Web-based Supplementary Materials for "More powerful genetic association testing via a new statistical framework for integrative genomics"

by Sihai D. Zhao, T. Tony Cai, and Hongzhe Li

Web Appendix A: Integration in the presence of direct effects

Power for SNPs with direct effects

In Section 3 we showed that our method can have more power to detect o-eSNPs. Here we discuss its power to detect SNPs whose functional mechanisms have non-regulatory components. For simplicity we again consider only continuous Y_i, $\mathbf{G}_i$, and a single SNP S_i in the ordinary linear model, where the variables have all been centered. We now consider the outcome model $Y_i = \mathbf{G}_i^T \boldsymbol{\alpha}_G + \alpha_S S_i + \epsilon_{i1}$, where the direct effect α_S is nonzero. The transcript model remains $\mathbf{G}_i^T \boldsymbol{\alpha}_G = \beta_S S_i + \epsilon_{i2}$. We again compare to the usual approach of fitting $Y_i = \beta_S^* S_i + N(0, \sigma^{*2})$.

We are interested in comparing the power of tests based on $\hat{\beta}_S$ and $\hat{\beta}_S^*$ under this direct effect model, so we study $\beta_S / \mathrm{var}\,(\hat{\beta}_S)$ and $\beta_S^* / \mathrm{var}\,(\hat{\beta}_S^*)$. First, we still have $\mathrm{var}\,(\hat{\beta}_S^*) = (\sigma_1^2 + \sigma_2^2)/\mathrm{var}\,(S_i)$. To calculate $\mathrm{var}\,(\hat{\beta}_S)$ note that Step 1 of our integrative procedure is equivalent to marginalizing over S_i in the true outcome model, which gives $\mathrm{E}(Y_i \mid \mathbf{G}_i) = \mathbf{G}_i^T \boldsymbol{\alpha}_G + \alpha_S \mathrm{E}(S_i \mid \mathbf{G}_i)$ and $\mathrm{var}\,(Y_i \mid \mathbf{G}_i) = \alpha_S^2 \mathrm{var}\,(S_i \mid \mathbf{G}_i) + \sigma_1^2$. Without knowing more about the distribution of S_i given $\mathbf{G}_i$ it is hard to draw further conclusions. But when S_i is weakly correlated with $\mathbf{G}_i$, the outcome model is approximately correctly specified and

$$\mathrm{var}\,(\hat{\beta}_S) \approx \sigma_2^2/\mathrm{var}\,(S_i) + \{\sigma_1^2 + \alpha_S^2 \mathrm{var}\,(S_i \mid \mathbf{G}_i)\} \boldsymbol{\Sigma}_{SG} \boldsymbol{\Sigma}_{GG}^{-1} \boldsymbol{\Sigma}_{GS} / \mathrm{var}\,(S_i)^2. \tag{1}$$

Second, $\beta_S^* = \alpha_S + \beta_S$. Thus when α_S is large relative to β_S with the same sign, standard analysis will be more powerful. Otherwise, denoting $c = \boldsymbol{\Sigma}_{SG} \boldsymbol{\Sigma}_{GG}^{-1} \boldsymbol{\Sigma}_{GS} / \mathrm{var}\,(S_i)$, $\mathrm{var}\,(\hat{\beta}_S) < \mathrm{var}\,(\hat{\beta}_S^*)$ if $\alpha_S^2 < \sigma_1^2(1 - c)/\mathrm{var}\,(S_i \mid \mathbf{G}_i)c$. Since c is small when S_i and $\mathbf{G}_i$ are weakly correlated and σ_1^2 tends to be large, our approach will still have more power unless α_S is very large, though when it is so large that $\alpha_S \mathrm{E}(S_i \mid \mathbf{G}_i)$ is not close to zero the outcome model will not be approximately correctly specified.

Alternative integrative procedures

One reviewer raised the question of whether accounting for these direct effects in Step 1 of our procedure might improve the power of our integrative approach. We consider testing β_S

in two alternative integrative models:

$$\begin{aligned} Y_i &= \mathbf{G}_i^T \boldsymbol{\alpha}_G + \alpha_S S_i + \epsilon_{i1} \\ \mathbf{G}_i^T \boldsymbol{\alpha}_G + \alpha_S S_i &= \beta_S S_i + \epsilon_{i2}, \end{aligned} \tag{2}$$

which we refer to as the *total effect* approach, because β_S now encompasses both the direct and regulatory effects of S_i on Y_i, and

$$\begin{aligned} Y_i &= \mathbf{G}_i^T \boldsymbol{\alpha}_G + \alpha_S S_i + \epsilon_{i1} \\ \mathbf{G}_i^T \boldsymbol{\alpha}_G &= \beta_S S_i + \epsilon_{i2}, \end{aligned} \tag{3}$$

which we refer to as the *conditional* approach, because β_S is the regulatory effect of S_i conditional on the presence of the direct effect.

To analyze (2) and (3) we require the covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\alpha}}_G^T, \hat{\alpha}_S)^T$, which equals

$$\sigma_1^2 \begin{pmatrix} \boldsymbol{\Sigma}_{GG} & \boldsymbol{\Sigma}_{GS} \\ \boldsymbol{\Sigma}_{SG} & \mathrm{var}\,(S_i) \end{pmatrix}^{-1} = \sigma_1^2 \begin{pmatrix} \boldsymbol{\Sigma}_{GG}^{-1} + k^{-1}\boldsymbol{\Sigma}_{GG}^{-1}\boldsymbol{\Sigma}_{GS}\boldsymbol{\Sigma}_{SG}\boldsymbol{\Sigma}_{GG}^{-1} & -k^{-1}\boldsymbol{\Sigma}_{GG}^{-1}\boldsymbol{\Sigma}_{GS} \\ -k^{-1}\boldsymbol{\Sigma}_{SG}\boldsymbol{\Sigma}_{GG}^{-1} & k^{-1} \end{pmatrix},$$

where $k = \mathrm{var}\,(S_i) - \boldsymbol{\Sigma}_{SG}\boldsymbol{\Sigma}_{GG}^{-1}\boldsymbol{\Sigma}_{GS}$. We again let $c = \boldsymbol{\Sigma}_{SG}\boldsymbol{\Sigma}_{GG}^{-1}\boldsymbol{\Sigma}_{GS}/\mathrm{var}\,(S_i)$ so that $k = (1-c)\mathrm{var}\,(S_i)$. The inverse of the covariance matrix has determinant $\det(\boldsymbol{\Sigma}_{GG})k$, which must positive because the matrix is positive-definite. Therefore $k > 0$, which implies $c < 1$. Under (2), $\beta_S = \beta_S^*$, and our calculations in Section 3 imply that

$$\mathrm{var}\,(\hat{\beta}_S) = \sigma_2^2/\mathrm{var}\,(S_i) + \sigma_1^2\{c/\mathrm{var}\,(S_i) + k^{-1}(1-c)^2\} = \mathrm{var}\,(\hat{\beta}_S^*).$$

Thus the total effect approach provides no advantage over standard association analysis.

Under (3) similar calculations imply that

$$\mathrm{var}\,(\hat{\beta}_S) = \sigma_2^2/\mathrm{var}\,(S_i) + \sigma_1^2\{c/\mathrm{var}\,(S_i) + k^{-1}c^2\} = \sigma_2^2/\mathrm{var}\,(S_i) + \sigma_1^2 c/(1-c)\mathrm{var}\,(S_i).$$

Compared to the variance (1) of our original integrative procedure without the direct effect, the conditional approach will be less powerful if $\alpha_s^2 < \sigma_1^2 c/(1-c)\mathrm{var}\,(S_i \mid \mathbf{G}_i)$. Thus if the direct effect is large, the conditional approach may have more power, but it will always be worse than our original procedure for SNPs with no direct effect.

Simulations and data analysis

We applied the conditional approach to the simulated data studied in the main paper. Table 1 reports the type I errors and shows that the conditional approach also maintains the nominal error rate. Figures 1 and 2 show that for continuous Y_i its power is nearly identical to that of our integrative procedure without the direct effect, but for binary Y_i it is noticeably less powerful even in Example 2, which was simulated with a direct effect. Table 2 shows that it is much less powerful in high dimensions, most likely because the data were simulated without direct effects.

Finally, we also used the conditional approach to analyze the yeast data studied in the main paper. We discovered no significant SNPs after Bonferroni correction, but this may be because we only tested pairs of genes and SNPs located in *cis*. Our analysis of the conditional approach suggests that it may outperform our original integrative analysis formulation when the direct effect is large, but *cis*-SNPs tend not to have large direct effects. Our *cis*-pairwise approach may not be optimal for applying the conditional approach to high-dimensional genomic data.

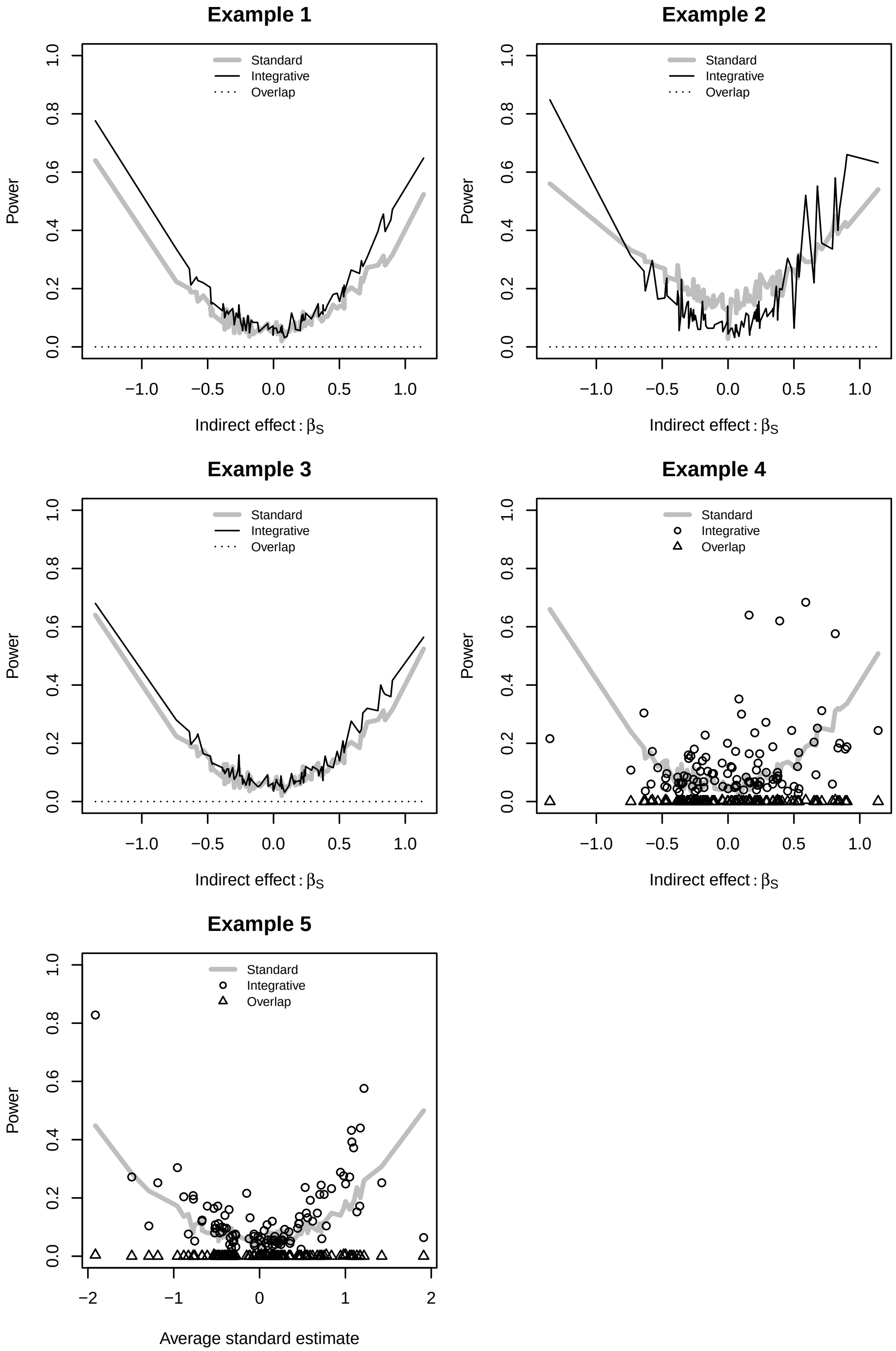

Figure 1: Average power curves for linear outcomes using conditional integrative approach. Integration: proposed method; Standard: standard univariate regression analysis; Overlap: overlap method.

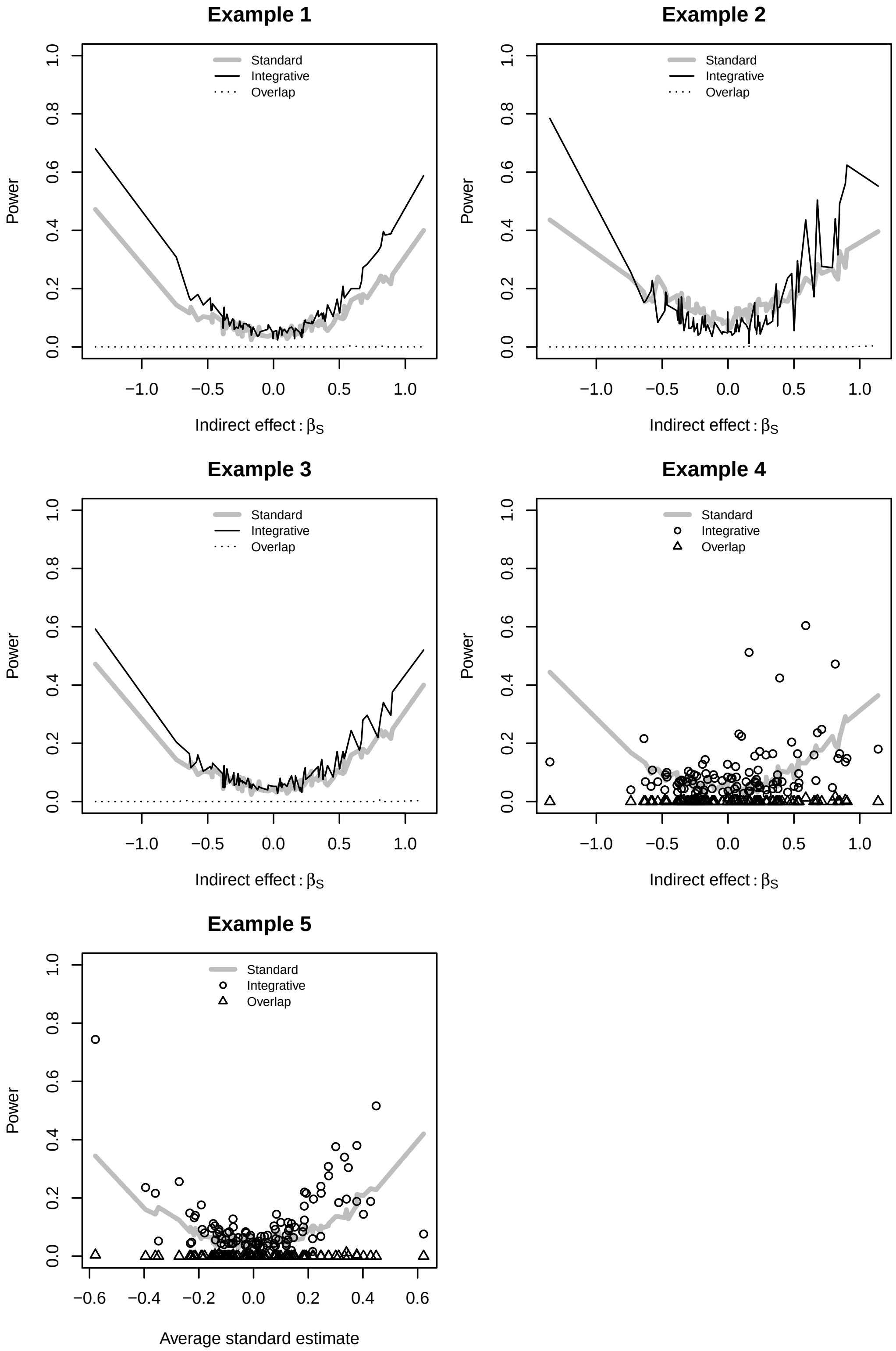

Figure 2: Average power curves for binary outcomes using conditional integrative approach. Integration: proposed method; Standard: standard univariate regression analysis; Overlap: overlap method.

Table 1: Average type I errors at nominal 0.05 level using conditional integrative approach. Integration: proposed method; Standard: standard univariate regression analysis; Overlap: overlap method.

		Linear			Binary	
Example	Integrative	Standard	Overlap	Integrative	Standard	Overlap
1	0.064	0.052	0.000	0.052	0.040	0.000
2	0.044	0.028	0.000	0.052	0.060	0.000
3	0.036	0.052	0.000	0.052	0.040	0.000
4	0.044	0.048	0.000	0.036	0.024	0.012
5	0.064	0.056	0.000	0.064	0.028	0.000

Table 2: SNP detection in high-dimensions (Example 6), after Bonferroni correction to give a family-wise error rate of 0.05. We simulated a total of 14 o-eSNPs. Integration: proposed method, 20,000 tests; Standard: standard univariate regression analysis, 10,000 tests. Performance metrics (SD): TP = true positive rate, FD = false discovery rate; Median size is reported (interquartile range).

Outcome	Method	TP	FD	Size
Continuous	Integration	29.71(7.81)	1.87(6.1)	4(2)
	Standard	1.14(2.85)	5.2(22.25)	0(0)
Binary	Integration	8.17(6.22)	0.2(3.16)	1(1)
	Standard	0.14(1)	0(0)	0(0)