Web-based Supplementary Materials for Estimation of False Discovery Rate Using Sequential Permutation P-Values by Tim Bancroft, Chuanlong Du, and Dan Nettleton

Web Appendix A: A Proof a Theorem 1

This appendix contains the proof of the Theorem 1 stated in Section 2 of the main manuscript. By equation (1) in the main paper, the support of p is $S(h,n)$. Thus, it is sufficient to show that $\Pr(p \leqslant \alpha) = \alpha$ for all $\alpha \in S(h,n)$. First, consider the case where $\alpha = j/n$ for some $j \in \{1,\ldots,h\}$. Then

$$\begin{aligned}
\Pr(p \leqslant \alpha) &= \Pr(p \leqslant j/n) \\
&= \Pr(p = 1/n) + \cdots + \Pr(p = j/n) \\
&= \Pr(G = 0 \text{ in a sample of size } n-1 \text{ from } \psi) + \cdots + \\
&\quad \Pr(G = j-1 \text{ in a sample of size } n-1 \text{ from } \psi) \\
&= \Pr(Z \text{ is smallest in a sample of total size } n \text{ from } \psi) + \cdots + \\
&\quad \Pr\left(Z \text{ is } j^{\text{th}} \text{ smallest in a sample of total size } n \text{ from } \psi\right) \\
&\overset{H_0}{=} 1/n + \cdots + 1/n = j/n = \alpha.
\end{aligned}$$

If $\alpha = h/j$ for any $j \in \{n-1,\cdots,h\}$, then

$$\begin{aligned}
\Pr(p \leqslant \alpha) &= \Pr(p \leqslant h/j) \\
&= \Pr(Z \text{ is among the } h \text{ smallest values when included} \\
&\quad \text{with the first } j-1 \text{ sampled from } \psi) \\
&= \Pr(Z \text{ is smallest in a sample of total size } j \text{ from } \psi) + \cdots + \\
&\quad \Pr\left(Z \text{ is } h^{\text{th}} \text{ smallest in a sample of total size } j \text{ from } \psi\right) \\
&\overset{H_0}{=} 1/j + \cdots + 1/j = h/j = \alpha. \square
\end{aligned}$$

Web Appendix B: A Simple Example of the Histogram-Based Estimator

Suppose we have a collection of $m = 100$ SP p-values produced using $h = 4$ and $n = 10$. Web Table 1 displays the possible p-values, their null probabilities, their hypothetical observed frequencies, and expected null frequencies (rounded to the nearest tenth for improved readability) for iterations 0,1,2, 14, and 15 of the algorithm. Estimates of m_0 for each displayed iteration are presented near the top of the table.

[Table 1 about here.]

After initializing $\hat{m}_0(0)$ to 100, the algorithm begins by computing the iteration 0 expected null frequency $e_{0j} = \hat{m}_0(0)s_j$ for each $j = 1, \ldots, 10$. As an example, e_{05} is given by $\hat{m}_0(0)s_5 = 100(4/9 - 4/10) = 4.4$. To determine $\hat{m}_0(1)$, we begin searching from the smallest p-value to the largest for the first observed frequency that does not exceed its expected null frequency. This occurs at the p-value 3/10. Thus, we add the differences between the observed number of p-values and the expected number of null p-values for p-values 1/10 and 2/10 to obtain $(23 - 10) + (13 - 10) = 16$ as an estimate of $m - m_0$. Therefore, $\hat{m}_0(1)$ is set equal to $100 - 16 = 84$. In terms of the formal definition of the algorithm given by (5) and (6) in the main paper, we have $o_1 = 23 > 10 = 100(1/10) = \hat{m}_0(0)s_1 = e_{01}$, $o_2 = 13 > 10 = 100(1/10) = \hat{m}_0(0)s_2 = e_{02}$, and $o_3 = 9 \leqslant 10 = 100(1/10) = \hat{m}_0(0)s_3 = e_{03}$, which implies $j_0 = 3$ and

$$\hat{m}_0(1) \;=\; m - \sum_{j=1}^{3-1}(o_j - e_{0j}) = 100 - \{(23 - 10) + (13 - 10)\} = 84.$$

Given that the current estimate of m_0 has been updated from 100 to 84, we then recalculate the number of each of the possible p-values that would be expected to originate from tests with a true null hypothesis. These expectations are given in the column labeled e_{1j} by $84s_j$ $(j = 1, \ldots, 10)$. Next, we find the smallest p-value whose observed frequency is less than its updated expected null frequency. In this case, the relevant p-value is 4/10. The excess for p-values less than 4/10 is $23 - 8.4 = 14.6$, $13 - 8.4 = 4.6$, and $9 - 8.4 = 0.6$, respectively. The

iterative algorithm next updates m_0 to $\hat{m}_0(2) = 100 - (14.6 + 4.6 + 0.6) = 80.2$. Continuing to iterate, the estimated number of true nulls at iteration i decreases monotonically to approximately 78.6.

Theorem 2 stated in Subsection 4.2 of the main paper can be used to obtain this limiting value directly without iteration. Note that $o_1/(o_1 + \cdots + o_{10}) = 23/100 > 0.1 = s_1/(s_1 + \cdots + s_{10})$, $o_2/(o_2 + \cdots + o_{10}) = 13/77 > 0.1/0.9 = s_2/(s_2 + \cdots + s_{10})$, $o_3/(o_3 + \cdots + o_{10}) = 9/64 > 0.1/0.8 = s_3/(s_3 + \cdots + s_{10})$, and $o_4/(o_4 + \cdots + o_{10}) = 1/55 \leqslant 0.1/0.7 = s_4/(s_4 + \cdots + s_{10})$. Thus $J = 4$ and $\hat{m}_0 = \sum_{i=4}^{10} o_i / \sum_{i=4}^{10} s_i = (1 + 1 + 8 + 7 + 8 + 13 + 17)/(1/10 + 2/45 + 1/18 + 1/14 + 2/21 + 2/15 + 1/5) = 55/0.7 \approx 78.57143$.

Web Appendix C: A Proof of Theorem 2

This appendix contains the proof of the Theorem 2 stated in Subsection 4.2 of the main manuscript. First, four facts used to facilitate the proof are given. Then, the proof is shown followed by the proofs of the four facts. To simplify notation throughout the proofs, we will let $o_{j:n} = o_j + \cdots + o_n$ and $s_{j:n} = s_j + \cdots + s_n$. It follows that $o_{1:n} = m$ and $s_{1:n} = 1$.

Four Facts Used in the Proof of Theorem 2

1. $o_{1:n}/s_{1:n} > o_{2:n}/s_{2:n} > \ldots > o_{J:n}/s_{J:n}$.
2. If $j_i < J$, then (a) $j_i \leqslant j_{i+1}$, (b) $j_{i+1} \leqslant J$, and (c) $j_i < j_{i^*}$ for some $i^* > i$.
3. If there exists i^* such that $j_{i^*} = J$, then $j_i = J$ for all $i \geqslant i^*$.
4. Suppose $\{a_i\}_{i \geqslant 0}$ is an infinite sequence of real numbers. If there exists $\lambda \in (0, 1]$, an integer i^*, and a real number a such that $a_{i+1} = \lambda a + (1 - \lambda)a_i$ whenever $i \geqslant i^*$, then $\lim_{i \to \infty} a_i = a$.

Proof of Theorem 2

We first show that j_i converges to J in a finite number of iterations.

Case I ($J = 1$):

By the definition of J, $J = 1 \Rightarrow o_1/o_{1:n} \leqslant s_1/s_{1:n} \Rightarrow o_1/m \leqslant s_1 \Rightarrow o_1 \leqslant \hat{m}_0(0)s_1$. Because $j_i \equiv \min\{j = 1, \ldots, n : o_j \leqslant \hat{m}_0(i)s_j\}$, $j_0 = 1 = J$. Fact 3 then implies $j_i = J = 1$ for all $i \geqslant 0$, i.e. $\{j_i\}$ converges to 1 at iteration 0.

Case II ($J > 1$):

The definition of J and Fact 1 imply $\frac{o_J}{s_J} \leqslant o_{J:n}/s_{J:n} < o_{1:n}/s_{1:n} = \hat{m}_0(0)$. Thus, $o_J < \hat{m}_0(0)s_J$. By definition of j_0, it follows that $j_0 \leqslant J$. This leaves two subcases to consider.

Subcase I ($j_0 = J$):

By Fact 3, if $j_0 = J$, then $j_i = J$ for all $i \geqslant 0$, i.e. $\{j_i\}$ converges to J at iteration 0.

Subcase II: ($j_0 < J$):

Given that $j_0 < J$, Fact 2 implies that the sequence $\{j_i\}$ (a) is nondecreasing, (b) is bounded above by J, and (c) will eventually increase until $j_{i^*} = J$ for some i^*, at which point, convergence to J is achieved by Fact 3.

Whatever the case, we have shown there exists a i^* such that $j_i = J$ for all $i \geqslant i^*$. Therefore, for all $i \geqslant i^*$,

$$\begin{aligned}\hat{m}_0(i+1) &= m - \sum_{j=1}^{J-1}(o_j - e_{ij}) = m - \sum_{j=1}^{J-1}(o_j - \hat{m}_0(i)s_j) \\ &= o_{J:n} + \hat{m}_0(i)(1 - s_{J:n}) = s_{J:n}o_{J:n}/s_{J:n} + (1 - s_{J:n})\hat{m}_0(i).\end{aligned}$$

By Fact 4, $\lim_{i\to\infty} \hat{m}_0(i) = o_{J:n}/s_{J:n} = \sum_{j=J}^{n} o_j / \sum_{j=J}^{n} s_j$. □

Proof of Fact 1

By definition of J, $j < J \Rightarrow o_j/o_{j:n} > s_j/s_{j:n}$. Therefore, we have

$$o_j > s_j o_{j:n}/s_{j:n} \text{ for all } j < J. \tag{1}$$

Now note that

$$\begin{aligned}o_{j:n}/s_{j:n} &= (1/s_{j:n})o_j + (1/s_{j:n})o_{j+1:n} \\ &= (1/s_{j:n})o_j + (1/s_{j:n})s_{j+1:n}o_{j+1:n}/s_{j+1:n}\end{aligned}$$

$$= (1/s_{j:n})o_j + (s_{j+1:n}/s_{j:n})o_{j+1:n}/s_{j+1:n}$$

$$= (1/s_{j:n})o_j + (1 - s_j/s_{j:n})o_{j+1:n}/s_{j+1:n}. \quad (2)$$

By combining (1) and (2), we have

$$o_{j:n}/s_{j:n} > (s_j/s_{j:n})o_{j:n}/s_{j:n} + (1 - s_j/s_{j:n})o_{j+1:n}/s_{j+1:n} \text{ for all } j < J. \quad (3)$$

Subtracting $(s_j/s_{j:n})o_{j:n}/s_{j:n}$ from both sides of (3) and dividing by $1-s_j/s_{j:n}$ yields $o_{j:n}/s_{j:n} > o_{j+1:n}/s_{j+1:n}$ for all $j < J$. □

Proof of Fact 2 (a)

For all $i \geqslant 0$, the definition of $\hat{m}_0(i+1)$ implies

$$\hat{m}_0(i+1) = o_{j_i:n} + \hat{m}_0(i)(1 - s_{j_i:n}). \quad (4)$$

Next, note that (1) and $j_i < J$ imply

$$o_{j_i:n} < o_{j_i} s_{j_i:n}/s_{j_i}. \quad (5)$$

By definition of j_i,

$$o_{j_i} \leqslant \hat{m}_0(i) s_{j_i}. \quad (6)$$

Taken together (5) and (6) imply

$$o_{j_i:n} < \hat{m}_0(i) s_{j_i:n}. \quad (7)$$

Combining (4) and (7) yields $\hat{m}_0(i+1) < \hat{m}_0(i)$. Therefore,

$$\{j = 1, \ldots, n : o_j \leqslant \hat{m}_0(i+1)s_j\} \subseteq \{j = 1, \ldots, n : o_j \leqslant \hat{m}_0(i)s_j\},$$

which implies $j_i \leqslant j_{i+1}$. □

Proof of Fact 2 (b)

By the definition of J,

$$o_J \leqslant s_J o_{J:n}/s_{J:n}. \quad (8)$$

Combining (8) and Fact 1 yields

$$o_J < s_J o_{j_i:n}/s_{j_i:n}. \tag{9}$$

By (9) and (1), we have $o_J < s_J o_{j_i}/s_{j_i}$, which implies

$$o_J < s_J \hat{m}_0(i) \tag{10}$$

by (6). By (4),

$$\hat{m}_0(i+1)s_J = o_{j_i:n}s_J + \hat{m}_0(i)(1 - s_{j_i:n})s_J. \tag{11}$$

Applying (9) and (10) to (11) yields

$$\hat{m}_0(i+1)s_J > o_J s_{j_i:n} + o_J(1 - s_{j_i:n}) = o_J.$$

Thus, we have $o_J < \hat{m}_0(i+1)s_J$, which implies

$$J \in \{j = 1, \ldots, n : o_j \leqslant \hat{m}_0(i+1)s_j\}.$$

Thus, j_{i+1}, the minimum of $\{j = 1, \ldots, n : o_j \leqslant \hat{m}_0(i+1)s_j\}$, can be no larger than J. □

Proof of Fact 2 (c)

If Fact 2(c) were false, then Fact 2(a) would imply $j_\ell = j_i$ for all $\ell \geqslant i$. By the same argument used to conclude the proof of Theorem 2, $j_\ell = j_i$ for all $\ell \geqslant i$ would imply

$$\lim_{i\to\infty} \hat{m}_0(i) = o_{j_i:n}/s_{j_i:n}. \tag{12}$$

However, by (5), $o_{j_i:n}/s_{j_i:n} < o_{j_i}/s_{j_i}$. Thus, (12) would imply $\hat{m}_0(\ell)s_{j_i} < o_{j_i}$ for ℓ sufficiently large, which implies

$$j_i \notin \{j = 1, \ldots, n : o_j \leqslant \hat{m}_0(\ell)s_j\}.$$

It follows that j_ℓ cannot equal j_i for sufficiently large ℓ. We have reached a contradiction, and the proof of Fact 2(c) follows.

Proof of Fact 3

Suppose $j_i = J$. Then (4) implies

$$\hat{m}_0(i+1) = o_{J:n} + \hat{m}_0(i)(1 - s_{J:n}). \tag{13}$$

By (8), definition of j_i, and $j_i = J$, we have $o_{J:n} \geqslant o_J s_{J:n}/s_J$ and $\hat{m}_0(i) \geqslant o_J/s_J$. Applying these inequalities to (13) yields $\hat{m}_0(i+1) \geqslant o_J/s_J$. Thus,

$$J \in \{j = 1, \ldots, n : o_j \leqslant \hat{m}_0(i+1)s_j\}, \text{ which implies } j_{i+1} \leqslant J. \tag{14}$$

Next, using the definitions of j_i and J, Fact 1, and $J = j_i$, we have for all $j < J$,

$$o_j > \hat{m}_0(i)s_j \text{ and } o_j/s_j > o_{j:n}/s_{j:n} > o_{J:n}/s_{J:n} = o_{j_i:n}/s_{j_i:n}.$$

Then, for all $j < J$,

$$\begin{aligned} \hat{m}_0(i+1) &= o_{j_i:n} + \hat{m}_0(i)(1 - s_{j_i:n}) \\ &< s_{j_i:n}o_j/s_j + (1 - s_{j_i:n})o_j/s_j \\ &= o_j/s_j. \end{aligned}$$

This implies $o_j > \hat{m}_0(i+1)s_j$ for all $j < J$, which by the definition of j_{i+1}, implies $j_{i+1} \geqslant J$. Together with (14), this implies $j_{i+1} = J$. □

Proof of Fact 4

Let $b_n = a_{i^*+n}$ for all $n \geqslant 0$. Then $b_1 = \lambda a + (1-\lambda)b_0$, and an induction argument shows that $b_n = \lambda a \sum_{\ell=0}^{n-1}(1-\lambda)^\ell + (1-\lambda)^n b_0$ for all $n \geqslant 1$. Now $\lambda \in (0, 1]$ implies that

$$\lim_{n\to\infty}(1-\lambda)^n = 0 \text{ and } \sum_{\ell=0}^{\infty}(1-\lambda)^\ell = \tfrac{1}{\lambda}.$$

Hence, $\lim_{n\to\infty} b_n = a$, and $\lim_{i\to\infty} a_i = a$ because $\{a_i\}_{i\geqslant i^*} = \{b_n\}_{n\geqslant 0}$. □

Web Appendix D: An Algorithm for Binning Sequential Permutation P-values

For practical choices of h and n, some sequential permutation p-values are unlikely to occur. For example, the sequential permutation p-value $h/(n-1)$ has probability

$$\frac{h}{n-1} - \frac{h}{n} = \frac{h}{n(n-1)}$$

under the null. More generally, the sequential permutation p-value $h/(n-k)$ has null probability

$$\frac{h}{(n-k+1)(n-k)} \text{ for } k = 1, \ldots, n-h.$$

For k small, these null probabilities are quite small. Furthermore, the probabilities of such p-values are likely to be small under most alternatives. Thus, even when thousands of tests are conducted, some small sequential permutation p-values in the support $S(h, n)$ are likely to be unobserved.

This phenomenon can be seen in the ALL data analysis discussed in Subsection 5.1 of the main paper. Web Figure 1 shows null probabilities (solid vertical lines) and observed distribution (dashed line) of the sequential permutation p-values for p-values $\{h/(n-k) : k = 1, \ldots, 20\}$. Several of the p-values are unobserved, including $h/(n-1) = 10/999$. With no binning of the p-values, o_J in Theorem 2 of the main paper would be small (10/999 in this case), which would produce a relatively large estimate of m_0, namely $11,992$. With binning of the p-values, the estimate computed in Subsection 5.1 of the main paper $(9,548)$ is considerably lower.

[Figure 1 about here.]

In general, failing to bin the p-values or using bins with low probabilities produces an estimator of m_0 that is too conservative. Thus, we use the following algorithm to obtain as many bins as possible that each have null probability no smaller than 0.05 (the bin probability recommended by Nettleton et al., 2006).

(1) Sum the null probabilities for the smallest p-values until the sum first reaches or exceeds 0.05. The p-values whose null probabilities contributed to this sum form the first bin.
(2) Starting with the smallest p-value not already in a bin, sum the null probabilities of the smallest p-values until the sum first reaches or exceeds 0.05. The new bin is defined by the set of all p-values whose null probabilities contributed to this sum. If the sum of the

null probabilities of all p-values not already in a bin is less than 0.05, these remaining p-values form the last bin.

(3) Repeat step 2 until each p-value has been assigned to exactly one bin.

Web Appendix E: The Choice of h and n

By Theorem 1 of Section 2 in the main paper, a sequential permutation (SP) p-value can be used to conduct a valid test at any desired significance level in $(0, 1]$ regardless of the choice of h and n. Small values of h and n obviously reduce the expense of computing a permutation p-value, but values that are too small can lead to an important loss of power for an SP p-value relative to an exact permutation (EP) p-value. For example, the smallest possible SP p-value is $1/n$. Thus, a test based on an SP p-value will have power 0 for any significance level below $1/n$. Of course, the same can be said for a traditional Monte Carlo permutation (MP) p-value based on a simple random sample of $n-1$ draws from the permutation distribution.

Given that the EP p-value based on an examination of the entire permutation distribution is p, it is straightforward to show that the conditional mean and conditional variance of the MP p-value are approximately $(n-1)p/n+1/n$ and $(n-1)p(1-p)/n^2$, respectively, provided that the number of distinct values in the test statistic's permutation distribution is large. Thus, for large n and small p, the conditional bias of the MP p-value as an estimator of p is approximately $1/n$, and the approximate standard error is $\sqrt{p/n}$. The positive bias $1/n$ clearly indicates the potential for power loss when using an MP p-value rather than an EP p-value. Thus, n is typically taken to be as large as is computationally affordable.

What about power loss of the SP test relative to the MP test? Fortunately, there is no power loss at all for a significance level α test if h is chosen to be $n\alpha$. To see this, note that rejection of the null will occur for the SP test if and only if the SP p-value is less than or equal to h/n, i.e., if and only if G in (1) is less than h. Now if $G < h$, then the SP p-value takes the same value as the traditional MP p-value, namely $(G+1)/n$. On the other hand, if

$G = h$, the SP p-value will be at least $h/(n-1) > \alpha$ while the MP p-value would be at least $(h+1)/n > \alpha$ if sampling were to continue to include a total of $n-1$ random draws from the permutation distribution. Thus, for the case of $h = n\alpha$, the SP tests stops sampling from the permutation distribution as soon as the decision of the traditional Monte Carlo permutation test at significance level α has been determined. Thus, it makes sense to choose $h = n\alpha$. (For a different perspective on power of the SP test, see Silva et al. (2009) who examined the power of the sequential test as a function of n for fixed h and α.)

Unfortunately, the prescription to choose n large and $h = n\alpha$ is not directly helpful when FDR control at level γ is desired in a multiple testing situation. The problem is that the p-value threshold for significance α that provides an estimated FDR at or below γ is unknown and depends on the distribution of observed p-values as defined in equation (4) of the main paper. This means that we do not know the value of α that will be used to conduct each individual hypothesis test until after we have obtained all the p-values.

In many cases, it may be reasonable to specify an upper bound for α and to choose h/n equal to that upper bound. As discussed in Subsection 5.1 of the main paper, the MP and SP tests will give the same rejection set for any p-value significance threshold less than or equal to h/n. We have considered $h = 10$, $n = 1,000$ and $h = 100$, $n = 10,000$. These choices each guarantee the same rejection set for the MP and SP tests for any p-value significance threshold less than 0.01, which is often a reasonable upper bound for α when FDR control around 5% is desired.

Returning again to the choice of n, our advice has been to choose n as large as is computationally affordable to increase power towards the power of the exact permutation test. An advantage of the sequential approach over a traditional Monte Carlo approach is that the sequential approach makes larger choices of n feasible. For example, suppose the total number of test statistics that we are willing to compute based on time and available

resources is $N_T = 10^7$. Furthermore, suppose we must conduct $m = 10,000$ tests. Then the MP approach would use $n = N_T/m = 1,000$. Because the SP approach does not compute n statistics for each of the m tests, we can take n to be larger than $N_T/m = 1,000$ without exceeding our time and resource limitations.

What value of n we can afford to select for the SP analysis given a value for N_T? An answer depends on the unknown number of true null hypotheses m_0. As discussed in Subsection 5.2 of the main paper, Besag and Clifford (1991) provide an approximation ($\hat{L}$) for the expected number of permutations sampled at termination of the sequential procedure ($E(L)$) for tests with a true null hypothesis. If we take the conservative assumption that all tests with a false null hypothesis will require the computation of n statistics (like each MP test), an upper bound on the expected number of statistics computed for the SP approach is

$$m_0(\hat{L} + 1) + (m - m_0)n.$$

If we select $h = n\alpha$, as suggested above, then

$$\hat{L} = h + h\log\{(n - 1/2)/(h + 1/2)\} \approx n\{\alpha - \alpha\log(\alpha)\}$$

for n reasonably large (e.g., 1,000 or more). Putting this altogether, we can choose

$$n = N_T[m_0\{\alpha - \alpha\log(\alpha) - 1\} + m]^{-1}$$

as a computationally affordable value of n for the SP procedure. For our example, with $N_T = 10^7$, $m = 10,000$, $m_0 = 7,500$, and $\alpha = 0.01$, we could easily afford to choose $h = 34$ and $n = 3,400$ and expect to have more power and less computational expense than the traditional Monte Carlo approach with $n = 1,000$.

Web Appendix F: Tables Summarizing Additional Results Obtained by Repeated Analysis of the ALL Data

[Table 2 about here.]

[Table 3 about here.]

Web Appendix G: Tables Summarizing Additional Results for the Expression Quantitative Trait Locus Analysis of the Barley Data

[Table 4 about here.]

[Table 5 about here.]

[Table 6 about here.]

Web Appendix H: Tables Summarizing Additional Results from the Simulations Described in Section 6 of the Main Paper

[Table 7 about here.]

[Table 8 about here.]

[Table 9 about here.]

[Table 10 about here.]

[Table 11 about here.]

[Table 12 about here.]

[Table 13 about here.]

[Table 14 about here.]

[Table 15 about here.]

[Table 16 about here.]

[Table 17 about here.]

[Table 18 about here.]

Web Appendix I: A Simulation Based on Microarray-Derived Dependent Data

In many modern multiple testing problems, tests have an unknown and presumably complex dependence structure. Testing for differential gene expression is a common example where dependence across genes within experimental units leads to dependence across tests. To evaluate the performance of the SP p-values relative to competing methods when tests are dependent, we revisit the ALL dataset discussed in Section 5 of the main paper. Recall that expression values were recorded for 12,625 genes on two groups of 21 males and 5 males. In this section, we focus on the data from the 21 males with a translocation between chromosomes 9 and 22 to obtain a gene expression matrix A with 12,625 rows (one for each gene) and 21 columns (one for each subject). This data is inherently dependent across genes as the 12,625 gene expression values for each subject are correlated.

To be consistent with the simulation scheme in Section 6 of the main paper, $m = 10,000$ genes were randomly selected from all 12,625 genes and used as the entire dataset for every run of this simulation. For each run of the simulation, 16 of the 21 males were randomly selected yielding a gene expression sub-matrix B of 10,000 rows and 16 columns. At the gene specific level, this subset can be thought of as a simple random sample from a finite homogeneous population. Next, if we further divide these 16 randomly selected males into two sub-groups of the first eight males and last eight males, respectively, we create two sub-groups for which the null hypothesis for each gene is true.

To generate data where some null hypotheses are true and some null hypotheses are false, we alter the gene expression sub-matrix B as follows. First, we randomly select $m - m_0$ genes from the subset of the $m = 10,000$ genes. Next, for each of those $m - m_0$ genes, we add to each of the eight males in the second sub-group a random effect generated from $\Gamma(\lambda, 1)$ with mean λ multiplied by the gene-specific standard deviation of all 21 males. This multiplier will yield a noncentrality parameter similar to the simulation study in Section 6 of the main

paper. In this new 10,000 by 16 simulated dataset, $m - m_0$ genes are differentially expressed across sub-groups (first eight columns vs. the last eight columns) and dependencies among genes within experimental units mimic those in the original data.

For the $m_0 = 7,500$ and $\lambda = 2$ scenario, $N = 1,000$ datasets were simulated using the strategy described above. EP, MP, and SP p-values were computed based on a two-sample t-statistic for each gene in each of the $N = 1,000$ simulated datasets. The SP p-values using $h = 10$ and $n = 1,000$ were computed from a subset of the permutations used by the MP approach with $n = 1,000$. As in Sections 5 and 6 of the main paper, the methods of Storey and Tibshirani (2003) (ST) and Nettleton et al. (2006) (NHCW) were used to estimate m_0 from EP and MP p-values, while the method proposed in Section 4 of the main paper was used to estimate m_0 from SP p-values. For each simulated dataset, the estimate of m_0 and the number of rejected null hypotheses (R), the estimated FDR, and $V/\max\{1, R\}$ were recorded for the p-value significance threshold 0.01.

[Table 19 about here.]

[Table 20 about here.]

A comparison of Web Tables 8 and 19 shows that the independent and dependent data simulations produced similar results. The main difference is an sharp increase in variability across simulation replications as indicated by the much larger standard error of the mean figures in Web Table 19 compared to those in Web Table 8 and the wider gap between quantiles in Web Table 20 compared to Web Table 14. This phenomenon of increasing variability under dependence has been seen in other dependent-data simulation studies (for example, Langaas et al., 2005; Nettleton et al., 2006).

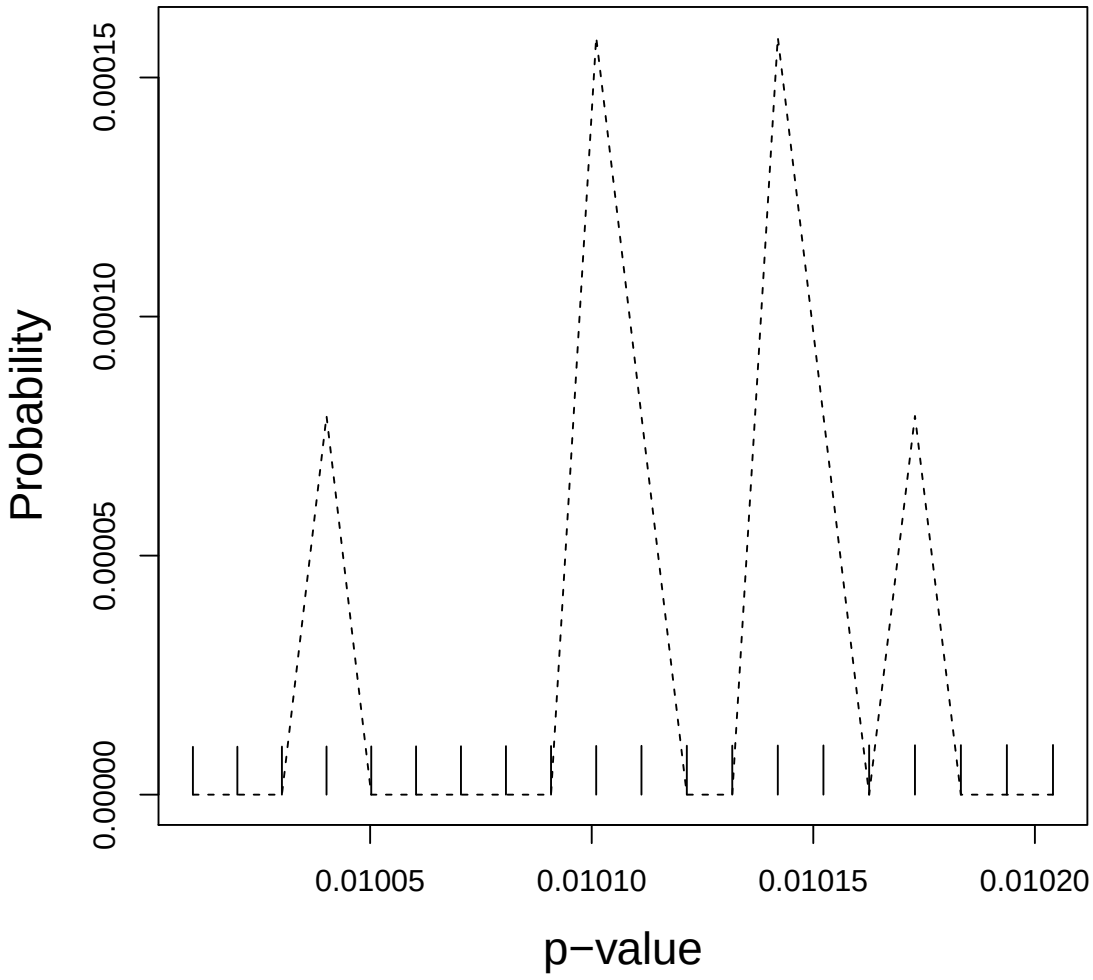

Figure 1. The null (solid) and observed (dashed) SP p-value distributions for the ALL analysis with $h = 10$ and $n = 1,000$ for p-values $10/999, \ldots, 10/980$.

Table 1

Possible p-values, the probability of each p-value under the null hypothesis, hypothetical observed frequencies, rounded expected frequencies under the null hypothesis, and estimates of m_0 for iterations 0, 1, 2, 14, and 15 of an extension of the algorithm of Mosig et al. (2001) for the special case of $m = 100$ SP p-values with $h = 4$ and $n = 10$.

		Iteration i	0	1	2	$\cdots$	14	15
		$\hat{m}_0(i)$	100	84.0	80.2	$\cdots$	78.6	78.6
j	$S_j(h,n)$	s_j	o_j	e_{0j}	e_{1j}	e_{2j}	$\cdots$	e_{14j}	e_{15j}
1	1/10	1/10	23	10.0	8.4	8.0	$\cdots$	7.9	7.9
2	2/10	1/10	13	10.0	8.4	8.0	$\cdots$	7.9	7.9
3	3/10	1/10	9	10.0	8.4	8.0	$\cdots$	7.9	7.9
4	4/10	1/10	1	10.0	8.4	8.0	$\cdots$	7.9	7.9
5	4/9	2/45	1	4.4	3.7	3.6	$\cdots$	3.5	3.5
6	4/8	1/18	8	5.6	4.7	4.5	$\cdots$	4.4	4.4
7	4/7	1/14	7	7.1	6.0	5.7	$\cdots$	5.6	5.6
8	4/6	2/21	8	9.5	8.0	7.6	$\cdots$	7.5	7.5
9	4/5	2/15	13	13.3	11.2	10.7	$\cdots$	10.5	10.5
10	1	1/5	17	20.0	16.8	16.0	$\cdots$	15.7	15.7

Table 2

Means, standard deviations, and quantiles of the 1,000 estimates of m_0 obtained by repeated analysis of the ALL data discussed in Section 5.1 of the main paper. The single results for the EP methods are included under the mean column for comparison with the Monte Carlo approaches.

	mean	std	$q_{0.05}$	$q_{0.95}$
EP/ST	9031			
EP/NHCW	9060			
MP/ST	9184	54.26	9093	9274
MP/NHCW	9600	421.31	9020	10189
SP	9663	332.91	9152	10171

Table 3

Means, standard deviations, and quantiles of the 1,000 estimates of FDR obtained by repeated analysis of the ALL data using p-value significance threshold 0.001 as discussed in Section 5.1 of the main paper. The single results for the EP methods are included under the mean column for comparison with the Monte Carlo approaches.

	mean	std	$q_{0.05}$	$q_{0.95}$
EP/ST	0.0393			
EP/NHCW	0.0394			
MP/ST	0.0435	0.0015	0.0410	0.0458
MP/NHCW	0.0454	0.0025	0.0415	0.0497
SP	0.0457	0.0022	0.0423	0.0493

Table 4

The estimated number of true null hypotheses ($\hat{m}_0$), the number of rejected hypotheses at p-value threshold 0.005 ($R(0.005)$), and the estimated FDR at p-value threshold 0.005 ($\widehat{FDR}(0.005)$) for MP and SP analyses of the barley eQTL dataset.

Method	n	h	$\hat{m}_0$	R	$\widehat{\text{FDR}}$
MP/ST	1000	1000	19783	3430	0.0288
MP/NHCW	1000	1000	16240	3430	0.0237
MP/ST	10000	10000	19584	3475	0.0282
MP/NHCW	10000	10000	16285	3475	0.0234
SP	1000	10	16340	3449	0.0237
SP	10000	100	16304	3470	0.0235

Table 5
The estimated number of true null hypotheses ($\hat{m}_0$), the number of rejected hypotheses at p-value threshold 0.01 ($R(0.01)$), and the estimated FDR at p-value threshold 0.01 ($\widehat{FDR}(0.01)$) for MP and SP analyses of the barley eQTL dataset.

Method	n	h	$\hat{m}_0$	R	$\widehat{\text{FDR}}$
MP/ST	1000	1000	19783	3838	0.0515
MP/NHCW	1000	1000	16240	3838	0.0423
MP/ST	10000	10000	19584	3865	0.0507
MP/NHCW	10000	10000	16285	3865	0.0421
SP	1000	10	16340	3857	0.0424
SP	10000	100	16304	3851	0.0423

Table 6

The estimated number of true null hypotheses ($\hat{m}_0$), the number of rejected hypotheses at p-value threshold 0.02 ($R(0.02)$), and the estimated FDR at p-value threshold 0.02 ($\widehat{FDR}(0.02)$) for MP and SP analyses of the barley eQTL dataset.

Method	n	h	$\hat{m}_0$	R	$\widehat{\text{FDR}}$
MP/ST	1000	1000	19783	4450	0.0889
MP/NHCW	1000	1000	16240	4450	0.0730
MP/ST	10000	10000	19584	4465	0.0877
MP/NHCW	10000	10000	16285	4465	0.0729
SP	1000	10	16340	4445	0.0735
SP	10000	100	16304	4462	0.0731

Table 7

Means and standard errors of the means (sem) computed from $N = 1,000$ simulation runs for the estimated number of true null hypotheses ($\hat{m}_0$), the number of rejected null hypotheses (R), estimated FDR ($\widehat{FDR}$), and false positive fraction ($V/\max\{1, R\}$) for p-value significance threshold 0.01. The number of tests and the number of true null hypotheses were $m = 10,000$ and $m_0 = 7,500$, respectively. Data were generated as described in Section 6 of the main paper with $\lambda = 1$.

	$\hat{m}_0$		R		$\widehat{\text{FDR}}$		$V/\max\{1, R\}$	
Method	mean	sem	mean	sem	mean	sem	mean	sem
EP/ST	8576	10	750	0.7627	0.1150	0.0002	0.1000	0.0003
EP/NHCW	8674	4	750	0.7627	0.1163	0.0001	0.1000	0.0003
MP/ST	8653	10	743	0.7554	0.1166	0.0002	0.1005	0.0003
MP/NHCW	8673	4	743	0.7554	0.1168	0.0001	0.1005	0.0003
SP	8667	5	743	0.7554	0.1167	0.0001	0.1005	0.0003

Table 8

Means and standard errors of the means (sem) computed from $N = 1,000$ simulation runs for the estimated number of true null hypotheses ($\hat{m}_0$), the number of rejected null hypotheses (R), estimated FDR ($\widehat{FDR}$), and false positive fraction ($V/\max\{1, R\}$) for p-value significance threshold 0.01. The number of tests and the number of true null hypotheses were $m = 10,000$ and $m_0 = 7,500$, respectively. Data were generated as described in Section 6 of the main paper with $\lambda = 2$.

	$\hat{m}_0$		R		$\widehat{\text{FDR}}$		$V/\max\{1, R\}$	
Method	mean	sem	mean	sem	mean	sem	mean	sem
EP/ST	7851	9	1526	0.8153	0.0518	0.0001	0.0490	0.0002
EP/NHCW	7908	3	1526	0.8153	0.0522	0.0000	0.0490	0.0002
MP/ST	7920	9	1516	0.8221	0.0523	0.0001	0.0492	0.0002
MP/NHCW	7907	3	1516	0.8221	0.0522	0.0000	0.0492	0.0002
SP	7906	4	1516	0.8221	0.0522	0.0000	0.0492	0.0002

Table 9

Means and standard errors of the means (sem) computed from $N = 1,000$ simulation runs for the estimated number of true null hypotheses ($\hat{m}_0$), the number of rejected null hypotheses (R), estimated FDR ($\widehat{FDR}$), and false positive fraction ($V/\max\{1, R\}$) for p-value significance threshold 0.01. The number of tests and the number of true null hypotheses were $m = 10,000$ and $m_0 = 7,500$, respectively. Data were generated as described in Section 6 of the main paper with $\lambda = 3$.

	$\hat{m}_0$		R		$\widehat{\text{FDR}}$		$V/\max\{1, R\}$	
Method	mean	sem	mean	sem	mean	sem	mean	sem
EP/ST	7580	9	2077	0.6641	0.0368	0.0000	0.0360	0.0001
EP/NHCW	7622	3	2077	0.6641	0.0370	0.0000	0.0360	0.0001
MP/ST	7650	9	2069	0.6695	0.0370	0.0000	0.0362	0.0001
MP/NHCW	7623	3	2069	0.6695	0.0369	0.0000	0.0362	0.0001
SP	7620	3	2069	0.6695	0.0368	0.0000	0.0362	0.0001

Table 10

Means and standard errors of the means (sem) computed from $N = 1,000$ simulation runs for the estimated number of true null hypotheses ($\hat{m}_0$), the number of rejected null hypotheses (R), estimated FDR ($\widehat{FDR}$), and false positive fraction ($V/\max\{1, R\}$) for p-value significance threshold 0.01. The number of tests and the number of true null hypotheses were $m = 10,000$ and $m_0 = 9,000$, respectively. Data were generated as described in Section 6 of the main paper with $\lambda = 1$.

	$\hat{m}_0$		R		$\widehat{\text{FDR}}$		$V/\max\{1, R\}$	
Method	mean	sem	mean	sem	mean	sem	mean	sem
EP/ST	9430	10	360	0.5365	0.2633	0.0005	0.2481	0.0007
EP/NHCW	9483	3	360	0.5365	0.2648	0.0004	0.2481	0.0007
MP/ST	9509	9	357	0.5366	0.2669	0.0005	0.2493	0.0007
MP/NHCW	9484	3	357	0.5366	0.2662	0.0004	0.2493	0.0007
SP	9482	3	357	0.5366	0.2659	0.0004	0.2493	0.0007

Table 11
Means and standard errors of the means (sem) computed from $N = 1,000$ simulation runs for the estimated number of true null hypotheses ($\hat{m}_0$), the number of rejected null hypotheses (R), estimated FDR ($\widehat{FDR}$), and false positive fraction ($V/\max\{1, R\}$) for p-value significance threshold 0.01. The number of tests and the number of true null hypotheses were $m = 10,000$ and $m_0 = 9,000$, respectively. Data were generated as described in Section 6 of the main paper with $\lambda = 2$.

	$\hat{m}_0$		R		$\widehat{\text{FDR}}$		$V/\max\{1, R\}$	
Method	mean	sem	mean	sem	mean	sem	mean	sem
EP/ST	9136	10	671	0.5785	0.1374	0.0002	0.1342	0.0004
EP/NHCW	9171	3	671	0.5785	0.1379	0.0001	0.1342	0.0004
MP/ST	9218	10	666	0.5824	0.1386	0.0002	0.1343	0.0004
MP/NHCW	9171	3	666	0.5824	0.1379	0.0001	0.1343	0.0004
SP	9169	3	666	0.5824	0.1378	0.0001	0.1343	0.0004

Table 12

Means and standard errors of the means (sem) computed from $N = 1,000$ simulation runs for the estimated number of true null hypotheses ($\hat{m}_0$), the number of rejected null hypotheses (R), estimated FDR ($\widehat{FDR}$), and false positive fraction ($V/\max\{1, R\}$) for p-value significance threshold 0.01. The number of tests and the number of true null hypotheses were $m = 10,000$ and $m_0 = 9,000$, respectively. Data were generated as described in Section 6 of the main paper with $\lambda = 3$.

	$\hat{m}_0$		R		$\widehat{\text{FDR}}$		$V/\max\{1, R\}$	
Method	mean	sem	mean	sem	mean	sem	mean	sem
EP/ST	9044	10	891	0.4872	0.1025	0.0001	0.1005	0.0003
EP/NHCW	9051	2	891	0.4872	0.1026	0.0001	0.1005	0.0003
MP/ST	9129	10	887	0.4823	0.1030	0.0001	0.1004	0.0003
MP/NHCW	9050	2	887	0.4823	0.1021	0.0001	0.1004	0.0003
SP	9051	2	887	0.4823	0.1021	0.0001	0.1004	0.0003

Table 13

Quantiles computed from $N = 1,000$ simulation runs for the estimated number of true null hypotheses ($\hat{m}_0$), the number of rejected null hypotheses (R), estimated FDR ($\widehat{FDR}$), and false positive fraction ($V/\max\{1,R\}$) for p-value significance threshold 0.01. The number of tests and the number of true null hypotheses were $m = 10,000$ and $m_0 = 7,500$, respectively. Data were generated as described in Section 6 of the main paper with $\lambda = 1$.

	$\hat{m}_0$		R		$\widehat{\text{FDR}}$		$V/\max\{1,R\}$	
	$q_{0.05}$	$q_{0.95}$	$q_{0.05}$	$q_{0.95}$	$q_{0.05}$	$q_{0.95}$	$q_{0.05}$	$q_{0.95}$
EP/ST	8063	9054	710	792	0.1062	0.1244	0.0837	0.1179
EP/NHCW	8473	8842	710	792	0.1095	0.1236	0.0837	0.1179
MP/ST	8126	9143	705	782	0.1078	0.1259	0.0829	0.1193
MP/NHCW	8462	8844	705	782	0.1103	0.1240	0.0829	0.1193
SP	8419	8853	705	782	0.1098	0.1241	0.0829	0.1193

Table 14

Quantiles computed from $N = 1,000$ simulation runs for the estimated number of true null hypotheses ($\hat{m}_0$), the number of rejected null hypotheses (R), estimated FDR ($\widehat{FDR}$), and false positive fraction ($V/\max\{1, R\}$) for p-value significance threshold 0.01. The number of tests and the number of true null hypotheses were $m = 10,000$ and $m_0 = 7,500$, respectively. Data were generated as described in Section 6 of the main paper with $\lambda = 2$.

	$\hat{m}_0$		R		$\widehat{\text{FDR}}$		$V/\max\{1, R\}$	
	$q_{0.05}$	$q_{0.95}$	$q_{0.05}$	$q_{0.95}$	$q_{0.05}$	$q_{0.95}$	$q_{0.05}$	$q_{0.95}$
EP/ST	7397	8318	1484	1571	0.0486	0.0553	0.0403	0.0580
EP/NHCW	7722	8060	1484	1571	0.0503	0.0541	0.0403	0.0580
MP/ST	7461	8396	1474	1559	0.0489	0.0557	0.0402	0.0584
MP/NHCW	7711	8065	1474	1559	0.0503	0.0541	0.0402	0.0584
SP	7693	8072	1474	1559	0.0502	0.0541	0.0402	0.0584

Table 15

Quantiles computed from $N = 1,000$ simulation runs for the estimated number of true null hypotheses ($\hat{m}_0$), the number of rejected null hypotheses (R), estimated FDR ($\widehat{FDR}$), and false positive fraction ($V/\max\{1,R\}$) for p-value significance threshold 0.01. The number of tests and the number of true null hypotheses were $m = 10,000$ and $m_0 = 7,500$, respectively. Data were generated as described in Section 6 of the main paper with $\lambda = 3$.

	$\hat{m}_0$		R		$\widehat{\text{FDR}}$		$V/\max\{1,R\}$	
	$q_{0.05}$	$q_{0.95}$	$q_{0.05}$	$q_{0.95}$	$q_{0.05}$	$q_{0.95}$	$q_{0.05}$	$q_{0.95}$
EP/ST	7117	8056	2043	2113	0.0345	0.0392	0.0293	0.0431
EP/NHCW	7484	7728	2043	2113	0.0361	0.0379	0.0293	0.0431
MP/ST	7194	8118	2034	2104	0.0347	0.0393	0.0289	0.0435
MP/NHCW	7485	7731	2034	2104	0.0359	0.0377	0.0289	0.0435
SP	7463	7739	2034	2104	0.0358	0.0377	0.0289	0.0435

Table 16

Quantiles computed from $N = 1,000$ simulation runs for the estimated number of true null hypotheses ($\hat{m}_0$), the number of rejected null hypotheses (R), estimated FDR ($\widehat{FDR}$), and false positive fraction ($V/\max\{1,R\}$) for p-value significance threshold 0.01. The number of tests and the number of true null hypotheses were $m = 10,000$ and $m_0 = 9,000$, respectively. Data were generated as described in Section 6 of the main paper with $\lambda = 1$.

	$\hat{m}_0$		R		$\widehat{\text{FDR}}$		$V/\max\{1,R\}$	
	$q_{0.05}$	$q_{0.95}$	$q_{0.05}$	$q_{0.95}$	$q_{0.05}$	$q_{0.95}$	$q_{0.05}$	$q_{0.95}$
EP/ST	8942	10000	332	387	0.2388	0.2894	0.2127	0.2861
EP/NHCW	9311	9601	332	387	0.2445	0.2878	0.2127	0.2861
MP/ST	9005	10000	330	385	0.2417	0.2916	0.2132	0.2865
MP/NHCW	9315	9602	330	385	0.2448	0.2879	0.2132	0.2865
SP	9301	9608	330	385	0.2448	0.2879	0.2132	0.2865

Table 17

Quantiles computed from $N = 1,000$ simulation runs for the estimated number of true null hypotheses ($\hat{m}_0$), the number of rejected null hypotheses (R), estimated FDR ($\widehat{FDR}$), and false positive fraction ($V/\max\{1,R\}$) for p-value significance threshold 0.01. The number of tests and the number of true null hypotheses were $m = 10,000$ and $m_0 = 9,000$, respectively. Data were generated as described in Section 6 of the main paper with $\lambda = 2$.

	$\hat{m}_0$		R		$\widehat{\text{FDR}}$		$V/\max\{1,R\}$	
	$q_{0.05}$	$q_{0.95}$	$q_{0.05}$	$q_{0.95}$	$q_{0.05}$	$q_{0.95}$	$q_{0.05}$	$q_{0.95}$
EP/ST	8629	9648	643	701	0.1276	0.1481	0.1145	0.1547
EP/NHCW	9018	9281	643	701	0.1312	0.1447	0.1145	0.1547
MP/ST	8713	9743	637	696	0.1285	0.1490	0.1148	0.1554
MP/NHCW	9015	9280	637	696	0.1310	0.1445	0.1148	0.1554
SP	9006	9284	637	696	0.1307	0.1448	0.1148	0.1554

Table 18

Quantiles computed from $N = 1,000$ simulation runs for the estimated number of true null hypotheses ($\hat{m}_0$), the number of rejected null hypotheses (R), estimated FDR ($\widehat{FDR}$), and false positive fraction ($V/\max\{1, R\}$) for p-value significance threshold 0.01. The number of tests and the number of true null hypotheses were $m = 10,000$ and $m_0 = 9,000$, respectively. Data were generated as described in Section 6 of the main paper with $\lambda = 3$.

	$\hat{m}_0$		R		$\widehat{\text{FDR}}$		$V/\max\{1, R\}$	
	$q_{0.05}$	$q_{0.95}$	$q_{0.05}$	$q_{0.95}$	$q_{0.05}$	$q_{0.95}$	$q_{0.05}$	$q_{0.95}$
EP/ST	8526	9555	866	916	0.0959	0.1092	0.0855	0.1161
EP/NHCW	8943	9129	866	916	0.0992	0.1060	0.0855	0.1161
MP/ST	8624	9672	861	913	0.0964	0.1095	0.0854	0.1160
MP/NHCW	8934	9131	861	913	0.0989	0.1054	0.0854	0.1160
SP	8922	9132	861	913	0.0989	0.1053	0.0854	0.1160

Table 19

Means and standard errors of the means (sem) computed from $N = 1,000$ simulation runs for the estimated number of true null hypotheses ($\hat{m}_0$), the number of rejected null hypotheses (R), estimated FDR ($\widehat{FDR}$), and false positive fraction ($V/\max\{1, R\}$) for p-value significance threshold 0.01. The number of tests and the number of true null hypotheses were $m = 10,000$ and $m_0 = 7,500$, respectively. Data were simulated from ALL microarray data with $\lambda = 2$. The MP and SP approaches were based on $n = 1,000$ and $h = 10$, $n = 1,000$, respectively.

	$\hat{m}_0$		R		$\widehat{\mathrm{FDR}}$		$V/\max\{1, R\}$	
Method	mean	sem	mean	sem	mean	sem	mean	sem
1	7785	49	1515	4.481	0.0527	0.0004	0.0488	0.0017
2	7538	36	1515	4.481	0.0509	0.0003	0.0488	0.0017
3	7847	49	1505	4.457	0.0530	0.0004	0.0490	0.0017
4	7544	36	1505	4.457	0.0509	0.0003	0.0490	0.0017
5	7550	35	1505	4.457	0.0509	0.0003	0.0490	0.0017

Table 20

Quantiles computed from $N = 1,000$ simulation runs for the estimated number of true null hypotheses ($\hat{m}_0$), the number of rejected null hypotheses (R), estimated FDR ($\widehat{FDR}$), and false positive fraction ($V/\max\{1,R\}$) for p-value significance threshold 0.01. The number of tests and the number of true null hypotheses were $m = 10,000$ and $m_0 = 7,500$, respectively. Data were simulated from ALL microarry data with $\lambda = 2$.

	$\hat{m}_0$		R		$\widehat{\text{FDR}}$		$V/\max\{1,R\}$	
	$q_{0.05}$	$q_{0.95}$	$q_{0.05}$	$q_{0.95}$	$q_{0.05}$	$q_{0.95}$	$q_{0.05}$	$q_{0.95}$
EP/ST	5089	10000	1346	1760	0.0298	0.0713	0.0080	0.1610
EP/NHCW	5240	8593	1346	1760	0.0306	0.0632	0.0080	0.1610
MP/ST	5124	10000	1339	1762	0.0300	0.0712	0.0083	0.1633
MP/NHCW	5267	8596	1339	1762	0.0302	0.0632	0.0083	0.1633
SP	5269	8600	1339	1762	0.0309	0.0633	0.0083	0.1633