이학박사 학위논문

Pitfalls of using shared controls in meta-analysis of genetic association studies

메타분석에서 표본중첩 현상이 존재 할 경우의 문제점과 해결책

2018년 8월

서울대학교 대학원

화학부 물리화학전공

김 은 지

Abstract

Pitfalls of using shared controls in meta-analysis of genetic association studies

Eunji Kim
Physical Chemistry, Department of Chemistry
The Graduate School
Seoul National University

Recent genome-wide association studies frequently augment sample size by using publicly available shared controls. Nowadays, many disease consortia or groups of investigators are combining their efforts to understand the genetic basis of diseases by collecting summary statistics from participating studies and performing a meta-analysis. However, if participating studies share publicly available controls, those can induce correlations between statistics, which prohibit the use of the standard meta-analysis methods. Fortunately, recently developed meta-analysis methods can systematically account for these correlations and are widely used in single-disease or cross-disease meta-analyses. In this paper, we identify and report a phenomenon that using shared controls in multiple participating studies in a meta-analysis can dramatically reduce power even with the use of these systematic methods. We demonstrate that meta-analysis power decreases as we add shared controls to multiple studies in a meta-analysis, in contrast with the common expectation that additional samples

should increase power. We investigate why this phenomenon occurs and provide three possible solutions to prevent this power reduction: (1) removing sample overlaps to make summary statistics completely independent, (2) exclusively using shared controls and no study-specific controls, or (3) using our newly proposed method FOLD (Fully-powered method for OverLapping Data).

To increase samples participating in an analysis, meta-analysis combines summary statistics from multiple independent studies. If multiple studies participating in a meta-analysis utilize the same public dataset as controls, the summary statistics from these studies are no more independent and become correlated. Lin and Sullivan proposed the correlation estimator based on the shared and unshared sample sizes and suggested an optimal test statistic to account for the correlations (AJHG 2010). Their method was shown to achieve similar power to the gold standard method, splitting, which refers to the method that splits shared individuals into the studies prior to meta-analysis when we have access to the genotype data. Many different methods were proposed after Lin and Sullivan, but most of these methods were based on the similar correlation estimator. In this paper, we report a phenomenon that

the use of the standard method suggested by Lin and Sullivan can lead to unbalanced power for detecting protective alleles (OR < 1) and risk alleles (OR > 1). Specifically, when we assumed that the controls were shared, the power for detecting protective minor alleles (OR<1) were lower than the power for detecting risk minor alleles (OR>1). For example, for detecting a minor allele of frequency 10% and of OR=0.85, simulating meta-analysis of 5 studies showed that the standard method only achieved 61.6% power whereas splitting achieved 67.0%. By contrast, when we flipped the effect direction (OR=1.17), the existing method conversely achieved higher power (71.8%) than splitting. The degree of asymmetry was exacerbated as the minor allele frequency (MAF) decreased. To our knowledge, we are the first to report this phenomenon. After investigating on this phenomenon, we identified that the power asymmetry problem occurred because the standard correlation estimator did not exactly predict the true correlation. The existing estimator was approximated under the simple assumption of the null hypothesis of no effect, but under the alternative hypothesis, the true correlation is dependent on MAF and effect size. Thus, the errors in estimator could lead to substantially unbalanced power. To overcome the power asymmetry problem, we developed a method that uses an accurate correlation

estimator, called PASTRY (A method to avoid Power ASymmeTRY). Our method is based on the correlation estimator that was designed to be accurate under the alternative hypothesis. We show that using our method, one can effectively achieve symmetry on power for testing risk and protective alleles.

Keyword: Meta-analysis, genome-wide association studies (GWAS), overlapping samples, correlation, power asymmetry, case-control study.

Student Number: 2011-30903

Table of Contents

List of figures

Lists of tables

Chapter 1. Optimal strategy to account for overlapping controls in cross-disease meta-analysis of genetic association studies.

Introduction

Recent genome-wide association studies frequently augment sample size to increase power by using publicly available shared controls (1-11). Nowadays, many disease consortia or groups of investigators are combining their efforts to understand the genetic basis of diseases. Due to the privacy issue that forbids sharing of genotype data, the common practice is to collect summary statistics from participating studies and perform a meta-analysis. However, if participating studies share publicly available controls, those can induce correlations between statistics. These correlations prohibit the use of the standard meta-analysis methods that assume independency between statistics.

Fortunately, recent studies developed meta-analysis methods that systematically account for these correlations, which opened opportunities for flexible use of shared controls in meta-analysis design (12-15). These methods are widely used to

account for disease meta-analyses that aim to uncover pleiotropic loci (16-18). Moskvina et al. (18) used Lin-Sullivan method to combine Alzheimer's disease and Parkinson's disease, Dichgans et al. (16) used Zaykin-Kozbur method to combine ischemic stroke and coronary artery disease, and Kar et al. (17) used decoupling method to combine breast, ovarian, and prostate cancers.

In this paper, we identify and report a phenomenon that using shared controls for multiple participating studies in a meta-analysis can dramatically reduce power even with the use of these systematic methods. For example, if we augment 1,000 shared controls to five studies of 1,000 cases and 1,000 controls that participate in a meta-analysis, the meta-analysis power of a widely used method drops from 86% to 72% in contrast with the common expectation that additional samples should increase power. This power drop phenomenon was persistent for all existing systematic methods. We investigated the origin of this phenomenon and identified that the power is reduced because shared controls and unshared controls differ in terms of the information they contain. Simulations and real data analyses demonstrate that in order to prevent power reduction, three strategies are possible: (1) removing sample overlaps by splitting shared subjects into

individual studies to make summary statistics completely independent, (2) exclusively using shared controls and no study-specific controls, or (3) using our new summary-statistics-based method FOLD (Fully-powered method for OverLapping Data). The key idea of FOLD is to combine multiple statistics that are calculated using homogeneous samples in terms of their information. For investigators choosing the first option, we also provide a companion method PowerSplit that determines the optimal splitting design.

Results

1. Adding shared controls can reduce power of meta-analysis

In this simulation, we evaluated the power of the existing meta-analysis methods as we added shared controls to a meta-analysis. We assumed five independent studies (K=5) with 1,000 cases and 1,000 controls (N^+=1000 and N^-=1000). We assumed a SNP with minor allele frequency of 0.3 and relative risk of 1.22. We evaluated the power of Lin-Sullivan method (12), Zaykin-Kozbur method (15), ASSET (14), and decoupling method (13). All these methods were designed to systematically account for correlations induced by shared controls. We added N_A shared controls that were assumed to be shared by all five studies and gradually increased N_A from 100 to 5,000. Figure 1 shows that as we added shared controls, the meta-analysis power dropped, which was consistent for all methods. For example, Lin-Sullivan method showed 86% power without shared controls, but 69% power with 500 shared controls and 72% power with 1,000 shared controls. As we added even more shared controls, the power began to slowly recover. Lin-Sullivan method recovered the original power after adding a half of the total number of original controls (N_A=2,900). Decoupling method had the same power to

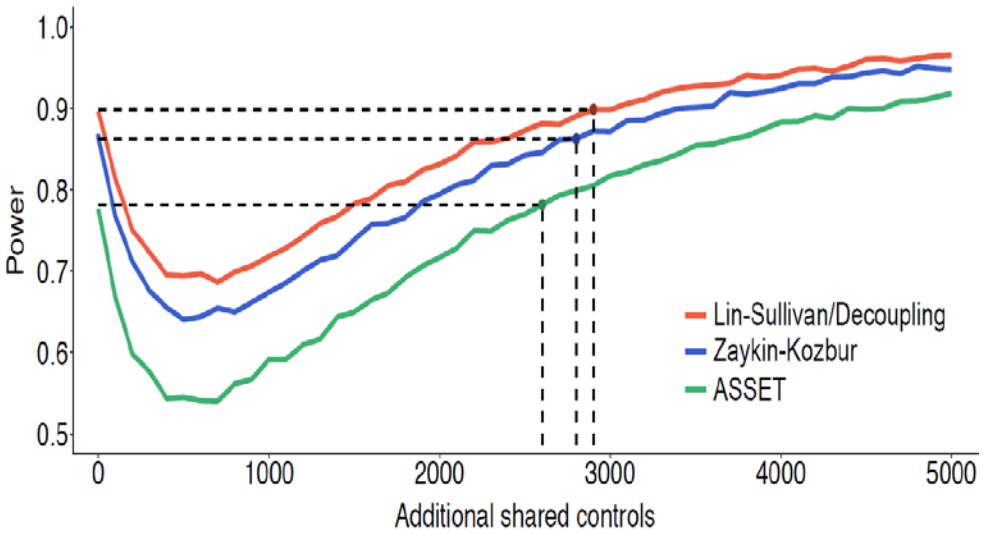


Figure 1. Powers of existing meta-analysis methods after adding shared controls to independent studies.

We plotted powers of methods after adding shared controls to five independent studies each consisting of 1000 cases and 1000 controls. We assumed that shared controls were shared by all five studies. We assumed a relative risk of 1.22. Dashed lines denote the numbers of additional shared controls that recovered the original power of the methods measured before adding shared controls. Lin-Sullivan method and decoupling method were analytically equivalent, because we applied fixed effects model after decoupling.

Lin−Sullivan, because we assumed the application of fixed effects model after decoupling, which is equivalent to Lin−Sullivan (13). Note that a relative power comparison between these methods is not of our interest in this simulation; ASSET shows the lowest power only because we assumed a fixed effect size across the studies while ASSET is designed to detect heterogeneous effects.

2. Power loss occurs in partial overlap design

This phenomenon that meta−analysis power drops as we add shared controls is related to the proportion of shared controls (η) among all controls within each study. To demonstrate this, we simulated five studies (N^+=6,000 and N^-=6000). We assumed a SNP with relative risk of 1.16. Beginning from no overlap (η=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 total number of controls in meta−analysis. This way, the power of splitting approach was maintained at the same level regardless of η. Here splitting approach refers to a strategy that splits genotype data of shared controls into individual studies to remove sample overlap. Splitting approach for Lin−Sullivan method means that we apply Lin−Sullivan method after splitting, and splitting approach for each of the other methods is similarly defined.

Figure 2 shows that when all controls were shared ($\eta = 1$), which we call *full overlap design*, power drop did not occur. We conjecture that this is because all controls are shared and therefore they are homogeneous in terms of their information. The powers of the methods dropped compared to their corresponding splitting approaches when a subset of controls was shared ($\eta < 1$), which we call *partial overlap design*. The power drop was more dramatic when a small portion is shared (η close to 0) than when a small portion was unshared (η close to 1). Interestingly, in the full overlap design ($\eta = 1$), the powers of existing methods were sometimes slightly higher than their corresponding splitting approaches. This difference was notable in ASSET, which achieved 68% power with splitting and 77% without splitting. This is possibly because as long as there is no heterogeneity issue ($\eta = 1$), comparing the whole controls to each case group can be more effective than splitting controls. Further investigations are necessary to confirm the cause and extent of this observation.

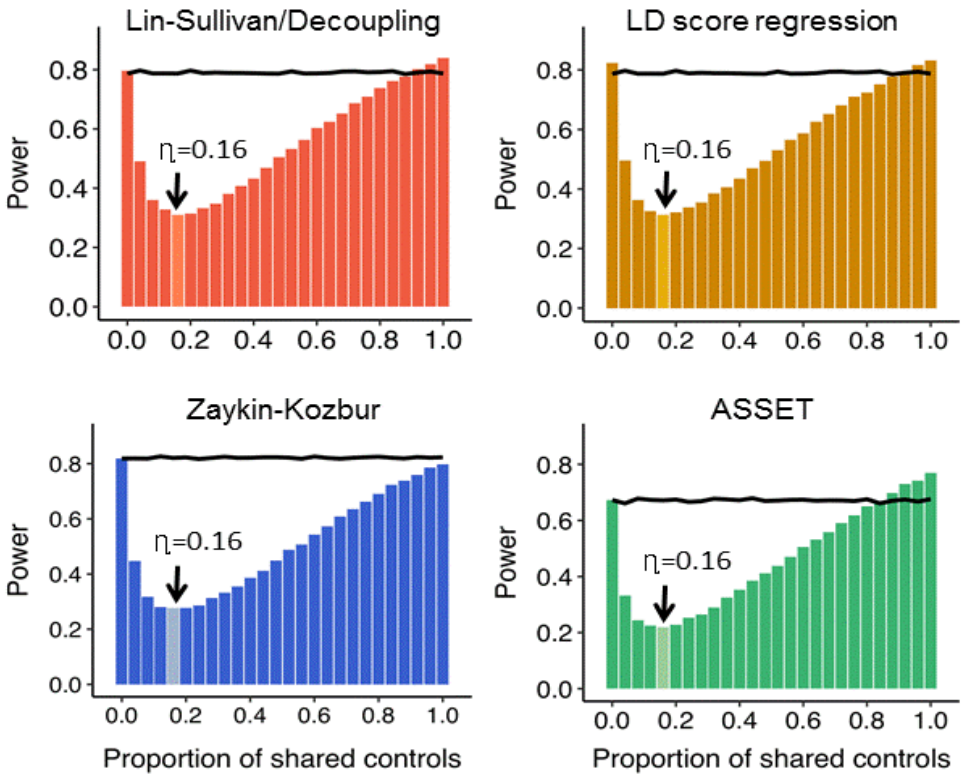


Figure 2. Powers of existing meta-analysis methods for differing proportions of shared controls. We assumed a meta-analysis of five studies and varied the proportion of shared controls among all controls within each study (η). The black lines denote the power of splitting, which refers to a strategy that splits overlapping samples, obtains independent summary statistics, and uses them in the corresponding method. We kept the total number of controls constant in order to maintain the power of splitting. Arrows indicate at which η the power was minimized. Lin-Sullivan method and decoupling method were analytically equivalent, because we applied fixed effects model after decoupling.

The power drop was also related to the number of studies that shared controls in a meta-analysis. Figure 3 shows the power of Lin-Sullivan method as we varied the number of studies from 2 to 10. To maintain similar power, we adjusted relative risk between 1.14 and 1.18. The power reduction became more severe as more studies shared controls. This is because as more studies share controls, the difference between the shared controls and unshared controls in terms of their information becomes greater.

3. FOLD maintains power

Our proposed method FOLD prevents power loss by combining a set of summary statistics that are calculated by using homogeneous subjects in terms of their information. Using the similar simulation framework, we first measured the false positive rate of FOLD, which was well controlled (Figure 4). Then we simulated partial overlap design varying η from 0 to 1. FOLD maintained a near identical power to the splitting approach regardless of η or the number of studies in the simulated meta-analysis (Figure 5).

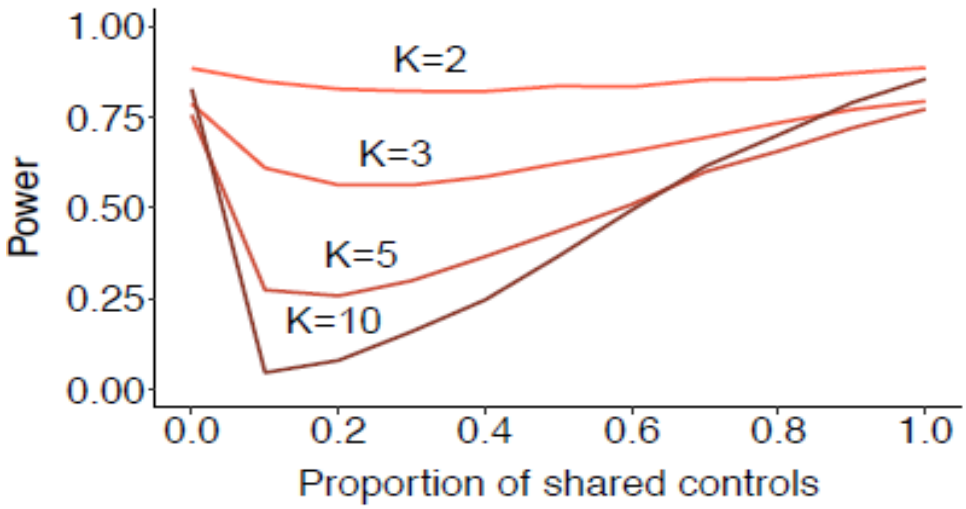


Figure 3. Powers of Lin-Sullivan method for differing numbers of studies in a meta-analysis.

We varied the number of studies in the simulated meta-analysis (K) as well as the proportion of shared controls among all controls within each study (η). We adjusted relative risks between 1.14 and 1.17 in order to maintain similar power for differing K.

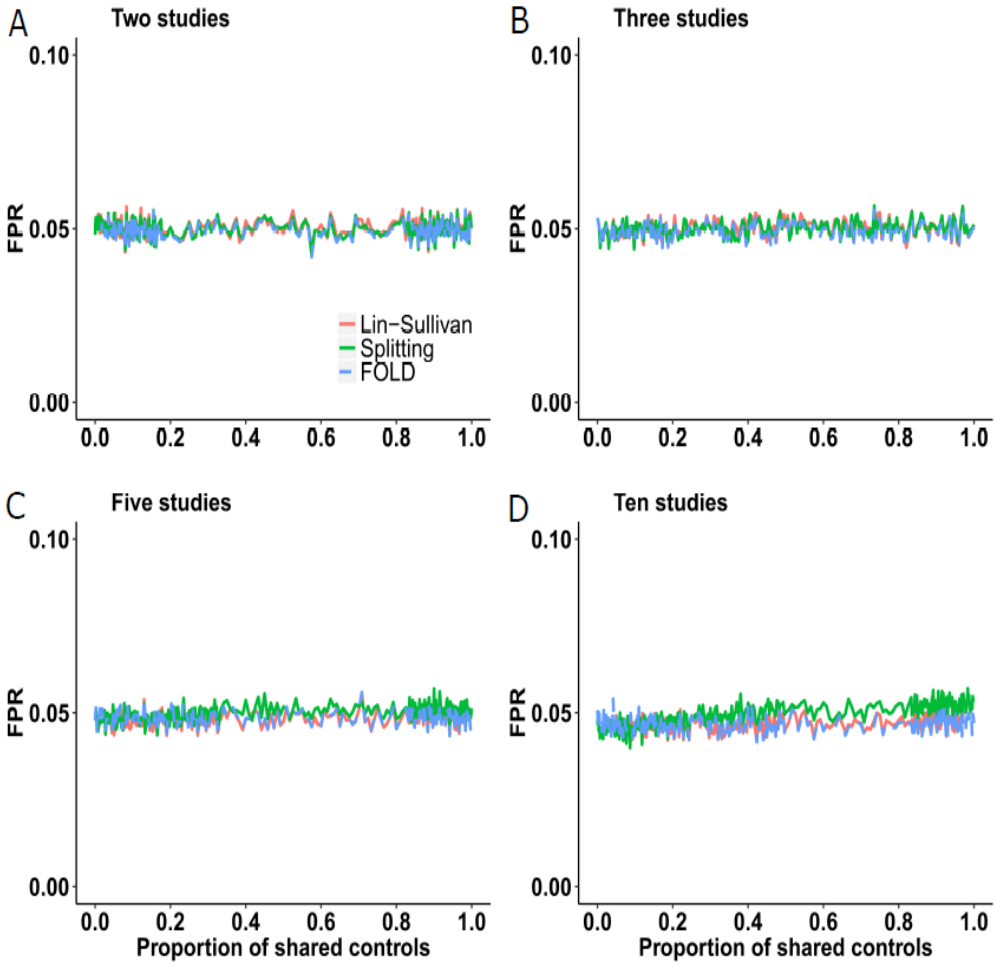


Figure 4. False positive rates of FOLD according to differing proportions of shared controls and numbers of studies in meta-analysis. The X-axis indicates the proportion of shared controls among all controls within each study and the Y-axis indicates the false-positive rate given the significance threshold $\alpha = 0.05$. We considered four different numbers of studies: two (A), three (B), five (C), and ten (D). We also added splitting and Lin-Sullivan method for comparison.

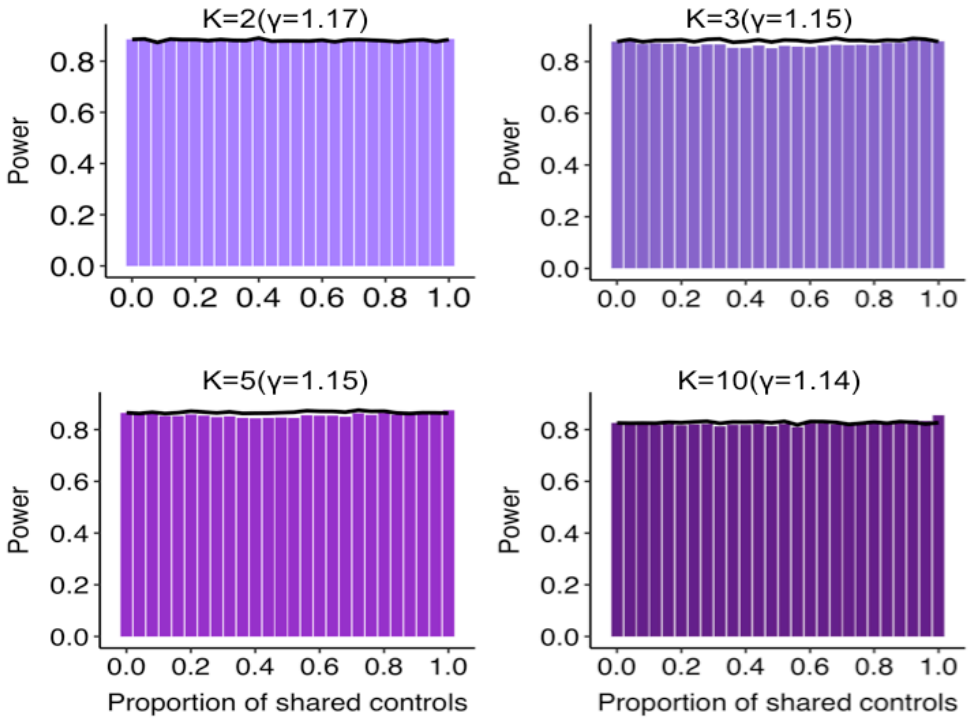


Figure 5. Powers of FOLD. We measured the power of our proposed method FOLD as we varied the number of studies in the simulated meta-analysis (K) as well as the proportion of shared controls among all controls within each study (η). The black lines denote the powers of splitting approach.

3.1. WTCCC data analysis

We compared splitting, Lin−Sullivan method (LS), and FOLD using the Wellcome Trust Case Control Consortium (WTCCC) data (5). Here, splitting refers to a method that applies the inverse−variance weighted average method after splitting. The three methods have the following relationship. (1) With no overlapping ($\eta=0$), three methods are equivalent. (2) With full overlapping ($\eta=1$), LS and FOLD are equivalent. We assumed a cross−disease analysis combining results for Crohn's disease (CD), rheumatoid arthritis (RA), and type 1 diabetes (T1D). We focused on eight candidate pleiotropic loci: two SNPs from the combined analysis result of the original WTCCC study excluding MHC and *PTPN22*, and six additional loci from the ImmunoBase (http://www.immunobase.org).

We first examined the full overlap design ($\eta=1$) by simply assuming that all controls were shared. LS and FOLD (which are equivalent under this design) showed comparable p−values to splitting at the eight loci (**Table 1**). The average difference in $\log_{10}$p was 0.441 between LS (=FOLD) and splitting. LS (=FOLD) gave slightly more significant p−values than splitting, consistently to our

CHR	Gene	SNP	Position	$p_{LS=FOLD}$	p_{Split}
12	*SH2B3*	rs1769673	110949538	$9.62X10^{-10}$	$1.16X10^{-9}$
18	*PTPN2*	rs2542151[a]	12769947	$4.20X10^{-8}$	$1.52X10^{-7}$
2	*AFF3*	rs9653442	100283885	$4.50X10^{-4}$	$9.82X10^{-4}$
2	*CD28,CTL*	rs3087243	204564425	$4.08X10^{-3}$	$6.62X10^{-3}$
6	*TNFAIP3*	rs6920220	138048197	$8.88X10^{-5}$	$3.51X10^{-4}$
10	*IL2RA*	rs10795791	6148346	$1.57X10^{-5}$	$3.98X10^{-4}$
17	*IKZF3,ME*	rs2872507	35294289	$1.50X10^{-2}$	$2.33X10^{-2}$
21	*UBASH3A*	rs11203203	42709255	$8.62X10^{-3}$	$1.46X10^{-2}$

Table 1. Association results of a cross-disease meta-analysis combining CD, RA, and T1D from the WTCCC data assuming a full overlap design.

We designed the analysis so that all controls were shared by the three diseases. We examined the two loci reported by WTCCC, excluding MHC and PTPN22. We present the p-values of Lin-Sullivan method (denoted as $p_{LS=FOLD}$, because Lin-Sullivan method and FOLD are equivalent in this situation) and the p-values of splitting (p_{Split}). Splitting was conducted with FOLD-Split.

simulation results that showed slightly lower power of splitting at η=1.

We then examined the partial overlap design (η <1). We used some controls as disease-specific and some as shared so that η is approximately 0.5. When we considered 433,901 SNPs excluding MHC and *PTPN22* region, QQ-plot did not show inflation for all three methods (**Figure 6**). The genomic control factors were 1.00 for splitting, 1.00 for LS, and 1.03 for FOLD. At the eight candidate pleiotropic loci, LS attenuated the statistical significances of splitting at seven loci (**Table 2**). The average $\log_{10}$p difference was -0.79; thus, LS showed nearly one order of magnitude less significant p-values. In contrast, FOLD gave nearly identical results to splitting (**Table 2**). The average $\log_{10}$p difference between FOLD and splitting was close to zero (-0.004). As a result, the association p-values were more significant in FOLD than in LS (**Figure 7A**).

3.2. PGC data analysis

Using the Psychiatric Genomics Consortium (PGC) data (19), we simulated cross-disease meta-analysis combining five psychiatric

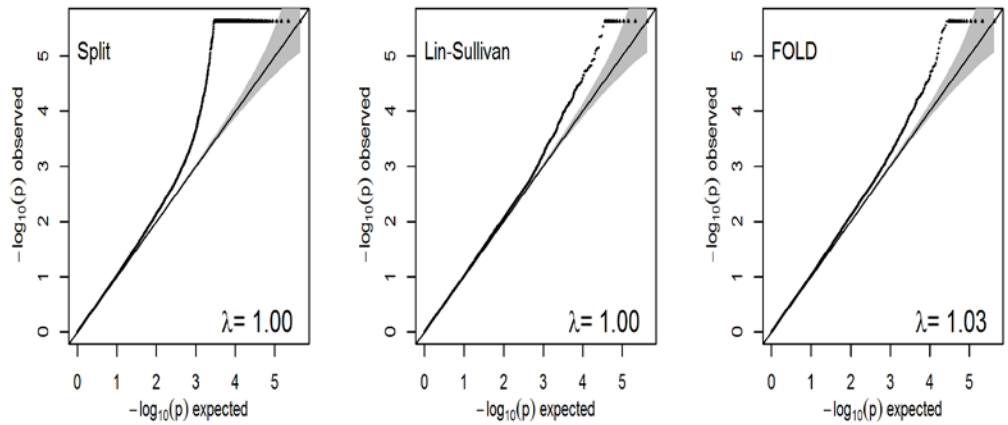


Figure 6. Quantile-quantile plots with splitting, Lin-Sullivan, and FOLD for WTCCC data.

We performed cross-disease meta-analysis of the CD, RA, and T1D datasets from WTCCC data, designing the analysis such that half of the controls in each disease were shared among the three diseases. We plotted 433,901 SNPs that passed quality controls, excluding the MHC and PTPN22 regions. The genomic control inflation factors (λ) were 1.00, 1.00, and 1.03 for splitting, Lin-Sullivan, and FOLD, respectively.

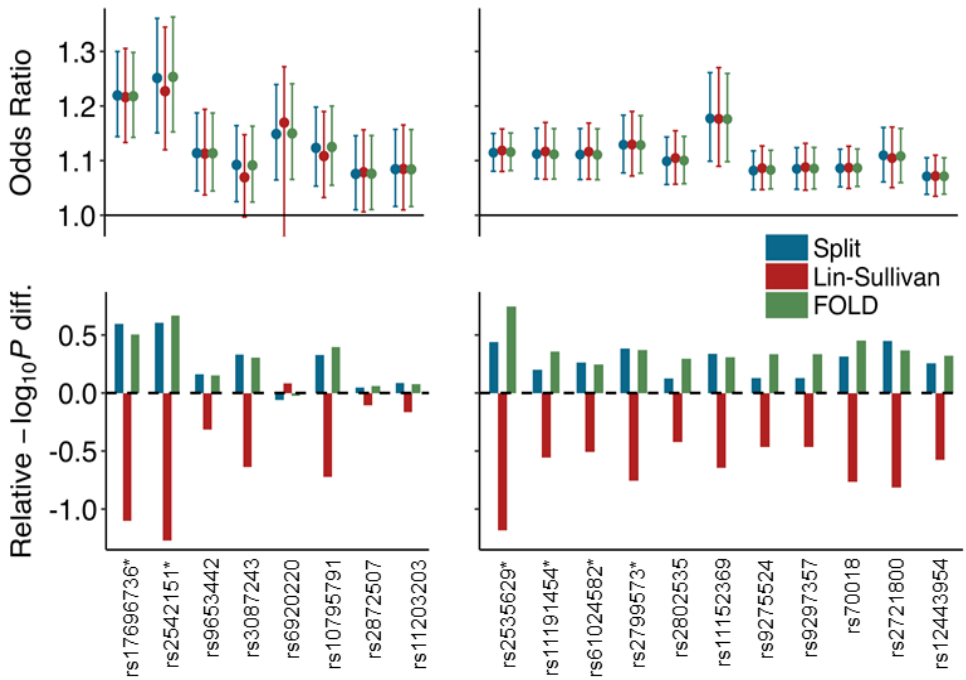


Figure 7. Cross-disease meta-analysis results in WTCCC and PGC data. (LHS) We performed cross-disease meta-analysis of CD, RA, and T1D using the WTCCC data, designing the analysis such that half of the controls in each study were shared among the three diseases ($\eta \approx 0.5$). We examined two loci reported by WTCCC (denoted with *) and six additional pleiotropic loci obtained from ImmunoBase. LS denotes Lin-Sullivan method. 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. (RHS) We performed cross-disease meta-

analysis of five psychiatric disorders using the PGC data. We examined four pleiotropic loci ($p<5\times10^{-8}$) reported by the original study (denoted with *) and seven additional loci that were moderately significant ($p<1\times10^{-6}$).

CHR	Gene	SNP	p_{Split}	p_{LS}	p_{FOLD}
12	*SH2B3*	rs17696736[a a]	$1.16X10^{-9}$	$5.82X10^{-8}$	$1.43X10^{-9}$
18	*PTPN2*	rs2542151[a]	$1.52X10^{-7}$	$1.15X10^{-5}$	$1.32X10^{-7}$
2	*AFF3*	rs9653442	$9.82X10^{-4}$	$2.95X10^{-3}$	$1.01X10^{-3}$
2	*CD28,CTLA*	rs3087243	$6.62X10^{-3}$	$6.17X10^{-2}$	$7.04X10^{-3}$
6	*TNFAIP3*	rs6920220	$3.51X10^{-4}$	$2.52X10^{-4}$	$3.25X10^{-4}$
10	*IL2RA*	rs10795791	$3.98X10^{-4}$	$4.50X10^{-3}$	$3.41X10^{-4}$
17	*IKZF3,MED1*	rs2872507	$2.33X10^{-2}$	$3.30X10^{-2}$	$2.26X10^{-2}$
21	*UBASH3A*	rs11203203	$1.46X10^{-2}$	$2.61X10^{-2}$	$1.50X10^{-2}$

Table 2. Cross-disease meta-analysis results combining CD, RA, and T1D from the WTCCC data. We performed cross-disease meta-analysis of the CD, RA, and T1D datasets from WTCCC data, designing the analysis such that half of the controls in each disease were shared by the three diseases. We examined two loci reported by WTCCC (denoted with [a]) and six additional pleiotropic loci obtained from ImmunoBase. We present the p-values of splitting (p_{Split}), Lin-Sullivan method (p_{LS}), and FOLD (p_{FOLD}). Splitting was conducted using PowerSplit.

disorders (autism spectrum disorders, attention deficit-hyperactivity disorder, bipolar disorder, major depressive disorder, and schizophrenia). We downloaded summary statistics of the original association results. The original analysis was based on independent studies without sample overlap. We wanted to simulate a design that augments additional shared controls to this data. To this end, for each tested SNP, we approximately reconstructed a 2 $\times$2 allele count table from the reported summary data (See *Material and methods* section). Then we randomly generated additional N_A=2000 shared controls and augmented them to the analysis. We examined the p-values of the eleven candidate pleiotropic loci that were at least moderately significant in the original meta-analysis ($p<10^{-6}$) (19).

Table 2 shows that under this partial overlap design, LS attenuated the statistical significances of splitting at the eleven loci, with the average $\log_{10}$p difference being -0.317. In contrast, FOLD gave nearly identical results to splitting, with the average $\log_{10}$p difference being 0.062. As a result, the association p-values were more significant in FOLD than in LS (Figure 7B).

4. Splitting strategy comparison

We proposed an additional method PowerSplit that facilitates splitting. We compared the performance of it to two other splitting strategies: (1) equal splitting that equally distributes shared controls to studies and (2) case-based splitting that distributes shared controls proportionally to the case sample sizes. We assumed a meta-analysis of five studies and a SNP with relative risk 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 used this simulation design to reflect real situations with varying sample sizes. We sampled the study-specific control size from Uniform(1,3000) and assumed 1,000 shared controls (N_A=1,000). After deciding sample sizes, we simulated genotypes and applied different splitting strategies to measure their power. Finally, we gradually increased N_A from 1,000 to 5,000. **Figure 7A** shows that PowerSplit achieved the best power among all splitting approaches, followed by case-based splitting and equal splitting. For example, at N_A=3,000, the power of PowerSplit was 78%, whereas the powers of case-based splitting and equal splitting

were 73% and 66% respectively.

We then wanted to examine how different the splitting result of PowerSplit is from the splitting results of the other approaches in a specific situation. We assumed a meta-analysis combining four studies, of which case/control sample sizes are 4,000/5,000, 5,000/3,500, 2,500/1,000, and 2,500/500 respectively. We assumed that we wanted to distribute 10,000 shared controls. Equal splitting equally distributed them, 2,500 controls per study. Case-based splitting distributed shared controls proportionally to the case size, 2,857, 3,571, 1,786 and 1,786 controls to the four studies respectively. PowerSplit assigned 714, 3,643, 2,572 and 3,071 controls to the four studies respectively (Figure 8B). When we assumed a relative risk of 1.08, the powers were 75.9%, 76.2% and 77.4% for equal splitting, case-based splitting, and PowerSplit respectively.

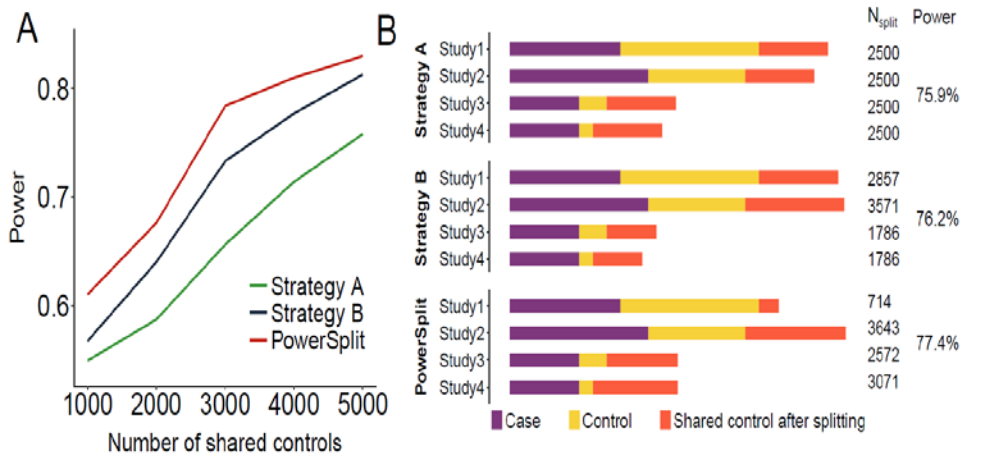


Figure 8. Comparison of three different splitting approaches. (A) Power of different methods. Strategy A is a strategy that equally distributes shared controls into studies (Equal-splitting). Strategy B is a strategy that distributes shared controls proportionally to the number of cases in each study (Case-based splitting). (B) Splitting results of the methods. Given four studies with differing sample sizes an10,000 shared controls, we distributed shared controls into the studies using each method. N_{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 in each study, and PowerSplit denotes our proposed splitting method based on the effective number of samples.

Materials and methods

1. Current approaches for meta-analysis with overlapping samples.

1.1 Splitting approach

A straightforward solution to account for overlapping subjects is to split the genotype data of shared samples into individual studies. Splitting exhibits two limitations. First, it may not be possible to access and split genotype data. This led to the development of the other methods introduced below that are based on summary statistics. Second, the optimal sample proportions for splitting are unknown. We address this issue in our framework.

1.2 Lin and Sullivan's method

Lin and Sullivan (12) were the first to introduce a method that can account for the correlation between statistics caused by overlapping samples. Lin-Sullivan method involves two main steps: (1) analytical approximation of correlations between statistics and (2)

construction of an optimal meta-analysis statistic that considers correlations.

Let X_1, …, X_M denote the observed effect sizes in M studies in the meta-analysis and let V_1, …, V_M be their variances. Lin and Sullivan derived a correlation formula:

$$r_{kl} \approx \frac{n_{kl-}\sqrt{\frac{n_{k+}n_{l+}}{n_{k-}n_{l-}}}+n_{kl+}\sqrt{\frac{n_{k-}n_{l-}}{n_{k+}n_{l+}}}}{\sqrt{n_k n_l}} \qquad (1)$$

where n_k, n_l, and n_{kl} are the total number of samples in the k^{th} and l^{th} studies and number of overlapping samples between the two studies (k and l), respectively. Subscripts + and - denote case and control subjects, respectively. Given the correlation matrix

$$C = [r_{ij}]_{M\times M}$$

one can calculate the covariance matrix, Ω.

Next, Lin and Sullivan proposed an optimal fixed-effects model meta-analysis statistic:

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

where e is an M × 1 vector with each component equal to 1. The variance of this statistic is

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

Therefore, one can obtain a z-score, $\frac{X_{LS}}{\sqrt{Var(X_{LS})}}$, as well as a p-value. When all studies are independent, Lin-Sullivan method becomes equivalent to the traditional inverse-variance-weighted average method.

1.3 Zaykin-Kozbur method

Independently from Lin-Sullivan method, Zaykin and Kozbur (1) also derived the correlation formulae for unsigned statistics as well as for signed statistics (e.g. z-score) for when samples overlap between studies. The correlation of signed statistics was equivalent to the derivation of Lin and Sullivan. The correlation of the unsigned statistic is,

$$\rho_{kl} = \left(\frac{1}{1+\frac{(n_{k-}-n_{kl-})}{n_{kl-}}}\right)\left(\frac{1}{1+\frac{(n_{l-}-n_{kl-})}{n_{kl-}}}\right)\left(\frac{1}{1+\frac{(n_{l-}-n_{kl-})}{n_{l+}}+\frac{n_{kl-}}{n_{l+}}}\right)\left(\frac{1}{1+\frac{(n_{k-}-n_{kl-})}{n_{k+}}+\frac{n_{kl-}}{n_{k+}}}\right)$$

where n_{k+}, n_{l+}, n_{k-}, n_{l-} and n_{kl-} are the number of cases, controls in the k^{th} and l^{th} studies, and number of overlapping control samples between the two studies (k and l), respectively.

Based on these results, Zaykin and Kozbur proposed several meta-analysis approaches. For example, suppose that the p-values are given and the direction of the effects is known. First, Zaykin and Kozbur applied inverse normal transformation to the p-values to obtain weighted z-scores

$$Z_k = w_k \Phi^{-1}(1 - q_k)$$

where q_k is defined as,

$$q_k = \begin{cases} p_k/2 & if\ effect > 0 \\ 1 - p_k/2 & otherwise. \end{cases}$$

where p_k is the two-sided p-value of the k^{th} study and "effect" refers to the direction of the effect size.

The weight for the k^{th} study is

$$w_k = \sqrt{N_k (R^{-1})_{kk}}$$

where R^{-1}_{kk} refers to the k^{th} diagonal of the inverse correlation matrix and N_k is the total number of samples in the k^{th} study. Next, they combine z-scores using

$$p^* = 1 - \Phi\left(\frac{\sum Z_k}{\sqrt{\sum W_k^2 + 2\sum_{k<l} W_k W_l R_{kl}}} \right)$$

where subscripts k and l refer to the k^{th} and l^{th} study. Finally, the combined two-sided p-value is

$$p^C = \begin{cases} 2p^* & if\ p^* < 0.5 \\ 2(1 - p^*) & otherwise \end{cases}$$

1.4 ASSET

Bhattacharjee et al.(2) proposed a meta-analysis method called ASSET. The focus of their study was to account for the possible presence and absence of effects (i.e. effect size heterogeneity between studies) to increase statistical power. To obtain high power when effects exist in only a subset of studies, ASSET iterates all possible subsets of studies, identifies the most significant subset, and corrects for multiple testing. To account for sample overlap in the meta-analysis, ASSET adapted Lin-Sullivan method into their framework. For every subset of studies, ASSET applies Lin-Sullivan method to the subset rather than applying the traditional fixed effects model. Additionally, Bhattacharjee et al. updated the correlation formula of Lin and Sullivan to a more general form that allows sharing of control individuals as case individuals in another study. Specifically, the updated formula is

$$\rho_{kl} = \sqrt{\frac{n_{k-}n_{k+}}{n_k}}\sqrt{\frac{n_{l-}n_{l+}}{n_l}}\left[\frac{n_{kl-}}{n_{k-}n_{l-}} - \frac{n_{k+l-}}{n_{k+}n_{l-}} - \frac{n_{k-l+}}{n_{k-}n_{l+}} + \frac{n_{kl+}}{n_{k+}n_{l+}}\right]$$

where n_k, n_{k+}, and n_{k-} are the total numbers of subjects, cases, and controls in the K^{th} study respectively, and where n_l, n_{l+}, and n_{l-} are similarly defined for the l^{th} study. n_{kl-} and n_{kl+} denote the numbers of

shared cases and shared controls between the k^{th} and l^{th} study, respectively. n_{k-l+} and n_{k+l-} represent the shared numbers between the case of the k^{th} study and the control of the l^{th} study, and the shared numbers between the control of the k^{th} study and the case of the l^{th} study, respectively.

1.5 Decoupling method

Han et al.(3) proposed a decoupling method that transforms the correlated data into independent data, which allows for the use of standard meta-analysis strategies to account for overlapping subjects. The main idea of their decoupling approach is to "uncorrelate" or "decouple" the studies so that studies obey the fundamental assumption of meta-analyzing methods, the independency between studies. As a result, the variance of a study increases as the correlations to other studies increase. They begin with the correlation matrix of Lin and Sullivan (R), and then transform the covariance structure of the data as follows:

$$\Omega_{decoupled} = Diag(e^T(Diag(s) \cdot R \cdot Diag(s))^{-1})^{-1}$$

where $Diag(\ )$ refers to a diagonal matrix whose diagonals are components inside the parentheses, s denotes the standard errors of the estimates, and e is a vector with each component equal to 1.

Next, they update the standard error of each study

$$s_{decoupled}[i] = \sqrt{\Omega_{decoupled}[i,i]}$$

i=1,⋯,K where K is the number of studies in the meta-analysis and $\Omega_{decoupled}[i,i]$ denotes the ith diagonal of the covariance matrix. Now that decoupled data can be considered independent, any standard meta-analysis methods can be applied such as the fixed effects model or random effects model. Han et al. proved that decoupling approach gives equivalent results to Lin-Sullivan method when the fixed effects model is applied to the decoupled data. Notably, decoupling translates non-zero correlations into added variances while giving the same results. This provides important insight into why existing approaches do not have optimal power in partial overlap designs, as described below.

2. FOLD (Fully-powered method for OverLapping Data)

The motivation for the development of our method came from the finding that the existing meta-analysis approaches designed for overlapping samples (12-15) reduced power in partial overlap designs. We demonstrate that this power drop is caused by

heterogeneity of information that each sample contains by using a toy example here. For simplicity, we assume normally distributed random variables. Suppose that we test whether the mean of the distribution that generates samples differs from zero. Consider two sets of samples, $\mathbf{x} = (x_1, \cdots, x_n)$ and $\mathbf{y} = (y_1, \cdots, y_n)$, that are to be combined. If all samples are identically and independently originated from $N(\beta, 1)$, we can combine all samples and obtain the average $\hat{\beta} = \frac{1}{2n}(\sum_i x_i + \sum_i y_i)$, where $\hat{\beta} \sim N\left(\beta, \frac{1}{2n}\right)$. Alternatively, we can obtain equivalent results by combining summary statistics; we can calculate $\widehat{\beta_1} = \frac{1}{n}\sum_i x_i$ and $\widehat{\beta_2} = \frac{1}{n}\sum_i y_i$ and meta-analyze these values as $\hat{\beta} = \frac{1}{2}(\widehat{\beta_1} + \widehat{\beta_2})$, which can be considered as the inverse-variance weighted average. Now, assume that $\mathbf{x}$ and $\mathbf{y}$ were correlated such that each pair of samples originated from a multivariate normal distribution, $(x_i, y_i) \sim MVN((\beta, \beta), R)$, where $R = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$ and r>0. Then, $\vec{\beta} = (\widehat{\beta_1}, \widehat{\beta_2}) \sim MVN((\beta, \beta), V)$, where $V = \frac{1}{n}R$. Analogous to Lin-Sullivan method, given $\bar{\beta}$, an optimal summary estimate would be $\hat{\beta} = \frac{\vec{\beta} V^{-1} e}{e^T V^{-1} e}$ (which is the same as $\frac{1}{2}(\widehat{\beta_1} + \widehat{\beta_2})$ for this simple example), where $\boldsymbol{e}$ is a vector of ones. The variance of $\hat{\beta}$ is $\frac{1}{e^T V^{-1} e}$, which is calculated to be $\frac{1+r}{2n}$. Thus, we can consider the additional variance $\frac{r}{2n}$ as a penalty for having correlations between samples. Another way of describing this is that the *2n* correlated samples contain less

information than *2n* independent samples with respect to the summary statistic, as if we have $\frac{2n}{1+r}$ independent samples. Note that this is exactly the intuition of the decoupling approach of Han *et al.* (13). The decoupling approach transforms $V = \frac{1}{n}\begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$ into a new V' whose non-diagonals become zero and whose diagonals increase, such that $V' = \frac{1}{n}\begin{bmatrix} 1+r & 0 \\ 0 & 1+r \end{bmatrix}$. As demonstrated by Han *et al.*, using decoupled variances for fixed-effects meta-analysis is equivalent to using Lin-Sullivan method. Thus, the decoupling approach can be thought of as translating the lower amount of information in dependent studies into increased variances.

Next, suppose that half of **x** and **y** are independent and the other half are correlated. That is, $(x_i, y_i) \sim MVN((\beta, \beta), I)$ for i = 1, …,n/2, and $(x_i, y_i) \sim MVN((\beta, \beta), R)$ for i = n/2+1, …,n. Then, $\vec{\beta} = (\widehat{\beta_1}, \widehat{\beta_2}) \sim MVN((\beta, \beta), V)$, where $V = \frac{1}{n}\begin{bmatrix} 1 & r/2 \\ r/2 & 1 \end{bmatrix}$. Again, application of Lin-Sullivan method gives a variance of $\frac{1+r/2}{2n}$. However, this approach may not be optimal in this situation because the independent samples contain more information than the dependent samples, as demonstrated by the decoupling approach described above. Thus, samples are heterogeneous in terms of their information. This situation is related to *heteroscedasticity (26)*,

which describes a situation in which samples have different variances. If the data for the dependent samples is transformed using the decoupling approach (13) into a dataset whose dependencies are removed, the samples that are dependent are expected to show larger variances. Thus, the two types of samples would be heteroscedastic with respect to the decoupled data if we remove dependencies. An optimal strategy in this situation would be to perform meta-analysis separately for independent samples (x_1, …, $x_{n/2}$, y_1, …, $y_{n/2}$) and dependent samples ($x_{n/2+1}$, …, x_{2n}, $y_{n/2+1}$, …, y_{2n}) and perform another meta-analysis on the two results. The variance for the meta-analysis summary of independent samples would be $\frac{1}{n}$ and the variance for the meta-analysis summary of dependent samples would be $\frac{1+r}{n}$. Thus, using the inverse-variance weighted average, the final variance would be $\frac{1+r}{n(2+r)}$. Note that the variance has now decreased compared to the naïve application of Lin-Sullivan 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 a suboptimal performance, explaining the power loss observed using existing approaches in partial overlap designs.

2.1 FOLD framework

We propose a meta-analysis framework for dealing with overlapping samples called FOLD. The main idea of FOLD is to obtain multiple summary statistics from each study, such that each statistic is calculated using homogeneous samples in terms of their information. To describe an example, suppose that we combine two case-control studies A and B (Fig. 9). Assume that a subset of controls is shared by A and B. We first calculate the log odds ratio by comparing the cases of A to the controls specific to A. This estimate is referred to as $\hat{\beta}_{A,Spe}$ (where "spe" denotes specific). We then calculate the log odds ratio by comparing the cases of A to the shared controls. This estimate is referred to as $\hat{\beta}_{A,Share}$. Similarly, we can obtain estimates from study B, $\hat{\beta}_{B,Spe}$ and $\hat{\beta}_{B,Share}$. Now, consider the vector of these four estimates ($\hat{\beta}_{A,Spe}$, $\hat{\beta}_{A,Share}$, $\hat{\beta}_{B,Spe}$, $\hat{\beta}_{B,Share}$). The correlation matrix of this vector is

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

r_{AB}' is the correlation between $\hat{\beta}_{A,Share}$ and $\hat{\beta}_{B,Share}$ driven by their shared controls. Because all controls, rather than a subset, are

shared between $\hat{\beta}_{A,Share}$ and $\hat{\beta}_{B,Share}$, heterogeneity is not present with respect to the information contained in each sample. r_A is the correlation between $\hat{\beta}_{A,Spe}$ and $\hat{\beta}_{A,Share}$ driven by the re-use of the study A cases in the calculation of these two statistics. r_B is similarly the correlation between $\hat{\beta}_{B,Spe}$ and $\hat{\beta}_{B,Share}$ driven by the re-use of the study B cases. Finally, we combine these four estimates while accounting for their correlation structure R (Fig. 9). We refer to this entire procedure as FOLD.

In practice, investigators commonly combine more than two studies in a meta-analysis. It is also possible that the controls are shared in a complicated manner; for example, for three 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. Classify controls into T groups where each group is homogeneous in how it is shared between studies. We call these T "configurations of sharing".
2. From each of the K studies, obtain T summary statistics, each calculated using each control group versus all cases.
3. Define the KT × KT correlation matrix between KT statistics using Eq. 1.
4. Apply Lin-Sullivan method to combine the KT statistics.

It is possible to generalize the procedure above to situations in which cases as well as controls are shared between studies. Note that although T can be as large as 2^K-1 in theory, it is much smaller in practice. Also, because not all configurations of sharing may occur in all studies, each study typically has fewer than T configurations.

2.2 FOLD gives smaller variance

Here, we analytically demonstrate that the FOLD effect size estimator achieves smaller variance compared to that of Lin-Sullivan method under the case-control study design in which studies share a subset of controls (partial overlap design). Let X_{ki} denote the explanatory variables (e.g., SNP genotype dosages) at the i^{th} subject in study k, and let $\theta_k = (\alpha_k, \beta_k)$ be the parameters of the logistic regression model for study k, where α_k and β_k denote the intercept and regression parameters, respectively. The corresponding "information matrix" (4) of study k is

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

where N_k denotes the sample size of study k. We assume the null

($\beta_k = 0$) to approximate the variance of the β_k estimator. Thus,

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

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 $\begin{bmatrix} 1 & X_{ki}^T \\ X_{ki} & X_{ki} X_{ki}^T \end{bmatrix}$. Assume that we consider just one explanatory variable (e.g., SNP). The variance of the β_k estimator is approximated using the inverse of this quantity, such that

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

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

$$\mathrm{var}(\hat{\beta}_k) \approx \frac{(1+\pi_k)^2}{N_k \pi_k} h$$

Now suppose that we have two studies, k and l. Let the subscripts kl− and kl+ indicate overlapping control and case samples,

respectively. Then, from Equation (1):

$$r_{kl} = \frac{N_{kl}\text{-}\sqrt{\pi_k \pi_l} + N_{kl}\text{+}\sqrt{\frac{1}{\pi_k \pi_l}}}{\sqrt{N_k N_l}}$$

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 η and 1−η denote the fractions of shared and study−specific subjects, respectively, which for simplicity we assume to be the same for cases and controls. That is, we assume that both a subset of cases and subset of controls are shared by the two studies, with the same fraction η. Then,

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

The inverse of the correlation matrix is

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

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

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

In contrast, FOLD divides the subjects of each study into groups (configurations) based on sharing with other studies. Because both cases and controls are shared by the two studies in our design, the

application of FOLD is equivalent to the following five-step procedure. (1) Obtain $\hat{\beta}_{k,shared}$ by comparing shared cases to shared controls in study k. (2) Obtain $\hat{\beta}_{k,spec}$ by comparing study k-specific cases to study k-specific controls. (3) Similarly, obtain $\hat{\beta}_{l,shared}$ and $\hat{\beta}_{l,spec}$. (4) Combine the two estimates based on shared subjects ($\hat{\beta}_{k,shared}$ and $\hat{\beta}_{l,shared}$) into $\hat{\beta}_{shared}$ while accounting for their correlation. Note that $\hat{\beta}_{k,shared}$ and $\hat{\beta}_{l,shared}$ 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 one). Thus, combining these estimates is equivalent to using only one of estimates ($\hat{\beta}_{shared}=\hat{\beta}_{k,shared}$). (5) Finally, combine $\hat{\beta}_{k,spec}$, $\hat{\beta}_{l,spec}$, and $\hat{\beta}_{shared}$, which are independent. This simple procedure of FOLD indicates the rationale for intentionally constructing a study design that shares both cases and controls for our example, because if the two studies shared only controls but not cases, the same cases would be used for both specific and shared controls. Thus, the four estimates would have three different correlations, similarly to our example in Fig. 9, making it more challenging to determine the analytical derivation of variance.

To obtain the variance of the FOLD estimator, note that

$$\mathrm{Var}(\hat{\beta}_{Shared}) = \frac{h(1+\pi)^2}{\eta N \pi} = \frac{C}{\eta N}$$

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

Thus, combining these three β estimates using the inverse-variance-weighted average gives us the final variance:

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

This is rational because there are $(2-\eta)$N unique individuals.

Comparing the variances of the FOLD and Lin-Sullivan estimators shows that

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

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.

3. PowerSplit

In addition to FOLD, we developed a companion approach PowerSplit to facilitate splitting. Han et al. (13) recently showed that how splitting is conducted can affect power. Lin and Sullivan (12) suggested splitting controls proportionally to the case sizes, which performs well in full overlap design. However, in real

situations, each study may contain study-specific controls of varying sizes. Moreover, controls can differ by which subset of studies they are shared. We developed PowerSplit to perform optimal splitting in these complicated scenarios. We focus on the fact that the variance of the commonly used statistic is inversely proportional to the effective sample size, $n_{Eff,k} = \frac{4n_k^- n_k^+}{(n_k^- + n_k^+)}$, where n^+_k and n^-_k denote the number of cases and controls in study k respectively. If we use the inverse-variance-weighted average method for meta-analysis, the variance of the final estimator is inversely proportional to the total effective sample size,

$$n_{Eff,Total} = \sum_{i=1}^{K} n_{Eff,i}$$

PowerSplit performs splitting to maximize $n_{Eff,Total}$. Suppose that we have K studies with n_{shared} shared controls. PowerSplit performs the following procedure:

1. Calculate $n_{Eff,i}$ of each study i excluding n_{shared} individuals.
2. Assign each of n_{shared} individuals to a study by repeating steps a-c. Assign individuals shared by fewer studies first.
    a. Try assigning the individual to each of the K studies and calculate the increase in $n_{Eff,Total}$.
    b. Assign the individual to the study that maximizes $n_{Eff,Total}$.

c. Update $n_{Eff,k}$ to reflect the assignment.

Essentially, our method is a greedy algorithm that maximizes the target function ($n_{Eff,Total}$) at each step. We assign individuals shared by fewer studies first because those individuals are less flexible in the choice of assignment. By making decisions for them first, assignment of individuals shared by more studies can flexibly complement the previous assignments if sample size balancing is required.

4. Power simulations

We use the following power simulation scheme. We assume that we combine data of K diseases. That is, we have K studies, with one study per each single disease. To simulate each study, we assumed N^+ case and N^- control samples ($2N^+$ case chromosomes and $2N^-$ control chromosomes) and assumed a variant with minor allele frequency (MAF) of $p = 0.3$. Given a relative risk γ, which we assume to be the same for all diseases (which is an assumption that we discuss later), the case MAF will be $p^+ = \gamma p/((\gamma-1)p+1)$ and control MAF will be $p^- \approx p$, assuming a low prevalence. Thus, we can construct a 2×2 table by randomly sampling minor allele counts x^+_k from Binomial($2N^+$, p^+) for cases and x^-_k from

Binomial($2N^-$, p^-) for controls, respectively, where Binomial(n, p) refers to a binomial distribution for n trials with probability p. From this 2 × 2 table, we can calculate the log odds ratio and its approximated variance. If we wish to simulate independent studies with no overlapping samples, we can repeat this to generate K studies.

Because we want to simulate shared controls in cross-disease analysis, we modify the simulation scheme as follows. To simplify the simulations, we assume that if some control subjects are shared by multiple studies, they are shared by all K studies. (A more complicated design can be developed in which controls are shared by a subset of studies, where the specific subset can vary between subjects.) Let η denote the proportion of shared controls among all controls contained in each study. We assume that the case sample sizes are equal and η is equal for all studies. Given η, to ensure that the total number of unique individuals is constant, which will maintain the power of the splitting method, we adjust the control sample size from N^- to $M^- = \frac{K}{K(1-\eta)+\eta}N^-$. The splitting method in this situation is equivalent to sampling $x^-_{Spec,k}$ from Binomial($2(1-\eta)M^-$, p^-) for study-specific controls and $x^-_{Split,k}$ from Binomial($2\eta M^-/K$, p^-) for split shared controls for each study

k. The total minor allele count of controls in study k is $\bar{x}_{Spec,k}+\bar{x}_{Split,k}$ (which is essentially the same as sampling from Binomial($2\bar{N}$, $\bar{p}$)). For existing meta-analysis methods, the minor allele count of shared controls before splitting should be used, which is $\bar{x}_{Shared} = \Sigma\bar{x}_{Split,k}$. The total minor allele count of the controls in study k is the sum of $\bar{x}_{Spec,k}+\bar{x}_{Shared}$. The power of methods is evaluated as the proportion of 10,000 simulations whose p-values exceed 5×10^{-8}.

4.1 WTCCC data

We obtained data from the Wellcome Trust Case Control Consortium (WTCCC) (5). These data include one shared control group for the analysis of seven different diseases. In their study(5), 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 1** of The Wellcome Trust Case Control Consortium (5)). In addition to these two loci, we further investigated pleiotropic loci known to be associated with multiple autoimmune diseases. We

obtained data for known associated loci from ImmunoBase (www.immunobase.org) and 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 subsequent studies, we assumed that they likely represent true-positives.

We evaluated the significances of these eight loci using two different designs. First, we tried a full overlap design. 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, 2,938 shared controls were available (1,480 individuals from the 1958 British Birth Cohort and 1,458 individuals from the UK Blood Service Control Group). We compared all these controls to cases of CD (N = 1,748), RA (N = 1,860), and T1D (N = 1,963) in the logistic regression framework.

Second, we tried a partial overlap design. 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 three diseases. Thus, approximately half of the control samples used for each disease analysis was shared by the three diseases ($\eta \approx 0.5$). Applying FOLD to this design involved calculating six estimates ($\hat{\beta}_{CD,Spec}$, $\hat{\beta}_{CD,Shared}$, $\hat{\beta}_{RA,Spec}$, $\hat{\beta}_{RA,Shared}$, $\hat{\beta}_{T1D,Spec}$, $\hat{\beta}_{T1D,Shared}$).

4.2 PGC data

We obtained summary statistics data (OR and standard error) from the PGC. We needed the 2x2 tables that generated the summary statistics, which were unavailable. Therefore, we reconstructed the 2x2 table that would likely have given the summary statistics. We obtained MAF from HapMap3(6). Given the numbers of cases and controls and marginal MAF, the degree of freedom of table became one. Thus, we enumerated 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.

Discussion

In this paper, we described a possible pitfall of using shared controls in meta-analysis of genetic association studies. We identified and reported a phenomenon that adding shared controls to multiple studies in a meta-analysis can unexpectedly hurt the statistical power. Possible strategies to avoid power reduction include (1) using splitting approach, (2) using only shared controls without study-specific controls, and (3) using our new method FOLD. The key idea of FOLD is to combine a set of statistics that are calculated using homogeneous samples in terms of their information. We employed Lin-Sullivan method to combine the statistics in FOLD, but other approaches can also be used (Fig. 10). We also presented a companion method Power Split that determines the optimal splitting design.

We expect that cross-disease meta-analyses can be more vulnerable to this pitfall than single-disease meta-analyses. To uncover pleiotropic risk loci, investigators are now combining association results of multiple diseases (16-18, 20, 21). However,

investigators of a single disease typically maximize sample size without strictly avoiding sample overlap to analyses of different diseases. For example, investigators studying multiple diseases may utilize the same control group for each disease. Additionally, investigators may borrow healthy control samples from the analyses of different diseases. In cross-disease analysis, because of these complicated situations, partial overlap designs can easily occur. Recent cross-disease meta-analyses that utilized existing methods were all in partial overlap designs (16-18), suggesting that the use of splitting or FOLD might possibly increase the statistical power.

One challenge of using FOLD is obtaining multiple statistics from each study. This typically requires the investigator performing meta-analysis to contact the investigator of each study to ask for recalculation of statistics. Although this task can be burdensome, considering the dramatic power gain compared to existing approaches, the task can be worth the effort. Another challenge of using FOLD can be identifying overlapping individuals. In fact, this is a challenge in any methods designed for sample overlap. If genotype data cannot be shared because of privacy issues, identification of duplicates can be difficult. Fortunately, recent

methods allow detection of duplicates and close relatives without sharing genotype data (22, 23).

In meta-analyses of genetic association studies, fixed effects model methods are widely used. In our simulations, we followed this practice and assumed the constant effect size across studies. In cross-disease analysis (16-18, 20, 21), this assumption may not hold true. In such situations, specialized methods for heterogeneity can be powerful (14, 24, 25) where ASSET is one of those. Our evaluation of ASSET demonstrates that power drop from using shared controls is an independent issue from whether a method is designed for heterogeneity. How we should account for effect size heterogeneity for cross-disease analysis is an important issue that requires further investigation.

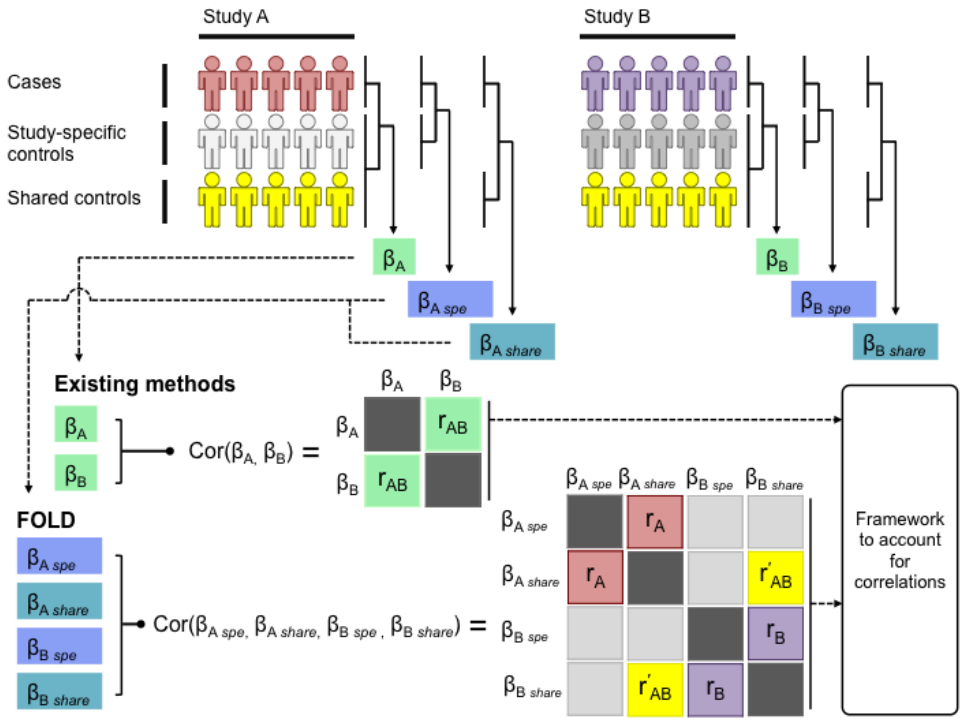


Figure 9. Analysis pipelines of existing approaches and our proposed strategy FOLD. β represents the estimated effect size. Subscripts of β denote the study (A and B) and design of control samples (spe, using study-specific controls; share, using shared controls). r_{AB} refers to the cross-study correlation between β_A and β_B in the existing approaches. r_{AB} refers to the cross-study correlation between the two statistics β_A;share and β_B;sharethat are calculated using shared controls only. r_A refers to the within-study correlation between the statistic calculated using study-specific controls (β_A;spe) and the statistic calculated using shared controls (β_A;share). r_B is defined similarly to r_A.

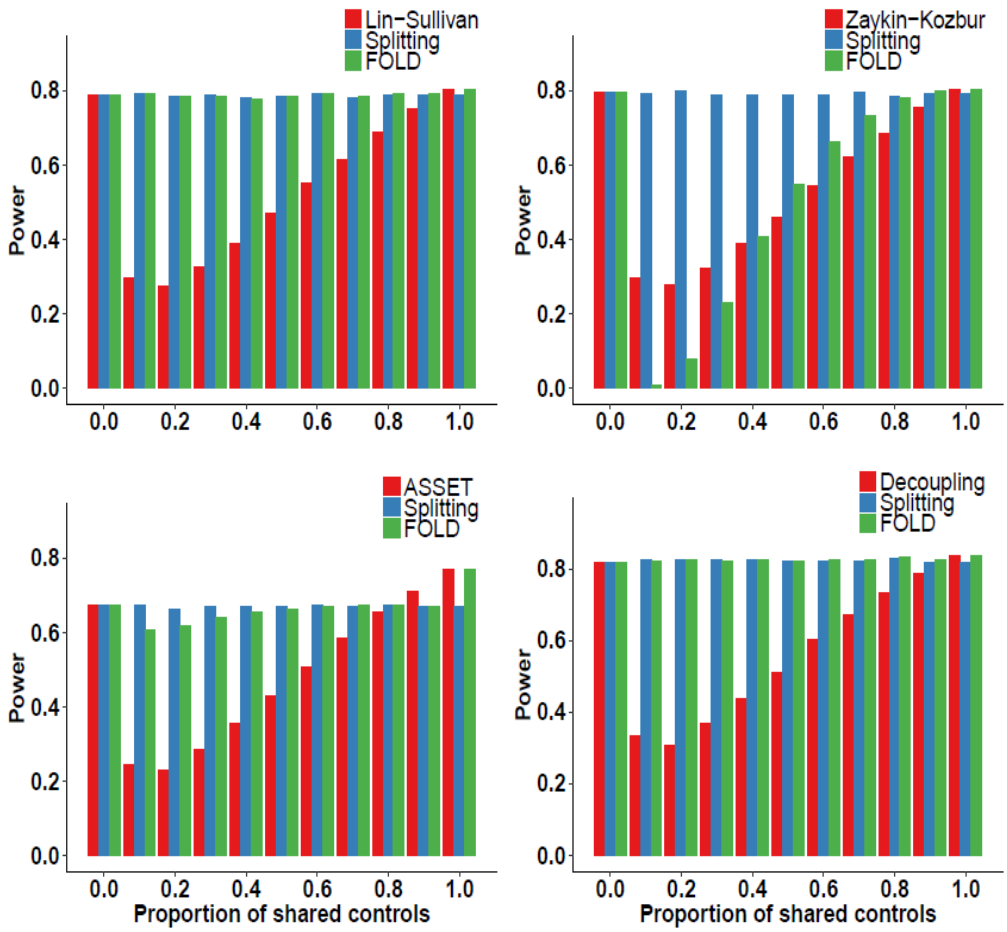


Figure 10. Powers of FOLD where differing methods were used to combine statistics in FOLD. At the second step of FOLD where we combine multiple statistics, we tried different methods other than the default choice of Lin−Sullivan method. We examined four methods, Lin−Sullivan method (a), ASSET (b), Zaykin−Kozbur method (c), and 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.

References

1. Weinhold N, *et al.* (2013) The CCND1 c.870G > A polymorphism is a risk factor for t(11;14)(q13;q32) multiple myeloma. *Nature Genetics* 45(5):522-U587.
2. Chubb D, *et al.* (2013) Common variation at 3q26.2, 6p21.33, 17p11.2 and 22q13.1 influences multiple myeloma risk. *Nature Genetics* 45(10):1221-U1366.
3. 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. *Nature Genetics* 47(4):381-U199.
4. Speedy HE, *et al.* (2014) A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia. *Nature Genetics* 46(1):56-+.
5. Wellcome Trust Case Control C (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447(7145):661-678.
6. Mukherjee S, *et al.* (2011) Including additional controls from public databases improves the power of a genome-wide association study. *Hum Hered* 72(1):21-34.
7. Crowther-Swanepoel D, *et al.* (2009) Genetic variation in

CXCR4 and risk of chronic lymphocytic leukemia. *Blood* 114(23):4843-4846.

8. Shete S, *et al.* (2009) Genome-wide association study identifies five susceptibility loci for glioma. *Nature Genetics* 41(8):899-U854.

9. Di Bernardo MC, *et al.* (2008) A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. *Nature Genetics* 40(10):1204-1210.

10. 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.

11. Orozco G, *et al.* (2014) Novel Rheumatoid Arthritis Susceptibility Locus at 22q12 Identified in an Extended UK Genome-Wide Association Study. *Arthritis Rheumatol* 66(1):24-30.

12. Lin DY & Sullivan PF (2009) Meta-analysis of genome-wide association studies with overlapping subjects. *Am J Hum Genet* 85(6):862-872.

13. Han B, *et al.* (2016) A general framework for meta-analyzing dependent studies with overlapping subjects in association mapping. *Hum Mol Genet* 25(9):1857-1866.

14. 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(5):821-835.

15. Zaykin DV & Kozbur DO (2010) P-value based analysis for shared controls design in genome-wide association studies. *Genet Epidemiol* 34(7):725-738.

16. Dichgans M, *et al.* (2014) Shared genetic susceptibility to ischemic stroke and coronary artery disease: a genome-wide analysis of common variants. *Stroke* 45(1):24-36.

17. Kar SP, *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*.

18. 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(10):1268-1276.

19. Cross-Disorder Group of the Psychiatric Genomics C (2013) Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* 381(9875):1371-1379.

20. Zhernakova A, *et al.* (2011) Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet* 7(2):e1002004.

21. Ellinghaus D, *et al.* (2016) Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat Genet* 48(5):510-518.

22. Hormozdiari F, *et al.* (2014) Privacy preserving protocol for detecting genetic relatives using rare variants. *Bioinformatics* 30(12):204-211.

23. He D, *et al.* (2014) Identifying genetic relatives without compromising privacy. *Genome Res* 24(4):664-672.

24. Morris AP (2011) Transethnic meta-analysis of genomewide association studies. *Genet Epidemiol* 35(8):809-822.

25. Han B & Eskin E (2011) Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. *American Journal of Human Genetics* 88(5):586-598.

26. Rao JNK (1973) On the Estimation of Heteroscedastic Variances. *Biometrics* 29(1):11-24.

CHAPTER 2: Achieving balanced power for detecting risk and protective alleles in meta-analysis of association studies with overlapping subjects.

Introduction

In a case-control genetic association study, investigators collect case and control samples and compare genotypes of a locus between cases and controls to pinpoint phenotype-associated locus. To increase the samples participating in an analysis, meta-analysis of genetic association studies became now prevalent, in which the summary statistics from multiple independent studies are combined. Recently, studies often utilize the public dataset as controls. If multiple studies participating in a meta-analysis utilize the same public dataset as controls (1), the summary statistics from these studies are no more independent and become correlated (2-5). Therefore, the use of standard meta-analysis frameworks leads to increased false positives.

Fortunately, recent advances in meta-analysis methods allow one to account for the correlations caused by overlapping

subjects. Lin and Sullivan proposed the correlation estimator based on the shared and unshared sample sizes (6). They also suggested an optimal test statistic to account for the correlations. Their method turns out to achieve the similar power to the splitting strategy, which refers to the method that splits the shared individuals into the studies prior to meta-analysis. Splitting is considered as the gold standard, but only doable in rare situations that we have access to the genotype data of the studies. After Lin and Sullivan proposed their approach, many different methods were proposed, but all these methods were based on the similar correlation estimator (7-9).

In this paper, we report a phenomenon that the use of the standard method suggested by Lin and Sullivan (6) can lead to unbalanced power for detecting protective alleles (OR < 1) and risk alleles (OR > 1). Specifically, when we assumed that the controls were shared, the power for detecting protective minor alleles (OR<1) were lower than the power for detecting risk minor alleles (OR>1). The degree of asymmetry was exacerbated as the minor allele frequency (MAF) decreased. For example, for detecting an allele of frequency 10% and of OR=0.85, simulating meta-analysis of 5 studies shows that the standard method only achieved 61.6%

power whereas splitting achieved 67.0%. Therefore, the existing method lost more than one thirds of power. By contrast, if we flip the effect direction (OR=1.17), the existing method conversely achieved higher power (71.8%) than splitting. Such power asymmetry is apparently not a characteristic of the test that we expected or intended. To our knowledge, we are the first to report this phenomenon.

After investigating on this phenomenon, we identified that the power asymmetry problem occurred because the standard correlation estimator did not exactly predict the true correlation. The existing estimator was approximated under the null hypothesis of no effect, but it was not investigated how accurate it is under the alternative hypothesis and how much impact the errors have on power. It turns out that the true correlation is largely dependent on MAF and effect size, which could lead to substantially unbalanced power (10-12). To overcome the power asymmetry problem, we developed a method that uses an accurate correlation estimator, called PASTRY (A method to avoid Power ASymmeTRY). We present two versions of our method: the analytical version (PASTRY) that is very fast and provides moderate accuracy, and the empirical version (PASTRY-emp) that is slow but provides

high accuracy.

RESULTS

1. Power asymmetry of existing method

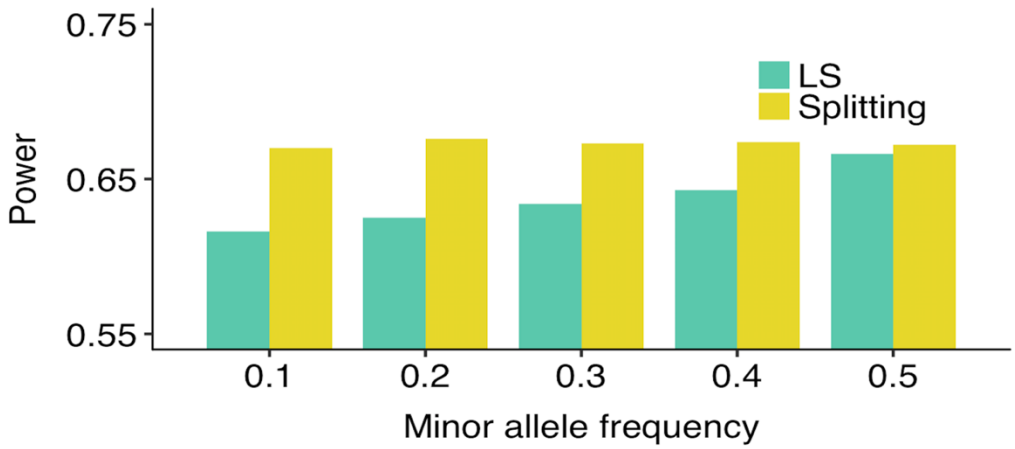
To assess the power of the existing method (Lin−Sullivan method), we used the simulation framework described in the Method section. In this simulation, we meta−analyze M studies that share N^- controls. We compared the power of the existing method to the splitting approach that equally splits shared controls into individual studies. Here, we varied the cut−off threshold to keep the overall power of the splitting strategy at about 60 − 70%.

In this paper, we assume that odds ratios (OR) are always calculated for the risk of minor allele relative to the risk of major allele and observed that the existing meta−analysis approach for dealing with sample overlap showed power asymmetry to detect risk and protective alleles. When the OR < 1(protective allele), power of the existing method decreased compared to the splitting approach. By contrast, when the OR > 1(risk allele), power of the existing method increased compared to the splitting approach. The

power asymmetry exacerbated as the minor allele frequency decreases and as the OR moves from 1. Figure 1 shows the power of the existing method compared to the splitting approach in a meta-analysis of ten studies with a variation in minor allele frequency. In case of the protective allele (OR < 1, RRs=0.85), the magnitude of power drop increased from 0.6% to 5.4% when the minor allele frequency of a protective allele decreases from 0.5 to 0.1. In case of the risk allele (OR < 1, RRs=1.17), the magnitude of power gain increased from 0.3% to 4.6% when the minor allele frequency of a risk allele decreases from 0.5 to 0.1.

Figure 2 shows the power of the existing method compared to the splitting approach at various relative risks of the allele. Our interest here is to observe the impact of relative risk in power without any other effect, thus we fixed a minor allele frequency as 10% for both the risk and protective allele and combined 5 studies. Moreover, in order to keep the power of the splitting approach we varied the the cut-off threshold. In case of the protective allele (OR < 1), the magnitude of power drop increased from 1.5% to 15.6% when the relative risk decreases from 0.95 to 0.75. In case of the risk allele (OR > 1), the magnitude of power gain increased from 1.4% to 8.5% when the relative risk increases from 1.05 to 1.25.

(A) Protective allele (OR < 1)



(B) Risk allele (OR > 1)

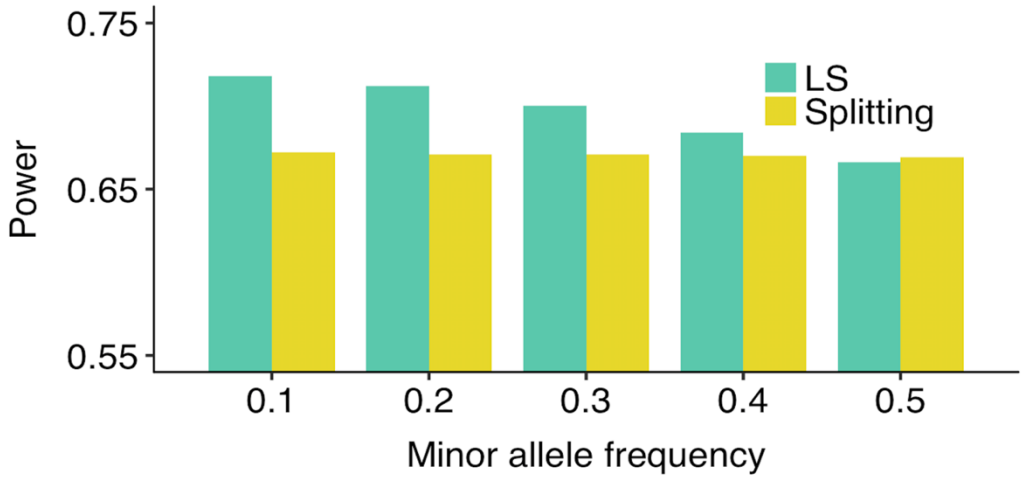
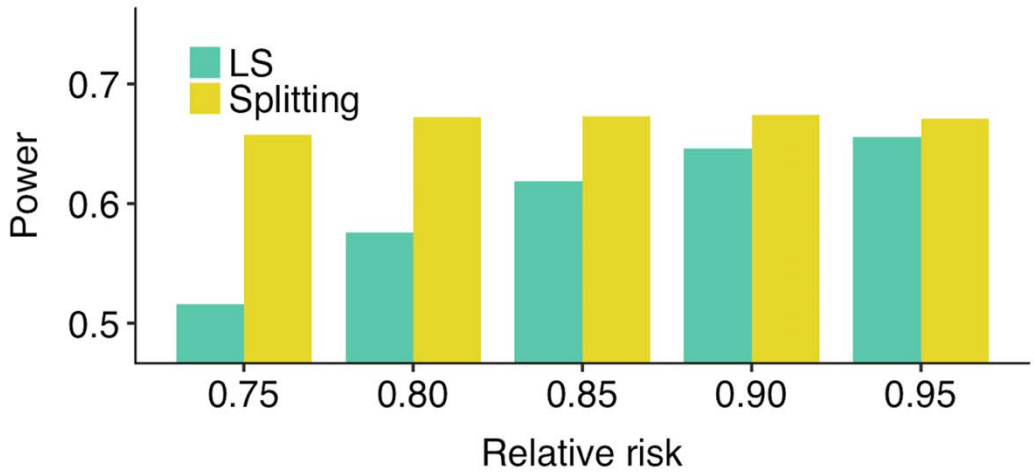


Figure. 11. Powers of Lin and Sullivan and Splitting method with minor allele frequency. *We measured the power of the existing method and splitting method as we varied the minor allele frequency from 0.1 to 0.5. We considered two different scenarios: (A) detecting protective minor alleles (OR < 1), (B) detecting risk minor alleles (OR > 1).*

(A) Protective allele (OR < 1)



(B) Risk allele (OR > 1)

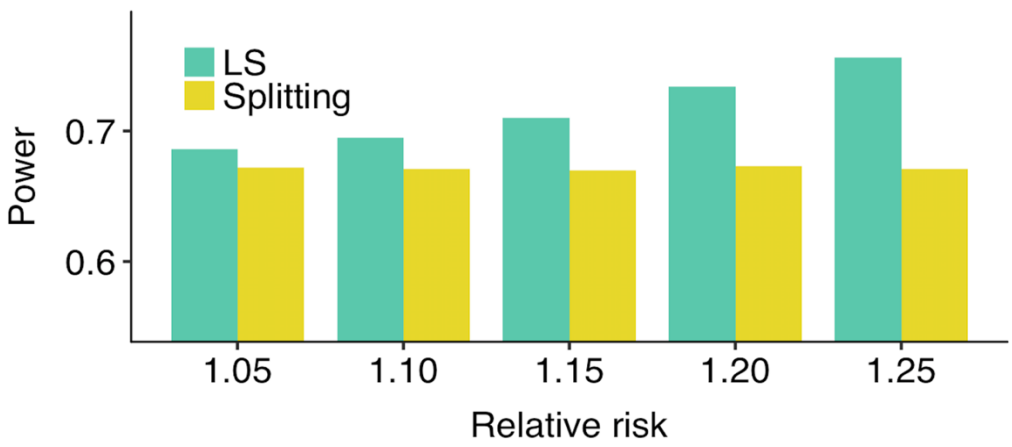


Figure. 12. Powers of Lin and Sullivan and Splitting method with relative risk. *We measured the power of the existing method and splitting method as we varied the relative risk from 0.75 to 1.25. We considered two different scenarios: (A) detecting protective minor alleles (OR < 1), (B) detecting risk minor alleles (OR > 1).*

2. Correlation estimator of existing method

Our hypothesis is that one of the reason behind the power asymmetry phenomenon is due to incorrect correlation estimations. Here, we examined the correlation accuracy of the existing estimator at various minor allele frequencies and relative risks.

2.1 Minor allele frequency

First, we investigated the amount of correlation inaccuracy of the existing estimator at various minor-allele frequencies, MAF. Here, we combined 5 studies with fixed relative risk of 0.75 and changed the minor allele frequencies from 0.1 to 0.5, of which the OR <1 and OR > 1 indicates the case of protective allele and risk allele in GWAS, respectively. We compared the correlation estimates of the existing estimator to the empirically assessed correlation by a large number of repetition (10 million), which we call "real correlation". Table 1 shows that the correlation estimates for the existing method was constant for all situations whereas the real correlation changed with varying minor allele frequency, MAF. As the MAF decreases the correlation inaccuracy increases compared to the real correlation. For instance, the existing estimator predicted correlation to be 0.167 at all situations, but the real correlation was

0.136 at a MAF of 0.1 of protective allele (OR=0.75), which was overestimated by 23%, and the real correlation was 0.282 at a MAF of 0.1 of risk allele (OR=1.25), which was underestimated by 13%.

2.2 Relative risk

To simulate realistic scenarios, we changed the relative risk from 0.75 to 1.25, of which the range includes the typically observed relative risks in GWAS. We used a MAF of 0.1 and compared the correlation estimates of the existing estimator to the empirically assessed correlation by a large number of repetition (10 million), which we call "real correlation". Table 2 shows that the correlation estimates for the existing method was constant for all situations whereas the real correlation changed with varying relative risks. As the relative risk increases the real correlation increases. For instance, the correlation estimated from the existing method was 0.167. However, the real correlation was 0.136 at a relative risk of 0.75, which was overestimated by 23%, and the real correlation was 0.1827 at a relative risk of 1.25, which was underestimated by 13%.

(A) OR=0.75

Minor allele frequency (MAF)	Real correlation	Correlation from the existing method
0.1	0.1359	0.16667
0.2	0.1424	0.16667
0.3	0.1484	0.16667
0.4	0.1563	0.16667
0.5	0.1643	0.16667

(B) OR=1.25

Minor allele frequency (MAF)	Real correlation	Correlation from the existing method
0.1	0.1916	0.16667
0.2	0.1848	0.16667
0.3	0.1779	0.16667
0.4	0.1714	0.16667
0.5	0.1653	0.16667

Table 3. Comparison of the real correlation with the correlation from the existing method with various minor allele frequencies. *Here we used two different settings: the case of (A) protective allele (OR=0.75) and (B) risk allele (OR=1.25).*

(A) Protective allele (OR<1)

Relative risk (RRs)	Real correlation	Correlation from the existing method
0.75	0.1359	0.16667
0.80	0.1426	0.16667
0.85	0.1487	0.16667
0.90	0.1551	0.16667
0.95	0.1610	0.16667

(B) Risk allele (OR>1)

Relative risk (RRs)	Real correlation	Correlation from the existing method
1.05	0.1723	0.16667
1.10	0.1776	0.16667
1.15	0.1827	0.16667
1.20	0.1874	0.16667
1.25	0.1916	0.16667

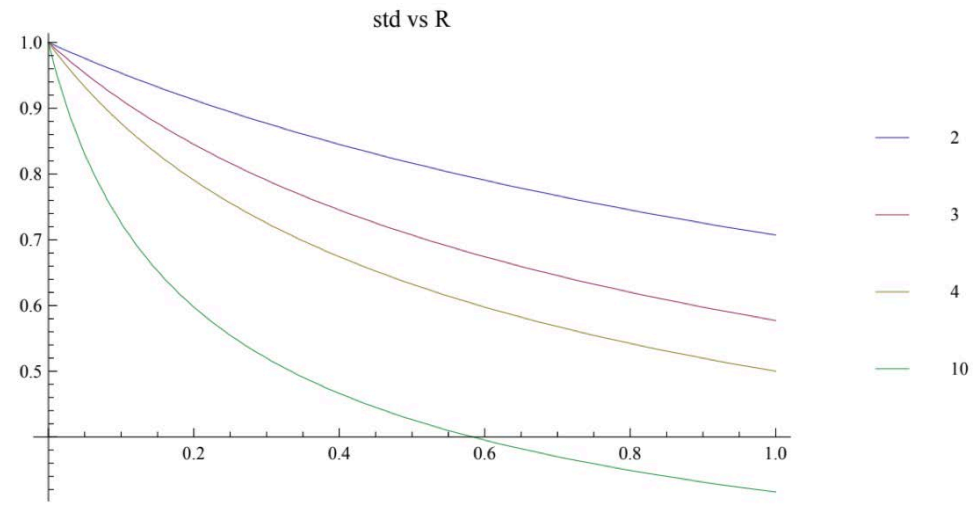
Table 4. Comparison of the real correlation with the correlation from the existing method with various relative risks from 0.75 to 1.25. *Here we used two different settings: the case of (A) protective allele (OR<1) and (B) risk allele (OR>1).*

3. Cumulative effect of correlation inaccuracy

Here, we investigated the effects of inaccurate correlation estimates in the final meta-analysis statistics. Figure 3 A shows the effect of meta-analysis standard deviation (Y axis) on the correlation between studies (X axis) with variation in number of studies. The increase in the number of studies of meta-analysis is observed to lead to significant changes in the final meta-analysis statistics. More specifically, Figure 3 B showed the rate of change of the standard deviation with respect to the number of studies in meta-analysis (The first derivative of the meta-analysis equation of standard deviation). The rate of change is extremely rapid when the number of studies are large (M=10), especially in the low correlation region ($r < 0.5$). One of the possible reasons is that we have to account the M-1 correlations to combine the M studies and their cumulative effect results in final meta-analysis statistics.

Therefore, if the correlation between the two studies is estimated with small inaccuracies, then their cumulative effects can lead to large differences in the final standard deviation when combining large numbers of studies. For example, if we estimate the correlation that is 0.01 less/greater than the real correlation

(A)



(B)

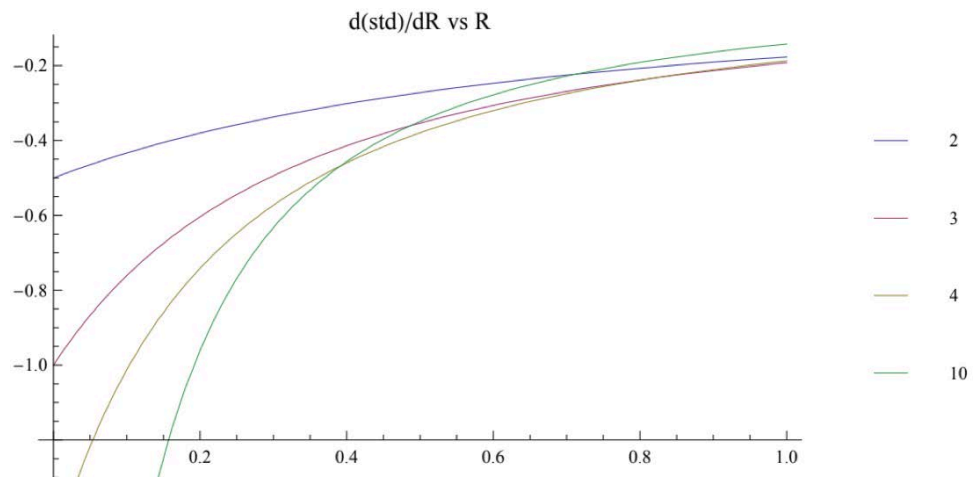


Figure 13. Relation between number of studies and standard deviation. *(A) effect of meta-analysis standard deviation (Y axis) on the correlation between studies (X axis) with variation in number of studies and (B) the rate of change of the standard deviation with respect to the number of studies in meta-analysis*

(real correlation=0.1), this difference is not problematic when we combine two studies. However, if we combine fifteen studies, the 0.01 difference in correlation leads to 4% increase and 3% decrease compared to the true final standard deviations after accounting true correlations (Figure 3 B). The current GWAS study contains larger number of controls than the cases, this corresponds to lower correlation ($r<0.5$), and as the fraction of controls increase the correlation gets even lower.

4. Performance of PASTRY

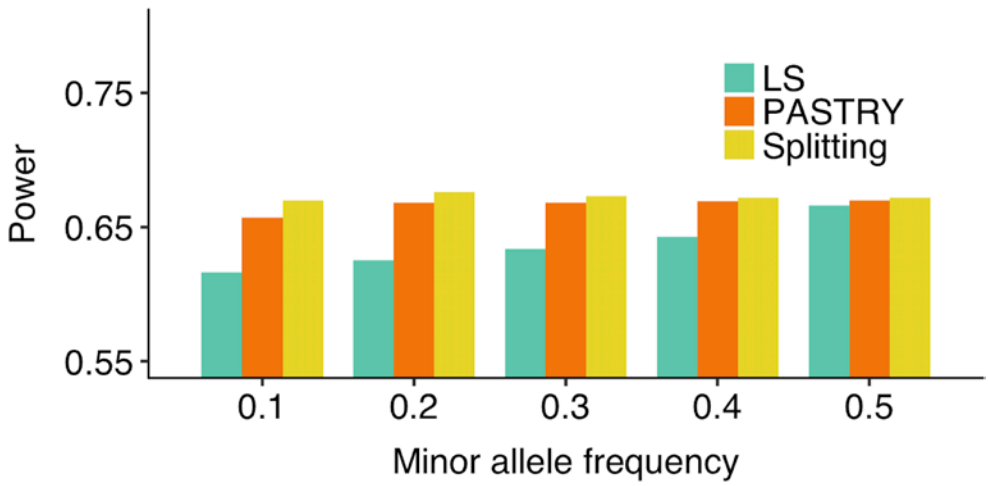
To evaluate the performance of our new method, PASTRY, we followed the same simulation framework as in the Method section. First of all, we observed that the false positive rate is controlled in all situations. In general, the PASTRY method is particularly useful when the relative risk of a particular allele is not small and when the minor allele frequency of the allele is low. Figure 4 show the powers of PASTRY and the existing estimator with the Splitting approach at various minor allele frequencies (MAF 0.1−0.5). Overall, the power of PASTRY is higher than the existing estimator. For instance, in the case of protective allele (OR < 1), the power difference between the PASTRY and the Splitting is only 1%

whereas the power difference between the existing estimator and the Splitting is 5% at MAF of 0.1 (PASTRY: 66%, Splitting: 67%, Lin-Sullivan: 62%).

Moreover, the powers of PASTRY and the existing estimator with the Splitting approach at various relative risks (RRs 0.75 - 1.25). The power of PASTRY is close to the Splitting. For instance, in the case of protective allele with RRs = 0.75, the power difference between the PASTRY and the Splitting is only 1% whereas the power difference between the existing estimator and the Splitting is 15% (PASTRY: 66%, Splitting: 67%, Lin-Sullivan: 52%). The correlation accuracy can be improved by PASTRY-emp which uses the real correlation from large iterations.

(A) Protective allele

(OR<1)



(B) Risk allele (OR > 1)

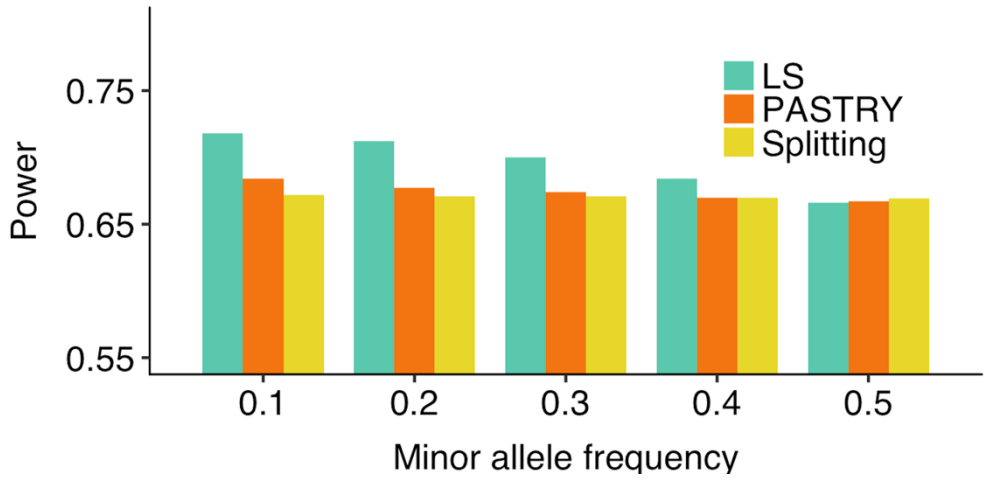
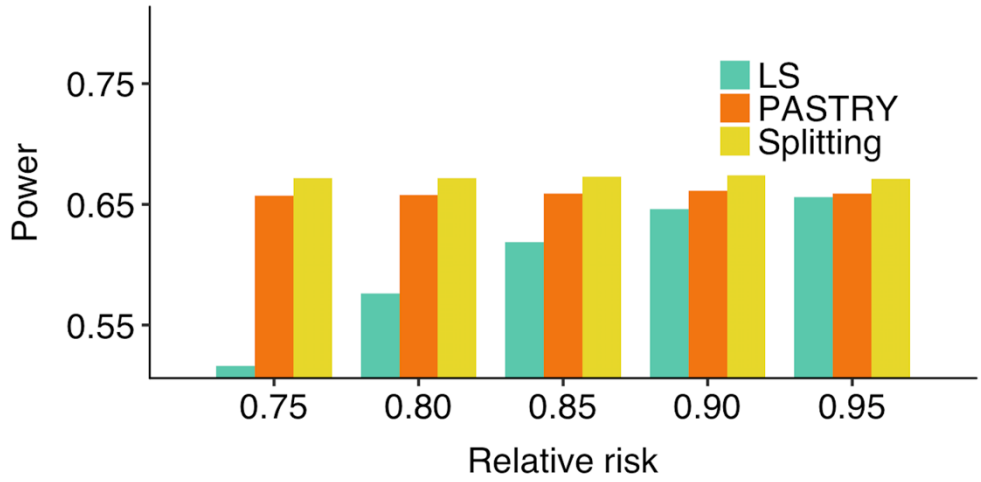


Figure 14. Powers of Lin-Sullivan, PASTRY and Splitting. *We measured the power of methods as we varied the minor allele frequency.*

(A) Protective allele (OR < 1)



(B) Risk allele (OR > 1)

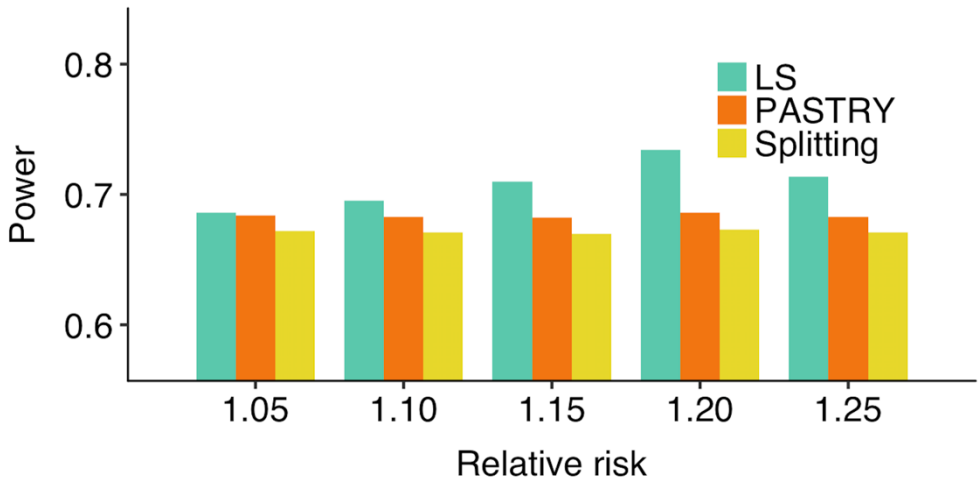


Figure 15. Powers of Lin-Sullivan, PASTRY and Splitting with various relative risks.

MATERIALS AND METHODS

1. Correlation from Lin and Sullivan's method

Lin and Sullivan (6) developed a method that can account for the correlation introduced by overlapping samples in meta-analyses.

Here we outline the general framework of Lin and Sullivan.

First, Lin and Sullivan derived correlation equation between two studies and then the final meta-analysis statistic is obtained after accounting the correlations.

Suppose that we have M studies with observed effect sizes of $X_1, \cdots, X_M$ and variances of $V_1, \cdots, V_M$ in the meta-analysis.

The correlation formulae of the existing method between study k and l is

$$r_{kl} \approx \frac{n_{kl-}\sqrt{\frac{n_{k+}n_{l+}}{n_{k-}n_{l-}}} + n_{kl+}\sqrt{\frac{n_{k-}n_{l-}}{n_{k+}n_{l+}}}}{\sqrt{n_k n_l}} \qquad (1)$$

This equation relates to case-control study thus, we used the subscript + and - to define case and control subjects. Note that the n_k and n_l denote the total number of samples in study k and l, and n_{kl+} and n_{kl-} denote the number of overlapping case and control

subjects between study k and l, respectively.

If we wish to combine M studies with overlapping samples, we can build a MxM correlation matrix C, where element [j,k] is the correlation between studies j and k, r_{jk}.

$$\mathbf{C} = [r_{ij}]_{M \times M}$$

Then, the variance−covariance matrix $\boldsymbol{\Omega}$ can be obtained using the above correlation matrix and standard deviations of studies. Then, the meta−analyzed effect size, X_{LS}, and the variance $Var(X_{LS})$ are obtained using

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

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

where e is the vector of ones.

2. Correlation estimator of PASTRY

We propose a new correlation estimator, PASTRY (A method to avoid Power ASymmeTRY), which is more accurate than the existing estimator for testing the SNPs with a small minor allele frequency and a non−negligible effect in meta−analysis involving

overlapping samples. Our estimator is also based on the score test, but our estimator is derived under the null hypothesis of effect, $\beta \neq 0$. Here, we analytically approximate the correlation between statistics using our new method, PASTRY.

Let y_{ki} and x_{ki} are the binary disease status (case-control) and the genotype on the i[th] subject of the k[th] study.

The log-likelihood model of the k[th] study:

$$l(\alpha,\beta) = \sum_{i}^{n} \beta_k^T X_{ki} y_{ki} - \sum_{i}^{n} \log\left(1 + e^{\alpha_k + \beta_k^T x_{ki}}\right)$$

where α_k and β_k are the intercept and regression parameter, respectively.

Then, the first and second derivative of log-likelihood model with respect to parameters (α_k and β_k) are:

$$U(\alpha,\beta) = \left(\frac{\partial l(\alpha,\beta)}{\partial \alpha}, \frac{\partial l(\alpha,\beta)}{\partial \beta}\right) = \sum_{i}^{n} \left(y_{ki} - \frac{e^{\alpha_k + \beta_k^T x_{ki}}}{1 + e^{\alpha_k + \beta_k^T x_{ki}}}\right) \tilde{X}_{ki}$$

$$\tilde{X}_{ki} = \begin{vmatrix} 1 \\ X \end{vmatrix}$$

$$I(\alpha,\beta) = \left(\frac{\partial^2 l(\alpha,\beta)}{\partial \alpha^2}, \frac{\partial^2 l(\alpha,\beta)}{\partial \beta^2}\right) = \sum_{i}^{n} \left(\frac{e^{\alpha_k + \beta_k^T x_{ki}}}{1 + e^{\alpha_k + \beta_k^T x_{ki}}}\right) \tilde{X}_{ki} \tilde{X}_{ki}^T$$

$$\tilde{X}_{ki} \tilde{X}_{ki}^T = \begin{bmatrix} 1 & X_{ki}^T \\ X_{ki} & X_{ki} X_{ki}^T \end{bmatrix}$$

After 2[nd] order Taylor series expansion, covariance between study k and l with n_{kl} overlapping samples can be estimated with

$$cov(\hat{\theta}_k, \hat{\theta}_l) \approx I^{-1}\left(\theta_k^{(t)}\right) cov\left(U_k\left(\theta_k^{(t)}\right), U_l\left(\theta_l^{(t)}\right)\right) I^{-1}\left(\theta_l^{(t)}\right)$$

where

$$cov\left(U_k\left(\theta_k^{(t)}\right), U_l\left(\theta_l^{(t)}\right)\right)3 \approx \sum_i^{n_{kl}} \left(Y_{ki} - \frac{e^{\alpha_k+\beta_k^T X_{ki}}}{1+e^{\alpha_k+\beta_k^T X_{ki}}}\right)\left(Y_{li} - \frac{e^{\alpha_l+\beta_l^T X_{li}}}{1+e^{\alpha_l+\beta_l^T X_{li}}}\right)\tilde{X}_{ki}\tilde{X}_{li}^T$$

Here, our method, PASTRY, assumes that the case/control status and regression parameter are dependent and overlapping subject who appears in both study k and l has the same case/control status and genotype in both studies ($X_{ki} = X_{li},\ Y_{ki} = Y_{li}$).

$$cov\left(U_k\left(\theta_k^{(t)}\right), U_l\left(\theta_l^{(t)}\right)\right) \approx \sum_i^{n_{kl1}} \left(1 - \frac{e^{\alpha_k+\beta_k^T X_{ki}}}{1+e^{\alpha_k+\beta_k^T X_{ki}}}\right)\sum_i^{n_{kl0}} \left(-\frac{e^{\alpha_k+\beta_k^T X_{ki}}}{1+e^{\alpha_k+\beta_k^T X_{ki}}}\right)\tilde{X}_{ki}\tilde{X}_{li}^T$$

where n_{kl1}, n_{kl0} denote number of overlapping case samples and control samples between study k and l, respectively.

3. Power simulations

We performed simulation to evaluate the false positive rate (FPR) and power and also assessed a correlation accuracy. We assumed that we combine data of M studies. To simulate each study, we assumed a low prevalence(γ) with N^+ case ($2N^+$ case chromosomes) and N^- control ($2N^-$ control chromosomes) samples. In each simulation, we varied the minor allele frequency and the

corresponding case MAF is $p^+ = \gamma p/((\gamma-1)p+1)$ and control MAF is $p^- \approx p$. The 2 x 2 table is constructed to obtain the log odds ratio and the variance by randomly sampling the number of minor allele for cases x^+_M and controls x^-_M with binomial distribution, Binomial (ncase, p^+) and Binomial (ncase, p^-), respectively. To simplify the simulations, we assumed that the number of cases are the same for each study and all control subjects are shared by K studies. The splitting method is equivalent to sampling $x^-_{Split,M}$ from Binomial($2N^-/M$, p^-) and for the existing meta-analysis method, the minor allele count of shared controls before splitting should be used, which is $x^-_{Shared} = \Sigma x^-_{Split,M}$.

To assess the power of methods, we iterate the simulation 0.1 million times with the GWAS threshold $5E^{-8}$.

DISCUSSION

Recently, investigators are combining genetic association studies of multiple different diseases using meta-analysis to uncover pleiotropic loci. In this cross-disease analysis, the use of standard meta-analysis strategies is often impractical because of shared controls across different diseases, which may induce correlations between statistics and incur false-positives if not properly

accounted for. To deal with overlapping subjects, previous studies developed a few meta-analysis approaches, which have been utilized in recent cross-disease meta-analysis studies. In this paper, we identify and report a phenomenon that these approaches can result in power asymmetry in overlap designs that cause unbalanced power for detecting protective alleles (OR < 1) and risk alleles (OR > 1).

We propose a correlation estimator, PASTRY, which accurately estimate correlation for overlapping samples in any design.

Our study observes that when the controls were shared, the power of the existing method for detecting protective minor alleles (OR<1) were lower than the power of the existing method for detecting risk minor alleles (OR>1) and under this design, one should consider using PASTRY. Our method is based on the correlation estimator that was designed to be accurate under the alternative hypothesis. We show that using our method, one can effectively achieve symmetry on power for testing risk and protective alleles.

References

1. Wellcome Trust Case Control C (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447(7145):661-678.
2. Purcell S., *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81:559-575.
3. Kavvoura1 F.K., Ioannidis J.P. (2008) Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. *Hum.Genet.*123:1-14.
4. de Bakker P.I., *et al.* (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.*17(R2):R122-R128.
5. Lin D.Y., Zeng D. (2009) Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data.*Genet. Epidemiol.* Published online October 21, 2009.
6. Lin DY & Sullivan PF (2009) Meta-analysis of genome-wide association studies with overlapping subjects. *Am J Hum Genet* 85(6):862-872.
7. Han B, *et al.* (2016) A general framework for meta-

analyzing dependent studies with overlapping subjects in association mapping. *Hum Mol Genet* 25(9):1857-1866.

8. 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(5):821-835.

9. Zaykin DV & Kozbur DO (2010) P-value based analysis for shared controls design in genome-wide association studies. *Genet Epidemiol.* 34(7):725-738.

10. Park JH, *et al.* (2011) Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susdeptibility variants. *PNAS.* 108(44):18026-18031.

11. Park JH, *et al.* (2010) Estimation of effect size distribution from genome-wide association studies and imputations for future discoveries. *Nat Genet*. 42:570-575.

12. Stringer S, Wray NR, Kahn RS et al (2011) Underestimated effect sizes in GWAS: fundamental limitations of single SNP analysis for dichotomous phenotypes. PLoS ONE 6:e27964

ABSTRACT IN KOREAN (국문초록)

메타분석에서 표본중첩 현상이 존재 할 경우의 문제점과 해결책

김은지
화학부 물리화학전공
서울대학교 대학원

인간의 유전체 염기서열이 규명된 이래로, 전장유전체를 대상으로 질병과 관련성을 밝히는 방법론 연구가 활발하게 이루어지고 있다. 전 게놈 관련분석 (GWAS) 은 2002년 개발된 방법으로, 병질환 및 약물 반응성에 대한 유전적 요인의 연관성을 분석한다. 관심있는 형질 또는 질환에 1차적으로 관련되어 있는 후보 유전자를 찾아내는 데 유용한 탐색 도구가 된다. 메타분석은 다른 그룹들에서 연구된 과학 연구의 결과를 결합하여 더 좋은 신뢰(Statistical power)의 결과를 얻는 통계 분석방법 이다. 전 게놈 관련분석 (GWAS) 에 대한 메타 분석은 점점 대중화되어 왔으며 최근에는 많은 메타 분석 방법이 제안됐다. 대부분의 메타 분석 방법은 하나의 연구에서 수집 된 개인이 다른 연구에 의해 다시 수집 될 가능성이 없기 때문에 연구가 독립적이라는 가정하에 여러 연구의 정보를 결합한다. 하지만 유전자형 또는 서열화 비용을 줄이기 위해 여러 연구에서 동일한 통제 개체를 이용하는 것이 점점 더 보편화되고 있다. 이로 인해 동일한 개인을 공유하는 연구가 의존적이며 메타 분석에서 중복되는 주제가 고려되지 않은 경우 가짜 연관이 발생할 수 있다. 메타분석이 더 좋은 신뢰(Statistical power)를 주어야 한다는 일반적인 기대와는 달리, 종속적인 여러 연구를 감안할 때 기존 메타 분석 성능이 감소 함이 보인다. 본 졸업논문에서는 이 현상이 발생하는 이유를 분석해, 새로운 방법 Fully-powered method for Overlapping Data (FOLD) 를 제안한다. 새로 제안한 방법으로 기존 메타분석에서 문제가 되던 종속적인 데이터를 완벽하게 독립적으로 만들었고, 이를 시뮬레이션과 실제 데이터를 통해 증명하였다. 또한, 신뢰도 비대칭 문제를 극복하기 위한 A method to avoid Power ASymmeTRy (PASTRY) 을 소개한다. 이 방법을 이용하면 연구간의 종속관계 즉 correlation을 기존방법보다 정확하게 얻고 기존 모델의 문제점이었던 신뢰도 비대칭을 해결 할 수 있다.