An association mapping framework to account for the sex difference in genetic architectures

Eun Yong Kang [1,†], Cue Hyunkyu Lee [5,6,†], Nicholas A. Furlotte [1], Jong Wha Joo [2], Emrah Kostem [1], Noah Zaitlen [3], Eleazar Eskin [1,4,*], Buhm Han [5,6,*]

1 Computer Science Department, University of California, Los Angeles, California, USA

2 Interdepartmental Program in Bioinformatics, University of California Los Angeles, California, USA

3 Department of Medicine, University of California San Francisco, California, USA

4 Department of Human Genetics, University of California, Los Angeles, California, USA

5 Asan Institute for Life Sciences, Asan Medical Center, Seoul, Republic of Korea

6 Department of Convergence Medicine, University of Ulsan College of Medicine, Seoul, Republic of Korea

∗ Corresponding Authors E-mails: eeskin@cs.ucla.edu, buhm.han@amc.seoul.kr

† These authors contributed equally.

Abstract

Over the past few years, genome-wide association studies have identified many trait-associated loci that have different effects on females and males, which increased attention to the genetic architecture differences between the sexes. The between-sex differences in genetic architectures can cause a variety of phenomena such as the differences in effect sizes at trait-associated loci, differences in the magnitudes of polygenic background effects, and differences in phenotypic variances. However, current approaches for dealing with sex such as including sex as a covariate cannot fully account for these phenomena and can be suboptimal in statistical power. We present a novel framework, MetaSex, that can comprehensively account for the genetic architecture differences between the sexes. Through simulations and applications to real data, we show that our framework has a superior performance than previous approaches.

1 Introduction

Genome-wide association studies (GWAS) have successfully identified numerous genetic loci associated with complex human traits. In recent years, increasing attention has been paid to the sex difference in genetic architectures in GWAS. A number of studies have found differences in effect sizes between males and females on loci associated with traits. (Boraska *et al.*, 2012; Chen *et al.*, 2013; Fox *et al.*, 2012; Kang *et al.*, 2014; Kostis *et al.*, 2012; Kubo *et al.*, 2013; Magi *et al.*, 2010; Mason and Lehert, 2012; Ohmen *et al.*, 2014; Peters *et al.*, 2013; Porcu *et al.*, 2013; Randall *et al.*, 2013). In particular, a large-scale meta-analysis of 46 studies of anthropomorphic phenotypes discovered seven loci with different effects between the sexes (Randall *et al.*, 2013).

It remains unclear how best to account for the sex difference in genetic architectures in association mapping. One traditional approach is to analyze each sex separately using sex-specific tests (SST). This approach is optimal for detecting sex-specific effects that only exist in one sex, but is not powerful for detecting effects that exist in both sexes. Another traditional approach is to analyze the whole sample and use sex as a covariate (CV). This approach is optimal for detecting effects that exist in both sexes in a constant effect size, but is not powerful for detecting sex-interacting effects that exist in both sexes in differing effect sizes.

In the present study, we first enumerate three possible phenomena that can be caused by the sex difference in genetic architectures. One is the effect size difference between the sexes at the associated locus, which was observed in previous studies (Randall *et al.*, 2013). Another is the effect size difference between the sexes at numerous loci spread throughout the genome with small effects, which can manifest as polygenic background effects that interact with sex. The final one is the phenotypic variance difference between the sexes, which can be caused by many factors such as sex acting as a biological environment (e.g. hormone difference) and sex interacting with external environments (e.g. lifestyle difference). In our analysis of the North Finland Birth Cohort (NFBC) dataset (Sabatti *et al.*, 2009), we often observed these phenomena in some human traits.

Here, we present a novel association mapping framework, MetaSex, that accounts for the sex difference in genetic architectures to effectively map associations in GWAS. Our framework comprehensively takes into account the three aforementioned phenomena by uniquely combining linear mixed model and meta-analysis. Our linear mixed model includes five variance components, where three are for capturing sex-interacting polygenic effects and two are for modeling distinct variances of the two sexes. We then combine the observed effect sizes of the two sexes using the random effects meta-analysis (Han and Eskin, 2011) that provides high power for detecting sex-interacting effects. Because this five variance component model can be impractically slow to be applied to millions of markers in GWAS, we propose an approximated model that splits the problem into two optimization problems each requiring only two variance components. Using simulations and real data, we demonstrate that our framework is powerful, comprehensive, and efficient.

2 Results

2.1 Overview of MetaSex

We first provide an overview of our proposed framework. We constructed a toy example with six individuals (three females and three males). The equation in Figure 1A shows the components in our model for testing a single SNP. In this equation, vector $\mathbf{y}$ is the observed phenotype measurements, where subscripts (f) and (m) denote females and males. μ denotes the phenotypic mean. $\mathbf{h}$ is the sex status indicator (female = 1 and male = 0), which is included as a covariate to account for the sex-specific phenotypic mean. The first column of $\mathbf{X}$ is the genotype vector of the SNP, whose effect size is β. The second column of $\mathbf{X}$ is the genotype-by-sex interaction term (SNP $\times$ $\mathbf{h}$), whose effect size is $\beta_{g\times s}$. $\mathbf{u_g}$ is a variance component that models the polygenic background effects from the genome-wide loci that affect both sexes. Consistently to the standard linear mixed model(Kang *et al.*, 2010, 2008; Zhou and Stephens, 2012), we assume that $\mathbf{u_g}$ follows a normal distribution with mean zero and variance-covariance matrix

$\sigma_g^2\mathbf{K}$ where $\mathbf{K}$ is the kinship matrix representing the relationship between individuals. $\mathbf{u_f}$ is an additional variance component that we introduced, which represents the female-specific polygenic effects. We assume that $\mathbf{u_f}$ have mean zero and variance $\sigma_{g,f}^2(\mathbf{K} \circ \mathbf{hh^T})$. Similarly, $\mathbf{u_m}$ is a variance component representing the male-specific polygenic effects, which have mean zero and variance $\sigma_{g,m}^2(\mathbf{K} \circ (\mathbf{1}-\mathbf{h})(\mathbf{1}-\mathbf{h})^\mathbf{T})$. We then modeled separate error terms for females and males, assuming that error variances can be different. $\mathbf{e_f}$ is a female-specific error term that follows a normal distribution with mean zero and variance $\sigma_{e,f}^2(\mathbf{I} \circ \mathbf{hh^T})$, where $\mathbf{I}$ is the identity matrix. Similarly, $\mathbf{e_m}$ is a male-specific error term that have mean zero and variance $\sigma_{e,m}^2(\mathbf{I} \circ (\mathbf{1}-\mathbf{h})(\mathbf{1}-\mathbf{h})^\mathbf{T})$.

Applying this full model to GWAS can be computationally challenging because there are five variance components to fit ($\mathbf{u_g}$, $\mathbf{u_f}$, $\mathbf{u_m}$, $\mathbf{e_f}$, and $\mathbf{e_m}$). Currently available linear mixed model methods for association mapping are optimized for models with two variance components (Kang *et al.*, 2010, 2008; Zhou and Stephens, 2012). If there is a third component, a state-of-the-art association mapping method uses a simple grid search (Lippert *et al.*, 2011). Thus, fitting five variance components may require a three-dimensional grid search, which can be prohibitively slow for GWAS.

To expedite the application of our model to GWAS, we propose an efficient decomposition of the model. Suppose that we restrict our scope to individuals of one sex. Then, the full model with five variance components collapses into a sex-specific model with two variance components (Figure 1B). Thus, the model can be efficiently solved using existing approaches (Kang *et al.*, 2010, 2008; Zhou and Stephens, 2012). In the decomposed model, we cannot distinguish the whole-sample polygenic component (σ_g^2) from the sex-specific polygenic components ($\sigma_{g,f}^2$ or $\sigma_{g,m}^2$), because they follow the exactly same distribution conditioned on one sex. However, this distinction is unimportant for association mapping, because we anyhow want to control for both.

Finally, given the sex-specific effect size estimates and the standard errors, ($\hat{\beta_m}$, se($\hat{\beta_m}$), $\hat{\beta_f}$, se($\hat{\beta_f}$)), we apply a series of statistical tests. We first apply SST, which is optimal for detecting sex-specific effects. Then, to effectively detect sex-interacting effects, we combine the two sex-specific estimates using the random effects meta-analysis method (RE) (Han and Eskin, 2011) (Figure 1C), which explicitly models heterogeneity. As a result, our framework involves three tests (Female SST, Male SST, and RE), requiring multiple testing correction. A powerful multiple testing strategy can be to adjust the significance threshold for each test to maximize power while controlling for overall false positive rate (Eskin, 2008). We identify and propose a set of thresholds for the three tests, what we call smart thresholding, that exactly controls the false positive rate to the GWAS threshold (5×10^{-8}) while maximizing power.

2.2 Power simulations

We performed simulations to evaluate the power of our MetaSex approach. Below, we also refer to our MetaSex method as RE+SST because the framework involves simultaneous testing of RE and SST. We compared our method to two other approaches: (1) CV, the traditional approach using sex as a covariate, and (2) GWAMA (Magi *et al.*, 2010), another meta-analysis approach designed to discover sex-interacting effects. Because CV and GWAMA are methods using the whole sample similarly to RE, we assumed the common practice in which investigators examine each sex separately using SST regardless of which method is used for the whole sample. Thus, we compared the power of our MetaSex (RE+SST) approach with that of CV+SST and GWAMA+SST, where A+B denotes a combination method that calls a result significant if either A or B method gives a significant result after correcting for multiple testing. We also compared the power of the bare SST to get a sense of how much power is increased by the methods using the whole sample.

To make a fair comparison between these methods, we corrected for multiple testing within each method in an equitable way. In MetaSex (RE+SST), CV+SST, and GWAMA+SST, we performed three tests, whereas in SST, we performed two tests. Therefore, for each of these methods, we generated 10 billion (10^{10}) null male/female statistic pairs and chose the 500th smallest p-value, which was the method-specific significance threshold to control the false positive rate (family-wise error rate) to 5×10^{-8}. The resulting significance thresholds were 1.70×10^{-8}

for CV+SST, 1.73×10^{-8} for GWAMA+SST, and 2.49×10^{-8} for SST. For MetaSex (RE+SST), we used our smart thresholding strategy that applied 2.41×10^{-8} for RE and 1.36×10^{-8} for each of the two SST (see Methods). These empirically calculated thresholds ensured that the false positive rates of all compared methods were well controlled. See Methods for the further details of our power simulations.

Simulating the sex difference in effect sizes

In the first power simulation, we simulated the effect size difference between the sexes, a phenomenon called "effect size heterogeneity". We assumed a SNP of minor allele frequency 0.3 and generated genotypes of 5,000 males and 5,000 females. Then we simulated continuous phenotypes of these individuals while assuming the same error variance for the two sexes. For each male individual, we generated a phenotype assuming a genetic effect size of 0.192 and variance of 1.0. For each female individual, we generated a phenotype assuming 10% of the male effect size (0.019) and variance of 1.0. We repeated this simulation 1,000 times and computed the power of a method as the proportion of simulations in which the test p-value was more significant than the given significance threshold. We then gradually increased the female effect size from 10% to 100% of the male effect size, to simulate differing levels of heterogeneity.

Figure 2 shows the power of the four approaches (RE+SST, CV+SST, GWAMA+SST, and SST) with respect to the effect size ratio between the two sexes. As expected, when the effect size ratio was very small (e.g., 0.1), meaning that the effect is almost sex-specific, SST showed the highest statistical power. As the effect size of the female study increased, the MetaSex (RE+SST) approach showed the highest statistical power, demonstrating that our approach can effectively detect sex-interacting effects. Even at the ratio of 1.0, a situation that the effect size was identical for both sexes (no heterogeneity), where CV was expected to be the most powerful, MetaSex (RE+SST) slightly outperformed CV+SST. This was because of our smart thresholding strategy that allowed a more liberal significance threshold for RE with the expense of a more stringent threshold for SST.

Simulating the sex difference in error variances

In the second power simulation, we simulated the error variance difference between the sexes while assuming a constant effect size (no heterogeneity). As in the first simulation, for each male and female individual, we generated a phenotype assuming a genetic effect size of 0.2 and variance of 1.0. Then, we gradually increased the error variance of the females from 1.0 to 4.0 (standard deviation from 1.0 to 2.0). Figure 3 shows the power of the four approaches (RE+SST, CV+SST, GWAMA+SST, and SST) with respect to the standard deviation ratio between the two sexes. Our proposed approach (RE+SST) outperformed other methods in all simulated situations. When we examined the second and the third best methods, we observed that GWAMA+SST outperformed CV+SST when the standard deviation ratio was large (≥ 1.4). This was because GWAMA, being a meta-analytic approach that estimates the variance of error terms in each sex separately, was robust to the variance difference between the sexes. Athough both our MetaSex (RE+SST) and GWAMA+SST were meta-analytic methods, our method consistently outperformed GWAMA+SST.

Power characteristics of the methods

We examined why using RE to complement SST was more powerful than using GWAMA or CV to complement SST. We evaluated the power of the individual methods (RE, CV, GWAMA, and SST) over a wide range of female/male genetic effect size pairs, varying each from small value (0) to large value (1.0). Note that although we examined a specific effect size range (0,1), the general tendencies in relative power are expected to be similar in different settings; for example, if the effect size range was larger and the variances were larger, the power results would be similar. Here, we assumed a error variance ratio of 1.2 (females/males) between the two sexes, because this was the average ratio of the phenotypic variances in 10 phenotypes of the NFBC data below.

Figure 4 shows the power of the four methods (RE, CV, SST, and GWAMA) in a two-dimensional space where x-axis is the male genetic effect size and y-axis is the female genetic effect size of a SNP. We plotted the 50% power lines of the four methods, so that each line denotes pairs of the male and female effect sizes where the method achieved an exact power of 50%. Because the power increased as the effect size increased, the closer the 50% line was to the bottom leftmost point on the graph, the more powerful the method was. As expected, when one of the effect sizes was close to zero (sex-specific effect: top left corner or bottom right corner), SST was the most powerful. When the effect sizes were at most moderately different between male and female studies (middle area), RE clearly outperformed other approaches. We measured the size of the area whose power was greater than 50%, which was the area outside of the each curve toward the top right corner. The sizes of the areas were 22.1% for SST, 23.5% for CV, 29.7% for GWAMA, and 29.9% for RE. Thus, RE and GWAMA achieved the largest similar areas. However, the difference was that the GWAMA power line was more steeply curved than the RE power line, which demonstrated that GWAMA tended to detect effects that were extremely different between the sexes. Therefore, what GWAMA found could have substantial overlap to what SST found. In contrast, RE and SST complemented each other resulting in the largest combined area in this plot. To quantify this difference, we measured the area greater than 50% not covered by the 50% power of the SST approach. The areas were 13.8% for GWAMA and 15.4% for RE. Thus, if we assume the common practice of applying SST first, RE can give us the biggest additional power.

2.3 Analysis of the North Finland Birth Cohort data

We analyzed the NFBC data (Sabatti *et al.*, 2009) which consisted of 5326 individuals (2546 males and 2780 females). This dataset provided 10 phenotypic measurements of the individuals.

Sex difference in phenotypic variances

We first investigated whether the phenotypic variances showed differences between the sexes. We applied the Levene's test, which tested for the equality of the variance between two groups (see Methods). Table 1 shows that five phenotypes (Triglycerides, HDL, LDL, BMI, and Diastolic blood pressure) have significant differences in phenotypic variance between males and females ($P < 0.005$, Bonferroni correction on 10 tests). The most significant difference was observed in Triglycerides ($P < 10^{-20}$).

Variance component analysis

To investigate why the phenotypic variances of some traits differed between the sexes, we performed a variance component analysis. We used the five-variance-component linear mixed model described in Figure 1A, where we excluded the SNP terms. Decomposing the variance components could reveal if the phenotypic variance difference came from the differences in polygenic background effects, or the differences in the error variances. We used the GCTA method to perform this analysis (Yang *et al.*, 2010). First, we generated the genetic relationship matrix, $\mathbf{K}$, implemented in GCTA framework. Second, we created two modified GRM matrices, $\mathbf{K_f}$ and $\mathbf{K_m}$, from $\mathbf{K}$ by having multiplications of $\mathbf{h}\mathbf{h}^T \circ \mathbf{K}$ and $(\mathbf{1}-\mathbf{h})(\mathbf{1}-\mathbf{h})^T \circ \mathbf{K}$, respectively. Unfortunately, GCTA did not allow us to separate the error term into two sex-specific terms as in the model in Figure 1A, because the default error term (with variance σ_e^2) for the whole sample is automatically included in the model. Thus, we added the sex-specific error term (with variance $\sigma_{e,ss}^2$) to one sex. We tried both males and females for this additional term, and chose the configuration with $\sigma_{e,ss}^2 > 0$.

Table 2 and Table S1 show the variance component estimates. The polygenic background effect (σ_g^2) was significantly non-zero in many traits ($P < 0.005$ for 5 traits; BMI, HDL, Height, LDL, and Systolic blood pressure). The sex-interacting polygenic background effects ($\sigma_{g,m}^2$ and $\sigma_{f,m}^2$) were non-zero in some traits, but none of them showed significance ($P > 0.005$) due to large standard errors. The variance of sex-specific error term ($\sigma_{e,ss}^2$) was significantly non-zero in Tryglyceride ($P = 6.23 \times 10^{-5}$) and BMI ($P = 1.2 \times 10^{-4}$). Thus, in some phenotypes, the phenotypic variance difference between the sexes was not completely explained by the genetic components alone, which suggested the need for explicitly modeling the sex difference in error variances as in our MetaSex method.

Full model versus approximated model

Although it is feasible to estimate five variance components one time using tools such GCTA, applying the full linear mixed model to millions of markers can be prohibitively slow because the variance component estimation needs to be repeated for each SNP. Therefore, we proposed an approximated model that decomposes the problem into two sex-specific linear mixed models (Figure 1B). Here, using the NFBC dataset, we examined if the variance components estimated by the approximated model were similar to those estimated by the full model.

To achieve this goal, we performed GCTA analyses in each sex-specific two-variance-component model described in Figure 1B. Table 3 shows the estimated variance components for the two sexes. The variance components can be categorized into two groups: the genetic variance and the random error variance. The genetic variaces in the full and sex-specific models are $\sigma_g^2 + \sigma_{g,f(m)}^2$ from Table 2 and $\sigma_{g,f(m)}^2$ from Table 3, respectively. The random error variances in the full and sex-specific models are $\sigma_e^2 + \sigma_{e,ss}^2$ from Table 2 (or σ_e^2, depending on which sex $\sigma_{e,ss}^2$ was added) and $\sigma_{e,f(m)}^2$ from Table 3, respectively. We compared if the genetic and random error variances are the same between the full and sex-specific models. Figure 5 shows that the estimated variance components were highly concordant between the full and sex-specific models. Most of the points followed the $y = x$ line (dashed line) well. The Pearson correlations were high ($r^2 > 0.9$) in all comparisons, except for the male genetic variance. The low correlation in the male genetic variance plot was driven by one outlier (Diastolic blood pressure), which may reflect instability in optimization procedure. If we excluded this outlier, the correlation was high ($r^2 = 0.979$). Overall, the estimates of the full and sex-specific models were highly concordant as expected,

Association mapping

We mapped associations to the 10 phenotypes in the NFBC dataset using our MetaSex approach. We used the efficient approximated model; that is, in each sex, we applied a sex-specific linear mixed model to account for the polygenic effects and the sex-interacting polygenic effect simultaneously (Figure 1B). Then, we applied RE and SST to the resulting male and female effect size estimates. For comparison, we also applied CV and GWAMA. In both CV and GWAMA, we similarly accounted for polygenic background effects using variance components. Finally, for a fair comparison, we calculated the genomic control factor λ separately for each method and corrected the resulting p-values of each method using this factor. (Table S2).

The challenge in this real dataset analysis was the lack of objective measures to compare performances of the methods, because we do not know which loci are true positives. What we could do was to examine loci that were genome-wide significant and to compare the p-values of different methods. Under the assumption that the loci exceeding the significance threshold have high chance of being true positives, a putatively better method can be a method that gave smaller p-values at those loci.

In this analysis, we discovered 16 loci that were associated with at least one of the 10 phenotypes at the threshold level $P < 5 \times 10^{-8}$ by at least one method. At these 16 loci, we calculated the p-values using RE, CV, GWAMA, and SST (Table S3). To compare the p-values of the methods, we chose CV as a reference method to be compared with. We plotted the $-log_{10}P$ difference between differing methods and the CV approach in Figure 6A ([$-log_{10}P$ of RE/GWAMA/SST] − [$-log_{10}P$ of CV]). Thus for each SNP, the positively larger the difference is, the better the method is compared with CV. As shown in Figure 6A, RE showed the best overall performance. RE gave smaller p-values than GWAMA at 14 out of 16 loci and better p-values than CV at 11 out of 16 loci. Even at loci where GWAMA or CV showed smaller p-values than RE, the difference from RE was small. Specifically, RE p-values were never larger by one order of magnitude than any of these methods at all 16 loci.

We further investigated on what characteristics of the loci caused these p-value differences between the methods. Figure 6B shows the phenotype variance ratio (PVR) between males and females after regressing out the genetic effect of the SNP tested. Figure 6C shows the ratio of the effect size estimates between males and females for each SNP. Because the power of the methods, as we showed in power simulations, depend on both the error variance ratio (which will affect PVR) and the effect size estimate ratio between males and females, for each SNP, we can interpret the p-values of the methods (Figure 6A) in terms of the PVR (Figure 6B) and the effect size ratio (Figure 6C).

If we look at the SNP rs7298683 (indicated by †), the effect size ratio between male and female studies was -0.0313, which means that the effect direction was opposite for the two sexes and that the absolute magnitude of the male effect size was 33 times larger than that of the female effect size. However, there was almost no difference in PVR between male and female studies (PVR of 0.965). In this case, the SST approach gave the smallest p-value, because SST was the best method to detect extreme effect size difference as we showed in simulations.

If we look at the SNP rs2167079 (indicated by $*$), the PVR (female/male) was 1.261 and the effect size ratio (female/male) was 0.43. Variance in females was larger, so the female effect size estimate were more uncertain than the male estimate. Thus, when combining information from the two sexes, an optimal method should give more weight to male estimate. Moreover, effect size was greater in males. Thus, an optimal method should weight the male estimate even further. Because CV ignores both variance difference and effect size difference, RE showed smaller p-values at this locus than CV. A similar interpretation of the result can be applied to the SNPs rs7120118 and rs693 (indicated by $\bullet$).

Now consider the SNP rs11668477 (indicated by ‡). The PVR (female/male) was 0.82 and the effect size ratio (female/male) was 0.5. In this case, when combining information from the two sexes, based on the variance, we should weight the female study more, but based on the effect size, we should weight the male study more. Thus, the impact of differing variance and the impact of differing effect size canceled out, giving CV the smallest p-value of all approaches, because CV can be considered as equally weighting the two sexes. A similar interpretation of the result can be applied to the SNPs rs2794520, rs560887, and rs10096633 (indicated by $\diamond$). However, as described, even in such situations, RE was not much worse than CV.

In summary, RE showed the best stable performance of all methods, except when the effect only existed in one sex where SST performed the best. Therefore, this analysis demonstrates that our MetaSex framework where RE and SST complement each other can cover many possible situations with high power.

3 Discussion

We here presented MetaSex, a novel framework that accounts for the sex difference in genetic architecture for powerful association mapping. Our method built upon a comprehensive model that included multiple variance components and expedited the optimization by using an approximation based on sex-specific models. We utilized the meta-analysis framework to achieve high power in a wide range of situations. We have shown superior performances of our approach compared with previous approaches in both simulations and real data analysis.

The high power of our approach was attributable to two factors: the effect size difference between the sexes and the error variance difference between the sexes. Previous studies have observed effect size differences at a number of loci (Randall *et al.*, 2013). However, few studies have reported phenotypic variance differences between the sexes, which can reflect the error variance difference. In our study, we showed that the phenotypic variance difference can be a real phenomenon in the existing dataset. The non-genetic cause of the phenotypic variance difference can be sex acting as an environment (e.g. hormone) or sex interacting with external environments (e.g. lifestyle). We demonstrated that accounting for the non-genetic causes by modeling differing error variances can increase power.

Although we tried to account for possible phenomena that can occur due to the sex difference in genetic association mapping, our model might still have limitations. We explictly modeled the sex difference in the effect sizes of the associated locus, magnitudes of the polygenic effects, and error variances. However, we did not model the sex difference in the phenotype distribution (i.e. shape), genetic interaction with covariates, or the liability distribution of binary traits. We expect that a large-scale study will be necessary to fully decipher the sex-interacting genetic architectures of human traits.

4 Methods

4.1 MetaSex

The standard linear mixed model that accounts for polygenic background effects can be written as

$$\mathbf{y} = \mu\mathbf{1} + \beta\mathbf{x} + \mathbf{u} + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ is a phenotype vector, μ is an intercept, $\mathbf{1}$ is a vector of ones, $\mathbf{x}$ is a genotype vector, β is the genetic effect, $\mathbf{u} \sim N(0, \sigma_g^2\mathbf{K})$ is a variance component that accounts for polygenic effects, and $\mathbf{e} \sim N(0, \sigma_e^2\mathbf{I})$ is the random error term. Typically, $\mathbf{K}$ is defined as the genotype similarity matrix between individuals. Recent studies have developed numerical optimization strategies that allow an efficient application of this standard linear mixed model to GWAS(Kang *et al.*, 2010, 2008; Zhou and Stephens, 2012).

We expand this model to a model that comprehensively accounts for sex differences. We assume that each of the four terms of the standard model (intercept, genetic effect, polygenic effect, and error variance) can have differences between the sexes. We expand the model as follows:

$$\mathbf{y} = \mu\mathbf{1} + \mu_s\mathbf{h} + \beta\mathbf{x} + \beta_{g\times s}\mathbf{x}\circ\mathbf{h} + \mathbf{u_g} + \mathbf{u_f} + \mathbf{u_m} + \mathbf{e_f} + \mathbf{e_m} \tag{2}$$

where $\mathbf{h}$ is a sex-indicating vector with elements 1 if an individual is female and 0 if male and $\mathbf{x}\circ\mathbf{h}$ denotes an element-wise product. There are five variance components in this model: $\mathbf{u_g} \sim N(0, \sigma_g^2\mathbf{K})$, $\mathbf{u_f} \sim N(0, \sigma_{g,f}^2(\mathbf{K}\circ\mathbf{hh^T}))$, $\mathbf{u_m} \sim N(0, \sigma_{g,m}^2(\mathbf{K}\circ(\mathbf{1}-\mathbf{h})(\mathbf{1}-\mathbf{h})^{\mathbf{T}}))$, $\mathbf{e_f} \sim N(0, \sigma_{e,f}^2\mathbf{I}\circ\mathbf{hh^T})$, and $\mathbf{e_m} \sim N(0, \sigma_{e,m}^2\mathbf{I}\circ(\mathbf{1}-\mathbf{h})(\mathbf{1}-\mathbf{h})^{\mathbf{T}})$. $\mathbf{u_g}$ is the standard random component that accounts for polygenic effects where $\mathbf{K}$ is the genotype similarity matrix. $\mathbf{u_f}$ is an additional random component that accounts for sex-interacting polygenic effects that are female-specific. Similarly, $\mathbf{u_m}$ is a random component that accounts for sex-interacting polygenic effects that are male-specific. We use two random error terms, $\mathbf{e_f}$ and $\mathbf{e_m}$ for females and males respectively, to account for the difference in error variances between the sexes.

Because this comprehensive model involves five variance components, application of this model to GWAS can be computationally challenging. For this reason, we applied the following approximation and split the model into two sex-specific models:

$$\begin{aligned}
\mathbf{y_f} &= \mu_f \mathbf{1} + \beta_f \mathbf{x_f} + \boldsymbol{\upsilon}_f + \boldsymbol{\varepsilon}_f \\
\mathbf{y_m} &= \mu_m \mathbf{1} + \beta_m \mathbf{x_m} + \boldsymbol{\upsilon}_m + \boldsymbol{\varepsilon}_m
\end{aligned} \tag{3}$$

where $\mathbf{y_f}$ is the phenotype vector of female individuals, β_f is the effect size in females, $\mathbf{x_f}$ is the genotype vector of female individuals, $\boldsymbol{\upsilon}_f \sim N(0, \rho_{g,f}^2 \mathbf{K_f})$ is the polygenic effect within females, and $\boldsymbol{\epsilon}_f \sim N(0, \rho_{e,f}^2 \mathbf{I_f})$ is the female-specific error term. $\mathbf{K_f}$ is the genotype similarity matrix between female individuals and $\mathbf{I_f}$ is an identity matrix defined for the female sample size. We can similarly define terms for males. This approximated model has the following relationships to the previous full model:

$$\begin{aligned}
\mu_f &= \mu + \mu_s \\
\mu_m &= \mu \\
\beta_f &= \beta + \beta_{g\times s} \\
\beta_m &= \beta \\
\rho_{g,f}^2 &= \sigma_g^2 + \sigma_{g,f}^2 \\
\rho_{g,m}^2 &= \sigma_g^2 + \sigma_{g,m}^2 \\
\rho_{e,f}^2 &= \sigma_{e,f}^2 \\
\rho_{e,m}^2 &= \sigma_{e,m}^2
\end{aligned}$$

These equalities hold because the approximated model can be considered as the same comprehensive model where we only look at a subset of samples (one sex). Intuitively, since we separate each sex into two models, the intercept is no more tied to be the same between the sexes. This freedom accounts for the phenotypic mean difference between the sexes. This is the same for the genetic effect size (β) and the error variance. The polygenic effect term for each sex simultaneously accounts for both the polygenic and the sex-interacting polygenic effects in the original model because, for each sex, the covariance matrix of the two terms become identical.

The benefit of this approximated model is that each model contains only two variance components. Currently available methods are well optimized for this two-variance-component model. (Kang *et al.*, 2010, 2008; Zhou and Stephens, 2012) The difference in this approximated model compared with the original model is that, in the original model, σ_g^2, $\sigma_{g,f}^2$ and $\sigma_{g,m}^2$ are estimated allowing the distinction between the three. Whereas, in this approximated model, the estimates $\rho_{g,f}^2$ and $\rho_{g,m}^2$ do not allow distinction between the whole-sample polygenic component and the sex-interacting polygenic component. However, this distinction is not crucial in association mapping where we want to control for both effects. Another difference is that the cross-sex elements in $\mathbf{K}$ are not used, which would give more accurate estimates of variance components. However, if the sample size in each sex is sufficiently large, the variance component estimates can be accurate and the approximated model will be almost identical to the original model. We also note that the cryptic relatedness between the sexes are not accounted in this approximated model.

Sex-specific test

After obtaining the effect size estimate and its standard error from each sex-specific model (for example, $\hat{\beta}_f$ and se($\hat{\beta}_f$) for females), we first apply the sex-specific test (SST). The null hypothesis of sex-specific test is that the variant has no effect in both sexes. Thus, the null hypothesis is $H_0 : \beta_m = 0$ and $\beta_f = 0$. We perform SST by obtaining a p-value from female-specific test (SST(F)) and a p-value from male-specific test (SST(M)). Since we perform two independent tests in SST, we correct for multiple testing. The reason that we apply SST first is not only because SST is optimal for detecting sex-specific effects, but also because in practice, investigators typically look at each sex separately in their data. By explicitly including this test in our framework, we can account for the additional multiple testing burden induced by this test.

Whole sample test using meta-analysis

Then, we perform whole sample test by combining information from both sexes. Our goal is to find a locus that has either common effect (effect that exists for both sexes with the same effect size), or interaction effect (effect that exists for both sexes with differing effect sizes). In the comprehensive model, our null hypothesis is $H_0 : \beta = 0$ and $\beta_{g\times s} = 0$. In our approximated model, this null hypothesis translates to an equivalent null hypothesis, $H_0 : \beta_m = 0$ and $\beta_f = 0$. What would be an optimal approach for simultaneously testing β_m and β_f will depend on the alternative models. If β_m and β_f are expected to be completely different (e.g. opposite directions of effects), simply adding chi-square statistics as is done in GWAMA method (Magi *et al.*, 2010) would be powerful. However, in practice, it is rare to observe opposite effect directions of one genetic variant for the two sexes. More common situations would be that the effects are in the same direction but in different magnitudes. Nevertheless, if the magnitudes of effects are extremely different such that one effect is relatively very close to zero, then the variant is likely to be already found by SST. Thus, we can specifically target effect size pairs whose directions are the same and whose magnitudes can be different, but none is very close to zero. To this end, we chose to use the random effects (RE) model meta-analysis method which assumes that the male and female effect sizes are random variables drawn from the same underlying distribution.

Specifically, in an RE model, we assume that the effect size of male and female studies, β_i ($i = m, f$), follow a distribution with the grand mean $\bar{\beta}$ and the variance τ^2 (DerSimonian and Laird, 1986; Han and Eskin, 2011):

$$\beta_i \sim N(\bar{\beta}, \tau^2).$$

The recently proposed RE model by Han and Eskin (Han and Eskin, 2011) tests the null hypothesis $H_0 : \bar{\beta} = 0$ and $\tau^2 = 0$ versus the alternative hypothesis $H_1 : \bar{\beta} \neq 0$ or $\tau^2 \neq 0$. Note that when we apply RE to our framework, this null hypothesis exactly corresponds to $H_0 : \beta_m = 0$ and $\beta_f = 0$. Let β_i and V_i be the effect size estimate and its variance respectively. As in the traditional random effect model (DerSimonian and Laird, 1986), Han-Eskin model uses the likelihood ratio framework considering each statistic (β_i and V_i pair) as a single observation. The difference from the traditional RE model is that the Han-Eskin model assumes no heterogeneity under the null hypothesis (Han and Eskin, 2011). This assumption is valid if the causes of heterogeneity do not exist under the null hypothesis, which is likely to be the case for GWAS. The likelihoods are then:

$$
\begin{aligned}
L_0 &= \prod_{i=m,f} \frac{1}{\sqrt{2\pi V_i}} \exp\left(-\frac{\beta_i^2}{2V_i}\right) \\
L_1 &= \prod_{i=m,f} \frac{1}{\sqrt{2\pi(V_i+\tau^2)}} \exp\left(-\frac{(\beta_i-\bar{\beta})^2}{2(V_i+\tau^2)}\right).
\end{aligned}
$$

The maximum likelihood estimates $\hat{\hat{\beta}}$ and $\hat{\tau}^2$ can be found by an iterative procedure suggested by Hardy and Thompson (Hardy and Thompson, 1996). Then the likelihood ratio test statistic can be built

$$
S_{\text{meta}} = -2\log(\lambda) = \sum \log\left(\frac{V_i}{V_i+\hat{\tau}^2}\right) + \sum \frac{\beta_i^2}{V_i} - \sum \frac{(\beta_i-\hat{\hat{\beta}})^2}{V_i+\hat{\tau}^2}, \tag{4}
$$

which asymptotically follows a half and half mixture of $\chi^2_{(1)}$ and $\chi^2_{(2)}$. A p-value after small sample adjustment can be efficiently calculated using pre-computed tabulated values (Han and Eskin, 2011).

Smart thresholding

In our MetaSex framework, we perform three tests: SST, which consists of SST(F) and SST(M), and whole sample test using RE. To account for multiple testing, we can use the Bonferroni correction, but that can be overly conservative because of the dependency between the test statistics of SST and RE. Instead, we can perform null simulations to empirically determine the significance threshold. Moreover, we can use a strategy similar to one published previously (Eskin, 2008), which uses different levels of significance thresholds for multiple tests to achieve higher power while controlling the overall false positive rate (family-wise error rate) to a fixed level.

To find an optimal threshold pair for RE and SST while controlling the false positive rate at 5.0×10^{-8}, we generated 10 billion null statistic pairs for the male studies and the female studies. Any pair of thresholds for RE and SST that rejected 500 null statistics would control the false positive rate at 5×10^{-8}. Then, we adjusted the thresholds for RE and SST while keeping the total number of rejections to 500. For example, a threshold pair can have one false positive for RE and 499 false positives for SST. Next, one can have two false positives for RE and 498 false positives for SST. There were 500 such threshold pairs that control the false positive rate of 5×10^{-8}. Among all 500 pairs of thresholds that gave the same false positive rate, we chose the threshold pair that gave us the maximum power. To calculate power, we needed a model assumption for the alternative hypothesis. We assumed the model in Figure 4, which uniformly sampled the female effect size and male effect size from a range between 0 and 1. Although this alternative model was just one possible model, we expect that it will cover a wide range of possible situations. Under this uniform prior assumption, we calculated the power of each pair. We found that using unequal thresholds, 2.41×10^{-8} for RE and 1.36×10^{-8} for each of SST(F) and SST(M), gave us the best power while still controlling the false positive rate. Note that although our pair of thresholds was optimized for a specific alternative model, even if the true alternative model would be different from the assumed model, our false positive rate can be still controlled; only the power will be affected. The users using our method can just use these pre-computed thresholds.

4.2 Existing approaches

CV

The standard approach for dealing with sex is to use sex as a covariate. We refer to this model as CV in short. The CV model is:

$$\mathbf{y} = \mu\mathbf{1} + \mu_s\mathbf{h} + \beta\mathbf{x} + \mathbf{u} + \mathbf{e}$$

which is equivalent to the traditional model in equation (1) with the only difference being the inclusion of the covariate denoting sex (μ_s). CV accounts for the phenotypic mean differences between the sexes. However, CV does not account for the followings: possible sex differences in the effect size (β), polygenic background effect ($\mathbf{u}$), and the error variance ($\mathrm{Var}(\mathbf{e})$).

GWAMA

GWAMA (Genome-Wide Association Meta-Analysis) is another meta-analytic approach proposed by Magi *et al.* (Magi *et al.*, 2010). In GWAMA, as in MetaSex, each sex is analyzed separately. Then, the χ^2 statistics of males and females are calculated by squaring the corresponding z-scores, that is,

$$\chi^2_m = z^2_m = (\frac{\beta_m}{\sqrt{V_m}})^2 \text{ and } \chi^2_f = z^2_f = (\frac{\beta_f}{\sqrt{V_f}})^2$$

The GWAMA statistic can be obtained by summing male χ^2 and female χ^2

$$S_{GWAMA} = \chi^2_m + \chi^2_f$$

The p-value can be obtained using χ^2 distribution with two degrees of freedom.

Because GWAMA is a meta-analytic approach that analyzes each sex separately and combines them, it shares some of the advantages with our MetaSex approach. That is, GWAMA framework can also account for between-sex differences in intercept and error variances.

4.3 Power calculation

To evaluate the power of methods, we performed simulations as follows. We assumed a specific effect size. Then based on an assumed standard error, we sampled an observed estimate of effect size. We performed this sampling for males and females separatly by N times. Given N male estimates and N female estimates, we can apply any of the tested methods. The statistical power was computed as the proportion of p-values that were more significant than a significance threshold. As described, we found the correct significance threshold by performing null simulations under the null hypothesis of no effects; Empirical null simulation was necessary because some methods involved multiple testing. For example, SST consists of two tests (SST(F) and SST(M)), and MetaSex (RE+SST) consists of three tests (SST(F), SST(M), and RE). As with MetaSex, each of GWAMA+SST and CV+SST consists of three tests. We used N of at least 10,000 in all of our simulations.

For some methods, specifically SST and GWAMA, it was also possible to calculate the exact power analytically. Let α be the desired significance threshold level. Let $\Phi(x;\mu,\sigma)$ be the cumulative density function (CDF) of a normal distribution evaluated at x given the mean μ and standard deviation σ of the distribution. Because SST consists of two independent tests (SST(F) and SST(M)), we should use the significance threshold α' that satisfies $\alpha = 1-(1-\alpha')^2$ for each of SST(F) and SST(M). Assuming a two-sided test, the z-score threshold corresponding to α' can be obtained by the inverse CDF of $\frac{\alpha'}{2}$ ($T_{\alpha'} = \Phi^{-1}(\frac{\alpha'}{2};0,1)$). Given the true effect sizes (β_m, β_f) and the standard errors of the estimates ($\sqrt{V_m}$, $\sqrt{V_f}$) that we assumed for males and females, let $\bar{z}_m = \frac{\beta_m}{\sqrt{V_m}}$ and $\bar{z}_f = \frac{\beta_f}{\sqrt{V_f}}$ be the expected z-scores of males and females. Then, the statistical power of the SST approach is

$$Power_{SST} = 1 - \left[\Big(1-\Phi(T_{\alpha'};\bar{z}_m,1) - (1-\Phi(-T_{\alpha'};\bar{z}_m,1))\Big) \times \Big(1-\Phi(T_{\alpha'};\bar{z}_f,1) - (1-\Phi(-T_{\alpha'};\bar{z}_f,1))\Big)\right]$$

For GWAMA, we can first define $F_{\chi^2}(x; m, k)$ as the CDF of a χ^2 distribution evaluated at x given the non-centrality parameter m and k degrees of freedom. Because the GWAMA statistic follows the central χ^2 distribution with two degrees of freedom under the null hypothesis, given the desired significance threshold α, the threshold of GWAMA statistic that corresponds to α can be expressed as $T_\alpha = F^{-1}_{\chi^2}(1 - \alpha; 0, 2)$. The non-centrality parameter for the GWAMA statistic is $\bar{z}_m^2 + \bar{z}_f^2$. Then the statistical power of the GWAMA approach can be computed analytically with the following formula:

$$Power_{GWAMA} = 1 - F_{\chi^2}(T_\alpha; \bar{z}_m^2 + \bar{z}_f^2, 2) \tag{5}$$

We observed that, as expected, the analytically calculated power was nearly identical to the empirical estimate.

4.4 Levene's test

Levene's test determines if there is a significant difference among the variances of multiple groups (Brown and Forsythe, 1974). The statistic is

$$W = \frac{(N-K)}{(k-1)} \frac{\sum_{i=1}^{k} N_i (Z_{i.} - Z_{..})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{i.})^2} \tag{6}$$

where k is the number of different groups to which the samples belong ($k = 2$ for our between-sex test), N is the total number of samples in all groups, and N_i is the number of sample in the ith group. Let Y_{ij} be the value of the measured variable for the jth sample from the ith group. We define $Z_{ij} = |Y_{ij} - \bar{Y_i}|$ ($\bar{Y_i}$ is a mean of ith group), $Z_{..} = \dfrac{1}{N} \displaystyle\sum_{i=1}^{k} \sum_{j=1}^{N_i} Z_{ij}$, and $Z_{i.} = \dfrac{1}{N} \displaystyle\sum_{j=1}^{N_i} Z_{ij}$. The resulting statistic W follows an F distribution with $k - 1$ and $N - 1$ degrees of freedom under the null hypothesis.

4.5 NFBC data

In our current study, we used the previously reported NFBC data (Sabatti *et al.*, 2009), which contained 5326 individuals (2546 males and 2780 females). To investigate the sex difference in genetic architectures of human traits, we examined 10 phenotypes: Triglycerides, High-density lipoprotein, Low-density lipoprotein, C-reactive protein, Glucose, Insulin, Body mass index, Systolic and Diastolic blood pressure, and Height. Detailed trait measurements and sample genotype collection have previously been described (Sabatti *et al.*, 2009).

References

Boraska, V., A. Jerončić, V. Colonna, L. Southam, D. R. Nyholt, *et al.*, 2012 Genome-wide meta-analysis of common variant differences between men and women. Hum Mol Genet **21**: 4805–15.

Brown, M. B., and A. B. Forsythe, 1974 Robust tests for the equality of variances **69**: 364–367.

Chen, Y. C., G. H. Dong, K. C. Lin, and Y. L. Lee, 2013 Gender difference of childhood overweight and obesity in predicting the risk of incident asthma: a systematic review and meta-analysis. Obes Rev **14**: 222–31.

DerSimonian, R., and N. Laird, 1986 Meta-analysis in clinical trials. Control Clin Trials **7**: 177–88.

Eskin, E., 2008 Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. Genome Res **18**: 653–60.

Fox, C. S., Y. Liu, C. C. White, M. Feitosa, A. V. Smith, *et al.*, 2012 Genome-wide association for abdominal subcutaneous and visceral adipose reveals a novel locus for visceral fat in women. PLoS Genet **8**: e1002695.

Han, B., and E. Eskin, 2011 Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. Am J Hum Genet **88**: 586–98.

Hardy, R. J., and S. G. Thompson, 1996 A likelihood approach to meta-analysis with random effects. Stat Med **15**: 619–29.

Kang, E. Y., B. Han, N. Furlotte, J. W. J. Joo, D. Shih, *et al.*, 2014 Meta-analysis identifies gene-by-environment interactions as demonstrated in a study of 4,965 mice. PLoS Genet **10**: e1004022.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S.-Y. Y. Kong, *et al.*, 2010 Variance component model to account for sample structure in genome-wide association studies. Nat Genet **42**: 348–54.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, *et al.*, 2008 Efficient control of population structure in model organism association mapping. Genetics **178**: 1709–1723.

Kostis, W. J., J. Q. Cheng, J. M. Dobrzynski, J. Cabrera, and J. B. Kostis, 2012 Meta-analysis of statin effects in women versus men. J Am Coll Cardiol **59**: 572–82.

Kubo, A., M. B. Cook, N. J. Shaheen, T. L. Vaughan, D. C. Whiteman, *et al.*, 2013 Sex-specific associations between body mass index, waist circumference and the risk of barrett's oesophagus: a pooled analysis from the international beacon consortium. Gut .

Lippert, C., J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, *et al.*, 2011 Fast linear mixed models for genome-wide association studies. Nature Methods **8**: 833.

Magi, R., C. M. Lindgren, and A. P. Morris, 2010 Meta-analysis of sex-specific genome-wide association studies. Genet Epidemiol **34**: 846–53.

Mason, B. J., and P. Lehert, 2012 Acamprosate for alcohol dependence: a sex-specific meta-analysis based on individual patient data. Alcohol Clin Exp Res **36**: 497–508.

Ohmen, J., E. Y. Kang, X. Li, J. W. Joo, F. Hormozdiari, *et al.*, 2014 Genome-wide association study for age-related hearing loss (ahl) in the mouse: A meta-analysis. J Assoc Res Otolaryngol .

Peters, S. A. E., R. R. Huxley, and M. Woodward, 2013 Comparison of the sex-specific associations between systolic blood pressure and the risk of cardiovascular disease: A systematic review and meta-analysis of 124 cohort studies, including 1.2 million individuals. Stroke .

Porcu, E., M. Medici, G. Pistis, C. B. Volpato, S. G. Wilson, *et al.*, 2013 A meta-analysis of thyroid-related traits reveals novel loci and gender-specific differences in the regulation of thyroid function. PLoS Genet **9**: e1003266.

Randall, J. C., T. W. Winkler, Z. Kutalik, S. I. Berndt, A. U. Jackson, *et al.*, 2013 Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. PLoS Genet **9**: e1003500.

Sabatti, C., S. K. Service, A.-L. L. Hartikainen, A. Pouta, S. Ripatti, *et al.*, 2009 Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat Genet **41**: 35–46.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, *et al.*, 2010 Common snps explain a large proportion of the heritability for human height. Nat Genet **42**: 565–9.

Zhou, X., and M. Stephens, 2012 Genome-wide efficient mixed-model analysis for association studies. Nat Genet **44**: 821–4.

5 Tables

Phenotype	Var(female)	Var(male)	Ratio (larger/smaller)	Levene's test p-value
Triglyceride	0.171	0.256	1.494	1.451e-21
HDL	0.134	0.107	1.251	2.546e-10
LDL	0.670	0.820	1.223	1.394e-05
BMI	0.0309	0.0189	1.635	6.142e-19
C-reactive protein	2.372	2.244	1.056	0.0877
Glucose	0.00647	0.00678	1.048	0.174
Insulin	0.1105	0.1172	1.061	0.1173
Systolic blood pressure	156.77	171.357	1.092	0.0079
Diastolic blood pressure	118.56	136.053	1.147	0.0012
Height	38.65	41.098	1.063	0.0154

Table 1. Phenotypic variances of 10 NFBC phenotypes in the two sexes.

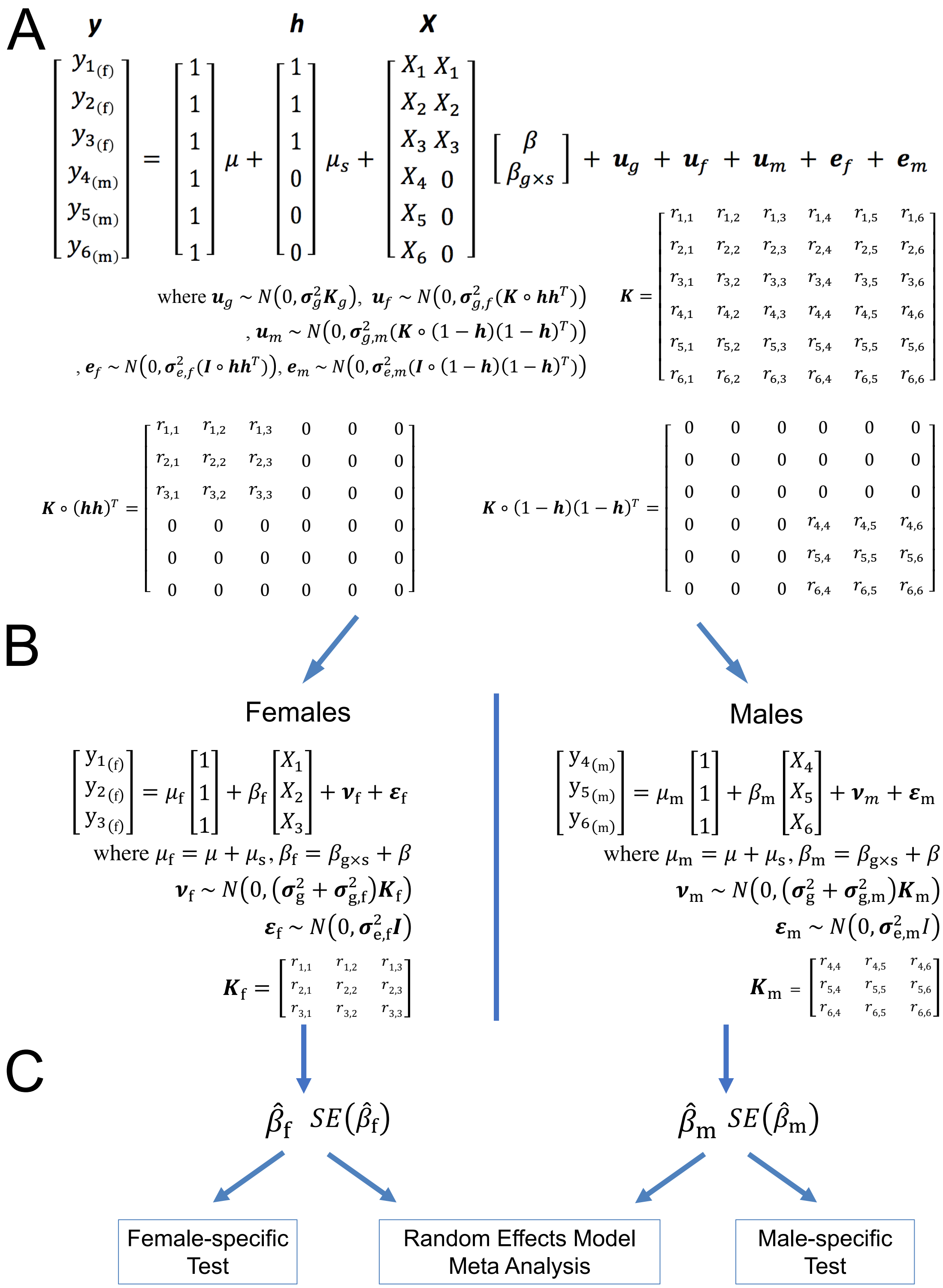
Phenotype	$\sigma_g^2(SE)$	$\sigma_{g,f}^2(SE)$	$\sigma_{g,m}^2(SE)$	$\sigma_{e,ss}^2(SE)$	$\sigma_e^2(SE)$
Triglyceride	0.0189(0.014)	0.0323(0.0219)	0.0031(0.0236)	0.1134(0.0283)	0.1205(0.0189)
HDL	0.0375(0.0072)	0(0.011)	0(0.0117)	0.0017(0.0137)	0.0694(0.0103)
LDL	0.2569(0.0511)	0.0237(0.0786)	0.042(0.082)	0.1186(0.0983)	0.3921(0.067)
BMI	0.0052(0.0014)	0(0.0022)	0(0.0023)	0.0107(0.0028)	0.0118(0.002)
C-reactive protein	0.222(0.1553)	0.1346(0.243)	0.0301(0.2641)	0.0227(0.3108)	1.9925(0.2321)
Glucose	0.0011(0.0005)	0.0002(0.0008)	0.0014(0.0008)	0.0009(0.0009)	0.0043(0.0007)
Insulin	0.0121(0.0082)	0.0139(0.0132)	0.004(0.0142)	0.0163(0.0167)	0.0847(0.0113)
Systolic blood pressure	37.1421(10.5002)	4.5774(16.7722)	1.9977(17.5993)	18.3323(21.4657)	114.189(14.6194)
Diastolic blood pressure	22.0399(8.3115)	0.0001(12.9059)	0.0001(13.6516)	11.6098(16.1558)	97.5402(11.0647)
Height	24.0673(2.6501)	1.2284(3.8886)	1.1577(4.2184)	3.2263(4.9783)	12.6085(3.326)

Table 2. Variance components of 10 NFBC phenotypes in the full five-variance-component model.

	Female		Male	
Phenotype	$\sigma_{g,f}^2(SE)$	$\sigma_{e,f}^2(SE)$	$\sigma_{g,m}^2(SE)$	$\sigma_{e,m}^2(SE)$
Triglyceride	0.0073(0.0159)	0.1642(0.0165)	0.0565(0.0267)	0.1997(0.0269)
HDL	0.0315(0.0122)	0.1032(0.0123)	0.0546(0.0119)	0.0531(0.0115)
LDL	0.2459(0.0631)	0.4228(0.062)	0.4762(0.0885)	0.3405(0.0846)
BMI	0.0067(0.003)	0.0243(0.0031)	0.0037(0.002)	0.0152(0.002)
C-reactive protein	0.2633(0.2215)	2.1086(0.227)	0.1493(0.221)	2.0952(0.2281)
Glucose	0.0021(0.0007)	0.0043(0.0007)	0.002(0.0007)	0.0048(0.0007)
Insulin	0.0176(0.0114)	0.0929(0.0116)	0.0261(0.013)	0.0911(0.0131)
Systolic blood pressure	50.6003(14.4485)	105.9015(14.3043)	24.6972(14.9829)	146.4733(15.3913)
Diastolic blood pressure	36.1682(11.0625)	82.26(10.9817)	0.0001(12.4726)	136.0544(13.0823)
Height	29.2825(3.5946)	8.8902(3.3024)	26.1134(4.3118)	14.7954(4.0722)

Table 3. Variance components of 10 NFBC phenotypes in the sex-specific models.

6 Figure Legends



$$\overset{\boldsymbol{y}}{\begin{bmatrix} y_{1(\mathrm{f})}\\ y_{2(\mathrm{f})}\\ y_{3(\mathrm{f})}\\ y_{4(\mathrm{m})}\\ y_{5(\mathrm{m})}\\ y_{6(\mathrm{m})}\end{bmatrix}} = \begin{bmatrix}1\\1\\1\\1\\1\\1\end{bmatrix}\mu + \overset{\boldsymbol{h}}{\begin{bmatrix}1\\1\\1\\0\\0\\0\end{bmatrix}}\mu_s + \overset{\boldsymbol{X}}{\begin{bmatrix}X_1 & X_1\\ X_2 & X_2\\ X_3 & X_3\\ X_4 & 0\\ X_5 & 0\\ X_6 & 0\end{bmatrix}}\begin{bmatrix}\beta\\ \beta_{g\times s}\end{bmatrix} + \boldsymbol{u}_g + \boldsymbol{u}_f + \boldsymbol{u}_m + \boldsymbol{e}_f + \boldsymbol{e}_m$$

$$\text{where } \boldsymbol{u}_g \sim N\left(0, \boldsymbol{\sigma}_g^2\boldsymbol{K}_g\right),\ \boldsymbol{u}_f \sim N\left(0, \boldsymbol{\sigma}_{g,f}^2(\boldsymbol{K}\circ\boldsymbol{hh}^T)\right)$$

$$,\ \boldsymbol{u}_m \sim N\left(0, \boldsymbol{\sigma}_{g,m}^2(\boldsymbol{K}\circ(1-\boldsymbol{h})(1-\boldsymbol{h})^T)\right)$$

$$,\ \boldsymbol{e}_f \sim N\left(0, \boldsymbol{\sigma}_{e,f}^2(\boldsymbol{I}\circ\boldsymbol{hh}^T)\right),\ \boldsymbol{e}_m \sim N\left(0, \boldsymbol{\sigma}_{e,m}^2(\boldsymbol{I}\circ(1-\boldsymbol{h})(1-\boldsymbol{h})^T)\right)$$

$$\boldsymbol{K} = \begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3} & r_{1,4} & r_{1,5} & r_{1,6}\\ r_{2,1} & r_{2,2} & r_{2,3} & r_{2,4} & r_{2,5} & r_{2,6}\\ r_{3,1} & r_{3,2} & r_{3,3} & r_{3,4} & r_{3,5} & r_{3,6}\\ r_{4,1} & r_{4,2} & r_{4,3} & r_{4,4} & r_{4,5} & r_{4,6}\\ r_{5,1} & r_{5,2} & r_{5,3} & r_{5,4} & r_{5,5} & r_{5,6}\\ r_{6,1} & r_{6,2} & r_{6,3} & r_{6,4} & r_{6,5} & r_{6,6}\end{bmatrix}$$

$$\boldsymbol{K}\circ(\boldsymbol{hh})^T = \begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3} & 0 & 0 & 0\\ r_{2,1} & r_{2,2} & r_{2,3} & 0 & 0 & 0\\ r_{3,1} & r_{3,2} & r_{3,3} & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\end{bmatrix} \qquad \boldsymbol{K}\circ(1-\boldsymbol{h})(1-\boldsymbol{h})^T = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & r_{4,4} & r_{4,5} & r_{4,6}\\ 0 & 0 & 0 & r_{5,4} & r_{5,5} & r_{5,6}\\ 0 & 0 & 0 & r_{6,4} & r_{6,5} & r_{6,6}\end{bmatrix}$$

$$\begin{bmatrix} y_{1(\mathrm{f})}\\ y_{2(\mathrm{f})}\\ y_{3(\mathrm{f})}\end{bmatrix} = \mu_\mathrm{f}\begin{bmatrix}1\\1\\1\end{bmatrix} + \beta_\mathrm{f}\begin{bmatrix}X_1\\X_2\\X_3\end{bmatrix} + \boldsymbol{\nu}_\mathrm{f} + \boldsymbol{\varepsilon}_\mathrm{f}$$

$$\text{where } \mu_\mathrm{f} = \mu + \mu_\mathrm{s}, \beta_\mathrm{f} = \beta_{\mathrm{g}\times\mathrm{s}} + \beta$$

$$\boldsymbol{\nu}_\mathrm{f} \sim N\left(0, \left(\boldsymbol{\sigma}_\mathrm{g}^2 + \boldsymbol{\sigma}_\mathrm{g,f}^2\right)\boldsymbol{K}_\mathrm{f}\right)$$

$$\boldsymbol{\varepsilon}_\mathrm{f} \sim N\left(0, \boldsymbol{\sigma}_\mathrm{e,f}^2\boldsymbol{I}\right)$$

$$\boldsymbol{K}_\mathrm{f} = \begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3}\\ r_{2,1} & r_{2,2} & r_{2,3}\\ r_{3,1} & r_{3,2} & r_{3,3}\end{bmatrix}$$

$$\begin{bmatrix} y_{4(\mathrm{m})}\\ y_{5(\mathrm{m})}\\ y_{6(\mathrm{m})}\end{bmatrix} = \mu_\mathrm{m}\begin{bmatrix}1\\1\\1\end{bmatrix} + \beta_\mathrm{m}\begin{bmatrix}X_4\\X_5\\X_6\end{bmatrix} + \boldsymbol{\nu}_m + \boldsymbol{\varepsilon}_\mathrm{m}$$

$$\text{where } \mu_\mathrm{m} = \mu + \mu_\mathrm{s}, \beta_\mathrm{m} = \beta_{\mathrm{g}\times\mathrm{s}} + \beta$$

$$\boldsymbol{\nu}_\mathrm{m} \sim N\left(0, \left(\boldsymbol{\sigma}_\mathrm{g}^2 + \boldsymbol{\sigma}_\mathrm{g,m}^2\right)\boldsymbol{K}_\mathrm{m}\right)$$

$$\boldsymbol{\varepsilon}_\mathrm{m} \sim N\left(0, \boldsymbol{\sigma}_\mathrm{e,m}^2 I\right)$$

$$\boldsymbol{K}_\mathrm{m} = \begin{bmatrix} r_{4,4} & r_{4,5} & r_{4,6}\\ r_{5,4} & r_{5,5} & r_{5,6}\\ r_{6,4} & r_{6,5} & r_{6,6}\end{bmatrix}$$

$$\hat{\beta}_\mathrm{f} \quad SE\left(\hat{\beta}_\mathrm{f}\right) \qquad\qquad \hat{\beta}_\mathrm{m} \quad SE\left(\hat{\beta}_\mathrm{m}\right)$$

Figure 1. Overview of the MetaSex method.

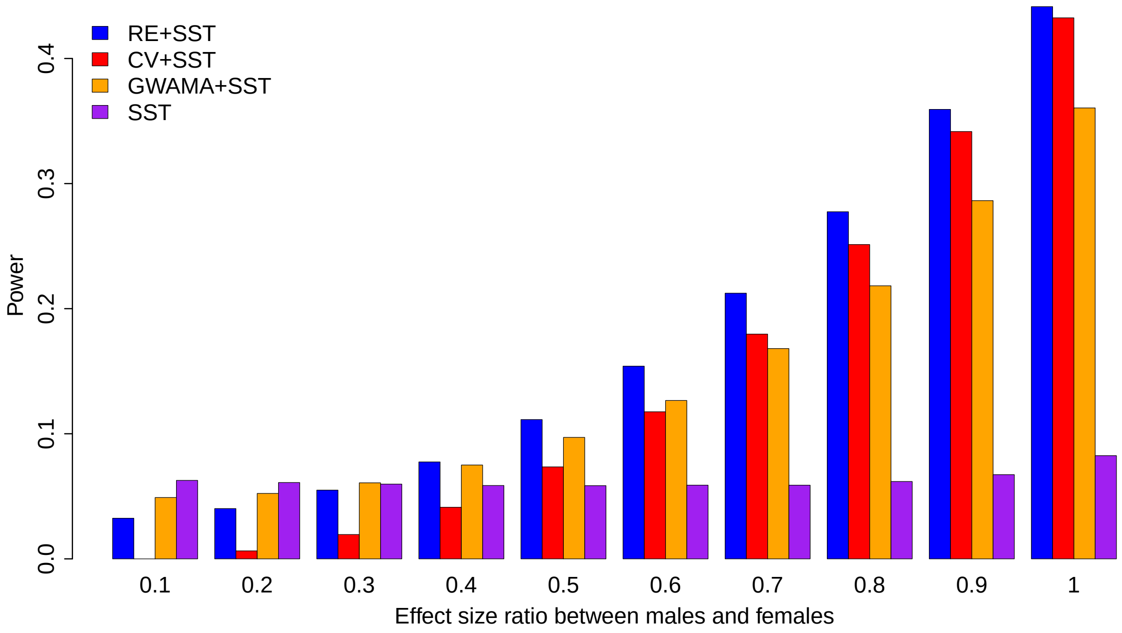


Figure 2. Power comparison between MetaSex (RE+SST), CV+SST, GWAMA+SST, and SST approaches while varying the effect size ratios of females and males. All methods were appropriately corrected for multiple testing.

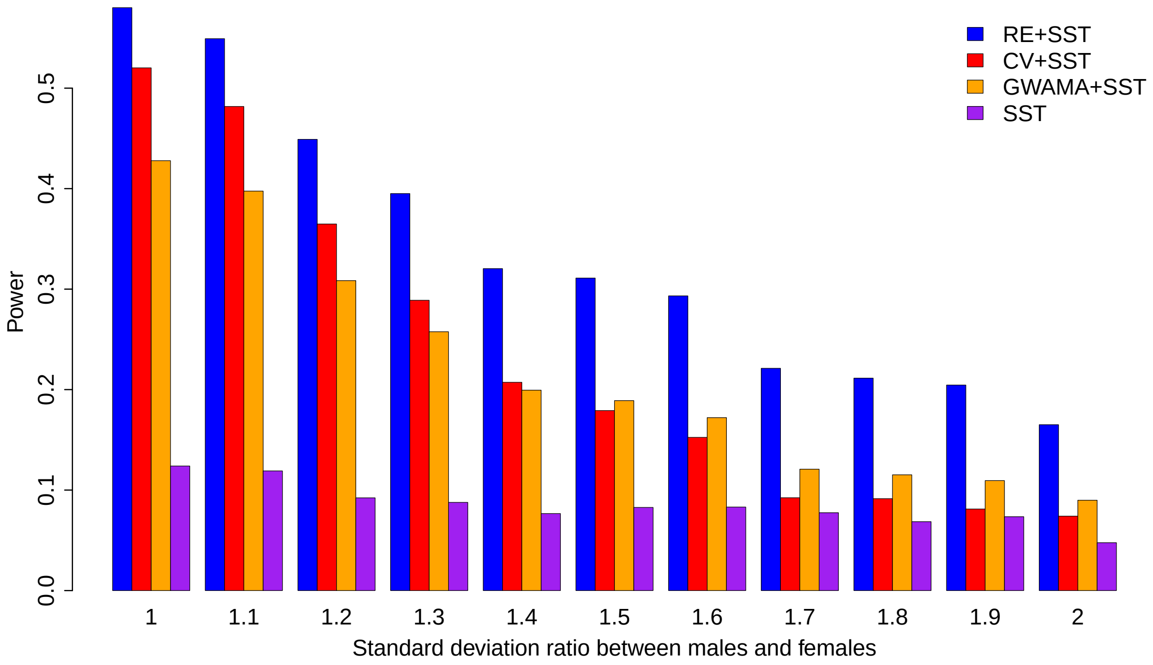


Figure 3. Power comparison between MetaSex (RE+SST), CV+SST, GWAMA+SST, and SST approaches while varying the error variance ratio in females and males. All methods were appropriately corrected for multiple testing.

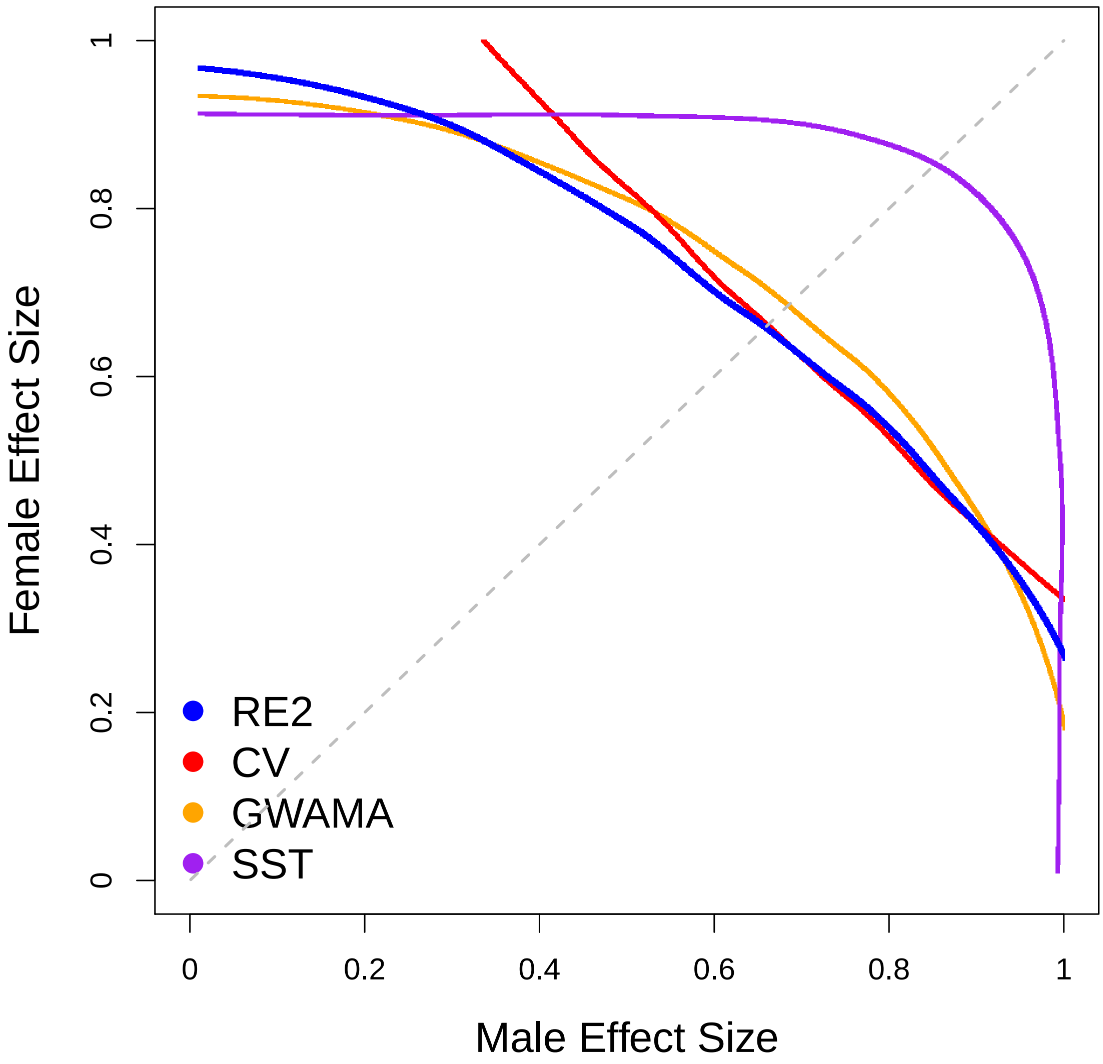


Figure 4. Power characteristics of RE, CV, GWAMA, and SST in a space varying the effect sizes of males and females. The lines denote the points where each method achieved 50% power. The diagonal line shows the points where the effect sizes of the two sexes are equal. The 50% power line is skewed due to the error variance ratio of 1.2 that we assumed.

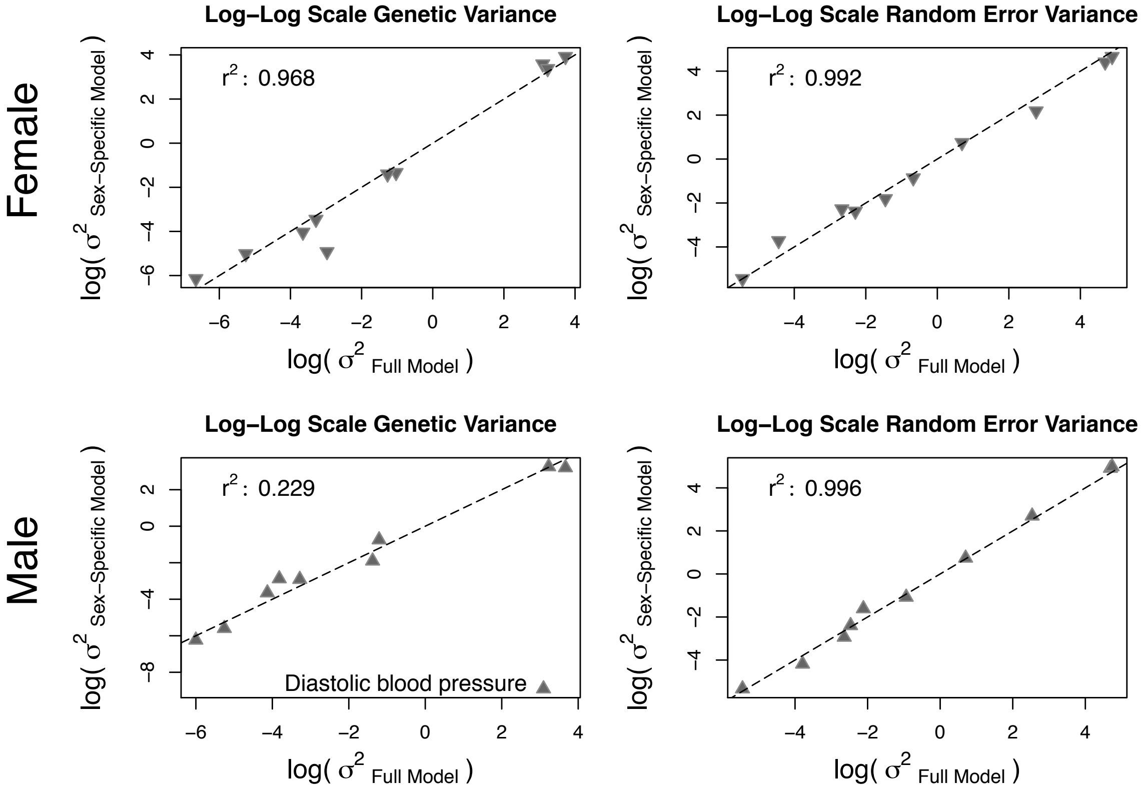


Figure 5. Comparison of the variance components between the full and the sex-specific models in 10 phenotypes of NFBC data. The points represent phenotypes and the dotted line is where the variance estimates of the two models are equal. We evaluated r^2 in the log scale and labeled an outlier(Diastolic blood pressure) observed in the male genetic variance plot.

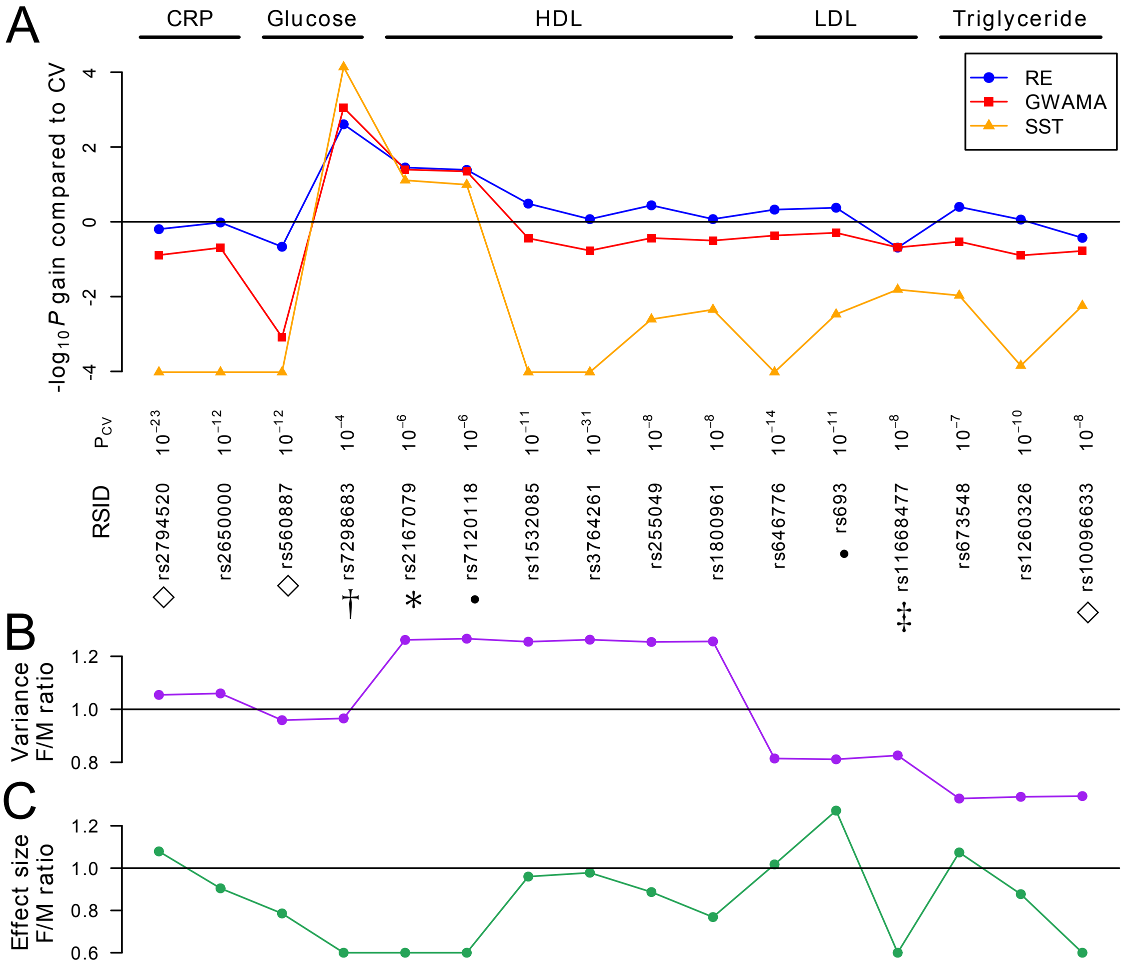


Figure 6. Association results of RE, CV, GWAMA, and SST at the 16 associated loci in any of 10 phenotypes of NFBC data. (A) Relative $-log_{10}P$ improvement of the other methods compared with the CV method, namely, [$-log_{10}P$ of RE/GWAMA/SST] $-$ [$-log_{10}P$ of CV]. The RSIDs of 16 significant SNPs as well as their CV p-values are shown at the bottom. (B) The ratio of phenotypic variance between males and females after regressing out the genetic effect of each SNP. (C) The ratio of the genetic effect size of each SNP between males and females.

7 Supplementary Tables

Phenotype	$\sigma_g^2(SE)$	$\sigma_{g,f}^2(SE)$	$\sigma_{g,m}^2(SE)$	$\sigma_{e,ss}^2(SE)$	$\sigma_e^2(SE)$	P_g	$P_{g,f}$	$P_{g,m}$	P_{ss}
Triglyceride	0.0189(0.014)	0.0323(0.0219)	0.0031(0.0236)	0.1134(0.0283)	0.1205(0.0189)	1.79^{-1}	1.4^{-1}	8.95^{-1}	$\mathbf{6.23^{-5}}$
HDL	0.0375(0.0072)	0(0.011)	0(0.0117)	0.0017(0.0137)	0.0694(0.0103)	$\mathbf{1.72^{-7}}$	1.00	1.00	8.99^{-1}
LDL	0.2569(0.0511)	0.0237(0.0786)	0.042(0.082)	0.1186(0.0983)	0.3921(0.067)	$\mathbf{4.88^{-7}}$	7.63^{-1}	6.08^{-1}	2.27^{-1}
BMI	0.0052(0.0014)	0(0.0022)	0(0.0023)	0.0107(0.0028)	0.0118(0.002)	$\mathbf{2.53^{-4}}$	1.00	1.00	$\mathbf{1.20^{-4}}$
C-reactive protein	0.222(0.1553)	0.1346(0.243)	0.0301(0.2641)	0.0227(0.3108)	1.9925(0.2321)	1.53^{-1}	5.80^{-1}	9.09^{-1}	9.42^{-1}
Glucose	0.0011(0.0005)	0.0002(0.0008)	0.0014(0.0008)	0.0009(0.0009)	0.0043(0.0007)	2.31^{-2}	7.83^{-1}	8.66^{-2}	3.56^{-1}
Insulin	0.0121(0.0082)	0.0139(0.0132)	0.004(0.0142)	0.0163(0.0167)	0.0847(0.0113)	1.40^{-1}	2.93^{-1}	7.79^{-1}	3.30^{-1}
Systolic blood pressure	37.1421(10.5002)	4.5774(16.7722)	1.9977(17.5993)	18.3323(21.4657)	114.189(14.6194)	$\mathbf{4.04^{-4}}$	7.85^{-1}	9.10^{-1}	3.93^{-1}
Diastolic blood pressure	22.0399(8.3115)	0.0001(12.9059)	0.0001(13.6516)	11.6098(16.1558)	97.5402(11.0647)	8.01^{-3}	1.00	1.00	4.72^{-1}
Height	24.0673(2.6501)	1.2284(3.8886)	1.1577(4.2184)	3.2263(4.9783)	12.6085(3.326)	$\mathbf{1.07^{-19}}$	7.52^{-1}	7.84^{-1}	5.17^{-1}

Table S1. Variance components of 10 NFBC phenotypes in the full five-variance-component model. For each trait, we tested if the variance of the polygenic background effect term(σ_g^2) is significantly non-zero (P_g) and if the variance of sex-specific error term($\sigma_{e,ss}^2$) is significantly non-zero (P_{ss}). The p-values less than the 0.005 are in bold font.

Phenotype	SST(F)	SST(M)	RE	CV	GWAMA
Triglyceride	1.007	0.991	0.995	1.001	1.006
HDL	1.002	0.995	1.008	1.003	0.993
LDL	1.001	0.996	1.011	1.002	1.004
BMI	0.999	0.996	0.992	0.994	0.994
C-reactive protein	0.994	0.989	0.998	0.993	0.994
Glucose	1.006	1.000	1.011	1.008	1.006
Insulin	1.004	0.996	1.006	1.005	1.001
Systolic blood pressure	0.997	0.997	1.012	1.006	1.000
Diastolic blood pressure	0.998	1.008	1.013	1.006	1.002
Height	1.004	1.002	1.028	1.003	1.001

Table S2. Genomic control factor in the genome-wide association mapping of 10 NFBC phenotypes.

Phenotype : C-reactive protein				**sex-specific**		**beta±std error**		**best**
rsid	**RE**	**CV**	**GWAMA**	**SST(F)**	**SST(M)**	**female**	**male**	**methods**
rs2794520	3.32e-23	**2.11e-23**	1.64e-22	1.28e-13	2.93e-11	0.314±0.042	0.291±0.043	RE,CV
rs2650000	**1.71e-12**	1.64e-12	8.09e-12	2.79e-06	8.65e-08	0.199±0.042	0.220±0.041	RE,CV

Phenotype : Glucose				**sex-specific**		**beta±std error**		**best**
rsid	**RE**	**CV**	**GWAMA**	**SST(F)**	**SST(M)**	**female**	**male**	**methods**
rs560887	5.30e-12	**1.14e-12**	1.39e-09	8.19e-06	4.83e-08	0.011±0.002	0.014±0.002	CV
rs7298683	2.01e-07	8.14e-05	7.23e-08	0.86718	**5.92e-09**	-0.0007±0.004	0.026±0.004	SST(M)

Phenotype : HDL				**sex-specific**		**beta±std error**		**best**
rsid	**RE**	**CV**	**GWAMA**	**SST(F)**	**SST(M)**	**female**	**male**	**methods**
rs2167079	**4.46e-08**	1.26e-06	5.06e-08	0.01838	9.78e-08	-0.024±0.010	-0.055±0.010	RE, GWAMA
rs7120118	**4.62e-08**	1.13e-06	5.07e-08	0.01679	1.15e-07	-0.024±0.010	-0.054±0.010	RE, GWAMA
rs1532085	**3.53e-12**	1.08e-11	2.98e-11	1.87e-06	2.25e-07	-0.048±0.010	-0.050±0.009	RE
rs3764261	**4.15e-32**	4.89e-32	2.88e-31	8.85e-16	3.63e-18	-0.090±0.011	-0.092±0.010	RE,CV
rs255049	**5.26e-09**	1.45e-08	3.94e-08	0.00014	5.85e-06	-0.047±0.012	-0.053±0.011	RE
rs1800961	**2.39e-08**	2.82e-08	8.96e-08	0.00051	6.29e-06	0.083±0.023	0.108±0.023	RE,CV

Phenotype : LDL				**sex-specific**		**beta±std error**		**best**
rsid	**RE**	**CV**	**GWAMA**	**SST(F)**	**SST(M)**	**female**	**male**	**methods**
rs646776	**1.97e-15**	4.17e-15	9.71e-15	8.36e-10	2.11e-07	0.169±0.027	0.166±0.031	RE
rs693	**1.25e-11**	2.97e-11	5.84e-11	8.81e-09	0.00013	-0.131±0.022	-0.103±0.026	RE
rs11668477	1.98e-08	**4.04e-09**	1.96e-08	0.00240	2.62e-07	0.090±0.029	0.179±0.034	CV

Phenotype : Triglyceride				**sex-specific**		**beta±std error**		**best**
rsid	**RE**	**CV**	**GWAMA**	**SST(F)**	**SST(M)**	**female**	**male**	**methods**
rs673548	**2.61e-08**	6.54e-08	2.21e-07	6.11e-06	0.00097	0.058±0.012	0.054±0.016	RE
rs1260326	**1.64e-10**	1.88e-10	1.49e-09	1.32e-06	2.10e-05	-0.057±0.011	-0.065±0.015	RE,CV
rs10096633	5.32e-08	**1.98e-08**	1.19e-07	0.00109	3.47e-06	0.062±0.019	0.113±0.024	CV

Table S3. Association mapping results of the NFBC data. The SNPs are presented for which any method gave $P < 5 \times 10^{-8}$ in any of 10 phenotypes. For each SNP, the most significant p-value among the four methods (RE, CV, GWAM, and SST) is in bold font. The "beta" column shows the effect sizes of male- and female-only studies and their estimated standard errors. For each significant association, we report the "best methods", the set of methods whose p-value was less than two times the most significant p-value for each association.