Powerful extreme phenotype sampling designs and score tests for genetic association studies

THEA BJØRNLAND[1], ANJA BYE[2], EINAR RYENG[3],
ULRIK WISLØFF[2], METTE LANGAAS[1]

[1] *Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway*
[2] *Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway*
[3] *Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway*

Abstract

We consider cross-sectional genetic association studies (common and rare variants) where non-genetic information is available, or feasible to obtain for N individuals, but where it is infeasible to genotype all N individuals. We consider continuously measurable Gaussian traits (phenotypes). Genotyping $n < N$ extreme phenotype individuals can yield better power to detect phenotype-genotype associations, as compared to randomly selecting n individuals. We define a person as having an extreme phenotype if the observed phenotype is above a specified threshold or below a specified thresholds. We consider a model where these thresholds can be tailored to each individual. The classical extreme sampling design is to set equal thresholds for all individuals. We introduce a design (z-extreme sampling) where personalized thresholds are defined based on the residuals of a regression model including only non-genetic (fully available) information. We derive score tests for the situation where only n extremes are analyzed (complete case analysis), and for the situation where the non-genetic information on $N - n$ non-extremes is included in the analysis (all case analysis). For the classical design, all case analysis is generally more powerful than complete case analysis. For the z-extreme sample, we show that all case and complete case tests are equally powerful. Simulations and data analysis also show that z-extreme sampling

is at least as powerful as the classical extreme sampling design and the classi-
cal design is shown to be at times less powerful than random sampling. The
method of dichotomizing extreme phenotypes is also discussed.

Key words: GWAS, rare variants, outcome-dependent sampling, residual-
based sampling, the HUNT study

1   Introduction

Extreme phenotype sampling is a a two-stage design for genetic association studies,
or other association studies where the covariate of interest is expensive or infeasible
to obtain for all individuals. Selective genotyping of extreme-phenotype individ-
uals has been proposed as a strategy for achieving good statistical power under
sample size limitations.[1–8] Inexpensive 'non-genetic' information (phenotype/trait
measurements, age, sex, lifestyle variables, etc.) is collected on N individuals in
the first stage. In the second stage, $n < N$ individuals are selected for genotyp-
ing. Criteria for selection are based on stage 1 observations with the intention to
increase the power to detect genotype-phenotype associations, as compared to ran-
domly selecting n individuals. For simple linear regression models, Huang and Lin
proposed likelihood methods for extreme samples that made full use of the genetic
data, that modelled the continuous rather than a discretized phenotype and prop-
erly accounted for the selective genotyping design.[9] Tao *et al* developed methods
for parameter estimation and a Wald test for two-stage sampling designs.[10] Derkach
et al derived score tests for similar designs.[11] Extreme phenotype sampling has also
been proposed for studies of rare genetic variants.[12–15]

Consider a common biallelic genetic variant g (SNP), an additive genetic model,
a Gaussian phenotype y and the model $y = \beta_0 + \beta_g g + \varepsilon$, where $\varepsilon \sim \mathrm{N}(0, \sigma^2)$.
The genetic covariate g takes values 0 (zero minor alleles), 1 (one minor allele) or
2 (two minor alleles). To establish whether $\beta_g \neq 0$, a powerful sample will have a
high proportion of individuals with genotypes 0 and 2. Assuming Hardy-Weinberg
equilibrium, if the minor allele frequency of a SNP is 0.25, then in any random
sample from the population approximately 56% of individuals have genotype 0,
37% will have genotype 1, and only 6% will have genotype 2. If indeed the genetic
variant has an additive effect on the phenotype, and there are no other covariates
influencing the phenotype, then in one end of the empirical phenotype distribution
we will find an increased proportion of subjects with genotype 2, while in the other
end we will mainly find subjects with genotype 0. The extreme sample can therefore
give better power than random sampling. For studies of rare genetic variants (minor
allele frequency below 5%) an extreme sample should have a higher proportion of
individuals with at least one minor allele if the genetic variant is causal.

If the phenotype is influenced by other 'non-genetic' covariates x (e.g. age, sex, lifestyle), $y = \beta_0 + \beta_1 x + \beta_g g + \varepsilon$, then it is not given that the phenotypic extremes will have a more favorable distribution of genotypes than a random sample. Therefore, we consider a design where extremes are defined based on a 'residual phenotype' ($z = y - \widehat{\beta}_0 - \widehat{\beta}_1 x$), i.e. the observed phenotype after adjustment for relevant non-genetic covariates. We refer to this design as the z-extreme sampling design.

Define *complete cases* as the $n < N$ individuals that are genotyped. Under extreme phenotype sampling, the most straightforward association analysis for the complete cases is to compare the two extreme groups using models for binary outcomes.[7,16] This approach has been shown to be less powerful than other methods.[9] A complete case test should rather be derived based on the probability distribution of the extremes.[9,17] Analysis of *all cases* (n complete cases and $N - n$ cases with missing genotype information) is based on methods for missing at random covariates.[10,11] When there are no non-genetic covariates in the model, complete case and all case analysis are equally powerful.[9,11] Otherwise, all case analysis is generally more powerful than complete case analysis. We prove that when extremes are defined based on the residual phenotype (z-extreme sampling), the complete case and all case score tests are equally powerful, regardless of the impact of non-genetic covariates.

When the sampling design is based on the construction of the residual phenotype z, it is not unnatural to consider testing for an association directly with this adjusted phenotype. Although these residuals by design are not independent, we show that score tests 'naively' applied to residual phenotype models ($z = \gamma_0 + \gamma_g g + \varepsilon$) will control the type-1 error rate at the desired significance level. If the genetic covariate is correlated with some of the non-genetic covariates in the model (x) then such residual phenotype tests are shown to be less powerful than tests based on the original regression model ($y = \beta_0 + \beta_1 x + \beta_g g + \varepsilon$).

In Section 2 of this paper we give some statistical background and specific results for random sampling. In Section 3 we present score tests for testing association with one common genetic variant in extreme samples. The multivariate version of this theory is given in the Supplementary Material. In Section 4 we consider tests of rare genetic variants. We derive association tests similar to the SKAT test[18,19] for rare genetic variants under extreme sampling. In Section 5 we illustrate the results of Sections 3 and 4 by simulations, in Section 6 we apply methods to data from the HUNT study (Helseundersøkelsen i Nord-Trøndelag),[20] and in Section 7 we discuss our findings.

2 Statistical background

We consider the linear model

$$\mathbf{Y} = \mathrm{X}\,\boldsymbol{\beta} + \mathbf{g}\beta_{\mathrm{g}} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{MVN}(\mathbf{0}, \sigma^2\,\mathrm{I}), \tag{1}$$

where $\mathbf{Y}$ is an N-vector of phenotype observations, X an $N \times (d+1)$ matrix of the intercept and d non-genetic covariates ($d+1 \ll N$), $\mathbf{g}$ an N-vector of genotype observations and I is the $N \times N$ identity matrix. The residual vector $\boldsymbol{\varepsilon}$ is multivariate normal (MVN) with mean $\mathbf{0}$ and covariance matrix $\sigma^2\,\mathrm{I}$. We test $H_0 : \beta_{\mathrm{g}} = 0$ against $H_1 : \beta_{\mathrm{g}} \neq 0$ (genotype-phenotype association) using score tests. We further generalize model (1) to

$$\mathbf{Y} = \mathrm{X}\,\boldsymbol{\beta} + \mathrm{G}\,\boldsymbol{\beta}_{\mathrm{g}} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{MVN}(\mathbf{0}, \sigma^2\,\mathrm{I}), \tag{2}$$

where the $N \times \nu$ genotype matrix G represents a selection of ν genetic variants ($\nu \ll N$), and derive score tests for testing $H_0 : \boldsymbol{\beta}_{\mathrm{g}} = \mathbf{0}$ against $H_1 : \boldsymbol{\beta}_{\mathrm{g}} \neq \mathbf{0}$.

Definition 2.1 (The score test for $H_0 : \boldsymbol{\beta}_{\mathrm{g}} = \mathbf{0}$) *Consider a log likelihood function $l(\boldsymbol{\beta}_{\mathrm{g}}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of 'nuisance parameters' and $\boldsymbol{\beta}_{\mathrm{g}}$ are the parameters of interest. When H_0 is assumed true ($\boldsymbol{\beta}_{\mathrm{g}} = \mathbf{0}$) the maximum likelihood estimators of nuisance parameters, denoted $\widehat{\boldsymbol{\theta}}$, are found by solving $\partial l / \partial \boldsymbol{\theta}(\boldsymbol{\beta}_{\mathrm{g}} = \mathbf{0}) = \mathbf{0}$. Define the score statistic*

$$\mathbf{U}_{\boldsymbol{\beta}_{\mathrm{g}}} = \frac{\partial l}{\partial \boldsymbol{\beta}_{\mathrm{g}}}\left(\boldsymbol{\beta}_{\mathrm{g}} = \mathbf{0}, \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}\right),$$

and the information matrix

$$\mathbf{F}(\boldsymbol{\beta}_{\mathrm{g}}, \boldsymbol{\theta}) = \begin{bmatrix} \mathbf{F}_{\boldsymbol{\beta}_{\mathrm{g}}\boldsymbol{\beta}_{\mathrm{g}}}(\boldsymbol{\beta}_{\mathrm{g}}, \boldsymbol{\theta}) & \mathbf{F}_{\boldsymbol{\beta}_{\mathrm{g}}\boldsymbol{\theta}}(\boldsymbol{\beta}_{\mathrm{g}}, \boldsymbol{\theta}) \\ \mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\beta}_{\mathrm{g}}}(\boldsymbol{\beta}_{\mathrm{g}}, \boldsymbol{\theta}) & \mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\beta}_{\mathrm{g}}, \boldsymbol{\theta}) \end{bmatrix} = -\mathrm{E}\begin{bmatrix} \frac{\partial^2 l}{\partial \boldsymbol{\beta}_{\mathrm{g}} \partial \boldsymbol{\beta}_{\mathrm{g}}^{\mathrm{T}}} & \frac{\partial^2 l}{\partial \boldsymbol{\beta}_{\mathrm{g}} \partial \boldsymbol{\theta}^{\mathrm{T}}} \\ \frac{\partial^2 l}{\partial \boldsymbol{\theta} \partial \boldsymbol{\beta}_{\mathrm{g}}^{\mathrm{T}}} & \frac{\partial^2 l}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \end{bmatrix}.$$

The variance of the score statistic is

$$\mathrm{Var}(\mathbf{U}_{\boldsymbol{\beta}_{\mathrm{g}}}) = \mathbf{F}_{\boldsymbol{\beta}_{\mathrm{g}}\boldsymbol{\beta}_{\mathrm{g}}}(\mathbf{0}, \widehat{\boldsymbol{\theta}}) - \mathbf{F}_{\boldsymbol{\beta}_{\mathrm{g}}\boldsymbol{\theta}}(\mathbf{0}, \widehat{\boldsymbol{\theta}})\mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}(\mathbf{0}, \widehat{\boldsymbol{\theta}})\mathbf{F}_{\boldsymbol{\theta}\boldsymbol{\beta}_{\mathrm{g}}}(\mathbf{0}, \widehat{\boldsymbol{\theta}}).$$

The score test statistic $T = \mathbf{U}_{\boldsymbol{\beta}_{\mathrm{g}}}^{\mathrm{T}}\,\mathrm{Var}(\mathbf{U}_{\boldsymbol{\beta}_{\mathrm{g}}})^{-1}\mathbf{U}_{\boldsymbol{\beta}_{\mathrm{g}}}$ is asymptotically χ^2_ν-distributed under the null. When β_{g} is a scalar then $\nu = 1$ and $T = U_{\beta_{\mathrm{g}}}^2 / \mathrm{Var}(U_{\beta_{\mathrm{g}}})$.

Definition 2.2 (Residual phenotype) *Consider the null model $\mathbf{Y} = \mathrm{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, and let $\widehat{\boldsymbol{\beta}} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathbf{Y}$ be the maximum likelihood estimator for $\boldsymbol{\beta}$. Define the residual phenotype as $\mathbf{Z} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathrm{X}\,\widehat{\boldsymbol{\beta}} = (\mathrm{I} - \mathrm{H})\mathbf{Y}$, where $\mathrm{H} = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}$.*

If the null hypothesis is true then $\mathbf{Z} \sim \mathrm{MVN}(\mathbf{0}, \sigma^2(\mathrm{I} - \mathrm{H}))$. The residual phenotypes $Z_1, \ldots, Z_N$ are therefore generally not i.i.d. The analogs to models (1) and (2) are the *residual phenotype models*

$$\mathbf{Z} = \mathbf{1}\gamma_0 + \mathbf{g}\gamma_{\mathrm{g}} + \boldsymbol{\varepsilon}_z, \quad \text{and} \quad \mathbf{Z} = \mathbf{1}\gamma_0 + \mathrm{G}\,\boldsymbol{\gamma}_{\mathrm{g}} + \boldsymbol{\varepsilon}_z. \tag{3}$$

The parameter γ_{g} ($\boldsymbol{\gamma}_{\mathrm{g}}$) can be interpreted as a surrogate for β_{g} ($\boldsymbol{\beta}_{\mathrm{g}}$) but we assume throughout that the main interest lies in testing $H_0 : \beta_{\mathrm{g}} = 0$ ($H_0 : \boldsymbol{\beta}_{\mathrm{g}} = \mathbf{0}$).

2.1 Score tests for the linear regression model

In this section only, assume that $(Y_i, \mathbf{x}_i, \mathrm{g}_i)$ is observed for all N individuals, randomly sampled from some (infinite) population of interest. The log likelihood function is $l = -N \log(\sigma) - \frac{1}{\sigma^2} \sum_{i=1}^{N} (Y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} - \mathrm{g}_i \beta_{\mathrm{g}})^2$. The score test statistic (Definition 2.1) for testing $H_0 : \beta_{\mathrm{g}} = 0$ in model (1) is given by $T = U_{\beta_{\mathrm{g}}}^2 / \mathrm{Var}(U_{\beta_{\mathrm{g}}})$, where

$$U_{\beta_{\mathrm{g}}} = \frac{1}{\widehat{\sigma}^2} \mathbf{g}^{\mathrm{T}} (\mathrm{I} - \mathrm{H}) \mathbf{Y}, \quad \mathrm{Var}(U_{\beta_{\mathrm{g}}}) = \frac{1}{\widehat{\sigma}^2} \mathbf{g}^{\mathrm{T}} (\mathrm{I} - \mathrm{H}) \mathbf{g}. \tag{4}$$

The hat matrix is defined under the null by $\mathrm{H} = \mathrm{X}(\mathrm{X}^{\mathrm{T}} \mathrm{X})^{-1} \mathrm{X}^{\mathrm{T}}$, and the maximum likelihood estimator for σ^2 is $\widehat{\sigma}^2 = \frac{1}{N} \mathbf{Y}^{\mathrm{T}} (\mathrm{I} - \mathrm{H}) \mathbf{Y}$. The test statistic T has an asymptotic χ_1^2 distribution when the null hypothesis is true, and the null hypothesis is rejected when $T > t_\alpha$, where t_α is a critical value from the χ_1^2 distribution. 'Naively' applying the same test to the residual phenotype model (3), we obtain $T_z = U_{\gamma_{\mathrm{g}}}^2 / \mathrm{Var}(U_{\gamma_{\mathrm{g}}})$, where

$$U_{\gamma_{\mathrm{g}}} = \frac{1}{\widehat{\sigma}_z^2} \mathbf{g}^{\mathrm{T}} (\mathrm{I} - \frac{1}{N} \mathbf{1} \mathbf{1}^{\mathrm{T}}) \mathbf{Z}, \quad \mathrm{Var}(U_{\gamma_{\mathrm{g}}}) = \frac{1}{\widehat{\sigma}_z^2} \mathbf{g}^{\mathrm{T}} (\mathrm{I} - \frac{1}{N} \mathbf{1} \mathbf{1}^{\mathrm{T}}) \mathbf{g}. \tag{5}$$

Note that since the assumption $\boldsymbol{\varepsilon}_z \sim \mathrm{MVN}(\mathbf{0}, \sigma^2 \mathrm{I})$ is generally incorrect, the asymptotic null distribution of T_z is not necessarily χ_1^2. However, an asymptotically valid test for $H_0 : \beta_{\mathrm{g}} = 0$ against $H_1 : \beta_{\mathrm{g}} \neq 0$ in model (1) when is to reject H_0 whenever $T_z > t_\alpha$, where t_α is a critical value from the χ_1^2-distribution. By *valid* we mean that the type-1 error rate is controlled at the desired level α; $\Pr(\text{reject } H_0 | H_0 \text{ true}) \leq \alpha$. The residual phenotype test is valid because $T_z \leq T$ (proof given in Supplementary Material Section 2.1), so that $T_z < t_\alpha$ whenever $T < t_\alpha$. Rejecting H_0 whenever $T_z > t_\alpha$ is then generally a less powerful test for $H_0 : \beta_{\mathrm{g}} = 0$ because the event $T_z > t_\alpha$ will occur less often than $T > t_\alpha$. When X and $\mathbf{g}$ are *uncorrelated* (sample correlation between each column of X and $\mathbf{g}$ is zero, $(\mathrm{X} - \frac{1}{N} \mathbf{1} \mathbf{1}^{\mathrm{T}} \mathrm{X})^{\mathrm{T}} (\mathbf{g} - \frac{1}{N} \mathbf{1} \mathbf{1}^{\mathrm{T}} \mathbf{g}) = \mathbf{0}$, then $T_z = T$. This result also generalizes to the score test for $H_0 : \boldsymbol{\beta}_{\mathrm{g}} = \mathbf{0}$ (model (2)) where the test statistics are compared to the χ_ν^2-distribution, see Supplementary Material Section 2.

2.2 Missing covariates

Consider a sample of size N (randomly generated from the infinite population of interest) where the genetic covariate is missing for $N - n$ individuals, while other non-genetic covariates and the response variable are observed for all N individuals.

Definition 2.3 (Complete case analysis) *Only the n individuals with observed genotype (g_i), non-genetic covariates ($\mathbf{x}_i$) and phenotype (y_i) are analyzed. The set of indexes $\{i : i \in \mathcal{C}\}$ denotes all completely observed individuals. Let $\mathbf{g}_\mathcal{C}$ be the*

n-vector of observed genotypes, $\mathbf{Y}_{\mathcal{C}}$ the complete case phenotype vector and $\mathrm{X}_{\mathcal{C}}$ the complete case covariate matrix.

Definition 2.4 (All case analysis) *For all case analysis we include the n complete case observations $(y_i, \mathbf{x}_i, g_i)$ for all $i \in \mathcal{C}$ and the $N - n$ observations of $(y_i, \mathbf{x}_i)$ for all $i \in \mathcal{M}$, where $\mathcal{M} = \{i \in \{1, \ldots, N\} : i \notin \mathcal{C}\}$.*

Let $\mathbf{R}$ be a random N-vector that indicates missing entries in the genetic covariate vector $\mathbf{g}$. The missing-mechanism is *missing completely at random* (MCAR) if $\Pr(\mathbf{R} = \mathbf{r}|\mathbf{Y}, \mathrm{X}, \mathbf{g}) = \Pr(\mathbf{R} = \mathbf{r})$ and *missing at random* (MAR) if $\Pr(\mathbf{R} = \mathbf{r}|\mathbf{Y}, \mathrm{X}, \mathbf{g}) = \Pr(\mathbf{R} = \mathbf{r}|\mathbf{Y}, \mathrm{X})$.[21] In the case of MCAR covariates, complete case analysis is performed by assuming that the complete case sample $(\mathbf{Y}_{\mathcal{C}}, \mathrm{X}_{\mathcal{C}}$ and $\mathbf{g}_{\mathcal{C}})$ is a random sample from the (infinite) population of interest. All case analysis is performed by imputing the mean genotype where genotypes are missing. The mean is estimated and imputed separately for genetically different subgroups. For sufficiently large samples, complete case and all case score tests are approximately equally powerful (Supplementary Material Section 5.5).

2.3 Extreme phenotype sampling

We define a general extreme phenotype sampling design where the classification rule (extreme or not extreme) can be tailored to each individual in the sample.

Definition 2.5 (Extreme phenotype sampling) *Individual i has an extreme phenotype if $Y_i < l_i$ or $Y_i > u_i$, where l_i and u_i are known thresholds. All individuals who are classified as extreme are selected for genotyping.*

Two special cases of this design are of particular interest:

Definition 2.6 (y-extreme sampling) *Individuals are y-extreme if $Y_i < l$ or $Y_i > u$, where thresholds l and u are equal for all $i \in \{1, \ldots, N\}$.*

Definition 2.7 (z-extreme sampling) *Individuals are z-extreme if $Z_i < l_z$ or $Z_i > u_z$, where thresholds l_z and u_z are equal for all $i \in \{1, \ldots, N\}$, and Z_i is the residual phenotype (Definition 2.2). By Definition 2.5, individual i has an extreme phenotype if $Y_i < l_i$ or $Y_i > u_i$, where $l_i = l_z + \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$ and $u_i = u_z + \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$.*

3 Single variant tests for extreme samples

We consider model (1) under extreme phenotype sampling (Definition 2.5) and score tests for $H_0 : \beta_{\mathrm{g}} = 0$ vs $H_1 : \beta_{\mathrm{g}} \neq 0$.

3.1 Complete case analysis

Using the conditional phenotype distribution $Y_i | (Y_i < l_i \cup Y_i > u_i)$, where classification rules l_i and u_i are determined before seeing the data, the likelihood for the complete cases (Definition 2.3) is

$$L_{\mathcal{C}} = \prod_{i \in \mathcal{C}} \frac{\frac{1}{\sigma} \phi \left(\frac{Y_i - \mu_i}{\sigma} \right)}{1 - \Phi \left(\frac{u_i - \mu_i}{\sigma} \right) + \Phi \left(\frac{l_i - \mu_i}{\sigma} \right)},$$

where ϕ is the density function and Φ is the cumulative probability distribution for the standard normal distribution, and $\mu_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \mathrm{g}_i \beta_{\mathrm{g}}$. Define (similarly to Tang[17]),

$$h_{ij} = \frac{-\phi \left(\frac{u_i - \mu_i}{\sigma} \right) \cdot \left(\frac{u_i - \mu_i}{\sigma} \right)^j + \phi \left(\frac{l_i - \mu_i}{\sigma} \right) \cdot \left(\frac{l_i - \mu_i}{\sigma} \right)^j}{1 - \Phi \left(\frac{u_i - \mu_i}{\sigma} \right) + \Phi \left(\frac{l_i - \mu_i}{\sigma} \right)}, \tag{6}$$

$$a_i = 1 - h_{i1} - h_{i0}^2, \quad b_i = -h_{i0} - h_{i2} - h_{i0}h_{i1}, \quad c_i = 2 - h_{i1} - h_{i3} - h_{i1}^2,$$

and let $\mathbf{h}_j$, $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ be the corresponding n-vectors for these expressions. Define $A = \mathrm{Diag}(\mathbf{a})$ and $\mathbf{W}_{\mathcal{C}} = \mathbf{Y}_{\mathcal{C}} + \sigma \mathbf{h}_0$. Let complete case maximum likelihood estimators of $\boldsymbol{\beta}$ and σ under the null ($\beta_{\mathrm{g}} = 0$) be denoted by a tilde $\sim$, and let $\widetilde{\mathbf{h}}_j$, $\widetilde{\mathbf{W}}_{\mathcal{C}}$, $\widetilde{A}$, $\widetilde{\mathbf{b}}$ and $\widetilde{\mathbf{c}}$ be the expressions $\mathbf{h}_j$, $\mathbf{W}_{\mathcal{C}}$, A, $\mathbf{b}$ and $\mathbf{c}$ evaluated at $\widetilde{\sigma}$, $\widetilde{\boldsymbol{\beta}}$ and $\beta_{\mathrm{g}} = 0$. The complete case score test statistic is defined by $T_{\mathcal{C}} = U_{\mathcal{C}}^2 / \mathrm{Var}(U_{\mathcal{C}})$, where

$$U_{\mathcal{C}} = \frac{1}{\widetilde{\sigma}^2} \mathbf{g}_{\mathcal{C}}^{\mathrm{T}} (\mathrm{I} - \mathrm{H}_{\mathcal{C}}) \widetilde{\mathbf{W}}_{\mathcal{C}}, \quad \mathrm{H}_{\mathcal{C}} = \mathrm{X}_{\mathcal{C}} (\mathrm{X}_{\mathcal{C}}^{\mathrm{T}} \mathrm{X}_{\mathcal{C}})^{-1} \mathrm{X}_{\mathcal{C}}^{\mathrm{T}}, \tag{7}$$

and

$$\mathrm{Var}(U_{\mathcal{C}}) = \frac{1}{\widetilde{\sigma}^2} \left(\mathbf{g}_{\mathcal{C}}^{\mathrm{T}} \widetilde{A} \mathbf{g}_{\mathcal{C}} - \mathbf{g}_{\mathcal{C}}^{\mathrm{T}} \begin{bmatrix} \widetilde{A} \mathrm{X}_{\mathcal{C}} & \widetilde{\mathbf{b}} \end{bmatrix} \begin{bmatrix} \mathrm{X}_{\mathcal{C}}^{\mathrm{T}} \widetilde{A} \mathrm{X}_{\mathcal{C}} & \mathrm{X}_{\mathcal{C}}^{\mathrm{T}} \widetilde{\mathbf{b}} \\ \widetilde{\mathbf{b}}^{\mathrm{T}} \mathrm{X}_{\mathcal{C}} & \mathbf{1}^{\mathrm{T}} \widetilde{\mathbf{c}} \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{X}_{\mathcal{C}}^{\mathrm{T}} \widetilde{A} \\ \widetilde{\mathbf{b}}^{\mathrm{T}} \end{bmatrix} \mathbf{g}_{\mathcal{C}} \right). \tag{8}$$

Details are given in Supplementary Material Section 4. When the null hypothesis is true, the test statistic $T_{\mathcal{C}}$ has an asymptotic χ_1^2 distribution.

3.1.1 y-extreme sampling

The score test statistic for the y-extreme sample follows directly from the above by inserting $l_i = l$ and $u_i = u$ into the functions h_{ij} (6). This leads to no simplifications unless $\mathrm{X}_{\mathcal{C}} = \mathbf{1}_n$. Then, the score test statistic $T_{\mathcal{C}}$ is equal to the test statistic one would find by assuming that the complete case sample ($\{i : i \in \mathcal{C}\}$) is a random sample.[17] Indeed, if the null hypothesis is true and there are no non-genetic covariates influencing the phenotype, then $Y_i \sim \mathrm{N}(\beta_0, \sigma^2) \; \forall i$, and all individuals have equal probability to be sampled for genotyping since $\Pr(Y_i < l \cup Y_i > u) = \Pr(Y_j < l \cup Y_j > u)$ for all i, j.

3.1.2 z-extreme sampling

Under z-extreme sampling (Definition 2.7) the expression for the complete case score test statistic can be simplified. Strictly, $\widehat{\boldsymbol{\beta}}$ should be estimated from a pilot study so that the z-extreme sampling rule is determined before seeing data. However, we may assume that $\widehat{\boldsymbol{\beta}}$ estimated from the full sample (large N) is unbiased and estimated with low variance and may therefore be regarded as deterministic. Recall that $\widetilde{\boldsymbol{\beta}}$ is the maximum likelihood estimator based on the complete cases. For large N and n, $u_i - \widetilde{\mu}_i = u_z + \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} - \mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} \approx u_z$, $l_i - \widetilde{\mu}_i = l_z + \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} - \mathbf{x}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} \approx l_z$, and $\widetilde{h}_{ij}$, $\widetilde{a}_i$, $\widetilde{b}_i$, $\widetilde{c}_i$ are (approximately) equal for all i. Equations (7) and (8) simplify to

$$U_{\mathcal{C}} \approx \frac{1}{\widetilde{\sigma}^2}\mathbf{g}_{\mathcal{C}}^{\mathrm{T}}(\mathrm{I} - \mathrm{H}_{\mathcal{C}})\mathbf{Y}_{\mathcal{C}}, \quad \mathrm{Var}(U_{\mathcal{C}}) \approx \frac{\widetilde{a}}{\widetilde{\sigma}^2}\mathbf{g}_{\mathcal{C}}^{\mathrm{T}}(\mathrm{I} - \mathrm{H}_{\mathcal{C}})\mathbf{g}_{\mathcal{C}}. \tag{9}$$

Observation 3.1 *Under z-extreme sampling $T_{\mathcal{C}} \approx T_n$, where T_n is the test statistic obtained by assuming that the complete case sample is a random sample; $T_n = U_n^2/\mathrm{Var}(U_n)$, where $U_n = \frac{1}{\widehat{\sigma}_n^2}\mathbf{g}_{\mathcal{C}}^{\mathrm{T}}(\mathrm{I} - \mathrm{H}_{\mathcal{C}})\mathbf{Y}_{\mathcal{C}}$, $\mathrm{Var}(U_n) = \frac{1}{\widehat{\sigma}_n^2}\mathbf{g}_{\mathcal{C}}^{\mathrm{T}}(\mathrm{I} - \mathrm{H}_{\mathcal{C}})\mathbf{g}_{\mathcal{C}}$, and $\widehat{\sigma}_n^2 = \frac{1}{n}\mathbf{Y}^{\mathrm{T}}(\mathrm{I} - \mathrm{H}_{\mathcal{C}})\mathbf{Y}_{\mathcal{C}}$.*

Observation 3.1 follows from the fact that $\widetilde{\sigma}^2 = \frac{\widehat{\sigma}_n^2}{\widetilde{a}}$, see Supplementary Material Section 4.4.2, and can be seen as a generalization of the result by Tang.[17] If the null hypothesis is true, then $Y_i - \mathrm{X}_i\boldsymbol{\beta} \sim \mathrm{N}(0, \sigma^2) \; \forall i$, and

$$\mathrm{Pr}((Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} < l_z) \cup (Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} > u_z)) = \mathrm{Pr}((Y_j - \mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta} < l_z) \cup (Y_j - \mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta} > u_z))$$

for all i, j. Replacing $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}}$, each individual in the sample will under H_0 have approximately equal probability to be sampled for genotyping.

Residual phenotype test If, under z-extreme sampling, we derive a test directly from the residual phenotype model (3), then we obtain

$$U_{z\mathcal{C}} = \frac{1}{\widetilde{\sigma}_z^2}\mathbf{g}_{\mathcal{C}}^{\mathrm{T}}(\mathrm{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\mathbf{Z}_{\mathcal{C}}, \quad \mathrm{Var}(U_{z\mathcal{C}}) = \frac{\widetilde{a}_z}{\widetilde{\sigma}_z^2}\mathbf{g}_{\mathcal{C}}^{\mathrm{T}}(\mathrm{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\mathbf{g}_{\mathcal{C}}, \tag{10}$$

and $T_{z\mathcal{C}} = U_{z\mathcal{C}}^2/\mathrm{Var}(U_{z\mathcal{C}})$. Since the model assumption $\boldsymbol{\varepsilon}_z \sim \mathrm{MVN}(\mathbf{0}, \sigma^2\,\mathrm{I})$ is generally incorrect (see Definition 2.2), the asymptotic null distribution of $T_{z\mathcal{C}}$ is not necessarily χ_1^2. Observation 3.2 is proven in Supplementary Material Section 4.4.3.

Observation 3.2 *Under z-extreme sampling, $T_{z\mathcal{C}} \approx T_{\mathcal{C}}$ if $\mathrm{X}_{\mathcal{C}}$ is column-wise uncorrelated with $\mathbf{g}_{\mathcal{C}}$. Otherwise, $T_{z\mathcal{C}} \leq T_{\mathcal{C}}$. If t_α is a critical value from the χ_1^2-distribution, then rejecting H_0 when $T_{z\mathcal{C}} > t_\alpha$ is a valid but conservative test.*

3.2 All case analysis

Assuming y_i and $\mathbf{x}_i$ are observed for all N individuals, the extreme sampling design (including y-extreme and z-extreme sampling) satisfies the MAR condition. For all case analysis (Definition 2.4) we use a likelihood model for a sample with MAR covariates[21] based on the following assumptions:

Assumption 3.1 *G is a discrete random variable (genotype) with sample space* $\{g_k; k = 1, \ldots, K\}$.

Assumption 3.2 *The distribution of G (genotype frequencies) can differ between different groups of individuals. Assume that there are J groups and let* $\Pr(G = g_k|j) = p_{jk}$, $k = 1, \ldots, K$, *be the point probabilities of G in group* j, $j = 1, \ldots, J$. *Groups are defined by unique covariate patterns of* X^* *(dummy coded categorical covariates), where* X^* *denotes* $d^* + 1 \leq d + 1$ *columns of* X *(including intercept).*

Let G_i denote the random genotype for individual i. For individuals $i \in \mathcal{C}$, g_i denotes the actual observation of G_i. Assumption 3.2 implies that genetic and non-genetic covariates $\mathbf{g}$ and X can be correlated. Generally, X can contain continuous covariates, but X^* is discrete (possibly including discretized and dummy coded continuous covariates). Let $\mathcal{C}_j$ denote the set of all n_j genotyped individuals $i \in \mathcal{C}$ who are also in group j, let $\mathcal{M}_j$ denote the set of all m_j missing-genotype individuals $i \in \mathcal{M}$ who are also in group j, and let $N_j = n_j + m_j$, $\sum_j n_j = n$, $\sum_j m_j = m$, $\sum_j N_j = N$ and $n + m = N$. Let n_{jk} be the random variable that counts observations of genotype g_k among genotyped individuals in group j ($i \in \mathcal{C}_j$). The all case likelihood is

$$L = \prod_{i \in \mathcal{C}} f_{Y_i|G_i=\mathrm{g}_i}(y_i; \mathbf{x}_i) \Pr(G_i = \mathrm{g}_i; \mathbf{x}_i) \prod_{i \in \mathcal{M}} \sum_k f_{Y_i|G_i=g_k}(y_i; \mathbf{x}_i) \Pr(G_i = g_k; \mathbf{x}_i)$$

$$\propto \prod_{i \in \mathcal{C}} \frac{1}{\sigma} \phi \left(\frac{Y_i - \mu_i}{\sigma} \right) \prod_j \prod_k p_{jk}^{n_{jk}} \prod_j \prod_{i \in \mathcal{M}_j} \sum_k \frac{1}{\sigma} \phi \left(\frac{Y_i - \mu_i(g_k)}{\sigma} \right) p_{jk},$$

where $\mu_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \mathrm{g}_i \beta_{\mathrm{g}}$ for $i \in \mathcal{C}$ and $\mu_i(g_k) = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + g_k \beta_{\mathrm{g}}$ for $i \in \mathcal{M}$. The maximum likelihood estimators of $\boldsymbol{\beta}$, σ and p_{jk} under the null are

$$\widehat{\boldsymbol{\beta}} = \left(\mathrm{X}^{\mathrm{T}} \mathrm{X} \right)^{-1} \mathrm{X}^{\mathrm{T}} \mathbf{Y}, \quad \widehat{\sigma}^2 = \frac{1}{N} \mathbf{Y}(\mathrm{I} - \mathrm{H}) \mathbf{Y}, \quad \widehat{p}_{jk} = n_{jk}^{\mathrm{obs}} / n_j, \tag{11}$$

and the estimated mean under the null is $\widehat{\boldsymbol{\mu}} = \mathrm{X} \widehat{\boldsymbol{\beta}} = \mathrm{H} \mathbf{Y}$. These closed-form expressions make model fitting under the null computationally efficient for all case analysis. The score test for the all case likelihood was derived by Derkach *et al*[11] for models where $\mathrm{X}^* = \mathrm{X}$ (Assumption 3.2). Our test is derived using a slightly different approach (see Supplementary Material Section 5), but the result is equivalent when $\mathrm{X}^* = \mathrm{X}$ and probabilities p_{jk} are unknown. Let $\mathbf{g}_M$ be an N-vector where for

individuals $i \in \mathcal{C}$, $\mathbf{g}_{M,i} = \mathrm{g}_i$, while for $i \in \mathcal{M}_j$ the mean genotype across individuals $i \in \mathcal{C}_j$ is imputed. The all case score test statistic is $T_A = U_A^2 / \mathrm{Var}(U_A)$, where

$$U_A = \frac{1}{\widehat{\sigma}^2} \mathbf{g}_M^{\mathrm{T}} (\mathrm{I} - \mathrm{H}) \mathbf{Y}, \tag{12}$$

and

$$\mathrm{Var}(\mathbf{U}_A) = \frac{1}{\sigma^2} \mathbf{g}_M^{\mathrm{T}} (\mathrm{I} - \mathrm{H}) \mathbf{g}_M - \frac{1}{\sigma^2} \sum_j n_j \, \mathrm{Var}(G|j)$$

$$+ \frac{1}{\sigma^4} \sum_j \mathrm{Var}(G|j) \left(\mathrm{E} \sum_{i \in \mathcal{C}_j} (Y_i - \mu_i(\beta_{\mathrm{g}} = 0))^2 - \frac{1}{n_j} \left(\mathrm{E} \sum_{i \in \mathcal{C}_j} (Y_i - \mu_i(\beta_{\mathrm{g}} = 0)) \right)^2 \right) \tag{13}$$

$$= \frac{1}{\sigma^2} \mathbf{g}_M^{\mathrm{T}} (\mathrm{I} - \mathrm{H}) \mathbf{g}_M - \frac{1}{\sigma^2} \sum_j \mathrm{Var}(G|j) \left(\sum_{i \in \mathcal{C}_j} h_{i1}(\beta_{\mathrm{g}} = 0) + \frac{1}{n_j} \left(\sum_{i \in \mathcal{C}_j} h_{i0}(\beta_{\mathrm{g}} = 0) \right)^2 \right), \tag{14}$$

where h_{i0} and h_{i1} are defined in Equation (6). The score test statistic has an asymptotic χ_1^2 distribution under the null.

3.2.1 y-extreme sampling

When there are no non-genetic covariates X, and therefore no groups (Assumption 3.2), then $U_A \approx \frac{1}{\widehat{\sigma}^2} \mathbf{g}_{\mathcal{C}}^{\mathrm{T}} (\mathrm{I} - \mathrm{H}_{\mathcal{C}}) \mathbf{Y}_{\mathcal{C}}$ and $\mathrm{Var}(U_A) \approx \frac{\widehat{a}}{\widehat{\sigma}^2} \mathbf{g}_{\mathcal{C}}^{\mathrm{T}} (\mathrm{I} - \mathrm{H}_{\mathcal{C}}) \mathbf{g}_{\mathcal{C}}$, where $\widehat{a} = a(\widehat{\sigma})$, see Equation (6). Then, the complete case and all case score test statistics are approximately equal. Details are given in Supplementary material Section 5.6.

3.2.2 z-extreme sampling

Under z-extreme sampling (Definition 2.7) the all case score statistic (12) can be approximated by

$$U_A \approx \frac{1}{\widehat{\sigma}^2} \mathbf{g}_{\mathcal{C}}^{\mathrm{T}} (\mathrm{I} - \mathrm{H}_{\mathcal{C}}) \mathbf{Y}_{\mathcal{C}}.$$

The variance (13) can be approximated by

$$\mathrm{Var}(U_A) \approx \frac{\widehat{a}}{\widehat{\sigma}^2} \mathbf{g}_{\mathcal{C}}^{\mathrm{T}} (\mathrm{I} - \mathrm{H}_{\mathcal{C}}) \mathbf{g}_{\mathcal{C}},$$

where a is given in Equation (6), and $\widehat{a}$ denotes that the function is evaluated at the all case maximum likelihood estimator for σ. Details are given in Supplementary Material Section 5.7.

Observation 3.3 *Under z-extreme sampling, $U_A \approx \frac{\widetilde{\sigma}^2}{\widehat{\sigma}^2} U_{\mathcal{C}} \approx U_{\mathcal{C}}$, where $U_{\mathcal{C}}$ is the compete case score statistic (7), U_A is the all case score statistic (12), $\widetilde{\sigma}$ is the complete case maximum likelihood estimator of σ and $\widehat{\sigma}$ is the all case estimator. Also, $\mathrm{Var}(U_A) \approx \frac{\widetilde{\sigma}^2}{\widehat{\sigma}^2} \frac{\widehat{a}}{\widetilde{a}} \mathrm{Var}(U_{\mathcal{C}}) \approx \mathrm{Var}(U_{\mathcal{C}})$. Under z-extreme sampling, the all case and complete case score tests are therefore approximately equal.*

Residual phenotype test Under z-extreme sampling it follows from Observation 3.3 that also the residual phenotype all case score test statistic T_{z_A} is approximately equal to the residual phenotype complete case score test $T_{z_{\mathcal{C}}}$ (10), see Supplementary Material Section 5.7.1.

4 Rare variant tests for extreme samples

In Supplementary Material Sections 4 and 5 we show that all results from the previous section generalize to testing $H_0 : \boldsymbol{\beta}_{\mathrm{g}} = \mathbf{0}$ in model (2) (i.e. simultaneously testing ν genetic variants, $1 < \nu \ll N$). Here we consider testing ν rare genetic variants and we derive two rare variant association tests for the extreme phenotype sampling design; the 'collapsing test'[22–25] and the 'variance component test'.[18, 19] Under extreme sampling, a binary response collapsing test was used by Peloso *et al.*[16] A complete case collapsing test was studied by Li *et al.*[12] A variance component test (the SKAT test) for complete case analysis of y-extreme samples was proposed by Barnett *et al.*[14] In the extreme sampling complete case setting, we find that estimation of the variance parameter σ^2 must be accounted for when deriving the score test (the parameters σ and $\boldsymbol{\beta}_{\mathrm{g}}$ are not orthogonal ($\mathbf{F}_{\sigma\boldsymbol{\beta}_{\mathrm{g}}} \neq 0$), see Supplementary Material Section 4.2). We therefore propose a different complete case variance component test. We also derive an all case variance component test for extreme sampling designs and compare y-extreme and z-extreme sampling.

4.1 The collapsing test

Let $\mathbf{w}$ be a ν-vector of chosen weights for the ν genetic variants in the region of interest. For the situation where all N individuals are genotyped, we construct a single 'collapsed' genetic covariate $\mathbf{g}_w$ by a weighted sum over all ν variants $\mathbf{g}_w = \mathrm{G}\,\mathbf{w}$, where G is the $N \times \nu$ genetic covariate matrix as defined in model (2). We then assume a linear model $\mathbf{Y} = \mathrm{X}\,\boldsymbol{\beta} + \mathbf{g}_w\beta_{\mathrm{g}} + \boldsymbol{\varepsilon}$, which is simply a modified version of model (1), and test $H_0 : \beta_{\mathrm{g}} = 0$. For extreme sampling designs, where only $\mathrm{G}_{\mathcal{C}}$ is observed, we define $\mathbf{g}_{w,\mathcal{C}} = \mathrm{G}_{\mathcal{C}}\,\mathbf{w}$ and treat this n-vector as the observed genetic covariate. The tests for $H_0 : \beta_{\mathrm{g}} = 0$ from the previous section (complete case and all case) apply.

4.2 The variance component test

We assume that $\boldsymbol{\beta}_g$ is a random vector with mean 0 and variance $\tau^2\Sigma$, where Σ is a symmetric weight matrix. Then we test $H_0 : \tau^2 = 0$ (which corresponds to $H_0 : \boldsymbol{\beta}_g = \mathbf{0}$), using the variance component score test statistic[18,19]

$$Q = \mathbf{U}_{\boldsymbol{\beta}_g}^{T}\Sigma\mathbf{U}_{\boldsymbol{\beta}_g}, \tag{15}$$

where $\mathbf{U}_{\boldsymbol{\beta}_g}$ is the score statistic from Definition 2.1. We let $\Sigma = WW$ where $W = \mathrm{Diag}(\mathbf{w})$, corresponding to SKAT with a linear weighted kernel.[18] For some score statistic $\mathbf{U}_{\boldsymbol{\beta}_g}$ (asymptotically multivariate normal) with mean $\mathbf{0}$ and covariance $\mathrm{Var}(\mathbf{U}_{\boldsymbol{\beta}_g})$, the asymptotic null distribution of Q can be estimated by the distribution of $\sum_{m=1}^{\nu}\lambda_m\chi_{1,m}^2$, where λ_m are the eigenvalues of $\Sigma\,\mathrm{Var}(\mathbf{U}_{\boldsymbol{\beta}_g})$ and $\chi_{1,m}^2$ are i.i.d χ_1^2 random variables.[26] In Supplementary Material Sections 4 and 5, complete and all case score statistics ($\mathbf{U}_{\mathcal{C}}$, $\mathbf{U}_A$) and variances ($\mathrm{Var}(\mathbf{U}_{\mathcal{C}})$, $\mathrm{Var}(\mathbf{U}_A)$) are derived for the different extreme sampling designs in the multivariate setting ($\nu > 1$ genetic variants). In Supplementary material Section 6, we show that the variance component test follows directly from these expressions.

4.2.1 z-extreme sampling

We make three observations regarding z-extreme sampling that follow from Observations 3.1, 3.2 and 3.3 for single variant tests. Proofs are given in Supplementary Material Section 6.

Observation 4.1 *Under z-extreme sampling, complete case analysis, the test that rejects $H_0 : \tau^2 = 0$ if $Q_n > t_\alpha$, where Q_n is the test statistic obtained by assuming that the complete case sample is a random sample and t_α is a critical value from the asymptotic null distribution of Q_n, is approximately equal to the test that rejects $H_0 : \tau^2 = 0$ if $Q_{\mathcal{C}} > t_{\alpha,\mathcal{C}}$, where $Q_{\mathcal{C}}$ is the complete case test statistic and $t_{\alpha,\mathcal{C}}$ is a critical value from the asymptotic distribution of $Q_{\mathcal{C}}$.*

Observation 4.2 *A variance component score test for complete case analysis under z-extreme sampling derived directly from the residual phenotype model (3) is valid, but conservative, for testing $H_0 : \tau^2 = 0$.*

Observation 4.3 *Under z-extreme sampling, the all case and complete case variance component score test statistics are approximately equal.*

Observations 4.1 and 4.3 also apply to y-extreme sampling for the particular case when there are no non-genetic covariates in the model ($X = \mathbf{1}$).

5 Simulations

5.1 Single variant tests

We simulated two non-genetic covariates; x_1 from a $N(2,1)$ distribution, and x_2 from a Bernoulli(0.5)-distribution. The covariate x_2 may for example represent population substructure. In each of 10000 simulations we drew a genetic covariate g_i such that $g_i \sim \mathrm{Binom}(2, p_0)$ if $x_{2,i} = 0$ and $g_i \sim \mathrm{Binom}(2, p_1)$ if $x_{2,i} = 1$, $p_0 = 0.4$ and $p_1 = 0.1$ or $p_0 = p_1 = 0.25$. Phenotypes were independently generated from the model

$$Y \sim \mathrm{N}\left(5 + \beta_1 x_1 + \beta_2 x_2 + \beta_g g, 4^2\right).$$

Simulations were performed for three sample sizes $N = 2000$, 10000 and 20000. The parameter β_g was 0 for estimates of type-1 error and otherwise chosen so that the power to detect an association in the full sample was between 80% and 90%, $\beta_g = 0.45$, 0.21, and 0.15, for $N = 2000$, 10000, and 20000, respectively. Coefficients of non-genetic covariates were $\beta_1 = 10$ and $\beta_2 = 5$, or $\beta_1 = 5$ and $\beta_2 = 2$, or $\beta_1 = 2$, and $\beta_2 = 1$. With these parameter choices the environmental covariates x_1 and x_2 were more important than the genetic covariate g for describing the response (R^2 from fitting the regression model with and without the genetic covariates varied minimally). The latter choice was motivated by the assumption that environmental variables are more important for predicting a complex trait, compared to the genotype of a common genetic variant.[27] For extreme sampling, we used estimated population quantiles $l = Y_{\left(\frac{n/2}{N}\right)}$, $u = Y_{\left(\frac{N-n/2}{N}\right)}$ or $l_z = Z_{\left(\frac{n/2}{N}\right)}$, $u_z = Z_{\left(\frac{N-n/2}{N}\right)}$ to ensure that n individuals were selected for genotyping in a symmetric fashion. Ideally, l and u should be determined before seeing the data, but for large N, l and u are unbiased estimates of the population quantiles and estimated with low variance. Therefore, these estimates may be regarded as deterministic and we use them in place of the true population quantiles. Under z-extreme sampling, we use thresholds $l_z + \mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}$ and $u_z + \mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}$ where also $\widehat{\boldsymbol{\beta}}$ is estimated based on the full population.

Estimated power when $n = N/2$, $\beta_1 = 10$ and $\beta_2 = 5$ is presented in Table 1. In addition to y-extreme and z-extreme sampling we estimated the power of randomly sampling n individuals. In Figure 1 we illustrate the estimated power of all methods for increasing n ($N = 10000$ fixed), for $p_0 = 0.4$, $p_1 = 0.1$, and different values of β_1 and β_2. Increasing n is equivalent to increasing l and decreasing u. Figures were generated using the R package ggplot2.[28]

y-extreme sampling For y-extreme sampling, the all case test is clearly more powerful than the complete case test. For the simulation with $\beta_1 = 10$ and $\beta_2 = 5$

(Figure 1, panel A) the complete case method is in fact less powerful than the random sample. When the non-genetic covariates have a strong effect, then the extreme-y individuals will be extreme due to the non-genetic covariates rather than the genetic covariate. As illustrated here, this can result in a particularly unfavorable sample with respect to detecting causal genetic variants. For lower non-genetic coefficient values, the power of y-extreme sampling designs improve (Figure 1, panels B and C). The extreme-y individuals are then more likely to be extreme due to the genetic effect.

z-extreme sampling The z-extreme sampling design is clearly more powerful than y-extreme sampling and the power of this design is not impacted by the effect of the non-genetic covariates (Figure 1). Three theoretical results of Section 3 are illustrated by simulations; (i) complete case and all case score tests are approximately equally powerful (Observation 3.3), (ii) the complete case score test statistic is approximately equal to the score test statistic we obtain by assuming that the complete case sample is a random sample (Observation 3.1) and (iii) residual phenotype tests are less powerful than tests based on the original regression model when the genetic covariate is correlated with other covariates, but equal to tests based on the original regression model when the genetic covariate is not correlated with other covariates (Observation 3.2). In our simulation set-up, g is uncorrelated with non-genetic covariates when $p_0 = p_1 = 0.25$, and correlated with non-genetic covariates when $p_0 = 0.4$ and $p_1 = 0.1$.

Result (i) can be seen from Table 1, where rows corresponding to z-extreme complete case and z-extreme all case analysis are almost identical. In Figure 1, the line corresponding to the z-extreme all case test is not visible as it lies directly underneath the line for the z-extreme complete case test.

Result (ii) is illustrated in Table 2. By Observation 3.3 we can either use the complete case score test, the all case score test (by result (i)), or we can use a score test for random samples (we used the function glm.scoretest in the R Statmod package[29]), to test $H_0 : \beta_g = 0$ under z-extreme sampling.

Result (iii) is illustrated in Table 1. For the column $p_0 = 0.4$, $p_1 = 0.1$, the rows 'z-extreme complete case (residual)' and 'z-extreme all case (residual)' show lower power than the rows 'z-extreme complete case' and 'z-extreme all case'. For the column $p_0 = p_1 = 0.25$, all these tests are equal. The simulations in Figure 1 are performed with $p_0 > p_1$ and the methods based on the model for the residual phenotype never reach the power of the full model, even when all N individuals are genotyped.

Binary response methods The simplest and most commonly approach to association testing under extreme sampling is to treat the extreme phenotypes as a binary responses.[7] We consider only the complete cases (Definition 2.3), define a binary response vector based on low and high extreme phenotypes (either y or z) and apply a logistic regression model. For the simulations in Table 1, the y-extreme binary response model has almost no power to detect associations while the z-extreme binary response model is quite powerful. As the effect of non-genetic covariates decreases, the y-extreme binary response method improves (Figure 1). We note the behavior of the binary response methods in Figure 1. As n increases the minor allele frequencies in each extreme tail become more similar, and it becomes harder to detect any differences between the two extremes. After some point, the sample size n is not large enough to detect this small difference with sufficient power. At $n = N/2$ (which corresponds to Table 1, $N = 10000$), the z-extreme binary response model has good power for this particular simulation set-up, but power decreases as n increases.

Type 1 error rates Estimated type-1 error rates for $n = N/2$ (symmetric sampling) are presented in Table 3. The residual phenotype tests are by design valid but conservative when $p_0 \neq p_1$. The y-extreme binary response test did not control the type-1 error rate when $N = 2000$ ($n = 1000$).

5.2   Rare variant tests

We performed a rare variant simulation study similar to Wu *et al.*[18] For 10000 individuals we generated a 1 Mb region of genetic information using the simulation program COSI (European population, standard out-of-Africa model).[30] For each of 10000 simulations, we randomly selected $N = 5000$ individuals and a 30 kb region. Then, we randomly chose 50% or 10% of variants with (observed) minor allele frequency below 3% to be causal. The effect of a causal variant with minor allele frequency m was set to $\beta_{\mathrm{g}} = \pm c \cdot |\log_{10}(m)|$, otherwise $\beta_{\mathrm{g}} = 0$. A certain percentage of the causal variants were set to have a negative effect sizes. The constant c was 0.2 when 50% of variants were causal and 1.2 when 10% of variants were causal (chosen so that the power of the full sample was approximately 80%). Phenotypes were simulated from

$$Y \sim \mathrm{N}\left(5 + 1x_1 + 2x_2 + \mathrm{G}\,\boldsymbol{\beta}_{\mathrm{g}}, 4^2\right),$$

where G is a genotype matrix of all variants in the selected region. We estimated power of the collapsing method and the variance component method for increasing n (N fixed). As proposed by Wu *et al*,[18] weights were $f_\beta(m, 1, 25)$ where $f_\beta()$ is the

15

density of the Beta-distribution and m is the minor allele frequency of the genetic variant. Power estimates are presented in Figure 2. For the collapsing methods, the function glm.scoretest (R Statmod package[29]) was used for full sample, random sample and binary response tests, while our implementation of complete case and all case tests for single variants were used for extreme samples. For the variance component methods, the the SKAT R package[31] was used for full sample, random sample and binary response tests, while our implementation of complete case and all case variance component tests were used for extreme samples. We used the Davies method[32] as implemented in the R package CompQuadForm[33] to obtain p-values for the variance component score test. The power estimates presented in Figure 2 are quite similar to that of the single-variant simulation (Figure 1). The z-extreme sample is more powerful than the y-extreme sample. Under z-extreme sampling the complete case and all case tests are equally powerful, and under y-extreme sampling the all case test is more powerful than the complete case test. Concerning choice of method for rare variant association testing, we confirm the simulation results of Wu *et al*,[18] namely that when many variants are causal and all effects in the same direction (Figure 2, panel A) the collapsing test is the more powerful, while when few variants are causal and effects are in opposite directions (Figure 2, panel B) the variant component test is the more powerful test. For the evaluation of type-1 error rates (significance level 2.5×10^{-6}, motivated by Lee *et al*[19]) we performed 6.5 million simulations under the null. All methods (collapsing method, complete and all case, variance component method, complete and all case) rejected the null 8 times, which gives an estimated type-1 error rate of $1.2 \times 10^{-6} (1.1 \times 10^{-7}, 4.7 \times 10^{-6})$ (Clopper-Pearson 95% confidence interval).

6 Application to data from the HUNT study

We assess the performance of different designs and methods using data from the HUNT Fitness study.[20,34] The study of genetic association with maximum oxygen uptake (VO2max) is an ongoing project that is run by co-authors Anja Bye, Einar Ryeng and Ulrik Wisløff. Here we use genome-wide observations of about 10^5 common variants (minor allele frequency ≥ 0.05) that are available for $N \approx 3000$ individuals. The study was designed so as not to include closely related participants. We assume that the sample is representative of the population of Nord-Trøndelag, Norway. The dataset consists of approximately 1500 men and 1500 women. There is a considerable difference in variance of VO2max between men and women, and we therefore base our analysis on a standardized VO2max variable where for each gender we subtract the mean and divide by the standard deviation. The null model

is then a linear regression model with (standardized) VO2max as the response, and age and physical activity as non-genetic covariates.

Since all N individuals have been genotyped, we can compare the results of extreme sampling designs to the results of analyzing the full sample. From full sample analysis it is known that a region on chromosome 1 is associated with VO2max. Plots of p-values (on $-\log 10$-scale) from this region on chromosome 1 are shown in Figure 3. The full sample size is quite low and as could be expected, the smallest p-values do not reach genome-wide significance ($5 \cdot 10^{-7}$ for 10^5 tests, family-wise error rate controlled at significance level 0.05, using Bonferroni), yet it is assumed that the peak in $-\log 10(p)$ represents a genetic association with VO2max. Here we attempt to replicate this finding using extreme phenotype sampling designs. We perform y-extreme and z-extreme sampling, $n = \frac{N}{2}$, so that approximately one quarter of individuals satisfy $y_i < l_i$ and one quarter satisfy $y_i > u_i$. The classification rules l_i and u_i were set using percentiles of the empirical phenotype or residual phenotype distributions. With both types of extreme sampling designs (y-extreme and z-extreme) we are to some extent able to replicate the genotype-phenotype association. The y-extreme sampling design has low power, both when using the all case and complete case score tests. The z-extreme sampling design is as expected more powerful, and the complete case and all case methods are equal (Observation 3.3). In fact, the z-extreme sampling results closely mimic the full sample results despite using only half of the available genetic information. Residual phenotype tests under z-extreme sampling can also be applied, and results are given in the Supplementary Material Section 7. These results are almost equal to those shown in Figure 3, i.e. from tests based on the original regression model (Observation 3.2).

We summarize the genome-wide test results in QQ plots. QQ plots of the p-values (on $-\log 10$-scale) for association tests on chromosomes 1 to 22 based on different extreme sampling methods are presented in Figure 4. In the construction of the QQ-plot we assume that the p-values from the score tests are exact, that the null model is true for all variants, and that all tests are independent. If so, each p-value is uniform$(0, 1)$-distributed and the ith smallest p-value $p_{(i)}$ follows a Beta$(i, m + 1 - i)$-distribution where m is the number of tests. Then $\mathrm{E}(p_{(i)}) = i/(m + i)$. The 2.5% and 97.5% quantiles of the distribution of the ordered p-values are also plotted. Deviations from the straight line can for example indicate associations with VO2max. The deviation seen from the full sample analysis is solely due to the region in chromosome 1 where the null model is rejected. As in Figure 3, we see that the z-extreme sampling design yields the results most similar to the full sample analysis.

Tests were performed separately for each chromosome. For the largest (chromosome 1), running times were 1.37 seconds for complete case analysis and 19.14

17

seconds for all case analysis. For reference, the full sample analysis (based on the function glm.scoretest in the Statmod package[29]) took 3.34 seconds. All computations were performed using R version 3.4.0[35] on a personal computer (MacBook Air (13", Early 2014) with 1.7 GHz Intel Core i7-4650U and 4 MB cache).

7 Discussion

Improving power with extreme phenotype sampling The aim with using an extreme phenotype sampling design is to improve power to detect associations between genetic variants and the phenotype. It is assumed that extreme phenotype individuals are more informative with respect to genetic influence than the average person. Our simulations show that when other non-genetic or environmental covariates have a strong impact on the phenotype, then the classical extreme phenotype sampling design (y-extreme sampling) is not necessarily more powerful than a random sample. This is because the extremes are likely have a high or low phenotype value due to the non-genetic covariates, and not due to a genetic effect. To resolve this issue, we have defined an extreme sampling design where each person has an individual classification rule for being extreme or not. We use the residuals of the null model ($\mathbf{Y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$), where non-genetic effects have been accounted for, to define extreme phenotypes (z-extreme sampling). Using this design, the extreme individuals are more likely to be extreme due to a genetic impact, and this can improve power to detect phenotype-genotype associations.

Hypothesis testing for extreme phenotype samples We have focused on complete case and all case analysis of selectively genotyped samples. In the former, only the extreme individuals are analyzed. A binary response method can be applied to the complete cases[7] but dichotomizing a continuous response typically results in loss of power.[9] We have focused on a complete case method that directly models the distribution of the extremes,[9] extending current methods to allow for individual classification rules. With all case analysis we include information on non-genetic variables for non-extremes. It has previously been shown that when there are no non-genetic covariates in the model, complete case and all case tests under y-extreme sampling are equally powerful.[9,11] When adding information on non-genetic covariates, the all case method is generally more powerful than complete case analysis for the classical y-extreme sampling design. For all case analysis we have restricted attention to a situation where the covariate to be tested is discrete, and potential confounders are discrete and completely observed. A more general all case model has been developed by Tao et al,[10] which for example can handle the issue of using

principal components to control for confounding due to population structures.

For the z-extreme sampling design, we showed that the all case score test reduces to the complete case test. This is noteworthy because the complete case test requires far fewer model assumptions: one does not need to model the distribution of missing covariates and any potential dependencies between covariates, and we are not restricted to discrete genetic covariates and discrete confounders. This comes in addition to the fact that the z-extreme sampling design in itself is more powerful. In complete case analysis it is trivial to include principal components (missing for non-extremes) and gene-environment or gene-gene interaction terms. The complete case score test for z-extreme samples can furthermore be shown to be equal to the score test for a random sample (under the null, any individual is equally likely to be classified as extreme).

For z-extreme samples, we considered the possibility of using the residual phenotype as the response of in a regression model that under the alternative would only include genetic covariates (residual phenotype tests). We proved that the score tests derived for a model with independent observations (assuming test statistic χ^2-distributed under the null) were also valid for the residual phenotype model.

Multiple variants and rare variant association tests   In the Supplementary Material we have derived score tests for simultaneously testing ν genetic variants. A special case is the single variant tests presented in Section 3. When simultaneously testing ν rare variants (Section 4) it is common to apply collapsing tests[22–25] or variance component score tests such as the SKAT test.[18] The complete and all case collapsing methods considered here follow directly from single variant tests, while the complete and all case variance component tests were derived using results from the multivariate score tests. It then follows that also in the rare variant setting, complete case and all case tests are equal under z-extreme sampling, and that residual phenotype tests are valid, but at times conservative. Simulations showed that the z-extreme sampling design is, as for single variant testing, the more powerful design. The relative performance of the collapsing and variance component tests for extreme samples were as previosuly found for random sampling.[18]

Practice guideline   The following guideline assumes that phenotype values y_i and non-genetic covariates $\mathbf{x}_i$ are observed for N individuals, representative of the population on which to make inference, and that $n < N$ of these are to be genotyped. Step 1: Decide the null model $\mathbf{Y} = \mathrm{X}\boldsymbol{\beta} + \varepsilon$ and calculate the residuals using standard statistical software. Step 2: From the empirical distribution of residuals, estimate thresholds l_z and u_z (for example so that approximately $n/2$ of individuals satisfy $z_i < l_z$ and $n/2$ satisfy $z_i > u_z$) and use these to classify individuals

as extreme or not. Step 3: After genotyping extreme individuals, perform complete case hypothesis testing (using standard statistical software if no additional covariates (e.g. interactions effects or principal components) have been added to the model, otherwise using complete case tests as provided by us). Estimate parameters by maximizing the complete case likelihood. All extreme sample tests, as well as functions for parameter estimation, are available as an R package.

8 Supplementary material

In the Supplementary Material we present all details in the derivation of score tests for the general setting of simultaneously testing ν genetic variants, as well as analytical results regarding equality of tests under the z-extreme sampling design. The single variant tests presented in Sections 2 and 3 of this paper represent the special case when $\nu = 1$. The rare variant tests of Section 4 can also be seen as a particular implementation of these multivariate score tests. All methods are implemented in R and available as an R package at https://github.com/theabjorn/extremesampling. The vignette includes examples on how to make use of all the tests presented in this paper.

Acknowledgments

Conflict of interest

None declared.

References

1. Lebowitz R, Soller M, Beckmann J. Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. *Theoretical and Applied Genetics* 1987; **73**(4):556–562.

2. Lander ES, Botstein D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 1989; **121**(1):185–199.

3. Darvasi A, Soller M. Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theoretical and Applied Genetics* 1992; **85**(2-3):353–359.

4. Slatkin M. Disequilibrium mapping of a quantitative-trait locus in an expanding population. *The American Journal of Human Genetics* 1999; **64**(6):1765–1773.

5. Chen Z, Zheng G, Ghosh K, Li Z. Linkage disequilibrium mapping of quantitative-trait loci by selective genotyping. *The American Journal of Human Genetics* 2005; **77**(4):661–669.

6. Wallace C, Chapman JM, Clayton DG. Improved power offered by a score test for linkage disequilibrium mapping of quantitative-trait loci by selective genotyping. *The American Journal of Human Genetics* 2006; **78**(3):498–504.

7. Van Gestel S, Houwing-Duistermaat JJ, Adolfsson R, van Duijn CM, Van Broeckhoven C. Power of selective genotyping in genetic association analyses of quantitative traits. *Behavior Genetics* 2000; **30**(2):141–146.

8. Xing C, Xing G. Power of selective genotyping in genome-wide association studies of quantitative traits. *BMC Proceedings*, vol. 3, BioMed Central, 2009; 1.

9. Huang B, Lin D. Efficient association mapping of quantitative trait loci with selective genotyping. *The American Journal of Human Genetics* 2007; **80**(3):567–576.

10. Tao R, Zeng D, Lin DY. Efficient semiparametric inference under two-phase sampling, with applications to genetic association studies. *Journal of the American Statistical Association* 2017; **112**(520):1468–1476.

11. Derkach A, Lawless JF, Sun L. Score tests for association under response-dependent sampling designs for expensive covariates. *Biometrika* 2015; **102**(4):988–994.

12. Li D, Lewinger JP, Gauderman WJ, Murcray CE, Conti D. Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. *Genetic Epidemiology* 2011; **35**(8):790–799.

13. Guey LT, Kravic J, Melander O, Burtt NP, Laramie JM, Lyssenko V, Jonsson A, Lindholm E, Tuomi T, Isomaa B, *et al.*. Power in the phenotypic extremes: a

simulation study of power in discovery and replication of rare variants. *Genetic Epidemiology* 2011; **35**(4):236–246.

14. Barnett IJ, Lee S, Lin X. Detecting rare variant effects using extreme pheno-type sampling in sequencing association studies. *Genetic Epidemiology* 2013; **37**(2):142–151.

15. Tao R, Zeng D, Franceschini N, North KE, Boerwinkle E, Lin DY. Analysis of sequence data under multivariate trait-dependent sampling. *Journal of the American Statistical Association* 2015; **110**(510):560–572.

16. Peloso GM, Rader DJ, Gabriel S, Kathiresan S, Daly MJ, Neale BM. Phenotypic extremes in rare variant study designs. *European Journal of Human Genetics* 2016; **24**(6):924–930.

17. Tang Y. Equivalence of three score tests for association mapping of quantitative trait loci under selective genotyping. *Genetic Epidemiology* 2010; **34**(5):522–527.

18. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* 2011; **89**(1):82–93.

19. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 2012; **13**(4):762–775.

20. Krokstad S, Langhammer A, Hveem K, Holmen T, Midthjell K, Stene T, Bratberg G, Heggland J, Holmen J. Cohort profile: the HUNT study, Norway. *International Journal of Epidemiology* 2013; **42**(4):968–977.

21. Little RJ, Rubin DB. *Statistical Analysis with Missing Data.* John Wiley & Sons, 2002.

22. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 2007; **615**(1):28–56.

23. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics* 2008; **83**(3):311–321.

24. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics* 2009; **5**(2):e1000 384.

25. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology* 2010; **34**(2):188–193.

26. Imhof J. Computing the distribution of quadratic forms in normal variables. *Biometrika* 1961; **48**(3/4):419–426.

27. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, *et al.*. Finding the missing heritability of complex diseases. *Nature* 2009; **461**(7265):747–753.

28. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. URL `http://ggplot2.org`.

29. Dunn PK, Smyth GK. Randomized quantile residuals. *J. Comput. Graph. Statist* 1996; **5**:236–244.

30. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. *Genome Research* 2005; **15**(11):1576–1583.

31. Lee S, Miropolsky L, Wu M. *SKAT: SNP-Set (Sequence) Kernel Association Test* 2017. URL `https://CRAN.R-project.org/package=SKAT`, R package version 1.3.0.

32. Davies RB. The distribution of a linear combination of χ^2 random variables. *Applied Statistics* 1980; **29**(3):323–333.

33. Duchesne P, de Micheaux PL. Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods. *Computational Statistics and Data Analysis* 2010; **54**:858–862.

34. Aspenes ST, Nauman J, Nilsen T, Vatten LJ, Wisløff U. Physical activity as a long-term predictor of peak oxygen uptake: the HUNT study. *Medicine and Science in Sports and Exercise* 2011; **43**(9):1675–1679.

35. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2016. URL `https://www.R-project.org`.

p_0, p_1	0.4, 0.1	0.25, 0.25
$N = 20000$, $\beta_{\mathrm{g}} = 0.15$		
Full sample	0.856 (0.849,0.863)	0.934 (0.929,0.939)
Random sample	0.575 (0.565,0.585)	0.690 (0.681,0.699)
y-extreme complete case	0.462 (0.452,0.471)	0.560 (0.550,0.569)
y-extreme all case	0.615 (0.606,0.625)	0.732 (0.724,0.741)
z-extreme complete case	0.830 (0.823,0.838)	0.914 (0.909,0.920)
z-extreme complete case (residual)	0.766 (0.757,0.774)	0.915 (0.909,0.920)
z-extreme all case	0.830 (0.823,0.838)	0.914 (0.909,0.920)
z-extreme all case (residual)	0.766 (0.758,0.774)	0.914 (0.909,0.920)
y-extreme binary	0.118 (0.111,0.124)	0.131 (0.125,0.138)
z-extreme binary	0.695 (0.686,0.704)	0.873 (0.866,0.879)
$N = 10000$, $\beta_{\mathrm{g}} = 0.21$		
Full sample	0.856 (0.849,0.863)	0.894 (0.888,0.900)
Random sample	0.567 (0.557,0.576)	0.626 (0.617,0.636)
y-extreme complete case	0.451 (0.441,0.461)	0.500 (0.491,0.510)
y-extreme all case	0.610 (0.601,0.620)	0.662 (0.653,0.672)
z-extreme complete case	0.833 (0.826,0.840)	0.868 (0.861,0.874)
z-extreme complete case (residual)	0.760 (0.752,0.769)	0.868 (0.861,0.874)
z-extreme all case	0.833 (0.826,0.841)	0.868 (0.861,0.875)
z-extreme all case (residual)	0.760 (0.751,0.768)	0.868 (0.861,0.874)
y-extreme binary	0.111 (0.105,0.117)	0.114 (0.108,0.120)
z-extreme binary	0.690 (0.680,0.699)	0.821 (0.814,0.829)
$N = 2000$, $\beta_{\mathrm{g}} = 0.45$		
Full sample	0.827 (0.819,0.834)	0.869 (0.863,0.876)
Random sample	0.527 (0.517,0.537)	0.586 (0.577,0.596)
y-extreme complete case	0.416 (0.407,0.426)	0.475 (0.465,0.485)
y-extreme all case	0.576 (0.567,0.586)	0.640 (0.631,0.650)
z-extreme complete case	0.799 (0.791,0.807)	0.850 (0.843,0.857)
z-extreme complete case (residual)	0.722 (0.713,0.731)	0.850 (0.843,0.857)
z-extreme all case	0.797 (0.789,0.805)	0.850 (0.843,0.857)
z-extreme all case (residual)	0.721 (0.712,0.730)	0.850 (0.843,0.857)
y-extreme binary	0.113 (0.107,0.120)	0.119 (0.113,0.125)
z-extreme binary	0.658 (0.649,0.667)	0.792 (0.784,0.800)

Table 1: Estimated power (with Clopper-Pearson 95% confidence intervals) for single variant tests at significance level $\alpha = 0.05$ and with sample size $n = N/2$; random sampling, y-extreme sampling and z-extreme sampling. Residual phenotype tests under the z-extreme sampling design are denoted by (residual). Coefficients $\beta_0 = 5$, $\beta_1 = 10$, $\beta_2 = 5$, $\beta_{\mathrm{g}} = 0$.

p_0, p_1	0.4, 0.1	0.25, 0.25
$N = 20000$, $\beta_\mathrm{g} = 0.15$		
z-extreme complete case (random)	0.836 (0.828,0.843)	0.881 (0.874,0.887)
z-extreme complete case	0.836 (0.829,0.843)	0.881 (0.874,0.887)
z-extreme all case	0.836 (0.829,0.843)	0.880 (0.874,0.887)
$N = 20000$, $\beta_\mathrm{g} = 0$		
z-extreme complete case (random)	0.049 (0.045,0.054)	0.051 (0.047,0.056)
z-extreme complete case	0.049 (0.045,0.054)	0.051 (0.047,0.056)
z-extreme all case	0.049 (0.045,0.054)	0.051 (0.047,0.056)
$N = 10000$, $\beta_\mathrm{g} = 0.21$		
z-extreme complete case (random)	0.825 (0.818,0.833)	0.871 (0.865,0.878)
z-extreme complete case	0.825 (0.818,0.833)	0.872 (0.865,0.878)
z-extreme all case	0.825 (0.818,0.833)	0.870 (0.864,0.877)
$N = 10000$, $\beta_\mathrm{g} = 0$		
z-extreme complete case (random)	0.052 (0.048,0.056)	0.051 (0.047,0.056)
z-extreme complete case	0.052 (0.048,0.057)	0.051 (0.047,0.056)
z-extreme all case	0.052 (0.048,0.057)	0.051 (0.047,0.056)
$N = 2000$, $\beta_\mathrm{g} = 0.45$		
z-extreme complete case (random)	0.796 (0.788,0.804)	0.842 (0.835,0.849)
z-extreme complete case	0.797 (0.789,0.805)	0.843 (0.835,0.850)
z-extreme all case	0.796 (0.788,0.804)	0.842 (0.834,0.849)
$N = 2000$, $\beta_\mathrm{g} = 0$		
z-extreme complete case (random)	0.048 (0.044,0.052)	0.047 (0.043,0.051)
z-extreme complete case	0.048 (0.044,0.053)	0.047 (0.043,0.052)
z-extreme all case	0.048 (0.044,0.053)	0.047 (0.043,0.051)

Table 2: Estimated power ($\beta_\mathrm{g} > 0$) and type-1 error rates ($\beta_\mathrm{g} = 0$) (with Clopper-Pearson 95% confidence intervals) for z-extreme sampling and single variant tests at significance level $\alpha = 0.05$, applying a score test for random samples, a complete case test with l_i and u_i as defined for z-extreme sampling, and applying the all case score test. Sample size $n = N/2$. Coefficients $\beta_0 = 5$, $\beta_1 = 10$, $\beta_2 = 5$.

p_0, p_1	0.4, 0.1	0.25, 0.25
$N = 20000, \beta_g = 0$		
Full sample	0.048 (0.044,0.052)	0.052 (0.048,0.057)
Random sample	0.046 (0.042,0.051)	0.052 (0.047,0.056)
y-extreme complete case	0.048 (0.044,0.053)	0.050 (0.046,0.054)
y-extreme all case	0.048 (0.044,0.052)	0.051 (0.047,0.055)
z-extreme complete case	0.048 (0.044,0.053)	0.052 (0.048,0.056)
z-extreme complete case (residual)	0.028 (0.025,0.031)	0.052 (0.048,0.056)
z-extreme all case	0.048 (0.044,0.053)	0.052 (0.048,0.056)
z-extreme all case (residual)	0.028 (0.025,0.031)	0.052 (0.048,0.056)
y-extreme binary	0.050 (0.046,0.055)	0.052 (0.048,0.057)
z-extreme binary	0.030 (0.027,0.033)	0.051 (0.047,0.056)
$N = 10000, \beta_g = 0$		
Full sample	0.049 (0.045,0.053)	0.052 (0.047,0.056)
Random sample	0.051 (0.047,0.056)	0.051 (0.047,0.056)
y-extreme complete case	0.051 (0.046,0.055)	0.052 (0.048,0.056)
y-extreme all case	0.052 (0.048,0.056)	0.051 (0.046,0.055)
z-extreme complete case	0.049 (0.045,0.053)	0.051 (0.047,0.055)
z-extreme complete case (residual)	0.027 (0.024,0.031)	0.051 (0.047,0.056)
z-extreme all case	0.049 (0.045,0.053)	0.051 (0.047,0.056)
z-extreme all case (residual)	0.027 (0.024,0.031)	0.051 (0.047,0.056)
y-extreme binary	0.054 (0.049,0.059)	0.050 (0.046,0.055)
z-extreme binary	0.031 (0.028,0.034)	0.050 (0.046,0.054)
$N = 2000, \beta_g = 0$		
Full sample	0.050 (0.045,0.054)	0.051 (0.046,0.055)
Random sample	0.050 (0.046,0.055)	0.050 (0.046,0.055)
y-extreme complete case	0.053 (0.049,0.058)	0.049 (0.045,0.054)
y-extreme all case	0.052 (0.048,0.057)	0.048 (0.044,0.052)
z-extreme complete case	0.049 (0.045,0.054)	0.051 (0.046,0.055)
z-extreme complete case (residual)	0.027 (0.024,0.030)	0.050 (0.046,0.055)
z-extreme all case	0.049 (0.045,0.053)	0.050 (0.046,0.054)
z-extreme all case (residual)	0.027 (0.024,0.030)	0.050 (0.046,0.055)
y-extreme binary	0.063 (0.058,0.068)	0.056 (0.051,0.061)
z-extreme binary	0.030 (0.026,0.033)	0.050 (0.046,0.054)

Table 3: Estimated type-1 error rates (with Clopper-Pearson 95% confidence intervals) for single variant tests at significance level $\alpha = 0.05$ and with sample size $n = N/2$; random sampling, y-extreme sampling and z-extreme sampling. Residual phenotype tests under the z-extreme sampling design are denoted by (residual). Coefficients $\beta_0 = 5$, $\beta_1 = 10$, $\beta_2 = 5$, $\beta_g = 0$.