Better together against genetic effect heterogeneity: a sex-combined interaction analysis of testosterone levels in the UK Biobank data

Boxi Lin,[1*] **Lei Sun**[1,2,*]

[1] *Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada*

[2] *Department of Statistical Science, University of Toronto, Toronto, ON, Canada*

Correspondence*:
Lei Sun
sun@utstat.toronto.edu

Boxi Lin

boxi.lin@mail.utoronto.ca

ABSTRACT

The effect of a genetic variant on a complex trait may differ between female and male, and in the presence of such genetic effect heterogeneity, sex-stratified analysis is often used. For example, genetic effects are sex-specific for testosterone levels, and sex-stratified analysis of testosterone in literature provided easy-to-interpret, sex-specific effect size estimates. However, from the perspective of association testing power, sex-stratified analysis may not be the best approach. As sex-specific genetic effect implies SNP$\times$Sex interaction effect, jointly testing SNP main and SNP$\times$Sex interaction effects may be more powerful than sex-stratified analysis or the standard main-effect testing approach. Moreover, since individual data may be unavailable, it is then of interest to study if the interaction analysis can be derived from sex-stratified summary statistics. We considered several different sex-combined methods and evaluated them through extensive simulation studies. We observed that a) the joint SNP main and SNP$\times$Sex interaction analysis is most robust to a wide range of genetic models, and b) this joint interaction testing result can be obtained by *quadratically* combining sex-stratified summary statistics (i.e. *squared* sum of the sex-stratified summary statistics). We then reanalysed the testosterone levels of the UK Biobank data using sex-combined interaction analysis, which identified 27 new loci that were missed by the sex-stratified approach and the standard sex-combined analysis. Finally, we provide supporting association evidence for nine new loci, uniquely identified by the sex-combined interaction analysis, from earlier association studies of either testosterone level or steroid biosynthesis pathway where testosterone is synthesized. We thus recommend sex-combined interaction analysis, particularly for traits with known sex differences, for most

powerful association testing, then followed by sex-stratified analysis for effect size estimation and interpretation.

Keywords: genome-wide association study, heterogeneity, sex-stratified analysis, sex-combined analysis, interaction, power, robustness

1 INTRODUCTION

The effect of a genetic variant on a complex trait may differ between female and male (Döring et al., 2008; Kong et al., 2008; Sinnott-Armstrong et al., 2021). For example, genetic effects are often sex-specific for hormone-related traits, such as testosterone expression (Sinnott-Armstrong et al., 2021), and severe disease of SARS-CoV-2 infection is known to display sexual dimorphism (Peckham et al., 2020).

A recent sex-stratified genome-wide association study (GWAS) of testosterone levels in the UK Biobank data (Bycroft et al., 2018) identified 79 and 127 independent genome-wide significant signals, respectively, in female and male (Sinnott-Armstrong et al., 2021). Not surprisingly, the top GWAS hits in male were largely non-overlapping with those from female, in terms of both genomic locations and the signaling pathways where the annotated genes are involved. Concerning the emerging GWAS of SARS-CoV-2 susceptibility and hospitalization, although most published work did not perform sex-stratified analysis (or consider SNP$\times$Sex interaction), Roberts et al. (2020) conducted a sex-stratified analysis (followed by a meta-analysis) and identified three genome-wide significant loci in at least one study.

Some earlier GWASs have also reported the existence of genetic effect heterogeneity between female and male. An example for genetic effect in the same direction but different magnitude is the serum uric acid concentrations (Döring et al., 2008): The association between rs7442295 in *SLC2A9* was stronger in female ($p = 2.6 \times 10^{-74}$, explaining 5.8% of the phenotypic variance) than male ($p = 7.0 \times 10^{-17}$, explaining 1.2% of the phenotypic variance). There are also examples of genetic effects in opposite directions. For example, the association between recombination rates and SNPs in *RNF212* were reported to be in opposite direction between female and male (Kong et al., 2008).

In the presence of genetic effect heterogeneity between female and male, a sex-stratified analysis approach is often preferred as in Sinnott-Armstrong et al. (2021), which provides easy-to-interpret, sex-specific *effect size estimates* (Li et al., 2010; Anttila et al., 2013; Traylor et al., 2015). However, from the viewpoint of maximizing *power of association testing*, sex-stratified analysis may not be the best analytical strategy for two reasons.

First, after a sex-stratified analysis, it is still tempting to consider sex-combined analysis by, for example, aggregating the association evidence from female and male using the traditional meta-analysis (Cochran,

1954; DerSimonian and Laird, 1986; Willer et al., 2010). However, the fixed-effect meta-approach, though could be used mechanically here, is conceptually at odds with the heterogeneity setting considered here (Greco et al., 2013). The application of the random-effect meta-approach is also challenging, as its implementation requires the between-study variance estimate. When the number of studies (or groups) is small, the estimate is not reliable, resulting in increased type I error rate (Borenstein et al., 2021; Friede et al., 2016).

Second, if the effect of a SNP indeed differs between female and male, it implies SNP×Sex interaction effect. This suggests that an interaction testing approach could be more powerful, but it requires investigation across different genetic models. Additionally, the traditional regression-based mega-approach requires the availability of individual data to include the SNP, Sex and SNP×Sex terms, along with other covariates in the regression model; the corresponding joint SNP main and SNP×Sex interaction 2 degrees of freedom (df) analysis will be termed as the interaction analysis, unless specified otherwise. Thus, it is another intriguing question if the 2 dfinteraction analysis can be derived from sex-stratified summary statistics.

To answer these questions and compare powers between sex-stratified and different sex-combined approaches, let Z_F and Z_M be the summary statistics from the sex-stratified analysis, and p_F and p_M the corresponding p-values, respectively for female and male. Now consider three different sex-combined approaches: (for completeness) the classical meta-analysis of using a weighted average of Z_F and Z_M (TSG_L), the minimal p-value of p_F and p_M (TSG_{min}), and an omnibus test of using a weighted sum of *squared* Z_F and Z_M (TSG_Q, 2 df) (Li and Lagakos, 2006; Derkach et al., 2014). For the mega-analysis of individual data, if available, let TI_{joint} be the 2 df test derived from jointly testing both the main and SNP×Sex interaction effects via regression. For completeness, let's also consider the most commonly used GWAS approach (Sullivan et al., 2013), testing the genetic main effect (T_{main}) in a regression model without the SNP×Sex interaction term, but with sex included as a covariate.

Without loss of generality, assume the genetic effect in female (b_F) is always greater or equal to that in male (b_M), where b_M can be zero or in a different direction from b_F. If $b_M = 0$, intuitively, the female-only TSG_F should be the most powerful test. Similarly, if $b_M = b_F$, then T_{main} should be the most powerful test. However, in both cases, it is unclear if the 2 df TI_{joint} test would be competitive, and if the 2 df TSG_Q behaves similarly. Such an understanding is important for recommending a test that performs reasonably well across different scenarios.

Focusing on gene×environmental interaction analysis of a binary trait, the important work of Kraft et al. (2007) concluded that "Although the joint test of genetic marginal effect and interaction is not the most powerful, it is nearly optimal across all penetrance models we considered." (Note that sex can be

statistically treated as an environmental covariate.) Additionally, in a different study setting where Chen et al. (2021) compared the 2 df genotypic test with the 1 df additive test, derived from a correctly specified 1 df additive model, the maximum power loss of the 2 df test was theoretically capped at 11.4% across all parameter values and sample sizes; this is because "the derivation is based on ncp [non-centrality parameter] and α alone." Thus, we can deduce that TI_{joint} is a more robust test than T_{main},

However, insights are lacking on the method comparison between TI_{joint} and the other tests, including TSG_F, TSG_{min}, TSG_L, and TSG_Q. The relative performance between (some of) these tests have been compared through multiple other empirical studies (Magi et al., 2010; Aschard et al., 2010; Clarke and Morris, 2010; Sung et al., 2016, 2014), and the perspective piece by Aschard (2016) offers great insights into the "contribution of interaction effects to the variance of quantitative outcomes". But these earlier studies focused on the testing of the gene×environment interaction term, rather than studying if the potential heterogeneous genetic effects can be *leveraged* through gene×environment (here SNP×Sex) interaction analysis, which is the main objective of our work here.

In the remainder of this report, first, we describe the six association testing methods considered and the simulation study design for method evaluation. Second, we present the simulation results, comparing the performance of the six tests across different scenarios, with varying degrees of genetic effect heterogeneity and male-to-female sample size ratio. We show that a) there is no uniformly most powerful test across all scenarios, b) TI_{joint} and TSG_Q are more robust than the others, and c) TI_{joint} and TSG_Q are practically the same for interaction analysis, complementing the known result that T_{main} and TSG_L are asymptotically equivalent for studying genetic main effect (Lin and Zeng, 2010). Third, we revisited the sex-stratified GWAS of testosterone levels in the UK Biobank data by Sinnott-Armstrong et al. (2021). After first replicating their sex-stratified findings, we then show that sex-combined *interaction* analysis, TI_{joint} and TSG_Q (if only summary statistics are available), are indeed more powerful than sex-stratified GWAS and the traditional sex-combined analysis (T_{main}, TSG_L or TSG_{min}), in terms of both the number and strength of the signals. Finally, we replicated some of the new findings using the GWAS catalogue resource (Buniello et al., 2019). We conclude the report with additional remarks and a discussion of future work.

2 MATERIALS AND METHOD

2.1 Notation and set-up

Let n_F and $n_M = k \cdot n_F$ be the sample sizes of, respectively, the female and male samples, where k is the male-to-female sample size ratio. Let $n = n_M + n_F$ be the total sample size of the combined sample. Let $\mathbf{Y} = (\mathbf{Y}_F^\top, \mathbf{Y}_M^\top)^\top$ be the phenotype vector of a trait of interest, $\mathbf{G} = (\mathbf{G}_F^\top, \mathbf{G}_M^\top)^\top$ be the genotypes

of a SNP under the study, and $\mathbf{S}$ be the $n \times 1$ sex-indicator vector. For each element in $\mathbf{S}$, without loss of generality, we define $S_i = 1$ for a female, $i = 1, ..., n_F$, and $S_j = 0$ for a male, $j = 1, ..., n_M$. We assume the phenotype data were generated from the following generalized linear model,

$$g(\mathbb{E}[\mathbf{Y}]) = a + b_g\mathbf{G} + b_s\mathbf{S} + b_{gs}\mathbf{G} \times \mathbf{S}.$$

In this work, we mainly focus on the quantitative trait with identity link, but the analysis is generalizable to other types of trait with appropriate link functions.

2.2 TSG_F and TSG_M, the female and male sex-stratified analysis

As a SNP's genetic effect could be sex-specific, we first consider the commonly used sex-stratified approach to testing association between $\mathbf{Y}$ and $\mathbf{G}$. In practice, important environmental covariates $\mathbf{Z}$'s are included in the analysis, but they are omitted here for notation simplicity but without loss of generality; results are characteristically the same unless (beyond gene×sex) gene×environmental interaction analysis is also of interest, which is beyond the scope of this work.

Let

$$TSG_F = Z_F$$

be the standard Wald test statistic, derived from the regression model, $\mathbf{Y}_F = a + b_{gF}\mathbf{G}_F + N(\mathbf{0}, \sigma^2\mathbf{I}_{n_F})$, by testing $H_0 : b_{gF} = 0$. (If the trait is binary, a logistic regression will be used.) Z_F follows $N(0, 1)$ asymptotically under H_0, and the asymptotic p-value of TSG_F is $p_F = 2 \times (1 - \Phi[|Z_F|])$, where Φ is the cumulative density function of the standard normal distribution. Similarly, we can obtain Z_M and p_M for the association analysis using the male sample.

2.3 Three sex-combined approaches based on sex-stratified summary statistics

2.3.1 TSG_L, the classical meta-analysis

As discussed earlier, the traditional fixed-effect meta-analysis is not recommended for the heterogeneity setting considered here, and the random-effect approach is not applicable to studies of two groups only (Borenstein et al., 2021). But for completeness of our evaluation, we include TSG_L, the classic inverse-variance weighted average of Z_F and Z_M,

$$TSG_L = \frac{1/\hat{\sigma}_F}{\sqrt{1/\hat{\sigma}_F^2 + 1/\hat{\sigma}_M^2}} Z_F + \frac{1/\hat{\sigma}_M}{\sqrt{1/\hat{\sigma}_F^2 + 1/\hat{\sigma}_M^2}} Z_M. \quad (1)$$

Under the null of no association, TSG_L is asymptotically $N(0, 1)$ distributed.

The subscript $_L$ reflects the fact that TSG_L combines the two Z statistics linearly. As a result, TSG_L could lose power when the genetic effects in female and male have opposite directions. This is akin to the burden type of tests for joint analyzing multiple rare variants (Derkach et al., 2014).

2.3.2 $\boldsymbol{TSG_Q}$ (2 df), an omnibus test

Instead of forming a directional test by linearly combining Z_F and Z_M, we could construct an omnibus test by calculating a quadratic sum as

$$TSG_Q = Z_F^2 + Z_M^2. \tag{2}$$

Under the null of no association, TSG_Q asymptotically follows the centralized χ_2^2 distribution. A sample-size-based weighted sum of Z_F^2 and Z_M^2 can be also considered, but does not alter the results qualitatively (results not shown). Additionally, TSG_Q appears to be numerically connected with the interaction analysis, which we discuss in Section 2.4.1.

2.3.3 $\boldsymbol{TSG_{min}}$, the minimal p-value approach

We could also consider the minimal p-value approach,

$$TSG_{min} = min(p_F, p_M).$$

Note that p_M and p_F are independent of each other, and each is Unif(0, 1) distributed under the null of no association. Thus, the cumulative distribution function of TSG_{min} is $F(x) = 1 - (1 - x)^2$, which is used to obtain p-value for TSG_{min}-based association analysis.

2.4 Two sex-combined mega-analysis approaches

When individual data are available, we could consider the following two mega-analysis approaches through regression, with or without the SNP×Sex interaction term; Sex is always included as a covariate by convention.

2.4.1 $\boldsymbol{TI_{joint}}$, the 2 df main and interaction testing approach.

We first consider the following regression model,

$$\mathbf{Y} = a + b_g\mathbf{G} + b_s\mathbf{S} + b_{gs}\mathbf{G} \times \mathbf{S} + N(\mathbf{0}, \sigma_{\boldsymbol{S}}^2\mathbf{I}_n),$$

where the interaction term is included to capture the potential genetic heterogeneity between female and male. Additionally and importantly, the error model is sex-specific, $\sigma_{\boldsymbol{S}}^2$. That is, sex-specific variances,

σ_F^2 and σ_M^2, are assigned to the female and male groups, respectively; for example, human height differs between female and male in both its mean and variances (Deng et al., 2019).

We use

$$TI_{joint} = \chi^2_{b_g, b_{gs}}$$

to denote the 2 df Wald test statistic derived from jointly testing $H_0 : b_g = b_{gs} = 0$. This test has one extra df than the traditional main-effect-only 1 df test. However this does not directly translate to loss of power, as the corresponding non-centrality parameters may differ (Chen et al., 2021). We provide numerical evidence in Sections 3 and 4.

2.4.2 T_{main}, the genetic main effect only testing approach (and adjusting for sex main effect).

For completeness, we also consider the standard approach of testing for genetic main effect only. Although such a main-effect only test can be derived from the interaction regression model above, the resulting test is difficult to interpret in the presence of an interaction effect; similarly it is difficult to interpret the interaction effect in the presence of the main effect (Brambor et al., 2006; Aschard, 2016).

Thus, we consider the following most commonly used regression model,

$$\mathbf{Y} = a + b'_g \mathbf{G} + b'_s \mathbf{S} + N(\mathbf{0}, \sigma'^2_{\boldsymbol{S}} \mathbf{I}_n),$$

with the modification on the error term to allow for sex-specific variance $\sigma'^2_{\boldsymbol{S}}$, as noted earlier. Additionally, important environmental factors should be included in the regression but omitted here to ease notation. We use

$$T_{main} = Z_{b'_g}$$

to denote the Wald test statistic derived from testing $H_0 : b'_g = 0$.

2.5 Simulation study design

We conducted extensive simulations to evaluate the performance of the six association tests, namely TSG_F, TSG_{min}, TSG_L, TSG_Q, T_{main}, and TI_{joint}, where TSG_{min}, TSG_L and TSG_Q are summary-statistics-based, and TSG_Q and TI_{joint} are 2 df tests. As genetic effect in male was assumed to be no larger than that in female, TSG_M was not explicitly examined but used to construct the summary statistics-based sex-combined analyses.

The alternative scenarios considered generally follow those in Magi et al. (2010), where genetic effects could be the same between female and male or sex-specific. Specifically, we considered the bi-allelic SNP of interest G is coded additively as in convention, and $\mathbb{P}(G = 2) = maf^2$, $\mathbb{P}(G = 1) = 2 \times maf \times$

$(1 - maf)$, and $\mathbb{P}(G = 0) = (1 - maf)^2$, where maf is the population minor allele frequency and takes values of $maf = 0.05,\ 0.1$ or 0.25. We also considered varying male-to-female sample size ratio, with $k = n_M/n_F$ taking values from $\{0.5, 1, 1.5, 2\}$, and $n_F = 5,000$. As power depends on both sample size and effect size, we did not empirically consider different n_F, as it does not affect the *relative* performance in power between the different methods. The sex-combined phenotype generating regression model can be reformulated, equivalently, as two separate simulation models for male and female, respectively, as

$$\text{male: } \mathbf{Y} = b_g\mathbf{G} + N(0,1), \text{ and}$$

$$\text{female: } \mathbf{Y} = b_s + (b_g + b_{gs})\mathbf{G} + N(0,1).$$

We fixed $b_s = 1$ so that sex explained about 20% of the total phenotypic variation. By varying b_g and b_{gs}, we were able to simulate phenotype data under different association scenarios including;

- The null scenario: No genetic effects in both female and male.

  To evaluate the empirical type I error rates, we fixed $b_g = b_{gs} = 0$ to generate data under the null hypothesis of no genetic effects of G on Y, in both female and male.

- Alternative scenario A1: Homogeneous genetic effect between female and male.

  In scenario A1, we fixed $b_{gs} = 0$ so that the true genetic model has no interaction effect, which is statistically equivalent with homogeneous genetic effect between female and male. This is the best-case scenario for T_{main} and the worst for TI_{joint}, but we will show that power of TI_{joint} is competitive under A1, and TSG_Q has the same performance as TI_{joint}.

  The range of b_g was $[0.06, 0.3]$, corresponding to the phenotypic variation explained by the genetic variable $(\beta_g^2 Var(G)/Var(y))$ in the range of $[0.14\%, 3.24\%]$. For completeness, we also let $b_g = 0$ so the power study also includes the null scenario, where we expect the empirical power to be at the nominal type I error level. Similar inclusion of the null was made when studying alternative scenarios A2 and A3 below.

- Alternative scenario A2: Female-only genetic effect.

  In scenario A2, we fixed $b_g = 0$ (no effect in male), and b_{gs} in the range of $[0.06, 0.3]$ (the effect in female). This setting is the best-case scenario for TSG_F. But again, we will show that TI_{joint} and TSG_Q are competitive, as compared with TSG_F, and under A2 TI_{joint} and TSG_Q also have the same empirical performance.

- Alternative scenario A3: Heterogeneous genetic effect between female and male.

Here we fixed genetic effect in female at $b_g + b_{gs} = 0.15$, and b_g in the range of $[-0.6, 0.3]$ for the genetic effect in male. When $b_g < 0$, the effect directions differ between female and male. When $b_g > 0$, although the effect directions are the same, the effect magnitudes differ, unless $b_g = 0.15$.

Additionally, we conducted a sensitivity analysis under the following two settings;

- Sensitive Study 1: Non-normal residual distributions.
  Instead of the standard normal residual, we generated ϵ from student-t_4 and χ^2_4 to compare the performance of the testing methods when there are, respectively, excess kurtosis and skewness in the residual distributions.
- Sensitive Study 2: Dominant genetic model.
  Instead of the standard additive genetic model, we generated phenotypes from dominant models, which statistically also covers recessive models by switching the choice of the baseline allele. The working models remained additive, following the common GWAS practice.

Finally, although we do not expect the relative method performance to change when analyzing a binary trait, for completeness, we also generated binary phenotype data from $Bernoulli(\boldsymbol{p})$, where $\boldsymbol{p} = g^{-1}(b_g\mathbf{G} + b_s\mathbf{S} + b_{gs}\mathbf{G} \times \mathbf{S})$ and g is the logit link function.

3 SIMULATION RESULTS

Results here focus on power across different alternative scenarios, as all six tests are standard statistical tests derived directly or indirectly from regression and expected to be accurate, particularly when the trait is normally distributed. But for completeness, we report the empirical type I error rates under the null of no association and including the sensitivity analyses.

3.1 Empirical type I error rates

We used $R = 5 \times 10^6$ replications to evaluate the empirical type I error rates at the nominal level of $\alpha = 10^{-5}$. Results in Figure S1 show that, as expected, all six tests are accurate, and have the nominal level $\alpha = 10^{-5}$ covered by 95% binomial proportion confidence interval $(\hat{\alpha} \pm z\sqrt{\hat{\alpha}(1-\hat{\alpha})/R})$, where $\hat{\alpha}$ is empirical type I error. We note that $\alpha = 5 \times 10^{-8}$ (Dudbridge and Gusnanto, 2008) is ideal but requires at least $R = 10^{10}$ replicates for accurate type I error evaluation, for each sample size considered, which is computational expensive, but $\alpha = 5 \times 10^{-8}$ was used for power study.

3.2 Power comparison

Figure 1 shows the empirical powers of the six testing methods, respectively, under the three alternative scenarios, A1, A2 and A3, where the empirical powers, at the genome-wide significance level of 5×10^{-8},

were obtained from 10^5 replicates. Result presentation here focuses on sample sizes of $n_F = n_M = 5,000$. Results for unbalanced sample size ratio with $k \in \{0.5, 1.5, 2\}$ are characteristically similar in that the relative method performance stays the same (Figures S2-S4).

In Scenario A1 where genetic effects are the same between female and male, it is easy to see that this is the best case-scenario for T_{main}, which jointly analyzes all samples available through a (correctly-specified) main-effect-only regression model. Similarly, TSG_F should have the lowest power, as the female-only analysis ignores the $n_M = k \cdot n_F$ of the total available n samples. Additionally, when there is no effect heterogeneity, meta-analysis (linearly combining Z_F and Z_M) is as efficient as mega-analysis (Lin and Zeng, 2010). Thus, TSG_L has the same power as T_{main} as expected; the two power curves overlap in Figure 1 Scenario A1.

As there is no interaction effect in Scenario A1, the 2 df TI_{joint} test, jointly testing both the main and interaction effects, is expected to be less powerful than T_{main}. However, results in Figure 1 Scenario A1 show that the loss of power is marginal, suggesting the robustness of TI_{joint} even under the scenario of no interaction effect. Additionally, TI_{joint} is noticeably more powerful than the minimum-p value approach of TSG_{min}. Finally, the power curve of TI_{joint}, interestingly, overlaps with that of TSG_Q (quadratically combining Z_F^2 and Z_M^2.)

In Scenario A2, where the genetic effect exists only in female, the female-only sex-stratified test TSG_F is most powerful as expected, and T_{main} is the least powerful with significant loss of power; power of TSG_L is practically identical to that of T_{main} as expected. Compared with Scenario A1, the performances of TSG_F and T_{main} (and TSG_L) here in Scenario A2 are reversed, suggesting that neither method is robust against different alternatives, which are unknown in practice. Also as expected is TSG_{min} being almost as powerful as TSG_F, as TSG_{min} captures the association signal through the female-only sex-stratified analysis; the slight loss of efficiency of TSG_{min} is a result of searching for the 'best' result between the two groups, akin to p-hacking penalization.

Importantly, TI_{joint} is also competitive in this case, with power almost identical with that of TSG_{min}, and only slightly smaller than that of the most powerful method of TSG_F. Interestingly, same as in Scenario A1, the power curve of TI_{joint} overlaps with that of TSG_Q in Scenario A2 as well.

In Scenario A3 where genetic effects exist in both female and male but differ in magnitude and/or direction, it reasonable to predict that no method dominates the others as confirmed by results in Figure 1. Specifically, (I) When the genetic effect in male is close to zero, the relative performance of the different methods is, as expected, similar to that observed scenario A2: The female-only sex-stratified test TSG_F has the highest power, but TSG_{min}, TSG_Q and TI_{joint} are competitive. (II) When the genetic effect in male is close to 0.15 (i.e. without effect heterogeneity between male and female), the relative performance

is similar to that in scenario A1: T_{main} and TSG_L have the highest power, but TSG_Q and TI_{joint} are competitive. (III) When the genetic effect in male is close to -0.15 (i.e. with severe effect heterogeneity to the extend of opposite effect directions), TSG_Q and TI_{joint} are clearly more powerful than the other methods.

Finally, across the range of parameter values under Scenario A3, the empirical powers of TI_{joint} and TSG_Q are the same as under A1 and A2. Thus, although theoretical justification is needed, our simulation results suggest that the 2 df interaction mega-analysis can be obtained from sex-stratified summary statistics through TSG_Q (quadratically combining Z_F^2 and Z_M^2). We will explore this as future work from the theoretical lens, following the earlier work of Aschard et al. (2010) and Aschard (2016) that focused on Score tests and regression models with residual variances independent of E (Sex in our case).

3.3 Sensitivity studies with model mis-specification and binary outcome

When the error distribution is t_4 or χ_4^2, given our simulation sample size ($n_F = 5,000$ with male-to-female sample size ratio of 0.5, 1, 1.5, or 2), the empirical type I error rates can be slightly inflated when MAF is low at 0.05, for all six tests examined (Figure S5). This is, however, not surprising as the convergence of the least square estimator's sampling distribution is slower when the residual distribution is non-normal, particularly asymmetrical (e.g. χ_4^2) (Lehmann, 1999). Subsequently, we did not consider power evaluation with non-normal residuals.

When the phenotype values are simulated from a dominant genetic model while the working model is additive, or the trait is binary, the type I error control results are similar to above (Figures S6 and S7).

Finally, across the three alternatives considered, the relative performances of the different tests for dominant generating models (Figure S8, the right panel) and for binary traits (Figure S9, left panel for additive model and right panel for dominant model) remain consistent with those observed from normally distributed traits and additive genetic models, as shown Figure 1 and Figures S2-S4.

4 APPLICATION TO GWAS OF THE TESTOSTERONE LEVELS IN THE UK BIOBANK DATA

Utilizing the UK Biobank data (Application number: 64875), we conducted GWAS of testosterone levels to compare the performance of sex-stratified and different sex-combined (main only vs. main and interaction joint) testing approaches.

First, we replicated the summary statistics, Z_F (named TSG_F for test of sub-group) Z_M (TSG_M), of the sex-stratified GWAS of Sinnott-Armstrong et al. (2021) in accordance with their sample quality

control and association testing procedures. With $\sim 320,000$ unrelated white British individuals as the study sample, we conducted GWAS using $\sim 800,000$ genotyped SNPs with $MAF > 0.01$; the sample size here is about 32 times larger than that of the simulation study, which allowed us to lower the MAF threshold from 5% to 1%. Our results are consistent with summary statistics provided in Sinnott-Armstrong et al. (2021) (https://figshare.com/collections/Supplementary_Data_For_Sinnott-Armstrong_and_Naqvi_eLife_2020/5304500/1). Second, we calculated summary statistics-based sex-combined test statistics, TSG_L (weighted sum of Z_F and Z_M) and TSG_Q (sum of Z_F^2 and Z_M^2), and their corresponding p-values, as described in Section 2; we omitted TSG_{min} from the application study, as our simulation studies has shown that it is sub-optimal as compared with TSG_Q. Lastly, we conducted the regression analyses using the individual data as described in Section 2.3.3, to obtain T_{main} and TI_{joint}.

Similar to the simulation results, here we observed that TSG_L and T_{main} are empirically equivalent (Figure S10, the right plot), which is expected given the theoretical results in Lin and Zeng (2010). Additionally, TSG_Q and TI_{joint} (Figure S10, the left plot) are also empirically equivalent, which is also not surprising given the work of Aschard et al. (2010) and Aschard (2016). However, additional theoretical work is needed as the earlier work focused on Score tests and covariate-independent residual variances.

The Miami plots in Figure 2 clearly demonstrate that TSG_Q was able to identify the top hits detected by either TSG_F (Figure 2a) or TSG_M (Figure 2b), as well as identify additional new loci. In contrast, as expected, TSG_L performs less well than TSG_Q in the presence of sex-effect heterogeneity. To provide better visualization for all significant loci, Figures S11 and S12 provide the Miami plots with $-log_{10}(p)$ truncated at 50 and 10, respectively.

To quantitatively evaluate and compare the performance between sex-stratified and sex-combined methods, the Venn diagram in Figure 3 summarizes the counts of independent loci. The independent loci were defined as the LD blocks constructed by Haploview using solid spline method (Barrett et al., 2004), with LD threshold r^2 = 0.1 and window size of 100kb. Figure 3 clearly shows that, unlike TSG_L, TSG_Q was able to retain most of the signals identified from either TSG_F or TSG_M, as well as discover new loci.

Additionally, the PP plots in Figure 4 show that for the genome-wide significant SNPs identified by TSG_Q or another method (sex-stratified or TSG_L), visually they all appear in the bottom triangle of each pair-wise comparison. This means that TSG_Q is more robust and can be much more powerful than the other testing methods, as for the associated variants, the $-\log_{10}(p)$ of TSG_Q are always almost as large as or substantially larger than the $-\log_{10}(p)$ of the other tests.

Indeed, although TSG_Q missed 12 and 35 independent loci identified, respectively, from the subgroup analyses TSG_M and TSG_F (Figure 3a), the PP plots in Figures 5c and 5d show that for these SNPs the maximum $-\log_{10}(p)$ of TSG_F and TSG_M was capped at 8.17, while the minimum $-\log_{10}(p)$ of TSG_Q was bounded at 7.30. That is, even TSG_Q may miss some signals strictly by the genome-wide significant threshold, the testing results of TSG_Q are comparable with that TSG_F and TSG_M, in terms of ranking of the associated SNPs. In contrast, the PP plots in Figures 5a and 5b show that among the novel signals identified by TSG_Q, the $-\log_{10}(p)$ of TSG_Q could be as high as 12.52 while the $-\log_{10}(p)$ of sex-stratified tests could be as low as 1.36, suggesting that TSG_Q is much more powerful than sex-stratified tests.

Finally, we evaluated whether the 27 novel genome-wide significant loci, uniquely identified by TSG_Q but missed by TSG_F, TSG_M and TSG_L in the UK Biobank data, are biologically relevant to the testosterone level phenotype. According to the GWAS catalog (Buniello et al., 2018), nine out of the 27 loci were previously reported to be associated with either testosterone level or steroid biosynthesis pathway where testosterone is synthesized (Haas et al. (2022); Ruth et al. (2020);Richardson et al. (2022);Hoffmann et al. (2018)). Furthermore, we used Ensembl Variant Effect Predictor (McLaren et al., 2016) to extract a list of genes affected by the top SNPs in these loci, of which multiple genes were previously reported to be associated with testosterone or the cholesterol level (Table 1).

5 DISCUSSION

In this report, we reviewed three classes of association testing methods that account for genetic effect heterogeneity, namely the sex-stratified approach, sex-combined approaches based on summary statistics, and sex-combined approaches based on individual data. We conducted extensive simulation studies and an application study, leveraging the sex-stratified GWAS summary statistics of Sinnott-Armstrong et al. (2021) which studied the testosterone levels in the UK Biobank data, and we gained several insights.

First, our simulation results confirmed that there is no uniformly most powerful test in this study setting. Instead, the relative power between these tests will depend on the underlying genetic models, and may substantially change between different scenarios. For example, T_{main} from the standard genetic main effect model is optimal when the true genetic effects are homogeneous between female and male. But, when the effects are in opposite directions with similar magnitudes, T_{main} has no power to detect association. Conversely, the female-only sex-stratified test TSG_F is optimal when the effect is only present in female, but it loses substantial power when the effect is also present in male.

Second, our simulation results revealed that, TI_{joint}, the 2 df joint test of both genetic main effect and SNP$\times$Sex interaction effect is most robust across all the scenarios considered. It remains powerful

even if there was no SNP×Sex interaction or the effect was only present in one group. For each scenario considered, TI_{joint} is either the most powerful test or only has a small amount of power loss as compared with the oracle method for that particular scenario, e.g. T_{main} under the homogeneous genetic effect scenario A1, and TSG_F under the female-only genetic effect scenario A2. The robustness of the joint test was previously noted, for example by Aschard et al. (2010): "In situations where the marginal test had better power than the joint test, the difference was small (less than 10% absolute power). Additionally, a recent work in a different study setting (Chen et al., 2021), comparing the standard 1 df additive test with the 2 df genotypic test, showed that "under additivity, the maximum power loss of a 2 df genotypic test is [theoretically] capped at 11.4% across all parameter values and sample sizes." Indeed, in our study setting when TI_{joint} was not the most powerful test, its maximum empirical power loss was 11.51% across all scenarios considered. In contrast, the maximum loss of power were 71.19% for TSG_F and 96.96% for T_{main}. Additionally, the power of TI_{joint} dominates the minimum p-value approach TSG_{min} in almost all cases considered, except when genetic effect is only present in female where TSG_{min} is only slightly more powerful (at most 4.23% more) than TI_{joint}.

Third, the power curves of TI_{joint} (jointly testing SNP main and SNP×Sex interaction) and TSG_Q ('meta' combining Z_F^2 and Z_M^2) overlapped. This is not surprising given the earlier work of Aschard et al. (2010), which examined Score tests derived from the standard regression model where residual variance is not sex-specific. The empirical evidence we observed here deserves further investigation, and a theoretical result for the equivalence between TI_{joint} and TSG_Q would complement to the known result between T_{main} (the traditional GWAS with no SNP×Sex interaction but including sex as a covariate) and TSG_L (the traditional meta-analysis combining Z_F and Z_M) (Lin and Zeng, 2010), as well as the earlier work . The work of Aschard et al. (2010)

We emphasize that TI_{joint} and TSG_Q (if only sex-stratified summary statistics are available) are recommended for powerful association testing, but not for effect size estimation. Thus, TI_{joint} and TSG_Q could be applied as an screening approach for GWAS to identify significant variants. Subsequent refined analysis on these variants, such as the sex-stratified analysis, could be conducted to provide a complete picture of genetic effect heterogeneity.

As sex can be modelled statistically as an environmental factor, our observation here complements the work in Kraft et al. (2007), which focused on TI_{joint} for gene×environmental interaction analysis and concluded that "the joint test [is] an attractive tool for large-scale association scans where the true gene×environment interaction model is unknown." Indeed, even if the competing tests include the different data-integration methods such as the meta-analysis (TSG_L), the minimal p-value approach (TSG_{min}), and the omnibus test (TSG_Q), TI_{joint} remains competitive. This insight also challenges the

current practice of analyzing samples from different populations separately. Instead, jointly analyzing all samples available and testing genetic main and gene-population effects simultaneously can be an useful alternative worth further investigation.

Finally, although we considered three different and complementing scenarios (A1, A2 and A3) for the true genetic model, simulation studies are never exhaustive, covering all possible alternatives. Indeed, when the residual distributions of phenotypic simulation model is non-normal and combined with the relatively small sample sizes, we observed inflated type I error at the extreme tail. Another future direction is to derive the non-centrality parameter of each test as a function of the minor allele frequency of a SNP, sex-specific effects of the SNP, sample size, male-to-female sample size ratio, as well as the random error. The analytical results will provide a complete picture for the power similarities and differences between the different tests.

DECLARATION OF INTEREST

The authors declare no competing interests.

AUTHOR CONTRIBUTIONS

BL developed the method, performed the analyses, summarized the results, and drafted the manuscript. LS conceptualized and supervised the study. Both authors read and approved the final manuscript.

FUNDING

This research was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC, RGPIN-04934) and the University of Toronto Data Sciences Institute (DSI) Catalyst Grant.

ACKNOWLEDGMENTS

We thank Dr. Andrew Paterson, Dr. Shelley Bull and Dr. Lisa Strug for their helpful comments and discussion. BL is a trainee of the CANSSI-ONTARIO STAGE (Strategic Training for Advanced Genetic Epidemiology) training program at the University of Toronto.

REFERENCES

Anttila, V., Winsvold, B. S., Gormley, P., Kurth, T., Bettella, F., McMahon, G., et al. (2013). Genome-wide meta-analysis identifies new susceptibility loci for migraine. *Nature Genetics* 45, 912–917. doi: 10.1038/ng.2676

Aschard, H. (2016). A perspective on interaction effects in genetic association studies. *Genetic epidemiology* 40, 678–688

Aschard, H., Hancock, D. B., London, S. J., and Kraft, P. (2010). Genome-wide meta-analysis of joint tests for genetic and gene-environment interaction effects. *Human Heredity* 70, 292–300. doi:10.1159/000323318

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2004). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263–265. doi:10.1093/bioinformatics/bth457

Borenstein, M., Hedges, L. V., Higgins, J., and Rothstein, H. R. (2021). *Introduction to meta-analysis* (Hoboken: Wiley), second edition edn.

Brambor, T., Clark, W. R., and Golder, M. (2006). Understanding interaction models: Improving empirical analyses. *Political Analysis* 14, 63–82. doi:10.1093/pan/mpi014

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2018). The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* 47, D1005–D1012. doi:10.1093/nar/gky1120

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2019). The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research* 47, D1005–D1012

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., et al. (2018). The UK biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209

Chen, B., Craiu, R. V., Strug, L. J., and Sun, L. (2021). The x factor: A robust and powerful approach to x-chromosome-inclusive whole-genome association studies. *Genetic epidemiology* 45, 694–709

Clarke, G. M. and Morris, A. P. (2010). A comparison of sample size and power in case-only association studies of gene-environment interaction. *Am. J. Epidemiol.* 171, 498–505

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics* 10, 101–129

Deng, W. Q., Mao, S., Kalnapenkis, A., Esko, T., Mägi, R., Paré, G., et al. (2019). Analytical strategies to include the x-chromosome in variance heterogeneity analyses: Evidence for trait-specific polygenic variance structure. *Genetic epidemiology* 43, 815–830

Derkach, A., Lawless, J. F., and Sun, L. (2014). Pooled association tests for rare genetic variants: a review and some new results. *Statistical Science* 29, 302–321

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Control. Clin. Trials* 7, 177–188

Dudbridge, F. and Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 32, 227–234

Döring, A., Gieger, C., Mehta, D., Gohlke, H., Prokisch, H., Coassin, S., et al. (2008). Slc2a9 influences uric acid concentrations with pronounced sex-specific effects. *Nature Genetics* 40, 430–436. doi:10.1038/ng.107

Friede, T., Röver, C., Wandel, S., and Neuenschwander, B. (2016). Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases. *Biometrical Journal* 59, 658–671. doi:10.1002/bimj.201500236

Greco, T., Zangrillo, A., Biondi-Zoccai, G., and Landoni, G. (2013). Meta-analysis: pitfalls and hints. *Heart Lung Vessel.* 5, 219–225

Haas, C. B., Hsu, L., Lampe, J. W., Wernli, K. J., and Lindström, S. (2022). Cross-ancestry genome-wide association studies of sex hormone concentrations in pre- and postmenopausal women. *Endocrinology* 163. doi:10.1210/endocr/bqac020

Hoffmann, T. J., Theusch, E., Haldar, T., Ranatunga, D. K., Jorgenson, E., Medina, M. W., et al. (2018). A large electronic-health-record-based genome-wide study of serum lipids. *Nature Genetics* 50, 401–413. doi:10.1038/s41588-018-0064-5

Klarin, D., , Damrauer, S. M., Cho, K., Sun, Y. V., Teslovich, T. M., et al. (2018). Genetics of blood lipids among ˜300, 000 multi-ethnic participants of the million veteran program. *Nature Genetics* 50, 1514–1523. doi:10.1038/s41588-018-0222-9

Kong, A., Thorleifsson, G., Stefansson, H., Masson, G., Helgason, A., Gudbjartsson, D. F., et al. (2008). Sequence variants in the rnf212 gene associate with genome-wide recombination rate. *New Series* 319, 1398–1401. doi:10.1126/science.ll51851

Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J., and Gauderman, W. J. (2007). Exploiting gene-environment interaction to detect genetic associations. *Human Heredity* 63, 111–119. doi:10.1159/000099183

Lehmann, E. L. (1999). *Elements of large-sample theory*. Springer Texts in Statistics (New York, NY: Springer)

Li, Q. H. and Lagakos, S. W. (2006). On the relationship between directional and omnibus statistical tests. *Scandinavian Journal of Statistics* 33, 239–246. doi:10.1111/j.1467-9469.2005.00489.x

Li, Y., Sheu, C. C., Ye, Y., de Andrade, M., Wang, L., Chang, S. C., et al. (2010). Genetic variants and risk of lung cancer in never smokers: A genome-wide association study. *The Lancet Oncology* 11, 321–330. doi:10.1016/S1470-2045(10)70042-5

Lin, D. Y. and Zeng, D. (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika* 97, 321–332. doi:10.1093/biomet/asq006

Magi, R., Lindgren, C. M., and Morris, A. P. (2010). Meta-analysis of sex-specific genome-wide association studies. *Genetic Epidemiology* 34, 846–853. doi:10.1002/gepi.20540

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., et al. (2016). The ensembl variant effect predictor. *Genome Biology* 17. doi:10.1186/s13059-016-0974-4

Peckham, H., de Gruijter, N. M., Raine, C., Radziszewska, A., Ciurtin, C., Wedderburn, L. R., et al. (2020). Male sex identified by global covid-19 meta-analysis as a risk factor for death and itu admission. *Nature communications* 11, 1–10

Richardson, T. G., Leyden, G. M., Wang, Q., Bell, J. A., Elsworth, B., Smith, G. D., et al. (2022). Characterising metabolomic signatures of lipid-modifying therapies through drug target mendelian randomisation. *PLOS Biology* 20, e3001547. doi:10.1371/journal.pbio.3001547

Roberts, G. H., Park, D. S., Coignet, M. V., McCurdy, S. R., Knight, S. C., Partha, R., et al. (2020). Ancestrydna covid-19 host genetic study identifies three novel loci. *medRxiv*

Ruth, K. S., , Day, F. R., Tyrrell, J., Thompson, D. J., Wood, A. R., et al. (2020). Using human genetics to understand the disease impacts of testosterone in men and women. *Nature Medicine* 26, 252–258. doi:10.1038/s41591-020-0751-5

Sinnott-Armstrong, N., Naqvi, S., Rivas, M., and Pritchard, J. K. (2021). Gwas of three molecular traits highlights core genes and pathways alongside a highly polygenic background. *eLife* 10, 1–35. doi:10.7554/eLife.58615

Sullivan, P. F., Daly, M. J., Ripke, S., Lewis, C. M., Lin, D. Y., Wray, N. R., et al. (2013). A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular Psychiatry* 18, 497–511. doi:10.1038/mp.2012.21

Sung, Y. J., Schwander, K., Arnett, D. K., Kardia, S. L. R., Rankinen, T., Bouchard, C., et al. (2014). An empirical comparison of meta-analysis and mega-analysis of individual participant data for identifying gene-environment interactions. *Genetic Epidemiology* 38, 369–378. doi:10.1002/gepi.21800

Sung, Y. J., Winkler, T. W., Manning, A. K., Aschard, H., Gudnason, V., Harris, T. B., et al. (2016). An Empirical Comparison of Joint and Stratified Frameworks for Studying G × E Interactions: Systolic Blood Pressure and Smoking in the CHARGE Gene-Lifestyle Interactions Working Group. *Genetic Epidemiology* 40, 404–415. doi:10.1002/gepi.21978

Traylor, M., Markus, H., and Lewis, C. M. (2015). Homogeneous case subgroups increase power in genetic association studies. *European Journal of Human Genetics* 23, 863–869. doi:10.1038/ejhg.2014.194

Willer, C. J., Li, Y., and Abecasis, G. R. (2010). Metal: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26, 2190–2191. doi:10.1093/bioinformatics/btq340

FIGURE CAPTIONS

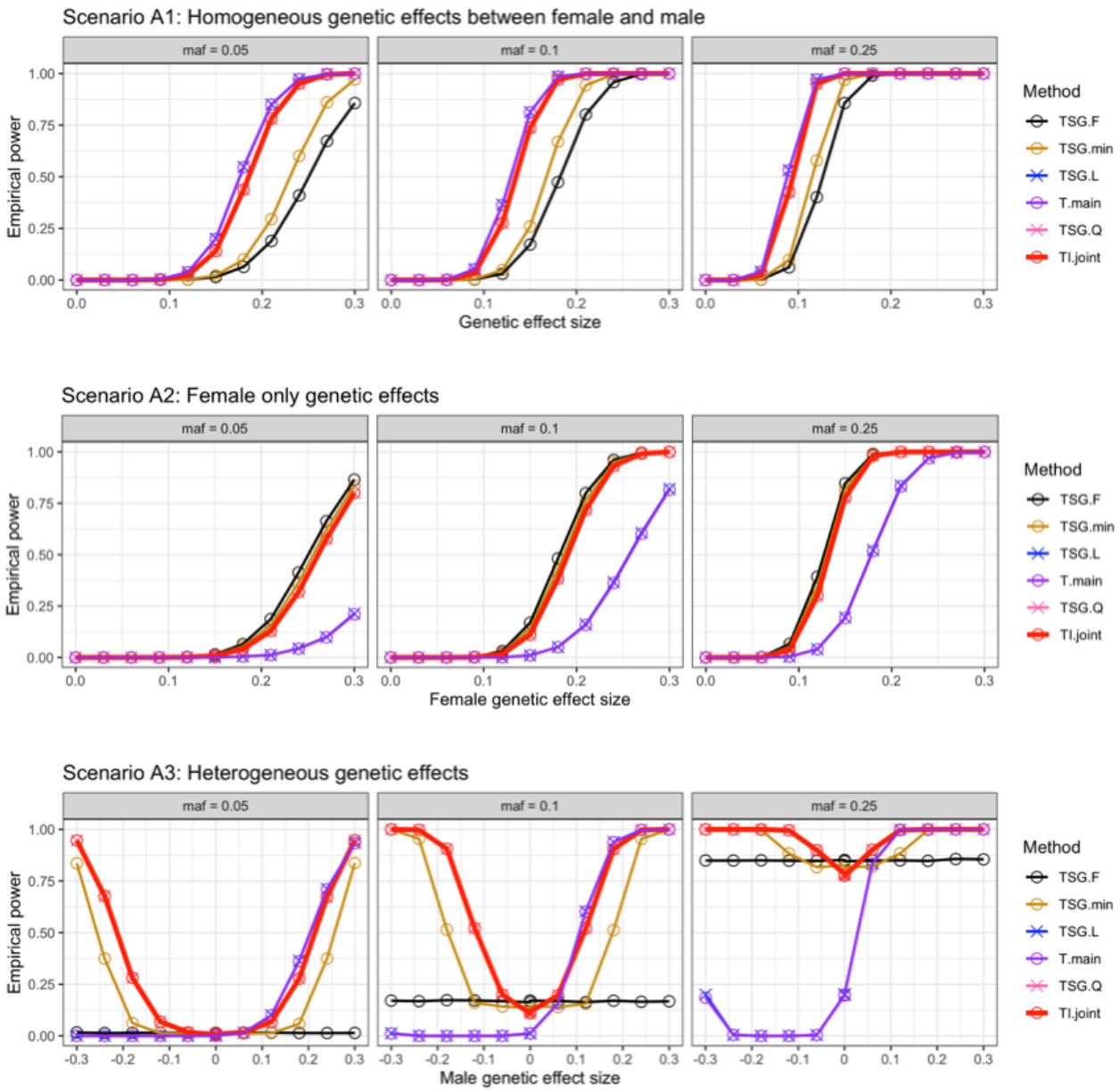

Figure 1. Power comparison at $\alpha = 5 \times 10^{-8}$ with female sample size $n_F = 5,000$ and $k = 1$ ($n_M = 5,000$) under the three alternative scenarios, A1, A2 and A3, stratified by the MAF. A1: Homogeneous genetic effect between female and male, and the genetic effect sizes ranged from 0 to 0.3; A2: Female-only genetic effect, and the genetic effect sizes ranged from 0 to 0.3; A3: Heterogeneous genetic effect between female and male, the genetic effect in female was kept at 0.15, while the effect in male ranged from -0.3 to 0.3. Six association testing methods were evaluated: (1) The female-only sex-stratified analysis (TSG_F); (2) The minimum p-value of sex-stratified analysis (TSG_{min}); (3) A meta-analysis combining Z-scores (TSG_L); (4) The main-effect-only model (T_{main}); (5) An omnibus test combining squared Z-scores (TSG_Q), the recommended test when only sex-stratified summary statistics are available; (6) The recommended joint test of both the main and interaction effects (TI_{joint}). Results for other $k = n_M/n_F$ male-to-female ratios are shown in Figures S2-S4.

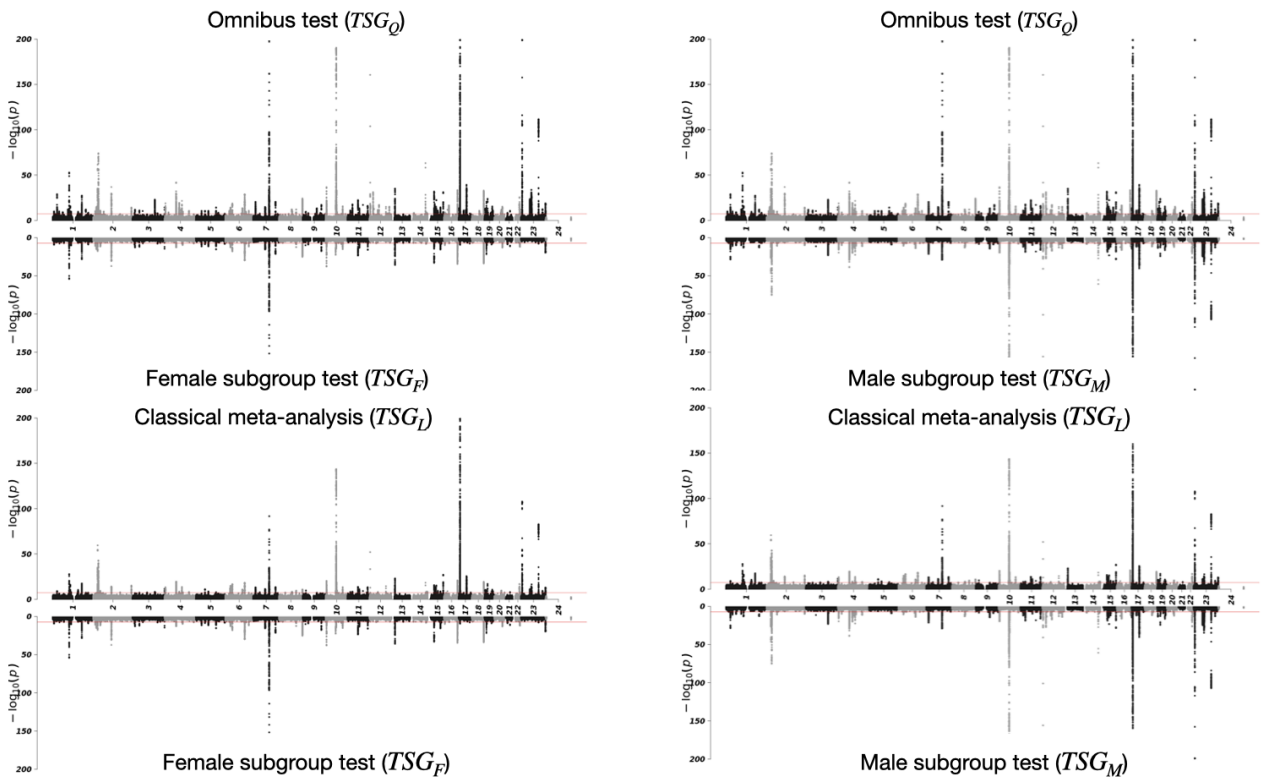

Figure 2. Miami plots comparing the $-\log_{10}(p)$ between tests based on (a) TSG_Q and TSG_F, (b) TSG_Q and TSG_M, (c) TSG_L and TSG_F, and (d) TSG_L and TSG_M, leveraging the sex-stratified GWAS summary statistics of testosterone level from Sinnott-Armstrong et al. (2021). Here TSG_F and TSG_M are sex-stratified GWAS results of Sinnott-Armstrong et al. (2021). The omnibus TSG_Q results were obtained using equation 2, and the traditional meta-analysis TSG_L results were obtained using equation 1. The red horizontal lines indicate the genome-wide significant threshold of 5×10^{-8} on the $-log_{10}$ scale. The maximum $-\log_{10} p$ was truncated at 200 for better visualization. Miami plots with $-log_{10}(p)$ truncated at 50 and 10 are shown in Figures S11 and S12, respectively.

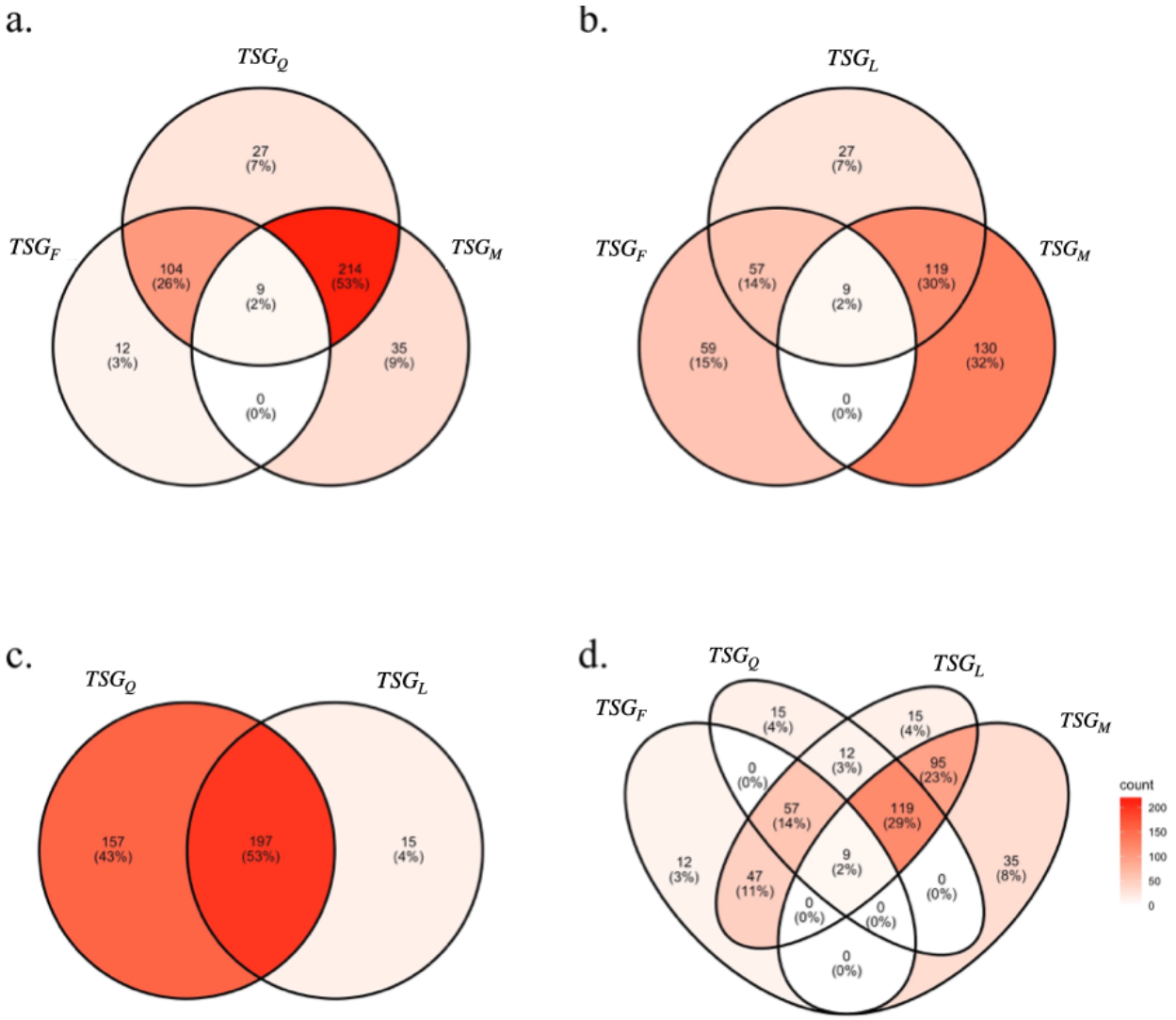

Figure 3. Venn diagrams showing the relationship between the independent genome-wide significant loci identified from (a) TSG_Q and the sex-stratified tests, TSG_F and TSG_M, (b) TSG_L and the sex-stratified tests, (c) TSG_Q and TSG_L, and (d) All of the four tests. Here TSG_F and TSG_M are sex-stratified GWAS results of Sinnott-Armstrong et al. (2021). The omnibus TSG_Q results were obtained using equation 2, and the traditional meta-analysis TSG_L results were obtained using equation 1. The independent loci were defined as the LD blocks constructed by Haploview using solid spline method (Barrett et al., 2004), with LD threshold $r^2 = 0.1$ and window size of 100kb.

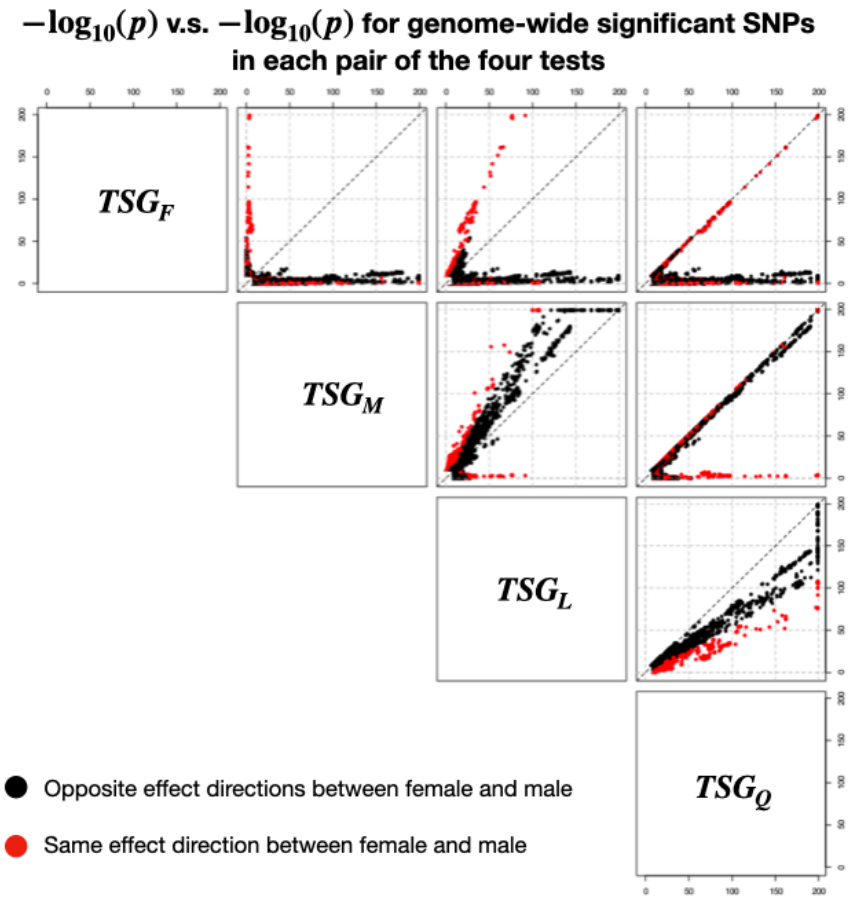

Figure 4. Pairwise PP plots comparing the $-\log_{10} p$-values between the female-only sex-stratified analysis (TSG_F), male-only sex-stratified analysis (TSG_M), the classical meta-analysis (TSG_L), and the omnibus test combining squared Z-scores (TSG_Q). Here TSG_F and TSG_M are sex-stratified GWAS results of Sinnott-Armstrong et al. (2021). The omnibus TSG_Q results were obtained using equation 2, and the traditional meta-analysis TSG_L results were obtained using equation 1. In each subplot, each point represents a SNP reaching genome-wide significant level of 5×10^{-8} *in any of the two corresponding testing methods*. The maximum $-\log_{10} p$ was truncated at 200 for better visualization. PP plots with $-log_{10}(p)$ truncated at 50 and 10 are shown in Figures S13 and S14, respectively.

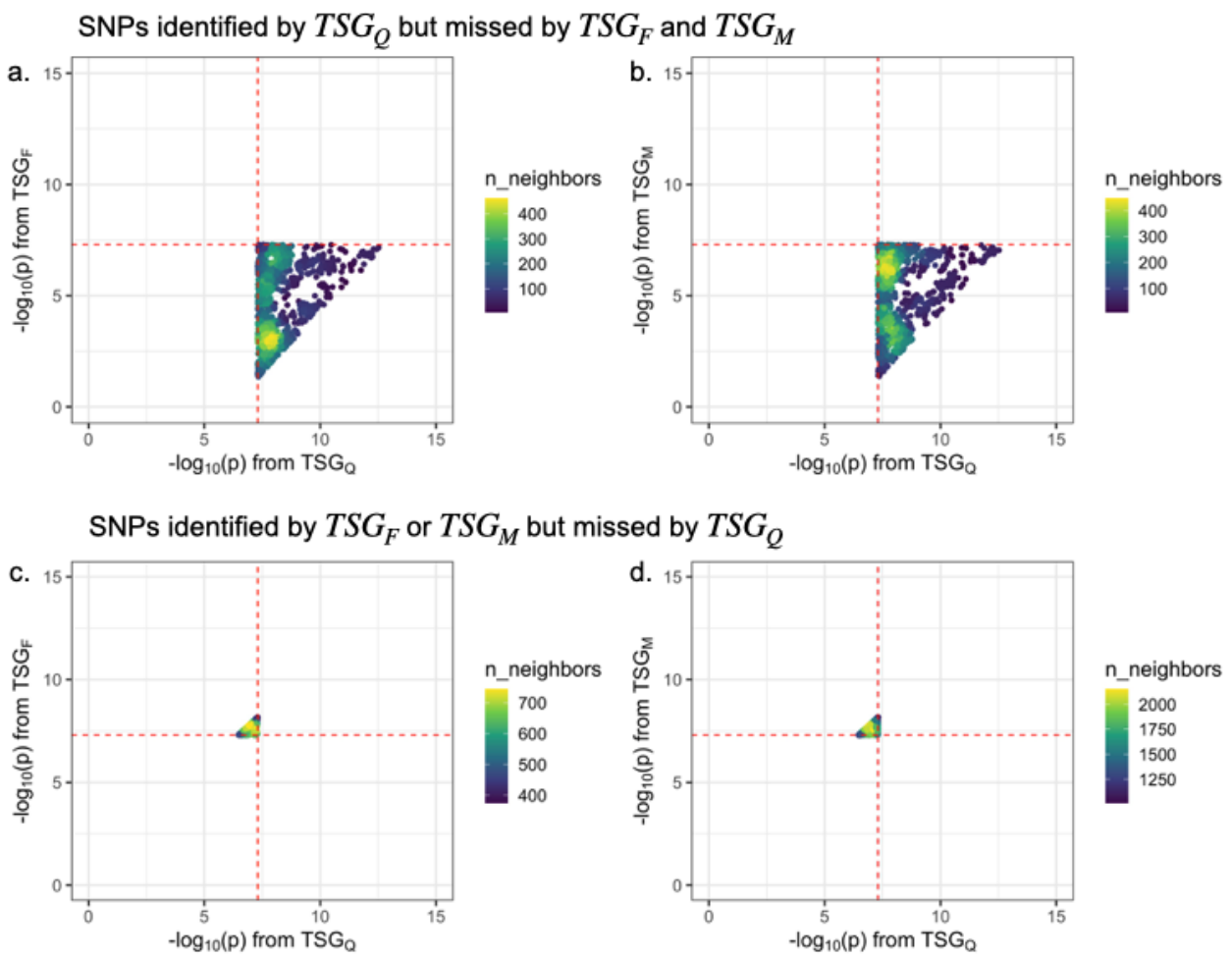

Figure 5. Restricted pairwise PP plots, considering (a) and (b): SNPs identified by TSG_Q but missed by TSG_F and TSG_M, (c) and (d): SNPs identified by TSG_F or TSG_M but missed by TSG_Q. Here TSG_F and TSG_M are sex-stratified GWAS results of Sinnott-Armstrong et al. (2021). The omnibus TSG_Q results were obtained using equation 2, and the traditional meta-analysis TSG_L results were obtained using equation 1. The red lines indicate the genome-wide significant threshold of 5×10^{-8} on the $-log_{10}$ scale. The color of points indicates $n_neighbors$, i.e. the number of its neighbouring points of the point. The p-values were *not* truncated for the set of SNPs considered.

rs ID	CHR:BP Location (hg19)	TSG_F	TSG_M	TSG_L	TSG_Q	Affected gene	Relevant SNP and gene association
rs4712976	6:25842203	8.46e-08	3.04e-06	6.57e-01	1.09e-11	*SLC17A3*	Bioavailable testosterone levels (Ruth et al., 2020); testosterone measurement (Haas et al., 2022)
rs7521292	1:155941685	8.24e-08	6.17e-05	3.63e-11	1.87e-10	*ARHGEF2*	
rs61741874	2:27462651	3.42e-04	1.08e-07	3.04e-10	1.22e-09	*CAD*	Total cholesterol levels (Hoffmann et al., 2018)
rs79267532	19:49530737	1.22e-07	4.97e-04	5.81e-10	1.94e-09		
rs28495024	7:99016683	2.7e-07	2.22e-04	3.23e-01	1.97e-09	*PTCD1*	
rs17580	14:94847262	5.64e-04	3.73e-07	2.33e-01	6.44e-09	*SERPINA1*	Sex hormone-binding globulin measurement (Haas et al., 2022); cholesterol levels (Klarin et al., 2018)
rs3130455	6:31125978	2.25e-04	1.59e-06	1.88e-09	1.1e-08	*TCF19*	
rs34090663	10:1047872	6.93e-08	6.54e-03	6.35e-02	1.2e-08	*GTPBP4*	
rs7816651	8:81822709	1.76e-03	2.67e-07	4.66e-09	1.34e-08		
rs73365510	14:106875035	4.91e-04	9.48e-07	2.9e-09	1.39e-08	*IGHV4-39, IGHVIII-38-1*	Testosterone measurement (Haas et al., 2022)
rs697885	12:116304043	2.95e-03	1.88e-07	1.06e-01	1.53e-08		
rs4397736	10:4971457	3.43e-05	1.77e-05	2.45e-09	1.87e-08	*AKR1C1*	
rs6920820	6:30969938	1.28e-07	6.07e-03	7.85e-02	2.03e-08	*MUC21,MUC22, HCG22*	total cholesterol measurement (Richardson et al., 2022)
rs10279715	7:40870935	9.66e-03	8.49e-08	1.81e-08	2.06e-08	*SUGCT*	Testosterone measurement (Ruth et al., 2020)
rs9304665	19:47602577	1.79e-03	6.35e-07	9.61e-09	3.14e-08	*ZC3H4*	
rs10779612	1:214318000	3.03e-02	5.26e-08	6.88e-08	3.54e-08		
rs534645300	16:30813455	1.35e-07	1.08e-02	5.86e-02	3.57e-08	*ZNF629*	Testosterone levels in postmenopausal women (Haas et al., 2022)
rs10750766	11:65473798	4.94e-03	2.87e-07	9.38e-02	3.67e-08	*KAT5, RN7SL309P*	Sex hormone-binding globulin measurement (Haas et al., 2022) ; Cholesterol (Richardson et al., 2022)
rs1030431	8:59311697	3.56e-04	3.71e-06	4.34e-01	3.82e-08	*UBXN2B*	Total cholesterol levels (Hoffmann et al., 2018)
rs1042114	1:29138975	1.49e-03	9.84e-07	2.11e-01	4.03e-08	*OPRD1*	
rs199669245	10:5034015	4.02e-06	3.7e-04	7.77e-09	4.27e-08	*AKR1C2*	
rs1910174	12:21197705	9.74e-03	1.91e-07	5.89e-02	4.55e-08	*LST3, SLCO1B7, SLCO1B7*	
rs111903103	11:118308499	1.09e-06	1.61e-03	1.43e-08	4.79e-08	*KMT2A*	

Table 1. Evidence for biological relevance of the SNPs and genes in the loci uniquely identified by TSG_Q in the UK Biobank data. Here TSG_F and TSG_M are sex-stratified association results with testosterone from the GWAS of Sinnott-Armstrong et al. (2021), which were also replicated by this study. The omnibus TSG_Q results were obtained using equation 2, and the traditional meta-analysis TSG_L results were obtained using equation 1. The p-values were calculated based on the summary statistics, TSG_F and TSG_M. Relevant SNP associations were based on previously reported association from GWAS catalog (Buniello et al., 2018); Affected genes and relevant gene association were extracted from Ensembl Variant Effect Predictor (McLaren et al., 2016). We ordered the rows based on TSG_Q column and excluded four SNPs that could be annotated to the GWAS Catalog.