Supplementary File for “FOLD: A method to optimize power in meta-analysis of genetic association studies with overlapping subjects”

Emma E. Kim, Seunghoon Lee, Cue Hyunkyu Lee, Hyunjung Oh, Kyuyoung Song and Buhm Han

This document contains Supplementary Text and Figures.

SUPPLEMENTARY TEXT

S1. Other Methods for Meta-Analysis with Overlapping Samples

Zaykin and Kozbur’s Method

Zaykin and Kozbur (Zaykin and Kozbur, 2010) derived both the correlation formula for unsigned statistics (e.g. \( \chi^2 \) statistic) and the correlation formula for signed statistics (e.g. \( z \)-score), where the latter was equivalent to the derivation of Lin and Sullivan (Lin and Sullivan, 2009). The correlation of the unsigned statistic was as follows:

\[
 r_{ij} = \left( \frac{1}{1 + \frac{(n_{i-} - n_{ij-})}{n_{ij-}}} \right) \left( \frac{1}{1 + \frac{(n_{j-} - n_{ij-})}{n_{ij-}}} \right) \left( \frac{1}{1 + \frac{(n_{j-} - n_{ij-})}{n_{ij+}}} + \frac{n_{ij-}}{n_{ij+}} \right) \left( \frac{1}{1 + \frac{(n_{i-} - n_{ij-})}{n_{ij+}}} + \frac{n_{ij-}}{n_{ij+}} \right)
\]

where \( n_{i+}, n_{i-}, n_{j+}, \) and \( n_{j-} \) are the numbers of cases and controls in the \( i \)th and \( j \)th studies, respectively, and \( n_{ij-} \) is the number of overlapping control samples between \( i \)th and \( j \)th studies. Zaykin and Kozbur defined weighted \( z \)-scores as

\[
 Z_i = w_i z_i
\]

where \( z_i \) is the \( z \)-score and \( w_i \) is the weight of the \( i \)th study such that

\[
 w_i = \sqrt{n_i (R^{-1})_{ii}}
\]

\((R^{-1})_{ii}\) refers to the \( i \)th diagonal of the inverse correlation matrix and \( n_i \) is the sample size of the \( i \)th study. Thereafter, they defined the standardized sum of the weighted \( z \)-scores
\[
\frac{\sum Z_i}{\sqrt{\sum W_i^2 + 2 \sum_{i<j} W_i W_j R_{ij}}}
\]

which is used to calculate \( P \)-value.

**ASSET**

Bhattacharjee et al. proposed a meta-analysis method called ASSET (Bhattacharjee et al., 2012). They sought to account for the effect size heterogeneity (presence/absence of effects) in order to increase the power. To account for sample overlap, Bhattacharjee et al. employed the Lin-Sullivan method and updated the correlation formula to a more general form that allows the sharing of a control as a case in another study. The updated formula was as follows:

\[
r_{ij} = \frac{n_i n_{i+} - n_j n_{j+}}{n_i \sqrt{n_j \left( n_i n_{j-} - n_{i+} n_{j-} - n_{i-} n_{j+} + n_{i+} n_{j+} \right)}}
\]

where \( n_i \), \( n_{i+} \) and \( n_{i-} \) are the total numbers of subjects, cases, and controls in the \( i \)th study respectively, and \( n_j \), \( n_{j+} \) and \( n_{j-} \) are similarly defined for the \( j \)th study. \( n_{ij-} \) and \( n_{ij+} \) denote the numbers of shared cases and shared controls between the \( i \)th and \( j \)th study, respectively. In addition, \( n_{i-j+} \) and \( n_{i+j-} \) represent the shared numbers between the cases of the \( i \)th study and the controls of the \( j \)th study, and the shared numbers between the controls of the \( i \)th study and the cases of the \( j \)th study, respectively. Refer to Bhattacharjee et al. (Bhattacharjee et al., 2012) for details of how they iterate all possible combinations of presence/absence of effects and apply the Lin-Sullivan method to each combination.

**Decoupling Method**

Han et al. (Han et al., 2016) proposed to “decouple” correlated statistics so that the statistics become independent, following which the standard meta-analysis methods can be applied. Given the correlation matrix \( R \), they calculated a transformed covariance structure:

\[
\Omega_{decoupled} = \text{Diag}(e^T \text{Diag}(s) \cdot R \cdot \text{Diag}(s))^{-1}
\]
where $s$ is the standard error of the estimates, $Diag(s)$ is a diagonal matrix whose diagonals are $s$, and $e$ is a vector of ones. Thereafter, they updated the standard error

$$s_{decoupled}[i] = \sqrt{\Omega_{decoupled}[i,i]} \quad i = 1, ..., K$$

where $K$ denotes the number of studies used in meta-analysis and $\Omega_{decoupled}[i,i]$ denotes the $i$th diagonal element of $\Omega_{decoupled}$. As the decoupled data can now be considered independent, the standard meta-analysis methods can be used. Han et al. proved that the decoupling approach is equivalent to the Lin-Sullivan method when the fixed effects model is applied to the decoupled data.

**LD Score Regression (LDSC)**

Bulik-Sullivan et al. proposed the LDSC method, which can distinguish the inflation driven by systematic confounders from the polygenic effects (Bulik-Sullivan et al., 2015). Therefore, LDSC can correct for the inflation caused by overlapping subjects after the meta-analysis is performed. In our simulations, we performed fixed effects meta-analysis while ignoring the correlations and corrected for the inflation using LDSC. We assumed an ideal condition where the LDSC estimate of the inflation factor was perfectly accurate.

**S2. Variance analysis of FOLD**

Here we analytically demonstrate that the FOLD estimator can achieve smaller variance as compared to the Lin-Sullivan estimator under a case/control study design wherein studies share a subset of controls. We employed the standard procedure for obtaining the estimator variance as described by Lin and Sullivan (Lin and Sullivan, 2009), as follows.

Let $X_{ki}$ denote the explanatory variables (e.g., SNP genotype dosages) at the $i$th subject in study $k$, and let $\theta_k = (\alpha_k, \beta_k)$ represent the parameters of the logistic regression model for
study $k$, where $\alpha_k$ and $\beta_k$ denote the intercept and regression parameters, respectively. The corresponding “Information matrix” (Lin and Sullivan, 2009) of study $k$ is

$$I_k(\theta_k) = \sum_{i=1}^{n_k} \frac{e^{\alpha_k + \beta_k X_{ki}}}{1 + e^{\alpha_k + \beta_k X_{ki}}} \begin{bmatrix} 1 & X_{ki} \ X_{ki}^T \end{bmatrix} - \frac{n_k}{(1 + e^{\alpha_k})^2} \begin{bmatrix} 1 & X_{ki} \ X_{ki} X_{ki}^T \end{bmatrix}$$

where $n_k$ denotes the sample size of study $k$. We assume the null hypothesis ($\beta_k = 0$) to approximate the variance of the estimators. Thus,

$$I_k(\theta_k) = n_k \frac{e^{\alpha_k}}{(1 + e^{\alpha_k})^2} \sum_{i=1}^{n_k} \frac{1}{n_k} \begin{bmatrix} 1 & X_{ki} \ X_{ki} X_{ki}^T \end{bmatrix}$$

which can be approximated as

$$I_k(\theta_k) \approx n_k \frac{e^{\alpha_k}}{(1 + e^{\alpha_k})^2} H$$

where $H$ denotes the expectation of $\begin{bmatrix} 1 & X_{ki} \ X_{ki} X_{ki}^T \end{bmatrix}$. When considering a single explanatory variable (e.g., SNP), the variance of the $\beta_k$ estimator is approximated using the inverse of this quantity, such that

$$Var(\hat{\beta}_k) \approx \left(n_k \frac{e^{\alpha_k}}{(1 + e^{\alpha_k})^2} \right)^{-1} h$$

where $h$ is the second diagonal element of $H^{-1}$. We can assume that $h$ is constant across studies under the assumptions that the minor allele frequencies are constant and that the SNP follows Hardy-Weinberg equilibrium. Note that $e^{\alpha_k} \approx \pi_k$, where $\pi_k$ is the number of cases divided by the number of controls in study $k$. Thus,

$$Var(\hat{\beta}_k) \approx \frac{(1 + \pi_k)^2}{n_k \pi_k} h$$

We now consider two studies, $k$ and $l$. Let the subscripts $kl -$ and $kl +$ indicate overlapping control and case samples, respectively. Then, from Equation 1:

$$r_{kl} = \frac{n_{kl-} \sqrt{\pi_k \pi_l} + n_{kl+} \sqrt{\frac{1}{\pi_k \pi_l}}}{\sqrt{n_k n_l}}$$
For simplicity, we assume that the two studies $k$ and $l$ have the same numbers of samples as well as the same proportions of cases over controls: $n_k = n_l = n$ and $\pi_k = \pi_l = \pi$. Let $\eta$ and $1 - \eta$ denote the fractions of the shared and study-specific subjects, respectively. We assume $\eta$ to be the same for both shared controls and shared cases. That is, we assume that both a subset of cases and subset of controls are shared by the two studies, with the same fraction $\eta$. Then,

$$r_{kl} = \eta \left( \pi + \frac{1}{\pi} \right)$$

The inverse of the correlation matrix is

$$R^{-1} = \frac{1}{1 - (r_{kl})^2} \begin{bmatrix} 1 & -r_{kl} \\ -r_{kl} & 1 \end{bmatrix}$$

Thus, we can derive the inverse of the covariance matrix $\Sigma_{LS}^{-1}$ and calculate the variance of the Lin and Sullivan estimator

$$\text{Var}(\hat{\beta}_{LS}) = \frac{1}{e^T \Sigma_{LS}^{-1} e} = \frac{h(1 + \pi)^2}{2n\pi} \left( 1 + \eta \left( \pi + \frac{1}{\pi} \right) \right) = \frac{C}{2n} \left( 1 + \eta \left( \pi + \frac{1}{\pi} \right) \right)$$

where

$$C = \frac{h(1 + \pi)^2}{\pi}$$

In contrast, FOLD divides the subjects of each study into groups (configurations) based on the sharing of the subjects with other studies. In our design, because both cases and controls are shared by the two studies, the application of FOLD is equivalent to the following five-step procedure. (1) Obtain $\hat{\beta}_{k,\text{Share}}$ by comparing the shared cases to the shared controls in study $k$. (2) Obtain $\hat{\beta}_{k,\text{spe}}$ by comparing study $k$-specific cases to study $k$-specific controls. (3) Similarly, obtain $\hat{\beta}_{l,\text{Share}}$ and $\hat{\beta}_{l,\text{Spe}}$. (4) Combine the two estimates based on the shared subjects ($\hat{\beta}_{k,\text{Share}}$ and $\hat{\beta}_{l,\text{Share}}$) into $\hat{\beta}_{\text{Share}}$, while accounting for their correlation. Note that $\hat{\beta}_{k,\text{Share}}$ and $\hat{\beta}_{l,\text{Share}}$ are the only correlated pair among all possible pairs in this design. In fact, these two estimates are the same redundant estimates (the correlation is 1). Thus, combining these estimates is equivalent to using only one of the estimates ($\hat{\beta}_{\text{Share}} = \hat{\beta}_{k,\text{Share}}$).
(5) Finally, combine $\hat{\beta}_{k,\text{Spec}}, \hat{\beta}_{l,\text{Spe}}$ and $\hat{\beta}_{\text{Share}}$ which are independent. To obtain the variance of the FOLD estimator, note that

$$\text{Var}(\hat{\beta}_{\text{Share}}) = \frac{h(1+\pi)^2}{\eta n \pi} = \frac{C}{\eta n}$$

$$\text{Var}(\hat{\beta}_{k,\text{Spec}}) = \text{Var}(\hat{\beta}_{l,\text{Spe}}) = \frac{h(1+\pi)^2}{(1-\eta)n \pi} = \frac{C}{(1-\eta)n}$$

Thus, the combination of these three $\beta$ estimators using the inverse-variance-weighted average provides the final variance:

$$\text{Var}(\hat{\beta}_{\text{FOLD}}) = \frac{C}{(2-\eta)n}$$

This is logical because there are $(2-\eta)n$ unique individuals.

A comparison of the variances of the FOLD and Lin-Sullivan estimators shows that

$$\text{Var}(\hat{\beta}_{\text{LS}}) - \text{Var}(\hat{\beta}_{\text{FOLD}}) = \frac{C\eta((1-\pi)^2 + \pi^2(1-\eta) + (1-\eta) + \pi)}{2(2-\eta)n \pi} > 0$$

given $0 < \eta < 1$ and $0 < \pi$. Therefore, the FOLD estimator is more efficient than the Lin-Sullivan estimator in this case/control study design.

**S3. Power simulations**

We used the following power simulation scheme. We assumed that we have $K$ studies. To simulate each study, we assumed $n^+$ case and $n^-$ control samples ($2n^+$ case chromosomes and $2n^-$ control chromosomes) and assumed a variant with a minor allele frequency (MAF) of $p = 0.3$. Given a relative risk $\gamma$, which we assumed to be the same for all studies (which is the fixed effects model assumption), the expected case MAF becomes $p = \frac{\gamma p}{(\gamma-1)p+1}$ and the expected control MAF becomes $p^+ \approx p$, assuming a low prevalence. Thus, we can construct a $2 \times 2$ table by randomly sampling minor allele counts $x_{k}^+$ from $\text{Binomial}(2n^+, p^+)$ for cases and $x_{k}^-$ from $\text{Binomial}(2n^-, p^-)$ for controls, respectively, where $\text{Binomial}(n, p)$ refers to a
binomial distribution for \( n \) trials with probability \( p \). From this \( 2 \times 2 \) table, we can also calculate the log odds ratio and its approximated variance. To simulate independent studies with no overlapping samples, we repeated this procedure to generate \( K \) studies.

To simulate shared controls in cross-disease analysis, we modified the simulation scheme as follows. To simplify the simulations, we assumed that all shared controls were shared by all \( K \) studies. Let \( \eta \) denote the proportion of shared controls among all controls contained within each study. We assumed that the case/control sample sizes as well as \( \eta \) were equal between all studies. Given \( \eta \), to ensure that the total number of unique individuals in the meta-analysis was constant, which will maintain the power of the splitting method, we adjusted the control sample size from \( n^- \) to \( m^- = \frac{K}{K(1-\eta)+\eta} n^- \). The splitting method in this situation was equivalent to sampling \( x_{k,\text{Spe}}^- \) from \( \text{Binomial}(2(1-\eta)m^-, p^-) \) for study-specific controls in study \( k \), and \( x_{k,\text{Split}}^- \) from \( \text{Binomial}(\frac{2\eta m^-}{K}, p^-) \) for controls that were shared but were now distributed to study \( k \) after splitting. The total minor allele count of controls in study \( k \) after splitting was \( x_{k,\text{Spe}}^- + x_{k,\text{Split}}^- \) (which was essentially the same as sampling from \( \text{Binomial}(2m^-, p^-) \)). This data was used as input data for the splitting method. Because we also wanted to evaluate other methods that account for overlapping subjects without splitting (e.g. the Lin and Sullivan method), the minor allele count of shared controls prior to splitting was calculated by merging the counts of split controls, which was \( x_{k,\text{Share}}^- = \sum x_{k,\text{Split}}^- \). The total minor allele count of the controls in study \( k \) before splitting was the sum of \( x_{k,\text{Spe}}^- + x_{k,\text{Share}}^- \). This data was used as input data for these methods that account for overlapping subjects. The power of the methods was evaluated as the proportion of 10 000 simulations whose \( P \)-values were less than \( 5 \times 10^{-8} \).

**S4. WTCCC and PGC data analysis**
**WTCCC data analysis**

We obtained data from the Wellcome Trust Case Control Consortium (WTCCC) (Wellcome Trust Case Control Consortium, 2007). These data included one shared control group for the analysis of seven different diseases. In their study (Wellcome Trust Case Control Consortium, 2007), the authors combined the genotype data of CD, RA, and T1D to identify loci putatively involved in shared pathways. This analysis resulted in two significant loci (rs17696736 and rs12769947) excluding the well-known risk factors, the major histocompatibility complex (MHC) and *PTPN22* \((P < 5 \times 10^{-7})\); See Supplementary Table 2 of The Wellcome Trust Case Control Consortium (Wellcome Trust Case Control Consortium, 2007)). In addition to these two loci, we further investigated pleiotropic loci known to be associated with multiple autoimmune diseases. From ImmunoBase (http://www.immunobase.org), we extracted loci that were reported to be associated with more than one of the three diseases (i.e., CD, RA, and T1D). This gave us six additional loci. Although these loci were not significant \((P > 5 \times 10^{-7})\) in the WTCCC data, they showed nominally significant association signals \((P < 0.05)\). Because these small signals became apparent in the subsequent studies (thus allowing them to be deposited in the Immunobase), we assumed that they may represent true-positives.

We evaluated the meta-analysis \(P\)-values of these eight loci using two different designs. First, we tried a full overlap design (Supplementary Table S2). Rather than combining actual genotype data as was conducted in the original study, we used a meta-analysis framework to combine the results of the three diseases into one. After quality control, 2938 shared controls were available (1480 individuals from the 1958 British Birth Cohort and 1458 individuals from the UK Blood Service Control Group). We compared all these controls to cases of CD \((n = 1748)\), RA \((n = 1860)\), and T1D \((n = 1963)\) in the logistic regression framework.
Second, we tried a partial overlap design (Supplementary Table S1). To this end, we altered the study design by re-distributing the control samples into disease-specific controls of the three diseases and shared controls. To avoid possible bias induced by the two different sources of controls (1958 British Birth Cohort and UK Blood Service Control Group), we performed this re-distribution for each source separately. After re-distribution, each disease included 734 disease-specific controls and 736 controls shared by the three diseases. Thus, approximately half of the controls within each disease analysis was shared by the three diseases ($\eta \approx 0.5$). Applying FOLD to this design involved calculating six estimates ($\hat{\beta}_{CD,\text{Spe}}, \hat{\beta}_{CD,\text{Share}}, \hat{\beta}_{RA,\text{Spe}}, \hat{\beta}_{RA,\text{Share}}, \hat{\beta}_{T1D,\text{Spe}}, \hat{\beta}_{T1D,\text{Share}}$).

**S4.2 PGC data analysis: 2×2 table approximation**

We obtained the summary statistics data (OR and standard error) of the association studies of five psychiatric diseases from the PGC website (Cross-Disorder Group of the Psychiatric Genomics, 2013). Although we needed the 2 × 2 tables that generated these summary statistics, these were unavailable. Therefore, we approximately reconstructed the 2 × 2 tables that would likely have given the summary statistics. We obtained MAF of the SNPs from HapMap3 (International HapMap Consortium, 2003). Given the numbers of cases and controls and the marginal MAF, the degree of freedom of a 2 × 2 table became one. Thus, we were able to enumerate all possible tables to search for a table that would give the closest $P$-value to the reported $P$-value while giving the same direction to the reported OR. The resulting table gave the control MAF, which was used to simulate additional shared controls (Supplementary Table S3).
References


Wellcome Trust Case Control Consortium (2007), Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, *Nature*, 447 (7145), 661-78.

**SUPPLEMENTARY FIGURES**

**Figure S1:** False positive rates of FOLD according to differing proportions of shared controls and numbers of studies in meta-analysis. The X-axis indicates the proportion of shared controls among all controls within each study and the Y-axis indicates the false positive rate given the significance threshold $\alpha = 0.05$. We considered four different numbers of studies: two (A), three (B), five (C), and ten (D). We also added splitting and the Lin-Sullivan method for comparison.
Figure S2: False positive rates of FOLD for various significance thresholds.
The X-axis indicates the proportion of shared controls among all controls within each study and the Y-axis indicates the false positive rates given the significance threshold \( \alpha = 0.01, 0.005, 0.001 \) and 0.0005. We assumed a meta-analysis of five studies.
**Figure S3:** Quantile-quantile plots of the association results using the splitting approach, the Lin-Sullivan method, and FOLD for the WTCCC data.
Figure S4: Power of FOLD where differing methods were used to combine multiple statistics in FOLD. At the second step of FOLD where we combine multiple statistics, we tried different methods other than the default choice of the Lin-Sullivan method. We examined four methods, the Lin-Sullivan method (a), ASSET (b), Zaykin-Kozbur method (c), and the decoupling method (d). The X-axis indicates the proportion of shared controls among all controls within each study and the Y-axis indicates power. We assumed a meta-analysis of five studies. As shown, all versions of FOLD except for the version using Zaykin-Kozbur method showed similar power to splitting, demonstrating that these methods can be used as a component of FOLD.
Figure S5: Comparison of three different splitting approaches. (A) Power of different methods. Strategy A is a strategy that equally distributes shared controls into studies (Equal-splitting). Strategy B is a strategy that distributes shared controls proportionally to the number of cases in the studies (Case-based splitting). (B) Splitting results of the methods. Given four studies with differing sample sizes, we distributed 10 000 shared controls into the studies using each method. $N_{\text{split}}$ denotes the size of shared controls distributed into each study. Case-based splitting denotes a strategy that distributes shared samples proportionally to the number of case samples, and FOLD-split denotes our proposed splitting method based on the effective number of samples.