Omnibus Inverse Normal Transformation Based Association Test Improves Power in Genome-Wide Association Studies of Quantitative Traits

Zachary R. McCaw,[1] Jacqueline M. Lane,[2] Richa Saxena,[2] Susan Redline,[3] Xihong Lin[1,4,*]

Abstract

Quantitative traits analyzed in Genome-Wide Association Studies (GWAS) frequently have non-normal distributions. Despite its widespread use, standard linear regression has reduced power and inflated type I error when applied to such traits. Association testing procedures have often incorporated the rank-based Inverse Normal Transformation (INT) to mitigate departures from normality. Existing INT-based association tests may be classified into two types: direct approaches (D-INT), in the which phenotype itself is transformed, and indirect approaches (I-INT), in which phenotypic residuals are transformed. In this paper, we formally describe the different INT-based association tests, and compare them with the standard, untransformed association test (UAT). We discuss their assumptions, underlying generative models, and connections. We show that, for non-normally distributed quantitative traits, INT-based tests provide robust control of the type I error, and have higher power than UAT. The relative powers of D-INT and I-INT depend on the underlying data generating process. Since neither approach is uniformly most powerful, we combine them into an adaptive, INT-based omnibus test (O-INT). Our omnibus test is robust to model misspecification, protects the type I error, and is well powered against a wide range of non-normal traits. In fact O-INT is nearly as powerful as selecting D-INT or I-INT based on prior knowledge of the underlying data generating mechanism. We apply the proposed methods to GWAS of spirometry traits in the UK Biobank (UKB), recovering previous reported loci and identifying novel associations.

[1] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA; [2] Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA; [3] Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, MA; [4] Department of Statistics, Harvard University, Cambridge, MA.
[*] Correspondence: xlin@hsph.harvard.edu

Introduction

Linear regression is the standard method for analyzing quantitative traits in Genome-wide Association Studies (GWAS). However, the operating characteristics of linear regression are sensitive to the underlying residual distribution. Many complex, quantitative traits have markedly non-normal residuals. An example is peak expiratory flow (PEF), a spirometry measurement whose distribution is skewed and asymmetric, even on the log scale. The rank-based inverse normal transformation (INT) is commonly applied during GWAS of non-normally distributed traits. In the direct, INT-based test (D-INT), the trait itself is transformed prior to association analysis. In the indirect, INT-based tests (I-INT), the trait is first regressed on covariates to obtain residuals, then the transformed residuals are tested for association. In this paper, we formally define and systematically compare the direct and indirect approaches. We describe the data generating process for which each approach is suited. We propose an adaptive, INT-based omnibus test (O-INT) that robustly controls the type I error, and is well-powered across the range of non-normal traits one is likely to encounter in practice.

Linear regression places few assumptions on the form of the residual distribution: in samples of sufficient size, only mean zero and finite variance are required. However, inference in linear regression is sensitive to asymmetry (skew) and diffuseness (kurtosis) of the residual distribution. In the presence of non-normal residuals, the statistical performance of linear regression is lacking. Specifically, in modest samples, linear regression is susceptible to inflated type I error. Although the type I error is better controlled in large samples, the sample sizes required to achieve nominal control of the type I error at genome-wide significance levels are often substantial. For instance, if a trait's residual distribution is approximately t with 3 degrees of freedom, then the sample size required for valid inference at $\alpha = 10^{-6}$ exceeds 10^5. Even when sample sizes are sufficiently large to protect the type I error, the power of linear regression to detect associations with non-normal traits diminishes rapidly as the residual distribution becomes increasingly skewed or kurtotic. Transforming the outcome to increase residual normality generally improves power.

Transformation of non-normal quantitative traits is a common practice in GWAS[4;32]. Although researchers often attempt to transform a trait such that its marginal distribution is approximately normal[10], transformation can benefit the analysis even if normality is not achieved.

The goal of transformation is to obtain residuals that are more symmetric and concentrated around zero. By doing so, the sampling distribution of the association test statistics more quickly approaches its asymptotic limit. Transformations commonly used in GWAS include the logarithm, the Box-Cox family, the warping transform, and the rank-based Inverse Normal Transform (INT). The logarithm has been applied to GWAS of neutrophil counts[30], plasma lipids[28], and pharmacokinetic coefficients[38]. The Box-Cox transformation has been applied to GWAS of vitamin D levels[2], microbiome composition[8], and circulating F-cell proportion[7]. The Box-Cox transformation includes the logarithm as a special case. Fusi and colleagues[17] proposed a warping transformation consisting of a linear combination of smooth step functions, and applied it to GWAS of metabolic traits. The Box-Cox and warping transformations are both parametric, requiring the estimation of additional tuning parameters.

INT is a non-parametric mapping that replaces the sample quantiles by quantiles from the standard normal distribution. Regardless of its initial distribution, the limiting distribution of a continuous phenotype after INT is standard normal. INT-based tests may be classified into direct and indirect approaches. In the former approach, INT is applied directly to the phenotype, and the INT-transformed phenotype is regressed on genotype and covariates. Covariates may include age, sex, and adjustments for population structure, such as ancestry principal components (PCs). D-INT has been applied to GWAS of BMI[34], circulating lipids[5], polysomnography signals[11], and many quantitative traits in the UK Biobank[1]. In the latter approach, the phenotype is first regressed on covariates to obtain residuals, then the INT-transformed phenotypic residuals are regressed on genotype, with or without adjustment for population structure. I-INT has been applied to GWAS of gene expression[15;16], serum metabolites[22], and spirometry measurements[33]. However, the relative performance of D-INT versus I-INT has not been studied in detail. For the non-normal quantitative traits encountered in practice, it is unclear which of the these methods will more robustly control the type I error, and which will provide better power. Neither have the connections between D-INT and I-INT been systematically examined.

As discussed by Beasley and colleagues[6], the question of whether INT-based methods have desirable operating characteristics in the GWAS context has not been critically evaluated. INT of the outcome in a regression model does not guarantee correct model specification. This

is because the standard linear model, considered parametrically, assumes normality for the residuals, not for the outcome. If the observed trait were generated by non-linear transformation of a latent normal trait, then the optimal transformation would simply invert the original, non-linear transformation. Interestingly, if the generative transformation is rank-preserving and the covariate effects are small, INT effectively provides the inverse. For example, if the observed trait is generated by exponentiating a latent normal trait, then the optimal transformation is the logarithm, yet D-INT of the observed trait is fully efficient in large samples. In contrast, if the mean of the observed trait is linear in covariates, but its conditional distribution is non-normal, then I-INT is asymptotically exact. In practice, the generative model for the trait is seldom known. It is therefore desirable to have a transformation strategy that is broadly applicable to quantitative traits, robustly effective for controlling for type I error rate, well powered to detect true associations, and which does not require parametric assumptions or tuning.

Here we present a systematic study of direct and indirect INT-based association tests, and provide recommendations on how to apply the INT in practice. We begin by formally defining D-INT and I-INT, studying their underlying assumptions and connections. Our derivation of I-INT agrees with recent work recommending double adjustment for covariates during INT-based association testing[35]. Through extensive simulations covering a range of non-normal phenotypes, we compare D-INT and I-INT with the standard untransformed association test (UAT). Our results show that the INT-based tests dominate the UAT in terms of statistical size and power. Whereas the UAT was sensitive to non-normality, resulting in failure to consistently control the type I error, the INT-based tests were robust. Even when the sample size was sufficient for the UAT to provide valid inference, the INT-based tests were considerably more powerful. However, the relative powers of D-INT and I-INT depend on the underlying data generating process, and neither approach was uniformly most powerful. D-INT performed best against non-linear transformations of a latent normal trait. I-INT performed best against traits with additive but non-normal residuals. Since the data generating model underlying the trait is seldom known in practice, we next propose an adaptive Omnibus test (O-INT) that synthesizes the direct and indirect approaches. O-INT robustly controls the type I error, and for any given trait is nearly as powerful as the more effective of D-INT and I-INT. We have implemented all

candidate INT-based tests (D-INT, I-INT, and O-INT) in the R package `RNOmni`.

We applied the UAT, D-INT, I-INT, and O-INT to GWAS of spirometry traits in the UK Biobank[36]. To demonstrate the power advantage provided by O-INT, we compare the results from the overall analysis ($N \approx 292K$) with those from a subgroup analysis ($N \approx$ 29K) among asthmatics. All associations identified by O-INT in the asthmatic subgroup were declared significant by UAT in the overall analysis. Hence UAT and O-INT tests agree as to the importance of these loci. However, the more-powerful O-INT test was able to detect them using only a fraction (9.7%) of the sample. In both the asthmatic subgroup and the overall analysis, O-INT realized discovery gains over the UAT for all phenotypes. Many of the top associations reported here have previously been associated with lung-related traits in the GWAS catalogue[29], or are proximal to loci that have.

The remainder of this paper is structured as follows. In the methods section, we present the UAT, formally define the INT, theoretically study the direct and indirect INT-based tests, and describe the omnibus test. In the results section, we present evidence from simulation studies that INT-based association tests dominate the UAT, and apply these tests to perform GWAS analysis of spirometry traits using the UK Biobank. We conclude with a discussion of the implications of our findings for GWAS of quantitative traits.

Material and Methods

Statistical Methods

Setting

Consider a sample of N independent subjects. For each subject, the following data are observed: a continuous (quantitative) phenotype y_i, genotype g_i at the locus of interest, and a vector $\boldsymbol{x}_i$ of p covariates. In our data application, the phenotype y_i is a spirometry measurement. Covariates may include an intercept, age, sex, and genetic principal components (PCs). Let $\boldsymbol{y}$ denote the $n \times 1$ vector of sample traits, $\boldsymbol{g}$ the $n \times 1$ vector of genotypes, and $\boldsymbol{X}$ the $n \times p$ covariate design matrix.

Untransformed Association Test

The untransformed association test (UAT) is derived from the normal linear model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_x + \boldsymbol{g}\beta_g + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\epsilon}$ is an $n \times 1$ residual vector following $\boldsymbol{\epsilon} \sim N\big(\boldsymbol{0}, \sigma^2\boldsymbol{I}\big)$, β_g is the effect of genotype, and $\boldsymbol{\beta}_x$ contains the regression coefficients for covariates. Define the *error projection matrix*, $\boldsymbol{P}_x = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. The *phenotypic residual* $\boldsymbol{e}_y = \boldsymbol{P}_x\boldsymbol{y}$ is the projection of $\boldsymbol{y}$ onto the orthogonal complement of the column space of $\boldsymbol{X}$. One can obtain the phenotypic residual by regressing the phenotype $\boldsymbol{y}$ on the covariates $\boldsymbol{X}$ according to the null model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_{0,x} + \boldsymbol{\varepsilon}$. The score statistic assessing $H_0 : \beta_g = 0$ has a simple expression in terms of the error projection matrix $\boldsymbol{P}_x$:

$$T_U = \boldsymbol{y}'\boldsymbol{P}_x\boldsymbol{g}(\boldsymbol{g}'\boldsymbol{P}_x\boldsymbol{g})^{-1}\boldsymbol{g}'\boldsymbol{P}_x\boldsymbol{y}/\sigma^2. \tag{2}$$

The Wald statistic assessing $H_0 : \beta_g = 0$ takes the same form as (2). The difference between the Wald and score statistics lies in how the *residual variance* σ^2 is estimated. For computational efficiency, we adopt the score test, in which the residual variance is estimated under the null hypothesis. This allows us to reuse the same estimate of σ^2 for association testing at each locus in the genome. Under $H_0 : \beta_g = 0$, an unbiased estimate of σ^2 is given by:

$$\hat{\sigma}^2 = \frac{\boldsymbol{e}_y'\boldsymbol{e}_y}{n-p} = \frac{\boldsymbol{y}'\boldsymbol{P}_x\boldsymbol{y}}{n-p}, \tag{3}$$

and the UAT score statistic T_U in (2) follows a χ^2_1 distribution. If the normal residual assumption is relaxed to allow for an arbitrary distribution with mean zero and finite variance σ^2, then for increasing sample size, T_U still follows an asymptotic χ^2_1 distribution. Consequently, for any continuous trait with constant residual variance, including those with non-normal residual distributions, standard inference using the UAT is eventually valid in samples of sufficient size. However, the sample size needed will depend on the skew and kurtosis of the trait, and may be impractically large. Moreover, as we will show, even in samples of sufficient size for valid inference, the UAT is generally less powerful than INT-based tests.

Rank-Based Inverse Normal Transformation

The rank-based inverse normal transformation (INT) is a non-parametric mapping from the sample quantiles to the quantiles of the standard normal distribution. INT is applicable to observations from any (absolutely) continuous distribution. The process may be broken into two steps. In the first, the observations are replaced by their fractional ranks. This is equivalent to transforming the observations by their empirical cumulative distribution function (ECDF) $\hat{F}(\cdot)$. If ω is any continuous random variable whose cumulative distribution function is $F_\omega(\cdot)$, then the transformed random variable $U = F(\omega)$ is uniformly distributed[12]. Since the empirical process $\hat{F}(\cdot)$ converges uniformly to $F(\cdot)$, in samples of sufficient size, $\hat{U} = \hat{F}(\omega)$ is uniformly distributed. After transformation by the ECDF, the observations reside on the probability scale. In the next step, these probabilities are mapped to Z-scores using the *probit* function $\Phi^{-1}(\cdot)$, which is the quantile function of the standard normal distribution. If U is uniformly distributed, then $\Phi^{-1}(U)$ follows the standard normal distribution[12]. Consequently, in large samples, $\mathrm{INT}(\omega) = \Phi^{-1}\left[\hat{F}(\omega)\right] = \Phi^{-1}(\hat{U})$ follows the standard normal distribution.

In practice, an offset is introduced when converting the observations to fractional ranks. This offset ensures that all fractional ranks are strictly between zero and one, which in turn guarantees that all Z-scores are finite after probit transformation. Suppose that a continuous measurement ω_i is observed for each of the N subjects. The modified INT is:

$$\mathrm{INT}(\omega_i) = \Phi^{-1}\left[\frac{\mathrm{rank}(\omega_i) - c}{n + 1 - 2c}\right], \tag{4}$$

where $c \in [0, 1/2]$ is the offset. Standard software uses $c = 3/8$, which is referred to as the *Blom* offset[6]. We adopt this value hereafter. Since other choices for c lead to Z-scores that are nearly linear transformations of one another, the choice of offset is considered immaterial[20].

Direct Inverse Normal Transformation (D-INT)

In Direct INT (D-INT), the transformation is applied directly to the phenotype. In this section, we first present the D-INT test, then discuss the underlying data generating process and approximations. The association model underlying D-INT is:

$$\mathrm{INT}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}_x + \boldsymbol{g}\beta_g + \boldsymbol{\epsilon}_D, \tag{5}$$

where $\boldsymbol{\epsilon_D} \sim N\big(\mathbf{0}, \sigma_D^2 \boldsymbol{I}\big)$. Comparison of (5) with (1) reveals that the only difference from UAT is the replacement $\boldsymbol{y}$ by INT($\boldsymbol{y}$). The resulting D-INT score statistic T_D for assessing $H_0 : \beta_g = 0$ takes the form:

$$T_D = \boldsymbol{z}'\boldsymbol{P}_x\boldsymbol{g}(\boldsymbol{g}'\boldsymbol{P}_x\boldsymbol{g})^{-1}\boldsymbol{g}'\boldsymbol{P}_x\boldsymbol{z}/\sigma_D^2, \tag{6}$$

where $\boldsymbol{z} = \text{INT}(\boldsymbol{y})$. The score statistic in (6) is immediately comparable to (2), replacing $\boldsymbol{y}$ by $\boldsymbol{z}$ and σ^2 by σ_D^2. The residual variance from the D-INT model (5) is estimated as:

$$\hat{\sigma}_D^2 = \frac{\boldsymbol{z}'\boldsymbol{P}_x\boldsymbol{z}}{n-p}.$$

A p-value is assigned with reference to the χ_1^2 distribution.

D-INT is adapted for data generating processes of the form:

$$\boldsymbol{y} = h\big(\boldsymbol{X}\boldsymbol{\beta}_x + \boldsymbol{g}\beta_g + \boldsymbol{\epsilon}_D^*\big), \tag{7}$$

where $\boldsymbol{\epsilon}_D^* \sim N(\mathbf{0}, \sigma_D^2\boldsymbol{I})$, and $h(\cdot)$ is a monotone (i.e. rank-preserving), non-linear transformation. An example is the log-normal phenotype, for which $h(t) = \exp(t)$. Under this data generating process, there exists a latent normally distributed outcome $\boldsymbol{y}^* = \boldsymbol{X}\boldsymbol{\beta}_x + \boldsymbol{g}\beta_g + \boldsymbol{\epsilon}_D^*$ of which the observed phenotype is a non-linear transformation $\boldsymbol{y} = h(\boldsymbol{y}^*)$. When $h(\cdot)$ is non-linear, the regression equation $E[y_i|\boldsymbol{x}_i, g_i]$ is also non-linear, and the residuals $\boldsymbol{\epsilon}_D^*$ have non-additive effects. However, there exists a transformed scale on which the mean model is linear and has additive normal residuals, namely $h^{-1}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}_x + \boldsymbol{g}\beta_g + \boldsymbol{\epsilon}_D$. Under the $H_0 : \beta_g = 0$, the efficient score for β_g is:

$$\tilde{\mathcal{U}}_{\beta_g} = \frac{1}{\sigma_D^2}\boldsymbol{g}'\boldsymbol{P}_x h^{-1}(\boldsymbol{y}).$$

Under the data generating process in (7), the *conditional distribution* of the transformed-scale residual $\epsilon_{D,i}^* = h^{-1}(y) - \boldsymbol{x}_i\boldsymbol{\beta}_x - g_i\beta_g$, given $\boldsymbol{x}_i$ and g_i, is $F_\epsilon(\epsilon_{D,i}^*|\boldsymbol{x}_i, g_i) = \Phi(\epsilon_{D,i}^*/\sigma_D)$. D-INT makes the approximation that $\text{INT}(y_i) = \Phi^{-1}[\hat{F}_y(y_i)] \approx \sigma_D^{-1}h^{-1}(y_i)$, where $\hat{F}_y(\cdot)$ is the ECDF for the *marginal distribution* of the observed phenotype $F_y(\cdot)$. The marginal CDF is related to the conditional CDF via:

$$\begin{aligned} F_y(y_i) &= \int F(y_i|\boldsymbol{x}_i, g_i)dF(\boldsymbol{x}_i, g_i) \\ &= \int \Phi\left\{\sigma_D^{-1}[h^{-1}(y) - \boldsymbol{x}_i\boldsymbol{\beta}_x - g_i\beta_g]\right\} dF(\boldsymbol{x}_i, g_i), \end{aligned} \tag{8}$$

where $F(\boldsymbol{x}_i, g_i)$ is the joint CDF of $\boldsymbol{x}_i$ and g_i. The empirical counterpart to the marginal-conditional relationship in (8) is:

$$\hat{F}_y(y_i) = \frac{1}{n}\sum_{i=1}^{n} \hat{F}_y(y_i|\boldsymbol{x}_i, g_i) = \frac{1}{n}\sum_{i=1}^{n} \Phi\big\{\sigma_D^{-1}[h^{-1}(y_i) - \boldsymbol{x}_i'\boldsymbol{\beta}_x - g_i\beta_g]\big\}.$$

Under the strong $H_0 : (\boldsymbol{\beta}_x = 0)$ *and* $(\beta_g = 0)$, $\hat{F}_y(\cdot)$ converges to $\Phi\{\sigma_D^{-1}h^{-1}(y_i)\}$. Thus, under the strong null, the D-INT approximation $\text{INT}(y_i) = \Phi^{-1}[\hat{F}_y(y_i)] \approx \sigma_D^{-1}h^{-1}(y_i)$ is asymptotically exact. Under the standard $H_0 : \beta_g = 0$, the D-INT approximation is accurate when $\boldsymbol{\beta}_x$ is small.

Indirect Inverse Normal Transformation (I-INT)

In Indirect INT (I-INT), the phenotype is first regressed on covariates to obtain residuals, then the INT-transformed phenotypic residuals are regressed on genotype. Since INT is non-linear, after transformation the phenotypic residuals are no longer orthogonal to covariates. Thus, secondary adjustment for covariates has been recommended[35]. In this section, we first consider the generative model and approximations underlying I-INT, then present the testing procedure.

To develop the I-INT, consider the data generating process:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_x + \boldsymbol{g}\beta_g + \boldsymbol{\varepsilon}, \tag{9}$$

where the residuals $\varepsilon_i \sim F_\varepsilon$ follow an arbitrary continuous distribution with mean zero and constant finite variance. Under (9), $F(y_i|\boldsymbol{x}_i, g_i) = F_\varepsilon(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_x - g_i\beta_g)$. Thus, the data generating processes in (5) and (9) are equivalent under the strong null $H_0 : (\boldsymbol{\beta}_x = 0)$ *and* $(\beta_g = 0)$.

Letting $f_\varepsilon(\cdot) = \partial F_\varepsilon / \partial \varepsilon(\cdot)$ denote the residual density, the model log likelihood is:

$$\ell(\beta_g, \boldsymbol{\beta}_x) = \ln f_\varepsilon(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_x - \boldsymbol{g}\beta_g).$$

We obtain a profile likelihood for β_g by observing that, for any residual distribution F_ε, the ordinary least squares estimate $\tilde{\boldsymbol{\beta}}_x = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{g}\beta_g)$ is *consistent* for $\boldsymbol{\beta}_x$. Thus, a profile likelihood for β_g can be consistently estimated by:

$$\ell_p(\beta_g) = \ell(\beta_g, \tilde{\boldsymbol{\beta}}_x) = \ln f_\varepsilon\big[\boldsymbol{P}_x(\boldsymbol{y} - \boldsymbol{g}\beta_g)\big].$$

Letting $\mathcal{U}_\varepsilon(\cdot) = \partial\ell(\cdot)/\partial\varepsilon$, a consistent estimate of the *efficient score* of β_g is:

$$\tilde{\mathcal{U}}_{\beta_g} = \boldsymbol{g}'\boldsymbol{P}_x\mathcal{U}_\varepsilon\big[\boldsymbol{P}_x(\boldsymbol{y} - \boldsymbol{g}\beta_g)\big].$$

Under $H_0 : \beta_g = 0$, the efficient score of β_g is consistently estimated by:

$$\tilde{\mathcal{U}}_{\beta_g} = \boldsymbol{g}'\boldsymbol{P}_x\mathcal{U}_\varepsilon\big(\boldsymbol{P}_x\boldsymbol{y}\big). \tag{10}$$

Recall that $\boldsymbol{e}_y = \boldsymbol{P}_x\boldsymbol{y}$ denotes the *phenotypic residual*, which is the projection of $\boldsymbol{y}$ onto the orthogonal complement of the column space for $\boldsymbol{X}$, and is the residual after regressing $\boldsymbol{y}$ on $\boldsymbol{X}$. Equation (10) corresponds to the efficient score for β_g from the model:

$$\boldsymbol{e}_y = \boldsymbol{X}\boldsymbol{\alpha}_x + \boldsymbol{g}\beta_g + \boldsymbol{\varepsilon}, \tag{11}$$

where the residual $\boldsymbol{\varepsilon}$ again follows an arbitrary distribution $F_\varepsilon(\cdot)$ with mean zero and finite variance. Since $\boldsymbol{e}_y$ resides in the error space of $\boldsymbol{X}$, the OLS estimate of $\boldsymbol{\alpha}_x$ in (11) is $\tilde{\boldsymbol{\alpha}}_x = \boldsymbol{0}$. In a similar manner, let $\boldsymbol{e}_g = \boldsymbol{P}_x\boldsymbol{g}$ denote the *genotypic residual*, which is the projection of $\boldsymbol{g}$ onto the orthogonal complement of the column space of $\boldsymbol{X}$, and can be obtained by regressing the genotypes $\boldsymbol{g}$ on the covariates $\boldsymbol{X}$ according to the model $\boldsymbol{g} = \boldsymbol{X}\boldsymbol{\gamma}_x + \boldsymbol{\varepsilon}_g$.

Equation (10) is also the score for β_g obtained by regressing the phenotypic residual $\boldsymbol{e}_y$ on the genotypic residual $\boldsymbol{e}_g$ according to the model:

$$\boldsymbol{e}_y = \boldsymbol{e}_g\beta_g + \boldsymbol{\varepsilon}. \tag{12}$$

Under the null hypothesis $H_0 : \beta_g = 0$, model (11) with $\boldsymbol{\alpha}_x$ evaluated at its least square estimator $\tilde{\boldsymbol{\alpha}}_x = \boldsymbol{0}$ and model (12) both reduce to $\boldsymbol{e}_y = \boldsymbol{\varepsilon}$, which implies that:

$$F(y_i|\boldsymbol{x}_i, g_i) = F_\varepsilon(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_x) = F_\varepsilon(\varepsilon_i).$$

Applying the probit transformation to both sides gives the identity $\Phi^{-1}\{F_y(y_i|\boldsymbol{x}_i, g_i)\} = \Phi^{-1}\{F_\varepsilon(\varepsilon_i)\}$, whose finite sample analogue is:

$$\mathrm{INT}(\boldsymbol{e}_y) = \mathrm{INT}(\boldsymbol{\varepsilon}) \dot{\sim} N\big(\boldsymbol{0}, \boldsymbol{I}\big). \tag{13}$$

The conclusion in (13) holds under the standard $H_0 : \beta_g = 0$ for any $\boldsymbol{\beta}_x$ and any residual distribution $F_\varepsilon(\cdot)$. This motivates the I-INT association model, in which the INT-transformed phenotypic residual $\boldsymbol{e}_y$ is regressed on the genotypic residual according to the model:

$$\mathrm{INT}(\boldsymbol{e}_y) = \boldsymbol{e}_g\beta_g + \boldsymbol{\epsilon}_I, \tag{14}$$

where $\boldsymbol{\epsilon}_I \sim N\big(\mathbf{0}, \sigma_I^2 \boldsymbol{I}\big)$. Comparing (14) with (13), the I-INT model is exact under the null hypothesis $H_0 : \beta_g = 0$. Model (14) effectively adjusts for covariates twice: once during regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ to obtain the phenotypic residual $\boldsymbol{e}_y$, and again during regression of $\boldsymbol{g}$ on $\boldsymbol{X}$ to obtain the genotypic residual $\boldsymbol{e}_g$.

Using the equivalence between (11) and (12), one can alternatively fit the I-INT model by regressing the INT-transformed phenotypic residual on both genotype and covariates according to the model:

$$\mathrm{INT}(\boldsymbol{e}_y) = \boldsymbol{X}\boldsymbol{\alpha}_x + \boldsymbol{g}\beta_g + \boldsymbol{\epsilon}_I. \tag{15}$$

This procedure is in agreement with recent recommendations to adjust for covariates twice during INT-based association testing[35]. Under the $H_0 : \beta_g = 0$, the (efficient) score from models (14) and (15) is

$$\mathcal{U}_{\beta_g} = \frac{1}{\sigma_I^2}\boldsymbol{e}_g'\mathrm{INT}(\boldsymbol{e}_y) = \frac{1}{\sigma_I^2}\boldsymbol{g}'\boldsymbol{P}_x\mathrm{INT}(\boldsymbol{P}_x\boldsymbol{y}). \tag{16}$$

Moreover, the score in (16) is asymptotically exact regardless of the form of $F_\varepsilon(\cdot)$. The resulting I-INT test statistic for testing $H_0 : \beta_g = 0$ takes the form:

$$T_I = \tilde{\boldsymbol{z}}'\boldsymbol{P}_x\boldsymbol{g}(\boldsymbol{g}'\boldsymbol{P}_x\boldsymbol{g})^{-1}\boldsymbol{g}'\boldsymbol{P}_x\tilde{\boldsymbol{z}}/\sigma_I^2, \tag{17}$$

where $\tilde{\boldsymbol{z}} = \mathrm{INT}(\boldsymbol{P}_x\boldsymbol{y})$. Comparison of (17) with (2) reveals that the only differences are the replacements $\boldsymbol{y}$ by $\tilde{\boldsymbol{z}}$ and σ^2 by σ_I^2. The residual variance from the I-INT model (14) under the null is:

$$\hat{\sigma}_I^2 = \frac{\tilde{\boldsymbol{z}}'\tilde{\boldsymbol{z}}}{n} = 1.$$

Omnibus Inverse Normal Transformation Test (O-INT)

As shown in the simulation studies, both D-INT and I-INT robustly control the type I error across a spectrum of non-normally distributed traits. However, neither D-INT nor I-INT is uniformly most powerful. Rather, their powers depend on the underlying data generating process, which is seldom known in practice. It is therefore of substantial interest to provide a test that is broadly well powered and robustly controls the type I error, regardless of the trait

distribution and the underlying data generating process. Our proposed omnibus test synthesizes the D-INT and I-INT approaches to achieve this goal.

To construct the omnibus statistic, we apply the Aggregated Cauchy Association Test (ACAT), in which the p-values from related hypothesis tests are converted to Cauchy random deviates then combined[24;25]. The CDF and inverse CDF of the standard Cauchy distribution are:

$$F_C(t) = \frac{1}{2} + \frac{1}{\pi}\arctan(t), \qquad F_C^{-1}(u) = \tan\big\{\pi(u-0.5)\big\}.$$

By odd parity of the tangent function:

$$-F_C^{-1}(u) = -\tan\big\{\pi(u-0.5)\big\} = \tan\big\{\pi(0.5-u)\big\}.$$

Let p_D and p_I denote the p-values from D-INT and I-INT. Define the O-INT test statistic as:

$$T_O = -\frac{1}{2}\big\{F_C^{-1}(p_D) + F_C^{-1}(p_I)\big\} = \frac{1}{2}\Big(\tan\big\{\pi(0.5-p_D)\big\} + \tan\big\{\pi(0.5-p_I)\big\}\Big). \tag{18}$$

Under the null hypothesis, p_D and p_I are each uniformly distributed. Thus, $F_C^{-1}(p_D)$ and $F_C^{-1}(p_I)$ are standard Cauchy random variables. Since the Cauchy distribution is symmetric and closed with respect to convolution, the omnibus statistic $T_O = -\{F_C^{-1}(p_D) + F_C^{-1}(p_I)\}/2$ is again standard Cauchy, even though p_D and p_I are, in general, positively correlated. Based on this fact, we can calculate a p-value for the O-INT statistic (18) according to:

$$p_O = P[T_O > t_O] = 1 - F_C(t_O) = \frac{1}{2} - \frac{1}{\pi}\arctan\big(t_O\big),$$

where t_O is the observed value of T_O.

Simulation Studies

Extensive simulations were conducted to evaluate the statistical size (i.e. type I error) and power of the UAT and three INT-based association tests (D-INT, I-INT, and O-INT). Size simulations were conducted using a data generating process in which genotype had no effect on the phenotype. Power simulations were conducted such that the proportion of phenotypic variation explained (PVE) by genotype was 0.5%.

Genotypes at loci on chromosome one were randomly sampled from the UK Biobank. The population of subjects whose genotypes were sampled is described in the data application section. Simulated covariates included age and sex. Age was drawn from a gamma distribution

with mean 50 and variance 10. Sex was drawn from a Bernoulli distribution with proportion 1/2. To emulate population structure, the top 3 PCs of the empirical genetic relatedness matrix were included among covariates.

For size simulations, a subject-specific linear predictor η_i was generated as $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}_x$. For power simulations, the linear predictor also included a contribution from genotype. Population regression coefficients were drawn independently on each simulation replicate from normal distributions with mean zero. The variance was tuned such that the PVE by age and sex was 20%, and the PVE by PCs was 5%. Note that the PVE by PCs was sufficiently large to substantially inflate the type I error if any PC were omitted from the linear predictor.

We considered multiple mechanisms for generating non-normal, quantitative traits. Broadly, these belonged to two categories: models with additive residuals, as in $y_i = \eta_i + \epsilon_i$, and non-linear transformations of such models, as in $y_i = h(\eta_i + \epsilon_i)$, where $h(\cdot)$ is a non-linear activation function. Here, we report on four representative traits; three models with additive residuals, and one model with multiplicative residuals. The additive models were: (1) a reference trait, generated by adding standard normal residuals to the linear predictor; (2) a skewed trait, generated by adding χ_1^2 residuals to the linear predictor; and (3) a kurtotic trait, generated by adding t_3 residuals to the linear predictor. In all cases, the residual distribution was centered and scaled to have mean zero and unit variance. For the multiplicative model, a log-normal phenotype was generated by adding standard normal residuals to the linear predictor then exponentiating. Specifically, $y_i = \exp(\eta_i + \epsilon_i)$, where $\epsilon_i \sim N(0, 1)$.

Size

For size simulations, a linear predictor of the following form was generated for each subject:

$$\eta_i = \text{Age}_i\beta_A + \text{Sex}_i\beta_S + \text{PC}_{1,i}\beta_1 + \text{PC}_{2,i}\beta_2 + \text{PC}_{3,i}\beta_3.$$

Since the linear predictor η_i included no contribution from genotype, the null hypothesis of $H_0 : \beta_g = 0$ was in fact true. The number of simulation replicates needed to accurately characterize the type I error increases as the significance level α decreases. To obtain estimates of the type I error accurate to at least $\alpha = 10^{-6}$, $R = 10^8$ simulation replicates were performed. Samples sizes of $n \in \{10^3, 10^4, 10^5\}$ were evaluated. On each simulation replicate, a block of 10^4

contiguous loci randomly was selected for association testing. The four phenotypes (normal, skewed, kurtotic, and log-normal) were generated independently for each subject. All loci in the block were individually tested for association with each phenotype using all four methods (UAT, D-INT, I-INT and O-INT), and the p-values were recorded. Size at level $\alpha = 10^{-6}$ was estimated as the proportion of p-values less than or equal to α.

Power

For power simulations, the subject-specific linear predictors included a contribution from genotype:

$$\eta_i = g_i\beta_g + \text{Age}_i\beta_a + \text{Sex}_i\beta_s + \text{PC}_{1,i}\beta_1 + \text{PC}_{2,i}\beta_2 + \text{PC}_{3,i}\beta_3.$$

Statistical power depends on the magnitude of the effect size β_g, the sample size n, and the significance level α. The effect size was simulated as a random effect such that the PVE by genotype was 0.5%. Samples size of $n \in \{10^3, 10^4\}$ were evaluated at significance level $\alpha = 10^{-6}$. Under these configurations, the power of the UAT to detect the association with the normal phenotype at sample size $n = 10^4$ was around 50%. Power was estimated across $R = 10^6$ simulation replicates. On each replicate, a randomly selected locus served as the causal locus. The four phenotypes (normal, skewed, kurtotic, and log-normal) were generated independently for each subject. The causal locus was tested for association with each phenotype using all four methods (UAT, D-INT, I-INT, O-INT), and the p-values were recorded. Power at level $\alpha = 10^{-6}$ was estimated as the proportion of p-values less than or equal to α. The non-centrality parameter was estimated by transforming the p-values to χ^2_1 random variables, then taking the average minus one.

UK Biobank Application

We conducted genetic association analyses of spirometry phenotypes among subjects from the UK Biobank (UKB). As detailed elsewhere[3;36], the UKB is a prospective cohort study of over 500K subjects recruited from across the United Kingdom between 2007 and 2010. Participants aged 40 to 69 agreed to provide bio-specimens (blood, urine, saliva), undergo physical examinations, and complete health and lifestyle questionnaires. To mitigate confounding due to

population structure, our study population was restricted to $N = 337,536$ unrelated subjects of white, British ancestry. Individuals who had withdrawn consent were excluded.

Spirometry measurements were recorded using a Vitalograph Pneumotrac 6800. As described on the UKB Portal (`https://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=100020`), participants were asked to record up to three blows, each lasting for at least 6 seconds, within a period of about 6 minutes. The phenotypes of primary interest were forced expiratory volume in 1 second (FEV1, UKB field: 3063), forced vital capacity (FVC, UKB field: 3062), and the peak expiratory flow (PEF, UKB field: 3064). The ratio of FEV1 to FVC was taken (FEV1/FVC), and the logarithm of PEF was taken (lnPEF) to reduce skew. Subjects lacking spirometry measurements were removed ($N = 29,617$), as were subjects who had smoked ($N = 3,878$), used an inhaler ($N = 2,249$), or used caffeine ($N = 6,431$) within 1 hour prior to measurement.

Covariates included age, sex, BMI, and height. Among the subjects with spirometry data, age and sex were available for everyone, while those missing either BMI ($N = 4,381$) or height ($N = 299$) were removed. The overall analysis consisted of $N = 291,729$ subjects meeting all inclusion criteria. A subgroup analyses was conducted among subjects with physician diagnosed asthma ($N = 28,589$), as reported in UKB field 6152.

UKB genotyping and quality control protocols are documented online (`http://www.ukbiobank.ac.uk/scientists-3/genetic-data/`). Briefly, genotypes for all subjects were ascertained using either the Affymetrix UK BiLEVE Axiom array, or the Affymetrix UK Biobank Axiom array. Among the subjects included in our overall analysis, 88.2% were genotyped using the Biobank platform. During quality control, genotype was set to missing for loci that exhibited unexpected differences in allele frequency across genotyping batches. To reduce both the computational burden and the risk of spurious associations, we focused on the approximately 800,000 directly genotyped, as opposed to imputed, loci. Of these, our analysis included 360,761 autosomal SNPs with a sample minor allele frequency exceeding 5%, and a per locus missingness of less than 10%.

Genetic association analyses were implemented in `R`[37] using the `RNOmni` package. The association model adjusted for age, sex, BMI, two orthogonal polynomials in height, genotyping array (UKB field: 22000), and the top 20 population structure PCs (UKB field: 22009). Each

locus was tested individually for association with the four spirometry phenotypes (FEV1, FVC, FEV1/FVC, lnPEF). The results were greedily 'clumped' in `PLINK` (1.90)[31] using a 1000 kb radius at an r^2 threshold of 0.2. To evaluate the presence of residual confounding, LD score regression (LDSC)[9] was performed in `Python` using `ldsc` (`https://github.com/bulik/ldsc`). LD scores were calculated using all unrelated subjects of white, British ancestry.

Results

Simulation Studies

We conducted extensive simulations comparing the statistical size and power of the INT-based tests (D-INT, I-INT, O-INT) with those of the untransformed association test (UAT). Each test was evaluated with respect to four independent phenotypes. The phenotype with additive normal residuals provided a baseline against which all tests were expected to perform well. The remaining three phenotypes were simulated to illustrate the primary ways in which the residual distribution of a quantitative trait might deviate from normality. Relative to normal residuals, the phenotype with additive χ^2_1 residuals exhibited excess skew, while the phenotype with additive t_3 residuals exhibited excess kurtosis. For these phenotypes, the standard linear model assumed by the UAT is correctly specified. However, the excess residual skew and kurtosis affect the *rate* at which the sampling distribution of the test statistic approaches its theoretical limit, and hence the finite sample operating characteristics. Finally, a heavily skewed log-normal phenotype was generated by exponentiating a latent normal phenotype. In this case, the residuals act on the observed phenotype in a multiplicative, rather than an additive, fashion. As such, the standard linear model is in fact misspecified, regardless of the sample size.

Size Simulations

The uniform QQ plots in Figure 1 summarize the distribution of p-values at sample size $N = 10^3$ for each combination of phenotype (row) and association test (column). All association tests performed well against the normal phenotype (row 1), providing uniformly distributed p-values. The same was observed for increasing sample size $N \in \{10^4, 10^5\}$ (see Figures S1-2). The UAT (column 1) exhibited inflated type I error against the non-normal phenotypes. Inflation was most severe at sample size $N = 10^3$, and attenuated with increasing sample size. In contrast,

the INT-based tests provided valid p-values when applied to the non-normal phenotypes, even at sample size $N = 10^3$. D-INT maintained the type I error well across all the scenarios, while the I-INT exhibited slight deflation in the case of the log-normal phenotype. Deflation was ameliorated with increasing sample size (Figures S1-2).

UAT incurred the most inflation against the log-normal (Figure 1, row 4). This is likely because the standard linear model is misspecified for a phenotype in which the residuals have multiplicative rather than additive effects. Among INT-based tests, I-INT exhibited slight deflation against the log normal phenotype. This is again likely due to model misspecification. However, unlike inflation, deflation of the p-values does not invalidate inference, but merely results in a conservative type I error rate. In contrast, p-values from D-INT and O-INT were uniformly distributed. The former observation is expected as D-INT model more nearly specifies the log-normal data generating mechanism.

For the skewed χ^2_1 phenotype (Figure 1, row 2), the UAT exhibited inflation at sample size $N = 10^3$, while the INT-based tests effectively maintained the type I error. By sample size $N = 10^4$, all methods performed well. For kurtotic t_3 phenotype (Figure 1, row 3), the UAT again exhibited inflated at sample size $N = 10^3$, and the inflation was still evident at sample size $N = 10^4$, notwithstanding the fact that the UAT is correctly specified for this phenotype. The INT-based methods performed well against the t_3 phenotype at all sample sizes.

Size estimates at $\alpha = 10^{-6}$ and sample sizes $N \in \{10^3, 10^4, 10^5\}$, across 10^8 simulation replicates, are presented in Table 1. For all non-normal phenotypes, the UAT has substantially inflated type I error at sample size $N = 10^3$. This includes the skewed χ^2_1 and kurtotic t_3 phenotypes, for which the UAT should provide asymptotically valid inference. Although the type I error rate approached its nominal level with increasing sample size, for the kurtotic and log-normal phenotypes the UAT still exhibited excess type I error at $N = 10^5$. For the non-normal phenotypes, D-INT generally provided nearly the nominal type I error, while I-INT was slightly conservative. Note, however, that this does not imply I-INT is less powerful (see below). For all phenotypes and sample sizes considered, the omnibus test provided nominal control of the type I error.

Power Simulations

Figure 2 presents the power for sample size $N = 10^4$ and α level 10^{-6}, as well as the relative efficiency of the INT-based tests to the UAT. Relative efficiency was calculated as the ratio of the χ^2_1 non-centrality parameters. This performance metric has the advantage of not depending on either the particular sample size or α level examined. For all phenotypes, the proportion of variation explained (PVE) by genotype at the causal locus was 0.5%. The UAT is theoretically most powerful for detecting associations with the normal phenotype. However, the INT-based tests were fully efficient, achieving a relative efficiency of one. Despite having inflated type I error under the null hypothesis, the UAT was consistently *least* powerful for detecting true associations with the non-normal phenotypes. Since, for these phenotypes, the relative efficiencies of the INT-based tests always exceeded one, this conclusion is expected to extend across sample sizes and significance thresholds. For the log normal phenotype, D-INT was most powerful, achieving twice the efficiency of the UAT. For this phenotype, the log transform is theoretically optimal. Since the log transform maps the log normal phenotype to a normal phenotype, the power of the log transform against the log normal phenotype is identically the power of the UAT against the normal phenotype. Comparing the power of D-INT against the log normal phenotype with that of UAT against the normal phenotype, we observe that D-INT attains optimal power.

For the skewed χ^2_1 phenotype, I-INT was most powerful, achieving over 6 times the efficiency of the UAT, and over twice the efficiency of D-INT. For the kurtotic t_3 phenotype, the efficiency gain provided by the INT-based tests were more modest yet still noteworthy, at 53% for the I-INT and 38% for D-INT. By synthesizing D-INT and I-INT, the omnibus approach (O-INT) aims to provide a test that is well powered regardless of the phenotype's generative model. The power and relative efficiency of O-INT were intermediate between those of D-INT and I-INT. Since O-INT is a compromise between two complementary methods, it cannot be expected to attain optimal efficiency. However, for all phenotypes, the efficiency of O-INT was within 2% of the most efficient test considered. Thus, O-INT achieves robustness to the underlying data generating mechanism with little to no loss of efficiency.

Run Time Analysis

We compared the computational complexity of INT-based test with UAT, which has the same computational cost as standard linear regression. Figure S4 plots the median CPU time for performing 10^2 association tests, across 10^3 simulation replicates, on log-log scale. The run times for the UAT and D-INT are nearly identical, the only additional cost of the latter begin the time required for the INT itself. The run time for I-INT, which must perform two linear regressions, is slightly increased, at approximately $1.8\times$ that of UAT. Finally, the run time of O-INT, which applies both D-INT and I-INT internally, is approximately $3.0\times$ that of UAT. Importantly, however, the run time for all association tests implemented in the `RNOmni` package scales linearly with sample size, which is the same rate as standard linear regression. Consequently, it is feasible to apply these methods for association testing in cohorts on the scale of UK Biobank. For example, applying O-INT to 10^6 loci in a cohort of $N = 5 \times 10^5$ subjects required around 28 CPU hours. `RNOmni` provides the option of distributing the run time across cores working in parallel.

Spirometry GWAS in UK Biobank Data

We performed GWAS of the following spirometry traits using subjects from the UK Biobank: forced expiratory volume (FEV1), forced vital capacity (FVC), the ratio of FEV1 to FVC (FEV1/FVC), and the logarithm of peak expiratory flow (lnPEF). The regression model adjusted for an intercept, age, sex, BMI, two orthogonal polynomials in height, genotyping array, and 20 genetic PCs. By conditioning on the heritable covariates BMI and height, our analyses estimate the direct effect of genotype on the spirometry trait, excluding any contribution mediated by BMI or height. Results from the INT-based tests are compared with those from the UAT, in the overall sample ($N = 291,729$) and in the subgroup of asthmatic subjects ($N = 28,589$).

Each spirometry trait exhibited residual non-normality after regression on covariates. QQ plots comparing the residuals with the standard normal distribution are shown in Figure S5. For FEV1, the central 98% of the residual distribution was approximately normal. However, there was an excess of positive outliers (kurtosis: 23.4). These subjects had higher FEV1 than

expected under normal residual variation. FVC was tightly correlated with FEV1 (Spearman's rho: 0.94). Consequently, the residual distribution of FVC mirrored that of FEV1, although the kurtosis was more extreme (kurtosis: 278). In contrast to the component measurements, the FEV1/FVC residuals were negatively skewed (skewness: -1.5), with primarily negative outliers. These subjects had lower FEV1, relative to their FVC, than expected under normal residual variation. The residual distribution of lnPEF was least normally distributed. Although lnPEF was only modestly correlated with FEV1/FVC (Spearman's rho: 0.24), the residuals distributions deviated from normality in qualitatively similar ways. The lnPEF residuals exhibited negative skew (skewness: -4.6), and excess kurtosis (kurtosis: 22.6).

Distribution of Association p-values

The QQ plots for the association p-values from GWAS of the spirometry are presented in Figure 3. These p-values were obtained by applying UAT and the INT-based association tests across 360,761 directly genotyped, common variants having a sample minor allele frequency of at least 5%. In all cases, the observed distribution of p-values differs dramatically from the uniform. Candidate explanations for the excess of small p-values include: (a) residual confounding of the genetic effect due to population structure; (b) failure of the testing procedure to control the type I error for the observed residual distributions; and (c) true polygenicity, or the presence of many non-zero genetic effects. In the following, we scrutinize these explanations in turn, demonstrating that neither population structure nor excess type I error provide satisfactory explanations for the apparent inflation of p-values.

Scatter plots comparing the first four genetic PCs of subjects included in the overall spirometry GWAS are presented in Figure S6. Each PC has been centered and scaled to have mean zero and unit variance. Separation of the subjects into distinct clusters would suggest the presence of related subgroups within the sample. Confounding would arise if these subgroups of genetically similar subjects shared common environmental exposures that influenced their spirometry traits. However, as expected from the ancestral homogeneity of the study population, there is no visual evidence for population stratification.

The intercept from LD score regression (LDSC) assesses inflation of the association test statistics due to confounding bias[9]. Briefly, in LDSC the test statistic for each locus is regressed

on a local measure of linkage disequilibrium. An intercept exceeding one suggests inflation, whereas an intercept falling below one suggests deflation. The results from applying LDSC in the overall sample are presented in Table 2, and the results for the asthmatic subgroup are presented in Table S1. In the asthmatic subgroup, the UAT had intercepts near the reference value of one, while the INT-based tests exhibited slight deflation. In the overall sample, all tests exhibited significant deflation. Generally, the deflation of UAT and D-INT was comparable, while that of I-INT was most pronounced. Overall, LDSC provided no evidence for inflation of the test statistics due to confounding bias, and in fact suggested that our analyses were somewhat conservative. Despite of slight deflation of INT-based analyses, as shown in the results below, these tests were substantially more powerful than the UAT.

Since we found no evidence of confounding due to population structure, under the null hypothesis of no genetic effects, the marginal distribution of each spirometry trait is independent of genotype. To validate nominal control of the type I error for the observed residual distributions, we repeated the association analyses using permuted genotypes. By permuting genotype, rather than the phenotype, the association between genotype and the spirometry traits is broken, while the associations between covariates and the the spirometry traits are preserved. Although permuting genotype inadvertently breaks the association between genotype and covariates, regression analyses are performed *conditional on* both genotype and covariates. Consequently, valid inference on the association between genotype at the phenotype does not depend on what association, if any, exists between the genotype and covariates. Another advantage of permuting genotypes is that the residual distributions presented in Figure S5 are preserved, since each subject retains their original covariates, including their genetic PCs. Testing for association between permuted genotype and the observed spirometry traits allows one to assess whether the type I error was maintained for the observed residual distributions. Uniform QQ plots for the association p-values from this analysis are presented in Figure 4. After permuting genotypes, the p-values are uniformly distributed, demonstrating that our testing procedure maintains the type I error under imposition of the null hypothesis.

Having eliminated (a) confounding due to population structure and (b) excess type I error as candidate explanations for the observed inflation in Figure 3, we conclude that true polygenicity

is a likely cause. Moreover, due to the large sample size, our analysis is well powered to detect genetic variants with small effect sizes. The presence of many, predominantly small genetic effects is expected to manifest as uniform QQ plots similar to those in Figure 3.

Empirical Discovery and Efficiency Gains

Association tests were conducted for 360,761 directly genotyped, common loci. Table 3 presents the average χ_1^2 statistics across all loci that reached genome-wide significance ($\alpha = 5 \times 10^{-8}$) in the overall sample according to at least one the association method. Average χ_1^2 for the asthmatic subgroup are presented in Table S2. The empirical *efficiency gain* (O-INT v. UAT) was defined as the ratio of the non-centrality parameters minus one, where the non-centrality parameters were estimated using loci that reached significance according to at least one association method:

$$\text{Efficiency Gain} = \frac{\left(\bar{\chi}^2_{1,\text{O-INT}} - 1\right)}{\left(\bar{\chi}^2_{1,\text{UAT}} - 1\right)} - 1.$$

In all cases the average χ_1^2 statistic of O-INT exceeded that of UAT. In the overall sample, the empirical efficiency gain ranged from 12% for FEV1 to 278% for lnPEF. In the asthmatic subgroup, the empirical efficiency gain ranged from 6% for the FEV1/FVC ratio to 222% for lnPEF. The estimates from the overall analysis are likely more stable, as comparatively fewer discoveries were made in the asthmatic subgroup.

Table 3 also presents the counts of genome-wide significant associations ($\alpha = 5 \times 10^{-8}$) in the overall sample after LD 'clumping,' to reduce redundant signals, within a 1000 kb window at an r^2 threshold of 0.2. Counts for the asthmatic subgroup are presented in Table S3. The empirical discovery gain was defined as:

$$\text{Discovery Gain} = \frac{n_{\text{OINT}} - n_{\text{UAT}}}{n_{\text{OINT}\cup\text{UAT}}},$$

where n_{OINT} is the number of associations identified by O-INT only, n_{UAT} is the number of associations identified by UAT only, if any, and $n_{\text{OINT}\cup\text{UAT}}$ is the number of associations identified by either O-INT or UAT. In all cases, O-INT discovered more independent (at $r^2 = 0.2$) associations with the target phenotype than UAT. In the overall sample, the empirical discovery gain ranged from 15% for FEV1 to 79 % for lnPEF. In the asthmatic subgroup, the

empirical discovery gain ranged from 12% for the FEV1/FVC ratio to 100% for FVC and lnPEF. For the latter two traits, UAT was unable to identify any genome-wide significant associations in the smaller asthmatic subgroup.

The discovery gains were more dramatic for those traits whose residuals were less normally distributed. However, the INT-based tests were uniformly more powerful than the UAT. Consistent with the simulations, the power of O-INT was always intermediate between that of D-INT and I-INT. For a given trait, the number of discoveries by O-INT was generally closer to the number of discoveries by the more powerful of D-INT and I-INT. In the following, we focus on comparing our proposed omnibus test O-INT with the standard UAT.

Manhattan Plots

Manhattan plots of the association p-values from O-INT, prior to LD clumping, are presented in Figure 5. We subset autosomal associations reported in the NHGRI GWAS Catalogue[29] to those whose trait description included at least one of the following key terms: apnea, airway, asthma, broncho-, COPD, cough, cystic fibrosis, emphysema, expiratory, FEV1, forced, FVC, immun-, influenza, oxygen, PEF, pneumo-, pressure, pulmonary, respiratory, sleep, smoking, spiro-, thrombo-, or tuberculosis. Among the 360,761 loci included in our analyses, 1,393 (0.4%) had previously been associated with a lung-related trait. Genome-wide significant variants previously associated with lung traits are highlighted in red on the Manhattan plots. Many of the newly reported associations from our analyses are interspersed among, and physical adjacent to, the previously reported loci, suggesting that they may tag the same underlying, causal variants. Among all loci included in our analyses that have previously been associated with some trait in the GWAS Catalogue (12,044 variants), the lung-related variants were significantly enriched among loci that reached genome-wide significance according to O-INT. The odds ratios and Fisher exact tests for enrichment were 3.5 ($p = 2.2 \times 10^{-25}$) for FEV1, 2.8 ($p = 3.5 \times 10^{-13}$) for FVC, 3.7 ($p = 1.6 \times 10^{-35}$) for FEV1/FVC, and 2.5 ($p = 8.3 \times 10^{-11}$) for lnPEF.

Biological Findings

FEV1

In the asthmatic subgroup, O-INT identified 5 significant associations with FEV1, while UAT

identified 3 significant associations. The discovery gain by O-INT was 40%. Among the 5 loci discovered by O-INT, 1 had previously been associated with a lung-related trait. The top association was rs6823809 ($p = 3.6 \times 10^{-10}$), an intronic variant of the NPNT protein coding gene on chromosome 4. This locus is 365 kb from rs10516541, which has previously been associated with FVC[40]. In the overall analysis, O-INT identified 398 significant associations, while UAT identified 331. The discovery gain by O-INT was 15%. Among the 398 loci discovered by O-INT, 36 had previously been associated with lung-related traits. The top association was rs34712979 ($p = 6.0 \times 10^{-60}$), a splice acceptor variant of the NPNT protein coding gene on chromosome 4. This variant has previously been associated with lung function[39], and respiratory disease[23].

FVC

In the asthmatic subgroup, O-INT identified 4 significant associations with FVC, while UAT identified none. The discovery gain by O-INT was therefore 100%. Among the 4 loci discovered by O-INT, 1 had previously been associated with a lung-related trait. The top association was rs4931726 ($p = 4.9 \times 10^{-11}$), an intronic variant of the LOC107984462 non-coding RNA on chromosome 12. This locus is 150 kb from rs191000678, which has previously been associated with FEV1[27]. In the overall analysis, O-INT identified 364 significant associations, while UAT identified 213. The discovery gain by O-INT was 38%. Among the 364 loci discovered by O-INT, 36 had previously been associated with lung-related traits. The top association was rs2169752 ($p = 1.8 \times 10^{-41}$), an intronic variant of the CCDC91 protein coding gene on chromosome 12. This locus is 43 kb from rs7977418, which has previously been associated with lung function[23].

FEV1/FVC

In the asthmatic subgroup, O-INT identified 16 significant associations with FEV1/FVC, while UAT identified 14 significant associations. The discovery gain by O-INT was 12%. Among the 16 loci discovered by O-INT, 7 had previously been associated with lung-related traits. The top association was rs6537293 ($p = 1.7 \times 10^{-23}$), an intronic variant of the LOC105377462 non-coding RNA on chromosome 4. This locus has previously been associated with the FEV1/FVC ratio[27]. In the overall analysis, O-INT identified 652 associations, while UAT identified 450. The discovery gain was 28%. Among the 652 loci discovered by O-INT, 58 had previously been associated with lung-related traits. The top association was rs2070600 ($p = 5.0 \times 10^{-160}$), a

missense variant in the AGER protein coding gene. This locus has previously been associated with COPD[19] and lung function[18;33;39].

lnPEF

In the asthmatic subgroup, O-INT identified 3 significant associations with lnPEF, while UAT identified none. The discovery gain by O-INT was thus 100%. Among the 3 loci discovered by O-INT, all had previously been associated with lung-related traits. The top association was rs6829956 ($p = 2.2 \times 10^{-13}$), another intronic variant of the LOC105377462 non-coding RNA on chromosome 4. This locus has previously been associated with the FEV1/FVC ratio[27]. In the overall analysis, O-INT identified 270 significant associations, while UAT identified only 39. The discovery gain was 79%. Among the 270 loci identified by O-INT, 21 had previously been associated with been associated with lung-related traits. The top association was again rs6829956 ($p = 7.5 \times 10^{-90}$). The second ranked association was rs13118083, an intronic variant of the LOC105377462 non-coding RNA on chromosome 4. This locus was also associated with the FEV1/FVC ratio[27].

Discussion

In this paper, we systematically investigated the utility of different INT-based association tests for GWAS of quantitative traits with non-normally distributed residuals. We formally presented the Direct (D-INT) and Indirect (I-INT) INT-based tests, and studied their assumptions, underlying data generative models, properties and connections. Neither D-INT nor I-INT was uniformly most powerful. D-INT is most effective when the observed trait could have arisen from a rank-preserving transformation of a latent normal trait. I-INT is most effective when the observed trait is linear in covariates, but the residual distribution is skewed or kurtotic. For non-normally distributed quantitative traits, both D-INT and I-INT provide better control of the type I error and higher power than the untransformed association test (UAT). We proposed an omnibus INT-based omnibus test (O-INT) as a robust, computationally efficient, and statistically powerful approach to GWAS of quantitative traits with non-normally distributed residuals. All INT-based tests (D-INT, I-INT, O-INT) have been implemented in the R package `RNOmni`, which is available on CRAN. We further demonstrated the utility of INT-based

association tests in GWAS of spirometry traits from the UK Biobank.

In D-INT, the INT is applied directly to the phenotype, and the transformed phenotype is regressed on genotype and covariates. Under the strong null hypothesis of no genetic or covariate effects, D-INT is asymptotically exact, and when the effect sizes are small, D-INT holds approximately. I-INT is a two-stage procedure. Different variants of I-INT have been considered in the literature. In all approaches, the phenotype is first regressed on covariates to obtain residuals. In the second stage, the INT-transformed residuals are regressed on genotype, with or without a secondary adjustment for genetic PCs. To provide guidance on which approach to use in practice, we formally derived I-INT, starting from the assumption that the observed phenotype follows a linear regression model with a non-normally distributed residual. Our derivations indicate that, during the second stage of I-INT, the transformed phenotypic residuals should be regressed on genotypic residuals, which are the residuals obtained by regressing genotype on covariates. This second stage is equivalent to regressing the INT-transformed phenotypic residuals on genotype while performing a secondary adjustment for covariates. Therefore, all covariates, including genetic PCs, should be included in both the first and second stage regressions. Under the standard null hypothesis of no genetic effects, I-INT is asymptotically exact, and holds approximately when effect sizes are small. Under the strong null hypothesis of neither genetic nor covariate effects, D-INT and I-INT are equivalent.

We conducted extensive simulations to evaluate the operating operating characteristics of these tests across a range of scenarios. For non-normally distributed quantitative traits, both D-INT and I-INT provide better control of the type I error and higher power than UAT. For traits with normally distributed residuals, the UAT is theoretically optimal, yet the INT-based tests were fully efficient. For traits with non-normally distributed residuals, the INT-based tests routinely controlled the type I error. In contrast, the UAT failed to control the type I error in moderate samples. Although better control was achieved as sample size increased, the INT-based tests still dominated the UAT in terms of power. The power gain was most dramatic for phenotypes with skewed residuals, yet was still significant for phenotypes with symmetric but kurtotic residuals.

Since neither the direct nor indirect approaches was uniformly most powerful, the proposed

an adaptive, INT-based omnibus test that synthesizes D-INT and I-INT. For non-normally distributed traits, O-INT robustly controls the type I error and increases power. Further, O-INT is nearly as powerful as the more effective INT-based test when the true data generating mechanism is known. Run time analysis demonstrated that the computational complexity of O-INT increased linearly in sample size, which is the same rate as standard linear regression. Thus O-INT is applicable in large sample settings. GWAS analysis of spirometry traits from the UK Biobank empirically demonstrated that O-INT is more powerful than standard procedures.

In this paper we focused on GWAS with independent samples. However, INT-based tests can be extended to the correlated data setting using linear mixed models[14;21;26]. We plan to develop INT-based tests for linear mixed models in which correlation across related subjects is modeled using random effects whose covariance pattern depends on the between-subject genetic relatedness matrix[14;21;26]. A similar modeling strategy can accommodate longitudinal phenotypes arising from repeated measurements on the same subjects across time[13].

Supplemental Data

Supplemental data include six figures and two tables.

Acknowledgments

This work was supported by grants R35 CA197449 and F31 HL140822 (to Z. M.), grants R35 CA197449, P01-CA134294, U01-HG009088, U19-CA203654, and R01-HL113338 (to X. L.), and R35 HL135818 (to S. R.).

Declaration of Interests

The authors declare no competing interests.

Web Resources

The association tests described in this paper (D-INT, I-INT, O-INT) are available in the R package on Github: `https://github.com/zrmacc/RNOmni` and on CRAN: `https://cran.r-project.org/web/packages/RNOmni/index.html`. Links to this and related software may also be found at `https://content.sph.harvard.edu/xlin/software.html`.

References

[1] Abbott, L., Bryant, S., Churchhouse, C., Ganna, A., Howrigan, D., et al. (2017). http://www.nealelab.is/uk-biobank/.

[2] Ahn, J., Yu, K., Stolzenberg-Solomon, R., Simon, K., McCullough, M., et al. (2010). Genome-wide association study of circulating vitamin d levels. *Human Molecular Genetics*, 19(13):2739–45.

[3] Allen, N. E., Sudlow, C., Peakman, T., Collins, R., et al. (2014). Uk biobank data: come and get it.

[4] Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781.

[5] Barber, M., Mangravite, L., Hyde, C., Chasman, D., Smith, J., et al. (2010). Genome-wide association of lipid-lowering response to statins in combined study populations. *PLOS One*, 5(2):e9763.

[6] Beasley, T., Erickson, S., and Allison, D. (2009). Rank-based inverse normal transformations are increasingly used, but are they merited? *Behavioral Genetics*, 39(5):580–95.

[7] Bhatnagar, P., Purvis, S., Barron-Casella, E., DeBaun, M., Casella, J., et al. (2011). Genome-wide association study identifies genetic variants influencing f-cell levels in sickle-cell patients. *Journal of Human Genetics*, 56(6):316–23.

[8] Blekhman, R., Goodrich, J., Huang, K., Sun, Q., Bukowski, R., et al. (2015). Host genetic variation impacts microbiome composition across human body sites. *Genome Biology*, 16:191.

[9] Bulik-Sullivan, B., Loh, P., Finucane, H., Ripke, S., Yang, J., et al. (2015). Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295.

[10] Buzkova, P. (2013). Linear regression in genetic association studies. *PLOS One*, 8(2):e56976.

[11] Cade, B., Chen, H., Stilp, A., Gleason, K., Sofer, T., et al. (2016). Genetic associations with obstructive sleep apnea traits in hispanic/latino americans. *American Journal of Respiratory and Critical Care Medicine*, 194(7):886–897.

[12] Casella, B. and Berger, R. (2002). *Statistical Inference.* Duxbury/Thomson Learning, Pacific Grove, CA, 2 edition.

[13] Chen, H., Huffman, J., Brody, J., et al. (2019). Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. *American Journal of Human Genetics*, 104(2):260–274.

[14] Chen, H., Wang, C., Conomos, M., et al. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *American Journal of Human Genetics*, 98(4):653–666.

[15] Consortium, G. et al. (2017). Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213.

[16] Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A., Zink, F., et al. (2008). Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423–428.

[17] Fusi, N., Lippert, C., Lawrence, N., and Stegle, O. (2014). Warped linear mixed models for the genetic analysis of transformed phenotypes. *Nature Communications*, 5:4890.

[18] Hancock, D., Soler Artigas, M., Gharib, S., Henry, A., Manichaikul, A., et al. (2012). Genome-wide joint meta-analysis of snp and snp-by-smoking interaction identifies novel loci for pulmonary function. *PLOS Genetics*, 8(12):e1003098.

[19] Hobbs, B., de Jong, K., Lamontagne, M., Bosse, Y., Shrine, N., et al. (2017). Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nature Genetics*, 49(3):426–432.

[20] J, T. (1962). The future of data analysis. *Annals of Mathematical Statistics*, 33(1):1–67.

[21] Kang, H., Sul, J., Service, S., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354.

[22] Kettunen, J., Tukiainen, T., Sarin, A., Ortega-Alonso, A., Tikkanen, E., et al. (2012). Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature Genetics*, 44(3):269–76.

[23] Kichaev, G., Bhatia, G., Loh, P., et al. (2019). Leveraging polygenic functional enrichment to improve gwas power. *The American Journal of Human Genetics*, 104(1):65–75.

[24] Liu, Y., Chen, S., Li, Z., Morrison, A. C., Boerwinkle, E., and Lin, X. (2019). Acat: A fast and powerful p value combination method for rare-variant analysis in sequencing studies. *The American Journal of Human Genetics*, 104(3):410–421.

[25] Liu, Y. and Xie, J. (2019). Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*.

[26] Loh, P., Tucker, G., Bulik-Sullivan, B., et al. (2015). Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47(3):284–290.

[27] Lutz, S., Cho, M., Young, K., Hersh, C., Castaldi, P., et al. (2015). A genome-wide association study identifies risk loci for spirometric measures among smokers of european and african ancestry. *BMC Genetics*, 16:138.

[28] Ma, L., Yang, J., Runesha, H., Tanaka, T., Ferrucci, L., et al. (2010). Genome-wide association analysis of total cholesterol and high-density lipoprotein cholesterol levels using the framingham heart study data. *BMC Medical Genetics*, 11:55.

[29] MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., et al. (2017). The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic Acids Research*, 45(D1):D896–D901.

[30] Okada, Y., Kamatani, Y., Takahashi, A., Matsuda, K., Hosono, N., et al. (2010). Common variations in psmd3-csf3 and plcb4 are associated with neutrophil count. *Human Molecular Genetics*, 19(10):2079–2085.

[31] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575.

[32] Reed, E., Nunez, S., Kulp, D., Qian, J., Reilly, M. P., and Foulkes, A. S. (2015). A guide to genome-wide association analysis and post-analytic interrogation. *Statistics in Medicine*, 34(28):3769–3792.

[33] Repapi, E., Sayers, I., Wain, L., Burton, P., Johnson, T., et al. (2010). Genome-wide association study identifies five loci associated with lung function. *Nature Genetics*, 42(1):36–44.

[34] Scuteri, A., Sanna, S., Chen, W., Uda, M., Albai, G., et al. (2007). Genome-wide association scan shows genetic variants in the fto gene are associated with obesity-related traits. *PLOS Genetics*, 3(7):e115.

[35] Sofar, T., Zheng, X., Gogarten, S., Laurie, C., Grinde, K., et al. (2019). A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genetic Epipemiology*, pages 1–13.

[36] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., et al. (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):e1001779.

[37] Team, R. C. (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[38] Varenhorst, C., Eriksson, N., Johansson, A., Barratt, B., Hagstrom, E., et al. (2015). Effect of genetic variations on ticagrelor plasma levels and clinical outcomes. *European Heart Journal*, 36(29):1901–1912.

[39] Wain, L., Shrine, N., Artigas, M., Erzurumluoglu, A., Noyvert, B., et al. (2017). Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. *Nature Genetics*, 49(3):416–425.

[40] Wilk, J., Shrine, N., Loehr, L., Zhao, J., Manichaikul, A., et al. (2012). Genome-wide association studies identify chrna5/3 and htr4 in the development of airflow obstruction. *American Journal of Respiratory and Critical Care Medicine*, 186(7):622–32.

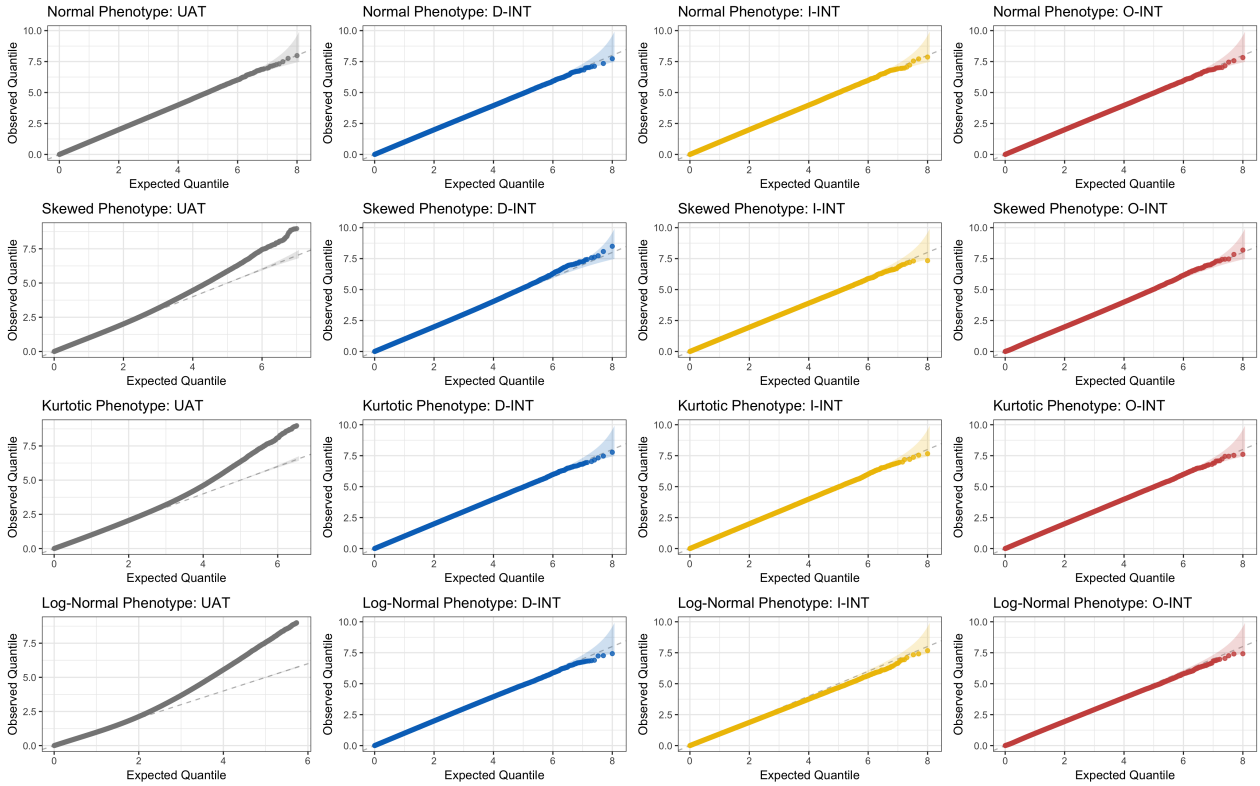

Figure 1: Simulation Studies of the Type I Error Rate at Sample Size $N = 10^3$ across $R = 10^8$ Simulation Replicates. Rows correspond to different phenotype distributions. The first is normal, the second skewed (χ^2_1), the third kurtotic (t_3), and the fourth log-normal. Columns correspond to different association tests. The first is the untransformed association test (UAT), the second is the direct inverse normal transformation (D-INT), the third is indirect INT (I-INT), and the fourth column is omnibus INT (O-INT).

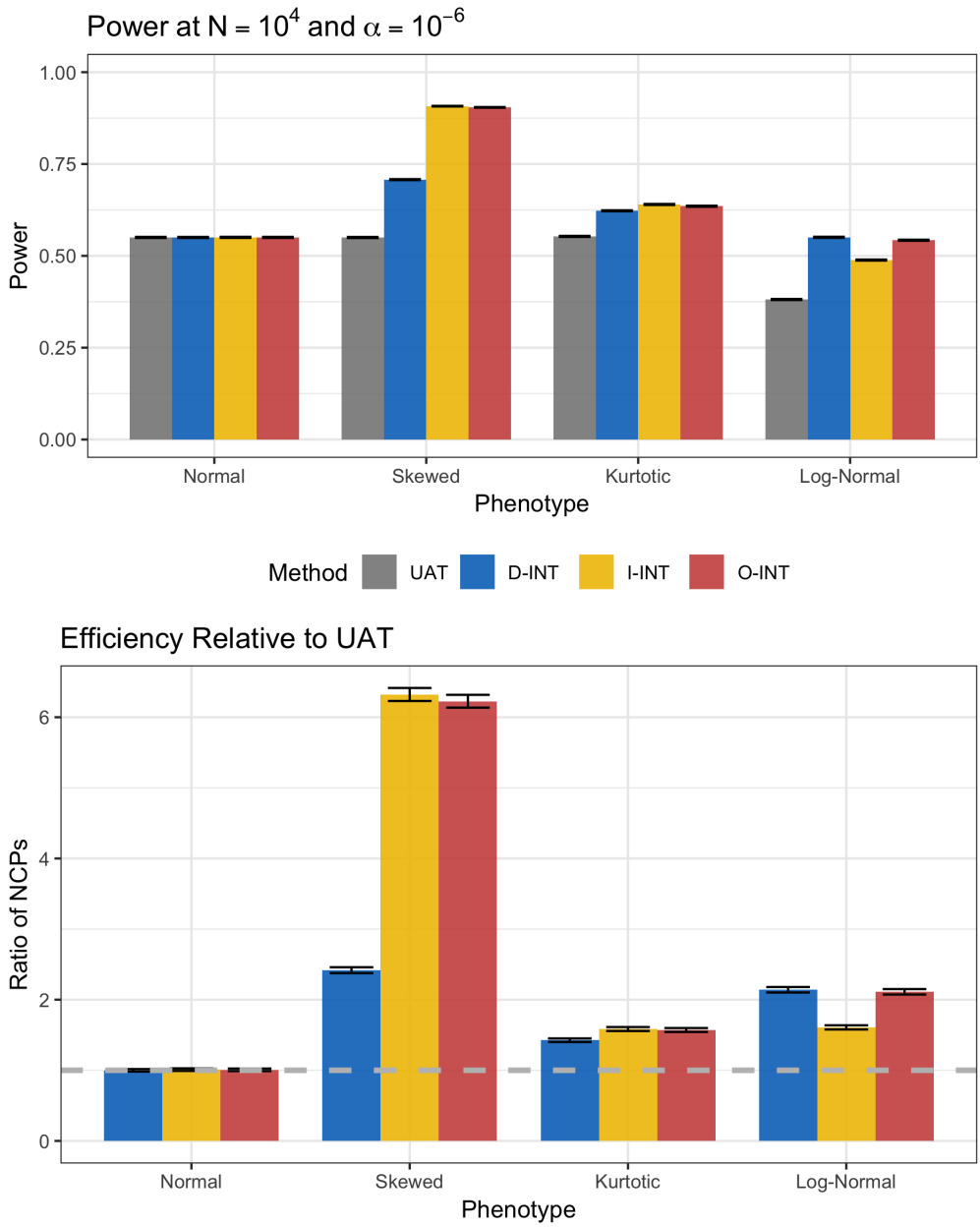

Figure 2: Simulation Studies of Power and Relative Efficiency at Sample Size $N = 10^4$ across $R = 10^6$ Simulation Replicates. The proportion of phenotypic variation explained by genotype is 0.5%. The upper panel is power, the proportion of null hypotheses correctly rejected at $\alpha = 10^{-6}$. The lower panel is relative efficiency, as the ratio of χ^2_1 non-centrality parameters, which is independent of sample size and α level. Gray is the UAT, blue the D-INT, yellow the I-INT, red the O-INT. Within figures, from left to right, the first phenotype has normal residuals; the second phenotype has χ^2_1; the third phenotype has t_3 residua; the log of the fourth phenotype has normal residuals.

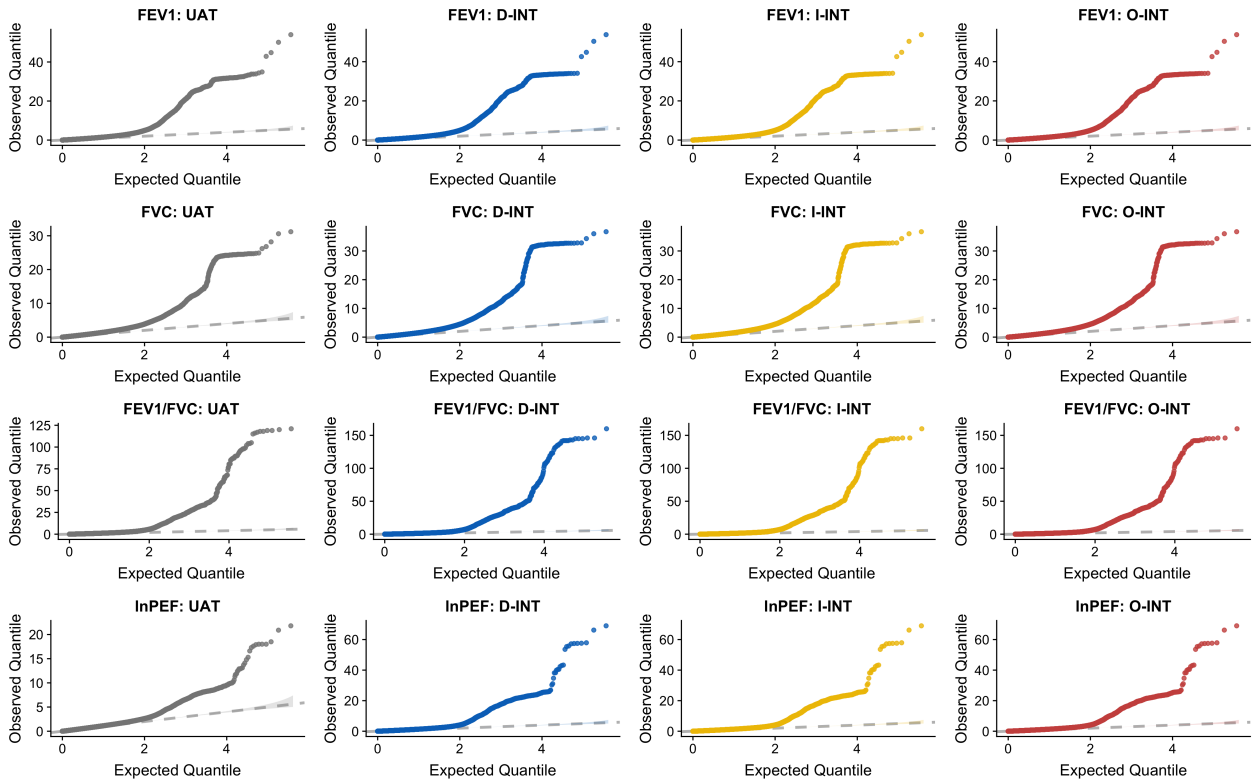

Figure 3: Distribution of p-values from Association Testing with Lung Function in the UK Biobank (N = 292K). Genotypes at 361K loci with common minor alleles were included. First column is the untransformed association test (UAT); second column is the direct inverse normal transformation (D-INT); third column is indirect INT (I-INT); fourth column is omnibus INT (O-INT). First row is forced expiratory volume (FEV1); second row is forced vital capacity (FVC); third row is the FEV1/FVC ratio; fourth row is the logarithm of peak expiratory flow (lnPEF).

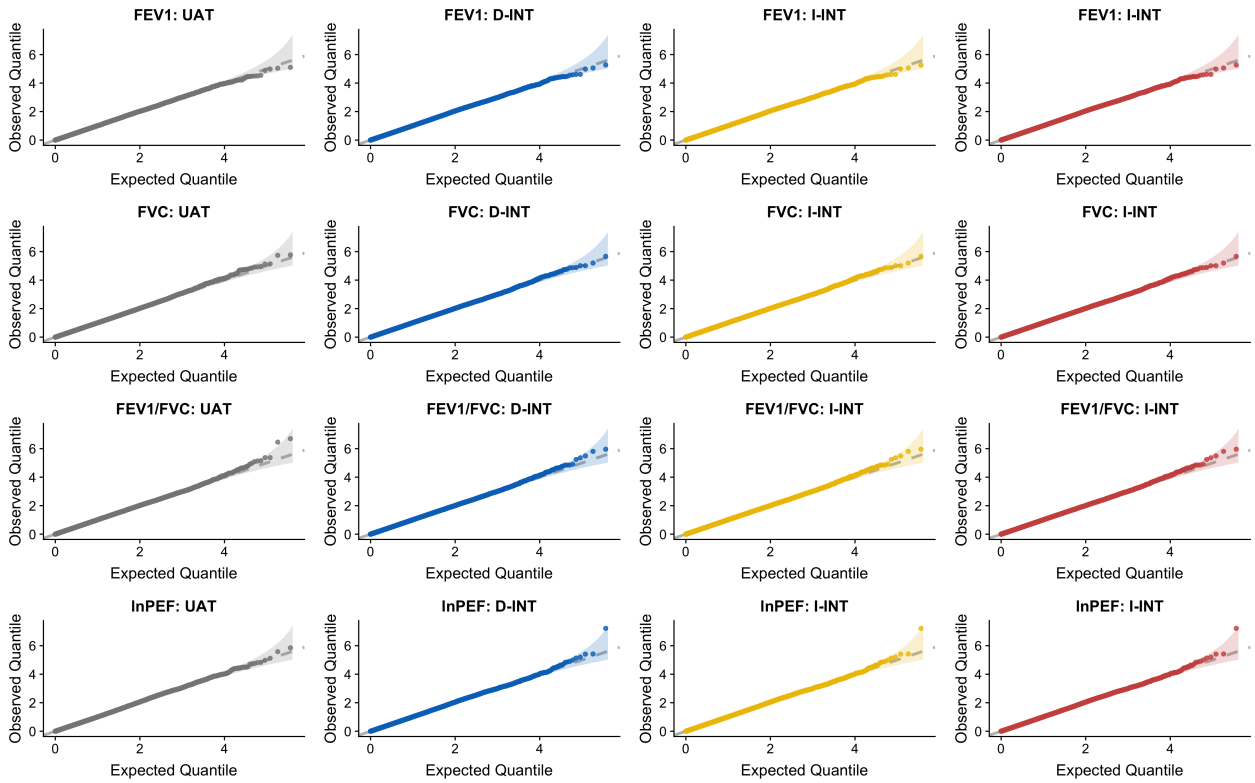

Figure 4: Distribution of p-values from Association Testing with Lung Function in the UK Biobank after Permuting Genotypes. The sample size was $N = 292$K. Genotypes at 361K loci with common minor alleles were included. First column is the untransformed association test (UAT); second column is the direct inverse normal transformation (D-INT); third column is indirect INT (I-INT); fourth column is omnibus INT (O-INT). First row is forced expiratory volume (FEV1); second row is forced vital capacity (FVC); third row is the FEV1/FVC ratio; fourth row is the logarithm of peak expiratory flow (lnPEF).

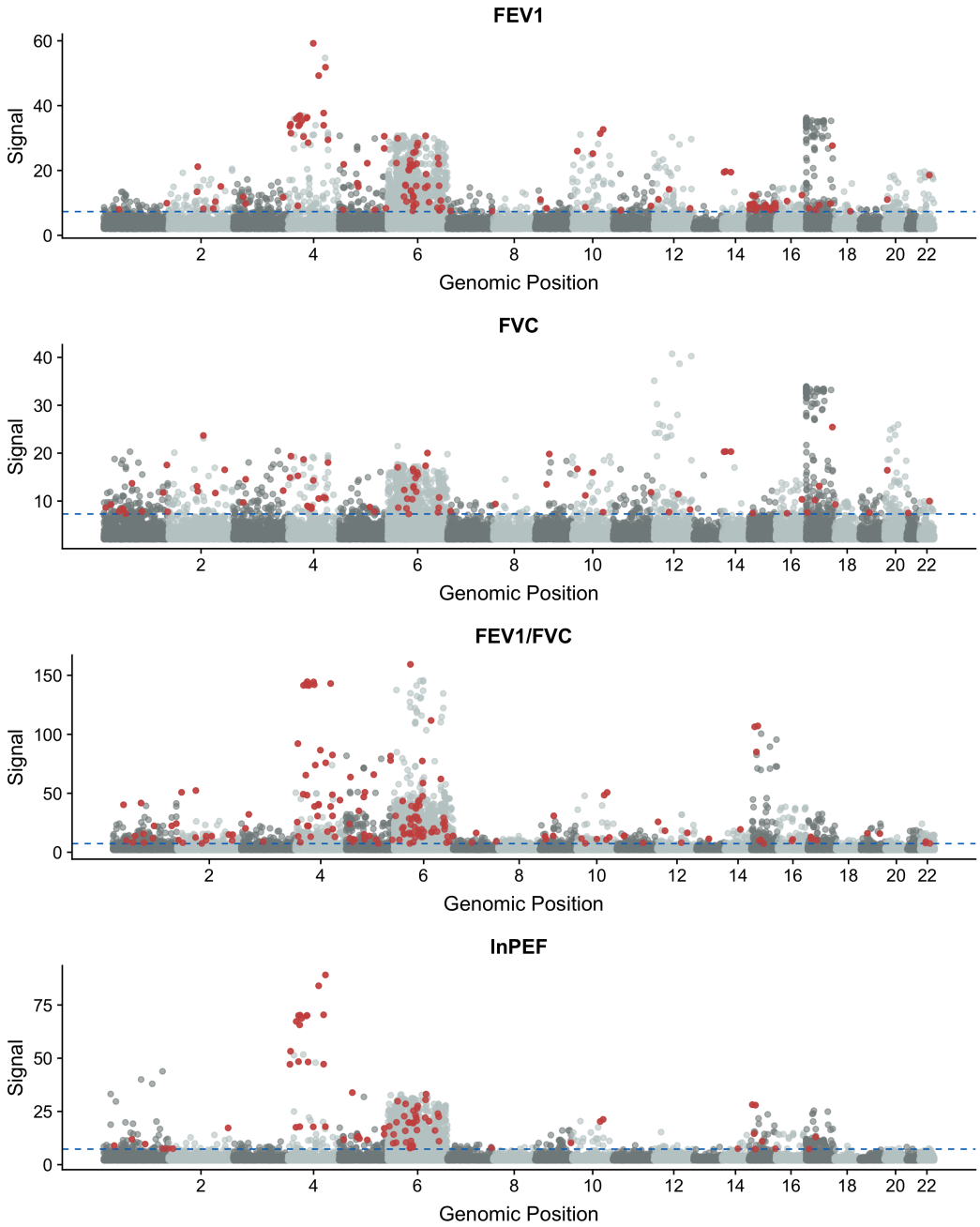

Figure 5: Manhattan Plots of p-values from Association Testing with Lung Function in the UK Biobank ($N = 292$K). Highlighted in red are significant associations that have previously been associated with lung-related traits in the GWAS Catalogue. First row is forced expiatory volume (FEV1); second row is forced vital capacity (FVC); third row is the FEV1/FVC ratio; fourth row is the logarithm of peak expiatory flow (lnPEF). Chromosomes are alternately colored, with odds in dark gray, and evens in light gray. The dashed line demarcates the threshold for genome-wide significance.

Table 1: Simulation Studies of the Empirical Size ($\times 10^6$) at $\alpha = 10^{-6}$ across $R = 10^8$ Simulation Replicates. Size simulations were conducted under the null hypothesis of no genetic effects. Each association test was applied to each phenotype at sample sizes ranging from $N = 10^3$ to $N = 10^5$. The phenotypic means depended on an intercept, age, sex, and 3 genetic principal components.

		Sample Size		
Phenotype	Test	$N = 10^3$	$N = 10^4$	$N = 10^5$
Normal	UAT	1.04	0.93	1.03
Normal	D-INT	0.84	0.87	1.02
Normal	I-INT	0.97	0.93	1.00
Normal	O-INT	0.91	0.93	0.99
Skewed	UAT	8.03	1.87	1.43
Skewed	D-INT	1.20	1.10	1.05
Skewed	I-INT	0.67	0.84	0.89
Skewed	O-INT	1.10	1.01	0.98
Kurtotic	UAT	15.89	5.54	3.12
Kurtotic	D-INT	0.94	0.91	0.95
Kurtotic	I-INT	1.00	0.88	1.00
Kurtotic	O-INT	0.96	0.90	0.97
Log-Normal	UAT	59.34	11.01	7.52
Log-Normal	D-INT	0.74	1.02	1.02
Log-Normal	I-INT	0.40	0.40	0.43
Log-Normal	O-INT	0.55	0.74	0.76

Table 2: LD Score Regression (LDSC) Analysis of Z-scores form Association Testing with Lung Function in the UK Biobank ($N = 292$K). LD scores were calculated within the UK Biobank, and LDSC was performed using summary statistics from association testing at all 361K loci with common minor alleles. An LDSC intercept significantly exceeding 1 is indicative of residual confounding. An estimate of the additive heritability h^2 explained by loci in the analysis is also provided.

Trait	Test	Intercept	SE(Int)	h^2	SE(h^2)
FEV1	UAT	0.84	0.02	0.24	0.01
FEV1	D-INT	0.82	0.02	0.25	0.01
FEV1	I-INT	0.81	0.02	0.27	0.01
FEV1	O-INT	0.81	0.02	0.27	0.01
FVC	UAT	0.86	0.02	0.21	0.01
FVC	D-INT	0.83	0.02	0.25	0.01
FVC	I-INT	0.81	0.02	0.28	0.01
FVC	O-INT	0.81	0.02	0.27	0.01
FEV1/FVC	UAT	0.86	0.02	0.22	0.01
FEV1/FVC	D-INT	0.86	0.02	0.27	0.01
FEV1/FVC	I-INT	0.84	0.02	0.28	0.01
FEV1/FVC	O-INT	0.84	0.02	0.28	0.01
lnPEF	UAT	0.95	0.01	0.06	0.004
lnPEF	D-INT	0.88	0.02	0.16	0.01
lnPEF	I-INT	0.86	0.02	0.18	0.01
lnPEF	O-INT	0.86	0.02	0.18	0.01

Table 3: Empirical Efficiency and Discovery Gains for Lung Function GWAS in the UK Biobank (N = 292K). Genome-wide significance was declared at $\alpha = 5 \times 10^{-8}$. The average χ_1^2 statistics are reported across all loci detected by at least one of the association tests. The empirical efficiency gain, comparing O-INT with UAT, the ratio of the estimated χ_1^2 non-centrality parameters minus 1. The counts of significant associations are reported after LD clumping within 1000 kb radii at $r^2 = 0.2$ to remove redundant signals. The discovery gain, comparing O-INT with UAT, is the ratio of the number of associations uniquely identified by O-INT to the total number of associations detected.

Average χ_1^2					
Trait	UAT	D-INT	I-INT	O-INT	Efficiency Gain (%)
FEV1	56.77	57.55	64.69	63.59	12
FVC	37.14	46.91	51.47	50.59	37
FEV1/FVC	63.21	83.76	83.00	83.63	33
lnPEF	17.72	52.44	65.40	64.11	278
Significant Associations					
Trait	UAT	D-INT	I-INT	O-INT	Discovery Gain (%)
FEV1	331	352	422	398	15
FVC	213	323	375	364	38
FEV1/FVC	450	653	649	652	28
lnPEF	39	202	270	251	79