Bioinformatics, 33(24), 2017, 3947–3954
doi: 10.1093/bioinformatics/btx463
Advance Access Publication Date: 20 July 2017
Original Paper

Genetic and population analysis

FOLD: a method to optimize power in meta-analysis of genetic association studies with overlapping subjects

Emma E. Kim,[1,2] Seunghoon Lee,[2] Cue Hyunkyu Lee,[3] Hyunjung Oh,[4] Kyuyoung Song[4] and Buhm Han[1,3,]*

[1]Asan Institute for Life Sciences, Asan Medical Center, Seoul 138-736, Korea, [2]Department of Chemistry, Seoul National University, Seoul 151-747, Korea, [3]Department of Convergence Medicine and [4]Department of Biochemistry and Molecular Biology, University of Ulsan College of Medicine, Seoul 138-736, Korea

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on December 6, 2016; revised on June 29, 2017; editorial decision on July 13, 2017; accepted on July 19, 2017

Abstract

Motivation: In genetic association studies, meta-analyses are widely used to increase the statistical power by aggregating information from multiple studies. In meta-analyses, participating studies often share the same individuals due to the shared use of publicly available control data or accidental recruiting of the same subjects. As such overlapping can inflate false positive rate, overlapping subjects are traditionally split in the studies prior to meta-analysis, which requires access to genotype data and is not always possible. Fortunately, recently developed meta-analysis methods can systematically account for overlapping subjects at the summary statistics level.

Results: We identify and report a phenomenon that these methods for overlapping subjects can yield low power. For instance, in our simulation involving a meta-analysis of five studies that share 20% of individuals, whereas the traditional splitting method achieved 80% power, none of the new methods exceeded 32% power. We found that this low power resulted from the unaccounted differences between shared and unshared individuals in terms of their contributions towards the final statistic. Here, we propose an optimal summary-statistic-based method termed as FOLD that increases the power of meta-analysis involving studies with overlapping subjects.

Availability and implementation: Our method is available at http://software.buhmhan.com/FOLD.

Contact: mail: buhm.han@amc.seoul.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In recent genetic association studies, researchers commonly use meta-analyses to achieve sufficient statistical power. A meta-analysis involves a combination of the summary statistics from multiple studies. These participating studies often share the same subjects because the researchers either used publicly available controls or recruited the same subjects accidentally at multiple sites (Chubb *et al.*, 2013; Crowther-Swanepoel *et al.*, 2009; Di Bernardo *et al.*, 2008; Kilpivaara *et al.*, 2009; Mukherjee *et al.*, 2011; Onengut-Gumuscu *et al.*, 2015; Orozco *et al.*, 2014; Shete *et al.*, 2009; Speedy *et al.*, 2014; Weinhold *et al.*, 2013; Wellcome Trust Case Control Consortium, 2007; Zhao *et al.*, 2007). In particular, the repeated use of the same controls is prevalent in cross-disease meta-analyses, where association results of multiple diseases are combined to uncover pleiotropic loci (Dichgans *et al.*, 2014; Kar *et al.*, 2016; Moskvina *et al.*, 2013). These overlapping subjects can induce correlations between the summary statistics and inflate the false positive rate of meta-analyses. A straightforward solution is to split the overlapping subjects in the studies prior to meta-analysis, which requires access to genotype data and is not always possible.

Fortunately, recently developed meta-analysis methods can systematically account for overlapping subjects at the summary statistics level (Bhattacharjee *et al.*, 2012; Bulik-Sullivan *et al.*, 2015; Han *et al.*, 2016; Lin and Sullivan, 2009; Zaykin and Kozbur, 2010). Most of these methods explicitly model correlations between the statistics, such as the Lin and Sullivan method (Lin and Sullivan, 2009), Zaykin and Kozbur method (Zaykin and Kozbur, 2010), ASSET (Association analysis for SubSETs) (Bhattacharjee *et al.*, 2012) and decoupling method (Han *et al.*, 2016). LD score regression (LDSC) (Bulik-Sullivan *et al.*, 2015) does not explicitly model correlations, but can distinguish inflation caused by overlapping subjects from the true polygenic effects, thus enabling appropriate control of the false positive rate.

In this study, we found that these methods designed to account for overlapping subjects at the summary statistics level can yield low power compared to the traditional splitting method. For instance, in our simulation involving a meta-analysis of five studies that share 20% of individuals, whereas the traditional splitting method achieved 80% power, none of these methods exceeded 32% power. Notably, the use of these methods often reduced power even below the level of an analysis based on data in which all overlapping samples were discarded. We discovered that this low power resulted from the unaccounted difference between the shared and unshared subjects in terms of their contributions towards the final statistic. To increase the power of the meta-analysis involving overlapping subjects, we developed an optimal summary-statistic-based method termed as FOLD (Fully powered method for OverLapping Data). In this method, we categorize subjects based on their contributions to the final statistic and then calculate the summary statistic per each category. We analytically show that the FOLD estimator can achieve smaller variance than the current methods. Moreover, we propose a companion method FOLD-split that determines the optimal splitting design in cases where genotype data are available and subjects can be split prior to meta-analysis.

2 Materials and methods

2.1 Existing methods for meta-analysis with overlapping subjects

A straightforward solution to account for overlapping subjects is to split the genotype data of overlapping subjects into individual studies before meta-analysis. However, this approach requires access to genotype data, and hence, the methods mentioned hereafter were developed.

2.1.1 Lin and Sullivan (LS) method

Lin and Sullivan (2009) were the first to introduce a summary-statistic-based method that can account for the correlations between statistics caused by overlapping subjects. The LS method analytically approximates the correlations between statistics and then constructs an optimal meta-analysis statistic while taking the correlations into account.

Let $X_1, \ldots, X_K$ denote the observed effect sizes in K studies in the meta-analysis, and $V_1, \ldots, V_K$ represent their variances. Lin and Sullivan (2009) derived a formula for correlation of X_i and X_j:

$$r_{ij} \approx \left(n_{ij-}\sqrt{\frac{n_{i+}n_{j+}}{n_{i-}n_{j-}}} + n_{ij+}\sqrt{\frac{n_{i-}n_{j-}}{n_{i+}n_{j+}}} \right) / \sqrt{n_i n_j} \qquad (1)$$

where n_i, n_j, and n_{ij} are the total number of samples in the ith and jth studies and the number of overlapping samples between the two studies, respectively. Subscripts $+$ and $-$ denote the case and control subjects, respectively. Given the correlation matrix of $X = (X_1, \ldots, X_K)$,

$$R = \left[r_{ij} \right]_{K \times K},$$

the covariance matrix, Ω, can easily be calculated.

Lin and Sullivan (2009) proposed a score test statistic assuming the fixed effects model:

$$X_{\mathrm{LS}} = \frac{e^T \Omega^{-1} X}{e^T \Omega^{-1} e}$$

where e is a $K \times 1$ vector with ones. The variance of this statistic is

$$\mathrm{Var}(X_{\mathrm{LS}}) = \frac{1}{e^T \Omega^{-1} e}$$

Therefore, one can obtain a z-score,$\frac{X_{\mathrm{LS}}}{\sqrt{\mathrm{Var}(X_{\mathrm{LS}})}}$, as well as a P-value. When there is no overlapping of subjects, the LS method becomes equivalent to the traditional fixed effects model (inverse-variance-weighted average) method.

2.1.2 Other methods for meta-analysis with overlapping samples

For details of Zaykin-Kozbur method (Zaykin and Kozbur, 2010), ASSET (Bhattacharjee *et al.*, 2012), decoupling method (Han *et al.*, 2016) and LDSC (Bulik and Sullivan *et al.*, 2015), see Section S1 of Supplementary Text.

2.2 FOLD

2.2.1 Motivation

We first provide a toy example to demonstrate the cause of the low power of the existing methods. For simplicity, we will consider samples drawn from a normal distribution with mean β, and test whether the mean is non-zero. We consider one dataset consisting of n samples, $\mathbf{x} = (x_1, \ldots, x_n)$ and another independent dataset of the same size, $\mathbf{y} = (y_1, \ldots, y_n)$. All $2n$ samples are independent draws from $N(\beta,1)$. The standard approach is to merge all the samples and calculate the statistic $\widehat{\beta} = \frac{1}{2n}\left(\sum_i x_i + \sum_i y_i\right)$, where $\widehat{\beta} \sim N\left(\beta,\ \frac{1}{2n}\right)$. Alternatively, in a meta-analytic setting, we may only have access to the summary statistics from the two datasets, namely $\widehat{\beta_1} = \frac{1}{n}\sum_i x_i$ and $\widehat{\beta_2} = \frac{1}{n}\sum_i y_i$. We can meta-analyze these values as $\widehat{\beta_{\mathrm{meta}}} = \frac{1}{2}\left(\widehat{\beta_1} + \widehat{\beta_2}\right)$, which can be considered as the inverse-variance weighted average and is equivalent to $\widehat{\beta}$ above. We now assume that $\mathbf{x}$ and $\mathbf{y}$ are correlated such that each sample pair originated from a multivariate normal distribution, $(x_i, y_i) \sim \mathrm{MVN}((\beta, \beta),\ R)$, where $R = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$ and $r > 0$. Then, let $\overrightarrow{\beta} = \left(\widehat{\beta_1}, \widehat{\beta_2}\right) \sim \mathrm{MVN}\left((\beta,\ \beta), V\right)$ be a vector containing the two summary statistics, where $V = \frac{1}{n}R$. Analogous to the LS method, given $\overrightarrow{\beta}$, an optimal meta-analysis statistic would be $\widehat{\beta_{LS}} = \frac{\overrightarrow{\beta}\, V^{-1} e}{e^T V^{-1} e}$, where e is a vector of ones. Note that $\widehat{\beta_{LS}}$ is the same as $\frac{1}{2}\left(\widehat{\beta_1} + \widehat{\beta_2}\right)$ regardless of r for this two-study example. The variance of $\widehat{\beta_{LS}}$ is $\frac{1}{e^T V^{-1} e}$, which is calculated as $\frac{1+r}{2n}$. Thus, we can consider the additional variance $\frac{r}{2n}$ as a penalty for having correlations between samples. In other words, the $2n$ correlated samples contain less information with respect to the final statistic, as if we have $\frac{2n}{1+r}$ independent samples. Note that this was the intuition of the decoupling approach of Han *et al.* (2016). The decoupling approach transforms the data to obtain a new variance V' whose non-diagonals become zero and whose diagonals increase such that $V' = \frac{1}{n}\begin{bmatrix} 1+r & 0 \\ 0 & 1+r \end{bmatrix}$. Thus, the decoupling approach can be thought of as translating the lower amount of information in dependent data into increased variances.

Next, we assume that half of $\mathbf{x}$ and $\mathbf{y}$ are independent and the other half are correlated. That is, $(x_i, y_i) \sim \text{MVN}((\beta, \beta),\ I)$ for $i = 1, \ldots, n/2$, and $(x_i, y_i) \sim \text{MVN}((\beta, \beta),\ R)$ for $i = n/2 + 1, \ldots, n$. Then, $\overrightarrow{\beta}$ follows $\text{MVN}((\beta,\ \beta), V)$ where $V = \frac{1}{n}\begin{bmatrix} 1 & r/2 \\ r/2 & 1 \end{bmatrix}$. The application of the LS method gives a variance of $\frac{1+r/2}{2n}$. However, this approach may not be optimal in this situation as the independent samples contain more information than the dependent samples as demonstrated by the decoupling approach. Thus, samples are heterogeneous in terms of their contributions to the final statistic. This situation is related to *heteroscedasticity* (Rao, 1973), which often describes a situation in which samples have different variances (Foulley and Quaas, 1995; Yin *et al.*, 2011). If the data for the dependent samples are transformed using the decoupling approach (Han *et al.*, 2016), the variances would increase. Thus, the two types of samples would be heteroscedastic with respect to the decoupled data. An optimal strategy in this situation is to perform meta-analysis separately for independent samples ($x_1, \ldots, x_{n/2}, y_1, \ldots, y_{n/2}$) and dependent samples ($x_{n/2+1}, \ldots, x_n, y_{n/2+1}, \ldots, y_n$), and perform another meta-analysis on the two results. The variance for the meta-analysis statistic of the independent samples is $\frac{1}{n}$ and the variance for the meta-analysis statistic of dependent samples is $\frac{1+r}{n}$. Thus, using the inverse-variance-weighted average, the final variance becomes $\frac{1+r}{n(2+r)}$. Note that the variance has now decreased as compared with the naïve application of the LS method, because $\frac{1+r/2}{2n} - \frac{1+r}{n(2+r)} = \frac{r^2}{4n(2+r)} > 0$. This shows that when a subset of samples exhibits a correlation structure, simply aggregating these samples can lead to suboptimal performance.

2.2.2 FOLD framework

We propose a summary-statistic-based meta-analysis framework that can account for the overlapping samples without losing power. The main idea is to categorize subjects in a study based on their contributions to the final statistic and then calculate the summary statistic for each category. For example, consider a combination of two case/control studies A and B (Fig. 1), which share a subset of controls. We first calculate the log odds ratio by comparing the study A cases to the controls specific to A. This estimator is referred to as $\widehat{\beta}_{A,\text{Spe}}$ (where 'Spe' denotes specific). We then calculate the log odds ratio by comparing the same study A cases to the shared controls. This estimator is referred to as $\widehat{\beta}_{A,\text{Share}}$. We similarly obtain $\widehat{\beta}_{B,\text{Spe}}$ and $\widehat{\beta}_{B,\text{Share}}$ from study B. We then consider the vector ($\widehat{\beta}_{A,\text{Spe}}$, $\widehat{\beta}_{A,\text{Share}}$, $\widehat{\beta}_{B,\text{Spe}}$, $\widehat{\beta}_{B,\text{Share}}$), of which the correlation matrix of this vector is as follows:

$$R = \begin{bmatrix} 1 & r_A & 0 & 0 \\ r_A & 1 & 0 & r_{AB}' \\ 0 & 0 & 1 & r_B \\ 0 & r_{AB}' & r_B & 1 \end{bmatrix}.$$

r_{AB}' is the correlation between $\widehat{\beta}_{A,\text{Share}}$ and $\widehat{\beta}_{B,\text{Share}}$, which is driven by their shared controls. Because all the controls rather than a specific subset are shared between $\widehat{\beta}_{A,\text{Share}}$ and $\widehat{\beta}_{\text{Share}}$, no heterogeneity is present with respect to the information contained in each control. r_A is the correlation between $\widehat{\beta}_{A,\text{Spe}}$ and $\widehat{\beta}_{A,\text{Share}}$, which is driven by the re-use of the study A cases in the calculation of these two statistics. Similarly, r_B is the correlation between $\widehat{\beta}_{B,\text{Spe}}$ and $\widehat{\beta}_{B,\text{Share}}$, which is driven by the re-use of the study B cases. Finally, we combine these four estimators while accounting for their correlation structure R (Fig. 1). We refer to this entire procedure as FOLD.

In practice, investigators commonly meta-analyze more than two studies. Moreover, subjects can be shared in a complicated manner; for example, for studies A, B and C, some controls can be shared by A and B, some controls can be shared by B and C, and some controls can be shared by all three. Therefore, we describe the general procedure of FOLD as follows:

1. Categorize the control subjects into T groups, where each group is homogeneous in terms of the sharing of subjects between studies. We refer to these as T 'configurations of sharing'.
2. From each of the K studies, obtain the T summary statistics, each of which is calculated using a specific control group versus all the cases.
3. Define the $KT \times KT$ correlation matrix between the KT statistics using Equation (1).
4. Apply the LS method to combine the KT statistics.

Although we assumed the sharing of controls, it is possible to generalize the procedure to the sharing of cases as well. Note that although T can be as large as $2^K - 1$ in theory, it is much smaller in practice. Moreover, because not all configurations of sharing may occur in all studies, each study typically has fewer than T configurations.

2.2.3 Variance analysis of FOLD

We analytically prove that FOLD is more efficient (yields smaller variance of estimator) than the LS method in case/control study designs under certain conditions (see Section S2 of Supplementary Text for details).

2.2.4 FOLD-split

In addition to FOLD, we developed a companion method called FOLD-split to facilitate splitting in conditions where splitting is possible. As Han *et al.* (2016) showed, the split proportions of overlapping subjects in multiple studies can markedly affect power. Previously, Lin and Sullivan (2009) suggested splitting the controls proportionally to the case sizes. However, an optimal splitting design may also depend on the study-specific controls and the configurations of subject sharing. We focus on the fact that the variances of the commonly used statistics are typically inversely proportional to the effective sample size $\text{n}_{\text{Eff},i} = \frac{4n_i^- n_i^+}{(n_i^- + n_i^+)}$, where n_i^+ and n_i^- denote the numbers of cases and controls in study i, respectively. If we use the inverse-variance-weighted average method for meta-analysis, the variance of the final estimator becomes inversely proportional to the total effective sample size,

$$n_{\text{Eff,Total}} = \sum\nolimits_{i=1}^{K} n_{\text{Eff},i}$$

where K denotes the total number of studies included in the meta-analysis. Thus, we set our goal to maximize this value. This problem is an integer nonlinear programming, because the numbers of split subjects are integers whereas the objective function is nonlinear. Let S_t be the set of studies that share controls belonging to a sharing configuration t. For each t, the conditions to satisfy are

$$\sum\nolimits_{i \in S_t} n_i^- \le n_t^-$$

$$n_i^- \ge 0$$

where n_i^- is the number of the shared controls split into study i, and n_t^- is the total number of the shared controls in t. In FOLD-split, we maximize $n_{\text{Eff,Total}}$ while satisfying these conditions. We solve

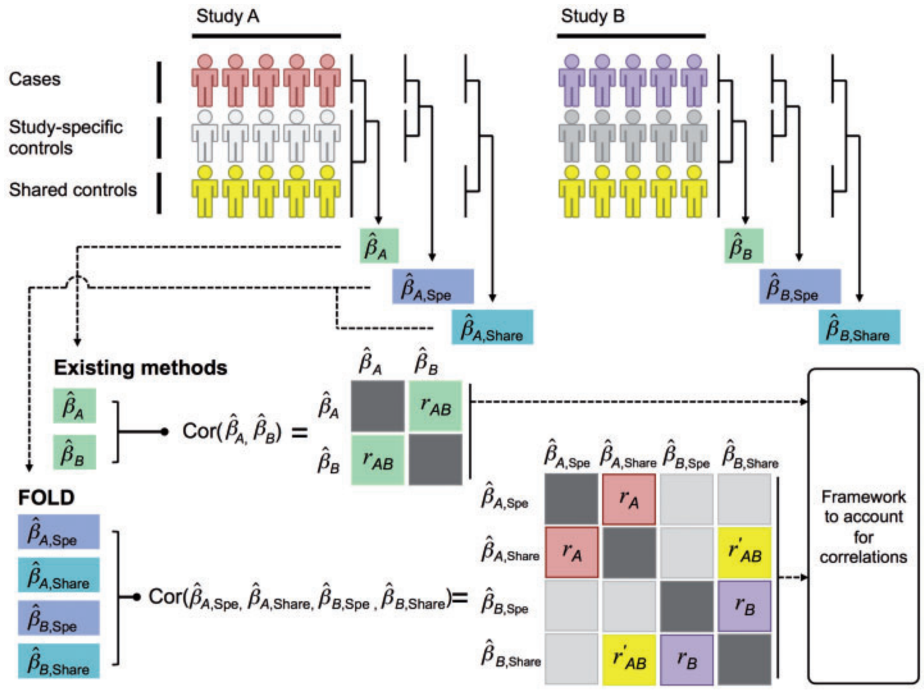

Fig. 1. Analysis pipelines of the existing meta-analysis methods and our proposed strategy FOLD. $\hat{\beta}$ represents the estimated effect size. Subscripts of $\hat{\beta}$ denote the study (A and B) and design of control samples (Spe, using study-specific controls; Share, using shared controls). In the existing methods, r_{AB} refers to the cross-study correlation between $\hat{\beta}_A$ and $\hat{\beta}_B$. In FOLD, r'_{AB} refers to the cross-study correlation between the two statistics $\hat{\beta}_{A,\text{Share}}$ and $\hat{\beta}_{B,\text{Share}}$that are calculated using shared controls only. r_A refers to the within-study correlation between the statistic calculated using study-specific controls ($\hat{\beta}_{A,\text{spe}}$) and the statistic calculated using shared controls ($\hat{\beta}_{A,\text{Share}}$). r_B is defined similarly to r_A. In the correlation matrix for FOLD, we applied a distinct color for each correlation element to denote which samples were overlapping and thus caused the correlation

this nonlinear optimization problem by utilizing the augmented Lagrange multiplier method (Ghalanos and Theussel, 2015).

2.2.5 Power simulations

See Section S3 of Supplementary Text for details of our power simulations.

2.2.6 WTCCC and PGC data

See Section S4 of Supplementary Text for details of our real data analyses using the Wellcome Trust Case Control Consortium (WTCCC) and the Psychiatric Genomics Consortium (PGC) data.

3 Results

3.1 Low power of existing methods designed to account for overlapping subjects

Using simulations, we comprehensively evaluated the power of the existing methods designed to account for overlapping subjects at the summary statistics level, including the LS method (Lin and Sullivan, 2009), Zaykin-Kozbur method (Zaykin and Kozbur, 2010), ASSET (Bhattacharjee *et al.*, 2012), decoupling method (Han *et al.*, 2016) and LDSC (Bulik-Sullivan *et al.*, 2015). We first simulated five studies each with 6000 cases and 6000 controls ($n^+ = 6000$ and $n^- = 6000$). We assumed a SNP with a relative risk of 1.16 and a minor allele frequency of 0.3. Then we modified the study design so that a certain proportion (η) of control samples of each study was shared by all five studies. Beginning with no overlap ($\eta = 0$), we gradually took some portions of controls from the five studies and used them as shared controls. Thus, we varied η from 0 to 1, while keeping the overall number of distinct control individuals contributing to the meta-analysis the same as 30 000. Thus, the power of the splitting approach was maintained at the same level, regardless of η. Here, the splitting approach refers to a strategy that splits the genotype data of shared controls into individual studies before meta-analysis. For example, the splitting approach for the LS method involves the application of the LS method after splitting (because the method can also be used for dataset without sample overlaps), and the splitting approach for each of the other methods is similarly defined.

Figure 2 shows that all these methods severely lose power, as compared to their corresponding splitting approaches. For example, at $\eta = 0.2$, the splitting approach for the LS method (equivalent to the standard fixed effects model meta-analysis) achieved 80% power, whereas without splitting, the LS method, Zaykin-Kozbur method, ASSET, decoupling method, and LDSC achieved only 31%, 28%, 23%, 31%, and 32% power, respectively. Interestingly, when all the controls were shared ($\eta = 1$), which we refer to as the *full overlap design*, no power drop was noted. However, the power of the methods markedly decreased when a subset of controls was shared ($\eta < 1$), which we term as the *partial overlap design*. This pattern of power drop suggested that the power was reduced due to the mixing of shared subjects and unshared subjects in the analysis. The power drop was more dramatic when a small portion was shared (η close to 0) as compared to that when a small portion was unshared (η close to 1). The power drop was the most severe at approximately $\eta = 0.16$ regardless of the method used, wherein the power of the method was less than half that of the corresponding splitting approaches. In this simulation, the decoupling method had the same power as the LS method, because we assumed the application of the fixed effects model after decoupling in which case the two methods are equivalent (Han *et al.*, 2016). Note that a relative

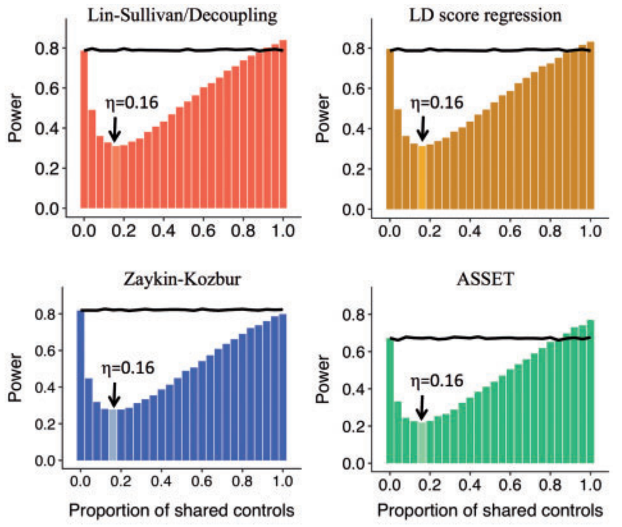

Fig. 2. Power of the existing meta-analysis methods for differing proportions of shared controls. We assumed a meta-analysis of five studies and varied the proportion of shared controls among all the controls within each study (η). The black lines denote the power of splitting, a strategy that splits overlapping samples before applying each meta-analysis method. Arrows indicate at which η the power was minimized. The LS method and decoupling method were analytically equivalent, because we applied the fixed effects model after decoupling (Color version of this figure is available at *Bioinformatics* online.)

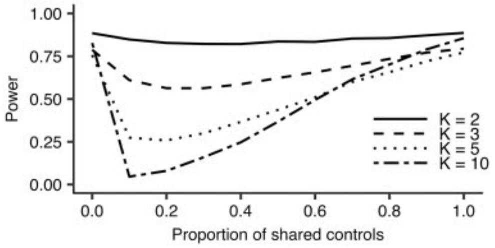

Fig. 3. Power of the LS method for differing numbers of studies in a meta-analysis. We varied the number of studies (K) as well as the proportion of shared controls among all the controls within each study (η). We adjusted the relative risk between 1.14 and 1.18 to maintain a similar power for differing K

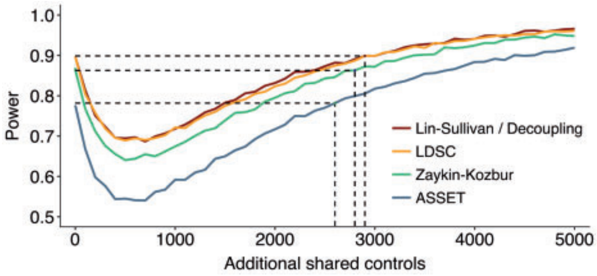

Fig. 4. Power of the existing meta-analysis methods after adding shared controls to independent studies. We plotted the power of the existing methods after adding shared controls to five independent studies, each consisting of 1000 cases and 1000 controls prior to the addition. We assumed that the shared controls were shared by all five studies. Dashed lines denote the shared control size that recovered the original power of the methods measured before adding shared controls (Color version of this figure is available at *Bioinformatics* online.)

power comparison between these methods is not of our interest in this simulation. ASSET showed the lowest power because we assumed a fixed effect size across the studies, as ASSET is designed to detect heterogeneous effects. Interestingly, in the full overlap design ($\eta = 1$), the powers of the existing methods were often slightly higher than those of their corresponding splitting approaches. For example, the LS method achieved 82% power with splitting and 84% without splitting. Further investigations are necessary to confirm the cause and extent of this observation.

The power reduction was also related to the number of studies that shared controls in the meta-analysis. Figure 3 shows the power of the LS method, as we varied the number of studies from 2 to 10. In this simulation, we adjusted the relative risks between 1.14 and 1.18 to maintain similar power for differing numbers of studies. The power reduction became more severe as more studies shared controls. This was possibly because, as more studies shared controls, the difference between the shared subjects and unshared subjects in terms of their contribution to the final statistic increased.

3.2 Adding shared controls can reduce the power of existing methods

Due to the power loss of the existing methods under the partial overlap design, these methods showed a counter-intuitive property that the addition of shared controls to the independent studies decreased the power, even though the total sample size increased. In this simulation, we assumed five independent studies ($n^+ = 1000$ and $n^- = 1000$). We also assumed a SNP with a minor allele frequency of 0.3 and a relative risk of 1.22. We added n_A shared controls that were assumed to be shared by all five studies, where we gradually increased n_A from 100 to 5000. We observed that as we added shared controls, the meta-analysis power dropped (Fig. 4). For example, the LS method had 86% power without shared controls, but showed 69% power with 500 shared controls and 72% power with 1000 shared controls. When we added even more shared controls, the power slowly recovered. The LS method recovered the original power after adding approximately one half of the total number of original controls ($n_A = 2900$). These results demonstrated that the use of the methods to correct for overlapping samples can reduce power even below the level of an analysis based on data in which all overlapping samples are discarded.

3.3 FOLD increases power

Using simulations, we first measured the false positive rate of FOLD, which was well controlled (Supplementary Figs S1 and S2). We then simulated the partial overlap design, where η varied from 0 to 1 similar to Figure 2. FOLD maintained identical power to the splitting approach regardless of η or the number of studies in the simulated meta-analysis (Fig. 5).

3.4 WTCCC data analysis

We compared the splitting approach, the LS method, and FOLD using Wellcome Trust Case Control Consortium (WTCCC) data (Wellcome Trust Case Control Consortium, 2007). In this analysis, splitting refers to a method that applies the standard fixed effects model (inverse-variance weighted average) after splitting. The three methods were in the following relationship. (i) With no overlap ($\eta = 0$), the three methods were equivalent. (ii) With full overlap ($\eta = 1$), LS and FOLD were equivalent. We assumed a cross-disease analysis combining the results for the three autoimmune diseases under the fixed effects model: Crohn's disease (CD), rheumatoid arthritis (RA) and type 1 diabetes (T1D). We focused on

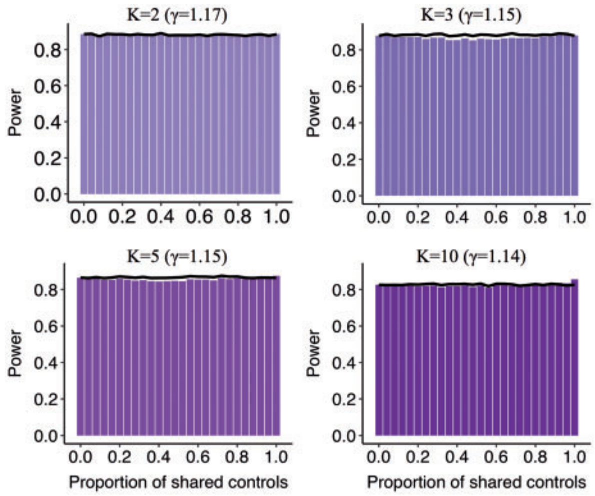

Fig. 5. Power of FOLD. We measured the power of our proposed method FOLD, as we varied the number of studies in the simulated meta-analysis (*K*) as well as the proportion of shared controls among all the controls within each study (η). The black lines denote the power of the splitting approach (Color version of this figure is available at *Bioinformatics* online.)

eight candidate pleiotropic loci: two SNPs from the combined analysis results of the original WTCCC study, excluding MHC and *PTPN22*, and six additional loci from ImmunoBase (http://www.immunobase.org). Refer to Section S4 of Supplementary Text for details of the analysis. In the original WTCCC study design, all controls were shared among diseases. We modified the study design and established a partial overlap design by using some controls as disease-specific and some as shared, such that η was approximately 0.5. When we considered 433,901 SNPs, excluding the MHC and *PTPN22* region, the QQ-plot did not show inflation for any of the methods (Supplementary Fig. S3). The genomic control factors were 1.00 for splitting, 1.00 for the LS method, and 1.03 for FOLD.

In this partial overlap design, at the eight candidate pleiotropic loci, the LS method attenuated the statistical significances of splitting at seven loci (Supplementary Table S1). The average $\log_{10}P$ difference was –0.79; thus, the LS method provided nearly one order of magnitude larger *P*-values. In contrast, FOLD yielded nearly identical results to splitting (Supplementary Table S1). The average $\log_{10}P$ difference between FOLD and splitting was close to zero (-0.004). As a result, the association *P*-values were notably smaller in FOLD than in the LS method (Fig. 6A).

We also established the full overlap design ($\eta = 1$), where we used all the controls for each disease and then combined the statistics using the LS method or FOLD. The LS method and FOLD were analytically equivalent in this design. FOLD (= LS) showed slightly smaller *P*-values than splitting at the eight loci (Supplementary Table S2), consistent with our simulation results where FOLD (= LS) showed slightly higher power than splitting at $\eta = 1$. The average difference in $\log_{10}P$ was 0.44 between FOLD and splitting.

3.5 PGC data analysis

Using the Psychiatric Genomics Consortium (PGC) data (Cross-Disorder Group of the Psychiatric Genomics, 2013), we simulated cross-disease meta-analysis combining five psychiatric disorders (autism spectrum disorders, attention deficit-hyperactivity disorder, bipolar disorder, major depressive disorder, and schizophrenia). We downloaded the summary statistics of the original meta-analysis, which came from studies without sample overlap. We sought to simulate a design that augments additional shared controls to these data. Hence, for each tested SNP, we approximately reconstructed a 2×2 allele count table from the reported summary data (Section S4 of Supplementary Text). We then randomly generated additional $n_A = 2000$ shared controls and augmented them to the analysis. We examined 11 candidate pleiotropic loci that showed *P*-values smaller than 10^{-6} in the original meta-analysis (Cross-Disorder Group of the Psychiatric Genomics, 2013).

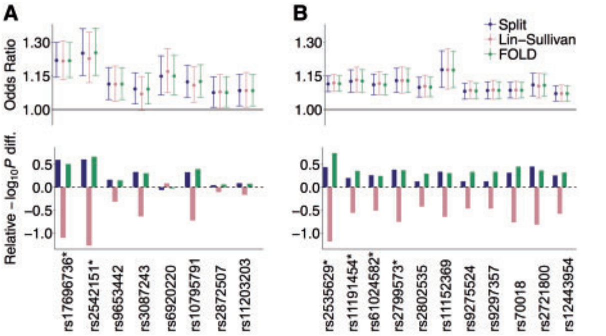

Fig. 6. Cross-disease meta-analysis results of WTCCC and PGC data. We performed cross-disease meta-analysis while accounting for overlapping subjects using WTCCC data **(A)** and PGC data **(B)**. In WTCCC data, we examined two loci reported by WTCCC (denoted with *) and six additional pleiotropic loci obtained from ImmunoBase. In PGC data, we examined four reported loci by PGC (denoted with *) and seven additional loci satisfying $P < 1 \times 10^{-6}$. The top panel shows the odds ratios and the 95% CIs. The bottom panel shows the difference between three methods in terms of statistical significances ($-\log_{10}P$), where zero is calibrated to the mean value of the three methods (Color version of this figure is available at *Bioinformatics* online.)

Supplementary Table S3 shows that under this partial overlap design, the LS method attenuated the statistical significance of splitting at the 11 loci, with the average $\log_{10}P$ difference being -0.72. In contrast, FOLD yielded nearly identical results to splitting, with the average $\log_{10}P$ difference being 0.05. As a result, the association *P*-values were notably smaller in FOLD than in LS (Fig. 6B).

3.6 Splitting strategy comparison

In addition to the FOLD approach, we also proposed the FOLD-split method to facilitate splitting. We compared its performance to two other splitting strategies: (i) equal splitting that equally distributes the shared controls to studies, and (ii) case-based splitting that distributes the shared controls proportionally according to the case sample sizes. We assumed a meta-analysis of five studies and a SNP with a relative risk of 1.16. We randomly sampled the case sample size from Uniform (1, 1000), Uniform (1, 2000), Uniform (1, 3000), Uniform (1, 4000) and Uniform (1, 5000) for the five studies respectively, where Uniform (x,y) refers to the uniform distribution between x and y. We also sampled the study-specific control sizes from Uniform (1, 3000) and assumed 1000 shared controls ($n_A = 1000$). Our goal was to reflect real situations with varying sample sizes. Given the sample sizes, we simulated genotypes and applied different splitting strategies to measure their power. Finally, we gradually increased n_A from 1000 to 5000. Supplementary Figure S4A shows that FOLD-split achieved the best power among all the splitting approaches, followed by case-based splitting and equal splitting. For example, at $n_A = 3000$, the power of FOLD-split was 78%, whereas the powers of case-based splitting and equal splitting were 73% and 66%, respectively.

We then sought to examine the difference between the splitting results of FOLD-split and those of other approaches under specific situations. We assumed a meta-analysis combining four studies, wherein the case/control sample sizes were 4000/5000, 5000/3500, 2500/1000 and 2500/500, respectively. We aimed to distribute 10 000 shared controls. Equal splitting equally distributed them as 2500 controls per study. Case-based splitting distributed the shared controls proportionally to the case size, as 2857, 3571, 1786 and 1786 controls to the four studies, respectively. FOLD-split assigned 714, 3643, 2572 and 3071 controls to the four studies, respectively (Supplementary Fig. S4B). When we assumed a relative risk of 1.08, the powers were estimated as 75.9%, 76.2% and 77.4% for equal splitting, case-based splitting, and FOLD-split, respectively.

4 Discussion

In this article, we identified and reported a phenomenon wherein existing meta-analysis methods for overlapping subjects experience a markedly reduced power compared with the traditional splitting method. To recover this loss of power, we proposed a solution termed as FOLD which categorizes samples based on the sharing of the subjects and calculates multiple statistics. To combine the multiple statistics in FOLD, we employed the LS method, although the use of other frameworks was also suitable (Supplementary Fig. S5). Moreover, we described the FOLD-split method to determine the optimal splitting design for conditions where splitting is possible. To our knowledge, the power decrease of existing methods has not been previously reported. We found that the original developers of the existing methods only evaluated the power under the full overlap design in their simulations (Bhattacharjee *et al.*, 2012; Bulik-Sullivan *et al.*, 2015; Han *et al.*, 2016; Lin and Sullivan, 2009; Zaykin and Kozbur, 2010), which is possibly why this phenomenon was overlooked.

Overlapping subjects are a particularly important issue in cross-disease meta-analyses, which is a recently developed study design that combines multiple diseases. Since it is possible to reuse controls for multiple diseases, the issue of overlapping subjects can easily occur. Recently, Moskvina *et al.* (2013) used the LS method to combine Alzheimer's disease and Parkinson's disease, Dichgans *et al.* (2014) used the Zaykin-Kozbur method to combine ischemic stroke and coronary artery disease, and Kar *et al.* (2016) used the decoupling method to combine three types of cancer. These studies were in the partial overlap designs, which suggested that the use of splitting or FOLD could have possibly enhanced the statistical power.

In this study, we focused on the fixed effects (FE) model that assumes a constant effect size between studies, similar to previous studies (Lin and Sullivan, 2009; Zaykin and Kozbur, 2010). Recently, the application of random effects (RE) model was shown to be powerful when the effect size heterogeneity existed (Han and Eskin, 2011). In cross-disease analysis, the use of RE could be suitable because we can expect heterogeneity. To employ RE in FOLD, we can utilize RE implementation that can account for correlations (Lee *et al.*, 2017). However, for optimal performance, we will need a flexible RE framework that can account for the fact that the statistics from the same study ($\widehat{\beta}_{B,\text{Spe}}$ and $\widehat{\beta}_{B,\text{Share}}$ in Fig. 1) do not exhibit heterogeneity. We expect that we will be able to fully utilize RE in FOLD as the RE framework becomes more flexible in the future.

We compared the performance of FOLD to existing methods, and the required information for each method differs. LDSC does not require any information of shared subjects prior to the analysis, and can be the most convenient if we do not have that information. However, LDSC cannot be easily applied to data with small sample size, population without appropriate reference panel, or genotyping platforms with uneven density (Bulik-Sullivan *et al.*, 2015). All other methods require the information of the numbers of shared subjects to calculate the correlation matrix. Our method FOLD additionally requires multiple statistics from each study, calculated based on the sharing of the subjects. However, this can occasionally be extremely difficult, particularly in a retrospective meta-analysis design, and may require contacting the original investigators. Hence, applying FOLD may be as difficult as splitting. If FOLD and splitting have the same difficulty, the use of splitting has an advantage that the statistics become independent and therefore can be straightforwardly used in future analyses. However, the use of FOLD can be preferred if the researchers want deterministic results, because the results of splitting can change depending on how the individuals are randomly split. Despite the difficulties of obtaining additional information, we emphasize that the use of FOLD or splitting to prevent power loss may be worthwhile, given the gain in power achieved. Another challenge is to identify the hidden duplicate individuals between studies without sharing genotype data, which can be performed using recently described methods (He *et al.*, 2014; Hormozdiari *et al.*, 2014).

Funding

This research was supported by the National Research Foundation of Korea (NRF) grant (grant number 2016R1C1B2013126) and the Bio & Medical Technology Development Program of the NRF (grant number 2017M3A9B6061852) funded by the Korean government, Ministry of Science and ICT (MSIT).

Conflict of Interest: none declared.

References

Bhattacharjee,S. *et al.* (2012) A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am. J. Hum. Genet.*, **90**, 821–835.

Bulik-Sullivan,B.K. *et al.* (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, **47**, 291–295.

Chubb,D. *et al.* (2013) Common variation at 3q26.2, 6p21.33, 17p11.2 and 22q13.1 influences multiple myeloma risk. *Nat. Genet.*, **45**, 1221–1225.

Cross-Disorder Group of the Psychiatric Genomics Consortium (2013) Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet*, **381**, 1371–1379.

Crowther-Swanepoel,D. *et al.* (2009) Genetic variation in CXCR4 and risk of chronic lymphocytic leukemia. *Blood*, **114**, 4843–4846.

Di Bernardo,M.C. *et al.* (2008) A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. *Nat. Genet.*, **40**, 1204–1210.

Dichgans,M. *et al.* (2014) Shared genetic susceptibility to ischemic stroke and coronary artery disease: a genome-wide analysis of common variants. *Stroke*, **45**, 24–36.

Foulley,J.L. and Quaas,R.L. (1995) Heterogeneous variances in Gaussian linear mixed model. *Genet. Sel. Evol.*, **27**, 211–228.

Ghalanos,A. and Theussel,S. (2015), *Rsolnp: General Non-linear Optimization Using Augmented Langrange Multiplier Method*. R package version 1.16.

Han,B. *et al.* (2016) A general framework for meta-analyzing dependent studies with overlapping subjects in association mapping. *Hum. Mol. Genet.*, **25**, 1857–1866.

Han,B. and Eskin,E. (2011) Random-effects model aimed at discovering associations in meta-analysis of genome wide association studies. *Am. J. Hum. Genet.*, **88**, 586–598.

He,D. *et al.* (2014) Identifying genetic relatives without compromising privacy. *Genome Res.*, **24**, 664–672.

Hormozdiari,F. *et al.* (2014) Privacy preserving protocol for detecting genetic relatives using rare variants. *Bioinformatics*, **30**, 204–211.

Kar,S.P. *et al*. (2016) Genome-wide meta-analyses of breast, ovarian, and prostate cancer association studies identify multiple new susceptibility loci shared by at least two cancer types, *Cancer Discov*., **6**, 1052–1067.

Kilpivaara,M. *et al*. (2009) A Germline Jak2 Snp Is Associated with Predisposition to the Development of Jak2 V617f-Positive Myeloproliferative Neoplasms. *Haematol. Hematol. J*., **94**, 420–420.

Lee,C.H. *et al*. (2017) Increasing the power of meta-analysis of genome-wide association studies to detect heterogeneous effects. *Bioinformatics*, **33**, i379–i388.

Lin,D.Y. and Sullivan,P.F. (2009) Meta-analysis of genome-wide association studies with overlapping subjects. *Am. J. Hum. Genet*., **85**, 862–872.

Moskvina,V. *et al*. (2013) Analysis of genome-wide association studies of Alzheimer disease and of Parkinson disease to determine if these 2 diseases share a common genetic risk. *Jama Neurol*., **70**, 1268–1276.

Mukherjee,S. *et al*. (2011) Including additional controls from public databases improves the power of a genome-wide association study. *Hum. Hered*, **72**, 21–34.

Onengut-Gumuscu,S. *et al*. (2015) Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet*., **47**, 381–386.

Orozco,G. *et al*. (2014) Novel Rheumatoid Arthritis Susceptibility Locus at 22q12 Identified in an Extended UK Genome-Wide Association Study. *Arthritis & Rheumatology*, **66**, 24–30.

Rao,J.N.K. (1973) On the estimation of heteroscedastic variances. *Biometrics*, **29**, 11–24.

Shete,S. *et al*. (2009) Genome-wide association study identifies five susceptibility loci for glioma. *Nat. Genet*., **41**, 899–904.

Speedy,H.E. *et al*. (2014) A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia. *Nat. Genet*., **46**, 56–60.

Weinhold,N. *et al*. (2013) The CCND1 c.870G>A polymorphism is a risk factor for t(11;14)(q13;q32) multiple myeloma. *Nat. Genet*., **45**, 522–525.

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**, 668–678.

Yin,J. *et al*. (2011) Kriging metamodel with modified nugget-effect: the heteroscedastic variance case. *Comput. Ind. Eng*., **61**, 760–777.

Zaykin,D.V. and Kozbur,D.O. (2010) P-value based analysis for shared controls design in genome-wide association studies. *Genet. Epidemiol*., **34**, 725–738.

Zhao,S. *et al*. (2007) Simple focal-length measurement technique with a circulat Dammann grating. *Appl. Opt*., **46**, 44–49.