Received: 11 September 2018 | Revised: 22 January 2020 | Accepted: 29 January 2020

DOI: 10.1111/sjos.12451

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Computationally efficient familywise error rate control in genome-wide association studies using score tests for generalized linear models

Kari Krizak Halle[1,2] | **Øyvind Bakke**[1] | **Srdjan Djurovic**[3,4] | **Anja Bye**[5] | **Einar Ryeng**[6] | **Ulrik Wisløff**[5] | **Ole A. Andreassen**[7,8] | **Mette Langaas**[1]

[1]Department of Mathematical Sciences, Norwegian University of Science and Technology

[2]Liaison Committee Between the Central Norway Regional Health Authority and the Norwegian University of Science and Technology

[3]Department of Medical Genetics, Oslo University Hospital

[4]NORMENT, Department of Clinical Science, University of Bergen

[5]The Cardiac Exercise Research Group, Department of Circulation and Medical Imaging, Norwegian University of Science and Technology

[6]Department of Cancer Research and Molecular Medicine, Norwegian University of Science Technology

[7]NORMENT Centre, University of Oslo

[8]Division of Mental Health and Addiction, Oslo University Hospital

Correspondence
Mette Langaas, Department of Mathematical Sciences, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway.
Email: mette.langaas@ntnu.no

Abstract

In genetic association studies, detecting phenotype–genotype association is a primary goal. We assume that the relationship between the data—phenotype, genetic markers and environmental covariates—can be modeled by a generalized linear model. The number of markers is allowed to be far greater than the number of individuals of the study. A multivariate score statistic is used to test each marker for association with a phenotype. We assume that the test statistics asymptotically follow a multivariate normal distribution under the complete null hypothesis of no phenotype–genotype association. We present the familywise error rate order k approximation method to find a local significance level (alternatively, an adjusted p-value) for each test such that the familywise error rate is controlled. The special case $k = 1$ gives the Šidák method. As a by-product, an effective number of independent tests can be defined. Furthermore, if environmental covariates and genetic markers are uncorrelated, or no environmental covariates are present, we show that covariances between score

statistics depend on genetic markers alone. This not only leads to more efficient calculations but also to a local significance level that is determined only by the collection of markers used, independent of the phenotypes and environmental covariates of the experiment at hand.

KEYWORDS

effective number of independent tests, FWER control, generalized linear model, GWAS, intersection approximation, The HUNT Study

1 | INTRODUCTION

A genetic marker is a DNA sequence at a known location on a chromosome. It may be a gene, but can, as in the data referred to in this article, be a single base-pair—a single-nucleotide polymorphism (SNP). In genome-wide association (GWA) studies, the aim is to test for association between genetic markers and a phenotype (an observable characteristic of an individual, e.g., a numerical variable, such as maximum oxygen uptake, or a categorical variable, such as presence of schizophrenia or bipolar disorder or absence of those diseases).

A large number of markers are tested, and multiple testing correction methods can be used to control the familywise error rate (FWER)—the probability of making one or more Type I errors—by specifying a local significance level (also known as genome-wide significance level) for the individual tests. In this work we present the *FWER order k approximation* method for finding a local significance level in multiple hypothesis testing (not to be confused with k-FWER control of the probability of k or more false rejections).

We assume independent individual observations in a case–control, cohort, or cross-sectional study. The phenotype of interest can be continuous or discrete. We consider biallelic genetic markers, giving three possible genetic variants, called genotypes. For each marker we assume the null hypothesis of "no association between phenotype and genotype" and a two-sided alternative. We model the data by a generalized linear regression model (GLM) with phenotype as response, marker genotypes as explanatory variables of interest, and possibly nongenetic explanatory variables, referred to as environmental covariates. In particular, a confounder such as population substructure (which may be associated with both phenotype and genotype) can be adjusted for by including principal components of the marker genotype covariance matrix as covariates (Price et al., 2006).The number of markers may be much greater than the number of data points, as the complete GLM is not fitted.

We use a score statistic for each marker separately for testing the marker's contribution to the model. The vector of score statistics asymptotically follows a multivariate normal distribution, with covariances that can be estimated from data (Conneely & Boehnke, 2007; Schaid, Rowland, Tines, Jacobson, & Poland, 2002; Seaman & Müller-Myhsok, 2005).

Let m be the number of markers. The FWER can be controlled at level α by using an appropriate local p-value cutoff, α_{loc}, for each of the m hypothesis tests. Inspired by the work of Moskvina and Schmidt (2008) and Dickhaus and Stange (2013), we will use an approximation to the m-dimensional asymptotic multivariate normal distribution of the score test statistics vector to calculate α_{loc}, or, alternatively to calculate FWER-adjusted p-values. In addition, α_{loc} together with α can be used to define an effective number of independent tests.

FWER order k approximation is more powerful than the Šidák method (which makes assumptions on the dependence structure among test statistics, allowing independence) and the Bonferroni method (which is valid for all dependence structures). It is more computationally efficient than the method of Conneely and Boehnke (2007), which is based on numerical integration in m dimensions, for which current algorithms have limited precision and are computationally intensive. The Westfall–Young permutation procedure is known to have asymptotically optimal power for a broad class of problems, including block-dependent and sparse dependence structure (Meinshausen, Maathuis, & Bühlmann, 2011). However, also this method is computationally intensive, and to have a valid permutation test, the assumption of exchangeability needs to be satisfied (Commenges, 2003). This assumption is in general not satisfied when environmental covariates are present in the GLM, but approximate methods exist in the case of a normally distributed response. There is, to our knowledge, no such simple approximation available in the general case, or in the special case of logistic regression. Finally, it should be mentioned that also Bayesian methods exist that include covariates and multiple markers (see Wakefield, 2009, for the logistic case).

We proceed (Section 2) to present the statistical background on the score test and derive expressions for the score test covariance matrix, which is of importance for the subsequent work. Next, our proposed method is presented in detail, together with characteristics of it (Section 3). The method is evaluated and compared to other methods by using two genetic datasets (Aspenes et al., 2011; Athanasiu et al., 2010; Djurovic et al., 2010; Loe, Rognmo, Saltin, & Wisløff, 2013), by simulations to assess asymptotic normality and validity of FWER approximations, and by using two artificial correlation structures (Section 4). Finally, a discussion and conclusion follow (Section 5).

2 | THE SCORE TEST IN GENERALIZED LINEAR MODELS FOR MULTIPLE HYPOTHESES

2.1 | Notation and data

We assume that a phenotype, m marker genotypes, and d environmental covariates are available from n independent individuals. Let $\boldsymbol{Y}$ be an n-dimensional vector having the phenotype Y_i of individual i as its ith entry, $i = 1, \dots, n$. Let X_e be an $n \times d$ matrix of rank d having environmental covariates (the first one being 1 to allow for an intercept in the model presented below) for individual i as its ith row, and let X_g be an $n \times m$ matrix having genetic covariates, or marker genotypes, for individual i as its ith row, each column corresponding to a genetic marker. We allow $m \gg n$.

We assume that the genetic data are from biallelic genetic markers with alleles a and A, and use the coding 0, 1, 2 for the genotypes aa, aA, and AA, respectively, in the genetic covariate matrix X_g (reflecting an additive genetic model), but other coding schemes are also possible (e.g., 0, 1, 1 to reflect a model in which A is dominant over a). We denote the total design matrix $X = (X_\mathrm{e}\ X_\mathrm{g})$, which has the total covariate vector for individual i as its ith row.

2.2 | Testing statistical hypotheses with the score test

We assume that the relationship between the phenotype $\boldsymbol{Y}$ and covariates X can be modeled by a GLM (McCullagh & Nelder, 1989) with an n-dimensional vector $\boldsymbol{\eta} = X_\mathrm{e}\boldsymbol{\beta}_\mathrm{e} + X_\mathrm{g}\boldsymbol{\beta}_\mathrm{g} = X\boldsymbol{\beta}$ of

linear predictors, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_e^T \ \boldsymbol{\beta}_g^T)^T$ is a $d+m$-dimensional parameter vector. Let η_i be the ith entry of $\boldsymbol{\eta}$, and let $\boldsymbol{\mu}$ be the n-dimensional vector having $\mu_i = EY_i$ as its ith entry. We assume that the link function of the GLM is canonical, which implies that the log-likelihood for individual i is $l_i = (Y_i\eta_i - b(\eta_i))/\phi + c(Y_i, \phi)$, where b and c are functions defining the exponential family of the phenotypes and ϕ the dispersion parameter. In general, $\mu_i = b'(\eta_i)$ and $\sigma_i^2 = \mathrm{Var}Y_i = \phi b''(\eta_i)$. For Y_i normally distributed, this reduces to $\sigma_i^2 = \sigma^2 = \phi$, and for Y_i Bernoulli distributed, $\sigma_i^2 = \mu_i(1-\mu_i)$ with $\phi = 1$.

As a starting point for a test statistic vector for testing whether components of $\boldsymbol{\beta}_g$ are equal to zero, we consider the full $d+m$-dimensional score vector

$$\boldsymbol{U} = \sum_{i=1}^{n} \nabla_{\boldsymbol{\beta}} l_i = \frac{1}{\phi} X^T(\boldsymbol{Y} - \boldsymbol{\mu}),$$

which is asymptotically normal with mean $\mathbf{0}$ and covariance matrix

$$V = \frac{1}{\phi^2} X^T \Lambda X,$$

where Λ is the diagonal matrix having σ_i^2 as its ii entry.

Under the complete null hypothesis $\boldsymbol{\beta}_g = \mathbf{0}$, the $\boldsymbol{\beta}_e$ parameters are still unknown and can be considered nuisance parameters, so $\boldsymbol{U}$ cannot be used directly as a test statistic vector. Therefore, we partition $\boldsymbol{U}$ into its environmental and genetic components, $\boldsymbol{U}^T = (\boldsymbol{U}_e^T \ \boldsymbol{U}_g^T)$, and replace $\boldsymbol{\beta}_e$ by its maximum likelihood estimate under the null hypothesis, which is determined by $\boldsymbol{U}_e = \mathbf{0}$ (partial derivatives of log-likelihood equal to zero). In effect, $\boldsymbol{\mu}$ is to be replaced by $\hat{\boldsymbol{\mu}}_e$, the fitted values in a model with only environmental covariates X_e present, giving the statistic

$$\boldsymbol{U}_{g|e} = \frac{1}{\phi} X_g^T(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_e). \tag{1}$$

Under the null hypothesis, $\boldsymbol{U}_{g|e}$ has the conditional distribution of $\boldsymbol{U}_g$ given $\boldsymbol{U}_e = \mathbf{0}$, which is asymptotically normal with mean $\mathbf{0}$ and covariance matrix

$$V_{g|e} = V_{gg} - V_{ge}V_{ee}^{-1}V_{eg} = \frac{1}{\phi^2} X_g^T \left(\Lambda - \Lambda X_e (X_e^T \Lambda X_e)^{-1} X_e^T \Lambda\right) X_g, \tag{2}$$

where V_{ee}, V_{eg}, V_{ge}, and V_{gg} are the upper left $d \times d$, upper right $d \times m$, lower left $m \times d$ and lower right $m \times m$ submatrices of V, respectively (see Smyth, 2003).

The covariance matrix $V_{g|e}$ will be singular if $m > n$, and also asymptotic normality would require $m \ll n$. This will, however, not present any problems for the FWER control method we present in the next section. We will not consider the complete null hypothesis, but the m individual null hypotheses $H_j : \beta_{gj} = 0$ for each component β_{gj} of $\boldsymbol{\beta}_g$, $j = 1, \ldots, m$, against two-sided alternatives. As test statistics, we use the standardized components of $\boldsymbol{U}_{g|e}$,

$$T_j = \frac{\boldsymbol{U}_{g|ej}}{\sqrt{V_{g|ejj}}}, \tag{3}$$

where $\boldsymbol{U}_{g|ej}$ denotes the jth entry of $\boldsymbol{U}_{g|e}$ and $V_{g|ejk}$ the jk entry of $V_{g|e}$. Under H_j, T_j is asymptotically standard normally distributed, and H_j will be rejected for large values of $|T_j|$. Under the

complete null hypothesis, $\boldsymbol{\beta}_g = \mathbf{0}$, the vector $\boldsymbol{T} = (T_1, T_2, \ldots, T_m)$ is asymptotically multivariate standard normally distributed with correlations

$$\mathrm{Cov}(T_j, T_k) = \frac{V_{g|ejk}}{\sqrt{V_{g|ejj} V_{g|ekk}}}. \tag{4}$$

Note that the dispersion parameter ϕ is canceled from $\boldsymbol{T}$ and the covariances. However, the σ_i^2 of Λ will have to be estimated.

2.3 | Special cases

2.3.1 | No environmental covariates

If no environmental covariates except the intercept are present in the GLM, then $X_e = \mathbf{1}$, the n-dimensional vector having all entries equal to 1, and $\Lambda = \sigma^2 I$ under the null hypothesis, where I is the $n \times n$ identity matrix. Then (1) and (2) reduce to

$$\boldsymbol{U}_{g|e} = \frac{1}{\phi} X_g^{\mathrm{T}} \left(I - \frac{1}{n} \mathbf{1}\mathbf{1}^{\mathrm{T}} \right) \boldsymbol{Y} \quad \text{and} \quad V_{g|e} = \frac{\sigma^2}{\phi^2} X_g^{\mathrm{T}} \left(I - \frac{1}{n} \mathbf{1}\mathbf{1}^{\mathrm{T}} \right) X_g,$$

so by (3) and (4),

$$T_j = \frac{\boldsymbol{x}_j^{\mathrm{T}} (I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}) \boldsymbol{Y}}{\sigma \sqrt{\boldsymbol{x}_j^{\mathrm{T}} (I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}) \boldsymbol{x}_j}}, \tag{5}$$

$$\mathrm{Cov}(T_j, T_k) = \frac{\boldsymbol{x}_j^{\mathrm{T}} (I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}) \boldsymbol{x}_k}{\sqrt{\boldsymbol{x}_j^{\mathrm{T}} (I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}) \boldsymbol{x}_j} \sqrt{\boldsymbol{x}_k^{\mathrm{T}} (I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}) \boldsymbol{x}_k}},$$

where $\boldsymbol{x}_j$ is the jth column of X_g, $j, k = 1, \ldots, m$. Hence, if σ is replaced by its estimate $\left(\frac{1}{n} \boldsymbol{Y}^{\mathrm{T}} (I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}) \boldsymbol{Y} \right)^{1/2}$, then T_j, the score test statistic for testing $\beta_{gj} = 0$, becomes $\sqrt{n}$ times the sample correlation between $\boldsymbol{x}_j$ and $\boldsymbol{Y}$, and $\mathrm{Cov}(T_j, T_k)$ is the sample correlation between $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$. Thus, for a GLM without adjustment for environmental covariates, correlations between score test statistics can be estimated by genotype sample correlations, which also estimate twice the composite linkage disequilibria if genotypes are coded 0, 1, 2 (Weir, 2008).

2.3.2 | Uncorrelated environmental and genetic covariates

Assume that each pair of an environmental covariate and a genetic covariate has near zero sample correlation, which should occur if each environmental covariate is uncorrelated with each genetic covariate. Then

$$V_{g|e} \approx \frac{\mathrm{tr}\Lambda}{n\phi^2} X_g^{\mathrm{T}} \left(I - \frac{1}{n} \mathbf{1}\mathbf{1}^{\mathrm{T}} \right) X_g$$

(see Appendix), which is the same expression as in the case of no environmental covariates with the exception that the common variance σ^2 of the responses is replaced by their average variance $\mathrm{tr}\Lambda/n = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2$, where, under the null hypothesis, the σ_i^2 are functions of the environmental covariates.

In this case, the correlations (4), which we will use to compute local significance levels, α_{loc}, do not depend on the environmental covariates, suggesting the possibility of a "standard" α_{loc} for a given set of markers, since the vast majority of markers is presumably independent of the covariates. Indeed, we repeated some of the calculations of α_{loc} for the VO_2-max data (Section 4), omitting the environmental covariates, and got almost identical results. This is further investigated by simulation in Section 4.5.

2.3.3 | The normal model

For Y_i normally distributed, $\Lambda = \sigma^2 I$, where I is the $n \times n$ identity matrix. Then (1) and (2) reduce to

$$\boldsymbol{U}_{\mathrm{g|e}} = \frac{1}{\sigma^2} X_{\mathrm{g}}^{\mathrm{T}}(I-H)\boldsymbol{Y} \quad \text{and} \quad V_{\mathrm{g|e}} = \frac{1}{\sigma^2} X_{\mathrm{g}}^{\mathrm{T}}(I-H)X_{\mathrm{g}}, \tag{6}$$

where $H = X_{\mathrm{e}}(X_{\mathrm{e}}^{\mathrm{T}}X_{\mathrm{e}})^{-1}X_{\mathrm{e}}^{\mathrm{T}}$ is the idempotent matrix projecting onto the column space of X_{e}. Then $I-H$ is the idempotent matrix projecting onto the orthogonal complement of the column space of X_{e}, and $(I-H)\boldsymbol{Y}$ are the residuals in the fitted linear model with only environmental covariates present. Note that σ^2 enters into the test statistics T_j, and needs to be replaced by an estimate; we have used $\frac{1}{n}\boldsymbol{Y}^{\mathrm{T}}(I-H)\boldsymbol{Y}$, the residual sum of squares of a fitted model with only environmental covariates present (the null hypothesis), divided by n.

2.3.4 | The logistic model

For Y_i Bernoulli distributed, $\phi = 1$ and the σ_i^2 of Λ are estimated by $\hat{\mu}_{ei}(1-\hat{\mu}_{ei})$, where $\hat{\mu}_{ei}$ are the fitted values under the null hypothesis, with only environmental covariates. Inference about $\boldsymbol{\beta}_{\mathrm{g}}$ is valid also if data are collected in a case–control study since the canonical (logit) link is used (Agresti, 2002, pp. 170–171) .

In the special case of no environmental covariates, that is, $X_{\mathrm{e}} = \mathbf{1}$, each score test statistic, T_j (5), is equal to the Cochran–Armitage trend test (Armitage, 1955; Cochran, 1954) statistic,

$$\frac{\sum_{i=0}^{2} s_i(n_0 x_i - n_1 y_i)}{\sqrt{n_0 n_1\left(\sum_{i=0}^{2} s_i^2 m_i - \frac{1}{n}\left(\sum_{i=0}^{2} s_i m_i\right)^2\right)}},$$

where s_i are the possible values of the genetic covariates, n_0 and n_1 the number of 0 and 1 phenotypes Y_i, respectively, x_i the number of observations having phenotype 1 and genotype i at the marker, y_i the number of observations having phenotype 0 and genotype i, and $m_i = x_i + y_i$. The Cochran–Armitage test is used in disease–genotype association testing with scores $(s_0, s_1, s_2) = (0, s, 1)$ (Sasieni, 1997; Slager & Schaid, 2001), for example, with $s = \frac{1}{2}$ for an additive genetic model.

3 | FAMILYWISE ERROR RATE CONTROL AND APPROXIMATIONS

We now turn to the topic of how to control the FWER by intersection approximations.

3.1 | Multiple hypothesis familywise error rate control

We have a collection of m null hypotheses, $H_j : \beta_{gj} = 0$ (no association between phenotype and genotype at marker j), $j = 1, \dots, m$, against two-sided alternatives. We will present a method for multiple testing correction that controls the FWER—the probability of making one or more Type I errors. We adopt the notation of Moskvina and Schmidt (2008) and denote by O_j the event that the null hypothesis H_j is not rejected, and by $\bar{O}_j$ its complement. Then, if all m null hypotheses are true,

$$\text{FWER} = P(\bar{O}_1 \cup \cdots \cup \bar{O}_m) = 1 - P(O_1 \cap \cdots \cap O_m). \tag{7}$$

In our case, O_j is an event of the form $|T_j| < c$, where the test statistic T_j (3) is asymptotically standard normally distributed. We will consider single-step multiple testing methods, and choose the same cutoff c for each j. We denote the *local significance level* by $\alpha_{\text{loc}} = 2\Phi(-c) = P(\bar{O}_j)$ for all j, the asymptotic probability of false rejection of H_j, where Φ is the univariate standard normal cumulative distribution function. When the joint distribution of the test statistics is known under the complete null hypothesis, or can be estimated, FWER control at the α significance level can be achieved by solving the inequality FWER $\leq \alpha$ for α_{loc}, based on either the union or intersection formulation of (7).

When the FWER is calculated under the complete null hypothesis, so-called weak FWER control is achieved. However, in our situation, subset pivotality is satisfied, meaning that the distribution of any subvector $(T_j)_{j\in J}$ is identical under $\cap_{j\in J} H_j$ and under the complete null hypothesis $\cap_{j=1}^{m} H_j$, for all subsets $J \subseteq \{1, 2, \dots, m\}$ (Westfall & Young, 1993, p. 42). In particular, a subvector of $\boldsymbol{U}_{\text{g|e}}$ (1) and a submatrix of $V_{\text{g|e}}$ (2) corresponding to J only involve genotypes of markers corresponding to J. Then strong FWER control is achieved, meaning that FWER $\leq \alpha$ regardless of which null hypotheses are true (Westfall & Young, 1993; Westfall & Troendle, 2008).

In principle, step-down methods could be considered: After application of a method controlling FWER, remove rejected hypotheses and redo the method, and repeat until no further rejections occur (Goeman & Solari, 2010). However, in a GWAS framework, very few, if any, rejections are expected; hence, in practice, the number of hypotheses will hardly change after application of the method, and the second application will not have any effect.

When m is large, FWER (7) involves high-dimensional integrals over the acceptance or rejection regions, the evaluation of which are suggested by Conneely and Boehnke (2007). To avoid these costly evaluations, we may instead control FWER by considering bounds based on (7). For example, the Bonferroni method is based on the Boole inequality applied to the union formulation of (7),

$$\text{FWER} = P(\bar{O}_1 \cup \cdots \cup \bar{O}_m) \leq \sum_{j=1}^{m} P(\bar{O}_j) = \sum_{j=1}^{m} \alpha_{\text{loc}} = m\alpha_{\text{loc}},$$

from which it is seen that $\alpha_{\text{loc}} = \alpha/m$ guarantees FWER $\leq \alpha$.

The focus in this work will be on the intersection formulation of (7). Background theory will be given next and new application in Section 3.3.

3.2 | Intersection approximations

Following Glaz and Johnson (1984), we define kth order product-type approximations to $P(O_1 \cap \cdots \cap O_m)$ by

$$\begin{aligned}\gamma_k &= P(O_1 \cap \cdots \cap O_k) \prod_{j=k+1}^{m} P(O_j | O_{j-k+1} \cap \cdots \cap O_{j-1}) \\ &= \frac{\prod_{j=k}^{m} P(O_{j-k+1} \cap \cdots \cap O_j)}{\prod_{j=k+1}^{m} P(O_{j-k+1} \cap \cdots \cap O_{j-1})},\end{aligned} \tag{8}$$

$k = 1, \ldots, m$, where probabilities are evaluated under the complete null hypothesis. This is similar to the general product rule for the probability of an intersection of events applied to $\gamma_m = P(O_1 \cap \cdots \cap O_m)$, but with dimension of distributions limited to k. The idea is that the γ_k should constitute increasingly better approximations of γ_m as k increases, and that calculation of γ_k is less costly than calculation of γ_m when $k < m$.

The approximations depend on the order of the components of $\boldsymbol{T} = (T_1, \ldots T_m)$. To realize most of the potential gains in test power due to correlation, we have used the order in which the markers are positioned along the genome, assuming that the largest correlations occur between close markers.

In our case, $\gamma_1 = \prod_{j=1}^{m} P(|T_j| < c) = (1 - \alpha_{\text{loc}})^m$ and $\gamma_m = P(|T_1| < c, |T_2| < c, \ldots, |T_m| < c) = 1 - \text{FWER}$. Since $\boldsymbol{T}$ is asymptotically multivariate normally distributed with mean $\mathbf{0}$ under the complete null hypothesis, $\gamma_1 \leq \gamma_m$ asymptotically in n for any correlation structure (Šidák, 1967). Choosing α_{loc} such that $\text{FWER} = 1 - \gamma_m \leq 1 - \gamma_1 = 1 - (1 - \alpha_{\text{loc}})^m = \alpha$ keeps FWER at the α level. It is well known that the α_{loc} found by this method, the Šidák method, is slightly larger than the α_{loc} found by the Bonferroni method, thus the Šidák method will give slightly higher power.

For general k, if $\gamma_k \leq \gamma_m$, then $\text{FWER} = 1 - \gamma_m \leq 1 - \gamma_k = \alpha$ can be used to control FWER by solving the last equation for α_{loc} (choosing the greatest solution if not unique; we have, however, never observed a γ_k that is not monotonically decreasing in α_{loc}). If $\gamma_k \leq \gamma_l$, then continuity of γ_k and of γ_l as functions of α_{loc} implies that the α_{loc} making $1 - \gamma_l = \alpha$ is no less than the α_{loc} making $1 - \gamma_k = \alpha$, so that the power obtained by the lth approximation is no less than the power obtained by the kth approximation.

The ideal property $\gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_k \leq \gamma_m$ for all α_{loc} is ensured if $|\boldsymbol{T}| = (|T_1|, \ldots, |T_m|)$ is monotonically sub-Markovian of order k (MSM_k) with respect to $(-\infty, c)^k$ for all c, $2 \leq k \leq m - 1$, as defined by Block, Costigan, and Sampson (1992). Examples of correlation structures making $|\boldsymbol{T}|$ satisfy the MSM_k properties are given in Section 4.6. There is, however, for general correlation structures no guarantee that $\gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_m$ for *all* α_{loc}, just as there is no guarantee that a higher order Taylor expansion of a function is closer to the true value than a lower order expansion at *all* points. The overall trend is nevertheless that the γ_k are increasing in k for the α_{loc} we consider. An assessment for covariance matrices estimated from real data is given in Section 4.3. We did not find violations.

A summary of concepts of positive dependence, like MSM, was given by Dickhaus (2014, pp. 58–61).

3.3 | Controlling FWER using *k*th order approximation for score tests

We recall that the vector $\boldsymbol{T}$ of score test statistics is under the complete null hypothesis asymptotically standard multivariate normal with covariances given by (4). We denote by O_j the event $|T_j| < c$ of nonrejection of H_j, which has probability $P(O_j) = 1 - \alpha_{\text{loc}}$ under the null hypothesis, with $\alpha_{\text{loc}} = 2\Phi(-c)$. We will detail how to find α_{loc} given by the second-order approximation, γ_2. Let $r_j = \text{Cov}(T_{j-1}, T_j)$. Then, after some calculation of the bivariate normal integral over a square having $(\pm c, \pm c)$ as corners, we get

$$P(O_{j-1} \cap O_j) = 1 - \alpha_{\text{loc}} - \sqrt{\frac{2}{\pi}} \int_{-c}^{c} e^{-x^2/2}\, \Phi\left(\frac{r_j x - c}{\sqrt{1 - r_j^2}} \right) dx,$$

giving

$$\begin{aligned}
\gamma_2 &= P(O_1 \cap O_2) \prod_{j=3}^{m} P(O_j | O_{j-1}) = \frac{\prod_{j=2}^{m} P(O_{j-1} \cap O_j)}{\prod_{j=3}^{m} P(O_{j-1})} \\
&= \frac{\prod_{j=2}^{m} \left(1 - \alpha_{\text{loc}} - \sqrt{\frac{2}{\pi}} \int_{-c}^{c} e^{-x^2/2}\, \Phi\left(\frac{r_j x - c}{\sqrt{1 - r_j^2}} \right) dx \right)}{(1 - \alpha_{\text{loc}})^{m-2}} \\
&= (1 - \alpha_{\text{loc}}) \prod_{j=2}^{m} \left(1 - \sqrt{\frac{2}{\pi}} \frac{1}{1 - \alpha_{\text{loc}}} \int_{-c}^{c} e^{-x^2/2}\, \Phi\left(\frac{r_j x - c}{\sqrt{1 - r_j^2}} \right) dx \right). \qquad (9)
\end{aligned}$$

For a desired upper bound α on FWER, the equation $1 - \gamma_2 = \alpha$ is solved with respect to α_{loc}, which can be done numerically using for example a bisection algorithm. Note that α_{loc} enters into $c = -\Phi^{-1}(\alpha_{\text{loc}}/2)$.

We can control FWER by higher order approximations by solving the equation $1 - \gamma_k = \alpha$ for α_{loc} in a similar way, which we will henceforth refer to as FWER order k approximation. By (8), γ_k can be written as a ratio of a product of k-dimensional and a product of $k-1$-dimensional multivariate normal integrals.

Good numerical methods for calculating multivariate normal integrals exist for small dimensions (Genz & Bretz, 2009). We used the `pmvnorm` function of the R (R Core Team, 2015) package mvtnorm (Genz et al., 2016), which can calculate multivariate normal probabilities with some accuracy for dimensions up to 1,000. The Miwa algorithm (Miwa, Hayter, & Kuriki, 2003) of `pmvnorm` can be used for small dimensions and is deterministic, whereas the default Genz–Bretz algorithm (Genz, 1992; Genz, 1993; Genz & Bretz, 2002) includes simulations that lead to small inaccuracies. We used standard R functions to compute second-order approximations (9) and the Miwa algorithm to illustrate order 3 and 4 approximations (Section 4), and the Genz–Bretz algorithm to obtain the "true" reference α_{loc} for blocks of 800 or 1,000 markers (Sections 4.3 and 4.5) and for two constructed correlation structures of 100 test statistics (Section 4.6).

The procedure to find α_{loc} does not depend on the exact form of the test statistic, only that the vector $(T_1, \ldots, T_m)$ of test statistics is asymptotically standard multivariate normal under the complete null hypothesis and $|T_j| \geq c$ leads to rejection. For example, (9) appeared in an allelic test procedure by Moskvina and Schmidt (2008).

In practice, instead of calculating α_{loc}, it may be preferable to calculate FWER-adjusted p-values: Replace α_{loc} with p, the unadjusted p-value for an individual test, in the calculation of γ_k (e.g., in (9) for $k = 2$). Then $1 - \gamma_k$ is an FWER-adjusted p-value for the test, in the sense that if $1 - \gamma_k \leq \alpha$ (rejection based on adjusted p-value), then $p \leq \alpha_{\text{loc}}$ (rejection based on local significance level). See Section 4.3 for an example.

3.4 | FWER control with independent blocks

A common assumption is independence of genotypes for markers from different chromosomes. Within a chromosome, genetic markers can belong to different haplotype blocks, being highly correlated within a block and independent or nearly independent between blocks (Gabriel et al., 2002).

Assume that the m markers to be tested, and $\{O_1, \ldots, O_m\}$, can be partitioned into b independent blocks, $\{O_1, \ldots, O_{m_1}\}$, $\{O_{m_1+1}, \ldots, O_{m_2}\}, \ldots, \{O_{m_{b-1}+1}, \ldots, O_m\}$, so that the events O_{j_1} and O_{j_2} are independent if they belong to different blocks. Let $\gamma_k^{(l)}$ be the kth-order approximation given by (8) for the intersection of the events belonging to the lth block, and let γ_k be the overall kth order approximation. Then it is easy to verify that $\gamma_k = \prod_{l=1}^{b} \gamma_k^{(l)}$. We will calculate α_{loc} based on the overall approximation, which is what we recommend. It is, however, also possible to calculate a different α_{loc} per block (Stange, Loginova, & Dickhaus, 2016).

3.5 | The effective number of independent tests

The concept of an effective number of independent tests, M_{eff}, in multiple testing problems has been described and discussed by many authors, including Nyholt (2004), Gao, Starmer, and Martin (2008), Moskvina and Schmidt (2008), Li and Ji (2005), Galwey (2009), and Chen and Liu (2011). All except Moskvina and Schmidt (2008) first estimate M_{eff}, and then use M_{eff} in place of m in the Šidák formula to calculate $\alpha_{\text{loc}} = 1 - (1 - \alpha)^{1/M_{\text{eff}}}$. (Alternatively, the Bonferroni formula could also be used to calculate $\alpha_{\text{loc}} = \alpha/M_{\text{eff}}$.)

These methods do not use the concept of FWER in the derivation of M_{eff}, and there is no mathematical justification that FWER is controlled, let alone that M_{eff} is independent of α. All methods start with the linkage disequilibrium or composite linkage disequilibrium matrix, and there is no mention of any dependence of the M_{eff} estimate on the test statistics used for the hypothesis tests.

The method of Moskvina and Schmidt (2008), on the other hand, is based on an allelic test and controls the FWER using second-order intersection approximations. The main output of their method is α_{loc}, as it is for our method. Solving for M_{eff} in the above Šidák formula, we can define $M_{\text{eff}} = \ln(1 - \alpha)/\ln(1 - \alpha_{\text{loc}})$ as a by-product. Note that M_{eff} depends on both α_{loc} and the FWER threshold α.

3.6 | The maxT permutation method

We will compare FWER order k approximation with the Westfall and Young (1993) maxT permutation method, and therefore give a brief review of the latter.

FWER, the probability that one or more of the m null hypotheses are falsely rejected, can be formulated FWER $= P(\max_j |T_j| \geq c)$ under the complete null hypothesis. In the maxT method, the critical value c is found empirically by permutation of the response variable in order to generate a sample from the distribution of the $\max_j |T_j|$ statistic. If the FWER is to be controlled at the α level and b permutations are made, c is estimated by the $(1-\alpha)b$th smallest value of the $\max_j |T_j|$, which is an estimate of the $1-\alpha$ quantile of $\max_j |T_j|$. The probability that the kth smallest value of a random sample of size b is greater than the $1-\alpha$ quantile is equal to the binomial cumulative distribution function with parameters b and $1-\alpha$ evaluated at $k-1$, which can be used to construct a confidence interval for c (Conover, 1980, p. 114; Thompson, 1936). To compare with our approximation method, for which asymptotic normality is assumed, we used $\alpha_{\text{loc}} = 2\Phi(-c)$, and a confidence interval for α_{loc} is obtained by transforming the bounds of the above interval the same way.

Success of the permutation method relies on exchangeability of the responses. In a GLM that includes environmental covariates, this is in general not the case, since the expected values of the responses are not equal under the null model. This is so even for the classical multiple regression normal model (Commenges, 2003). In models without environmental covariates (only intercept), however, the responses are exchangeable and the maxT method gives FWER control.

In the case that the responses follow a normal model, there exist several approximate permutation methods. For our purposes, the Still–White method (see Winkler, Ridgway, Webster, Smith, & Nichols, 2014) is appropriate (see also Hummel, Meister, & Mansmann, 2008, for a similar approach). The original linear model is replaced by one in which the environmental covariates (except the intercept) are removed, and the responses are replaced by the residuals when only environmental covariates are fitted in the original model. Thus, H is replaced by $\frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}$ and $\boldsymbol{Y}$ by $(I-H)\boldsymbol{Y}$ in (6). Because $H\mathbf{1} = \mathbf{1}$, the numerator of the test statistic (3) and the σ^2 estimate will be unchanged. In the denominator, on the other hand, at a marker having genotype vector $\boldsymbol{x}_{\text{g}}$, $\boldsymbol{x}_{\text{g}}^{\mathrm{T}}(I-H)\boldsymbol{x}_{\text{g}}$ will be replaced by $\boldsymbol{x}_{\text{g}}^{\mathrm{T}}(I-\frac{1}{m}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{x}_{\text{g}}$. The latter (total sum of squares of $\boldsymbol{x}_{\text{g}}$) is greater than the former (residual sum of squares if $\boldsymbol{x}_{\text{g}}$ is regressed on X_{e}). Hence, although the assumptions of a multiple regression model are not satisfied when using residuals as responses (they are dependent and heteroscedastic), the score test statistic of the new model has, for all markers, an absolute value that is less than or equal to that of the original model, with equality if and only if the coefficient of determination when regressing the marker on the environmental covariates, is zero. We expect the vast majority of the markers to be independent of the environmental covariates, so that the loss of power will be small. In the new model, without environmental covariates, the maxT method can be used.

We are not aware of any well-established approximate methods using permutations in presence of covariates in nonnormal GLMs, but recently Hemerik, Goeman, and Finos (2019) presented a method based on flipping the sign of score contributions.

4 | VALIDITY, POWER, AND EFFICIENCY OF THE FWER APPROXIMATION

4.1 | Datasets: TOP and VO_2-max

In the case–control GWA study TOP, data were collected with the aim to detect SNPs associated with schizophrenia or bipolar disorder (Athanasiu et al., 2010; Djurovic et al., 2010). The preprocessed TOP data contained genetic information on 672,972 SNPs (Affymetrix Genome-Wide

Human SNP Array 6.0), all with minor allele frequency (MAF) > 0.01, for 1148 cases and 420 controls. Our data included individuals sampled until March 2013, and therefore the sample size is larger than in the cited papers.

However, in our analysis, the number of cases was reduced to 420 by drawing a random sample from the 1,148 because the normal approximation is in general poor for small tail probabilities of test statistics based on unbalanced binary data. This is so even when the total number of observations is large, leading to tests exceeding their nominal size (Langaas & Bakke, 2014; see also discussion in Section 4.4). The reduction of data in this case was done solely for the purpose of demonstrating FWER approximation, and we would in general recommend that binary experiments are designed to be balanced if normal approximation is to be used, or else use methods that do not rely on normal approximation.

Some genotype data were missing from the TOP data, and mean imputation was done for 0.04% of the genotypes. Genotype–phenotype association was assessed by fitting a logistic regression without any environmental covariates, so that score test correlations equal genotype correlations (Section 2.3.1).

The VO_2-max data came from participants of the HUNT Study (Nord-Trøndelag Health Study, ntnu.edu/hunt; Aspenes et al., 2011; Loe et al., 2013). A cross-sectional GWA study was performed to find SNPs associated with maximum oxygen uptake. The preprocessed VO_2-max data consisted of 123,497 SNPs (Illumina Cardio-MetaboChip; Moore et al., 2012) with MAF > 0.01 for 2,802 individuals. The VO_2-max data were analyzed using a normal linear regression model, including age, sex, and physical activity score as covariates. Due to missing data, mean imputation was done for 0.7% of the genotypes.

4.2 | Relative power of methods

We computed α_{loc} for the TOP and VO_2-max data by FWER-order 1–3 approximation, controlling FWER at the .05 level, assuming independence between chromosomes (Section 3.4). In addition, the Bonferroni method and the maxT permutation method using 10^6 permutations were applied. For the TOP data, which did not include environmental covariates, the maxT method was done by permuting the binary response vector. The VO_2-max data, assumed to follow a normal linear regression model, included environmental covariates, so the maxT method was performed by fitting the environmental covariates and then permuting the residuals (see Section 3.6).

For both sets, α_{loc} controlling the FWER at level .05 was smallest for the Bonferroni method, slightly larger for the order 1 approximation (Šidák), and further increasing, giving higher power, through the order 2 and 3 approximations (Table 1), and greatest for maxT. At present, computation of α_{loc} using very high-order approximations, or, indeed the full multivariate normal distribution, is unfeasible, both because it is time consuming and due to randomness in the Genz–Bretz algorithm, making it difficult to solve for α_{loc} (Section 3.3). However, in the next section, FWER bounds are found using a specific α_{loc} based on large independent blocks (instead of solving for α_{loc} given a specified FWER bound), in effect giving a comparison of FWER approximations with the full multivariate normal distribution.

4.3 | Validity and potential of FWER approximation

First, we will investigate monotonicity of the γ_k when k increases. Next, we will compare the FWER approximation method with the full multivariate normal distribution, as announced in

TABLE 1 Local significance level α_{loc} controlling FWER at level .05 calculated by the Bonferroni, the FWER-order 1–3 approximation and the maxT permutation method for the TOP and VO_2-max data, ratio of α_{loc} to Bonferroni α_{loc}, effective number M_{eff} of tests, and computing times

					Computing time		
Data	**Method**	$10^7\alpha_{\text{loc}}$	**Ratio**	$10^5 M_{\text{eff}}$	**Total**	**Read**	**Setup (s)**
TOP	Bonferroni	.743	1.00	6.90	13 min	13 min	25
	Order 1 (Šidák)	.762	1.03	6.73	13 min	13 min	25
	Order 2	.864	1.16	5.93	18 min	13 min	28
	Order 3	.880	1.18	5.83	31 hr	13 min	39
	maxT	1.52	2.05	3.36	391 hr	Confidence interval: [1.51,1.54]	
VO_2-max	Bonferroni	4.05	1.00	1.27	3 min	2 min 30 s	7
	Order 1 (Šidák)	4.15	1.03	1.23	3 min	2 min 30 s	7
	Order 2	4.69	1.16	1.09	3 min	2 min 30 s	18
	Order3	5.11	1.26	1.00	6 hr	2 min 30 s	29
	maxT	6.69	1.65	0.77	224 hr	Confidence interval: [6.63, 6.75]	

Note: The "Total" column shows total CPU time, including time for reading data (response, environmental covariates, and genetic markers) and imputation of markers ("Read"), and for calculating score test statistics and estimating the necessary correlations between them ("Setup"). "Read" and "Setup" show actual elapsed time summed over computer cores, and are thus upper bounds for CPU time. For maxT, 95% confidence intervals for α_{loc} are shown instead of data read/imputing and setup time.

Section 4.2. Finally, we will make a per chromosome comparison of FWER approximation, full multivariate normal, and maxT.

As mentioned earlier, it is at present unfeasible to calculate α_{loc} to achieve a given level of FWER control when the order of the approximation is large. Instead, we will in all of this section fix $\alpha_{\text{loc}} = 1.52 \times 10^{-7}$ for the TOP data and $\alpha_{\text{loc}} = 6.69 \times 10^{-7}$ for the VO_2-max data, which were the α_{loc} controlling FWER at level .05 using the maxT permutation method for the complete data (see Section 4.2), and then compare achieved FWER bounds. Alternatively, this can be viewed as considering the two α_{loc} unadjusted p-values and the FWER bounds FWER-adjusted p-values (see end of Section 3.3).

In general, we expect (i) that $\gamma_k \leq \gamma_m$ for $k \leq m$, where m is the number of genetic markers, so that the approximations give conservative FWER control and (ii) that the γ_k are increasing in k, meaning that the approximations give higher power for larger k.

The largest dimension that can be handled by the `pmvnorm` function (see Section 3.3) is 1,000. We checked the two properties above by dividing the markers into blocks of 1,000 along each chromosome for the VO_2-max data and into blocks of 800 for the TOP data (for the estimated covariance matrices for test statistics to be nonsingular, the number of markers cannot exceed the number of observations). This resulted in 833 blocks of 800 markers, and 22 shorter blocks at the end of the chromosomes, for TOP, and 113 blocks of 1,000 markers, and 22 shorter blocks, for VO_2-max. The upper FWER bounds, $1 - \gamma_k$, given by order $k = 1, 2, 3$ approximation, and without approximation using the Genz–Bretz algorithm, $k = 800$ (TOP) or $k = 1000$ (VO_2-max), using the values of α_{loc} given above, were calculated for each separate block.

Both properties were satisfied for the TOP and the VO_2-max data (Figure 1; for clarity, only complete blocks of 800 or 1,000 markers are shown, not the 22 shorter blocks). This is a very strong indication that properties (i) and (ii) hold for 1–3 order approximations for the complete datasets.

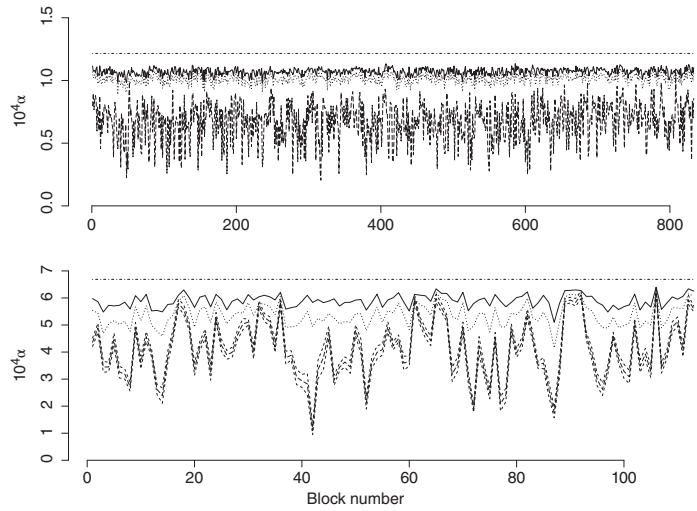


FIGURE 1 FWER bounds, α, for each of 833 blocks of 800 SNPs from the TOP data (top) and for each of 113 blocks of 1000 SNPs from the VO_2-max data (bottom). For TOP, $\alpha_{\text{loc}} = 1.52 \times 10^{-7}$ was used, and for VO_2-max, $\alpha_{\text{loc}} = 6.69 \times 10^{-7}$. The upper horizontal line (dash-dotted) shows α obtained by the FWER order 1 approximation (Šidák). The next curves (solid and dotted) show α for order 2 and 3 approximations. The lower dashed curve shows the FWER obtained by the complete multivariate normal distribution of the 800-dimensional (TOP) or 1,000-dimensional (VO_2-max) test statistic vector, having correlation structure given by (4). The two enveloping curves (also dashed, only shown for VO_2-max for clarity) indicate the estimated absolute error reported by the `pmvnorm` function of the R (R Core Team, 2015) package mvtnorm (Genz et al., 2016)

We proceed to compare FWER approximation with the full multivariate normal distribution and now assume multivariate normality. We first note that γ_m, the probability of making no Type I errors when all null hypotheses are true, is greater than or equal to the product of the 855 mostly 800-dimensional multivariate normal γ_k referred to above for TOP. Similarly, γ_m for VO_2-max is greater than or equal to the product of the 135 mostly 1,000-dimensional γ_k. This follows from the Gaussian correlation conjecture, which was famously proven by T. Royen in 2014 (Royen, 2014; Latała & Matlak, 2017).

Calculating the products, we get FWER control bounds $1 - \gamma_m$ at least as good as 0.055 for TOP and 0.050 for VO_2-max. We assume that these bounds are very close to the true $1 - \gamma_m$, an assumption that is strengthened by the fact that they are very close to the targeted bound of 0.05 using the maxT method. Another fact pointing in the same direction is that order k approximation gives virtually identical results for each chromosome, whether calculated for the whole chromosome or via multiplication over blocks of 800 or 1,000, which is the same as assuming independence between blocks (Section 3.4), $k = 2, 3$. In comparison, using order $k = 1$, 2, and 3 approximations for the blocks instead, the FWER bounds were 0.097, 0.086, and 0.082, respectively, for the TOP data, and 0.079, 0.070, and 0.065 for the VO_2-max data.

Finally, we did a similar analysis per chromosome, also including maxT bounds (Figure 2). Again, the FWER bounds decreased as the order of the approximation increased, and even smaller bounds were achieved using the full multivariate normal distribution (approximated by multiplying over blocks of 800 or 1,000 markers) or maxT. Also, as expected, for fixed α_{loc}, FWER control depends heavily on the number of markers tested for at each chromosome (Figure 2). However, in addition, chromosomes with general high correlation levels benefit more from methods taking the correlations into account than do chromosomes with lower correlations. For example,

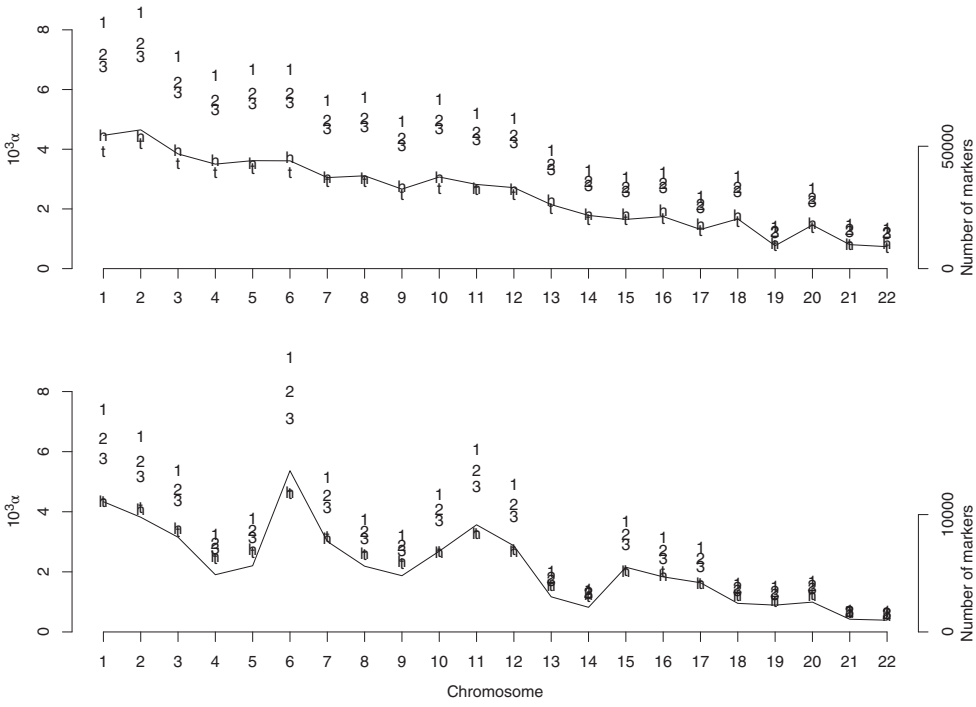
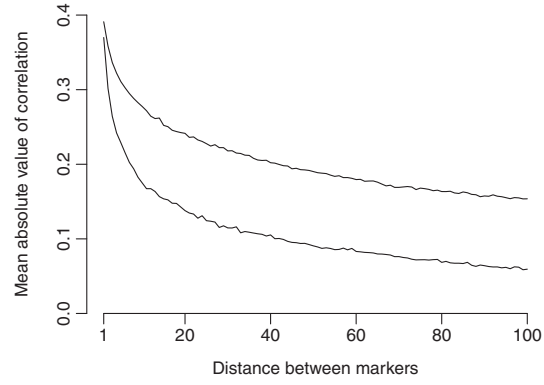


FIGURE 2 FWER bounds, α, per chromosome for the TOP data (top) and for the VO_2-max data (bottom). For TOP, $\alpha_{\text{loc}} = 1.52 \times 10^{-7}$ was used, and for VO_2-max, $\alpha_{\text{loc}} = 6.69 \times 10^{-7}$. Bounds for order 1, 2, and 3 approximations and bounds obtained by high-order calculations and by the maxT methods are shown (1, 2, 3, h and t, respectively; use left vertical axis). Also shown are the number of markers tested at each chromosome (line; use right vertical axis). As expected, FWER control depends on the number of markers, but also note that highly correlated chromosomes, such as 6 for VO_2-max benefit more from taking correlations into account than do less correlated chromosomes, such as 9

FIGURE 3 Mean absolute value of correlations between markers of distance 1–100 on chromosome 6 (upper curve) and 9 (lower curve) of the VO_2 data



mean correlation between markers of a fixed distance is in general higher in the VO_2-max data for chromosome 6 than for chromosome 9 (Figure 3), and this fact is reflected in Figure 2.

4.4 | Assessment of asymptotic normality

Our FWER order k method relies on asymptotic normality of the test statistic (3), and we have performed a simulation study to assess this assumption on univariate data without environmental

covariates, using the assumed distributions of the responses of our two datasets and using sample sizes motivated by those sets. Normality will ensure that the Type I error probability is within the nominal level, and our focus will be to check this rather than to assess normality directly.

To assess whether the Type I error probabilities were within nominal bounds, we studied whether the proportion of rejections (p-value $\leq \alpha$) in a large number of simulations under the null hypothesis was less than or equal to α for various α, with a particular interest in the very small α used in GWA analysis.

In the TOP data available to us at the time of analyzing the data, there were 420 controls and 1,148 cases. We drew genotypes randomly with MAF = 0.05, assuming Hardy–Weinberg equilibrium (giving genotype probabilities 0.0025, 0.0950, 0.9025), for the controls and cases, and assuming the null hypothesis of no effect of genotype. Then the score test was performed and a p-value calculated for the two-sided test of no association between the disease phenotype and the genotype, assuming the univariate normal distribution of the test statistic.

This procedure was performed 10^9 times, and the proportion of p-values smaller than various cutoffs α in a range from 5×10^{-8} to 5×10^{-4} were recorded, and also 95% confidence intervals for the true Type I error rate were calculated (using the common normal approximation method for binomial probability) (Figure 4, panel a). There was a severe inflation of Type I errors, and approximate normality may not be assumed in this case.

We repeated the simulations with a balanced dataset with 420 controls and 420 cases (Figure 4, panel b). Approximate normality would lead to Type I error rates close to α, which was not the case. However, the error rates were now consistently less than the nominal levels, so the assumption of normality will lead to conservative FWER control.

The VO_2-max data have a sample size of 2,802, and were analyzed using a multiple linear regression model in Section 4.2. We repeated a similar procedure as described above for the TOP data, but drew standard normal responses and performed the score test for a model with intercept in addition to genotypes drawn randomly with MAF = 0.05. This (drawing both genotypes and normal responses) was repeated 10^9 times (Figure 4, panel c). Type I error rates were now equal to α. Decreasing the sample size to 1,401 also gave valid, but conservative p-values (panel d), and thus conservative FWER control.

This simulation study, which is by no means exhaustive, indicates that we would trust approximate normality of the score test statistic in the extreme tails when the responses are normal. For binary responses we would only trust the (conservative) normality of the test statistic in the extreme tails for balanced samples. The problem of nonnormality of score test statistics for unbalanced case–control data has been investigated by Dey, Schmidt, Abecasis, and Lee (2017).

4.5 | Evaluation of effect of environmental covariates and further comparisons with full multivariate normal and maxT

To further study the properties of the FWER approximation method, and in particular to evaluate the effect of environmental covariates, a small simulation study was performed.

After removing duplicate neighboring SNPs from chromosome 1 of the preprocessed VO_2-max data so only one remained, the average neighbor correlation in a sliding window of width 1,000 ranged from 0.29 to 0.43. We chose the window having average neighbor correlation closest to the average, 0.36, as marker genotypes in the simulation. The genetic covariate matrix thus had dimension 2,802 (observations) times 1,000 (SNPs). The environmental covariates sex (x_{sex},

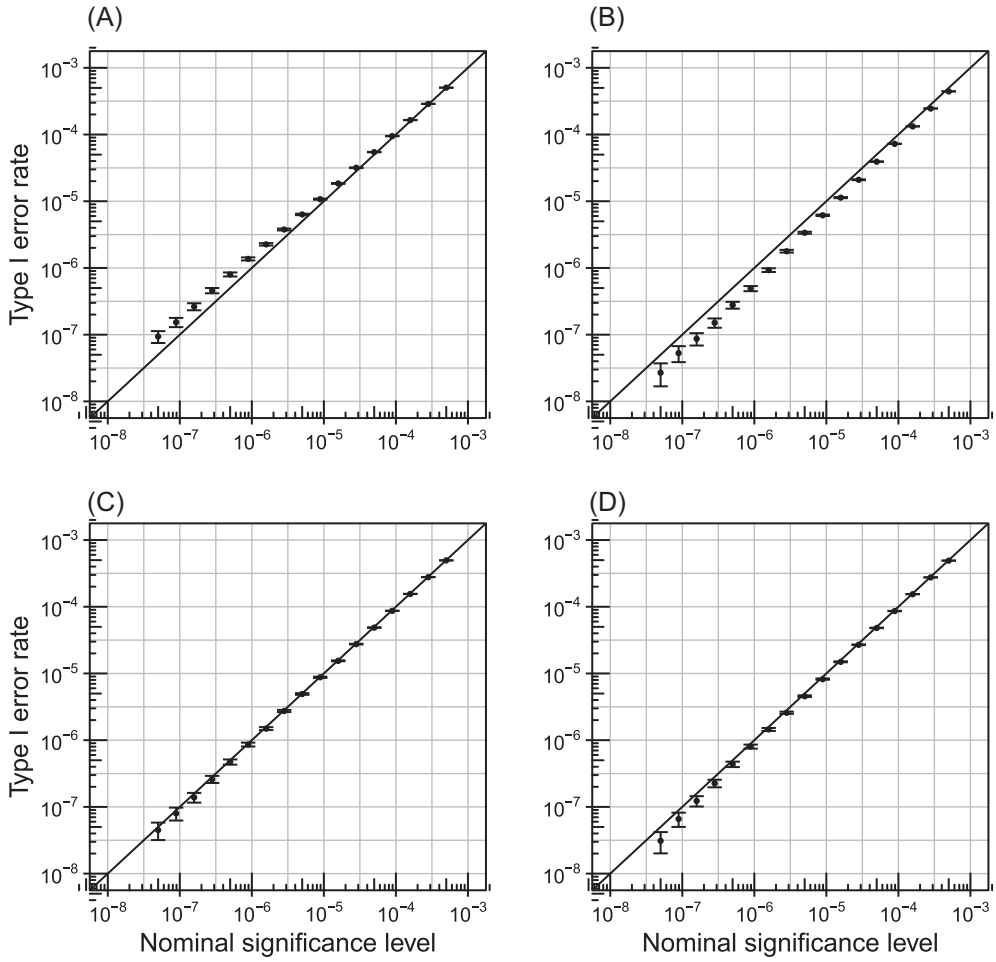


FIGURE 4 Nominal significance level versus Type I error rate assuming asymptotic normality of the score test statistic for (a) an unbalanced binomial design (1,148 cases, 420 controls), (b) a balanced binomial design (420 cases, 420 controls), (c) a normal model with sample size 2,802, and (d) a normal model with sample size 1,401. 95% confidence intervals for the Type I error probabilities are indicated

male = 0, female = 1), age (x_{age}, range 19.2–84.4 years) and activity level (x_{act}, range 0–15) from the VO_2-max data were used. These covariates were not very correlated with the 1,000 SNPs, with average absolute value of correlations of 0.02, 0.02, 0.02, respectively, and maximal absolute value of correlations 0.06, 0.20, 0.07, respectively. This means that we would assume that our FWER approximation method would not require knowledge of these covariates (Section 2.3.2).

We started by simulating data from a core model $Y = 171.3 - 32.5x_{sex} - 0.9x_{age} + 2.7x_{act} + \epsilon$, where ϵ was drawn from the univariate normal distribution with mean 0 and SD 17.7. These choices of parameter values were motivated from the multiple linear regression model null fit to the VO_2-max data. To find a local significance level that provides FWER error control at level .05, we used the FWER order 2, 3, and 1,000 approximation method. We also used 10 independent blocks of size 100 with FWER order 100 approximation (Section 3.4), and the maxT permutation method (5×10^5 permutations) using Still–White for handling covariates (Section 3.6). To study the effect of including covariates in the methods for FWER control, we both fitted the

multiple linear regressions with all covariates (sex, age, and activity level) and without covariates (intercept only).

In a second set of simulations, four environmental variables x_1, x_2, x_3, and x_4 were created to be correlated with four of the 1,000 SNPs, with correlations 0.50, 0.29, 0.45, 0.41, respectively. The data were generated using $Y = 171.3 - 32.5x_{\text{sex}} - 0.9x_{\text{age}} + 2.7x_{\text{act}} + x_1 + x_2 + x_3 + x_4 + \epsilon$, where ϵ was again drawn from the univariate normal distribution with mean 0 and SD 17.7. The same methods for FWER control were studied, and in addition to fitting a model with all covariates, a model without the correlated environmental covariates (only with sex, age, and activity level) was fitted. Again, Still–White was used to handle covariates for the maxT method.

The simulations confirm the findings of Sections 4.2 and 4.3 for normal data (Table 2). In addition, the simulations indicate that environmental covariates are of lesser importance for determining local significance levels, whether they are correlated with a few of the genetic markers or not. This is not to say that environmental covariates are not important when performing tests, only that they do not seem to be important for correlation structure between the test statistics, at least not for normal data. For all simulations, the 95% confidence interval for the maxT α_{loc} also covers the full multivariate normal solution.

4.6 | Effect of correlation strength

For certain structured correlation matrices, theoretical results exist for the multivariate normal integral for γ_m, including AR(1) and compound symmetry, which have off-diagonal ij entries $\rho^{|i-j|}$ and ρ, respectively (Genz & Bretz, 2009). Figure 3 shows that the average mean absolute value of correlation between markers of different distance is not constant (as is the case for compound symmetry, which we would regard as an extreme hypothetical situation where our method, only taking close neighbors into account, would be at disadvantage). Instead there is a decrease in the mean absolute value of correlation between markers when distance between them increases, and AR(1) is one possible way the correlation could decrease.

Consider $m = 100$ markers and a multivariate normal test statistic vector $\boldsymbol{T}$. For the two cases that the correlation structure of $\boldsymbol{T}$ is AR(1) or compound symmetry, we investigated the effect

TABLE 2 Local significance level α_{loc} controlling FWER at level .05 calculated by FWER order 2–3 approximation, by full multivariate normal assuming 10 independent blocks of 100 markers, by full 1,000-variate normal, and by the maxT permutation method, with 95% confidence interval for the latter, using simulated data

	$10^5\alpha_{\text{loc}}$					
Simulation	**2**	**3**	**Blocks**	**1,000**	**maxT**	**maxT Confidence interval**
1	5.77	6.12	7.04	7.19	7.23	[7.15, 7.33]
2	5.77	6.12	7.05	7.20	7.29	[7.20, 7.37]
3	5.77	6.12	7.03	7.18	7.25	[7.15, 7.35]
4	5.77	6.12	7.04	7.21	7.26	[7.18, 7.36]

Note: Simulations 1 and 2 were done according to a model having sex, age, and activity level as covariates. In Simulation 1, all three covariates were included when calculating α_{loc}; in Simulation 2, only the intercept. Simulations 3 and 4 were done according to a model having four environmental covariates correlated with four genetic markers, in addition to sex, age, and activity. In Simulation 3, all seven covariates were included when calculating α_{loc}; in Simulation 4, only sex, age, and activity. The Bonferroni α_{loc} is 5×10^{-5} and the Šidák α_{loc} is 5.13×10^{-5} for all simulations.

of positive ρ on α_{loc} found by order 1–4 approximations to control FWER at the .05 level. Also, the "true" α_{loc} was calculated without approximation (based on the true joint distribution, that is solving $1 - \gamma_{100} = 0.05$ for α_{loc}).

The inverses of both kinds of matrices contain only negative off-diagonal entries, which ensures a property called MTP_2 (Karlin & Rinott, 1981) for the density of $|\boldsymbol{T}|$, which implies that the product-type approximations γ_k of Section 3.2 are nondecreasing in k (Glaz & Johnson, 1984), making the α_{loc} of the FWER order k approximations nondecreasing in k.

For AR(1), the effect of ρ on α_{loc} was small for $\rho < 0.4$ (Figure 5), so in this case, there would be no gain in using FWER approximation instead of Šidák. For larger ρ, FWER order 2 approximation provides an improvement over Šidák. The improvements through order 3 and 4 up to the α_{loc} based on the true joint distribution, were smaller. Also for compound symmetry, order 2, 3, and 4 provide improvement over Šidák for larger ρ, but with this strong correlation structure, the α_{loc} of the true joint distribution is significantly larger.

In the extreme case $\rho = 1$, for all k γ_k equals the probability of nonrejection of a single null hypothesis; see (8). Hence, in this case all FWER approximations as well as using the true joint distribution will yield an α_{loc} equal to the significance level of the test (.05 in Figure 5).

4.7 | Efficiency and computational details

The computations of Sections 4.2 and 4.3 are well suited for parallelization, as they can be performed separately for each chromosome. For the maxT method, the permutations can additionally be distributed among cores. The computations were done on a computing cluster consisting of 42 computing nodes, most of them 2×10-core Intel Xeon E5-2630 v4 2.20 GHz Dell PE630 computers.

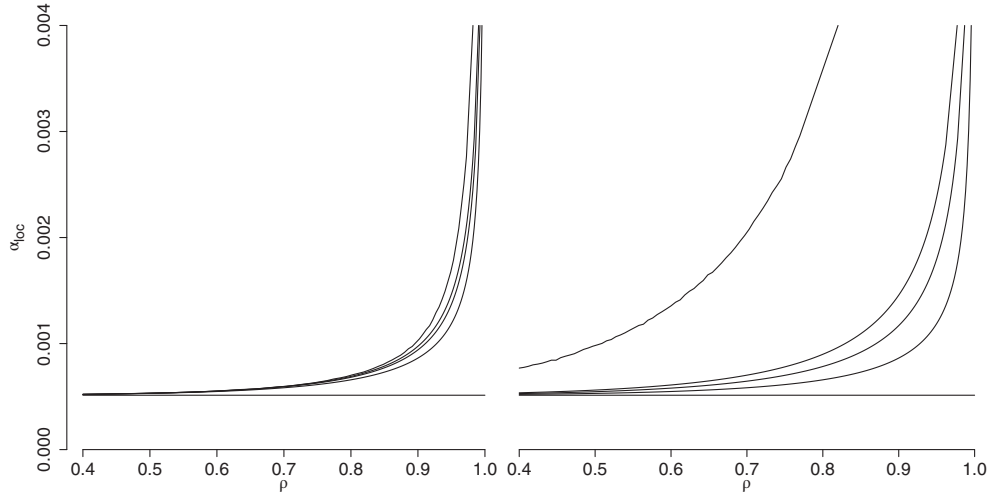


FIGURE 5 Local significance level α_{loc} controlling FWER at the .05 level as a function of the parameter ρ of an AR(1) (left) and a compound symmetry (right) correlation matrix for 100 markers. The horizontal line corresponds to Šidák correction (FWER order 1 approximation), then α_{loc} is increasing with the order of the approximation (order 2–4; the three curves in the middle). The uppermost curve shows α_{loc} based on the true joint distribution

For order 2 approximation, the total CPU time (summed over all cores) was dominated by the time for reading marker genotypes from files, meaning that the total time for analyzing GWAS data using order 2 approximation is not much larger than using Bonferroni or Šidák correction (Table 1). For order 3 approximation, where the Miwa algorithm was used (see Section 3.3), the CPU time was a couple of orders of magnitudes larger, and for maxT with 10^6 permutations 3–4 orders of magnitudes larger.

When solving $1 - \gamma_k = 0.05$ for α_{loc}, $k = 2, 3$, the bisection function `uniroot` of R was used. The argument `tol` was set to 10^{-14}. As lower end-point of the search interval, 0.05/672,972 was used for TOP and 0.05/123,497 for VO_2-max (Bonferroni bounds), and as upper end-points 10^{-7} and 10^{-6}, respectively. For order 2 approximations, `uniroot` needed four iterations for both datasets. For order 3, five iterations were needed for TOP and six for VO_2-max.

A problem that will arise in practice when doing order k approximation is correlations of ± 1 between statistics involved in the k- and $k-1$-dimensional integrals, leading to a singular correlation matrix. In this case, a factor

$$\frac{P(O_{j-k+1} \cap \cdots \cap O_j)}{P(O_{j-k+1} \cap \cdots \cap O_{j-1})} = P(O_j | O_{j-k+1} \cap \cdots \cap O_{j-1})$$

of γ_k (8) involving an offending statistic will be equal to 1, so the factor is simply set to 1.

Also for the simulation study of Section 4.4, parallelization was used. Each of the four situations took between 380 and 570 CPU hours.

5 | DISCUSSION AND CONCLUSION

The FWER order k approximation method offers FWER control when the vector of test statistics is approximately multivariate normal and is well suited for GLMs when approximate normality of the score test statistics is satisfied. We have applied it in a GWA study setting using score test statistics. In particular, we recommend the order 2 approximation, which offers improvement over Bonferroni or Šidák correction, without adding significantly to computing time. Further improvements toward the true multivariate distribution are possible using higher order approximations. Currently, available algorithms are time-consuming. However, the research into better and faster integration of multivariate normal densities is ongoing, and Botev (2017) provides an interesting new approach, applicable for dimensions smaller than or equal to 100. We expect improvements in algorithms and hardware to make higher order approximations feasible in the future.

Conneely and Boehnke (2007) introduced a method to calculate Bonferroni–Holm-type FWER adjusted p-values from score tests in GLMs with multiple responses (traits) and multiple genetic models. The core of their method is the multivariate integral arising from (7), and their correlation matrix between score statistics for one normal trait, or one nonnormal trait without environmental covariates present, coincides with our correlation matrix (4). However, for a nonnormal trait and environmental covariates present, the scaling of individual observations provided by the Λ matrix in (2) is not present in their method, and their correlation matrix differs from ours.

The adjusted p-values of Conneely and Boehnke were calculated by numerical integration using the `pmvnorm` function in the R package mvtnorm (see Section 3.3) for dimensions in the order of hundreds. Larger dimensions, as observed in GWA studies, were not handled, but

Conneely and Boehnke suggested to break the analysis up into independent blocks of hundreds of tests each and calculate adjusted p-values for each block, and then adjust the blockwise adjusted p-values by Šidák or Holm-type methods. We suggest that adjusted p-values or α_{loc} can instead be calculated by multiplying probabilities of the independent blocks, as we have done in Section 4.3.

In the maxT method (Section 3.6) of Westfall and Young (1993), there is no general way of including environmental covariates in a nonnormal GLM. The method is extremely time-consuming for GWA studies and nondeterministic (see confidence intervals in Table 1). When applicable (normal responses or no environmental covariates), however, maxT is expected to give better power than low-order FWER approximations.

Parametric bootstrap is an alternative to the maxT method that can also be used in GWA studies (Seaman & Müller-Myhsok, 2005), and it does not depend on the exchangeability assumption. However, the method is as time-consuming as maxT, and estimation of nuisance parameters makes the control of FWER uncertain.

In conclusion, the FWER order k approximation method is a considerable generalization of the intersection approximations by Moskvina and Schmidt (2008) and Dickhaus and Stange (2013) and can be used to control FWER for GWA data modeled by GLMs when normality of the test statistics is assumed (thus excluding unbalanced binomial designs); in particular, phenotypes can be discrete or continuous, and environmental covariates can be included. The method takes correlation structure of markers and test statistics into account and provides a local significance level, α_{loc}, for the individual tests, meaning that the null hypothesis of no association between phenotype and genetic marker should be rejected if the (unadjusted) p-value of a test is less than α_{loc}. We have applied the method to GWA data and shown that it is a powerful alternative to the Bonferroni and Šidák methods—methods that does not take correlation structure into account—especially in situations where permutation methods cannot be used. We found a substantial increase in α_{loc} over Bonferroni and Šidák already at the inexpensive order 2 approximation.

ACKNOWLEDGEMENTS

The authors would like to thank Per Kristian Hove (Norwegian University of Science and Technology) for help using a computing cluster for the computations, Jelle J. Goeman (Leiden University Medical Centre, Leiden, the Netherlands) and anonymous reviewers for valuable comments. The PhD position of the first author was funded by the Liaison Committee between the Central Norway Regional Health Authority and the Norwegian University of Science and Technology. The Nord-Trøndelag Health Study (The HUNT study) is a collaboration between HUNT Research Centre (Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology), Trøndelag County Council, Central Norway Regional Health Authority, and the Norwegian Institute of Public Health.

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

The statistical analysis was performed using R (R Core Team, 2015), and the preprocessing of the genetic data was done using PLINK (Purcell et al., 2007). An R package is available at https://github.com/oyvind-bakke/fwerapprox containing functions to compute score statistics (3) and

correlations (4), and perform FWER order k approximations (8), including a fast function for order 2 (9).

ORCID

Øyvind Bakke https://orcid.org/0000-0003-4592-7734
Mette Langaas https://orcid.org/0000-0002-5714-0288

REFERENCES

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, *11*, 375–386.

Aspenes, S. T., Nilsen, T. I. L., Skaug, E.-A., Bertheussen, G. F., Ellingsen, Ø., Vatten, L., & Wisløff, U. (2011). Peak oxygen uptake and cardiovascular risk factors in 4631 healthy women and men. *Medicine & Science in Sports & Exercise*, *43*, 1465–1473.

Athanasiu, L., Mattingsdal, M., Kähler, A. K., Brown, A., Gustafsson, O., Agartz, I., et al. (2010). Gene variants associated with schizophrenia in a Norwegian genome-wide study are replicated in a large European cohort. *Journal of Psychiatric Research*, *44*, 748–753.

Block, H. W., Costigan, T., & Sampson, A. R. (1992). Product-type probability bounds of higher order. *Probability in the Engineering and Informational Sciences*, *6*, 349–370.

Botev, Z. (2017). The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*, 125–148.

Chen, Z., & Liu, Q. (2011). A new approach to account for the correlations among single nucleotide polymorphisms in genome-wide association studies. *Human Heredity*, *72*, 1–9.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, *10*, 417–451.

Commenges, D. (2003). Transformations which preserve exchangeability and application to permutation tests. *Journal of Nonparametric Statistics*, *15*, 171–185.

Conneely, K. N., & Boehnke, M. (2007). So many correlated tests, so little time! Rapid adjustment of p values for multiple correlated tests. *The American Journal of Human Genetics*, *81*, 1158–1168.

Conover, W. J. (1980). *Practical nonparametric statistics* (2nd ed.). New York, NY: Wiley.

R Core Team. (2015). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Dey, R., Schmidt, E. M., Abecasis, G. R., & Lee, S. (2017). A fast and accurate algorithm to test for binary phenotypes and its application to phewas. *The American Journal of Human Genetics*, *101*, 37–49.

Dickhaus, T. (2014). *Simultaneous statistical inference*. Berlin, Germany / Heidelberg: Springer-Verlag.

Dickhaus, T., & Stange, J. (2013). Multiple point hypothesis test problems and effective numbers of tests for control of the family-wise error rate. *Calcutta Statistical Association Bulletin*, *65*, 123–144.

Djurovic, S., Gustafsson, O., Mattingsdal, M., Athanasiu, L., Bjella, T., Tesli, M., et al. (2010). A genome-wide association study of bipolar disorder in Norwegian individuals, followed by replication in Icelandic sample. *Journal of Affective Disorders*, *126*, 312–316.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., et al. (2002). The structure of haplotype blocks in the human genome. *Science*, *296*, 2225–2229.

Galwey, N. W. (2009). A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, *33*, 559–568.

Gao, X., Starmer, J., & Martin, E. R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, *32*, 361–369.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, *1*, 141–150.

Genz, A. (1993). Comparison of methods for the computation of multivariate normal probabilities. *Computing Science and Statistics*, *25*, 400–405.

Genz, A., & Bretz, F. (2002). Comparison of methods for the computation of multivariate t probabilities. *Journal of Computational and Graphical Statistics*, *11*, 950–971.

Genz, A., & Bretz, F. (2009). *Computation of multivariate normal and t probabilities Lecture Notes in Statistics* (Vol. *195*). Berlin, Heidelberg: Springer-Verlag.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2016). mvtnorm: Multivariate normal and t distributions. (R package version 1.05). https://CRAN.R-project.org/package=mvtnorm.

Glaz, J., & Johnson, B. M. (1984). Probability inequalities for multivariate distributions with dependence structures. *Journal of the American Statistical Association*, *79*, 436–440.

Goeman, J. J., & Solari, A. (2010). The sequential rejection principle of familywise error control. *The Annals of Statistics*, *38*, 3782–3810.

Hemerik, J., Goeman, J., & Finos, L. (2019). Robust testing in generalized linear models by sign-flipping score contributions. arXiv preprint arXiv:1909.03796.

Hummel, M., Meister, R., & Mansmann, U. (2008). Globalancova: Exploration and assessment of gene group effects. *Bioinformatics*, *24*, 78–85.

Karlin, S., & Rinott, Y. (1981). Total positivity properties of absolute value multinormal variables with application to confidence interval estimates and related probabilistic inequalities. *The Annals of Statistics*, *9*, 1035–1049.

Langaas, M., & Bakke, Ø. (2014). Robust methods to detect disease-genotype association in genetic association studies: Calculate p-values using exact conditional enumeration instead of simulated permutations or asymptotic approximations. *Statistical Applications in Genetics and Molecular Biology*, *13*, 675–692.

Latała, R., & Matlak, D. (2017). *Royen's proof of the Gaussian correlation inequality*. In *Geometric aspects of functional analysis* (pp. 265–275). New York, NY: Springer.

Li, J., & Ji, L. (2005). Adjusting multiple testing in multi locus analyses using the eigenvalues of the correlation matrix. *Heredity*, *95*, 221–227.

Loe, H., Rognmo, Ø., Saltin, B., & Wisløff, U. (2013). Aerobic capacity reference data in 3816 healthy men and women 20–90 years. *PLoS One*, *8*, e64319.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman and Hall.

Meinshausen, N., Maathuis, M. H., & Bühlmann, P. (2011). Asymptotic optimality of the Westfall-Young permutation procedure for multiple testing under dependence. *Annals of Statistics*, *39*, 3369–3391.

Miwa, A., Hayter, J., & Kuriki, S. (2003). The evaluation of general non-centered orthant probabilities. *Journal of the Royal Statistical Society, Series B*, *65*, 223–234.

Moore, A., Enquobahrie, D. A., Sanchez, S. E., Ananth, C. V., Pacora, P. N., & Williams, M. A. (2012). A genome-wide association study of variations in maternal cardiometabolic genes and risk of placental abruption. *International Journal of Molecular Epidemiology and Genetics*, *3*, 305–313.

Moskvina, V., & Schmidt, K. M. (2008). On multiple-testing correction in genome-wide association studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, *32*, 567–573.

Nyholt, D. R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *The American Journal of Human Genetics*, *74*, 765–769.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*, 904–909.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., … Sham, P. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*, 559–575.

Royen, T. (2014). A simple proof of the Gaussian correlation conjecture. *Far East Journal of Theoretical Statistics*, *48*, 139–145.

Sasieni, P. D. (1997). From genotypes to genes: Doubling the sample size. *Biometrics*, *53*, 1253–1261.

Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M., & Poland, G. A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *The American Journal of Human Genetics*, *70*, 425–434.

Seaman, S. R., & Müller-Myhsok, B. (2005). Rapid simulation of *P* values for product methods and multiple-testing adjustment in association studies. *The American Journal of Human Genetics*, *76*, 399–408.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, *62*, 626–633.

Slager, S. L., & Schaid, D. J. (2001). Case–control studies of genetic markers: Power and sample size approximations for Armitage's test for trend. *Human Heredity*, *52*, 149–153.

Smyth, G. K. (2003). *Pearson's goodness of fit statistic as a score test statistic*. In D. R. Goldstein (Ed.), *Science and statistics: A festschrift for Terry Speed IMS Lecture Notes—Monograph Series* (Vol. *40*, pp. 115–126). Beachwood, OH: Institute of Mathematical Statistics.

Stange, J., Loginova, N., & Dickhaus, T. (2016). Computing and approximating multivariate chi-square probabilities. *Journal of Statistical Computation and Simulation*, *86*, 1233–1247.

Thompson, W. R. (1936). On confidence ranges for the median and other expectation distributions for populations of unknown distribution form. *The Annals of Mathematical Statistics*, *7*, 122–128.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: Comparison with p-values. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, *33*, 79–86.

Weir, B. S. (2008). Linkage disequilibrium and association mapping. *Annual Review Genomics and Human Genetics*, *9*, 129–142.

Westfall, P. H., & Troendle, J. F. (2008). Multiple testing with minimal assumptions. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *50*, 745–755.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing*. New York, NY: John Wiley and Sons, Inc.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *Neuroimage*, *92*, 381–397.

How to cite this article: Halle KK, Bakke Ø, Djurovic S, et al. Computationally efficient familywise error rate control in genome-wide association studies using score tests for generalized linear models. *Scand J Statist*. 2020;47:1090–1113. https://doi.org/10.1111/sjos.12451

APPENDIX

Uncorrelated environmental and genetic covariates

Two n-dimensional vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ of observations have zero sample correlation if their centered observations are orthogonal, $0 = (\boldsymbol{x}_1 - \bar{x}_1\mathbf{1})^{\mathrm{T}}(\boldsymbol{x}_2 - \bar{x}_2\mathbf{1}) = \boldsymbol{x}_1^{\mathrm{T}}(I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{x}_2$. If X_1 and X_2 are two matrices, then near zero sample correlation of each combination of a column of X_1 and a column of X_2 can be written compactly as $X_1^{\mathrm{T}}(I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})X_2 \approx O$, or $X_1^{\mathrm{T}}X_2 \approx \frac{1}{n}X_1^{\mathrm{T}}\mathbf{1}\mathbf{1}^{\mathrm{T}}X_2$, where O denotes a null matrix.

Assume now that all sample correlations of a column of X_{g} and a column of ΛX_{e} are near zero, $X_{\mathrm{g}}^{\mathrm{T}}(I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\Lambda X_{\mathrm{e}} \approx O$. In addition, we assume that all correlations of genetic first-order multiplicative interactions and $\Lambda\mathbf{1}$ are near zero, $(\boldsymbol{x}_j \circ \boldsymbol{x}_k)^{\mathrm{T}}\Lambda\mathbf{1} \approx \frac{1}{n}(\boldsymbol{x}_j \circ \boldsymbol{x}_k)^{\mathrm{T}}\mathbf{1}\mathbf{1}^{\mathrm{T}}\Lambda\mathbf{1} = \frac{1}{n}(\mathrm{tr}\Lambda)\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_k$ for all columns $\boldsymbol{x}_j$, $\boldsymbol{x}_k$ of X_{g}, where $\circ$ denotes entrywise multiplication. Note that, under the null hypothesis, Λ is a function of environmental covariates only.

Next, note that the statistic $\boldsymbol{U}_{\mathrm{g|e}}(1)$ remains unchanged if the columns of $X_{\mathrm{g}}^{\mathrm{T}}$ are centered, that is, $X_{\mathrm{g}}^{\mathrm{T}}$ is replaced by $X_{\mathrm{g}}'^{\mathrm{T}} = X_{\mathrm{g}}^{\mathrm{T}}(I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})$. This is because $\mathbf{1}^{\mathrm{T}}(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_{\mathrm{e}}) = 0$, which follows from the estimation equations, $X_{\mathrm{e}}^{\mathrm{T}}(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_{\mathrm{e}}) = 0$, for $\hat{\boldsymbol{\mu}}_{\mathrm{e}}$. Making this substitution, by the first assumption above, (2) reduces to $\phi^2 V_{\mathrm{g|e}} = X_{\mathrm{g}}'^{\mathrm{T}}\Lambda X_{\mathrm{g}}'$.

The jk entry of this matrix is $(\boldsymbol{x}_j \circ \boldsymbol{x}_k)^{\mathrm{T}}\Lambda\mathbf{1}$, where $\boldsymbol{x}_j$ denotes the jth column of X_{g}'. By the second assumption above, this is approximately equal to $\frac{1}{n}(\mathrm{tr}\Lambda)\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_k$, so that $\phi^2 V_{\mathrm{g|e}} \approx \frac{1}{n}(\mathrm{tr}\Lambda)X_{\mathrm{g}}'^{\mathrm{T}}X_{\mathrm{g}}' = \frac{1}{n}(\mathrm{tr}\Lambda)X_{\mathrm{g}}^{\mathrm{T}}(I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})X_{\mathrm{g}}$.