



Genetics and population analysis

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

Yongwen Zhuang^{1,2}, Brooke N. Wolford³, Kisung Nam⁴, Wenjian Bi⁵, Wei Zhou ⁶,
Cristen J. Willer^{3,7,8}, Bhramar Mukherjee^{2,9,10} and Seunggeun Lee ^{4,*}

¹Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI, USA, ²Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA, ³Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA, ⁴Graduate School of Data Science, Seoul National University, Seoul, Korea, ⁵Department of Medical Genetics, School of Basic Medical Sciences, Peking University, Beijing, China, ⁶Massachusetts General Hospital, Broad Institute, Boston, MA, USA, ⁷Department of Internal Medicine, Division of Cardiology, University of Michigan Medical School, Ann Arbor, MI, USA, ⁸Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI, USA, ⁹Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, MI, USA and ¹⁰Michigan Institute of Data Science, University of Michigan, Ann Arbor, MI, USA

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on December 5, 2021; revised on May 22, 2022; editorial decision on June 13, 2022

Abstract

Motivation: 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.

Results: We propose a new method for genetic association test with family disease history, 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, accounts for case–control imbalance by using empirical saddlepoint approximation, and is flexible to incorporate any existing adjusted phenotypes, such as phenotypes from the LT-FH method. We show through simulation studies and analysis of UK Biobank data of white British samples and the Korean Genome and Epidemiology Study of Korean samples that the proposed method is robust and yields better calibration compared to existing methods while gaining power for detection of variant–phenotype associations.

Availability and implementation: The summary statistics and code generated in this study are available at <https://github.com/styvon/TAPE>.

Contact: lee7801@snu.ac.kr

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

The Genome-wide and phenome-wide studies are facilitated by the recent development of large-scale biobanks, such as the UK Biobank (UKB) (Bycroft *et al.*, 2018), BioBank Japan (Nagai *et al.*, 2017) and the Korean Genome and Epidemiology Study (KoGES) (Kim *et al.*, 2017). Individuals in the biobanks are samples from a target population and large numbers of phenotypes are collected for each individual, which allows phenome-wide scans. 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 (Gudbjartsson *et al.*, 2008; Kong *et al.*, 2009; Thornton and McPeck, 2007; Zhong *et al.*, 2016). 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 (Liu et al., 2017). The LT-FH method (Huijoe et al., 2020) 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 Fig. S1) can lead to type I error inflation. Huijoe et al. (2020) 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. 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 UKB, 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 Figs S2 and S3). 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, mixed-model-based Test with Adjusted Phenotype and Empirical saddlepoint approximation (TAPE), which controls for increased phenotype correlation and case-control imbalance. In standard mixed-model methods, only a dense genetic relatedness matrix (GRM) 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 LMM (Bi et al., 2020; Davison and Hinkley, 1988; Feuerverger, 1989). We show through simulation studies and analysis of UKB that the proposed method is robust and yields better calibration compared to existing methods while gaining power for detection of variant-phenotype associations.

2 Materials and methods

2.1 Overview of methods

The TAPE method takes a three-step framework (Fig. 1): (i) 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; (ii) fit a two-variance-components null LMM to obtain parameter estimates; and (iii) test for genetic association using score test with empirical saddlepoint approximation.

In Step 1, the phenotypes are adjusted using inferred risk of individuals. TAPE-WP uses a weighted proportion of the affected close relatives to the control, which can be viewed as an extension of the GWAX method⁸ to further differentiate disease risk of controls based on family disease history configurations. TAPE-LTFH uses the liability of diseases generated from the existing LT-FH method as the adjusted phenotypes.

In Step 2, we fit the null LMM with two random effects, the first uses the sparse kinship matrix (Jiang et al., 2019), and the second uses the dense GRM. These two-variance-components can capture both increased correlation in phenotypes due to phenotype adjustment procedure and distance genetic relatedness. To make the method scalable, average information restricted maximum likelihood (AI-REML) (Gilmour et al., 1995), with preconditioned conjugate gradient (PCG) method (Hestenes and Stiefel, 1952) similar to that used in BOLT-LMM (Loh et al., 2015) and SAIGE, is used.

In Step 3, a 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

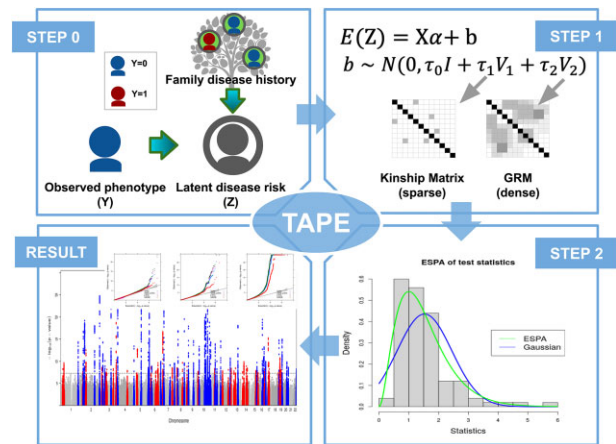


Fig. 1. Analytical framework of TAPE. In Step 1, latent disease risk of individuals is estimated from observed phenotypes and family disease history using a weighted proportion of the affected close relatives to the individual. In Step 2, a null LMM is fit with covariates and two random effects with the sparse kinship matrix and the dense GRM as covariance structures. In Step 3, P -values score test is performed for each genetic variant using empirical saddlepoint approximation

distribution, we approximate the distribution by empirical saddlepoint approximation (Feuerverger, 1989), which uses an empirically estimated cumulant-generating function (CGF) to calculate P -value. The empirical saddlepoint approximation is utilized when the test statistic exceeds 2 SD of the mean. Time complexity for this step is $O(MN)$.

2.2 Phenotype adjustment

We first introduce the proposed phenotype adjustment procedure (Step 1) in TAPE. For TAPE-WP, we assume a sample of N individuals where each individual has N_{R_i} relatives with phenotypic information ($i \in \{1, \dots, N\}$), F_{ij} denotes the kinship coefficient between individual i and relative j ($i \in \{1, \dots, N\}; j \in \{1, \dots, N_{R_i}\}$), D_{ij} denotes 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 $r_i = \frac{\sum_{j=1}^{N_{R_i}} F_{ij} \mathbb{I}(D_{ij}=1)}{\sum_{j=1}^{N_{R_i}} F_{ij}}$. If

$Y_i = 0$ and all N_{R_i} relatives of the i th individual are cases, the latent disease risk is $Z_i = \rho$. For the analysis in this article, 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 Notes S1 and S2 for details.

2.3 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 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 the GRM. 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 Fig. S1). Such phenotypic concordance is not sufficiently captured by GRM alone and can lead to mis-calibration as pointed out by Hujuel *et al.* (2020). It is also shown that incorporating pedigree structure as a variance component in LMMs improves association outcomes (Tucker *et al.*, 2015; Zaitlen *et al.*, 2013).

2.4 Parameter estimation for the null model

In Step 2, we fit a LMM under the following null hypothesis

$$E(Z_i) = X_i \alpha + b_i.$$

Treating Z as a quantitative trait, the marginal log likelihood of (α, τ) 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, $\Sigma = \tau_0 I + \sum_{k=1}^2 \tau_k V_k$, $P = \Sigma^{-1} - \Sigma^{-1} X (X^T \Sigma^{-1} X)^{-1}$. Parameters (α, β, τ) are estimated iteratively with a working model $\tilde{Z} = X \tilde{\alpha} + \tilde{b}$ for iteration l . Let $\tilde{\Sigma} = \tilde{\tau}_0 I + \sum_{k=1}^2 \tilde{\tau}_k V_k$ be the working variance matrix. The score function with respect to τ are:

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

For each iteration, variance components $\tilde{\tau}$ are updated using AI-REML algorithm (Gilmour *et al.*, 1995), in which the Hessian is approximated by an average information matrix, AI, with its entries expressed as:

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

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

Both the score and AI matrix involve $\tilde{\Sigma}^{-1}$, which is computationally heavy when N is large. To reduce the computational burden, the PCG method (Hestenes and Stiefel, 1952) with Jacobi preconditioner is adopted, which avoids directly calculating matrix inverse by finding solutions of linear systems and involves only matrix multiplication. Since $\tilde{\Sigma}$ is a linear combination of V_1 and V_2 , matrix multiplication with regard to each part can be calculated separately. For V_1 , the computation cost is further lowered by using the sparsity. For V_2 , we reduce the memory usage by calculating its elements in runtime instead of using a pre-computed $N \times N$ GRM matrix. The overall time complexity for null model estimation is $\mathcal{O}(B(M_{\text{GRM}} + C_{\text{sparse}})N^{1.5})$, 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 $\mathcal{O}(N^{0.5})$ (Loh *et al.*, 2015). To avoid double-fitting the candidate variant in the model and GRM, leave-one-chromosome-out scheme was implemented.

2.5 Single variant association test with empirical SPA

In Step 3, we use the score test to calculate the P -value for association of each genetic variant. The score statistic for testing $H_0 : \beta_j = 0$ 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 $\text{Var}(T) = G_j^T \hat{P} G_j$. For computational

efficiency, $\text{Var}(T)$ can be approximated using $\text{Var}(T)^* = \tilde{G}_j^T \tilde{G}_j$ combined with a calibration factor $r = \frac{\text{Var}(T)}{\text{Var}(T)^*}$ estimated using a subset of single-nucleotide polymorphism (SNP) data (Loh *et al.*, 2015; Svischcheva *et al.*, 2012; Zhou *et al.*, 2018). The variance-adjusted statistic after calibration is $T' = \frac{\tilde{G}_j^T (Z - \hat{\mu})}{\sqrt{\tilde{r} \tilde{G}_j^T \tilde{G}_j}}$. For the proposed method,

30 independent SNPs were randomly chosen to obtain the estimated calibration factor \tilde{r} . The number of SNPs were chosen such that the estimated value is stable within a given variation threshold. The TAPE program is capable of adaptively increase the number of SNPs to reach a stable estimate.

When Z is unbalanced and a variant has low minor allele count, using a Gaussian distribution to calculate a P -value of T' can result in type I error inflation. Saddlepoint approximation is shown to improve over Gaussian approximation in such conditions by utilizing the entire CGF (Daniels, 1954; Jensen, 1995). Fixing 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 CGF.

Alternatively, we use the empirical version of saddlepoint approximation (Davison and Hinkley, 1988; Feuerverger, 1989) as a non-parametric estimator for the distribution of the test statistic (Bi *et al.*, 2020). The empirical estimator for CGF of T' is $\hat{K}(\xi) = \log(\frac{1}{N} \sum_{i=1}^N e^{\xi t_i})$, where t_i is the residual of the i th individual from Step 2. The empirical approximation of the first and second derivative is $\hat{K}'(\xi) = \frac{\sum_{i=1}^N e^{\xi t_i} t_i}{\sum_{i=1}^N e^{\xi t_i}}$ and $\hat{K}''(\xi) = \frac{\sum_{i=1}^N e^{\xi t_i} t_i^2}{\sum_{i=1}^N e^{\xi t_i}} - \hat{K}'(\xi)^2$, respectively. Suppose ξ^* is a value satisfying the equation $\hat{K}'(\xi^*) = q$, the P -value can be calculated by the following formula (Kuonen, 1999)

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

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

2.6 Simulation studies of type I error control and power

We considered two types of relatedness structures. The first one consists of 5000 independent individuals and 2500 sibling pairs ($p_r = 50\%$). The second one is a mixture of independent individuals and families with eight members in each family. The pedigree for the eight-member family was shown in Supplementary Figure S5. Binary phenotypes for sample individuals and parents were simulated from Bernoulli(μ_i) with μ_i from a logistic mixed model

$$\text{logit}(\mu_i) = \alpha_0 + X_i + G_i \beta + b_i,$$

where for individual i ($i = 1, \dots, 3N$), 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, τK) 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 $K_{ij} = 2\phi_{ij}$.

Simulation results for TAPE-WP and TAPE-LTFH were compared with two other methods: (i) GWAS with original binary phenotypes by SAIGE (Zhou *et al.*, 2018); and (ii) original LT-FH method that uses BOLT-LMM with LT-FH phenotypes (Hujuel *et al.*, 2020) for all individuals (hereafter denoted as LT-FH), which is shown to increase association power over GWAS (Liu *et al.*, 2017).

Type I error rates were evaluated with 10^9 independent null SNPs, and sample size 10000 at case-control ratio of 1:99, 5:95 and 10:90. Phenotypes were generated given $\tau = 1$, corresponding to liability-scale heritability 0.23 (Zhou *et al.*, 2018). To investigate

type I error rates by minor allele frequency (MAF), SNPs were generated with MAF 0.001, 0.01 and 0.1, respectively, for each simulated dataset.

Power of the tests was assessed using simulated datasets with 10 000 individuals and 100 000 variants with MAF 0.1 for each setting with 1% variants selected as causal variants. We calculated both the average χ^2 statistics for causal SNPs and the empirical power at empirical α level from SAIGE, LT-FH, TAPE-LTFH and TAPE-WP. 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 simulation scenario.

2.7 Computation time evaluation

Computation time was evaluated with $M = 100000$ variants and sample size N ranging from 10 000 to 408 898 sampled from white British individuals in UKB 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. All evaluations were computed on an Intel(R) Xeon(R) Gold 6152 CPU.

2.8 UK biobank data

Over 21 million genetic variants imputed from the Haplotype Reference Consortium (McCarthy et al., 2016) and with $MAF \geq 0.01\%$ 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 (Zhou et al., 2018). 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 UKB (Bycroft et al., 2017; Zhou et al., 2018). Kinship coefficients were estimated using the KING software (Manichaikul et al., 2010), and the sparse kinship matrix was constructed using those with estimated kinship no larger than third-degree relatedness. Calibration of the testing method was evaluated by the attenuation ratio obtained from stratified LD score regression (LDSC). The attenuation ratio is defined as (LDSC intercept-1)/(average $\chi^2 - 1$), with smaller values indicating better control of false positives.

2.9 KoGES data

For the association analysis among a sample population of 72 298 Korean individuals, over 8 million genetic variants were imputed from 1000 Genome project phase 3 + Korean reference genome (397 samples) and with $MAF > 1\%$ (Kim et al., 2017). 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 \times 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.

3 Results

3.1 Simulation study results

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, 0.01 and 0.001, respectively. We first simulated a population consisting of 2500 pairs of siblings and 5000 independent individuals (Table 1). The empirical type I error rates of LT-FH were significantly inflated under more unbalanced case-control ratio and lower MAF, while results from TAPE-WP and SAIGE were well calibrated. TAPE-LTFH also yielded better controlled type I error rates

Table 1. Empirical type I error rates for TAPE-WP, TAPE-LTFH, LT-FH and SAIGE, estimated using 10^9 independent SNPs and a sample size of 10 000 ($\alpha = 5 \times 10^{-8}$)

Case:control	MAF	TAPE-WP	TAPE-LTFH	LTFH	SAIGE
2500 pairs of siblings and 5000 independent individuals					
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
625 8-member families and 5000 independent individuals					
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

Note: Two types of population structure were considered: (i) sample consists of 2500 pairs of siblings and 5000 independent individuals; and (ii) sample consists of 625 8-member families and 5000 independent individuals.

than that of LT-FH, especially when the case-control ratio is more unbalanced (1:99). Further, we evaluated type I error rates with a more complex relatedness structure, i.e. a population consisting of 625 eight-member families and 5000 independent individuals (Table 1). 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 individual 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 under the eight-member family pedigree scenario. For TAPE-nok, the sparse kinship matrix was not included as an LMM variance component and inflated empirical type I error was observed (Supplementary Fig. S4). For TAPE-misk, the true kinship matrix of an eight-member family pedigree was replaced with a slightly mis-specified one (Supplementary Fig. S5) in Steps 1 and 2. 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 of causal SNPs (Fig. 2) and the proportion of causal SNPs significant at empirical α level (Supplementary Fig. S6) for simulated datasets with sample size 10 000 under different genetic effects and case-control ratio. For each dataset, 100 000 independent variants with MAF 0.1 were simulated in which 1% were causal, and we generated 100 datasets for each setting. TAPE-WP and TAPE-LTFH achieve greater detection power over SAIGE, with a 21.0% and 26.5% average increase in average χ^2 statistics, and a 18.3% and 22.4% average increase in proportion of causal SNPs detected, respectively.

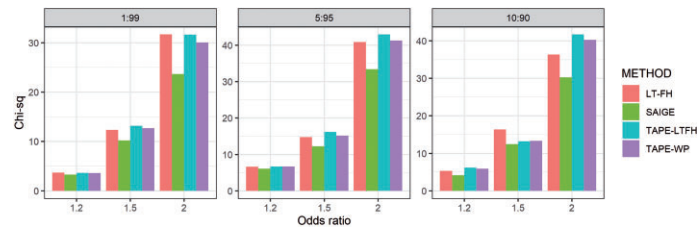


Fig. 2. Average χ^2 values of causal variants with $N = 10\,000$ (5000 independent individuals and 2500 pairs of siblings), comparing TAPE-WP, TAPE-LTFH, LT-FH and SAIGE. For each dataset, 100 000 independent variants were simulated and 1% variants were selected as causal variants with four different effect sizes. A total of 100 datasets were generated to calculate average χ^2 values. MAFs of variants were 0.1

LT-FH also had increased χ^2 over SAIGE by 27.5% and had a 22.1% average increase in detection rate, but 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 eight members (Supplementary Fig. S5). TAPE-WP achieves greater detection power over SAIGE GWAS results, with a 24.1% average increase in average χ^2 statistics and a 30.6% average increase in proportion of causal SNPs detected. TAPE-LTFH has higher overall power than both LT-FH and SAIGE under such relatedness structure (Supplementary Fig. S6), with a 27.5% average increase in average χ^2 statistics and a 40.8% average increase in proportion of significant SNPs detected over SAIGE, whereas for LT-FH the increase is 25.6% for average χ^2 statistics and 36.6% for proportion of significant SNPs detected.

In general, TAPE-WP and TAPE-LTFH 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 gains detection power. In addition, TAPE-LTFH achieved higher detection power than LT-FH, especially under the simulation scenario with more complex relatedness structure.

3.2 Computation time

Computation time was evaluated using randomly selected samples from 408 898 white British individuals in UKB 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 Fig. S7). Computation time for TAPE-LTFH is similar to that for TAPE-WP 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 S3. 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.

3.3 Analysis of binary traits in biobank data

We analyzed 10 binary disease outcomes with available parental disease status in the UKB (Table 2).

Figures 3 and 4 present 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 S8 and S9. For diseases, such as Parkinson's disease (Fig. 4), the observed quantile distribution of $-\log_{10}(p)$ corresponding to SNPs with $MAF < 0.01$ for LT-FH method in the Q-Q plot curved off in the middle of the graph, indicating potential type I error inflation due to unaccounted-for relatedness structure. Similar problematic patterns can also be found in LT-FH Q-Q plots for lung cancer, depression, chronic bronchitis, colorectal cancer and cerebral ischemia

Table 2. Summary of 10 traits in UKB

Trait	Phecode	Case:control	Parental prevalence
Parkinson's disease	332	1:360	0.0186
Dementias	290.1	1:406	0.0609
Lung cancer	165.1	1:181	0.0604
Depression	296.2	1:33	0.0462
Type II diabetes	250.2	1:20	0.0845
Hypertension	401	1:4	0.2388
Chronic bronchitis	496.2	1:136	0.0785
Colorectal cancer	153	1:87	0.0499
Ischemic heart disease	411	1:11	0.2373
Cerebral ischemia	433.3	1:138	0.1348

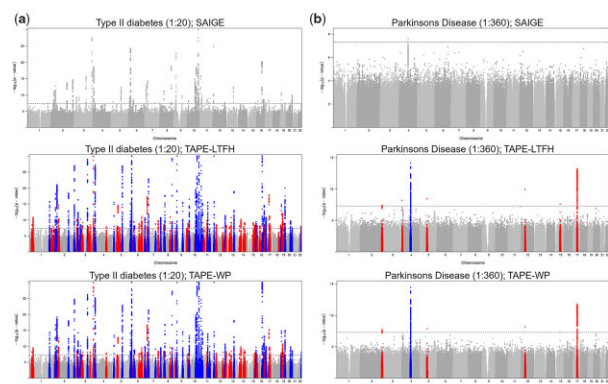


Fig. 3. Manhattan plot for the UKB association test results from SAIGE (first row), TAPE-LTFH (second row) and TAPE-WP (third row) among white British ($N = 408\,898$). Left: type II diabetes (Phecode 250.2); right: Parkinson's disease (Phecode 332). Significant clumped variants are identified using a window width of 5 Mb and a linkage disequilibrium threshold of 0.1

(Supplementary Fig. S8), but not in plots for TAPE-WP and TAPE-LTFH.

To assess the calibration of testing methods, we performed stratified LDSC with the baselineLD model to obtain the attenuation ratios (Finucane *et al.*, 2015) (Supplementary Table S2). For traits with more unbalanced case-control, TAPE-WP consistently yields relatively lower attenuation ratios than TAPE-LTFH, while LT-FH 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-WP, 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-WP, 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

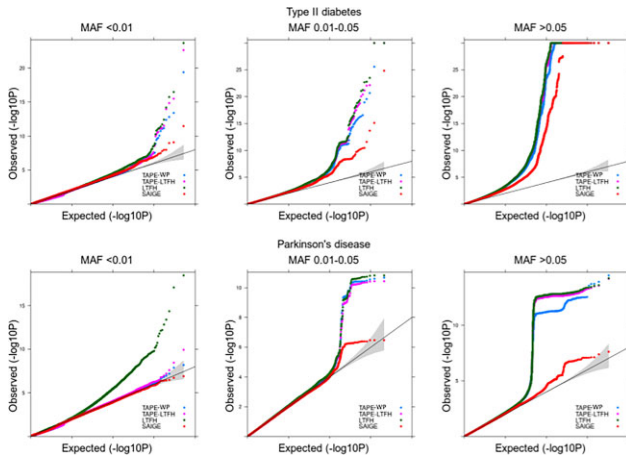


Fig. 4. Q-Q plot for the UKB association test results from SAIGE, LT-FH, TAPE-LTFH and TAPE-WP among white British ($N = 408\,898$), categorized by MAF. Up: type II diabetes (Phcode 250.2); bottom: Parkinson's disease (Phcode 332)

(Huijool *et al.*, 2020). On the contrary, TAPE-WP is able to generate better calibrated results under such situations, followed by TAPE-LTFH. Due to the above-mentioned potential type I error inflation of LT-FH method for samples with relatedness, we only compare the proposed method with SAIGE, which is also capable of handling related samples. [Supplementary Table S1](#) lists the number of significant variants and significant clumped variants at $\alpha = 5 \times 10^{-8}$ detected by TAPE-WP, TAPE-LTFH and SAIGE. Significant clumped variants were further identified by clumping genome-wide significant variants with 5 Mb window size and linkage disequilibrium threshold $r^2 = 0.1$ using PLINK software (Purcell *et al.*, 2007). TAPE-WP identified 84 more genome-wide significant clumped variants than SAIGE for type II diabetes, and 5 more for Parkinson's disease. For TAPE-LTFH, a total of 111 more genome-wide significant clumped variants were identified for type II diabetes, and 7 more for Parkinson's disease as compared to SAIGE. For all 10 diseases analyzed, a total of 663 genome-wide significant clumped variants were identified by TAPE-WP, including 33 clumped variants with $MAF < 1\%$; whilst a total of 344 clumped variants were identified by SAIGE, of which 25 were with $MAF < 1\%$. For TAPE-LTFH, a total of 726 genome-wide significant clumped variants were identified, including 63 clumped variants with $MAF < 1\%$.

For additional analysis, we applied TAPE-WP, 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 S4](#). For diabetes (case:control=1:12), TAPE-WP 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-WP and TAPE-LTFH identified three genome-wide significant clumped variants (rs760077, rs35972942 and 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 (Du *et al.*, 2020; Tanikawa *et al.*, 2018; Yan *et al.*, 2020), but not among Korean samples. Manhattan plots and Q-Q plots are presented in [Supplementary Figures S10 and S11](#).

4 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. Previous studies have shown that additional information from family disease history can help improve test power, yet challenges remain (i) to control for type I error inflation induced by increased correlation of phenotypes; and (ii) to account for unbalanced distribution of phenotypes

after being adjusted by family disease information. Our TAPE framework uses both a dense GRM and a sparse kinship matrix in the LMM to account for sample relatedness and family history-induced correlations. Empirical saddlepoint approximation 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.

For the null model, both sparse kinship matrix and GRM are included in the TAPE framework 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 (Speed and Balding, 2014; Widmer *et al.*, 2014), 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 (Huijool *et al.*, 2020), and indicates a possible solution to control for phenotypic correlation introduced by incorporating family disease information. When estimating variance parameters, the TAPE framework 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, TAPE-LTFH uses LT-FH phenotypes in the proposed two-variance-component mixed model. We show by simulation studies that TAPE-LTFH can better control for type I error inflation than LT-FH and achieves higher power. 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. When there is relatively high sample relatedness in the target population, TAPE-WP is recommended since it consistently controls type I error better than TAPE-LTFH and LT-FH while being capable of incorporating family disease history from the whole population. For studies with exploratory or discovery purposes, we would recommend TAPE-LTFH, as it increases power for detecting possible associations, and can keep type I error inflation on a controllable level.

For both TAPE-LTFH and LT-FH in the study, we used the posterior mean genetic liability from the LT-FH method proposed by Huijool *et al.* as the outcome variable for all individuals (including related ones), which is computed conditioning on test samples' binary phenotypes and family disease history. Since the original LT-FH study suggested against the use of LT-FH phenotypes for related individuals, we also evaluated the performance of two corresponding methods, TAPE-LTFHc and LT-FHc, which adjust phenotypes based on the presence of genetically related individuals in the data (details in [Supplementary Note S3](#)). Simulation studies ([Supplementary Note S3](#)) and UKB data analysis ([Supplementary Note S4](#)) show that this approach can help lowering type I error rates for both TAPE-LTFHc and LT-FHc, but at the cost of a decrease in detection power.

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 the TAPE framework. 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, the TAPE framework 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.

Acknowledgements

The authors acknowledge the National Institutes of Health (NIH) and the National Research Foundation of Korea (NRF) for the support. UK Biobank data were accessed under the accession number UKB: 45227. Data in this study were from the Korean Genome and Epidemiology Study (KoGES; 4851-302). National Research Institute of Health, Centers for Disease Control and Prevention, Ministry for Health and Welfare, Republic of Korea.

Authors' contributions

Y.Z., C.J.W. and S.L. designed the experiments. Y.Z., B.N.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.

Funding

This research was supported by the National Institutes of Health [grants R01-LM012535, R01-HG008773 to 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 to K.N. and S.L.).

Conflict of Interest: none declared.

Availability of data and code

The data and code underlying this article are available at <https://github.com/syvon/TAPE>.

References

- Bi, W. *et al.* (2020) A fast and accurate method for genome-wide time-to-event data analysis and its application to UK Biobank. *Am. J. Hum. Genet.*, **107**, 222–233.
- Bycroft, C. *et al.* (2017) Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv preprint at*, <https://doi.org/10.1101/166298>.
- Bycroft, C. *et al.* (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature*, **562**, 203–209.
- Daniels, H.E. (1954) Saddlepoint approximations in statistics. *Ann. Math. Stat.*, **25**, 631–650.
- Davison, A.C. and Hinkley, D.V. (1988) Saddlepoint approximations in resampling methods. *Biometrika*, **75**, 417–431.
- Du, M. *et al.* (2020) Remote modulation of lncRNA GCLET by risk variant at 16p13 underlying genetic susceptibility to gastric cancer. *Sci. Adv.*, **6**, eaay5525.
- Feuerverger, A. (1989) On the empirical saddlepoint approximation. *Biometrika*, **76**, 457–464.
- Finucane, H.K. *et al.*; RACI Consortium. (2015) Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.*, **47**, 1228–1235.
- Gilmour, A.R. *et al.* (1995) Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, **51**, 1440–1450.
- Gudbjartsson, D.F. *et al.* (2008) Many sequence variants affecting diversity of adult human height. *Nat. Genet.*, **40**, 609–615.
- Hestenes, M.R. and Stiefel, E. (1952) Methods of conjugate gradients for solving linear systems. *J. Res. Natl. Bur. Stand.*, **49**, 409–436.
- Hujoel, M.L. *et al.* (2020) Liability threshold modeling of case-control status and family history of disease increases association power. *Nat. Genet.*, **52**, 541–547.
- Jensen, J.L. (1995) *Saddlepoint Approximations*. Oxford University Press, Oxford.
- Jiang, L. *et al.* (2019) A resource-efficient tool for mixed model association analysis of large-scale data. *Nat. Genet.*, **51**, 1749–1755.
- Kim, Y. *et al.*; KoGES group. (2017) Cohort profile: the Korean Genome and Epidemiology Study (KoGES) consortium. *Int. J. Epidemiol.*, **46**, e20.
- Kong, A. *et al.*; DIAGRAM Consortium. (2009) Parental origin of sequence variants associated with complex diseases. *Nature*, **462**, 868–874.
- Kuonen, D. (1999) Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika*, **86**, 929–935.
- Liu, J.Z. *et al.* (2017) Case-control association mapping by proxy using family history of disease. *Nat. Genet.*, **49**, 325–331.
- Loh, P.-R. *et al.* (2015) Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.*, **47**, 284–290.
- Manichaikul, A. *et al.* (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics*, **26**, 2867–2873.
- McCarthy, S. *et al.*; Haplotype Reference Consortium. (2016) A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.*, **48**, 1279–1283.
- Nagai, A. *et al.*; BioBank Japan Cooperative Hospital Group. (2017) Overview of the BioBank Japan project: study design and profile. *J. Epidemiol.*, **27**, S2–S8.
- Purcell, S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
- Speed, D. and Balding, D.J. (2014) MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.*, **24**, 1550–1557.
- Svishcheva, G.R. *et al.* (2012) Rapid variance components-based method for whole-genome association analysis. *Nat. Genet.*, **44**, 1166–1170.
- Tanikawa, C. *et al.* (2018) Genome-wide association study identifies gastric cancer susceptibility loci at 12q24.11-12 and 20q11.21. *Cancer Sci.*, **109**, 4015–4024.
- Thornton, T. and McPeck, M.S. (2007) Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *Am. J. Hum. Genet.*, **81**, 321–337.
- Tucker, G. *et al.* (2015) Two-variance-component model improves genetic prediction in family datasets. *Am. J. Hum. Genet.*, **97**, 677–690.
- Widmer, C. *et al.* (2014) Further improvements to linear mixed models for genome-wide association studies. *Sci. Rep.*, **4**, 1–13.
- Yan, C. *et al.* (2020) Meta-analysis of genome-wide association studies and functional assays decipher susceptibility genes for gastric cancer in Chinese populations. *Gut*, **69**, 641–651.
- Zaitlen, N. *et al.* (2013) Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet.*, **9**, e1003520.
- Zhong, S. *et al.* (2016) CERAMIC: case-control association testing in samples with related individuals, based on retrospective mixed model analysis with adjustment for covariates. *PLoS Genet.*, **12**, e1006329.
- Zhou, W. *et al.* (2018) Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.*, **50**, 1335–1341.