Material and Methods

Standard Meta-Analysis Method

A common approach in meta-analysis is to sum the Z-scores across studies, weighting them appropriately using the sample sizes (Stouffer et al. 1949). Suppose we have K studies, with $Z_k,\ k = 1, \ldots K$, being the Z-score from the k^{th} study and N_k the corresponding sample size. A standard meta-analysis uses weights $w_k,\ k = 1, \ldots K$, to combine the estimates as follows:

$$Z = \sum_{k=1}^{K} w_k\, Z_k \qquad \ldots \text{Equation (1)}$$

The Z_k's are assumed to be have standard normal distribution under the null hypothesis of no association between trait and genetic marker. Hence, the variance of the combined Z-score is:

$$\text{Var}(Z) = \sum_{k=1}^{K} w_k^2 \qquad \ldots \text{Equation (2)}$$

The weights are usually chosen based on per-study sample size so that larger studies have more weight (eg. $w_k = \frac{\sqrt{N_k}}{\sqrt{\sum_l N_l}}$). When the Z-scores are independent, these weights ensure that the combined Z-score is distributed as $N(0,1)$ under the null. However, when the studies have overlapping samples, the variance (2) becomes:

$$\text{Var}(Z) = \sum_{k=1}^{K} w_k^2 + 2 \sum_{k=1}^{K} \sum_{l=k+1}^{K} w_k w_l\, \text{Cov}\,(Z_k, Z_l) \qquad \ldots \text{Equation (3)}$$

where the covariance terms $\text{Cov}(Z_k, Z_l)$ depend on overlap between each pair of studies. Thus, using standard weights no longer leads to a $N(0,1)$ test statistic under the null. To account

for this, we estimate this covariance and adjust the weights accordingly. The optimal weights can be shown to be (Lin and Sullivan 2009):

$$[w_1, \dots w_K] = e^T\Omega^{-1}/e^T\Omega^{-1}e \qquad \dots \text{Equation (4)}$$

where e is a K x 1 vector of 1's and Ω is the estimated covariance matrix of $(Z_1, \dots Z_K)$.

The covariance matrix Ω can be calculated easily if individual level data are available, or if the exact number of overlapping samples between each pair of studies is known. We consider the more general case where the number of overlapping samples is not known and use the pair-wise correlation between Z-scores to estimate the overlap and adjust the weights as in (4).

Meta-Analysis Correcting for Sample Overlap

We develop a method to estimate the sample overlap and correct for it (**Figure 4.1**) using the correlation between Z-scores from each pair of studies. First, the Z-scores are stratified according to sample size at each marker because differences in the number of typed samples at each site could reflect success – or lack thereof – in genotyping across different studies. Second, we truncate the Z-scores using a cutoff value c (|Z| < c) to remove the effect of strongly associated loci. Finally, we estimate the correlation from these stratified truncated observations, and use it to estimate the covariance matrix in (4) and meta-analyze using the modified weights.

Correcting for Overlap in Meta-Analysis

Suppose there are K studies in a meta-analysis, and the Z-scores are combined in a weighted sum where w_k is the weight for the k^{th} study. If we can estimate the covariance between Z-scores of each pair of studies in the meta-analysis, we can meta-analyze adjusting for covariance using modified weights as in (4) as follows:

$$\hat{Z} = \frac{1}{\sqrt{\sum_k w_k^2 + \sum_k \sum_{l \neq k} w_k w_l \hat{r}_{kl}}} \sum_{k=1}^{K} w_k Z_k \qquad \dots \text{Equation (5)}$$

where $\hat{r}_{kl}$ is the estimated correlation between the Z-scores of the k^{th} and l^{th} studies under the null.

Using Truncated Z-scores to Estimate Covariance

We assume that (a) effect sizes at trait associated loci do not vary from study to study, a condition that should be approximately true given our assumption that all studies are of the same ancestry and (b) the degree of overlap is uniform across markers after accounting for sample size stratification. Furthermore, we assume that the Z-scores for a pair of studies have a bivariate normal distribution. Suppose that the trait under consideration is independent of genetic effects. Then the Z-scores are standard normally distributed for each study, and sample correlation of paired Z-scores can be used to estimate the correlation parameter of the bivariate normal distribution.

$$\begin{pmatrix} Z_i \\ Z_j \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{ij} \\ \rho_{ij} & 1 \end{pmatrix}\right) \qquad \text{... Equation (6)}$$

However, Z-scores at trait associated loci are expected to show positive correlation even in independent samples and using the sample correlation between observed Z-scores would lead to an over-estimation of the correlation. We also expect most traits for GWAS to be complex polygenic traits where there may be many variants with small effect sizes.

To exclude potentially causal loci, we use a cutoff c, and use markers with Z-scores in the interval $(-c, c)$ to estimate the correlation. For example, using $c = 1$ uses about 68% of the markers while excluding the more significant loci. We assume a truncated normal distribution on the Z-scores to estimate the maximum likelihood estimate of correlation, and use this to estimate the overlap. The likelihood of the observed Z-scores between studies i and j is:

$$L = \prod_{m} \frac{\phi(\boldsymbol{Z_{im}}, \boldsymbol{Z_{jm}} \mid \rho_{ij})}{P(|\boldsymbol{Z_{im}}| < c, |\boldsymbol{Z_{jm}}| < c \mid \rho_{ij})} \qquad \text{... Equation (7)}$$

where m ranges over all the markers present in both studies, and the Z-scores are assumed to follow a bivariate normal distribution with mean 0, variance 1 and correlation ρ_{ij}.

The estimated correlation obtained from (7) is then used in (5) to meta-analyze the studies by modifying the weights to for overlap.

Stratification Based on Sample Size of Marker

For a pair of studies, if all markers are present in both studies, the overlap number is the same for each marker. However, it may happen that sample size varies across markers as some markers may be present only in a sub-cohort of a study. For example, **Figure 4.2** describes a simple scenario where a pair of studies have a cohort overlapping (cohort 2). Markers absent in this overlapping cohort 2 would have an overlap of 0, and so they should be meta-analyzed without correcting for overlap. However, markers present in the overlapping cohort should be meta-analyzed after correcting for an overlap the size of cohort 2.

Two problems arise if the overlapping number varies by marker. First, the estimated total covariance is biased downward by the markers where there is no overlap and we may apply an insufficient adjustment at many markers, leading to false signals. Secondly, when applying a constant correction for overlap, we may over-correct at markers with no overlap and lose power.

Ideally, clustering methods such as k-means clustering can be used to stratify the total sample size at each marker and works well when comparing a pair of studies. When many studies are included in a meta-analysis there may be a broad range of sample sizes (**Figure 4.3**) and using less refined clustering improves computational efficiency. Thus, we use markers that have at least 50% of total sample size, and bin them using relatively broad bin sizes. Then we

estimate the correlation at each stratified level using (7) to estimate the overlap for that group of markers, and then correctly meta-analyze using (5).

Using Pair-wise Correlation of Z-scores to Estimate Effective Overlap Size

Consider a pair of studies with sample sizes n_1 and n_2, and suppose that the trait under investigation is independent of genetic effects. Then, we expect the Z-scores to be distributed as $N(0,1)$ for both studies. Let n_{12} be the number of samples overlapping between the two studies. Now, the Z-scores for each study can be considered as a weighted sum of the Z-scores for the overlapping and non-overlapping parts. Assuming the weights are proportional to the sample size as follows:

$$Z_1 = \sqrt{(1-p_1)}Z_A + \sqrt{p_1}Z_C \qquad \text{... Equation (8)}$$

$$Z_2 = \sqrt{(1-p_2)}Z_A + \sqrt{p_2}Z_C \qquad \text{... Equation (9)}$$

where the weights used are $p_1 = n_{12}/n_1$ and $p_2 = n_{12}/n_2$, that is, the overlap proportions in each study and Z_A, Z_B, Z_C are standard normal variables. Then,

$$\text{Cov}(Z_1, Z_2) = E(Z_1 Z_2) = \sqrt{p_1 p_2} \qquad \text{... Equation (10)}$$

Thus, as the Z-scores have variance 1,

$$\text{Cor}(Z_1, Z_2) = \sqrt{p_1 p_2} = n_{12}/\sqrt{n_1 n_2} \qquad \text{... Equation (11)}$$

Hence, the effective overlapping number can be estimated using the sample correlation r_{12} between the Z-scores of the 2 studies as follows:

$$\hat{n}_{12} = \sqrt{n_1 n_2} r_{12} \qquad \text{... Equation (12)}$$

In case of GWAS where the trait is not independent of genetic effects, the estimated correlation from (7) can be used in (12) to get an estimate of the effective sample size.

Observe that (12) estimates the effective sample overlap which may be different from the actual sample overlap. For example, for two case-control studies k and l, the estimated correlation corresponds to:

$$\mathrm{Cor}(Z_k, Z_l) \approx \frac{\left(n_{kl0}\sqrt{\frac{n_{k1}n_{l1}}{n_{k0}n_{l0}}} + n_{kl1}\sqrt{\frac{n_{k0}n_{l0}}{n_{k1}n_{l1}}}\right)}{\sqrt{n_k n_l}} \quad \ldots \text{Equation (13)}$$

where 1 refers to cases and 0 to controls (Lin and Sullivan 2009).

Hence, the estimated effective overlap sample size ($\hat{n}_{kl} = \sqrt{n_k n_l}\hat{r}_{kl}$) may correspond to a range of actual overlap numbers. We can readily derive two extreme possibilities. First, when the overlap is restricted to the cases, $\sqrt{\frac{n_{k1}n_{l1}}{n_{k0}n_{k1}}}\hat{n}_{kl}$ is a point estimate of the number of overlapping samples. Second, when the overlap is restricted to the controls, $\sqrt{\frac{n_{k0}n_{l0}}{n_{k1}n_{l1}}}\hat{n}_{kl}$ is an alternative point estimate of the overlap.

Similar issues may arise in GWAS for quantitative traits if overlap proportions vary by phenotype values. For example, if overlap is concentrated in participants with extremely high phenotype, the estimated effective overlap may be an over-estimate. Note that while the estimated correlation may correspond to a range of overlap proportions, the adjustments to the weights in (5) are still valid.

Meta-Analysis of Multiple Studies

Multiple studies are meta-analyzed sequentially, that is, each new study is meta-analyzed with the result from meta-analyzing the previous studies. For each marker for a pair of studies i and j, we meta-analyze them as described above and calculated the following quantities:

$$\text{Total Weight } W = \sqrt{w_i^2 + w_j^2 + 2 * w_i w_j r_{ij}}$$

Effective Sample Size $N = n_i + n_j - n_{ij}$

$$Z = \frac{1}{W}(w_i Z_i + w_j Z_j)$$

Observe that this ensures that the order the studies are analyzed in doesn't affect the results.