Web-based Supplementary Materials for "Controlling False Discoveries in Multidimensional Directional Decisions, with Applications to Gene Expression Data on Ordered Categories"

Wenge Guo

Biostatistics Branch, National Institute of Environmental Health Sciences

Research Triangle Park, NC 27709, U.S.A.

email: wenge.guo@gmail.com

and

Sanat K. Sarkar

Department of Statistics, Temple University, Philadelphia, PA 19122, U.S.A.

email: sanat@temple.edu

and

Shyamal D. Peddada

Biostatistics Branch, National Institute of Environmental Health Sciences

Research Triangle Park, NC 27709, U.S.A.

email: peddada@niehs.nih.gov

Web Appendix A: Proof of Theorem 1

We begin by calculating the pure directional FDR (dFDR), $E\left\{\frac{S}{R\vee 1}\right\}$. Let $I_1 = \{1 \leqslant j \leqslant m : \boldsymbol{\delta}_j \neq \mathbf{0}\}$ be the set of indices of false null hypotheses, then S can be expressed as

$$S = \sum_{j\in I_1} I\left(\bigcup_{i=1}^{q}\left(\tilde{P}_{ij} \leqslant \frac{R}{qm}\alpha, T_{ij}\delta_{ij} \leqslant 0\right)\right),$$

where $I(\cdot)$ is indicator function. Thus

$$\begin{aligned} dFDR = E\left\{\frac{S}{R\vee 1}\right\} &= E\left\{\frac{E\left(S \mid R\right)}{R\vee 1}\right\} \\ &= E\left[\frac{\sum_{j\in I_1} Pr\left\{\bigcup_{i=1}^{q}\left(\widetilde{P}_{ij} \leqslant \frac{R}{qm}\alpha, T_{ij}\delta_{ij} \leqslant 0\right) \,\middle|\, R\right\}}{R\vee 1}\right] \\ &= \sum_{r=1}^{m}\sum_{j\in I_1}\frac{1}{r}Pr\left\{\bigcup_{i=1}^{q}\left(\widetilde{P}_{ij} \leqslant \frac{r}{qm}\alpha, T_{ij}\delta_{ij} \leqslant 0, R = r\right)\right\} \\ &\leqslant \sum_{r=1}^{m}\sum_{j\in I_1}\sum_{i=1}^{q}\frac{1}{r}Pr\left\{\widetilde{P}_{ij} \leqslant \frac{r}{qm}\alpha, T_{ij}\delta_{ij} \leqslant 0, R = r\right\}. \end{aligned} \tag{A.1}$$

The inequality follows from the Bonferroni inequality.

For any given i and j, without loss of generality, we assume $\delta_{ij} \geqslant 0$. When $\delta_{ij} > 0$, we have

$$\begin{aligned} & Pr\left\{\widetilde{P}_{ij} \leqslant \frac{r}{qm}\alpha, T_{ij}\delta_{ij} \leqslant 0, R = r\right\} \\ = \; & Pr\left\{\widetilde{P}_{ij} \leqslant \frac{r}{qm}\alpha, T_{ij} \leqslant 0, R = r\right\} \\ \leqslant \; & Pr\left\{F_{ij}(T_{ij}, 0) \leqslant \frac{r}{2qm}\alpha, R = r\right\} \\ = \; & Pr\left\{T_{ij} \leqslant F_{ij}^{-1}\left(\frac{r}{2qm}\alpha, 0\right), R = r\right\}, \end{aligned} \tag{A.2}$$

where $F_{ij}^{-1}(\cdot, 0)$ is the inverse function of $F_{ij}(\cdot, 0)$. The inequality in the above calculations follows from the definition of $\widetilde{P}_{ij}$ and the assumption $F_{ij}(0, 0) = \frac{1}{2}$.

Noting that $\mathbf{T}_j = (T_{1j}, \cdots, T_{qj}), j = 1, \cdots, m$, are independent of each other, the last probability in (A.2) can be simplified to

$$\begin{aligned} & Pr\left\{T_{ij} \leqslant F_{ij}^{-1}\left(\frac{r}{2qm}\alpha, 0\right)\right\} \cdot Pr\left(R^{(-j)} = r - 1\right) \\ = \; & F_{ij}\left(F_{ij}^{-1}\left(\frac{r}{2qm}\alpha, 0\right), \delta_{ij}\right) \cdot Pr\left(R^{(-j)} = r - 1\right) \end{aligned}$$

$$\begin{aligned} &\leqslant F_{ij}\left(F_{ij}^{-1}\left(\frac{r}{2qm}\alpha,0\right),0\right)\cdot Pr\left(R^{(-j)}=r-1\right)\\ &= \frac{r}{2qm}\alpha\cdot Pr\left(R^{(-j)}=r-1\right), \end{aligned} \tag{A.3}$$

where $R^{(-j)}$ denotes the number of rejections in the stepup procedure with critical constants $\alpha_k = \frac{k+1}{m}\alpha, k = 1,\cdots,m-1$ based on $\{P_1,\cdots,P_m\}\setminus\{P_j\}$. The above inequality follows from the assumption that $F_{ij}(\cdot,\delta_{ij})$ is stochastically increasing in $\delta_{ij}\geqslant 0$.

Similarly, when $\delta_{ij}=0$, we have

$$\begin{aligned} &Pr\left\{\widetilde{P}_{ij}\leqslant\frac{r}{qm}\alpha, T_{ij}\delta_{ij}\leqslant 0, R=r\right\}\\ =\ &Pr\left\{\widetilde{P}_{ij}\leqslant\frac{r}{qm}\alpha, R=r\right\}\\ =\ &Pr\left\{\widetilde{P}_{ij}\leqslant\frac{r}{qm}\alpha, R^{(-j)}=r-1\right\}\\ \leqslant\ &\frac{r}{qm}\alpha\cdot Pr\left(R^{(-j)}=r-1\right). \end{aligned} \tag{A.4}$$

The last inequality follows from the fact that the two-sided p-value $\widetilde{P}_{ij}$ satisfies the condition (2) when $\delta_{ij}=0$.

Using (A.2)–(A.4) in (A.1), we have

$$dFDR\leqslant\sum_{r=1}^{m}\sum_{j\in I_1}\sum_{i=1}^{q}\frac{\alpha}{qm}Pr\left(R^{(-j)}=r-1\right)=\frac{m_1}{m}\alpha. \tag{A.5}$$

Noting that the pooled p-values $P_j, j=1,\cdots,m$, satisfy the condition (2), then for independent p-values P_j's, the usual FDR of the q-dimensional directional BH procedure satisfies the following inequality,

$$FDR\leqslant\frac{m_0}{m}\alpha\ ; \tag{A.6}$$

see Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001) or Sarkar (2002). Combining (A.5) and (A.6), we have

$$\begin{aligned} &mdFDR = FDR + dFDR\\ \leqslant\ &\frac{m_0}{m}\alpha+\frac{m_1}{m}\alpha=\alpha, \end{aligned} \tag{A.7}$$

and hence the proof is complete. ■

Web Appendix B: Some Additional Simulation Results

In addition to evaluating the performance of Procedure 1, we also evaluated the performance of Procedure 2 in the same simulation study. Web Figure 7 presents the simulated FDR, dFDR and mdFDR and Web Figure 8 presents the average power of Procedure 2 plotted against the number of false null hypotheses for $m = 1000$, $q = 5$, $\alpha = 0.05$ and $\rho = 0$, $0.2, 0.5$ and 0.8. Comparing Figure 1 with Web Figure 7 and Web Figure 1 with Web Figure 8, we find that both procedures, Procedure 1 and Procedure 2, perform similarly.

We also used a simulation study to evaluate the performance of Procedure 2 under dependence within genes. We generated m independently distributed $(q+1)$-dimensional random normal vectors $\mathbf{Z}_1, \ldots, \mathbf{Z}_m$, where the components $Z_{ij}, j = 1, \cdots, q+1$ in each $\mathbf{Z}_i$ with $Z_{ij} \sim N(\mu_{ij}, 1)$, are dependent with compound symmetry structure or autoregressive order one structure (AR(1)), respectively, and have a correlation parameter ρ. Let $\delta_{ij} = (\mu_{i,j+1} - \mu_{ij})/\sqrt{2}$, $i = 1, \ldots, m; j = 1, \ldots, q$. Out of the m parameter vectors $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_{iq})$, $i = 1, \ldots, m$, m_0 were set to a null vector each, and all the δ_{ij}'s in 50%, 25% and 25% of the remaining $m - m_0$ $\boldsymbol{\delta}_i$'s were selected randomly from the intervals $(-0.75, 0.75)$, $(-4.25, -2.75)$ and $(2.75, 4.25)$ respectively. For each $i = 1, \cdots, m$, and $j = 1, \cdots, q$, the statistic $T_{ij} = (Z_{i,j+1} - Z_{ij})/\sqrt{2}$ for testing $H_{0i}^j : \delta_{ij} = 0$ vs. $H_{1i}^j : \delta_{ij} \neq 0$ and the corresponding two-sided p-value $\widetilde{P}_{ij} = 2\{1 - \Phi(|T_{ij}|)\}$ were then computed, where $\Phi(\cdot)$ is the standard normal cdf. The pooled p-values were calculated according to the Simes' test for each $i = 1, \ldots, m$, and Procedures 2 were applied to their respective lists of pooled p-values for testing the m null hypotheses described in (1). Similar to the above simulation study, the simulated values of the FDR, dFDR and mdFDR were obtained by repeating the simulation steps 10,000 times. Web Figures 9 and 11 provide the simulated FDR, dFDR and mdFDR of Procedure 2 plotted against the number of false null hypotheses for $m = 1,000$, $q = 5$, $\alpha = 0.05$ and $\rho = 0$, $0.1, 0.2$ and 0.3 under dependence within genes according

to compound symmetry structure and AR(1) structure, respectively. The average power of Procedure 2 under the above dependence structures, are provided in Web Figures 10 and 12, respectively. As seen from Web Figures 9 and 11, the simulated mdFDR of Procedure 2 is severely affected by dependence within genes, but it is still below the pre-specified level. In addition, as seen from Web Figures 10 and 12, there is no monotone relationship between the average power of Procedure 2 and correlation parameter ρ.

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

[Figure 8 about here.]

[Figure 9 about here.]

[Figure 10 about here.]

[Figure 11 about here.]

[Figure 12 about here.]

LIST OF FIGURES

9 Performance of Procedure 2 under dependence within genes with compound symmetry structure in terms of its control of the FDR, dFDR and mdFDR for $m = 1000, q = 5, \alpha = 0.05$, and $\rho = 0, 0.1, 0.2$ and 0.3.

10 Power of Procedure 2 under dependence within genes with compound symmetry structure for $m = 1000, q = 5, \alpha = 0.05$ and $\rho = 0, 0.1, 0.2$ and 0.3.

11 Performance of Procedure 2 under dependence within genes with AR(1) structure in terms of its control of the FDR, dFDR and mdFDR for $m = 1000, q = 5, \alpha = 0.05$, and $\rho = 0, 0.1, 0.2$ and 0.3.

12 Power of Procedure 2 under dependence within genes with AR(1) structure for $m = 1000, q = 5, \alpha = 0.05$ and $\rho = 0, 0.1, 0.2$ and 0.3.

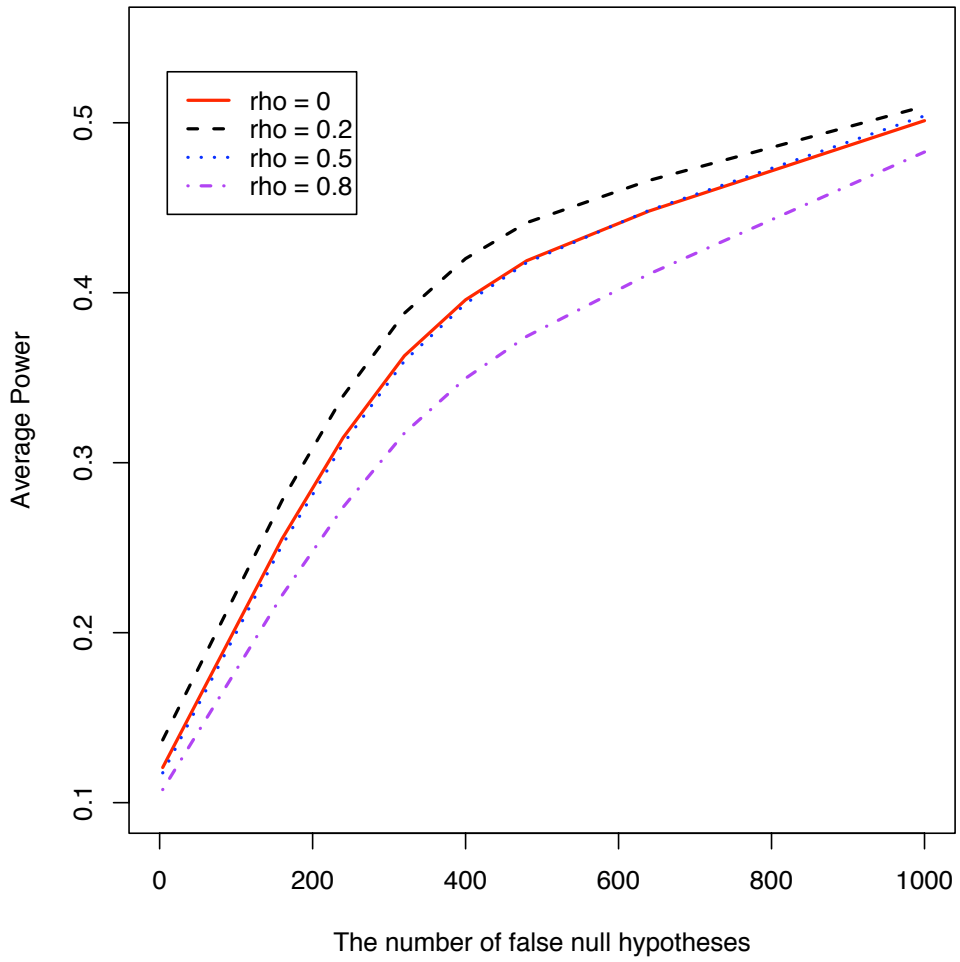

Figure 1. Power of Procedure 1 under dependence across genes for $m = 1000, q = 5, \alpha = 0.05$ and $\rho = 0, 0.2, 0.5$ and 0.8.

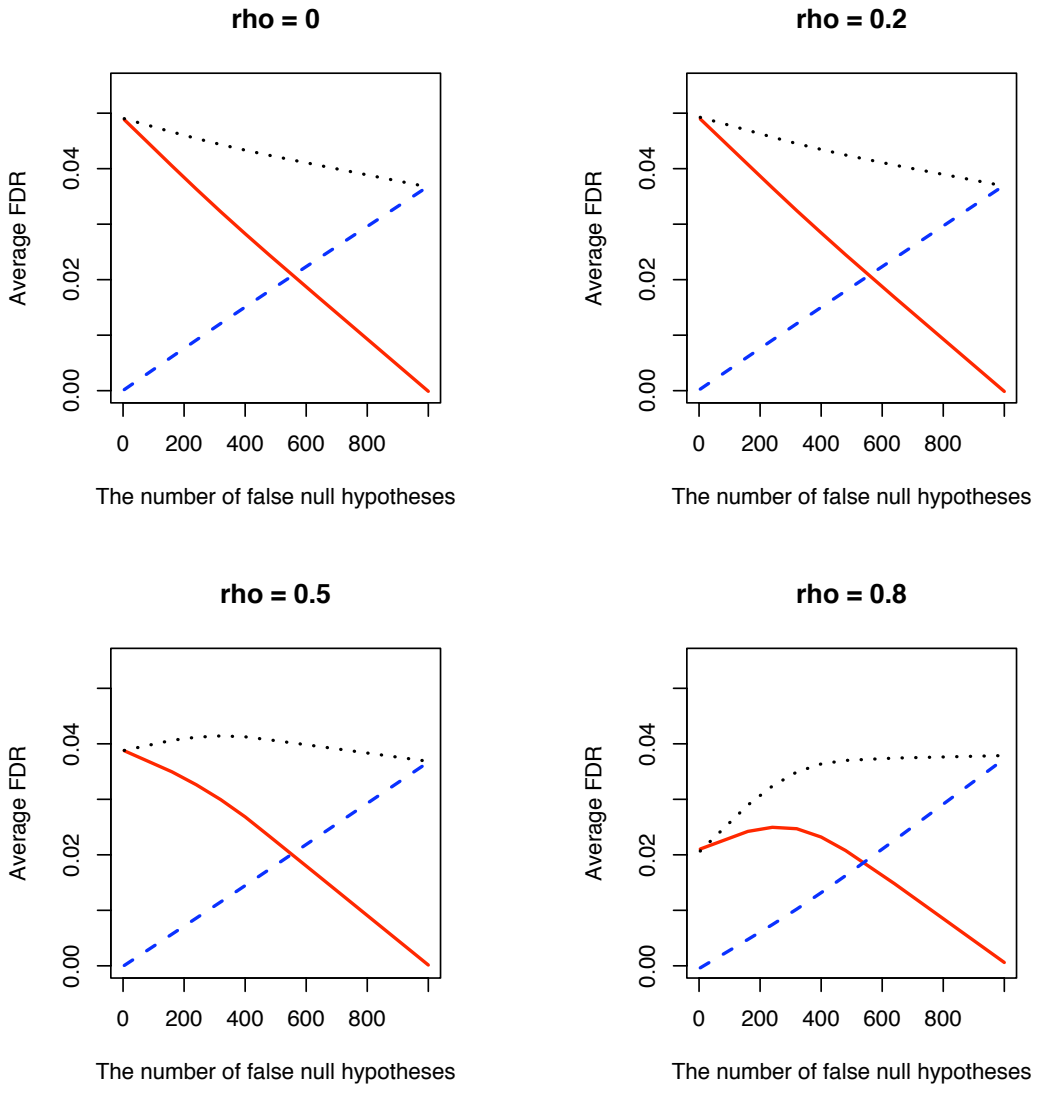

Figure 2. Performance of Procedure 1 under dependence across genes in terms of its control of the FDR (solid), dFDR (dashed) and mdFDR (dotted) for $m = 1000, q = 5, \alpha = 0.05$, $\rho = 0, 0.2, 05$ and 0.8, and $\delta = (100, 0, \ldots, 0)$.

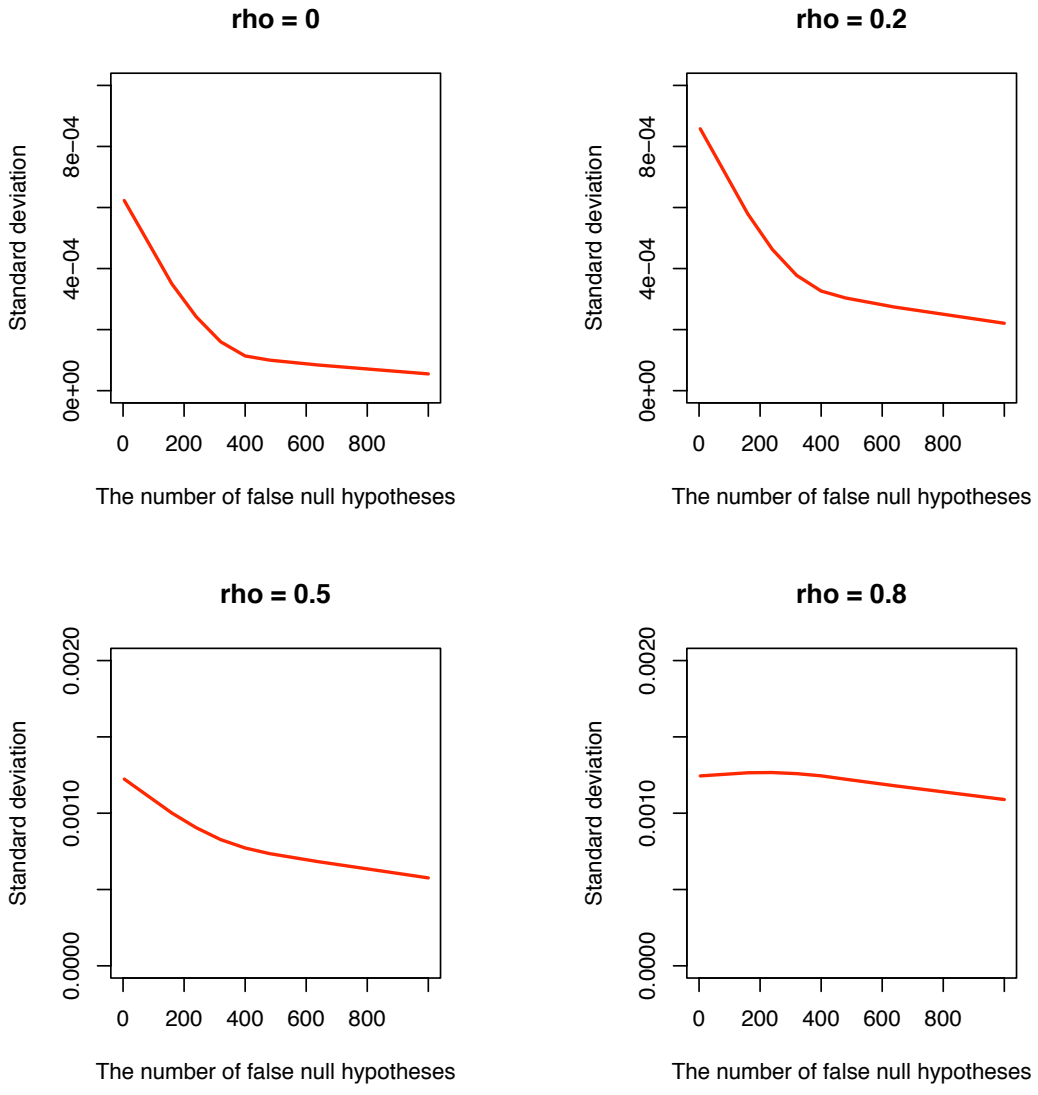

Figure 3. Standard deviation of the mdFDR of Procedure 1 under dependence across genes for $m = 1000, q = 5, \alpha = 0.05$, $\rho = 0, 0.2, 05$ and 0.8, and $\delta = (100, 0, \ldots, 0)$.

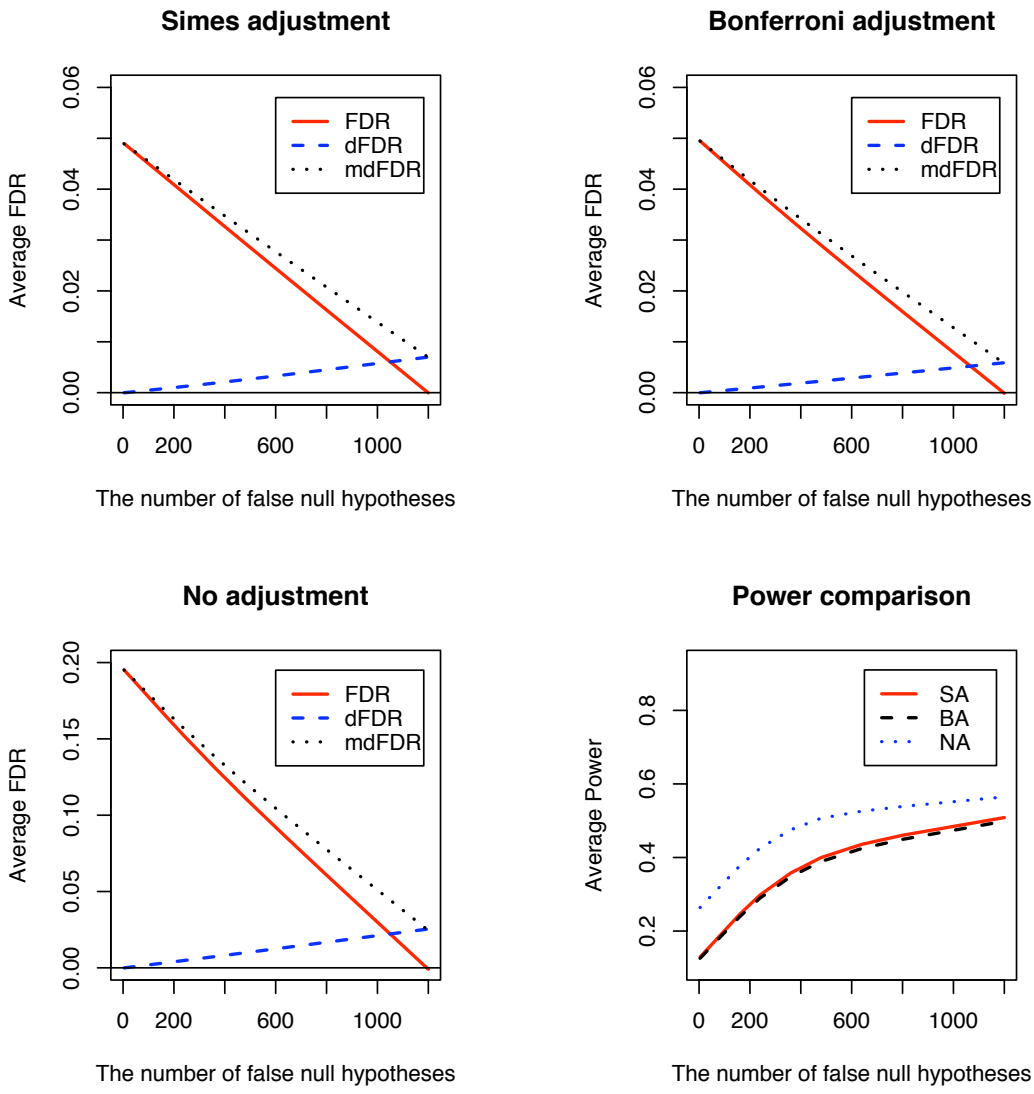

Figure 4. A numerical comparison of Procedures 1 and 2 and the 'no-adjustment' procedure in terms of the control of the FDR, dFDR and mdFDR and also power for $m = 1200, q = 4, \rho = 0,$ and $\alpha = 0.05$.

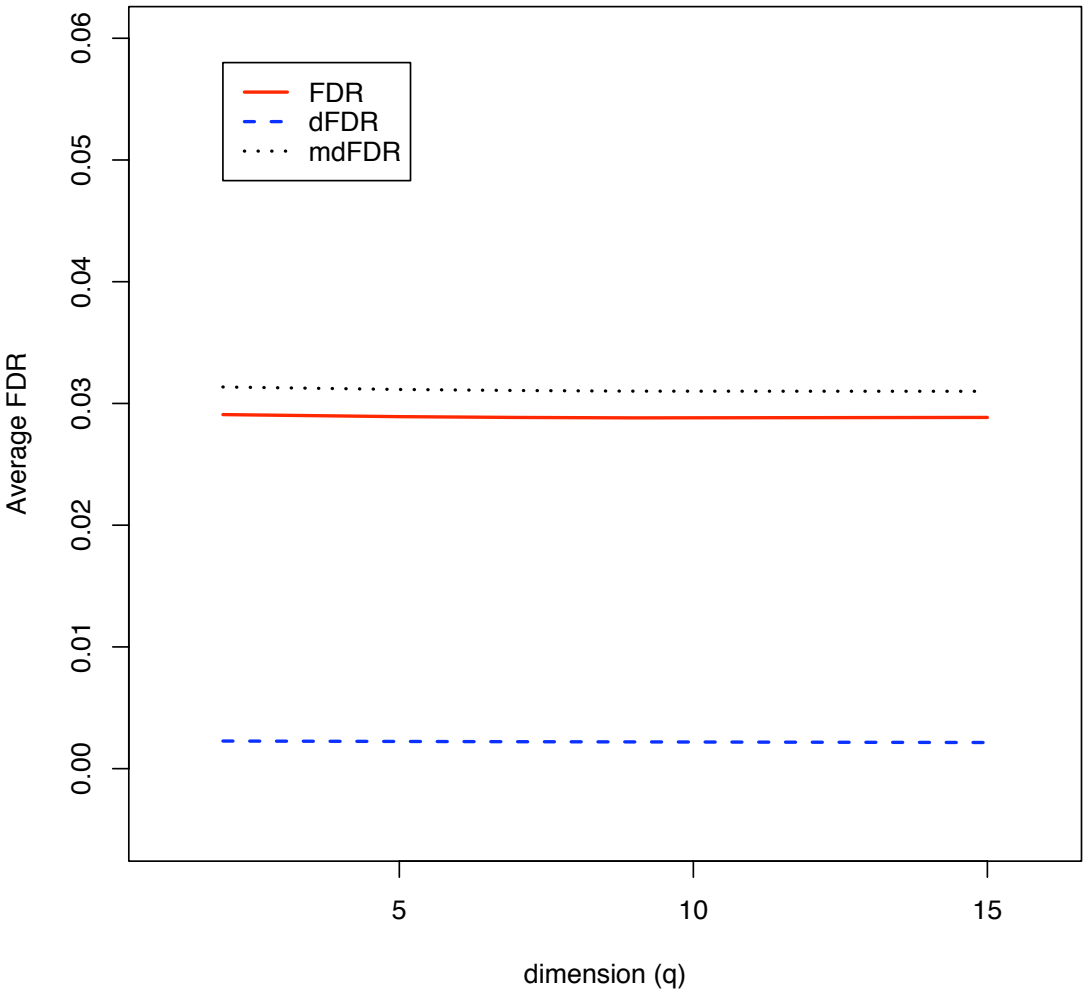

Figure 5. Performance of Procedure 1 with respect to the dimension q in terms of its control of the FDR, dFRD and mdFDR for $m = 1000, m_0 = 600, \alpha = 0.05$ and $\rho = 0$.

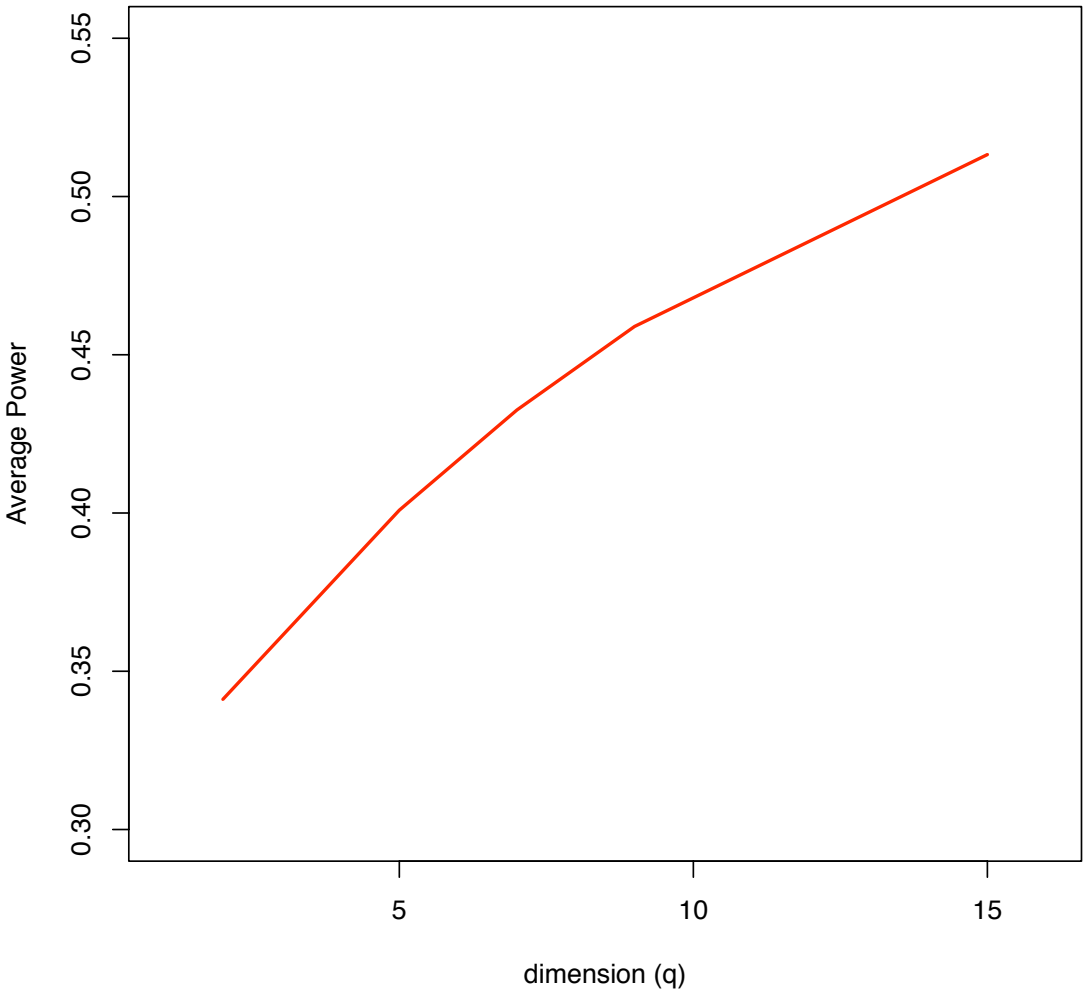

Figure 6. Power performance of Procedure 1 with respect to the dimension q for $m = 1000, m_0 = 600, \alpha = 0.05$ and $\rho = 0$ when testing $H_{0i} : \boldsymbol{\delta}_i = \mathbf{0}$ vs. $H_{1i} : \boldsymbol{\delta}_i \neq \mathbf{0}$, where $i = 1, \cdots, 1000$, $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_{iq})$ and all the δ_{ij}'s in 200, 100 and 100 of the 400 non-null $\boldsymbol{\delta}_i$'s were selected randomly from the intervals $(-0.75, 0.75)$, $(-4.25, -2.75)$ and $(2.75, 4.25)$ respectively.

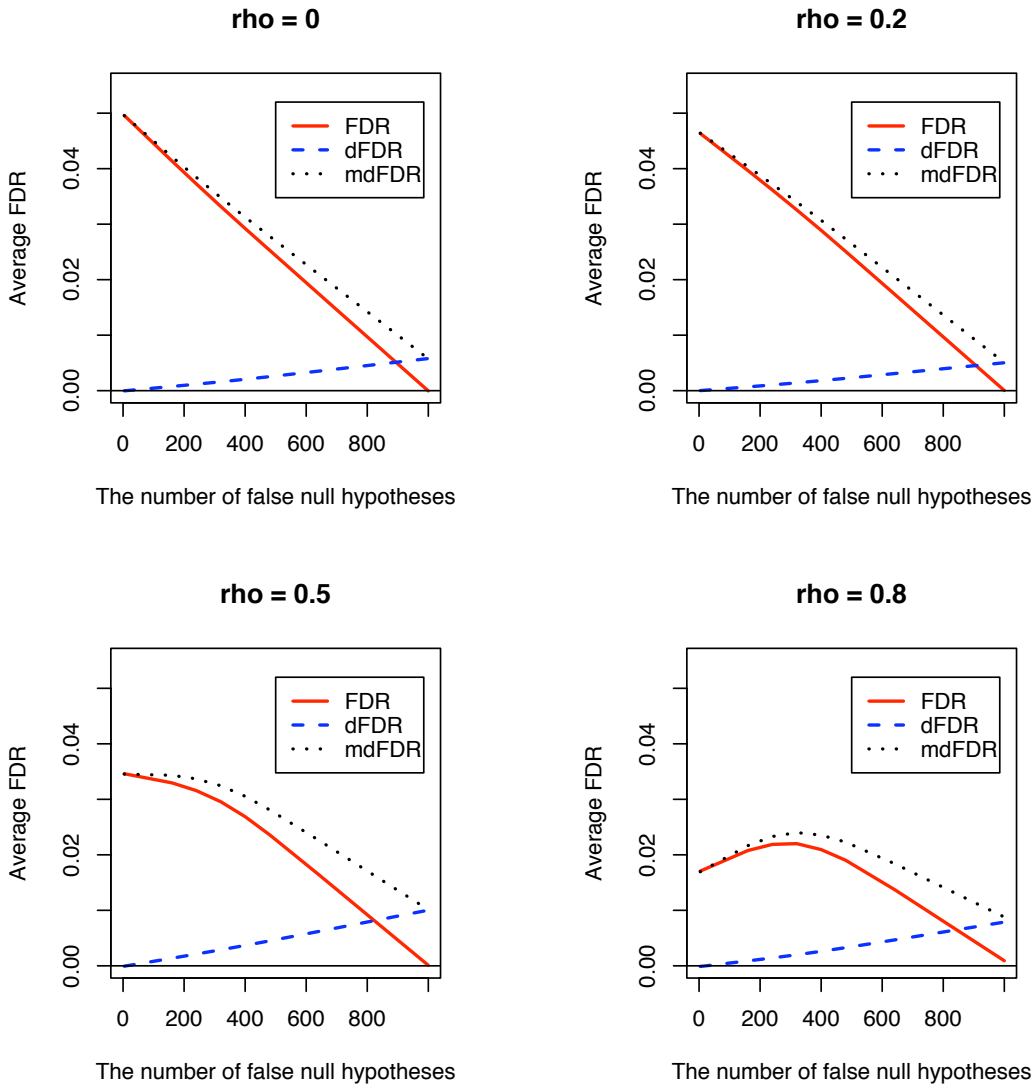

Figure 7. Performance of Procedure 2 under dependence across genes in terms of its control of the FDR, dFDR and mdFDR for $m = 1000, q = 5, \alpha = 0.05$, and $\rho = 0, 0.2, 05$ and 0.8.

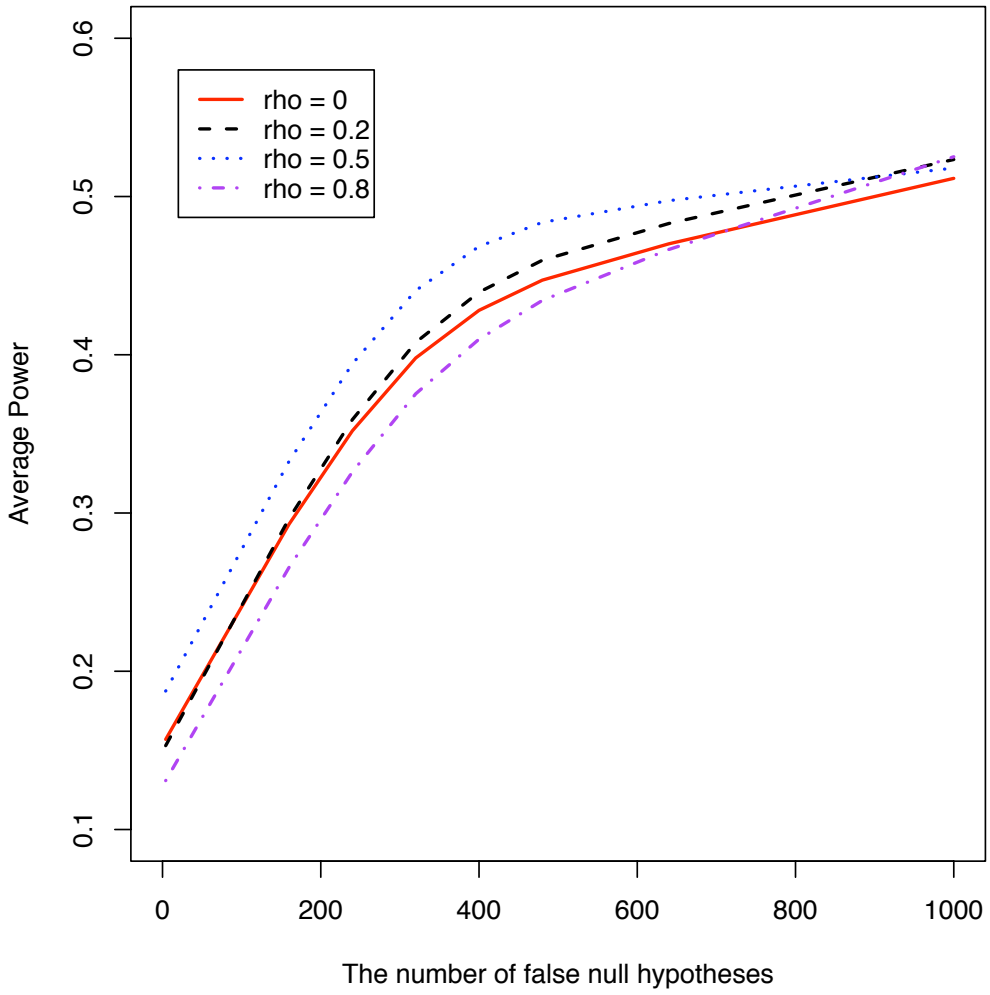

Figure 8. Power of Procedure 2 under dependence across genes for $m = 1000, q = 5, \alpha = 0.05$ and $\rho = 0, 0.2, 0.5$ and 0.8.

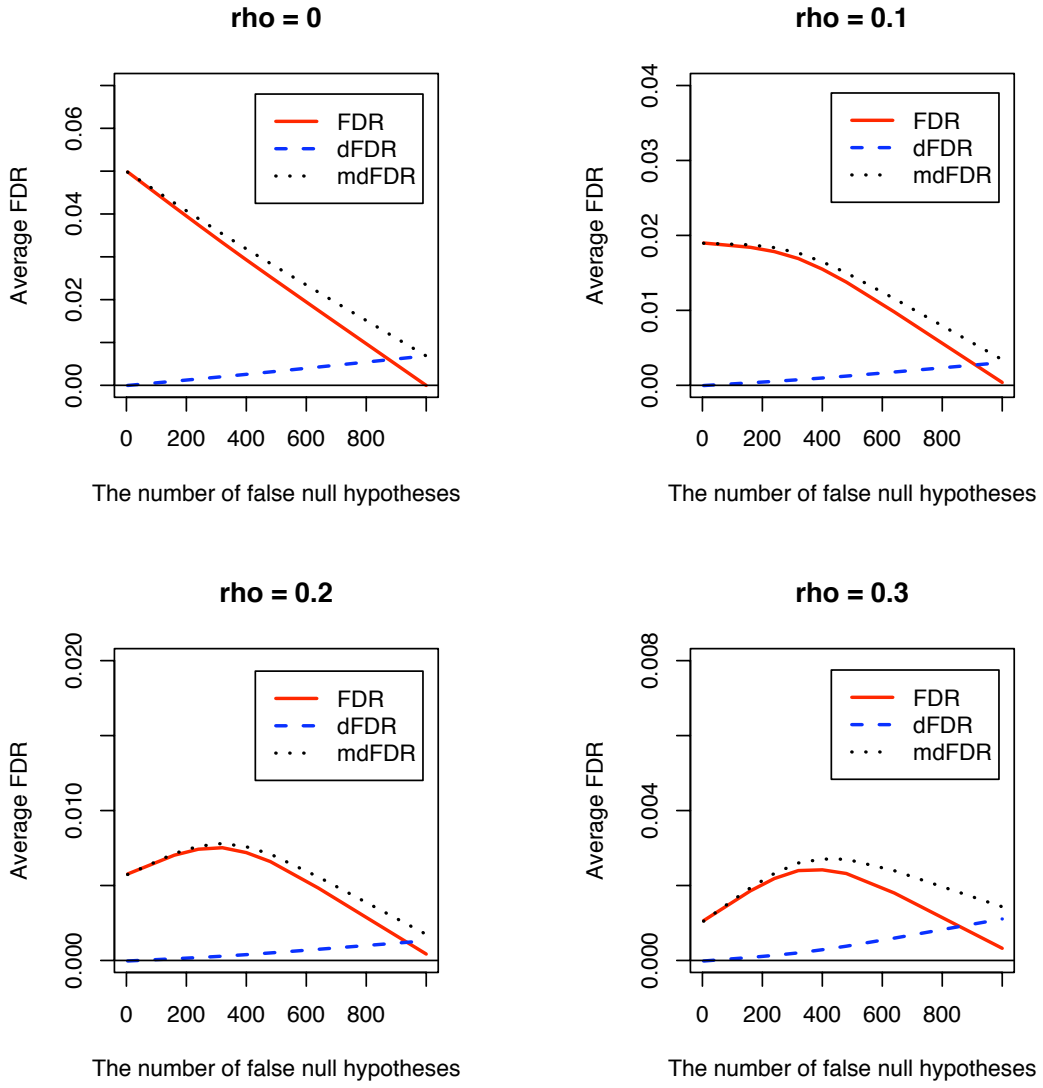

Figure 9. Performance of Procedure 2 under dependence within genes with compound symmetry structure in terms of its control of the FDR, dFDR and mdFDR for $m = 1000, q = 5, \alpha = 0.05$, and $\rho = 0, 0.1, 0.2$ and 0.3.

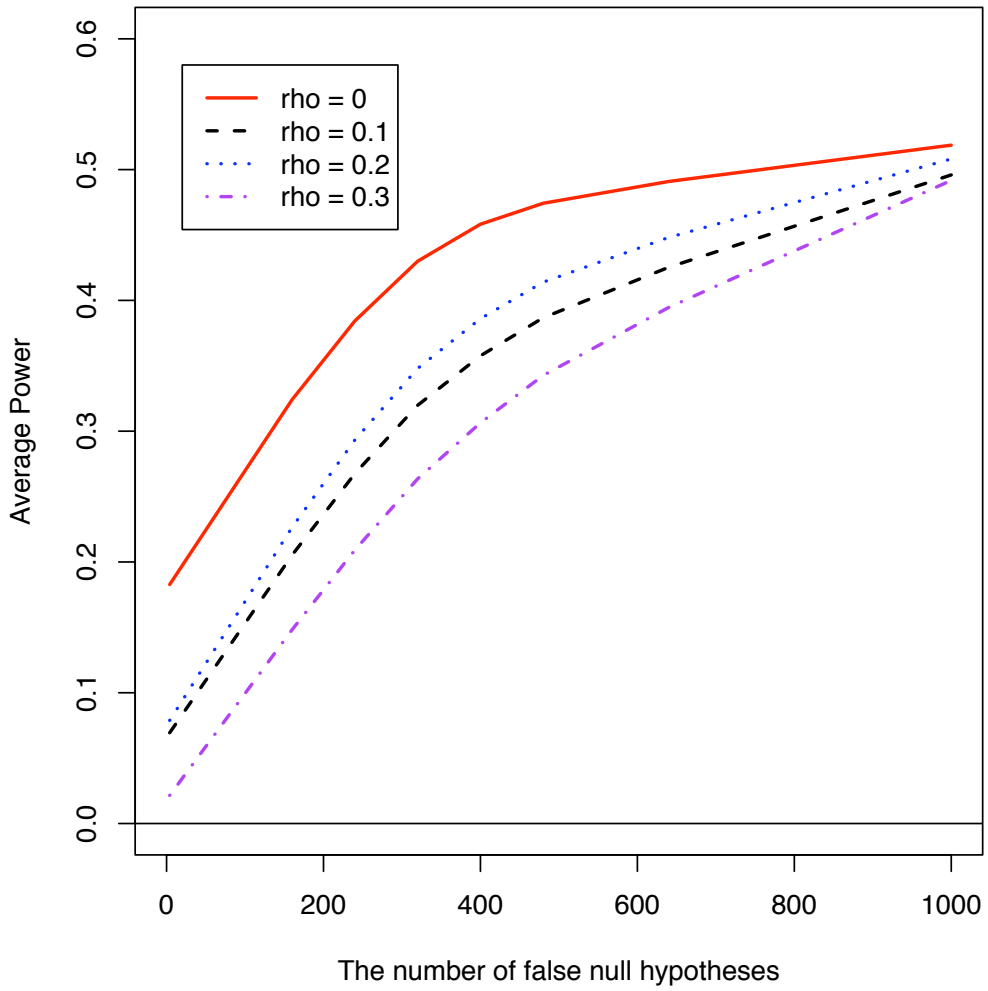

Figure 10. Power of Procedure 2 under dependence within genes with compound symmetry structure for $m = 1000, q = 5, \alpha = 0.05$ and $\rho = 0, 0.1, 0.2$ and 0.3.

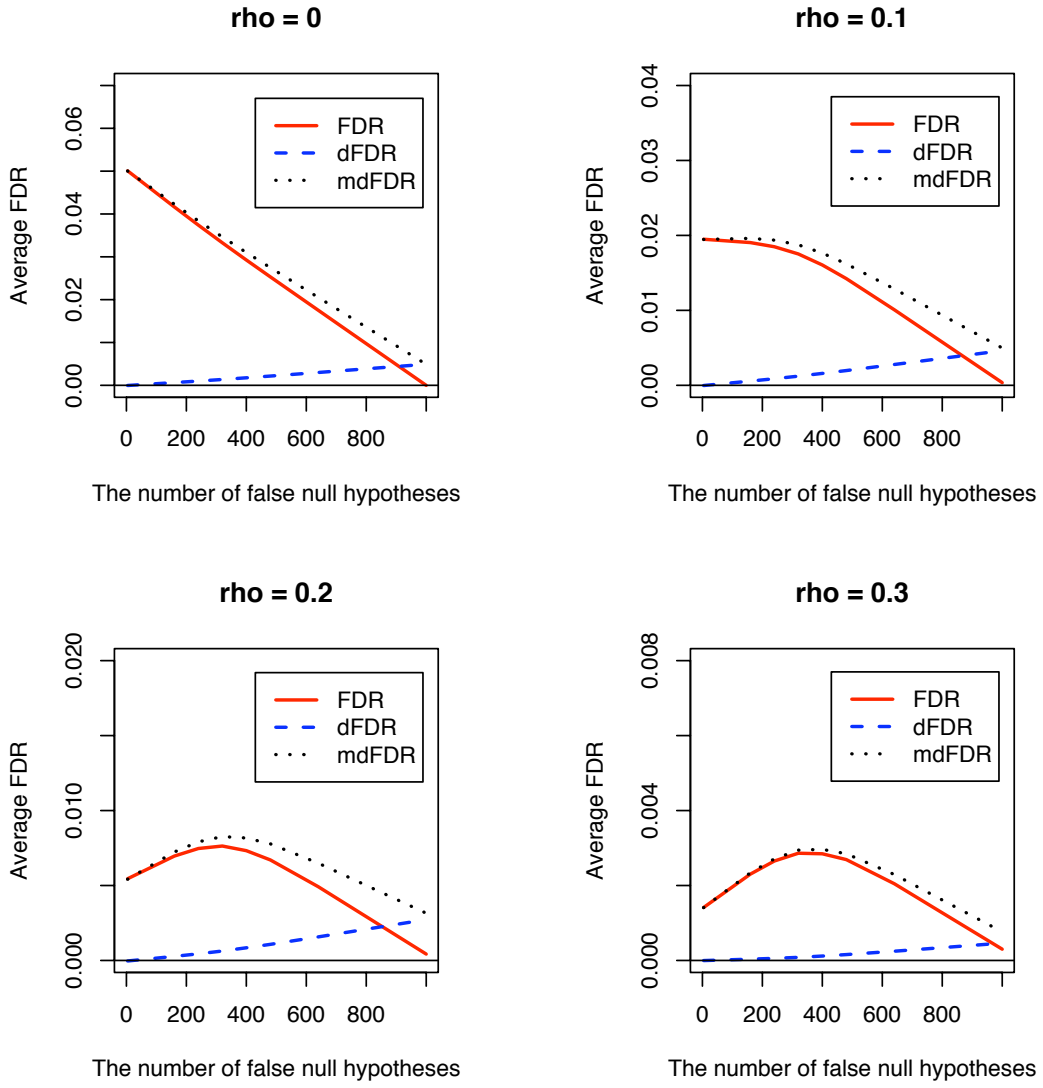

Figure 11. Performance of Procedure 2 under dependence within genes with AR(1) structure in terms of its control of the FDR, dFDR and mdFDR for $m = 1000, q = 5, \alpha = 0.05$, and $\rho = 0, 0.1, 0.2$ and 0.3.

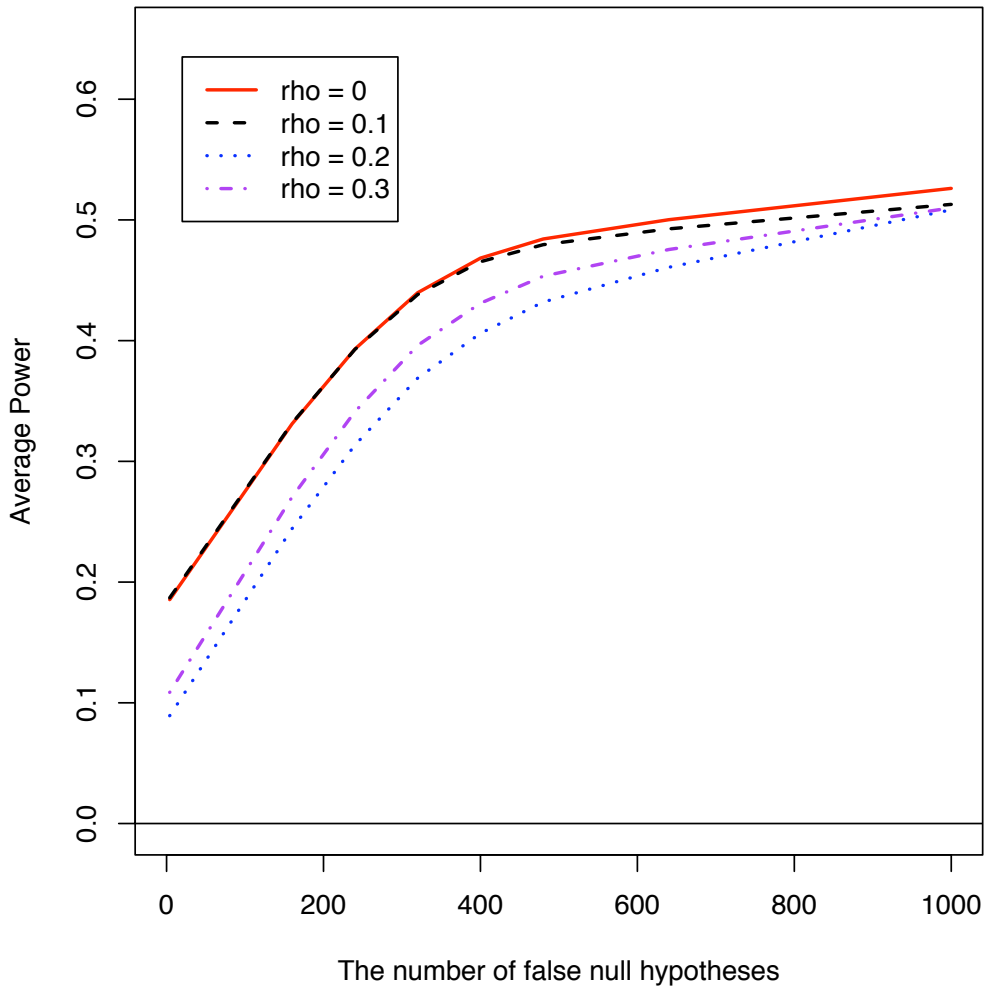

Figure 12. Power of Procedure 2 under dependence within genes with AR(1) structure for $m = 1000, q = 5, \alpha = 0.05$ and $\rho = 0, 0.1, 0.2$ and 0.3.