Supporting Information

FOR

USAT: A Unified Score-based Association Test for Multiple Phenotype-Genotype Analysis

BY

Debashree Ray[1], James S. Pankow[2], Saonli Basu[1]

[1]*Division of Biostatistics, School of Public Health, University of Minnesota, U.S.A.*

[2]*Division of Epidemiology & Community Health, School of Public Health, University of Minnesota, U.S.A.*

Appendix S1

Proof of Theorem 1

Without loss of generality, we assume that $\boldsymbol{Y}$ and $\boldsymbol{X}$ are centered. For testing $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$, the Wilk's Lambda test statistic is $\det\boldsymbol{E}/\det(\boldsymbol{H}+\boldsymbol{E}) = \det(\frac{1}{n}\boldsymbol{E})/\det(\frac{1}{n}\boldsymbol{H}+\frac{1}{n}\boldsymbol{E})$, where $\boldsymbol{H} = \hat{\boldsymbol{\beta}}(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}}'$, $\boldsymbol{E} = \boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{\beta}}(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}}'$, n is the number of unrelated individuals, and $\hat{\boldsymbol{\beta}} = \boldsymbol{Y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is the least squares estimate of the vector of genetic effects $\boldsymbol{\beta}$. Note that $\boldsymbol{X}'\boldsymbol{X} = \sum_{i=1}^{n} X_i^2$ is a random variable (not a matrix), where $\mathrm{E}(X_i^2) = 2f(1-f) = \mathrm{Var}(X_i)\ \forall\, i$. Using our distributional assumptions about centered $\boldsymbol{X}$ and $\boldsymbol{\mathcal{E}}$, it can be shown that $\frac{1}{n}\boldsymbol{H} \xrightarrow{\mathrm{P}} 2f(1-f)\boldsymbol{\beta}\boldsymbol{\beta}'$ and $\frac{1}{n}\boldsymbol{E} \xrightarrow{\mathrm{P}} \boldsymbol{\Sigma}$ as $n \to \infty$. Here, $\xrightarrow{\mathrm{P}}$ denotes convergence in probability as $n \to \infty$.

For the CS residual covariance matrix $\boldsymbol{\Sigma}$, we know that the eigen vector corresponding to the largest eigenvalue $\lambda_1 = \sigma^2\{1 + (K-1)\rho\}$ is $\boldsymbol{v}_1 \propto \boldsymbol{1}$, while the eigen vectors corresponding to $\lambda_2 = ... = \lambda_K = \sigma^2(1-\rho)$ are respectively $\boldsymbol{v}_2, ..., \boldsymbol{v}_K$ such that $\boldsymbol{1}'\boldsymbol{v}_k = 0\ \forall\, k = 2, ..., K$. For the eigen vectors to be orthonormal, we must have $\boldsymbol{v}_1 = c_K\boldsymbol{1}$ where $c_K^2 = 1/K$. Thus, we can write, $\boldsymbol{\Sigma} = \lambda_1 c_K^2 \boldsymbol{1}\boldsymbol{1}' + \sum_{i=2}^{K} \lambda_i \boldsymbol{v}_i\boldsymbol{v}_i'$ and $\boldsymbol{\Sigma}^{-1} = \frac{1}{\lambda_1} c_K^2 \boldsymbol{1}\boldsymbol{1}' + \sum_{i=2}^{K} \frac{1}{\lambda_i}\boldsymbol{v}_i\boldsymbol{v}_i'$.

Consider the testing of $H_0 : \boldsymbol{\beta} = 0$ against two possible alternatives: $H_{a,u} : \beta_1 = ... = \beta_u \neq 0, \beta_{K-u} = ... = \beta_K = 0$ (partial association) and $H_{a,K} : \beta_1 = ... = \beta_K \neq 0$ (complete association). Under the alternative $H_{a,K}$ (complete association), $|\boldsymbol{I}+\boldsymbol{H}\boldsymbol{E}^{-1}|$ is given by

$$\begin{aligned}\left|\boldsymbol{I}+\frac{\boldsymbol{H}_K}{n}\left(\frac{\boldsymbol{E}}{n}\right)^{-1}\right| &\underset{n\to\infty}{\overset{P}{\to}} \left|\boldsymbol{I}_K+(2f(1-f)\beta_1^2\mathbf{1}\mathbf{1}')\left(\frac{1}{\lambda_1}c_K^2\mathbf{1}\mathbf{1}'+\sum_{i=2}^{K}\frac{1}{\lambda_i}\boldsymbol{v}_i\boldsymbol{v}_i'\right)\right| \\ &= 1+\frac{2f(1-f)\beta_1^2}{\lambda_1}K\end{aligned}$$

Under the alternative $H_{a,u}$ (partial association),

$|\boldsymbol{I}+\boldsymbol{H}\boldsymbol{E}^{-1}| \overset{H_{a,u}}{=} |\boldsymbol{I}+\frac{\boldsymbol{H}_u}{n}\left(\frac{\boldsymbol{E}}{n}\right)^{-1}|$

$$\begin{aligned}&\underset{n\to\infty}{\overset{P}{\to}} \left|\boldsymbol{I}_K+2f(1-f)\begin{pmatrix}\beta_1^2\mathbf{1}_u\mathbf{1}_u' & \mathbf{0}\\ \mathbf{0}' & \boldsymbol{O}\end{pmatrix}\begin{pmatrix}\boldsymbol{\Sigma}_{11(u\times u)} & \boldsymbol{\Sigma}_{12(u\times\overline{K-u})}\\ \boldsymbol{\Sigma}'_{12(\overline{K-u}\times u)} & \boldsymbol{\Sigma}_{22(\overline{K-u}\times\overline{K-u})}\end{pmatrix}^{-1}\right| \\ &= \left|\boldsymbol{I}+2f(1-f)\begin{pmatrix}\beta_1^2\mathbf{1}_u\mathbf{1}_u' & \mathbf{0}\\ \mathbf{0}' & \boldsymbol{O}\end{pmatrix}\begin{pmatrix}\boldsymbol{\Sigma}^{11} & \boldsymbol{\Sigma}^{12}\\ \star & \star\end{pmatrix}\right| \\ &= \left|\boldsymbol{I}+2f(1-f)\begin{pmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \mathbf{0}' & \boldsymbol{O}\end{pmatrix}\right| \\ &= |\boldsymbol{I}_u+2f(1-f)\boldsymbol{A}| \\ &= 1+\frac{2f(1-f)\beta_1^2}{\sigma^2(1-\rho)}\frac{1+(K-u-1)\rho}{1+(K-1)\rho}u\end{aligned}$$

where $\boldsymbol{\Sigma}_{11}=\sigma^2(1-\rho)\boldsymbol{I}_u+\sigma^2\rho\mathbf{1}_u\mathbf{1}_u'$, $\boldsymbol{\Sigma}_{22}=\sigma^2(1-\rho)\boldsymbol{I}_{K-u}+\sigma^2\rho\mathbf{1}_{K-u}\mathbf{1}_{K-u}'$, $\boldsymbol{\Sigma}_{12}=\sigma^2\rho\mathbf{1}_u\mathbf{1}_{K-u}'$, $\boldsymbol{\Sigma}^{11}=(\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}')^{-1}$, $\boldsymbol{\Sigma}^{12}=-\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}$, $\boldsymbol{A}=\beta_1^2\mathbf{1}_u\mathbf{1}_u'\left(\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}'\right)^{-1}$, $\boldsymbol{B}=-\beta_1^2\mathbf{1}_u\mathbf{1}_u'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}$

So, $|\boldsymbol{I}+\boldsymbol{H}_u\boldsymbol{E}^{-1}|-|\boldsymbol{I}+\boldsymbol{H}_K\boldsymbol{E}^{-1}| \underset{n\to\infty}{\overset{P}{\to}} \frac{2f(1-f)\beta_1^2}{\sigma^2\{1+(K-1)\rho\}}\left(\frac{1+(K-u-1)\rho}{1-\rho}u-K\right)>0$ under the condition $\frac{u}{K}>\frac{\sigma^2\{1-\rho\}}{\sigma^2\{1+(K-u-1)\rho\}}$. It may be noted that the condition simplifies to $\rho>\frac{1}{u+1}$, which explains why we observe higher power for partial association and lower for complete association for $K=2$ traits once the within trait correlation ρ exceeds 1/2. ■

Proof of Theorem 2

Without loss of generality, let us assume that $\boldsymbol{Y}$ and $\boldsymbol{X}$ are centered. In particular, for $K=2$, $\frac{1}{n}\boldsymbol{H}\xrightarrow{\text{P}}2f(1-f)\begin{pmatrix}\beta_1^2 & \beta_1\beta_2\\ \beta_1\beta_2 & \beta_2^2\end{pmatrix}$ and $\frac{1}{n}\boldsymbol{E}\xrightarrow{\text{P}}\sigma^2\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}$ as $n\to\infty$.

Let us now consider the alternatives $H_{a1} : \beta_1 \neq 0, \beta_2 = 0$ (only 1 trait is associated), and $H_{a2} : \beta_1 \neq \beta_2 \neq 0$ (both traits are associated). Under H_{a1}, the $\boldsymbol{H}/n$ matrix becomes $\frac{1}{n}\boldsymbol{H}_1 \xrightarrow{P} 2f(1-f)\begin{pmatrix} \beta_1^2 & 0 \\ 0 & 0 \end{pmatrix}$ for large n. Let $\boldsymbol{H}_2$ be the $\boldsymbol{H}$ matrix under H_{a2}. So,

$$\begin{aligned}
& \det\left(\frac{\boldsymbol{H}_1}{n} + \frac{\boldsymbol{E}}{n}\right) - \det\left(\frac{\boldsymbol{H}_2}{n} + \frac{\boldsymbol{E}}{n}\right) \\
\xrightarrow{P} \; & \begin{vmatrix} \sigma^2 + 2f(1-f)\beta_1^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{vmatrix} - \begin{vmatrix} \sigma^2 + 2f(1-f)\beta_1^2 & \rho\sigma^2 + 2f(1-f)\beta_1\beta_2 \\ \rho\sigma^2 + 2f(1-f)\beta_1\beta_2 & \sigma^2 + 2f(1-f)\beta_2^2 \end{vmatrix} \\
= \; & 2f(1-f)\beta_2\sigma^2(2\rho\beta_1 - \beta_2) \\
> \; & 0 \text{ if } \{\beta_2 < 2\rho\beta_1 \;\&\; \beta_2 > 0\} \text{ or } \{\beta_2 > 2\rho\beta_1 \;\&\; \beta_2 < 0\}
\end{aligned}$$

This means, we expect the statistic $|\boldsymbol{E}|/|\boldsymbol{H}_1 + \boldsymbol{E}|$ under H_{a1} (when only 1 trait is associated) to be closer to 0 than the statistic $|\boldsymbol{E}|/|\boldsymbol{H}_2 + \boldsymbol{E}|$ under H_{a2} when $\{0 < \beta_2 < 2\rho\beta_1\}$ or $\{0 > \beta_2 > 2\rho\beta_1\}$. Thus, for $K = 2$, MANOVA is expected to have more power when 1 trait is associated than when both traits are associated if $0 < \beta_2 < 2\rho\beta_1$ or $0 > \beta_2 > 2\rho\beta_1$.

$\blacksquare$

Appendix S2

Acceptance Region for MANOVA based on $\boldsymbol{Z}$

Consider the MMLR model

$$\boldsymbol{Y}_{n\times K} = \boldsymbol{X}_{n\times 1}\boldsymbol{\beta}'_{1\times K} + \boldsymbol{\mathcal{E}}_{n\times K} \tag{1}$$

where $\boldsymbol{\beta}' = (\beta_1, ..., \beta_K)$ is the vector of fixed unknown genetic effects corresponding to the K correlated traits, and $\boldsymbol{\mathcal{E}}$ is the matrix of random errors. For testing that the SNP is not associated with any of the K traits, the null hypothesis of interest is $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$.

Assume $\boldsymbol{\mathcal{E}}$ is a normal data matrix from $N_K(\boldsymbol{0}, \boldsymbol{\Sigma})$. The log-likelihood $l(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ of the trait matrix $\boldsymbol{Y}$ is given by

$$l(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = -\frac{1}{2}n \log|2\pi\boldsymbol{\Sigma}| - \frac{1}{2}\text{tr}\left\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}')'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}')\right\} \tag{2}$$

where $\boldsymbol{\Sigma}$ is a positive definite matrix representing residual covariance among the traits. The MLE of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are $\hat{\boldsymbol{\beta}} = \boldsymbol{Y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{Y}'(\boldsymbol{I}_K - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Y}$ respectively. Under the null, $\boldsymbol{\beta} = \boldsymbol{0}$ and the MLE of $\boldsymbol{\Sigma}$ is $\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{n}\boldsymbol{Y}'\boldsymbol{Y}$. The likelihood ratio test (LRT) of H_0 based on the MMLR model with matrix normal errors is equivalent to MANOVA statistic $\boldsymbol{\Lambda}$ (Wilk's Lambda):

$$-2\log\boldsymbol{\Lambda} = 2\left(l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}) - l(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_0)\right) = n\log\frac{|\hat{\boldsymbol{\Sigma}}_0|}{|\hat{\boldsymbol{\Sigma}}|} = -n\log\frac{|\boldsymbol{E}|}{|\boldsymbol{H} + \boldsymbol{E}|} \tag{3}$$

where $\boldsymbol{H}$ and $\boldsymbol{E}$ are the hypothesis and the error sum of squares and cross product (SSCP) matrices respectively.

Let us now consider the following notations: $\dot{\boldsymbol{l}}(\boldsymbol{\beta}) = \frac{\partial}{\partial\boldsymbol{\beta}}l(\boldsymbol{\beta}, \boldsymbol{\Sigma})$; $\ddot{\boldsymbol{l}}(\boldsymbol{\beta}) = \frac{\partial^2}{\partial\boldsymbol{\beta}^2}l(\boldsymbol{\beta}, \boldsymbol{\Sigma})$. The Fisher Information matrix under H_0 is $\boldsymbol{I}(\boldsymbol{0}) = -\,\text{E}_{\boldsymbol{\beta}=\boldsymbol{0}}(\ddot{\boldsymbol{l}}(\boldsymbol{\beta}))$. Using Taylor's Expansion upto order 2, we can write the LRT statistic as

$$-2\log\boldsymbol{\Lambda} = 2\left\{0 + \frac{1}{2}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{0})'\left(-\frac{1}{n}\ddot{\boldsymbol{l}}(\boldsymbol{\beta}^*)\right)(\hat{\boldsymbol{\beta}} - \boldsymbol{0})\right\},\ \text{where } |\boldsymbol{\beta}^* - \boldsymbol{0}| \leq |\hat{\boldsymbol{\beta}} - \boldsymbol{0}|$$

Observe that $\sqrt{n}(\hat{\boldsymbol{\beta}} - \mathbf{0})' \overset{D}{\to} \boldsymbol{Z} \sim N_K(\mathbf{0}, \boldsymbol{I}^{-1}(\mathbf{0}))$. If a particular component of the true $\boldsymbol{\beta}$ is large (small), we expect the corresponding component of $\hat{\boldsymbol{\beta}}$ and hence of $\boldsymbol{Z}$ to be large (small). Thus for $\boldsymbol{Z}$ to be larger than $\mathbf{0}$, we need to have the true $\boldsymbol{\beta}$ larger than $\mathbf{0}$. We can then write the asymptotically equivalent form of MANOVA Wilk's Lambda statistic in terms of a statistic involving $\boldsymbol{Z}$:

$$-2\log \boldsymbol{\Lambda} \overset{D}{\to} \boldsymbol{Z}'\boldsymbol{I}(\mathbf{0})\boldsymbol{Z} \overset{a}{\sim} \chi^2_K$$

Instead of drawing the acceptance region of Wilk's Lambda statistic, one can draw the acceptance region of the test statistic $\boldsymbol{Z}'\boldsymbol{I}(\mathbf{0})\boldsymbol{Z}$. The ellipse representing acceptance region for MANOVA is asymptotically equivalent to

$$\boldsymbol{\mathcal{E}}_c(\boldsymbol{z}; \boldsymbol{S}, \bar{\boldsymbol{z}}) \equiv \left\{\boldsymbol{z} : (\boldsymbol{z} - \bar{\boldsymbol{z}})'\boldsymbol{S}^{-1}(\boldsymbol{z} - \bar{\boldsymbol{z}}) \leq c^2\right\}$$

where $\boldsymbol{S} = (n-1)^{-1}\sum_{i=1}^{n}(\boldsymbol{z}_i - \bar{\boldsymbol{z}})(\boldsymbol{z}_i - \bar{\boldsymbol{z}})'$ and c^2 is the 95-th percentile of the distribution of $\boldsymbol{Z}$. The boundary of the ellipse $\boldsymbol{\mathcal{E}}_c$ is computed as a transformation of the unit circle, $\mathcal{U} = (\sin\theta, \cos\theta)$ for $\theta \in (0, 2\pi)$. Let $\boldsymbol{A} = \boldsymbol{S}^{1/2}$ be the Choleski square root of $\boldsymbol{S}$ in the sense that $\boldsymbol{S} = \boldsymbol{A}\boldsymbol{A}'$. Then, $\boldsymbol{\mathcal{E}}_c = \bar{\boldsymbol{z}} + c\boldsymbol{A}\boldsymbol{\mathcal{U}}$ is an ellipse centered at the mean $\bar{\boldsymbol{z}} = (\bar{z}_1, \bar{z}_2)$. The size of the ellipse reflects the standard deviations of z_1 and z_2 while the shape reflects their correlation. $\boldsymbol{Z}$ has a $N_K(\mathbf{0}, \boldsymbol{I}(\mathbf{0})^{-1})$ distribution due to which we expect $\bar{\boldsymbol{z}} \approx \mathbf{0}$ and $\boldsymbol{S} \approx \frac{1}{n}\sum \boldsymbol{z}\boldsymbol{z}' \overset{P}{\to} \boldsymbol{I}(\mathbf{0})^{-1} = \frac{1}{2p(1+p)}\boldsymbol{\Sigma}$ where p is the m.a.f. of the genetic variant. Thus, for drawing the theoretical acceptance region of MANOVA, we use the facts that $\bar{\boldsymbol{Z}} \overset{P}{\to} \mathbf{0}$ and $\boldsymbol{S} \overset{P}{\to} \frac{1}{2p(1+p)}\boldsymbol{\Sigma}$. For Figure 1 in the main manuscript, we assumed $\boldsymbol{\Sigma} = \sigma^2\{(1-\rho)\boldsymbol{I}_K + \rho\mathbf{1}\mathbf{1}'\}$ with $K = 2$. The theoretical acceptance region for MANOVA will then be asymptotically equivalent to $\boldsymbol{\mathcal{E}}_c\left(\boldsymbol{z}; \frac{\boldsymbol{\Sigma}}{2p(1+p)}, \mathbf{0}\right) \equiv \left\{\boldsymbol{z} : \boldsymbol{z}'\left(\frac{\boldsymbol{\Sigma}}{2p(1+p)}\right)^{-1}\boldsymbol{z} \leq c^2\right\}$.

Appendix S3

Details of the approximate p-value calculation for USAT

Let $T_M = -2\log\boldsymbol{\Lambda} \overset{a}{\sim} \chi^2_K$ be the MANOVA test statistic based on Wilk's lambda and $T_S \overset{approx}{\sim} a\chi^2_d + b$ be the SSU test statistic based on score vector from marginal normal models. For USAT, we first consider the weighted statistic $T_\omega = \omega T_M + (1-\omega)T_S$, where $\omega \in [0,1]$ is the weight. Both MANOVA and SSU are special cases of the class of statistics T_ω. Under H_0, for a given weight ω, T_ω is approximately a linear combination of chi-squared distributions. The computation of p-value p_ω of the test statistic T_ω does not require independence of the statistics T_M and T_S. A detailed explanation of the determination of p_ω is provided below.

Observe that one can write $T_M = \boldsymbol{U}'\boldsymbol{I}(\mathbf{0})^{-1}\boldsymbol{U}$, where $\boldsymbol{U}$ is the score vector under $H_0 : \boldsymbol{\beta} = \mathbf{0}$ from the MMLR model (1) and $\boldsymbol{I}(\mathbf{0}) = -\operatorname{E}_{\boldsymbol{\beta}=\mathbf{0}}\left(\frac{\partial}{\partial\boldsymbol{\beta}} l(\boldsymbol{\beta},\boldsymbol{\Sigma})\right) = \operatorname{Cov}(\boldsymbol{U})|_{\boldsymbol{\beta}=\mathbf{0}}$ is the Fisher Information matrix under H_0. On the other hand, $T_S = \boldsymbol{U}_M'\boldsymbol{U}_M$, where $\boldsymbol{U}_M$ is the marginal score vector under H_0 from the marginal models in equation (2) of main paper. As derived in the main manuscript, $\boldsymbol{U}_M = \boldsymbol{Y}'\boldsymbol{X}/\hat{\sigma}_0^2$, where $\boldsymbol{Y}$ is the $n \times K$ phenotype matrix, $\boldsymbol{X}$ is the $n \times 1$ genotype matrix and $\hat{\sigma}_0^2$ is the MLE of σ^2 under H_0. Similarly, one can show that $\boldsymbol{U} = \hat{\boldsymbol{\Sigma}}_0^{-1}\boldsymbol{Y}'\boldsymbol{X}$, where $\hat{\boldsymbol{\Sigma}}_0 = \boldsymbol{Y}'\boldsymbol{Y}/n$ is the MLE of $\boldsymbol{\Sigma}$ in MMLR model (1) under H_0. The estimated variance of the score vector $\boldsymbol{U}$ under H_0 is given by $\operatorname{Cov}(\boldsymbol{U})|_{\boldsymbol{\beta}=\mathbf{0}} = \boldsymbol{I}(\mathbf{0}) = (\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\Sigma}}_0^{-1}$. For a given weight ω, one can thus write

$$\begin{aligned}
T_\omega &= \omega T_M + (1-\omega)T_S \\
&= \omega\left(\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\sigma}_0^2\boldsymbol{U}_M\right)'\boldsymbol{I}(\mathbf{0})^{-1}\left(\hat{\boldsymbol{\Sigma}}_0^{-1}\hat{\sigma}_0^2\boldsymbol{U}_M\right) + (1-\omega)\boldsymbol{U}_M'\boldsymbol{U}_M \\
&= \boldsymbol{U}_M'\left(\omega\hat{\sigma}_0^4(\boldsymbol{X}'\boldsymbol{X})^{-1}\hat{\boldsymbol{\Sigma}}_0^{-1} + (1-\omega)\boldsymbol{I}_K\right)\boldsymbol{U}_M
\end{aligned}$$

where $\boldsymbol{I}_K$ is the identity matrix of order K. Denote $\boldsymbol{A} = \omega\hat{\sigma}_0^4(\boldsymbol{X}'\boldsymbol{X})^{-1}\hat{\boldsymbol{\Sigma}}_0^{-1} + (1-\omega)\boldsymbol{I}_K$, which is a $K \times K$ symmetric, non-negative definite matrix. Note that marginal score vector $\boldsymbol{U}_M$ has mean $\mathbf{0}$, estimated variance $\operatorname{Cov}(\boldsymbol{U}_M) = \boldsymbol{X}'\boldsymbol{X}\boldsymbol{Y}'\boldsymbol{Y}/(n\hat{\sigma}_0^4)$, and has an asymptotic K-variate normal distribution. Let $\boldsymbol{P}$ be a $K \times K$ orthonormal matrix that converts $\boldsymbol{B} = \text{Cov}(\boldsymbol{U}_M)^{1/2}\boldsymbol{A}\text{Cov}(\boldsymbol{U}_M)^{1/2} = \omega\boldsymbol{I}_K + (1-\omega)\text{Cov}(\boldsymbol{U}_M)$ to the diagonal form $\boldsymbol{\Gamma} = \text{diag}(\lambda_1, ...\lambda_K)$, where $\lambda_1 \geq 0, ..., \lambda_K \geq 0$. The weighted statistic T_ω can, then, be expressed as a non-negative quadratic form:

$$T_\omega = \boldsymbol{U}'_M\boldsymbol{A}\boldsymbol{U}_M = \boldsymbol{V}'_M\boldsymbol{\Gamma}\boldsymbol{V}_M = \sum_{j=1}^{K}\lambda_j\chi^2_{h_j}(\delta_j) \tag{4}$$

where $\boldsymbol{V}_M = \boldsymbol{P}\text{Cov}(\boldsymbol{U}_M)^{-1/2}\boldsymbol{U}_M \overset{a}{\sim} N(\boldsymbol{0}, \boldsymbol{I}_K)$, and $h_j = 1$, $\delta_j = 0$ for all $j = 1, 2, ..., K$. For a given $\omega \in [0, 1]$, the p-value p_ω of the statistic T_ω can, thus, be calculated by Liu et al. (2009) algorithm as:

$$p_\omega = \text{P}\left(T_\omega > t_\omega\right) \approx \text{P}\left(\chi^2_l(\delta) > t^*_\omega\sigma_\chi + \mu_\chi\right) \tag{5}$$

where t_ω is the observed value of T_ω statistic, $t^*_\omega = (t_\omega - \text{E}(T_\omega))/\sqrt{\text{Var}(T_\omega)}$, $\mu_\chi = \text{E}\left(\chi^2_l(\delta)\right) = l + \delta$, $\sigma_\chi = \sqrt{\text{Var}\left(\chi^2_l(\delta)\right)} = \sqrt{2(l + 2\delta)}$. The parameters δ and l are chosen such that the skewness of T_ω and $\chi^2_l(\delta)$ are same and the difference between the kurtoses of T_ω and $\chi^2_l(\delta)$ is minimized.

Apriori the optimal weight ω is not known. We propose our unified test USAT as

$$T_{USAT} = \min_{0\leq\omega\leq 1} p_\omega$$

Thus, the USAT test statistic is not exactly the best weighted combination of MANOVA and SSU. It is the minimum of the p-values of the different weighted combinations. For practical implementations of USAT, a grid of 11 ω values were considered: $\{\omega_1 = 0, \omega_2 = 0.1, ..., \omega_{10} = 0.9, \omega_{11} = 1\}$.

To find the p-value of our USAT test statistic, we need the null distribution of USAT. We propose an approximate p-value calculation using a one-dimensional numerical integration, which makes USAT suitable for application on a GWAS scale. Observe that the p-value of statistic T_{USAT} is

$$p_{USAT} \quad = \quad P(T_{USAT} \leq t_{USAT}) = 1 - P(T_{USAT} \geq t_{USAT})$$

$$
\begin{aligned}
&= 1 - P\left(\min_{\omega} p_{\omega} \geq t_{USAT}\right) = 1 - P\left(1 - \min_{\omega} p_{\omega} < 1 - t_{USAT}\right) \\
&= 1 - P\left(\max_{\omega}(1 - p_{\omega}) < 1 - t_{USAT}\right) \\
&= 1 - P\left(\{1 - p_{\omega_1} < 1 - t_{USAT}\}, \ldots, \{1 - p_{\omega_{11}} < 1 - t_{USAT}\}\right) \\
&= 1 - P\Big(\{(1 - p_{\omega_1})^{th} \text{ quantile} < (1 - t_{USAT})^{th} \text{ quantile}\}, \ldots, \\
&\quad \{(1 - p_{\omega_{11}})^{th} \text{ quantile} < (1 - t_{USAT})^{th} \text{ quantile}\}\Big) \\
&= 1 - P\left(T_{\omega_1} < q_{\min}(\omega_1), ..., T_{\omega_{11}} < q_{\min}(\omega_{11})\right) \\
&= 1 - P\left(T_S < \min_{\omega} \frac{q_{\min}(\omega) - \omega T_M}{1 - \omega}\right) \\
&= 1 - \int F_{T_S|T_M}\big(\delta_{\omega}(x)|x\big) f_{T_M}(x) dx
\end{aligned}
$$

where t_{USAT} is the observed value of USAT test statistic for a given dataset, $q_{\min}(\omega_b)$ is the $(1 - t_{USAT})$-th percentile of the distribution of T_{ω_b} for a given $\omega = \omega_b$, $F_{T_S|T_M}(.|x)$ is the conditional cdf of SSU statistic T_S given MANOVA statistic T_M, $f_{T_M}(.)$ is the pdf of MANOVA test statistic T_M, and $\delta_{\omega}(x) = \min_{\omega \in \{\omega_1, \ldots, \omega_{11}\}} \frac{q_{\min}(\omega) - \omega x}{1 - \omega}$.

Recall that T_S and T_M are two quadratic forms (QF), which are not independently distributed. The exact joint distribution of T_S and T_M is too complicated to compute (Khatri et al., 1977; Khatri, 1980). Our literature search did not yield any computationally feasible method for approximating the distribution $F_{T_S|T_M}(.|T_M = x)$ required to calculate p_{USAT}. In such a scenario, a simple and straightforward approximation seems to be the assumption of independence and thereby we get the approximate p-value

$$
p_{USAT} \approx 1 - \int_0^{\infty} F_{T_S}\big(\delta_{\omega}(x)|x\big) f_{T_M}(x) dx
$$

where $F_{T_S}(.)$ is the cdf of SSU test statistic T_S. This approximation of distribution $[T_S|T_M]$ by $[T_S]$ can yield conservative p-values at heavier tails of the null distribution of USAT test statistic. However, for extreme tails (regions in which we are interested when applying USAT at GWAS level), this conservativeness is not an issue (as demonstrated by USAT type I error analysis in main manuscript). Detailed study on the accuracy of this approximation is provided in the next section. In this context, it is worth noting that we have not assumed T_S and T_M to be independent throughout. For example, the information on their dependence has been incorporated in the calculation of p_ω (p-value of weighted statistic T_ω). The independence assumption has been made only in the last step of USAT p-value calculation.

Implementation of the approximate p-value method

For the integral $\int_0^\infty F_{T_S}\big(\delta_\omega(x)\big) f_{T_M}(x)dx$, we first need to evaluate

$$F_{T_S}\big(\delta_\omega(x)\big) = \mathrm{P}\left(T_S \le \delta_\omega(x)\right) \approx \mathrm{P}\left(a\chi^2_d + b \le \delta_\omega(x)\right) = \mathrm{P}\left(\chi^2_d \le \frac{\delta_\omega(x) - b}{a}\right)$$

This can be easily evaluated using function `pchisq()` in `R` (R Development Core Team, 2014). The integrand as a function of x can then be coded as `pchisq((delta.x-b)/a, df=d, ncp=0)*dchisq(x, df=K)`. The integration has been performed numerically using `R` function `integrate()`. When the optimal choice of ω lies near the boundary (i.e., close to 0 or 1) and the corresponding statistic (T_S or T_M depending upon whether optimal ω is close to 0 or 1) is highly significant (i.e., corresponding p-value is of the order of 10^{-8}), the function `integrate` can have low accuracy and can give rise to an integral value exceeding 1. In such a scenario, `R` function `quadinf()` from package `pracma` (Borchers, 2012) can give very accurate results. The cost of accuracy is longer computation time: `quadinf` takes almost twice as much time compared to `integrate`. For our simulated datasets as well as real dataset, we found the two functions giving very similar results in most situations except in the afore-mentioned scenario where `integrate` gave negative p-values for USAT. In such rare situations, we implemented the numerical integration using `quadinf`.

Details on the accuracy of the approximation involving independence assumption

Since the exact distribution of $[T_S|T_M]$ is not known, we studied the accuracy of our approximation (independence assumption in the last step of p_{USAT} calculation) using Monte Carlo samples. For this purpose, we first simulated two independent sets of $N = 10,000$ marginal score vectors $\boldsymbol{U}_M$ from multivariate $N_K(\mathbf{0}, \boldsymbol{C})$, where $\boldsymbol{C}$ is the score covariance matrix that directly depends on the trait covariance structure $\mathbf{\Sigma}$. Both CS(ρ) and AR1(ρ) correlation structures were considered for $\rho = 0.2, 0.5, 0.8$. We took three different choices of K as in our simulation studies: $K = 5,\ 10,\ 20$. For each set, we calculated the statistics T_S and T_M (i.e., we have samples of SSU and MANOVA statistics from their null distributions). Let us denote $T_S^{(j)}$ and $T_M^{(j)}$ to be the SSU and the MANOVA statistics from the j-th set of Monte Carlo samples, $j = 1, 2$. Note that the statistics $T_S^{(j)}$ and $T_M^{(j')}$ are correlated for $j = j'$ and uncorrelated for $j = j'$. Thus, for a given value of t_{USAT}, the Monte Carlo estimate (MCE) of the true probability $p_{USAT} = \mathrm{P}(T_S > \delta_\omega(t_{USAT}, T_M))$ is

$$p_{USAT}^{true} = \frac{1}{N}\sum_{i=1}^{N} I\left(T_S^{(1),i} > \delta_\omega\left(t_{USAT}, T_M^{(1),i}\right)\right)$$

where $I(.)$ is the indicator function, $T_S^{(1),i}$ is the SSU statistic based on i-th sample in the 1st set, $T_M^{(1),i}$ is similarly defined, and $\delta_\omega(.)$ is as defined earlier. The MCE of the approximate p_{USAT} (where independence of T_S and T_M was assumed) can be obtained as

$$p_{USAT}^{approx} = \frac{1}{N}\sum_{i=1}^{N} I\left(T_S^{(1),i} > \delta_\omega\left(t_{USAT}, T_M^{(2),i}\right)\right)$$

Note that one can also obtain this approximate p_{USAT} using our p-value calculation method directly. Next we plotted these three different estimates of p_{USAT} against a range of values of t_{USAT}. In the following Figures S1−S4, the black solid curve corresponds to p_{USAT}^{true} (MCE of true p-value), blue solid curve corresponds to p_{USAT}^{approx} (MCE of approximate p-value) and the red solid curve corresponds to p_{USAT} computed directly from our approximate p-value calculation approach.

Figures S1 and S2 show the plots of these different estimates of p_{USAT} against t_{USAT} in $[0, 1]$ range using weights $\omega = 0, 0.1, 0.2, ..., 0.9, 1$ for CS and AR1 correlation structures respectively. As expected, the approximate p-values from Monte Carlo samples and the approximate p-values from our method are similar (the blue and the red curves are overlapping). We also observe that our approximation causes the USAT p-values to be conservative, more so at the heavier tails of the null distribution of USAT. With increase in strength of correlation parameter ρ or increase in the number of traits K, this conservativeness decreases. For very small values of t_{USAT} (the region we are interested in when applying USAT on a genome-wide scale), our approach does not seem to be conservative.

To study the effect of approximation at the extreme tail (near 0), we considered t_{USAT} values in the range of $[0, 10^{-3}]$. Precisely, the chosen t_{USAT} values were $0, 10^{-5}, 2 \times 10^{-5}, ..., 10^{-3}$. In order to consider t_{USAT} values of the order of 10^{-5}, we simulated two independent sets of $N = 10^7$ Monte Carlo samples of T_S and T_M. Using these samples, we calculated p_{USAT}^{true} and p_{USAT}^{approx} as before. Generating 10^7 samples for as many as 20 traits is computationally intensive. To reduce computation time, we considered only a minimal set of weights: $\omega = \left\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\right\}$. Figures S3 and S4 show the plots of true and approximate p_{USAT} against t_{USAT} for CS(ρ) and AR1(ρ) correlation structures respectively. The approximate p-value curve (blue) seems to lie below the true p-value curve (black) for higher values of K and ρ, indicating that the approximation is slightly inflated at the extreme tail. The magnitude of inflatedness seems to depend on the correlation structure as well. USAT is less inflated at stringent error levels for AR1 correlation structure compared to a CS structure. In Table 2 of main manuscript, although USAT maintains correct type I error for low error levels, we observe slightly inflated type I error for $K = 20$ traits at $\alpha = 10^{-4}$.

More on the performance of the p-value approximation method: The following Table S1 (an extension of Table 2 of main manuscript) provides estimated type I error rates of USAT for $K = 5$ traits for a stringent error level of $\alpha = 10^{-5}$. To reduce computational burden, we considered only 3×10^6 datasets and hence provided $100(1-\alpha)\%$ confidence intervals for the error estimates. Although we saw that USAT generally maintains proper type I error rate at moderately low error levels (Table 2), here we observe that USAT produces somewhat inflated type I errors at stringent value of level α.

To have an idea about the effect of approximation on power in a real GWAS, Table S2 provides approximate USAT p-values along with empirical p-values for a few SNPs from the ARIC Study (refer Section 3.5 of main manuscript). For a given SNP, the empirical

Table S1: Estimated type I errors of the approximate p-value calculation approach for our USAT test. The p-values were calculated using 3×10^6 null datasets with $10,000$ unrelated individuals. Type I error rate was calculated as the proportion of datasets that had approximate p-value $\leq \alpha$. The $100(1-\alpha)\%$ confidence intervals for the estimates are provided in square braces.

K	5		
ρ	0.2	0.4	0.6
$\alpha = 10^{-5}$	2.96×10^{-5} $[1.56 \times 10^{-5}$, $4.35 \times 10^{-5}]$	2.51×10^{-5} $[1.22 \times 10^{-5}$, $3.80 \times 10^{-5}]$	2.07×10^{-5} $[0.90 \times 10^{-5}$, $3.24 \times 10^{-5}]$

Table S2: Empirical USAT p-values alongwith approximate USAT p-value (calculated from the approximate p-value method in Section 2.5) for a few randomly chosen SNPs from the ARIC data. For a given SNP, the empirical USAT p-value is calculated using 10^8 permutations of the ARIC data.

chr	SNP	position	m.a.f.	USAT p	Empirical USAT p
6	rs7753319	106997646	0.421	0.30×10^{-5}	1.65×10^{-5}
7	rs7793197	147499191	0.175	0.45×10^{-5}	2.10×10^{-5}
18	rs11660607	33269184	0.287	0.96×10^{-4}	1.55×10^{-4}
20	rs3790223	19405611	0.322	0.60×10^{-5}	2.45×10^{-5}

USAT p-value is calculated by considering 10^8 permuted datasets. Table S2 corroborates our findings from Table S1.

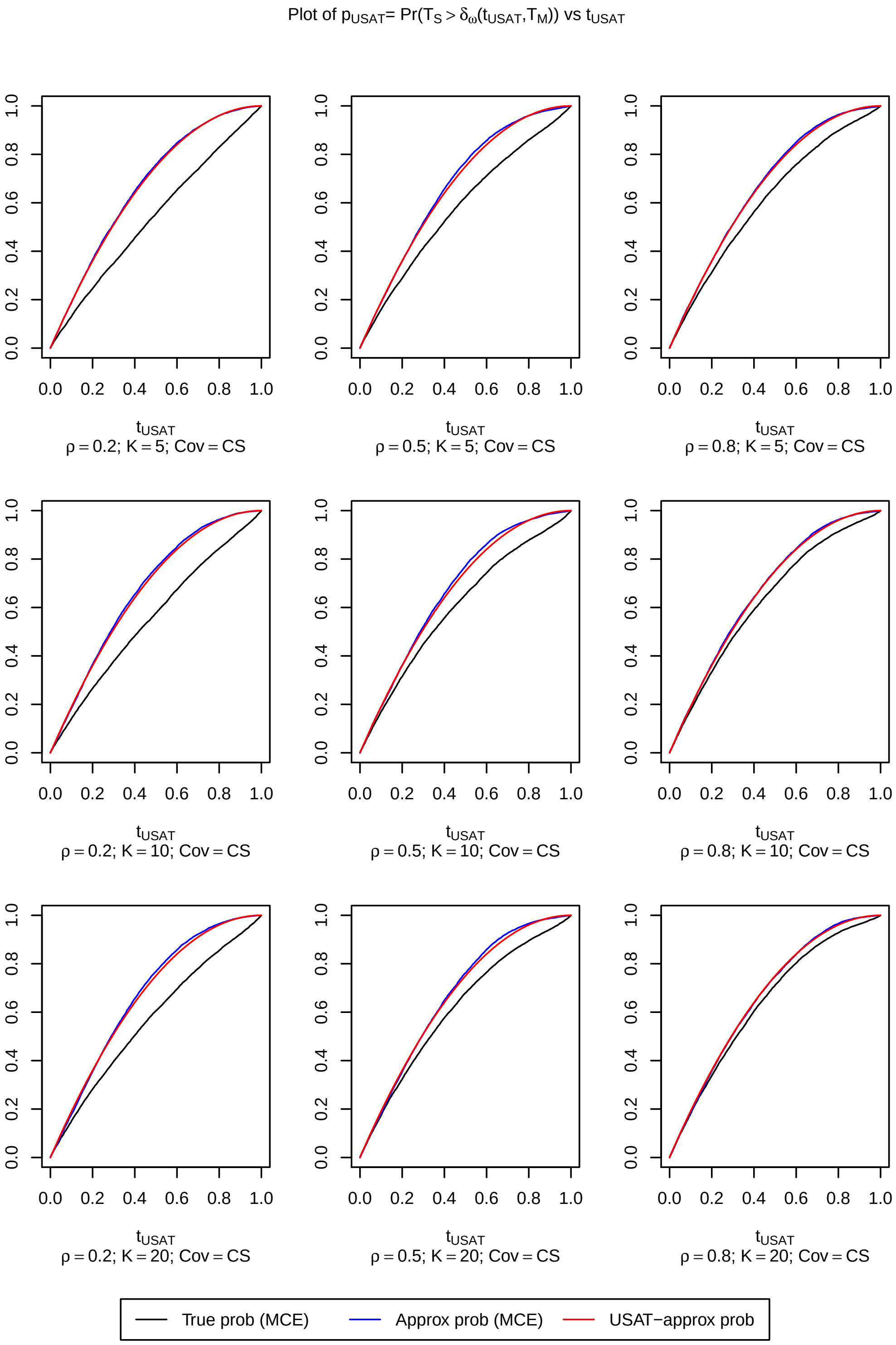


Figure S1: Comparison of approximate and true p-value of USAT based on Monte Carlo samples for CS(ρ) correlation structure. The different parameter values are: $N = 10,000$ samples, weight $\omega \in \{0, 0.1, ..., 0.9, 1\}$, $t_{USAT} \in \{0, 0.01, 0.02, ..., 0.99, 1\}$, $K \in \{5, 10, 20\}$ traits and $\rho \in \{0.2, 0.5, 0.8\}$. The black solid curve corresponds to p^{true}_{USAT} (MCE of true p-value), blue solid curve corresponds to p^{approx}_{USAT} (MCE of approximate p-value) and the red solid curve corresponds to p_{USAT} computed directly from our approximate p-value calculation approach. The approximate curves lie above the true curve indicating conservativeness of the approximation.

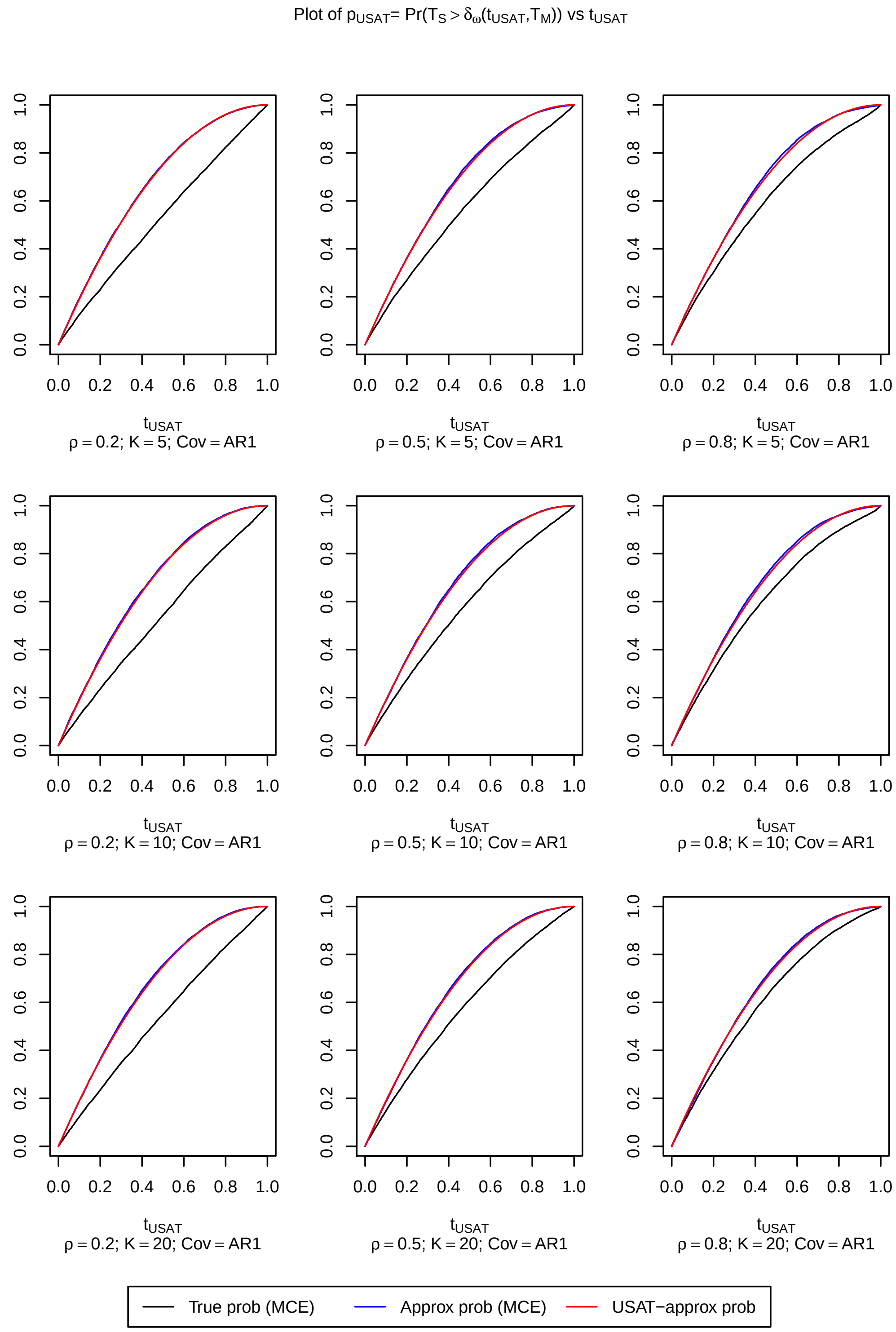


Figure S2: Comparison of approximate and true p-value of USAT based on Monte Carlo samples for AR1(ρ) correlation structure. The different parameter values are: $N = 10,000$ samples, weight $\omega \in \{0, 0.1, ..., 0.9, 1\}$, $t_{USAT} \in \{0, 0.01, 0.02, ..., 0.99, 1\}$, $K \in \{5, 10, 20\}$ traits and $\rho \in \{0.2, 0.5, 0.8\}$. The black solid curve corresponds to p_{USAT}^{true} (MCE of true p-value), blue solid curve corresponds to p_{USAT}^{approx} (MCE of approximate p-value) and the red solid curve corresponds to p_{USAT} computed directly from our approximate p-value calculation approach. The approximate curves lie above the true curve indicating conservativeness of the approximation.

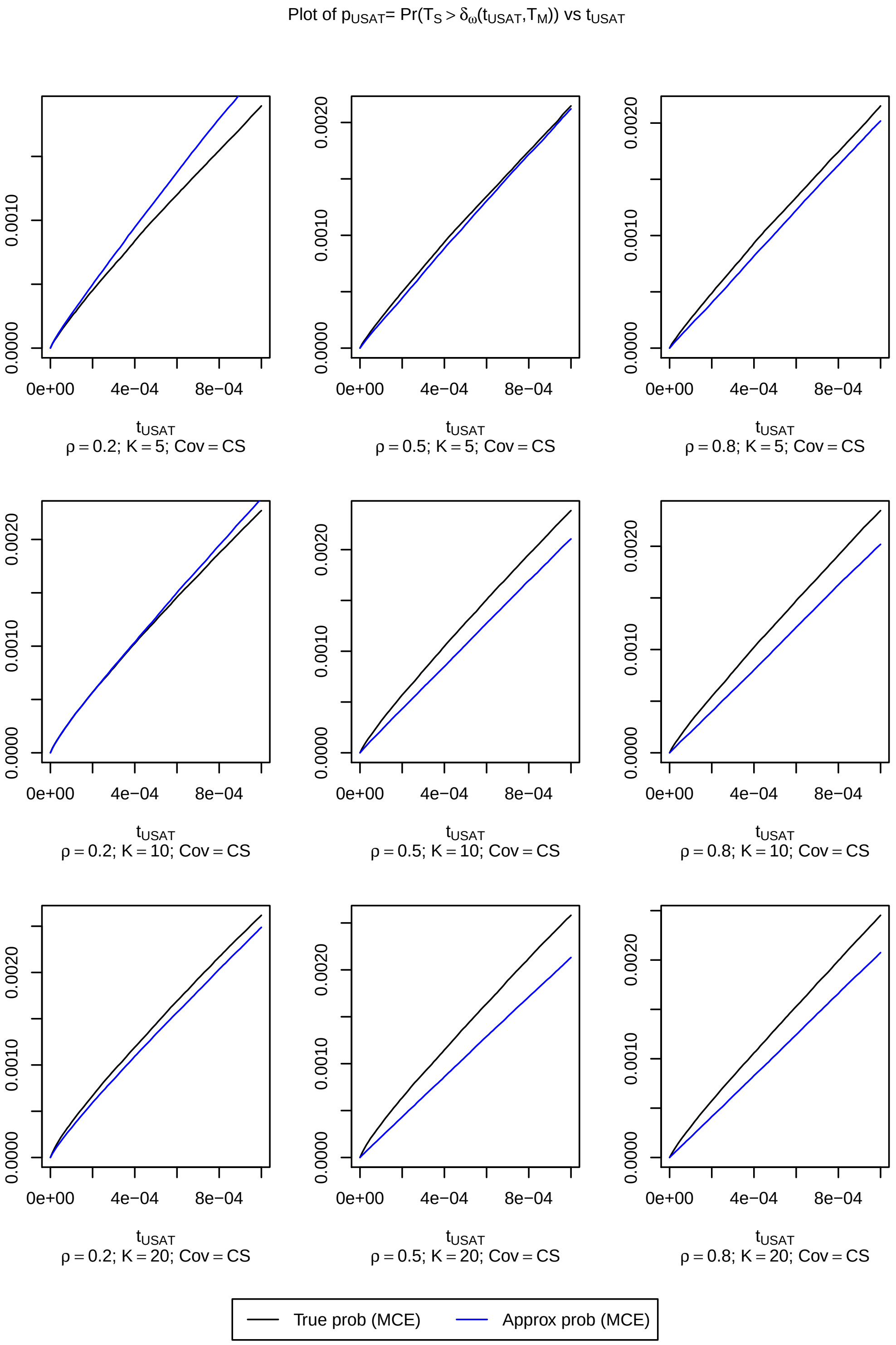


Figure S3: Comparison of approximate and true p-value of USAT based on Monte Carlo samples for CS(ρ) correlation structure. The different parameter values are: $N = 10^7$ samples, weight $\omega \in \{0, 1/4, 1/2, 3/4, 1\}$, $t_{USAT} \in \{0, 10^{-5}, 2\times10^{-5}, ..., 10^{-3}\}$, $K \in \{5, 10, 20\}$ traits and $\rho \in \{0.2, 0.5, 0.8\}$. The black solid curve corresponds to p_{USAT}^{true} (MCE of true p-value) and the blue solid curve corresponds to p_{USAT}^{approx} (MCE of approximate p-value). Curve corresponding to p_{USAT} computed directly from our approximate p-value calculation approach is not plotted to avoid clutter. In some situations, the approximate curve lies below the true curve indicating slight inflatedness of the approximation at the extreme tail.

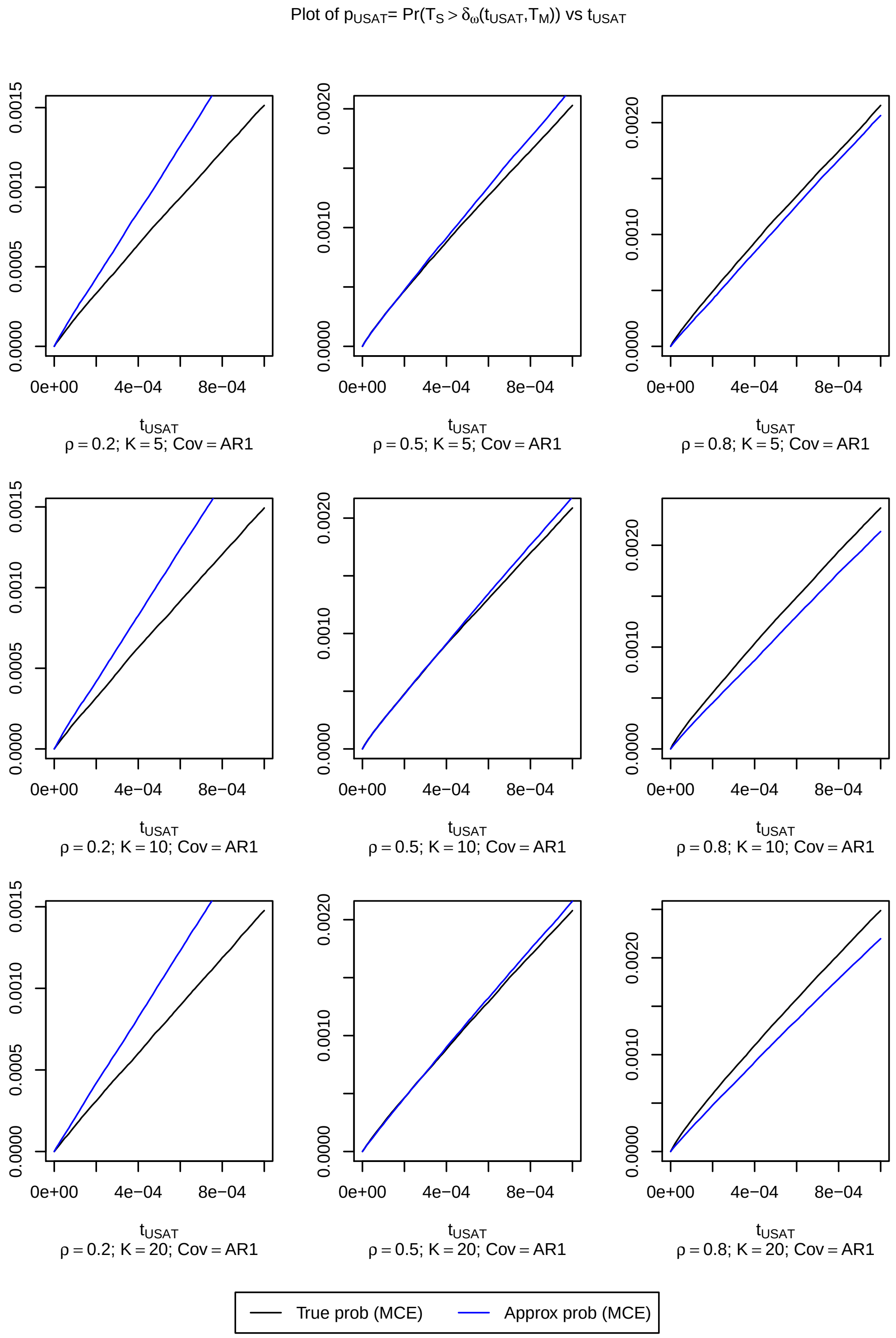


Figure S4: Comparison of approximate and true p-value of USAT based on Monte Carlo samples for AR1(ρ) correlation structure. The different parameter values are: $N = 10^7$ samples, weight $\omega \in \{0, 1/4, 1/2, 3/4, 1\}$, $t_{USAT} \in \{0, 10^{-5}, 2 \times 10^{-5}, ..., 10^{-3}\}$, $K \in \{5, 10, 20\}$ traits and $\rho \in \{0.2, 0.5, 0.8\}$. The black solid curve corresponds to p^{true}_{SAT} (MCE of true p-value) and the blue solid curve corresponds to p^{approx}_{USAT} (MCE of approximate p-value). Curve corresponding to p_{USAT} computed directly from our approximate p-value calculation approach is not plotted to avoid clutter. In most situations, the approximate curve lies above the true curve indicating conservativeness of the approximation. However, for strongly correlated traits, we observe inflatedness at the extreme tail.

Appendix S4

Simulation 4: Other correlation structures

Apart from the compound symmetry (CS) structure, we also considered AR1(ρ) and other structures for correlation in our simulation studies. Details on how the datasets were simulated can be found in Section 3 of our main paper.

Correlation Structure I: uncorrelated traits : We assumed that none of the traits was correlated with another. From Figure S5, we see that performances of all methods are similar except minP/TATES. All the methods, including MANOVA, have steadily rising power curves with increase in proportion of associated traits. This confirms that MANOVA's lack of power in detecting pleiotropy in certain situations is primarily due to the correlatedness of all the traits.

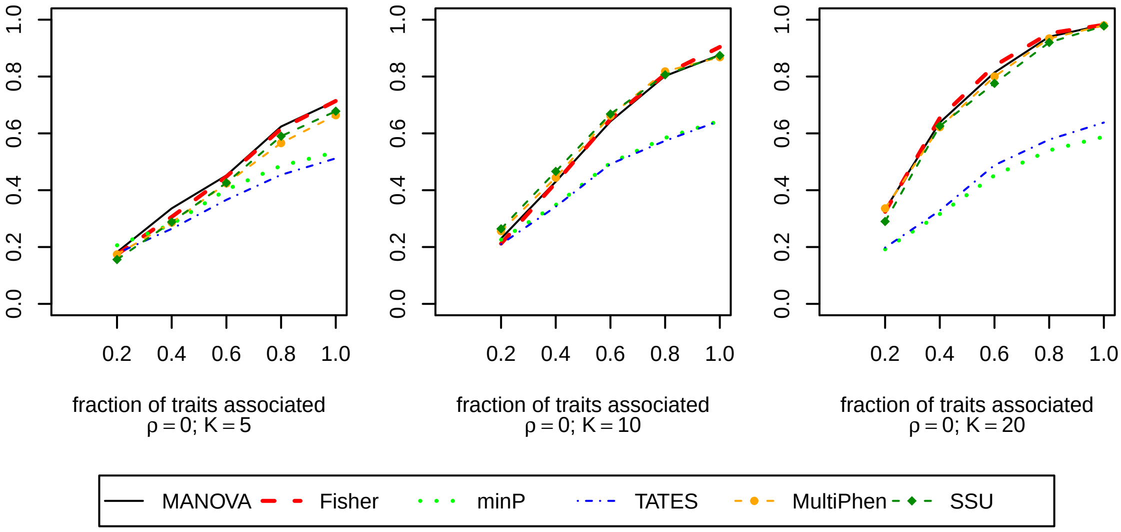


Figure S5: Correlation structure I (uncorrelated): Empirical power curves of the different association tests for $K = 5, 10, 20$ traits and within trait correlation $\rho = 0$ based on $N = 500$ datasets. The correlation structure assumes all traits to be uncorrelated. Same direction and same size effects (effect size of 0.395; proportion of variance explained is 0.5%) are used when 2 or more traits are associated. The power is plotted along y-axis while the fraction of traits associated with the genetic variant is plotted along x-axis.

Correlation Structure II: Here we assumed that first 80% of the K traits were correlated (with a compound symmetry structure) and the rest 20% were uncorrelated. For our simulation study, we considered $K = 5, 10, 20$ traits and positive correlation param-

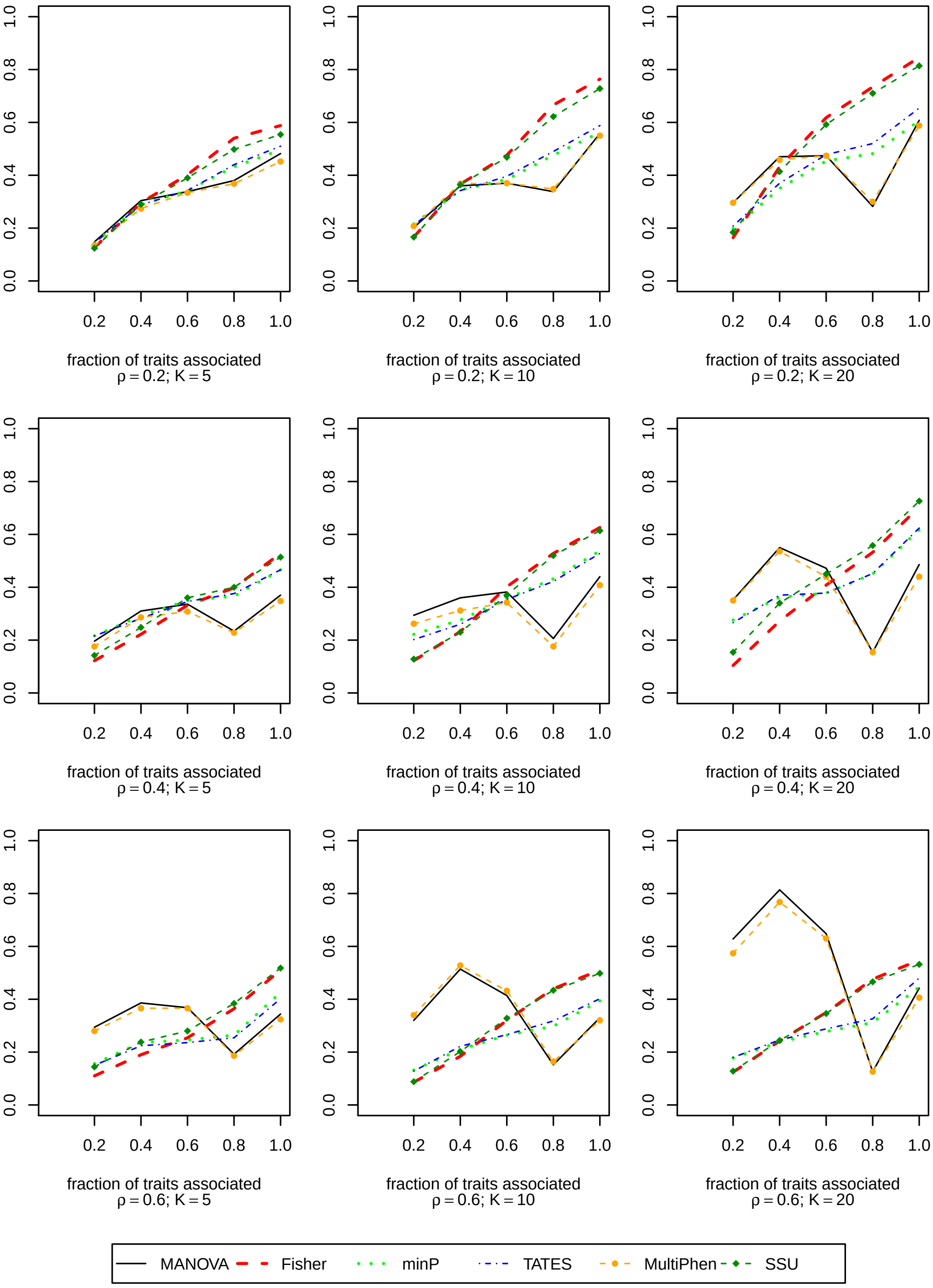


Figure S6: **Correlation structure II**: Empirical power curves of the different association tests for $K = 5, 10, 20$ traits and different within trait correlation values $\rho = 0.2, 0.4, 0.6$ based on $N = 500$ datasets. This correlation structure assumes that the first 80% of the traits are correlated (Compound Symmetry structure with correlation ρ) and the last 20% of the traits are independent of the others. Same direction and same size effects (effect size of 0.395; proportion of variance explained is 0.5%) are used when 2 or more traits are associated. The power is plotted along y-axis while the fraction of traits associated with the genetic variant is plotted along x-axis. Upto the point 0.8 on the x-axis, all the traits are correlated.

eter $\rho = 0.2, 0.4, 0.6$. In such a situation we noticed that as correlation increased among the associated traits, the power of MANOVA dropped. Figure S6 shows that the lowest point in the MANOVA power curve occurs at 0.8 on the axis, which means MANOVA has the least power in detecting association when all the correlated traits are associated. At point 1.0 on the x-axis, when all the traits are associated but not all are correlated, the performance of MANOVA improves but not as good as the methods that do not explicitly consider the covariance matrix in the test statistic.

An important observation from Figure S6 is that MANOVA is not expected to suffer from power loss at 'complete association' (when all traits are associated) if all associated traits are not correlated (refer Appendix S5 for theoretical result).

Correlation Structure III: AR1(ρ) : For given K traits, we assumed the covariance structure $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{R}(\rho) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{K-1} \\ \rho & 1 & \rho & \dots & \rho^{K-2} \\ \vdots & & & \ddots & \vdots \\ \rho^{K-1} & \rho^{K-2} & \rho^{K-3} & \dots & 1 \end{pmatrix}$. Figure S7 shows that for a given ρ, MANOVA performs better with increase in K and with increase in the fraction of associated traits. This is so because at a higher fraction (on the x-axis), the AR1 correlation among traits becomes negligible and the latter traits are effectively uncorrelated (the behavior we saw in Figures S5 & S6). Observe that for a given ρ, the power at or near 'complete association' (where all traits are associated) increases with increase in K since for the latter traits, the correlation rapidly goes towards 0. With increase in the parameter ρ and for small K, we start observing MANOVA's lack of power as the latter pairwise correlations are not effectively zero.

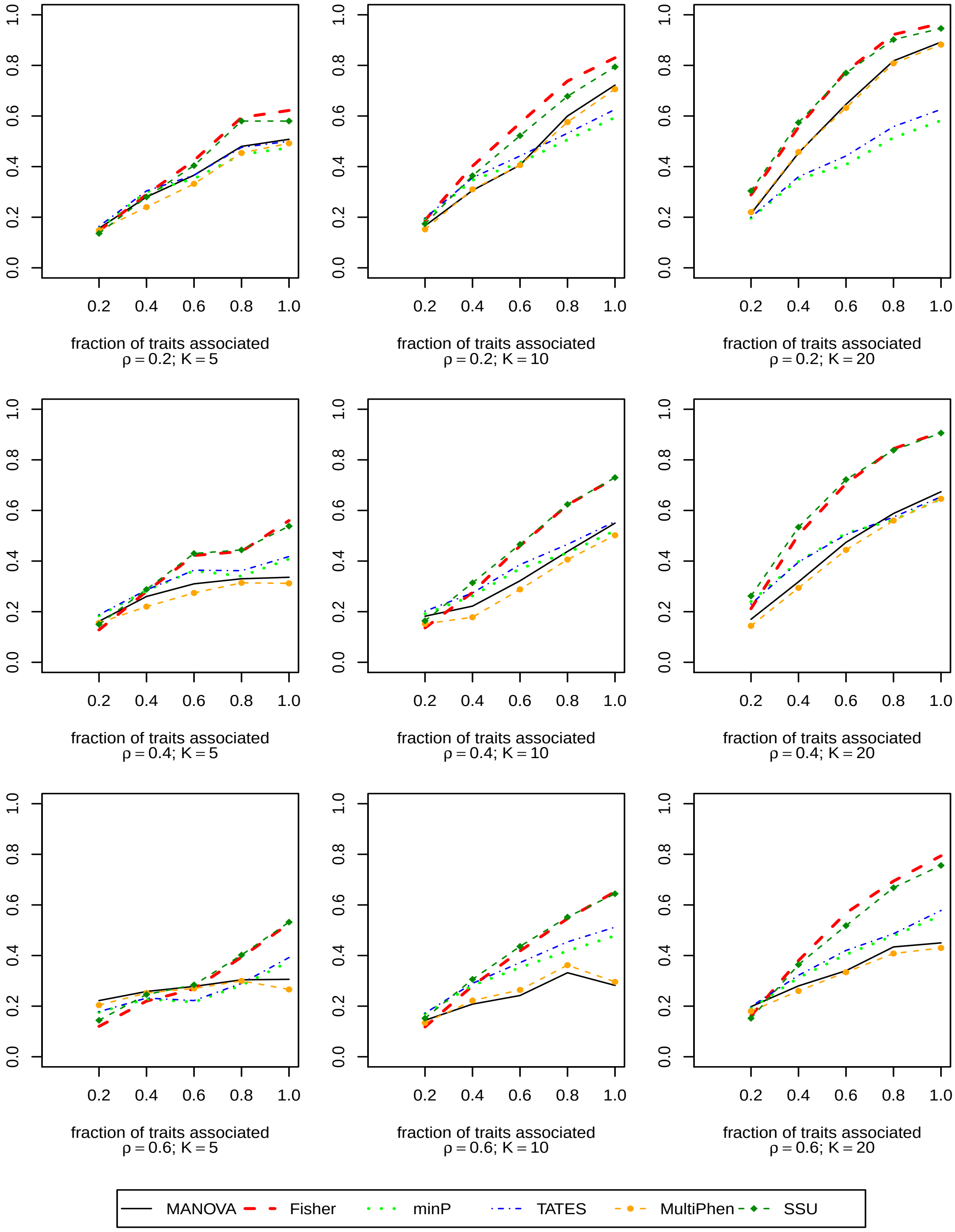


Figure S7: Correlation structure III (AR1): Empirical power curves of the different association tests based on $N = 500$ datasets for $K = 5, 10, 20$ traits and AR1(ρ) correlation structure with $\rho = 0.2, 0.4, 0.6$. Same direction and same size effects (effect size of 0.395; proportion of variance explained is 0.5%) are used when 2 or more traits are associated. The power is plotted along y-axis while the fraction of traits associated with the genetic variant is plotted along x-axis.

Appendix S5

Figure S6 shows that if all the traits are not correlated, MANOVA does not experience power loss for testing H_0 even when all the traits are associated. This behavior is theoretically explained by the following theorem for the special case of CS residual correlation structure for the correlated traits.

Theorem. *Without loss of generality, let $\boldsymbol{Y}$ and $\boldsymbol{X}$ be the centered phenotype matrix and the centered genotype vector respectively. Consider the MMLR model*

$$\boldsymbol{Y}_{n\times K} = \boldsymbol{X}_{n\times 1}\boldsymbol{\beta}'_{1\times K} + \boldsymbol{\mathcal{E}}_{n\times K},\ \ vec(\boldsymbol{\mathcal{E}}) \sim N_{nK}(\boldsymbol{0}, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma}_{K\times K} = \begin{pmatrix} \boldsymbol{\Sigma}_{11(m\times m)} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$, $\boldsymbol{\Sigma}_{11} = \sigma^2\left((1-\rho)\boldsymbol{I}_m + \rho\boldsymbol{1}\boldsymbol{1}'\right)$, $\sigma^2 > 0$, $\rho\ (> 0)$ *is the within trait correlation such that* $\boldsymbol{\Sigma}_{11}$ *is a positive definite covariance matrix,* $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}'_{21} = \mathrm{O}_{m\times(K-m)}$, $\boldsymbol{\Sigma}_{22} = \sigma^2\boldsymbol{I}_{K-m}$ *and* $\boldsymbol{\beta}' = (\beta_1, ..., \beta_K)$ *is the vector of genetic effects. Assume that the genetic effects of the associated traits are equal in size and positive. Consider two scenarios of association: 'partial association' (when the SNP is associated with u ($< K$) traits), and 'complete association' (when all K traits are associated).*

For testing $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$, MANOVA is not expected to suffer from power loss at 'complete association' compared to 'partial association' with u ($> m$) associated traits.

Proof. For the $m \times m$ CS residual covariance sub-matrix $\boldsymbol{\Sigma}_{11}$, we know that the eigen vector corresponding to the largest eigenvalue $\lambda_{(m)1} = \sigma^2\{1 + (m-1)\rho\}$ is $\boldsymbol{v}_1 \propto \boldsymbol{1}$, while the eigen vectors corresponding to $\lambda_{(m)2} = ... = \lambda_{(m)m} = \sigma^2(1-\rho)$ are respectively $\boldsymbol{v}_2, ..., \boldsymbol{v}_m$ such that $\boldsymbol{1}'\boldsymbol{v}_k = 0\,\forall\, k = 2, ..., m$. For the eigen vectors to be orthonormal, we must have $\boldsymbol{v}_1 = c_m\boldsymbol{1}$ such that $\sqrt{c_m^2 + ... + c_m^2} = 1 \iff c_m^2 = 1/m$. Thus, we can write,

$$\boldsymbol{\Sigma}_{11(m\times m)} = \lambda_{(m)1}c_m^2\boldsymbol{1}\boldsymbol{1}' + \sum_{i=2}^{m}\lambda_{(m)i}\boldsymbol{v}_i\boldsymbol{v}_i' \text{ and } \boldsymbol{\Sigma}_{11}^{-1} = \frac{1}{\lambda_{(m)1}}c_m^2\boldsymbol{1}\boldsymbol{1}' + \sum_{i=2}^{m}\frac{1}{\lambda_{(m)i}}\boldsymbol{v}_i\boldsymbol{v}_i'$$

Consider the 2 alternatives $H_{a,u} : \beta_1 = ... = \beta_u \neq 0, \beta_{K-u} = ... = \beta_K = 0$ (partial association) and $H_{a,K} : \beta_1 = ... = \beta_K \neq 0$ (complete association) against the null hypothesis $H_0 : \beta_1 = ... = \beta_K = 0$. Here, for the partial association case, $u\ (> m)$ is the number of traits associated and m is the number of correlated traits. In the following, the notation $\overset{P}{\to}$ denotes convergence in probability as $n \to \infty$.

Under the alternative $H_{a,K}$ (complete association), it can be shown that

$$\begin{aligned}
\left|\boldsymbol{I} + \frac{\boldsymbol{H}_K}{n}\left(\frac{\boldsymbol{E}}{n}\right)^{-1}\right| &\overset{P}{\to} \left|\boldsymbol{I}_K + (2pq\beta_1^2)\begin{pmatrix} \mathbf{1}_m\mathbf{1}_m' & \mathbf{1}_m\mathbf{1}_{K-m}' \\ \mathbf{1}_{K-m}\mathbf{1}_m' & \mathbf{1}_{K-m}\mathbf{1}_{K-m}' \end{pmatrix}\begin{pmatrix} \boldsymbol{\Sigma}_{11}^{-1} & \mathrm{O} \\ \mathrm{O} & \frac{1}{\sigma^2}\boldsymbol{I}_{K-m} \end{pmatrix}\right| \\
&= \left|(\boldsymbol{I}_m + a\mathbf{1}_m\mathbf{1}_m') - (b\mathbf{1}_m\mathbf{1}_{K-m}')(\boldsymbol{I}_{K-m} + b\mathbf{1}_{K-m}\mathbf{1}_{K-m}')^{-1}(a\mathbf{1}_{K-m}\mathbf{1}_m')\right| \\
&\quad \times|\boldsymbol{I}_{K-m} + b\mathbf{1}\mathbf{1}'| \\
&= |(\boldsymbol{I}_m + a\mathbf{1}_m\mathbf{1}_m') - ac(K-m)\mathbf{1}_m\mathbf{1}_m'| \times |\boldsymbol{I}_{K-m} + b\mathbf{1}\mathbf{1}'| \\
&= 1 + b(K-m) + am
\end{aligned}$$

where $a = \frac{2pq\beta_1^2}{\sigma^2\{1+(m-1)\rho\}}$, $b = \frac{2pq\beta_1^2}{\sigma^2}$, $(\boldsymbol{I}_{K-m} + b\mathbf{1}_{K-m}\mathbf{1}_{K-m}')^{-1} = \boldsymbol{I} - c\mathbf{1}\mathbf{1}'$, $c = \frac{b}{1+(K-m)b}$.

For $u(> m)$ associated traits, let us now partition the residual covariance matrix as

$$\boldsymbol{\Sigma}_{K\times K} = \begin{pmatrix} \boldsymbol{S}_{11(u\times u)} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{12}' & \boldsymbol{S}_{22} \end{pmatrix} \text{ where } \boldsymbol{S}_{11} = \begin{pmatrix} \boldsymbol{\Sigma}_{11(m\times m)} & \mathrm{O} \\ \mathrm{O} & \sigma^2\boldsymbol{I}_{u-m} \end{pmatrix}, \boldsymbol{S}_{12} = \mathrm{O}_{u\times(K-u)}, \boldsymbol{S}_{22} = \sigma^2\boldsymbol{I}_{K-u}$$

Under the alternative $H_{a,u}$ (partial association) where $0 < m < u < K$, one can show that

$$\begin{aligned}
\left|\boldsymbol{I} + \frac{\boldsymbol{H}_u}{n}\left(\frac{\boldsymbol{E}}{n}\right)^{-1}\right| &\overset{P}{\to} \left|\boldsymbol{I}_K + 2pq\begin{pmatrix} \beta_1^2\mathbf{1}_u\mathbf{1}_u' & \mathrm{O} \\ \mathrm{O} & \mathrm{O} \end{pmatrix}\begin{pmatrix} \boldsymbol{S}_{11}^{-1} & \mathrm{O} \\ \mathrm{O} & \frac{1}{\sigma^2}\boldsymbol{I}_{K-u} \end{pmatrix}\right| \\
&= \left|\boldsymbol{I}_u + 2pq\beta_1^2\mathbf{1}_u\mathbf{1}_u'\begin{pmatrix} \boldsymbol{\Sigma}_{11}^{-1} & \mathrm{O} \\ \mathrm{O} & \frac{1}{\sigma^2}\boldsymbol{I}_{u-m} \end{pmatrix}\right| \\
&= 1 + b(u-m) + am, \text{ where } a = \frac{2pq\beta_1^2}{\sigma^2\{1+(m-1)\rho\}}, b = \frac{2pq\beta_1^2}{\sigma^2}
\end{aligned}$$

$$\therefore\ \left|\boldsymbol{I}_K + \boldsymbol{H}_K\boldsymbol{E}^{-1}\right| - \left|\boldsymbol{I}_K + \boldsymbol{H}_u\boldsymbol{E}^{-1}\right| \overset{P}{\to} b(K-u) > 0$$

■

Appendix S6

Details on ARIC Study phenotypes and covariate choices

ARIC has collected measures on many type 2 diabetes (T2D) related traits at 4 separate visits over a 9-year period. A diagnosis of T2D is considered positive if fasting plasma glucose concentration is ≥ 126 mg/dL, or casual plasma glucose level is ≥ 200 mg/dL, or 2-hour plasma glucose value after a standard glucose challenge is ≥ 200 mg/dL (WHO, 2003). All analytes were determined at central laboratories according to standard protocols: plasma glucose by a hexokinase assay, and insulin by radioimmunoassay (^{125}Insulin Kit; Cambridge Medical Diagnosis, Billerica, MA). Sedentary lifestyle and obesity are major risk factors for T2D. In addition to general obesity, the distribution of body fat (or abdominal obesity, as estimated by waist-to-hip circumference ratio) contributes to T2D risk. For our analysis, we focused on the Caucasian participants and the following 3 T2D related quantitative traits measured at visit 4 ($1996-98$): fasting glucose; 2-hour glucose from an oral glucose tolerance test; fasting insulin. The pairwise correlations among these 3 traits were within $(0.2, 0.35)$. These traits are substantially affected by treatment with diabetes medications, and so statistical analysis results are not generally interpretable in the same way they can be interpreted in non-diabetic individuals. Other available traits were Body Mass Index (BMI) and waist circumference (WC). WC was measured at the umbilical level. BMI was calculated as weight/height^2 (kg/m^2), and obesity was defined as a BMI ≥ 30 kg/m^2. BMI, being a major risk factor for T2D, is traditionally adjusted as a covariate in association analysis of glycemic traits (Manning et al., 2012; Scott et al., 2012; Dupuis et al., 2010, for example). Manning et al. (2012) notes that "adiposity may also hinder the identification of genetic variants influencing insulin resistance by introducing variance in the outcome that is not attributable to genetic variation, suggesting that adjustment for adiposity *per se* may be necessary".

Due to a high pairwise correlation of 0.9 between WC and BMI, we chose to adjust BMI only. When BMI is adjusted, inclusion of WC or any other adiposity trait in the multivariate response makes it difficult to interpret analysis results. We chose not to include the adiposity traits (BMI, WC, waist-hip ratio, hip circumference) along with the glycemic traits (fasting glucose, 2-hour glucose, fasting insulin) in the response vector because not many SNPs have been reported to jointly influence both adiposity traits and glycemic traits. As in most studies of T2D, BMI was used as a covariate along with age and sex. Individuals with diagnosed or treated diabetes at visit 4 were removed. Since USAT requires complete phenotype data, individuals with missing traits were excluded too, leaving $5,816$ in our analytic sample.

Appendix S7

Covariate Adjustment for USAT

The ARIC data analysis using USAT required covariate adjustment (predictors other than SNP). This version of USAT requires covariate adjustment for both SSU test and MANOVA. Once the adjusted MANOVA and SSU test statistics are available, one can easily compute approximate p-value for USAT (refer section 2.5 of the main paper for the p-value calculation method). Let $\boldsymbol{Z}_{n\times q}$ be the matrix of q covariates (other than SNP) for n unrelated individuals. Without loss of generality, the phenotype matrix $\boldsymbol{Y}$, the genotype vector $\boldsymbol{X}$ and the covariate matrix $\boldsymbol{Z}$ are centered (but not scaled). The following paragraphs outline the details of such covariate adjustment.

MANOVA with covariate adjustment

The MMLR model for the association test of K traits and the SNP (after adjusting for other covariates):

$$\boldsymbol{Y}_{n\times K} = \boldsymbol{X}_{n\times 1}\boldsymbol{\beta}'_{1\times K} + \boldsymbol{1}' \otimes \boldsymbol{Z\Phi} + \boldsymbol{\mathcal{E}}_{n\times K}$$

where $\boldsymbol{\beta}' = (\beta_1, ..., \beta_K)$ is the vector of fixed unknown genetic effects corresponding to the K correlated traits, and $\boldsymbol{\mathcal{E}}$ is the matrix of random errors. For testing that the SNP is not associated with any of the K traits, the null hypothesis of interest is $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$. For testing H_0, the LRT is equivalent to the MANOVA test statistic, which is the ratio of generalized variances $\boldsymbol{\Lambda} = |\boldsymbol{E}|/|\boldsymbol{H} + \boldsymbol{E}|$. Here, $\boldsymbol{H} + \boldsymbol{E}$ is the covariance matrix of the K residual vectors where the k-th residual vector is obtained by fitting the model for k-th trait under H_0. $\boldsymbol{E}$ is the covariance matrix of the K residual vectors where the k-th residual vector is obtained by fitting the full model for k-th trait. Under H_0, Wilk's Lambda $-2\log\boldsymbol{\Lambda}$ has an approximate asymptotic χ^2_K distribution under H_0.

SSU Test with covariate adjustment

For k-th trait vector, we assume the marginal normal model :

$$\boldsymbol{Y}_k = \beta_k \boldsymbol{X} + \boldsymbol{Z\Phi} + \boldsymbol{\epsilon}_k, \;\; \boldsymbol{\epsilon}_k \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$$

β_k is the parameter associated with the SNP effect on the k-th trait. $\boldsymbol{\Phi}$ is the $q \times 1$ vector of parameters associated with the q covariates. The null hypothesis associated with k-th marginal model is $H_{0k} : \beta_k = 0$. We need to obtain the MLE $\hat{\boldsymbol{\Phi}}$ under the global null $H_0 : \cap_{k=1}^{K} H_{0,k}$. Under $H_{0,k}$, the k-th marginal model is

$$\boldsymbol{Y}_k = \boldsymbol{Z\Phi} + \boldsymbol{\epsilon}_k, \;\; \boldsymbol{\epsilon}_k \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$$

The MLE of $\boldsymbol{\Phi}$ from k-th model is $\hat{\boldsymbol{\Phi}}_{(k)} = (\boldsymbol{Z'Z})^{-1}\boldsymbol{Z'Y}_k$. Thus, MLE of $\boldsymbol{\Phi}$ under H_0 is

$$\hat{\boldsymbol{\Phi}} = \frac{1}{K} \sum_{k=1}^{K} (\boldsymbol{Z'Z})^{-1}\boldsymbol{Z'Y}_k$$

The MLE of σ^2 under H_0 is given by

$$\hat{\sigma}_0^2 = \frac{1}{nK} \sum_{k=1}^{K} (\boldsymbol{Y}_k - \boldsymbol{Z}\hat{\boldsymbol{\Phi}})'(\boldsymbol{Y}_k - \boldsymbol{Z}\hat{\boldsymbol{\Phi}})$$

The log-likelihood for the k-th genetic effect from the k-th marginal model is given by

$$l(\beta_k) \propto -\frac{1}{2\sigma^2}(\boldsymbol{Y}_k - \beta_k\boldsymbol{X} - \boldsymbol{Z\Phi})'(\boldsymbol{Y}_k - \beta_k\boldsymbol{X} - \boldsymbol{Z\Phi})$$

Marginal score for parameter β_k under H_{0k}:

$$U_k = \dot{l}(\beta_k)\Big|_{H_0} = \frac{1}{\sigma^2}(\boldsymbol{Y}_k - \beta_k\boldsymbol{X} - \boldsymbol{Z\Phi})'\boldsymbol{X}\Big|_{H_0} = \frac{1}{\hat{\sigma}_0^2}(\boldsymbol{Y}_k - \boldsymbol{Z}\hat{\boldsymbol{\Phi}})'\boldsymbol{X}$$

Under the null, the variances and covariances of the marginal scores are:

$$\mathrm{Var}(U_k) = \frac{1}{\sigma^4}\boldsymbol{X}' \,\mathrm{Var}(\boldsymbol{Y}_k)\boldsymbol{X}\Big|_{H_0} = \frac{1}{\sigma^2}\boldsymbol{X'X}\Big|_{H_0}$$

$$\mathrm{Cov}(U_k, U_j) = \frac{1}{\sigma^4}\,\mathrm{E}(\boldsymbol{Y}'_k\boldsymbol{X} \times \boldsymbol{Y}'_j\boldsymbol{X})\Big|_{H_0} = \frac{1}{\sigma^4}\boldsymbol{X}'\,\mathrm{E}(\boldsymbol{Y}_k\boldsymbol{Y}'_j)\boldsymbol{X}\Big|_{H_0} = \frac{\rho}{\sigma^2}\boldsymbol{X}'\boldsymbol{X}\Big|_{H_0}\ \forall\, j \neq k$$

Thus, under H_0, the score vector from the marginal normal model for $\boldsymbol{Y}$ is

$$\boldsymbol{U}_M = \left(\boldsymbol{Y} - \boldsymbol{1}' \otimes \boldsymbol{Z}\hat{\boldsymbol{\Phi}}\right)' \boldsymbol{X}/\hat{\sigma}_0^2$$

with covariance

$$\mathrm{Cov}(\boldsymbol{U}_M) = \frac{1}{\sigma^4}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\Sigma}\Big|_{H_0} = \frac{1}{\hat{\sigma}_0^4}(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{\hat{\sigma}_0^4}(\boldsymbol{X}'\boldsymbol{X})\frac{\left(\boldsymbol{Y} - \boldsymbol{1}' \otimes \boldsymbol{Z}\hat{\boldsymbol{\Phi}}\right)'\left(\boldsymbol{Y} - \boldsymbol{1}' \otimes \boldsymbol{Z}\hat{\boldsymbol{\Phi}}\right)}{n}$$

The SSU test based on the marginal normal score vector $\boldsymbol{U}_M$ is

$$T_S = \boldsymbol{U}'_M\boldsymbol{U}_M \overset{approx}{\sim} a\chi_d^2 + b$$

where parameters a, b, d are estimated as

$$a = \frac{\sum \delta_i^3}{\sum \delta_i^2}, b = \sum \delta_i - \frac{(\sum \delta_i^2)^2}{\sum \delta_i^3}, d = \frac{(\sum \delta_i^2)^3}{(\sum \delta_i^3)^2}, \{\delta_i\}_{i=1}^K \text{ are the ordered eigenvalues of } \mathrm{Cov}(\boldsymbol{U}_M)$$

Appendix S8

Table S3: List of all SNPs that exceed the genome-wide significance threshold 5×10^{-8} for the multivariate methods USAT and MANOVA. SNPs with m.a.f. $< 5\%$ were screened out. It is to be noted that most of these SNPs are in high linkage disequilibrium (LD). p values for the univariate analysis of the individual traits are also provided. SNPs in bold are the ones detected solely by MANOVA but not by USAT. The abbreviations used are FG (Fasting Glucose), 2-hr GL (2-hour glucose from an oral glucose tolerance test), FI (Fasting Insulin)

			MANOVA	USAT	Univariate Analysis p		
chr	SNP	position	p	p	FG	2-hr GL	FI
2	rs1260326	27584444	3.77×10^{-15}	4.44×10^{-15}	1.24×10^{-4}	6.26×10^{-6}	1.24×10^{-5}
2	rs780094	27594741	9.99×10^{-16}	1.67×10^{-15}	7.34×10^{-5}	7.10×10^{-6}	4.65×10^{-6}
2	rs780093	27596107	9.99×10^{-16}	1.67×10^{-15}	7.34×10^{-5}	7.10×10^{-6}	4.65×10^{-6}
2	rs1260333	27602128	4.72×10^{-11}	8.14×10^{-11}	4.99×10^{-4}	3.84×10^{-4}	7.59×10^{-5}
2	rs2911711	27604050	4.72×10^{-11}	8.14×10^{-11}	4.99×10^{-4}	3.84×10^{-4}	7.59×10^{-5}
2	rs4665987	27609329	9.67×10^{-10}	2.31×10^{-9}	1.56×10^{-2}	4.49×10^{-6}	1.46×10^{-2}
2	rs4665991	27619788	1.23×10^{-9}	2.57×10^{-9}	1.75×10^{-2}	4.83×10^{-6}	1.44×10^{-2}
2	rs4665382	27637305	1.20×10^{-9}	2.54×10^{-9}	1.52×10^{-2}	5.13×10^{-6}	1.60×10^{-2}
2	rs10208529	27639692	1.20×10^{-9}	2.54×10^{-9}	1.52×10^{-2}	5.13×10^{-6}	1.60×10^{-2}
2	rs4665383	27645059	1.20×10^{-9}	2.54×10^{-9}	1.52×10^{-2}	5.13×10^{-6}	1.60×10^{-2}
2	rs1919127	27654997	1.20×10^{-9}	2.54×10^{-9}	1.52×10^{-2}	5.13×10^{-6}	1.60×10^{-2}
2	rs1919128	27655263	1.20×10^{-9}	2.54×10^{-9}	1.52×10^{-2}	5.13×10^{-6}	1.60×10^{-2}
2	rs12478841	27665226	1.06×10^{-9}	2.40×10^{-9}	1.63×10^{-2}	4.96×10^{-6}	1.31×10^{-2}
2	rs6760250	27665756	9.94×10^{-10}	2.34×10^{-9}	1.49×10^{-2}	6.01×10^{-6}	1.07×10^{-2}
2	rs13022873	27669014	9.94×10^{-10}	2.34×10^{-9}	1.49×10^{-2}	6.01×10^{-6}	1.07×10^{-2}
2	rs12467476	27679219	9.86×10^{-10}	2.33×10^{-9}	1.53×10^{-2}	6.00×10^{-6}	1.02×10^{-2}
2	rs2384656	27685559	9.86×10^{-10}	2.33×10^{-9}	1.53×10^{-2}	6.00×10^{-6}	1.02×10^{-2}
2	rs4666002	27694144	8.64×10^{-10}	1.43×10^{-9}	1.61×10^{-2}	5.56×10^{-6}	9.02×10^{-3}
2	rs3749147	27705422	6.52×10^{-9}	1.31×10^{-8}	3.65×10^{-2}	6.56×10^{-6}	1.86×10^{-2}
2	rs13002853	27706749	6.52×10^{-9}	1.31×10^{-8}	3.65×10^{-2}	6.56×10^{-6}	1.86×10^{-2}
2	rs13431652	169461661	1.85×10^{-13}	5.48×10^{-13}	2.24×10^{-12}	9.57×10^{-1}	2.85×10^{-1}
2	**rs12475700**	169461922	4.52×10^{-8}	8.76×10^{-8}	1.07×10^{-8}	3.62×10^{-1}	6.73×10^{-1}
2	rs1402837	169465600	4.91×10^{-9}	1.15×10^{-8}	2.78×10^{-10}	5.18×10^{-2}	8.07×10^{-1}
2	rs573225	169465787	4.55×10^{-14}	5.81×10^{-14}	9.75×10^{-13}	9.83×10^{-1}	2.33×10^{-1}
2	rs560887	169471394	5.55×10^{-16}	1.33×10^{-15}	1.24×10^{-14}	8.87×10^{-1}	2.93×10^{-1}
2	rs563694	169482317	1.54×10^{-14}	2.91×10^{-14}	4.12×10^{-14}	3.50×10^{-1}	3.54×10^{-1}

. . . continued

			MANOVA	USAT	Univariate Analysis p		
chr	SNP	position	p	p	FG	2-hr GL	FI
2	rs537183	169482892	1.54×10^{-14}	2.91×10^{-14}	4.12×10^{-14}	3.50×10^{-1}	3.54×10^{-1}
2	rs502570	169483205	1.54×10^{-14}	2.91×10^{-14}	4.12×10^{-14}	3.50×10^{-1}	3.54×10^{-1}
2	rs475612	169484992	3.54×10^{-13}	7.12×10^{-13}	3.74×10^{-13}	3.63×10^{-1}	5.12×10^{-1}
2	rs557462	169485841	1.54×10^{-14}	2.91×10^{-14}	4.12×10^{-14}	3.50×10^{-1}	3.54×10^{-1}
2	rs478333	169487402	8.61×10^{-10}	1.42×10^{-9}	3.33×10^{-10}	4.16×10^{-1}	6.41×10^{-1}
2	rs496550	169487958	8.61×10^{-10}	1.42×10^{-9}	3.33×10^{-10}	4.16×10^{-1}	6.41×10^{-1}
2	rs473351	169488142	2.58×10^{-11}	6.05×10^{-11}	1.38×10^{-11}	2.75×10^{-1}	5.29×10^{-1}
2	rs575671	169489064	2.58×10^{-11}	6.05×10^{-11}	1.38×10^{-11}	2.75×10^{-1}	5.29×10^{-1}
2	rs519887	169489131	8.50×10^{-10}	1.41×10^{-9}	3.45×10^{-10}	4.40×10^{-1}	6.41×10^{-1}
2	rs486981	169490395	2.72×10^{-14}	4.05×10^{-14}	1.15×10^{-13}	6.17×10^{-1}	3.89×10^{-1}
2	rs484066	169490727	2.01×10^{-12}	2.32×10^{-12}	6.10×10^{-12}	8.54×10^{-1}	4.87×10^{-1}
2	rs569805	169491126	2.72×10^{-14}	4.05×10^{-14}	1.15×10^{-13}	6.17×10^{-1}	3.89×10^{-1}
2	rs579060	169491285	2.45×10^{-14}	3.77×10^{-14}	9.95×10^{-14}	6.04×10^{-1}	3.93×10^{-1}
2	rs17540154	169492739	3.46×10^{-9}	6.48×10^{-9}	3.41×10^{-10}	2.33×10^{-1}	9.19×10^{-1}
2	rs508506	169493201	3.55×10^{-14}	4.86×10^{-14}	8.64×10^{-14}	5.74×10^{-1}	4.97×10^{-1}
2	rs503931	169493695	8.50×10^{-10}	1.41×10^{-9}	3.45×10^{-10}	4.40×10^{-1}	6.41×10^{-1}
2	rs551754	169495932	8.50×10^{-10}	1.41×10^{-9}	3.45×10^{-10}	4.40×10^{-1}	6.41×10^{-1}
2	rs497692	169497262	8.41×10^{-10}	1.41×10^{-9}	3.27×10^{-10}	4.24×10^{-1}	6.46×10^{-1}
2	rs494874	169497552	1.42×10^{-13}	5.06×10^{-13}	1.70×10^{-13}	5.38×10^{-1}	6.48×10^{-1}
2	rs552976	169499684	1.55×10^{-13}	5.19×10^{-13}	1.87×10^{-13}	5.31×10^{-1}	6.37×10^{-1}
2	rs472614	169500667	1.26×10^{-8}	2.65×10^{-8}	3.55×10^{-9}	5.72×10^{-1}	8.17×10^{-1}
2	rs565412	169502529	9.14×10^{-9}	1.57×10^{-8}	4.16×10^{-9}	7.18×10^{-1}	7.34×10^{-1}
2	rs567074	169502677	3.45×10^{-10}	5.76×10^{-10}	1.51×10^{-10}	6.13×10^{-1}	8.04×10^{-1}
2	rs479682	169502933	7.13×10^{-9}	1.37×10^{-8}	3.33×10^{-9}	6.89×10^{-1}	7.06×10^{-1}
2	rs480562	169503017	7.46×10^{-9}	1.40×10^{-8}	3.43×10^{-9}	6.84×10^{-1}	7.07×10^{-1}
2	rs2685803	169504531	7.46×10^{-9}	1.40×10^{-8}	3.43×10^{-9}	6.84×10^{-1}	7.07×10^{-1}
2	rs2544367	169504534	4.54×10^{-9}	7.54×10^{-9}	2.46×10^{-9}	7.19×10^{-1}	6.85×10^{-1}
2	rs2685805	169505306	4.54×10^{-9}	7.54×10^{-9}	2.46×10^{-9}	7.19×10^{-1}	6.85×10^{-1}
2	rs1581397	169505898	4.05×10^{-9}	7.05×10^{-9}	2.40×10^{-9}	7.37×10^{-1}	6.66×10^{-1}
2	rs2685814	169506865	3.62×10^{-9}	6.63×10^{-9}	2.19×10^{-9}	7.48×10^{-1}	6.71×10^{-1}
2	**rs6709087**	169507256	3.87×10^{-8}	8.13×10^{-8}	2.55×10^{-9}	2.67×10^{-1}	6.74×10^{-1}
2	rs853789	169509734	2.00×10^{-15}	2.66×10^{-15}	8.50×10^{-15}	6.36×10^{-1}	4.84×10^{-1}

. . . continued

			MANOVA	USAT	Univariate Analysis p		
chr	SNP	position	p	p	FG	2-hr GL	FI
2	*rs*860510	169509874	3.62×10^{-9}	6.63×10^{-9}	2.19×10^{-9}	7.48×10^{-1}	6.71×10^{-1}
2	*rs*853788	169510151	3.62×10^{-9}	6.63×10^{-9}	2.19×10^{-9}	7.48×10^{-1}	6.71×10^{-1}
2	*rs*853787	169510498	2.00×10^{-15}	2.66×10^{-15}	8.50×10^{-15}	6.36×10^{-1}	4.84×10^{-1}
2	*rs*853786	169510556	3.62×10^{-9}	6.63×10^{-9}	2.19×10^{-9}	7.48×10^{-1}	6.71×10^{-1}
2	*rs*862662	169510575	1.52×10^{-10}	2.42×10^{-10}	1.01×10^{-10}	6.80×10^{-1}	7.17×10^{-1}
2	*rs*853785	169510840	3.62×10^{-9}	6.63×10^{-9}	2.19×10^{-9}	7.48×10^{-1}	6.71×10^{-1}
2	*rs*853784	169511920	5.48×10^{-9}	1.21×10^{-8}	3.37×10^{-9}	8.18×10^{-1}	7.11×10^{-1}
2	*rs*853783	169513757	5.48×10^{-9}	1.21×10^{-8}	3.37×10^{-9}	8.18×10^{-1}	7.11×10^{-1}
2	*rs*853781	169514567	3.43×10^{-10}	5.74×10^{-10}	2.19×10^{-10}	7.37×10^{-1}	7.53×10^{-1}
2	*rs*853780	169515728	7.62×10^{-9}	1.42×10^{-8}	4.68×10^{-9}	7.98×10^{-1}	6.76×10^{-1}
2	*rs*1101533	169516768	7.62×10^{-9}	1.42×10^{-8}	4.68×10^{-9}	7.98×10^{-1}	6.76×10^{-1}
2	*rs*853779	169517918	4.39×10^{-9}	7.39×10^{-9}	2.88×10^{-9}	7.91×10^{-1}	6.67×10^{-1}
2	*rs*853778	169519470	1.28×10^{-9}	2.62×10^{-9}	1.52×10^{-9}	9.11×10^{-1}	5.79×10^{-1}
2	*rs*853773	169522593	1.63×10^{-9}	2.96×10^{-9}	2.44×10^{-9}	8.12×10^{-1}	7.64×10^{-1}
15	***rs*17271144**	59920213	4.47×10^{-8}	8.71×10^{-8}	6.21×10^{-2}	4.17×10^{-6}	1.06×10^{-1}
15	***rs*3743297**	59937076	4.59×10^{-8}	8.84×10^{-8}	1.09×10^{-2}	2.87×10^{-5}	2.49×10^{-1}
15	***rs*1981916**	59958771	3.14×10^{-8}	7.42×10^{-8}	8.50×10^{-3}	2.89×10^{-5}	2.56×10^{-1}
15	***rs*2414755**	59959721	3.14×10^{-8}	7.42×10^{-8}	8.50×10^{-3}	2.89×10^{-5}	2.56×10^{-1}
15	***rs*2042608**	60019672	3.02×10^{-8}	7.29×10^{-8}	1.71×10^{-3}	1.16×10^{-4}	0.59
15	***rs*7170293**	60023665	1.98×10^{-8}	6.28×10^{-8}	5.77×10^{-3}	3.20×10^{-5}	2.68×10^{-1}
15	***rs*1425270**	60025002	2.25×10^{-8}	6.54×10^{-8}	1.25×10^{-2}	1.28×10^{-5}	2.68×10^{-1}
15	***rs*7166891**	60026596	1.98×10^{-8}	6.28×10^{-8}	5.77×10^{-3}	3.20×10^{-5}	2.68×10^{-1}
15	***rs*7172145**	60026989	1.98×10^{-8}	6.28×10^{-8}	5.77×10^{-3}	3.20×10^{-5}	2.68×10^{-1}
15	*rs*4587915	60029254	1.44×10^{-8}	2.82×10^{-8}	9.60×10^{-3}	1.12×10^{-5}	3.28×10^{-1}
15	***rs*8027751**	60035012	3.81×10^{-8}	8.07×10^{-8}	8.78×10^{-3}	2.73×10^{-5}	3.33×10^{-1}
15	***rs*3784634**	60046929	2.52×10^{-8}	6.81×10^{-8}	1.54×10^{-2}	9.17×10^{-6}	3.40×10^{-1}
15	*rs*8034335	60074748	1.19×10^{-8}	2.59×10^{-8}	1.43×10^{-2}	5.58×10^{-6}	3.35×10^{-1}
15	*rs*8034216	60074820	1.19×10^{-8}	2.59×10^{-8}	1.43×10^{-2}	5.58×10^{-6}	3.35×10^{-1}
15	*rs*17271305	60120272	6.87×10^{-9}	1.35×10^{-8}	6.29×10^{-3}	1.17×10^{-5}	2.86×10^{-1}
15	*rs*17271340	60135177	8.99×10^{-9}	1.55×10^{-8}	1.68×10^{-2}	3.71×10^{-6}	3.08×10^{-1}
15	*rs*8039105	60146377	8.71×10^{-9}	1.53×10^{-8}	1.65×10^{-2}	3.75×10^{-6}	3.05×10^{-1}
15	***rs*4502156**	60170447	3.38×10^{-8}	7.65×10^{-8}	1.31×10^{-4}	1.62×10^{-3}	0.20

. . . continued

			MANOVA	USAT	Univariate Analysis p		
chr	SNP	position	p	p	FG	2-hr GL	FI
15	***rs*7163757**	60178900	1.68×10^{-8}	5.98×10^{-8}	7.98×10^{-4}	3.52×10^{-4}	4.88×10^{-2}
15	***rs*7173964**	60184234	2.06×10^{-8}	6.36×10^{-8}	1.15×10^{-3}	3.04×10^{-4}	4.47×10^{-2}
15	*rs*8037894	60181556	8.20×10^{-9}	1.48×10^{-8}	4.09×10^{-4}	4.89×10^{-4}	3.12×10^{-2}
15	***rs*6494307**	60181982	1.68×10^{-8}	5.98×10^{-8}	7.98×10^{-4}	3.52×10^{-4}	4.88×10^{-2}
15	***rs*7167878**	60183481	1.68×10^{-8}	5.98×10^{-8}	7.98×10^{-4}	3.52×10^{-4}	4.88×10^{-2}
15	***rs*7172432**	60183681	1.68×10^{-8}	5.98×10^{-8}	7.98×10^{-4}	3.52×10^{-4}	4.88×10^{-2}

Appendix S9

Table S4: List of interesting SNPs that barely missed the genome-wide threshold (5×10^{-8}) for USAT. SNPs with m.a.f. $< 5\%$ were screened out. The MANOVA and the univariate analyses p-values are also provided. The SNPs listed here are the ones left after LD screening. In a group of highly correlated SNPs (i.e., SNPs with estimated absolute pairwise correlation coefficient > 0.8 with another SNP), one SNP was kept as a representative. The abbreviations used are FG (Fasting Glucose), 2-hr GL (2-hour glucose from an oral glucose tolerance test), FI (Fasting Insulin).

For convenience, the optimal ω has been reported. It represents the adaptive weight given to MANOVA statistic by the USAT approach. One must note that when SSU and MANOVA p-values are close, the optimal weight ω in USAT is not really identifiable. One can expect SSU and MANOVA to behave similarly at 'partial association' when number of traits is few and they are weakly correlated (refer Figure 4).

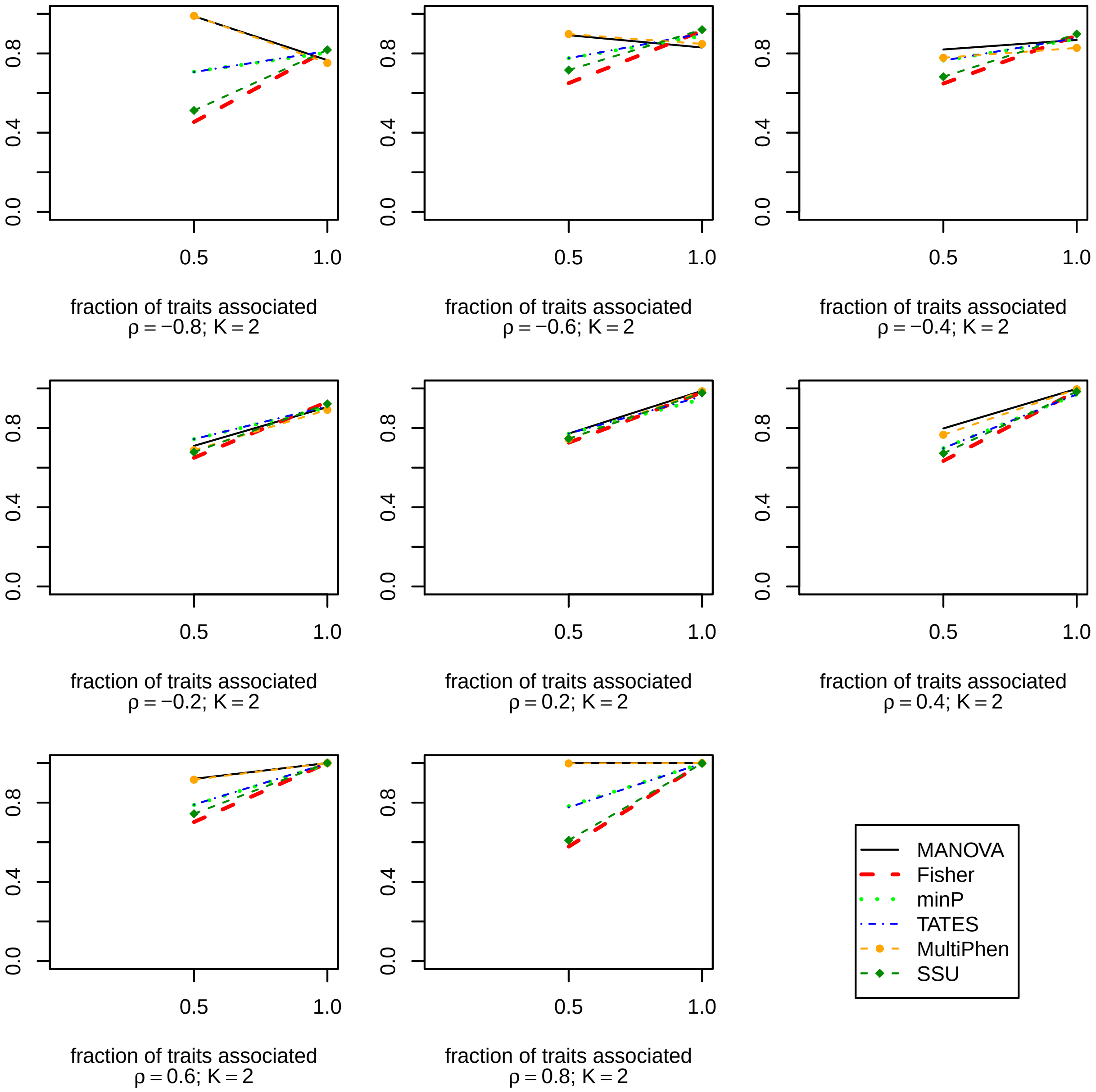
			MANOVA	USAT		Univariate Analysis p		
chr	SNP	m.a.f.	p	p	ω	FG	2-hr GL	FI
1	*rs*12095642	0.282	6.10×10^{-4}	4.71×10^{-5}	0.00	6.30×10^{-1}	8.08×10^{-2}	9.07×10^{-5}
1	*rs*10920639	0.166	1.36×10^{-3}	4.63×10^{-5}	0.00	1.15×10^{-1}	4.66×10^{-1}	9.77×10^{-5}
2	*rs*12622958	0.249	9.61×10^{-4}	9.61×10^{-5}	0.00	3.61×10^{-1}	4.37×10^{-2}	1.69×10^{-4}
3	*rs*9836499	0.367	1.26×10^{-3}	9.76×10^{-5}	0.00	4.26×10^{-1}	8.81×10^{-1}	3.51×10^{-4}
3	*rs*2336664	0.366	1.21×10^{-3}	9.50×10^{-5}	0.00	4.24×10^{-1}	8.94×10^{-1}	3.36×10^{-4}
3	*rs*6790846	0.095	5.68×10^{-3}	9.32×10^{-5}	0.00	9.06×10^{-1}	3.29×10^{-1}	6.69×10^{-4}
5	*rs*3798012	0.062	1.27×10^{-4}	7.84×10^{-6}	0.55	2.92×10^{-1}	7.81×10^{-1}	6.37×10^{-6}
5	*rs*7718567	0.132	4.49×10^{-4}	9.02×10^{-5}	0.70	5.85×10^{-1}	2.93×10^{-1}	1.23×10^{-4}
5	*rs*10213852	0.059	5.00×10^{-4}	9.05×10^{-5}	0.50	8.00×10^{-1}	5.32×10^{-1}	6.33×10^{-5}
5	*rs*10515261	0.097	3.00×10^{-3}	4.89×10^{-5}	0.00	2.31×10^{-2}	3.93×10^{-1}	6.29×10^{-4}
5	*rs*11135532	0.199	6.84×10^{-4}	4.90×10^{-5}	0.00	3.83×10^{-2}	1.30×10^{-1}	8.48×10^{-5}
5	*rs*1438733	0.255	1.72×10^{-3}	5.02×10^{-5}	0.00	8.32×10^{-1}	9.95×10^{-1}	1.47×10^{-4}
6	*rs*7753319	0.421	7.68×10^{-4}	3.13×10^{-6}	0.00	7.00×10^{-1}	2.20×10^{-1}	3.42×10^{-4}
6	*rs*6906163	0.139	2.21×10^{-3}	9.41×10^{-5}	0.00	3.66×10^{-1}	7.61×10^{-1}	2.24×10^{-4}
7	*rs*7793197	0.175	3.07×10^{-3}	4.56×10^{-6}	0.00	6.43×10^{-1}	1.06×10^{-1}	4.77×10^{-4}
10	*rs*2671692	0.367	6.96×10^{-4}	5.05×10^{-6}	0.00	5.83×10^{-1}	7.66×10^{-1}	6.15×10^{-5}
10	*rs*4376833	0.219	9.40×10^{-4}	8.93×10^{-5}	0.00	5.54×10^{-1}	4.47×10^{-1}	5.73×10^{-5}
12	*rs*7962136	0.186	1.38×10^{-3}	5.00×10^{-5}	0.00	4.53×10^{-1}	1.31×10^{-1}	1.36×10^{-4}
12	*rs*11829673	0.051	2.39×10^{-4}	8.91×10^{-5}	0.65	3.78×10^{-1}	6.08×10^{-1}	9.11×10^{-5}
13	*rs*7998882	0.122	1.60×10^{-4}	9.71×10^{-5}	0.85	8.86×10^{-1}	1.45×10^{-1}	9.18×10^{-5}
15	*rs*16957165	0.094	1.98×10^{-4}	9.43×10^{-5}	0.80	2.29×10^{-2}	5.81×10^{-3}	1.45×10^{-4}
15	*rs*931892	0.102	2.03×10^{-4}	8.59×10^{-5}	0.75	1.65×10^{-2}	2.21×10^{-2}	6.78×10^{-5}
16	*rs*11149640	0.374	1.50×10^{-3}	3.39×10^{-5}	0.00	5.93×10^{-1}	9.09×10^{-1}	1.21×10^{-4}

. . . continued

			MANOVA	USAT		Univariate Analysis p		
chr	SNP	m.a.f.	p	p	ω	FG	2-hr GL	FI
18	rs1443598	0.092	1.81×10^{-4}	2.80×10^{-6}	0.65	6.91×10^{-1}	9.61×10^{-1}	2.62×10^{-5}
18	rs11660607	0.287	3.16×10^{-3}	9.59×10^{-5}	0.00	4.26×10^{-1}	5.07×10^{-1}	2.03×10^{-4}
18	rs12604897	0.139	1.05×10^{-3}	4.61×10^{-5}	0.00	2.19×10^{-1}	3.27×10^{-1}	6.23×10^{-5}

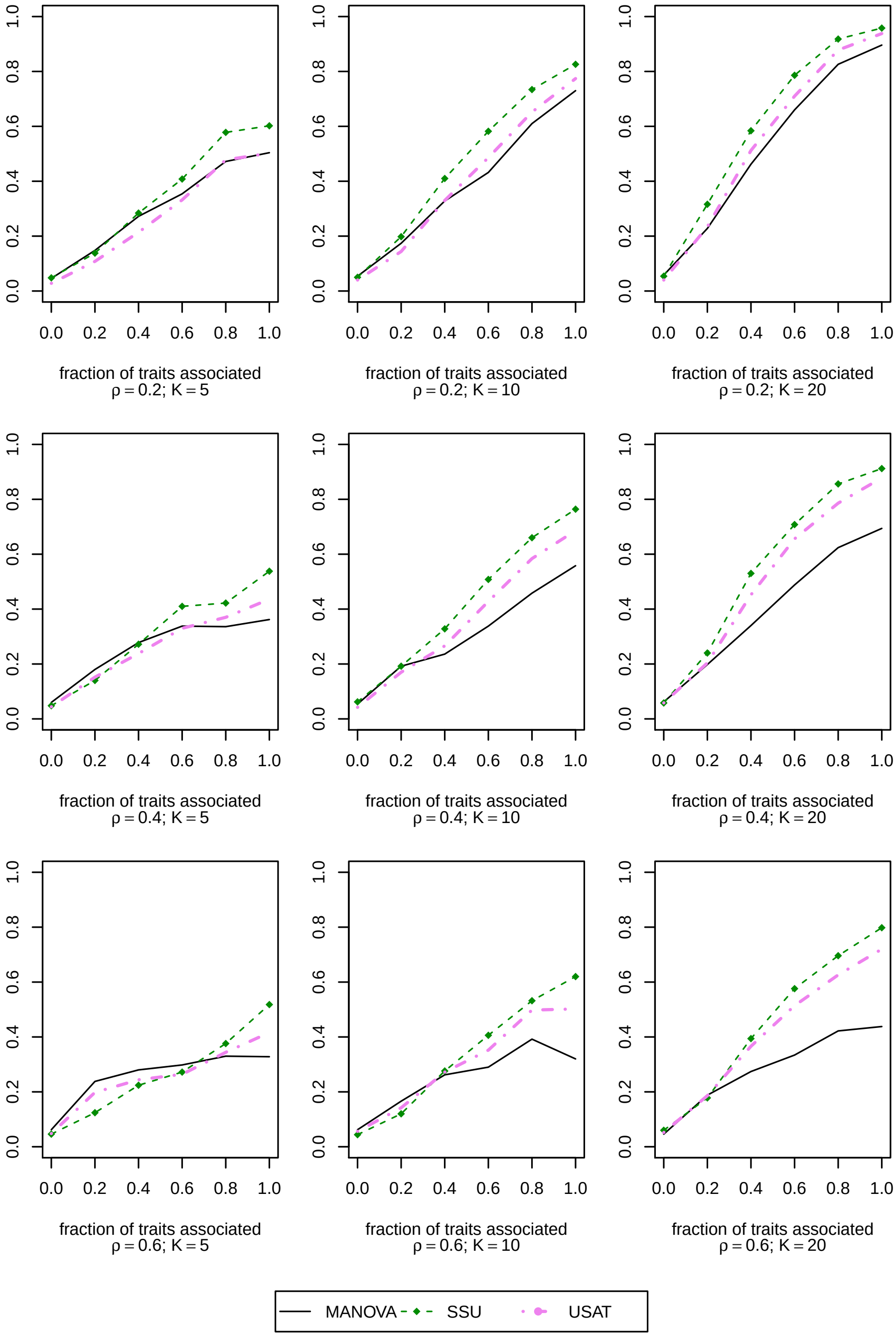
Appendix S10

Figure S8: Empirical power curves of the different existing association tests for $K = 2$ traits and different within trait correlation values $\rho = -0.8, -0.6, -0.4, -0.2, 0.2, ..., 0.8$ based on $N = 500$ datasets with $n = 4,000$ unrelated subjects. Opposite direction but same size genetic effect used when both traits are associated (i.e., datasets are generated from an alternative model $H_{a2,2} : \beta_1 = -\beta_2 > 0$). Effect size of 0.25 (proportion of variance explained is 0.2%) is used for the associated traits. The power is plotted along y-axis while the fraction of traits associated with the genetic variant is plotted along x-axis.



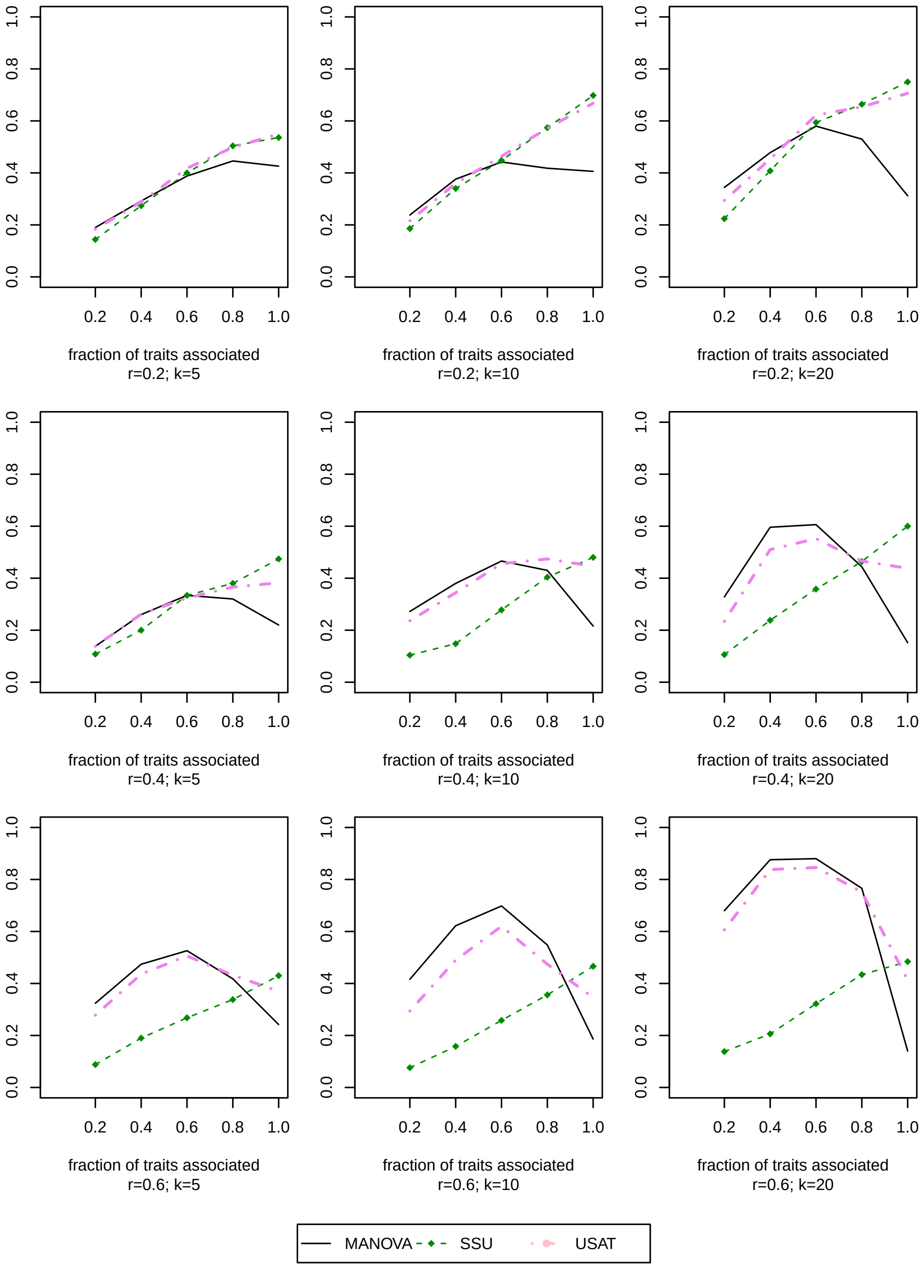
Appendix S11

Figure S9: Asymptotic power curves of the SSU and MANOVA tests along with our novel approach USAT for AR1(ρ) within-trait correlation structure. $K = 5, 10, 20$ traits have been simulated at different within trait correlation values $\rho = 0.2, 0.4, 0.6$. For each value of K and ρ, there were $N = 500$ datasets of $n = 400$ unrelated individuals. Same effect size of 0.395 (proportion of variance explained is 0.5%) was used for the traits that are associated. The power is plotted along y-axis while the fraction of traits associated with the genetic variant is plotted along x-axis.



Appendix S12

Figure S10: Empirical power curves of the SSU and MANOVA tests along with our novel approach USAT for CS(ρ) within-trait correlation structure. $K = 5, 10, 20$ traits have been simulated at different within trait correlation values $\rho = 0.2, 0.4, 0.6$. For each value of K and ρ, there were $N = 500$ datasets of $n = 400$ unrelated individuals. A single SNP with minor allele frequency (m.a.f.) 0.05 was simulated. Same effect size of 0.725 (proportion of variance explained is 0.5%) was used for the traits that are associated. The power is plotted along y-axis while the fraction of traits associated is plotted along x-axis. This figure shows that the relative behavior of MANOVA and the SSU test does not vary much with change in m.a.f. Since our proposed test USAT is derived from an optimal weighted combination of MANOVA and the SSU test, the performance of USAT compared to MANOVA or SSU also does not vary much with change in m.a.f.



References

H.W. Borchers. pracma: Practical numerical math functions. r package version 0.9.6. 2012. URL `http://CRAN.R-project.org/package=pracma`.

J. Dupuis, C. Langenberg, I. Prokopenko, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet*, 42 (2): 105–116, 2010.

C.G. Khatri. The necessary and sufficient conditions for dependent quadratic forms to be distributed as multivariate gamma. *J Multivar Anal*, 10:233–242, 1980.

C.G. Khatri, P.R. Krishnaiah, and P.K. Sen. A note on the joint distribution of correlated quadratic forms. *J Stat Plan Inference*, 1:299–307, 1977.

H. Liu, Y. Tang, and H.H. Zhang. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Comput Stat Data Anal*, 53:853–856, 2009.

A.K. Manning, M-F Hivert, R.A. Scott, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet*, 44 (6):659–669, 2012.

R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

R.A. Scott, V. Lagou, R.P. Welch, et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat Genet*, 44 (9):991–1005, 2012.

WHO. *Screening for Type 2 Diabetes: Report of a World Health Organization and International Diabetes Federation meeting.* Geneva, 2003.