT-LMM. All evaluations were computed on an Intel(R) Xeon(R) Gold 6152 CPU.

UK Biobank data

Over 21 million genetic variants imputed from the Haplotype Reference Consortium (HRC)[18] and with minor allele frequency (MAF) $\geq 0.01\%$ were used for the association analysis among a sample population of 408,898 white British individuals. NCBI Build 37/UCSC hg19 was adopted for genomic coordinates. A total of 10 binary traits with available parental disease status were analyzed, where the binary traits for genotyped individuals were defined by the PheWAS codes[17] aggregated from ICD9 and ICD10 codes in the UK Biobank. Parental phenotypes were extracted from data fields for self-reported paternal and maternal

illness. We included sex, age and first 10 principal components as covariates to adjust for. GRM was constructed using 93,511 genotyped variants suggested by UK Biobank[17, 33]. Kinship coefficients were estimated using the KING software[19], and the sparse kinship matrix was constructed using those with estimated kinship no larger than third-degree relatedness.

The KoGES data

For the association analysis among a sample population of 72,298 Korean individuals, over 8 million genetic variants were imputed from 1,000 Genome project phase 3 + Korean reference genome (397 samples) and with minor allele frequency (MAF) > 1%[3]. Two binary traits (diabetes and gastric cancer) with different case-control ratios were analyzed. Phenotypes for both genotyped individuals and their relatives are self-reported survey data. We adjusted for sex, age, first 10 principal components, and 34 indicator variables of batch information (cohort × collection year). GRM was constructed using 327,540 genotyped variants. The sparse GRM was constructed using SAIGE with pairwise relatedness coefficients larger than 0.1.

FIGURES & TABLES

Figure 1. Analytical framework of TAPE. In Step 0, latent disease risk of individuals is estimated from observed phenotypes and family disease history using a weighted proportion of the affected close relatives to the individual. In Step 1, a null linear mixed model is fit with covariates and two random effects with the sparse kinship matrix and the dense genetic relatedness matrix (GRM) as covariance structures. In Step 2, p-values score test is performed for each genetic variant using empirical saddlepoint approximation. GWAS results from TAPE yields higher detection power while maintaining good calibration among related individuals.

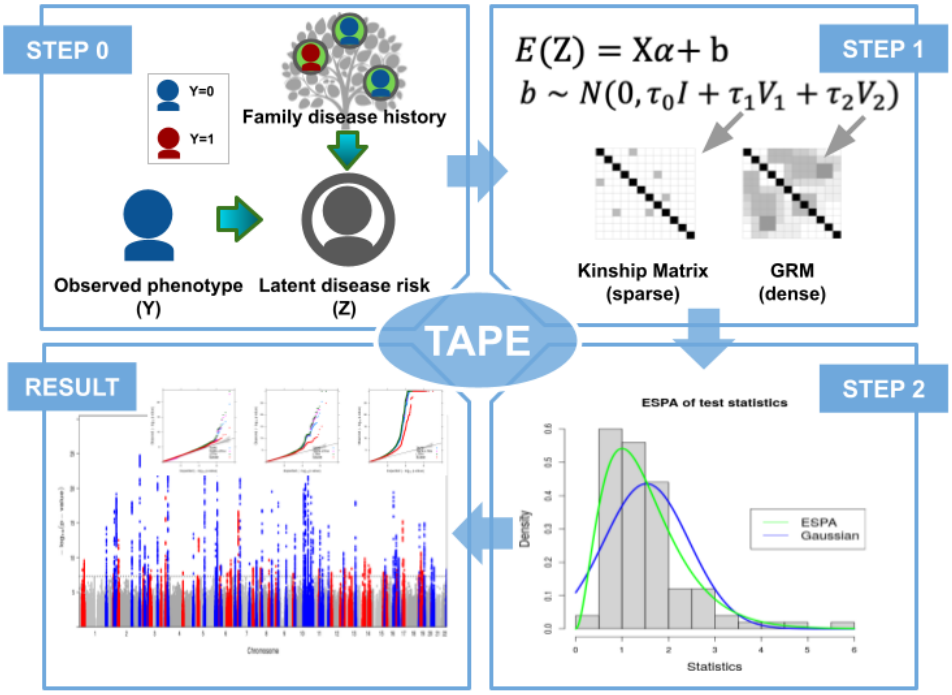

Figure 2. Average χ^2 values of causal variants with sample size N=10,000 (5,000 independent individuals and 2,500 pairs of siblings), comparing TAPE, TAPE-LTFH, LTFH and SAIGE. For each dataset, 100,000 independent variants were simulated and 1% variants were selected as causal variants with 4 different effect sizes. A total of 100 datasets were generated to calculate average χ^2 values. MAFs of variants were 0.1.

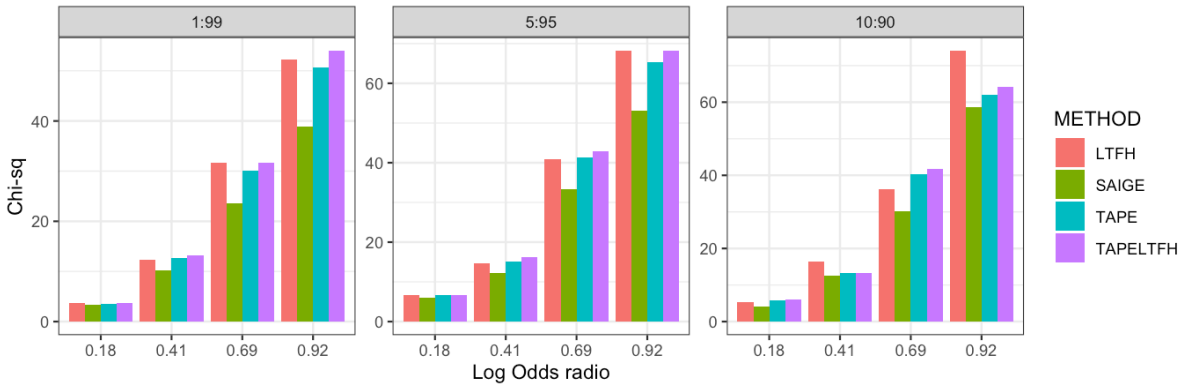

Figure 3. Manhattan plot for the UK Biobank association test results from SAIGE (first row), LT-FH (second row), TAPE-LTFH (third row) and TAPE (fourth row) among white British (N=408,898). **a**: Type II diabetes (Phecode 250.2); **b**: Parkinson's disease (Phecode 332). For plots from TAPE, red marks clumped significant variants from TAPE that were not detected by SAIGE; blue marks clumped significant variants detected by both TAPE and SAIGE. Significant clumped variants are identified using a window width of 5Mb and a linkage disequilibrium threshold of 0.1.

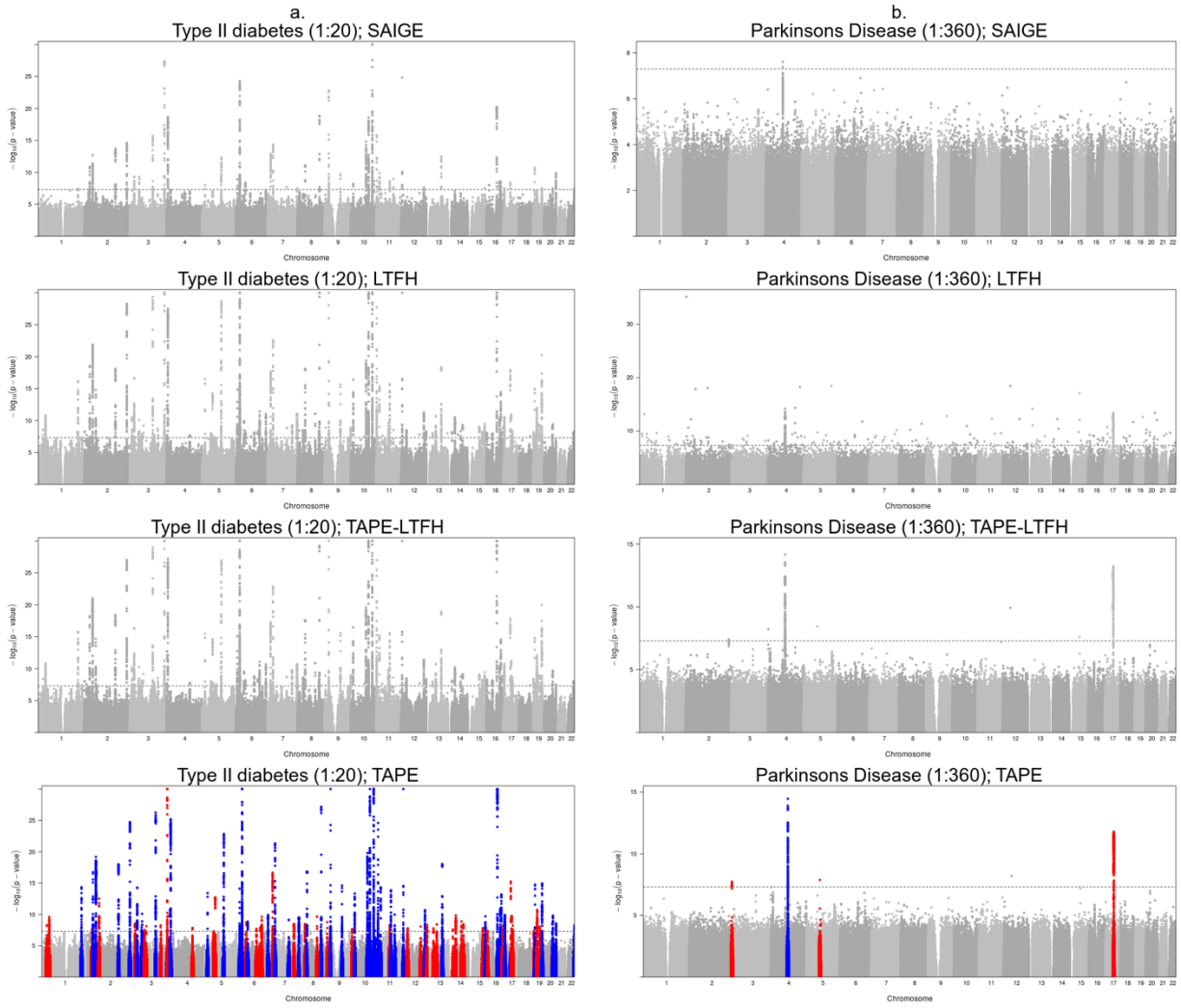

Figure 4. Q-Q plot for the UK Biobank association test results from SAIGE, LT-FH, TAPE-LTFH and TAPE among white British (N=408,898), categorized by MAF. Up: Type II diabetes (Phecode 250.2); Bottom: Parkinson's disease (Phecode 332)

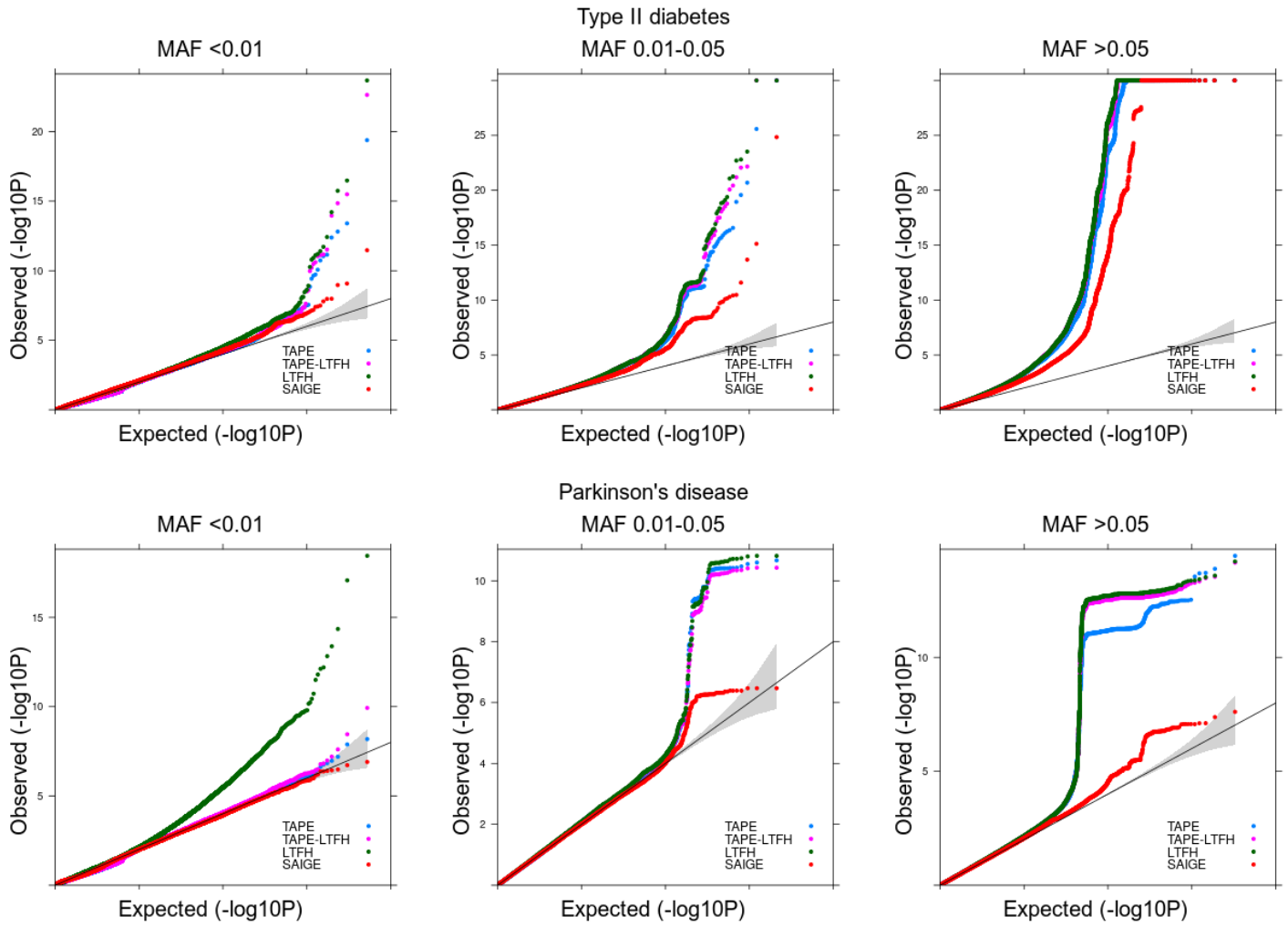

Table 1. Empirical type I error rates for TAPE, TAPE-LTFH, LT-FH and SAIGE, estimated using 10^9 independent SNPs and a sample size of 10,000 ($\alpha = 5 \times 10^{-8}$). **a**: Sample consists of 2,500 pairs of siblings and 5,000 independent individuals; **b**: Sample consists of 625 8-member families and 5,000 independent individuals.

a.

Case:Control	MAF	TAPE	TAPE-LTFH	LTFH	SAIGE
1:99	0.001	4.977e-08	1.019e-07	5.928e-06	4.418e-08
5:95	0.001	5.115e-08	8.275e-08	1.252e-06	4.368e-08
10:90	0.001	5.476e-08	7.452e-08	5.489e-07	4.641e-08
1:99	0.01	5.455e-08	1.069e-07	1.409e-07	3.963e-08
5:95	0.01	5.143e-08	1.158e-07	1.940e-07	4.341e-08
10:90	0.01	5.459e-08	9.086e-08	1.141e-07	4.980e-08
1:99	0.10	5.007e-08	1.275e-07	1.500e-07	3.964e-08
5:95	0.10	5.213e-08	1.639e-07	1.238e-07	4.355e-08
10:90	0.10	6.416e-08	7.782e-08	7.232e-08	4.650e-08

b.

Case:Control	MAF	TAPE	TAPE-LTFH	LTFH	SAIGE
1:99	0.001	3.329e-08	9.028e-08	4.446e-06	3.832e-08
5:95	0.001	3.051e-08	6.563e-08	8.171e-07	4.245e-08
10:90	0.001	2.967e-08	5.145e-08	3.751e-07	4.721e-08
1:99	0.01	3.742e-08	9.792e-08	4.818e-07	4.547e-08
5:95	0.01	3.156e-08	7.906e-08	1.463e-07	4.311e-08
10:90	0.01	2.978e-08	6.215e-08	8.811e-08	4.324e-08
1:99	0.10	3.113e-08	7.730e-08	1.000e-07	3.895e-08
5:95	0.10	3.050e-08	7.983e-08	6.025e-08	4.232e-08
10:90	0.10	3.163e-08	6.372e-08	5.857e-08	4.546e-08

Table 2. Summary of 10 traits in UK Biobank

Trait	Phecode	Case:Control	Parental Prevalence
Parkinson's Disease	X332	1:360	0.0186
Dementias	X290.1	1:406	0.0609
Lung Cancer	X165.1	1:181	0.0604
Depression	X296.2	1:33	0.0462
Type II Diabetes	X250.2	1:20	0.0845
Hypertension	X401	1:4	0.2388
Chronic Bronchitis	X496.2	1:136	0.0785
Colorectal Cancer	X153	1:87	0.0499
Ischemic Heart Disease	X411	1:11	0.2373
Cerebral Ischemia	X433.3	1:138	0.1348