Question: What is the correlation of zscores of two genes, in the case that zscores are weighted meta zscores from different cohorts?

Assume we consider gene j and gene k, and,

$$Cor(Z_j, Z_k) = Cov(Z_j, Z_k) = \rho_{jk}$$

Let's consider the case where the zscore comes from meta analysis of C cohorts, i.e.

$$\tilde{Z}_j = \sum_{c=1}^{C} w_c Z_{cj}$$

, where Z_{cj} is the zscore of gene j using cohort c, w_c is the weights, typically based on sample size ratio. Let's assume $\sum_{c=1}^{C} w_c^2 = 1$.

Say, we consider the case where the number of corhorts used for $\tilde{Z}_j$ and $\tilde{Z}_k$ are different. For simplicity, assume 2 cohorts for gene j and 3 cohorts for gene k. Then,

$$\begin{aligned}\tilde{Z}_j &= w_1 Z_{1j} + w_2 Z_{2j} \\ \tilde{Z}_k &= w_1^* Z_{1k} + w_2^* Z_{2k} + w_3^* Z_{3k}\end{aligned}$$

, where $w_1^* = \sqrt{\frac{N_1}{N_1+N_2+N_3}}$ is for the weight of cohort1, but different from $w_1 = \sqrt{\frac{N_1}{N_1+N_2}}$ as the total sample size has changed.

Next, let's see the correlation between the meta-ed zscores.

$$\begin{aligned}Var(\tilde{Z}_j) &= Var(w_1 Z_{1j} + w_2 Z_{2j}) = 1 \\ Var(\tilde{Z}_k) &= Var(w_1^* Z_{1k} + w_2^* Z_{2k} + w_3^* Z_{3k}) = 1\end{aligned}$$

, because $w_1^2 + w_2^2 = 1$, and $Var(Z_{1j}) = Var(Z_{2j}) = 1$, and also $cov(Z_{1j}, Z_{2j}) = 0$ as cohort1 and cohort2 are independent. Same for $\tilde{Z}_k$.

Then, the correlation becomes,

$$\begin{aligned}Cor(\tilde{Z}_j, \tilde{Z}_k) &= Cov(\tilde{Z}_j, \tilde{Z}_k) \\ &= Cov(w_1 Z_{1j} + w_2 Z_{2j}, w_1^* Z_{1k} + w_2^* Z_{2k} + w_3^* Z_{3k}) \\ &= (w_1 w_1^* + w_2 w_2^*)\rho_{jk} \\ &< \rho_{jk}\end{aligned}$$

Conclusion

0. The correlations among eQTLGen zscores are complicated.

1. For meta-ed zscores, their correlations is always smaller than the original correlation. Denote $Cor(\tilde{Z}_j, \tilde{Z}_k) = a\rho_{jk}$, $a < 1$ and depends on the overlap of cohorts used for the meta-ed zscores.

2. For different gene pairs, $Cor(\tilde{Z}_j, \tilde{Z}_k) \neq Cor(\tilde{Z}_m, \tilde{Z}_n)$ can occur, if the overlap of the cohorts for these gene pairs are different.

3. For different SNPs. The underlying reason why we can calculate $Cor(Z_j, Z_k)$ for two genes using the sample correlation of zscores of independent SNPs $Cor(\begin{bmatrix} Z_j^{snp_1} \\ \vdots \\ Z_j^{snp_M} \end{bmatrix}, \begin{bmatrix} Z_k^{snp_1} \\ \vdots \\ Z_k^{snp_M} \end{bmatrix})$, is that we assume these SNPs' zscores follow the same distribution and have same correlation, i.e. $cor(Z_j^{snp_1}, Z_k^{snp_1}) = \cdots = cor(Z_j^{snp_M}, Z_k^{snp_M})$. However, this can also be violated if the zscores of these SNPs are calculated from different cohorts, leading to different a.