An empirical study of sex-stratified vs. sex-combined analysis in the presence of genetic effect heterogeneity

Boxi Lin
Dalla Lana School of Public Health
University of Toronto
Toronto, ON, Canada
boxi.lin@mail.utoronto.ca

Lei Sun
Department of Statistical Science
University of Toronto
Toronto, ON, Canada
sun@utstat.toronto.edu

Abstract

The effect of a genetic variant on a complex trait may differ between male and female, e.g. genetic effects may be sex-specific for testosterone levels. In the presence of genetic effect heterogeneity between female and male, sex-stratified analysis is often used, which provides easy-to-interpret, sex-specific effect size estimates. However, from power of association testing perspective, sex-stratified analysis may not be the best approach. As sex-specific genetic effect implies SNP-sex interaction effect, jointly testing SNP main and SNP-sex interaction effects may be more powerful than sex-stratified analysis or the standard main-effect testing approach. When individual data are not available, 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-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* sex-stratified summary statistics). We then provide additional supporting evidence by utilizing the publicly available sex-stratified GWAS summary statistics of testosterone levels of the UK Biobank data.

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 male and female (Döring et al., 2008; Kong et al., 2008; Sinnott-Armstrong et al., 2021). For example, genetic effects may be 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-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 male and female. An example for genetic effect in the same direction but with 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 direction. For example, the association between recombination rates and SNPs in *RNF212* were in opposite direction between male and female (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* (Y. Li et al., 2010; Anttila et al., 2013; Traylor et al., 2015). However, from the perspective 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 aggregating the association evidence from female and male using the traditional meta-analysis (Cochran, 1954; DerSimonian & Laird, 1986; Willer et al., 2010). However, the fix-effect meta-approach, though could be used here technically, is conceptually at odds with the heterogeneity setting considered here (Greco et al., 2013). The application of the random-effects meta-approach is also challenging, as the calculation of the variance of the meta-effect relies on the estimate of between-study variance. When the number of studies is small, the estimate is not reliable, resulting in increased type I error (Borenstein et al., 2021; Friede et al., 2016).

Second, if the effect of a SNP differs between male and female, then it implies that there is a SNP-sex interaction effect. This suggests that an interaction testing approach could be more powerful, but it requires investigation across different genetic models. Additionally, this regression 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 (d.f.) analysis will be termed as the interaction analysis, unless specified otherwise. Thus, it is another intriguing question if the 2 d.f. interaction analysis can be derived from sex-stratified summary statistics.

To answer these questions and compare powers of the sex-stratified and 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. 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 average of *squared* Z_f and Z_m (TSG_Q, also 2 d.f.) (Q. H. Li & Lagakos, 2006; Derkach et al., 2014). For the mega-analysis of individual data, if available, let TI_{joint} be the 2 d.f. 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 d.f. TI_{joint} test would be competitive, and if the 2 d.f. 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 seminal 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 B. Chen et al. (2021) compared a 2 d.f. test with the 1 d.f. test, derived from a correctly specified 1 d.f. model, the maximum power loss of the 2 d.f. test is 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 & Morris, 2010; Sung et al., 2016; 2014). 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 (e.g. SNP-sex) interaction analysis, which is the main objective of our work here.

In the remainder of this report, first, we provide the details of the six association testing methods and our simulation study design. 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 ratios. We show that a) there is no uniformly most powerful test across all scenarios, b) TSG_Q and TI_{joint} are more robust than the others,

and c) TSG_Q and TI_{joint} are practically the same. Third, we utilize the sex-stratified GWAS summary statistics from Sinnott-Armstrong et al. (2021) to show that TSG_Q is more robust than sex-stratified analysis and TSG_L, the traditional meta-analysis. Finally, we conclude the report with additional remarks and a discussion of future work.

2 Method

2.1 Notations and settings

Let n_f and $n_m = k \cdot n_f$ be the sample sizes of, respectively, the female sample and the male sample, where k reflects the sample size ratio between male and female. 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 generated from the 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 paper we mainly focus on the quantitative trait with identity link. However, the analysis could be generalized to other types of trait with appropriate link functions.

2.2 TSG_f and TSG_m, the sex-stratified analysis

As a SNP's genetic effect could be sex-specific, we 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 gene-environmental (beyond gene-sex) interaction analysis is also of interest, which is beyond the scope of this brief report.

Let

$$TSG_f = Z_f$$

be the standard score 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 for TSG_f is derived by $p_f = 2 \times (1 - \Phi[|Z_f|])$, where Φ is the cumulative density function of the standard normal distribution. Similarly, we can obtain Z_m and p_m.

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 fix-effect meta-analysis is not recommended for the heterogeneity setting considered here, and the random-effects 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 classical sample-size weighted average of Z_f and Z_m,

$$TSG_L = \frac{\sqrt{n_f}}{\sqrt{n_f + n_m}} Z_f + \frac{\sqrt{n_m}}{\sqrt{n_f + n_m}} Z_m. \tag{1}$$

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

We use a subscript $_L$ to reflect the fact that we linearly combine the two Z statistics. 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 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.

2.3.3 TSG_Q (2 d.f.), 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, asymptotically TSG_Q 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).

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.

2.4.1 TI_{joint}, a 2 d.f. main and interaction testing approach.

We first consider the following regression model,

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

where the interaction term is included to capture the potential genetic heterogeneity between male and female.

We use

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

to denote the 2 d.f. score test statistic from jointly testing $H_0 : b_g = b_{gs} = 0$. This test has one extra d.f. than the traditional main-effect-only 1 d.f. test, but this does not directly translate to loss of power, as the corresponding non-centrality parameters may differ (B. Chen et al., 2021).

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

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

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

which is the most commonly used model in GWAS (Sullivan et al., 2013); as noted earlier, environmental factors would be included in the regression model in practice.

We use

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

to denote the score test statistic 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}. We focused on evaluating power across different alternative scenarios, as all the tests are standard statistical tests derived from regression with controlled type I error rates. But for completeness, we report the empirical type I error rates under the null of no association.

The alternative scenarios considered generally follow those in Magi et al. (2010), where genetic effects are the same or differ between female and male in the context of meta-analysis. Specifically, we generated phenotypes from $\mathbf{Y} = b_g\mathbf{G} + b_s\mathbf{S} + b_{gs}\mathbf{G} \times \mathbf{S} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, I_n)$. $\mathbf{G}$ represents additively coded bi-allelic genotypes generated under the Hardy–Weinberg proportions (HWP) where the genotype frequencies $\mathbb{P}(G = 2) = maf^2$ and

$\mathbb{P}(G = 1) = 2 \times maf \times (1 - maf)$. Here maf denotes the population minor allele frequency, and we examined $maf = 0.05,\ 0.1$ and 0.25. We fixed $b_S = 1$ such that sex explained about 20% of the total phenotypic variation. We also considered varying sample size ratio between male and female, with $k = n_m/n_f$ taking values from $\{0.5, 1, 1.5, 2\}$, and $n_f = 10^3$.

The joint phenotype generating regression model can be reformulated, equivalently, as two separate simulation models for female and male, respectively, as

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

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

By varying b_g and b_{gs}, we simulated phenotype data under the following scenarios for association testing:

- 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 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. 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 $[0.05, 0.5]$, corresponding to the phenotypic variation explained by the genetic variable ($\beta_g^2 Var(G)/Var(y)$) in the range of $[0.02\%, 4.47\%]$. For completeness, we also let $b_g = 0$ so the power study includes the null scenario, where we expect the empirical power to be at the nominal type I error level. Similar inclusions of the null were made when studying alternative scenarios A2 and A3 below.

- Alternative scenario A2: Female-only genetic effect.

  In scenario A2, we fixed $b_g = 0$ and chose b_{gs} in the range of $[0.05, 0.5]$. In this setting, G has genetic effect size of b_{gs} in female but no effect in male, the best-case scenario for TSG_f. Again, we will show that power of TI_{joint} and TSG_Q are competitive under A2, and they have the same empirical performance.

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

  Here we fixed genetic effect in female at $b_g + b_{gs} = 0.2$, and chose b_g in the range of $[-0.2, 0.2]$ 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 $b_g = 0.2$.

To study the robustness of these tests to other model misspecifications, we repeated the above analyses with phenotypes simulated under the following three settings:

- Supplementary Sensitive Study 1: Non-normal residual distributions

In addition to the standard normal residual, we generated ϵ from student-t_4 and χ^2_4, which allows us to compare the performance of the testing methods when there is excess kurtosis or skewness, respectively, in the residual distributions.

- Supplementary Sensitive Study 2: Dominant genetic model

  In addition to the 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 model remains additive following the standard practice.

- Supplementary Sensitive Study 3: Binary response

  We further generated binary phenotypes 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

3.1 Empirical type I error rates

We used $R = 10^7$ replications to evaluate the empirical type I error rates of the six association tests, at the nominal level of $\alpha = 5 \times 10^{-6}$. Results in Supplementary Figure S1 show that, as expected, when the genetic variants are common ($MAF = 0.1$ or 0.25), all six tests are accurate, and have the nominal level $\alpha = 5 \times 10^{-6}$ 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. For rare variants ($MAF = 0.05$; rare relative to the sample size considered here) we observed a slightly inflated type I error for TSG_F due to its reduced sample size. Nevertheless, we kept TSG_F in our subsequent power comparison for completeness of the presentation but not affecting the conclusion, as TSG_F is not a recommended approach and rare variants are not the focus of our 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. Result presentation here focuses on sample sizes of $n_f = n_m = 1,000$. Results for unbalanced sample size ratio with $k \in \{0.5, 1.5, 2\}$ are characteristically similar in that the performance ordering of the six tests stays the same (Supplementary 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 regression model. Additionally, when there is no effect heterogeneity, meta-analysis (linearly combining Z_f and Z_m) is as efficient as mega-analysis (Lin & Zeng, 2010). Thus, TSG_L has the same power as T_{main} as expected; the two power curves overlap in Figure 1 Scenario A1.

Under Scenario A1 when there is no heterogeneity, it is also easy to see that TSG_f should have the lowest power, as the female-only analysis ignores the $n_m = k \cdot n_f$ portion of the available n samples, which is supported by our empirical results.

As there is no interaction effect in Scenario A1, the 2 d.f. 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 effects only exists 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 is a result of searching for the best result between the two groups, akin to p-hacking.

Interestingly, 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. Same as in Scenario A1, the power curve of TI_{joint} overlaps with that of TSG_Q in Scenario A2 as well.

Figure 1, Scenario A3 shows the power results when the genetic effect exists in both female and male but differ in magnitude and/or direction. In this study setting, as we fixed the effect size for female, the power of the female-only sex-stratified test TSG_f is a constant across the different effect sizes for the male.

Results in Scenario A3 show that (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 0.2 (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 -0.2 (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 d.f. interaction mega-analysis can be obtained from sex-stratified summary statistics through TSG_Q (quadratically combining Z_f^2 and Z_m^2), which we will explore as future work from the theoretical lens.

3.3 Model mis-specification and binary outcome

When the error distribution is t_4 or χ^2_4, given our simulation sample size ($n_f = 1,000$ and $2,000$), the empirical type I error rates are inflated for some of the tests, at the extreme tail of $\alpha = 5 \times 10^{-6}$ (Supplementary Figure S5). This is, however, not surprising as the convergence of the ordinary least square estimator's sampling distribution is slower when the residual distribution is non-normal (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, all six tests could still control type I error rates well (Supplementary Figure S6, the right panel). Additionally, the relative performance of the different tests remains consistent with the original simulation studies, where both the simulating and working models are additive (Supplementary Figure S7, the right panel).

Finally, when the trait is binary, the relative method performance of the different tests did not change as compared with a continuous trait (Supplementary Figures S6 and S7). We note that this is expected, because under common regularity conditions, the score test statistics (not parameter estimation) derived from models from the exponential family have identical form (C.-F. Chen, 1983; Zhang & Sun, 2021).

4 Application to GWAS on testosterone levels in UK Biobank

We utilized the sex-stratified GWAS of testosterone levels (Sinnott-Armstrong et al., 2021) to compare the performance of sex-stratified and sex-combined testing approaches. We first obtained the sex-stratified GWAS summary statistics, TSG_{Female} and TSG_{Male}, provided in the paper (`https://doi.org/10.6084/m9.figshare.c.5304500.v1`). We then calculated sex-combined test statistics, TSG_L and TSG_Q and their corresponding p-values by combining TSG_{Female} and TSG_{Male} through formula 1 and 2.

Without the individual data, we could not conduct the regression analysis to obtain T_{main} and TI_{Joint} directly. However, it is known that TSG_L and T_{main} are equivalent (Lin & Zeng, 2010), and TSG_Q and TI_{joint} appear to be empirically equivalent based on our simulation studies. Thus, we expect the performances and conclusions for T_{main} and TI_{Joint} would be similar for TSG_L and TSG_Q, respectively, although a formal regression-based mega-analysis is required to confirm this. We did not consider TSG_{min} in the application study, as our simulation studies show that it is sub-optimal compared with TSG_Q.

The Miami plots in Figure 2a and 2b clearly demonstrate that TSG_Q was able to detect the top hits identified by either TSG_{Female} or TSG_{Male} and identify additional 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, Supplementary Figures S8 and S9 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, we used Venn diagram (Figure 3) to summarize 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 recover most of the signals identified from either TSG_{Female} or TSG_{Male}, as well as discover new loci.

Additionally, the P-P 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 lower half of the figure, with some SNPs slightly off the main diagonal. This means that 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, respectively, identified from TSG_{Male} and TSG_{Female} (Figure 3), the P-P plots in Figures 5(c) and 5(d) show that for these SNPs the maximum $-\log_{10}(p)$ of TSG_{Female} and TSG_{Male} 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_{Female} and TSG_{Male}. In contrast, the P-P plots in Figures 5(a) and 5(b) 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.

5 Discussion

In this brief 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 of 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 d.f. joint test of both genetic main effect and SNP-sex interaction effect is most robust across all 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, e.g. T_{main} under the homogeneous genetic effect scenario A1, and TSG_f under the female-only genetic effect scenario A2. A recent work in a different study setting (B. Chen et al., 2021), comparing the standard 1 d.f. additive test with the 2 d.f. 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 empirical evidence 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 & Zeng, 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.

Additionally, TI_{joint} could be applied in pharmacogenomics (PGx) studies. PGx investigates the impact of individual's genetic basis on the variation in their responses to drug (Stromeyer, 1959). Particularly in randomized clinical trial, although the gene-treatment interaction is often the primary interest and typically inferred by the case-only design or 1 d.f. regression analysis (Little et al., 2005, Ritchie, 2012), jointly testing genetic main effect and gene-treatment interaction may increase the power in screening genes.

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.

Acknowledgments

We thank Dr. Andrew Paterson and Dr. Shelley Bull 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., ... Palotie, A. (2013, 8). Genome-wide meta-analysis identifies new susceptibility loci for migraine. *Nature Genetics*, *45*, 912-917. doi: doi:10.1038/ng.2676

Aschard, H. (2016, December). A perspective on interaction effects in genetic association studies. *Genet. Epidemiol.*, *40*(8), 678–688.

Aschard, H., Hancock, D. B., London, S. J., & Kraft, P. (2010). Genome-wide meta-analysis of joint tests for genetic and gene-environment interaction effects. *Human Heredity*, *70*(4), 292–300. doi: doi:10.1159/000323318

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2004, 08). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, *21*(2), 263-265. Retrieved from https://doi.org/10.1093/bioinformatics/bth457 doi: doi:10.1093/bioinformatics/bth457

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

Brambor, T., Clark, W. R., & Golder, M. (2006). Understanding interaction models: Improving empirical analyses. *Political Analysis*, *14*(1), 63–82. doi: doi:10.1093/pan/mpi014

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., ... Marchini, J. (2018, October). The UK biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209.

Chen, B., Craiu, R. V., Strug, L. J., & Sun, L. (2021). The x factor: A robust and powerful approach to x-chromosome-inclusive whole-genome association studies. *Genetic epidemiology*, *45*(7), 694–709.

Chen, C.-F. (1983). Score tests for regression models. *Journal of the American Statistical Association*, *78*(381), 158-161. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/01621459.1983.10477945 doi: doi:10.1080/01621459.1983.10477945

Clarke, G. M., & Morris, A. P. (2010, February). A comparison of sample size and power in case-only association studies of gene-environment interaction. *Am. J. Epidemiol.*, *171*(4), 498–505.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*(1), 101–129.

Derkach, A., Lawless, J. F., & Sun, L. (2014). Pooled association tests for rare genetic variants: a review and some new results. *Statistical Science*, *29*(2), 302–321.

DerSimonian, R., & Laird, N. (1986, September). Meta-analysis in clinical trials. *Control. Clin. Trials*, *7*(3), 177–188.

Döring, A., Gieger, C., Mehta, D., Gohlke, H., Prokisch, H., Coassin, S., ... Meisinger, C. (2008, 4). Slc2a9 influences uric acid concentrations with pronounced sex-specific effects. *Nature Genetics*, *40*, 430-436. doi: doi:10.1038/ng.107

Friede, T., Röver, C., Wandel, S., & Neuenschwander, B. (2016, October). Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases. *Biometrical Journal*, *59*(4), 658–671. Retrieved from https://doi.org/10.1002/bimj.201500236 doi: doi:10.1002/bimj.201500236

Greco, T., Zangrillo, A., Biondi-Zoccai, G., & Landoni, G. (2013). Meta-analysis: pitfalls and hints. *Heart Lung Vessel.*, *5*(4), 219–225.

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

Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J., & Gauderman, W. J. (2007, 2). Exploiting gene-environment interaction to detect genetic associations. *Human Heredity*, *63*, 111-119. doi: doi:10.1159/000099183

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

Li, Q. H., & Lagakos, S. W. (2006, 6). On the relationship between directional and omnibus statistical tests. *Scandinavian Journal of Statistics*, *33*, 239-246. doi: doi:10.1111/j.1467-9469.2005.00489.x

Li, Y., Sheu, C. C., Ye, Y., de Andrade, M., Wang, L., Chang, S. C., ... Yang, P. (2010, 4). Genetic variants and risk of lung cancer in never smokers: A genome-wide association study. *The Lancet Oncology*, *11*, 321-330. doi: doi:10.1016/S1470-2045(10)70042-5

Lin, D. Y., & Zeng, D. (2010, 6). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*, *97*, 321-332. doi: doi:10.1093/biomet/asq006

Little, J., Sharp, L., Khoury, M. J., Bradley, L., & Gwinn, M. (2005). The epidemiologic approach to pharmacogenomics. *American Journal of PharmacoGenomics*, *5*(1), 1–20. Retrieved 2022-04-15, from http://link.springer.com/10.2165/00129785-200505010-00001 doi: doi:10.2165/00129785-200505010-00001

Magi, R., Lindgren, C. M., & Morris, A. P. (2010, 12). Meta-analysis of sex-specific genome-wide association studies. *Genetic Epidemiology*, *34*, 846-853. doi: doi:10.1002/gepi.20540

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

Ritchie, M. D. (2012, October). The success of pharmacogenomics in moving genetic association studies from bench to bedside: study design and implementation of precision medicine in the post-GWAS era. *Human Genetics*, *131*(10), 1615–1626. doi: doi:10.1007/s00439-012-1221-z

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

Sinnott-Armstrong, N., Naqvi, S., Rivas, M., & Pritchard, J. K. (2021, 2). Gwas of three molecular traits highlights core genes and pathways alongside a highly polygenic background. *eLife*, *10*, 1-35. doi: doi:10.7554/eLife.58615

Stromeyer, R. (1959, July). *Moderne probleme des magazinbaues in Deutschland und seinen Nachbarländern* . Rainald Stromeyer. *The Library Quarterly*, *29*(3), 218–220. Retrieved 2022-04-15, from https://www.journals.uchicago.edu/doi/10.1086/618679 doi: doi:10.1086/618679

Sullivan, P. F., Daly, M. J., Ripke, S., Lewis, C. M., Lin, D. Y., Wray, N. R., ... Viktorin, A. (2013, 4). A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular Psychiatry*, *18*, 497-511. doi: doi:10.1038/mp.2012.21

Sung, Y. J., Schwander, K., Arnett, D. K., Kardia, S. L. R., Rankinen, T., Bouchard, C., ... Rao, D. C. (2014, May). An empirical comparison of meta-analysis and mega-analysis of individual participant data for identifying gene-environment interactions. *Genetic Epidemiology*, *38*(4), 369–378. doi: doi:10.1002/gepi.21800

Sung, Y. J., Winkler, T. W., Manning, A. K., Aschard, H., Gudnason, V., Harris, T. B., ... Cupples, L. A. (2016, July). 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*(5), 404–415. doi: doi:10.1002/gepi.21978

Traylor, M., Markus, H., & Lewis, C. M. (2015, 6). Homogeneous case subgroups increase power in genetic association studies. *European Journal of Human Genetics*, *23*, 863-869. doi: doi:10.1038/ejhg.2014.194

Willer, C. J., Li, Y., & Abecasis, G. R. (2010, 7). Metal: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, *26*, 2190-2191. doi: doi:10.1093/bioinformatics/btq340

Zhang, L., & Sun, L. (2021, March). A generalized robust allele-based genetic association test. *Biometrics*(biom.13456).

Figures

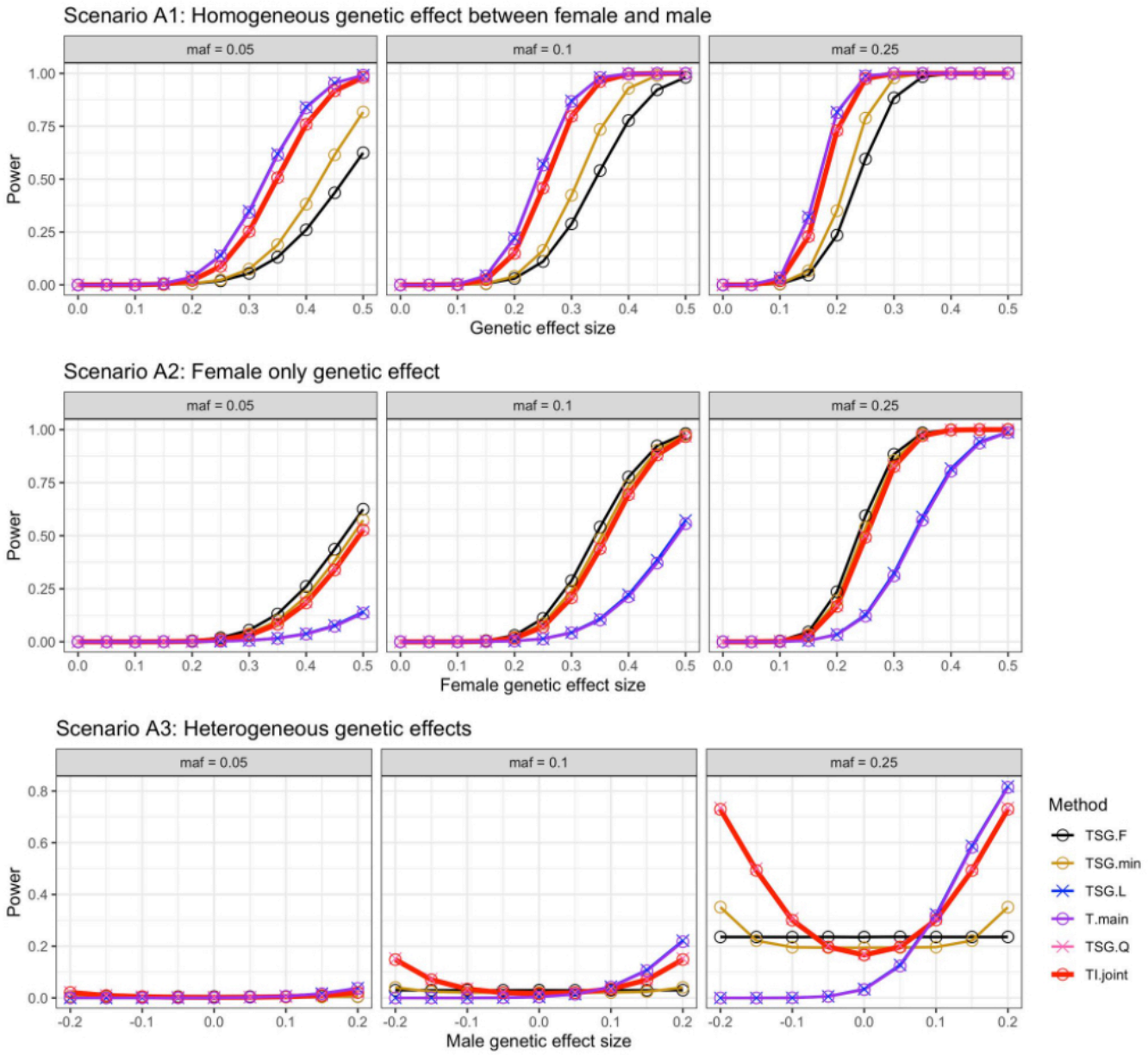

Figure 1: **Power at $\alpha = 5 \times 10^{-6}$ and female sample sizes $n_f = 1,000$ under the three scenarios across different effect sizes, stratified by minor allele frequencies and male-to-female sample size.** The three alternative scenarios are: A1. Homogeneous genetic effect between female and male; A2. Female-only genetic effect; A3. Heterogeneous genetic effect between female and male. The genetic effect sizes ranged from 0 to 0.5, Six association testing methods were evaluated: (1). The female-only sex-stratified analysis (TSG_f); (2). The minimum p-value of female and male sex-stratified analysis (TSG_{min}); (3). A meta-analysis combining Z-scores (TSG_L); (4). An omnibus test combining squared Z-scores (TSG_Q); (5). The main-effect-only model (T_{main}); (6). The recommended joint test of both the main and interaction effects.

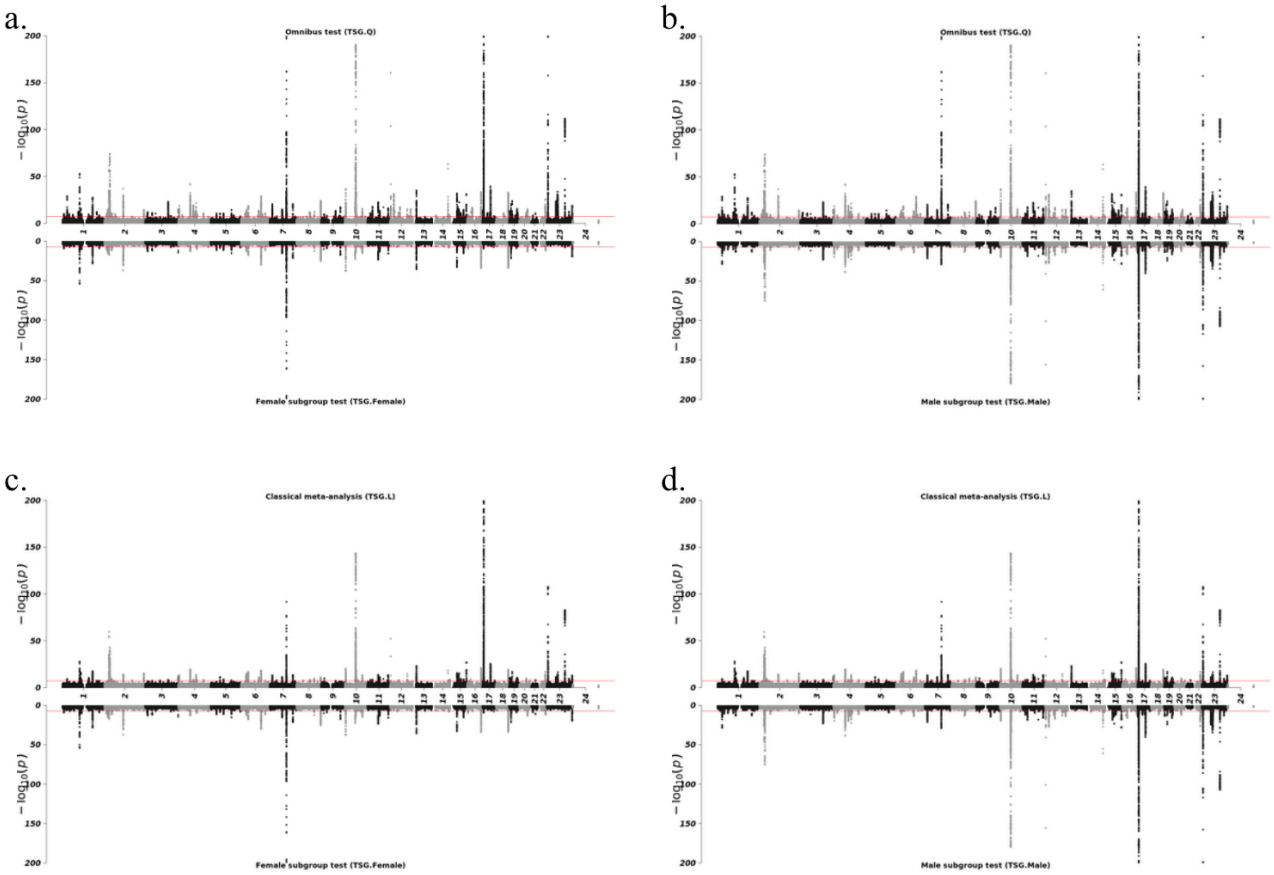

Figure 2: **Miami plots comparing the $-\log_{10}(p)$ between tests based on (a) TSG_Q and TSG_{Female}, (b) TSG_Q and TSG_{Male}, (c) TSG_L and TSG_{Female}, and (d) TSG_L and TSG_{Male}, leveraging the sex-stratified GWAS summary statistics of testosterone level from Sinnott-Armstrong et al. (2021).** Here TSG_{Female} and TSG_{Male} 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 Supplementary Figures S5 and S6, respectively.

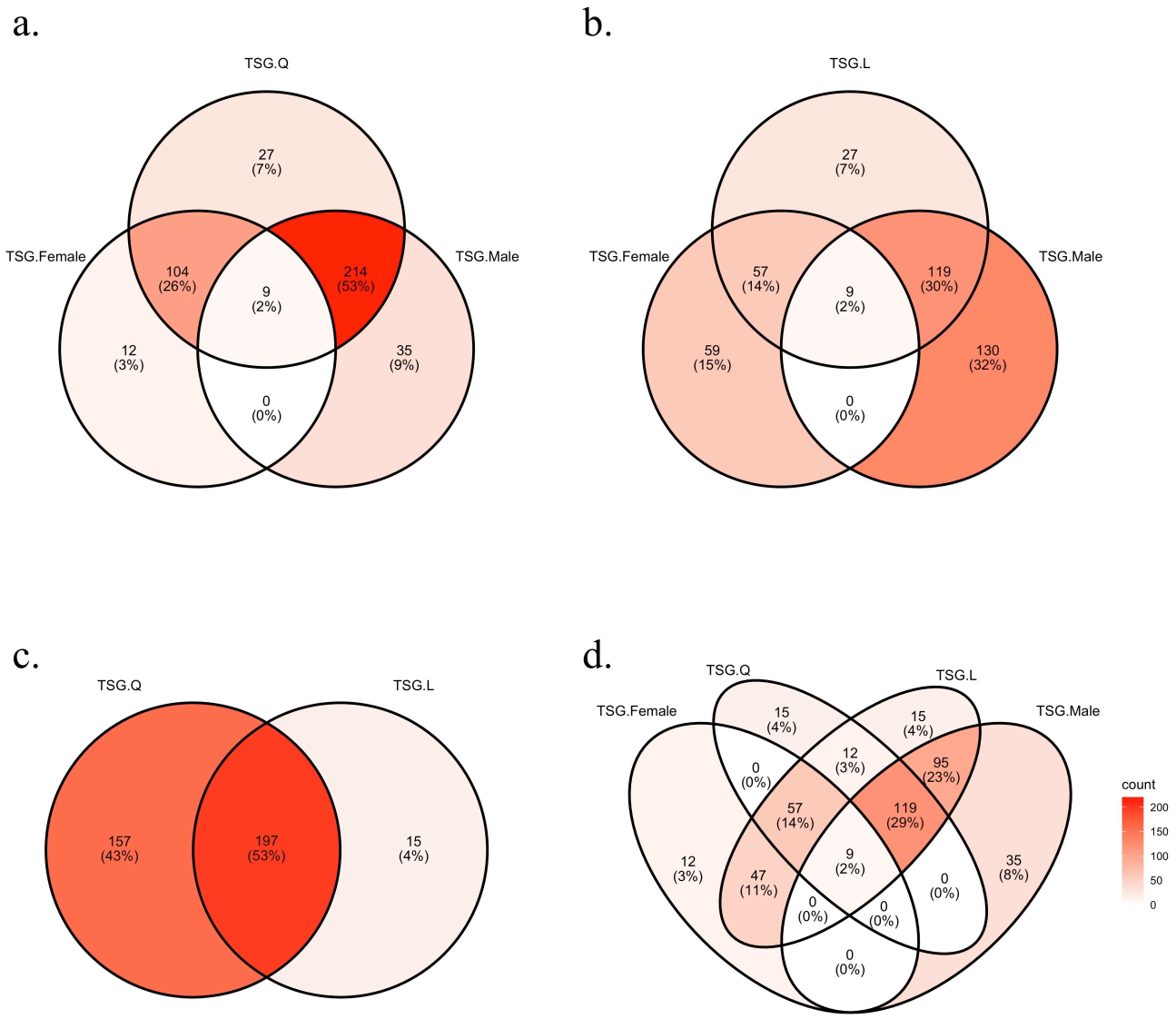

Figure 3: **Venn diagrams showing the relationship between the independent genome-wide significant loci identified from (a) TSG_Q and the sex-stratified tests, TSG_{Female} and TSG_{Male}, (b) TSG_L and the sex-stratified tests, (c) TSG_Q and TSG_L, and (d) All of the four tests.** Here TSG_{Female} and TSG_{Male} 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.

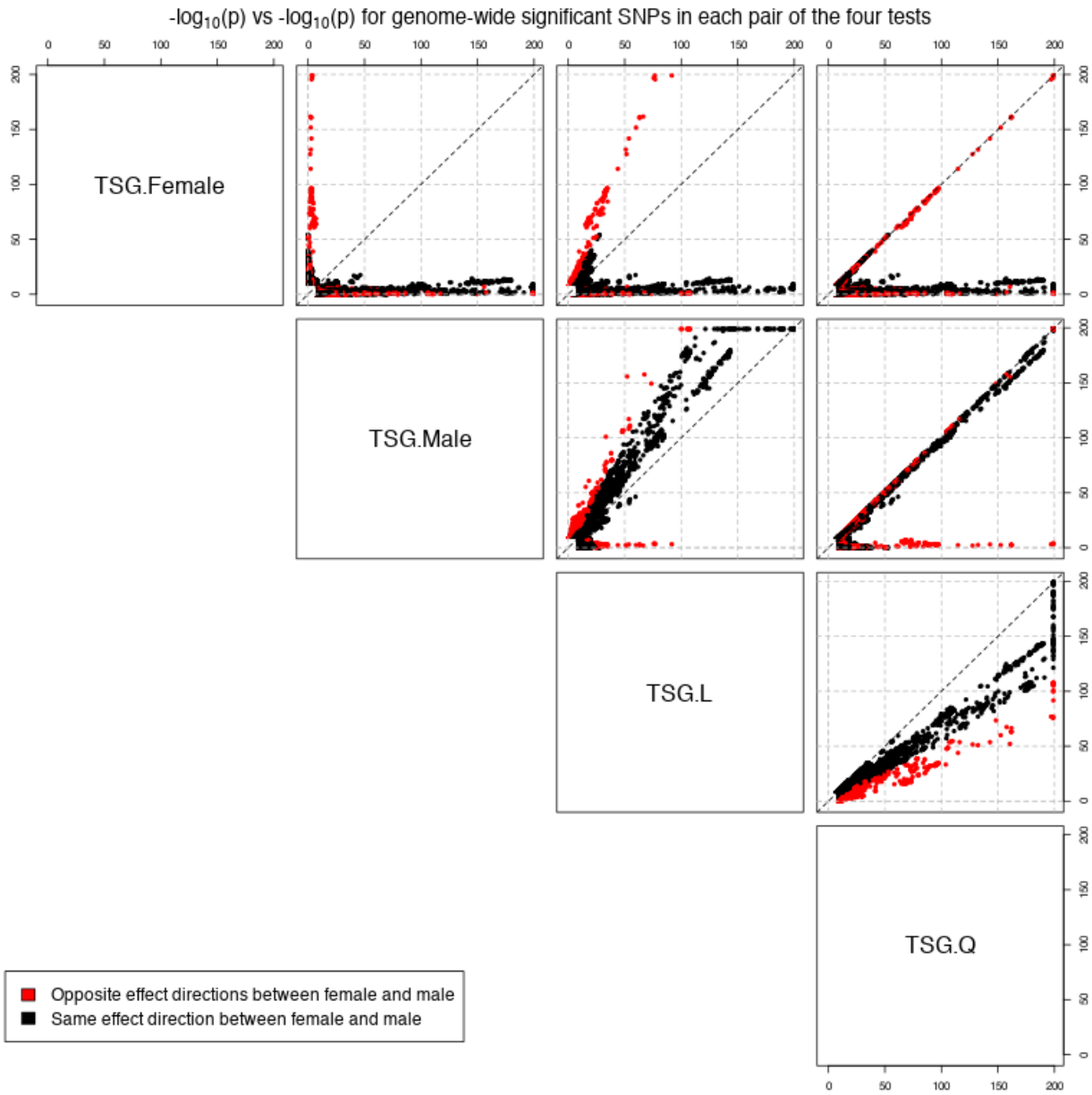

Figure 4: **Pairwise P-P plots comparing the $-\log_{10} p$-values between the female-only sex-stratified analysis (TSG_{Female}), male-only sex-stratified analysis (TSG_{Male}), the classical meta-analysis (TSG_L), and the omnibus test combining squared Z-scores (TSG_Q).** In each sub-plot, each point 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. P-P plots with $-log_{10}(p)$ truncated at 50 and 10 are shown in Supplementary Figures S7 and S8, respectively.

Figure 5: **Restricted pairwise P-P plots, considering (a) and (b): SNPs identified by TSG_Q but missed by either TSG_{Female} or TSG_{Male}, (c) and (d): SNPs identified by either TSG_{Female} or TSG_{Male} but missed by TSG_Q.** Here TSG_{Female} and TSG_{Male} 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.

Supplementary Material

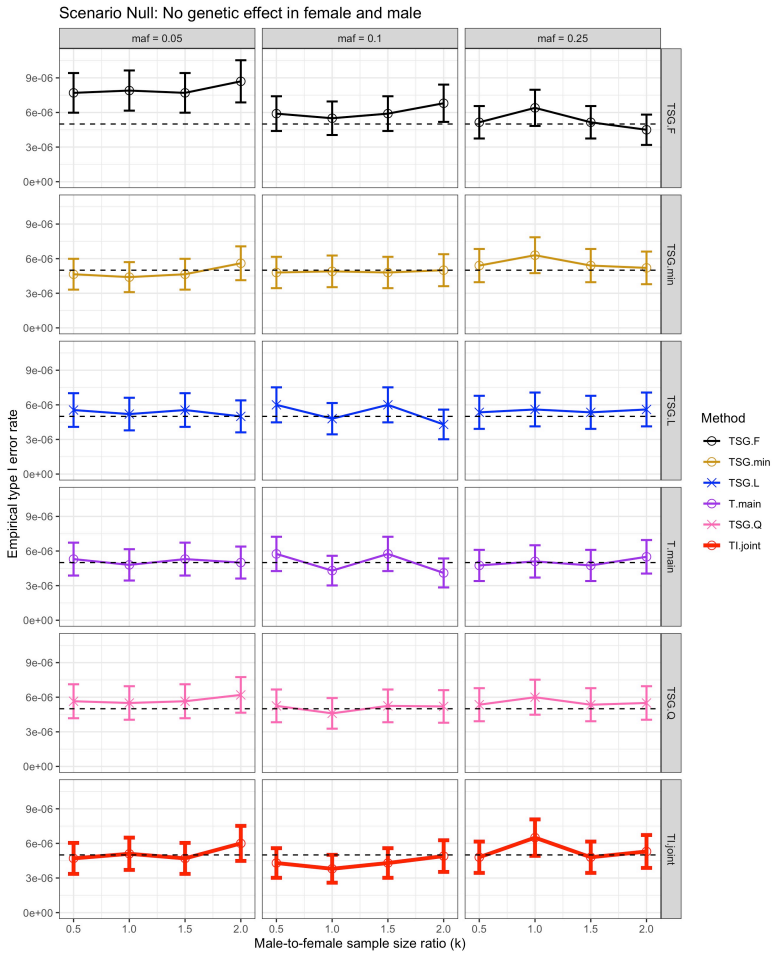

Figure S1. The empirical type I error rates at nominal $\alpha = 5 \times 10^{-6}$ based on $R = 10^7$ replications and their 95% confidence intervals under the null scenario, with $n_f = 1,000$, across different $k = n_m/n_f$ and MAF. Six association testing methods were evaluated: (1). The female-only subgroup analysis (TSG_f); (2). The minimum p-value of female and male subgroup analysis (TSG_{min}); (3). A meta-analysis combining Z-scores (TSG_L); (4). An omnibus test combining squared Z-scores (TSG_Q); (5). The main-effect-only model (T_{main}); (6). The recommended joint test of both the main and interaction effects. The dashed line indicates nominal type I error rate at 5×10^{-6}. The 95% Binomial proportion confidence interval is constructed by formula: $\left(\hat{\alpha} - 1.96\sqrt{(1-\hat{\alpha})\hat{\alpha}/R}, \hat{\alpha} + 1.96\sqrt{(1-\hat{\alpha})\hat{\alpha}/R}\right)$, where $\hat{\alpha}$ is the empirical type I error rate.

Note: Readers may notice that the nominal level was not at the center of the confidence intervals (CIs) for all the tests except for TSG_L and T_{main} at MAF=0.25 and $k = 1$, nevertheless it was still covered by the CIs. The slight inflation is possibly due to randomness. As we ran the type I error simulation at k = 1 and maf = 0.25 using a different seed, and we observed that the inflation disappeared, with $\hat{\alpha}$ being 5.2×10^{-6}, 5.9×10^{-6}, 5.9×10^{-6}, 6.1×10^{-6}, 5.1×10^{-6}, 5.2×10^{-6} respectively for the six tests (with the order appeared in the Figure S1 legend).

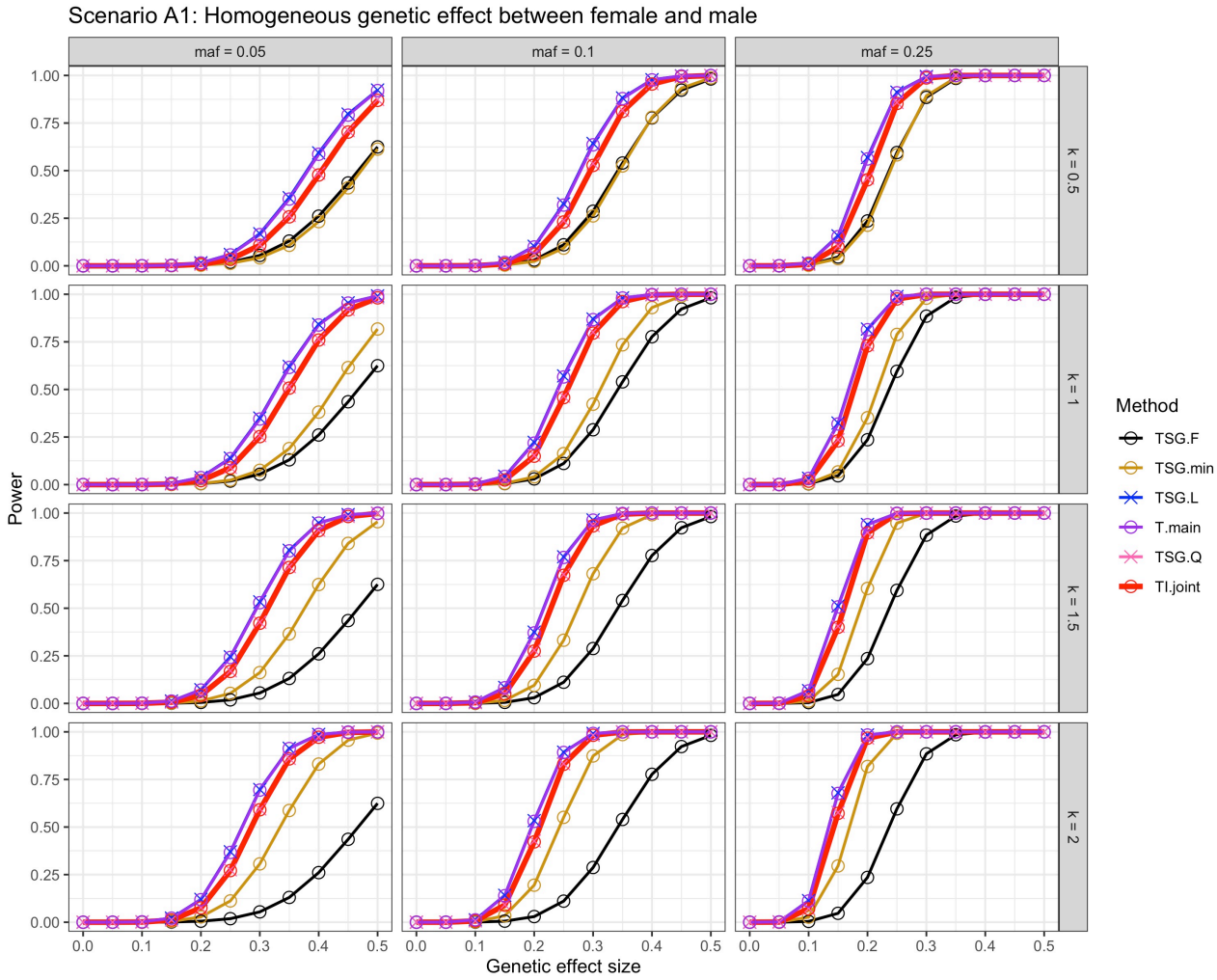

Figure S2. Power at $\alpha = 5 \times 10^{-6}$ and female sample sizes $n_f = 10^3$ under the scenario A1. Homogeneous genetic effect across different effect sizes, stratified by minor allele frequencies and male-to-female sample size. The genetic effect sizes ranged from 0 to 0.5, Six association testing methods were evaluated: (1). The female-only subgroup analysis (TSG_f); (2). The minimum p-value of female and male subgroup analysis (TSG_{min}); (3). A meta-analysis combining Z-scores (TSG_L); (4). An omnibus test combining squared Z-scores (TSG_Q); (5). The main-effect-only model (T_{main}); (6). The recommended joint test of both the main and interaction effects.

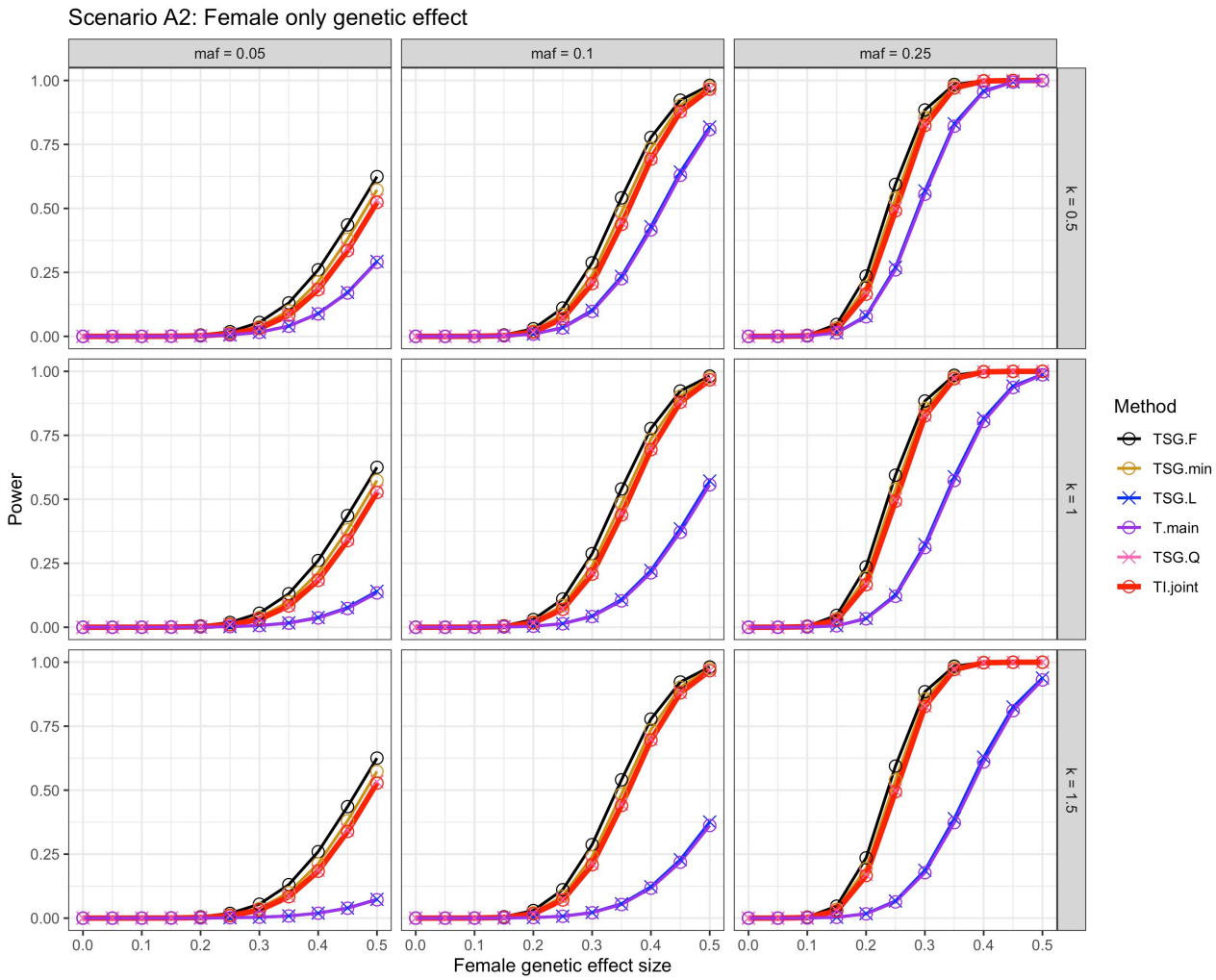

Figure S3. Power at $\alpha = 5 \times 10^{-6}$ and female sample sizes $n_f = 10^3$ under the scenario A2: Female-only genetic effect across different effect sizes, stratified by minor allele frequencies and male-to-female sample size. The genetic effect sizes ranged from 0 to 0.5, Six association testing methods were evaluated: (1). The female-only subgroup analysis (TSG_f); (2). The minimum p-value of female and male subgroup analysis (TSG_{min}); (3). A meta-analysis combining Z-scores (TSG_L); (4). An omnibus test combining squared Z-scores (TSG_Q); (5). The main-effect-only model (T_{main}); (6). The recommended joint test of both the main and interaction effects.

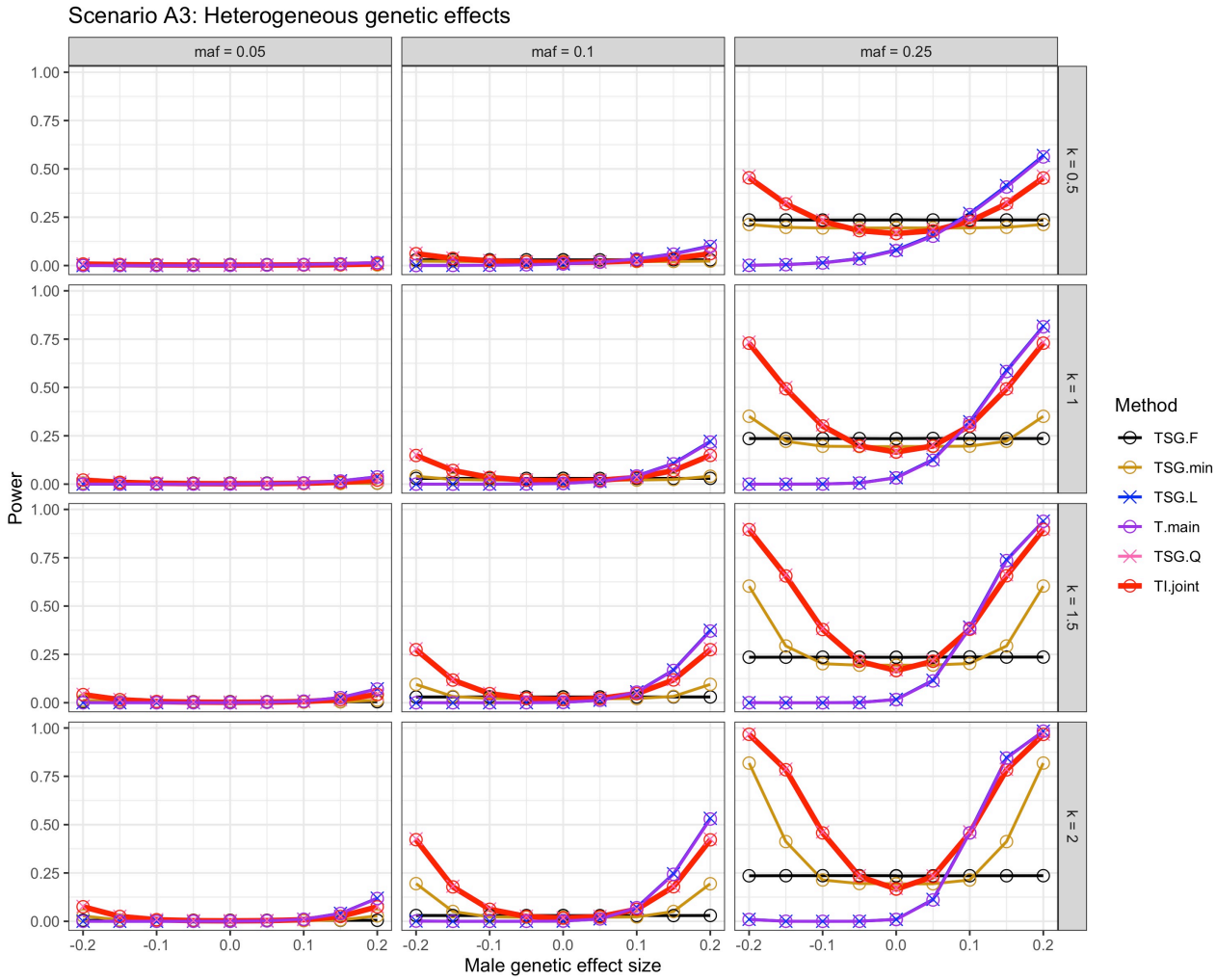

Figure S4. Power at $\alpha = 5 \times 10^{-6}$ and female sample sizes $n_f = 10^3$ under the scenario A3. Heterogeneous genetic effect between female and male, stratified by minor allele frequencies and male-to-female sample size. The genetic effect sizes ranged from 0 to 0.5, Six association testing methods were evaluated: (1). The female-only subgroup analysis (TSG_f); (2). The minimum p-value of female and male subgroup analysis (TSG_{min}); (3). A meta-analysis combining Z-scores (TSG_L); (4). An omnibus test combining squared Z-scores (TSG_Q); (5). The main-effect-only model (T_{main}); (6). The recommended joint test of both the main and interaction effects.

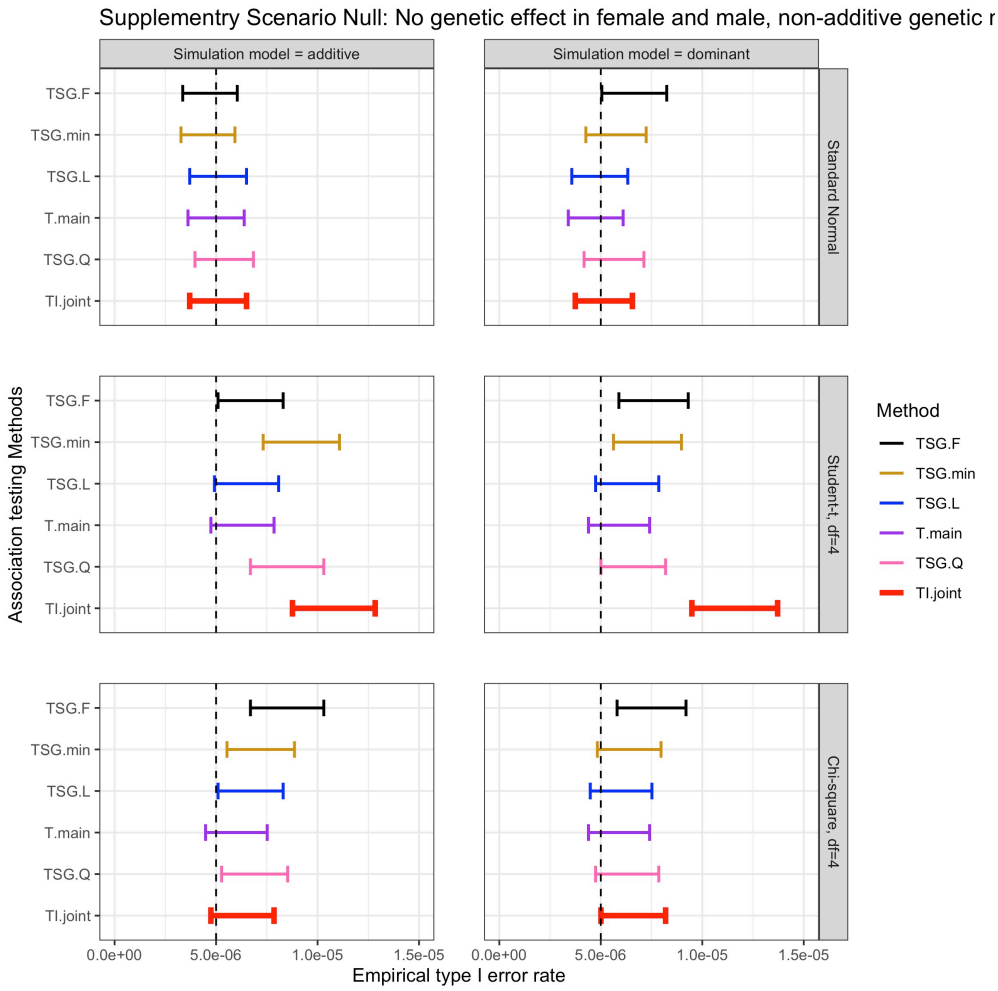

Figure S5. The empirical type I error rates at nominal $\alpha = 5 \times 10^{-6}$ based on $R = 10^7$ replications and their 95% confidence intervals under the null scenario, with $n_f = n_m = 1,000$ and $MAF = 0.25$, stratified by genetic models and residual distributions. The rows correspond to the residual distributions generated from Standard Normal, Student t_4 and χ^2_4 respectively. The columns correspond to the additive and dominant genetic models respectively, although we assume additive model for association testing. Six association testing methods were evaluated: (1). The female-only subgroup analysis (TSG_f); (2). The minimum p-value of female and male subgroup analysis (TSG_{min}); (3). A meta-analysis combining Z-scores (TSG_L); (4). An omnibus test combining squared Z-scores (TSG_Q); (5). The main-effect-only model (T_{main}); (6). The recommended joint test of both the main and interaction effects. The dashed line indicates nominal type I error rate at 5×10^{-6}. The 95% Binomial proportion confidence interval is constructed by formula: $\left(\hat{\alpha} - 1.96\sqrt{\frac{(1-\hat{\alpha})\hat{\alpha}}{R}}, \hat{\alpha} + 1.96\sqrt{\frac{(1-\hat{\alpha})\hat{\alpha}}{R}}\right)$, where $\hat{\alpha}$ is the empirical type I error rate.

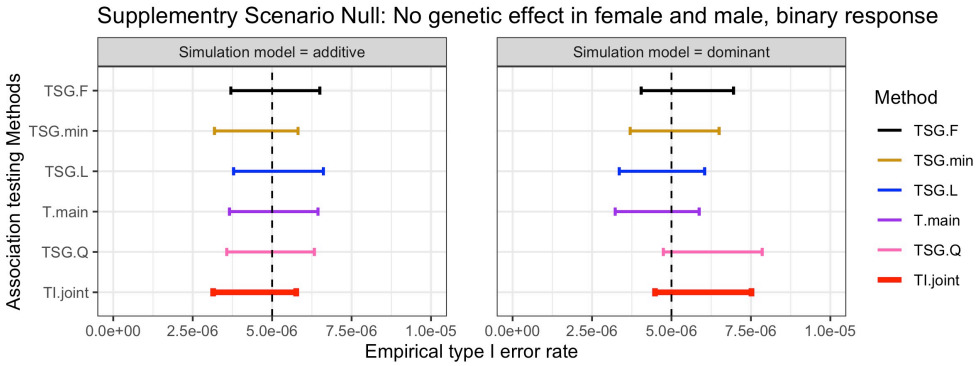

Figure S6. The empirical type I error rates with binary response at nominal $\alpha = 5 \times 10^{-6}$ based on $R = 10^7$ replications and their 95% confidence intervals under the null scenario, with $n_f = n_m = 1,000$ and $MAF = 0.25$. We simulated data from additive (Left) and dominant genetic models (Right) respectively, although we assume additive model for association testing. Six association testing methods were evaluated: (1). The female-only subgroup analysis (TSG_f); (2). The minimum p-value of female and male subgroup analysis (TSG_{min}); (3). A meta-analysis combining Z-scores (TSG_L); (4). An omnibus test combining squared Z-scores (TSG_Q); (5). The main-effect-only model (T_{main}); (6). The recommended joint test of both the main and interaction effects. The dashed line indicates nominal type I error rate at 5×10^{-6}. The 95% Binomial proportion confidence interval is constructed by formula: $\left(\hat{\alpha} - 1.96\sqrt{\frac{(1-\hat{\alpha})\hat{\alpha}}{R}}, \hat{\alpha} + 1.96\sqrt{\frac{(1-\hat{\alpha})\hat{\alpha}}{R}}\right)$, where $\hat{\alpha}$ is the empirical type I error rate.

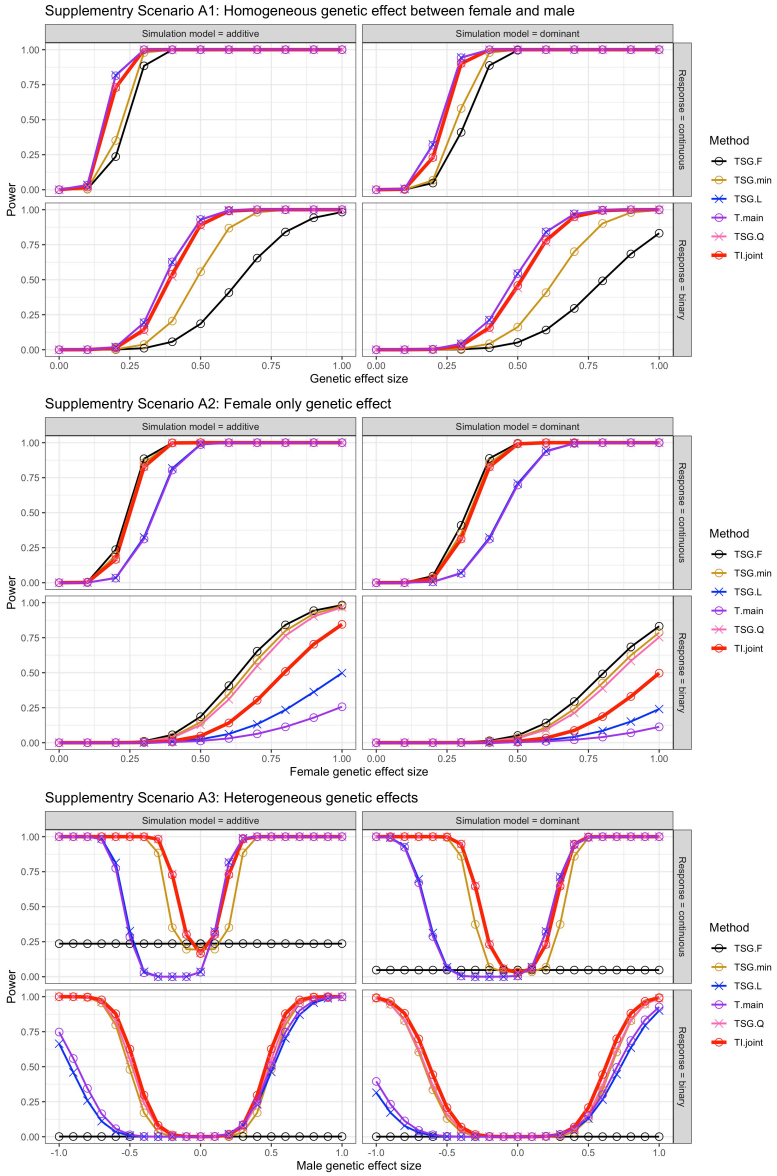

Figure S7. Power at $\alpha = 5 \times 10^{-6}$ under the three alternative scenarios, stratified by genetic models and response types. Sample sizes $n_f = n_m = 10^3$, and $MAF = 0.25$. The genetic effect sizes ranged from 0 to 0.5. The columns correspond to the additive and dominant genetic models respectively, although we assume additive model for association testing. Within each panel, the top row is for continuous response, and the bottom row is for binary response. Six association testing methods were evaluated: (1). The female-only subgroup analysis (TSG_f); (2). The minimum p-value of female and male subgroup analysis (TSG_{min}); (3). A meta-analysis combining Z-scores (TSG_L); (4). An omnibus test combining squared Z-scores (TSG_Q); (5). The main-effect-only model (T_{main}); (6). The recommended joint test of both the main and interaction effects.

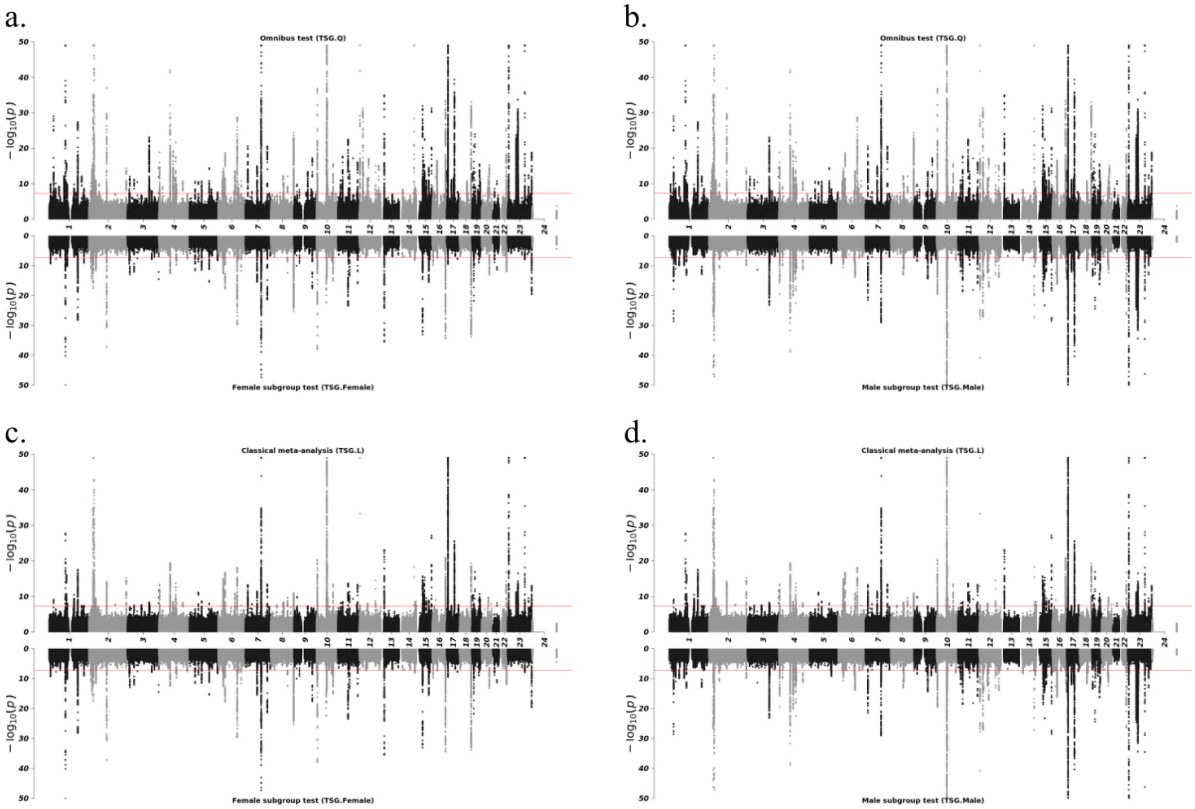

Figure S8. Miami plots comparing the $-\log_{10}(p)$ between tests based on (a.) TSG_Q and TSG_{Female}, (b.) TSG_Q and TSG_{Male}, (c.) TSG_L and TSG_{Female}, (d.) TSG_L and TSG_{Male} of GWAS on testosterone level from Sinnott-Armstrong et al., 2021. Here TSG_{Female}, TSG_{Male} are the test statistics of female- and male only subgroup analysis respectively. The omnibus test combining squared Z-scores (TSG_Q) and meta-analysis combining Z-scores (TSG_L) were calculated by combining TSG_{Female} and TSG_{Male}. The red line indicates genome-wide significant threshold at 5×10^{-8}. The maximum $-\log_{10} p$ is truncated at 50.

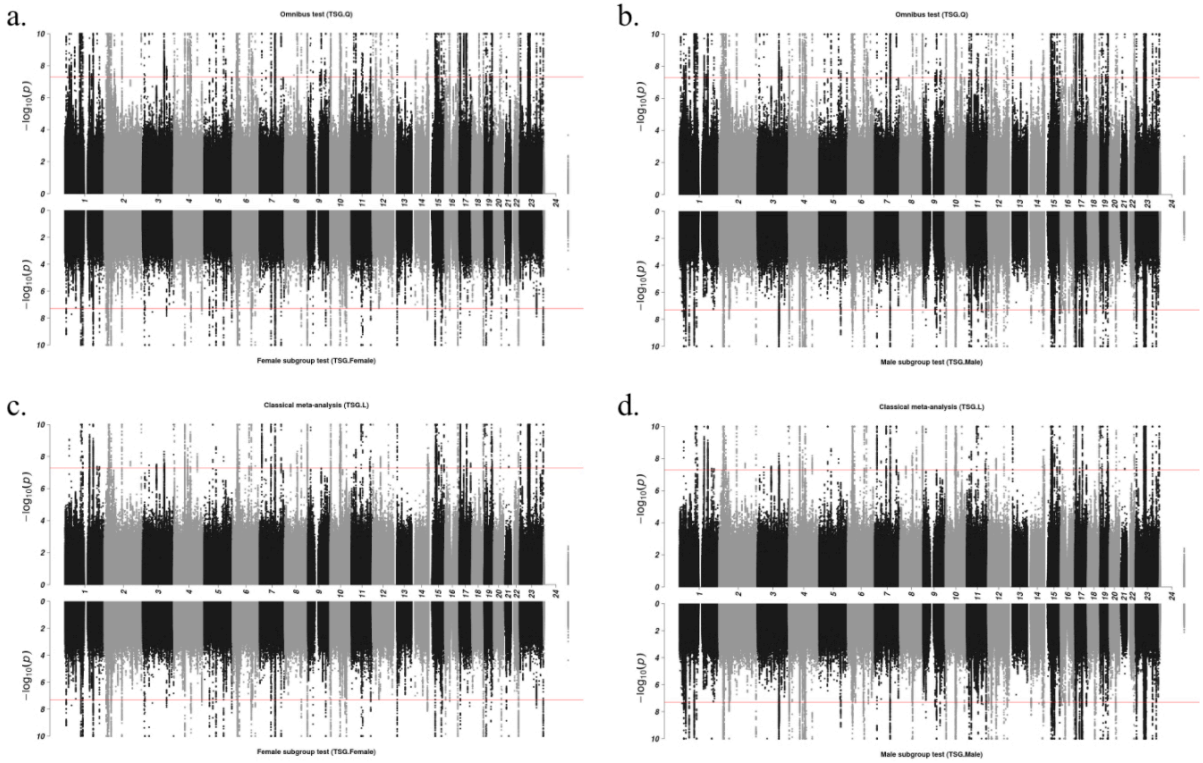

Figure S9. Miami plots comparing the $-\log_{10}(p)$ between tests based on (a.) TSG_Q and TSG_{Female}, (b.) TSG_Q and TSG_{Male}, (c.) TSG_L and TSG_{Female}, (d.) TSG_L and TSG_{Male} of GWAS on testosterone level from Sinnott-Armstrong et al., 2021. Here TSG_{Female}, TSG_{Male} are the test statistics of female- and male only subgroup analysis respectively. The omnibus test combining squared Z-scores (TSG_Q) and meta-analysis combining Z-scores (TSG_L) were calculated by combining TSG_{Female} and TSG_{Male}. The red line indicates genome-wide significant threshold at 5×10^{-8}. The maximum $-\log_{10} p$ is truncated at 10.

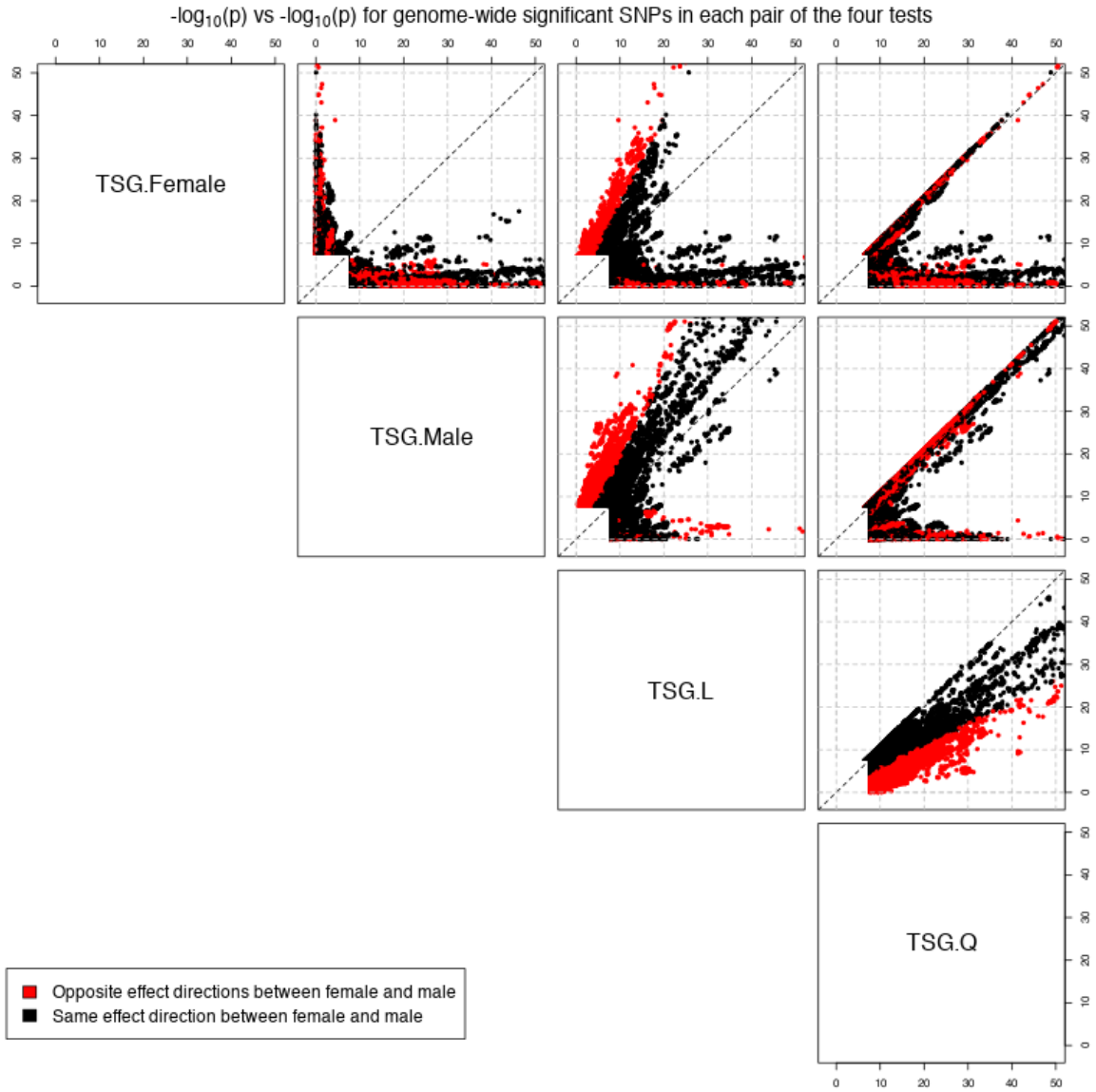

Figure S10. Pairwise P-P plots comparing the $-\log_{10} p$-values between the female-only subgroup analysis (TSG_{Female}), male-only subgroup analysis (TSG_{Male}), the classical meta-analysis (TSG_L), and the omnibus test combining squared Z-scores (TSG_Q). In each sub-plot, the points represent the variants reported genome-wide significant in any of the two corresponding testing methods. The p-values for TSG_{Female} and TSG_{Male} were obtained from summary statistics of Sinnott-Armstrong et al., 2021, while TSG_Q and TSG_L were calculated by combining TSG_{Female} and TSG_{Male}. The p-values are truncated at 50.

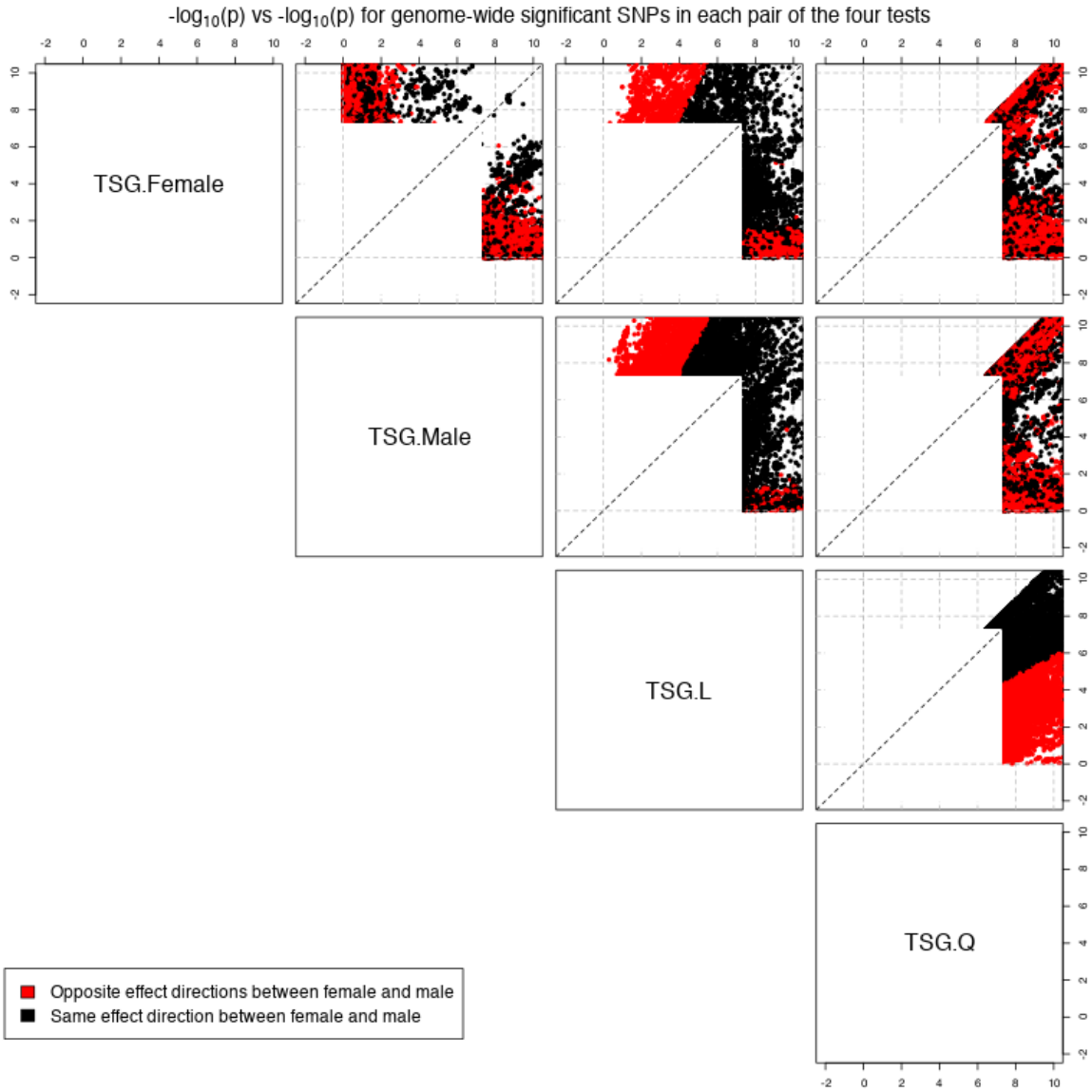

Figure S11. Pairwise P-P plots comparing the $-\log_{10} p$-values between the female-only subgroup analysis (TSG_{Female}), male-only subgroup analysis (TSG_{Male}), the classical meta-analysis (TSG_L), and the omnibus test combining squared Z-scores (TSG_Q). In each sub-plot, the points represent the variants reported genome-wide significant in any of the two corresponding testing methods. The p-values for TSG_{Female} and TSG_{Male} were obtained from summary statistics of Sinnott-Armstrong et al., 2021, while TSG_Q and TSG_L were calculated by combining TSG_{Female} and TSG_{Male}. The p-values are truncated at 10.