SMALL SAMPLE MULTIPLE TESTING

WITH APPLICATION TO CDNA MICROARRAY DATA

A Dissertation

by

ERIC POOLE HINTZE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2005

Major Subject: Statistics

SMALL SAMPLE MULTIPLE TESTING

WITH APPLICATION TO CDNA MICROARRAY DATA

A Dissertation

by

ERIC POOLE HINTZE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Michael Sherman
Committee Members,	F. Michael Speed
	Marina Vannucci
	Rajesh C. Miranda
Head of Department,	Simon Sheather

August 2005

Major Subject: Statistics

ABSTRACT

Small Sample Multiple Testing with Application
to cDNA Microarray Data. (August 2005)
Eric Poole Hintze, B.S., Brigham Young University;
M.S., Brigham Young University
Chair of Advisory Committee: Dr. Michael Sherman

Many tests have been developed for comparing means in a two-sample scenario. Microarray experiments lead to thousands of such comparisons in a single study. Several multiple testing procedures are available to control experiment-wise error or the false discovery rate. In this dissertation, individual two-sample tests are compared based on accuracy, correctness, and power. Four multiple testing procedures are compared via simulation, based on data from the lab of Dr. Rajesh Miranda. The effect of sample size on power is also carefully examined. The two sample *t*-test followed by the Benjamini and Hochberg (1995) false discovery rate controlling procedure result in the highest power.

ACKNOWLEDGMENTS

I thank my advisor, Dr. Michael Sherman, for his positive advice throughout the development of the dissertation. I thank Dr. F. Michael Speed for his encouragement and help with computational resources. I thank Dr. Marina Vannucci and Dr. Rajesh Miranda for their help in the understanding of microarrays. I also thank Dr. Miranda for allowing me to use his data in the dissertation.

I especially thank my wife, Valora, for her constant support, love, and patience. I thank my children, Spencer, Ashley Jo, and Brooklyn, for making my dissertation experience a complete adventure.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

1. INTRODUCTION

A common problem in genomics is determining which among the thousands of genes in the DNA of an organism are differentially expressed when a treatment is given. Until recently, the number of genes in humans was commonly cited to be around 100,000. The human genome projects of Celera and the public consortium of scientists both found the number of genes in humans instead to be around 40,000, “give or take a few thousand” (Pennisi, 2001). Forty-thousand might still be considered a formidable number of genes to examine were it not for developments in DNA and RNA technology. These developments allow scientists to gather expression data about thousands of genes at a time using microarrays.

On a microarray there are several thousand probes of known identity, each corresponding to a gene of interest. The mRNA expressed in an individual is obtained from some of the individual’s cells (i.e., blood or tissue) and converted to fluorescently labeled cDNA. When the cDNA is exposed to a microarray, its segments bind preferentially to the probes to which they complement. The cDNA corresponding to the mRNA from genes which are expressed in higher quantities will hybridize to the corresponding probe in higher quantities. The amount of hybridized material for each probe can then be measured using the intensity of fluorescence from each probe when exposed to laser scanning. The result is several thousand intensities that correspond in some degree to the amount of mRNA expression for each of those genes in that individual. This can be repeated among several individuals who have and have not received a treatment of interest.

By comparing the intensities of expression of control individuals and treatment individuals, specific genes which are differentially expressed may be determined using a statistical test. Due to the variation in gene expression from individual to individual and the sheer number of comparisons, it is clear that if genes are compared without

---

This dissertation follows the style of the *Journal of the American Statistical Association*.

adjustment for multiplicity, some (or many) will be declared different by chance alone. It is desirable for researchers to limit the number of genes which are wrongly declared to be differentially expressed, but at the same time find as many of the genes as possible for which expression truly is different.

Superior methods for determining which genes are to be declared differentially expressed should be successful in two ways. First, the method should do well at assigning low *individual* unadjusted *P*-values to genes which are expressed differently in the control and treatment groups. Conversely, high *P*-values should be assigned to those genes which are not differentially expressed. Second, the method should make efficient use of the individual *P*-values so as to identify the most differentially expressed genes while appropriately limiting the number of false positives.

One distinguishing feature of microarray experiments is sample size. Although the price of producing microarray data has steadily decreased, the number of replicates in a microarray experiment is typically in the range of 2 to 5 (Yang and Speed, 2003).

Another common feature in microarray experiments is dependent gene expression. There are many groups of genes which are co-regulated and thus have correlated (sometimes highly) expression levels. This casts doubt on the assumption of independence of tests.

The purpose of this dissertation is two-fold. First, I compare several common two-sample testing methods under several distributional conditions in terms of *accuracy* and *correctness*, which are defined in the text. The focus is determining the proper test for comparing microarray expression levels with small sample sizes (e.g., n = 3 or 5). The primary methods compared are the *t*-test (Fisher, 1925), Welch's (1947, 1949) *t*-test, the permutation test, and two bootstrap *t*-tests. I discuss briefly some other nonparametric tests such as the two sample median test, Fisher's (1934) exact test, the Wilcoxon (1945) rank test (also called Wilcoxon signed rank test or Wilcoxon-Mann-Whitney test [Mann and Whitney, 1947]), and the use of trimmed means.

The second major objective of the dissertation is determining how the accuracy and correctness of the individual tests affect the identification of differentially expressed

genes when thousands of comparisons are made simultaneously. These are measured in terms of family-wise error rate, false discovery rate, and power. Multiple testing adjustment procedures compared will be the Bonferroni correction procedure, Benjamini and Hochberg's (1995) false discovery rate control procedure, and Westfall and Young's (1993) single step maxT resampling based procedure. I also explore the construction of the empirical null hypothesis, recently described by Efron (2004).

Comparisons of these methods are made via simulation and microarray data provided by the lab of Dr. Rajesh Miranda (Texas A&M University, Departments of Anatomy and Neurobiology). An experiment was run in the Miranda lab to examine the effect of CD133 on gene expression. It is of interest to determine which among 22,276 genes are up- or down-regulated when cells are injected with CD133, which is known to cause cells to differentiate. Three experimental units received the CD133 treatment and 3 were controls. Microarray measurements of expression intensity were made for each of the 22,276 genes for the 6 experimental units.

The details of the two sample tests, multiple testing adjustment procedures, microarray data normalization, and simulations are described in the sections that follow.

2. TWO SAMPLE TESTS

2.1. Background

The problem of comparing two population means has been studied extensively over the past hundred years. The null hypothesis is that the means are equal,

$$H_0 : \mu_1 = \mu_2,$$

with three common alternative hypotheses,

$$H_a : \mu_1 < \mu_2,$$

$$H_a : \mu_1 \neq \mu_2, \text{ or}$$

$$H_a : \mu_1 > \mu_2,$$

which are selected according to the nature of the experiment or study. The most common test statistic used for evaluation of the chosen hypotheses based on samples $y_{11},\ldots,y_{1n_1}$ and $y_{21},\ldots,y_{2n_2}$ is the *t*-statistic

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

with $s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$, where s_1^2 and s_2^2 are the usual sample variances. If samples $y_{11},\ldots,y_{1n_1}$ and $y_{21},\ldots,y_{2n_2}$ come from two normal populations with equal variances then the statistic *t* is known to follow "Student's" *t* distribution with degrees of freedom equal to $n_1 + n_2 - 2$ (Fisher, 1925). An appropriate *P*-value for the test may then be calculated as the probability of *t* being as or more extreme than the one obtained, based on this distribution. Problems may arise, however, when the two underlying distributions are not normally distributed and/or have differing variances. These problems are often amplified when the sample sizes also differ. Unfortunately in practice little is usually known about the true underlying distributions from which the two samples come, particularly when sample sizes are small.

To understand the effect of nonnormality on *t*, consider two samples y_1 and y_2 from populations with mean zero and unit variance, but with (possibly) differing skewness, say $\gamma_1(y_1)$ and $\gamma_1(y_2)$, and/or (possibly) differing kurtosis, say $\gamma_2(y_1)$ and $\gamma_2(y_2)$. The distribution of each of the two populations can be described nearly by the first four terms of the Edgeworth expansion:

$$f(x) = \phi(x) - \frac{\gamma_1}{3!}\phi^{(3)}(x) + \frac{\gamma_2}{4!}\phi^{(4)}(x) + \frac{\gamma_1^2}{72}\phi^{(6)}(x)$$

where $\phi(x)$ is the p.d.f. of the standard normal distribution and $\phi^{(r)}(x)$ its *r*th derivative. Building upon the work of Geary (1947), Gayen (1950) used this expansion to derive the first four raw (not central) moments of *t*:

$$\mu_1'(t) \cong \frac{1}{v_1^{1/2}}\left\{-\frac{1}{2}\left(\gamma_1(y_1)-\gamma_1(y_2)\right)\left(\frac{1}{v_2}-\frac{2}{v_2^2}\right)+\frac{1}{2}\left(\gamma_1(y_1)+\gamma_1(y_2)\right)\frac{\left(n_2-n_1\right)}{2n_1n_2}\left(\frac{1}{v_2}+\frac{3}{4v_2^2}\right)\right\},$$

$$\begin{aligned}\mu_2'(t) \cong \frac{1}{v_1}\Bigg\{&\left(1+\frac{2}{v_2}+\frac{6}{v_2^2}\right)v_1+\left(\gamma_2(y_1)-\gamma_2(y_2)\right)\frac{\left(n_2^2-n_1^2\right)}{n_1n_2}\frac{1}{v_2^2}\\&+\left(\gamma_2(y_1)+\gamma_2(y_2)\right)\left(\frac{\left(n_2-n_1\right)^2}{2n_1n_2}-1\right)\frac{v_1}{v_2^2}+\left(\gamma_1(y_1)-\gamma_1(y_2)\right)^2\frac{\left(16-5v_1\right)}{8v_2^2}\\&+\left(\gamma_1^2(y_1)-\gamma_1^2(y_2)\right)\frac{\left(n_2^2-n_1^2\right)}{4n_1n_2}\frac{27}{v_2^3}-\left(\gamma_1(y_1)+\gamma_2(y_2)\right)^2\frac{5v_1}{8v_2^2}\Bigg\},\end{aligned}$$

$$\mu_3'(t) \cong \frac{1}{v_1^{3/2}}\left\{\frac{1}{2}\left(\gamma_1(y_1)-\gamma_1(y_2)\right)\left(\frac{1}{n_1^2}+\frac{1}{n_2^2}-\frac{9}{n_1n_2}\right)+\frac{1}{2}\left(\gamma_1(y_1)+\gamma_1(y_2)\right)\left(\frac{1}{n_1^2}-\frac{1}{n_2^2}\right)\right\},$$

$$\begin{aligned}\mu_4'(t) \cong \frac{1}{v_1^2}\Bigg\{&\left(3+\frac{18}{v_2}+\frac{102}{v_2^2}\right)v_1^2+\left(\gamma_2(y_1)-\gamma_2(y_2)\right)\frac{\left(n_2-n_1\right)}{2n_1n_2}\left(\frac{\left(n_2-n_1\right)^2}{n_1n_2}\left(\frac{v_1}{v_2}-\frac{12v_1}{v_2^2}\right)\right.\\&\left.-\frac{18v_1}{v_2}-\frac{66v_1}{v_2^2}\right)+\left(\gamma_2(y_1)+\gamma_2(y_2)\right)\left(\frac{\left(n_2-n_1\right)^2}{2n_1n_2}\left(\frac{v_1^2}{v_2}-\frac{12v_1^2}{v_2^2}\right)-2\left(\frac{v_1^2}{v_2}-\frac{15v_1^2}{v_2^2}\right)\right)\\&+\left(\gamma_1(y_1)-\gamma_1(y_2)\right)^2\left(-\frac{12v_1}{v_2}+\frac{2v_1\left(3v_1+26\right)}{v_2^2}+\frac{\left(n_2-n_1\right)^2}{n_1^2n_2^2}\left(48-\frac{4}{v_2}-\frac{27}{4n_1n_2}\right)\right)\end{aligned}$$

$$+\left(\gamma_1^2(y_1)-\gamma_1^2(y_2)\right)\frac{\left(n_1^2-n_2^2\right)}{n_1 n_2}\left(\left(\frac{4v_1}{v_2}-\frac{81v_1}{v_2^2}\right)+\frac{27(n_2-n_1)^2}{n_1^2 n_2^2}\left(\frac{1}{2v_2}+\frac{1}{v_2^2}\right)\right)$$

$$+\left(\gamma_1(y_1)+\gamma_1(y_2)\right)^2\left(\frac{6v_1^2}{v_2^2}-\frac{27}{4}\frac{\left(n_2^2-n_1^2\right)^2}{n_1^3 n_2^3}\frac{1}{v_2^2}\right)\Bigg\}.$$

where n_1 and n_2 are the sample sizes and

$$v_1=\frac{1}{n_1}+\frac{1}{n_2}, \qquad v_2=n_1+n_2-2$$

From these approximate moments, approximate central moments of *t* can be constructed using (see, e.g., Stuart and Ord, 1994)

$$\mu_2(t)=\mu_2'(t)-\mu_1'(t)^2,$$

$$\mu_3(t)=\mu_3'(t)-3\mu_2'(t)\mu_1'(t)+2\mu_1'(t)^3,$$

$$\mu_4(t)=\mu_4'(t)-4\mu_3'(t)\mu_1'(t)+6\mu_2'(t)\mu_1'(t)^2-3\mu_1'(t)^4,$$

which can then be used to calculate the approximate skewness and kurtosis of *t* using

$$\gamma_1(t)=\frac{\mu_3(t)}{\mu_2^{3/2}(t)}=\frac{\mathrm{E}(t-\mu_t)^3}{\sigma_t^3},$$

$$\gamma_2(t)=\frac{\mu_4(t)}{\mu_2^2(t)}-3=\frac{\mathrm{E}(t-\mu_t)^4}{\sigma_t^4}-3.$$

The skewness and kurtosis of a true 'Student's' *t*-distribution are

$$\gamma_1(t)=0, \qquad \gamma_2(t)=6/(\eta-4), \qquad \eta>4$$

where η represents the degrees of freedom. Comparisons between the estimated and expected theoretical distributions can be made to determine the effects of nonnormality of the two sampled distributions on the distribution of the test statistic *t*. Alternatively, the distribution of *t* can be simulated.

In general, if $n_1 \cong n_2$ and if one can assume $\gamma_1(y_1) \cong \gamma_1(y_2)$ and $\gamma_2(y_1) \cong \gamma_2(y_2)$, then the skewness and kurtosis of the sampled distributions will have very little effect on the distribution of the *t*-statistic. Pearson and Please (1975) simulated 2,000 pairs of samples of equal sample size, skewness, and kurtosis of size 10 and 25 from

distributions ranging in skewness ($\gamma_1(y_1) = \gamma_1(y_2)$) from 0 to 0.8, and kurtosis ($\gamma_2(y_1) = \gamma_2(y_2)$) from -1 to 1.4. Almost without exception the proportion of the 2,000 tests with *t*-statistics in the outer 5% of the tails was between 0.04 and 0.06. Similar results were found by Pearson (1929). When sample size, skewness, and kurtosis are not approximately equal, less is known of what the distribution of the *t*-statistic may be, mostly because of the enormity of the number of possible combinations of n_1, n_2, $\gamma_1(y_1), \gamma_1(y_2), \gamma_2(y_1),$ and $\gamma_2(y_2)$. A reasonable idea might be obtained by estimating these parameters and using Gayen's approximate moments described above to get reasonable estimates of $\gamma_1(t)$ and $\gamma_2(t)$. Unfortunately, samples of size five or even ten do not allow for accurate estimation of skewness and kurtosis of underlying distributions. Some larger sample examples are given in Geary (1947) and Gayen (1950).

An oft-used technique for correcting for non-normality is the use of transformations such as the logarithmic or square root transformations. Transformations may be particularly useful if there appear to be marked differences in variances among the samples. A difficulty that arises in transforming two groups occurs when one group appears to benefit from a transformation while the other does not.

Nonparametric and/or robust estimation techniques are also often employed when the distributional assumptions of the common *t*-test are not met. The two sample median test, Fisher's (1934) exact test, the Wilcoxon (1945) rank test (also called Wilcoxon signed rank test or Wilcoxon-Mann-Whitney test [Mann and Whitney, 1947]), Pitman's (1937) permutation test, the bootstrap method (Efron, 1979, 1982), and the use of trimmed means are examples of such techniques, the Wilcoxon rank test being the most popular. For details, see, for example, Miller (1986), or Ostle and Malone (1988). These methods are often considered to be useful when outliers are present.

Recent simulation studies indicate that (at least) some non-parametric rank procedures (i.e., Wilcoxon's sign rank test) perform very poorly when variance heterogeneity is a problem, even for equal sample sizes. The inferiority in performance is even more pronounced when the underlying distributions are skewed, which is the

usual reason for using such tests (see Zimmerman, 2004). Use of rank tests is not recommended when equality of variance is in question.

Differing variances among the two populations sampled can also be a formidable challenge in a two sample comparison, especially when the sample sizes differ. Miller (1986) notes that for the usual *t*-statistic, "the variance for the larger sample tends to dominate the denominator of the *t*-statistic." Transformations can be useful in correcting the problem of unequal variance. Another approach is to use a different *t*-test. When the populations compared have unequal variance, but are both normally distributed, the resulting test of $H_0 : \mu_1 = \mu_2$ is known as the Behrens-Fisher problem. The statistic usually recommended for testing in this scenario is one developed by Welch (1947, 1949). The statistic is

$$t_W = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

or Welch's *t*-statistic. The use of this statistic relies on the asymptotic convergence of the sample variances to the true variances, and is certainly appropriate for large samples. For small or moderate samples t_W approximates 'Student's' *t*-distribution, with estimated degrees of freedom

$$\hat{v} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}.$$

The statistic t_W usually outperforms *t* (higher power when nominal α is preserved) when the variances of the sampled populations differ considerably (Welch, 1947, 1949). When $n_1 = n_2$, *t* and t_W are equivalent, except for the degrees of freedom used in the test.

Using the first four terms of the Edgeworth series, Bhattacharjee (1968) derived the approximate distribution of *t* and t_W based on n_1, n_2, $\gamma_1(y_1), \gamma_1(y_2)$, $\gamma_2(y_1)$, $\gamma_2(y_2)$, and σ_1 and σ_2, thus generalizing the work of Geary (1947) and Gayen (1950) to allow

for unequal variances. Bhattacharjee's illustrations show a wide range of effects of various combinations of these parameters on the two-sided tail area of t and t_W as compared to the customary 'Student's' t distribution two-sided tail area. It should be noted that Bhattacharjee uses degrees of freedom $n_1 + n_2 - 2$ for both t and t_W, and not adjusted degrees of freedom for t_W. Bhattacharjee concludes "If the populations are non-normal and the variances are unequal, the normal theory tests on the basis of the criteria [t and t_W], may under certain circumstances give misleading results. The effect may, however, be minimized by taking samples with equal number of observations."

In the context of microarray data, the sample sizes are usually very small. The chips used for a microarray experiment are currently very expensive so that only a small number (perhaps 2-5) of individuals are typically in each of the treatment and control groups. The small sample size presents difficulty in determining the distribution of the individual test statistics to be used. The equality of variance assumption for expression levels between the two groups may also be questionable. Consequently, it is important to use a method which is not tied to the central limit theorem or the equality of variance assumption. Candidates for test statistic null distributions would then be the one suggested by Welch (1947) or null distributions based on resampling methods.

Some of the relative merits of the bootstrap and permutation resampling techniques for testing the difference of two means of samples with unknown underlying distributions are discussed briefly in Efron and Tibshirani (1993), Good (2000), and Troendle et al. (2004). The following is a summary of how the tests are run and how they compare to each other and to some traditional methods.

2.2. General Comparison of the *t*-test, Welch's *t*-test, the Bootstrap Within Test, the Bootstrap Across Test, and the Permutation Test

We first consider the computation details of the permutation and bootstrap methods.

2.2.1 Permutation Achieved Significance Levels

A suitable statistic which properly compares the means must first be chosen, usually t or t_W. Recall that t and t_W are equal when the sample sizes n_1 and n_2 are equal. There are two ways to obtain the permutation distribution of t or t_W. In the first, all $n_1 + n_2 = N$ individuals are pooled and then randomly assigned to two groups, each of size n_1 and n_2, without replacement. The test statistic is then computed for the reassigned data, and called t^* or t_W^*. This process of resampling is repeated B times to obtain $\boldsymbol{t}^* = t_1^*, t_2^*, t_3^*, \ldots, t_B^*$, or $\boldsymbol{t}_W^* = t_{W1}^*, t_{W2}^*, t_{W3}^*, \ldots, t_{WB}^*$. The distribution of $t_1^*, t_2^*, t_3^*, \ldots, t_B^*$, or $t_{W1}^*, t_{W2}^*, t_{W3}^*, \ldots, t_{WB}^*$ is assumed to approximate the true distribution of t or t_W. Alternatively, if the sample sizes of the two groups are sufficiently small, all $\binom{N}{n_1}$ permutations of size n_1 and n_2 may be enumerated. The resulting $t_1^*, t_2^*, t_3^*, \ldots, t_{\binom{N}{n_1}}^*$, or $t_{W1}^*, t_{W2}^*, t_{W3}^*, \ldots, t_{W_{\binom{N}{n_1}}}^*$ will then serve as the sampling distribution of t or t_W. For a two-sided alternative, when resampling is used to obtain the null distribution of t or t_W, the permutation achieved significance level (ASL), following the naming given by Efron and Tibshirani (1993), can be defined in two ways:

$$\text{ASL}_{\text{perm},1} = \#(|\boldsymbol{t}^*| \geq |t|)/B \quad \text{or}$$

$$\text{ASL}_{\text{perm},2} = (1 + \#(|\boldsymbol{t}^*| \geq |t|))/(1 + B)$$

The choice of $\text{ASL}_{\text{perm},1}$ or $\text{ASL}_{\text{perm},2}$ depends on whether or not one wants to consider the observed t or t_W as part of the resampling distribution. We shall see that this choice becomes important when ASL is less than about 0.005.

If the complete distribution of all $\binom{N}{n_1}$ permutations of size n_1 and n_2 are enumerated, then we have

$$\text{ASL}_{\text{perm},3} = \#(|\boldsymbol{t}^*| \geq |t|) / \binom{N}{n_1}$$

The choice of "≥" rather than ">" in the above equations is somewhat arbitrary, but does affect the value of $\text{ASL}_{\text{perm},1}$, $\text{ASL}_{\text{perm},2}$, or $\text{ASL}_{\text{perm},3}$ because the resampled or enumerated distribution of *t* or t_W is not continuous.

2.2.2. Bootstrap Achieved Significance Levels

The bootstrap distribution of *t* or t_W is obtained in a similar manner to that of the resampling based permutation distribution. However, in this case, the observations are first centered for each group so that the distribution of *t* or t_W is reconstructed in a manner that reflects the null hypothesis. Resampling of the centered observations can then be done within each group or pooled across the groups. Whether resampling within groups or across groups, bootstrap resampling is carried out with replacement. Statistics for each resample are obtained as in the permutation method, forming estimated distributions $\boldsymbol{t}^* = t_1^*, t_2^*, t_3^*, \ldots, t_B^*$, or $\boldsymbol{t}_W^* = t_{W1}^*, t_{W2}^*, t_{W3}^*, \ldots, t_{WB}^*$. Thus, there are four possible designations for the ASL,

$$\text{ASL}_{\text{boot,within},1} = \#(|\boldsymbol{t}^*| \geq |t|) / B,$$

$$\text{ASL}_{\text{boot,within},2} = (1 + \#(|\boldsymbol{t}^*| \geq |t|)) / (1 + B),$$

$$\text{ASL}_{\text{boot,across},1} = \#(|\boldsymbol{t}^*| \geq |t|) / B, \text{ or}$$

$$\text{ASL}_{\text{boot,across},2} = (1 + \#(|\boldsymbol{t}^*| \geq |t|)) / (1 + B).$$

The choice among the ASL definitions depends on the type of resampling done (within groups or across groups) and whether or not we want to consider the observed *t* or t_W as part of the resampling distribution. Since resampling is done with replacement, complete enumeration of all resamplings of the $n_1 + n_2 = N$ individuals is prohibitively large, even for small sample sizes, so that a complete enumeration definition for the ASL is not included here.

2.2.3. Comparison of Null Distributions

The histograms in Appendix A give an idea of how the individual resampled *t* distributions based on the permutation and bootstrap resampling methods appear under the null hypothesis of equal means. For figures A-2 through A-7, two random samples of size five were generated, each from a normal distribution of mean 0 and standard deviation 1. The samples were resampled B = 100,000 times under each of the permutation and bootstrap methods. The resulting *t*-statistics from the 100,000 resamplings are shown as histograms with the true *t* distribution with 8 degrees of freedom overlaid. For figures A-8 through A-10 the sample sizes were increased from five per group to ten per group. It is readily apparent that the resampling based *t*-distributions differ substantially from the known *t* distribution, particularly for samples of size five. The permutation *t* distribution looks the least like the true distribution.

Because it is already well-known that the test based on the known *t* distribution is ideal for samples from identical normal distributions, figures A-11 through A-28 focus on the *t*-statistic null distributions when one of the underlying populations differs from standard normal. The figures are in groups of three. The first figure of a group (e.g., figure A-11) shows the underlying population distributions. The second figure (e.g., figure A-12) shows the true *t*-distribution based on 10,000,000 samples of size 5 from each of the distributions of the first figure. These are followed by a third figure (e.g., figure A-13) examining resampling based *t*-distributions created from single samples from the two distributions in question. Several underlying population distributions are examined in these figures ranging from differing variance to differing shape or both.

The non-normal distribution used is based on the Chi-Square distribution with 1 or 3 degrees of freedom. When the Chi-Square distribution is used, each value has the mean subtracted followed by division by the appropriate number to give the desired mean and variance.

The histograms of Appendix A illustrate some important aspects of the null distributions produced by the three resampling methods. First, the permutation and bootstrap across methods always generate a null distribution which is symmetric,

regardless of what the true null distribution should be. Second, the bootstrap within method and particularly the permutation method seem to produce distributions that are less stable than those produced by the bootstrap across method. Third, the true null *t* distributions do not depart substantially from the common *t* distribution unless one of the underlying distributions is highly skewed.

2.2.4. Accuracy and Correctness

2.2.4.1. Definitions and ASL Formulation

This is a good point to discuss the concepts of *accuracy* and *correctness*, following the terminology of Efron and Tibshirani (1993). A test is *accurate* if

$$\text{Prob}(ASL \leq \alpha) = \alpha$$

when the null hypothesis is true, and

$$\text{Prob}(ASL \leq \alpha) = \text{ Expected Power}$$

when the null hypothesis is false. The expected power is based on a known most powerful test. Thus, a test is accurate if the nominal level and power are preserved. A test is more accurate than another if $\text{Prob}(ASL \leq \alpha)$ is closer to α under the null hypothesis and if $\text{Prob}(ASL \leq \alpha)$ is closer to the expected power under the alternative hypothesis.

Correctness of a method indicates that the observed ASL is close to the *P*-value a known optimal method would give for each data set. A test is more correct than another if it yields ASLs which are closer to the correct method *P*-values. Each correct method *P*-value is based on a known distribution. For example, if two samples are known to come from normal distributions with equal variance, the known optimal method for comparing the means is the two-sample *t*-test. The *t*-statistics from this method are known to follow Student's *t* distribution. For a given data set, ASLs from any other method (i.e., permutation test or bootstrap test) can be compared to the known correct *P*-value of the two-sample *t*-test. An ASL which is close to this *P*-value is more correct than an ASL which is further away.

Accuracy when the null hypothesis is false and correctness can only be evaluated for a method if a known most powerful test is available. It is for this reason that the bootstrap, permutation, and *t*-tests are first compared using samples with normal distributions of equal variance (of 1). The known Student's *t* and noncentral *t* distributions can be used to evaluate the bootstrap and permutation resampling methods for accuracy and correctness. For other null and alternative distributions (i.e., nonnormal or unequal variance), the unknown optimal test statistic distributions are estimated through simulation.

2.2.4.2. Estimated Test Size Comparison

A comparison of the estimated size for the two bootstrap methods, the permutation method, the known-size common *t*-test, and Welch's *t*-test can be found in Appendix B. Each graph represents 20,000 simulated two-sample data sets. The means for the distributions from which each of the samples are taken are both zero, corresponding to the null hypothesis of equal means. Other parameters such as sample size in each group, variance, distribution types, and number of resamplings *B* are specific to each graph.

For the bootstrap and permutation tests, there are four possible definitions for the ASL. Because of the discrete nature of the compared sampling distributions, each definition may result in a different estimated size. The four definitions are

$$\text{ASL} = \#(|\boldsymbol{t}^*| \geq |t|)/B$$

$$\text{ASL} = \#(|\boldsymbol{t}^*| > |t|)/B$$

$$\text{ASL} = (1+\#(|\boldsymbol{t}^*| \geq |t|))/(1+B)$$

$$\text{ASL} = (1+\#(|\boldsymbol{t}^*| > |t|))/(1+B)$$

Figures B-1 through B-8 of Appendix B allow us to compare the effects of the definition of ASL and the choice of *B*. The graphs are created by finding the proportion of ASLs below small increments of alpha for each method and plotting them against those increments. The same 20,000 simulated data sets were used for all methods and for

all four definitions with B = 999 (figures B-1 through B-4). A new set of 20,000 simulated data sets was used for B = 1,000 (figures B-5 through B-8).

The choice of B and adding one to the numerator and denominator of the ASL definition appear to have very little effect on the estimated sizes. The estimated size based on the bootstrap across resampling method follows the *t*-test estimated size closely. When samples are of size 5 per group, the permutation method yields estimated sizes which are slightly below or slightly above the *t*-test estimated sizes, depending on whether equality is included in the ASL definition. Including equality produces the conservative result. For subsequent comparisons I use the definition of ASL:

$$\text{ASL} = (1 + \#(|\boldsymbol{t}^*| \geq |t|))/(1 + B)$$

because it is conservative for the small sample permutation method and since it is natural to include the observed t as one of the t^*'s. Also, B = 999 will be used.

2.2.4.3. Two Sample Test Accuracy Comparison

Figures B-9 through B-15 of Appendix B can be used to compare the accuracy of five tests: the standard *t*-test, Welch's *t*-test, the permutation test, the within sample bootstrap test, and the across sample bootstrap test. The specifics of the distributions from which the samples are taken are shown below each graph. The black 45 degree line represents the true level. Estimated levels are given explicitly for known α = .01, .05, and .10 below each graph in a separate table of the appendix.

Welch's *t*-test is seen to preserve the nominal error rate in all scenarios except those for which both the distribution shape and the variance of the two underlying distributions differ. The bootstrap within test is generally conservative while the permutation test, bootstrap across test, and *t*-test are general anti-conservative when the underlying distributions are not equal.

2.2.4.4. Power

The comparison of tests based on estimated power is done in the same way as that used for comparing estimated size, except that the means of the underlying

distributions from which the data are sampled are not equal. The results are found in Appendix C. The means differ by the amount shown below each graph. Here, again, the graphs are created by finding the proportion of ASLs below small increments of alpha and plotting them against those increments for each method. The same 20,000 simulated data sets are used for all methods. Care should be taken when interpreting these figures. Usually, the power of a test is evaluated for a given test size. In this case, the figures of Appendix B indicate that for many of the tests the size is very different from the nominal α. Power should be compared only after consulting the corresponding estimated size for the same test.

A comparison of the powers for the different tests follows a general trend of higher powers for the permuation, bootstrap across, and *t*-tests, although nominal sizes are seldom maintained for these tests. Welch's *t*-test and the bootstrap within test are more conservative. This follows the pattern seen in the estimated sizes for the tests. Among the two tests which maintain the correct size for most distribution scenarios, Welch's *t*-test clearly has higher power than the bootstrap within test. If little or nothing is known about the underlying distributions from which small samples are taken, or if the underlying distributions are known to differ in variance or distribution, Welch's *t*-test is the recommended test based on accuracy and power.

2.2.4.5. Comparison of Correctness

We turn now from accuracy to correctness. Recall that correctness implies that the individual ASLs are close to the known correct *P*-values, where the known correct P-values are based on an optimal test. The correctness can be gauged by the mean square error of the observed ASLs from the known correct *P*-values for each method.

When the sampling distribution of the test statistic *t* is known, the *P*-value obtained from a specific realized *t* is obtained directly from that known sampling distribution. For example, suppose two samples of size 5 result in the two-sample test statistic $t = 2.306$. If this value is compared to Student's *t* distribution with 8 degrees of freedom, the two-sided test *P*-value is 0.05. If two completely different samples result in

a test statistic that is also $t = 2.306$, the P-value for the test will still be 0.05. This is not the case when the sampling distribution of the test statistic is estimated from the data as in the bootstrap and permutation methods. Here, different samples typically result in different estimated sampling distributions. For example, using a resampling based distribution for t, an ASL for one two-sample data set with observed $t = 2.306$ may be 0.045 while for another two-sample data set with $t = 2.306$ the ASL might be 0.058, since the sample itself is used to create the sampling distribution. This concept can be visualized using the example shown in Figures 1-4. A single data set (of two samples) is used to produce all four graphs on the left of Figures 1-4, but by different methods. A similar data set is used to create all four graphs on the right. Each of the two data sets consists of two random samples of size five from a standard normal distribution. The histograms in the graphs represent the distribution used for each of the four methods for obtaining two-sided significance levels. In Figure 1, the distribution is the known Student's t distribution. In Figures 2-4, the t distributions were created using bootstrap or permutation resampling. Each was produced from 10,000 resamples from each data set. If another 10,000 resamples were taken from the same data sets, the distributions would change. This change, however, may be considered negligible due to the finiteness of the number of permutation and bootstrap resamples when samples of size 5 are used (see Table 1), and because the number of resamples is substantial. Although the two data sets used in this example are random samples and do not result in t-statistics of 2.306, I assume that the resampled distributions of Figures 2-4 are typical of data sets which do result in a t-statistics of 2.306.

The objective of each of the bootstrap and permutation methods is to produce sampling distributions which are close to the known t distribution. In this example, this closeness in distribution to the known t distribution is determined by finding the proportion of the distribution outside -2.306 and 2.306. The correct proportion is known to be 0.05, based on the known Student's t distribution with 8 degrees of freedom. The difference between the achieved proportion from each of the resampling based

distributions and the known proportion is a measure of the correctness of the method being used.

Student's *t* distribution with 8 degrees of freedom

(a) (b)

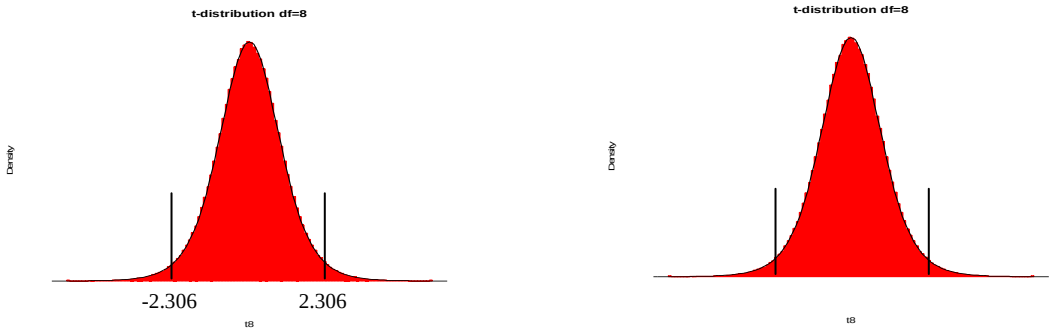

Figure 1. First in a Series of Four Figures Depicting the Variation in Area Outside 2.306 When 4 Different Tests Are Used, t-distribution. The graph on the left represents a distribution from two samples of size five. The graph on the right represents another two samples. The same samples are used in Figures 1-4. The two-sided p-value for (a) is 0.025 + 0.025 = 0.05. The two-sided p-value for (b) is 0.025 + 0.025 = 0.05.

Permutation *t* Distribution two samples of size 5 (twice)

(a) (b)

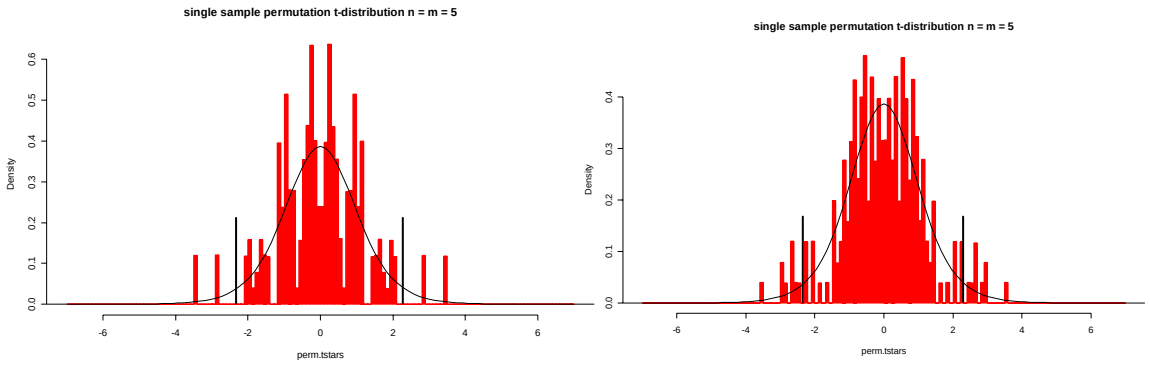

Figure 2. Second in a Series of Four Figures Depicting the Variation in Area Outside 2.306 When 4 Different Tests Are Used, Permutation Distribution. The left graph represents a distribution from two samples of size five. The right graph represents another two samples. The same samples are used in Figures 1-4. The two-sided ASL for (a) is 0.023 + 0.022 = 0.045. The two-sided ASL for (b) is 0.027 + 0.027 = 0.054.

Bootstrap Within *t* distribution two samples of size 5 (twice)

(a) (b)

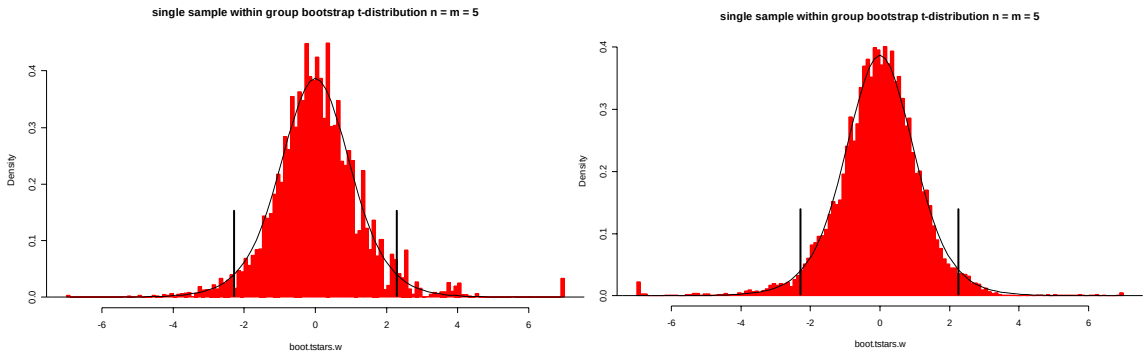

Figure 3. Third in a Series of Four Figures Depicting the Variation in Area Outside 2.306 When 4 Different Tests Are Used, Bootstrap Within t Distribution. The graph on the left represents a distribution from two samples of size five. The graph on the right represents another two samples. The same samples are used in Figures 1-4. The two-sided ASL for (a) is 0.021 + 0.028 = 0.049. The two-sided ASL for (b) is 0.031 + 0.027 = 0.058.

Bootstrap Across *t* distribution two samples of size 5 (twice)

(a) (b)

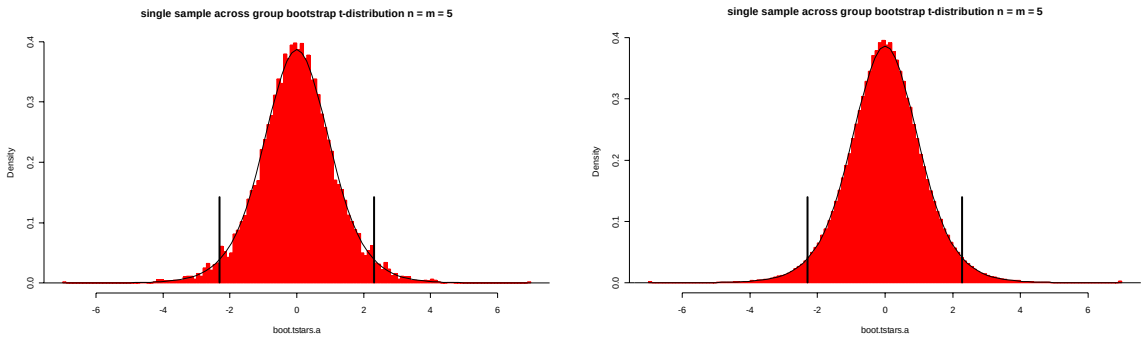

Figure 4. Fourth in a Series of Four Figures Depicting the Variation in Area Outside 2.306 When 4 Different Tests Are Used, Bootstrap Across t Distribution. The graph on the left represents a distribution from two samples of size five. The graph on the right represents another two samples. The same samples are used in Figures 1-4. The two-sided ASL for (a) is 0.024 + 0.025 = 0.049. The two-sided ASL for (b) is 0.024 + 0.024 = 0.048.

The graphs in Figures 1-4 illustrate the variation in ASL that occurs when a resampling method is used. Based on only two data sets, it appears that the bootstrap across *t* distribution is the most correct of the 3 resampling methods. The histograms in Figure 4 are closest to the correct distribution and the ASLs are closest to 0.05.

The results of a simulation study found in Appendix D show more rigorously the correctness of each of the methods for samples of size 5. First, 20,000 two-sample data sets were simulated from a normal distribution with mean zero and variance one. For each data set, the resampling distribution was produced using $B = 999$ resamples for each resampling method. Cutoff values for determining ASLs were chosen as seen in Table D-1, based on the known Student's *t* distribution with 8 degrees of freedom. The correctness is measured as the mean square error of the ASLs from each known correct proportion (0.001, 0.01, 0.05, and 0.10). The most correct methods are those which yield the smallest mean square errors.

In Figure D-4 it is seen that the mean square error for the bootstrap across method is lowest, followed by the permutation method and then the bootstrap within method. This is the result anticipated based on the figures of Appendix A.

2.3. Two Sample Test Discussion and Recommendations

In terms of accuracy, for samples of size 5, the Welch's *t*-test generally performs much better than the other methods. Except under the most extreme underlying distributions, Welch's *t*-test preserves the nominal error rate. The bootstrap within test preserves the error rate but is usually far too conservative. The *t*-test, permutation test, and bootstrap across test are anti-conservative for even mild differences in shape or variance among underlying distributions. When the underlying distributions differ in both shape and variance, none of the examined tests is accurate for samples of size 5.

In terms of correctness, the ASL mean square error for the bootstrap across test is the lowest, followed by the permutation test. The ASL mean square error for the bootstrap within test is much higher than the other two.

Based on these simulation experiments, I recommend Welch's *t*-test for a single test comparing the means of two samples of small sample size from unknown underlying distributions. If the underlying distributions are known to be at least close to normally distributed with equal variance, the common *t*-test is the preferred test because of the gain in power.

There is one other aspect of resampling based two-sample testing procedures that makes them undesirable, particularly when multiple comparison correction is to be done. The formula for an individual ASL is, again,

$$\text{ASL} = (1 + \#(|\boldsymbol{t}^*| \geq |t|)) / (1 + B)$$

which has a minimum that is based on the size of *B*. That is, if *B* = 999, the smallest possible ASL is 1/(1 + 999) = 0.001. ASLs much smaller than this are required when hundreds or thousands of tests need be adjusted for simultaneously. The size of *B* is limited by the number of possible resampling permutations, which can be seen in Table 1.

Table 1 shows the number of unique resampling statistics that can be obtained from the three resampling methods for per group sample sizes of 2 to 10. If *n* is the sample size in each group, then the number of permutations as defined in Section 2.2.1 is given by $\binom{2n}{n}$. The numbers of unique bootstrap within and bootstrap across resampling statistics as defined in Section 2.2.2 were derived to be $\binom{2n-1}{n}\binom{2n-1}{n}$ and $\binom{3n-1}{n}\binom{3n-1}{n}$, respectively.

Table 1. Number of Possible Unique Resampling Statistics from Three Resampling Methods for the Two Sample Test Scenario

Number of Obs. in Each Group	Permutation	Bootstrap Within	Bootstrap Across
2	6	9	100
3	20	100	3,136
4	70	1,225	108,900
5	252	15,876	4,008,004
6	924	213,444	153,165,376
7	3,432	2,944,656	6,009,350,400
8	12,870	41,409,225	240,407,818,596
9	48,620	590,976,100	9,762,812,702,500
10	184,756	8,533,694,884	401,201,300,600,100

3. MULTIPLE TESTING ADJUSTMENT

3.1. Historical Perspective

Recognition for the need of appropriate adjustments in multiple testing has become widespread since the dissemination of the idea by Fisher (1935). A host of procedures have been developed to provide such adjustments for the various scenarios under which multiple testing occurs. Detailed treatment of most multiple comparison procedures can be found in Hochberg and Tamhane (1987) and Hsu (1996). Most multiple testing adjustment methods have centered around multiple testing in terms of the one-way layout scenario. However, with the increase in availability of data from increased computing power and novel techniques, multiple testing in more general situations with higher numbers of comparisons is occurring with greater frequency. When larger numbers of tests occur, more attention needs to be paid to issues such as bias, variance estimation, correlation, and distributional assumptions. That is, the effect of an incorrect assumption on the overall error rate for 20 – 30 tests may be only moderate while for 1000 tests the same incorrect assumption may affect the overall error rate dramatically.

To acquaint the reader with the development of multiple comparisons and testing in general, I offer a historical perspective and summary of the most commonly used procedures.

At the beginning of the nineteenth century, Legendre first proposed the minimization of the sum of squared errors as a method of estimating parameters (Legendre, 1805). This method, the method of least squares, was formalized shortly thereafter with the work of both Legendre and Gauss (Gauss, 1809). By around 1820, the concept of standard error and standard deviation as measures of variation emerged, largely due to the work of Laplace and Gauss (Cochran, 1976). Propelled by a desire to apply these and other mathematical tools to the social sciences, astronomy, agriculture, and later in studies of heredity, scientists throughout the 1800s made improvements and extensions to the method of least squares (Stigler, 1986). “It was this period which saw

the emergence and the beginning of the extensive use of the normal distribution both as a model and for approximating large-sample distributions of statistics and also the germination of seminal concepts like relative efficiency of estimators" (Chatterjee, 2003).

In the late nineteenth century, Sir Francis Galton, who is widely known for his work on the correlation coefficient, was asked by Charles Darwin for statistical advice concerning his height data for comparison of crossed and self-fertilized corn. Darwin had 15 replications for each group. Galton was aware that "averages of independent samples from a normal distribution are themselves normally distributed," but did not feel comfortable estimating the standard deviation nor the "law of distribution followed by the individual differences in height" from only 15 observations (Cochran, 1976). In 1908, William Sealy Gossett, under the pen name of 'Student', published "The probable error of a mean" (Student, 1908) in which the sampling distribution of $t = \frac{\sqrt{n}(\overline{X} - \mu)}{s}$, or 'Student's *t*', was derived, paving the way for the legitimate comparison of means when only small samples are available. As was the case with many previous fundamental statistical discoveries of the eighteenth and nineteenth centuries, the value of this finding was not quickly disseminated. In 1922, Gossett wrote to R. A. Fisher, with whom he corresponded frequently, "I am sending you a copy of Student's Tables as you are the only man that's ever likely to use them!" (Cochran, 1976) It was Fisher who opened the door to comparative experimentation of multiple levels and factors with his work at Rothamsted Experimental Station and ensuing publication *The Design of Experiments* (Fisher, 1935). In this cornerstone work, Fisher explained the now routine techniques of blocking, randomization, factorial design, and the analysis of variance. In this same volume, Fisher also proposed two of the earliest methods for making appropriate adjustment for multiplicity of tests, which are often inherent with analysis of variance.

The suggestion of Fisher (1935) was to first test the effect of a factor using an overall *F*-test. If the *F*-test indicates significant differences among the means, it is

followed by individual *t*-tests, with the mean square error from the analysis of variance as the estimate of variance, comparing each mean to each other mean. This came to be know as the "protected" LSD (least significant difference), the protection coming, in his view, from the rejection of the *F*-test. If the *F*-test for equality of means is not deemed significant, Fisher proposed the conservative Bonferroni adjustment to the individual *t*-tests, also using the estimated variance based on the pooled samples.

Shortly thereafter, at the suggestion of Gossett (see Pearson, 1939) Newman (1939) proposed a method for comparing multiple means based on the studentized range. This method was modified by Keuls (1952), and came to be known as the Newman-Keuls (sometimes Student-Newman-Keuls) multiple range test (see Harter, 1980).

Increased interest in multiple comparison procedures following World War II was evidenced by the work of John W. Tukey, David B. Duncan, Henry Scheffe, and Charles W. Dunnett. Tukey (1952, 1953) presented another method based on the studentized range. The equal sample size version is now known as Tukey's method or Tukey's HSD (honest significant difference) method. For unequal sample sizes it is known as the Tukey-Kramer procedure (Kramer, 1956). A publication of Tukey (1953), a manuscript of mimeographed notes that has been widely circulated privately and widely used, was not produced until 1994 (Braun, 1994). Tukey (1953) proved that his equal sample size method maintains the overall experimentwise error rate. That is, the probability of a Type I error for all tests *jointly*, is α. Unable to prove this for differing sample sizes, Tukey conjectured that the Tukey-Kramer procedure also preserved the overall experimentwise error rate (in the conservative direction). It was not until 1984 that Tukey's conjecture was proven correct by Hayter (1984) (The Tukey-Kramer method was shown to be conservative based on simulation studies by Dunnett [1980]).

Duncan (1947, 1951, 1952, 1955) developed a multiple range test which by the late 1970s was the most commonly used multiple comparison procedure, according to a *Science Citation Index* survey by Harter (1980). Duncan's multiple range test has since dropped in popularity based on the finding that it does not preserve the overall experimentwise error rate (see, for instance, Hsu [1996], pp. 129-130).

A method for jointly comparing all contrasts of means was developed by Scheffe (1953). Because this method is less powerful than Tukey's method when only pair-wise comparison of means is desired, this method has come to be recommended only when the primary comparisons of interest are contrasts other than pair-wise comparisons.

Dunnett (1955) proposed a multiple comparison procedure similar to the Tukey methods, but for situations when only comparison with a single control is desired.

Development during the 1960s and 1970s in the areas of probability inequalities (i.e., Sidak's (1967) inequality), unbalanced ANOVA methods, conditional confidence levels, empirical Bayes, and confidence bands in regression are outlined and discussed in Miller (1981). The empirical Bayes methods were set forth in a series of papers: Waller and Duncan (1969, 1974), Waller and Kemp (1975), Duncan (1975), and Dixon and Duncan (1975). The Ryan-Einot-Gabriel-Welsch (REGW) multiple range method was also developed during this period. Ryan (1960) proposed a conservative adjustment to α that was improved upon by Welsch (1972). This adjustment can be used in conjunction with the adjustment proposed by Einot and Gabriel (1975). Lehmann and Shaffer (1979) have shown this method "approximately maximizes the power subject to the [experimentwise error] control requirement" (for details, see Tamhane, 1995, pp. 607-610 or Hsu, 1996, pp. 128-129).

Multiple comparison procedures for finding the "best" treatment among several were developed by Hsu (1981, 1982) and Edwards and Hsu (1983) and improved in Hsu (1984). Comprehensive treatment of these developments can be found in Hsu (1996).

Because most of the above comparison procedures were developed for the one-way normal layout with equal variance model, distribution free and robust procedures for coping with nonnormal and heteroscedastic data were developed almost in parallel. It is well-known that *t*- and *F*- statistics are robust to non-severe departures from normality in the two-sample and one-way layout scenarios. However, as Hochberg and Tamhane (1987) note, "the problem of robustness becomes more serious in the case of multiple inferences." Steel (1959a, 1959b) was the first to develop nonparametric multiple comparison procedures. Based on signs and ranks, these are applied to

comparison of means with a control when the assumption of normality is not met to permit use of Dunnett's method. Similar all-pairwise nonparametric procedures based on signs and ranks were first developed and discussed in Steel (1960), Dwass (1960), Steel (1961), Nemenyi (1961), and Nemenyi (1963). The Kruskal-Wallis-type multiple comparison tests and similar tests (Friedman-type) for the two-way classification problem were also put forth in Nemenyi (1963). For a detailed description of the early development of nonparametric multiple testing procedures see Miller (1981).

More recently, Westfall and Young (1993) have applied the resampling ideas (such as the bootstrap first proposed by Efron [1979]) to the multiple comparison problem. Although computationally intensive, these methods are distribution-free, and can incorporate important correlation structure among the means.

Another perspective has found recent popularity in biological applications, particularly "gene finding." Instead of preserving the experimentwise error rate for multiple comparisons, Benjamini and Hochberg (1995) proposed a different error rate, called the *false discovery rate* (FDR). I refer to the summary given in Tamhane (1995): "Let T and F be the (random) number of true and false null hypotheses rejected. Then the FDR is defined as FDR = E [T / (T + F)], where 0/0 is defined as 0. ... When all null hypotheses are true, the FDR equals [the experimentwise error rate]. … Since control of the FDR is less stringent than control of the [experimentwise error rate], it generally results in more rejections."

3.2. Present Microarray Multiple Testing Problem

The following table (adapted to the subject of microarray data) is found in Benjamini and Hochberg's (1995) false discovery rate article.

Table 2. Benjamini and Hochberg's (1995) Table Used to Define the False Discovery Rate

	Declared Not Different	Declared Different	Total
Genes in the treatment and control groups *are not* differentially expressed	U	V	m_0
Genes in the treatment and control groups *are* differentially expressed	T	S	$m - m_0$
	m – R	R	m

Note: The table is adapted to the subject of microarray data.

In Table 2, the null hypotheses for the microarray scenario are that the expression levels for the treatment and control groups for each gene are equal. The number m is the total number of hypotheses tested (or total number of genes) and is assumed to be known in advance. Of the m hypotheses tested, m_0 are true. The variable R is the total number of genes declared significantly different. The random variables U, V, T, and S are unobservable.

Individual *P*-values (or test statistics) are calculated for each test followed by adjustments to account for multiplicity of tests. It is desirable that these adjustments minimize the number of genes that are falsely declared different (V) while maximizing the number of genes which are correctly declared different (S). To address this issue the researcher must know the comparative value of finding a gene to the price of a false positive. If a false positive is very expensive, methods which focus on minimizing V should be used. If the value of finding a gene is much higher than the cost of additional false positives, methods which focus on maximizing S should be employed. Further, adjustments for multiplicity should incorporate to some degree the correlation of expressions of genes within an individual. There are groups of genes which are expressed in tandem while some genes may be "turned off" when others are "turned on." Preferably, the method used to adjust for multiplicity of tests would incorporate an ability to account for this correlation.

Ge, Dudoit, and Speed (2003) outline the most common methods of control of false positive declarations:

- Per-comparison error rate (PCER), defined as

$$\text{PCER} = E(V)/m$$

- Per-family error rate (PFER), defined as

$$\text{PFER} = E(V)$$

- Family-wise error rate (FWER), defined as

$$\text{FWER} = \Pr(V > 0)$$

- False discovery rate (FDR) (Benjamini and Hochberg, 1995), defined as

$$\text{FDR} = E(\frac{V}{R}1_{\{R>0\}}) = E(\frac{V}{R} \mid R > 0)\Pr(R > 0)$$

- Positive false discovery rate (pFDR) (Storey, 2001, 2002), defined as

$$\text{pFDR} = E(\frac{V}{R} \mid R > 0)$$

These rates are generally considered to be computed under the complete null hypothesis of all genes being equally expressed between treatment and control groups. Ge, Dudoit, and Speed (2003) also show the following to be the general ordering of these rates:

$$\text{PCER} \leq \text{FDR} \leq \text{pFDR} \leq \text{FWER} \leq \text{PFER}$$

Dudoit, Shaffer, and Boldrick (2003) state, "Thus, for a fixed criterion α for controlling the Type I error rates, the order reverses for the number of rejections R: procedures that control the PFER are generally more conservative, that is, lead to fewer rejections, than those that control either the FWER or the PCER, and procedures that control the FWER are more conservative than those that control the PCER."

A review and discussion of procedures which control the FWER can by found in Ge, Dudoit, and Speed (2003) and Dudoit, Shaffer, and Boldrick (2003). They describe in detail the following procedures for obtaining adjusted *P*-values:

1. Bonferroni single-step adjusted *P*-values
2. Sidak single-step adjusted *P*-values

3. Sidak step-down adjusted *P*-values
4. Holm step-down adjusted *P*-values
5. Single-step minP adjusted *P*-values
6. Step-down minP adjusted *P*-values
7. Single-step maxT adjusted *P*-values
8. Step-down maxT adjusted *P*-values

When single-step methods are used, the adjusted *P*-values may be used to reflect the amount of evidence of expression difference. That is, lower adjusted *P*-values indicate more evidence of a difference. Step-down method adjusted *P*-values can only be used to indicate a significant difference, but do not allow one to quantify the amount of evidence of a difference, except that the *P*-value is below the specified overall level that is to be preserved.

Adjusted *P*-values which control the FDR (Benjamini and Hochberg, 1995) are discussed in Ge, Dudoit, and Speed, (2003). Benjamini and Tekutieli (2001) proposed a modification to the Benjamini and Hochberg (1995) *P*-value adjustment which controls the FDR while allowing for arbitrary dependence. This method is considerably more conservative than the original method. Details of adjusted *P*-values (or *q*-values) based on the proposal of Storey (2001, 2002) are also found in Ge, Dudoit, and Speed (2003).

3.3. Details of Adjustment Methods Compared

3.3.1. Bonferroni Adjustment

The Bonferroni adjustment is applied to all *m* unadjusted *P*-values (p_j) as

$$\tilde{p}_j = \min(mp_j, 1)$$

3.3.2. Benjamini and Hochberg's (1995) False Discovery Rate Control Procedure

Adjusted *P*-values are found as

$$\tilde{p}_{r_i} = \min_{k=i,\ldots,m} \{\min(\frac{m}{k} p_{r_k}, 1)\},$$

where $p_{r_1} \le p_{r_2} \le \cdots \le p_{r_m}$ are the observed ordered unadjusted *P*-values. The procedure is defined in Benjamini and Hochberg (1995). The corresponding adjusted *P*-value definition given here is found in Dudoit, Shaffer, and Boldrick, (2003).

3.3.3. Westfall and Young's (1993) Single Step maxT Resampling Based Procedure

The test statistics T_j are used to give adjusted *P*-values as

$$\tilde{p}_j = \Pr\left(\max_{1 \le l \le m} |T_l| \ge |t_j| \mid H_0^C\right)$$

where H_0^C is the complete null hypothesis.

The Bonferroni and false discovery rate adjustment methods are two step methods. In the first step the unadjusted *P*-values are calculated. In the second step the error rate adjustment is made. In the case of the maxT resampling procedure, both steps are incorporated into a single process in the following way (Westfall and Young, 1993):

1. A counting variable is initialized for each of the compared samples: $COUNT_i = 0$, i = 1,…, *m*.

2. A *t*-statistic t_i is calculated for each of the compared samples of the original data.

3. A new data set is generated from the estimated complete null distribution via appropriate resampling (permutation or bootstrap) of complete columns (individuals) of expression data. This preserves the internal correlation among genes.

4. A new *t*-statistic *t** is calculated for each of the compared samples of the resampled data set and the maximum absolute *t** is found.

5. The absolute value of each t_i from (2.) is compared to the maximum absolute *t** from (4.). If $|t_i| \ge t^*$, then $COUNT_i \leftarrow COUNT_i + 1$.

6. Steps 2-4 are repeated B times. The value of $\tilde{p}_i$ is estimated to be $\tilde{p}_i^{(B)} = \frac{1 + COUNT_i}{(1 + B)}$. Because adjusted *P*-values of zero are unrealistically small, one is added to the numerator to avoid adjusted *P*-values of zero. The minimum adjusted *P*-value should be limited by the number of resampling replicates. This lower limit is 1/*B*. One is added to the denominator to compensate for addition of one to the numerator.

3.3.4. Efron's (2004) Empirical Null Distribution Local False Discovery Rate Method

The primary novelty of Efron's (2004) method for large-scale simultaneous hypothesis testing involves the estimation of the null distribution from the abundance of test statistics or *P*-values available. Because it can be assumed that only a small fraction of the hypothesis tests are significant, the remaining majority of *P*-values can be used to create an empirical distribution of *P*-values against which all *P*-values may be compared. Following the notation of Efron (2004), suppose we are testing *N* null hypotheses,

$$H_1, H_2, \ldots, H_N;$$

with corresponding test statistics (not necessarily independent),

$$Y_1, Y_2, \ldots, Y_N;$$

with *P*-values $P_1, P_2, \ldots, P_N$. For convenience Efron uses *z*-values rather than Y_i's or P_i's,

$$z_i = \Phi^{-1}(P_i), \qquad i = 1, 2, \ldots, N,$$

where Φ is the standard normal cumulative distribution function. If the complete null hypothesis $H_1, H_2, \ldots, H_N$ is true, then $P_1, P_2, \ldots, P_N$ are $U(0,1)$ and the distribution of z_i is the standard normal distribution,

$$z_i \mid H_i \sim N(0,1).$$

Efron calls this the theoretical null hypothesis. However, as Efron discusses, there are compelling examples of how important unobserved covariates can "dilate the null hypothesis density," producing the need for an empirical null hypothesis density. He shows how this empirical density can be estimated using the central peak of the histogram of the observed z_i's.

The local false discovery rate (*lfdr*) as defined by Efron (2004), focuses "on densities rather than tail areas." We first suppose that each of the *N* *z*-values can be classified as "uninteresting" or "interesting." Uninteresting genes are generated according to the null hypothesis, while interesting genes are not. The prior probability of uninteresting genes is p_0 while for interesting genes it is $p_1 = 1 - p_0$. The density of z_i is assumed to be $f_0(z)$ if the gene is uninteresting, and $f_1(z)$ if the gene is interesting. That is,

$$p_0 = \Pr\{\text{Uninteresting}\}, \quad f_0(z) \text{ density if Uninteresting (Null)},$$

$$p_1 = \Pr\{\text{Interesting}\}, \quad f_1(z) \text{ density if Interesting (Nonnull)},$$

The curve from the natural spline fit to the histogram of all observed *z*-values can be used to estimate the mixture density

$$f(z) = p_0 f_0(z) + p_1 f_1(z).$$

Using Bayes theorem, the a posteriori probability of being an uninteresting gene given *z* is

$$\Pr\{\text{Uninteresting} \mid z\} = p_0 f_0(z) / f(z).$$

Efron defines the lfdr as

$$\text{lfdr}(z) \equiv f_0(z) / f(z),$$

"ignoring the factor p_0, so lfdr(z) is an upper bound on Pr{Uninteresting | *z*}. In fact, p_0 can be roughly estimated, but [he assumes] that p_0 is near 1, say $p_0 \geq .90$ so fdr(*z*) is not a flagrant overestimator."

To illustrate Efron's method, I have simulated *z*-values for 10,000 tests, 95% of which the underlying population values reflect a true null hypothesis and 5% reflecting a false null hypothesis. In Efron's terminology, 9,500 differences are *uninteresting* while 500 of the differences are *interesting*. Specifically,

$$z_1, \ldots, z_{9,500} \sim N(0,1) \quad \text{and} \quad z_{9,501}, \ldots, z_{10,000} \sim N(-5,1).$$

The histogram (see Figure 5) readily shows a central peak from which an estimate of the empirical null hypothesis can be obtained, assuming the shape of the null distribution of *z*-values is normal. The focus is given to the natural spline fits to the histogram counts for z-values within 1.5 units of the maximum natural spline fit. It is expected that the fits near this maximum reflect the true null hypothesis and follow a normal distribution. Thus, these fits can be used to produce an estimate of the mean and standard deviation for the empirical null hypothesis. To obtain an estimate of mean and standard deviation, the logarithms of the fits near the maximum are further fit with quadratic regression (the log of the normal density is quadratic), giving,

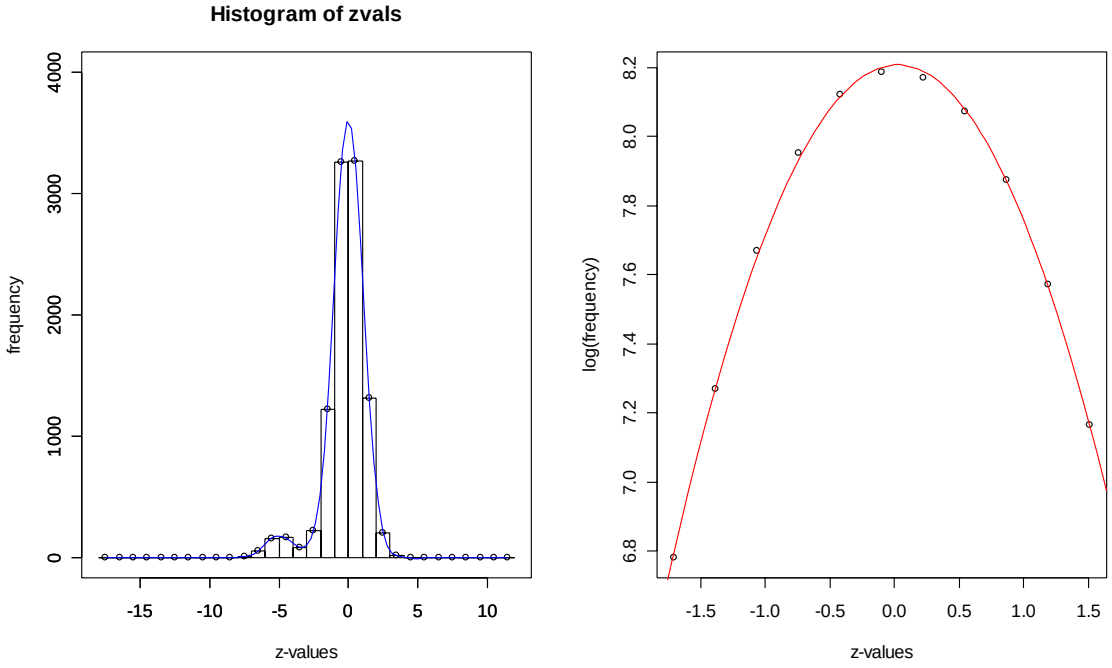

Figure 5. Empirical Distribution Estimation. Histogram (left) of 10,000 z-values (95% standard normal, 5% N(-5,1)) with natural spline fit overlaid. Quadratic regression fit (Right) to the log of the natural spline fit to the histogram counts for the z-values of the central peak.

say, $a_0 + a_1 x + a_2 x^2$. The mean is then estimated by

$$\delta_0 = -\frac{a_1}{2a_2}$$

and the standard deviation by

$$\sigma_0 = \left(-2a_2\right)^{-\frac{1}{2}}.$$

The resulting empirical null hypothesis density is then

$$f_0(z) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(z-\delta_0)^2}{2\sigma_0^2}}$$

or $N(\delta_0, \sigma_0^2)$. The local false discovery rate is then obtained as

$$\text{lfdr}(z) \equiv f_0(z)/f(z),$$

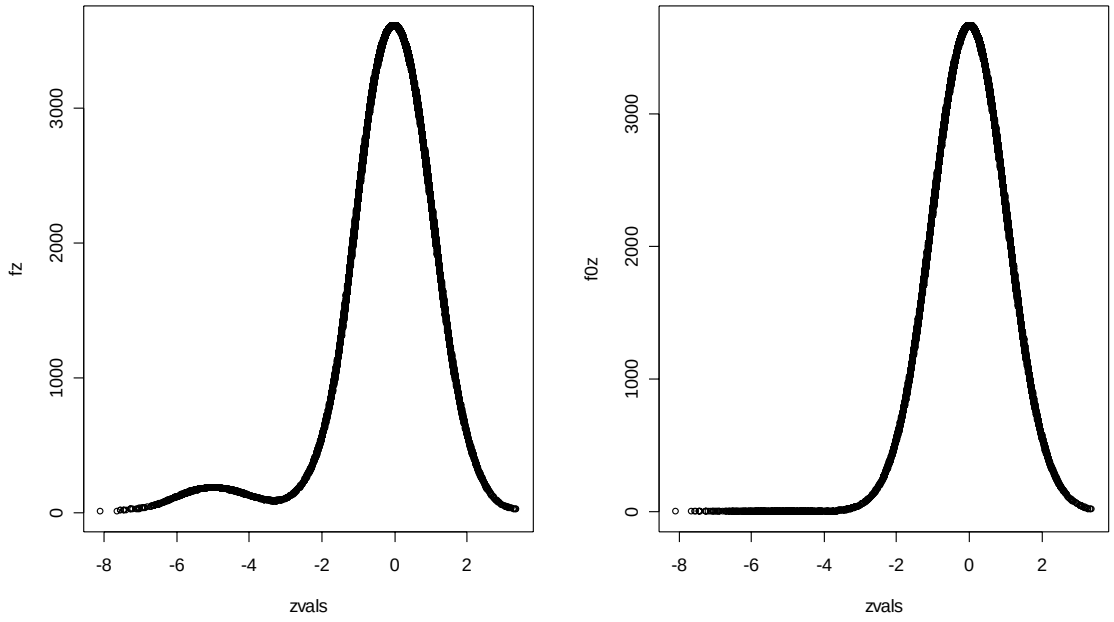

Figure 6. Plot of $f(z)$ and $f_0(z)$ for All of the 10,000 Simulated z-values. Points on the left graph are based on the natural spline fit to the histogram of all z-values. Points on the right graph come from the empirical null hypothesis density, which for these data is N(0.02060998, 1.030219^2).

where, again, $f(z)$ is the natural spline fit to the histogram of all *z*-values (see Figure 6). Cases where lfdr is less than a specified threshold are then reported as "interesting."

4. MICROARRAY DATA DETAILS

4.1. Data Normalization in Microarrays

A general description of how intensity data is obtained from microarrays was outlined in Section 1. Below is a detailed description of the data refinement process that occurred in the data provided by the lab of Dr. Rajesh Miranda to prepare it for the multiple comparison tests. Six microarrays were used in the experiment: 3 treatment microarrays and 3 control microarrays. Each microarray is further divided into 48 blocks which are printed by 48 different print-tips.

The 6 microarrays used were two-color cDNA microarrays. RNA from two sources was introduced into each of the microarrays. The first source was the RNA of primary interest, corresponding to the treatment or control. This RNA is labeled with Cy5 dye, called ‘red’ by convention. The second source of RNA is a universal reference, expected to be the same for all 6 microarrays except for a small amount of random variation. This RNA is labeled with Cy3 or ‘green’ dye. The red and green RNA compete for each of the approximately 23,000 probes on each array. Laser scanning then provides red and green intensities for each probe. These intensities are measured by an instrument which distinguishes about 65,000 intensity shades for each color (red or green).

The goal is ultimately to compare the red intensities of the treatment microarrays to the red intensities of the control microarrays. Instead of comparing the intensities directly, intensities relative to the green intensities are used. Before these relative intensities are compared, however, adjustment need be made to account for variation that arises from the technology used. Smyth and Speed (2003) state, “Imbalances between the red and green dyes may arise from differences between the labeling efficiencies or scanning properties of the two fluors complicated perhaps by the use of different scanner settings. If the imbalance is more complicated than a simple scaling of one channel relative to the other, as it usually will be, then the dye bias is a function of intensity and normalization will need to be intensity dependent. The dye bias will also generally vary

with spatial position on the slide. Positions may differ because of differences between the print tips on the array printer…" They go on to say that "differences between arrays may arise from differences in print quality, from differences in ambient conditions when the plates were processed or simply from changes in the scanner settings." The normalization spoken of as applied to the Miranda data follows.

Let R and G be the red and green intensities, respectively, for each gene in a given microarray. Dudoit et. al (2002) suggest using logged intensities rather than absolute intensities for the following reasons: "(i) the variation of logged intensities and ratios of intensities is less dependent on absolute magnitude; (ii) normalization is usually additive for logged intensities; (iii) taking logs evens out highly skewed distributions; and (iv) taking logs gives a more realistic sense of variation." They also note that "logarithms base 2 are used instead of natural or decimal logarithms as intensities are typically integers between 0 and $2^{16} - 1$." To incorporate information about the overall transcription abundance as well as relative intensity, two measures are used in the normalization. The measure of relative intensity is

$$M = \log_2(R/G) = \log_2(R) - \log_2(G)$$

The measure of overall brightness for each spot is

$$A = \log_2\sqrt{RG} = \frac{\log_2(R) + \log_2(G)}{2}$$

M and A are mnemonics for *m*inus and *a*dd, respectively. A scatterplot of M versus A (known as an MA-plot) is a good method for visualizing the relationship between dye-bias and intensity. It is recommended that this be done separately for each print-tip block within each microarray (Smyth and Speed, 2003).

A print-tip loess normalization as proposed by Yang *et al.* (2001) is achieved by subtracting the loess curve value from each corresponding M-value, or

$$N = M - loess_i(A),$$

where $loess_i(A)$ is the loess value at position A of the loess curve associated with the *i*th print-tip group. The loess curve is estimated via "re-descending M estimation with Tukey's biweight function (family="symmetric")" (Smyth and Speed, 2003 and

Cleveland *et al.*, 1992). Local linear regression parameters are estimated from the nearest 40% (span = 0.4) of the points to each A-value location. The resulting normalized *N*-values are the values used for statistical comparison of gene expression.

4.2. Miranda Data Summary

Normalized gene expression values were obtained for 22,276 genes on 6 individuals using the methods described above. These expression values were then used to provide estimates of the shape and spread of the underlying distributions which they represent. Estimates of shape and spread are then incorporated into the structure of the simulation study described in Section 5.

4.2.1. Distribution Shape

Assuming the vast majority of the genes are not differentially expressed, the gene expression standardized residuals are obtained as

$$r_{ij} = \frac{y_{ij} - \bar{y}_i}{sd(y_i)}, \quad i = 1, 2, \ldots, 22{,}276; \quad j = 1, 2, \ldots, 6$$

where $sd(y_i) = \sqrt{\sum_{j=1}^{6} \frac{(y_{ij} - \bar{y}_i)^2}{5}}$. The resulting 6*22,276 = 133,656 standardized residuals were used to determine the general underlying distribution shape of the expression data (see Figure 7). As a basis for comparison, 22,276 samples of size 6 were randomly generated from a standard normal distribution. The standardized residuals from these samples are shown in Figure 8.

When data were simulated from other distributions (not shown), the pattern of the standardized residuals was far different from that shown figures 7 and 8. It is thus apparent that the distribution of the microarray log-intensities may be approximated with the normal distribution.

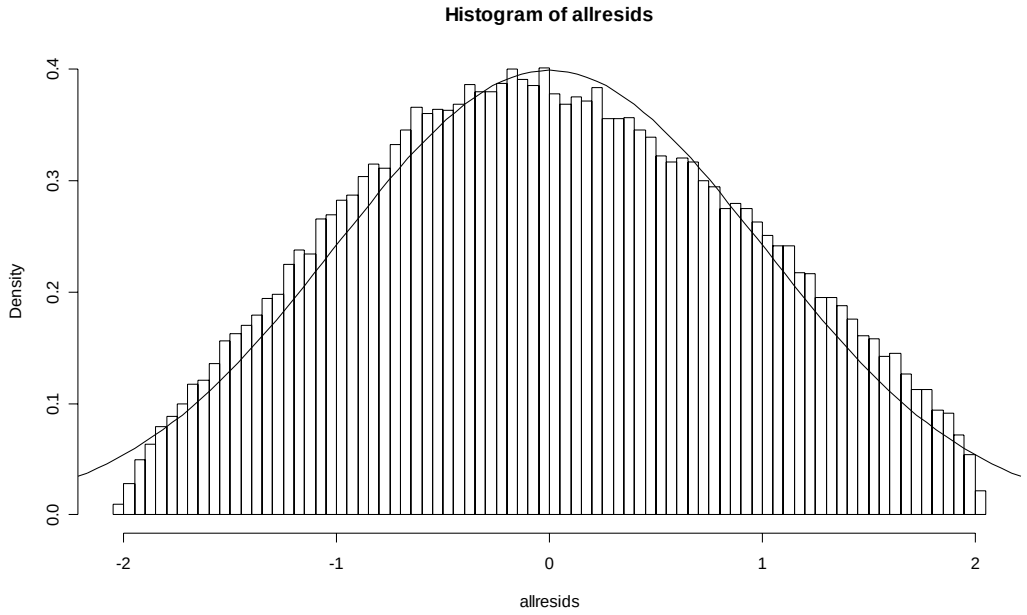

Figure 7. Histogram of 133,656 Standardized Residuals from Miranda Data. Standard normal density is overlaid.

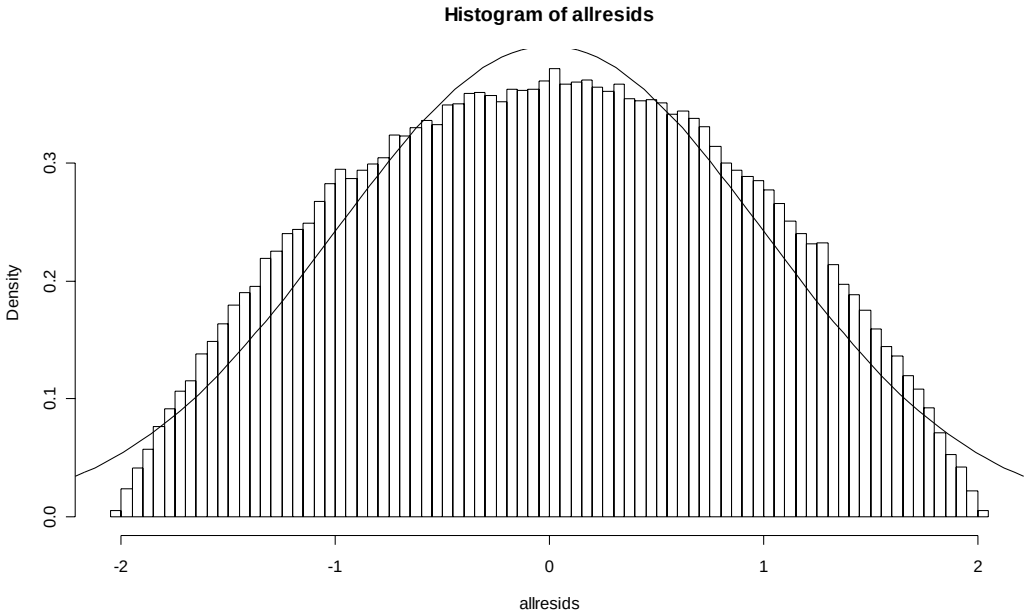

Figure 8. Histogram of 133,656 Standardized Residuals from Standard Normal Simulated Data. Standard normal density is overlaid.

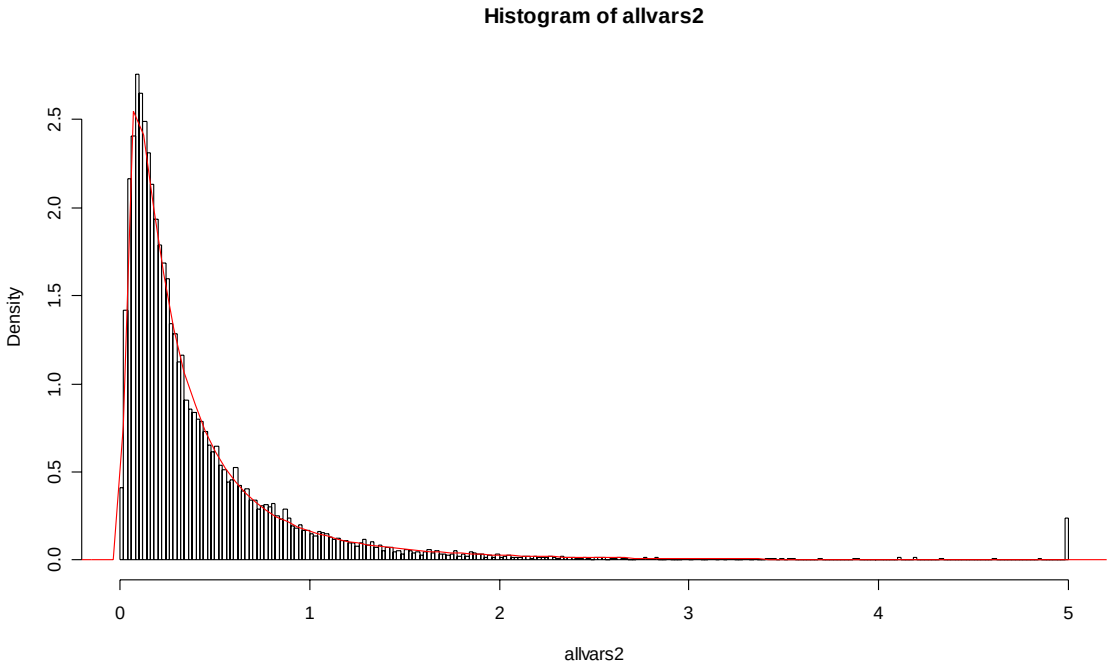

Figure 9. Histogram of Variances from Expression Data for All 22,276 Genes. Variances larger than 5 are grouped at 5. The density overlay is the lognormal distribution. The logarithm of this lognormal distribution has mean -1.365223 and standard deviation 1.057166, which were estimated from the log-transformed data.

4.2.2 Variance

The distribution of the sample variances for all 22,276 genes follows a lognormal distribution (see Figure 9). The lognormal distribution is used to produce variances for the simulation study described in Section 5. It is reasonable to assume that the variation in expression of differentially expressed genes may differ from the variation of the expression of those genes to which no treatment is imposed. That is, a change in mean expression may also result in a change in variation of expression. This aspect of expression is also incorporated into the simulation study.

4.2.3 Mean – Standard Deviation Relationship

The relationship between the mean and the variance can be seen in Figure 10.

The graph shows a slight positive linear trend and significant polynomial trends up to the thirteenth degree polynomial (see Appendix E for details).

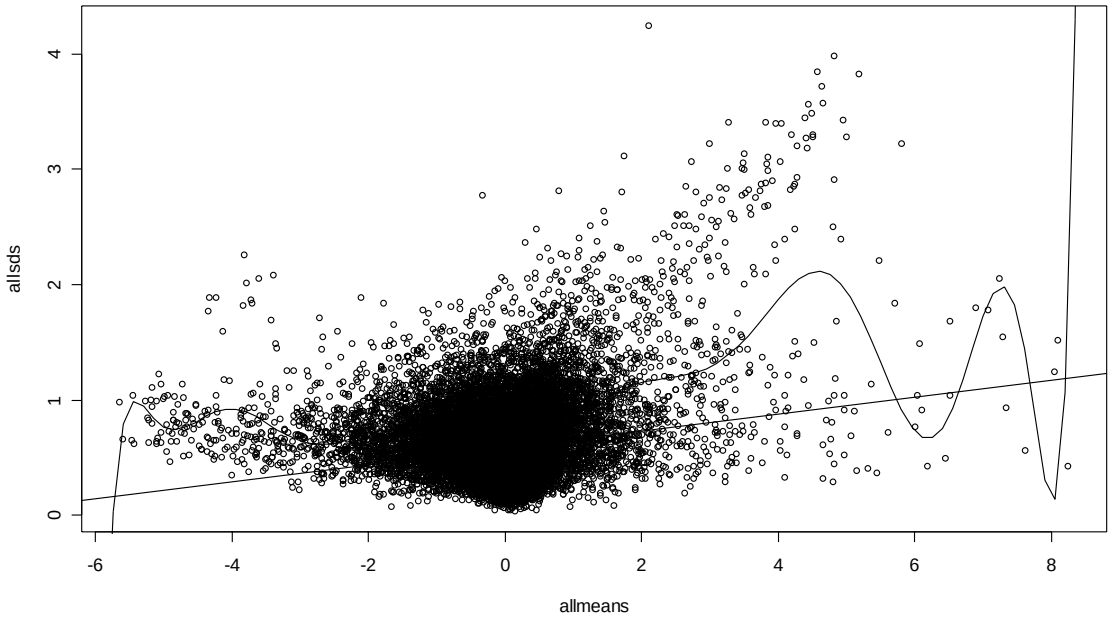

Figure 10. Scatter Plot of Standard Deviation vs. Mean for 22,276 Genes.

4.2.4. Miranda Data Results

None of the 22,276 genes of the Miranda study showed significant statistical evidence of differential expression, for any of the testing methods. Table 3 shows the lowest 10 raw *t*-test *P*-values and the corresponding Bonferroni and Benjamini and Hochberg false discovery rate adjusted *P*-values.

Table 3. Smallest 10 Raw, Bonferroni, and Benjamini and Hochberg False Discovery Rate Adjusted P-values for the Miranda Study

Gene ID	Raw P-value	Bonferroni P-value	FDR-adjusted P-value
11696	2.794248e-05	0.6224466	0.6224466
6188	7.466690e-05	1.0	0.8162653
5929	1.633017e-04	1.0	0.8162653
3271	2.492396e-04	1.0	0.8162653
680	2.555940e-04	1.0	0.8162653
8256	3.350913e-04	1.0	0.8162653
9892	4.023492e-04	1.0	0.8162653
14785	4.906419e-04	1.0	0.8162653
3021	5.008333e-04	1.0	0.8162653
3245	5.036014e-04	1.0	0.8162653

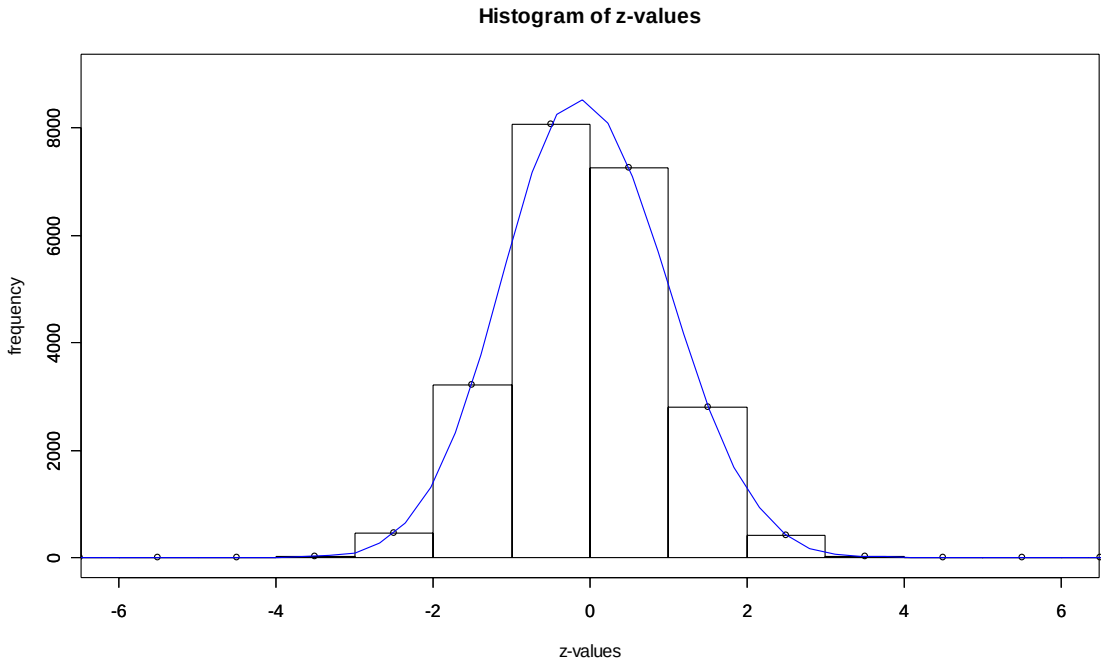

Figure 11. Histogram of 22,276 z-values from Miranda Expression Data. A natural spline fit to the histogram counts is overlaid.

Figure 11 shows a histogram of the 22,276 *z*-values following a standard normal transformation of the *P*-values. This histogram of *z*-values is used to estimate the empirical null hypothesis of Efron's method. The estimated mean and standard deviation of the empirical null hypothesis density are -0.06045423 and 1.031574, respectively. This indicates the empirical null hypothesis density differs very little from that of the theoretical null hypothesis. It does not appear that there are unobserved covariates causing dilation of the null hypothesis density. In this example, the advantages of Efron's method are expected to be minimal. The smallest 10 local false discovery rates are reported in Table 4.

The results of the three maxT procedures are not shown because the sample sizes are too small to give meaningful adjusted *P*-values.

Because it is not known which, if any, of the 22,276 genes are differentially expressed, it is impossible to assess the performance of the methods compared. The simulation study of Section 5 can be used to compare these methods directly as well as assess the power of the Miranda study.

Table 4. Smallest 10 Local False Discovery Rates for Miranda Data Using Efron's Method

Gene ID	Local False Discovery Rate
9471	0.2991060
11696	0.4190492
5822	0.4393112
3981	0.5428558
8342	0.5797304
19946	0.5902545
6188	0.6032793
3656	0.7765818
18619	0.8117799
717	0.8534339

5. SIMULATION STUDY

The purpose of the simulation study is to determine the effect of several factors on the power and the error rates associated with identifying differentially expressed genes. The factors are the number of genes in the study, the proportion of the genes that are differentially expressed, the magnitude of differential expression, the testing procedure used, the sample size, the level of the test, and the amount of correlation among genes. Information about the variation and underlying distribution of microarray expression from the Miranda data is incorporated into this study (see Sections 4.2.1 and 4.2.2).

5.1. Setup

The following are the parameters of the simulation study. Due to some run-time limitations a small number of combinations of these parameters are excluded.

5.1.1. Number of Genes

The numbers of genes that were simulated are 200, 2000, and 20,000. These reflect a variety of possible numbers of genes in microarray studies.

5.1.2. Proportion of Differentially Expressed Genes

In each simulation, either 1% or 10% of the genes are differentially expressed.

5.1.3. Magnitude of Differential Expression

Genes which are differentially expressed when a treatment is imposed will vary in the amount of differential expression. In these simulations, the mean for each of the differentially expressed treatment genes is randomly sampled from the distribution of Figure 12. The distribution of Figure 12 may reflect (although there is no way of knowing in advance) typical treatment effects. This distribution is based on comments from Dr. Rajesh Miranda that 2- to 4-fold differences in expression are biologically

realistic while larger-fold differences are not. Because the gene expression data are on a log base 2 scale, a two-fold ratio amounts to a one unit difference. Thus, the majority of the differences on the original expression scale are $2^{0.5}$ to 2^{2} or about 1.4 to 4.0. The distribution of Figure 12 was created by dividing samples from a Chi-square distribution with 10 degrees of freedom by 10 to achieve a mean of 1 (a two-fold difference).

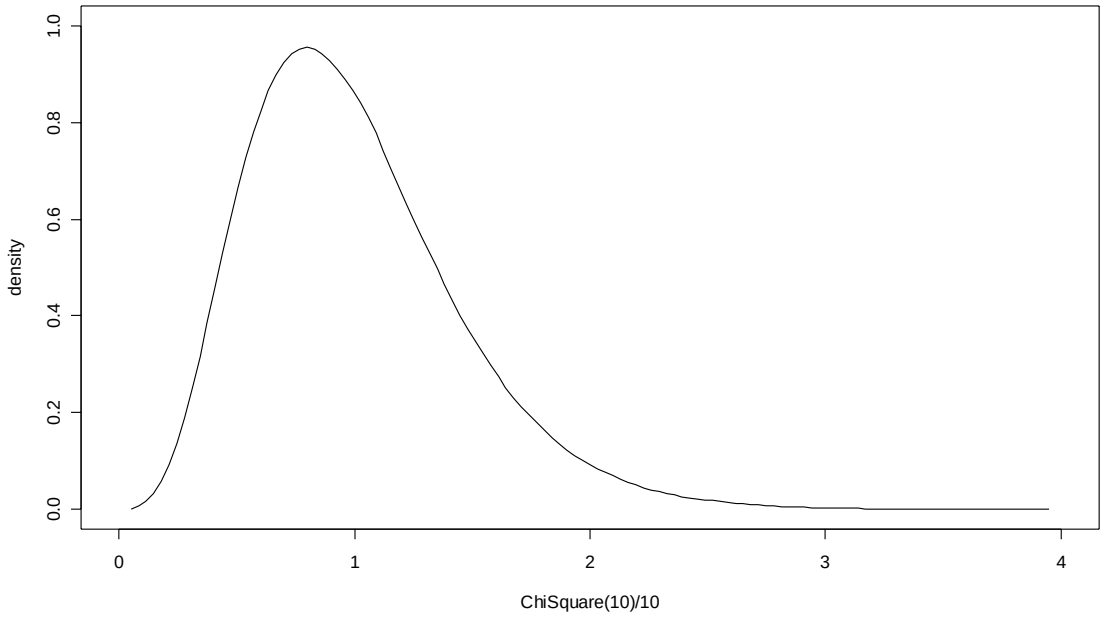

Figure 12. Distribution of Magnitudes of Difference for Simulated Differentially Expressed Genes.

5.1.4. Testing Methods

Only two individual tests are used to obtain unadjusted *P*-values: the traditional two sample *t*-test and Welch's *t*-test. The reasons for excluding the individual permutation, bootstrap within, and bootstrap across tests are discussed in Section 2.3. Multiple testing adjustment techniques used are the Bonferroni correction, Benjamini and Hochberg's false discovery rate controlling procedure, and Efron's local false

discovery rate with the empirical null distribution. Three maxT procedures (permutation, bootstrap within, bootstrap across) are included in only a few of the 200-gene simulations due to computational constraints. Table 5 shows a list of the procedures included in the simulations.

Table 5. Simulation Testing Procedures and Titles

Title	Description
TT	Two-sample *t*-test with no adjustment
WT	Welch's *t*-test with no adjustment
TT-Bonf	Two-sample *t*-test followed by Bonferroni adjustment
WT-Bonf	Welch's *t*-test followed by Bonferroni adjustment
TT-BH	Two-sample *t*-test followed by Benjamini and Hochberg's false discovery rate controlling procedure
WT-BH	Welch's *t*-test followed by Benjamini and Hochberg's false discovery rate controlling procedure
TT-Efron	Two-sample *t*-test followed by Efron's local false discovery rate procedure
WT-Efron	Welch's *t*-test followed by Efron's local false discovery rate procedure
maxT-P*	Permutation maxT procedure
maxT-W*	Bootstrap within maxT procedure
maxT-A*	Bootstrap across maxT procedure

*Included in only a small number of the simulation studies.

5.1.5. Sample Size

The sample sizes (per group) examined in the study are 3, 5, 15, and 100. Sample sizes of 3 and 5 are currently common to microarray studies. Sample sizes of 10 or more are currently considered to be prohibitively expensive. However, as we will see, in many cases larger sample sizes are required to achieve the desired power.

5.1.6. Test Level

The levels of the tests (or false discovery rates, when applicable) used are .01, .05, .10, .30, and .50. The choice of level (or false discovery rate) reflects the willingness of a researcher to make some incorrect decisions about non-differentially expressed genes (false positives).

5.1.7. Number of Simulations

The number of simulations used in each run is 1,000. The variation in estimated level and estimated power from this simulation size can be seen as error bars of +/- 2 standard errors in the figures of Appendix F and Appendix G.

5.1.8. Error Rates

The three error rates reported from the simulations are the average per-comparison error rate (PCER), the family-wise error rate (FWER), and the average false discovery rate (FDR). In each simulated data set, PCER is estimated by proportion of genes declared to be significantly different among those which are, in fact, no different. To compute FWER, it is first determined for each simulated data set whether or not there was a false rejection of expression equality. The proportion of data sets resulting in at least one false rejection is the estimate of FWER. FDR is computed for each data set as the proportion of total rejections which were false rejections. If there were no rejections, FDR is 0/0. FDR values of 0/0 were removed from the calculation of average FDR.

5.1.9. Correlation Among Genes

In a small number of the simulations, correlation was introduced among genes. This correlation is intended to reflect the tendency of some genes to be expressed in groups. One half of the genes were forced to be highly correlated in groups of 10. The correlation used was 0.707, corresponding to R-squared = 0.50.

5.2. Results

The results of the simulation study are seen in Appendix F and Appendix G. Table 6 gives a summary of the figures of Appendix F involving simulations in which the genes are assumed independent. Each figure shows the average PCER, FWER, FDR, and power for each of the testing methods of Table 3. Variation across simulations is

Table 6. Summary of Figures of Appendix F

Number of Genes	Percent Different	Sample Size (per group)	Figures
200	1%	3	F-1 to F-5
		5	F-6 to F-10
		15	F-11 to F-15
		100	F-16 to F-20
	10%	3	F-21 to F-25
		5	F-26 to F-30*
		15	F-31 to F-35
		100	F-36 to F-40
2,000	1%	3	F-41 to F-45
		5	F-46 to F-50
		15	F-51 to F-55
		100	F-56 to F-60
	10%	3	F-61 to F-65
		5	F-66 to F-70
		15	F-71 to F-75
		100	F-76 to F-80
20,000	1%	3	F-81 to F-85
		5	F-86 to F-90
		15	F-91 to F-95
		100	F-96 to F-100
	10%	3	F-101 to F-105
		5	F-106 to F-110
		15	F-111 to F-115
		100	F-116 to F-120

*Includes results for the maxT procedure.

shown by error bars, which represent 2 standard errors. Figures are in groups of 5 corresponding to the 5 testing levels compared (see Section 5.1.6 for details on the levels).

Because of the prohibitive number of calculations required to run the maxT procedures, simulations evaluating the maxT procedure were only run for three scenarios. Each of these three scenarios involved only 200 genes with 10% of the genes differentially expressed. One involved independent samples with 5 simulated individuals per group. The other two scenarios involved correlation among the genes as described in Section 5.1.9. Those two scenarios had sample sizes of 5 and 15.

The results of the correlated data simulations are found in Appendix G. A summary of the figures in Appendix G is shown in Table 7.

Table 7. Summary of Figures of Appendix G

Number of Genes	Percent Different	Sample Size (per group)	Figures
200	10%	5	G-1 to G-5*
		15	G-6 to G-10*
2,000	10%	3	G-11 to G-15
		5	G-16 to G-20
		15	G-21 to G-25
		100	G-26 to G-30

*Includes results for the maxT procedure.

5.3. Summary and Discussion

Figures G-1 through G-10 are perhaps the most useful for comparing the 11 multiple comparison error rate controlling procedures considered in the simulation study. These figures show the results for 200 simulated genes for 5 and 15 individuals per group. One half of the 200 genes were constrained to have a correlation of 0.707. Ten percent (20) of the 200 genes were simulated to be differentially expressed genes, according to the Chi-square distribution of Figure 12.

If we first consider the per-comparison error rate, the WT and TT procedures (see Table 3 for procedure descriptions) perform as expected. These procedures do not control for multiple testing and therefore should not be considered in power comparisons.

The procedures which were developed with the intent to control FWER are maxT-A, maxT-W, maxT-P, WT-Bonf, and TT-Bonf. This control is achieved at all levels for all five procedures except for maxT-A, which rejects equivalent gene expression slightly too often at levels 0.30 and 0.50 (see Figures G-4, G-5, G-9, and G-10). The power for the maxT-P, WT-Bonf, and TT-Bonf is nearly identical at all five levels, indicating that with this amount of correlation, these procedures give very similar results. The maxT-A has the highest power, which is consistent with its high FWER. For samples of size 5, the maxT-W procedure is overly conservative, resulting in a low power. Thus, when this amount of correlation is present, one might equally choose any of the maxT-P, WT-Bonf, and TT-Bonf procedures, perhaps favoring the one with the lowest FWER.

The procedures WT-Efron, TT-Efron, WT-BH, and TT-BH were developed to control the false discovery rate. In every figure the TT-BH has the highest power and maintains the specified false discovery rate. The WT-Efron and TT-Efron procedures do not control the false discovery rate under the scenarios of Figures G-1 and G-2. In fairness to Efron's method, however, it is noted that no covariates were introduced in the simulation which would cause dilation of the null hypothesis density. If such covariates were introduced, Efron's method may show improved relative performance.

For one that is willing to make the concession of increasing the error rate, the increase in power is substantial. For example, with 5 individuals per group, the power for the TT-BH procedure at 0.05 false discovery rate control is 0.20 (Figure G-2). The power for the same procedure at 0.50 control is 0.60.

Although a typical sample size for microarray studies is 3 individuals per group, it is clear from all the simulation studies that the power of such an experiment is very low. The highest power achieved for any of the 0.30 false discovery rate controlled procedures for sample sizes of 3 per group was 0.11 (see Figure F-64, TT-BH). For sample sizes of 5 per group, the highest power is 0.47 (see Figure F-69, TT-BH), a considerable increase. If it is the desire of the researcher to control the family-wise error rate rather than the false discovery rate, samples of size 3 will result in at most 0.02 power (Figure F-64, TT-Bonf). Samples of size five controlling the family-wise error rate at level 0.30 approach 0.08 power (Figure F-69, TT-Bonf). Figures 13-15 show the

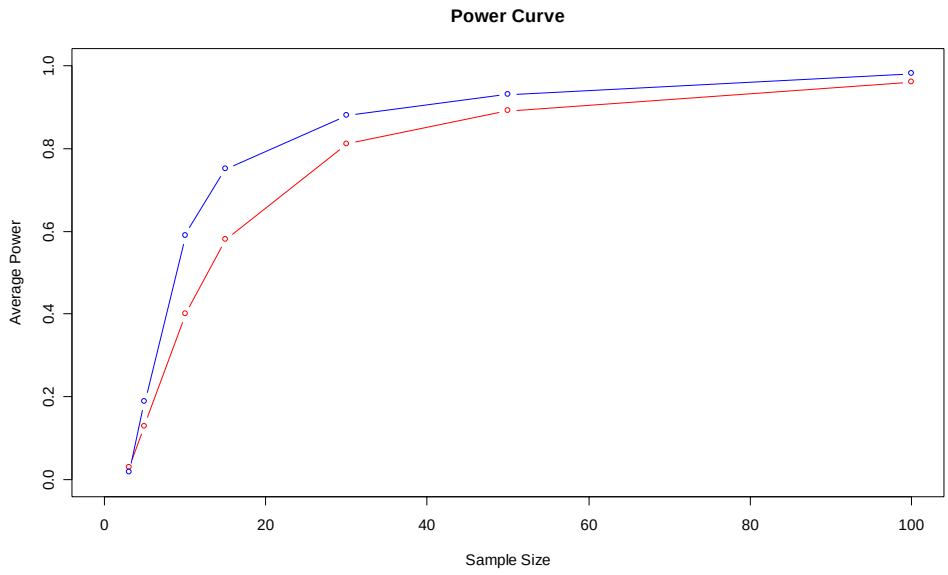

Figure 13. Power Curve for 200 Gene Scenarios. The blue line represents the power for the WT-BH procedure. The red line represents the power for the WT-Bonf procedure. Each power estimate comes from 1,000 simulations of 200 genes, of which 10% (20) are differentially expressed. The significance (or fdr) level used was 0.10.

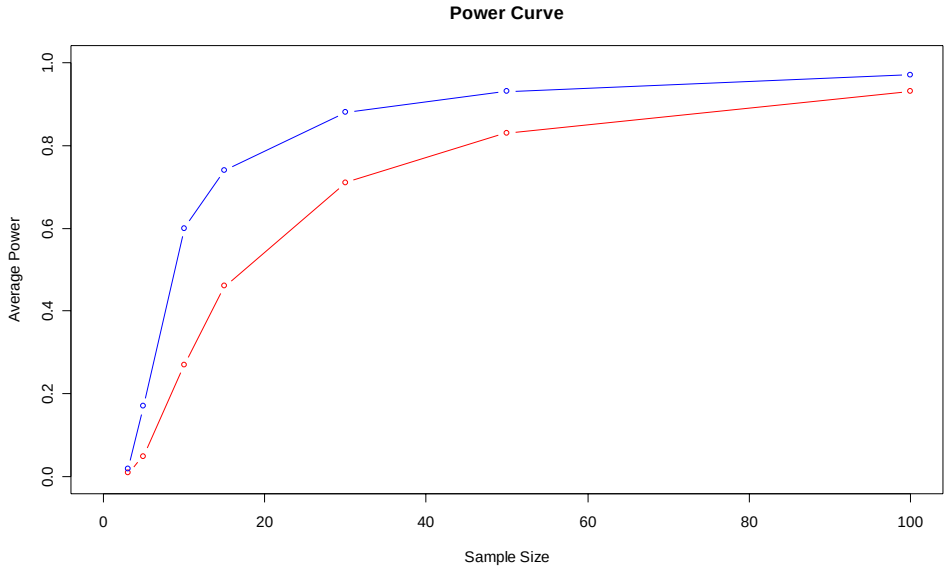

Figure 14. Power Curve for 2,000 Gene Scenarios. The blue line represents the power for the WT-BH procedure. The red line represents the power for the WT-Bonf procedure. Each power estimate comes from 1,000 simulations of 2,000 genes, of which 10% are differentially expressed. The significance (or fdr) level used was 0.10.

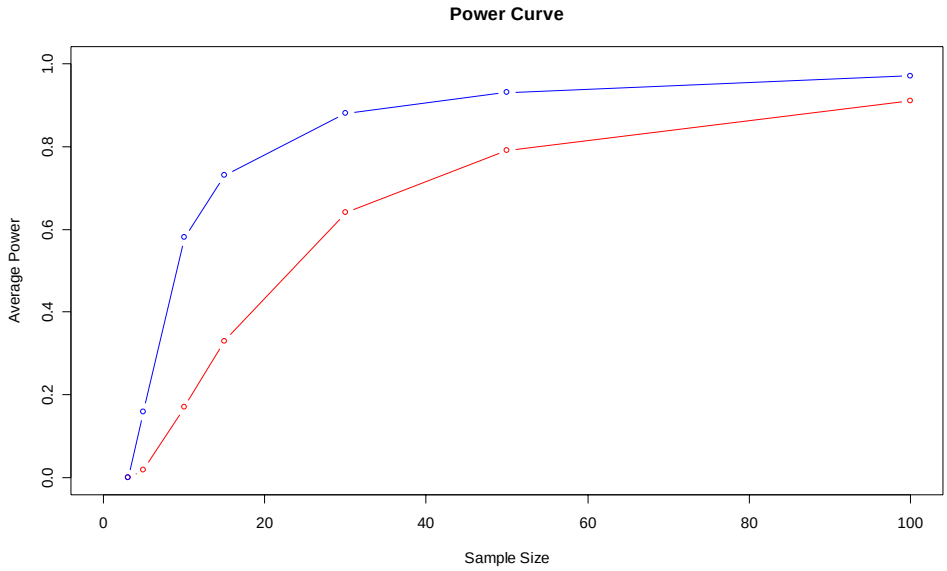

Figure 15. Power Curve for 20,000 Gene Scenarios. The blue line represents the power for the WT-BH procedure. The red line represents the power for the WT-Bonf procedure. Each power estimate comes from 1,000 simulations of 20,000 genes, of which 10% are differentially expressed. The significance (or fdr) level used was 0.10.

effect of sample size on power for 7 sample sizes (per group): 3, 5, 10, 15, 30, 50, and 100. The two representative procedures compared in the figures are WT-BH and WT-Bonf.

Figures 13 through 15 show that the power of the WT-BH false discovery rate controlling procedure is little affected by the number of genes in the experiment. The figures also show a gradual decrease in power as the number of genes increases for the WT-Bonf family-wise error rate controlling procedure.

The figures show a dramatic increase in power for any increase in sample size until about 15 per group, where the power curve begins to level off. Figure 16 (below) shows the power curves of the WT-BH, WT-Bonf, and WT-Efron procedures when correlated samples are present in the simulation. Although the intent, in part, of the WT-Efron procedure is to incorporate correlation information into the adjustment procedure while controlling false discovery rates, the correlation did not introduce a significant change in the null hypothesis density. With no dilation of the null hypothesis density (i.e., from other covariates), there is no advantage afforded by creating the empirical null hypothesis density. For this simulation scenario, the WT-Efron procedure is clearly inferior to the WT-BH procedure, which also maintains false discovery rate control. Simulation studies involving scenarios in which the null hypothesis density is different from a standard normal density are left for further research.

There are three primary ways in which the power of a microarray study may be increased. The most obvious way to increase power is to increase the sample size. Because microarray chips are very expensive, this may not be an option. Careful examination of the power curves of Figures 13-16 will aid the decision of the number of microarray chips to be included in a gene expression study.

A second way to increase power is to change from controlling FWER to controlling the false discovery rate. Further increase in power may be obtained by increasing the false discovery rate itself. Increasing the power in this way, of course, depends on the willingness of the researcher to allow false discoveries.

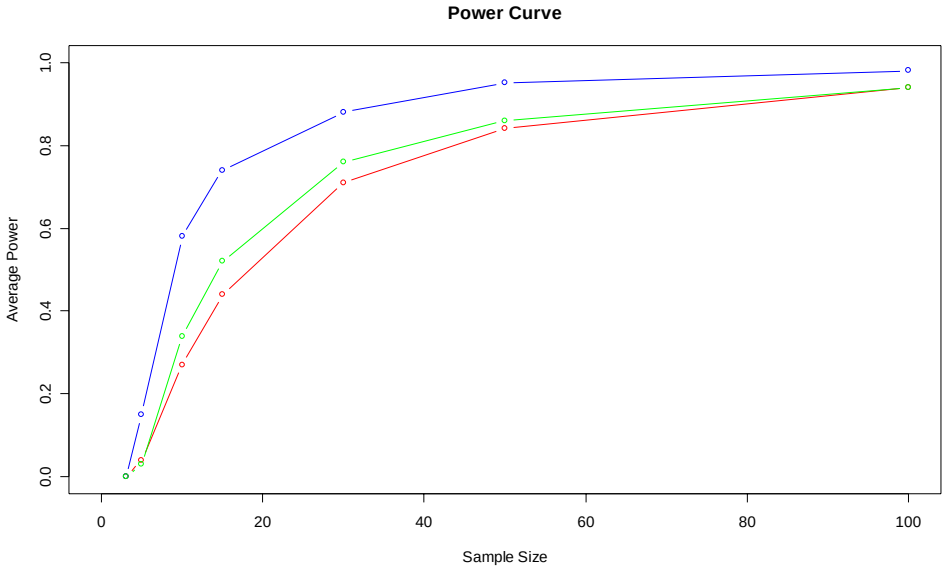

Figure 16. Power Curve for 2,000 Gene Scenarios with 1/2 Genes Correlated. The blue line represents the power for the WT-BH procedure. The red line represents the average power for the WT-Bonf procedure. The green line represents the average power for the WT-Efron procedure. Each power estimate comes from 1,000 simulations of 2,000 genes, of which 10% (200) are differentially expressed. The significance (or fdr) level used was 0.10. Correlated genes have 0.707 correlation in groups of 10.

Thirdly, power may be increased by decreasing the variation within the samples themselves. Considerable discussion of this topic is found in Gautier et al. (2004) and Bolstad et al. (2005).

6. CONCLUSIONS

When a two-sample test is carried out with small sample sizes in each group, there is very little information that can be obtained about the underlying distributions from which the samples come. Although commonly thought of as suitable non-parametric solutions for small sample size scenarios, the permutation, bootstrap, and rank tests are in general less accurate, less correct, and often less powerful than the common *t*-test and Welch's *t*-test. Further, when large scale multiple testing is to be done, the numbers of permutations or bootstrap replicates are too few to obtain sufficiently small achieved significance levels. The clear individual two-sample test favorites are the *t*-test and Welch's *t*-test. Welch's t-test is slightly more conservative and is more accurate when assumptions of equal variance or normal underlying distributions are not met. The *t*-test has higher power when variances are equal or unequal, but is less accurate when variances are unequal.

Very innovative approaches to large scale testing have been developed in the past 15 years, focusing on both FWER and the false discovery rate. From simulation studies based on an experiment of the Miranda lab, we can see that controlling the false discovery rate results in considerably higher power than FWER control. If FWER control is desired, the permutation maxT procedure performs similarly to the Bonferroni procedure, but the power for either is very low (< 0.10) for group samples of size 5 or smaller. For these procedures, increasing the number of genes results in a further decrease in the power.

More promising are the methods which control the false discovery rate. Benjamini and Hochberg's (1995) false discovery rate controlling procedure outperforms Efron's (2004) procedure under the simulation scenarios I examined. In fairness to Efron's procedure, though, it requires careful consideration of the distribution of the *P*-values and careful choices in histogram bin widths and spline parameters. These considerations were generalized in the simulation study, which may have been the cause of its poor performance. I emphasize that to use this procedure, one need have

experience in statistical programming and a basic understanding of the parameters involved in smoothing techniques for curve fitting.

Whatever the error control method used, for an experiment similar to the Miranda experiment, an extremely large increase in power is obtained by increasing the sample size in each group from the range of 3-5 to 10-15. Substantial power increases may also be possible by using improved methods to decrease internal variation. As and example, to achieve 0.80 power with false discovery rate control of 0.10, I recommend individual *t*-tests followed by Benjamini and Hochberg's (1995) procedure with 20 individuals in each of the control and treatment groups. Twenty-five to 30 individuals per group would be needed to obtain the same power for false discovery rate control of 0.01. These sample sizes should be appropriate regardless of the number of genes in the study.

REFERENCES

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Ser. B, 57, 289-300.

Bhattacharjee, G. P. (1968), "Non-normality and Heterogeneity in Two Sample *t*-Test," *Annals of the Institute of Statistical Mathematics*, 20, 239-253.

Bolstad, B. M., Collin, F., Simpson, K. M., Irizarry, R. A. and Speed, T. P. (2005), "Design and Low Level Analysis of Microarray Experiments," *International Review of Neurobiology*, to appear.

Braun, H. I. (1994), *The Collected Works of John W. Tukey, Vol. VIII, Multiple Comparisons: 1948-1983*, New York: Chapman & Hall.

Chatterjee, S. K. (2003), *Statistical Thought: A Perspective and History*, New York: Oxford University Press, Inc.

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992), "Local Regression Models," in *Statistical Models in S*, eds. J. M. Chambers and T. J. Hastie, Pacific Grove, CA: Wadsworth & Brooks/Cole.

Cochran, W. G. (1976), "Early Development of Techniques in Comparative Experimentation," in *On the History of Statistics and Probability*, ed. D. B. Owen, New York: Marcel Dekker, Inc.

Dixon, D. O., and Duncan, D. B. (1975), "Minimum Bayes Risk *t*-intervals for Multiple Comparisons, *Journal of the American Statistical Association*, 70, 822-831.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003), "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science*,18, 71-103.

Duncan, D. B. (1947), "Significance Tests for Differences Between Ranked Variates Drawn from Normal Populations," unpublished Ph. D. dissertation, Iowa State College, Dept. of Mathematics.

Duncan, D. B. (1951), "A Significance Test for Differences Between Ranked Treatments in an Analysis of Variance," *Virginia Journal of Science*, 2, 172-189.

Duncan, D. B. (1952), "On the Properties of the Multiple Comparisons Test," *Virginia Journal of Science*, 3, 49-67.

Duncan, D. B. (1955), "Multiple range and multiple F test," *Biometrics*, 11, 1-42.

Duncan, D. B. (1975), "*t* Tests and Intervals for Comparisons Suggested by the Data," *Biometrics*, 31, 339-359.

Dunnett, C. W. (1955), "A Multiple Comparison Procedure for Comparing Several Treatments with a Control," *Journal of the American Statistical Association*, 50, 1096-1121.

Dunnett, C. W. (1980), "Pairwise Multiple Comparisons in the Homogeneous Variances, Unequal Sample Size Case," *Journal of the American Statistical Association*, 75, 789-795.

Dwass, M. (1960), "Some *k*-sample Rank-order Tests, in *Contributions to Probability and Statistics*, ed. I. Olkin, Stanford, CA: Stanford University Press.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, 7, 1-26.

Efron, B. (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, Society for Industrial and Applied Mathematics CBMS-NSF Monographs, Philadelphia.

Efron, B. (2004), "Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis," *Journal of the American Statistical Association*, 99, 96-104.

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall.

Einot, I., and Gabriel, K. R. (1975), "A Study of the Powers of Several Methods in Multiple Comparisons," *Journal of the American Statistical Association*, 70, 574-583.

Fisher, R. A. (1925), "Applications of "Student's" Distribution," *Metron*, 5, 90-104.

Fisher, R. A. (1934), *Statistical Methods for Research Workers* (5th ed.), Edinburgh: Oliver & Boyd.

Fisher, R. A. (1935). *The Design of Experiments* (4th ed.), London: Oliver & Boyd.

Gabriel, K. R. (1969), "Simultaneous Test Procedures – Some Theory of Multiple Comparisons," *Annals of Mathematical Statistics*, 40, 224-250.

Gauss, C. F. (1809), *Theoria motus corporum celestium*. Hamburg: Perthes et Besser. Translated, 1857, as *Theory of Motion of the Heavenly Bodies Moving about the Sun in Conic Sections*, trans. C. H. Davis, Boston: Little, Brown. Reprinted, 1963, New York: Dover.

Gautier, L., Cope, L. M., Bolstad, B. M., and Irizarry, R. A. (2004), "Affy – Analysis of Affymetrix GeneChip Data at the Probe Level," *Bioinformatics*, 20(3), 307-315.

Gayen, A. K. (1950), "Significance of Difference Between the Means of Two Non-Normal Samples," *Biometrika*, 37, 399-408.

Ge, Y., Dudoit, S., and Speed, T. P. (2003), "Resampling-based Multiple Testing for Microarray Data Analysis, *Test*, 12, 1-77.

Geary, R. C. (1947), "Testing for Normality," *Biometrika*, 34, 209-242.

Good, P. (2000), *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses* (2nd ed.), New York: Springer-Verlag.

Harter, H. L. (1980), "History of multiple comparisons," in *Handbook of Statistics*, Vol. 1, ed. P. R. Krishnaiah, Amsterdam: North Holland, 617-622.

Hayter, A. J. (1984), "A Proof of the Conjecture that the Tukey-Kramer Multiple Comparisons Procedure is Conservative," *Annals of Statistics*, 12, 61-75.

Hochberg, Y., and Tamhane, A. C. (1987), *Multiple Comparison Procedures*, New York: John Wiley.

Hsu, J. C. (1981), "Simultaneous Confidence Intervals for All Distances from the 'Best'," *Annals of Statistics*, 9, 1026-1034.

Hsu, J. C. (1982), "Simultaneous Inference with Respect to the Best Treatment in Block Designs," *Journal of the American Statistical Association*, 77, 461-467.

Hsu, J. C. (1984), "Constrained Two-sided Simultaneous Confidence Intervals for Multiple Comparisons with the Best," *Annals of Statistics*, 12, 1136-1144.

Hsu, J. C. (1996), *Multiple Comparisons: Theory and Methods*, London: Chapman & Hall.

Kramer, C. Y. (1956), "Extension of Multiple Range Tests to Group Means with Unequal Numbers of Replications," *Biometrics*, 12, 309-310.

Kuels, M. (1952), “The Use of the “Studentized Range” in Connection with an Analysis of Variance,” *Euphytica*, 1, 112-122.

Legendre, A. M. (1805), *Nouvilles methods pour la determination des orbites des cometes,* Paris: Courcier. Reissued with a supplement, 1806, Second supplement published 1820, A portion of the appendix was translated, 1929, pp. 576-579 in *A Source Book in Mathematics*, ed. D. E. Smith, trans. by H.A. Ruger and H. M. Walker, New York: McGraw-Hill; reprinted 1959 in 2 vols., New York: Dover.

Lehmann, E. L. and Shaffer, J. P. (1979), “Optimum significance levels for multistage comparison procedures,” *Annals of Statistics*, 7, 27-45.

Mann, H. B., and Whitney, D. R. (1947), “On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other,” *Annals of Mathematical Statistics*, 18, 50-60.

Miller, R. G. (1981), *Simultaneous Statistical Inference* (2nd ed.), Berlin: Springer-Verlag.

Miller, R. G. (1986), *Beyond ANOVA, Basics of Applied Statistics*, New York: Wiley.

Nemenyi, P. (1961), “Some Distribution-free Multiple Comparison Procedures in the Asymptotic Case,” *Annals of Mathematical Statistics*, 32, 921-922.

Nemenyi, P. (1963), “Distribution-free Multiple Comparisons,” unpublished Ph.D. dissertation, Princeton University, Dept. of Mathematics.

Newman, D. (1939), “The Distribution of Range in Samples from a Normal Population, Expressed in Terms of an Independent Estimate of the Standard Deviation” *Biometrika*, 31, 20-30.

Ostle, B., and Malone, L. C. (1988), *Statistics in Research: Basic Concepts and Techniques for Research Workers* (4th ed), Ames, Iowa: Iowa State University Press.

Pearson, E. S. (1929), “The Distribution of Frequency Constants in Small Samples from Non-normal Symmetrical and Skew Populations,” *Biometrika*, 21, 259-286.

Pearson, E. S. (1939), “‘Student’ as Statistician,” *Biometrika*, 30, 210-250.

Pearson, E. S., and Please, N. W. (1975), "Relation Between the Shape of Population Distribution and the Robustness of Four Simple Test Statistics," *Biometrika*, 62, 223-241.

Pennisi, E. (2001), "The Human Genome," *Science*, 291, 1177-1180.

Pitman, E. J. G. (1937), "Significance Tests Which May Be Applied to Samples from Any Populations," *Journal of the Royal Statistical Society* (superseded by Series B), 4, 119-130.

Ryan, T. A. (1960), "Significance Tests for Multiple Comparison of Proportions, Variances and Other Statistics," *Psychological Bulletin*, 57, 318-328.

Scheffe, H. (1953), "A Method for Judging All Contrasts in the Analysis of Variance," *Biometrika*, 40, 87-104.

Smyth, G. K., and Speed, T. (2003), "Normalization of cDNA Microarray Data," *Methods,*31(4), 265-273.

Steel, R. G. D. (1959a), "A Multiple Comparison Sign Test: Treatments Versus Control," *Journal of the American Statistical Association*, 54, 767-775.

Steel, R. G. D. (1959b), "A Multiple Comparison Rank Sum Test: Treatments Versus Control," *Biometrics*, 15, 560-572.

Steel, R. G. D. (1960), "A Rank Sum Test for Comparing All Pairs of Treatments," *Technometrics*, 2, 197-207.

Steel, R. G. D. (1961), "Some Rank Sum Multiple Comparisons Tests," *Biometrics*, 17, 539-552.

Stigler, S. M. (1986), *The History of Statistics,* Cambridge, MA: Belknap Press.

Storey, J. D. (2001), "The Positive False Discovery Rate: A Bayesian Interpretation and the *q*-value," Technical Report 2001-12, Stanford University, Department of Statistics.

Storey, J. D. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society*, Ser. B, 64, 479-498.

Stuart, A., and Ord, J. K. (1994), *Kendall's Advanced Theory of Statistics: Volume 1, Distribution Theory*, New York: Oxford University Press.

Student (1908), "The Probable Error of a Mean," *Biometrika*, 6, 1-25.

Tamhane, A. C. (1995), "Multiple Comparison Procedures," in *Handbook of Statistics: Design and Analysis of Experiments, Volume 13*, eds. S. Ghosh and C. R. Rao, Amsterdam: North-Holland, 607-610.

Troendle, J. F., Korn, E. L., and McShane, L. M., (2004), "An Example of Slow Convergence of the Bootstrap in High Dimensions," *The American Statistician*, 58, 25-29.

Tukey, J. W. (1952), "Allowances for Various Types of Error Rates," unpublished invited address presented at Blacksburg meeting of Institute of Mathematical Statistics and Biometric Society.

Tukey, J. W. (1953), "The Problem of Multiple Comparisons," dittoed manuscript of 396 pages, Princeton University, Dept. of Statistics.

Waller, R. A., and Duncan, D. B. (1969), "A Bayes Rule for the Symmetric Multiple Comparisons Problem," *Journal of the American Statistical Association***,** 64, 1484-1503.

Waller, R. A., and Duncan, D. B. (1974), "A Bayes Rule for the Symmetric Multiple Comparisons Problem II," *Annals of the Institute of Statistical Mathematics*, 26, 247-264.

Waller, R. A., and Kemp, K. E. (1975), "Computations of Bayesian *t*-values for Multiple Comparisons," *Journal of Statistical Computation and Simulation*, 4, 169-171.

Welch, B. L. (1947), "The Generalization of "Student's" Problem When Several Different Population Variances Are Involved," *Biometrika*, 34, 28-35.

Welch, B. L. (1949), "Further Mote on Mrs. Aspin's Tables and on Certain Approximations to the Tabled Function," *Biometrika*, 36, 293-296.

Welsch, R. E. (1972), "A Modification of the Newman-Keuls Procedure for Multiple Comparisons," Working paper 612-72, M.I.T., Sloan School of Management, Boston, MA.

Westfall, P. H., and Young, S. S. (1993), *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*, New York: John Wiley & Sons, Inc.

Wilcoxon, F. (1945), “Individual Comparisons by Ranking Methods,” *Biometrics Bulletin* (superseded by *Biometrics*), 1, 80-83.

Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. P. (2001), “Normalization for cDNA Microarray Data,” in *Microarrays: Optical Technologies and Informatics*, eds. M. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty, volume 4266 of Proceedings of The International Society for Optical Engineering (SPIE).

Yang, Y. H., and Speed, T. P. (2003), “Design and analysis of comparative microarray experiments,” in *Statistical Analysis of Gene Expression Microarray Data*, ed. T. Speed, Boca Raton, FL: Chapman & Hall, CRC Press LLC.

Zimmerman, D. W. (2004), “Inflation of Type I Error Rates by Unequal Variances Associated with Parametric, Nonparametric, and Rank-Transformation Tests,” *Psicologica*, 25, 103-133.

APPENDIX A

RESAMPLED T DISTRIBUTIONS

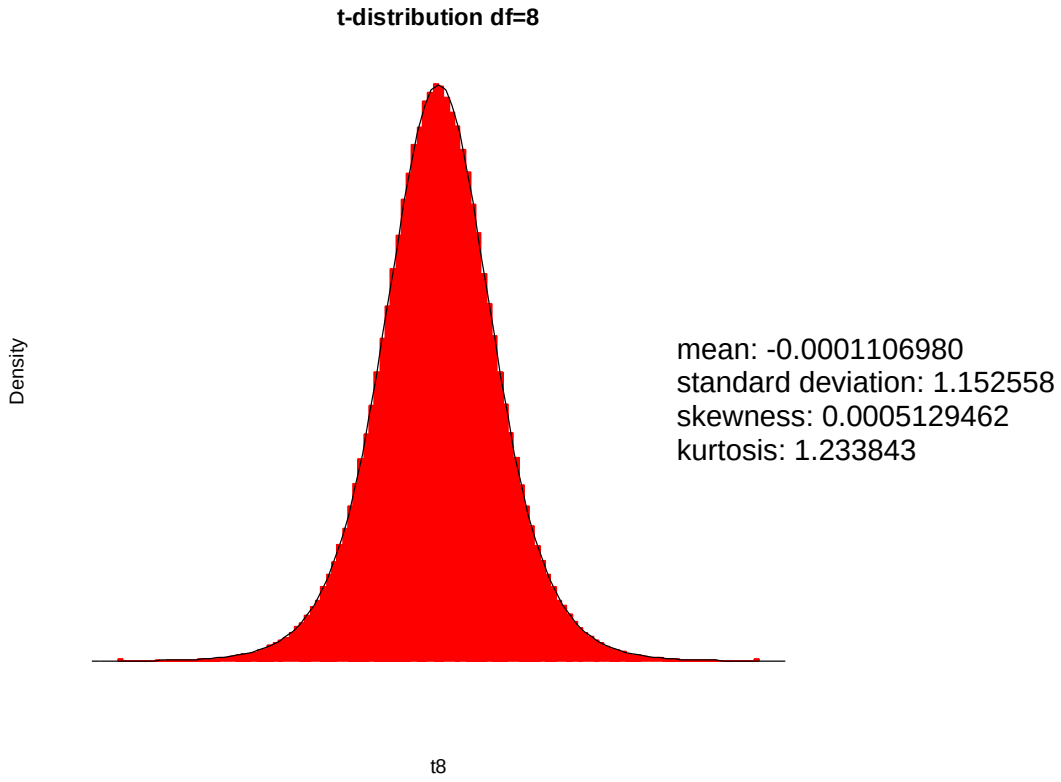

Figure A-1. Reference t Distribution. Histogram of 1,000,000 simulated t (df = 8) values with true curve overlaid.

Sample A-1: Two random samples of size 5 from standard normal distributions

$\boldsymbol{y_1} = (y_{11}, y_{12}, y_{13}, y_{14}, y_{15})$ = (0.4273074, 0.4357245, 0.4310095, 1.1086347, -0.4688814)

$\boldsymbol{y_2} = (y_{21}, y_{22}, y_{23}, y_{24}, y_{25})$ = (1.8306395, -0.364505, 0.6374127, 1.0841149, -0.273859)

Summary statistics and histograms in the figures that follow are produced after setting all values outside -7 and 7 to -7 and 7 respectively.

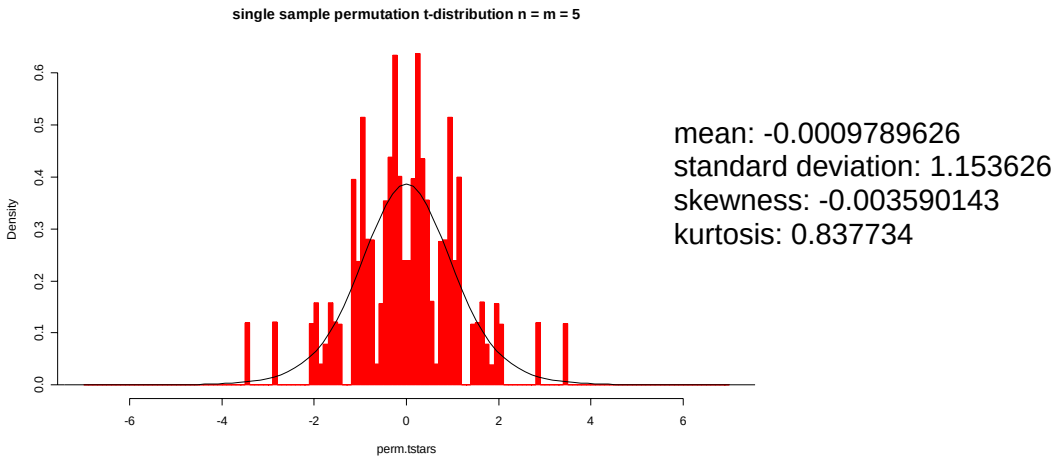

Figure A-2. Permutation t Distribution Based on 100,000 Resamples of Sample 1 with Known t Distribution (df = 8) Overlaid.

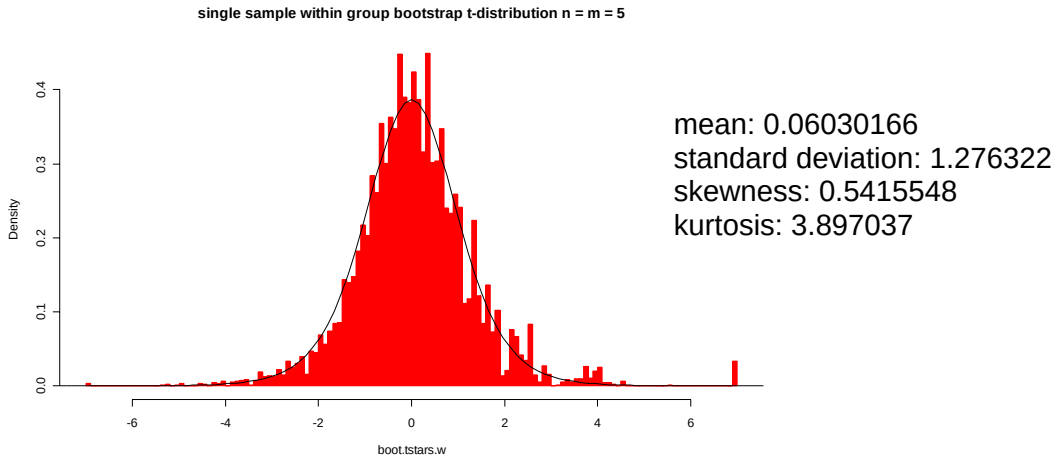

Figure A-3. Within Group Bootstrap t Distribution Based on 100,000 Resamples of Sample 1 with Known t (df = 8) Overlaid.

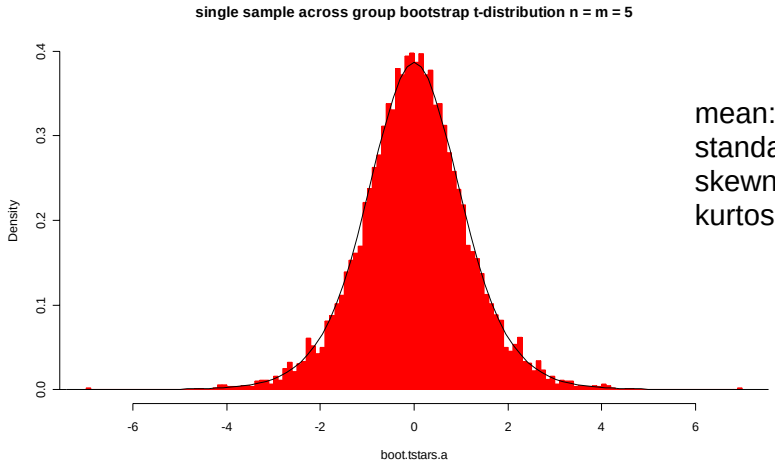

Figure A-4. Across Group Bootstrap t Distribution Based on 100,000 Resamples of Sample 1 with Known t (df = 8) Overlaid.

Sample A-2: Two random samples of size 5 from standard normal distributions

$\boldsymbol{y_1} = (y_{11}, y_{12}, y_{13}, y_{14}, y_{15})$ = (-0.987590, -1.002704, 2.731845, -1.293495, 1.533078)

$\boldsymbol{y_2} = (y_{21}, y_{22}, y_{23}, y_{24}, y_{25})$ = (-0.538513, -0.247843, -2.403953, 2.151142, -1.110683)

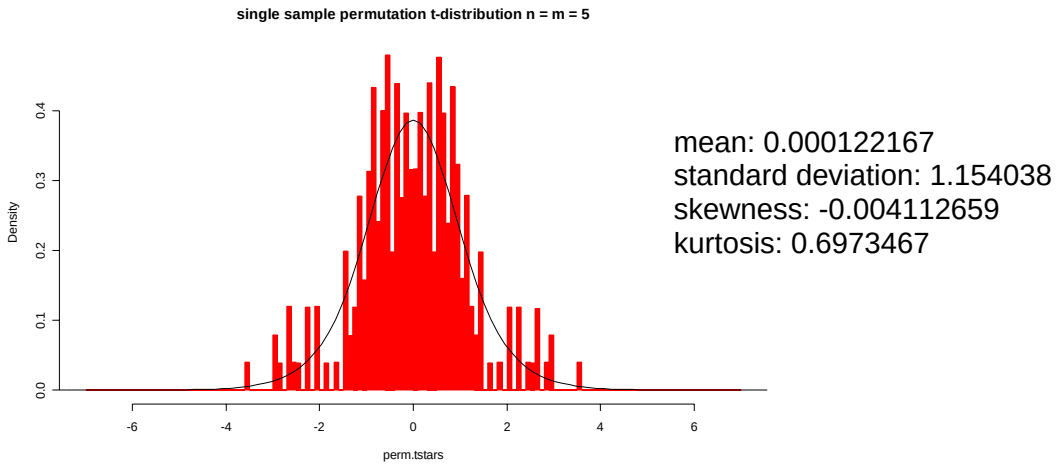

Figure A-5. Permutation t Distribution Based on 100,000 Resamples of Sample 2 with Known t Distribution (df = 8) Overlaid.

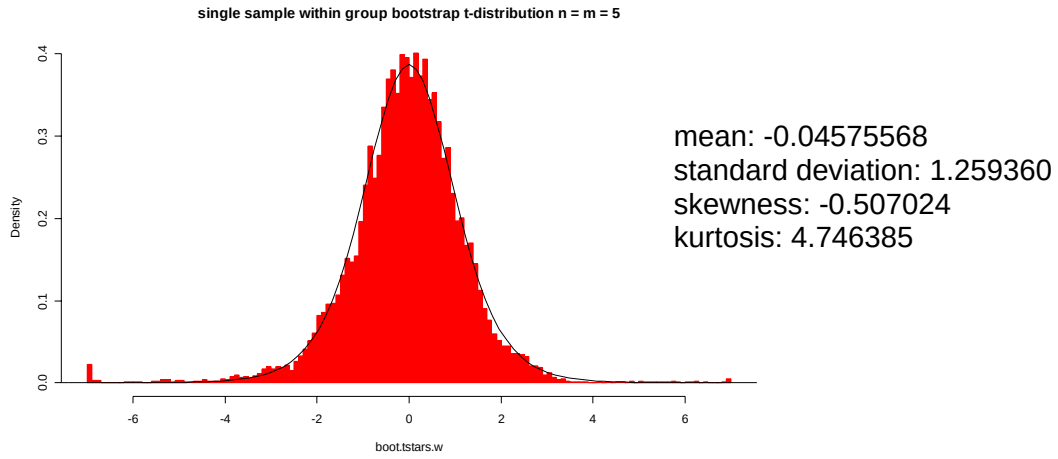

Figure A-6. Within Group Bootstrap t Distribution Based on 100,000 Resamples of Sample 2 with known t (df = 8) Overlaid.

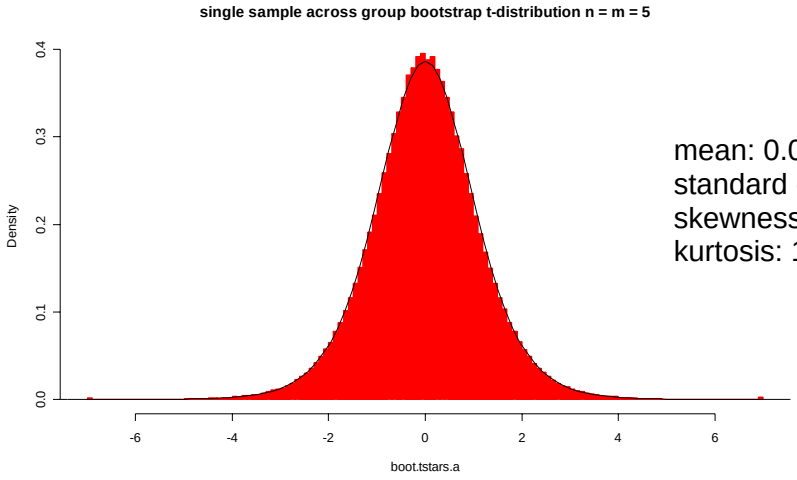

Figure A-7. Across Group Bootstrap t Distribution Based on 100,000 Resamples of Sample 2 with Known t (df = 8) Overlaid.

Sample A-3: Two random samples of size 10 from standard normal distributions.

$\boldsymbol{y_1} = (y_{11}, y_{12}, y_{13}, y_{14}, y_{15}, y_{16}, y_{17}, y_{18}, y_{19}, y_{110})$ = (0.2555992, -2.8272124, -0.9841037, 0.5630356, -0.8407057, 0.4847272, 1.1066374, 0.8250996, -0.7569346, -0.7513211)

$\boldsymbol{y_2} = (y_{21}, y_{22}, y_{23}, y_{24}, y_{25}, y_{26}, y_{27}, y_{28}, y_{29}, y_{210})$ =(0.6224373, -0.4200094, 0.4926407, 0.4720321, 0.7433177, 0.8619719, 0.1843880, -0.1592327, 0.8602116, -0.8879397)

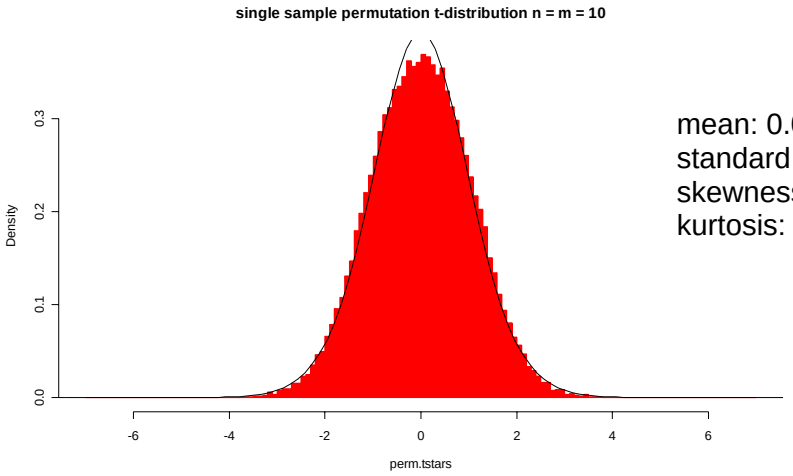

Figure A-8. Permutation t Distribution Based on 100,000 Resamples of Sample 3 with Known t Distribution (df = 18) Overlaid.

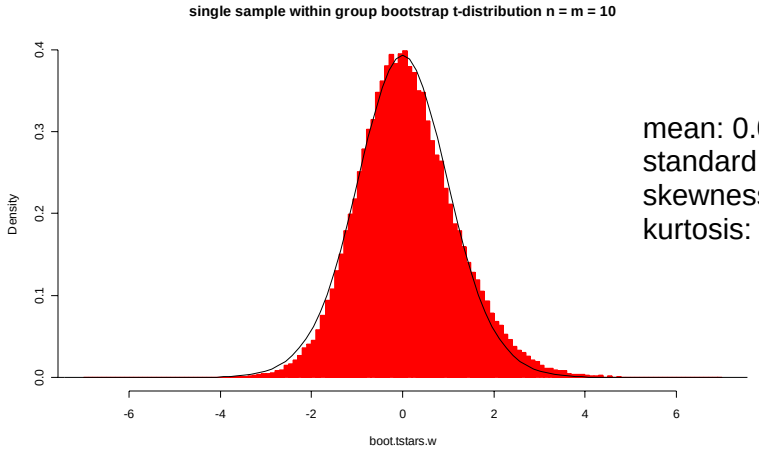

Figure A-9. Within group bootstrap t distribution based on 100,000 resamples of Sample 3 with known t (df = 18) overlaid.

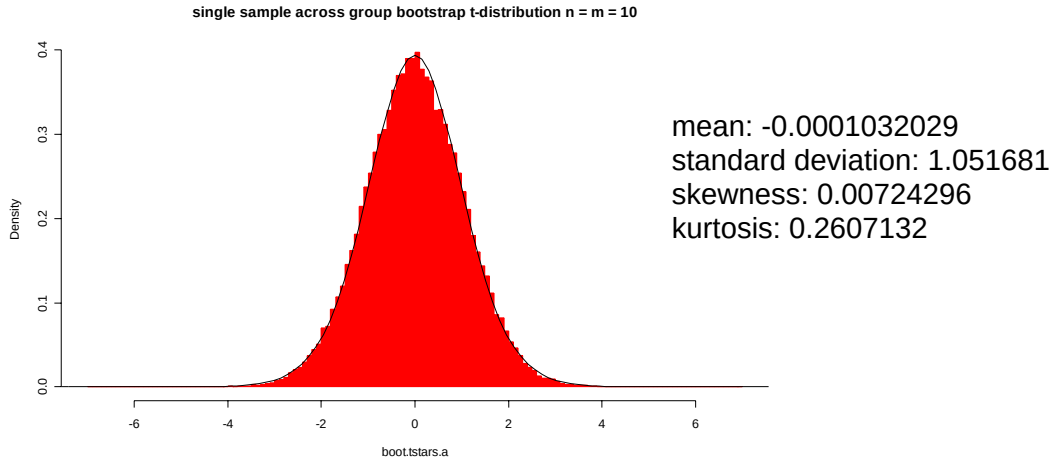

Figure A-10. Across Group Bootstrap t Distribution Based on 100,000 Resamples of Sample 3 with Known t (df = 18) Overlaid.

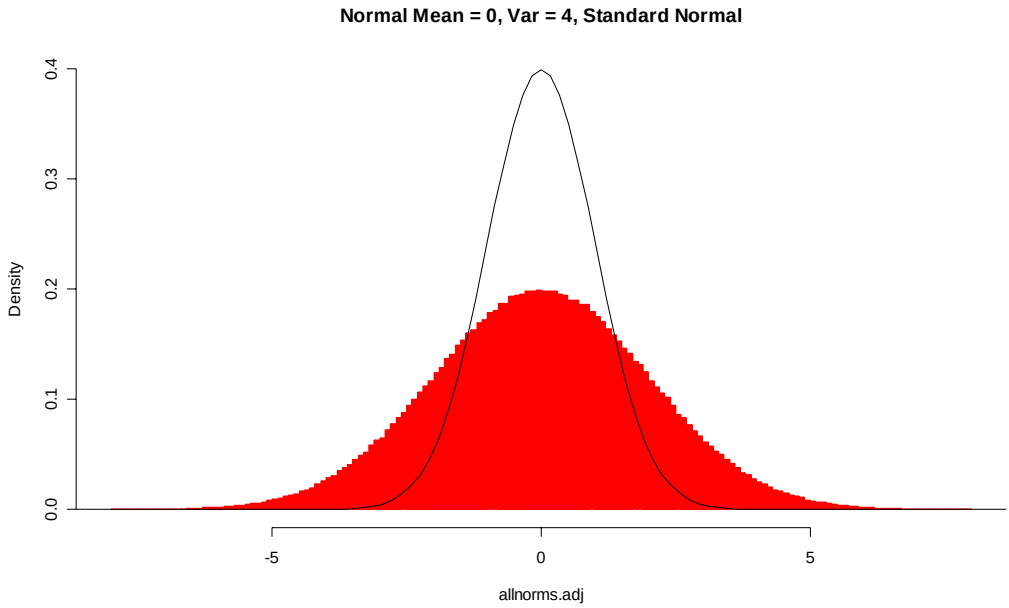

Figure A-11. Distributions Used to Form Null t Distribution: $N(0,1^2)$ and $N(0,2^2)$

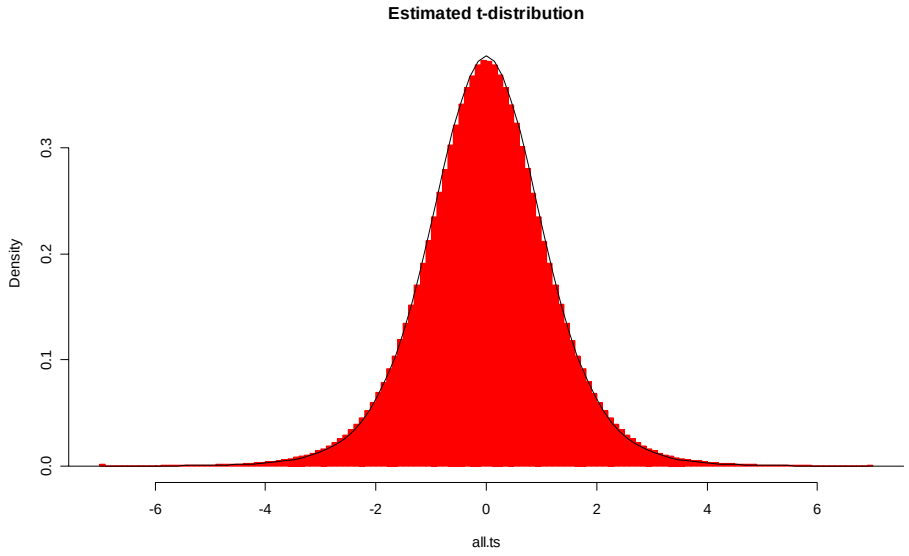

Figure A-12. Histogram is Estimated Null Distribution from 10,000,000 t-statistics from Samples of Size 5 from Distributions of Figure A-11 (the Student's t-distribution is overlaid).

Sample A-4: Two random samples of size 5. The first is from a standard normal distribution, the second from a normal distribution with mean 0 and variance 4.

$\boldsymbol{y_1} = (y_{11}, y_{12}, y_{13}, y_{14}, y_{15})$ = (-1.285275, -1.188968, -1.055633, -1.072713, -1.624037)

$\boldsymbol{y_2} = (y_{21}, y_{22}, y_{23}, y_{24}, y_{25})$ = (0.121921, 1.58351, -1.5408659, -2.1218782, -0.7939920)

Sample A-5: Two random samples of size 5. The first is from a standard normal distribution, the second from a normal distribution with mean 0 and variance 4.

$\boldsymbol{y_1} = (y_{11}, y_{12}, y_{13}, y_{14}, y_{15})$ = (0.465594, 1.1209567, 0.5293456, -0.6451020, -1.9100824)

$\boldsymbol{y_2} = (y_{21}, y_{22}, y_{23}, y_{24}, y_{25})$ = (-2.181948, -2.698829, 1.650957, -0.0511667, -1.3775296)

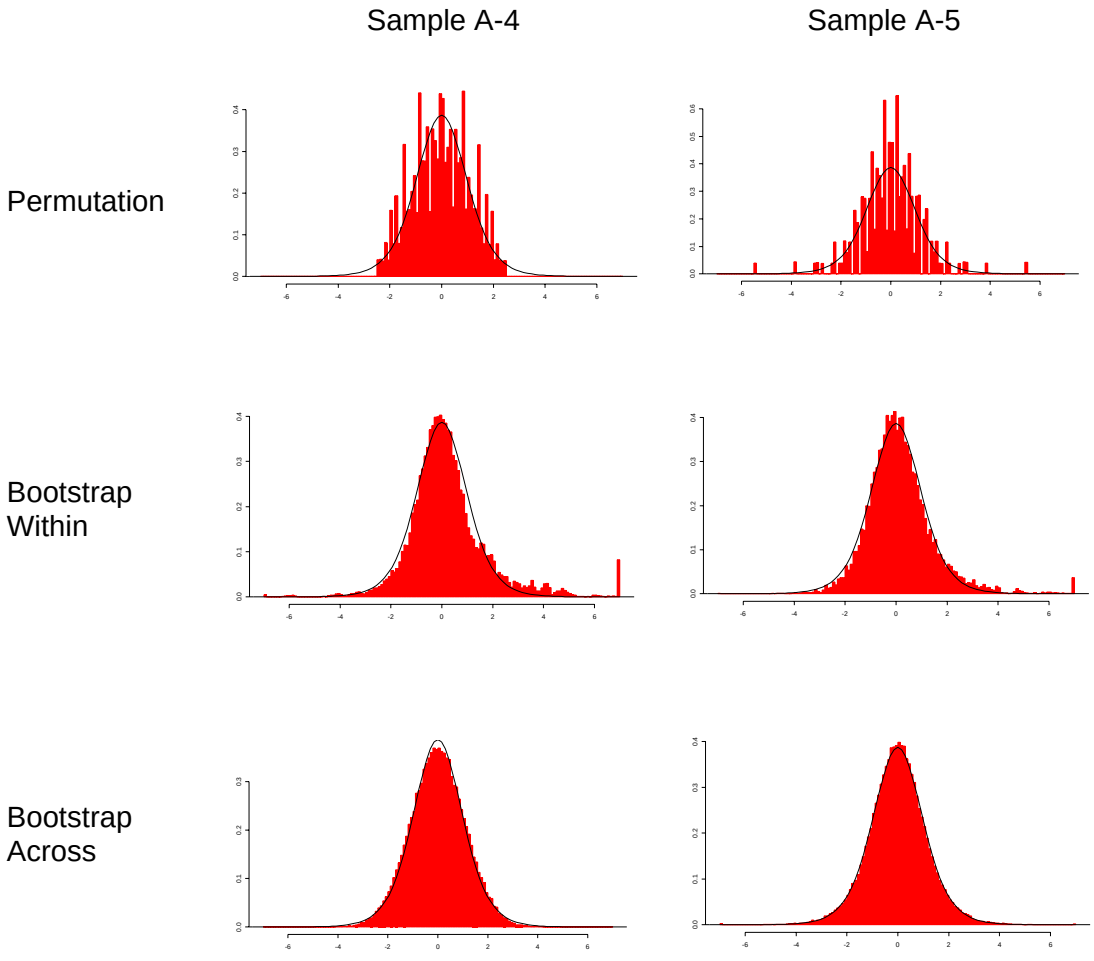

Figure A-13. Estimated Sampling Distributions Based on Samples A-4 and A-5 (B = 100,000 resamples in each). The t-distribution with df=8 is overlaid.

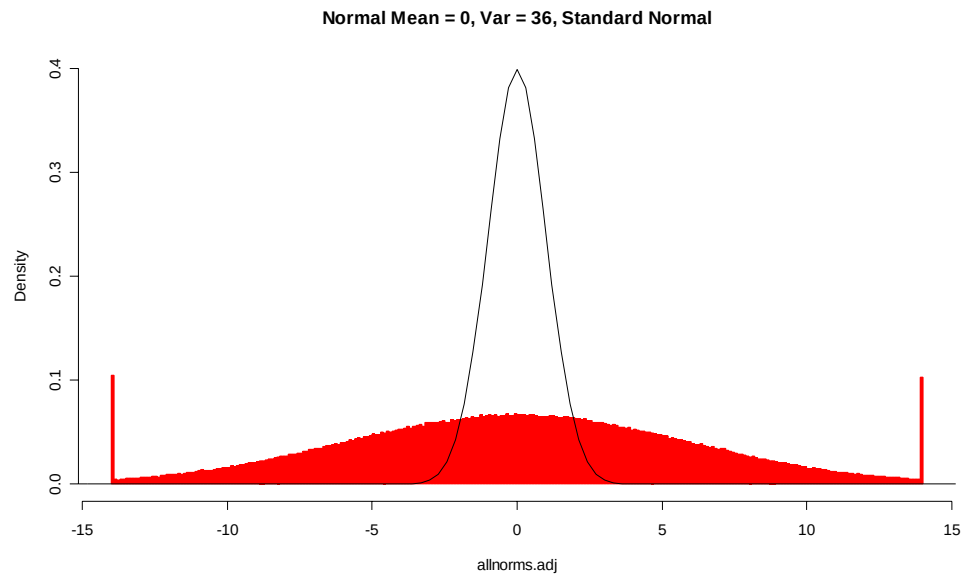

Figure A-14. Distributions Used to Form Null t Distribution: $N(0,1^2)$ and $N(0,6^2)$.

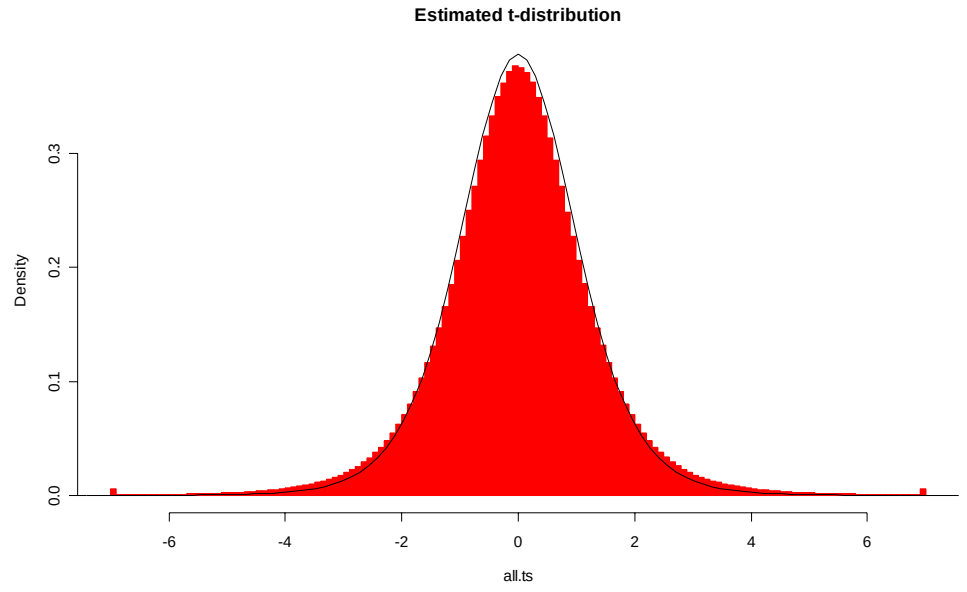

Figure 15. Histogram is Estimated Null Distribution from 10,000,000 t-statistics from Samples of Size 5 from Distributions of Figure A-14 (the Student's t-distribution is overlaid).

Sample A-6: Two random samples of size 5: the first from a standard normal distribution, the second from a normal distribution with mean 0 and variance 36.

$\boldsymbol{y_1} = (y_{11}, y_{12}, y_{13}, y_{14}, y_{15})$ = (-0.263812, -0.4774015, 0.1932803, 0.2453493, 0.2768207)

$\boldsymbol{y_2} = (y_{21}, y_{22}, y_{23}, y_{24}, y_{25})$ = (10.2066, 0.197866, -4.517884, -3.7252808, -17.0642388)

Sample A-7: Two random samples of size 5: the first from a standard normal distribution, the second from a normal distribution with mean 0 and variance 36.

$\boldsymbol{y_1} = (y_{11}, y_{12}, y_{13}, y_{14}, y_{15})$ = (-0.0868123, -1.6410296, -1.0190776, -0.20162, 0.2939446)

$\boldsymbol{y_2} = (y_{21}, y_{22}, y_{23}, y_{24}, y_{25})$ = (-8.932675, -7.409822, -3.906738, 11.550006, -1.342640)

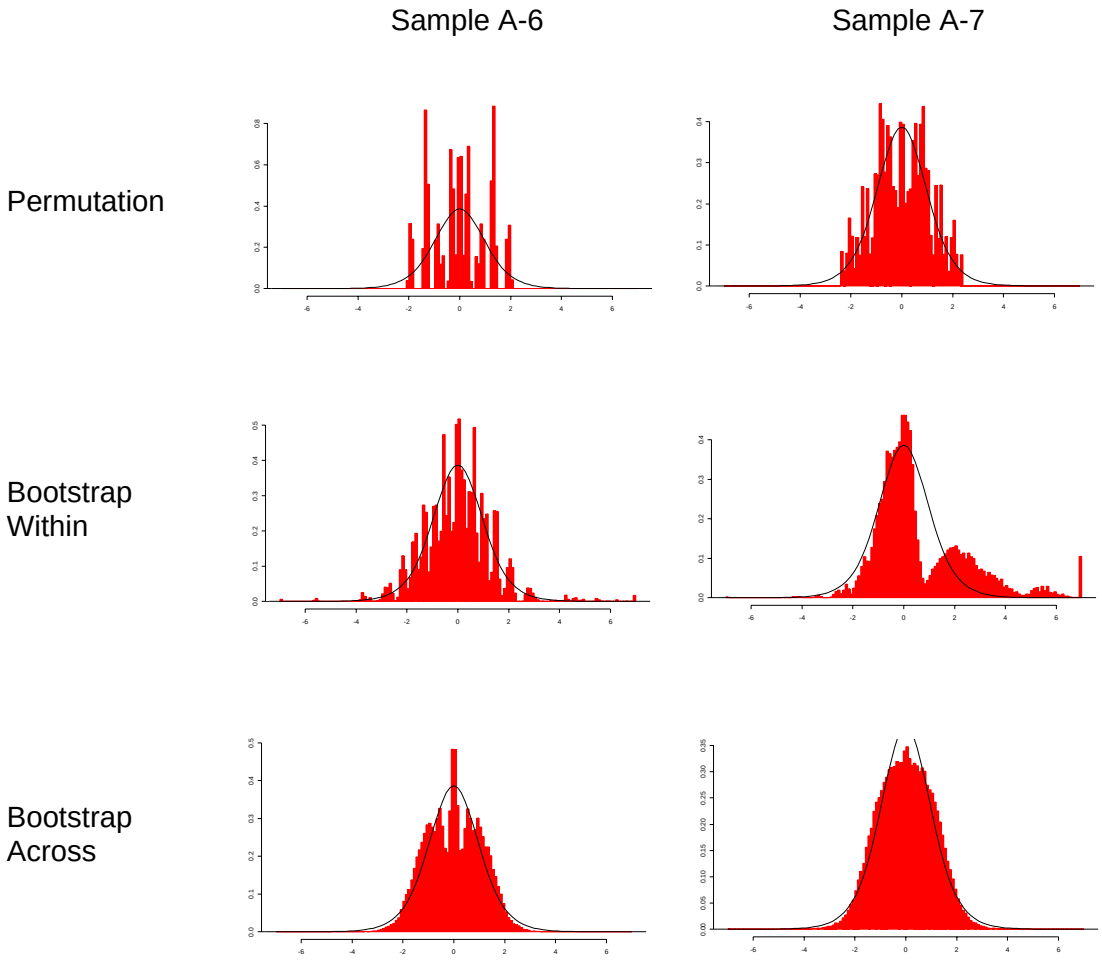

Figure A-16. Estimated Sampling Distributions Based on Samples A-6 and A-7 (B = 100,000 resamples in each). The t-distribution with df=8 is overlaid.

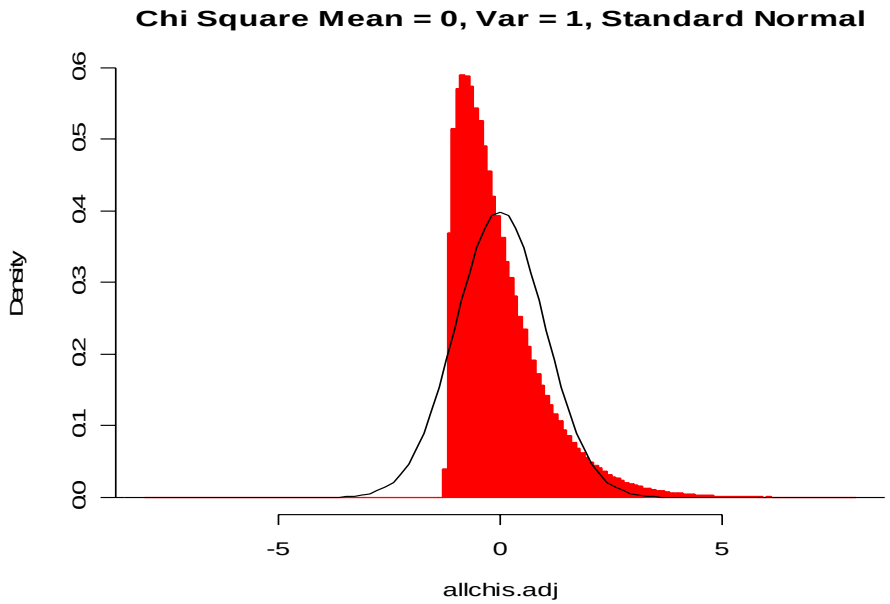

Figure A-17. Distributions Used to Form Null t Distribution: $N(0,1^2)$ and ChiSquare$(0,1^2)$. ChiSquare distribution has df = 3.

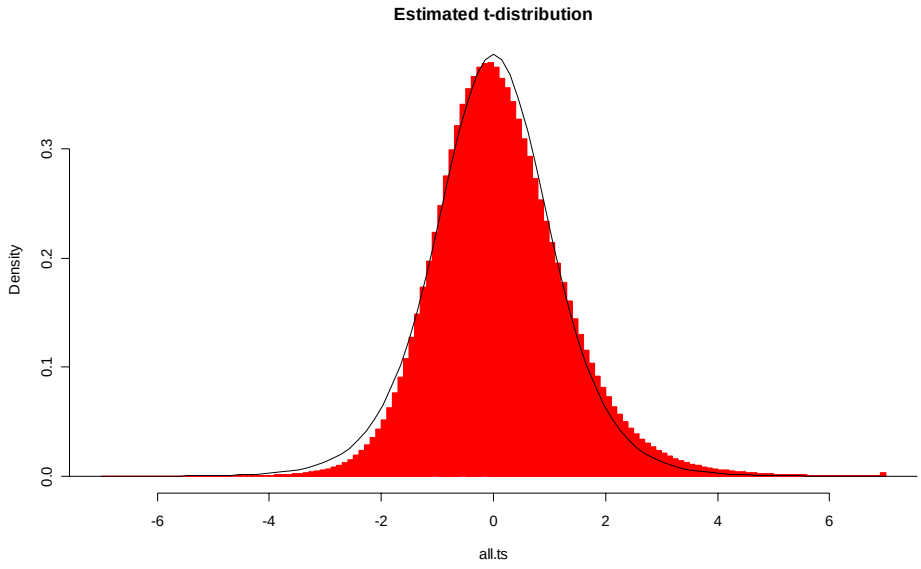

Figure A-18. Histogram is Estimated Null Distribution from 10,000,000 t-statistics from Samples of Size 5 from Distributions of Figure A-17 (the Student's t-distribution is overlaid).

Sample A-8: Two random samples of size 5, the first from a standard normal distribution, the second from a Chi-Square distribution (df = 3) shifted and scaled so that mean = 0 and variance = 1.

$\boldsymbol{y_1} = (y_{11}, y_{12}, y_{13}, y_{14}, y_{15})$ = (-0.3431115, -1.152665, 0.9054808, 0.4141999, -0.5974258)

$\boldsymbol{y_2} = (y_{21}, y_{22}, y_{23}, y_{24}, y_{25})$ = (-1.0305257, -1.075481, 0.603571, -0.6369602, -0.563531)

Sample A-9: Two random samples of size 5: the first from a standard normal distribution, the second from a Chi-Square distribution (df = 3) shifted and scaled so that mean = 0 and variance = 1.

$\boldsymbol{y_1} = (y_{11}, y_{12}, y_{13}, y_{14}, y_{15})$ = (-0.0456322, 1.6148466, -0.739835, 0.7162557, 0.7871868)

$\boldsymbol{y_2} = (y_{21}, y_{22}, y_{23}, y_{24}, y_{25})$ = (-0.3623759, -0.8840759, -0.729232, 0.727296, 0.8851749)

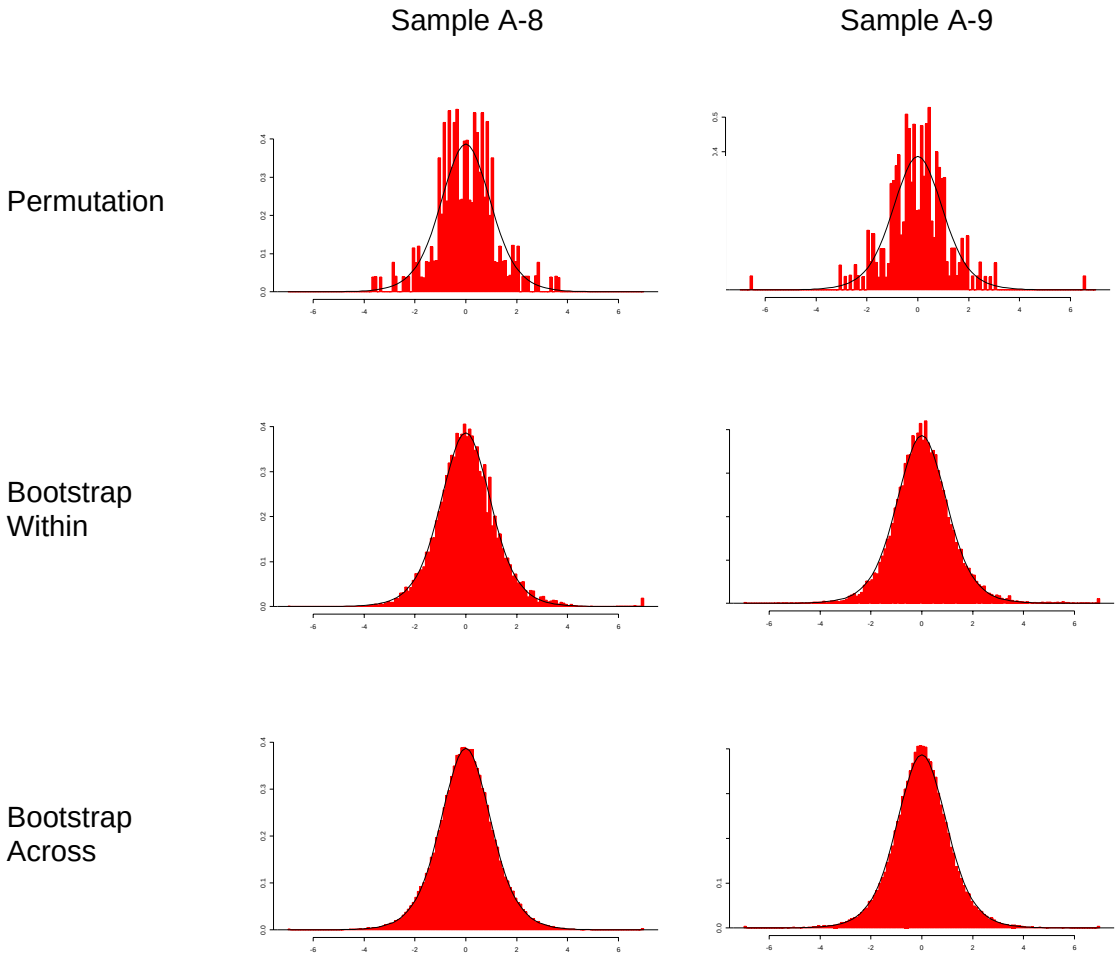

Figure A-19. Estimated Sampling Distributions Based on Samples A-8 and A-9 (B = 100,000 resamples in each). The t-distribution with df=8 is overlaid.

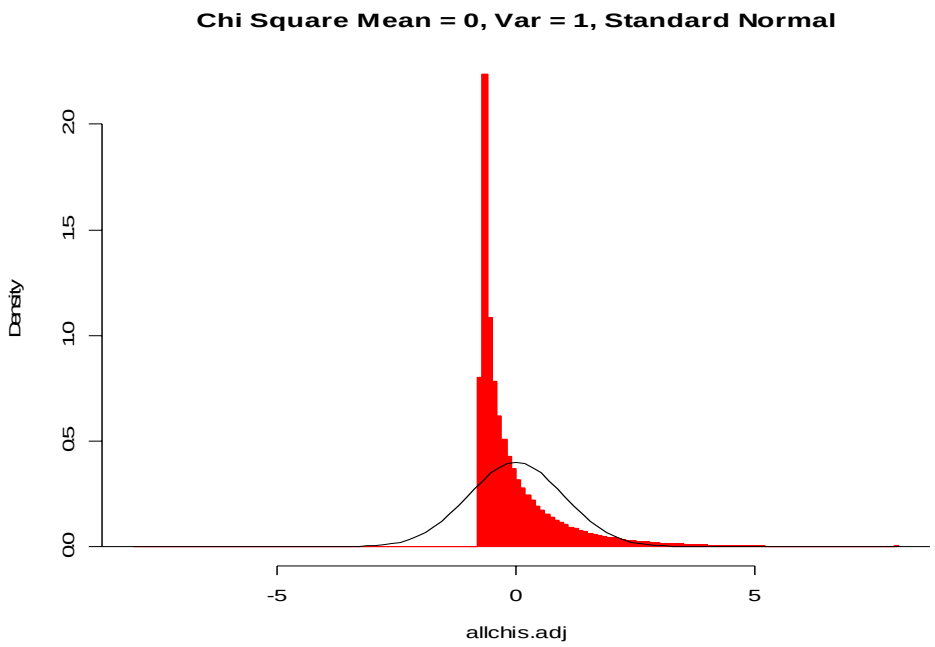

Figure A-20. Distributions Used to Form Null t Distribution: N(0,1^2) and ChiSquare(0,1^2). ChiSquare distribution has df = 1.

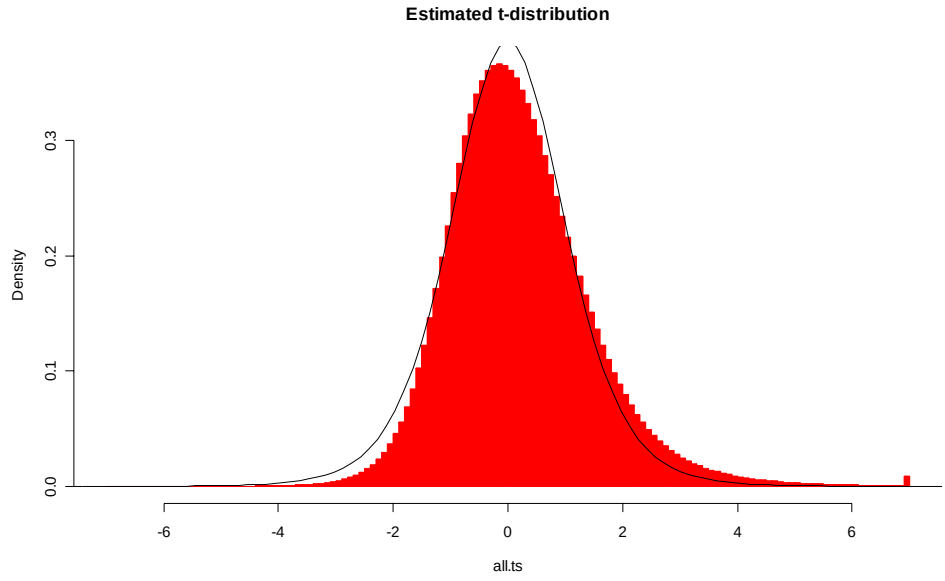

Figure A-21. Histogram is Estimated Null Distribution from 10,000,000 t-statistics from Samples of Size 5 from Distributions of Figure A-20 (the Student's t-distribution is overlaid).

Sample A-10: Two random samples of size 5, the first from a standard normal distribution, the second from a Chi-Square distribution (df = 1) shifted and scaled so that mean = 0 and variance = 1.

$\boldsymbol{y_1} = (y_{11}, y_{12}, y_{13}, y_{14}, y_{15})$ = (1.0819834, 0.045843, -1.3361416, -0.9334126, 0.5467037)

$\boldsymbol{y_2} = (y_{21}, y_{22}, y_{23}, y_{24}, y_{25})$ = (0.43116, -0.3004396, 5.2929116, -0.2470618, 1.2523760)

Sample A-11: Two random samples of size 5, the first from a standard normal distribution, the second from a Chi-Square distribution (df = 1) shifted and scaled so that mean = 0 and variance = 1)

$\boldsymbol{y_1} = (y_{11}, y_{12}, y_{13}, y_{14}, y_{15})$ = (0.1007925, 0.9009849, -1.599445, -1.2884828, -1.1696947)

$\boldsymbol{y_2} = (y_{21}, y_{22}, y_{23}, y_{24}, y_{25})$ = (-0.51345, 0.0391394, -0.4469858, -0.704290, -0.1865331)

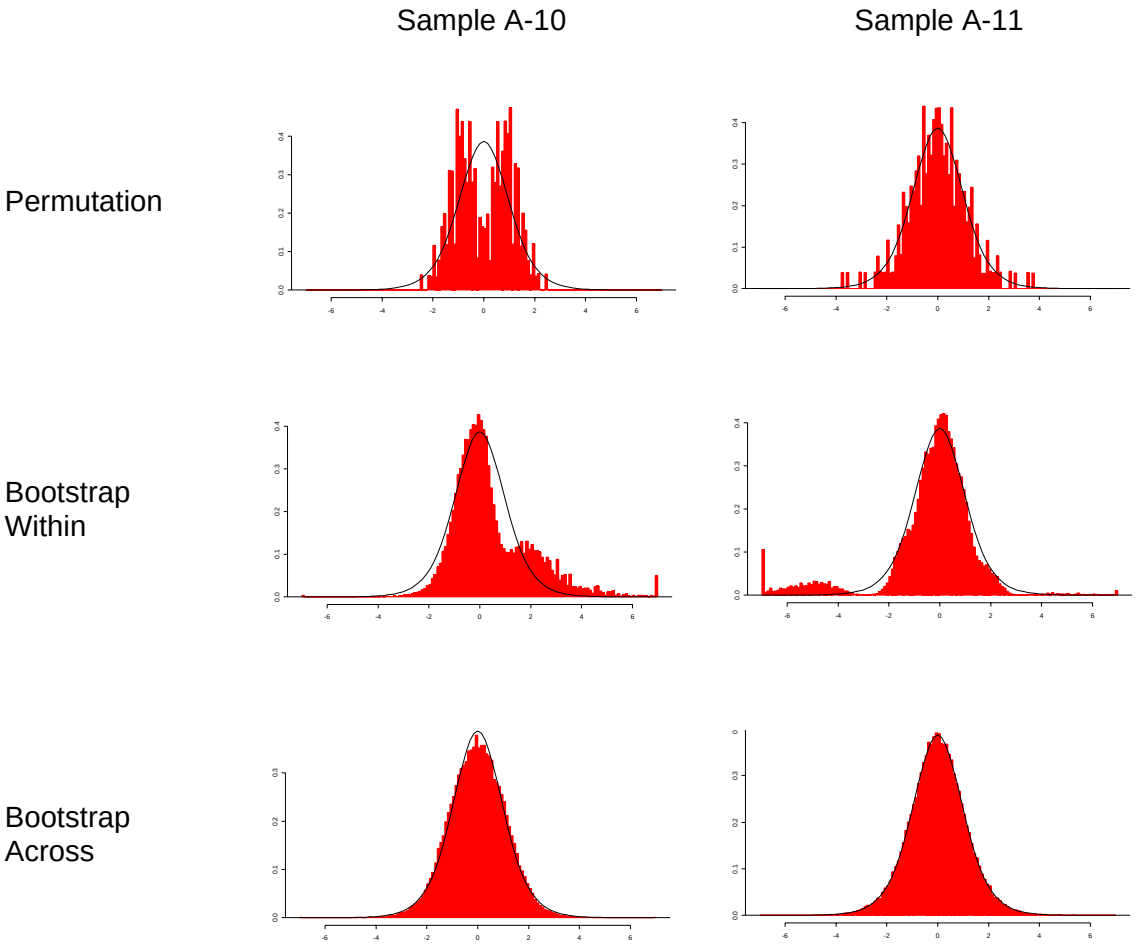

Figure A-22. Estimated Sampling Distributions Based on Samples A-10 and A-11 (B = 100,000 resamples in each). The t-distribution with df=8 is overlaid.

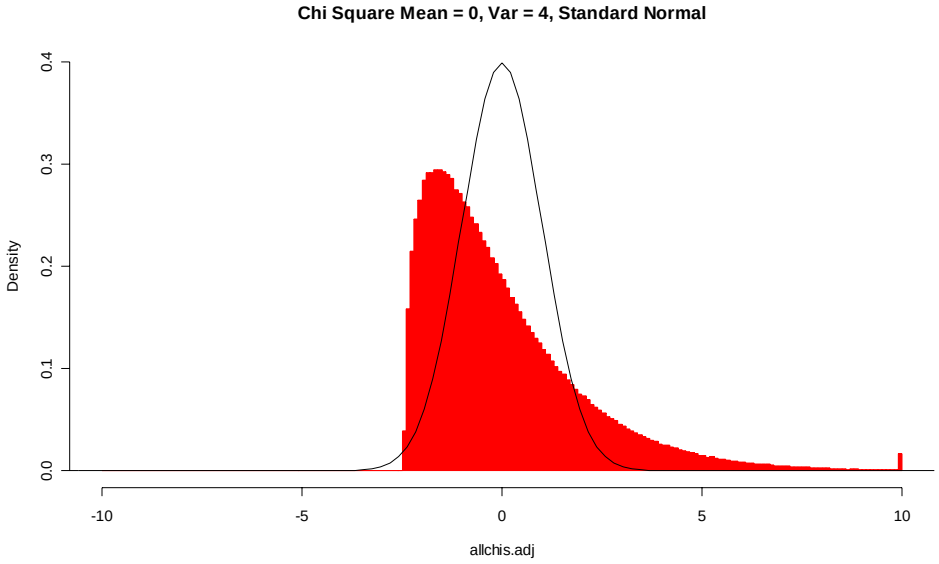

Figure A-23. Distributions Used to Form Null t Distribution: $N(0,1^2)$ and ChiSquare$(0,2^2)$. ChiSquare distribution has df = 3.

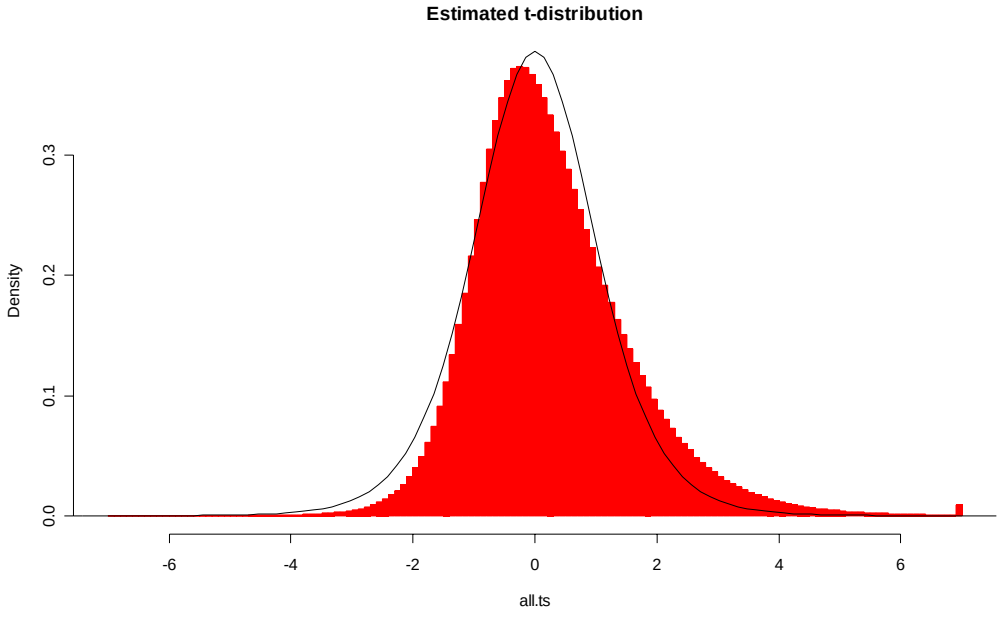

Figure A-24. Histogram is Estimated Null Distribution from 10,000,000 t-statistics from Samples of Size 5 from Distributions of Figure A-23 (the Student's t-distribution is overlaid).

Sample A-12: Two random samples of size 5, the first from a standard normal distribution, the second from a Chi-Square distribution (df = 1) shifted and scaled so that mean = 0 and variance = 4.

$\boldsymbol{y_1} = (y_{11}, y_{12}, y_{13}, y_{14}, y_{15})$ = (-1.721818, -0.988713, -0.00707245, -0.1294487, 0.886869)

$\boldsymbol{y_2} = (y_{21}, y_{22}, y_{23}, y_{24}, y_{25})$ = (-1.463912, 1.073482, 2.7693038, -0.9923066, -1.1085539)

Sample A-13: Two random samples of size 5: the first from a standard normal distribution, the second from a Chi-Square distribution (df = 1) shifted and scaled so that mean = 0 and variance = 4.

$\boldsymbol{y_1} = (y_{11}, y_{12}, y_{13}, y_{14}, y_{15})$ = (-0.9378573, -0.4159593, 0.5113173, 1.279988, -1.1341501)

$\boldsymbol{y_2} = (y_{21}, y_{22}, y_{23}, y_{24}, y_{25})$ = (2.2997002, -1.715483, 0.459342, -2.0425536, -1.6940275)

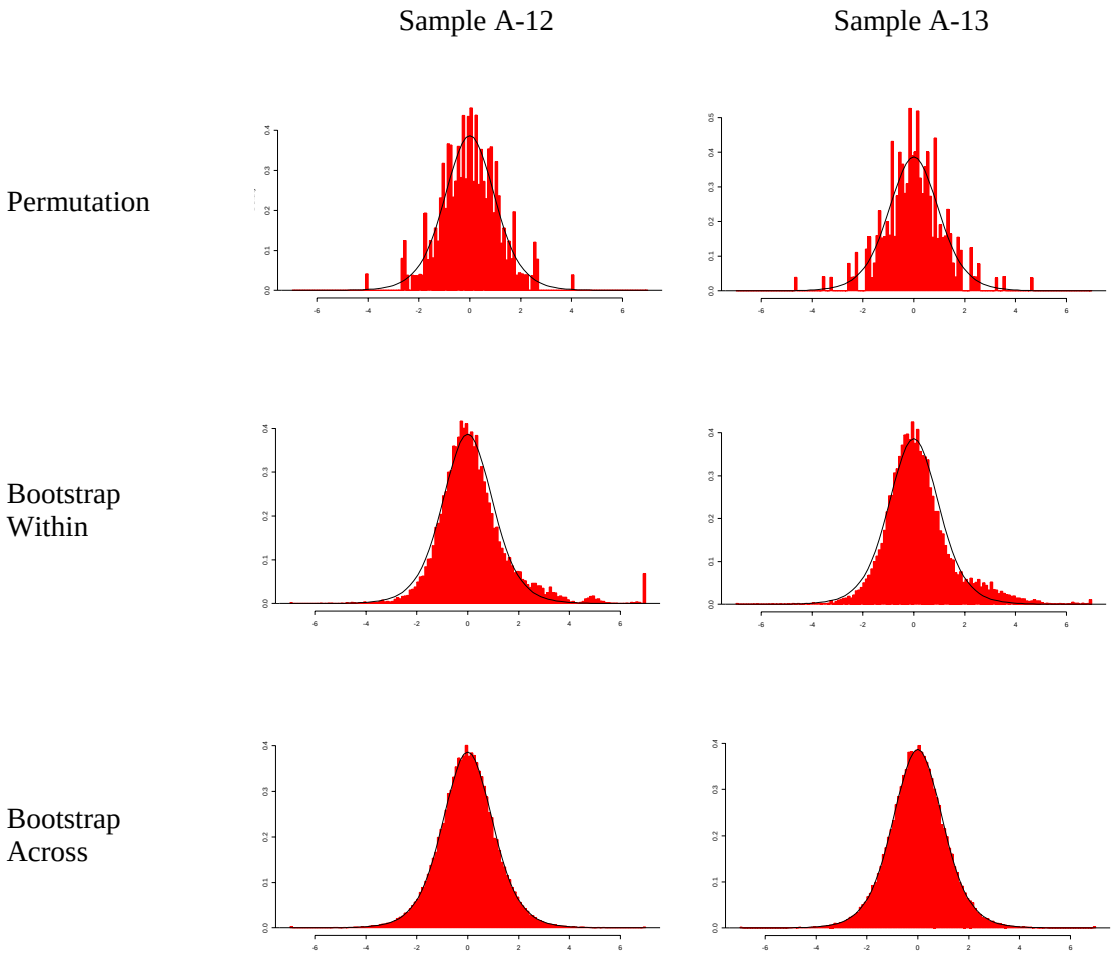

Figure A-25. Estimated Sampling Distributions Based on Samples A-12 and A-13 (B = 100,000 resamples in each). The t-distribution with df=8 is overlaid.

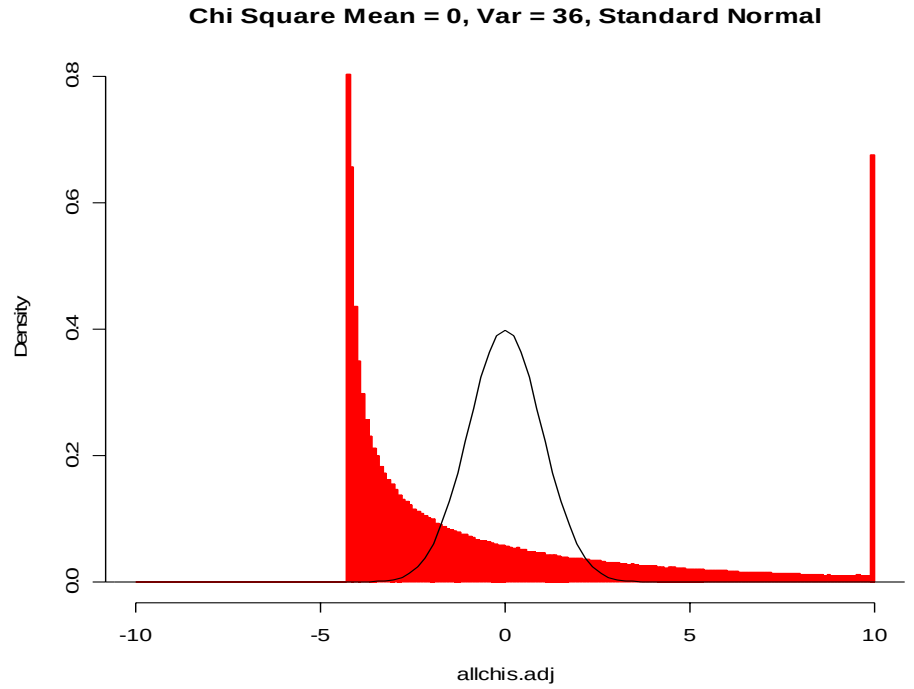

Figure A-26. Distributions Used to Form Null t distribution: $N(0,1^2)$ and ChiSquare$(0,6^2)$. ChiSquare distribution has df = 1. Values greater than 10 are condensed to a single bin at 10.

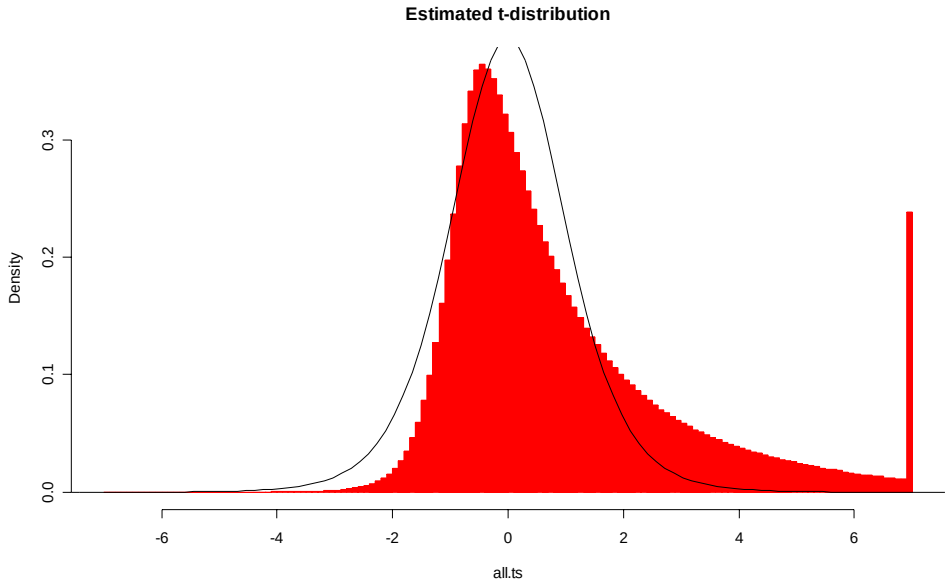

Figure A-27. Histogram is Estimated Null Distribution from 10,000,000 t-statistics from Samples of Size 5 from Distributions of Figure A-26 (the Student's t-distribution is overlaid). Values greater than 7 are condensed to a single bin at 7.

Sample A-14: Two random samples of size 5, the first from a standard normal distribution, the second from a Chi-Square distribution (df = 1) shifted and scaled so that mean = 0 and variance = 36.

$\boldsymbol{y_1} = (y_{11}, y_{12}, y_{13}, y_{14}, y_{15})$ = (1.4822938, 1.858962, -0.409045, 0.0976272, 0.64474533)

$\boldsymbol{y_2} = (y_{21}, y_{22}, y_{23}, y_{24}, y_{25})$ = (-1.988260, 4.929872, 4.236700, 1.674328, 3.915160)

Sample A-15: Two random samples of size 5: the first from a standard normal distribution, the second from a Chi-Square distribution (df = 1) shifted and scaled so that mean = 0 and variance = 36.

$\boldsymbol{y_1} = (y_{11}, y_{12}, y_{13}, y_{14}, y_{15})$ = (-1.160121, -0.3176905, 1.1170616, -1.880935, -0.8355052)

$\boldsymbol{y_2} = (y_{21}, y_{22}, y_{23}, y_{24}, y_{25})$ = (-1.289370, 9.665552, -2.677599, -4.173899, -3.996047)

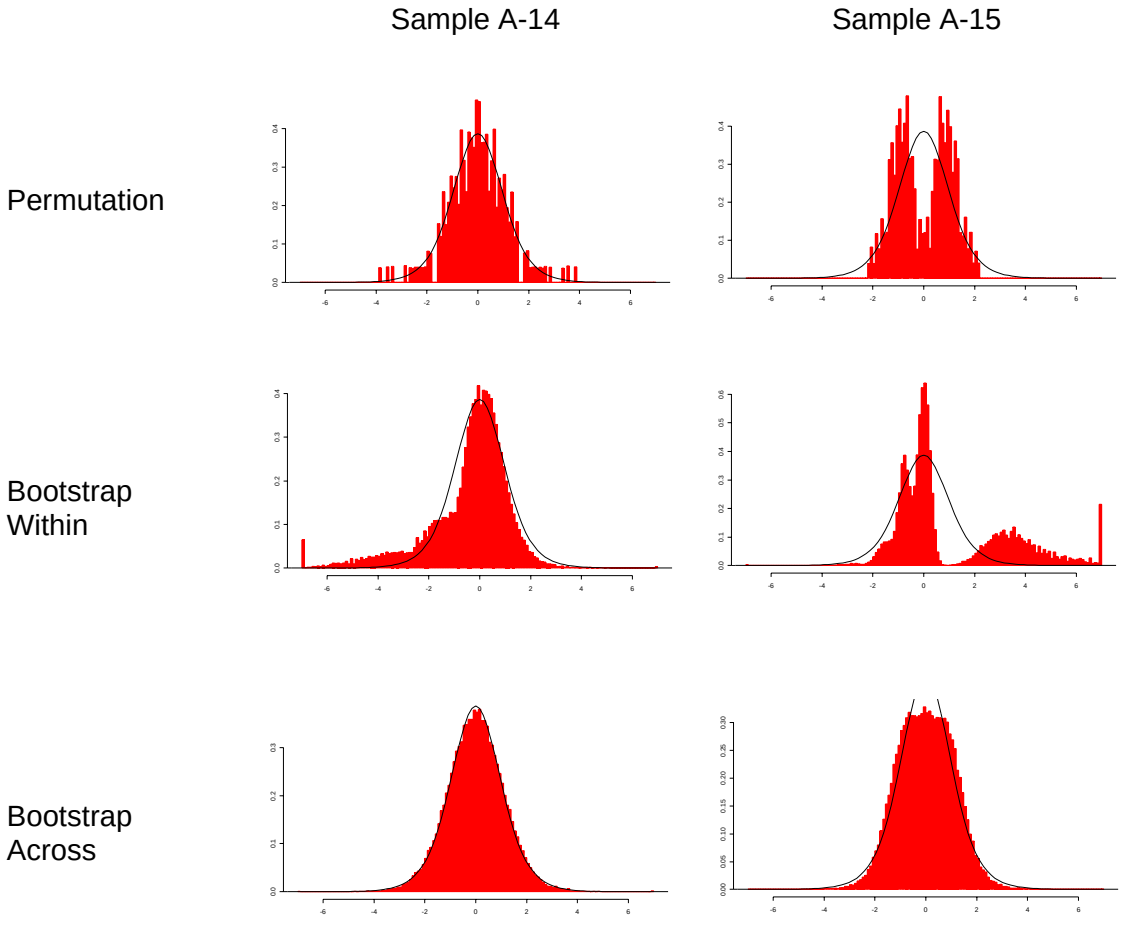

Figure A-28. Estimated Sampling Distributions Based on Samples A-14 and A-15 (B = 100,000 resamples in each). The t-distribution with df=8 is overlaid.

APPENDIX B

COMPARISON OF ESTIMATED ALPHA

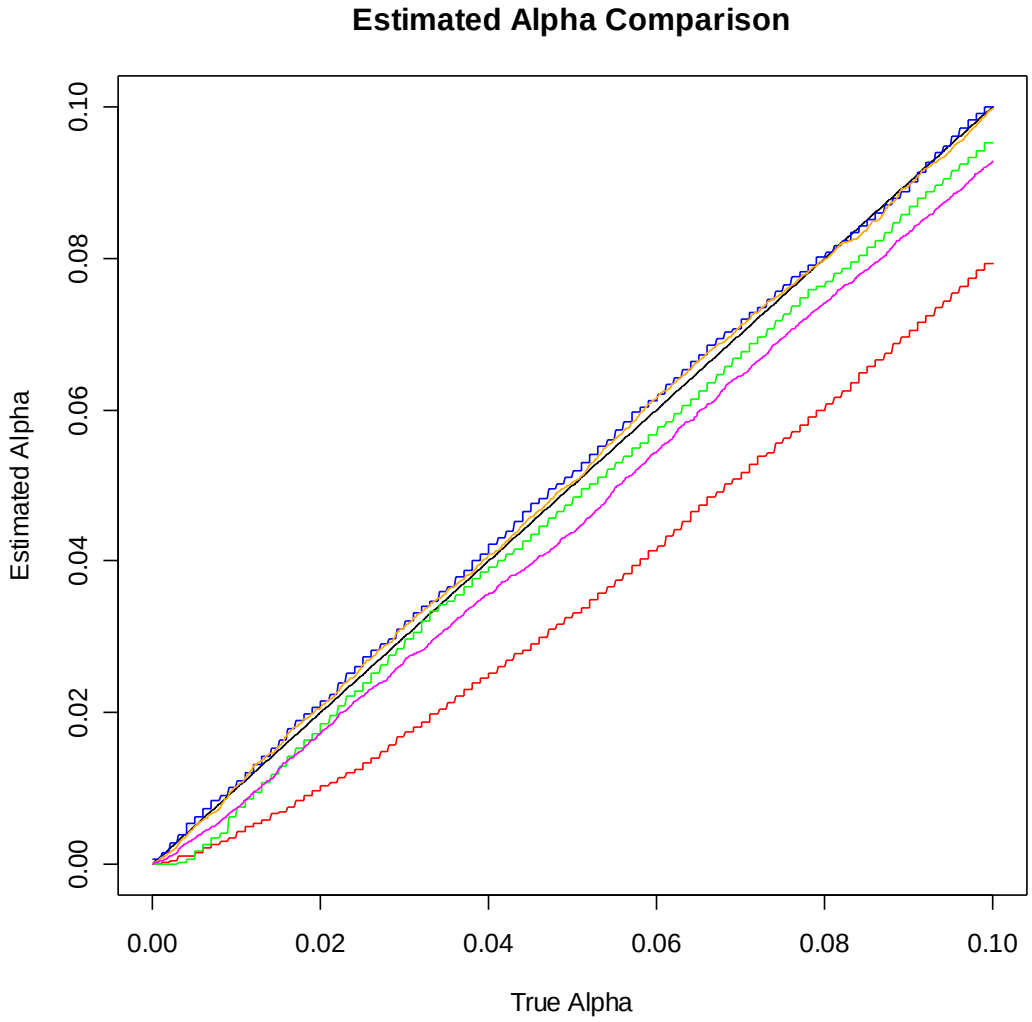

Figure B-1. Comparison of Estimated Alpha. $\mathrm{ASL} = \#(|\boldsymbol{t}^*| \geq |t|)/B$, $n_1 = n_2$ *= 5, both sampled distributions standard normal, B = 999. Orange: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap; Black: Exact Level Reference. Based on 20,000 simulated data sets the standard error of the estimated* α *is 0.00212 for* α *= 0.10, 0.00154 for* α *= 0.05, and 0.00070 for* α *= 0.01.*

Table B-1. Specific Values of Estimated α *from Figure B-1.*

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.00625	0.0035	0.01005	0.01025	0.00735
0.05	0.04740	0.0325	0.05110	0.05045	0.04375
0.10	0.09515	0.0794	0.09990	0.09965	0.09280

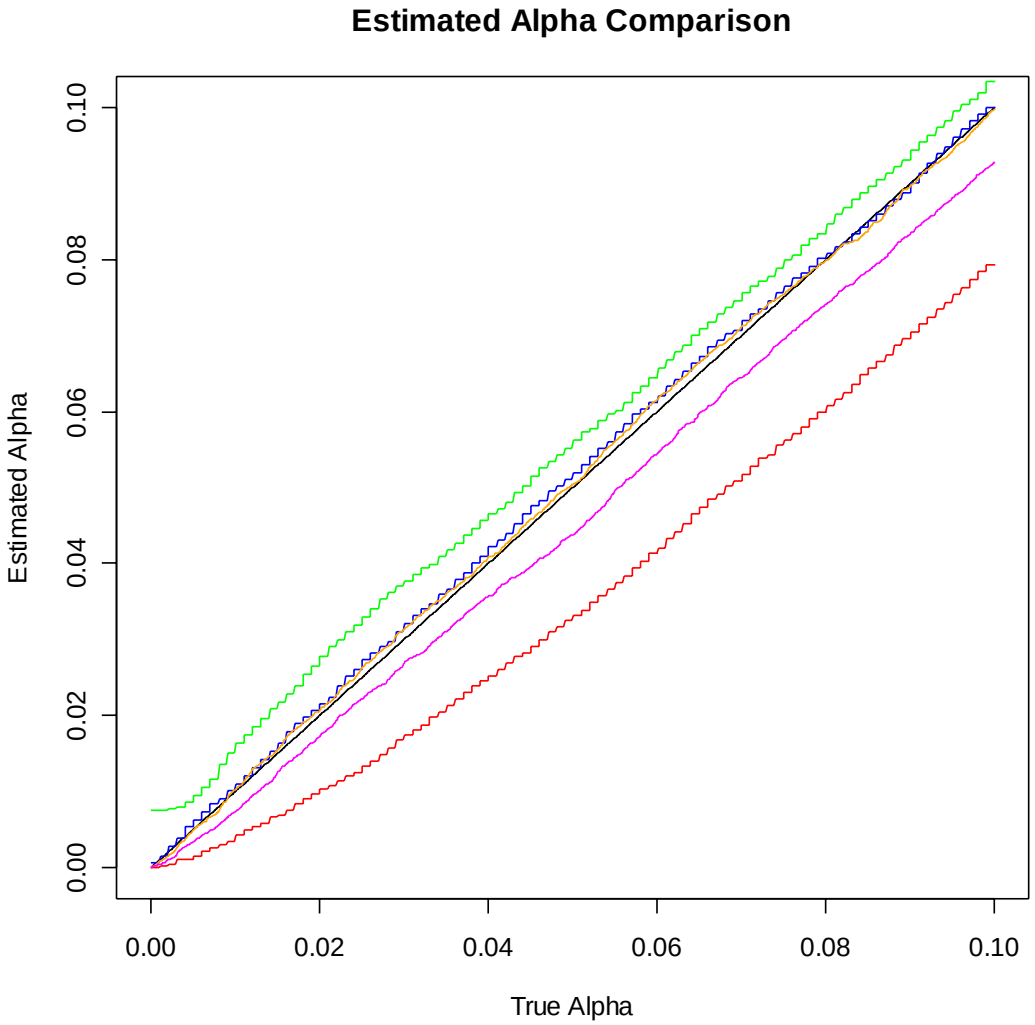

Figure B-2. Comparison of Estimated Alpha. $\mathrm{ASL} = \#(|\boldsymbol{t}^*| > |t|)/B$, $n_1 = n_2$ = 5, *both sampled distributions standard normal, B = 999. Orange: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap; Black: Exact Level Reference. Based on 20,000 simulated data sets the standard error of the estimated* α *is 0.00212 for* α = *0.10, 0.00154 for* α = *0.05, and 0.00070 for* α = *0.01.*

Table B-2. Specific Values of Estimated α *from Figure B-2.*

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.01515	0.0035	0.01005	0.01025	0.00735
0.05	0.05525	0.0325	0.0511	0.05045	0.04375
0.10	0.10345	0.0794	0.0999	0.09965	0.0928

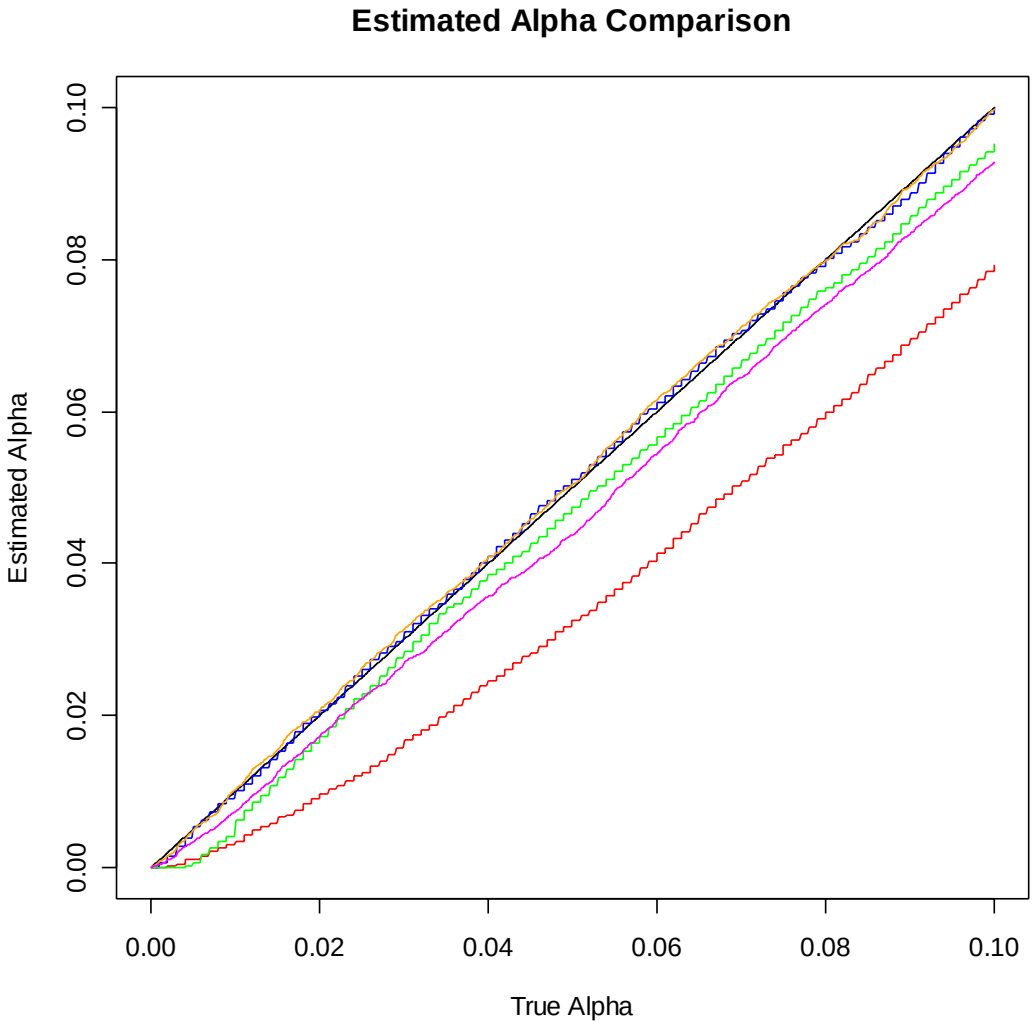

Figure B-3. Comparison of Estimated Alpha. $\mathrm{ASL} = (1+\#(|\boldsymbol{t}^*| \geq |t|))/(1+B)$, $n_1 = n_2$ = *5, both sampled distributions standard normal, B = 999. Orange: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap; Black: Exact Level Reference. Based on 20,000 simulated data sets the standard error of the estimated* α *is 0.00212 for* α *= 0.10, 0.00154 for* α *= 0.05, and 0.00070 for* α *= 0.01.*

Table B-3. Specific Values of Estimated α *from Figure B-3.*

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.00625	0.0035	0.01005	0.01025	0.00735
0.05	0.0474	0.0325	0.0511	0.05045	0.04375
0.10	0.09515	0.0794	0.0999	0.09965	0.0928

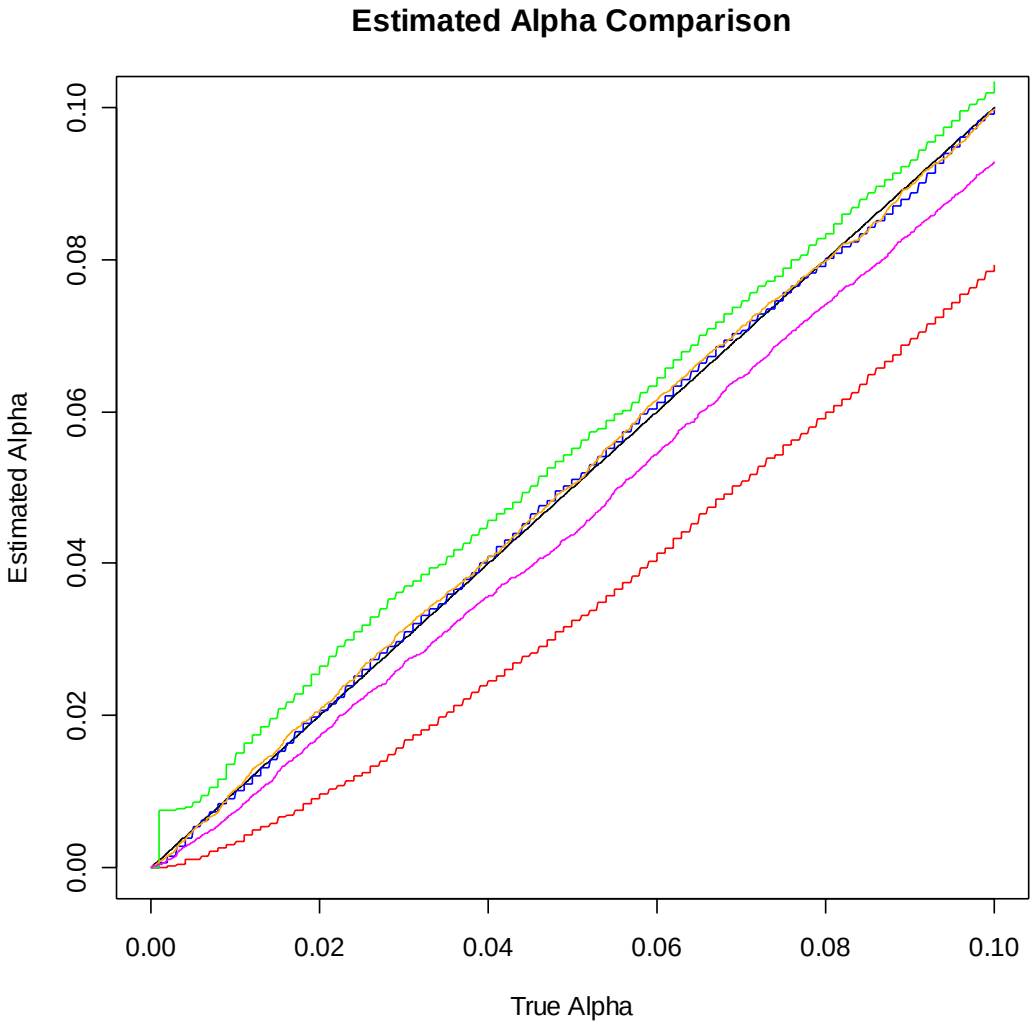

Figure B-4. Comparison of Estimated Alpha. $\mathrm{ASL} = (1+\#(|\boldsymbol{t}^*| > |t|))/(1+B)$, $n_1 = n_2$ = *5, both sampled distributions standard normal, B = 999. Orange: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap; Black: Exact Level Reference. Based on 20,000 simulated data sets the standard error of the estimated* α *is 0.00212 for* α *= 0.10, 0.00154 for* α *= 0.05, and 0.00070 for* α *= 0.01.*

Table B-4. Specific Values of Estimated α *from Figure B-4.*

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.01515	0.0035	0.01005	0.01025	0.00735
0.05	0.05525	0.0325	0.0511	0.05045	0.04375
0.10	0.10345	0.0794	0.0999	0.09965	0.0928

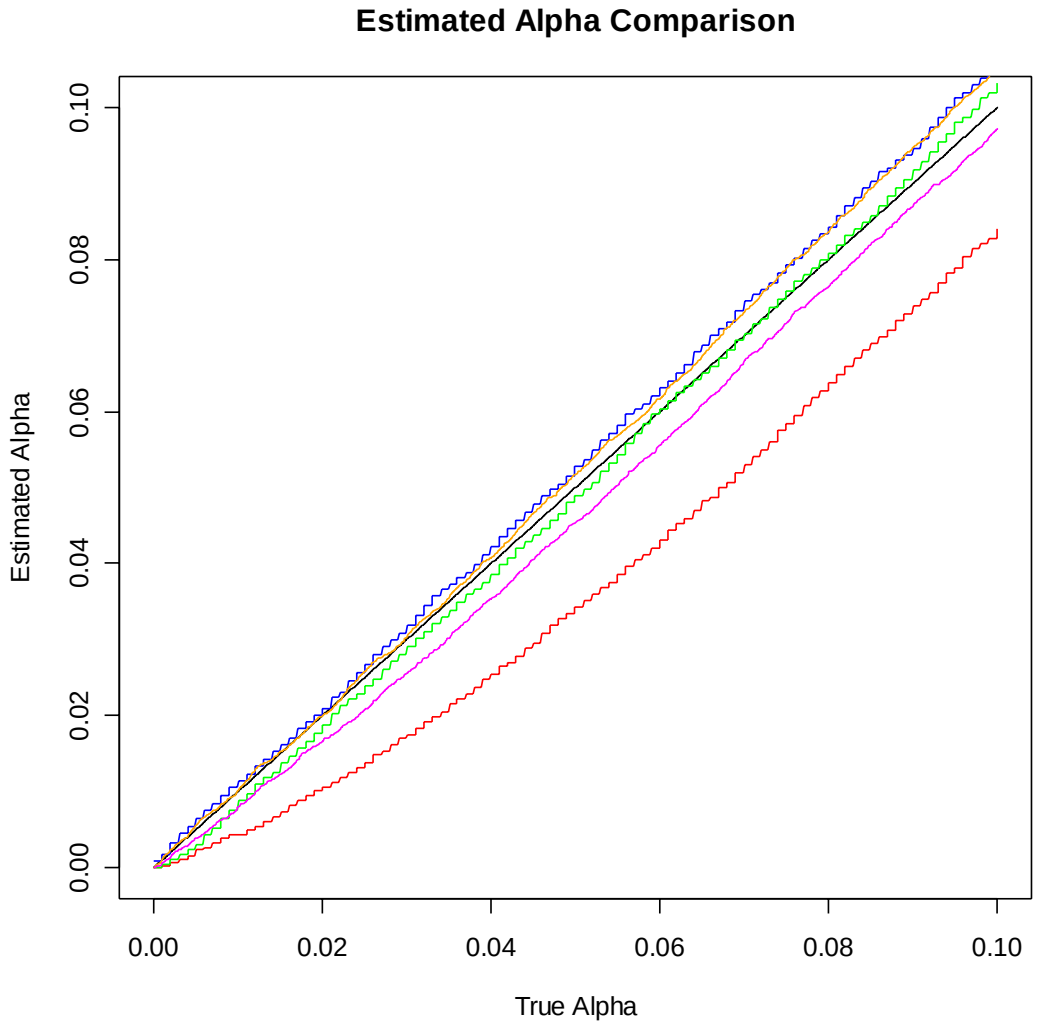

Figure B-5. Comparison of Estimated Alpha. $\mathrm{ASL} = \#(|t^*| \geq |t|)/B$, $n_1 = n_2$ *= 5, both sampled distributions standard normal, B = 1,000. Orange: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap; Black: Exact Level Reference. Based on 20,000 simulated data sets the standard error of the estimated* α *is 0.00212 for* α *= 0.10, 0.00154 for* α *= 0.05, and 0.00070 for* α *= 0.01.*

Table B-5. Specific Values of Estimated α *from Figure B-5.*

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.00895	0.00435	0.01155	0.01005	0.0079
0.05	0.04885	0.03435	0.0528	0.0518	0.04535
0.10	0.10315	0.08395	0.10585	0.1052	0.0972

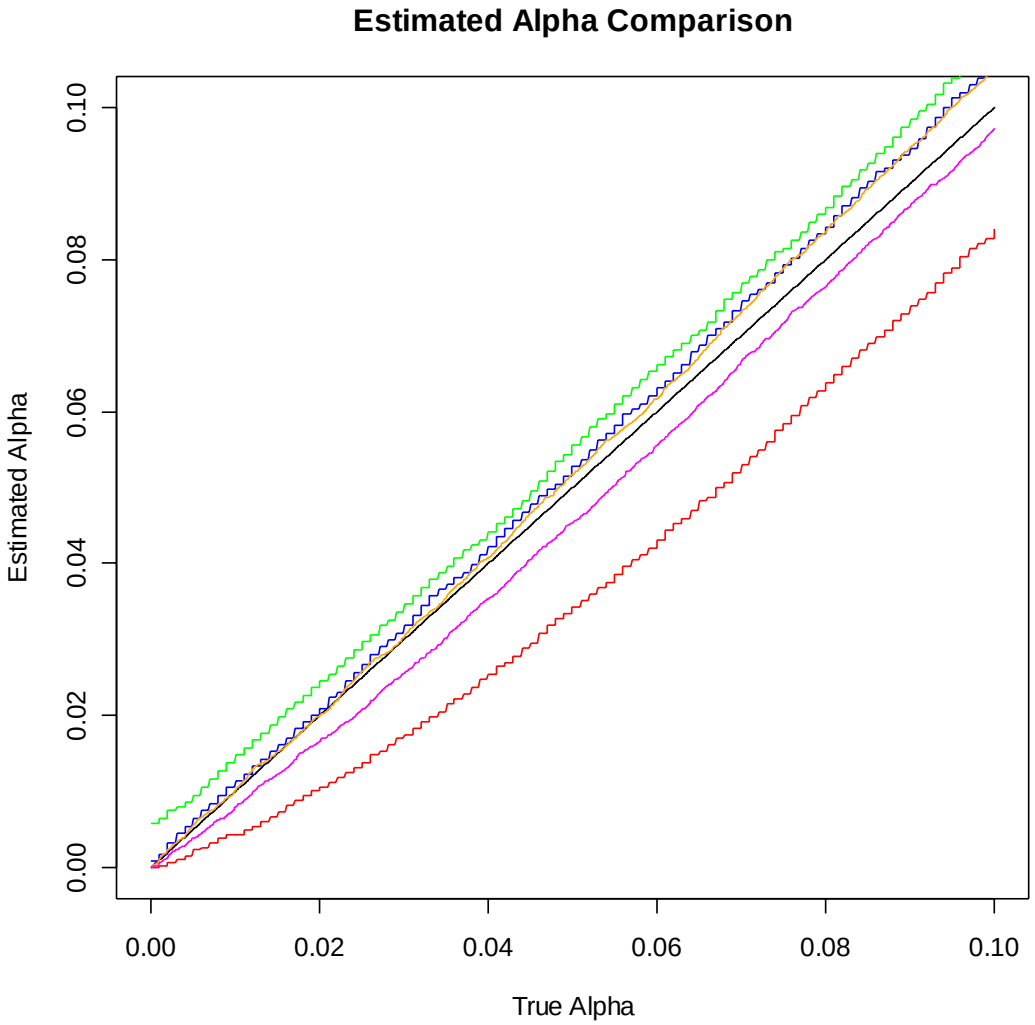

Figure B-6. Comparison of Estimated Alpha. $\mathrm{ASL} = \#(|\boldsymbol{t}^*| > |t|)/B$, $n_1 = n_2$ *= 5, both sampled distributions standard normal, B = 1,000. Orange: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap; Black: Exact Level Reference. Based on 20,000 simulated data sets the standard error of the estimated* α *is 0.00212 for* α *= 0.10, 0.00154 for* α *= 0.05, and 0.00070 for* α *= 0.01.*

Table B-6. Specific Values of Estimated α *from Figure B-6.*

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.01485	0.00435	0.01155	0.01005	0.0079
0.05	0.0555	0.03435	0.0528	0.0518	0.04535
0.10	0.10925	0.08395	0.10585	0.1052	0.0972

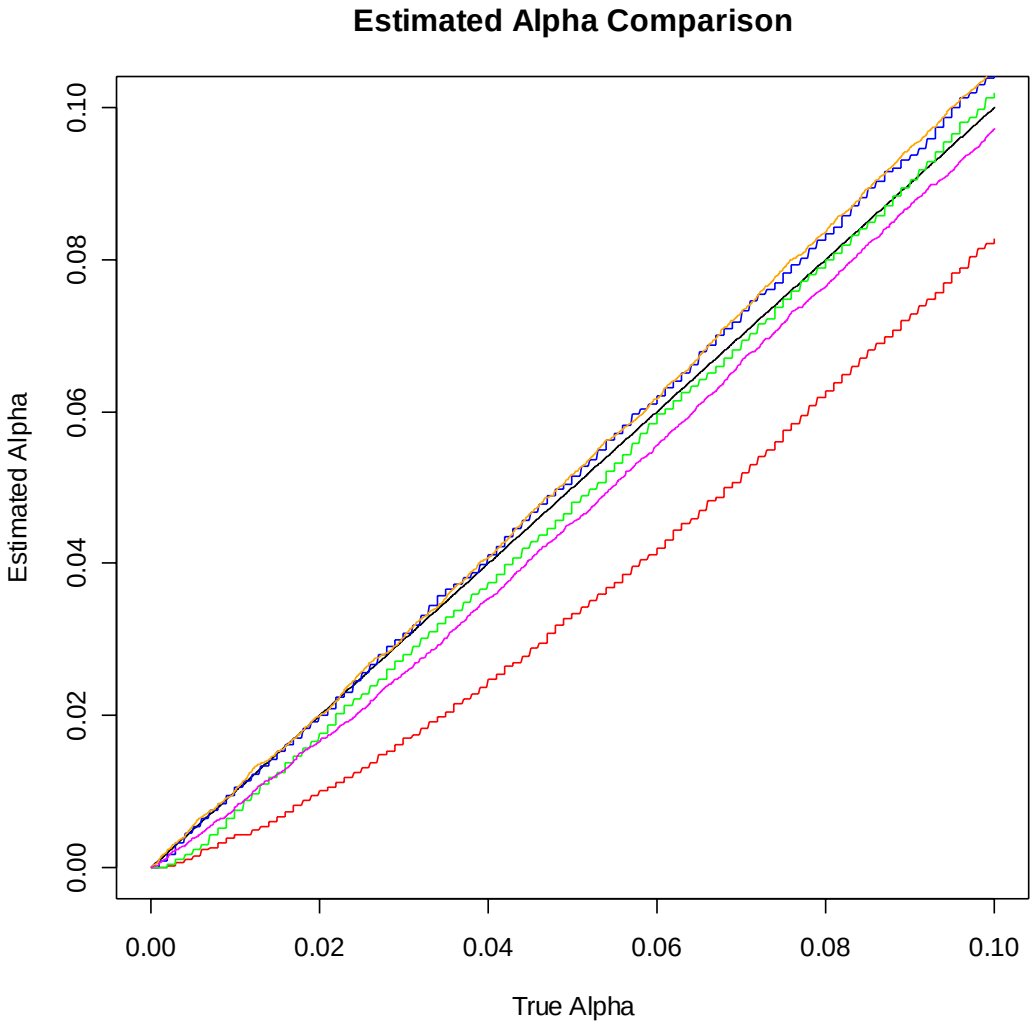

Figure B-7. Comparison of Estimated Alpha. $\mathrm{ASL} = (1 + \#(|\boldsymbol{t}^*| \geq |t|))/(1+B)$, $n_1 = n_2$ = *5, both sampled distributions standard normal, B = 1,000. Orange: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap; Black: Exact Level Reference. Based on 20,000 simulated data sets the standard error of the estimated* α *is 0.00212 for* α *= 0.10, 0.00154 for* α *= 0.05, and 0.00070 for* α *= 0.01.*

Table 7. Specific Values of Estimated α *from Figure B-7.*

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.00765	0.00425	0.01055	0.01005	0.0079
0.05	0.048	0.03345	0.0516	0.0518	0.04535
0.10	0.10195	0.0828	0.1049	0.1052	0.0972

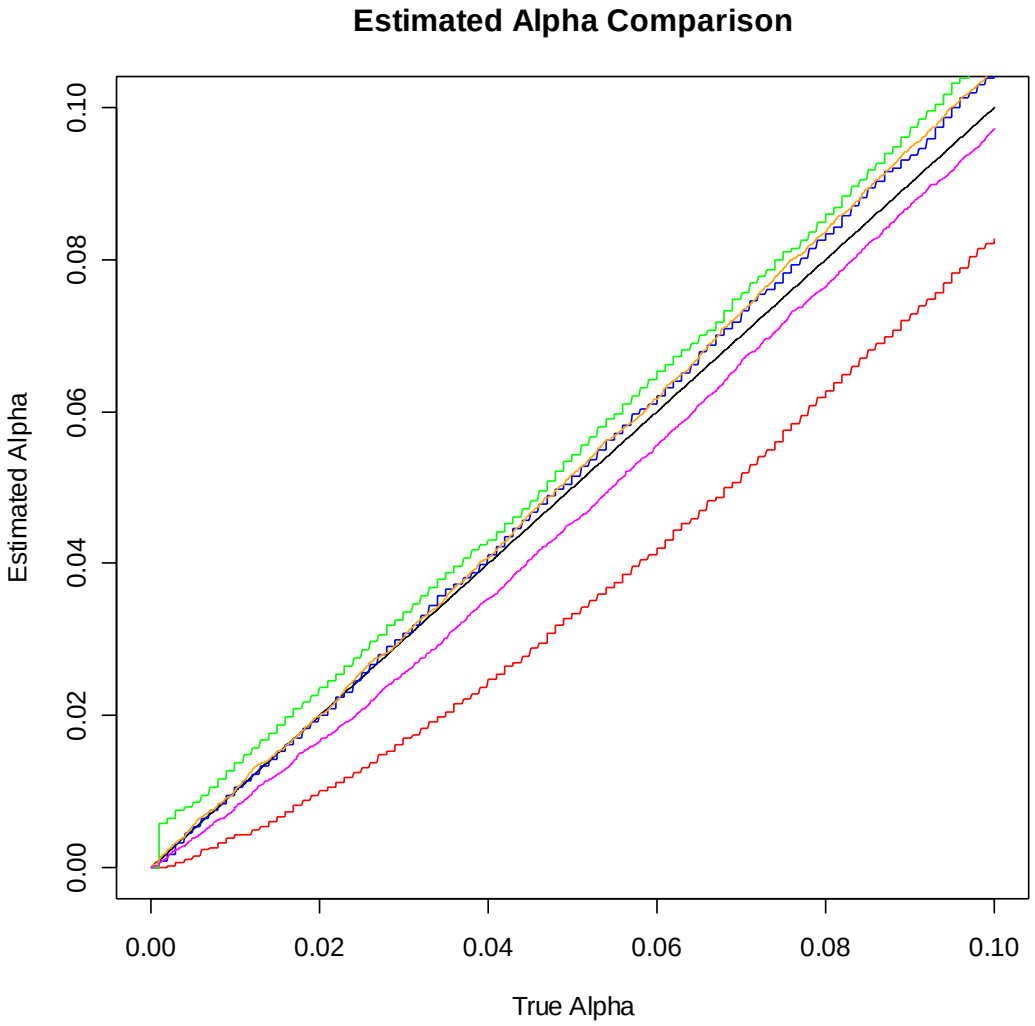

Figure B-8. Comparison of Estimated Alpha. $\mathrm{ASL} = (1 + \#(|\boldsymbol{t}^*| > |t|))/(1 + B)$, $n_1 = n_2$ = *5, both sampled distributions standard normal, B = 1,000. Orange: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap; Black: Exact Level Reference. Based on 20,000 simulated data sets the standard error of the estimated* α *is 0.00212 for* α *= 0.10, 0.00154 for* α *= 0.05, and 0.00070 for* α *= 0.01.*

Table B-8. Specific Values of Estimated α *from Figure B-8.*

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.01375	0.00425	0.01055	0.01005	0.0079
0.05	0.0544	0.03345	0.0516	0.0518	0.04535
0.10	0.1084	0.0828	0.1049	0.1052	0.0972

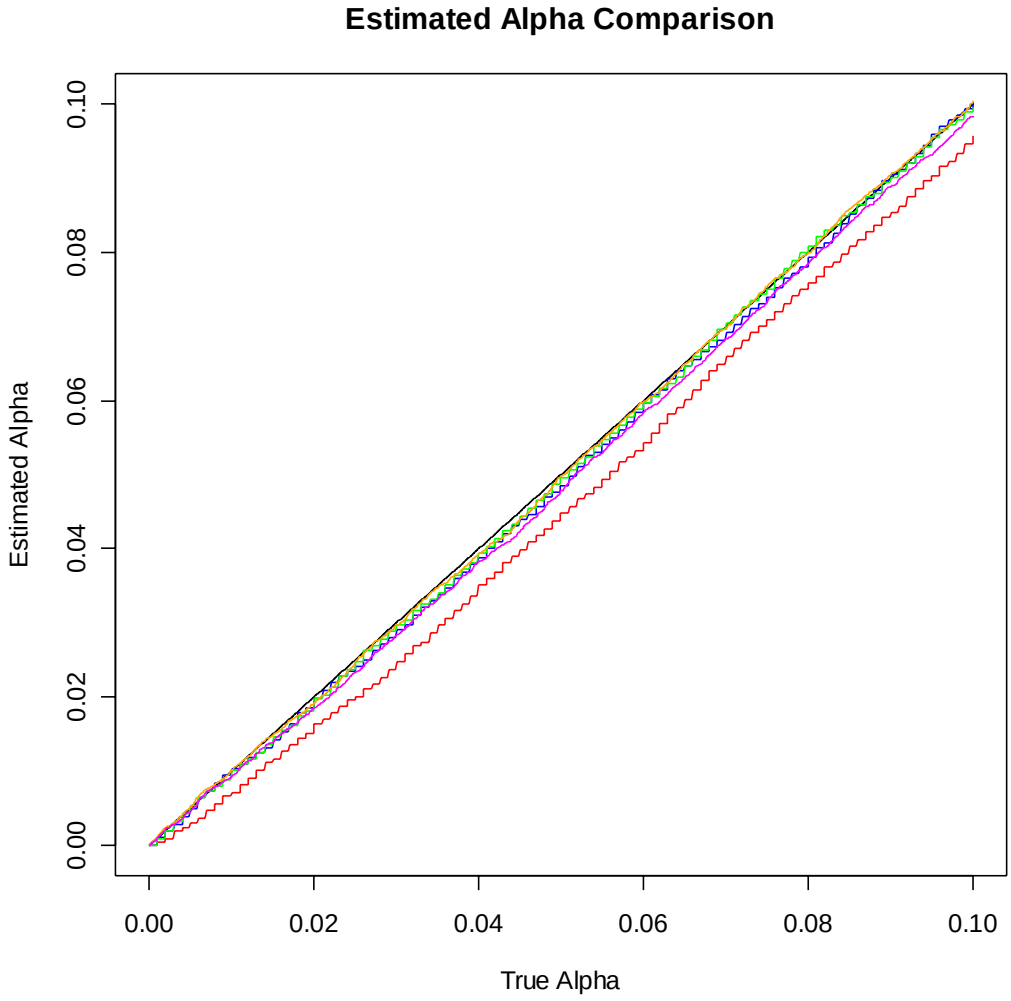

Figure B-9. Comparison of Estimated Alpha. $\mathrm{ASL} = (1 + \#(|t^*| \geq |t|))/(1+B)$, $n_1 = n_2$ = *10, both sampled distributions standard normal, B = 999. Orange: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap; Black: Exact Level Reference. Based on 20,000 simulated data sets the standard error of the estimated* α *is 0.00212 for* α *= 0.10, 0.00154 for* α *= 0.05, and 0.00070 for* α *= 0.01.*

Table B-9. Specific Values of Estimated α *from Figure B-9.*

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.01005	0.00715	0.01030	0.01005	0.00925
0.05	0.04950	0.04480	0.04850	0.04970	0.04775
0.10	0.09950	0.09560	0.10025	0.10030	0.09835

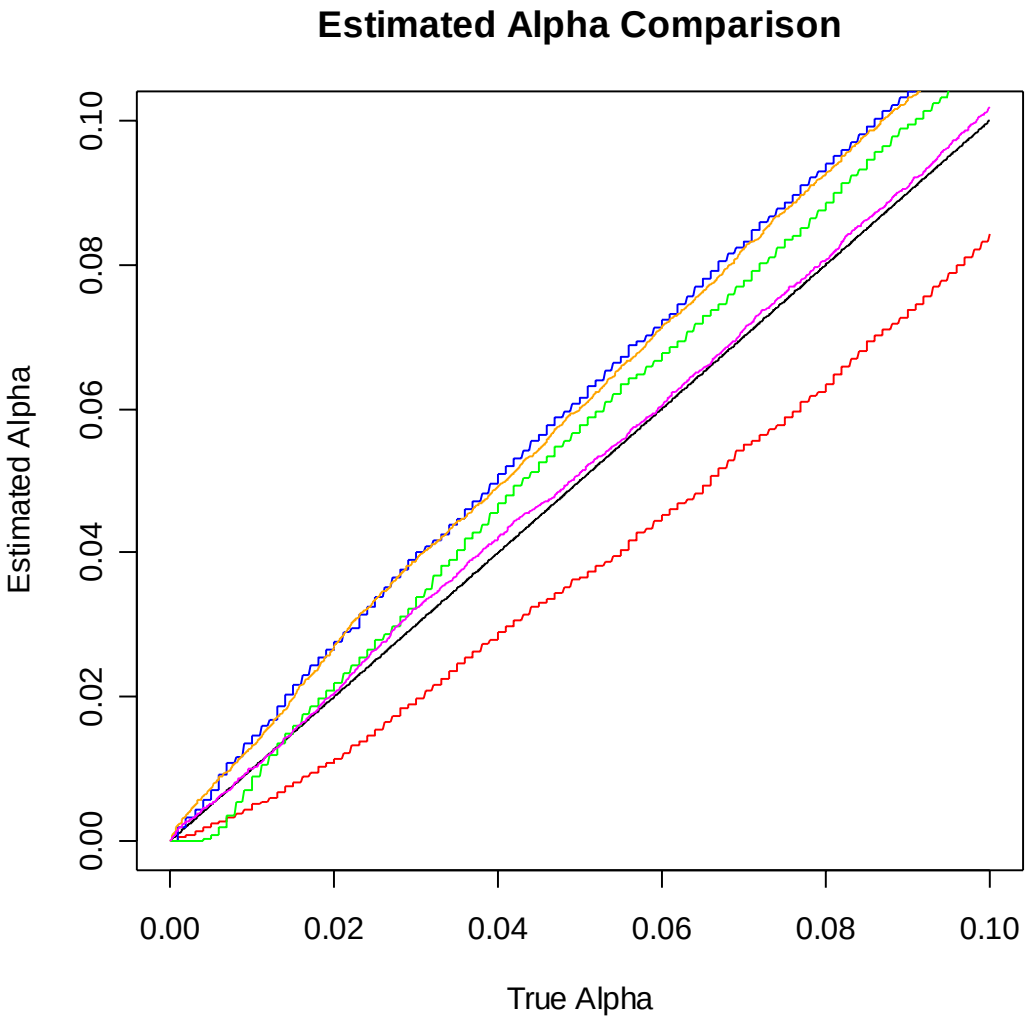

Figure B-10. Comparison of Estimated Alpha. $\text{ASL} = (1 + \#(|\boldsymbol{t}^*| \geq |t|))/(1+B)$, $n_1 = n_2$ = 5, *sampled distributions:* $N(0,1^2)$, $N(0,2^2)$, *B = 999. Orange: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap; Black: Exact Level Reference. Based on 20,000 simulated data sets the standard error of the estimated* α *is 0.00212 for* α = *0.10, 0.00154 for* α = *0.05, and 0.00070 for* α = *0.01.*

Table B-10. Specific Values of Estimated α *from Figure B-10.*

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.00885	0.005	0.0147	0.0132	0.01005
0.05	0.0576	0.0365	0.06165	0.06005	0.05105
0.10	0.1095	0.08425	0.11475	0.11345	0.1019

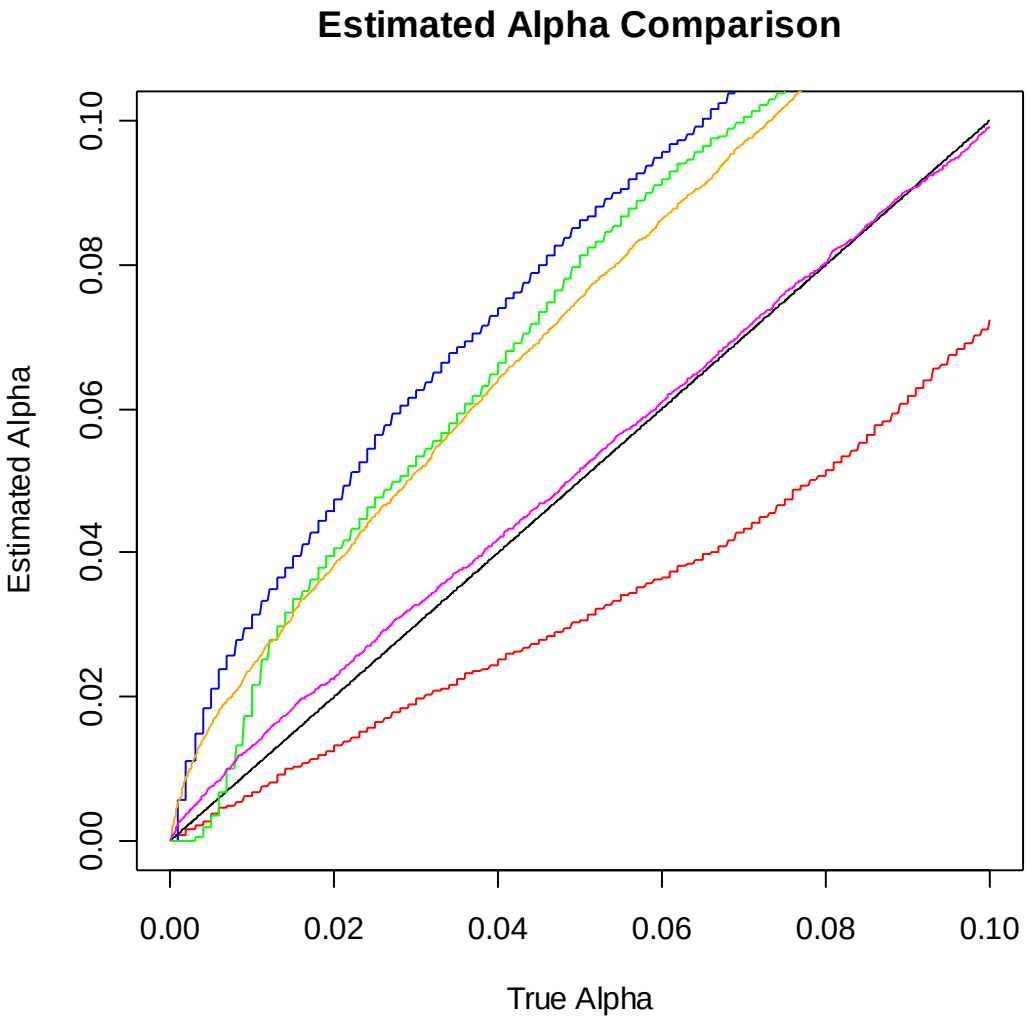

Figure B-11. Comparison of Estimated Alpha. $\mathrm{ASL} = (1+\#(|\boldsymbol{t}^*| \geq |t|))/(1+B)$, $n_1 = n_2$ = 5, *sampled distributions:* $N(0,1^2)$, $N(0,6^2)$, *B = 999. Orange: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap; Black: Exact Level Reference. Based on 20,000 simulated data sets the standard error of the estimated* α *is 0.00212 for* α = *0.10, 0.00154 for* α = *0.05, and 0.00070 for* α = *0.01.*

Table B-11. Specific Values of Estimated α *from Figure B-11.*

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.02165	0.00675	0.0313	0.0244	0.0132
0.05	0.0812	0.03075	0.0861	0.07535	0.05145
0.10	0.12515	0.07245	0.1348	0.1287	0.09925

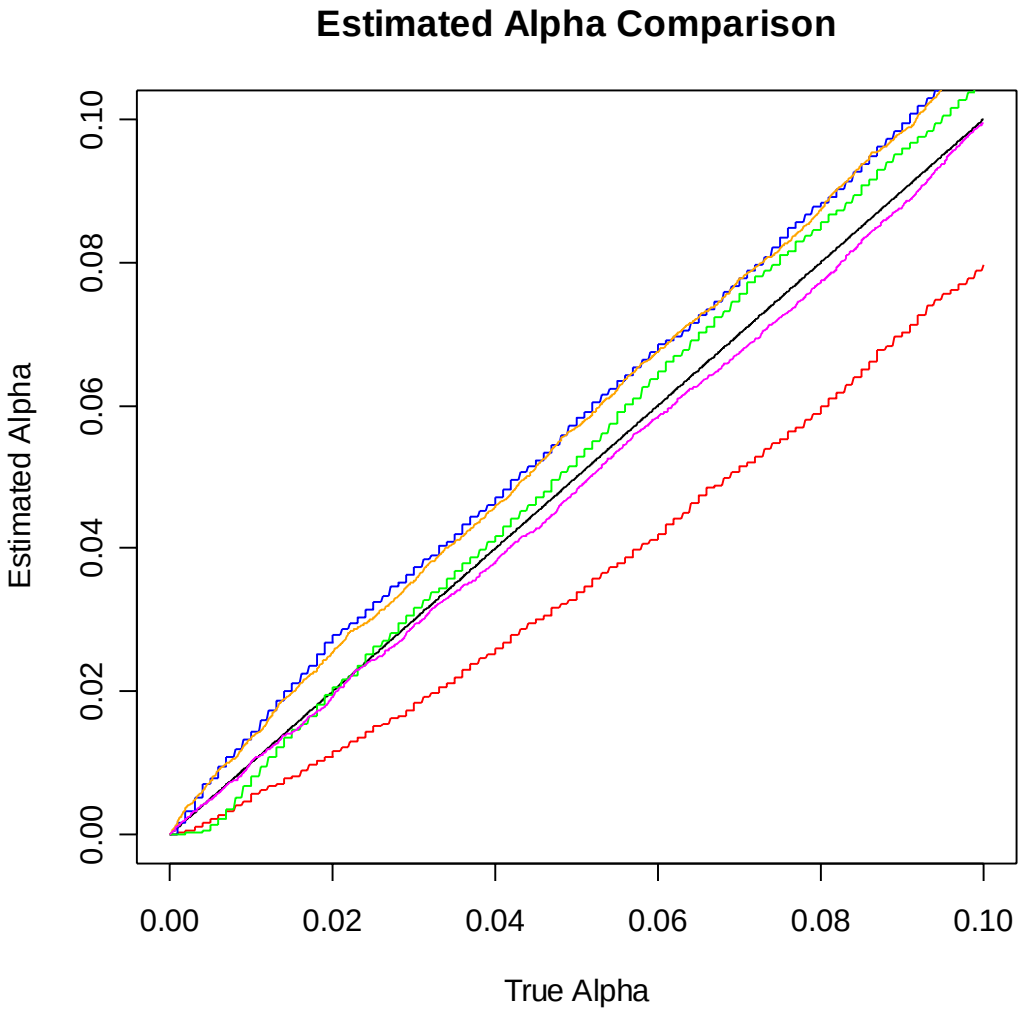

Figure B-12. Comparison of Estimated Alpha. $\mathrm{ASL} = (1 + \#(|\boldsymbol{t}^*| \geq |t|))/(1+B)$, $n_1 = n_2$ *= 5, sampled distributions:* $N(0,1^2)$, $ChiSquare(0,1^2)$ (df=3)*, B = 999. Orange: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap; Black: Exact Level Reference. Based on 20,000 simulated data sets the standard error of the estimated* α *is 0.00212 for* α *= 0.10, 0.00154 for* α *= 0.05, and 0.00070 for* α *= 0.01.*

Table B-12. Specific Values of Estimated α *from Figure B-12.*

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.00825	0.00555	0.0143	0.0135	0.01005
0.05	0.05285	0.0338	0.0583	0.057	0.048
0.10	0.106	0.0797	0.11155	0.10945	0.09935

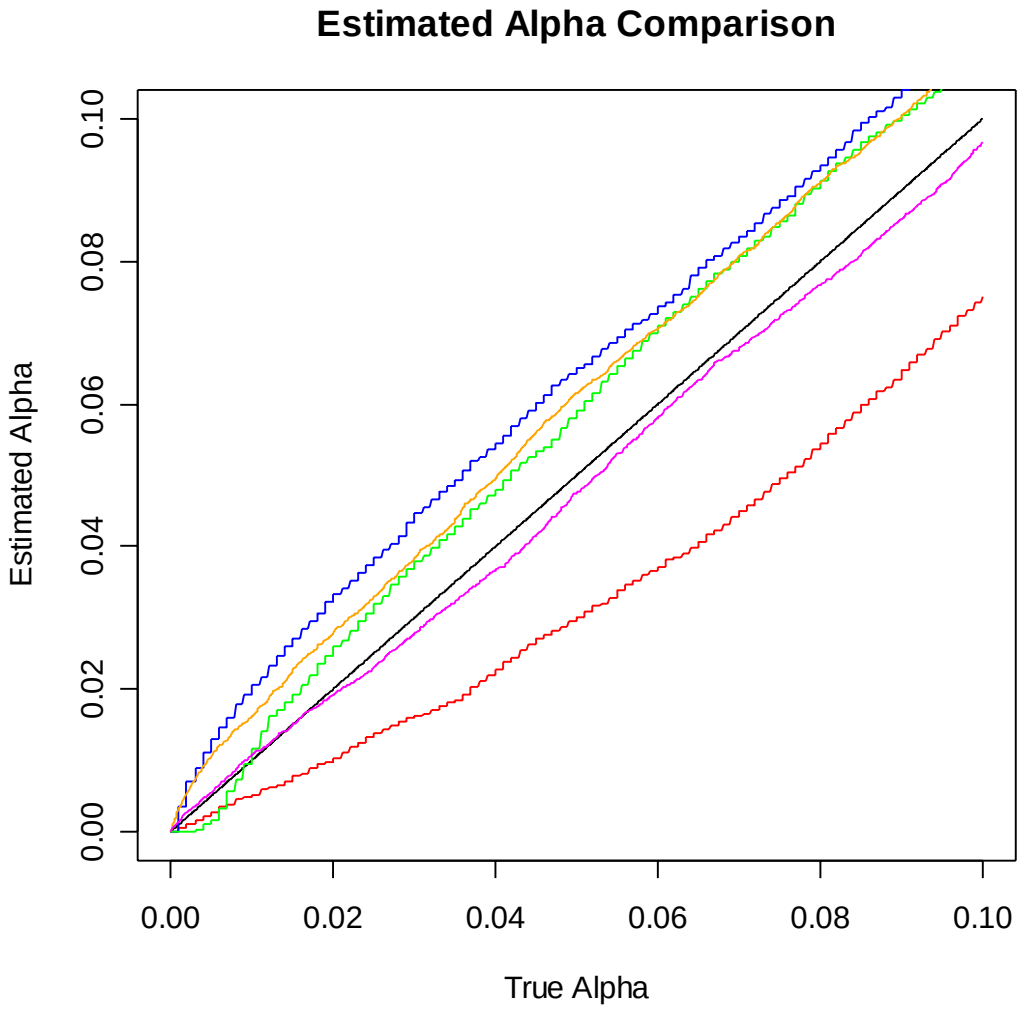

Figure B-13. Comparison of Estimated Alpha. $\mathrm{ASL} = (1+\#(|\boldsymbol{t}^*| \geq |t|))/(1+B)$, $n_1 = n_2$ *= 5, sampled distributions:* $N(0,1^2)$, $ChiSquare(0,1^2)$ (df=1), *B = 999. Orange: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap; Black: Exact Level Reference. Based on 20,000 simulated data sets the standard error of the estimated* α *is 0.00212 for* α *= 0.10, 0.00154 for* α *= 0.05, and 0.00070 for* α *= 0.01.*

Table B-13. Specific Values of Estimated α *from Figure B-13.*

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.0117	0.00525	0.0207	0.01615	0.0107
0.05	0.0592	0.0301	0.06515	0.06155	0.04755
0.10	0.10865	0.0751	0.11415	0.11185	0.09665

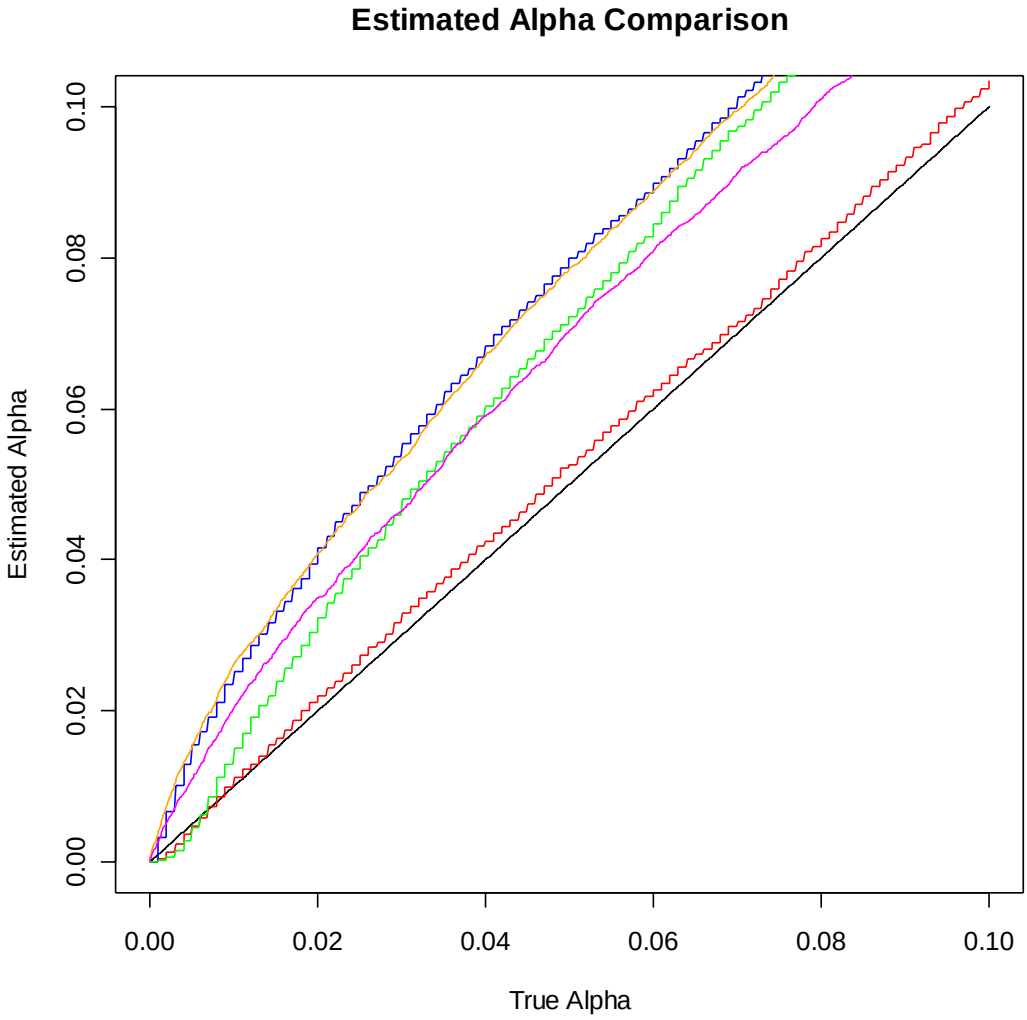

Figure B-14. Comparison of Estimated Alpha. $\mathrm{ASL} = (1+\#(|\boldsymbol{t}^*| \geq |t|))/(1+B)$, $n_1 = n_2$ *= 5, sampled distributions:* $N(0,1^2)$, $ChiSquare(0,2^2)$ (df=3), *B = 999. Orange: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap; Black: Exact Level Reference. Based on 20,000 simulated data sets the standard error of the estimated* α *is 0.00212 for* α *= 0.10, 0.00154 for* α *= 0.05, and 0.00070 for* α *= 0.01.*

Table B-14. Specific Values of Estimated α *from Figure B-14.*

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.0152	0.01125	0.0253	0.02605	0.0203
0.05	0.07225	0.05265	0.0799	0.0785	0.0703
0.10	0.1306	0.10345	0.1338	0.13215	0.12135

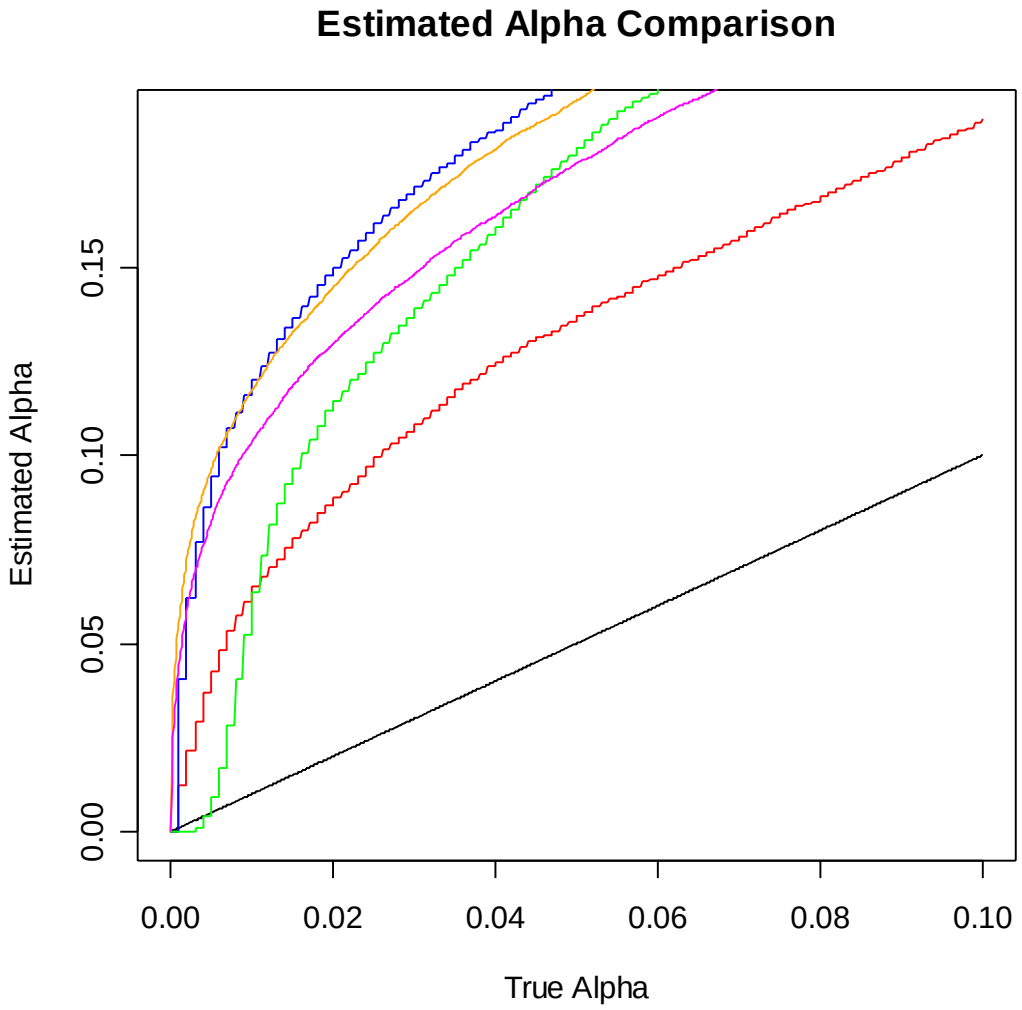

Figure B-15. Comparison of Estimated Alpha. $\mathrm{ASL} = (1 + \#(|\boldsymbol{t}^*| \geq |t|))/(1+B)$, $n_1 = n_2$ *= 5, sampled distributions:* $N(0,1^2)$, $ChiSquare(0,6^2)$ (df=1), *B = 999. Orange: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap; Black: Exact Level Reference. Based on 20,000 simulated data sets the standard error of the estimated* α *is 0.00212 for* α *= 0.10, 0.00154 for* α *= 0.05, and 0.00070 for* α *= 0.01.*

Table B-15. Specific Values of Estimated α *from Figure B-15.*

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.0634	0.06495	0.12025	0.11715	0.10355
0.05	0.1818	0.1369	0.2011	0.19425	0.1776
0.10	0.2382	0.1893	0.2495	0.24515	0.228

APPENDIX C

COMPARISON OF ESTIMATED POWER

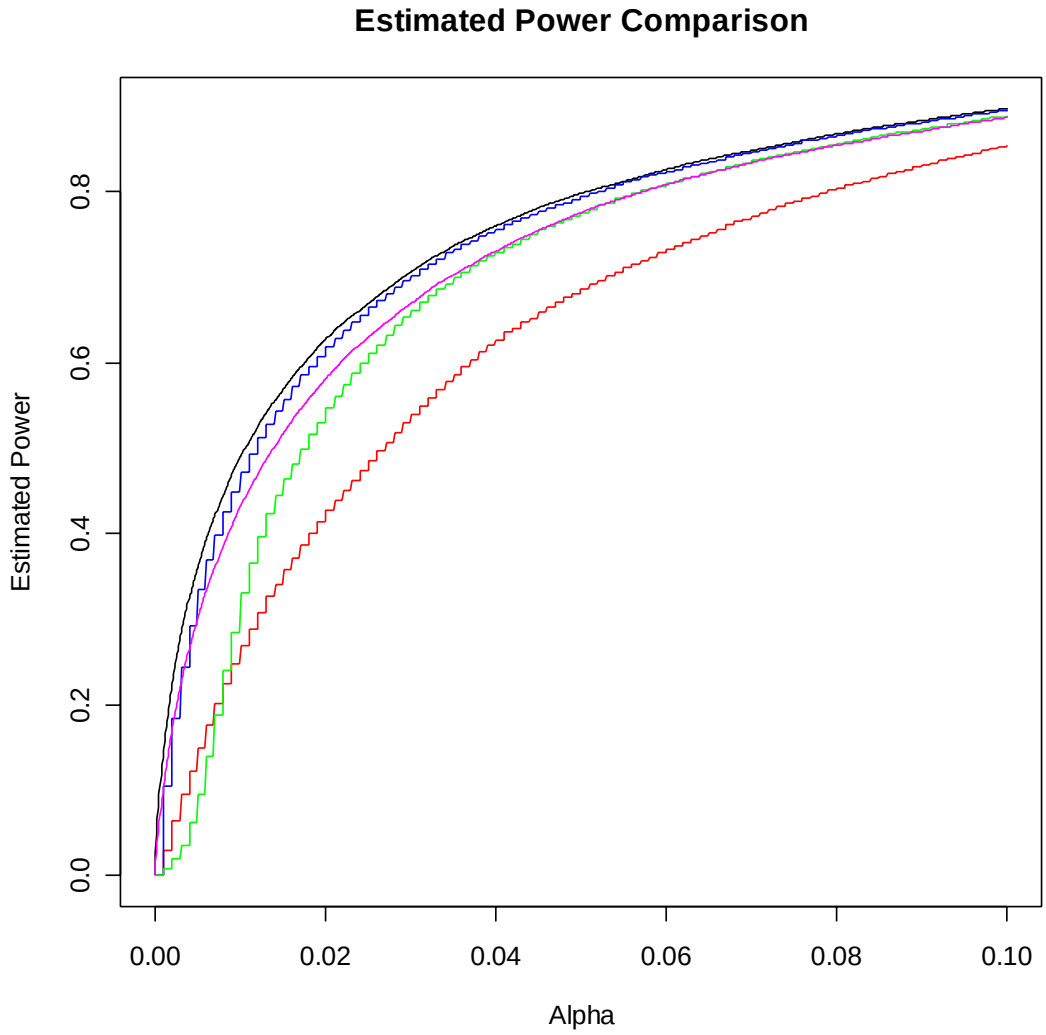

Figure C-1. Comparison of Estimated Power. $\mathrm{ASL} = (1 + \#(|\boldsymbol{t}^*| \geq |t|))/(1 + B)$, $n_1 = n_2$ *= 5, sampled distributions:* $N(0,1^2)$, $N(2.02443934,1^2)$, *B = 999. Black: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap. Based on 20,000 simulated data sets the standard error of the estimated power is 0.00354 for power= 0.50, 0.00283 for power= 0.80, and 0.00212 for power = 0.90.*

Table C-1. Specific Values of Estimated Power from Figure C-1.

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.3296	0.26835	0.47225	0.48915	0.4307
0.05	0.77565	0.6858	0.7939	0.79775	0.7751
0.10	0.88875	0.8547	0.89535	0.89675	0.8865

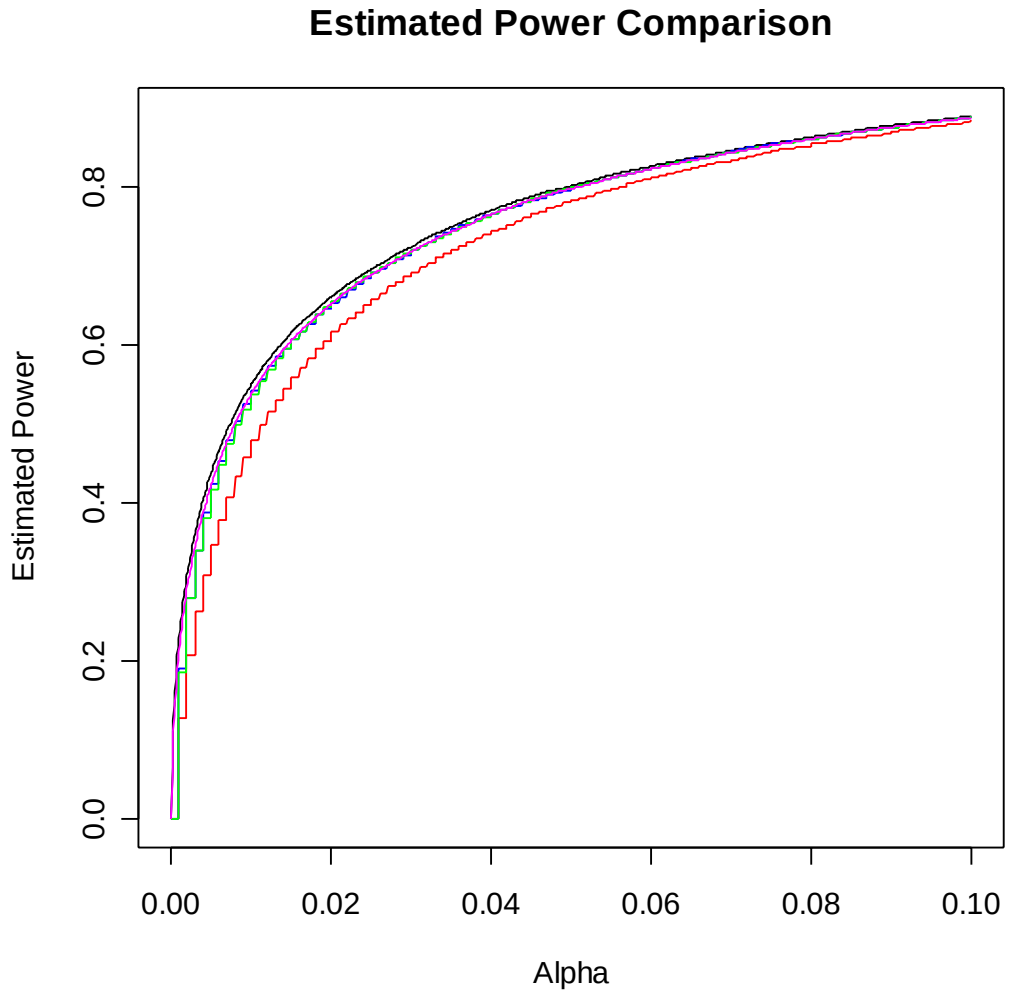

Figure C-2. Comparison of Estimated Power. $\mathrm{ASL} = (1 + \#(|\boldsymbol{t}^*| \geq |t|))/(1+B)$, $n_1 = n_2$ *= 10, sampled distributions:* $N(0,1^2)$, $N(1.32494739,1^2)$, *B = 999. Black: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap. Based on 20,000 simulated data sets the standard error of the estimated power is 0.00354 for power= 0.50, 0.00283 for power= 0.80, and 0.00212 for power = 0.90.*

Table C-2. Specific Values of Estimated Power from Figure C-2.

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.53665	0.4791	0.54135	0.5486	0.5366
0.05	0.79945	0.7824	0.79805	0.80045	0.7969
0.10	0.88775	0.88255	0.8872	0.8882	0.88645

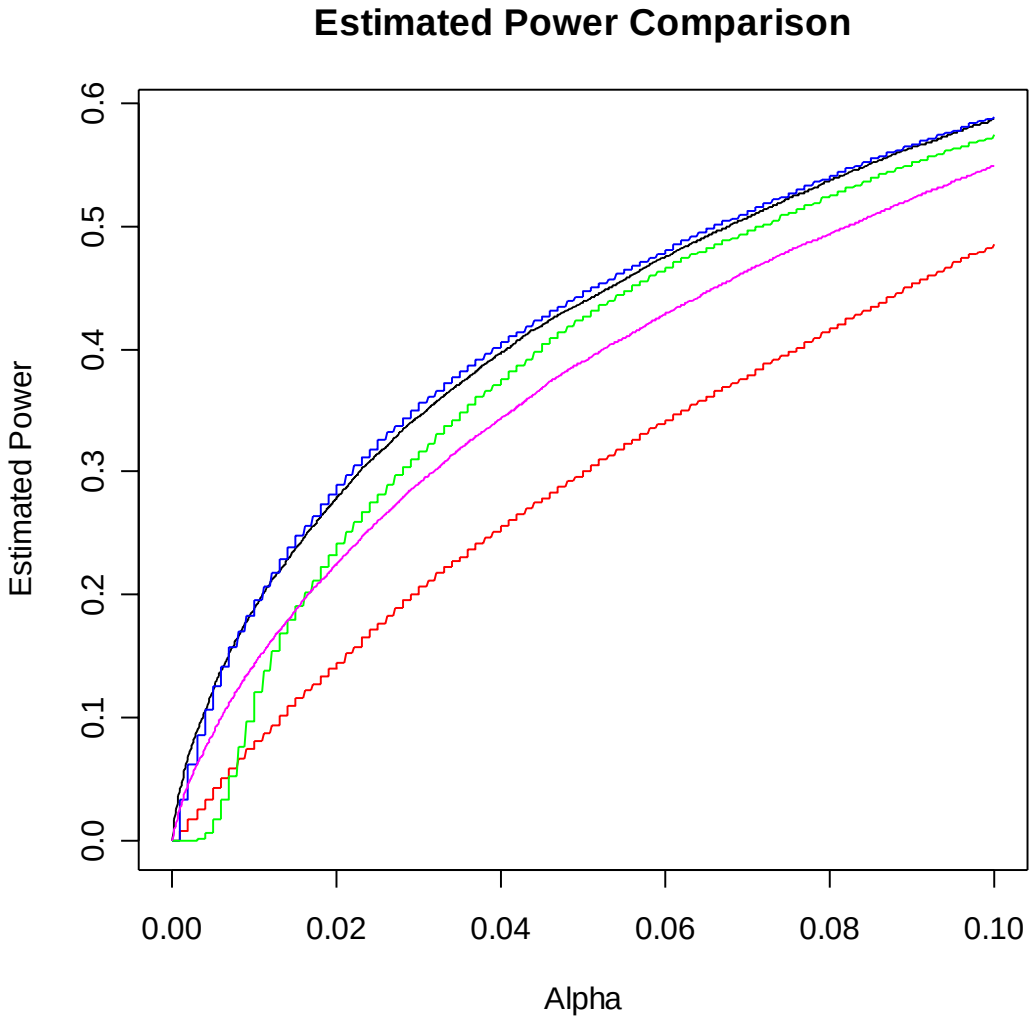

Figure C-3. Comparison of Estimated Power. $\mathrm{ASL} = (1 + \#(|\boldsymbol{t}^*| \geq |t|))/(1+B)$, $n_1 = n_2$ = 5, *sampled distributions:* $N(0,1^2)$, $N(2.02443934,2^2)$, *B = 999. Black: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap. Based on 20,000 simulated data sets the standard error of the estimated power is 0.00354 for power= 0.50, 0.00283 for power= 0.80, and 0.00212 for power = 0.90.*

Table C-3. Specific Values of Estimated Power from Figure C-3.

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.1201	0.0805	0.1949	0.18885	0.1431
0.05	0.42625	0.3005	0.4466	0.4378	0.3896
0.10	0.57345	0.4853	0.58885	0.58685	0.54895

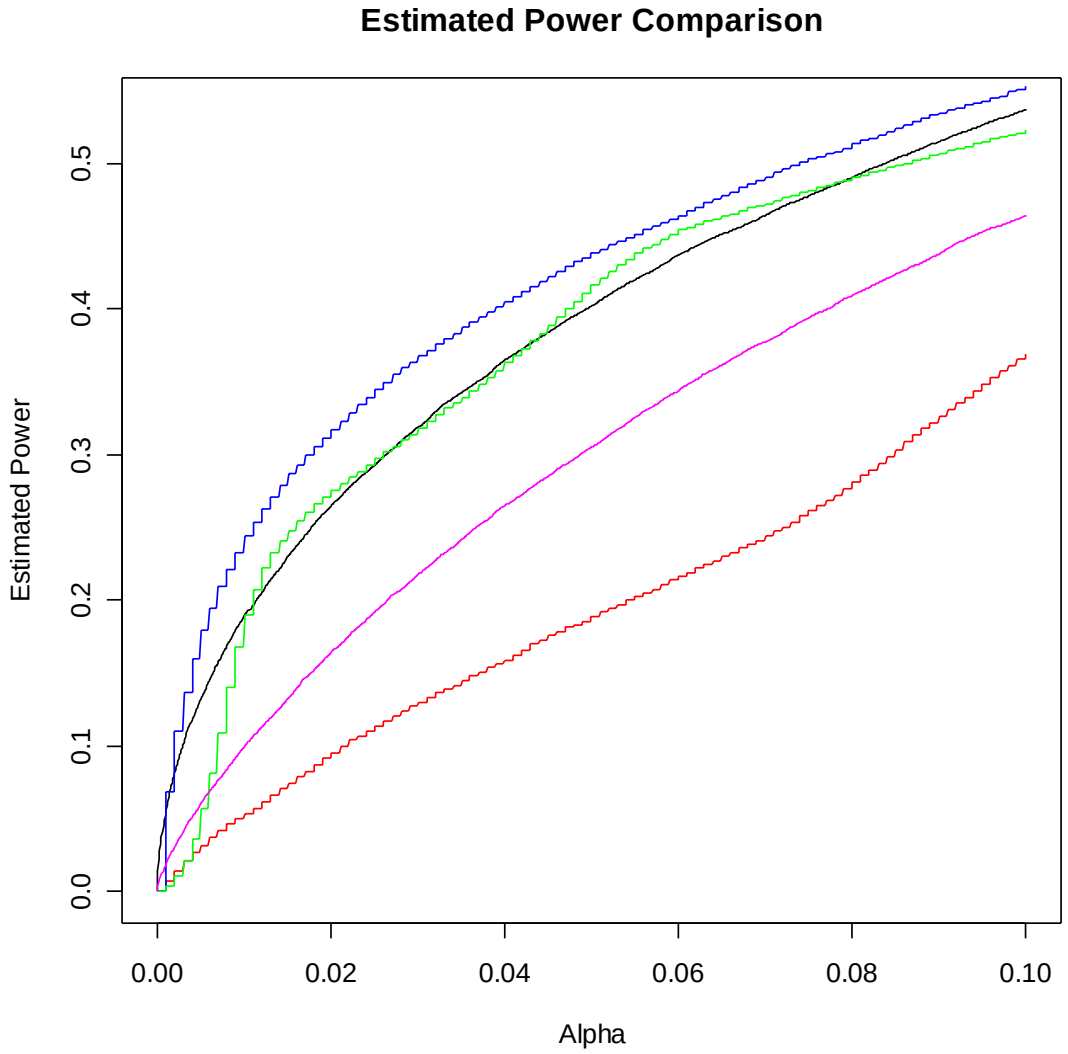

Figure C-4. Comparison of Estimated Power. $\mathrm{ASL} = (1 + \#(|t^*| \geq |t|))/(1 + B)$, $n_1 = n_2$ = 5, *sampled distributions:* $N(0,1^2)$, $N(5.00,6^2)$, *B = 999. Black: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap. Based on 20,000 simulated data sets the standard error of the estimated power is 0.00354 for power= 0.50, 0.00283 for power= 0.80, and 0.00212 for power = 0.90.*

Table C-4. Specific Values of Estimated Power from Figure C-4.

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.18965	0.0534	0.2439	0.1885	0.0989
0.05	0.41645	0.1886	0.43865	0.40205	0.30505
0.10	0.52285	0.3697	0.5526	0.53705	0.4646

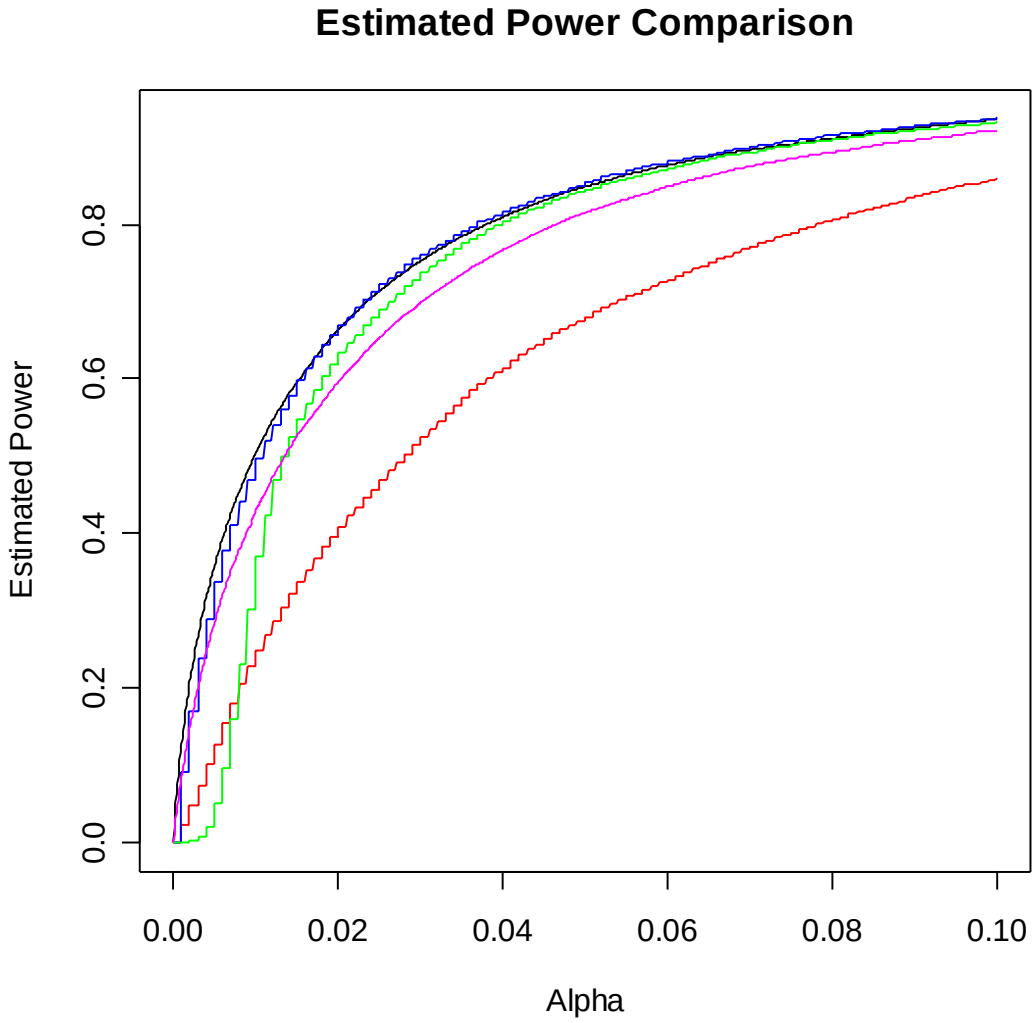

Figure C-5. Comparison of Estimated Power. $\mathrm{ASL} = (1 + \#(|\boldsymbol{t}^*| \geq |t|))/(1+B)$, $n_1 = n_2$ *= 5, sampled distributions:* $N(0,1^2)$, *ChiSquare*$(2.02443934,1^2)(\mathrm{df}=3)$, *B = 999. Black: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap. Based on 20,000 simulated data sets the standard error of the estimated power is 0.00354 for power= 0.50, 0.00283 for power= 0.80, and 0.00212 for power = 0.90.*

Table 20. Specific Values of Estimated Power from Figure 48.

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.3701	0.2489	0.49625	0.5037	0.4284
0.05	0.845	0.68025	0.85485	0.84915	0.81495
0.10	0.93275	0.86045	0.93745	0.9359	0.92205

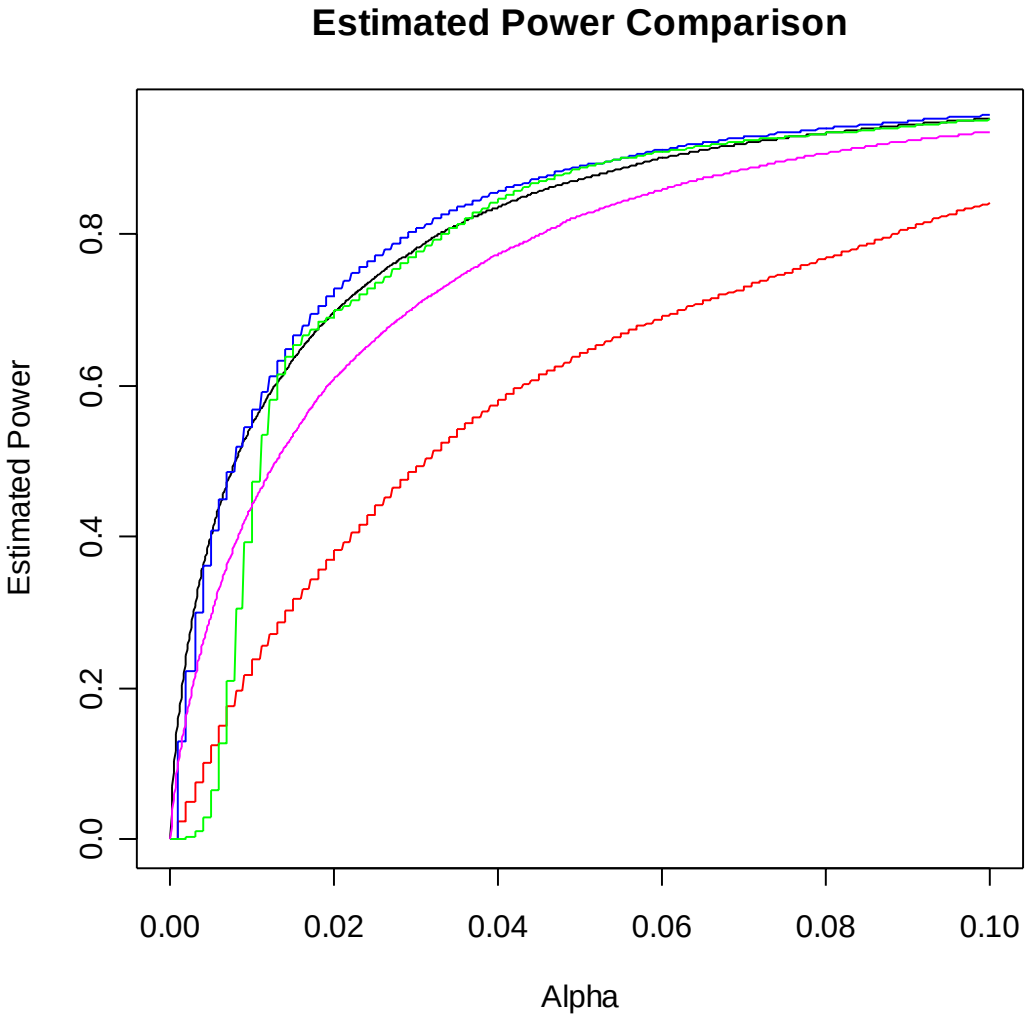

Figure C-6. Comparison of Estimated Power. $\mathrm{ASL} = (1 + \#(|\boldsymbol{t}^*| \geq |t|))/(1+B)$, $n_1 = n_2$ = *5, sampled distributions:* $N(0,1^2)$, *ChiSquare*$(2.02443934,1^2)$(df=1), *B = 999. Black: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap. Based on 20,000 simulated data sets the standard error of the estimated power is 0.00354 for power= 0.50, 0.00283 for power= 0.80, and 0.00212 for power = 0.90.*

Table C-6. Specific Values of Estimated Power from Figure C-6.

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.47395	0.2365	0.56895	0.5483	0.4414
0.05	0.88825	0.64325	0.89075	0.8731	0.8248
0.10	0.9526	0.84345	0.9587	0.95325	0.93625

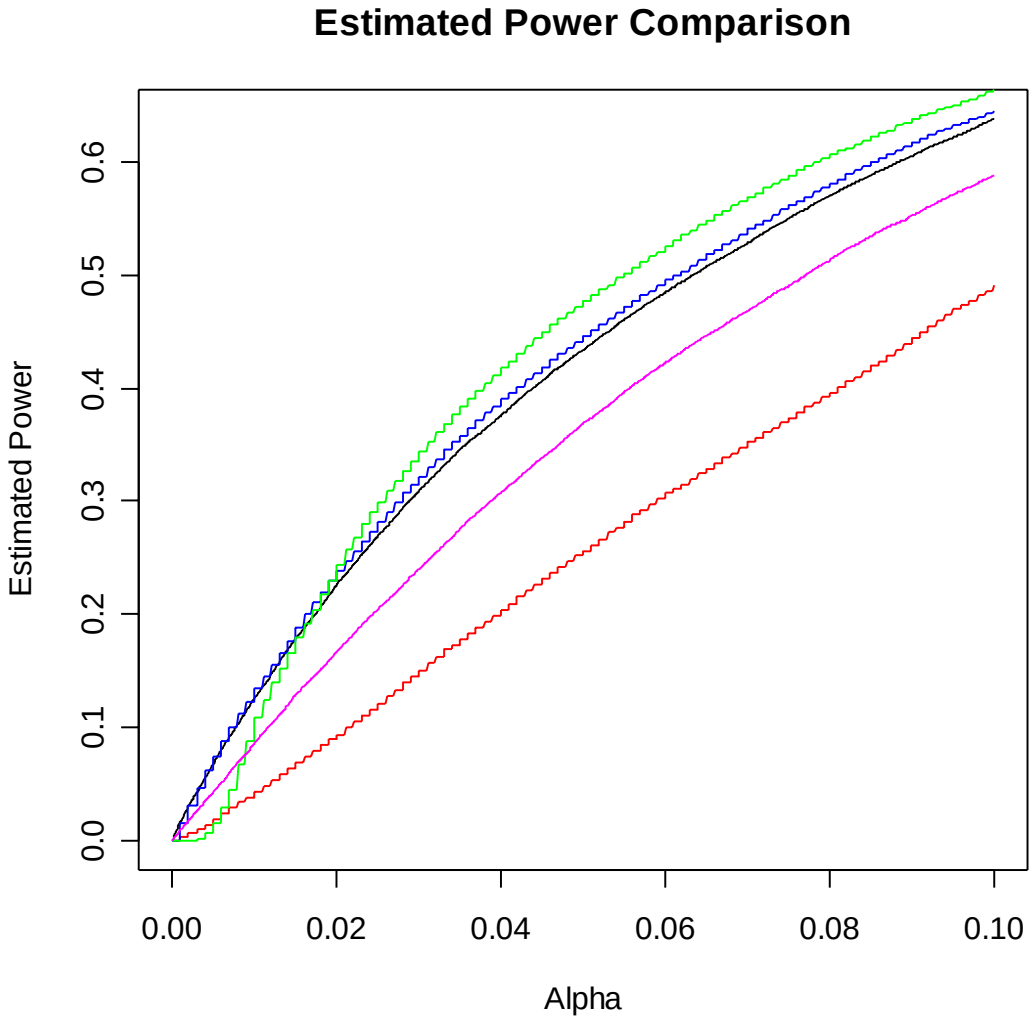

Figure C-7. Comparison of Estimated Power. $\mathrm{ASL} = (1 + \#(|\boldsymbol{t}^*| \geq |t|))/(1+B)$, $n_1 = n_2$ = 5, *sampled distributions:* $N(0,1^2)$, *ChiSquare*$(2.02443934,2^2)(\mathrm{df}=3)$, *B = 999. Black: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap. Based on 20,000 simulated data sets the standard error of the estimated power is 0.00354 for power= 0.50, 0.00283 for power= 0.80, and 0.00212 for power = 0.90.*

Table C-7. Specific Values of Estimated Power from Figure C-7.

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.10895	0.0437	0.1343	0.12565	0.086
0.05	0.4767	0.25655	0.4466	0.4337	0.3689
0.10	0.6635	0.4901	0.6452	0.6376	0.5879

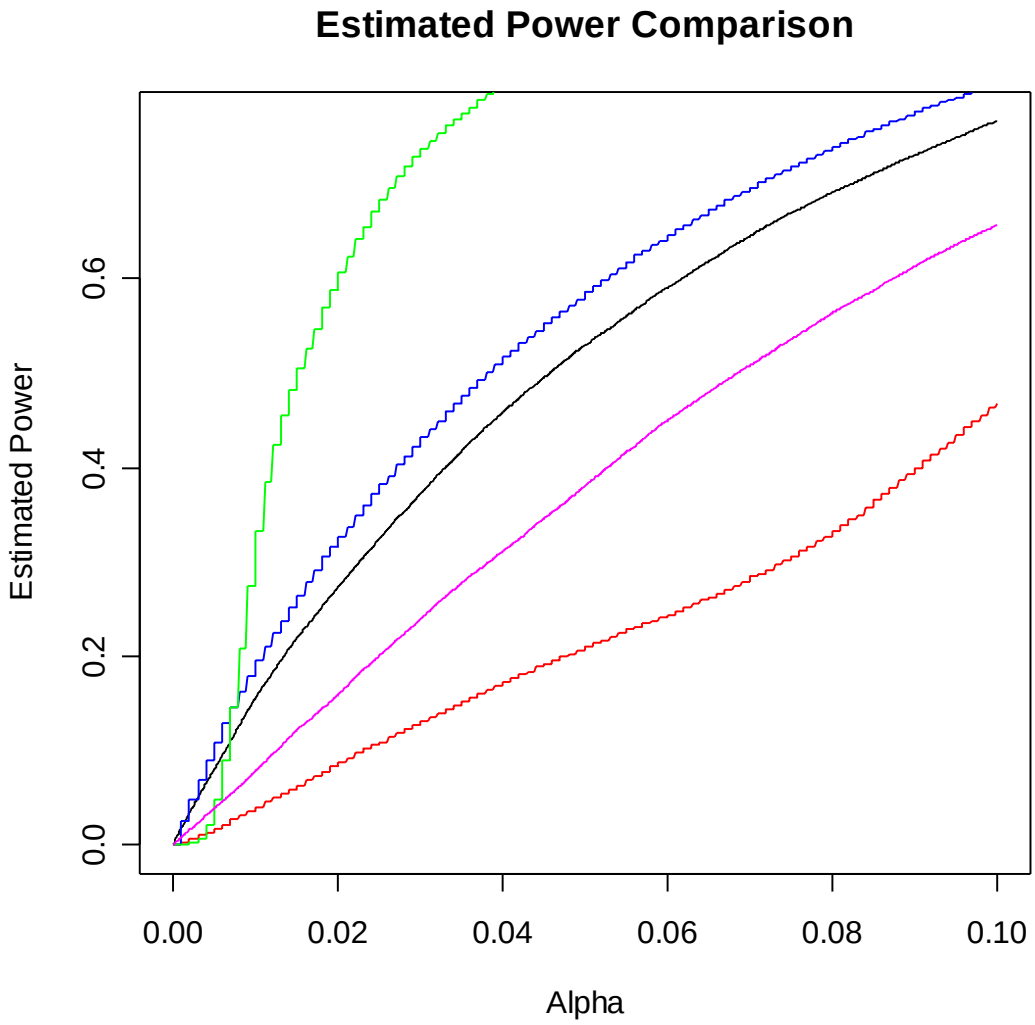

Figure C-8. Comparison of Estimated Power. $\mathrm{ASL} = (1 + \#(|t^*| \geq |t|))/(1 + B)$, $n_1 = n_2$ *= 5, sampled distributions:* $N(0,1^2)$, *ChiSquare*$(5.00,6^2)$(df=1), *B = 999. Black: t-test; Magenta: Welch's t-test, Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap. Based on 20,000 simulated data sets the standard error of the estimated power is 0.00354 for power= 0.50, 0.00283 for power= 0.80, and 0.00212 for power = 0.90.*

Table C-8. Specific Values of Estimated Power from Figure C-8.

True α	Permutation	Bootstrap Within	Bootstrap Across	*t*-test	Welch's *t*-test
0.01	0.33295	0.0396	0.195	0.15535	0.0779
0.05	0.85265	0.21	0.58575	0.52975	0.3808
0.10	0.94585	0.46885	0.80855	0.7673	0.6579

APPENDIX D

OBSERVED ASL COMPARISON

Each ASL was found as the proportion of resampled t^*'s as extreme or more extreme than the cutoff (Table D-1). These ASLs were then compared to the expected ASL based on the mean, 10% trimmed mean, standard deviation, and MSE of the ASLs. Only the results for sampling from two standard normal distributions are included here. The results for the many other distributions are very lengthy (not included) but follow the same patterns as those indicated in the graphs below.

Table D-1. Cutoff Values of the t-distribution for Two-Sided Expected ASLs

Two-sided Expected ASL	Cutoff
0.10	1.85954803752958
0.05	2.30600413520721
0.01	3.35538733132929
0.005	3.83251868533852

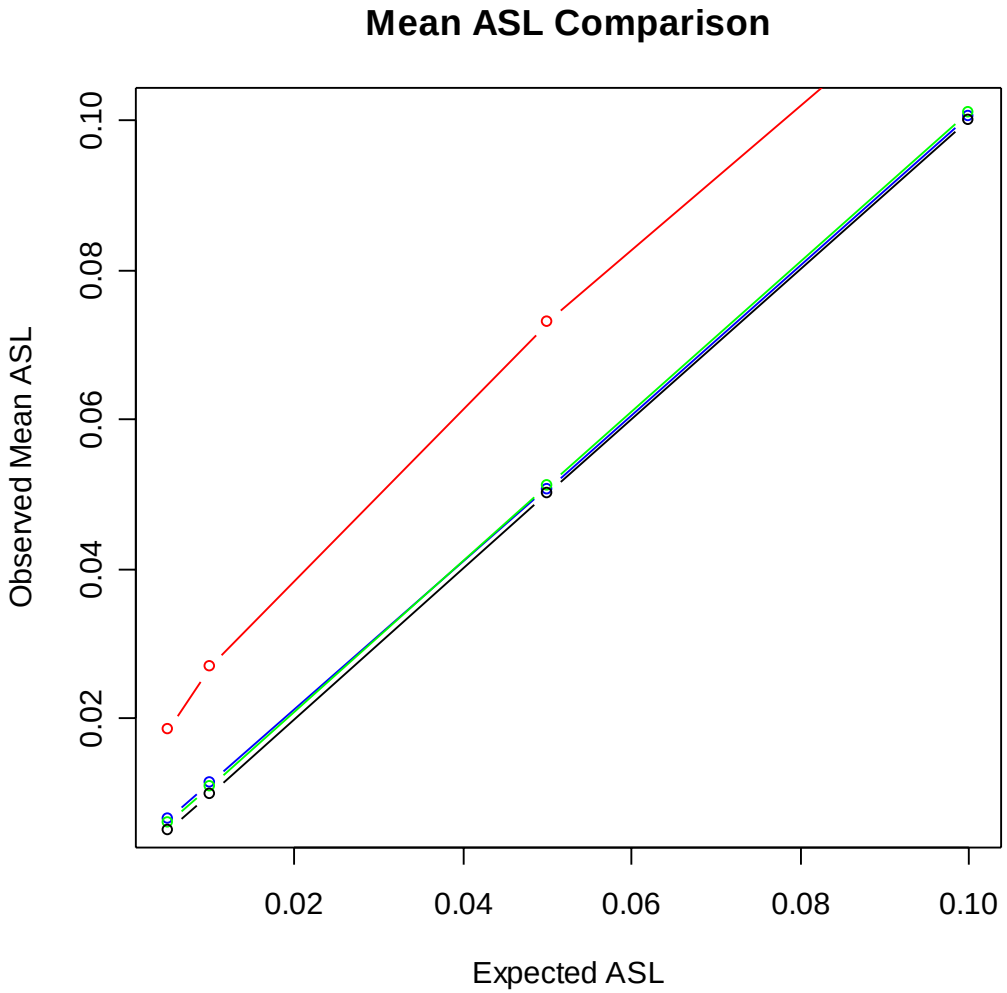

Figure D-1. Summary of Observed Mean ASL for Underlying Distributions with Equal Means, $n_1 = n_2 = 5$. Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap; Black: Expected ASL Reference.

Table D-2. Observed Mean ASL Values from Figure D-1.

Method	0.005	0.01	0.05	0.10
Within	0.01852	0.02682	0.07293	0.12110
Across	0.00661	0.01149	0.05073	0.10048
Permutation	0.00601	0.01101	0.05110	0.10091

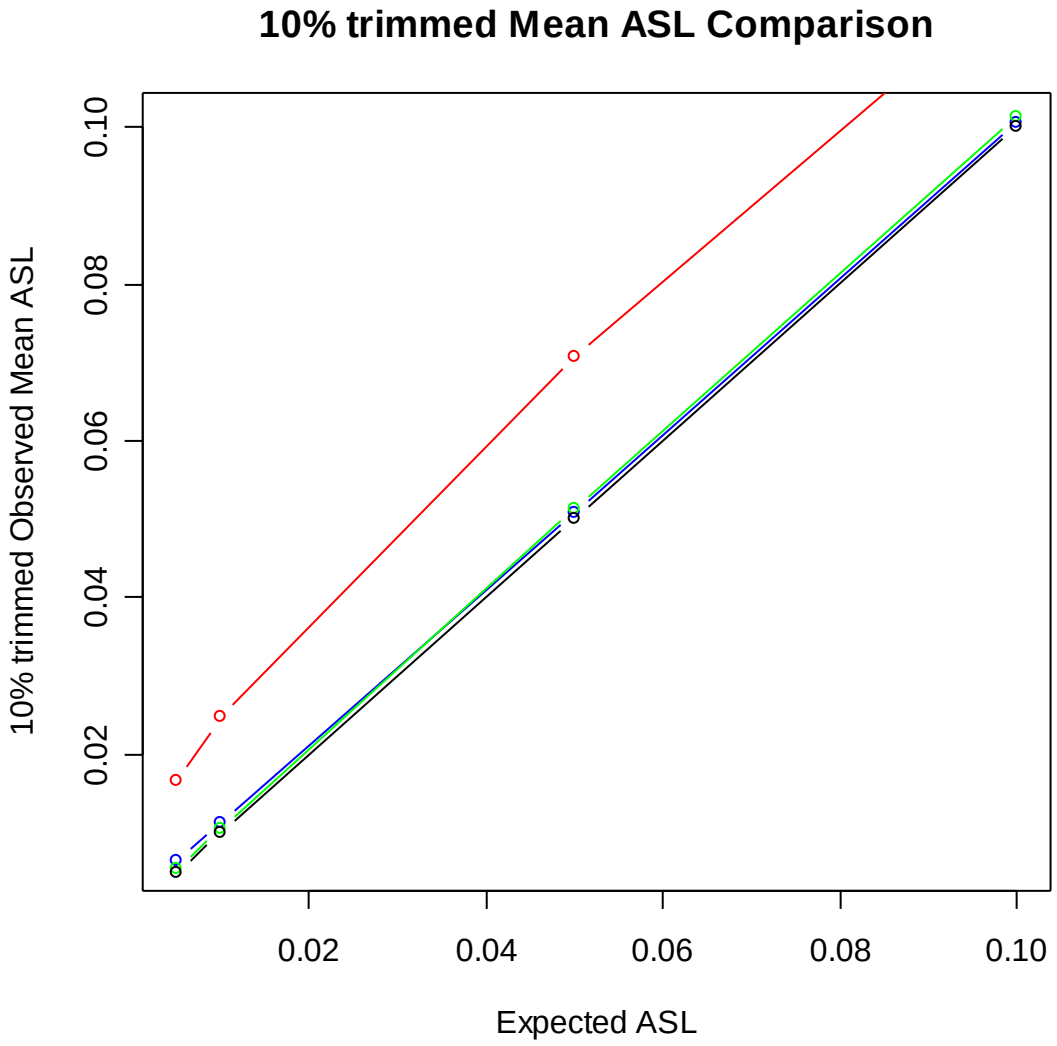

Figure D-2. Summary of Observed 10% Trimmed Mean ASL for Underlying Distributions with Equal Means, $n_1 = n_2 = 5$. Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap; Black: Expected ASL Reference.

Table D-3. Observed 10% Trimmed Mean ASL Values from Figure D-2.

Method	0.005	0.01	0.05	0.10
Within	0.01661	0.02471	0.07076	0.11863
Across	0.00648	0.01141	0.05083	0.10052
Permutation	0.00540	0.01044	0.05145	0.10124

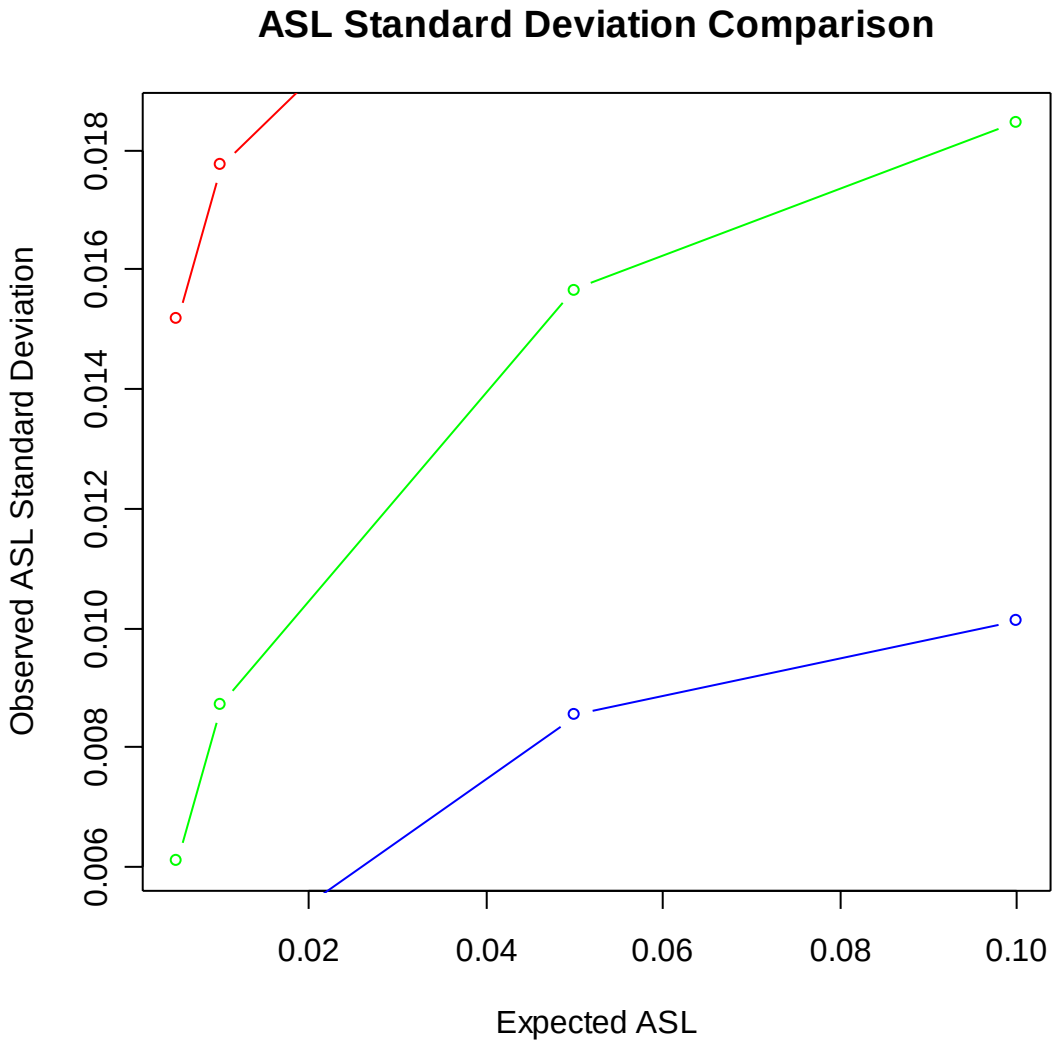

Figure D-3. Summary of Observed ASL Standard Deviation for Underlying Distributions with Equal Means, $n_1 = n_2 = 5$. Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap.

Table D-4. Observed ASL Standard Deviation Values from Figure D-3.

Method	0.005	0.01	0.05	0.10
Within	0.01517	0.01775	0.02352	0.02493
Across	0.00311	0.00437	0.00854	0.01013
Permutation	0.00611	0.00872	0.01566	0.01846

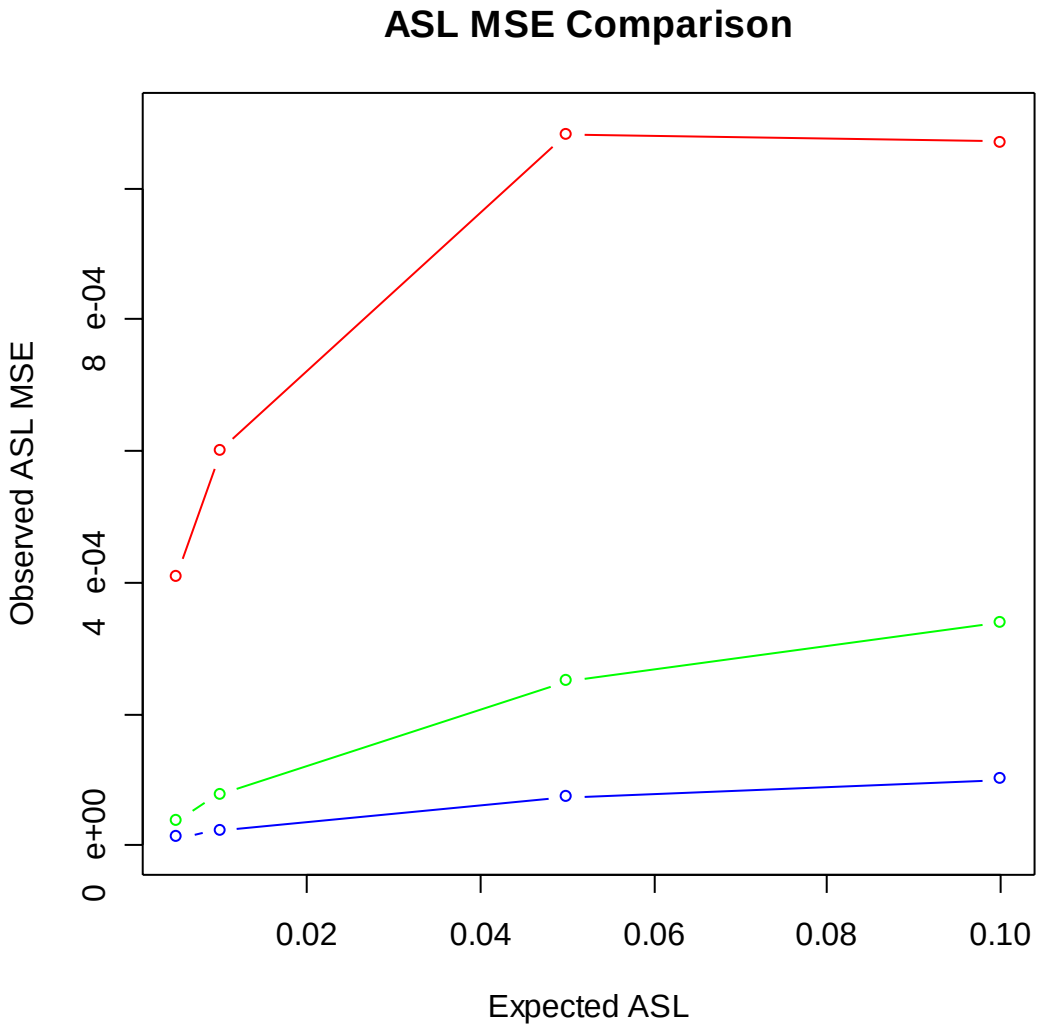

Figure D-4. Summary of Observed ASL MSE for Underlying Distributions with Equal Means, $n_1 = n_2 = 5$. Green: Permutation; Blue: Across group Bootstrap; Red: Within group Bootstrap.

Table D-5. Observed ASL MSE Values from Figure D-4.

Method	0.005	0.01	0.05	0.10
Within	0.00041	0.00060	0.00108	0.00107
Across	1.22833e-05	2.13346e-05	7.34789e-05	0.00010
Perm	3.83672e-05	7.71225e-05	0.00025	0.00034

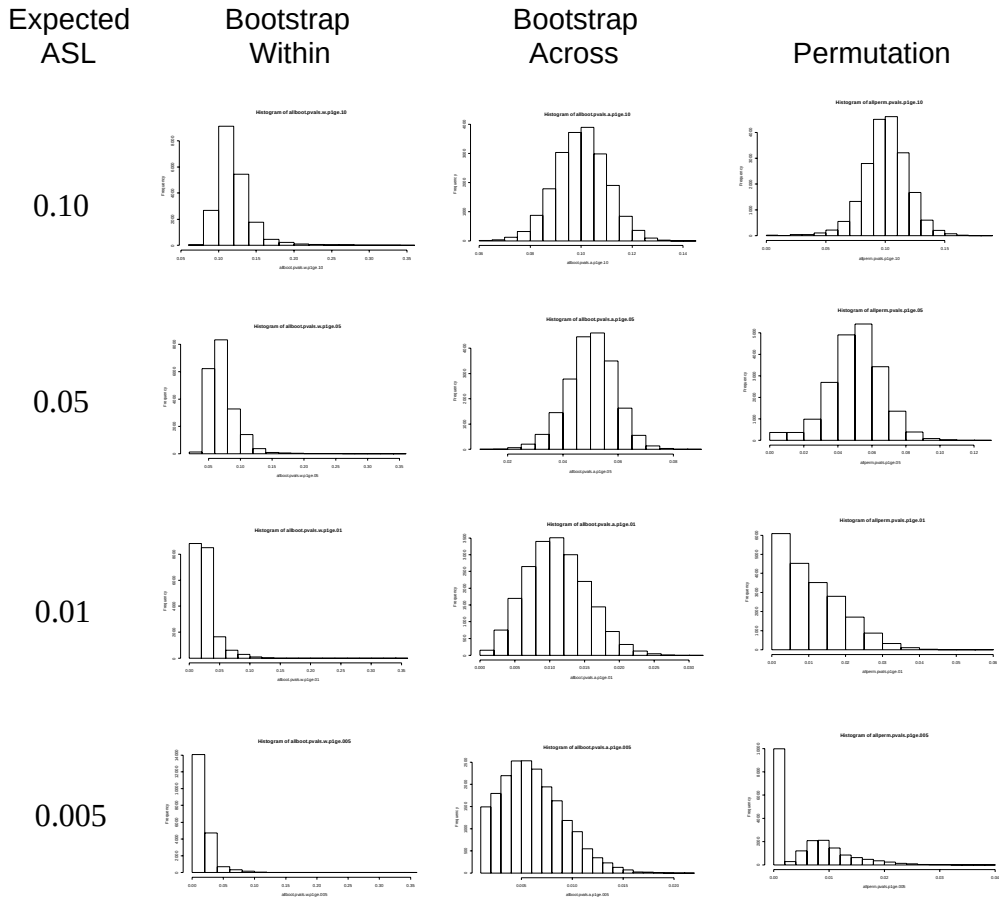

Figure D-5. ASL Distribution from Two-sample Tests with Sample Size 5 per Group

APPENDIX E

MEAN – STANDARD DEVIATION RELATIONSHIP OUTPUT

The following is the R output for thirteenth degree polynomial regression of standard deviation on the mean for 22,276 genes:

```
Call:
lm(formula = allsds ~ allmeans + I(allmeans^2) +
I(allmeans^3) +
    I(allmeans^4) + I(allmeans^5) + I(allmeans^6) +
I(allmeans^7) +
    I(allmeans^8) + I(allmeans^9) + I(allmeans^10) +
I(allmeans^11) +
    I(allmeans^12) + I(allmeans^13))

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.870e-01  2.472e-03 197.029  < 2e-16 ***
allmeans        3.513e-02  6.626e-03   5.301 1.16e-07 ***
I(allmeans^2)   3.271e-01  7.018e-03  46.605  < 2e-16 ***
I(allmeans^3)   3.435e-02  5.112e-03   6.719 1.87e-11 ***
I(allmeans^4)  -8.791e-02  2.824e-03 -31.134  < 2e-16 ***
I(allmeans^5)  -4.416e-03  1.060e-03  -4.167 3.09e-05 ***
I(allmeans^6)   1.054e-02  4.066e-04  25.936  < 2e-16 ***
I(allmeans^7)   9.629e-05  8.316e-05   1.158    0.247
I(allmeans^8)  -5.878e-04  2.689e-05 -21.862  < 2e-16 ***
I(allmeans^9)   1.384e-05  2.656e-06   5.211 1.90e-07 ***
I(allmeans^10)  1.533e-05  8.419e-07  18.203  < 2e-16 ***
I(allmeans^11) -7.913e-07  4.193e-08 -18.871  < 2e-16 ***
I(allmeans^12) -1.522e-07  1.006e-08 -15.133  < 2e-16 ***
I(allmeans^13)  1.147e-08  6.696e-10  17.135  < 2e-16 ***
---
Signif. codes:  ~0 `***'

Residual standard error: 0.2981 on 22262 degrees of freedom
Multiple R-Squared: 0.2472,     Adjusted R-squared: 0.2468
F-statistic: 562.3 on 13 and 22262 DF,  p-value: < 2.2e-16
```

APPENDIX F

SIMULATION RESULTS - UNCORRELATED

Table F-1. Summary of Figures F-1 to F-120 of Appendix F

Number of Genes	Percent Different	Sample Size (per group)	Figures
200	1%	3	F-1 to F-5
		5	F-6 to F-10
		15	F-11 to F-15
		100	F-16 to F-20
	10%	3	F-21 to F-25
		5	F-26 to F-30*
		15	F-31 to F-35
		100	F-36 to F-40
2,000	1%	3	F-41 to F-45
		5	F-46 to F-50
		15	F-51 to F-55
		100	F-56 to F-60
	10%	3	F-61 to F-65
		5	F-66 to F-70
		15	F-71 to F-75
		100	F-76 to F-80
20,000	1%	3	F-81 to F-85
		5	F-86 to F-90
		15	F-91 to F-95
		100	F-96 to F-100
	10%	3	F-101 to F-105
		5	F-106 to F-110
		15	F-111 to F-115
		100	F-116 to F-120

*Includes results for the maxT procedure.

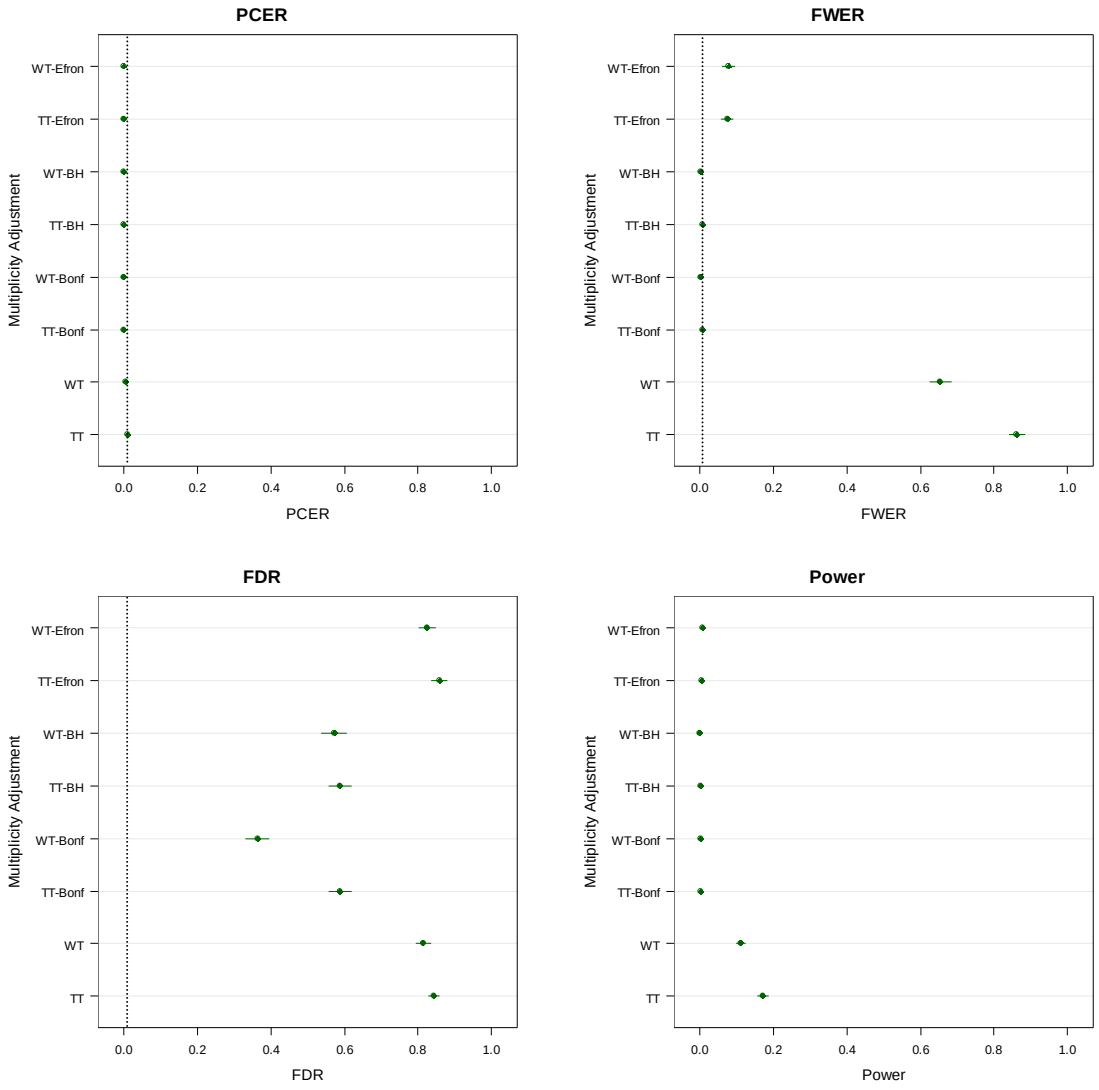

Figure F-1. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 1% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

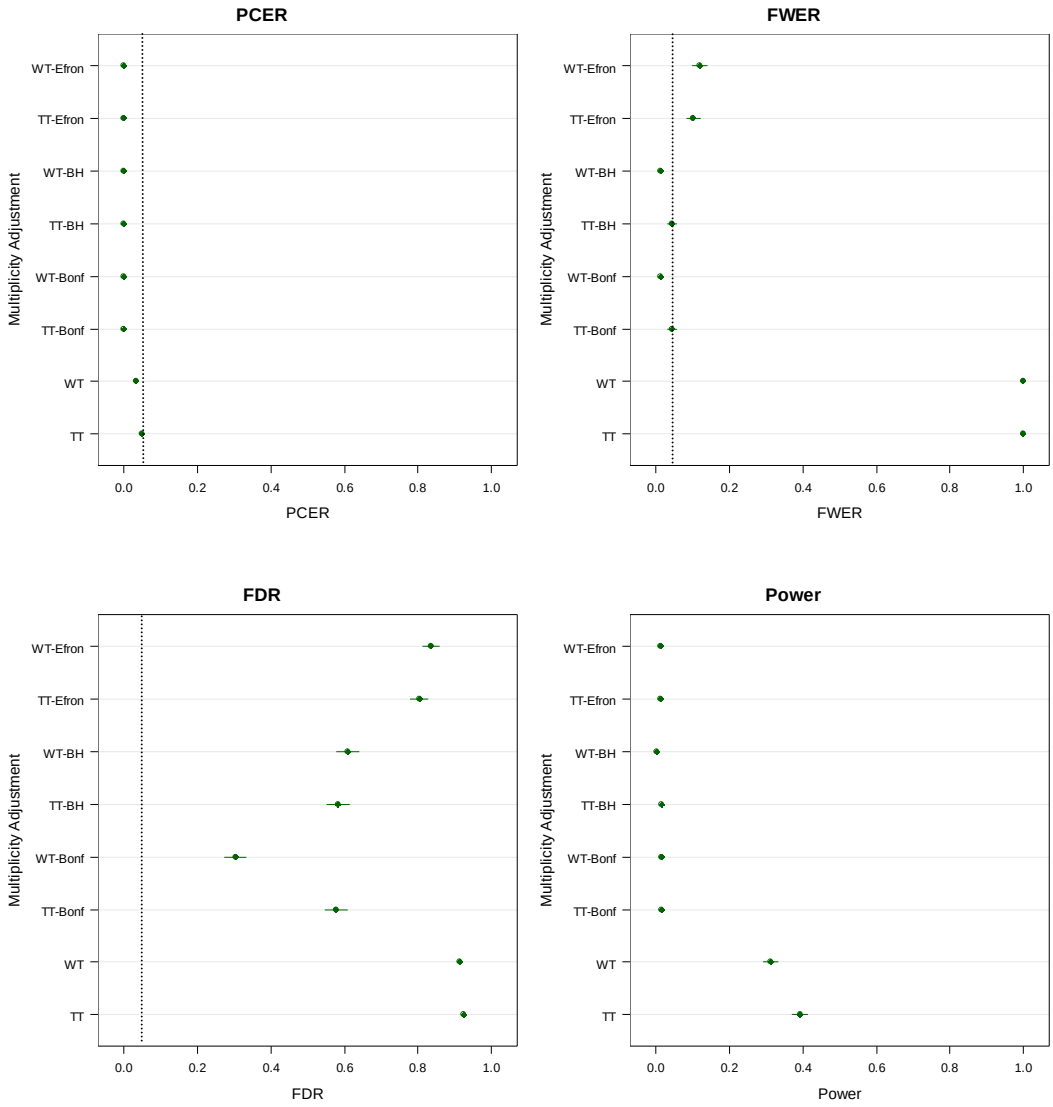

Figure F-2. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 1% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

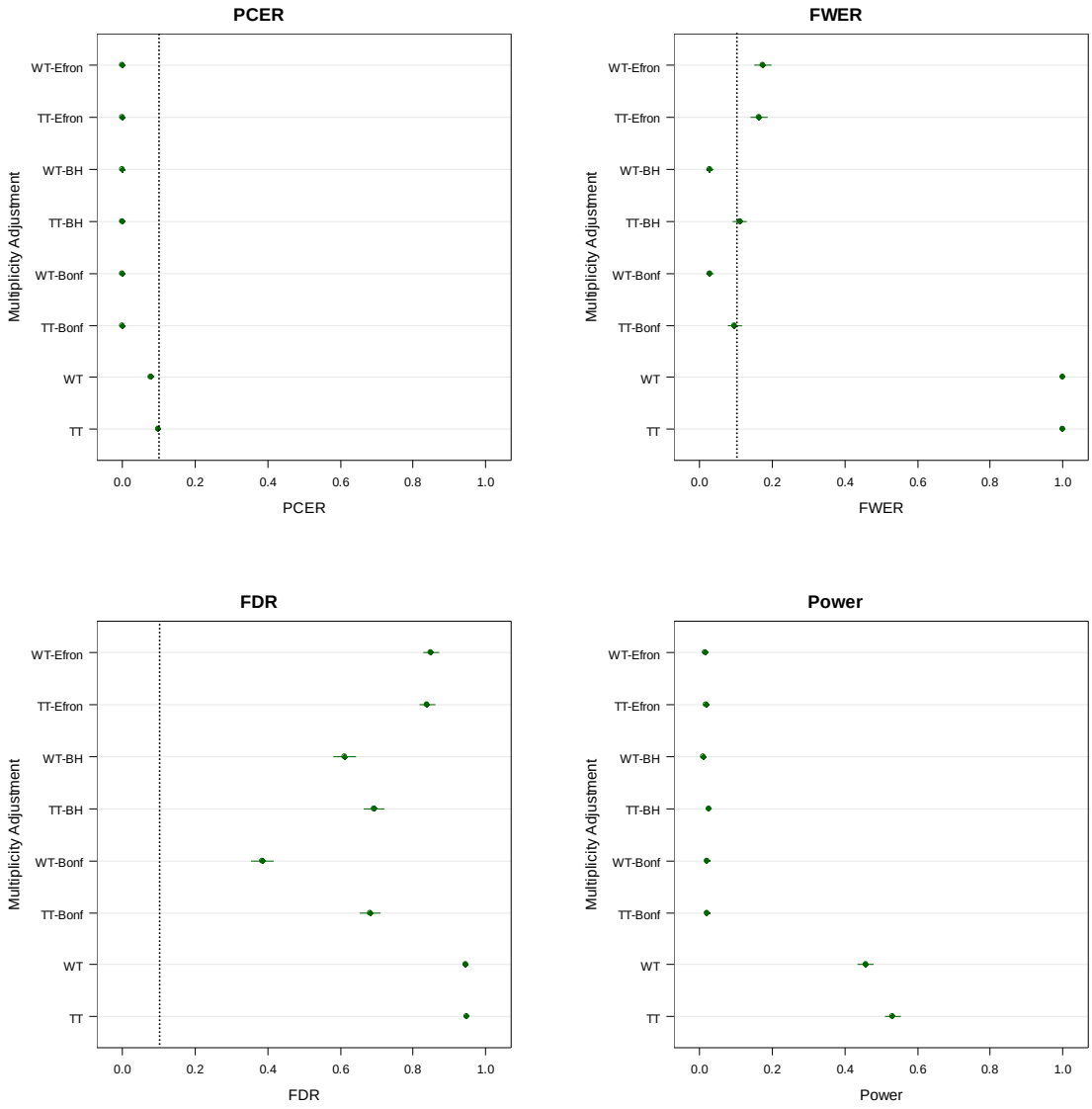

Figure F-3. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 1% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

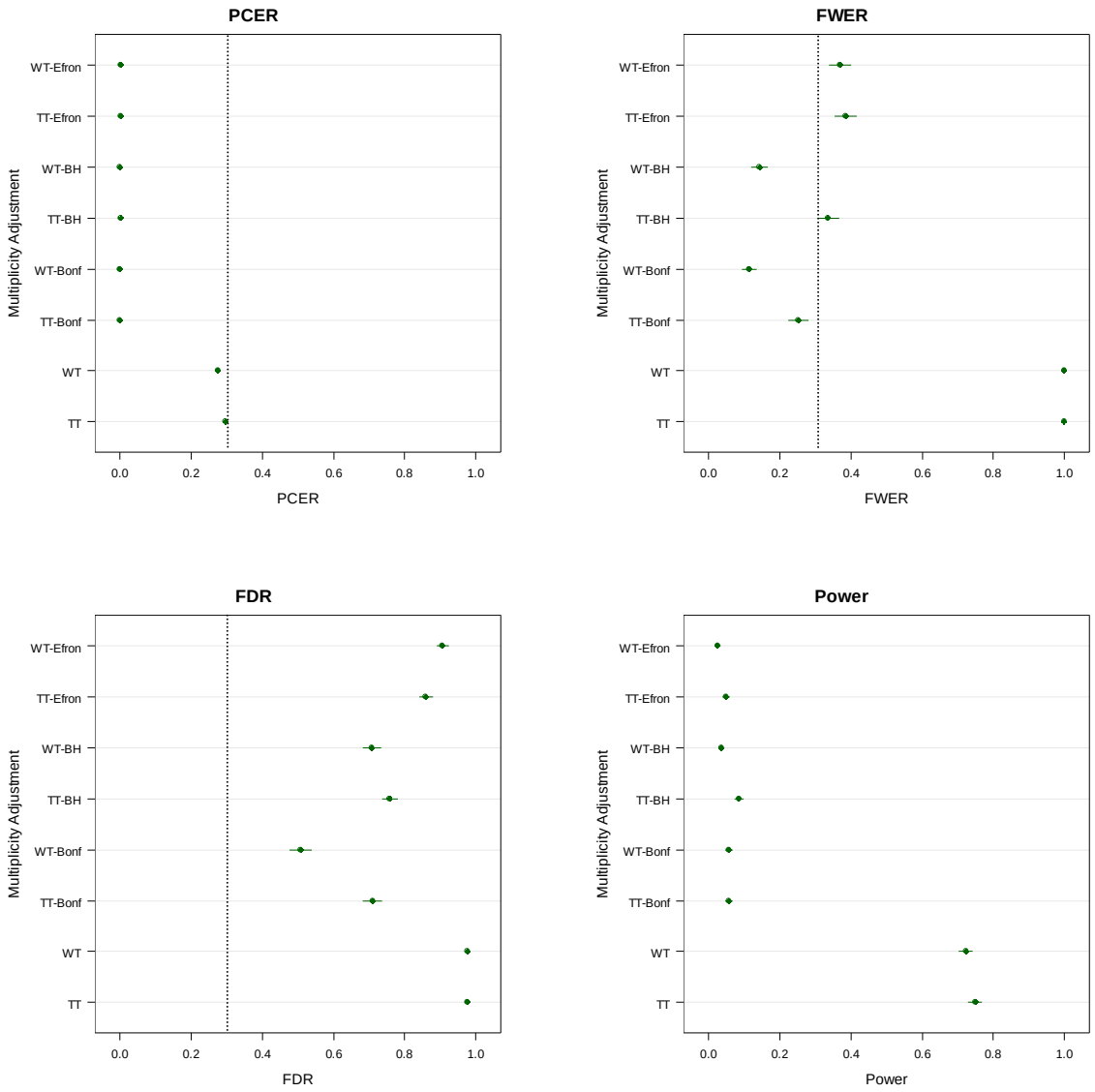

Figure F-4. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 1% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

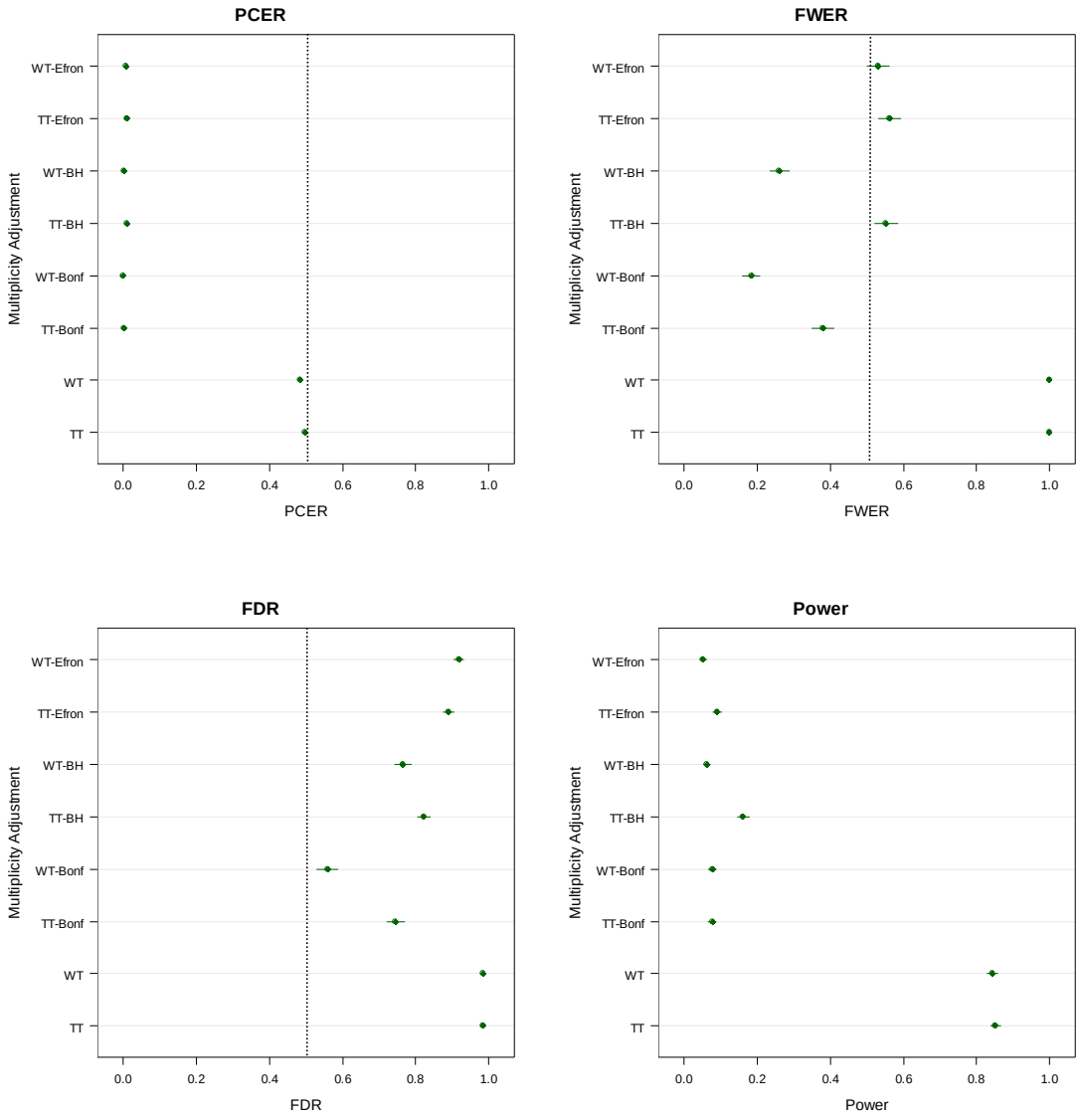

Figure F-5. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 1% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

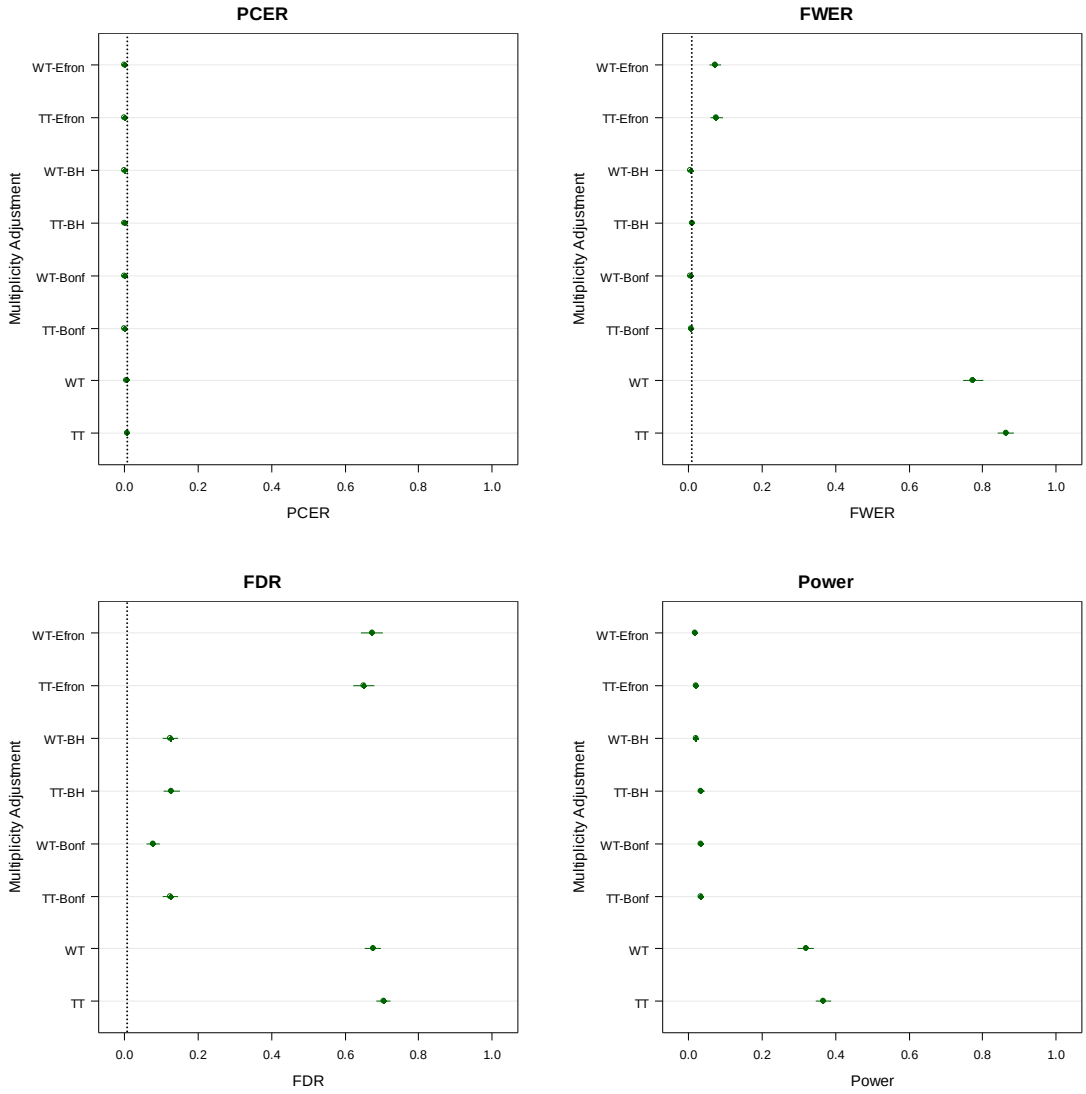

Figure F-6. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 1% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

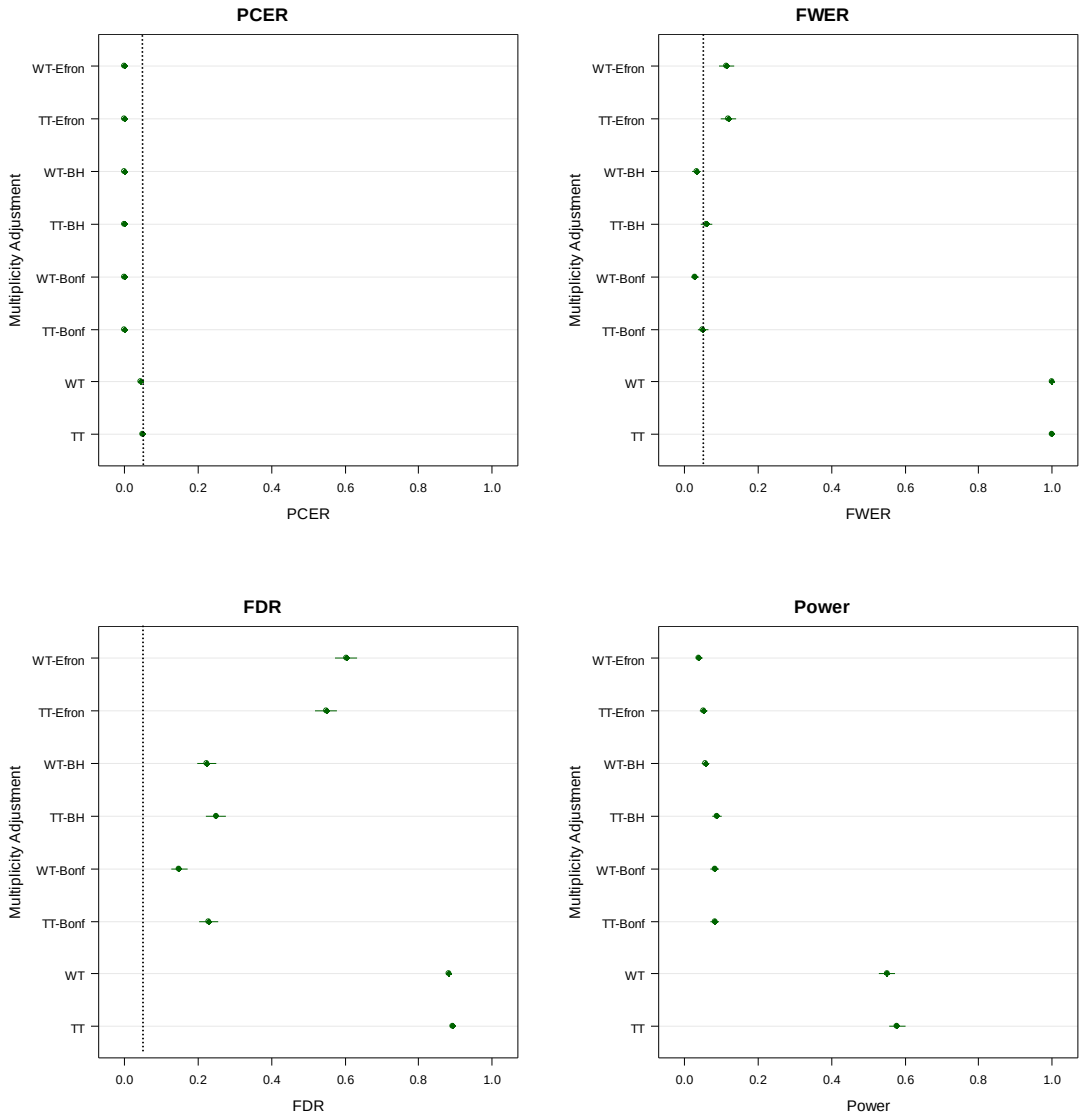

Figure F-7. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 1% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

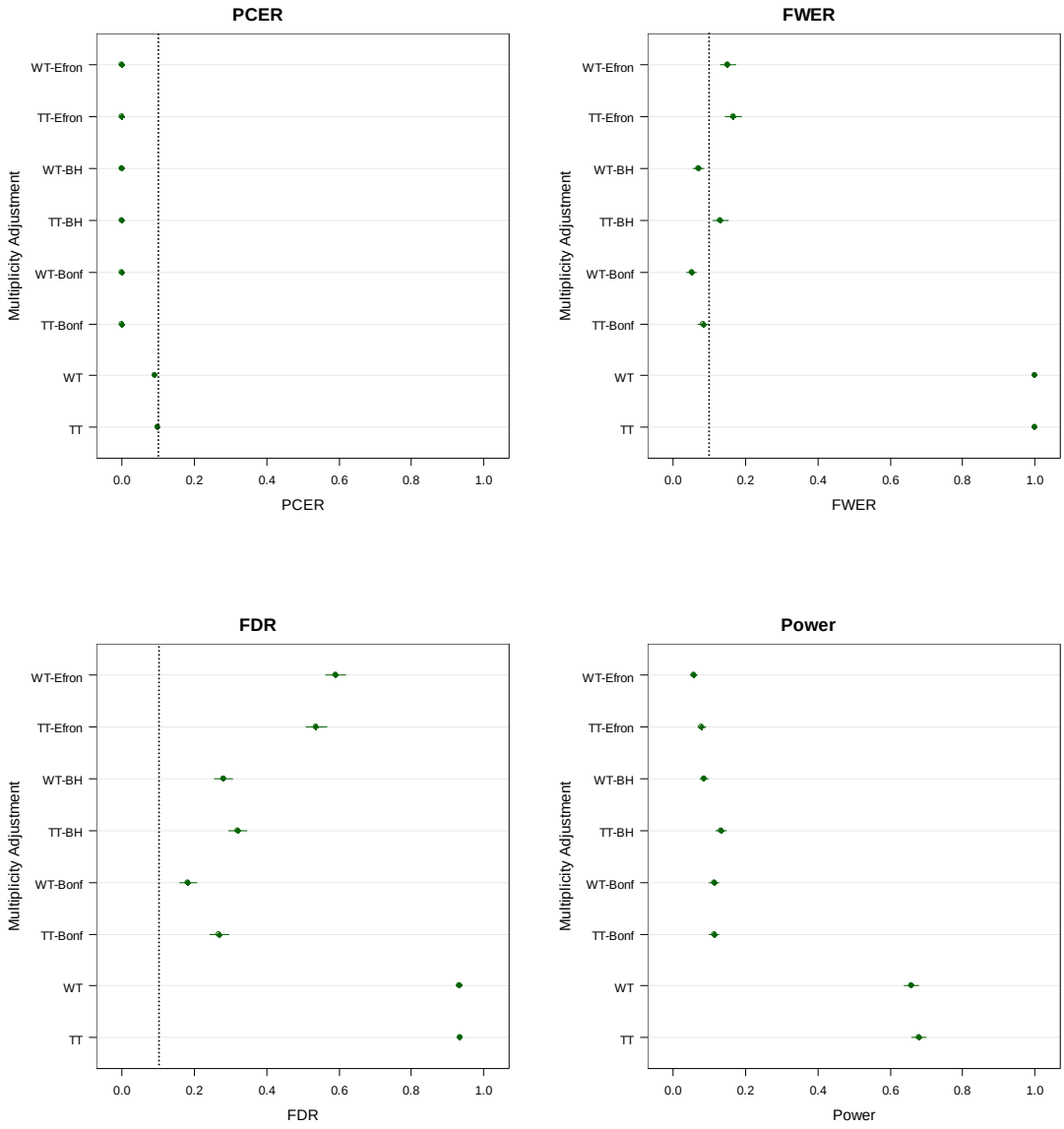

Figure F-8. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 1% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

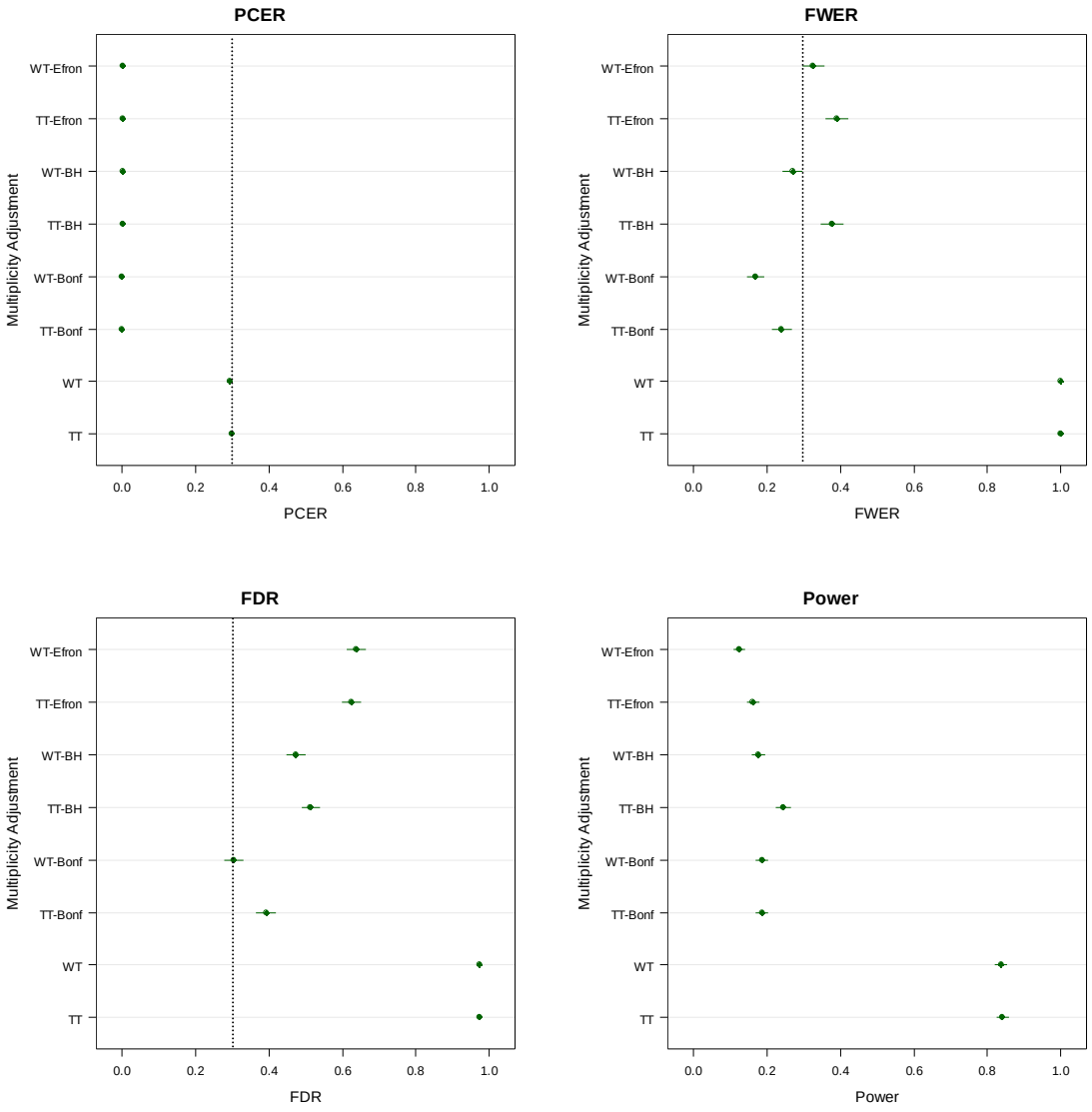

Figure F-9. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 1% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

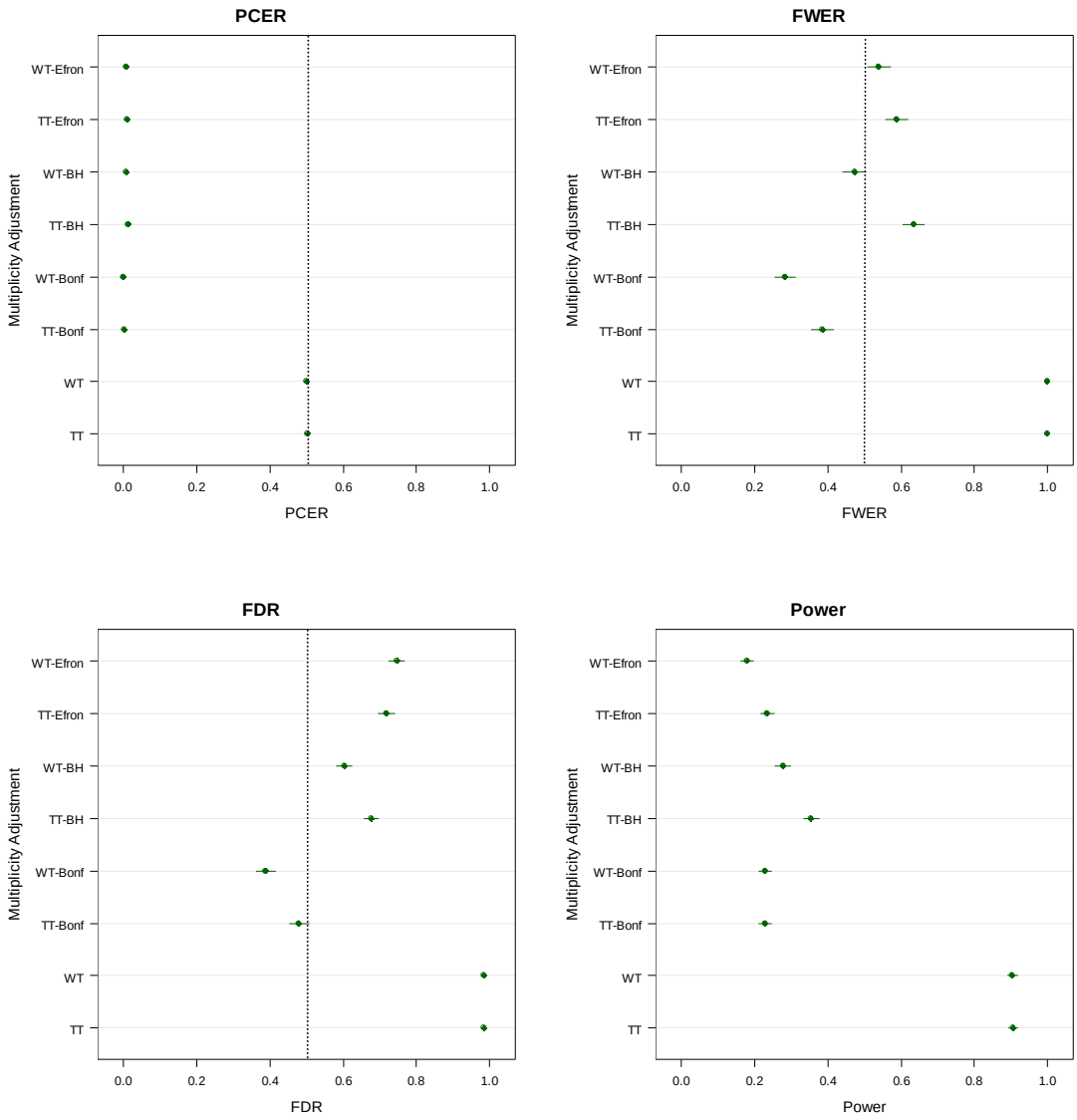

Figure F-10. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 1% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

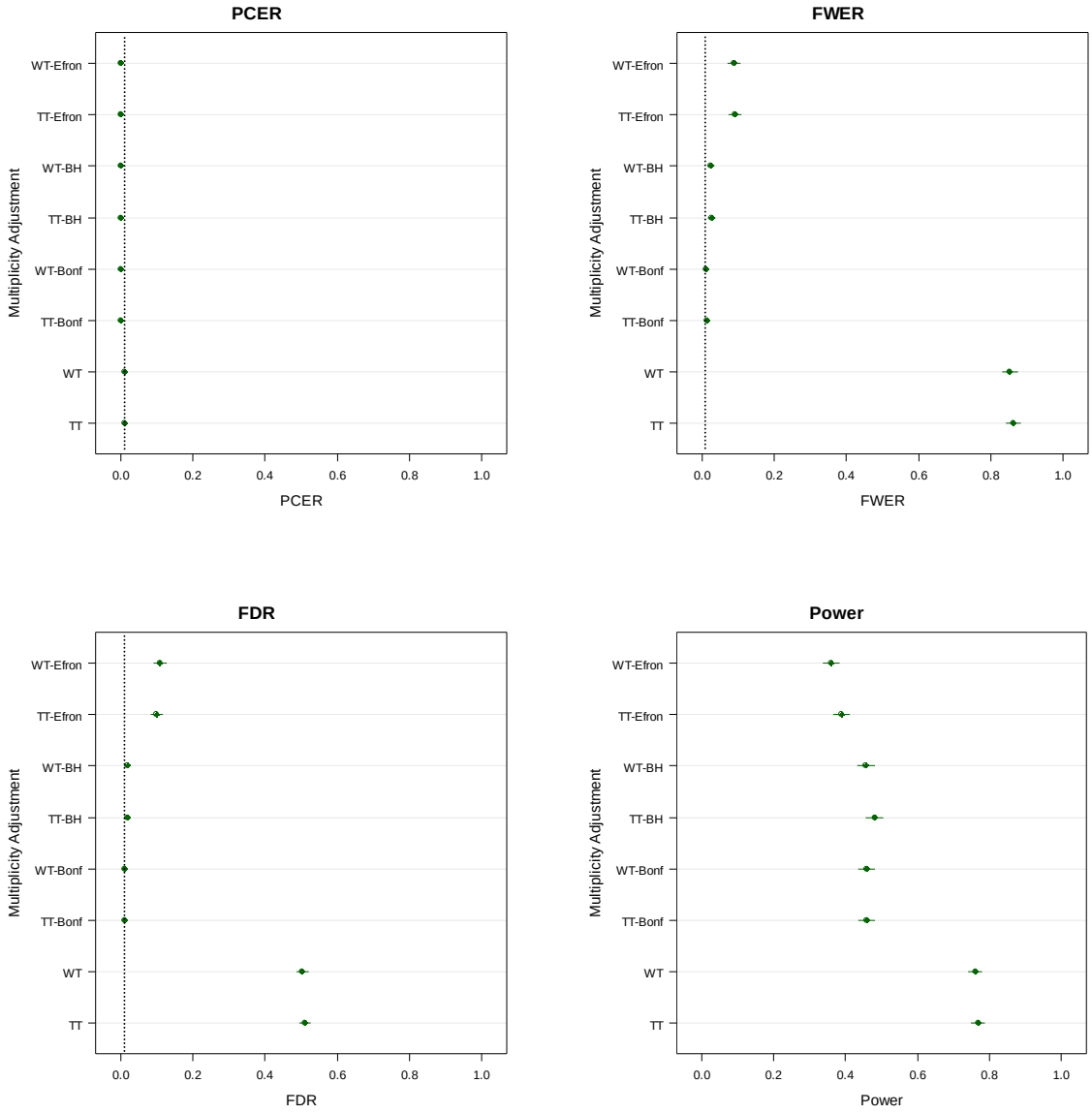

Figure F-11. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 1% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

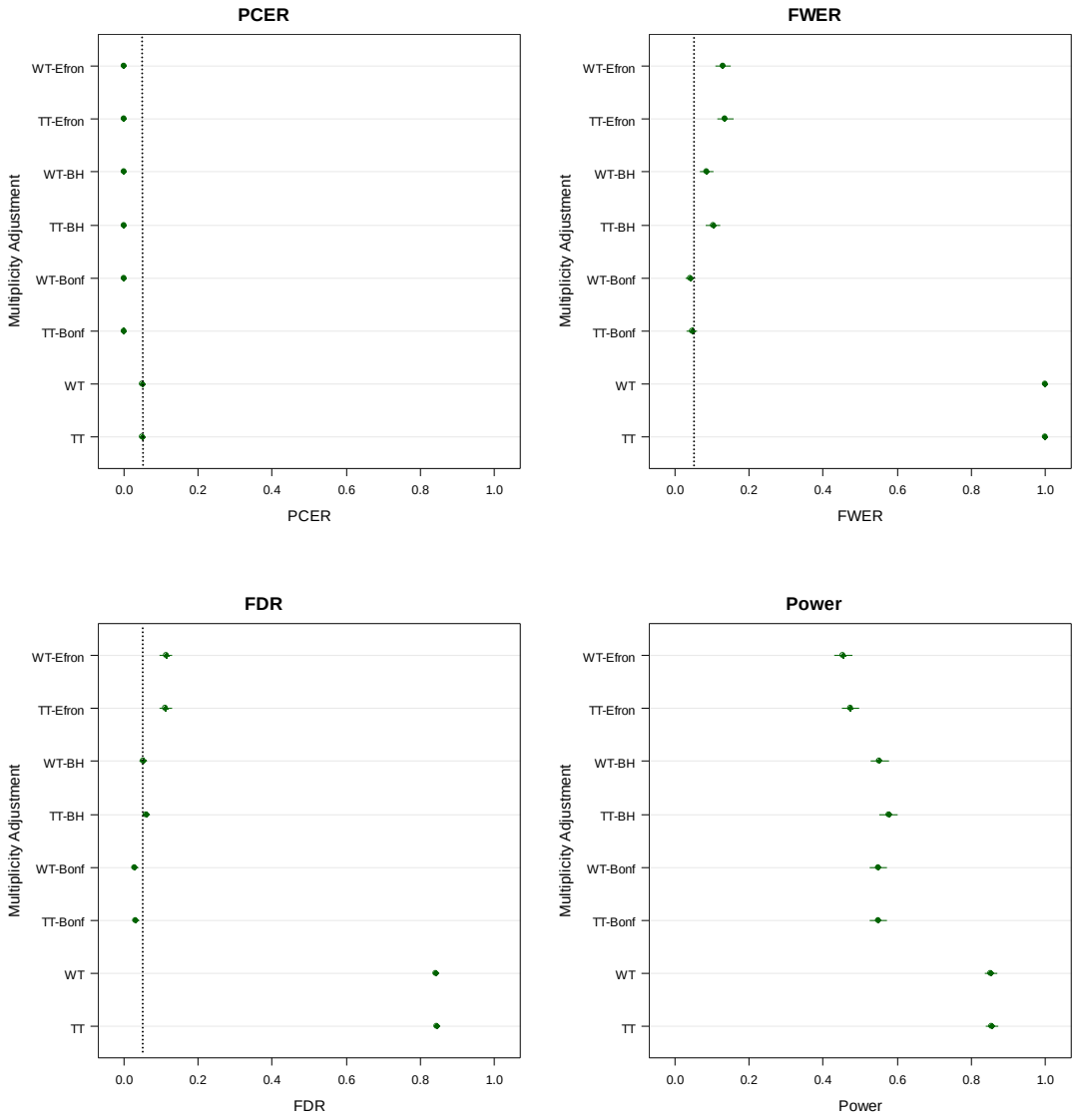

Figure F-12. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 1% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

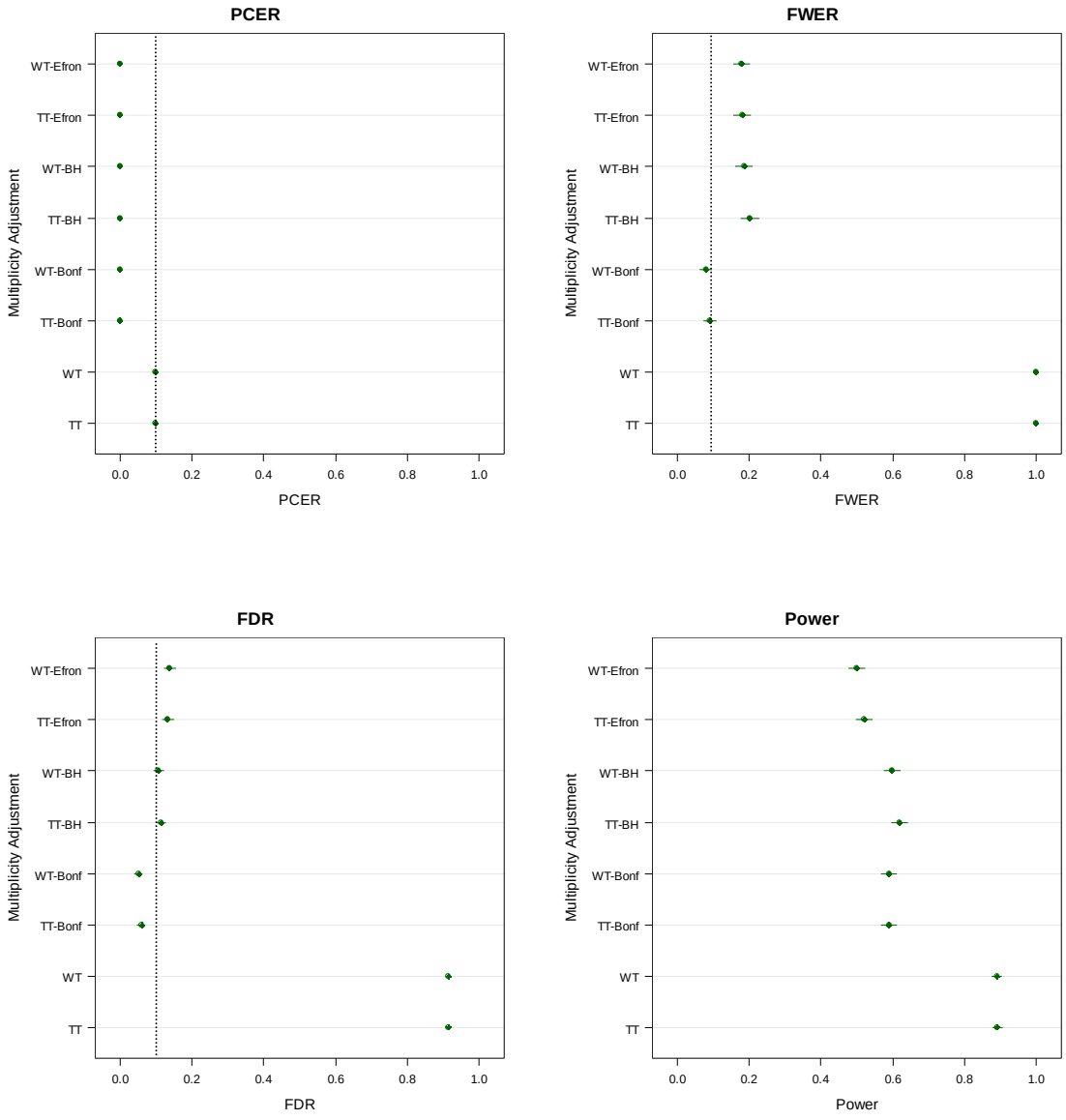

Figure F-13. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 1% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

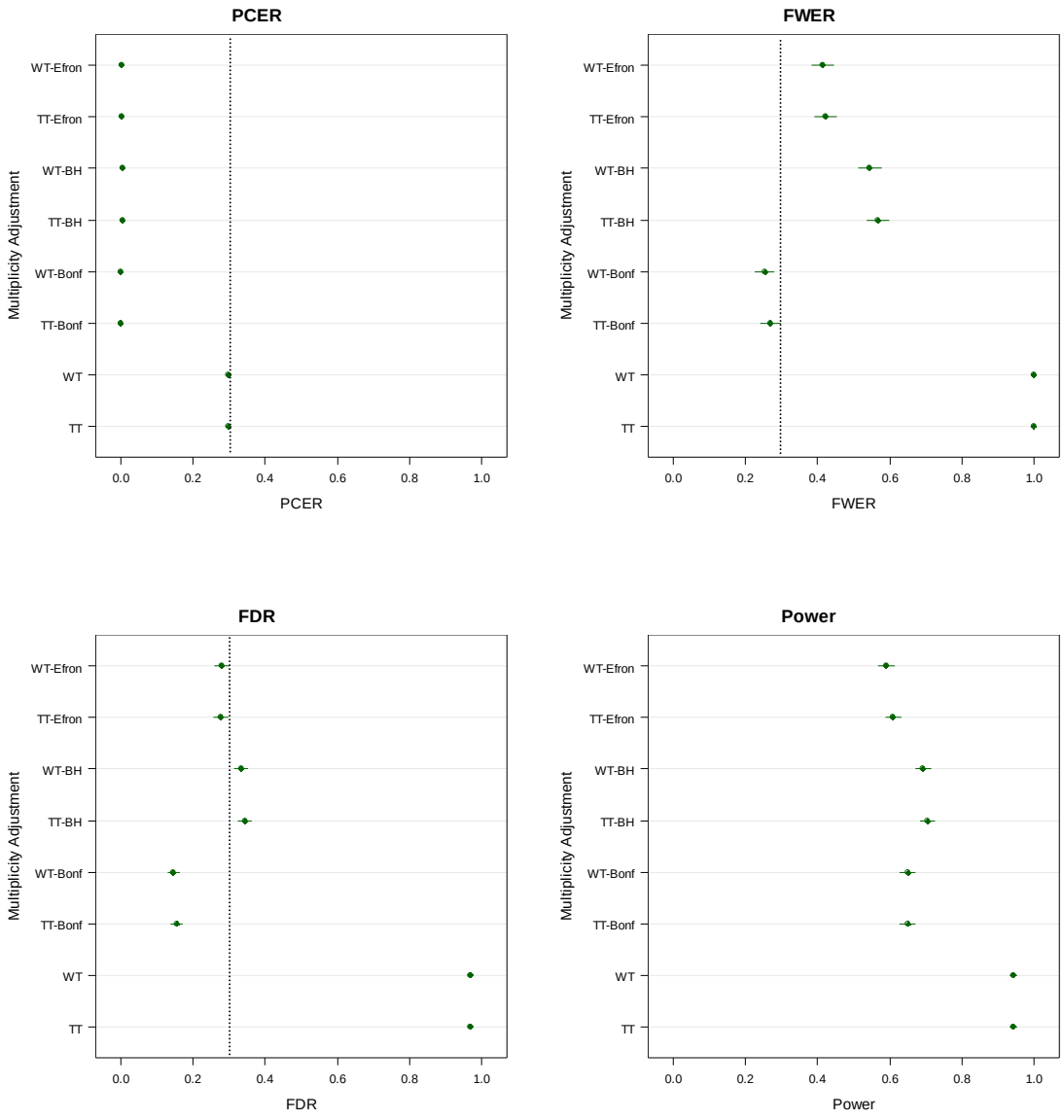

Figure F-14. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 1% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

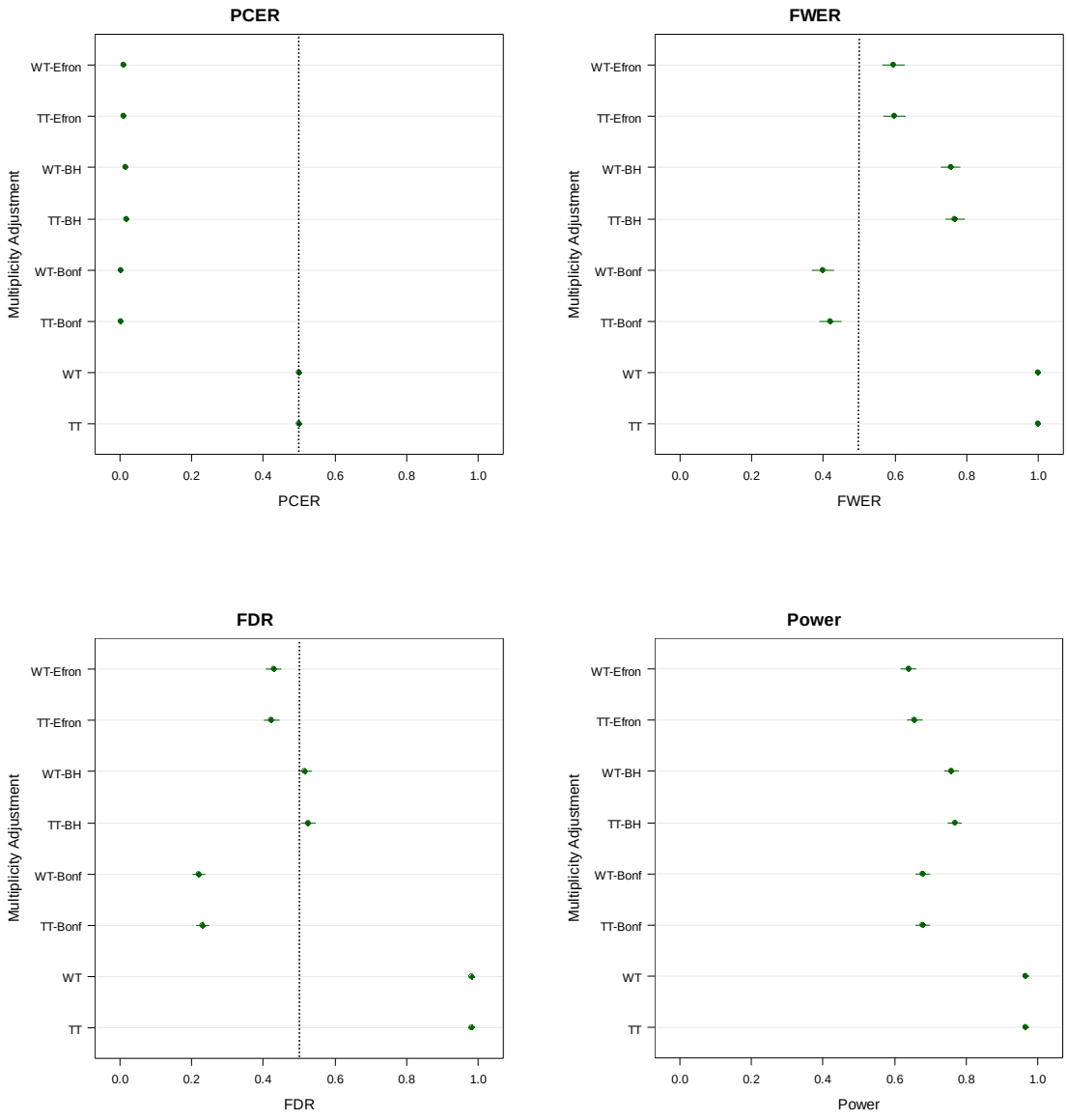

Figure F-15. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 1% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

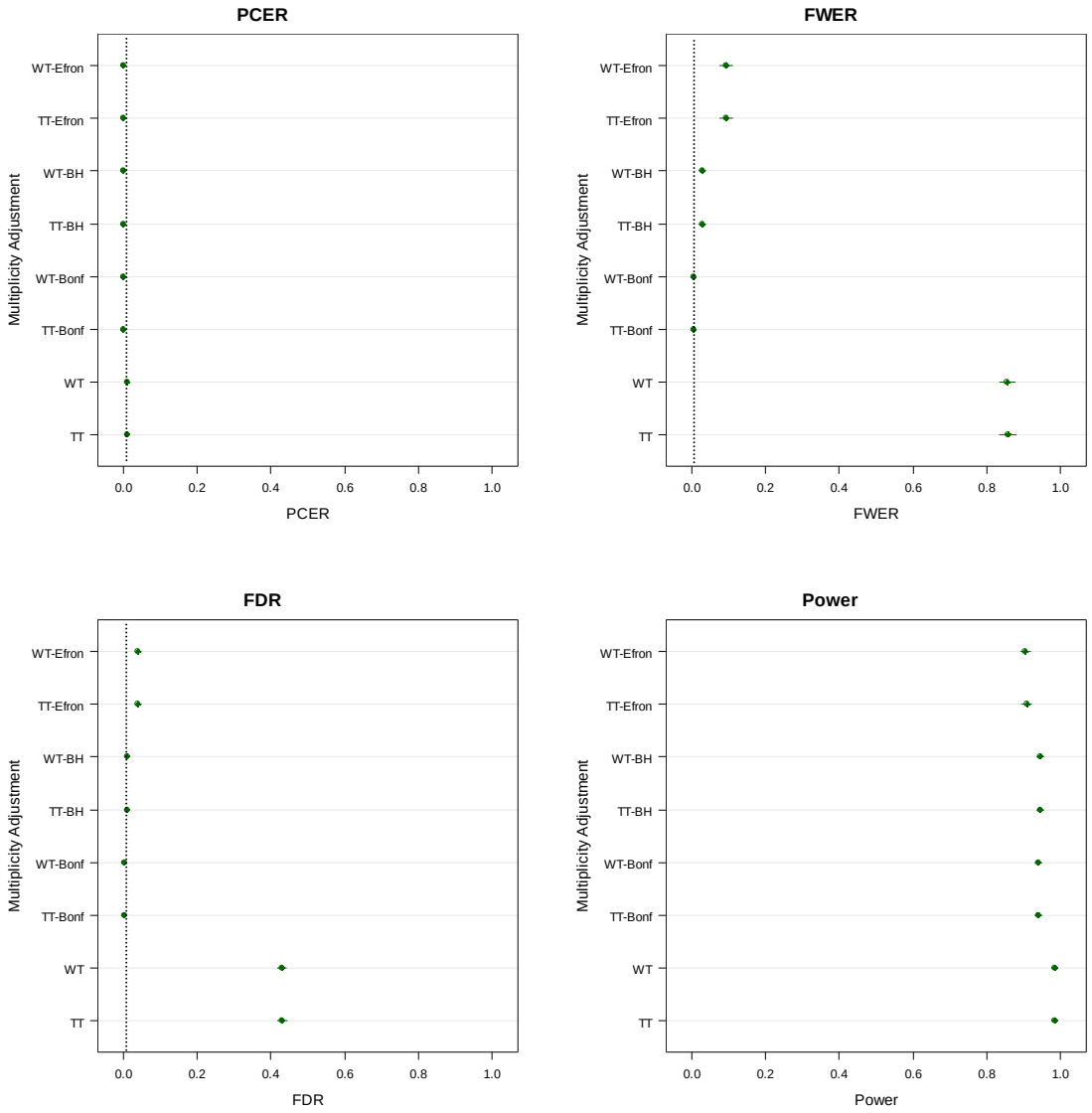

Figure F-16. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 1% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

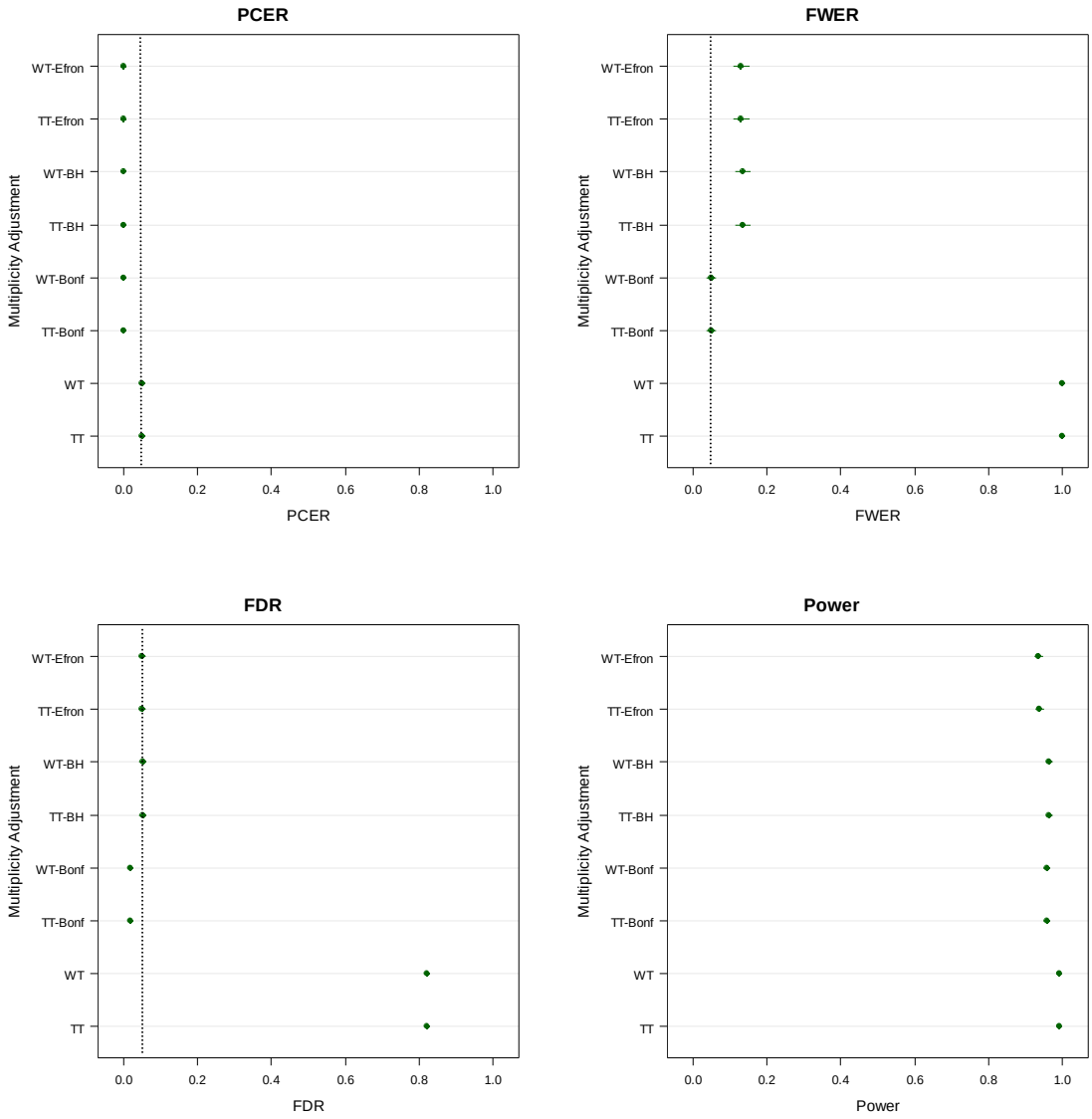

Figure F-17. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 1% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

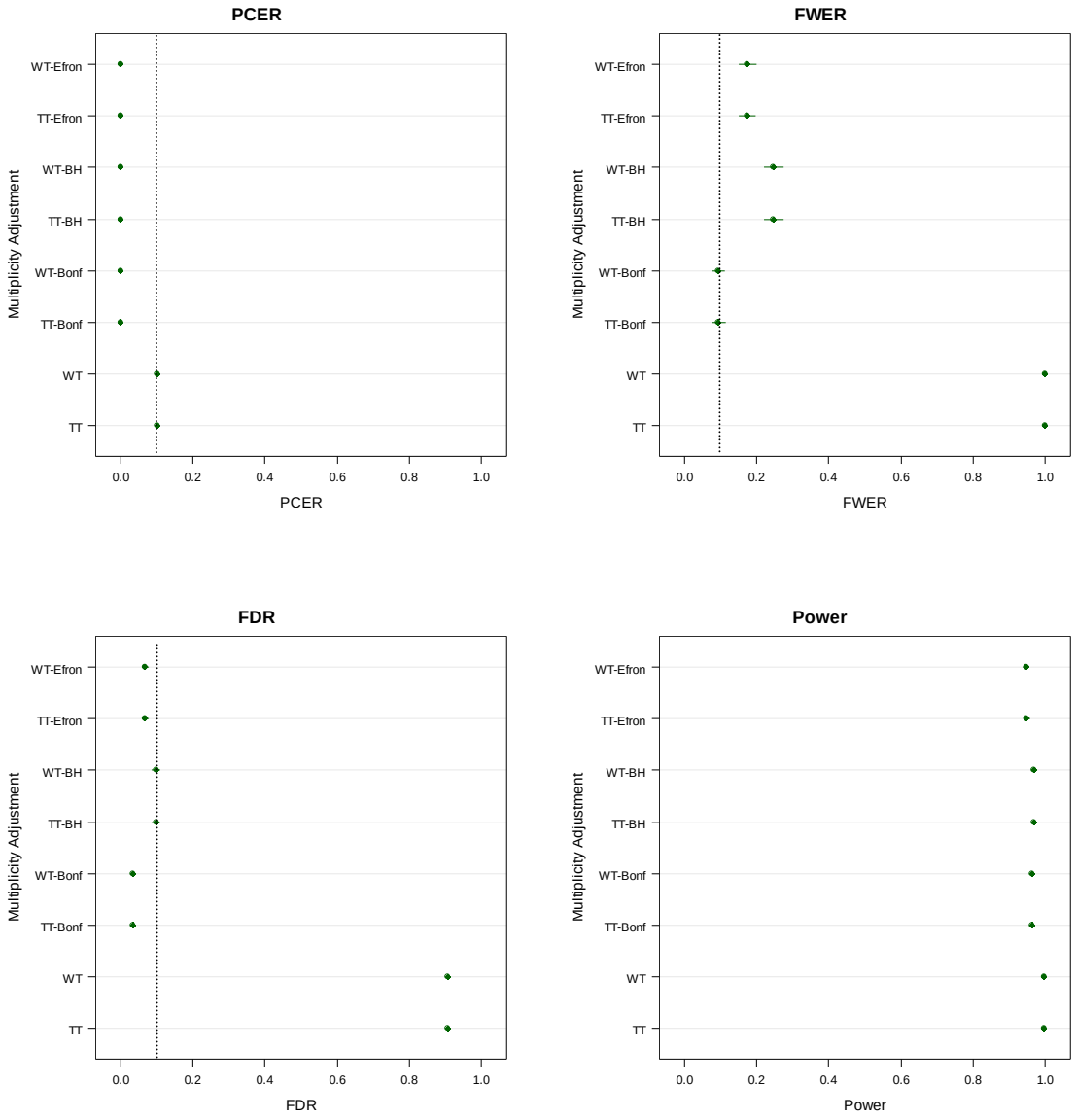

Figure F-18. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 1% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

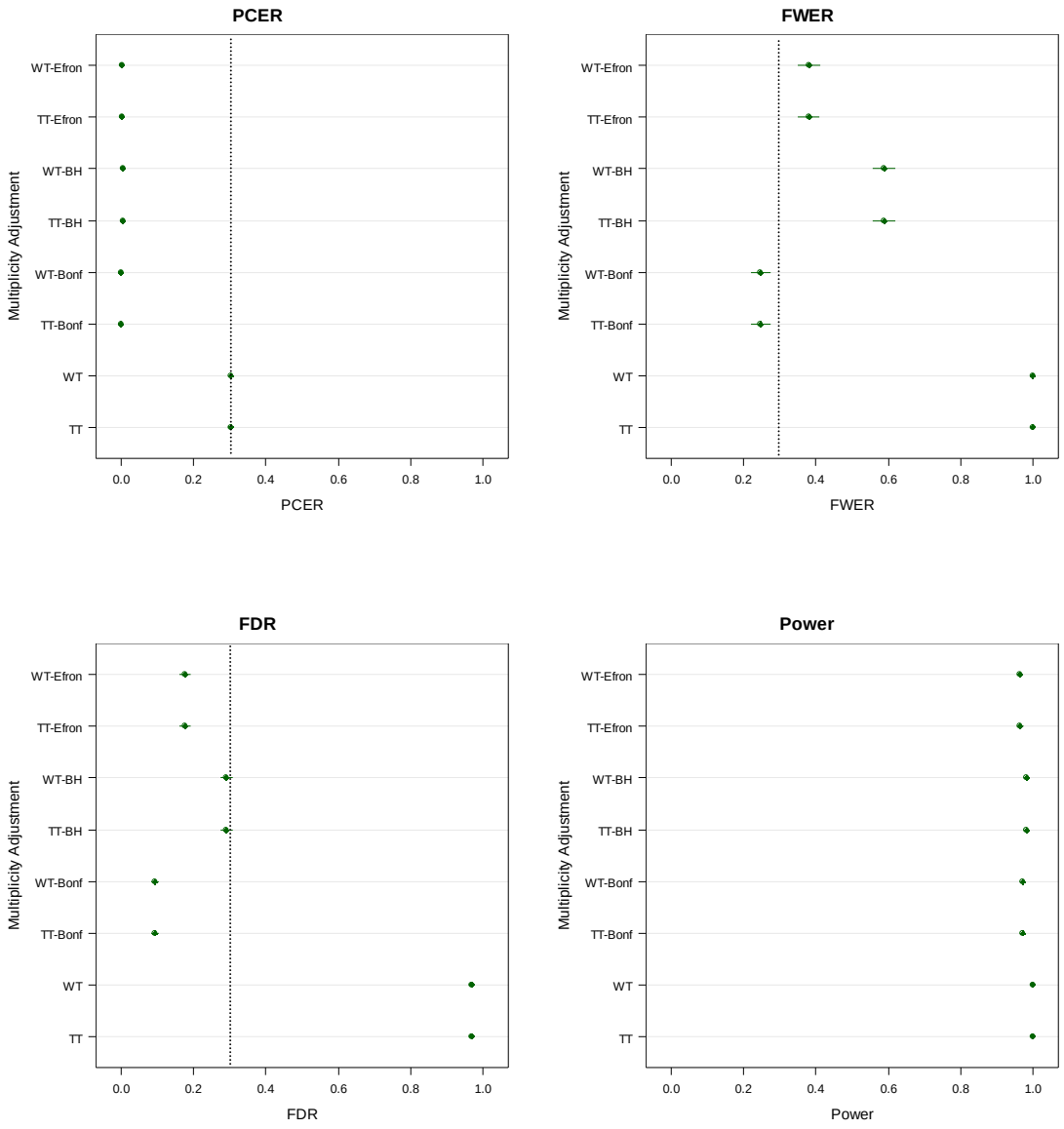

Figure F-19. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 1% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

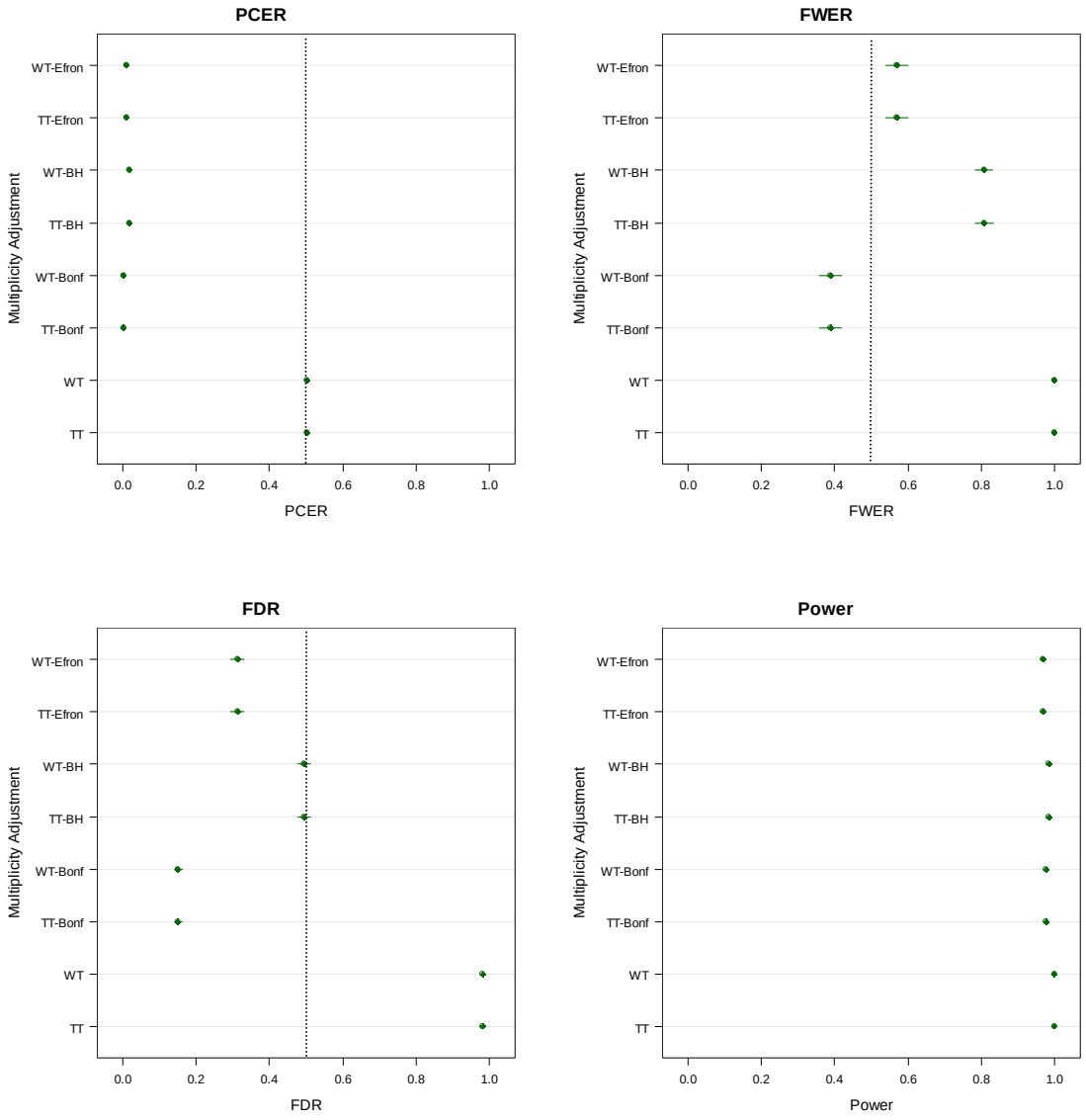

Figure F-20. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 1% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

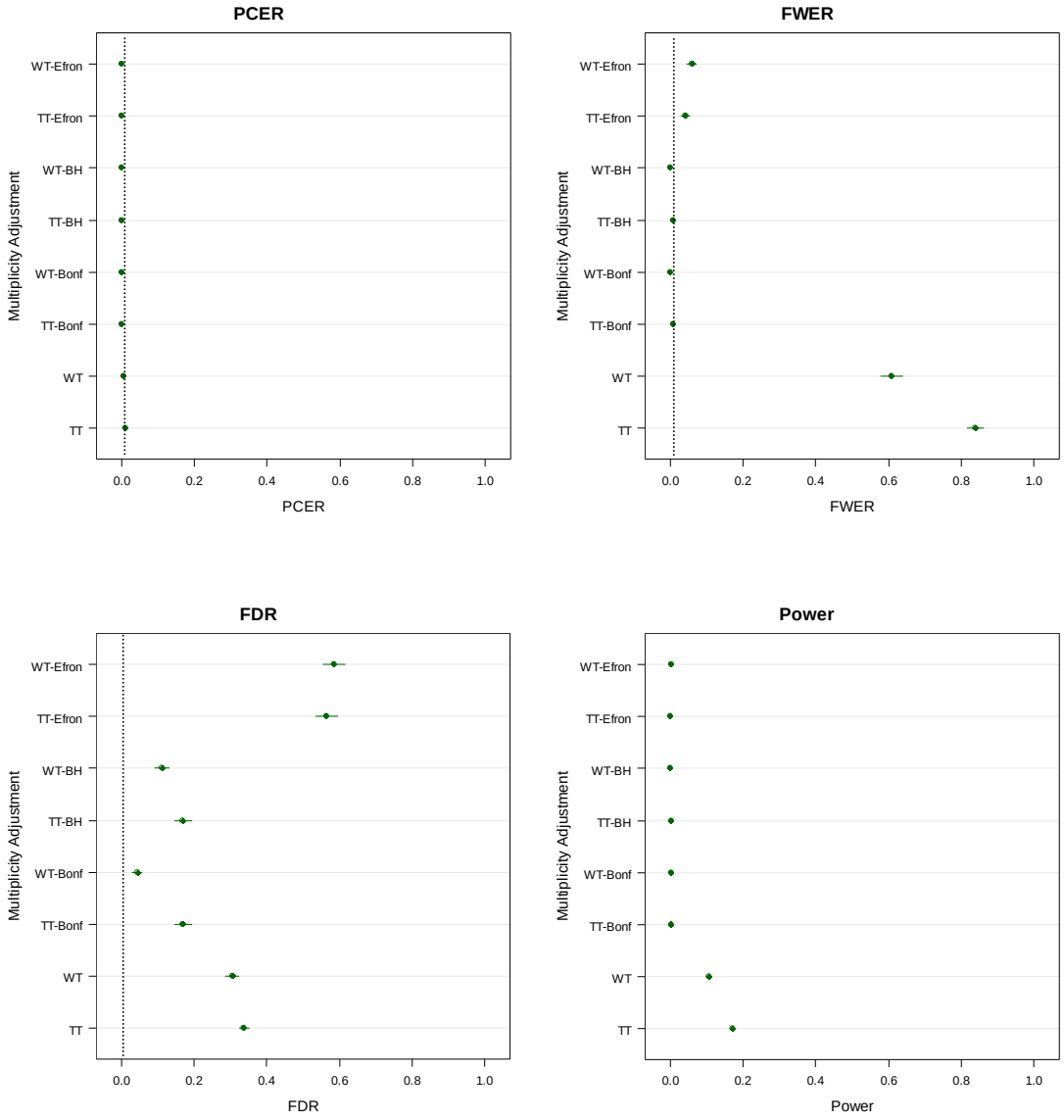

Figure F-21. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 10% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

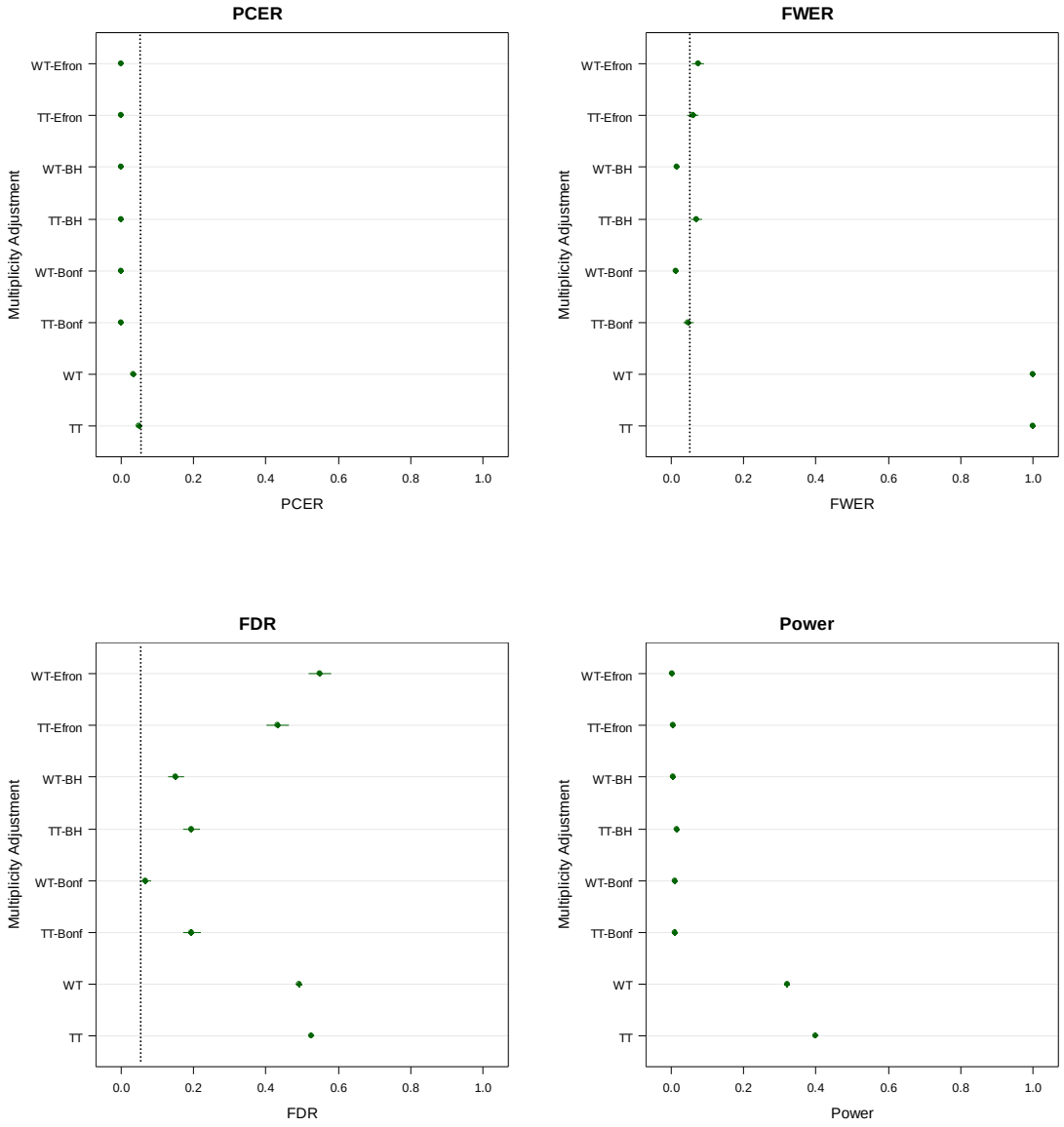

Figure F-22. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 10% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

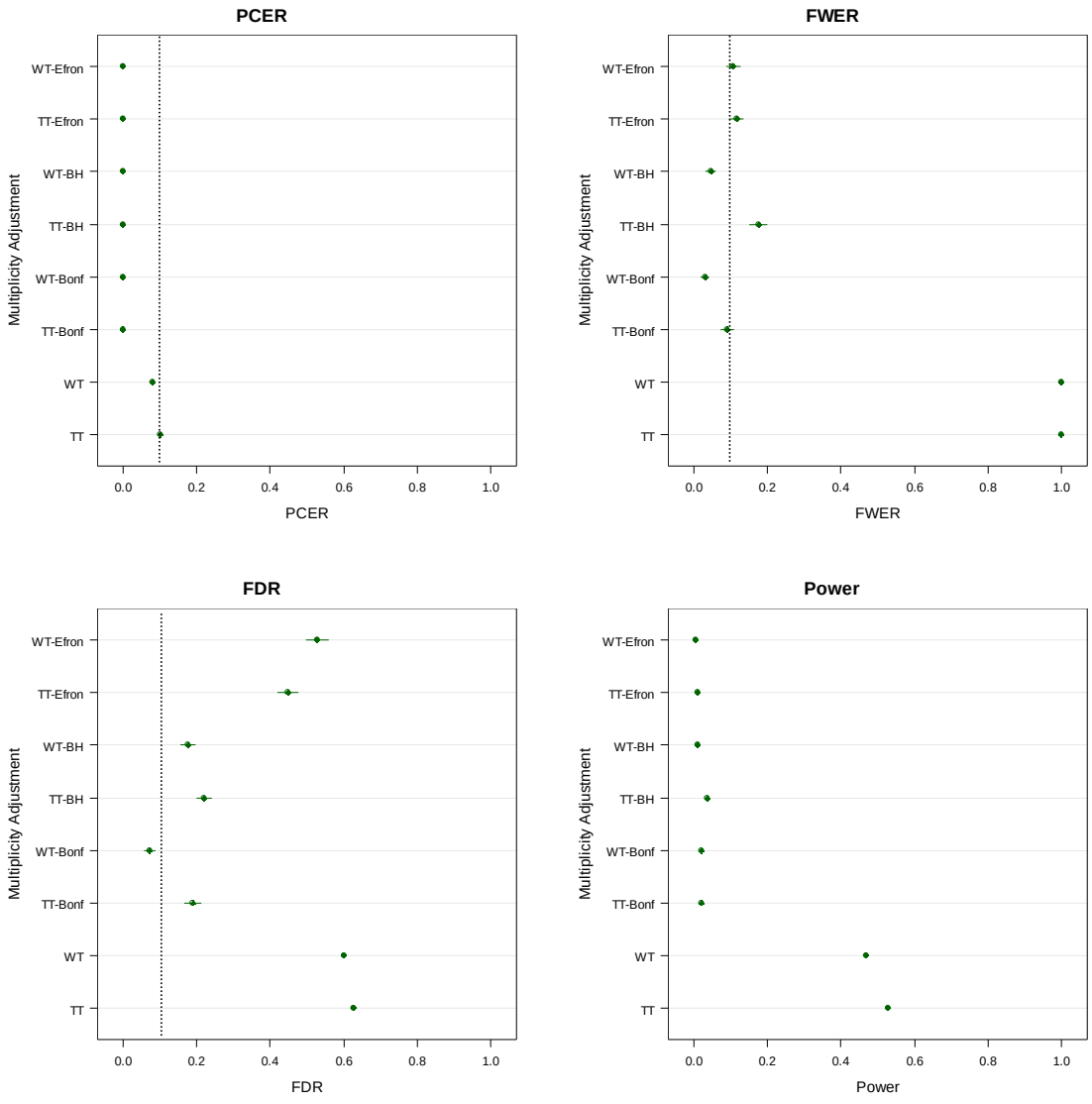

Figure F-23. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 10% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

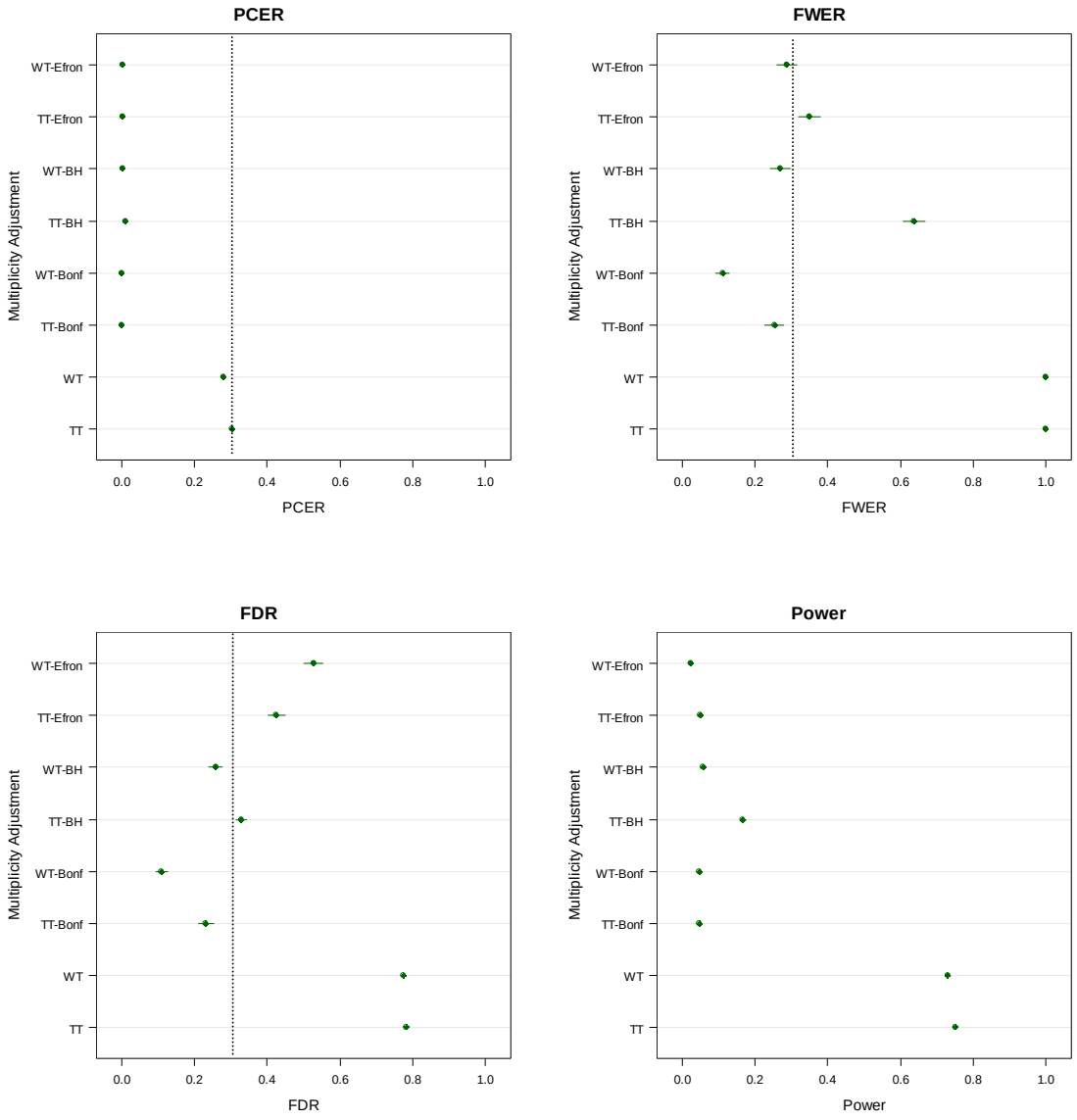

Figure F-24. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 10% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

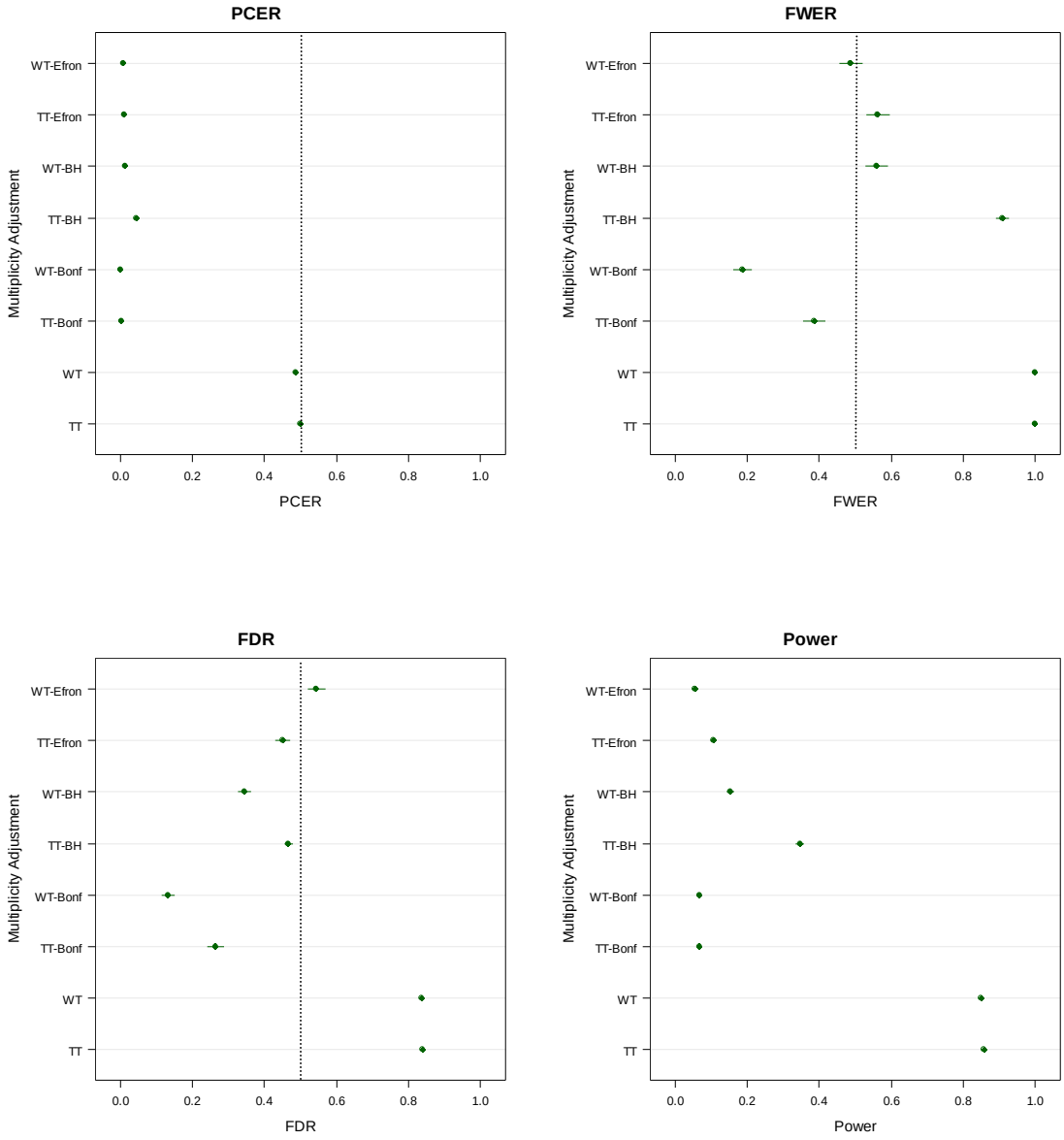

Figure F-25. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 10% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

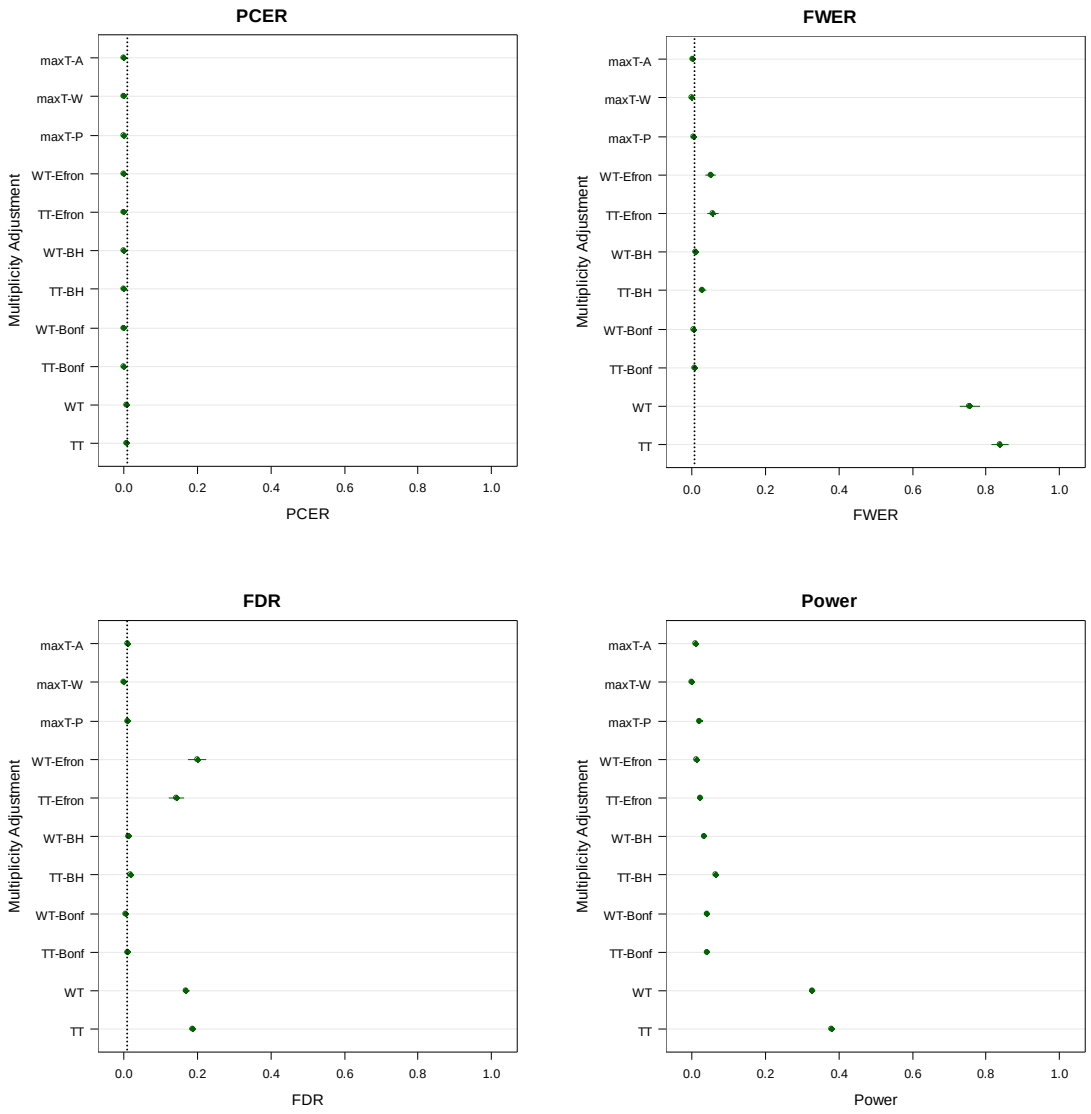

Figure F-26. Simulation Results (0.01); Error and Power Summary Comparing Eleven Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

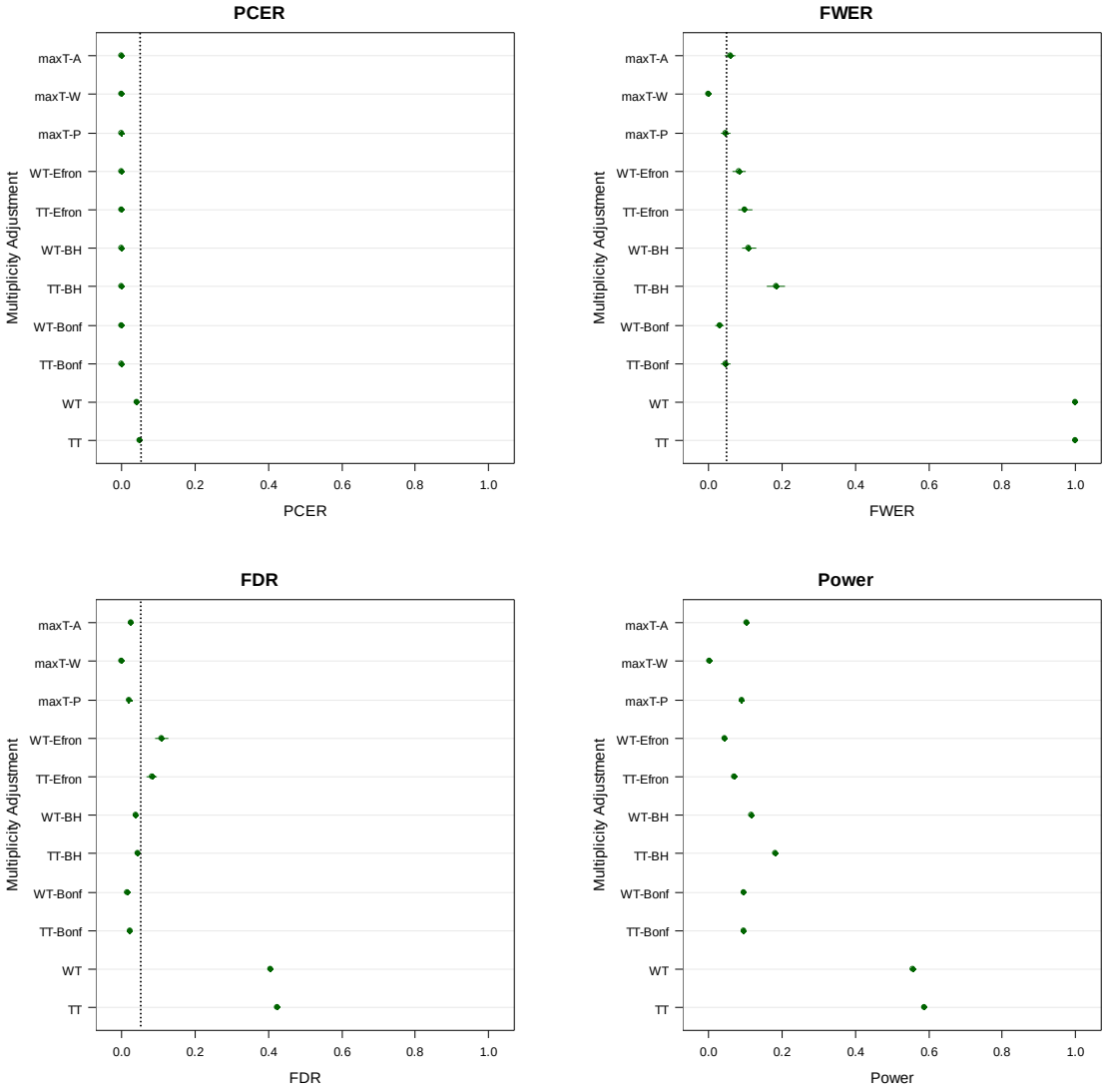

Figure F-27. Simulation Results (0.05); Error and Power Summary Comparing Eleven Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

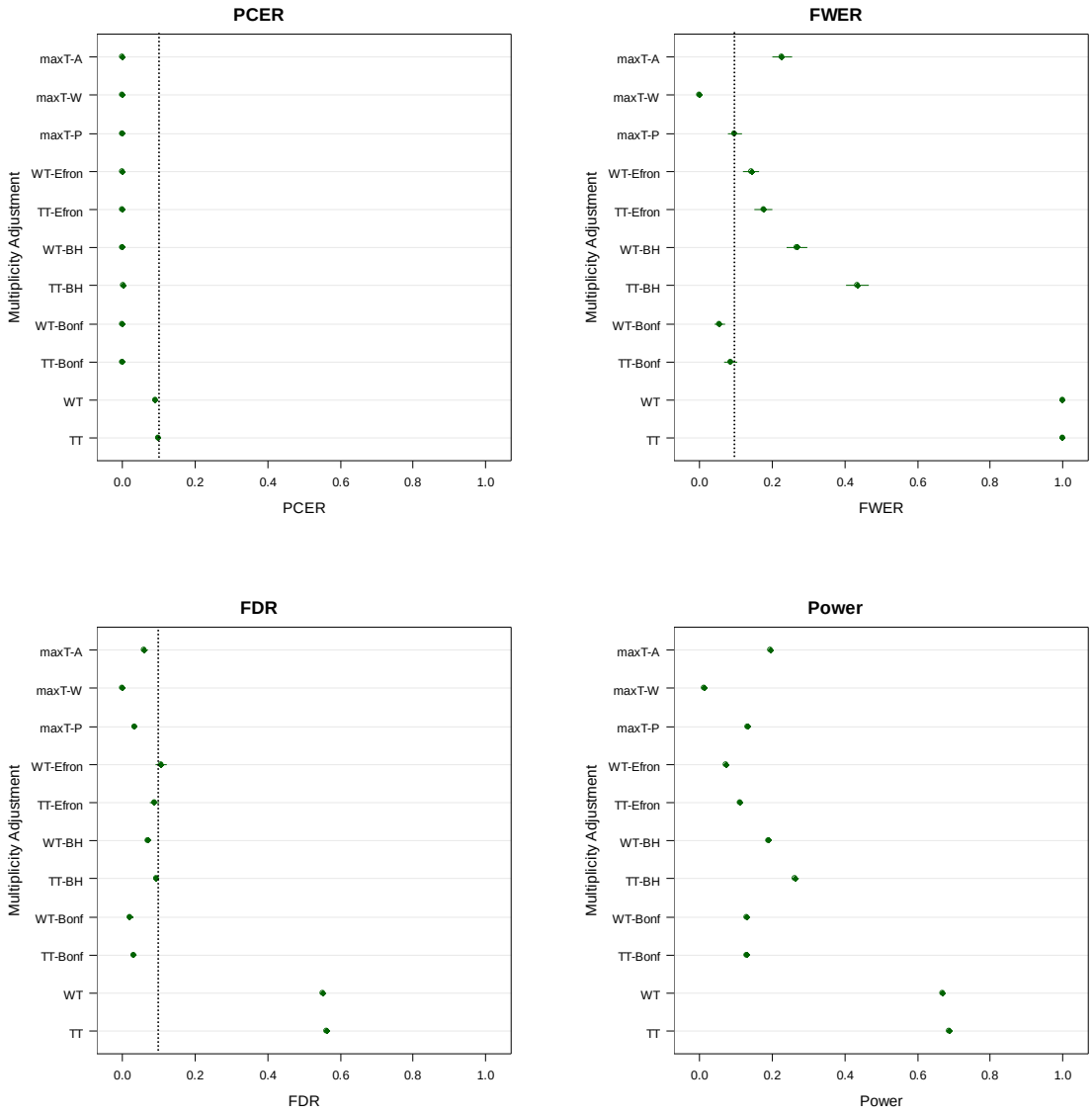

Figure F-28. Simulation Results (0.10); Error and Power Summary Comparing Eleven Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

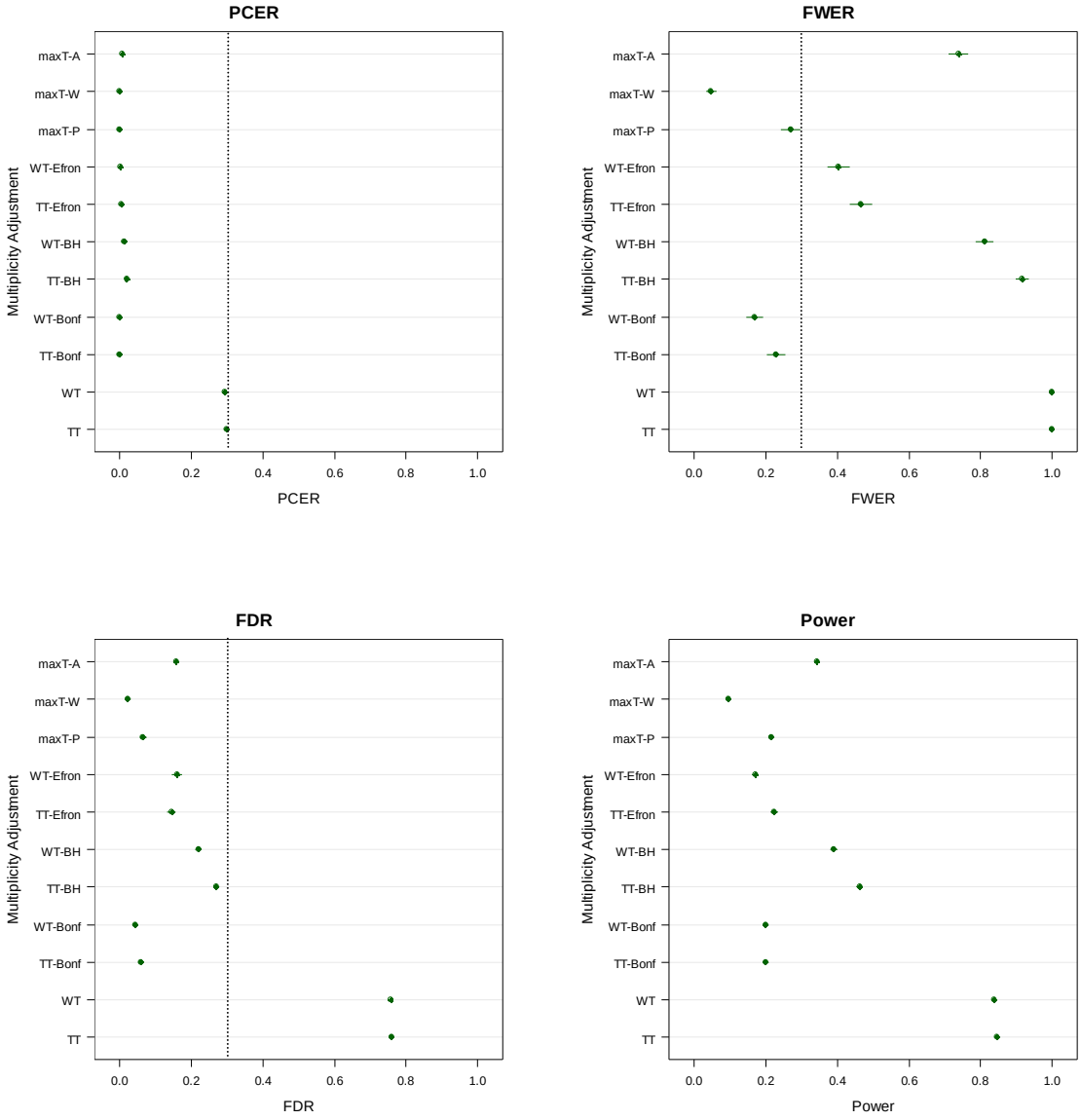

Figure F-29. Simulation Results (0.30); Error and Power Summary Comparing Eleven Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

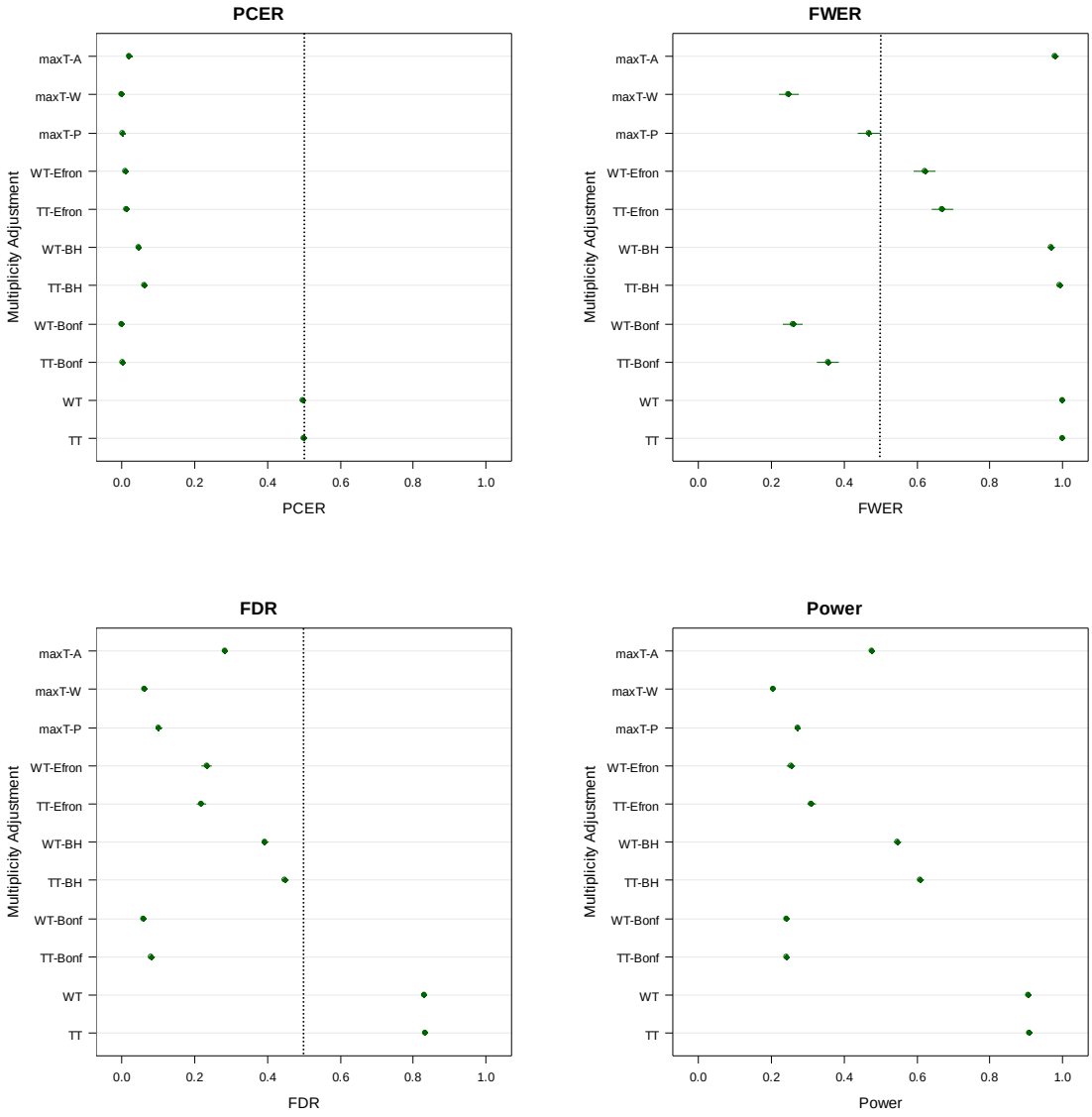

Figure F-30. Simulation Results (0.50); Error and Power Summary Comparing Eleven Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

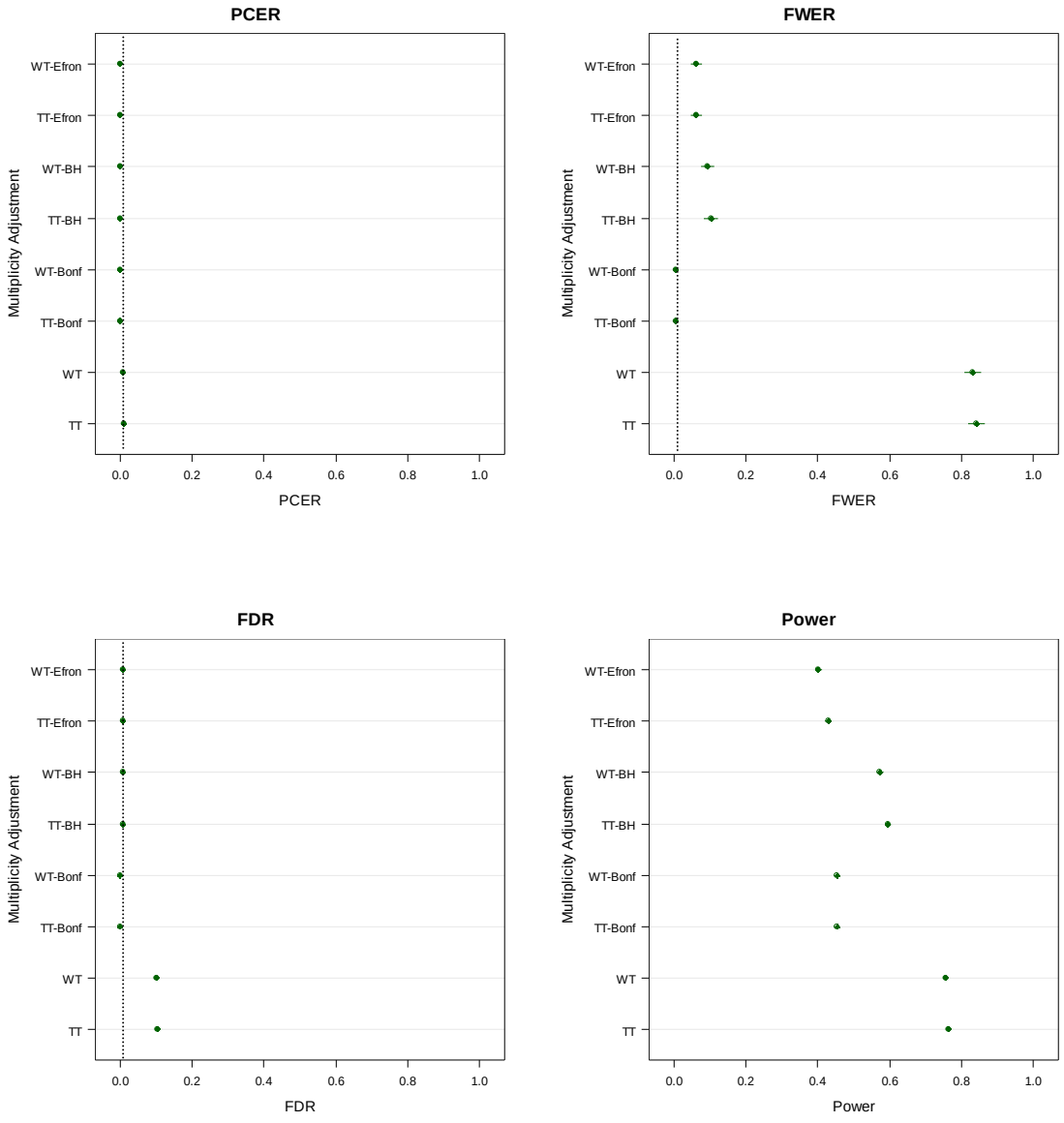

Figure F-31. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

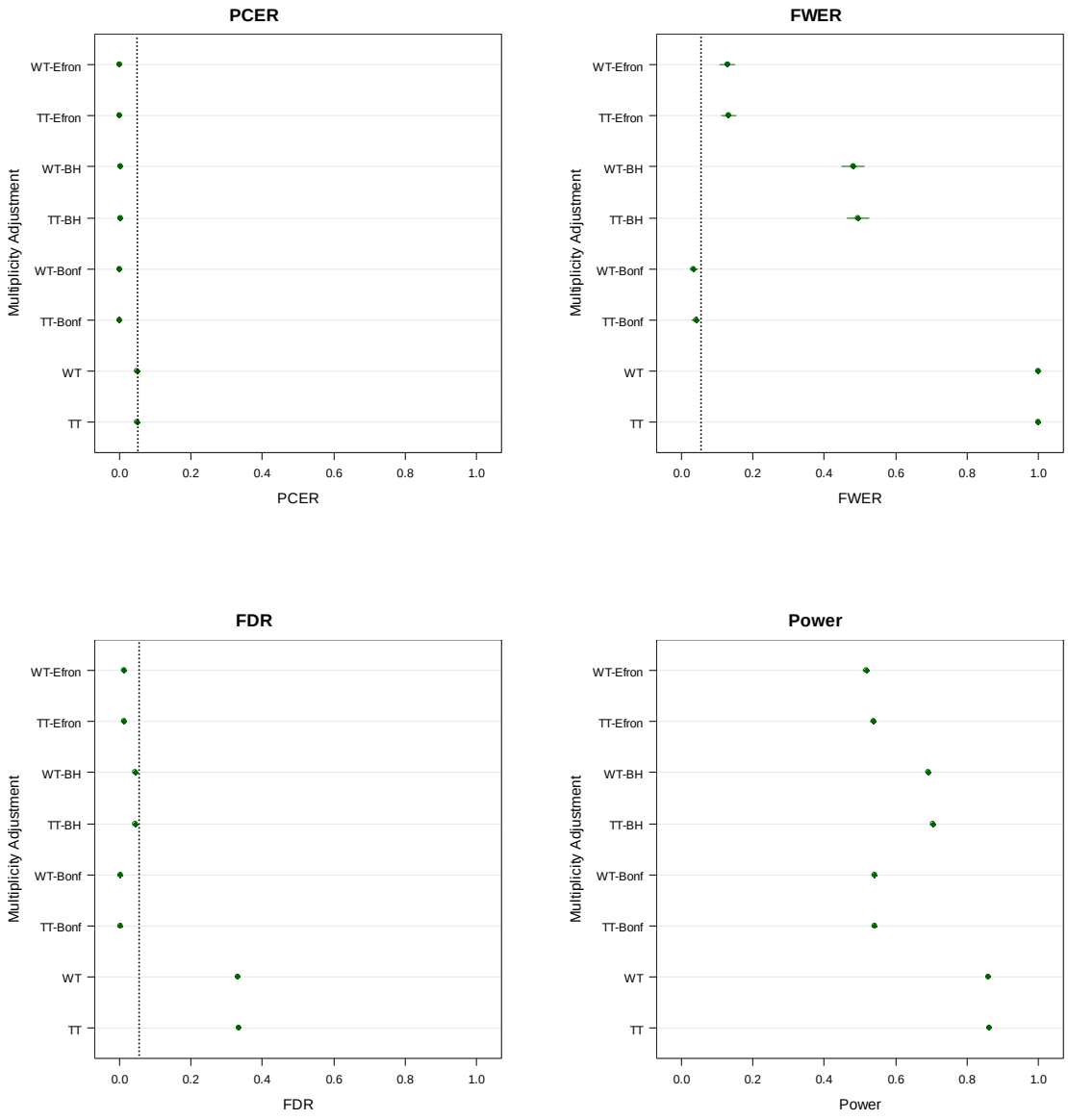

Figure F-32. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

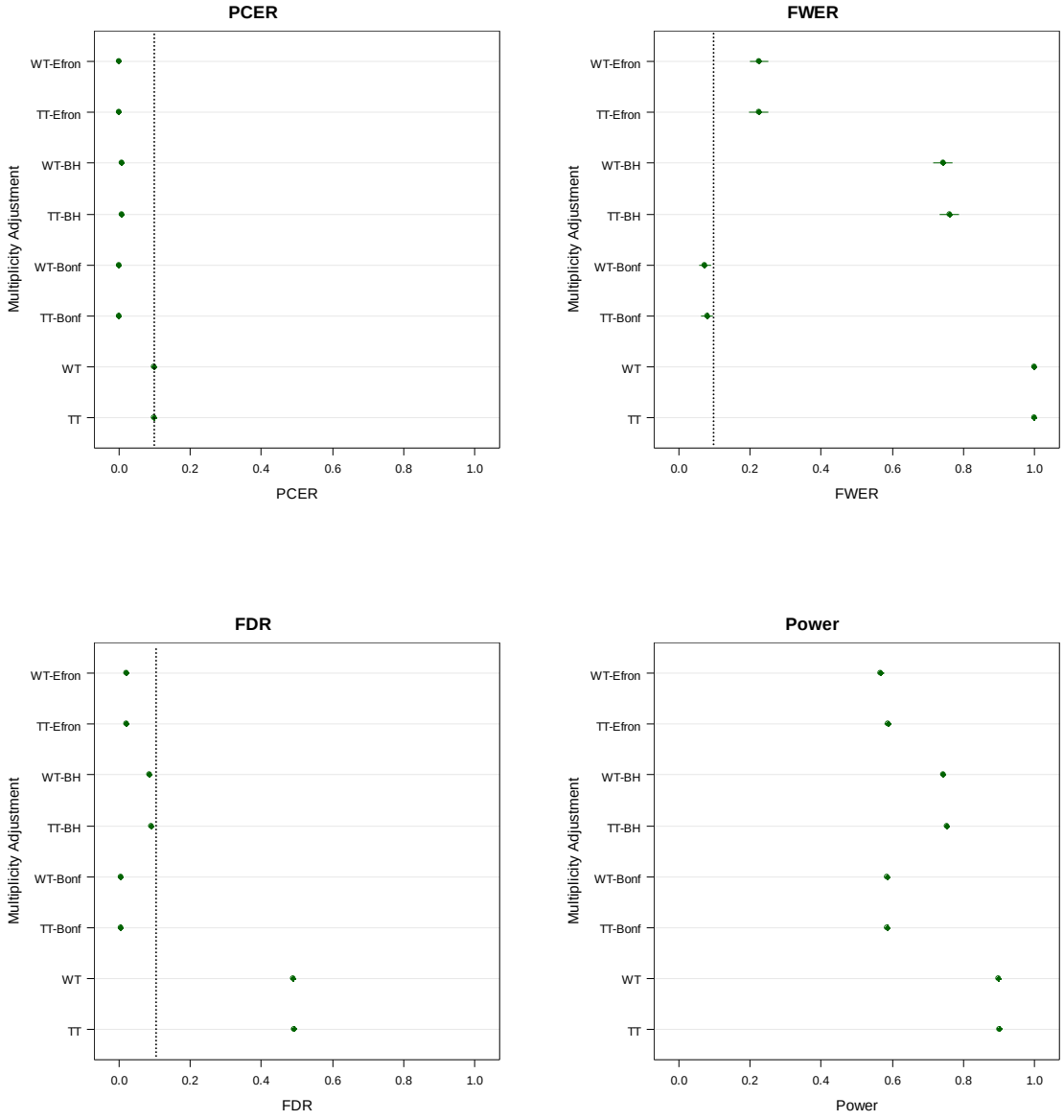

Figure F-33. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

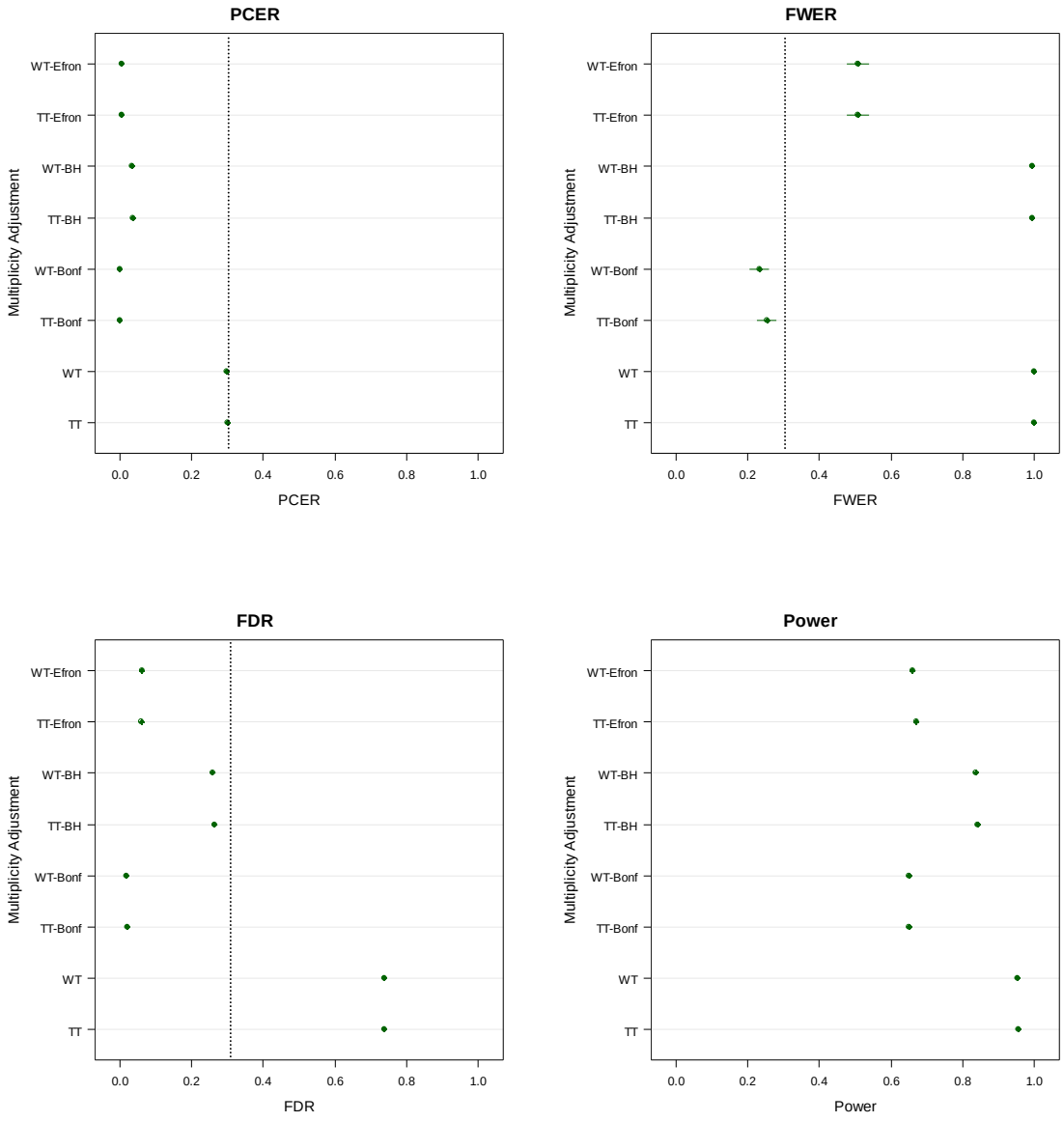

Figure F-34. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

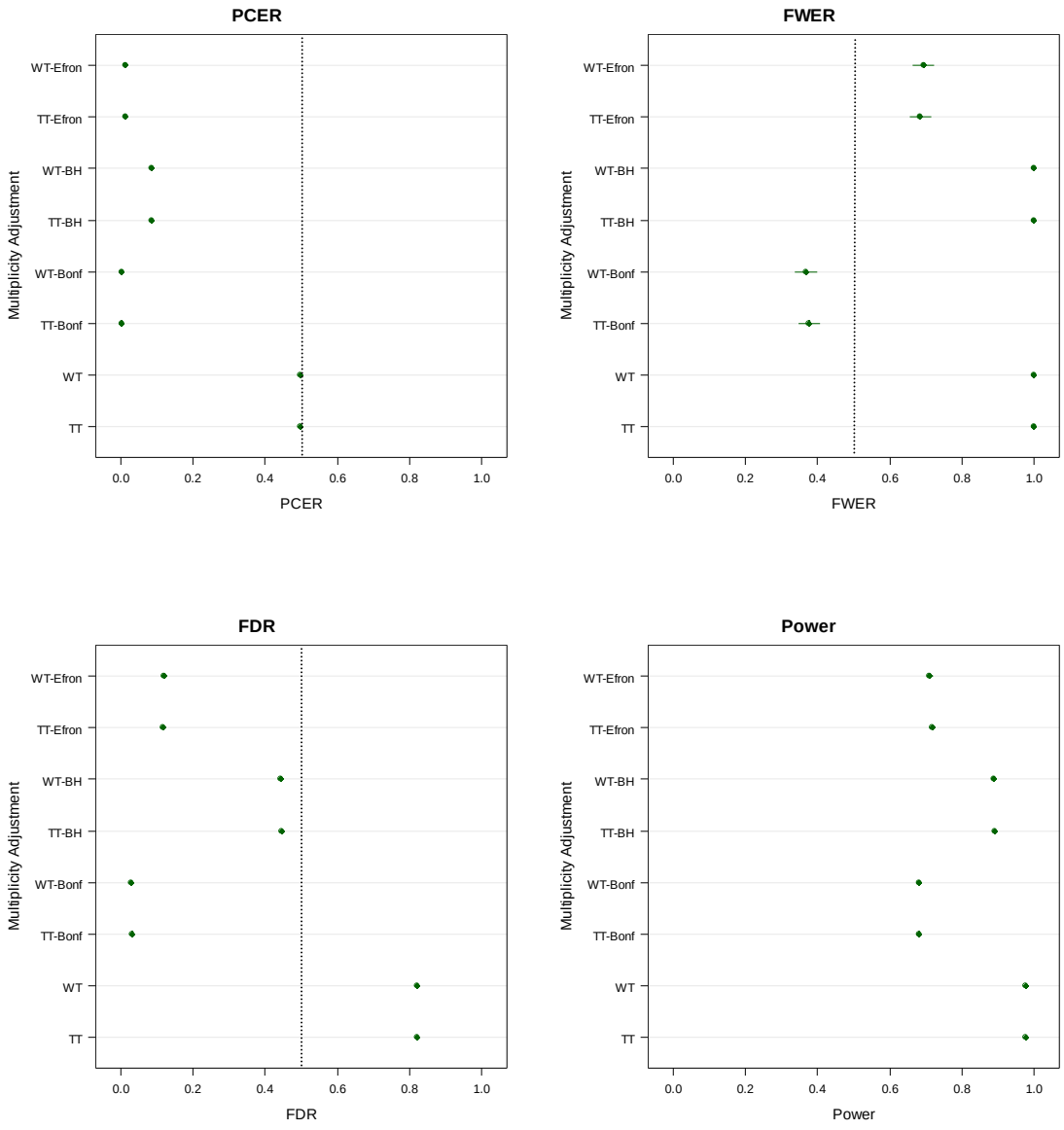

Figure F-35. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

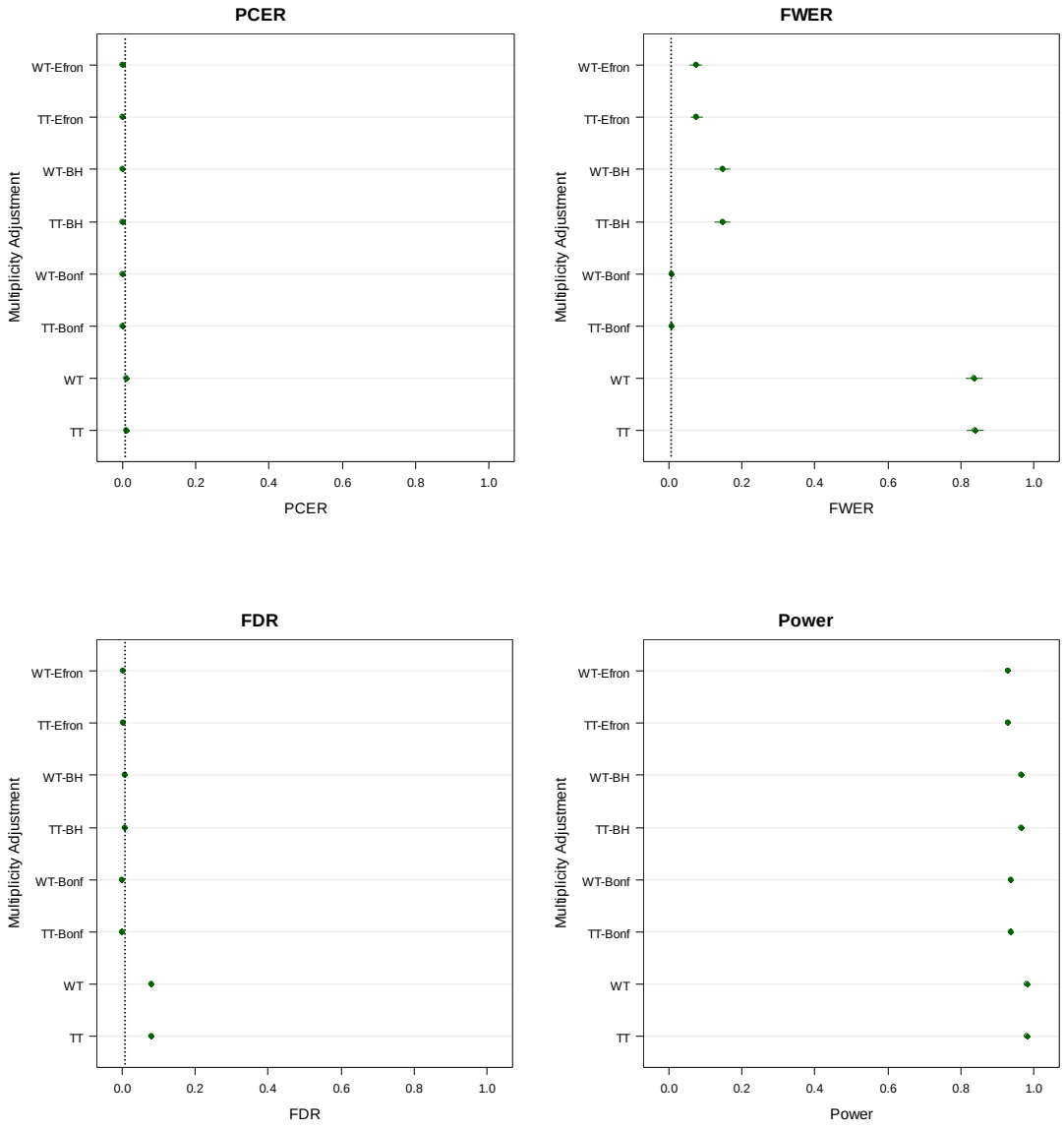

Figure F-36. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 10% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

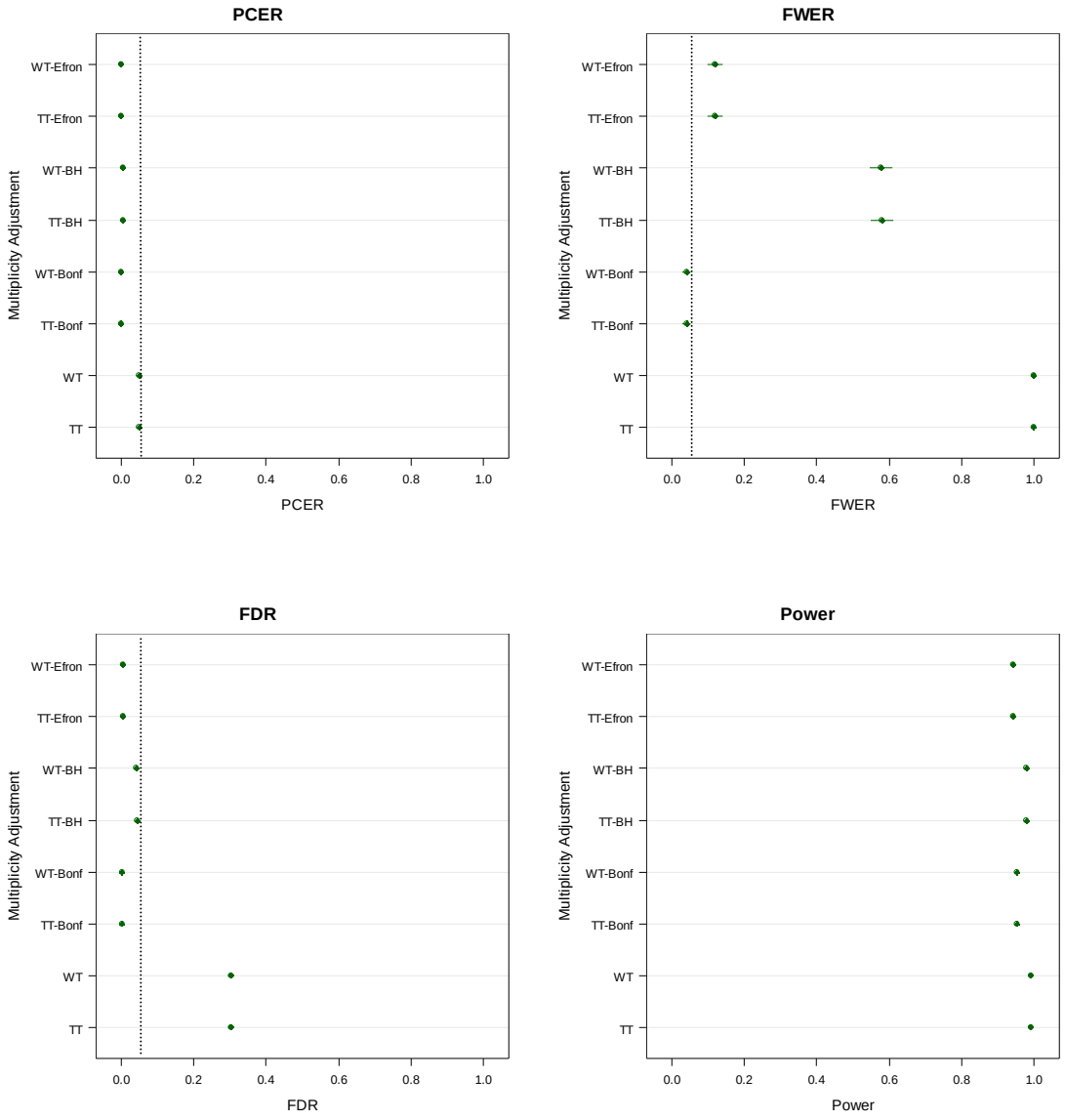

Figure F-37. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 10% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

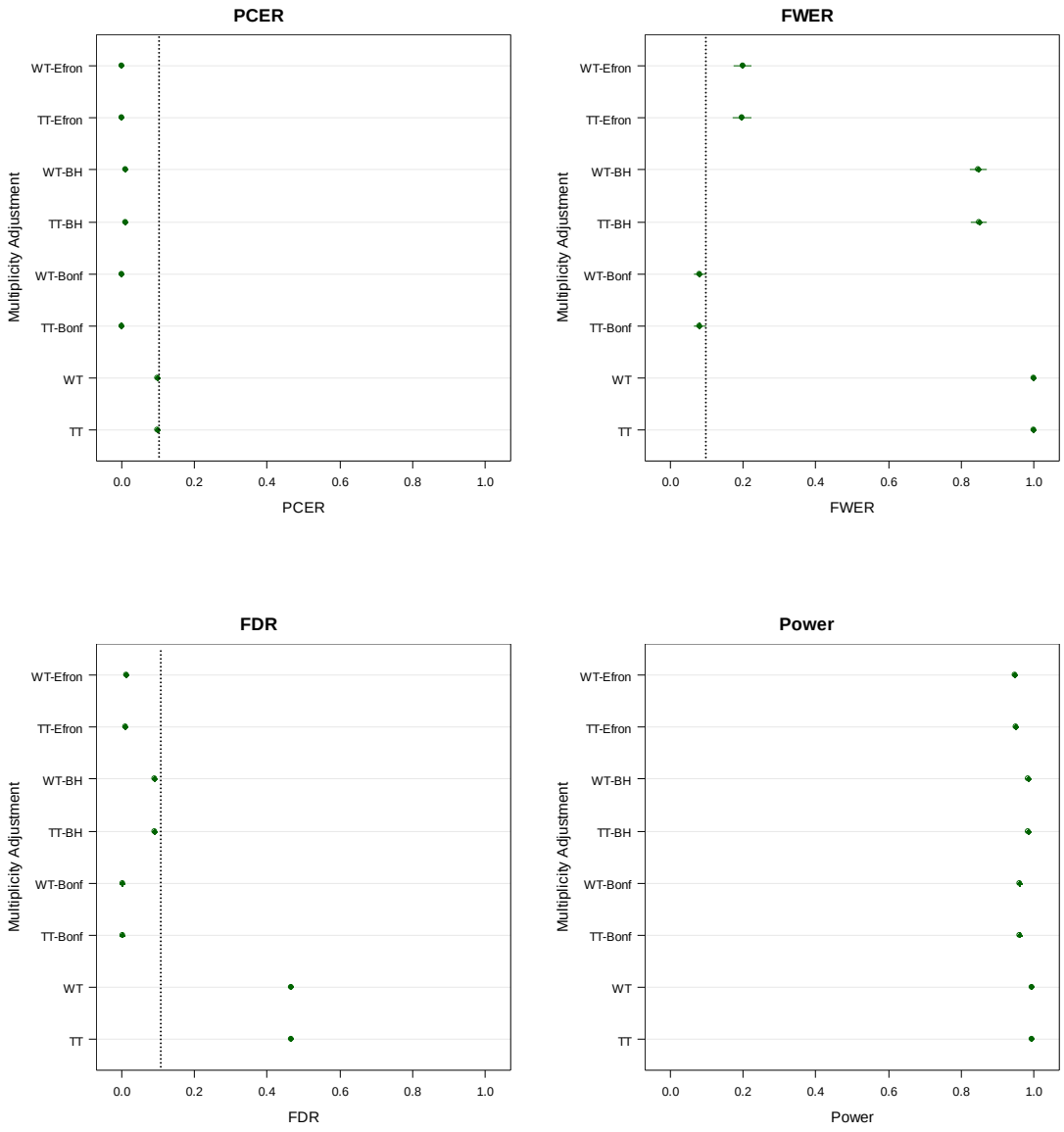

Figure F-38. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 10% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

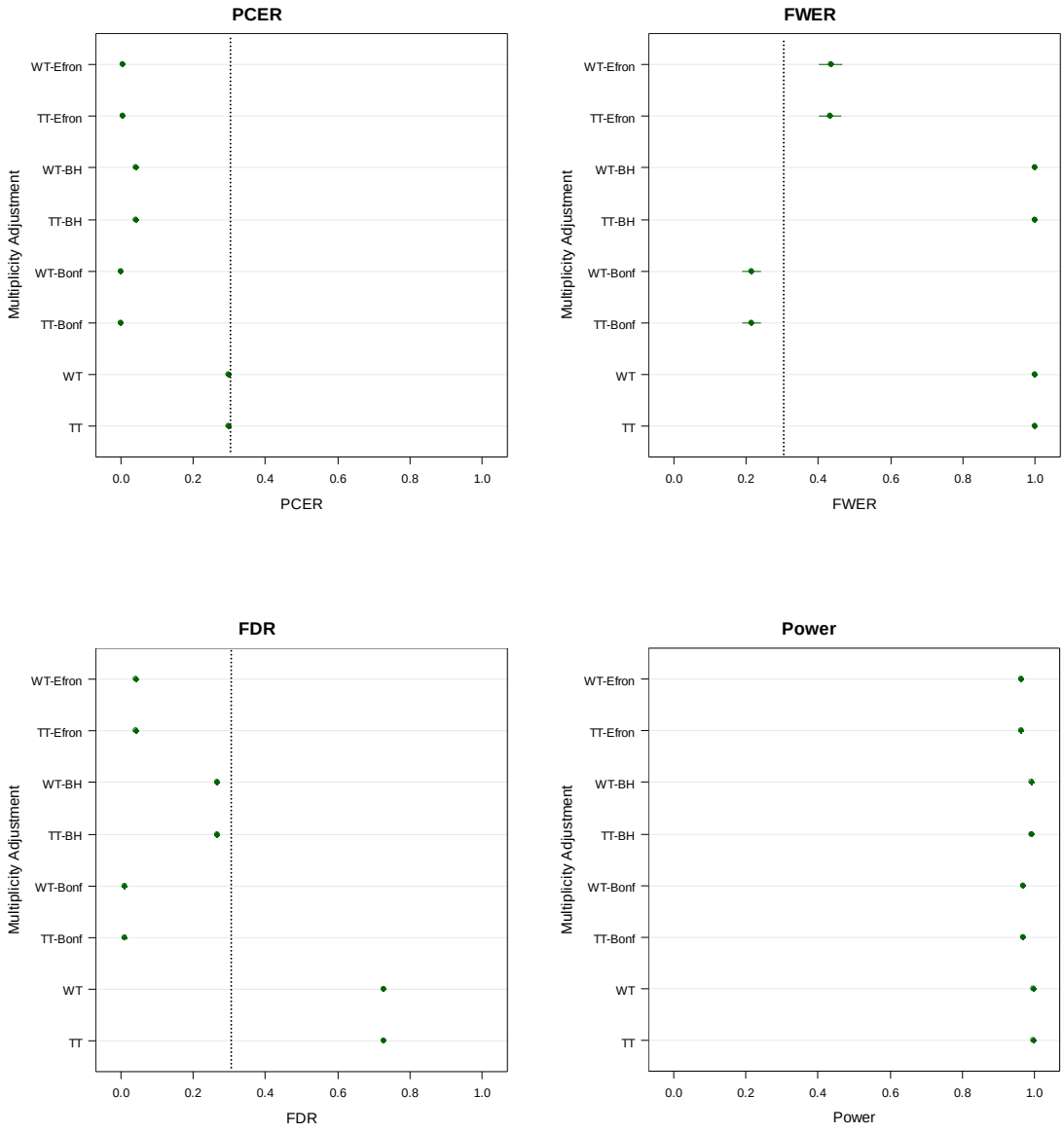

Figure F-39. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 10% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

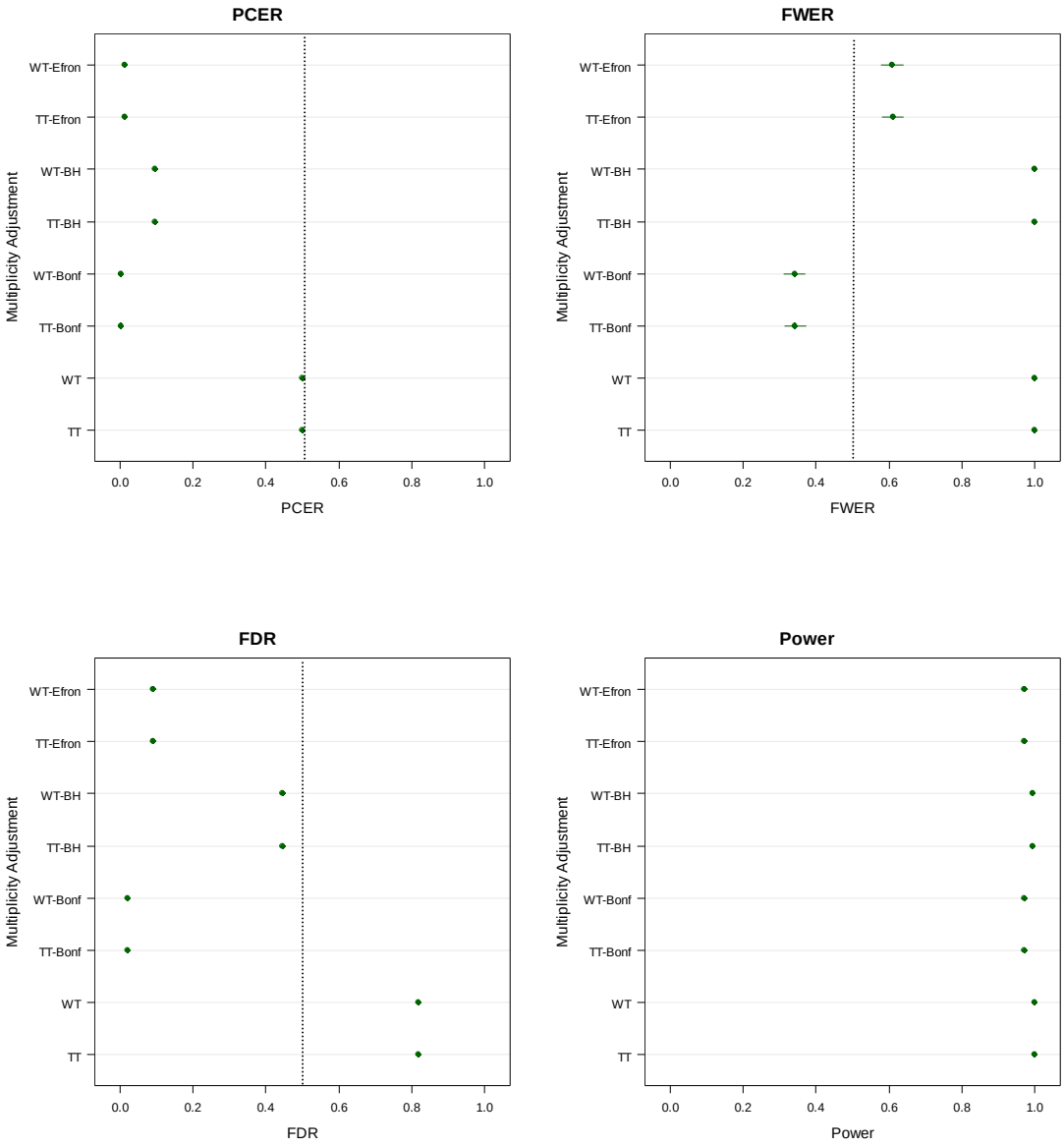

Figure F-40. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 10% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

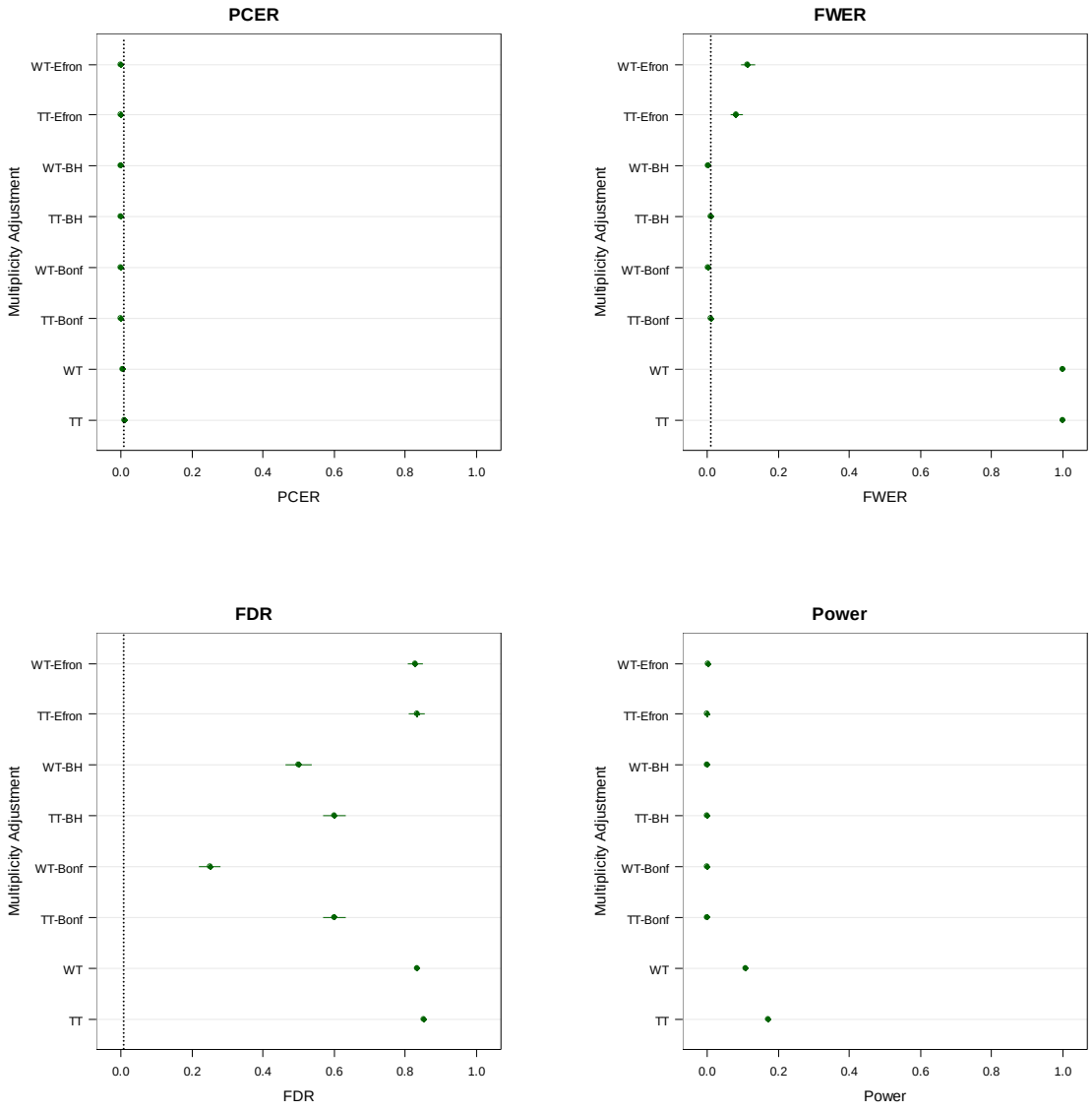

Figure F-41. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 1% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

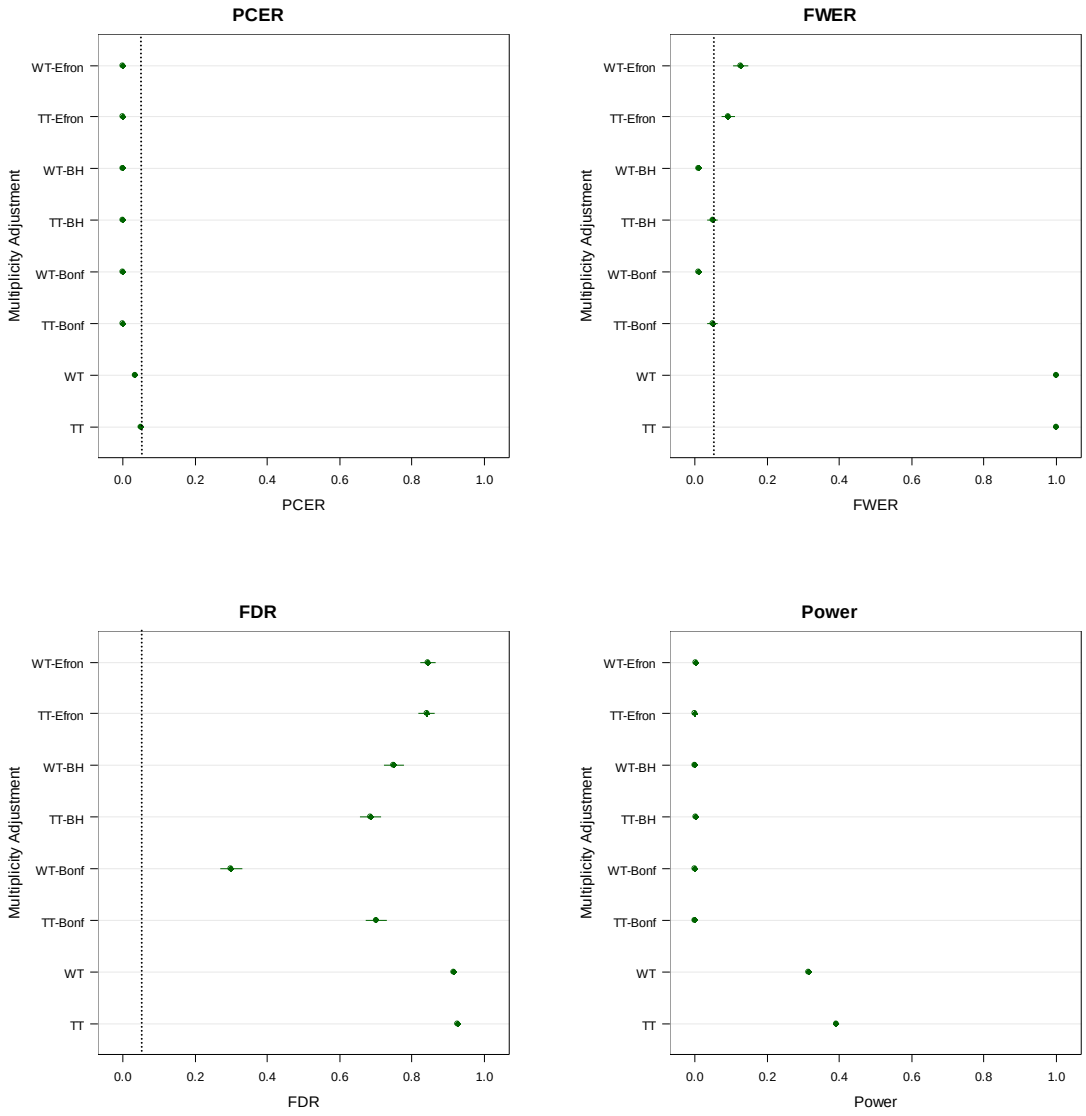

Figure F-42. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 1% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

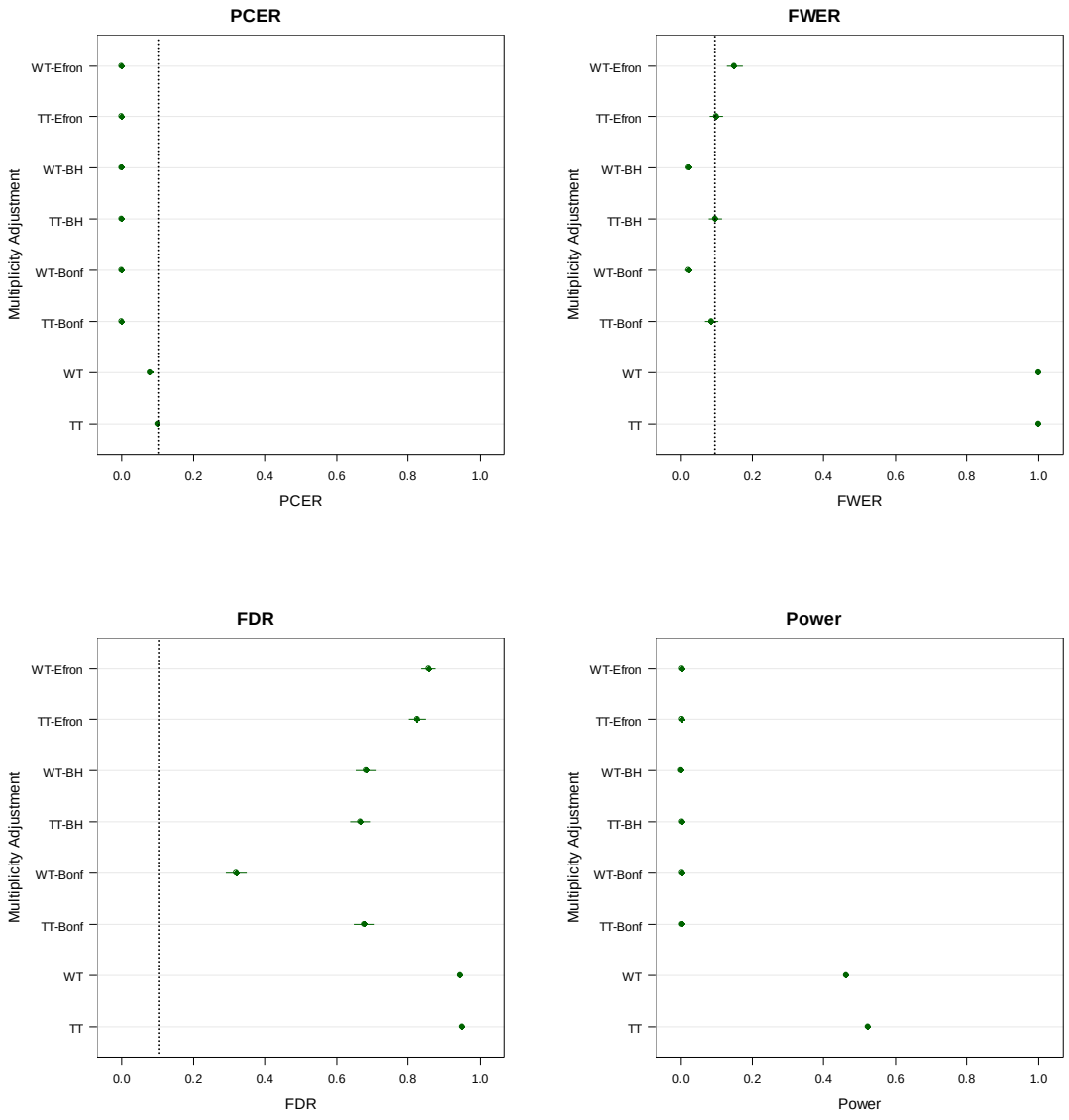

Figure F-43. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 1% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

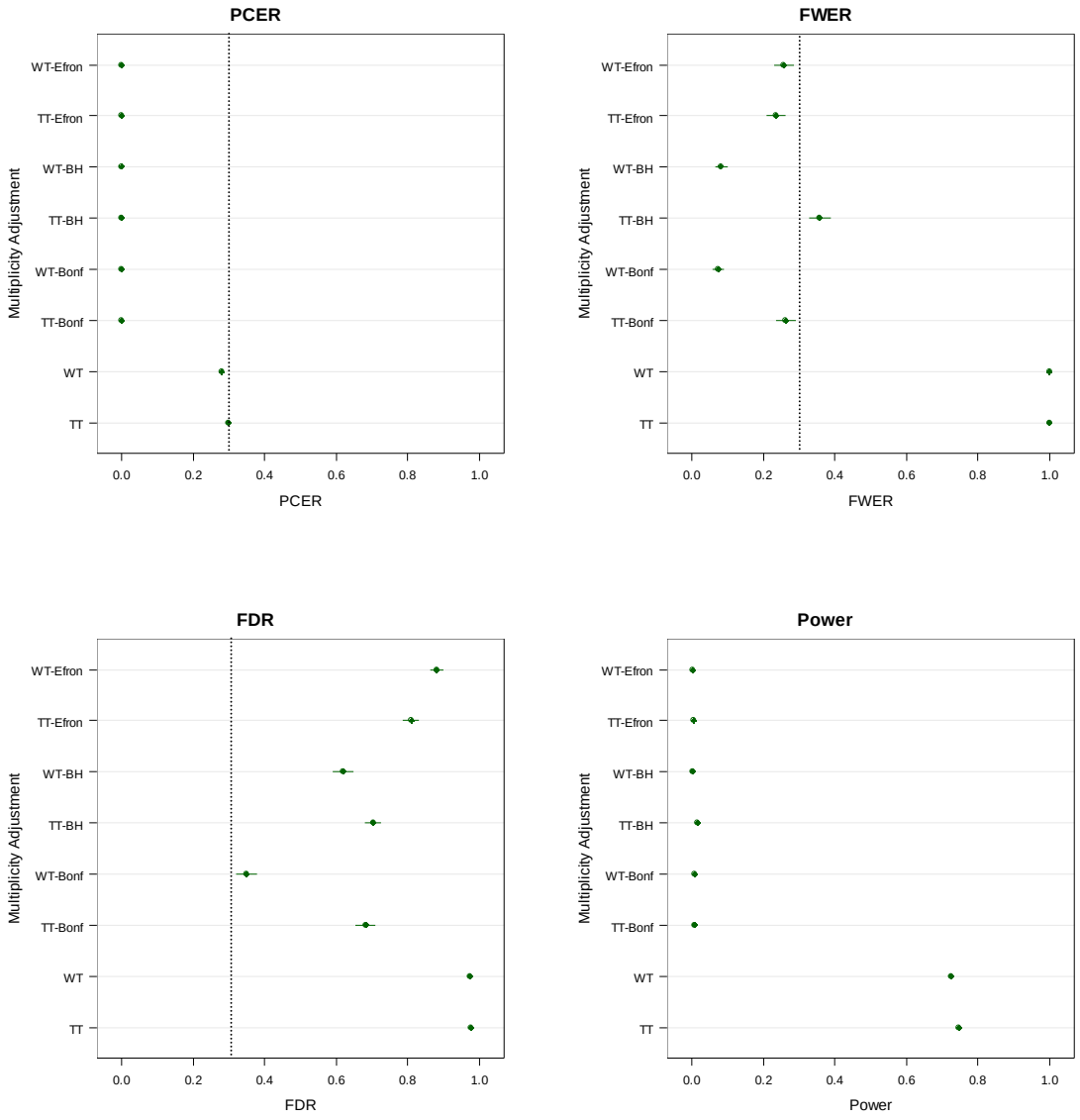

Figure F-44. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 1% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

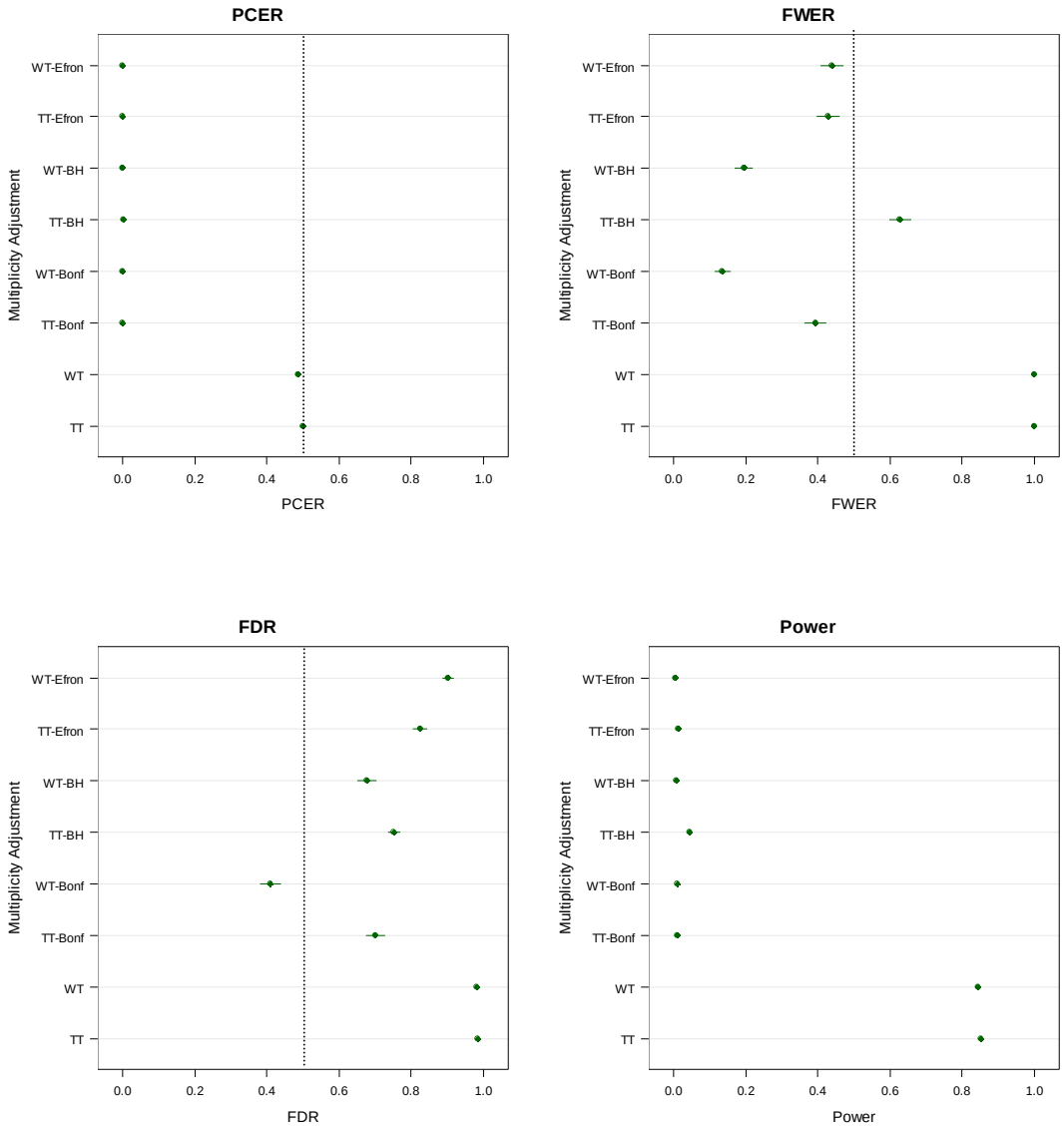

Figure F-45. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 1% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

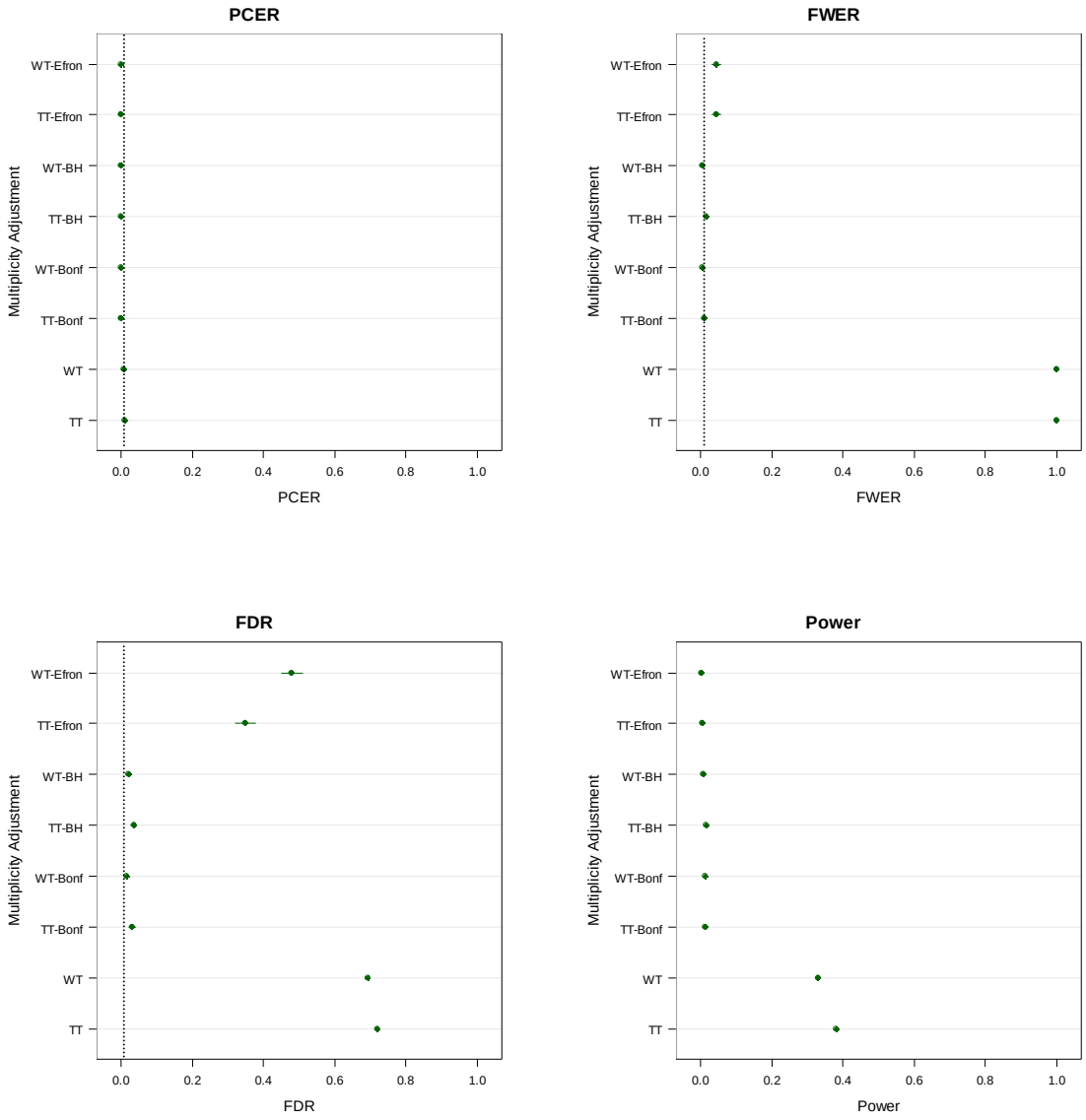

Figure F-46. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 1% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

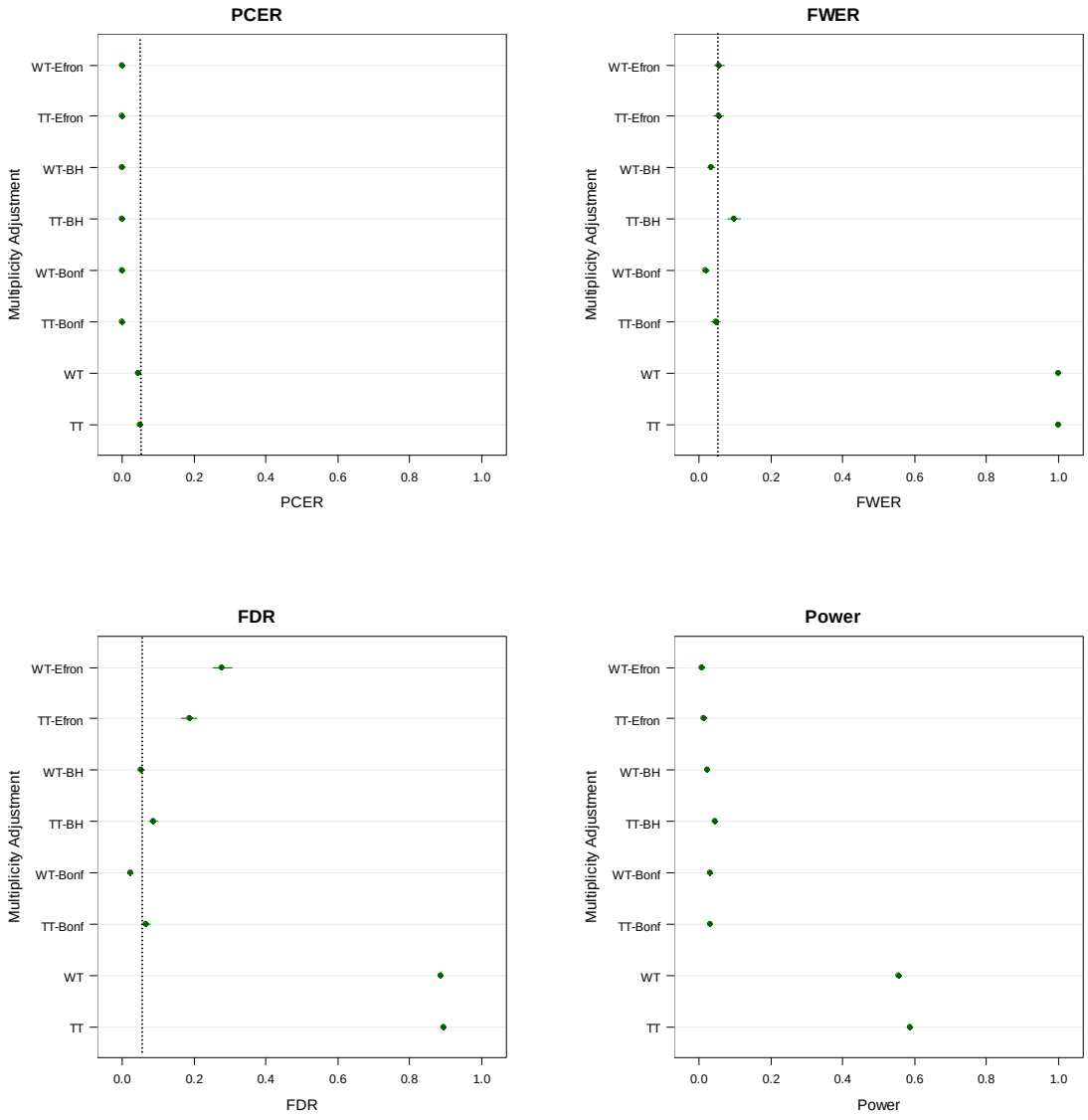

Figure F-47. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 1% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

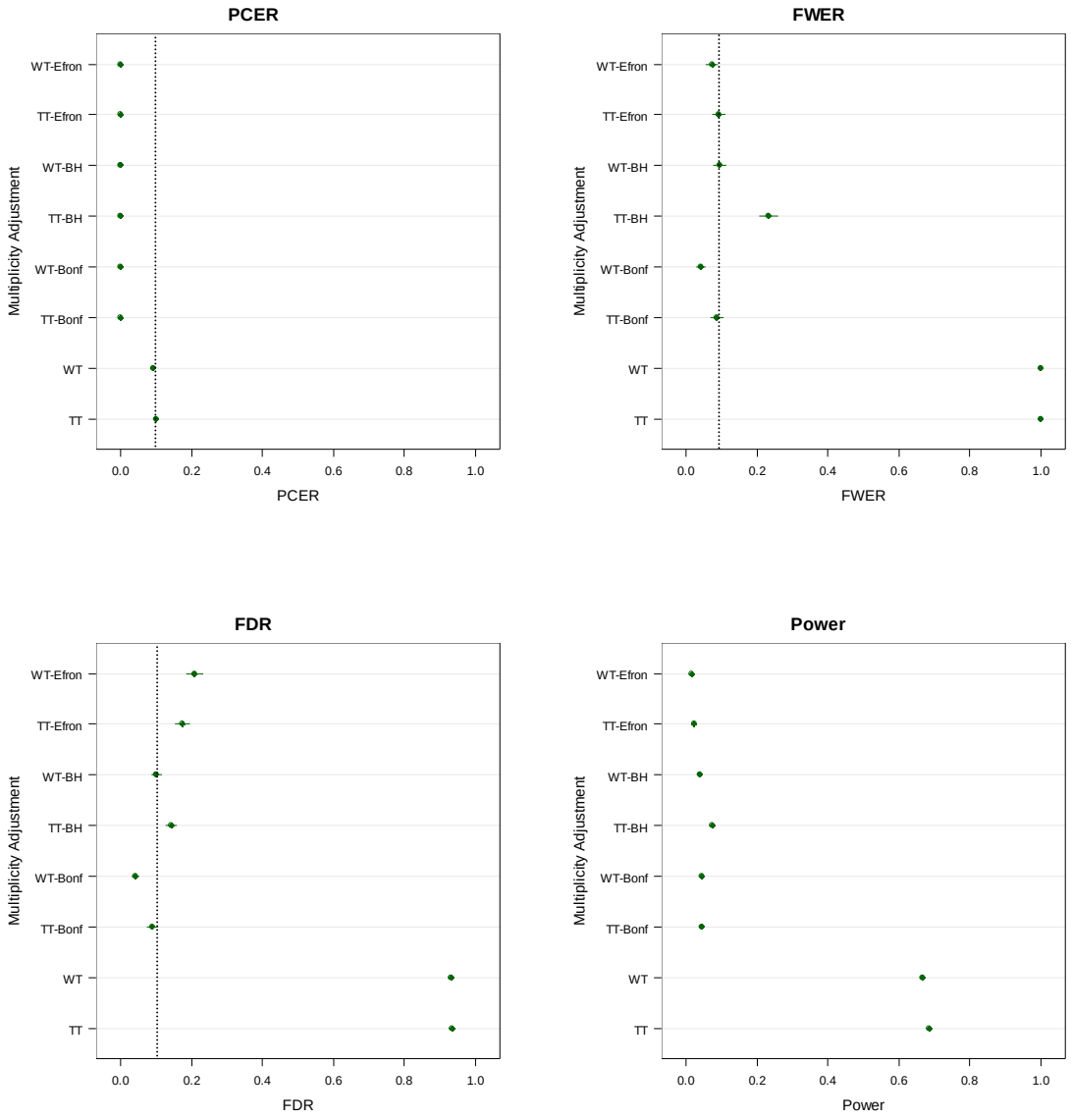

Figure F-48. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 1% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

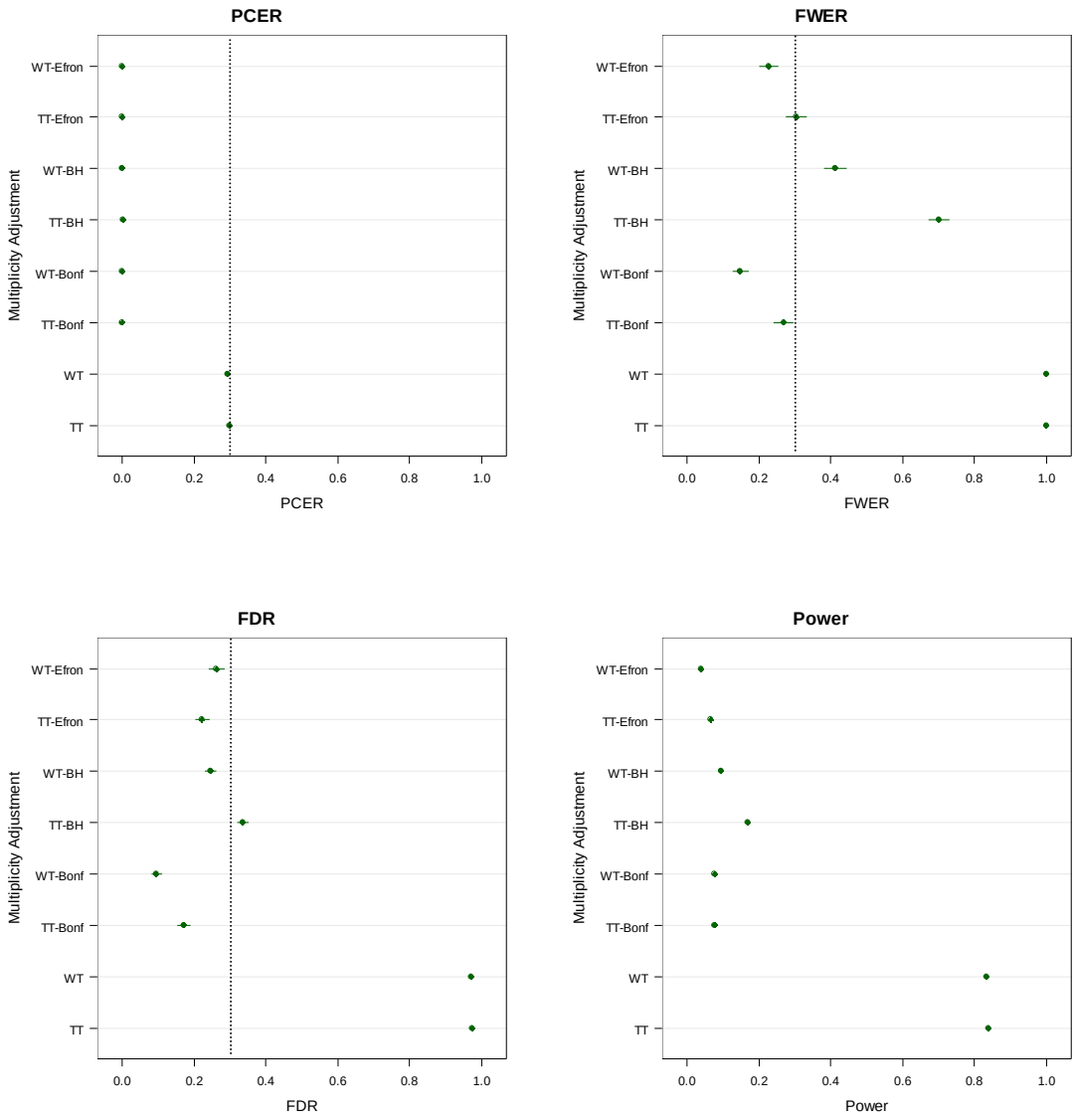

Figure F-49. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 1% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

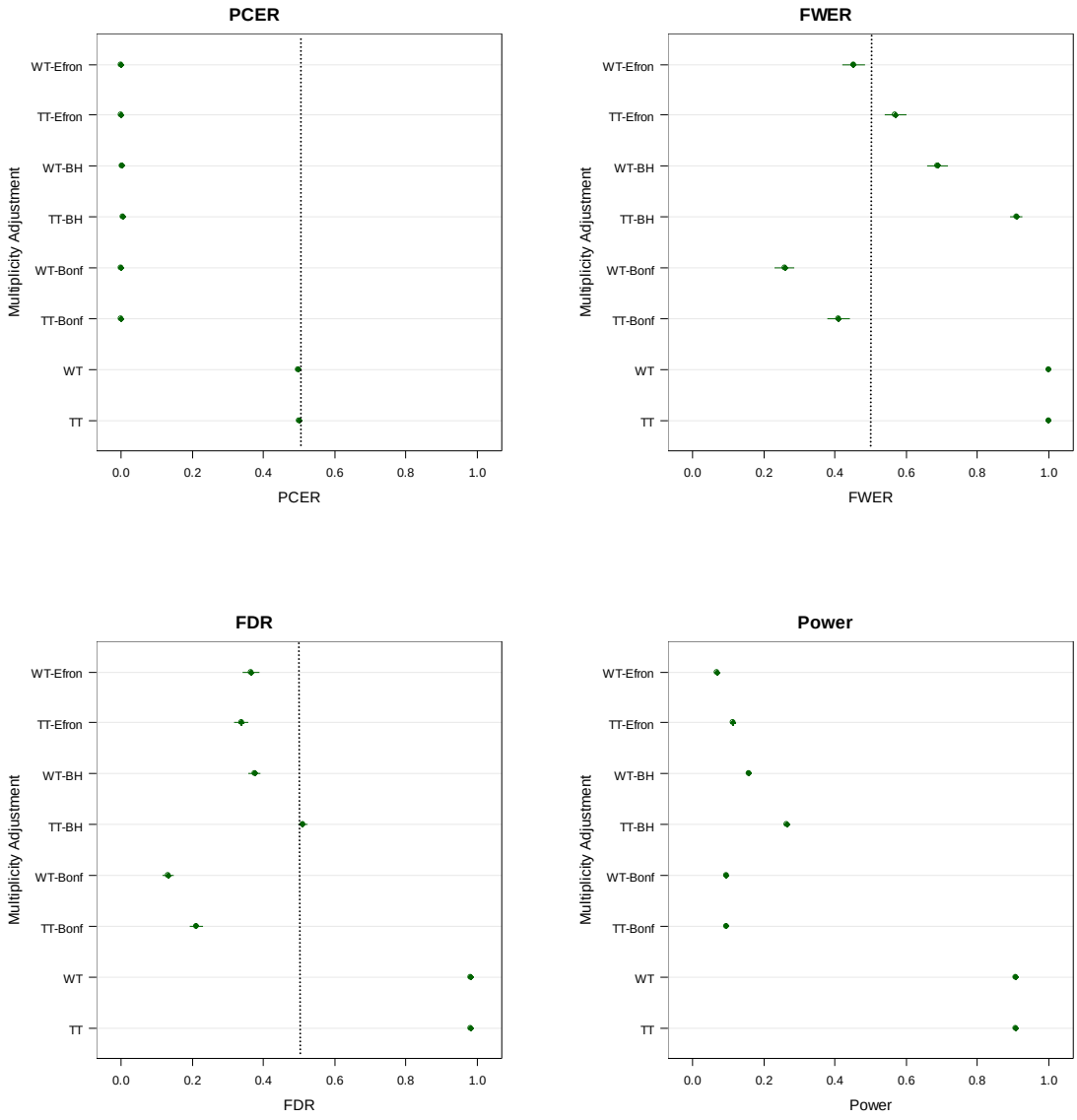

Figure F-50. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 1% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

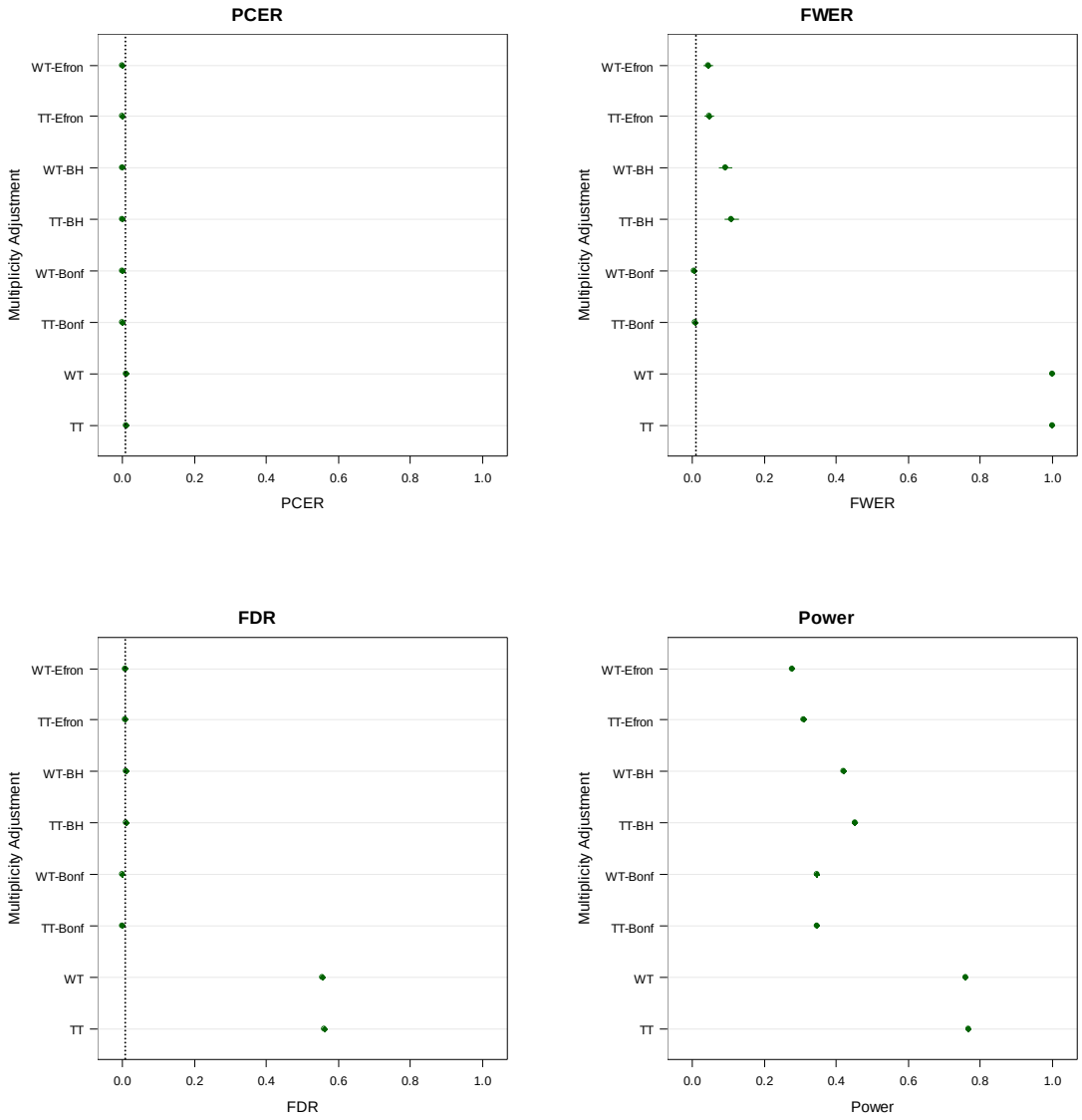

Figure F-51. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 1% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

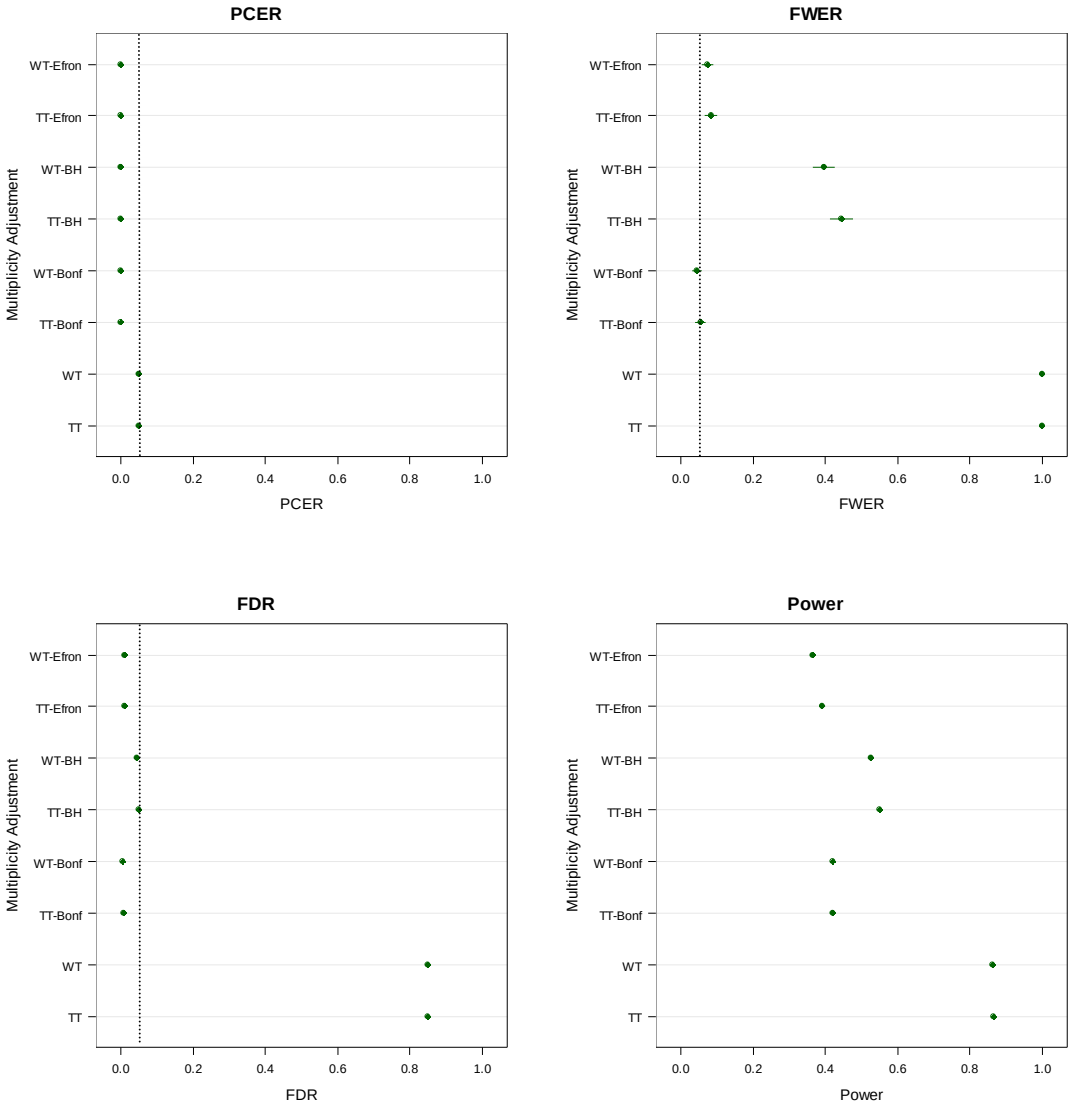

Figure F-52. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 1% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

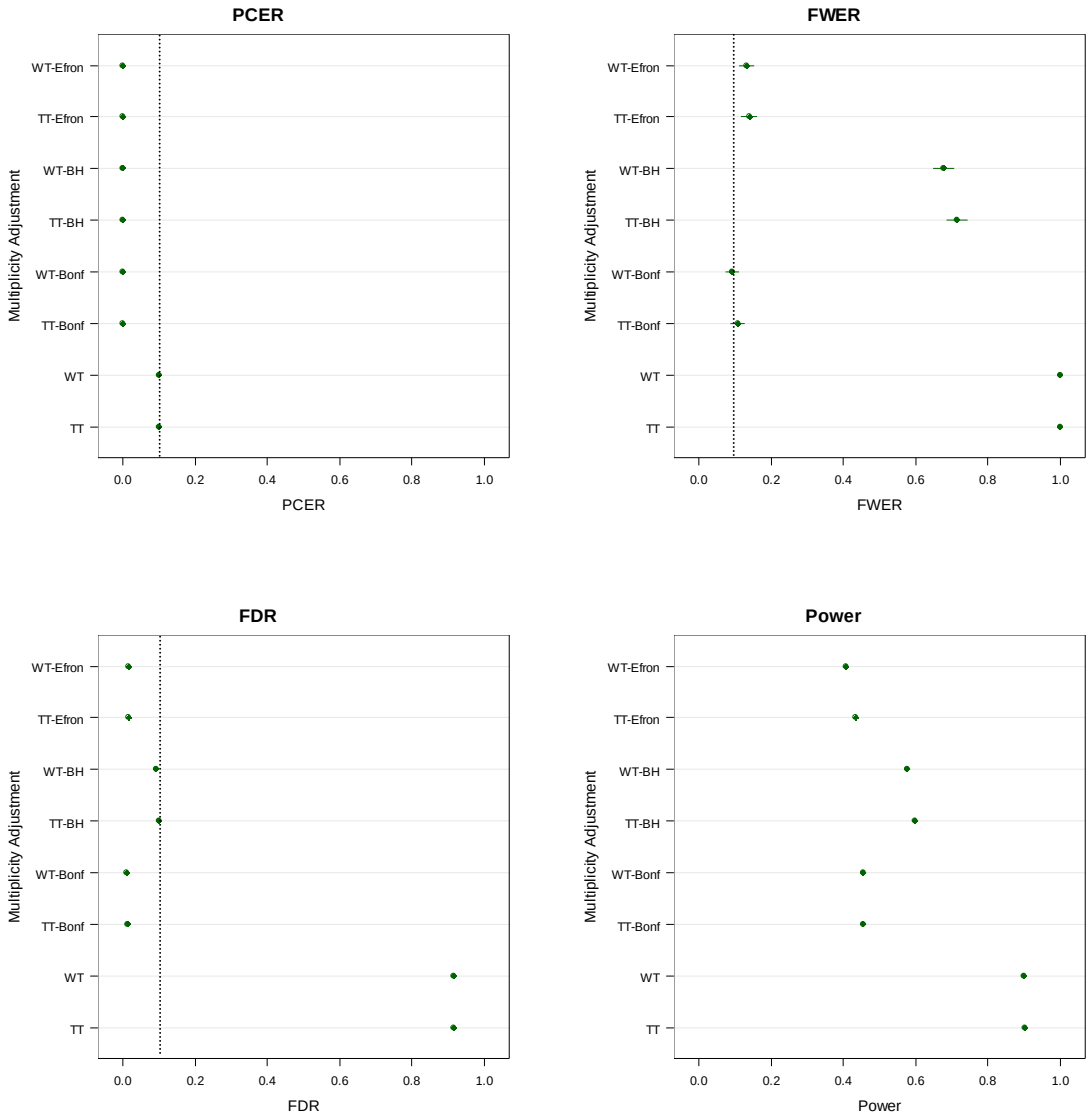

Figure F-53. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 1% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

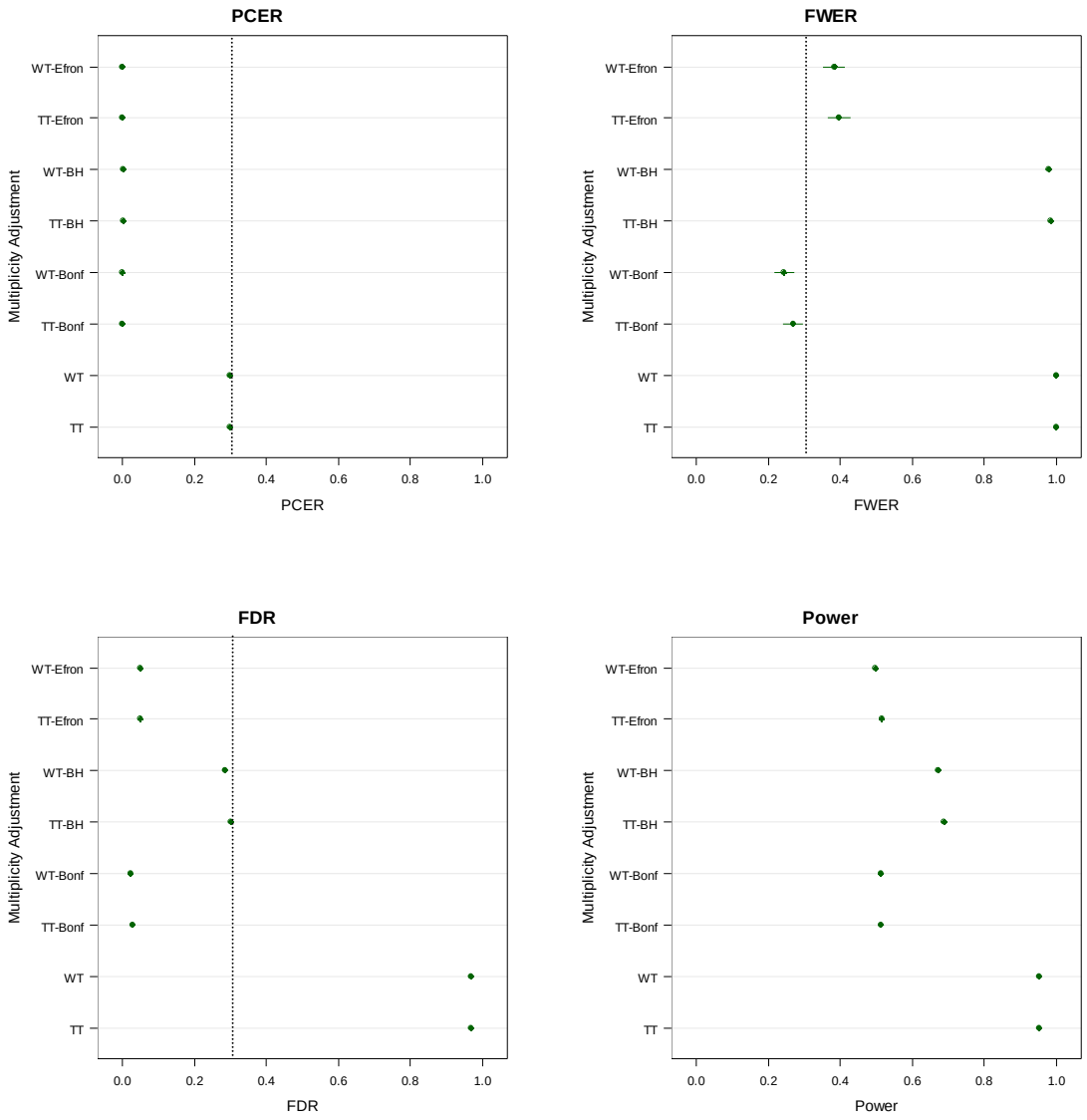

Figure F-54. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 1% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

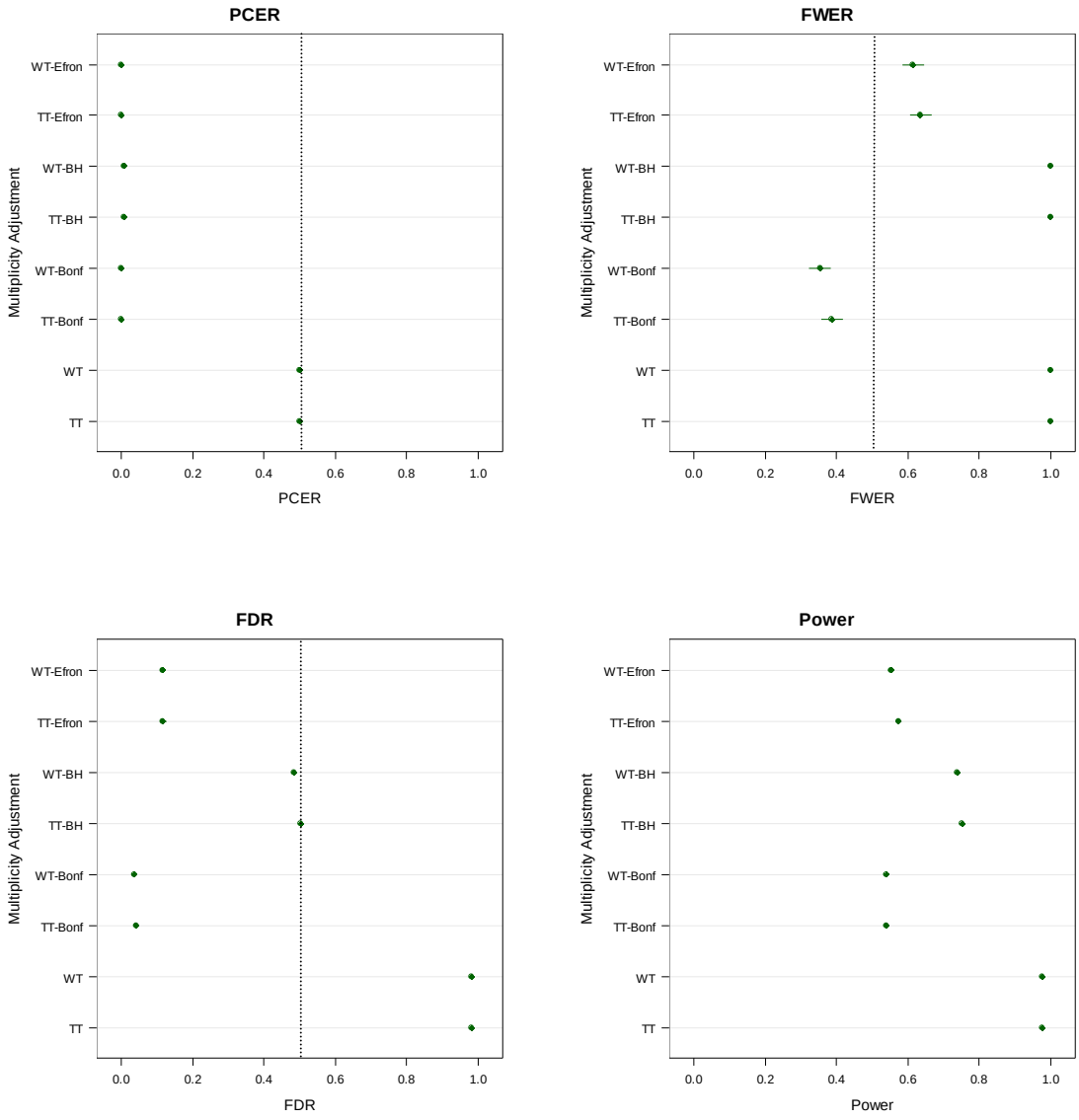

Figure F-55. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 1% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

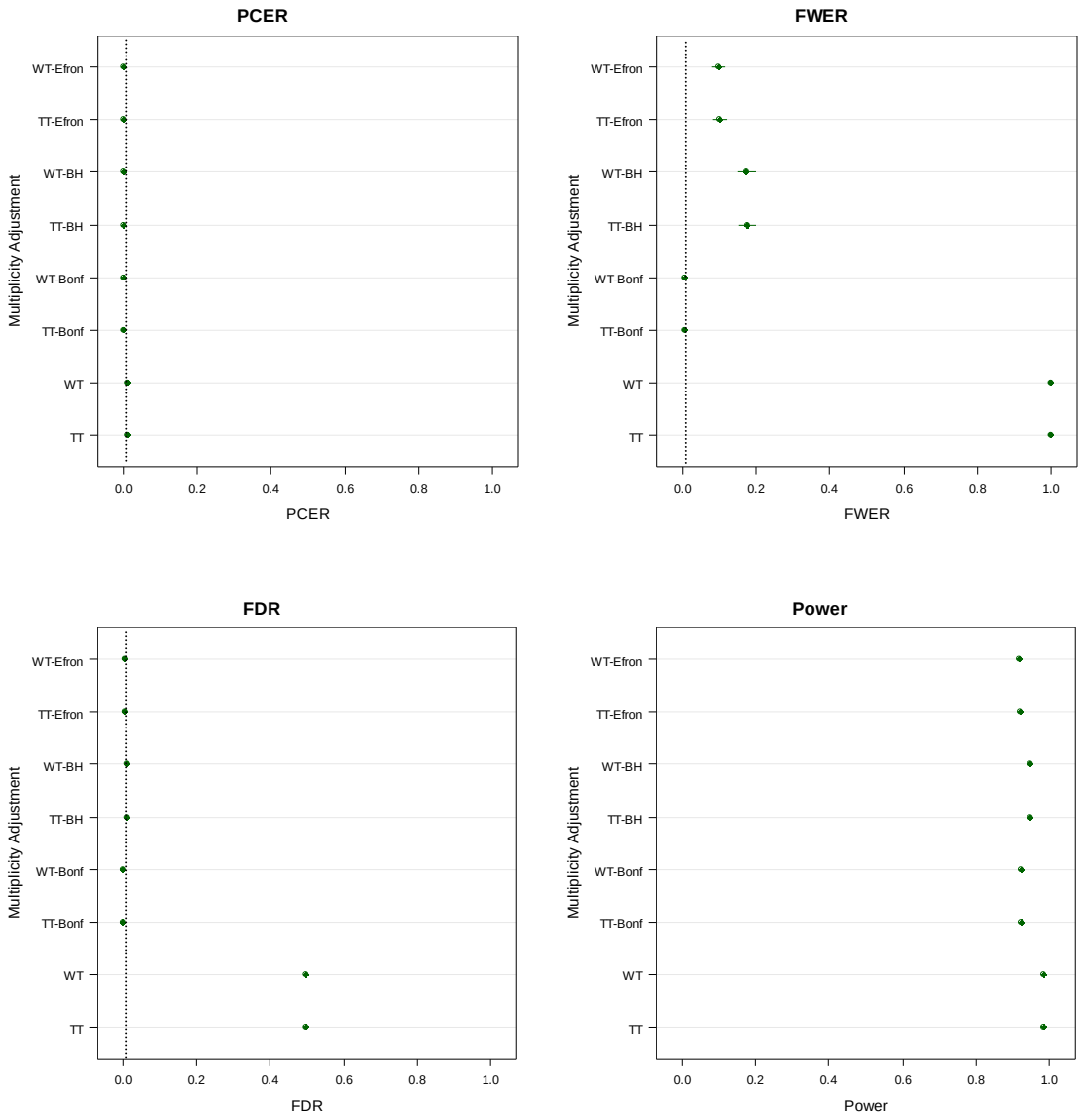

Figure F-56. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 1% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

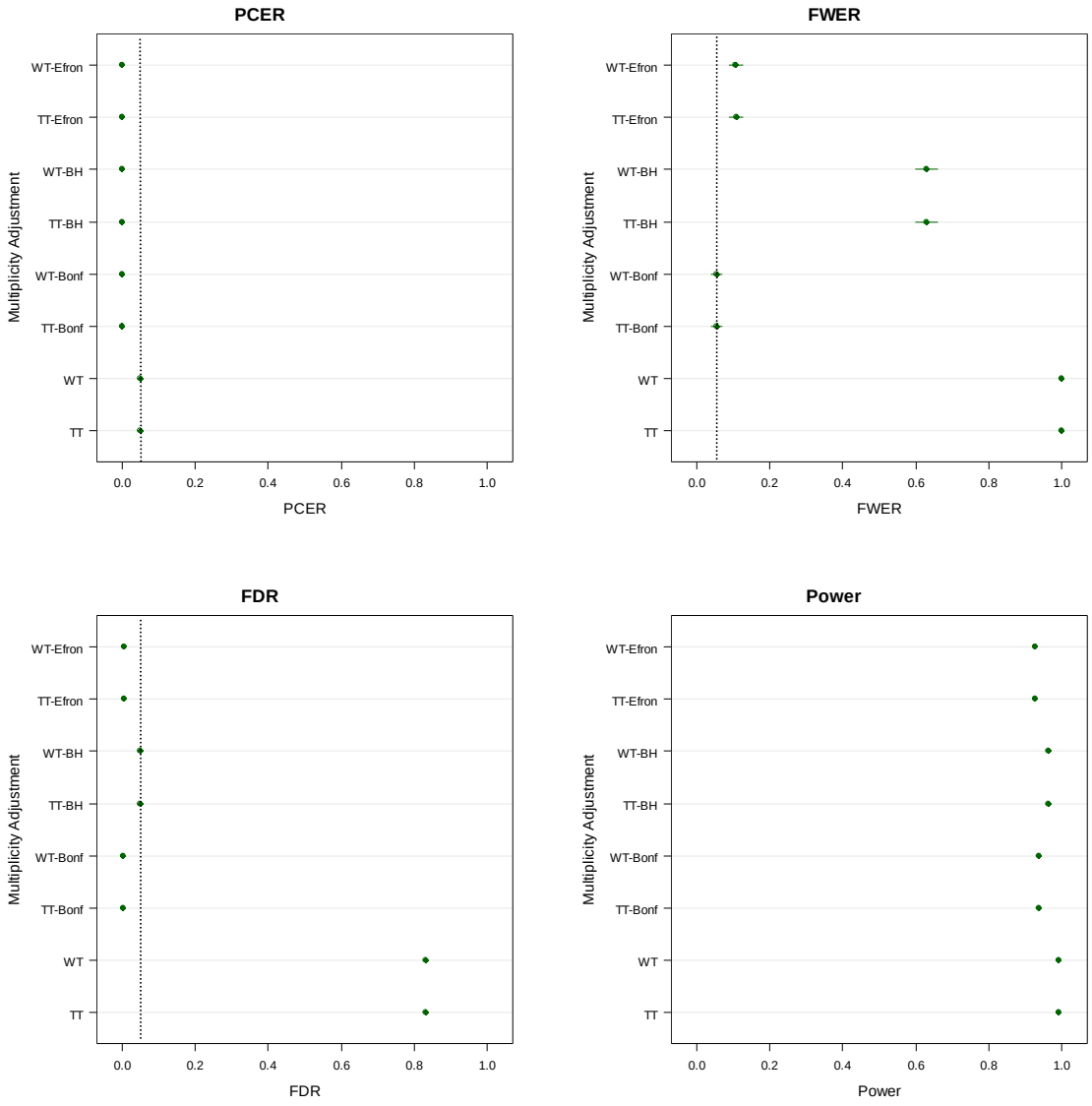

Figure F-57. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 1% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

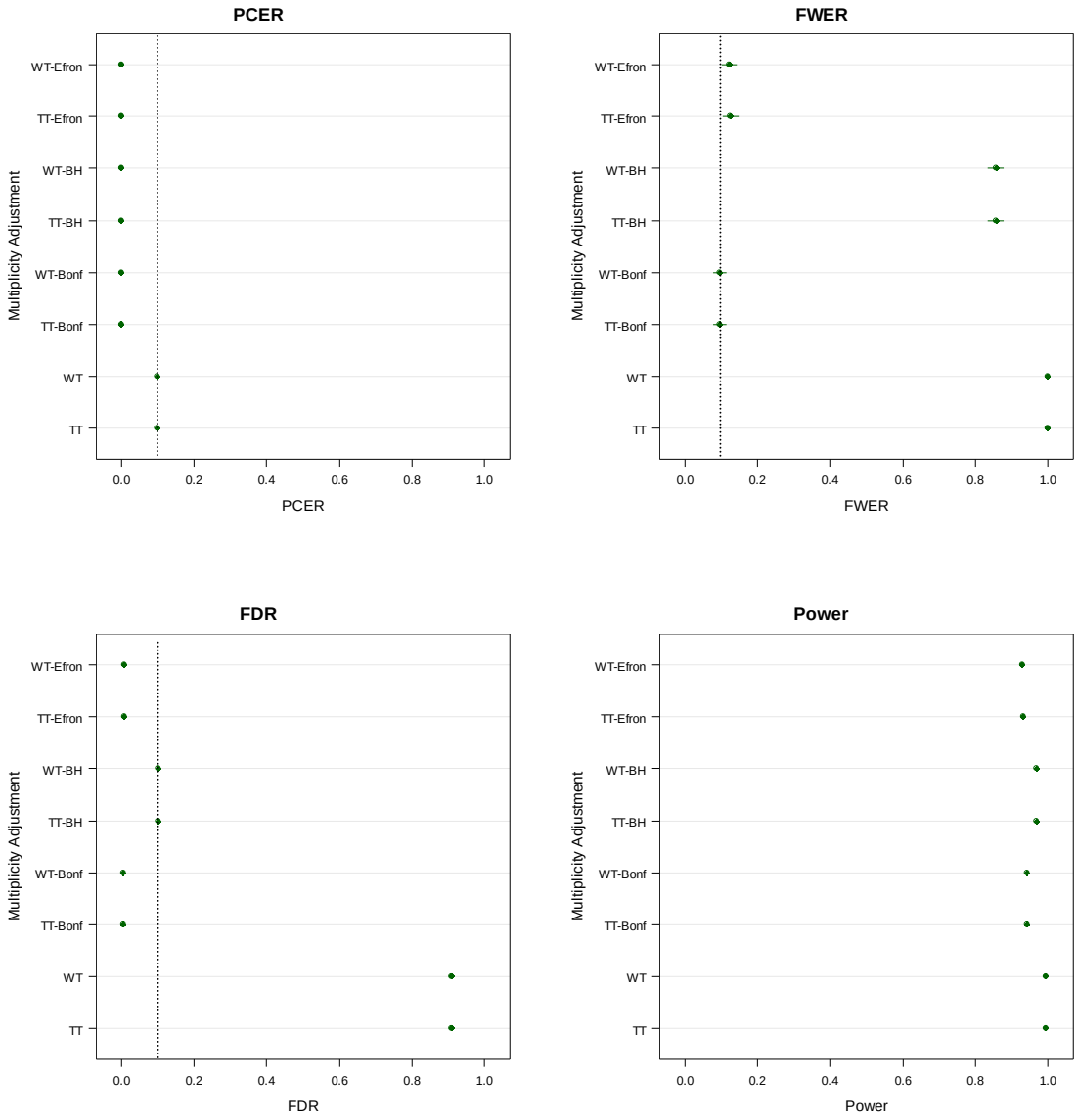

Figure F-58. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 1% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

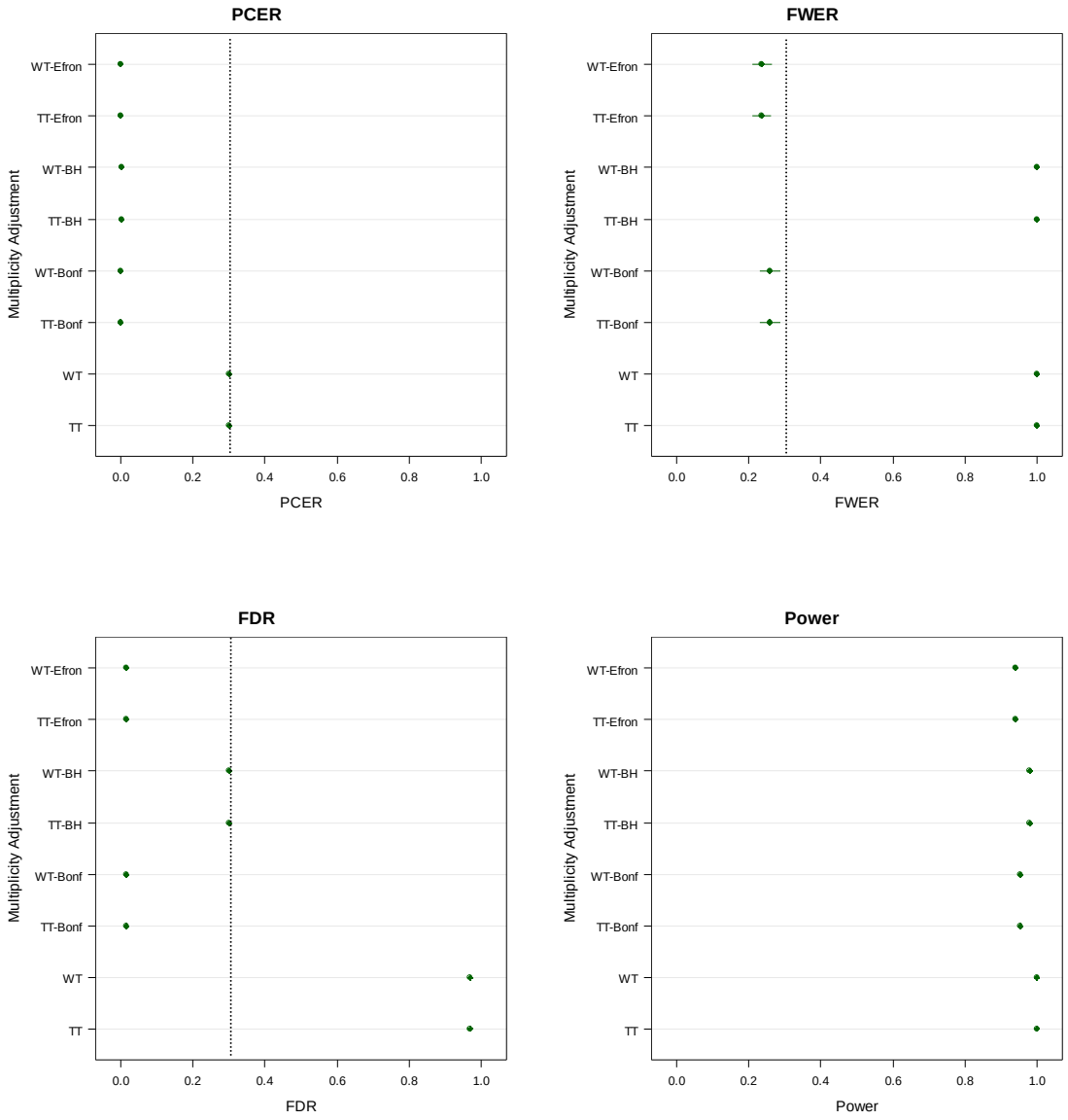

Figure F-59. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 1% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

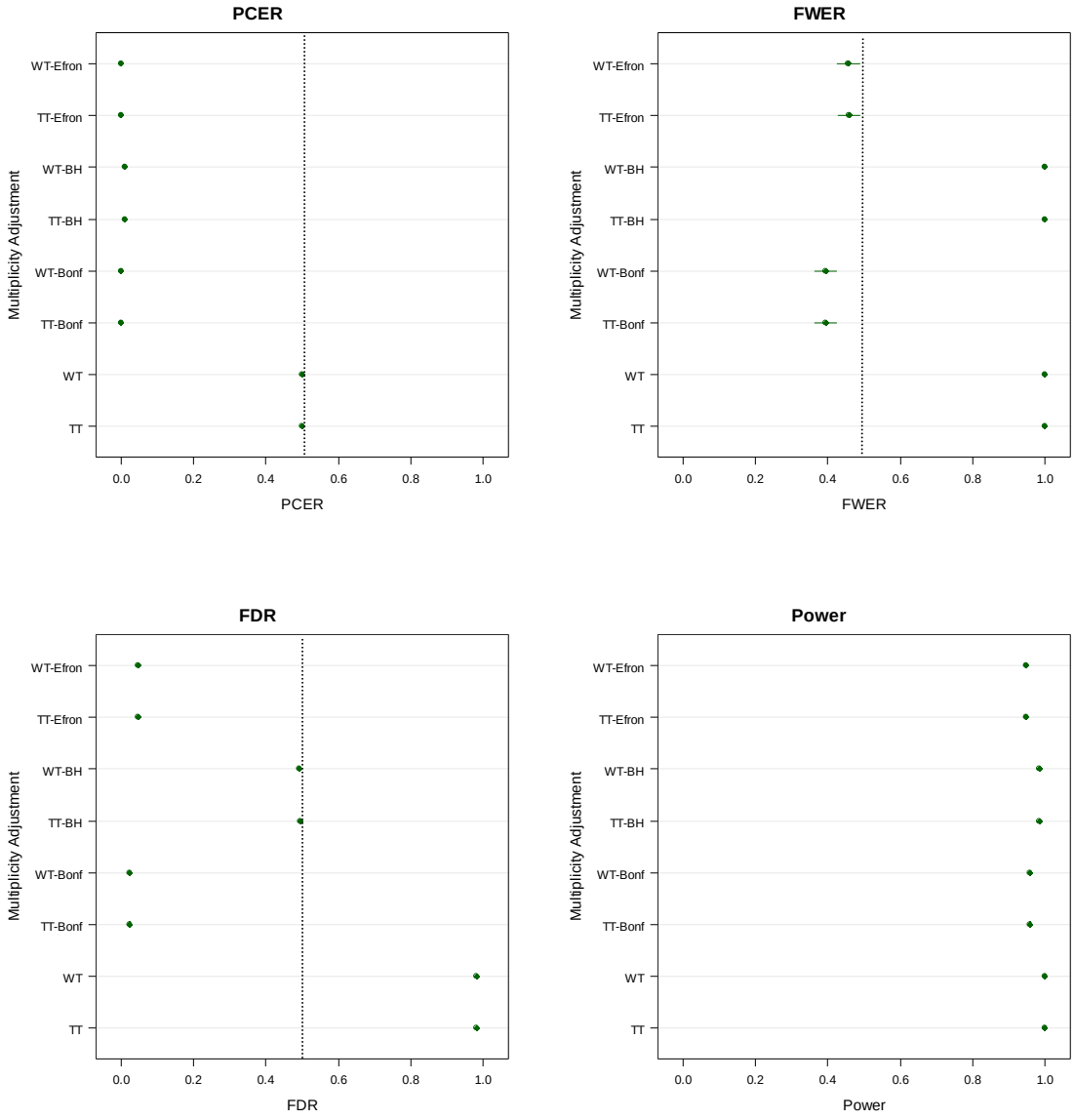

Figure F-60. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 1% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

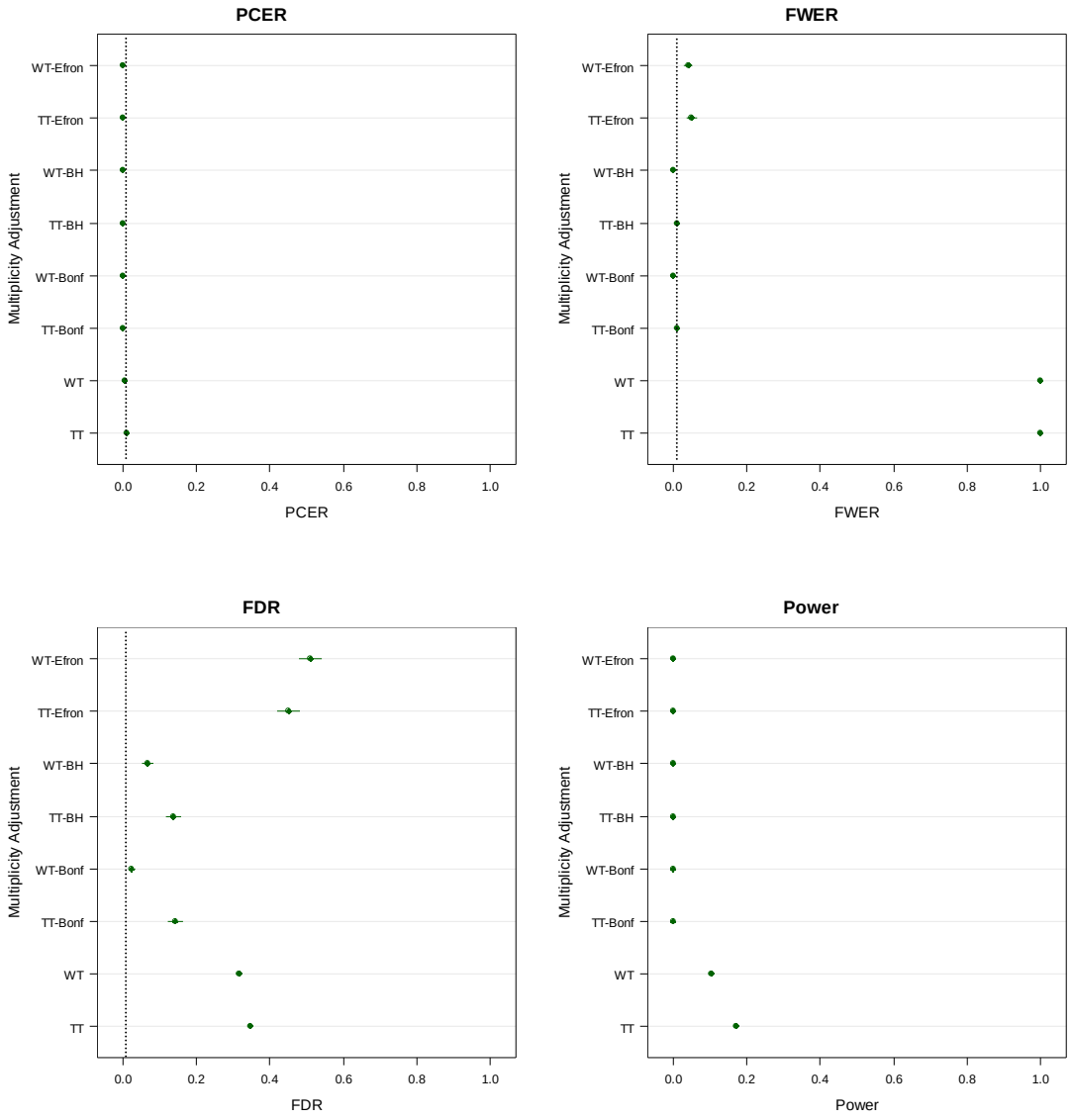

Figure F-61. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 10% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

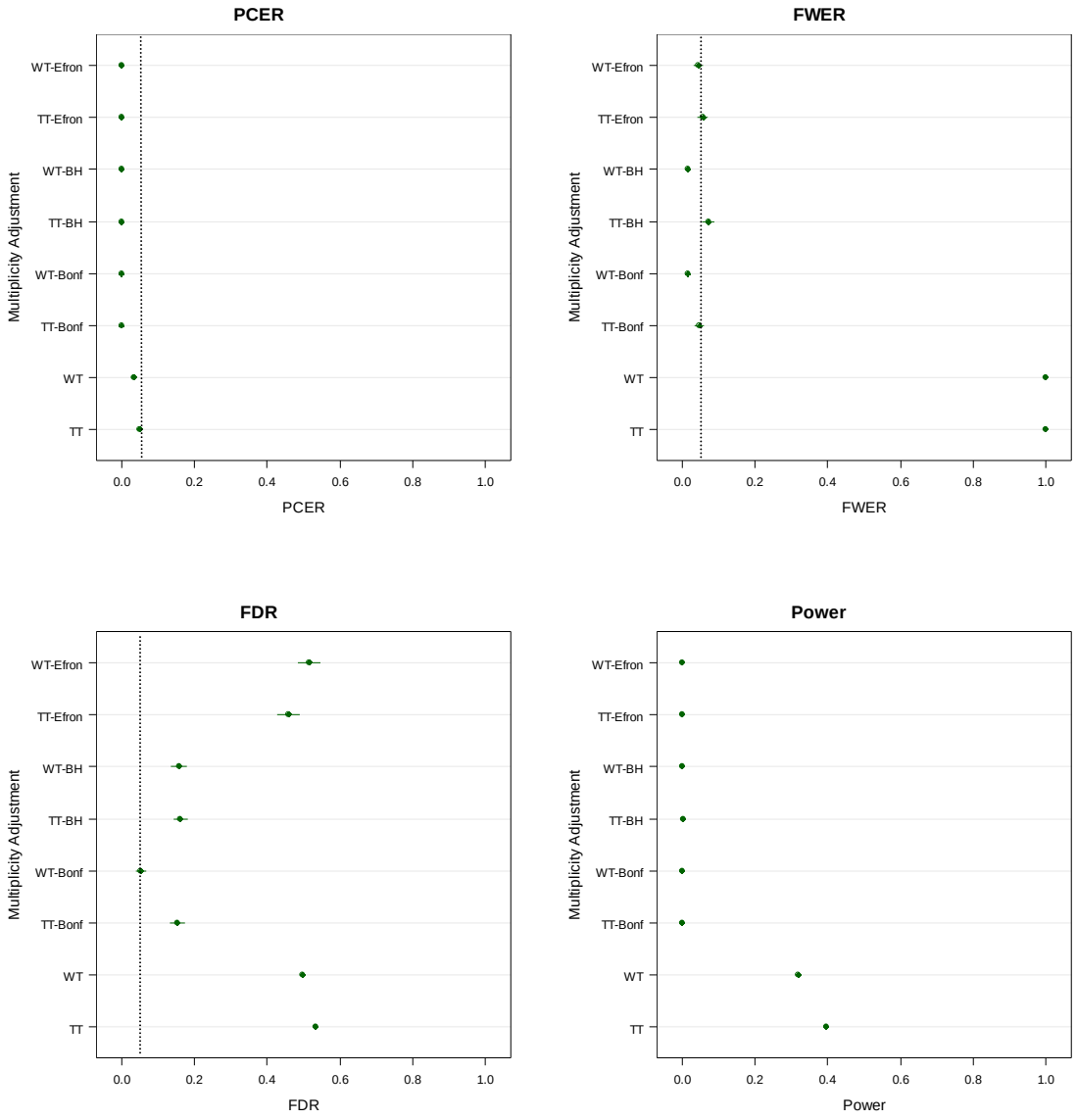

Figure F-62. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 10% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

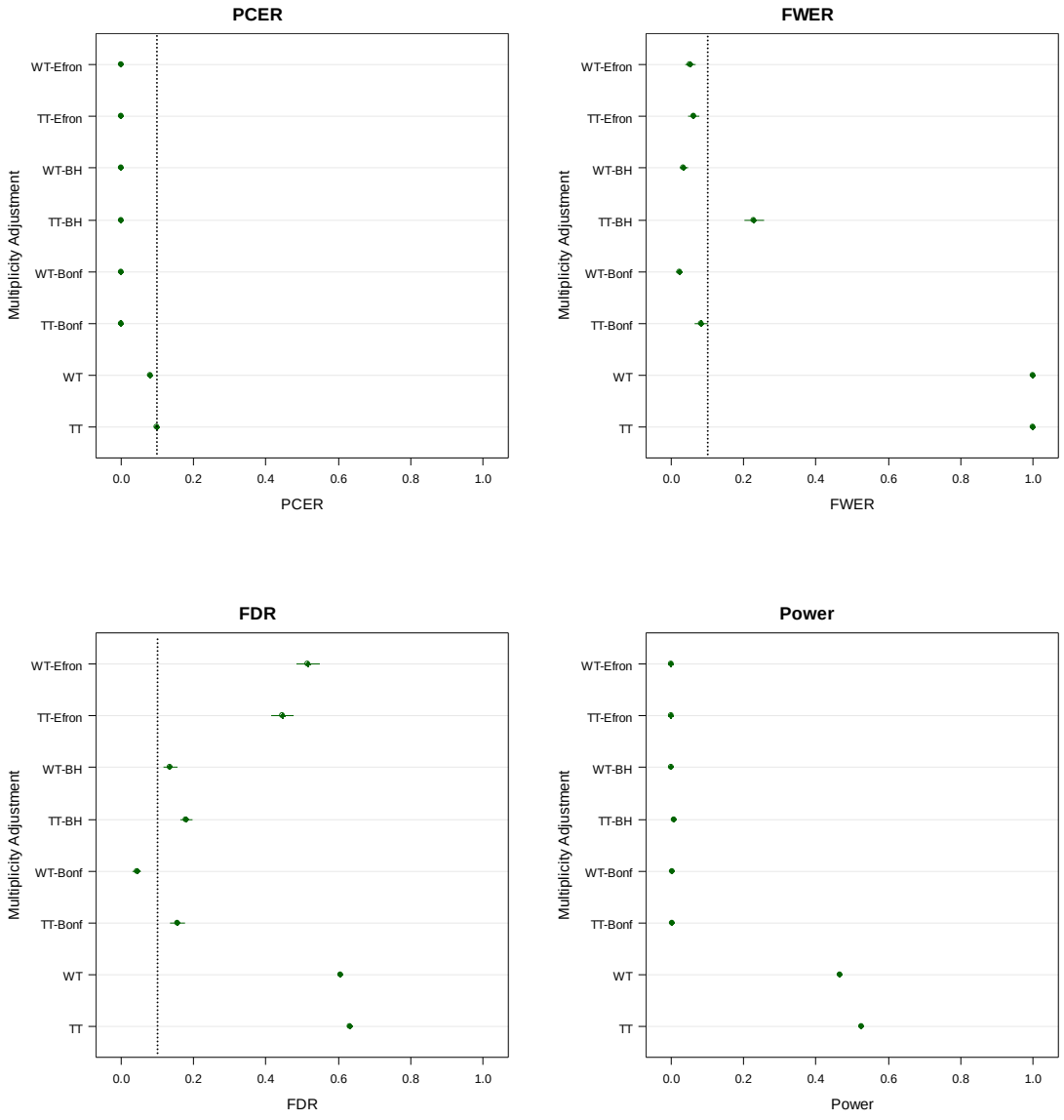

Figure F-63. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 10% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

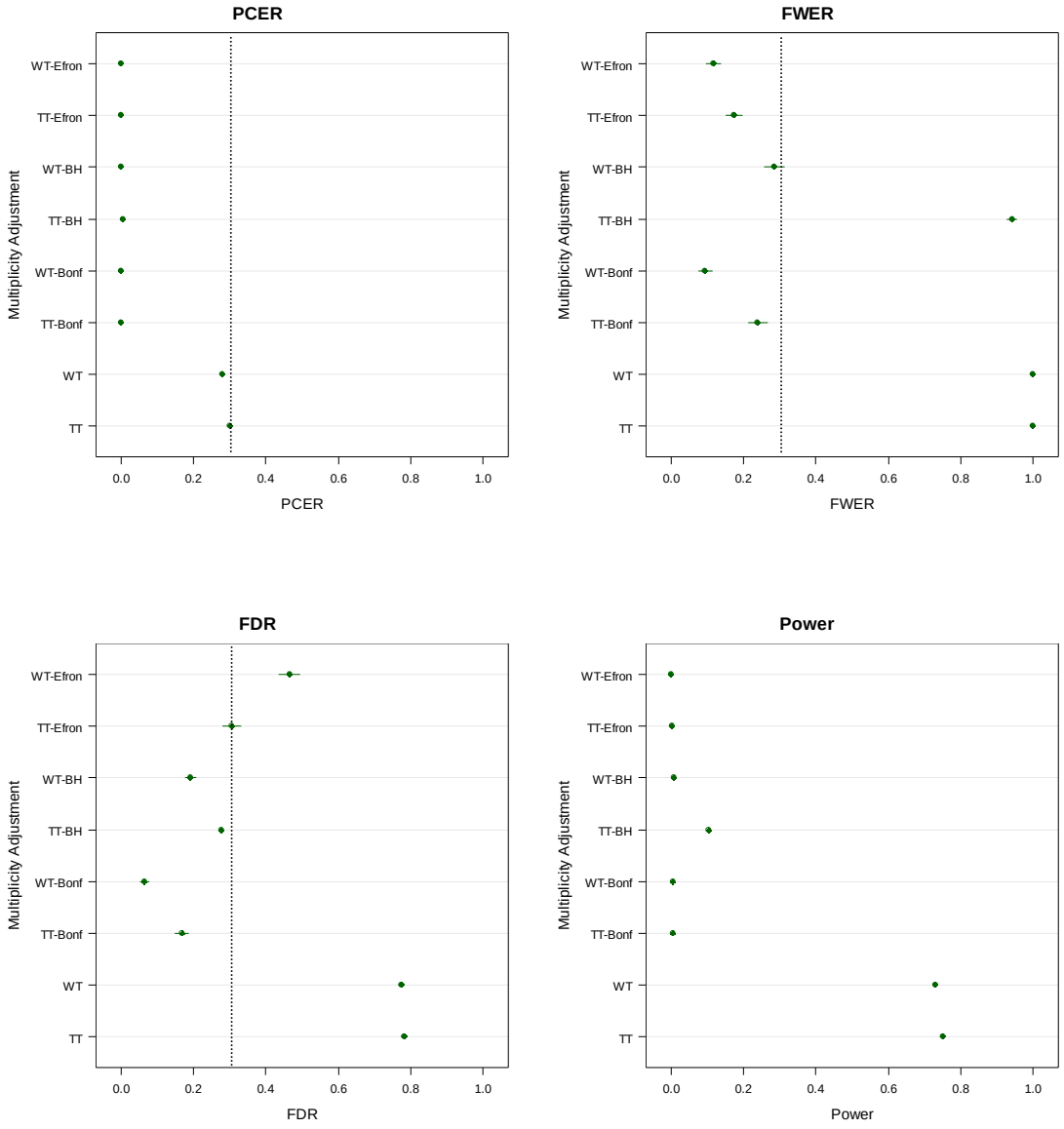

Figure F-64. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 10% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

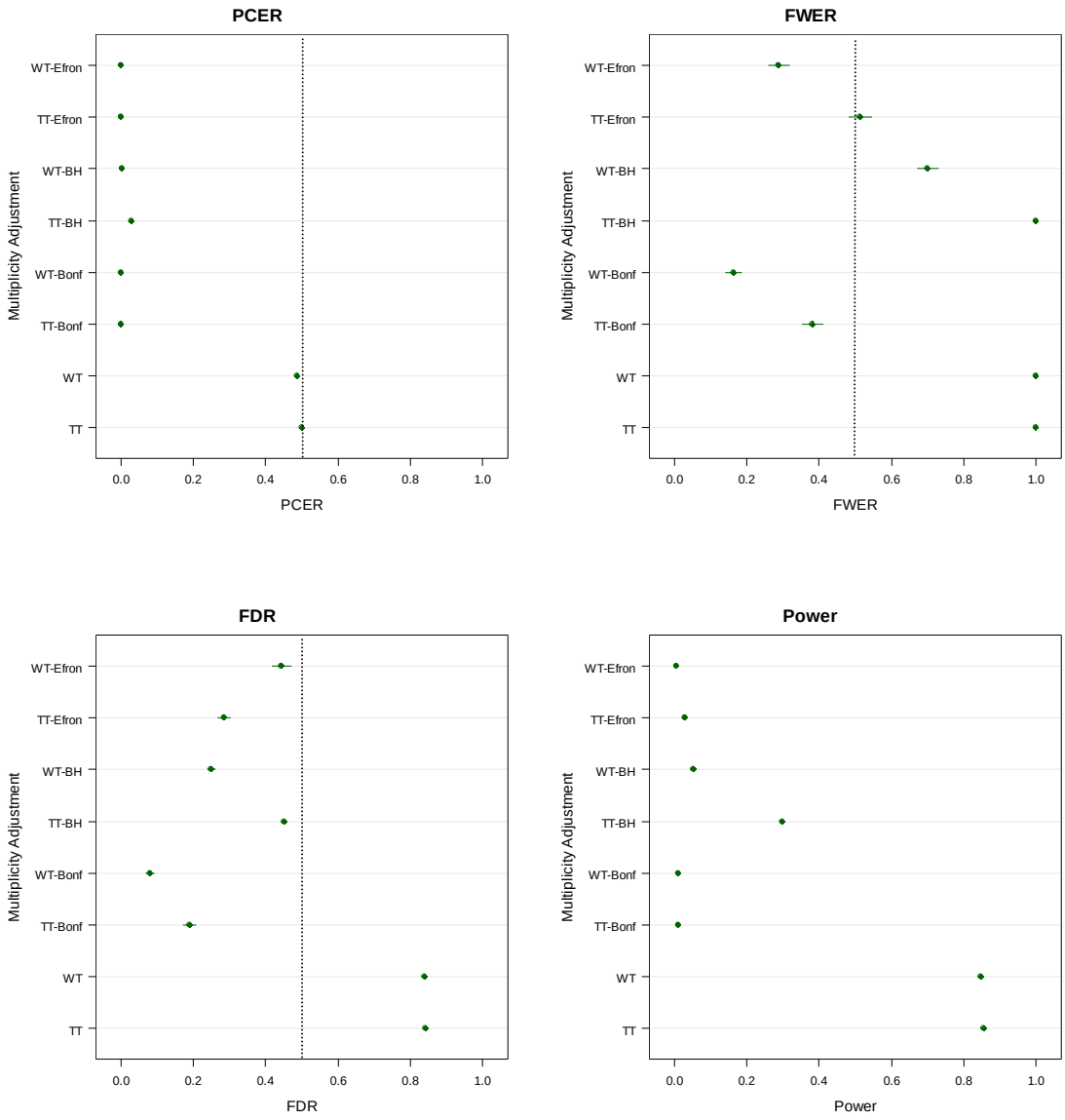

Figure F-65. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 10% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

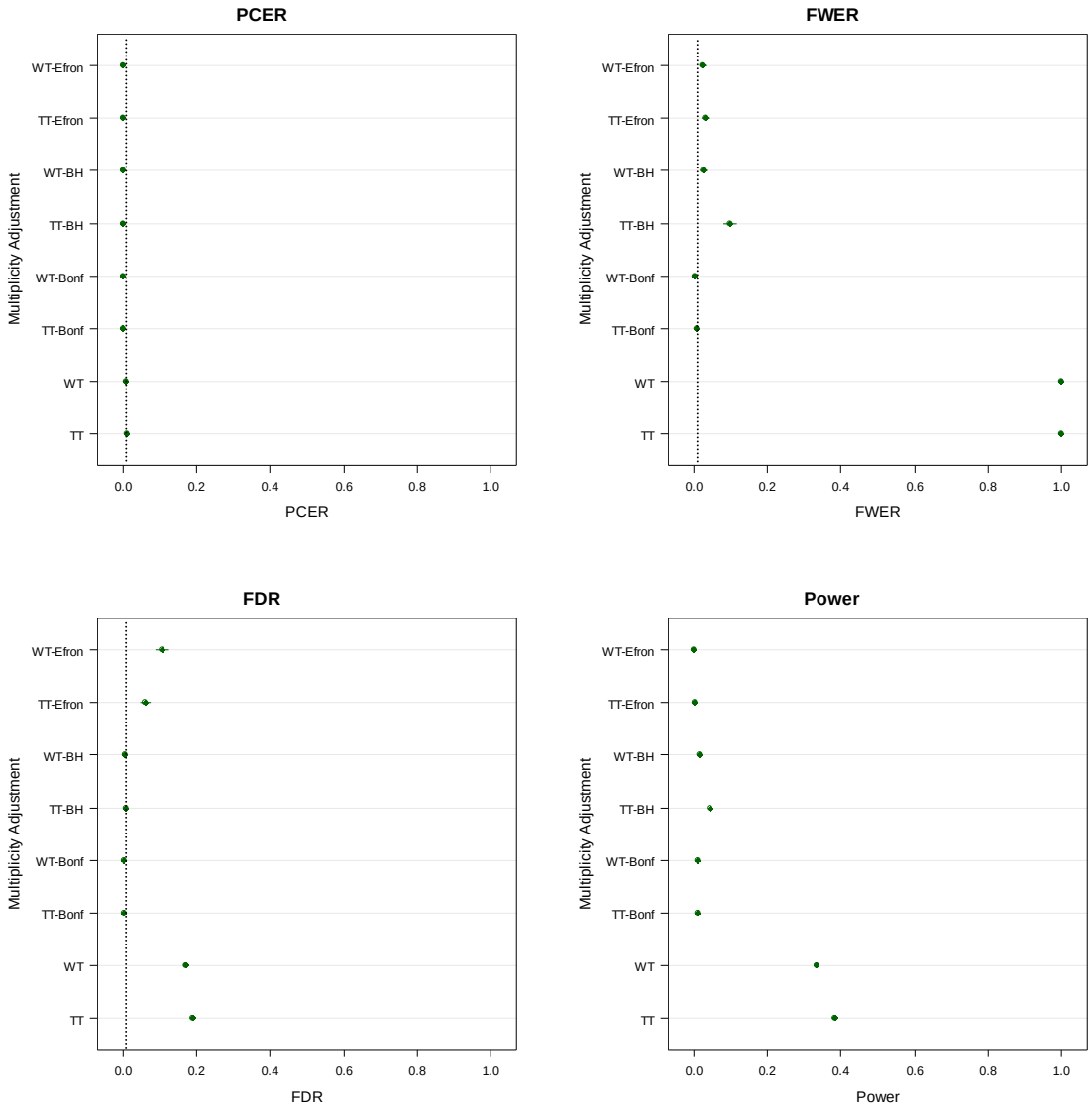

Figure F-66. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

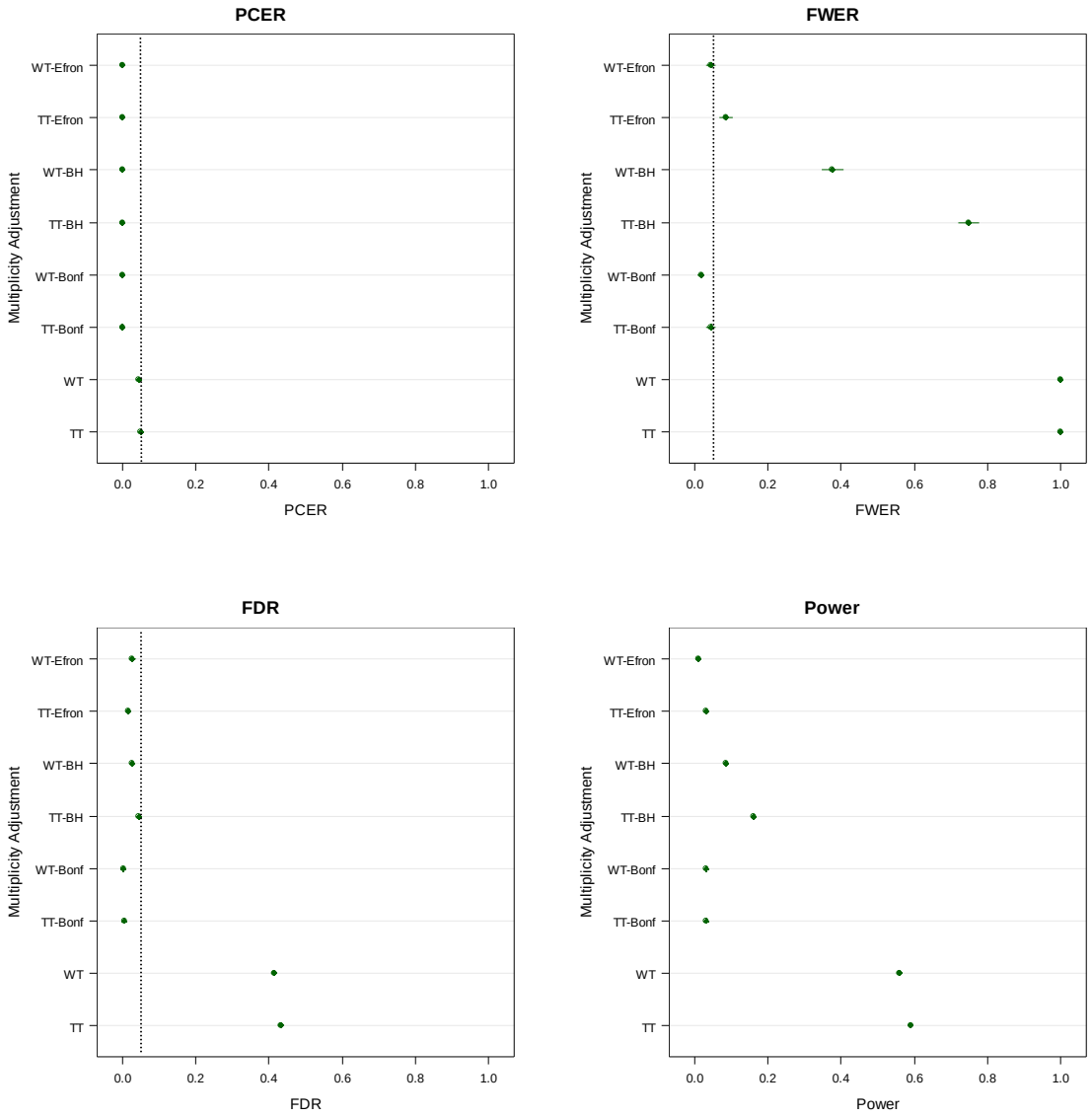

Figure F-67. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

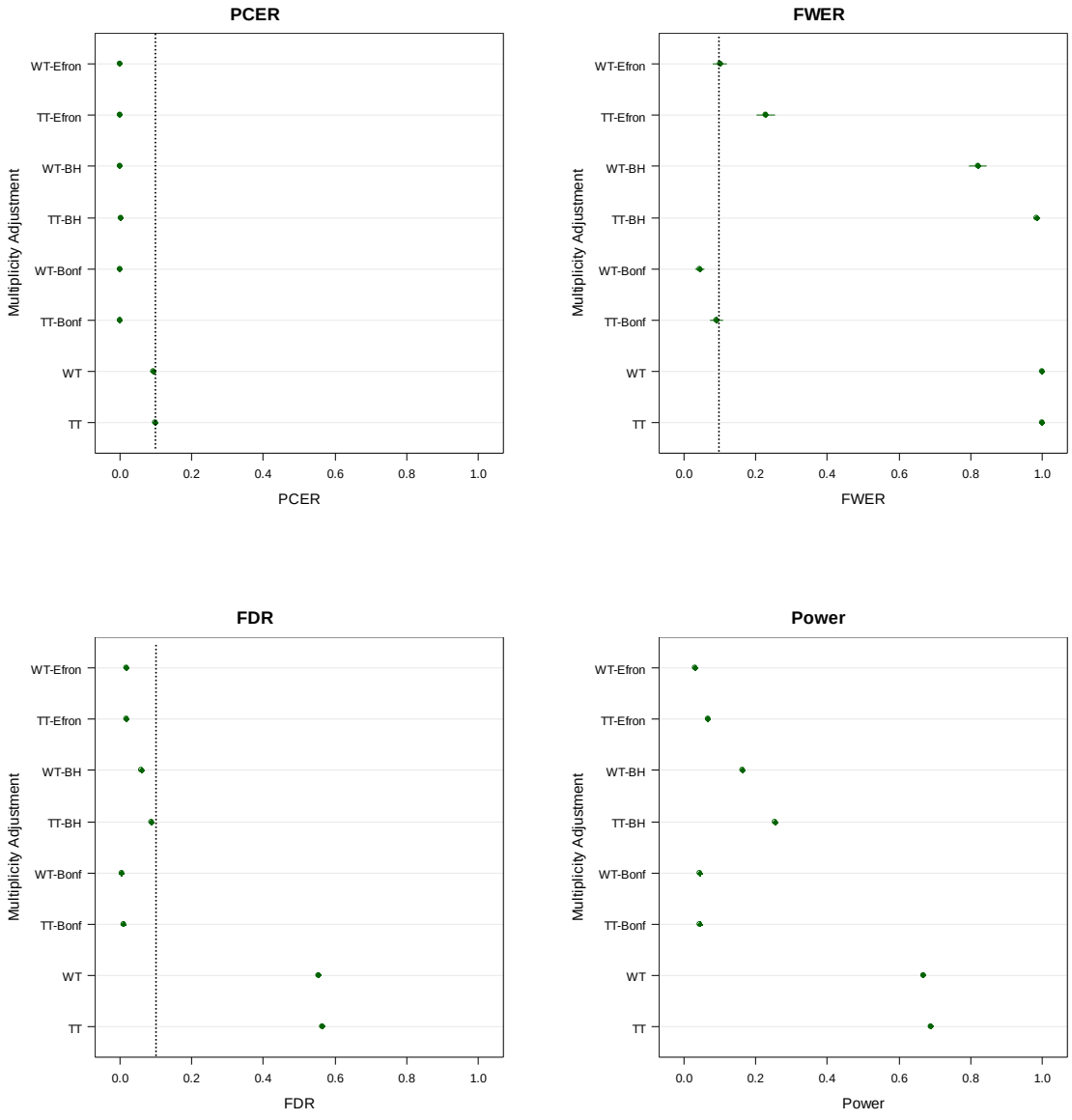

Figure F-68. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

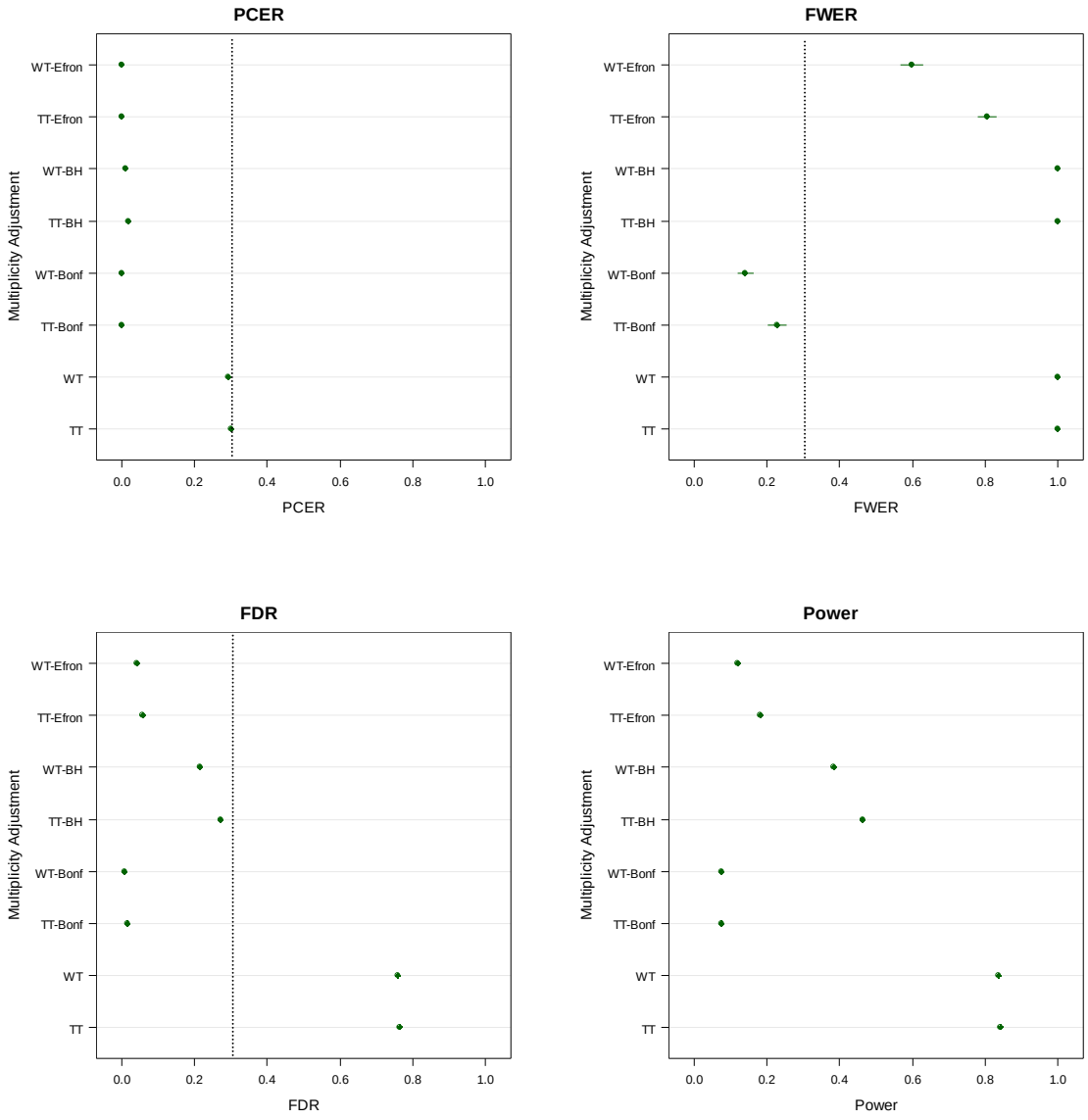

Figure F-69. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

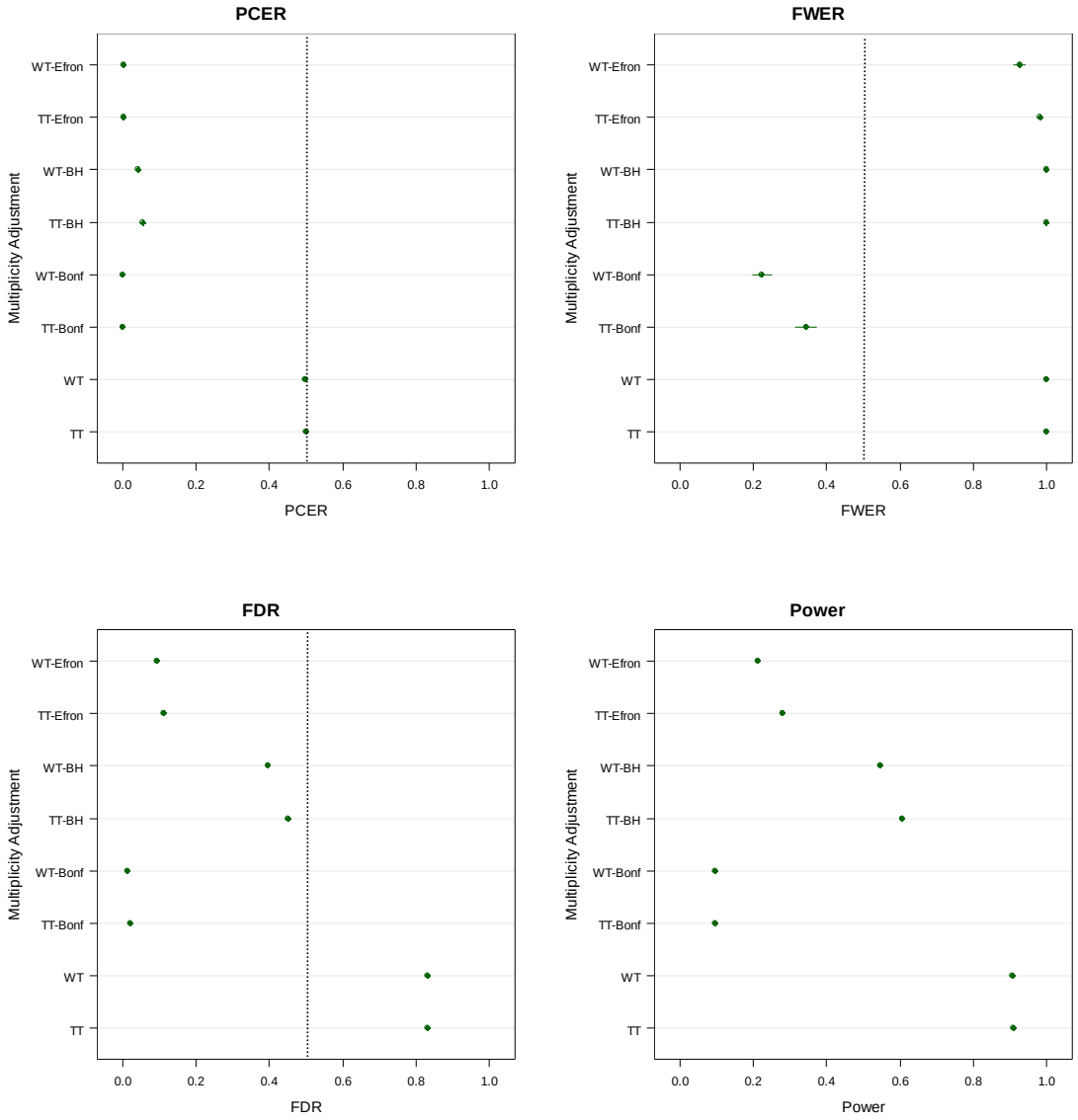

Figure F-70. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

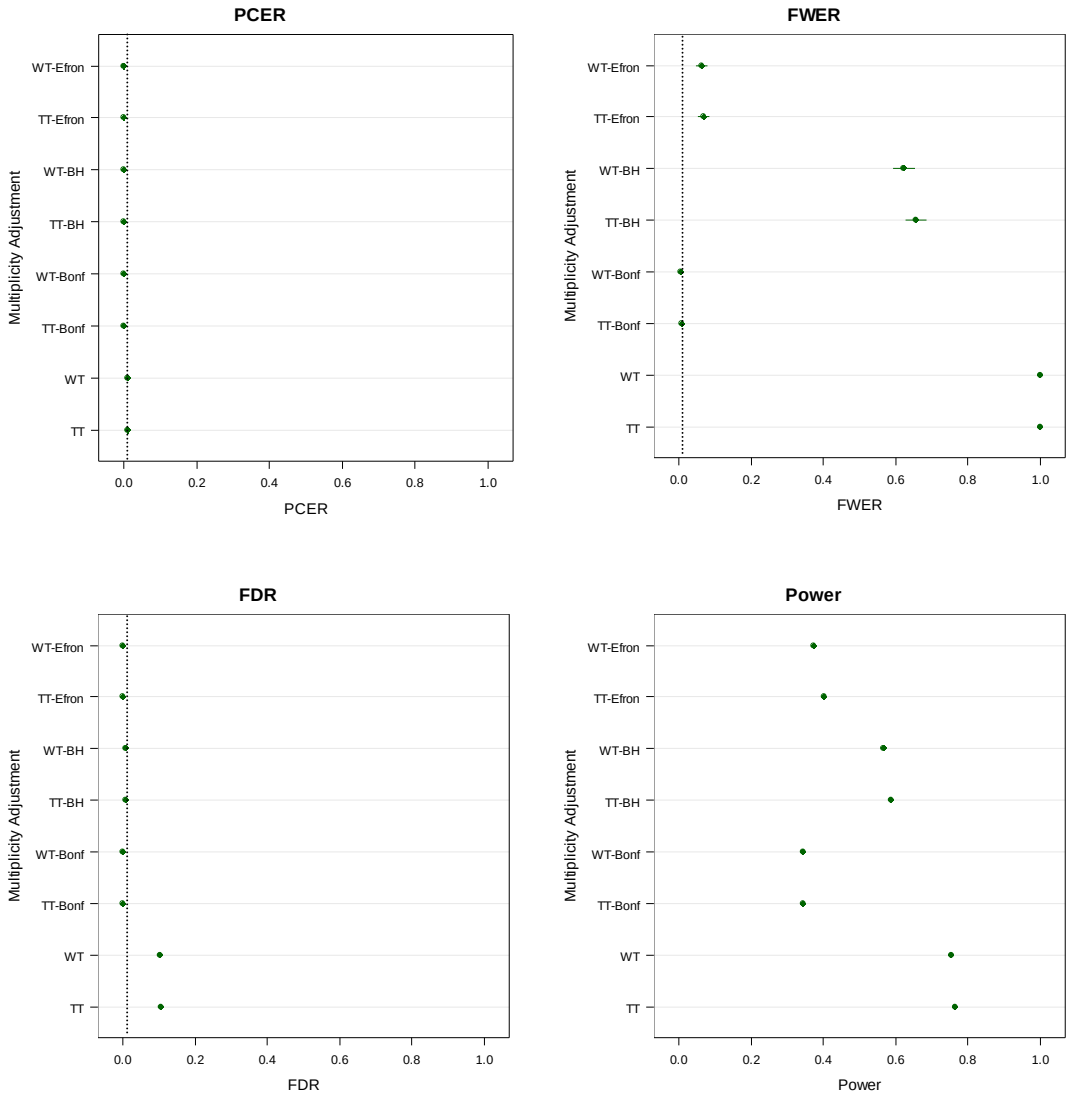

Figure F-71. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

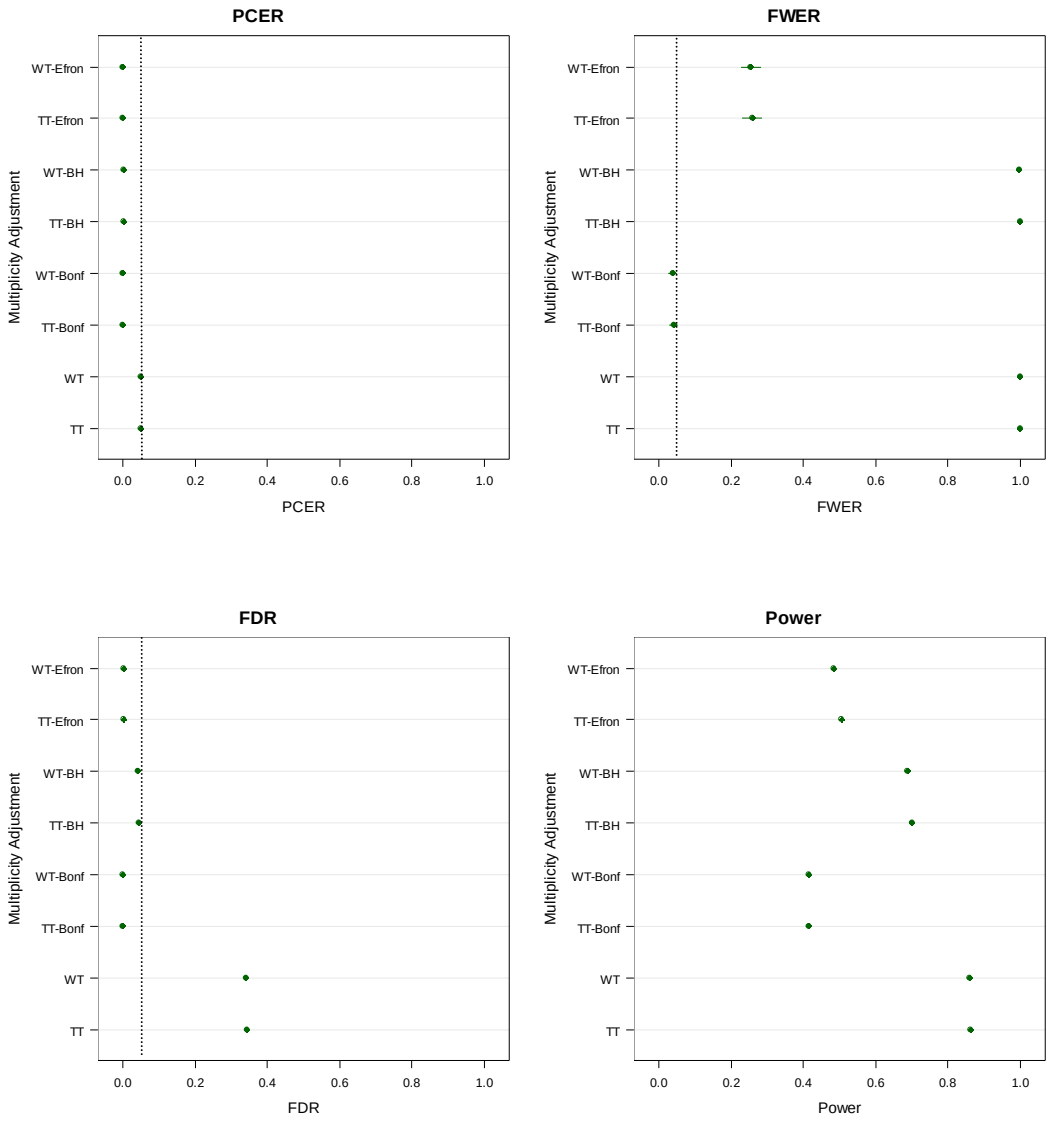

Figure F-72. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

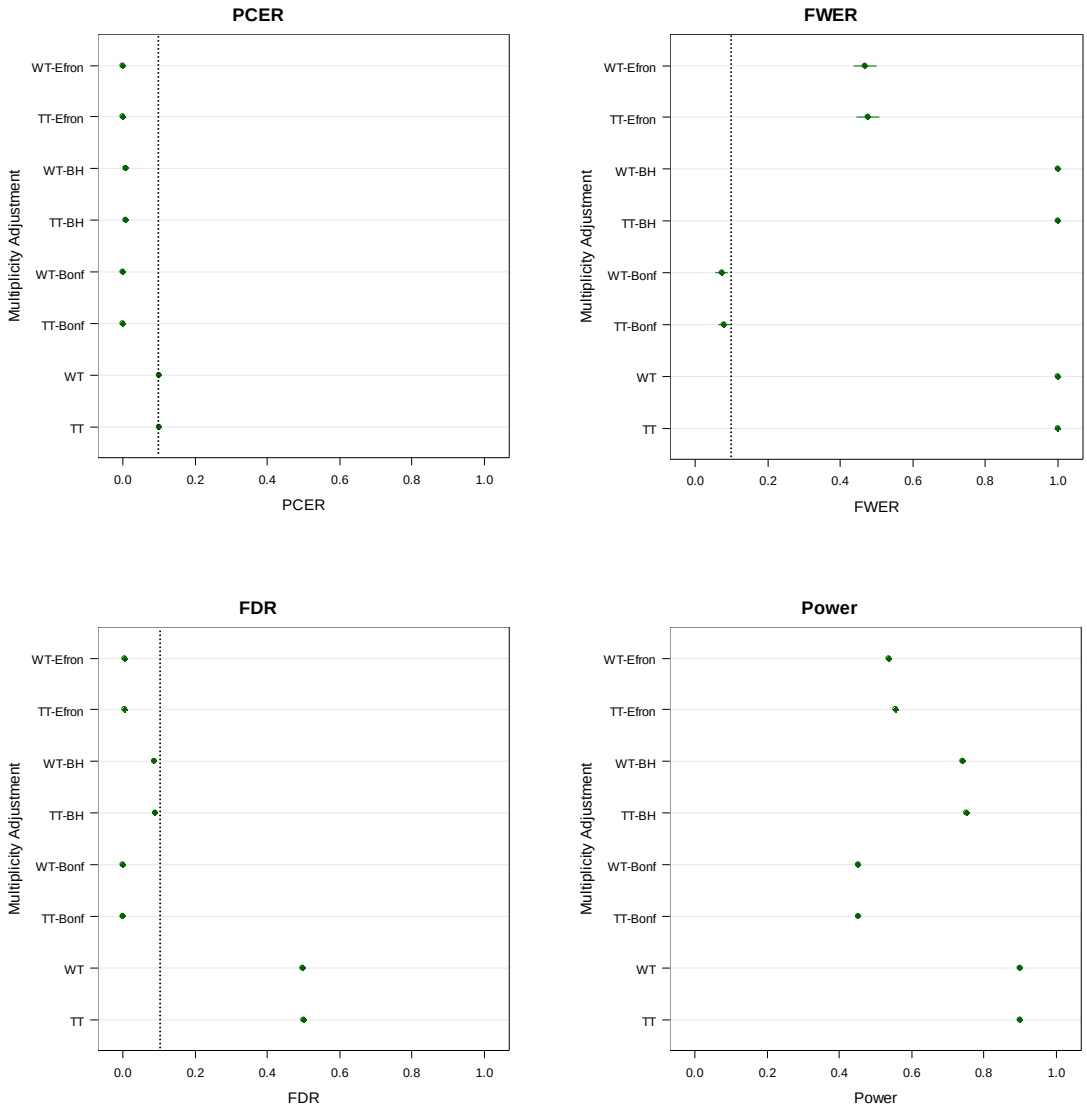

Figure F-73. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

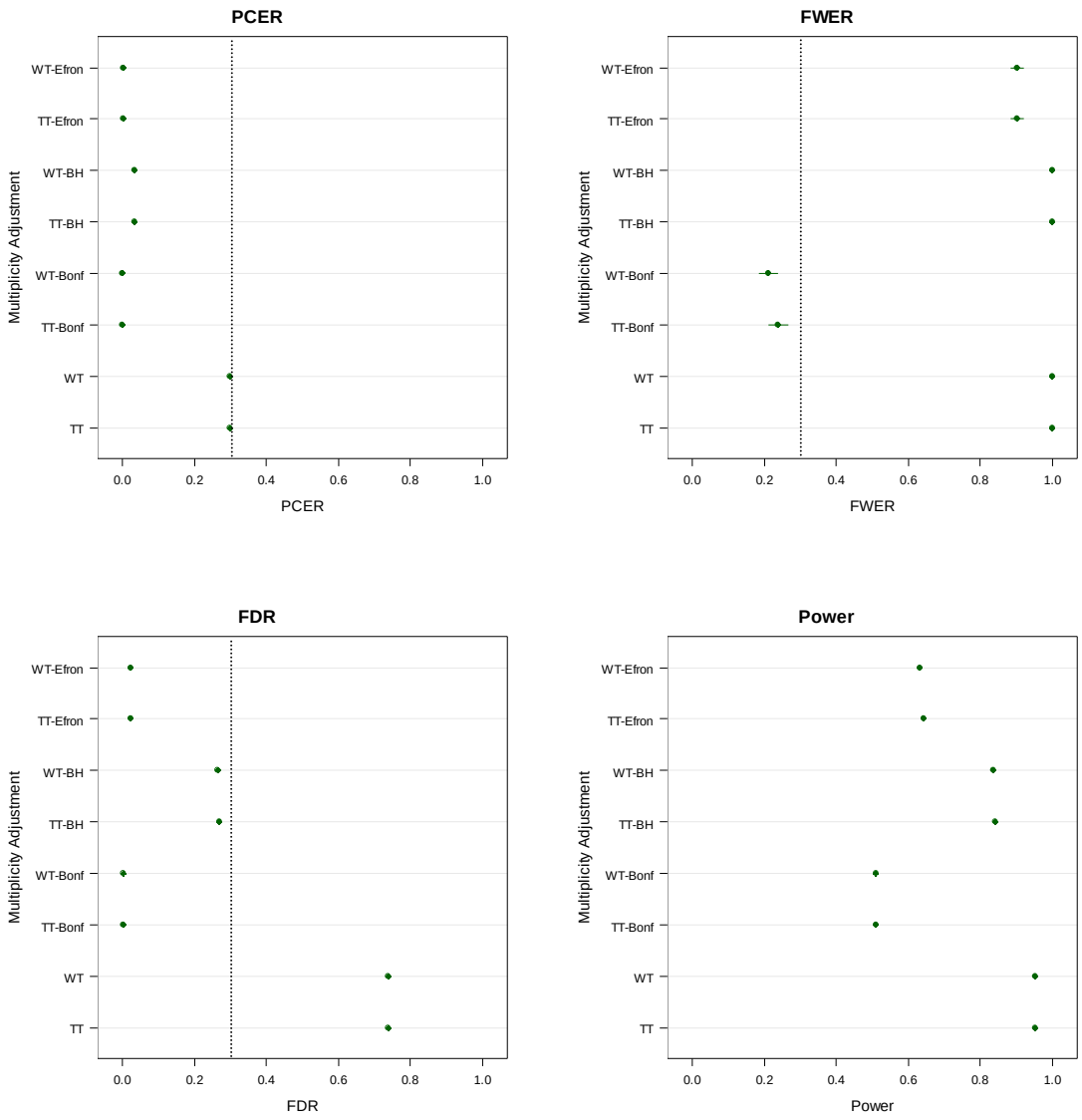

Figure F-74. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

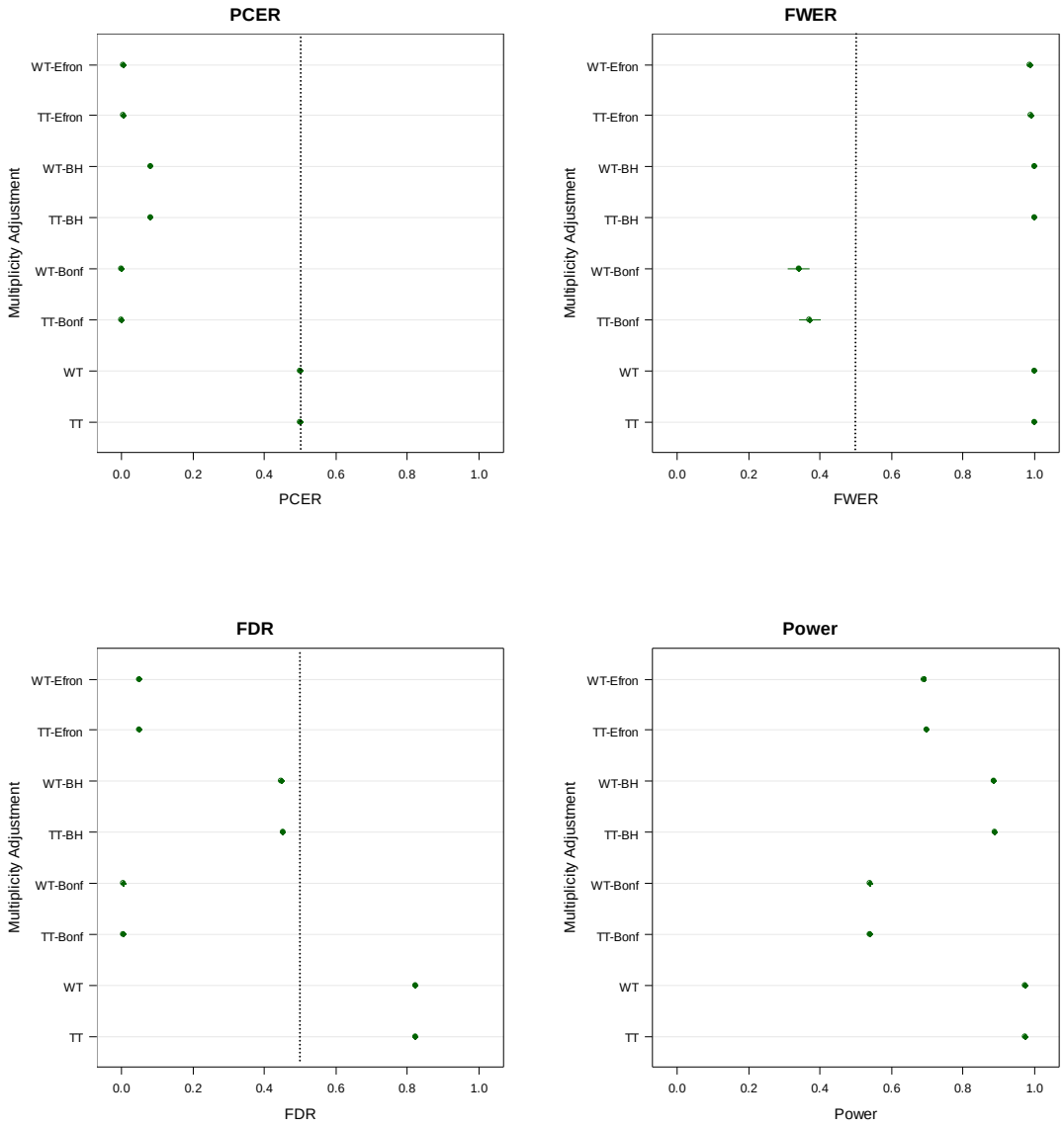

Figure F-75. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

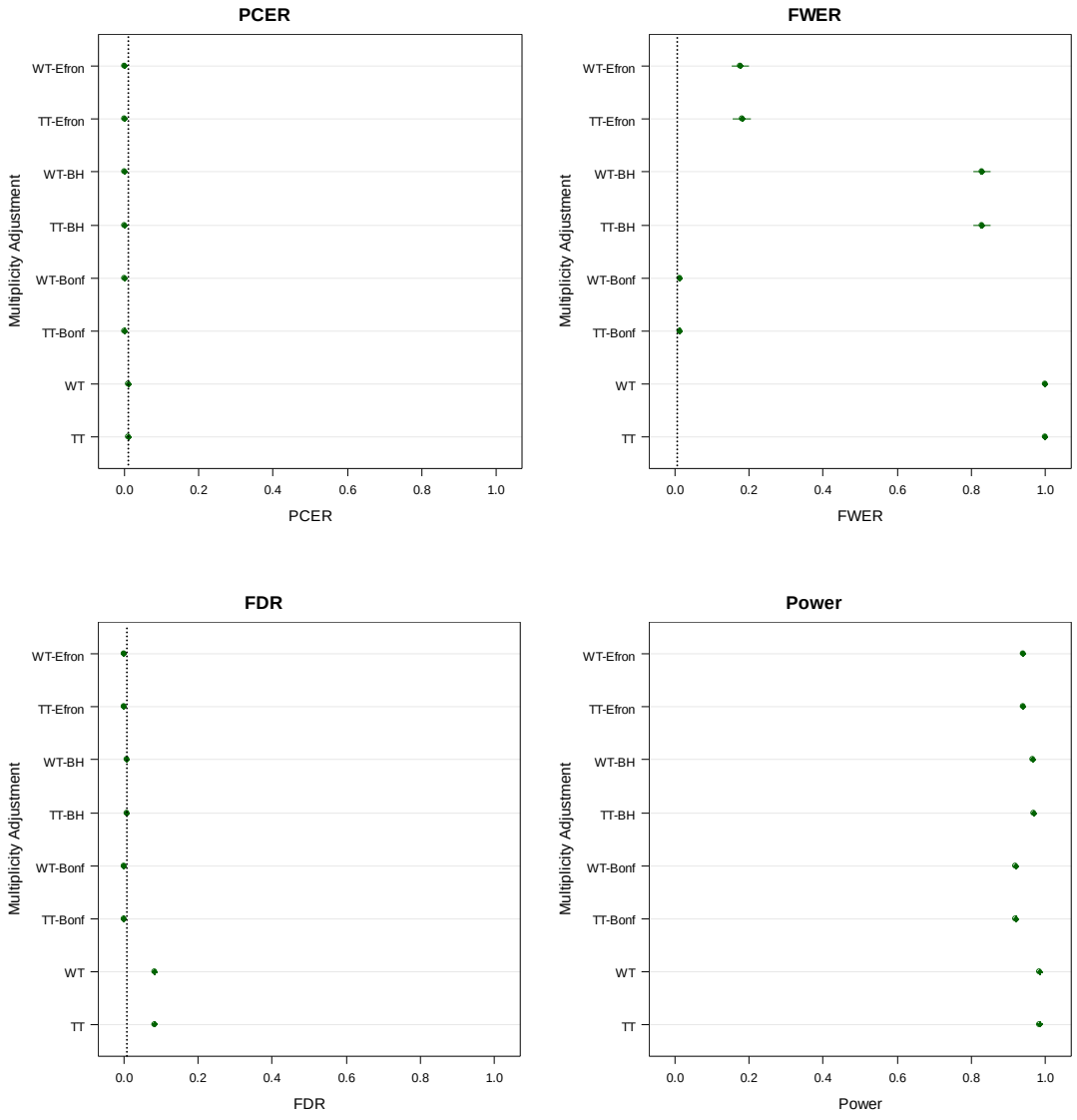

Figure F-76. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 10% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

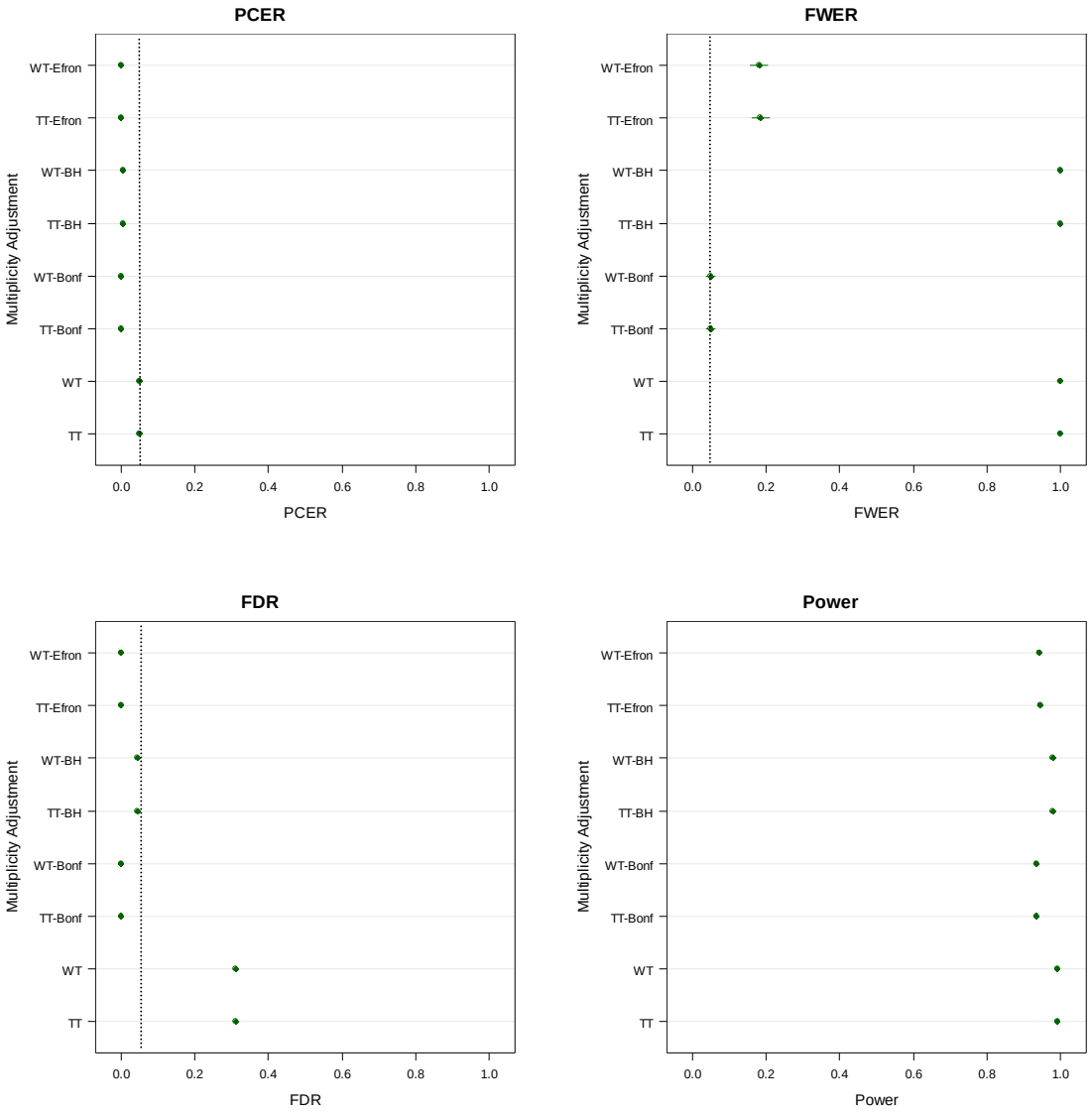

Figure F-77. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 10% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

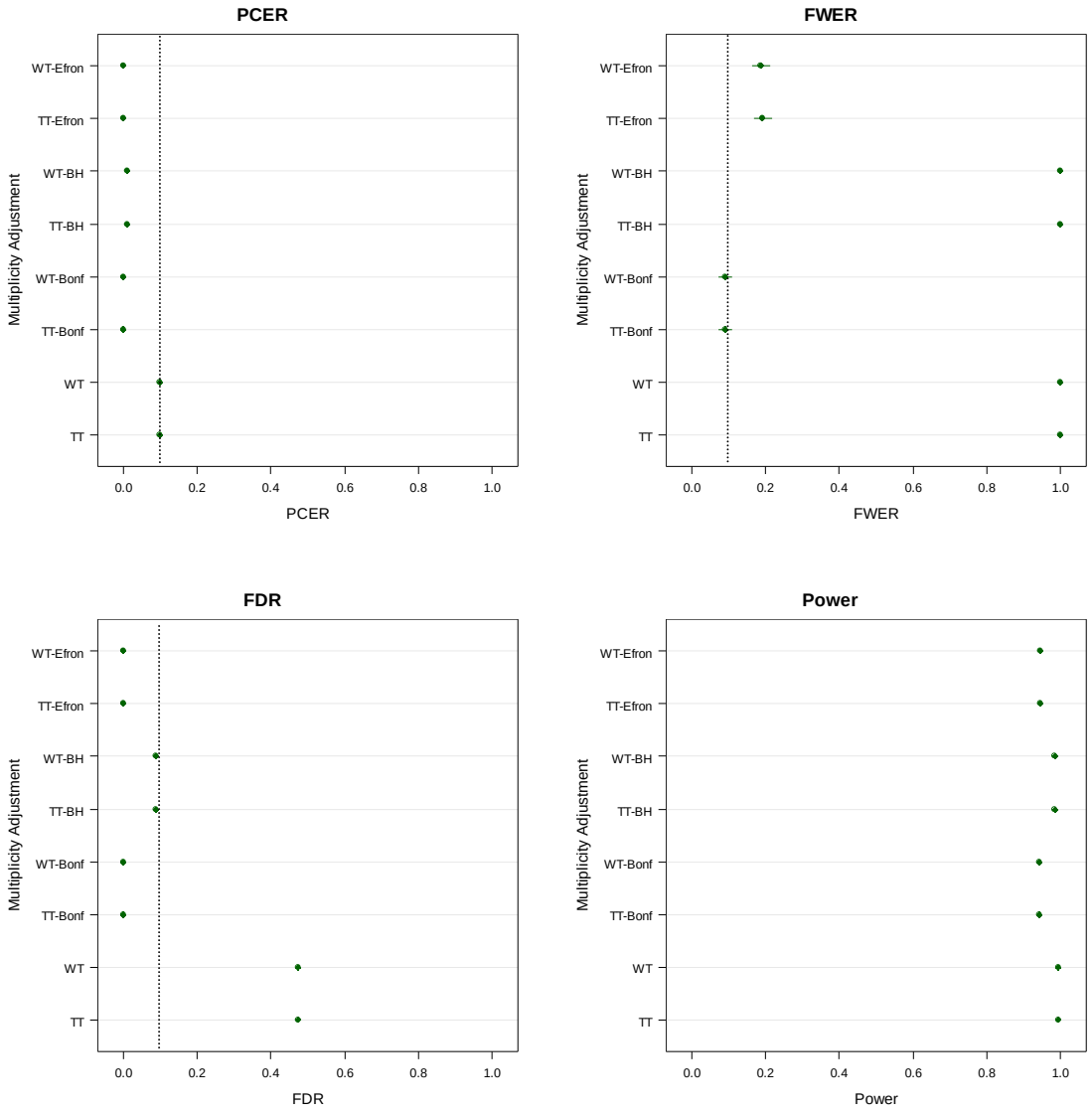

Figure F-78. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 10% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

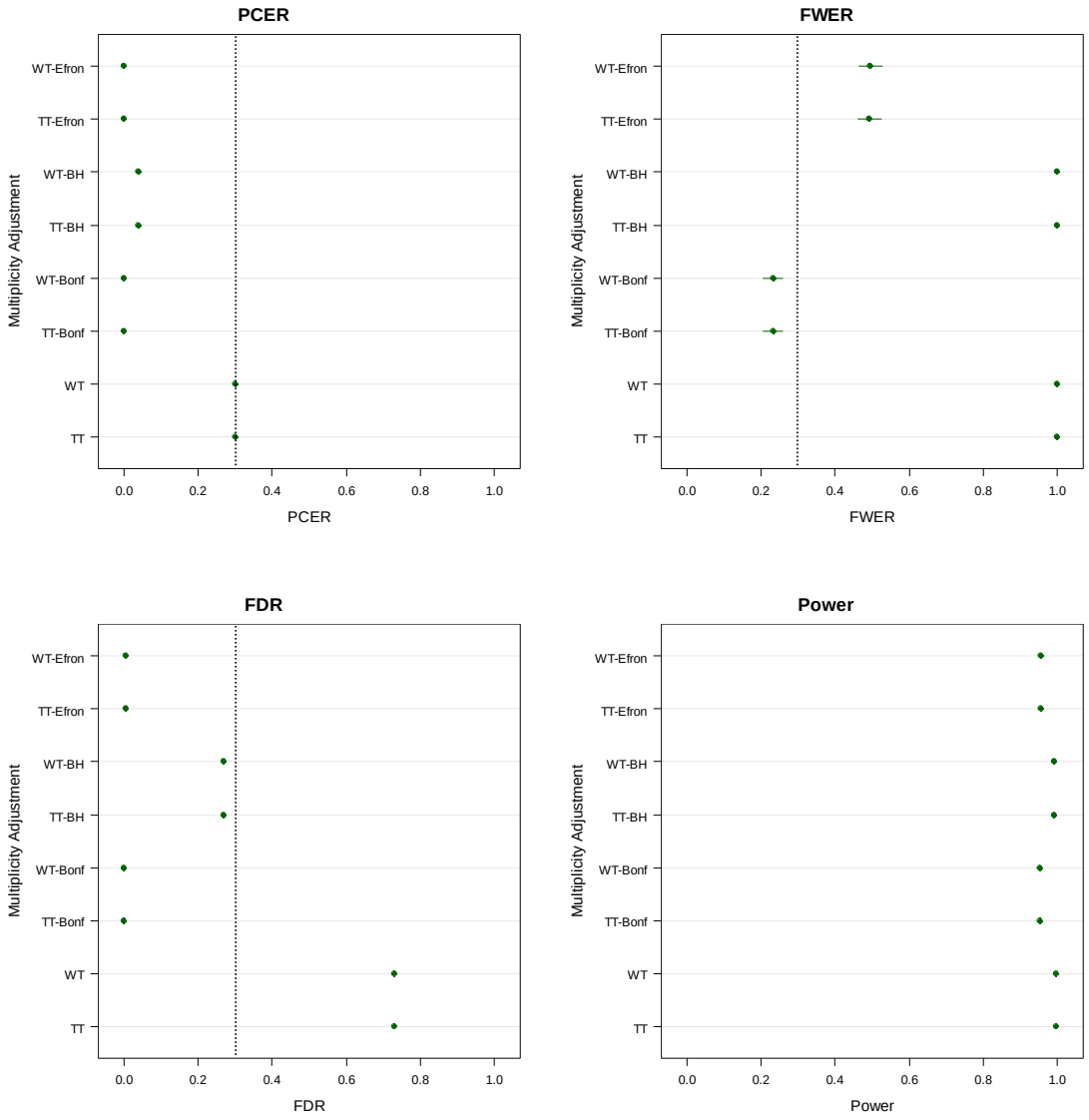

Figure F-79. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 10% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

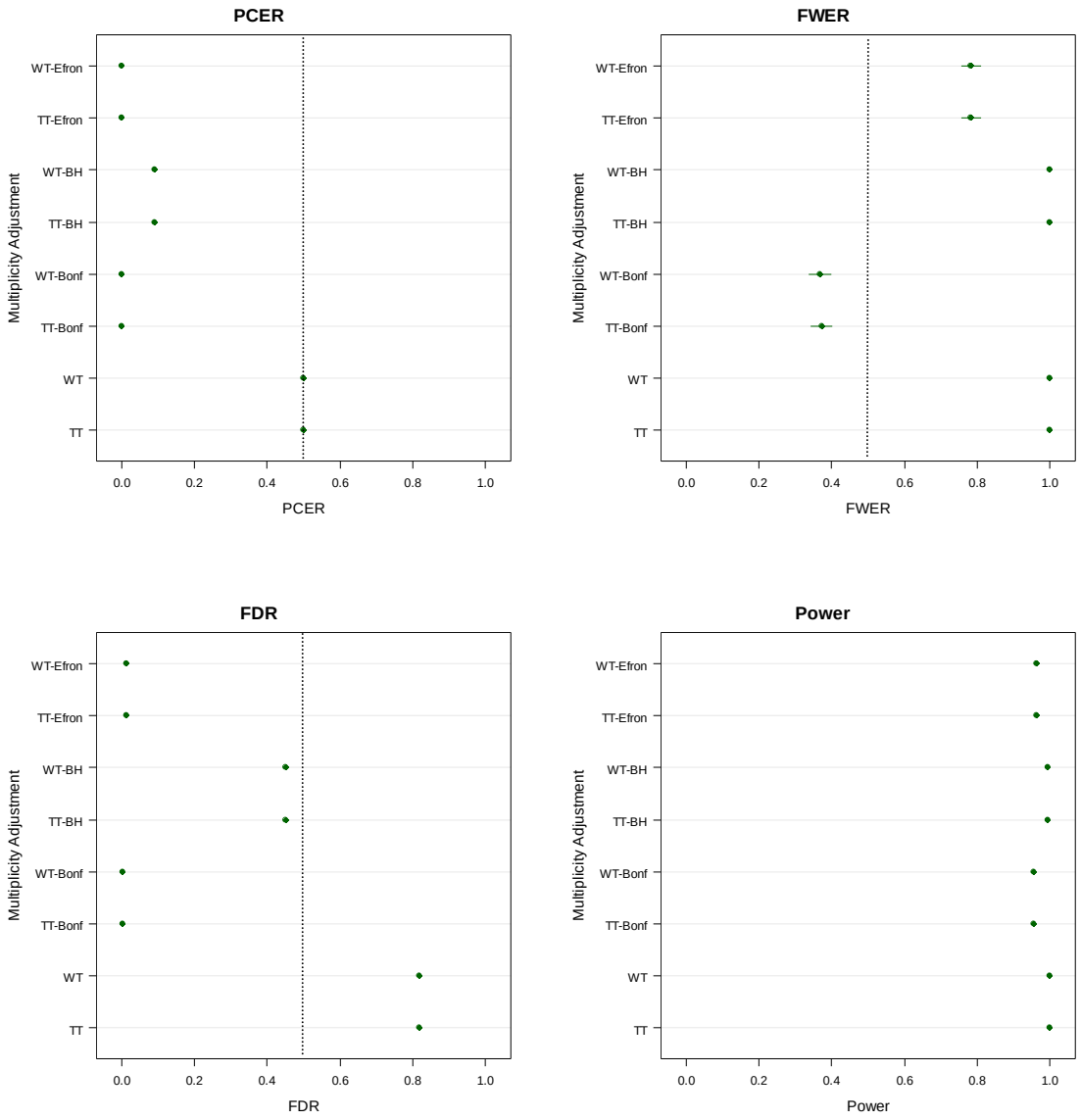

Figure F-80. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes, 10% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

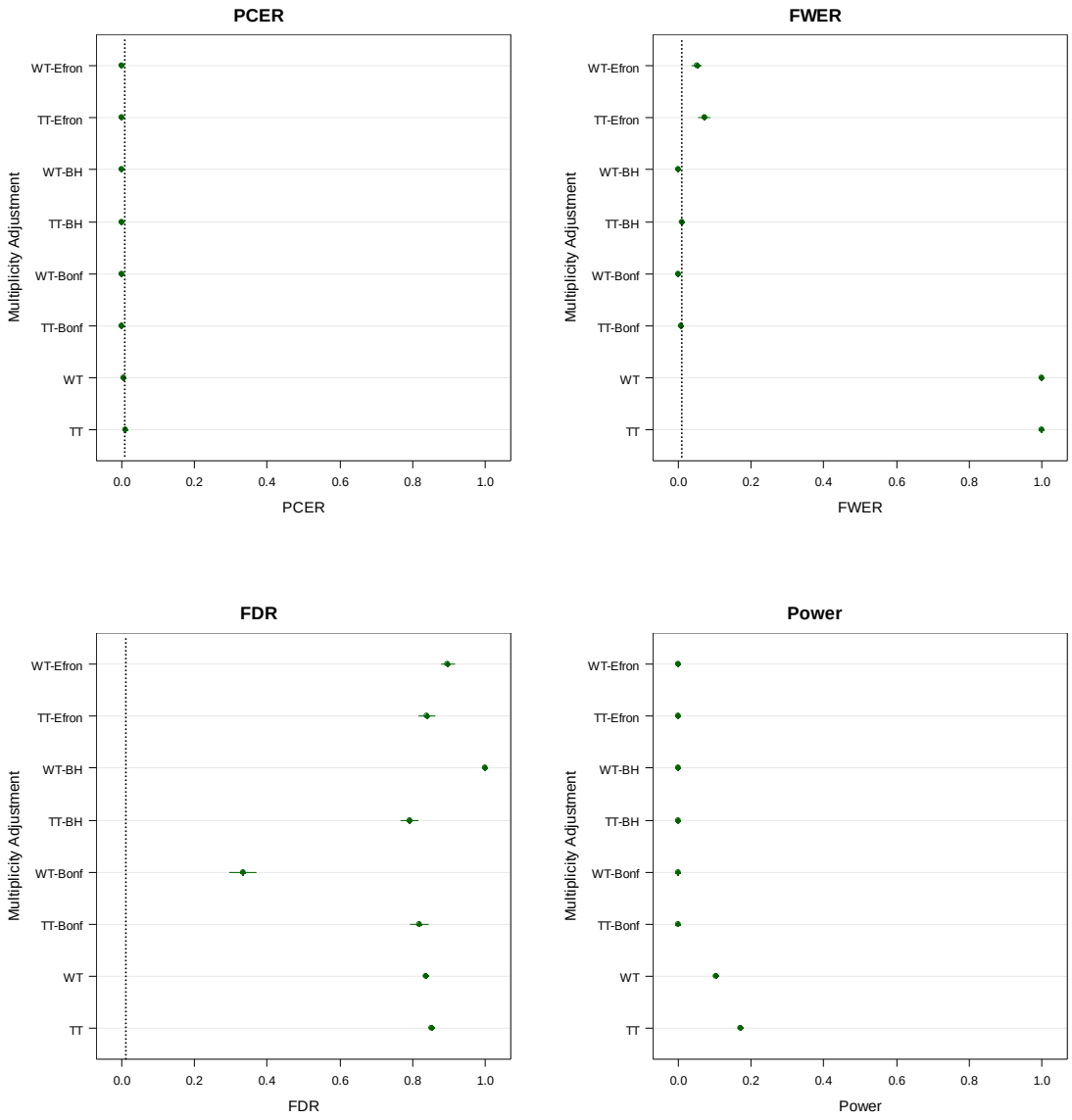

Figure F-81. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 1% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

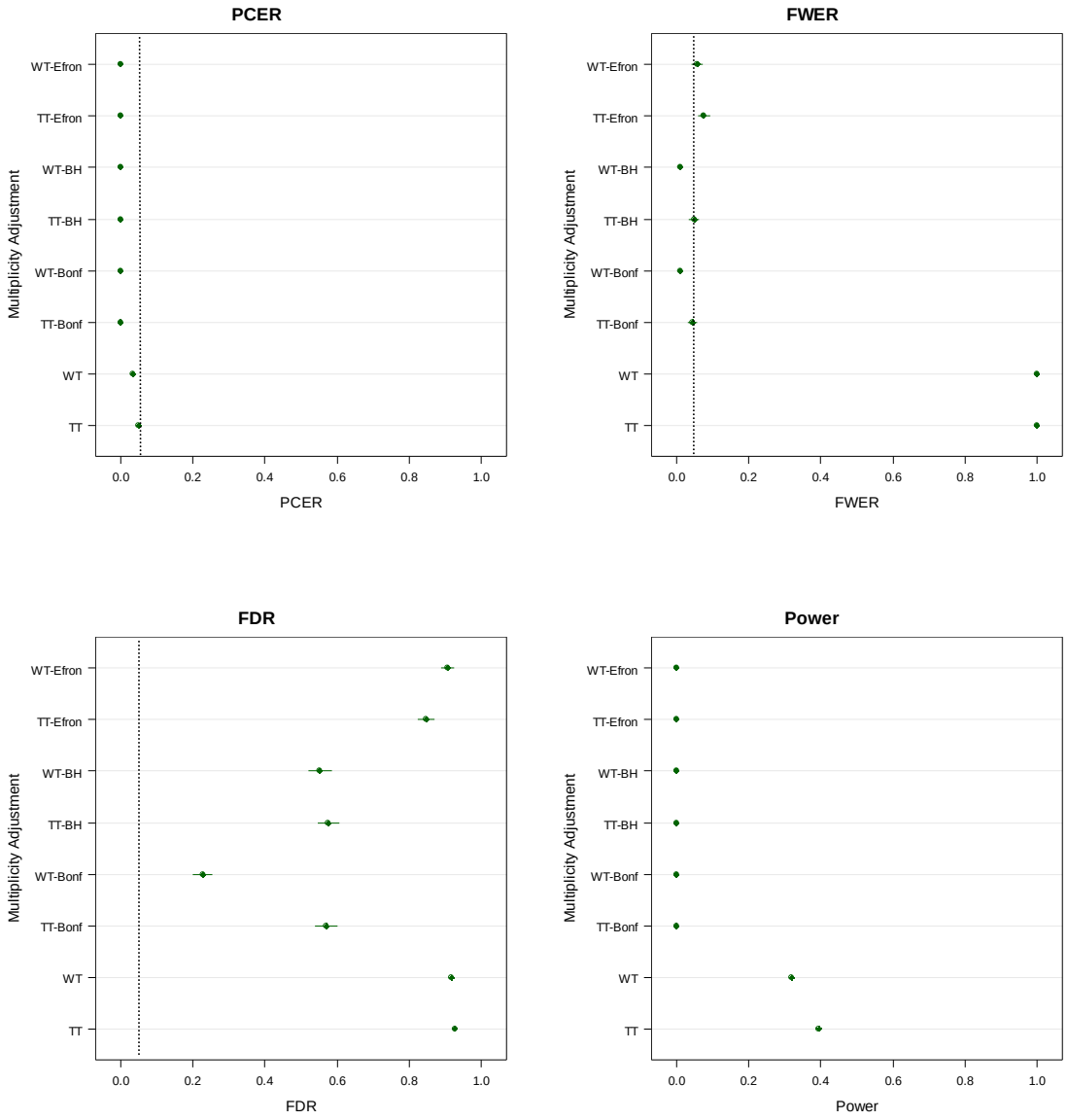

Figure F-82. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 1% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

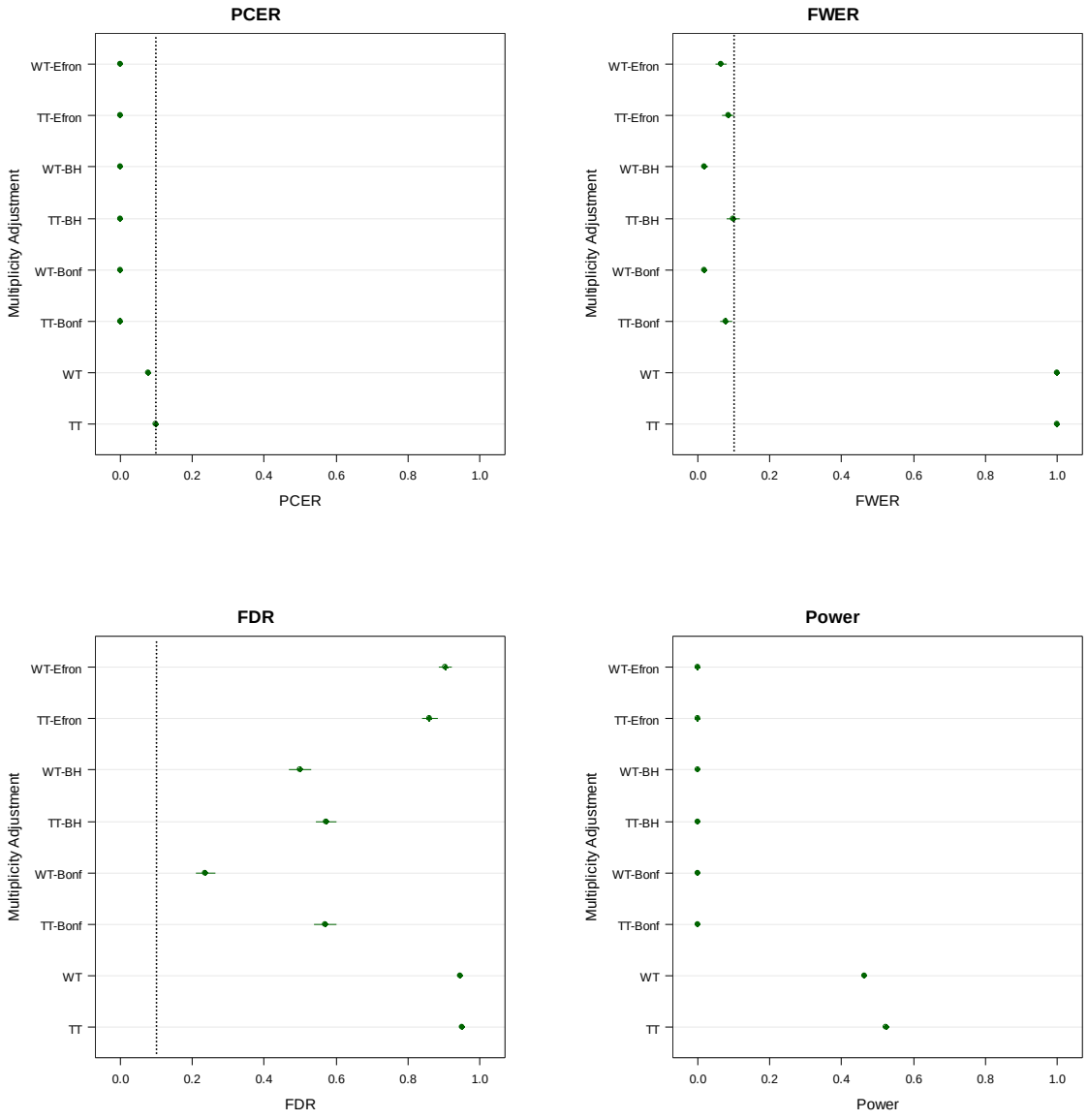

Figure F-83. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 1% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

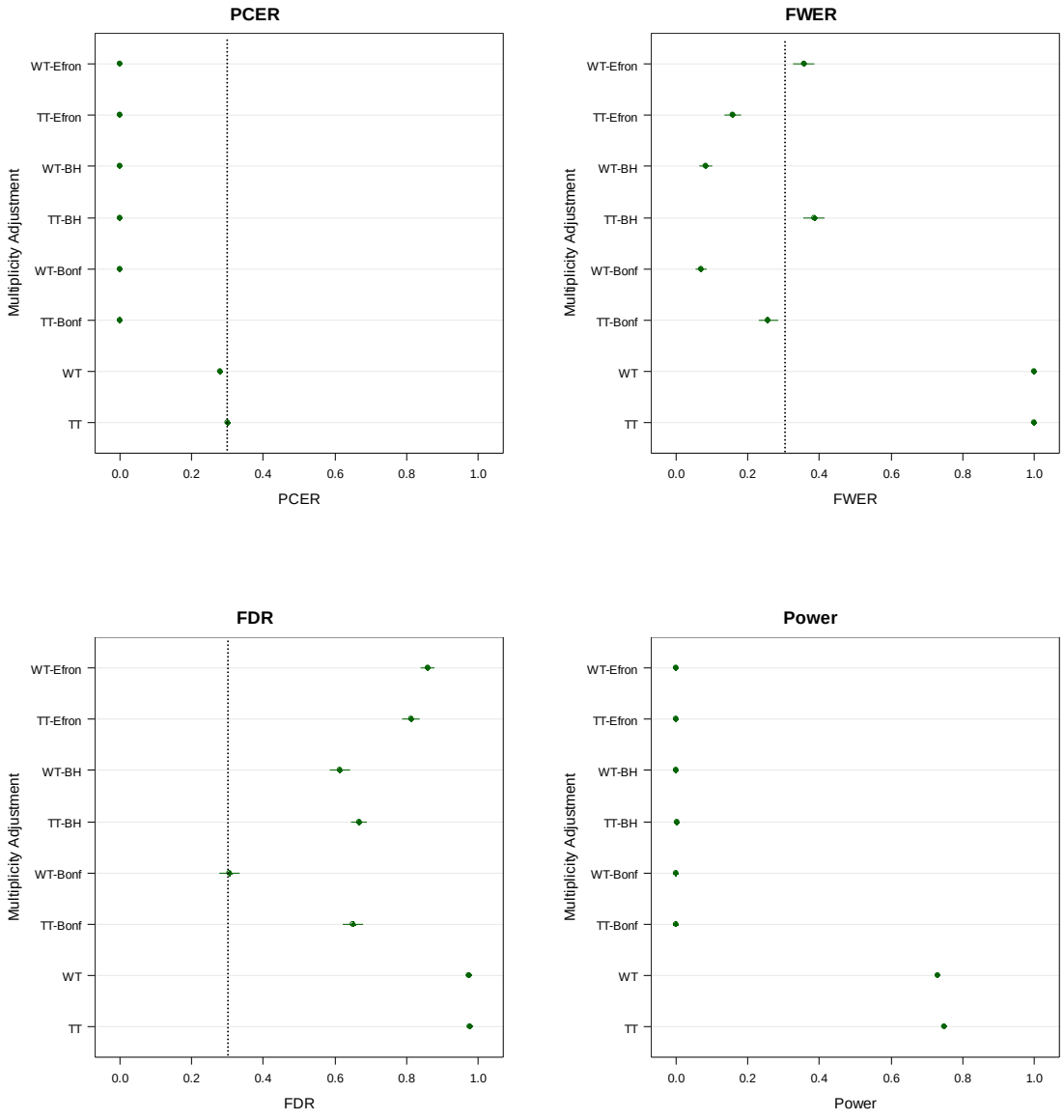

Figure F-84. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 1% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

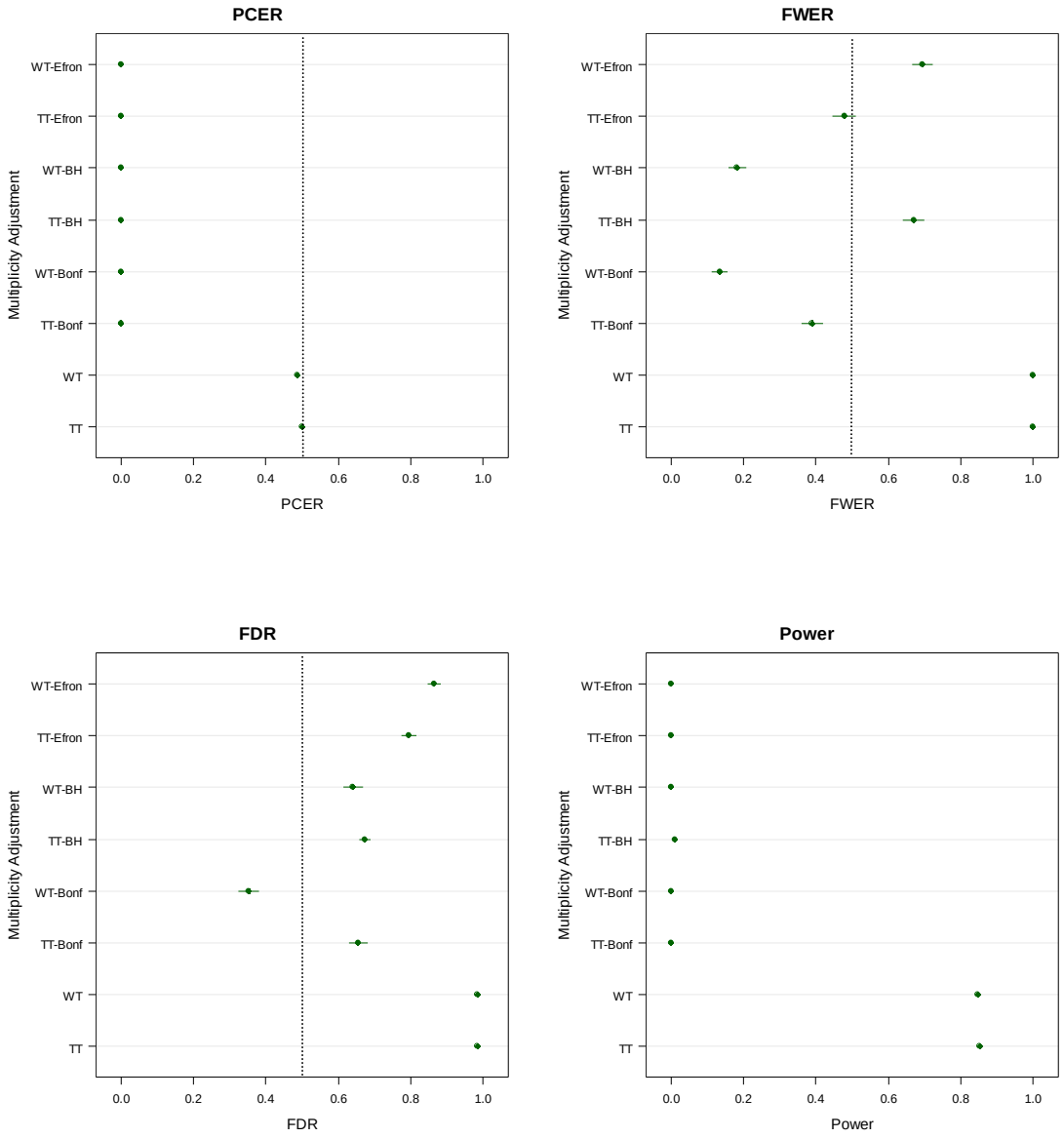

Figure F-85. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 1% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

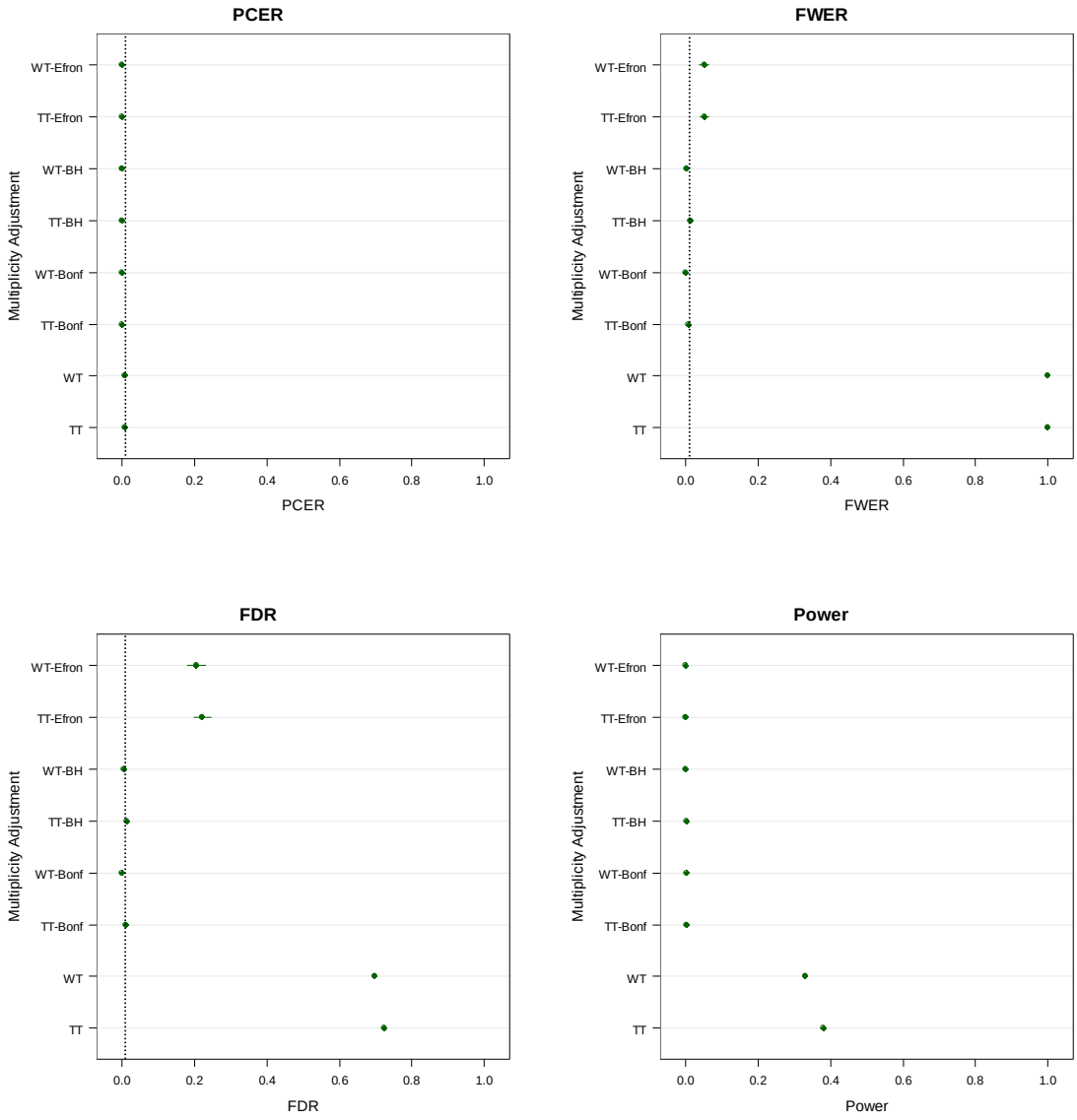

Figure F-86. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 1% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

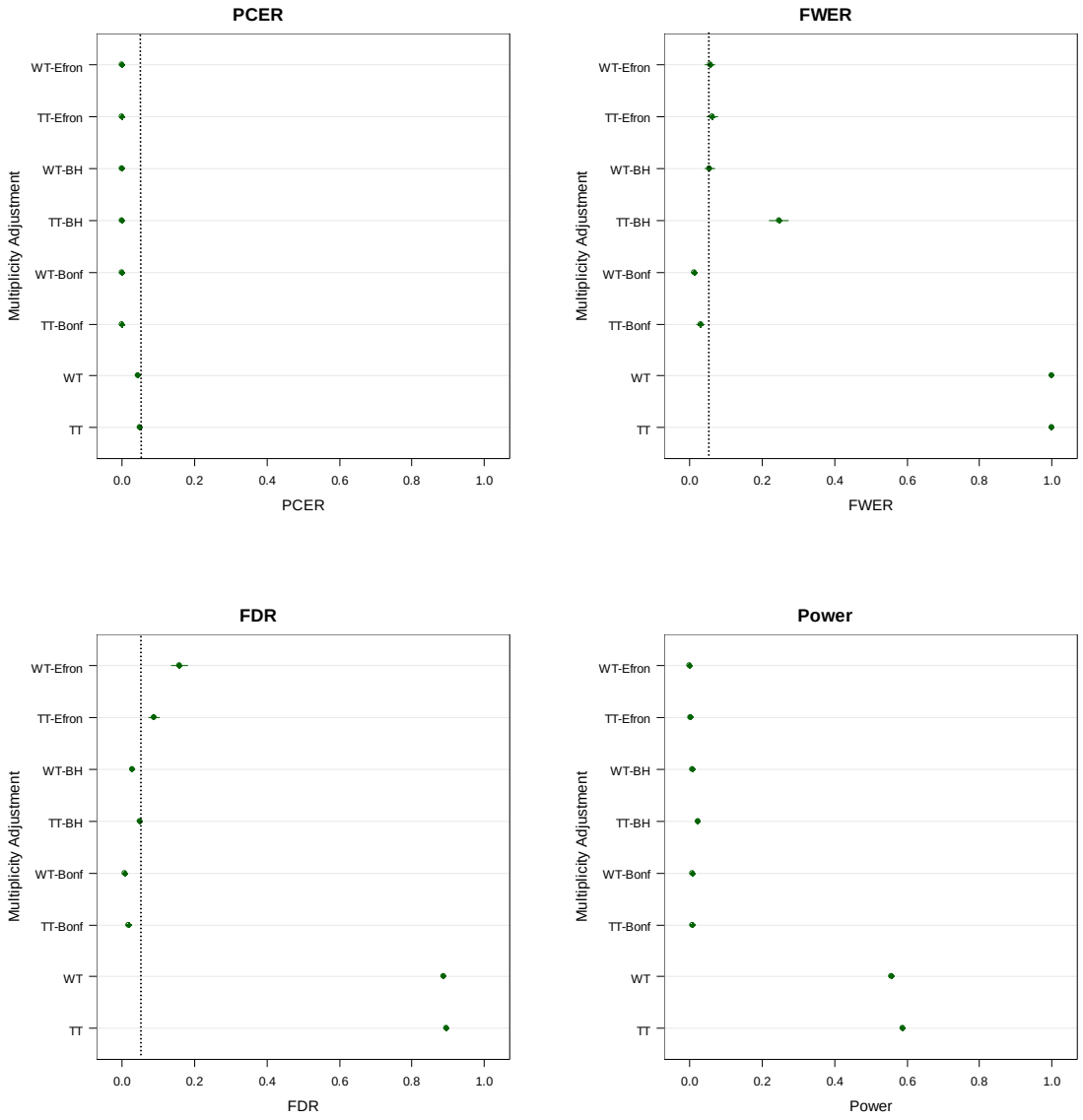

Figure F-87. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 1% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

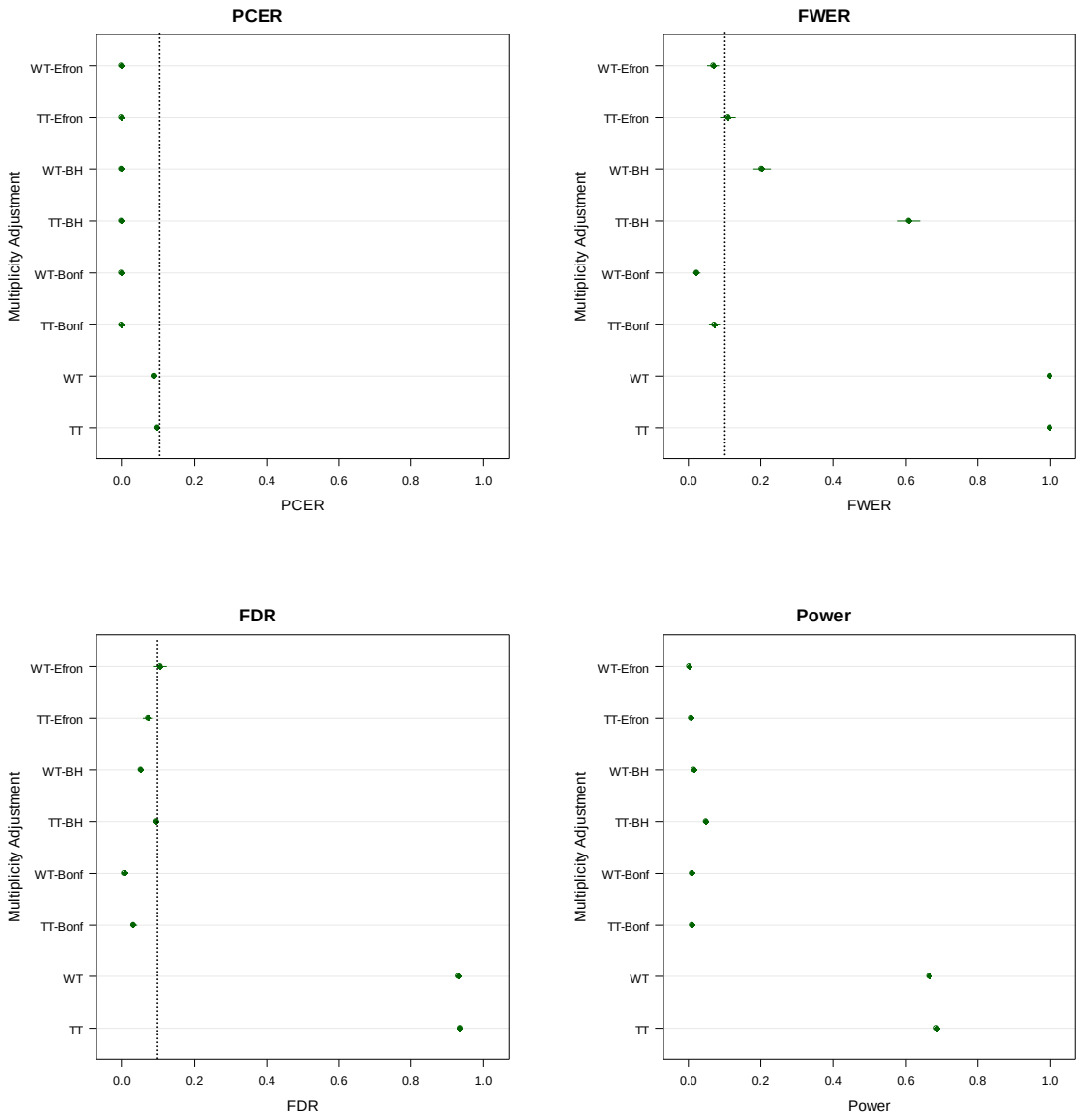

Figure F-88. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 1% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

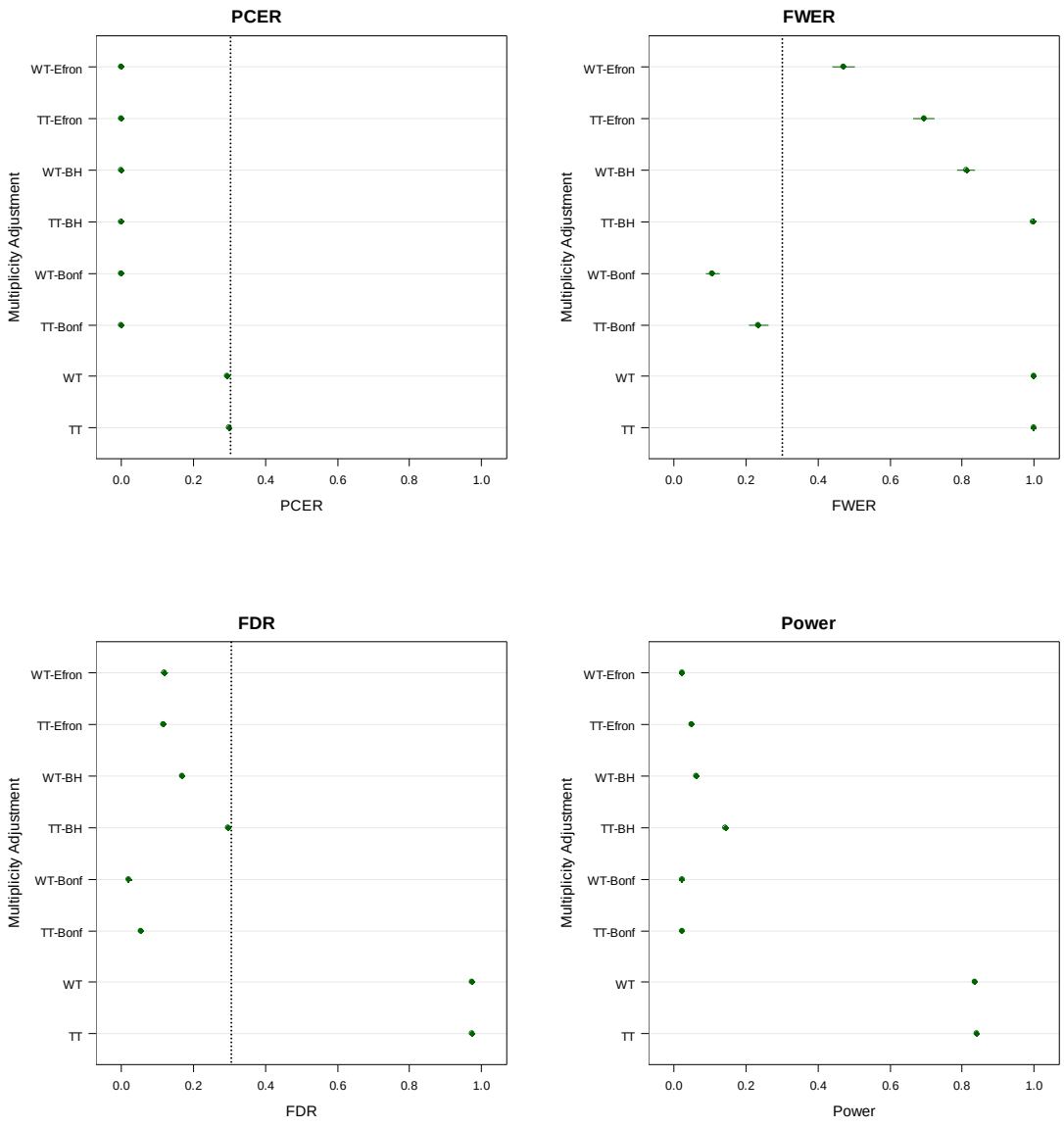

Figure F-89. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 1% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

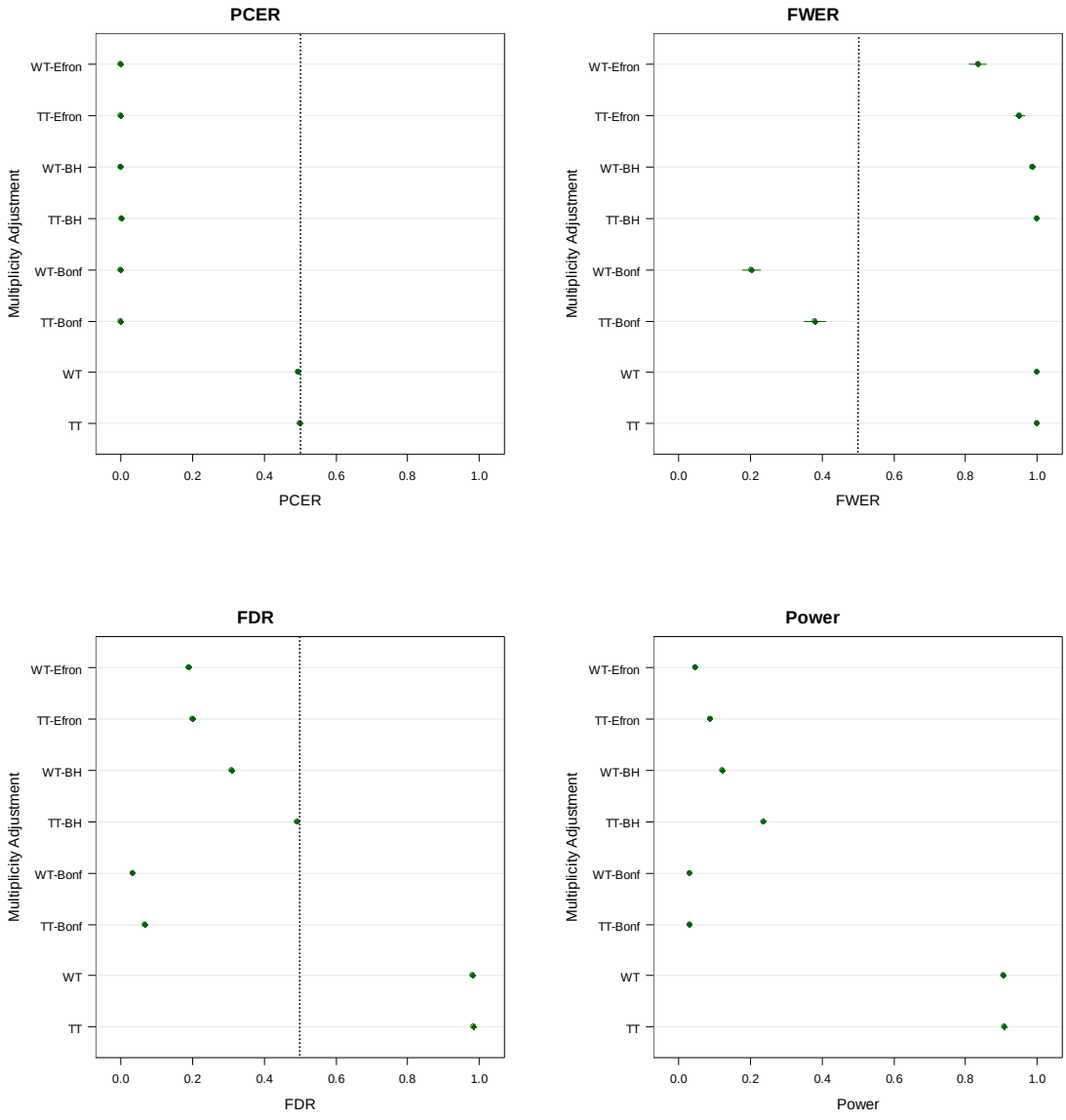

Figure F-90. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 1% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

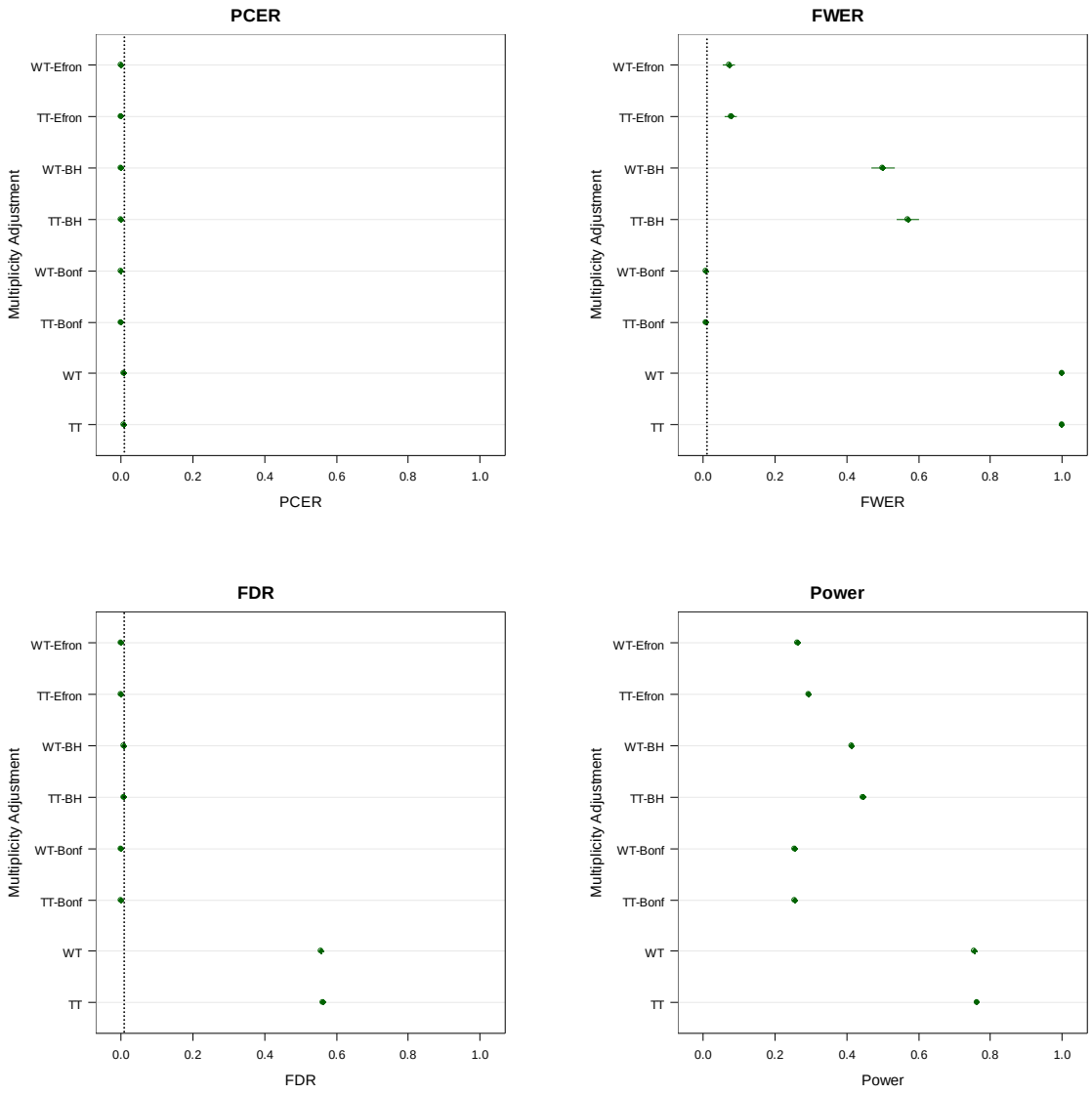

Figure F-91. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 1% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

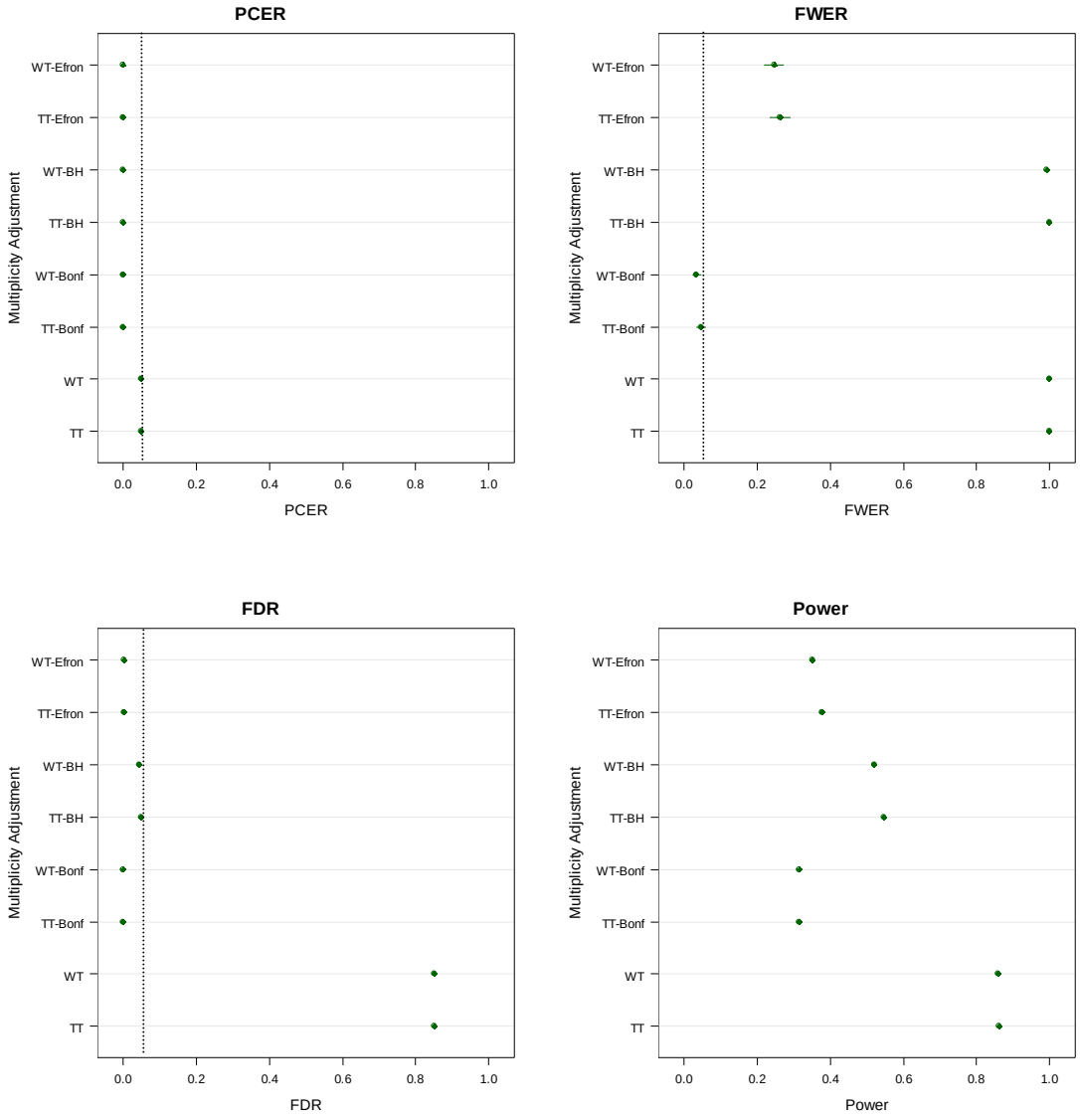

Figure F-92. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 1% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

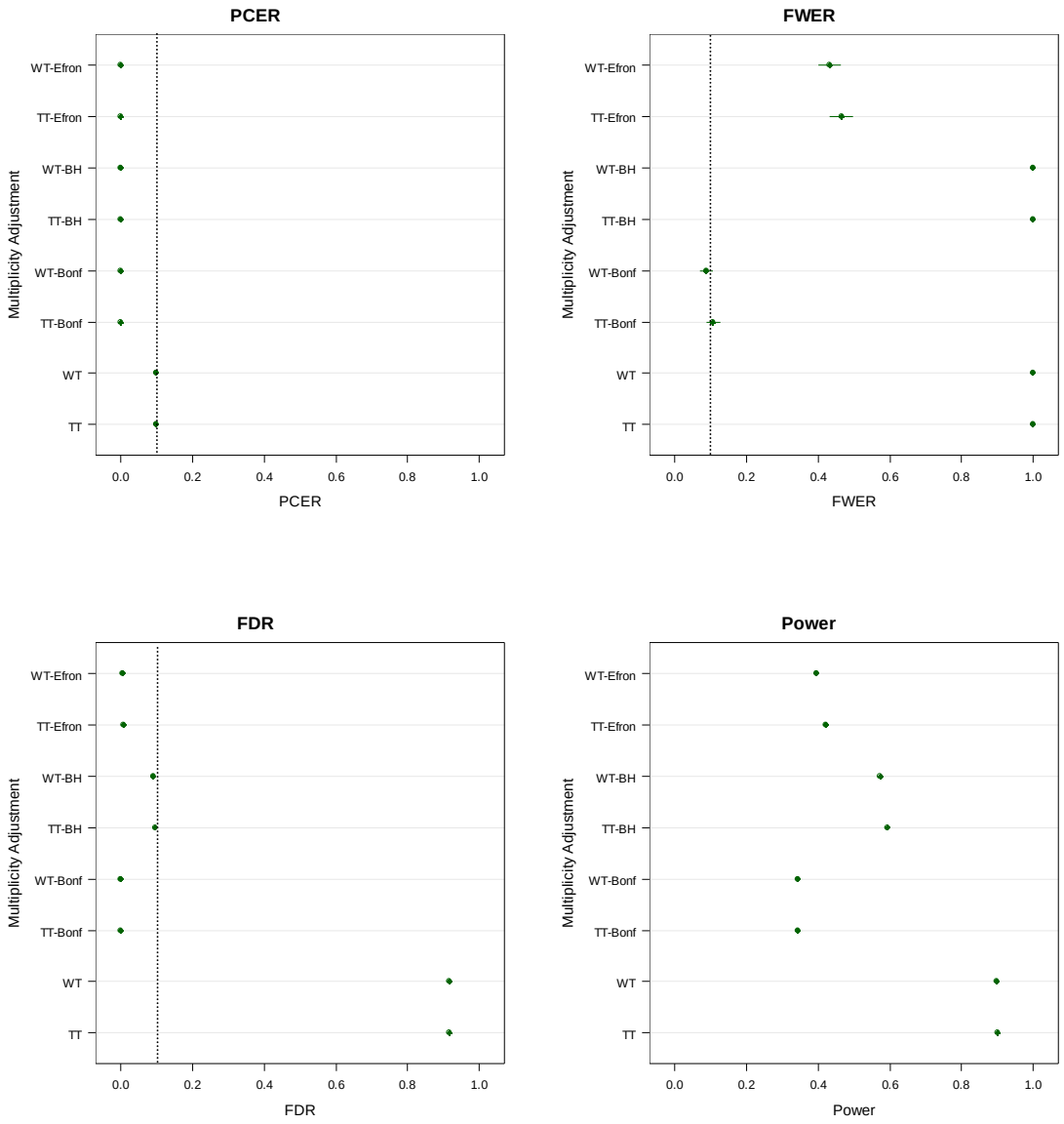

Figure F-93. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 1% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

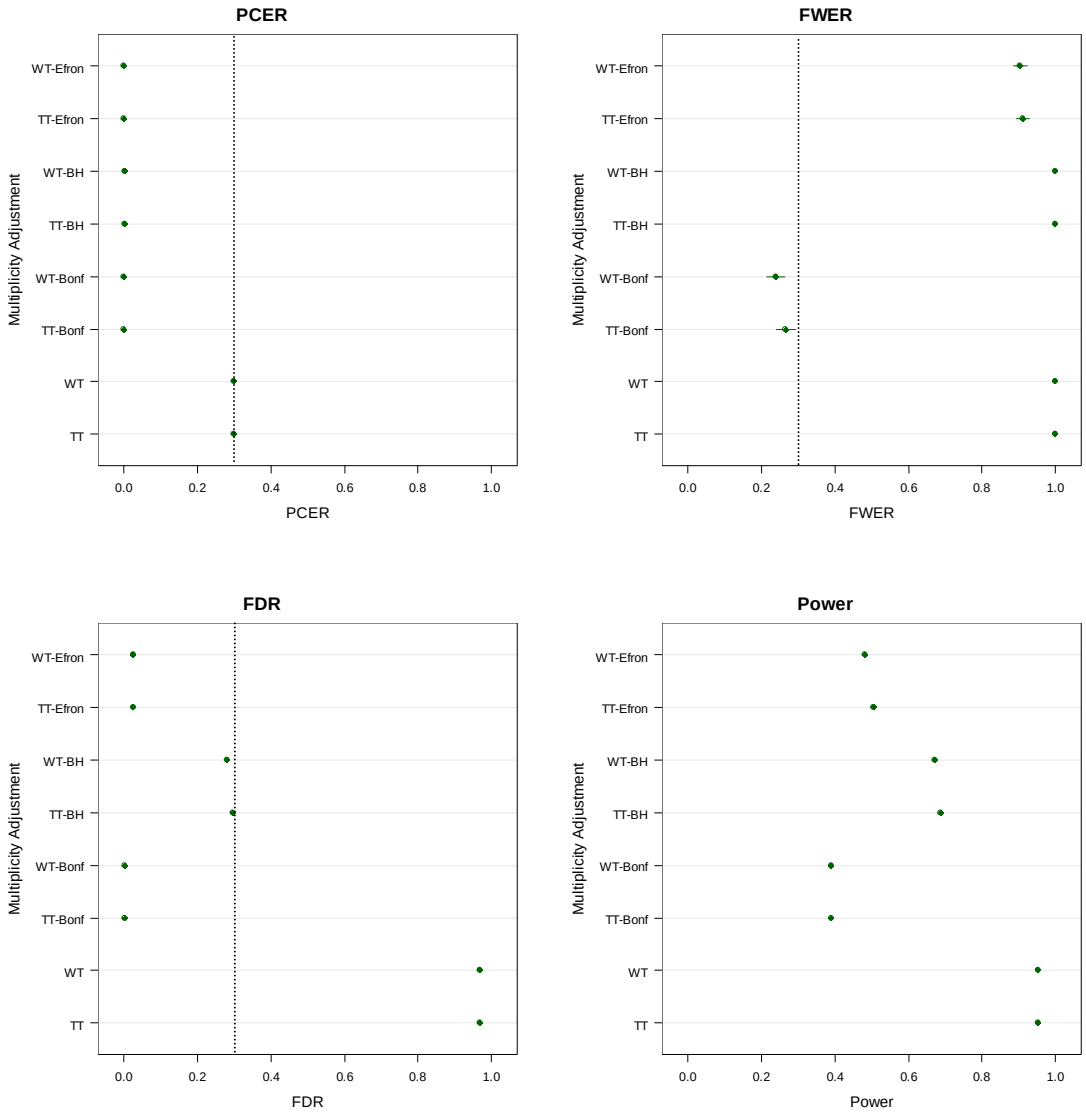

Figure F-94. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 1% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

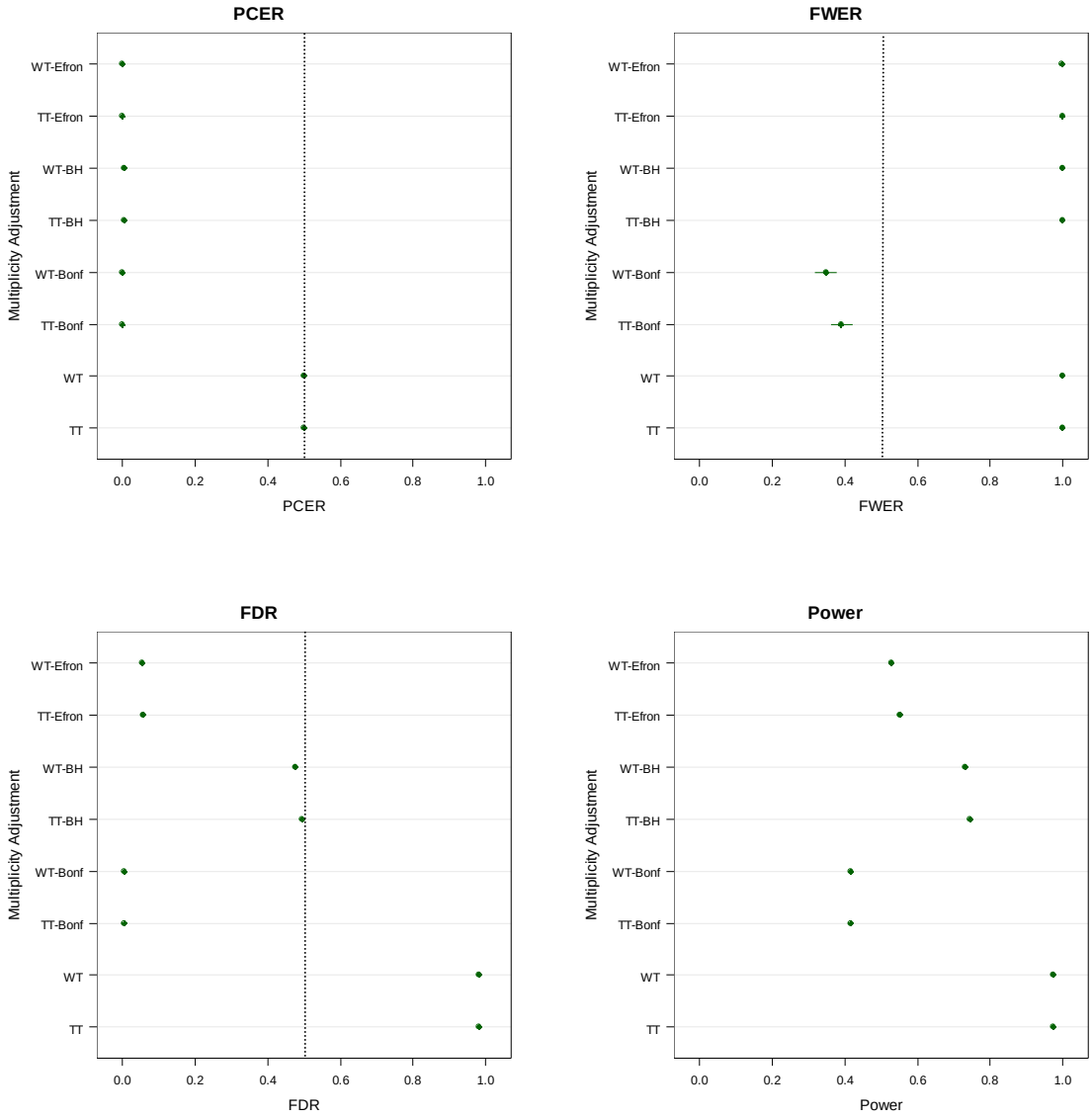

Figure F-95. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 1% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

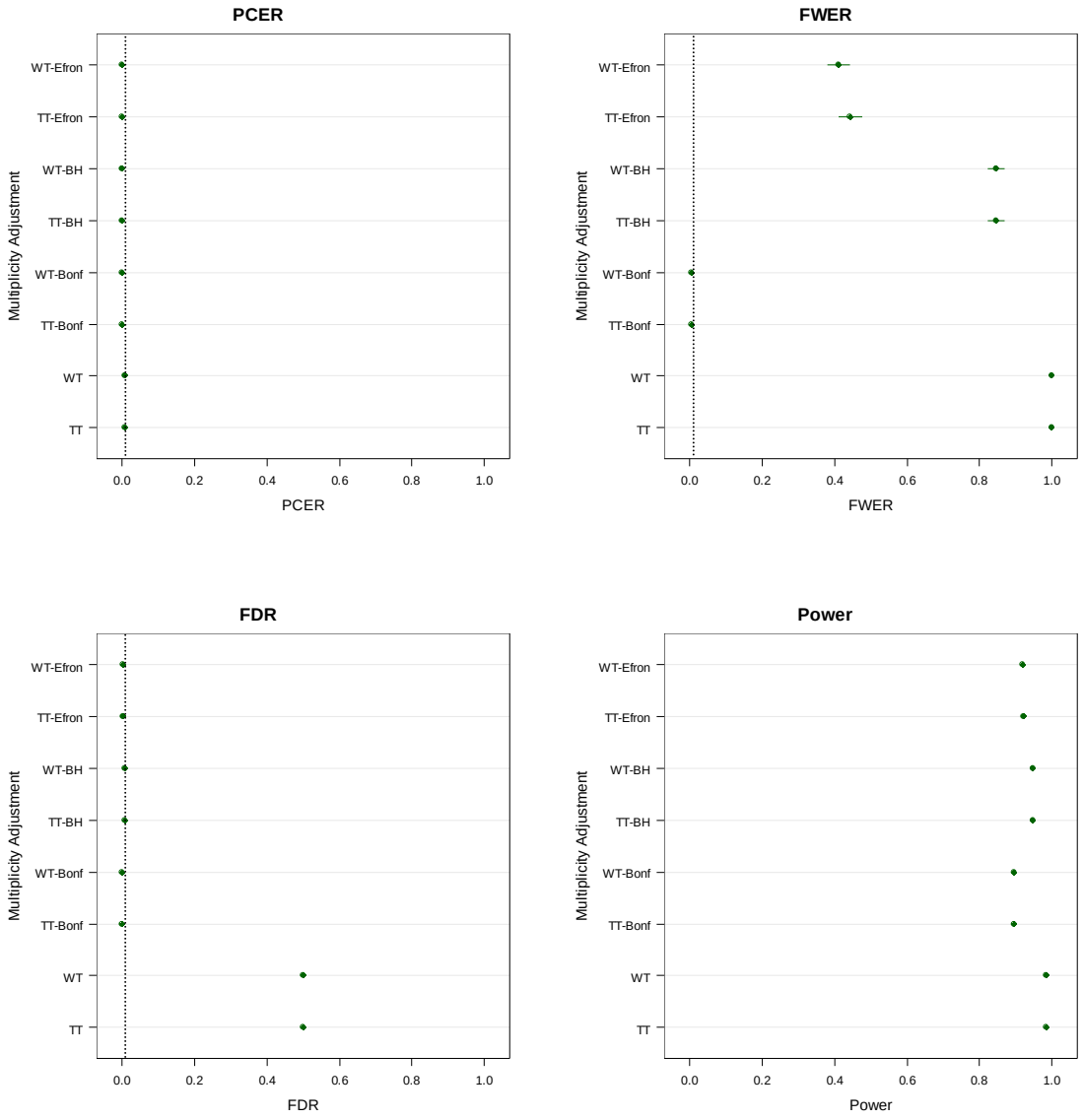

Figure F-96. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 1% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

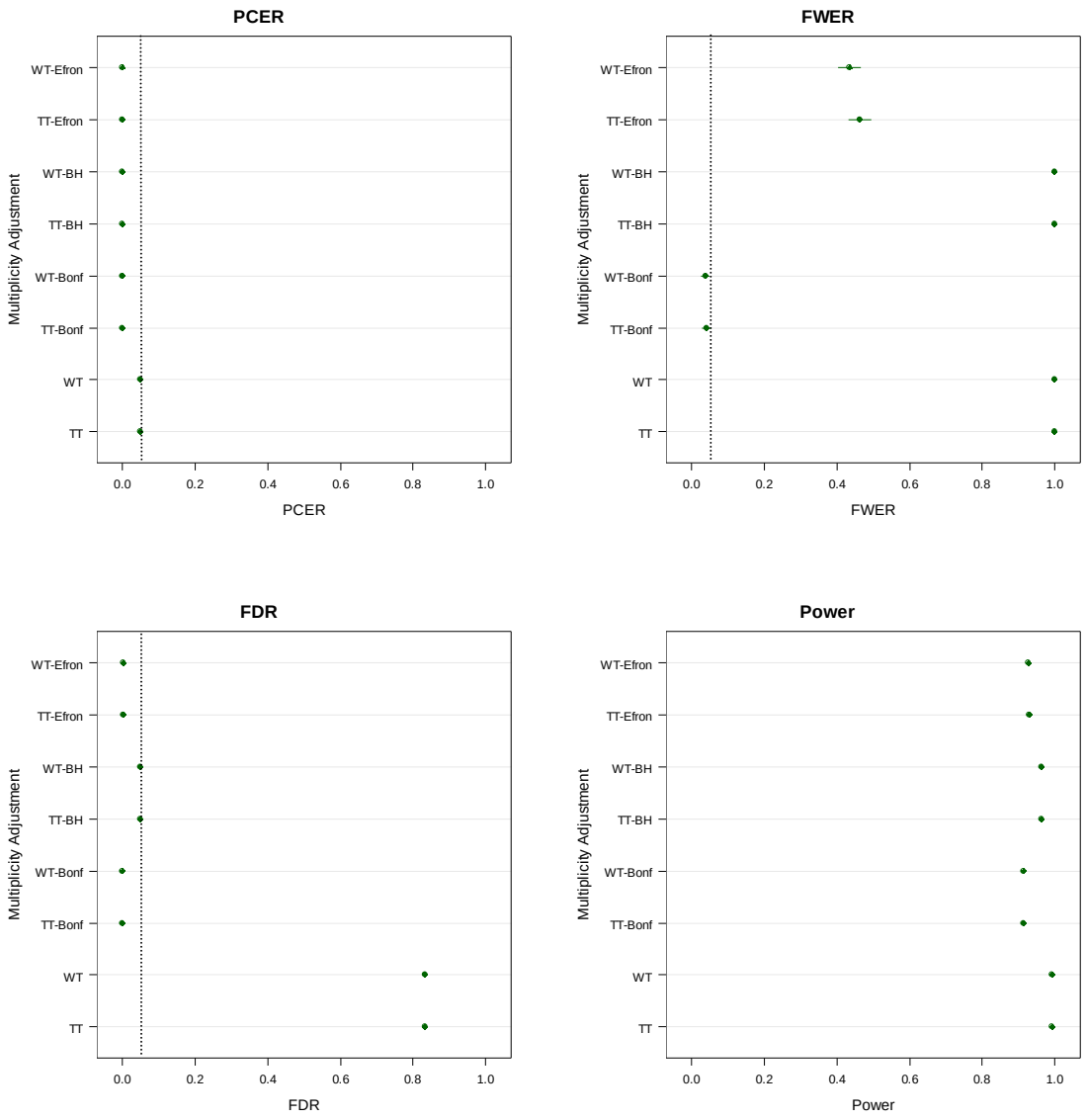

Figure F-97. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 1% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

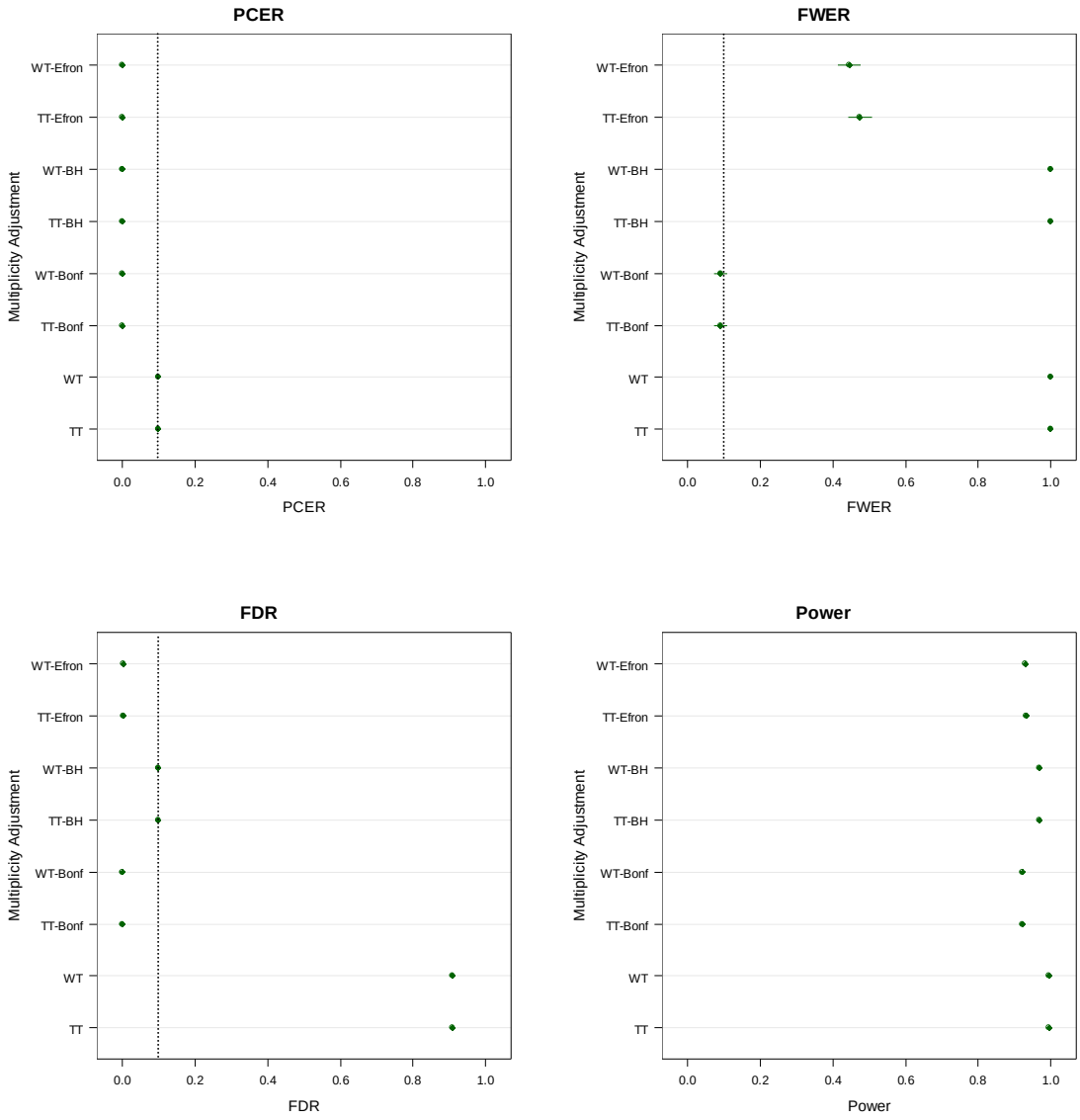

Figure F-98. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 1% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

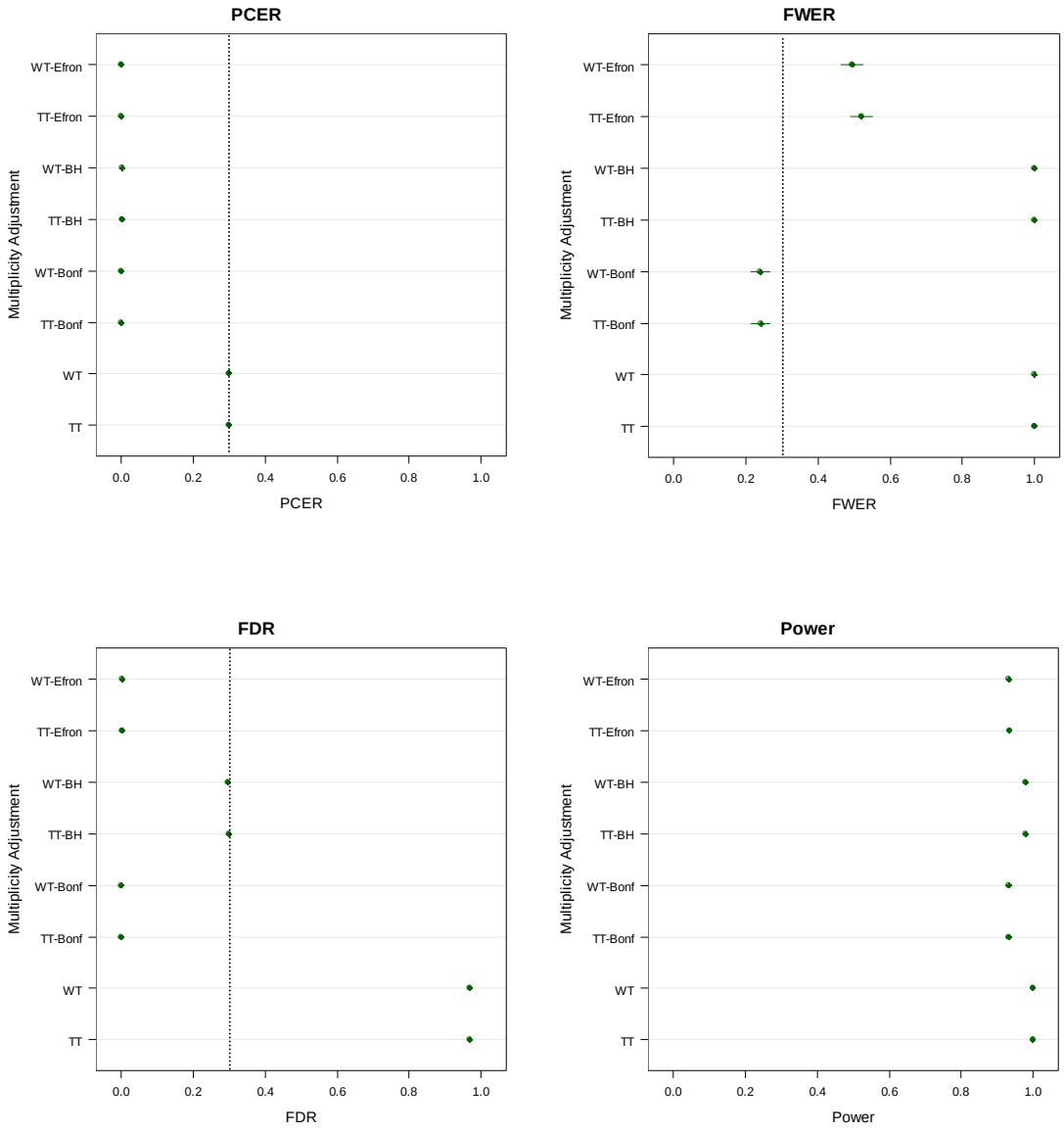

Figure F-99. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 1% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

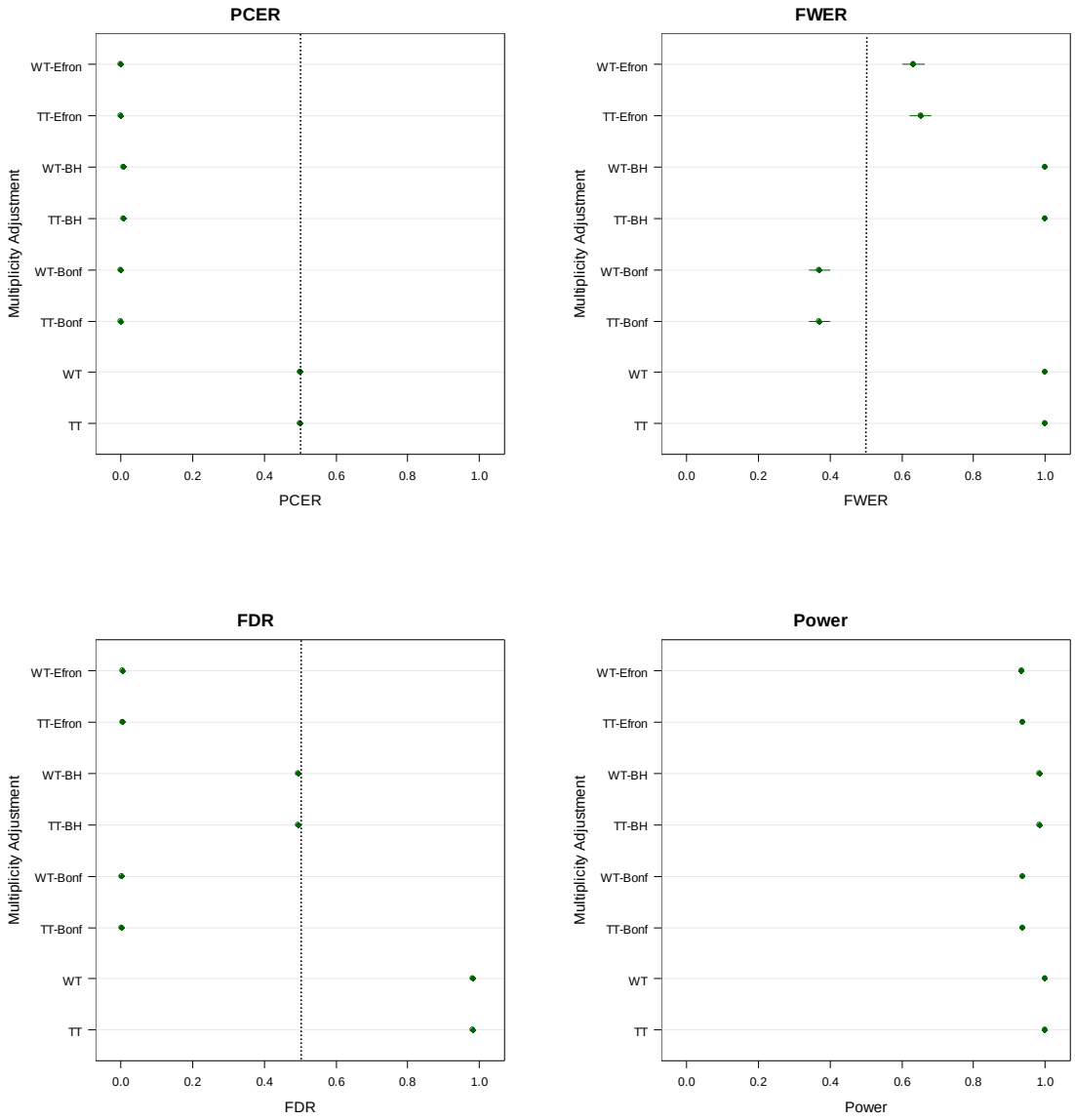

Figure F-100. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 1% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

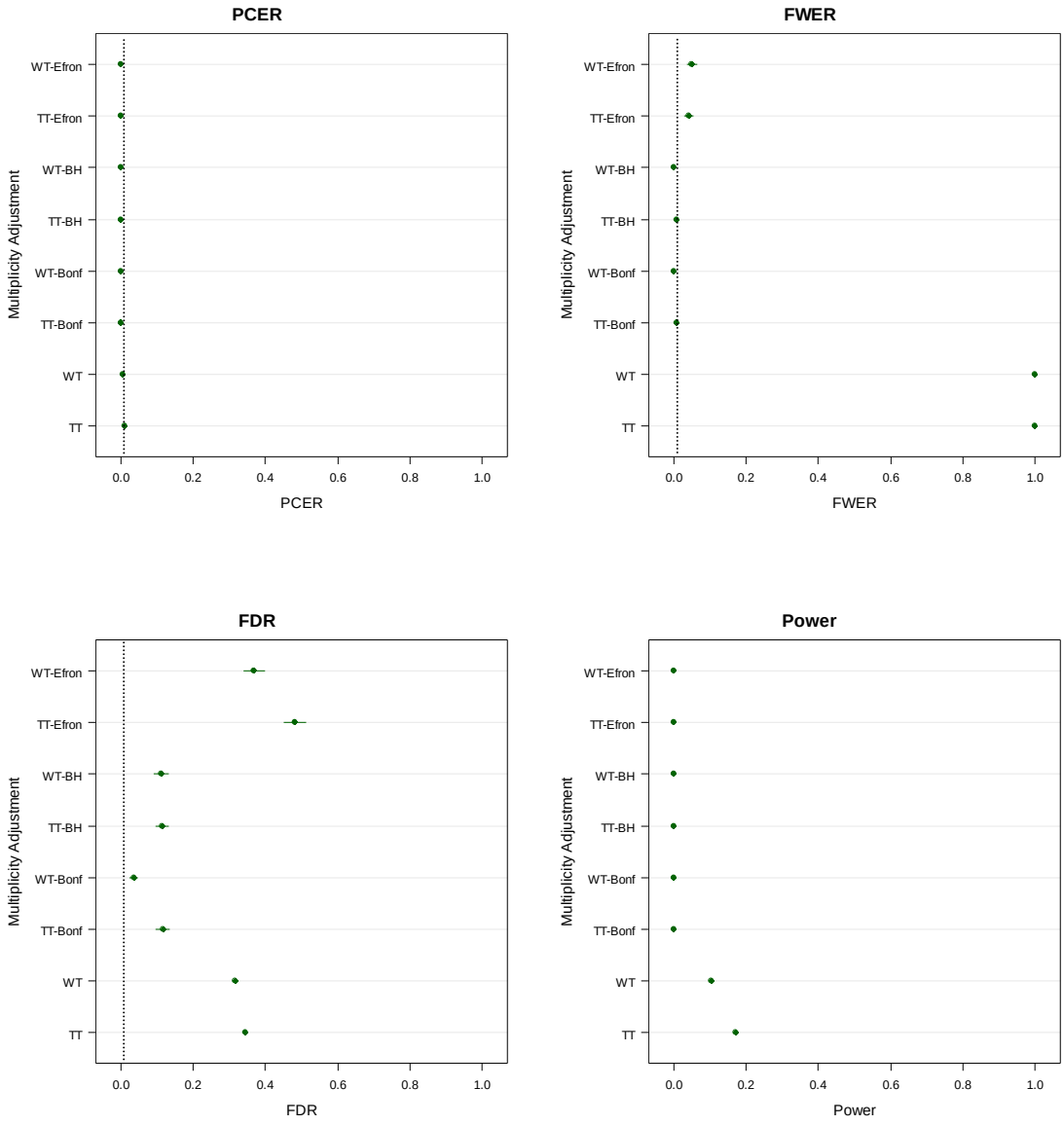

Figure F-101. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 10% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

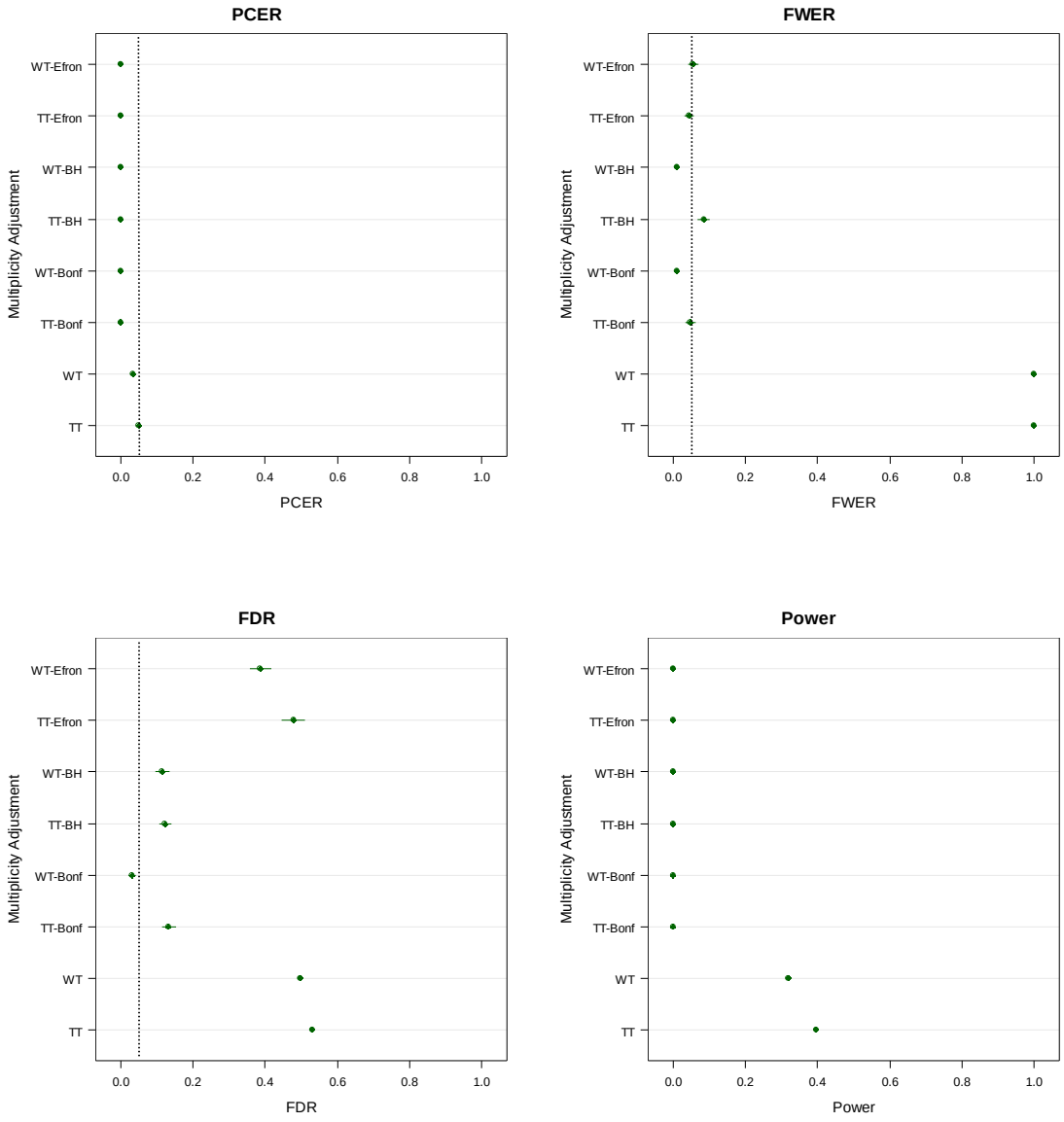

Figure F-102. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 10% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

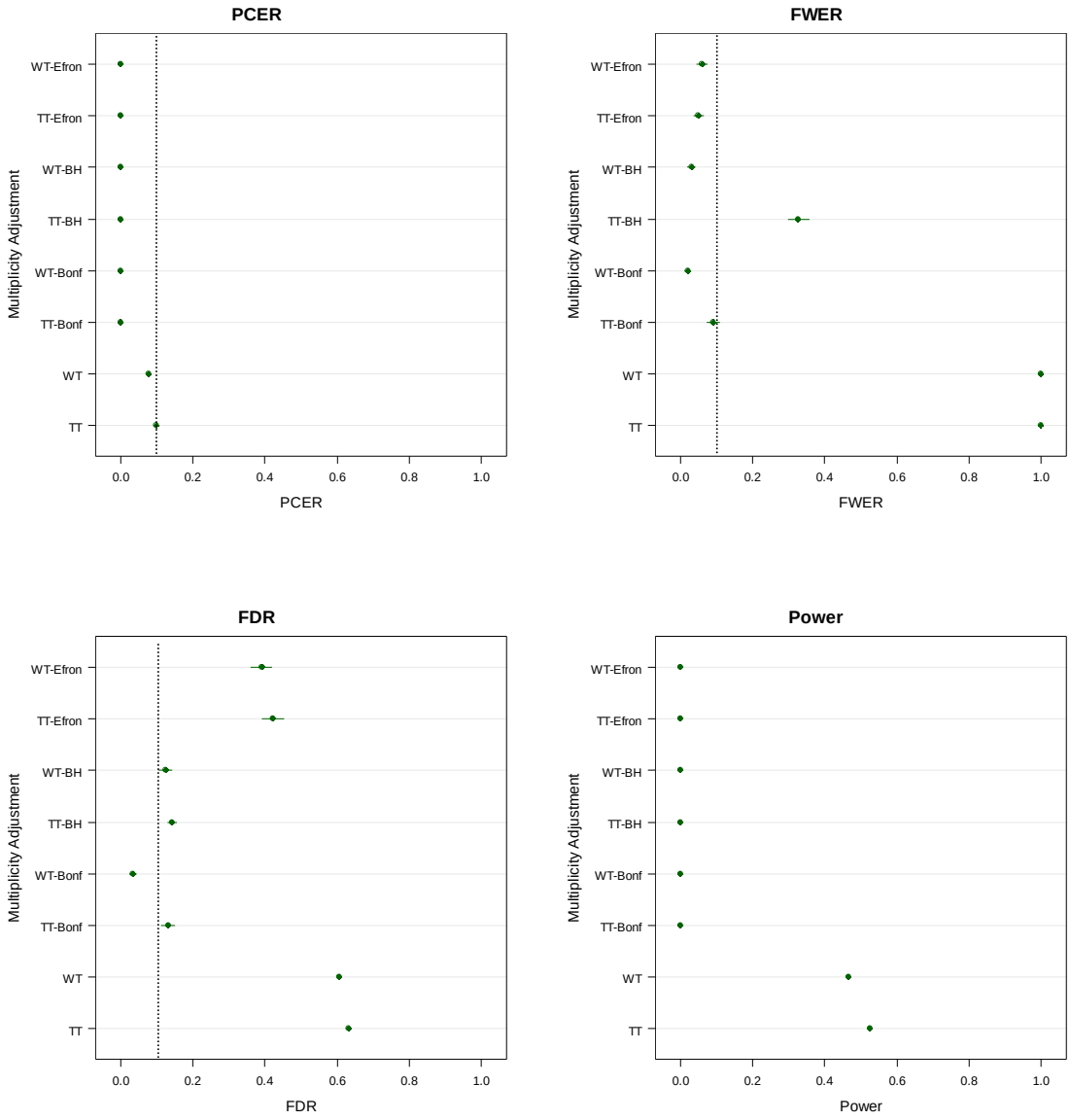

Figure F-103. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 10% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

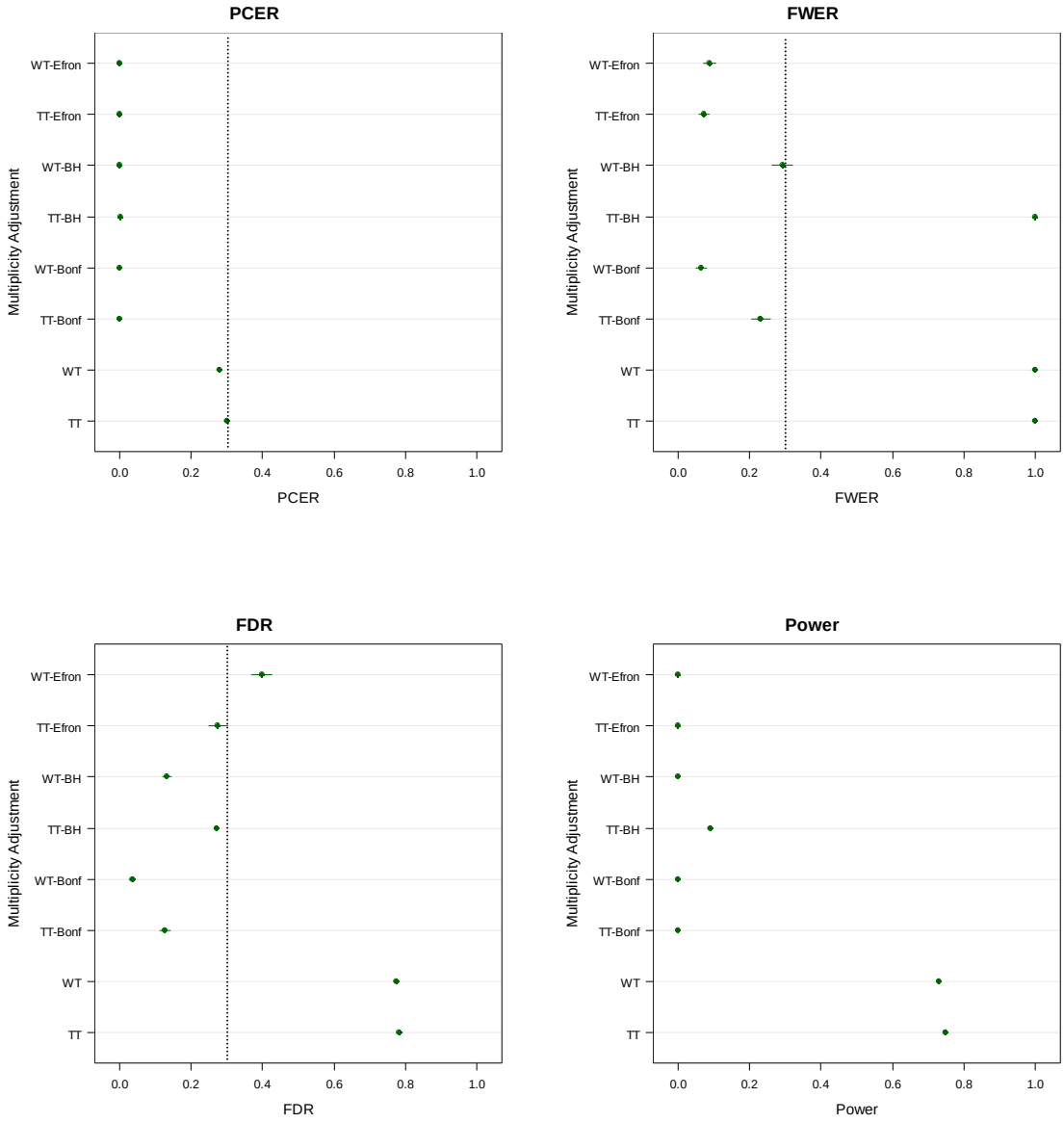

Figure F-104. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 10% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

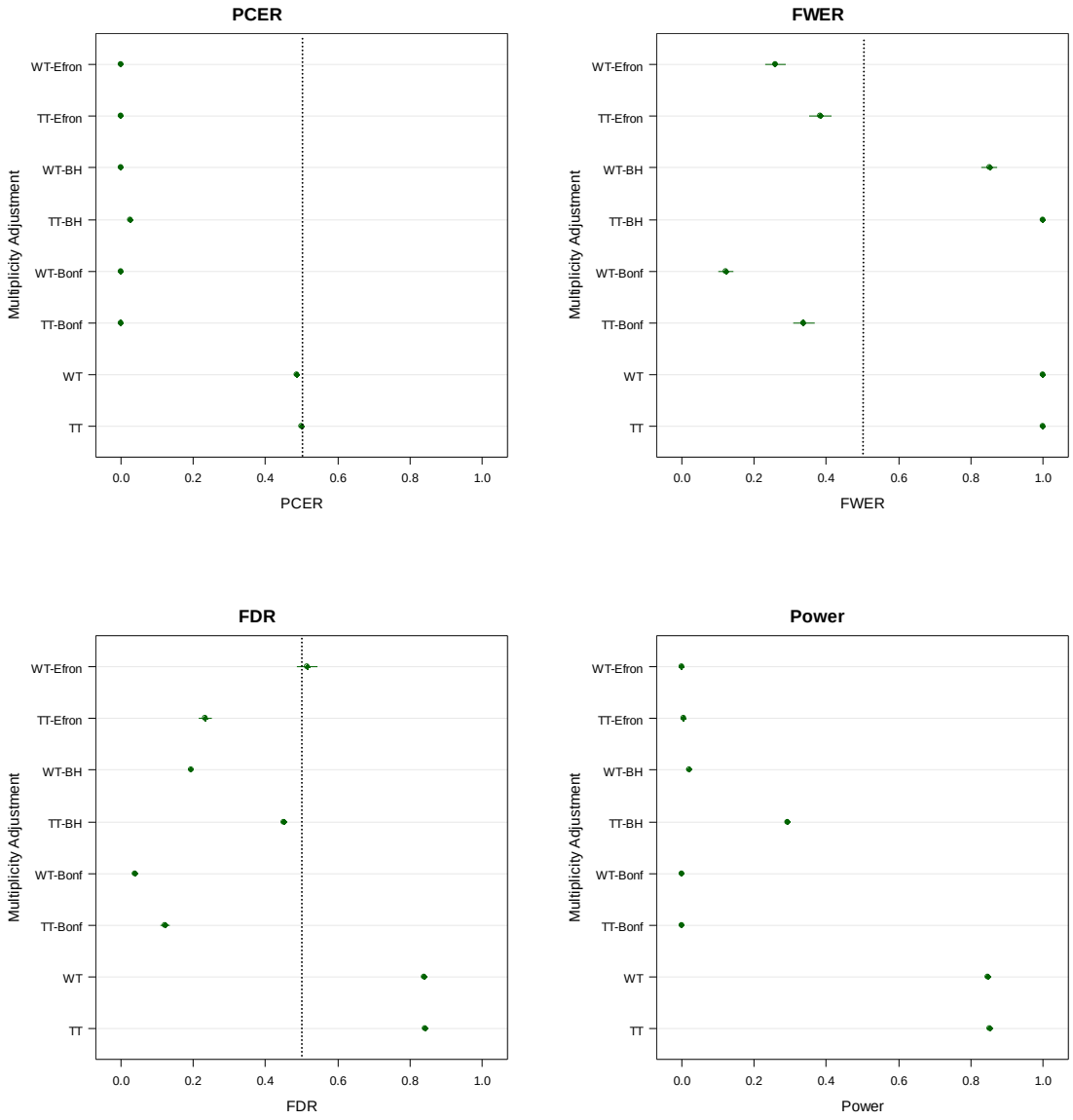

Figure F-105. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 10% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

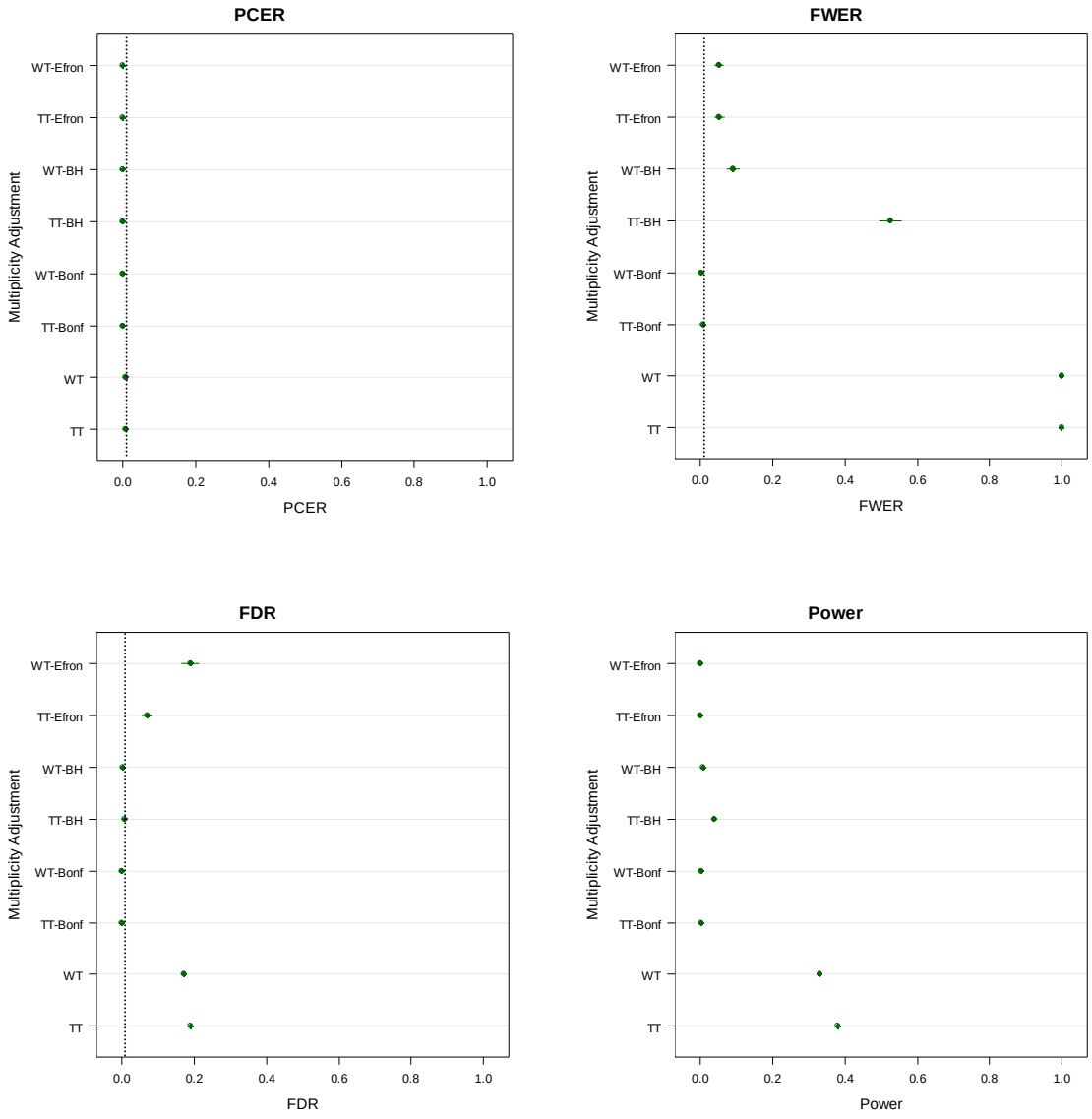

Figure F-106. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

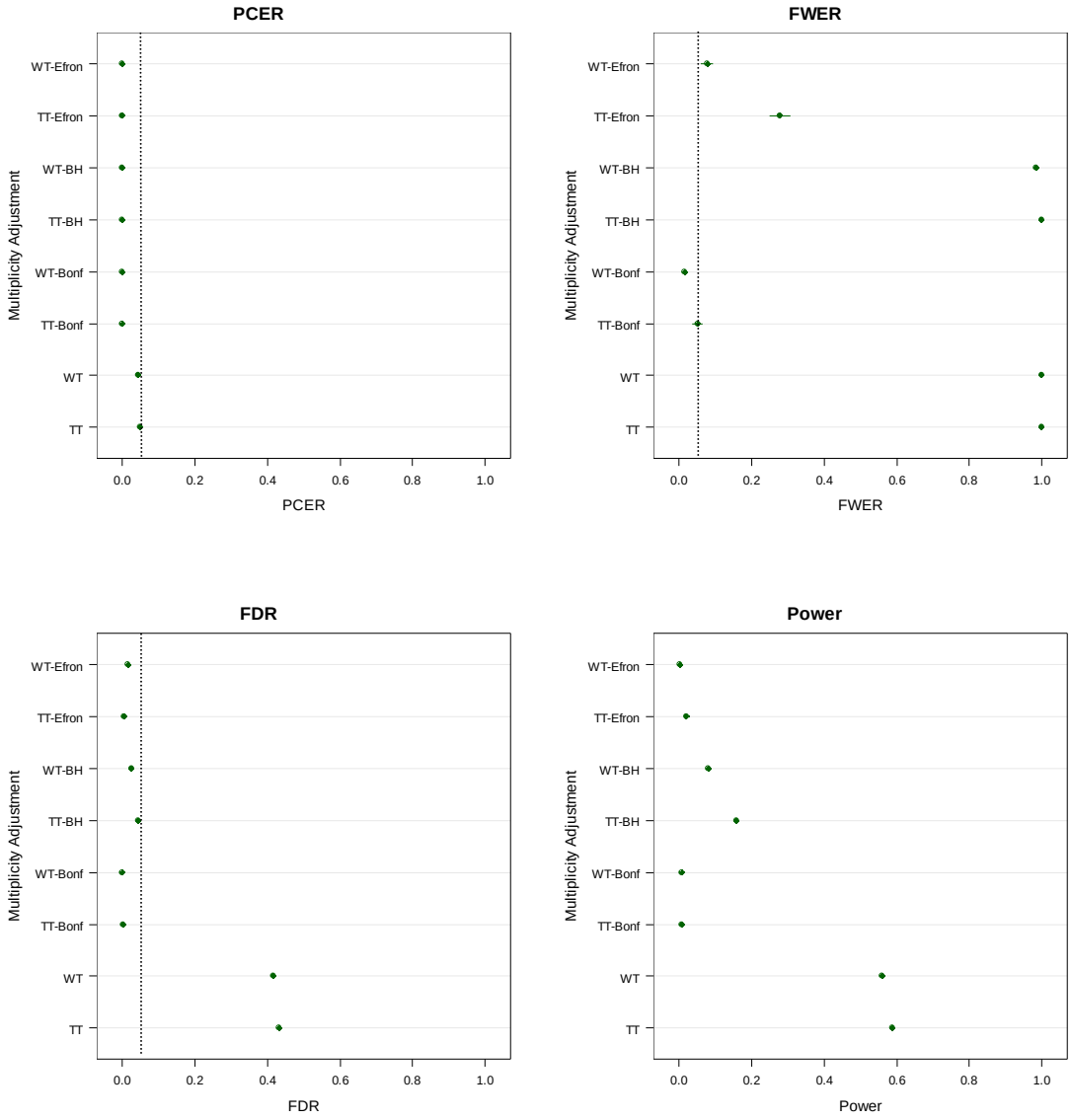

Figure F-107. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

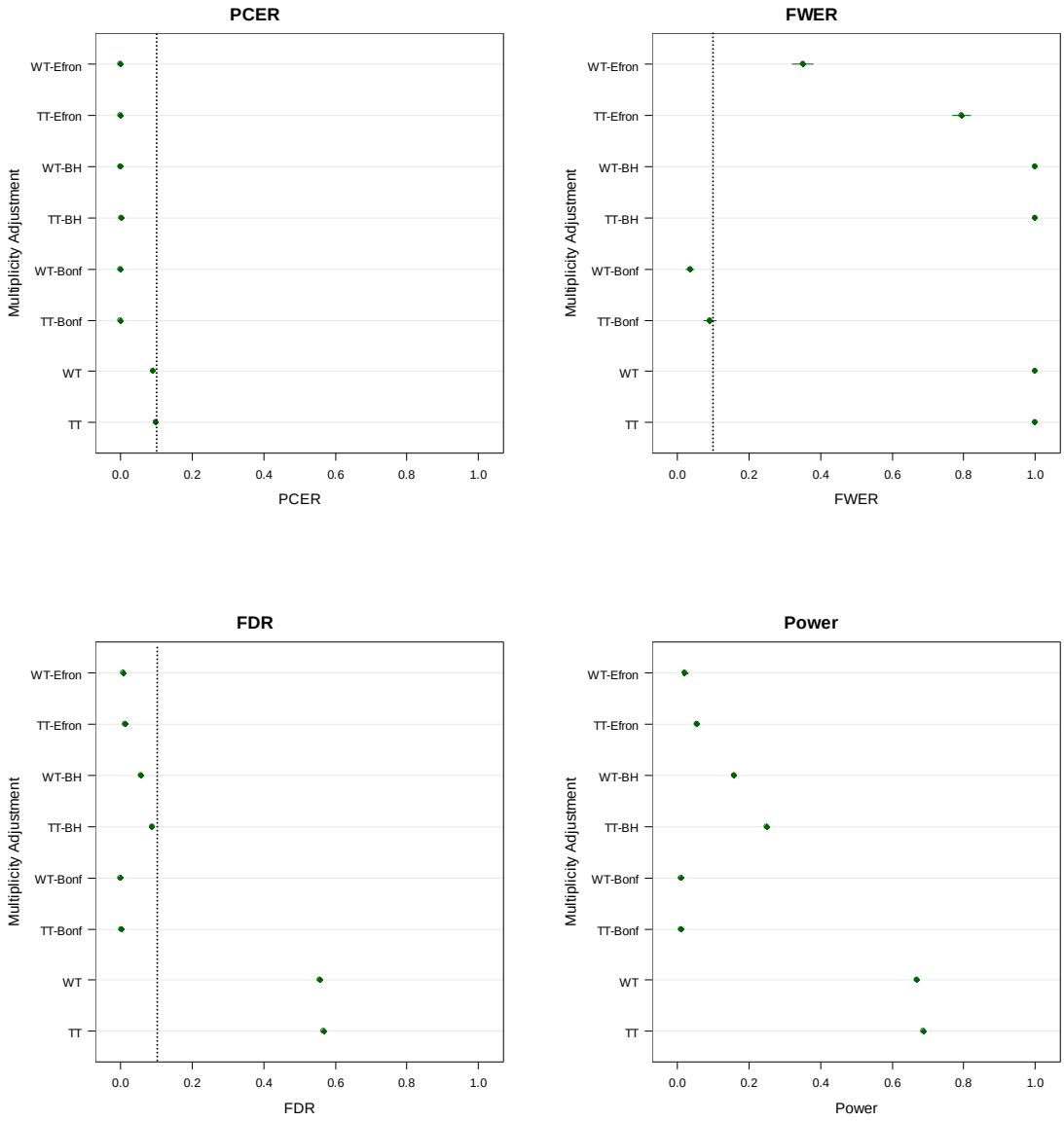

Figure F-108. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

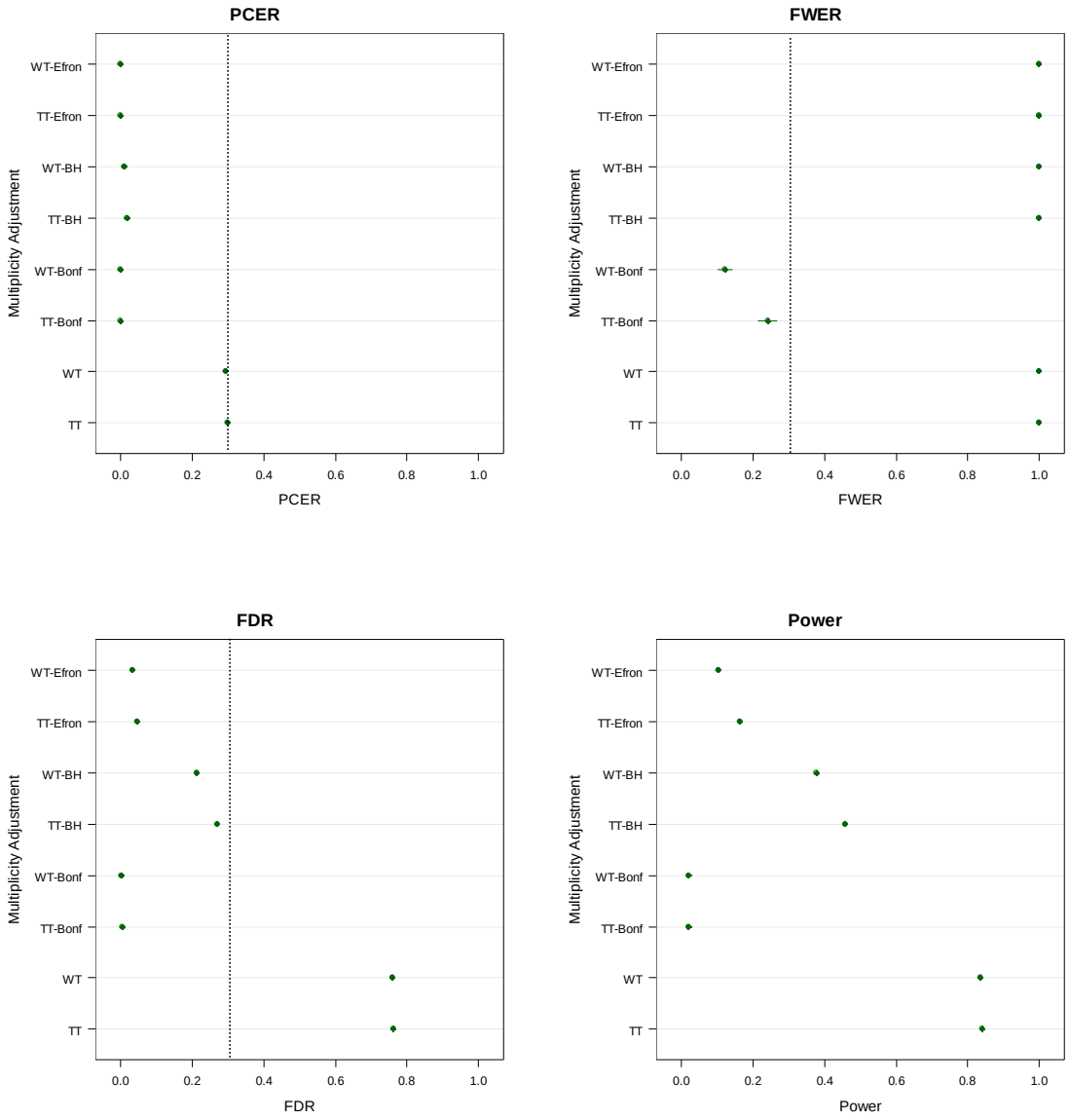

Figure F-109. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

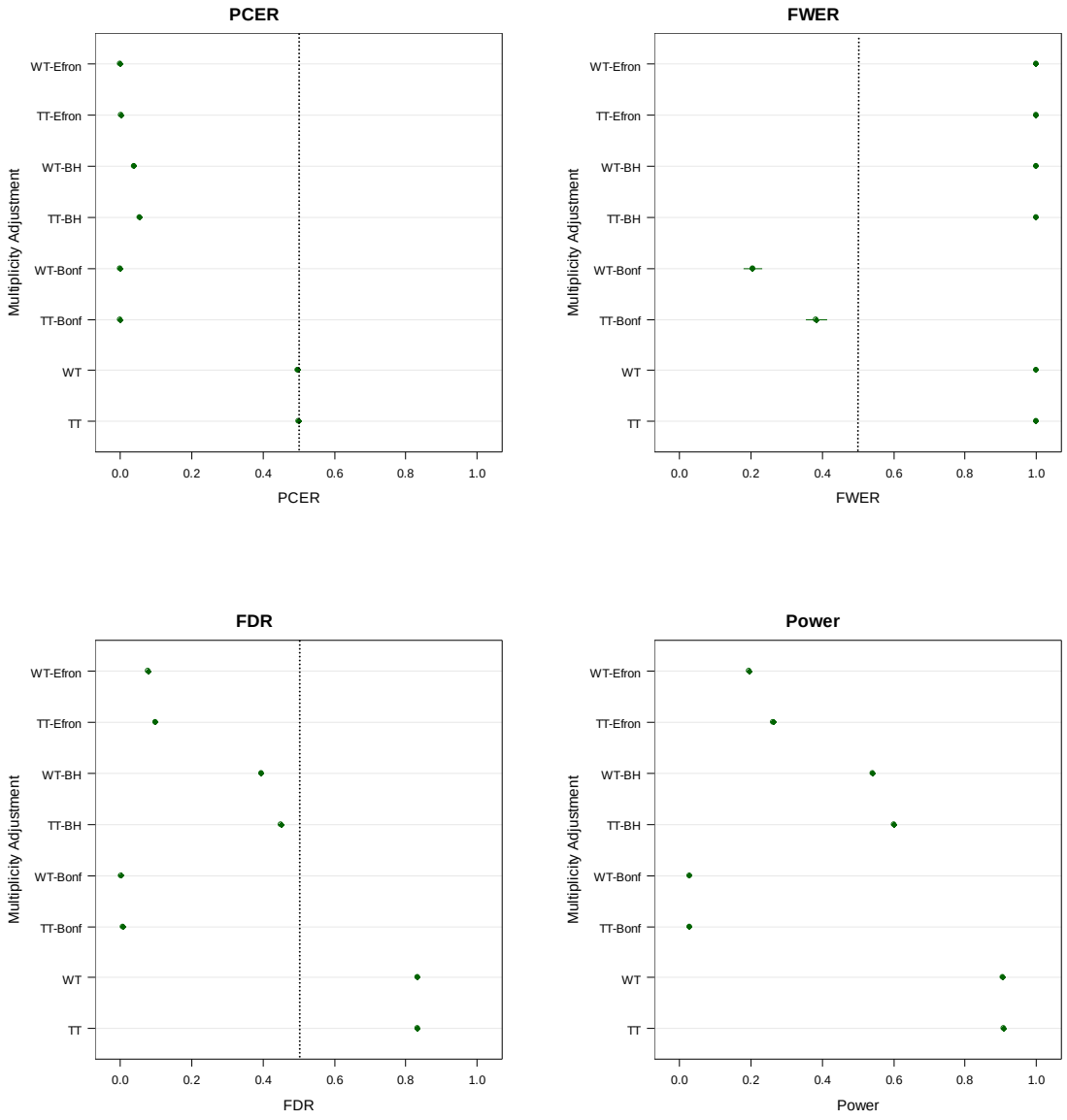

Figure F-110. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

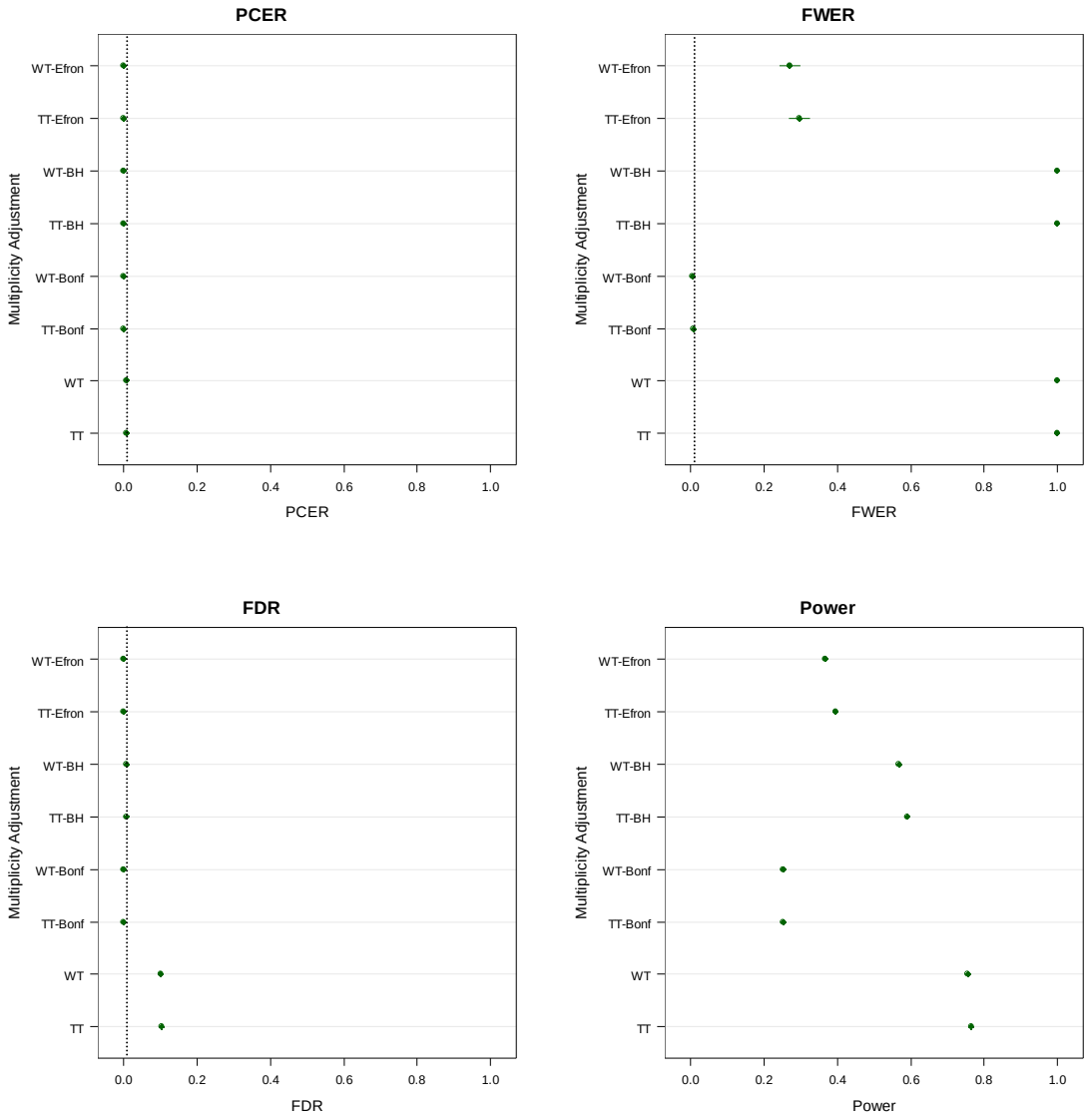

Figure F-111. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

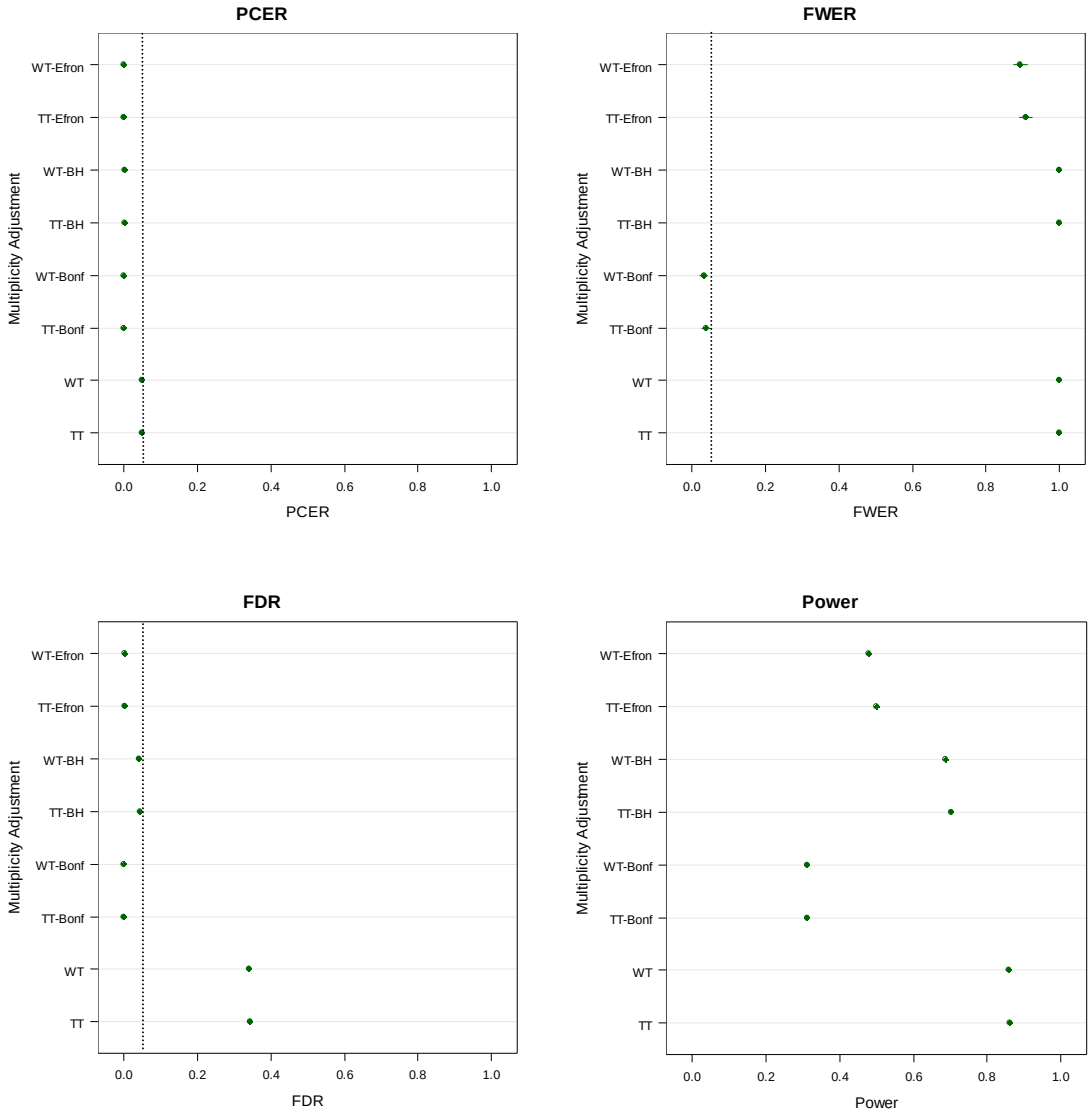

Figure F-112. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

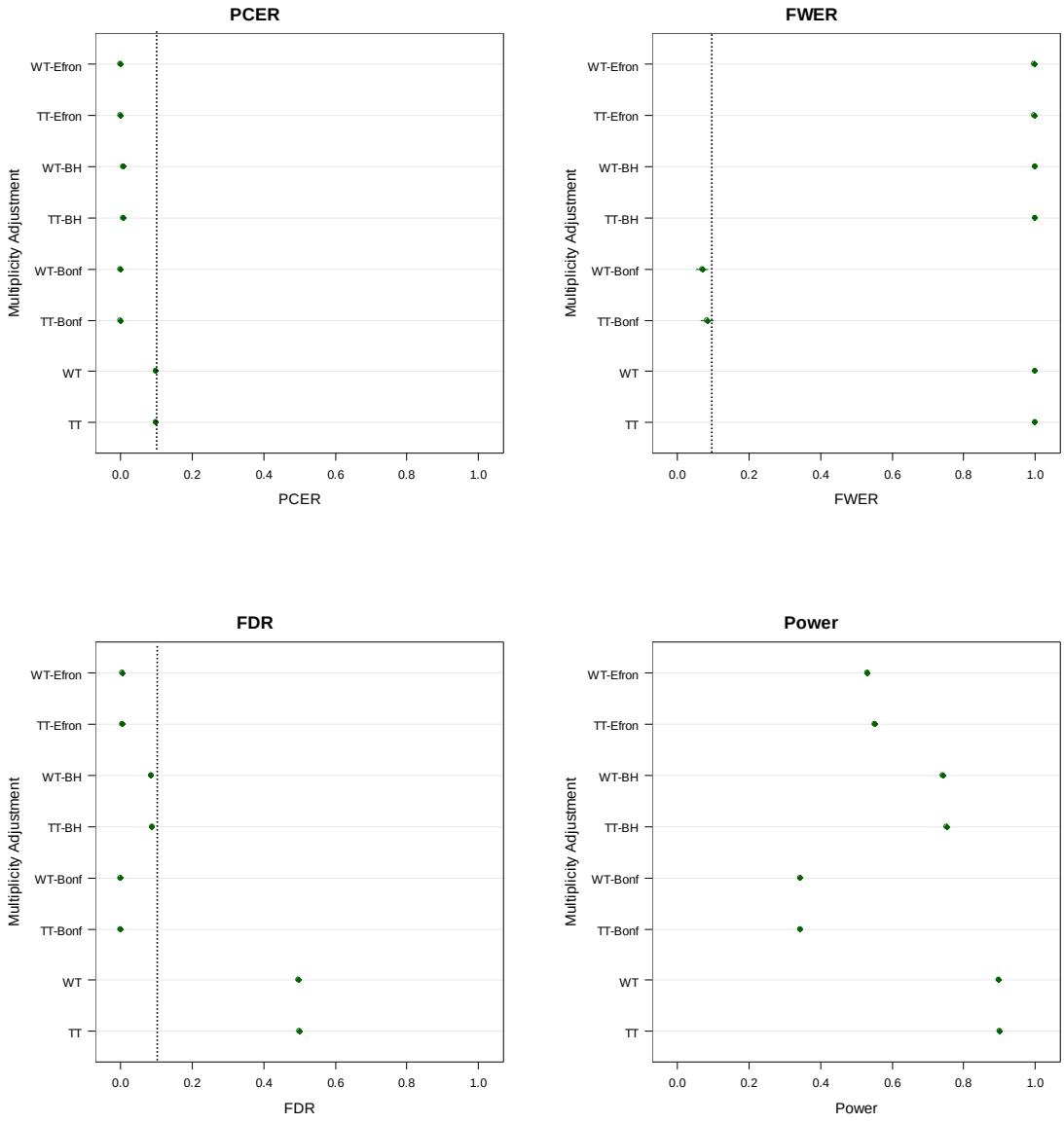

Figure F-113. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

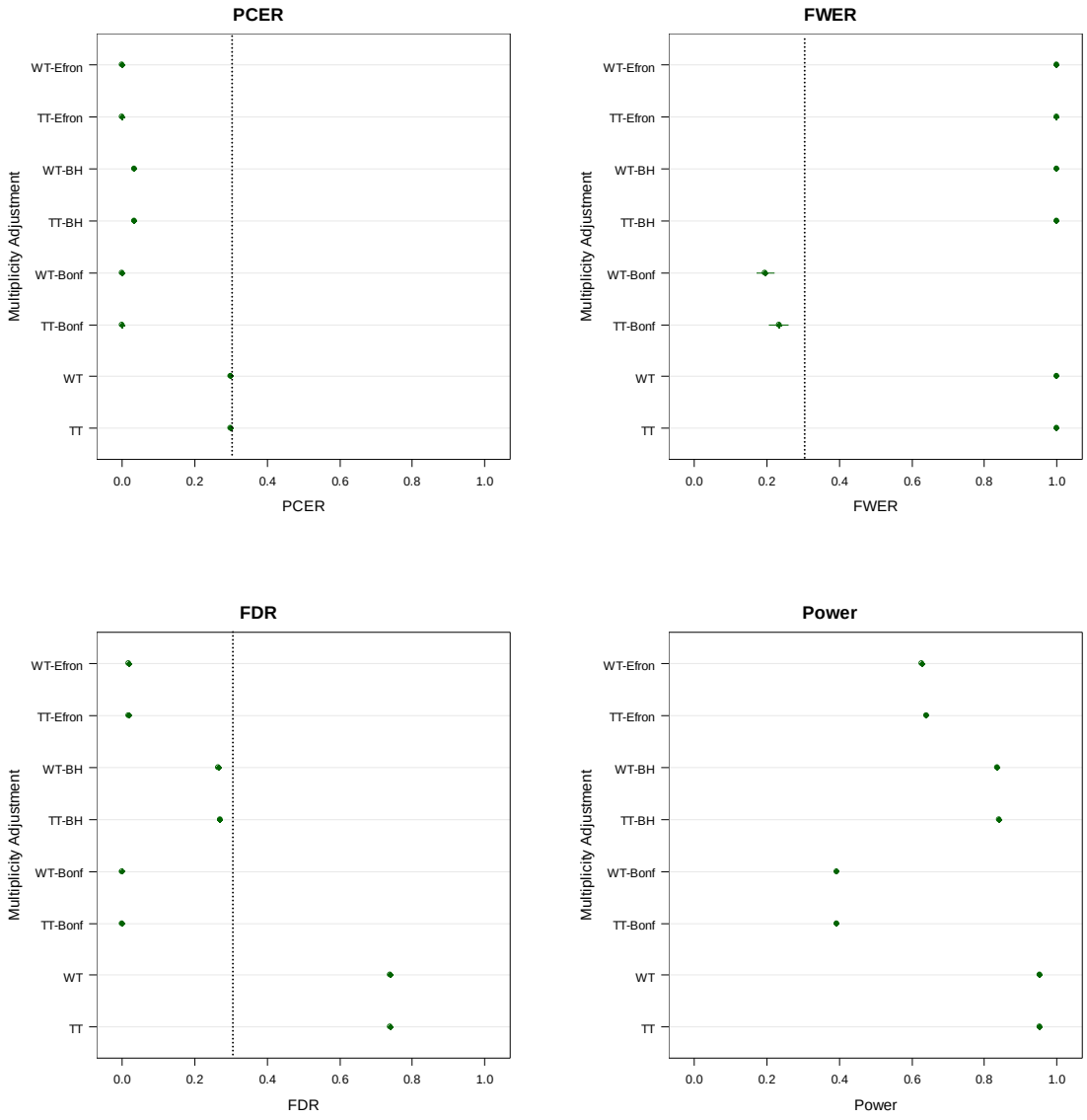

Figure F-114. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

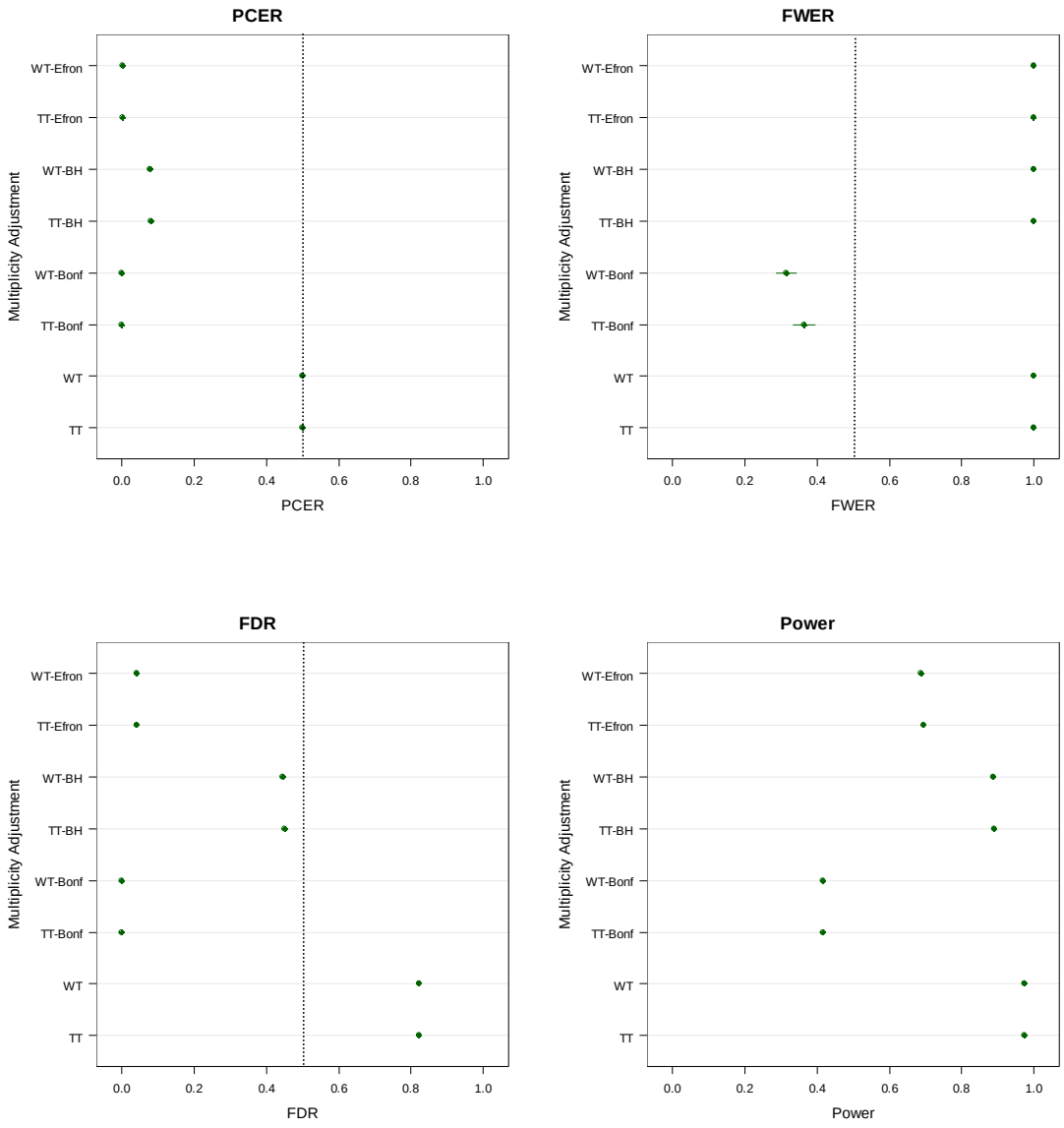

Figure F-115. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

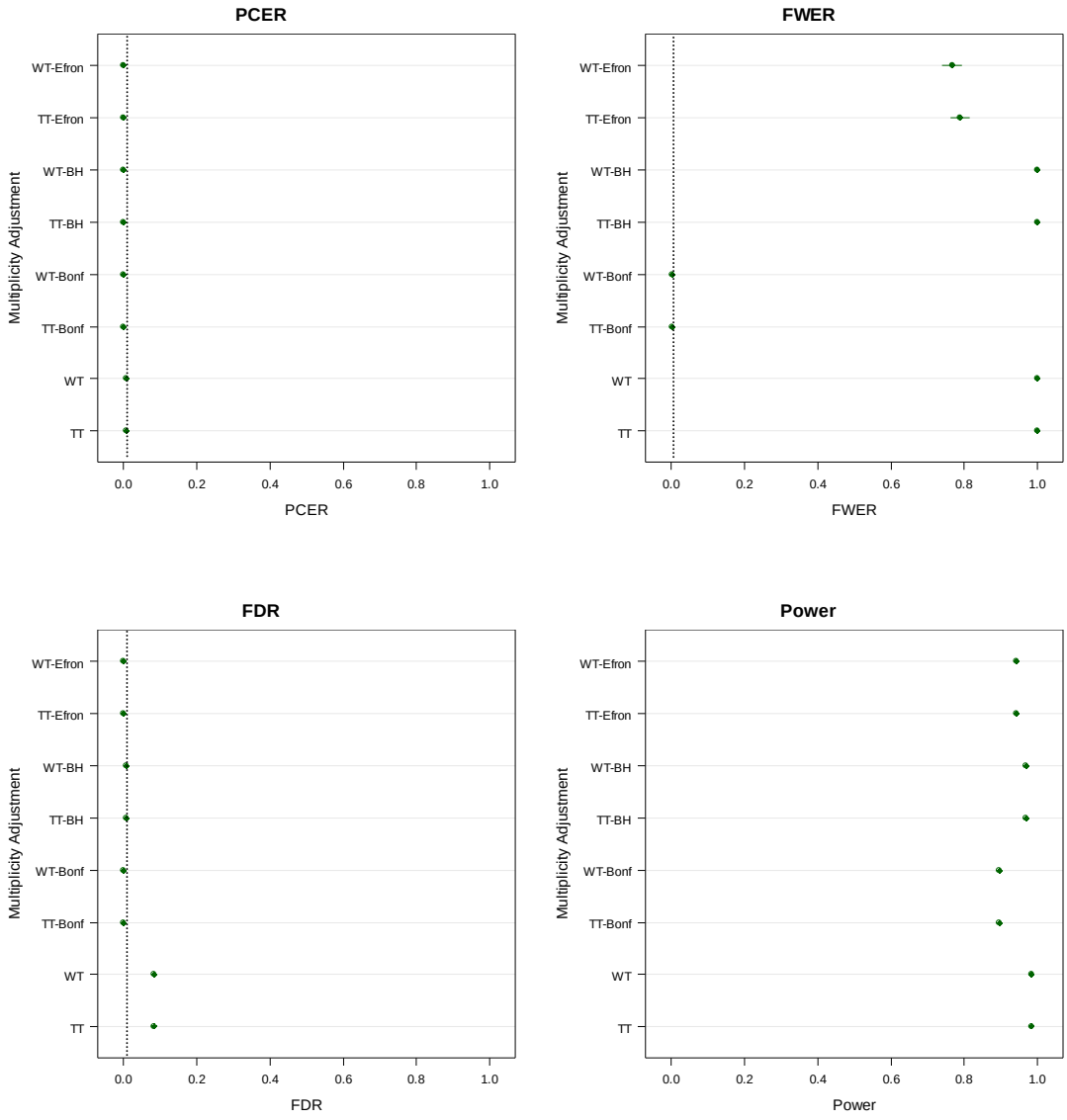

Figure F-116. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 10% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

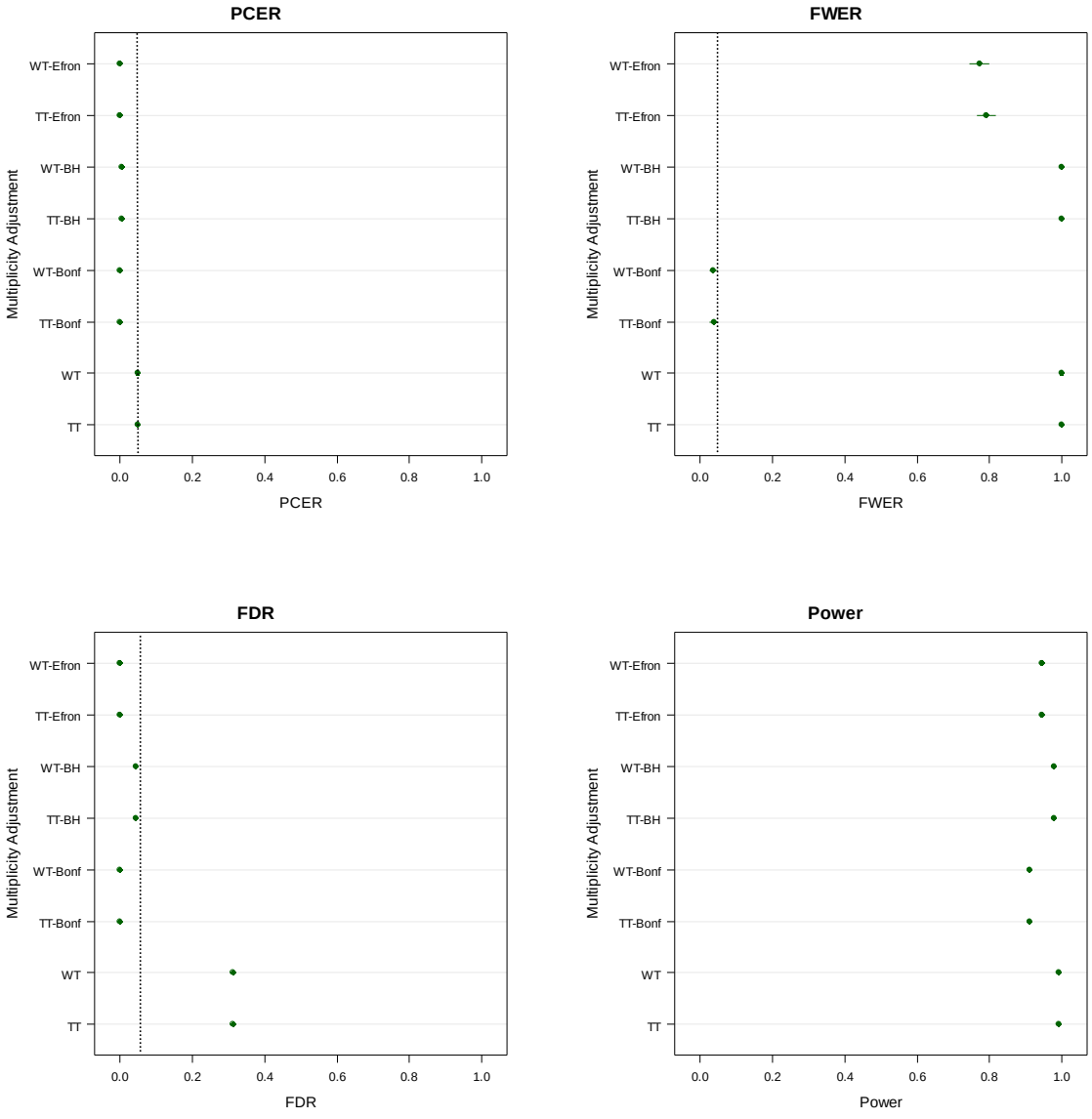

Figure F-117. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 10% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

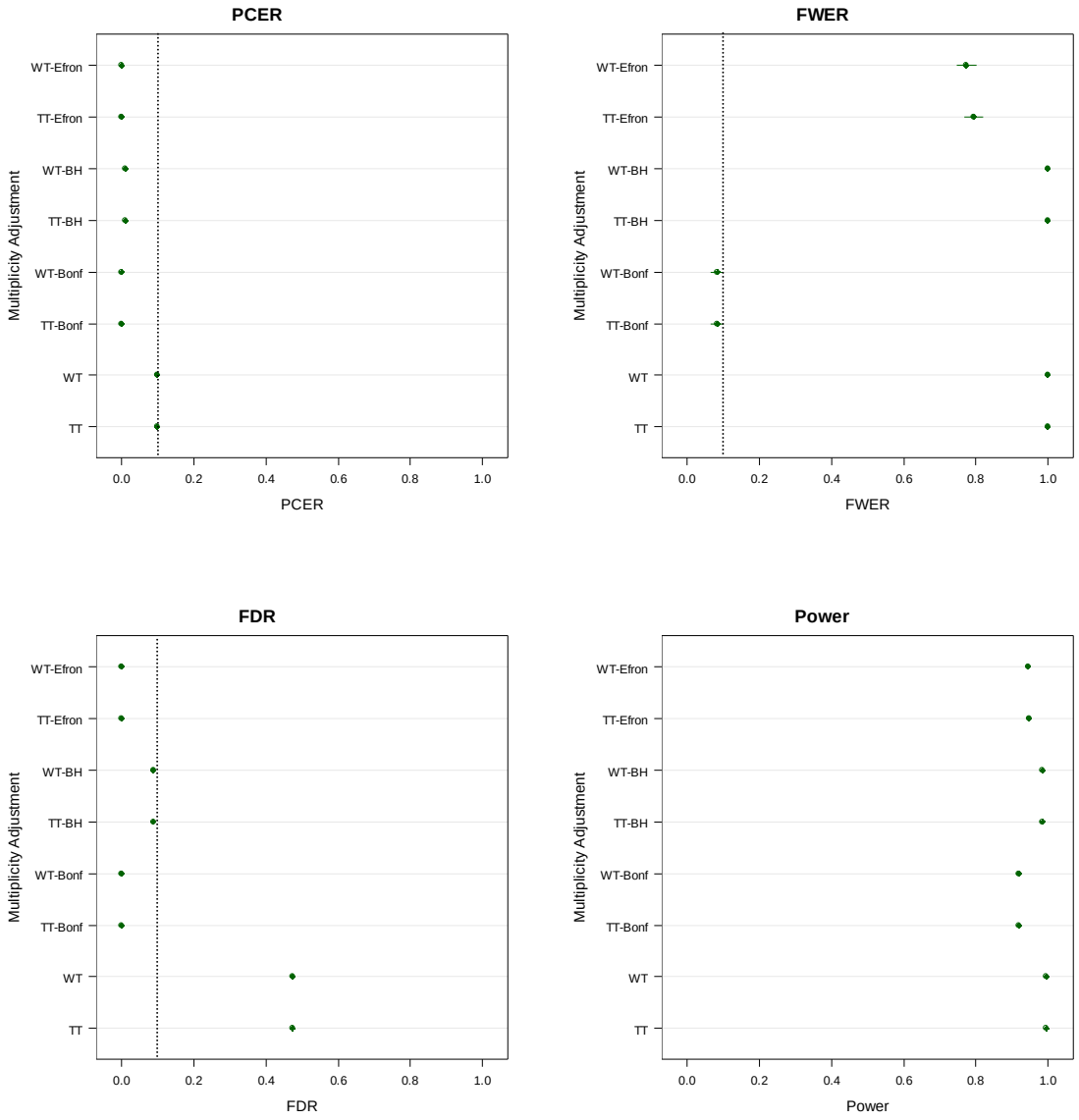

Figure F-118. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 10% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

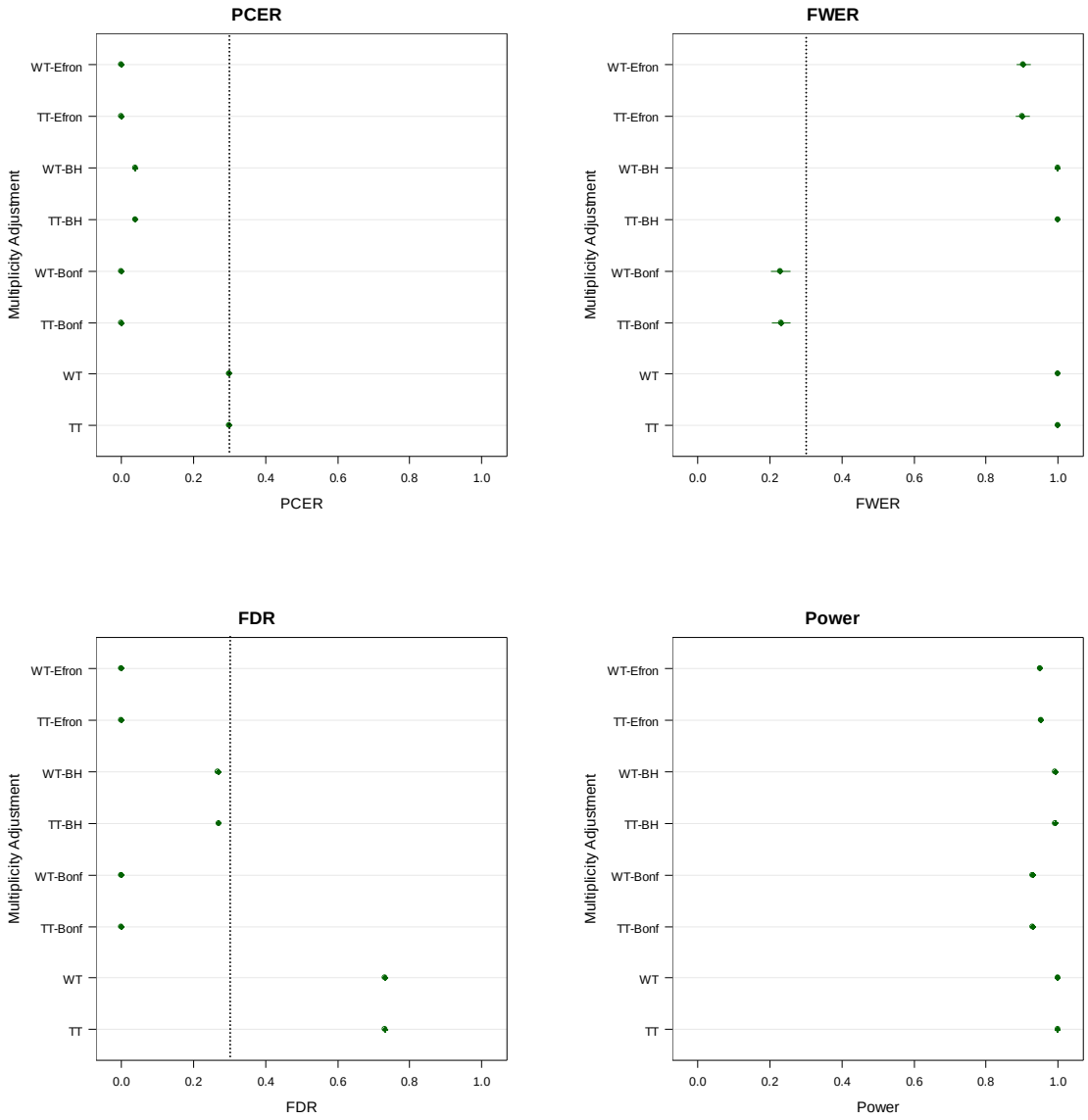

Figure F-119. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 20,000 simulated genes, 10% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

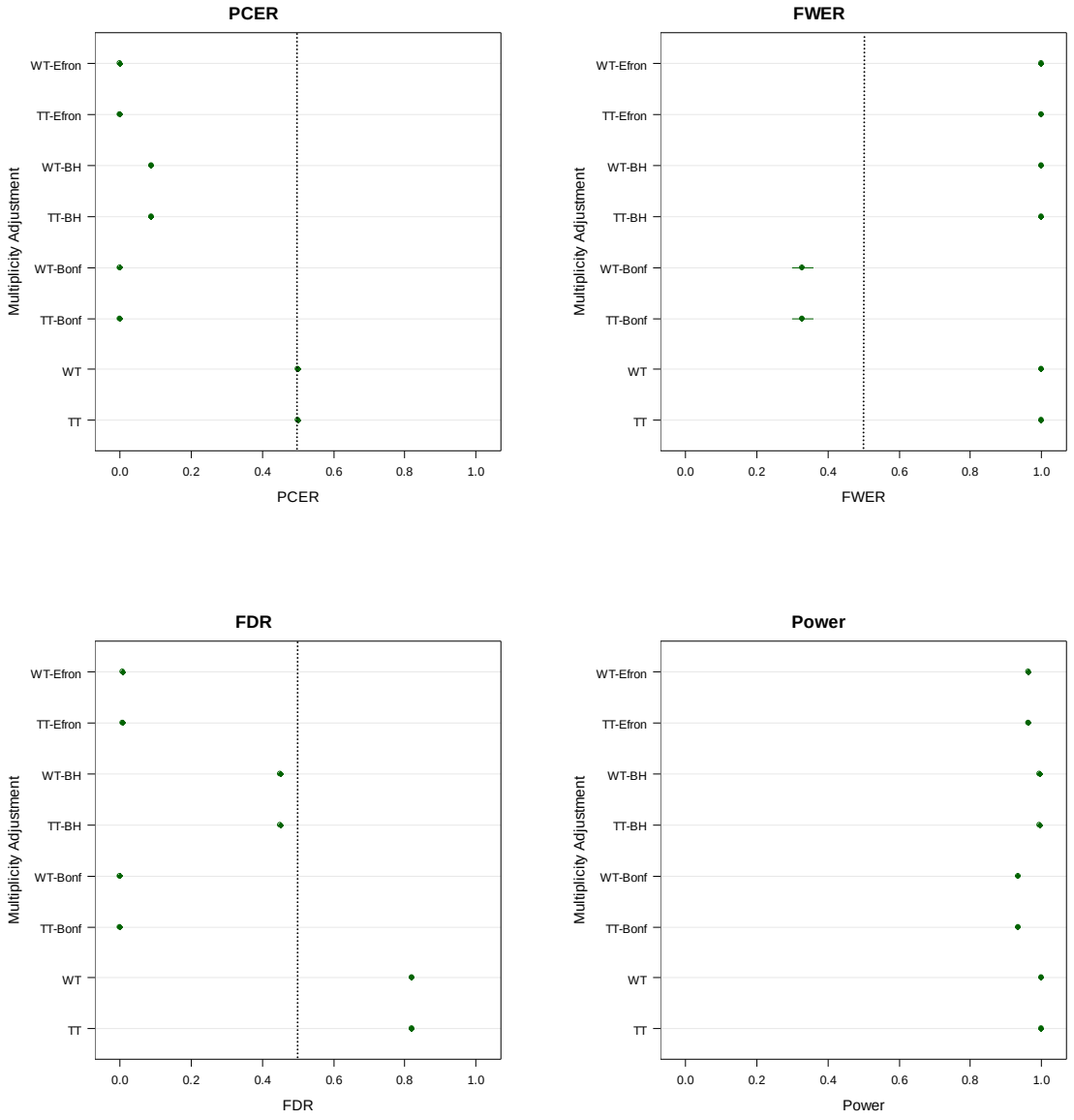

Figure F-120. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes, 1% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

APPENDIX G

SIMULATION RESULTS - CORRELATED

Table G-1. Summary of Figures G-1 to G-30 of Appendix G

Number of Genes	Percent Different	Sample Size (per group)	Figures
200	10%	5	G-1 to G-5*
		15	G-6 to G-10*
2,000	10%	3	G-11 to G-15
		5	G-16 to G-20
		15	G-21 to G-25
		100	G-26 to G-30

*Includes results for the maxT procedure.

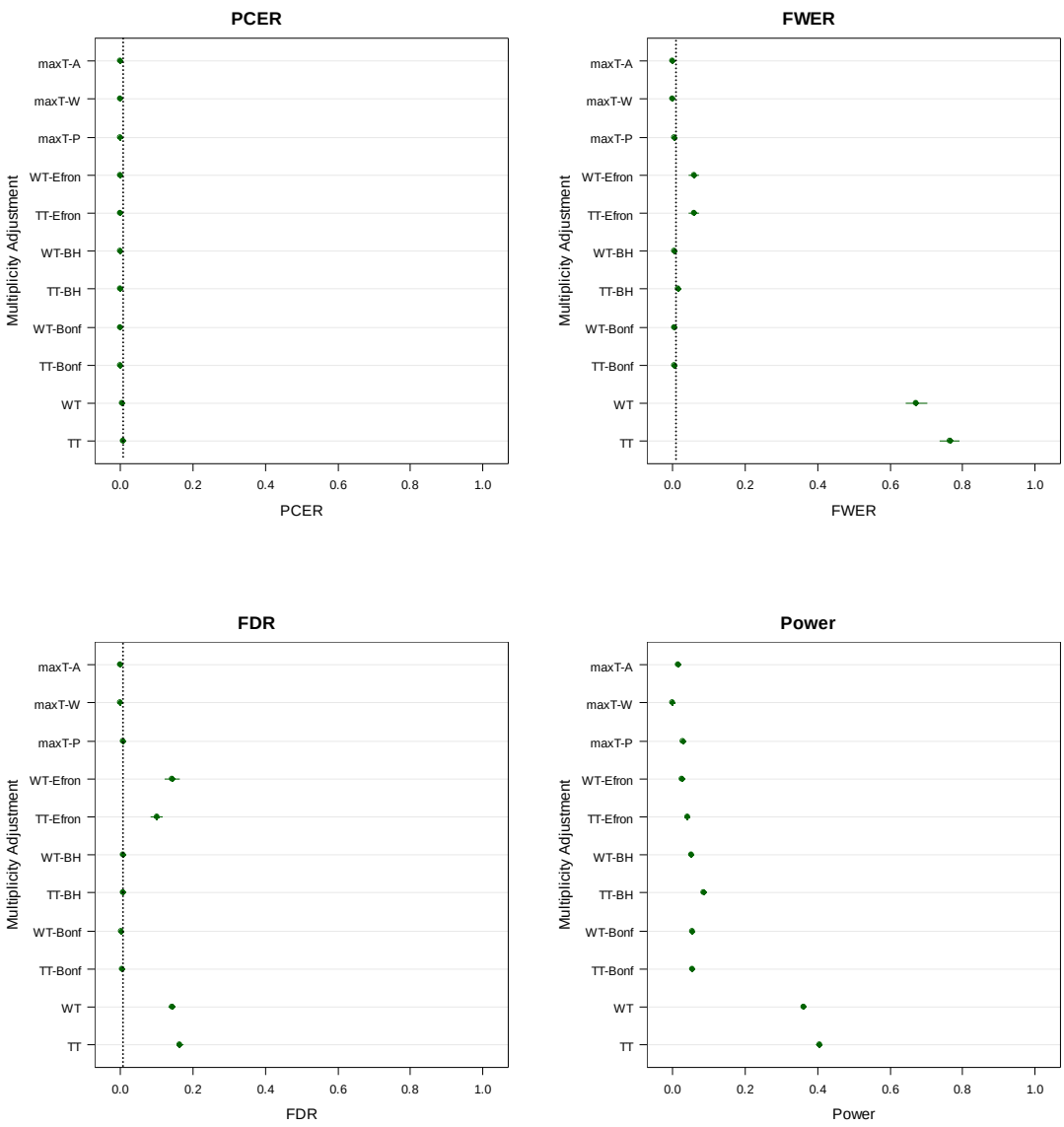

Figure G-1. Simulation Results (0.01); Error and Power Summary Comparing Eleven Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

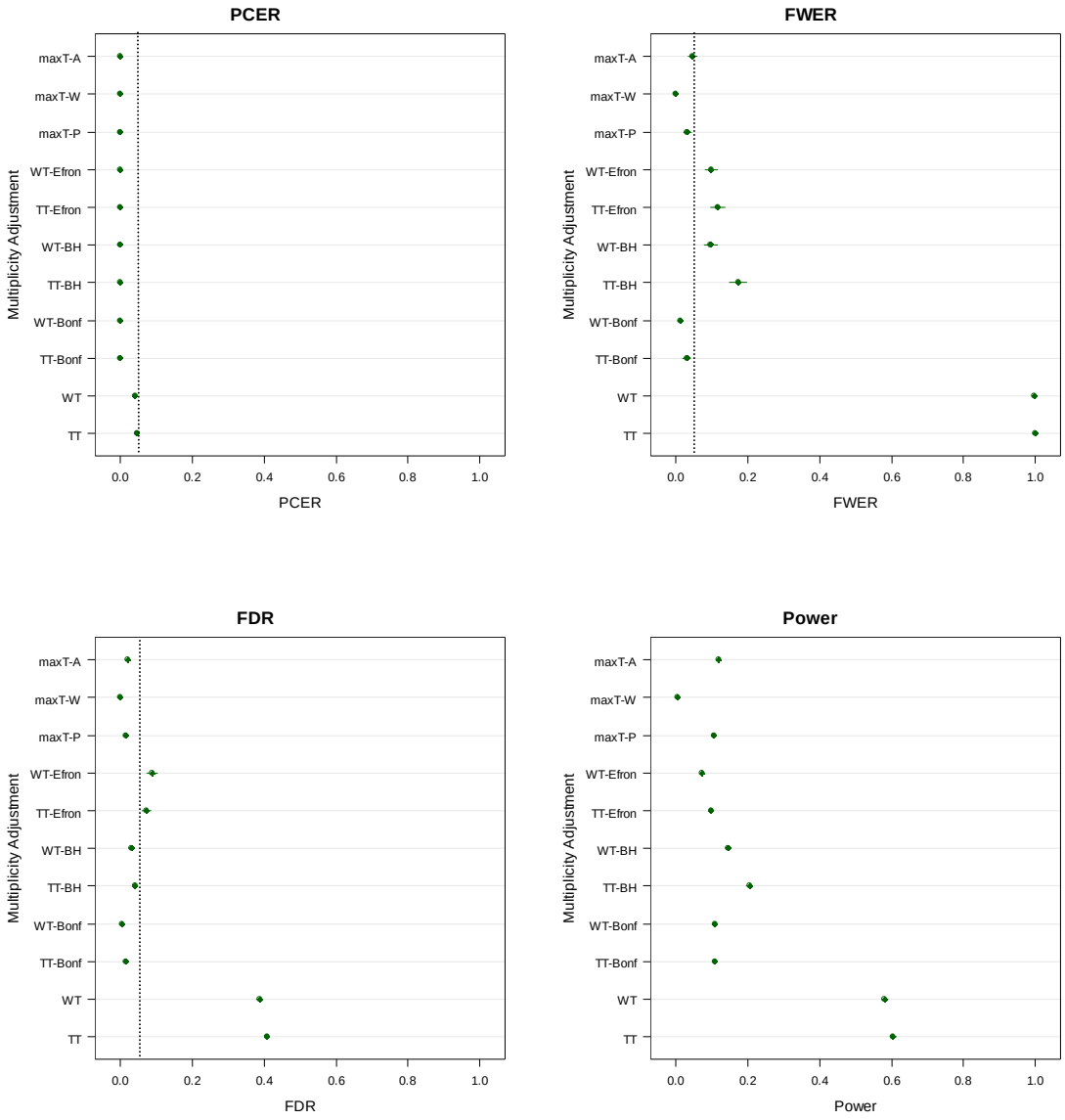

Figure G-2. Simulation Results (0.05); Error and Power Summary Comparing Eleven Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

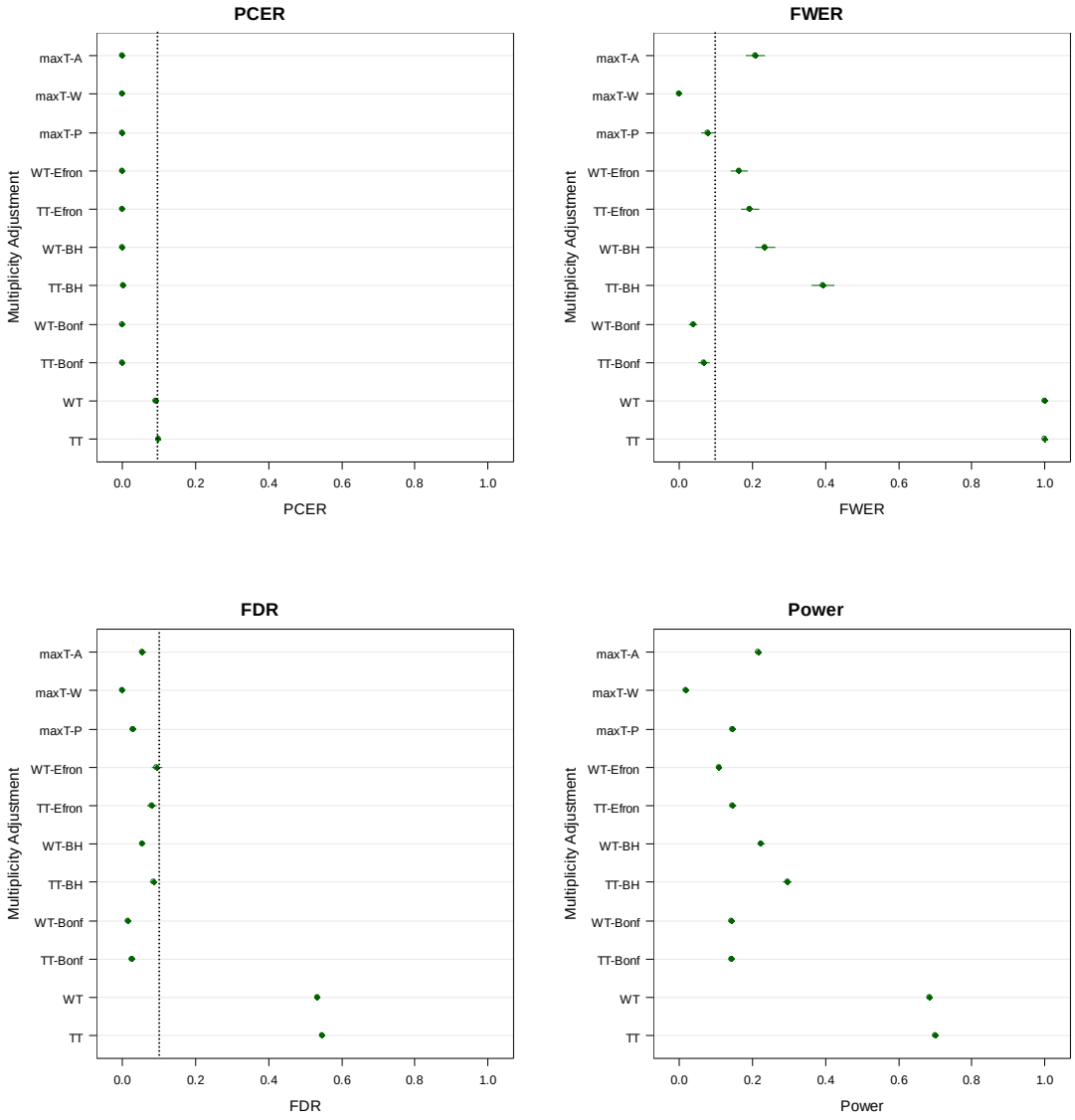

Figure G-3. Simulation Results (0.10); Error and Power Summary Comparing Eleven Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

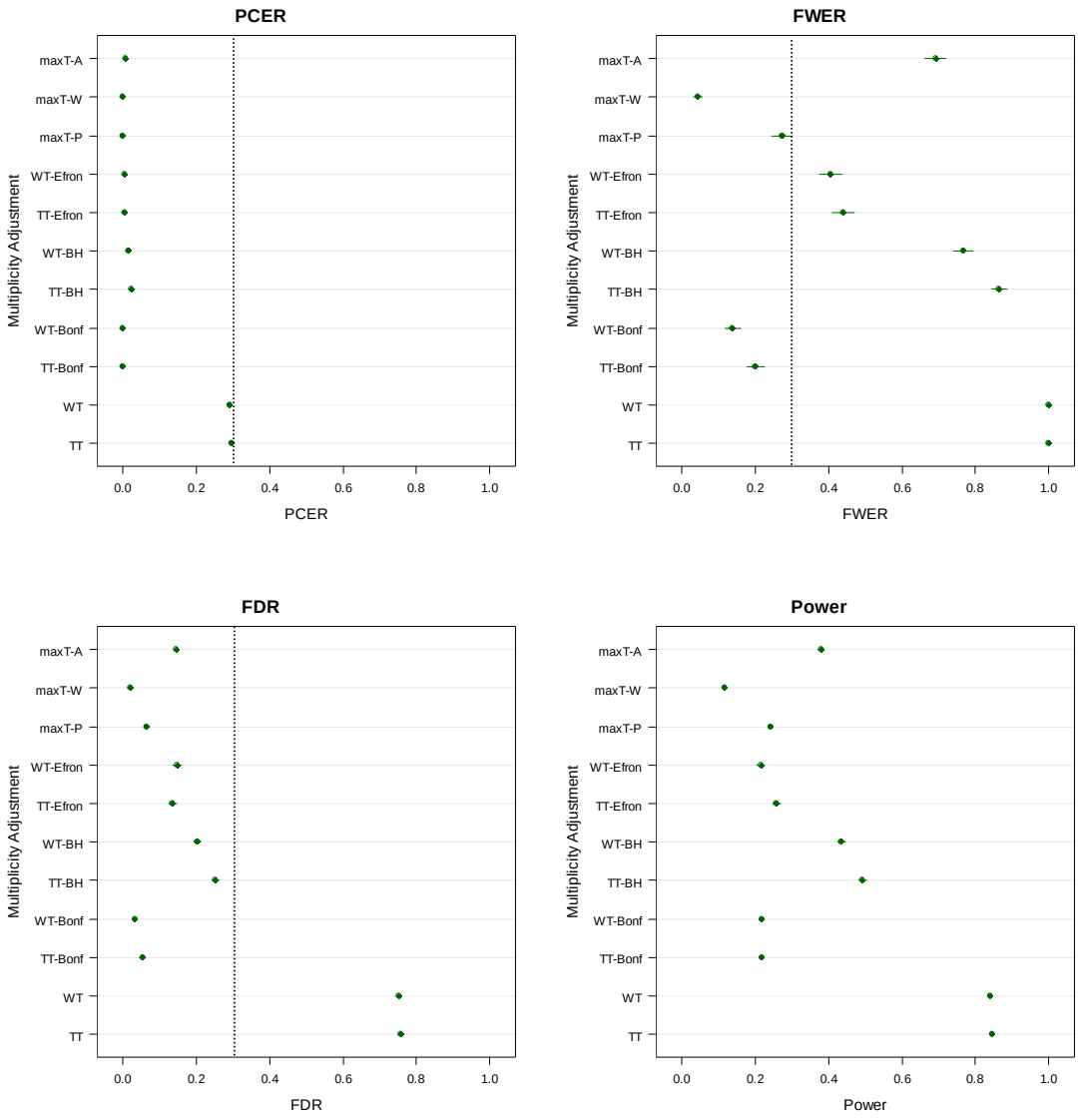

Figure G-4. Simulation Results (0.30); Error and Power Summary Comparing Eleven Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

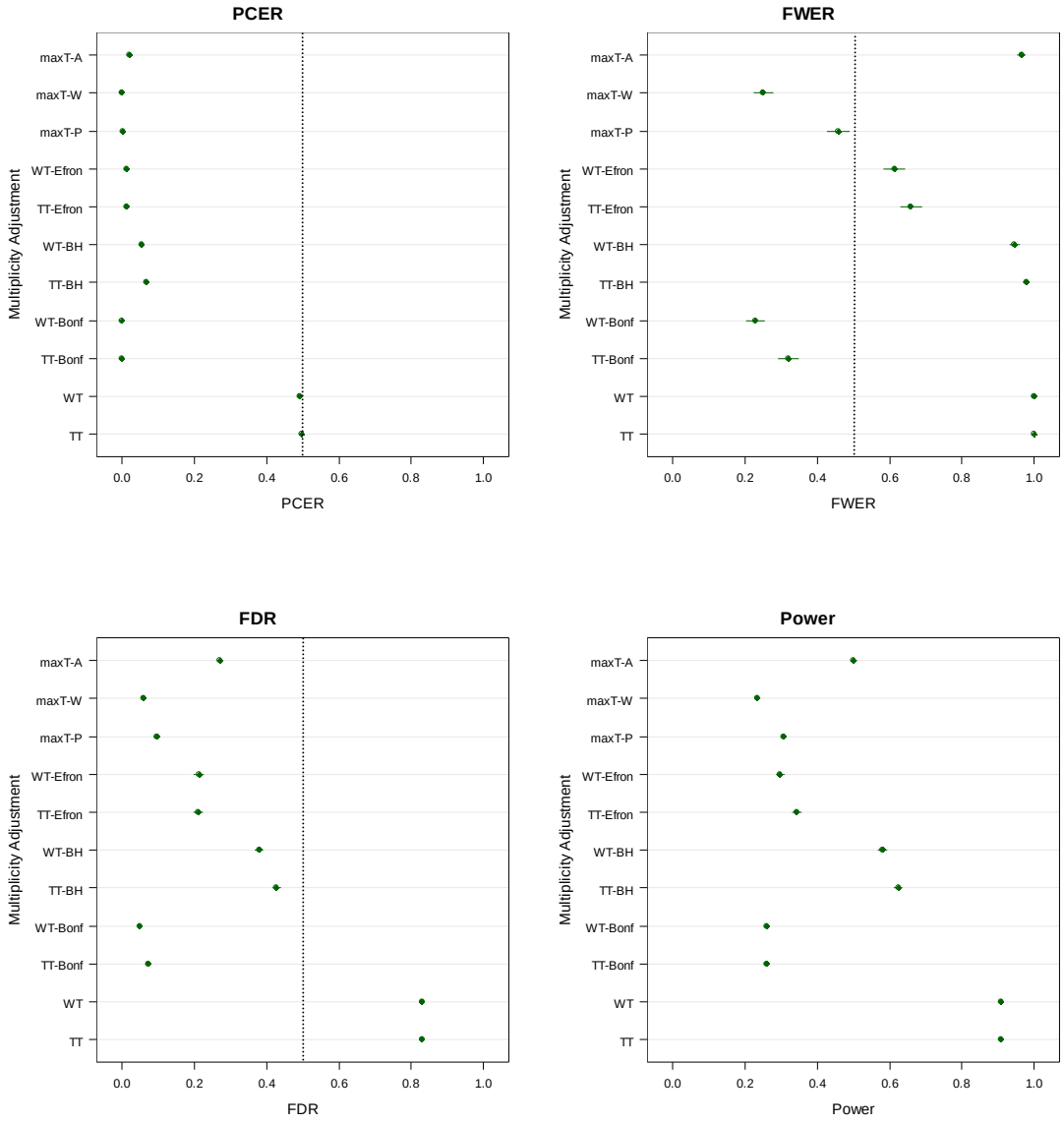

Figure G-5. Simulation Results (0.50); Error and Power Summary Comparing Eleven Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

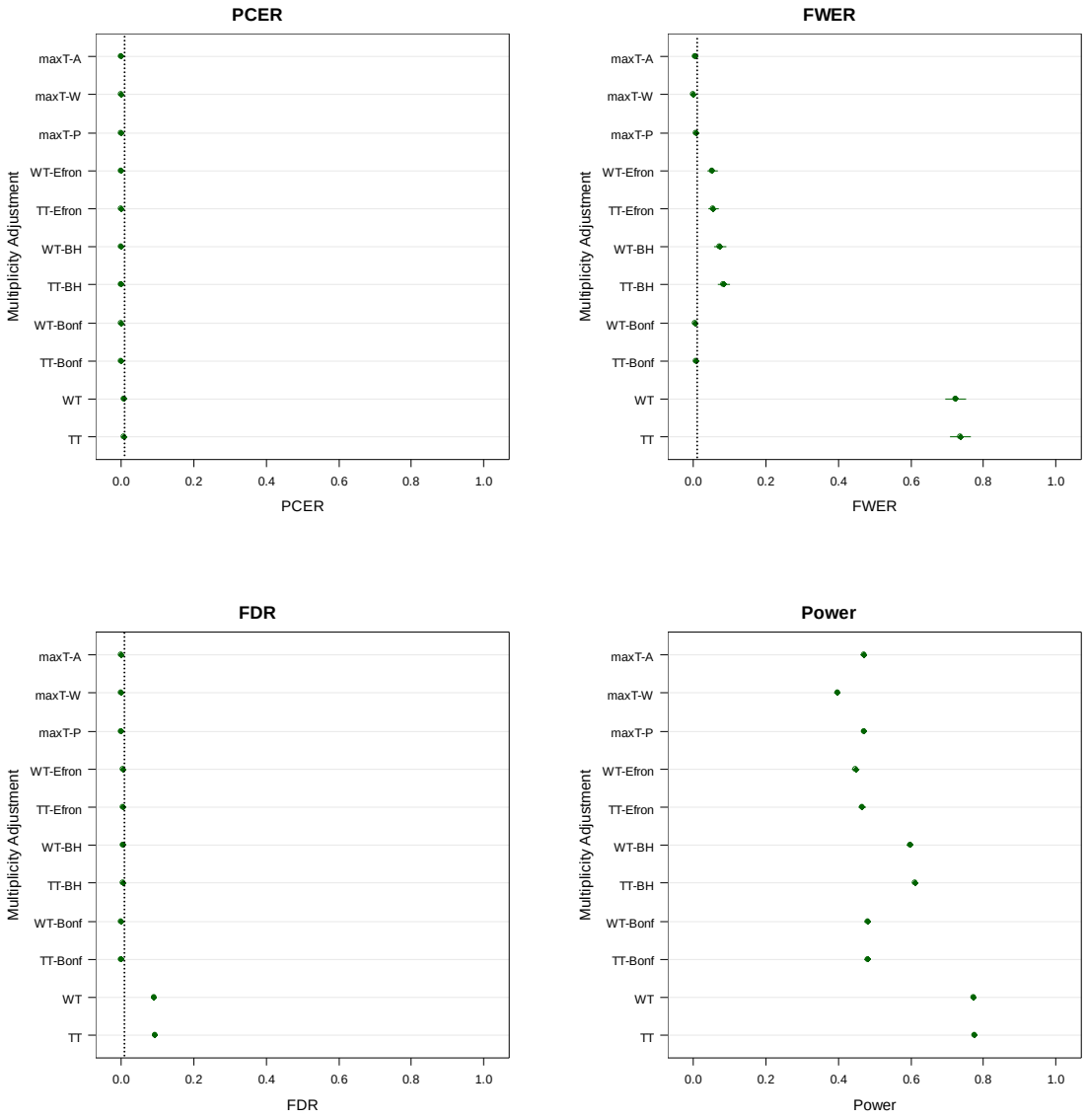

Figure G-6. Simulation Results (0.01); Error and Power Summary Comparing Eleven Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

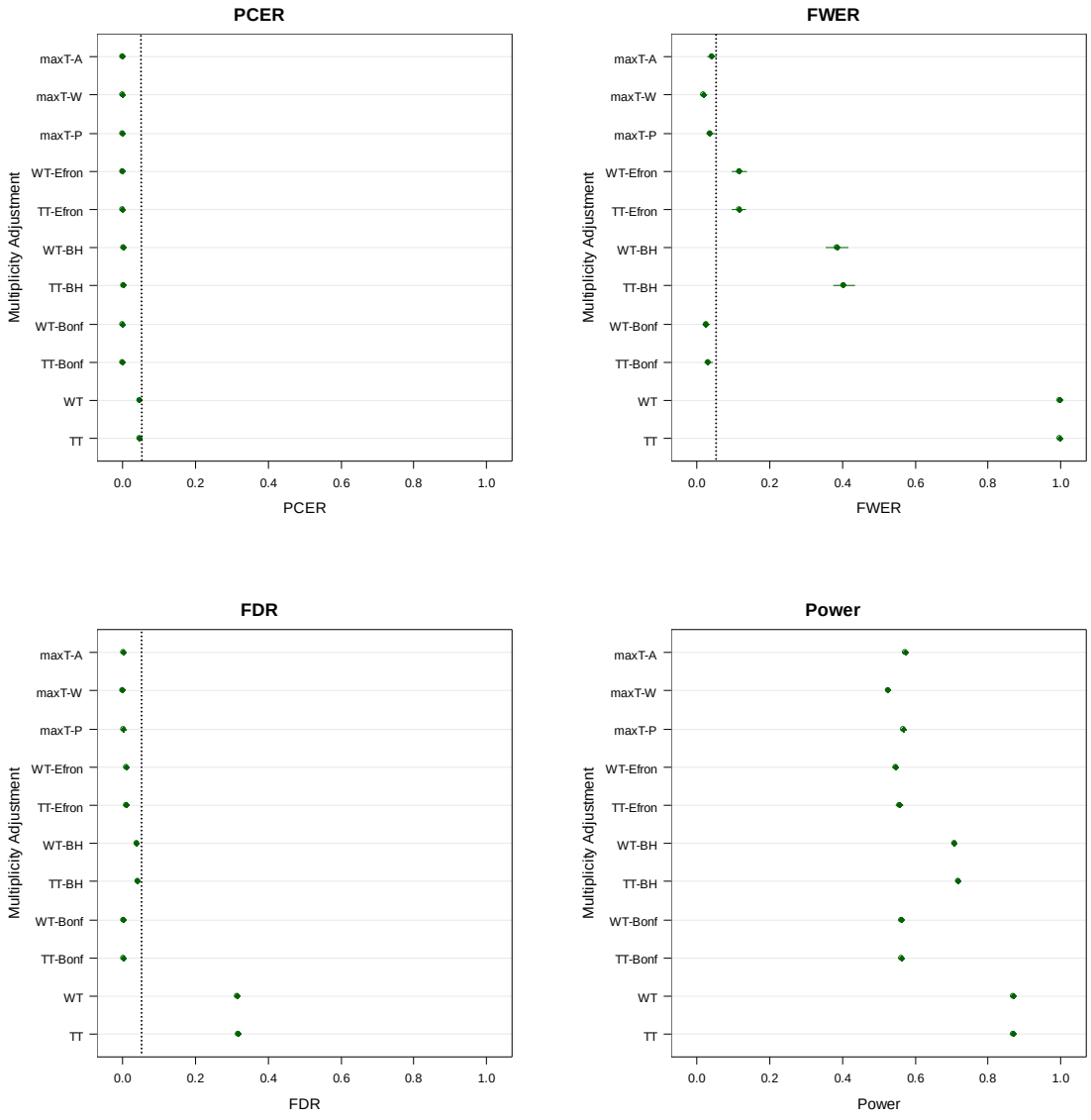

Figure G-7. Simulation Results (0.05); Error and Power Summary Comparing Eleven Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

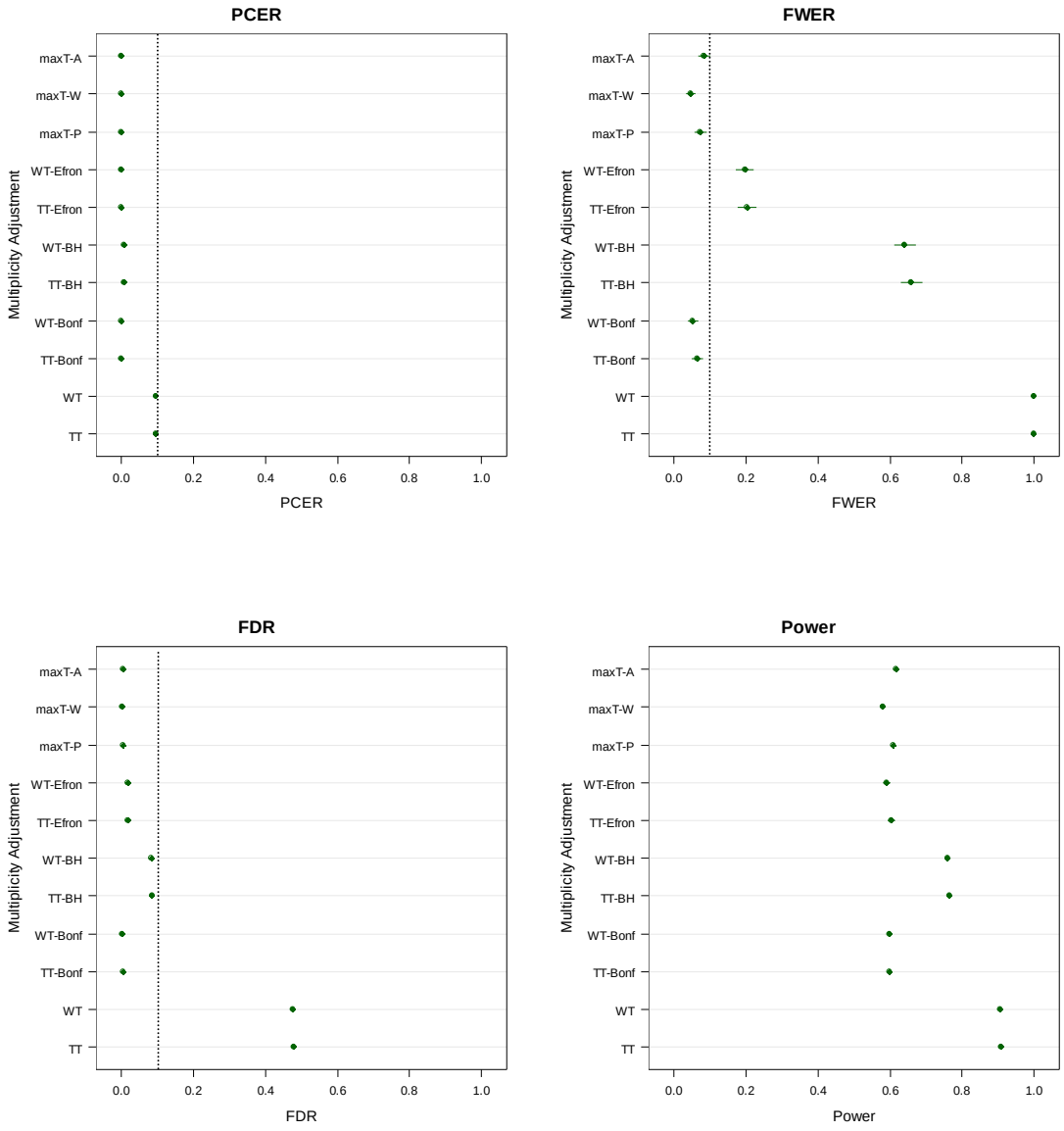

Figure G-8. Simulation Results (0.10); Error and Power Summary Comparing Eleven Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

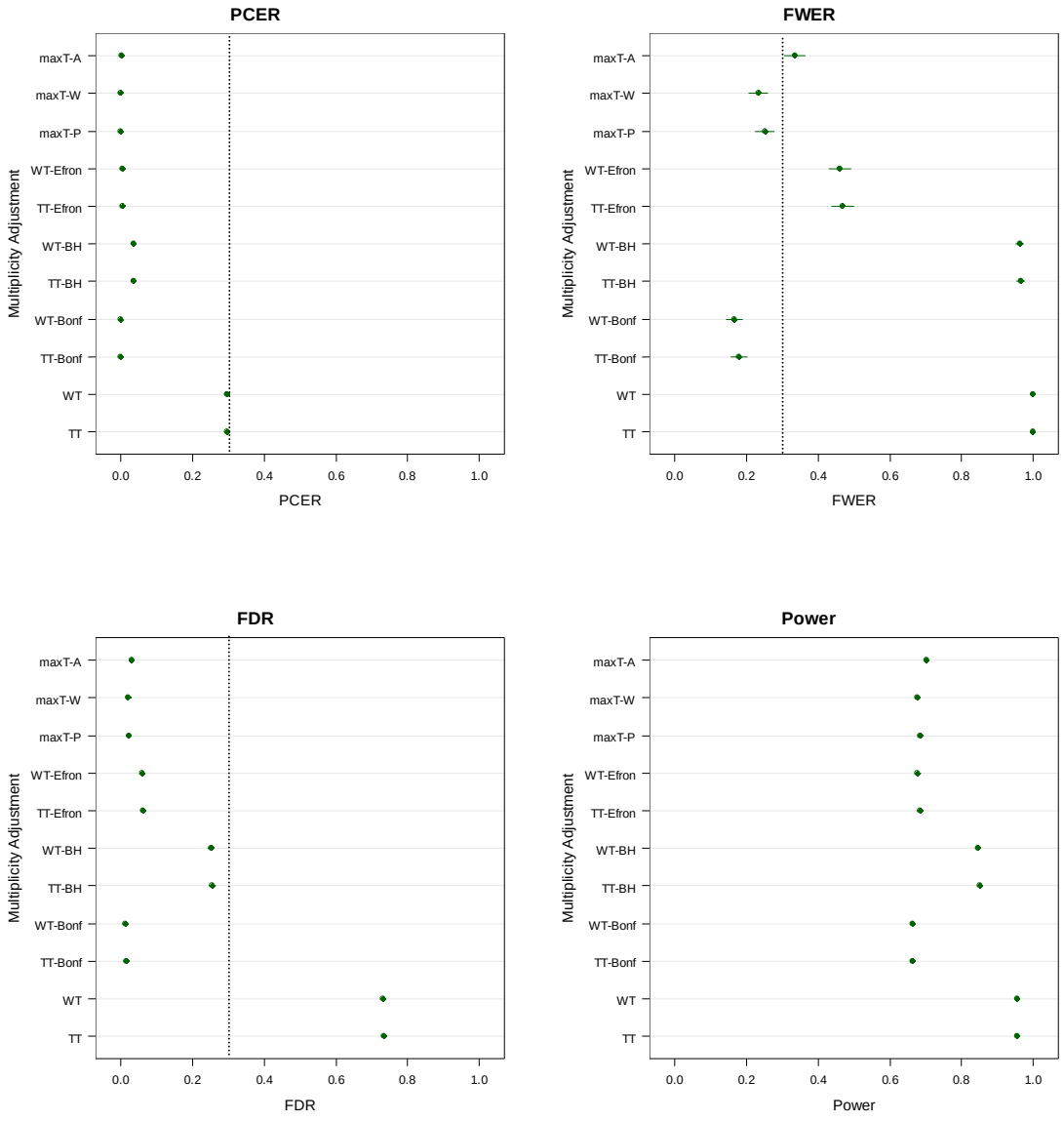

Figure G-9. Simulation Results (0.30); Error and Power Summary Comparing Eleven Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

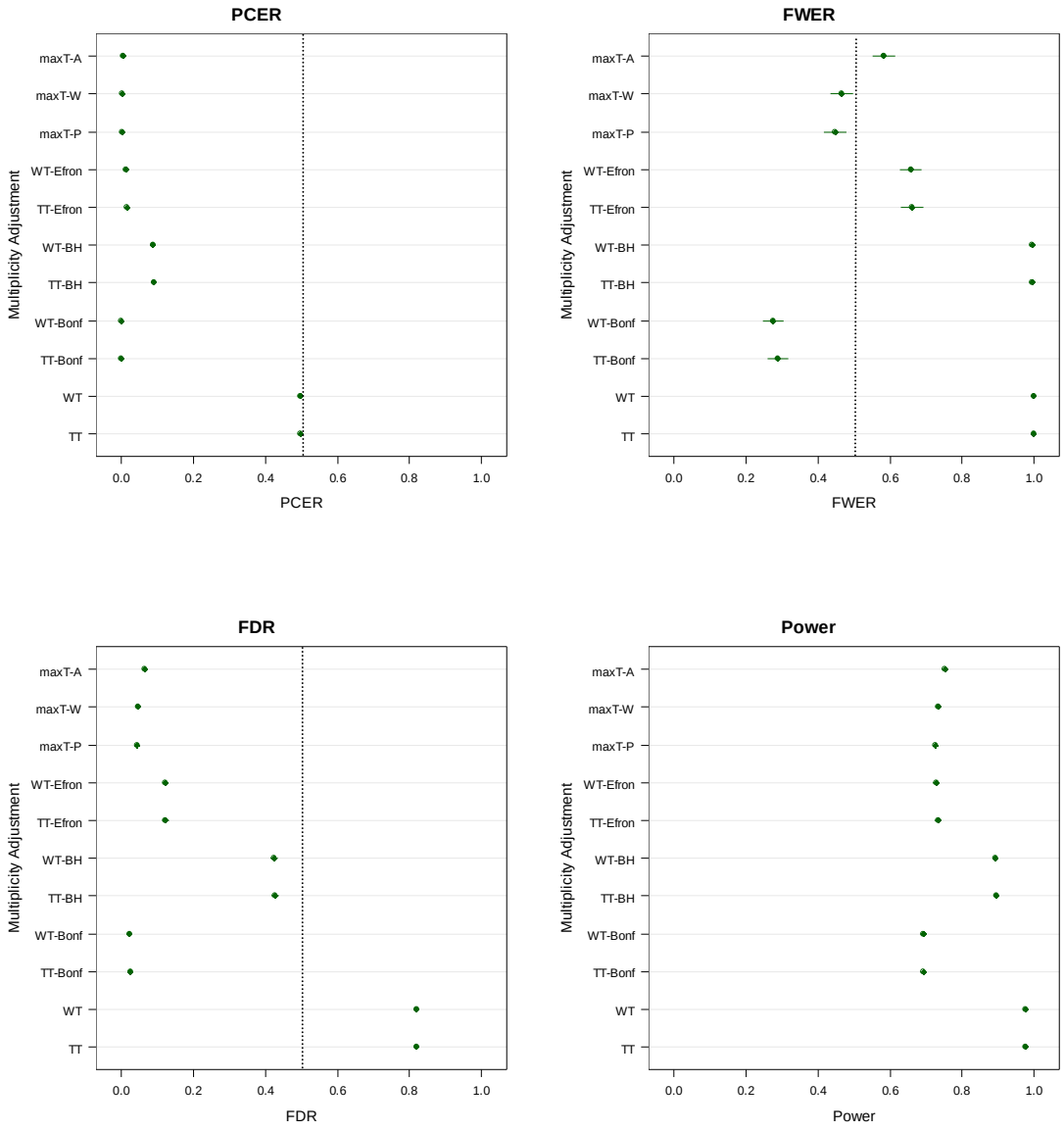

Figure G-10. Simulation Results (0.50); Error and Power Summary Comparing Eleven Methods of Adjustment for Multiplicity. Scenario: 200 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

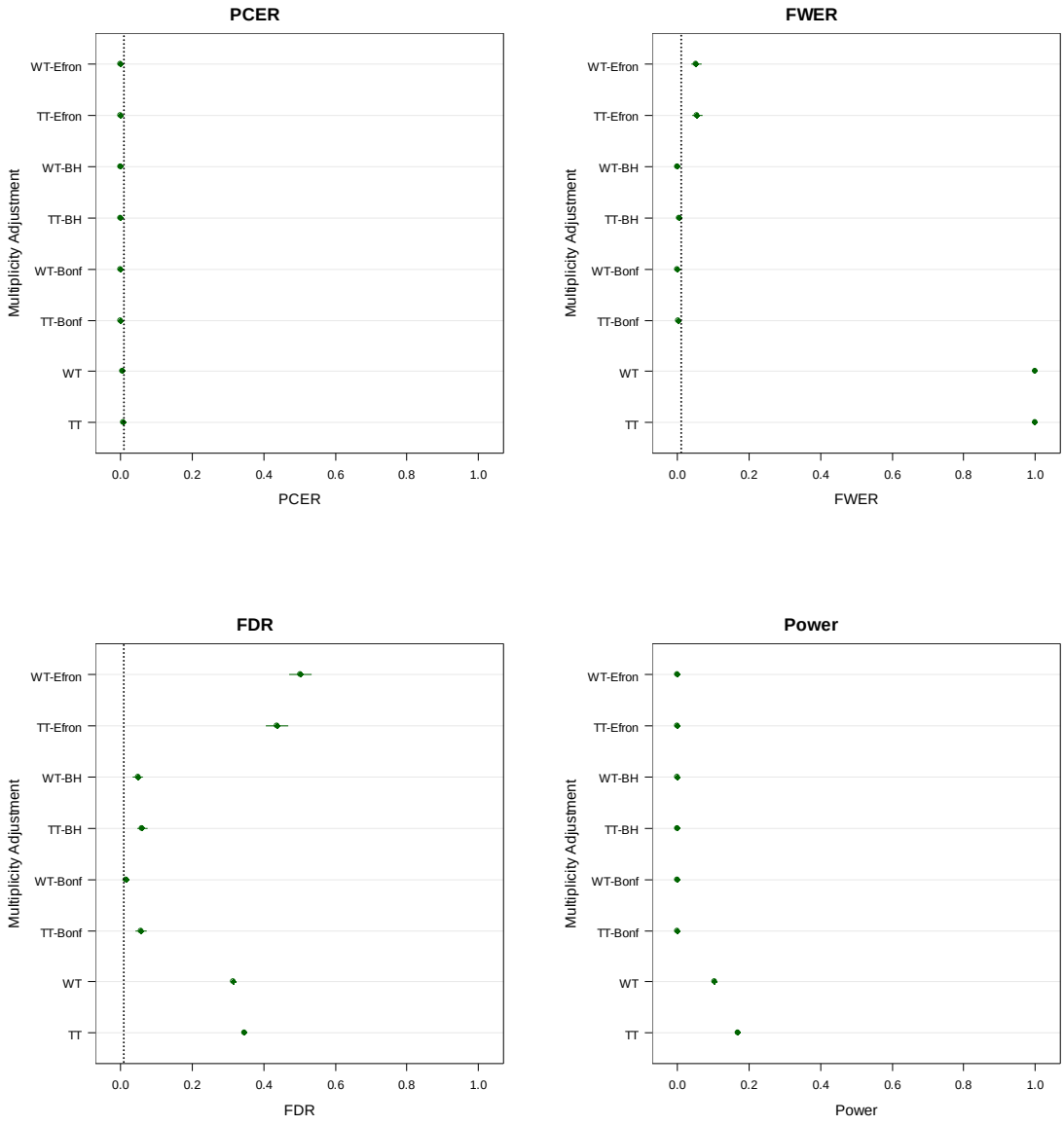

Figure G-11. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

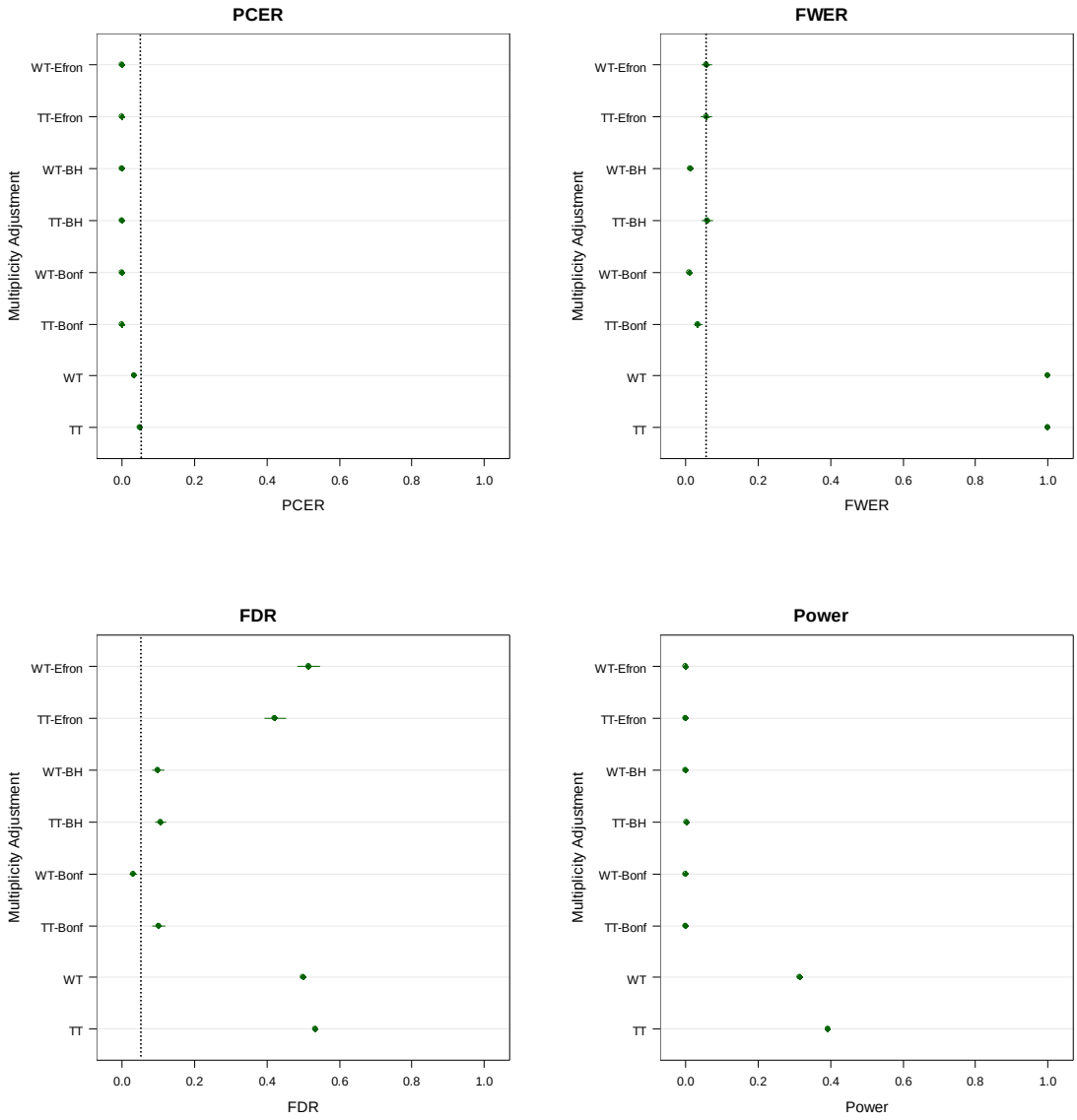

Figure G-12. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

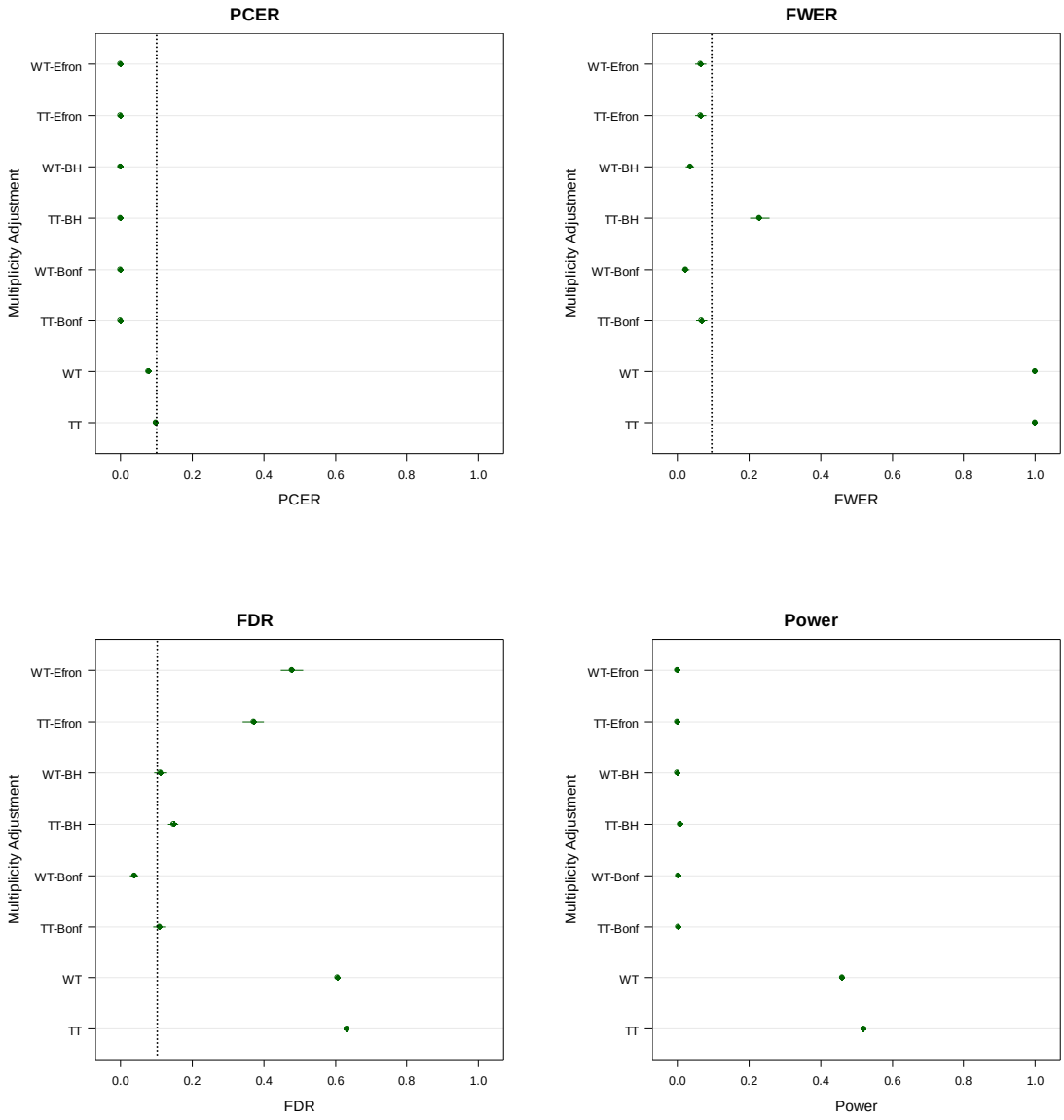

Figure G-13. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

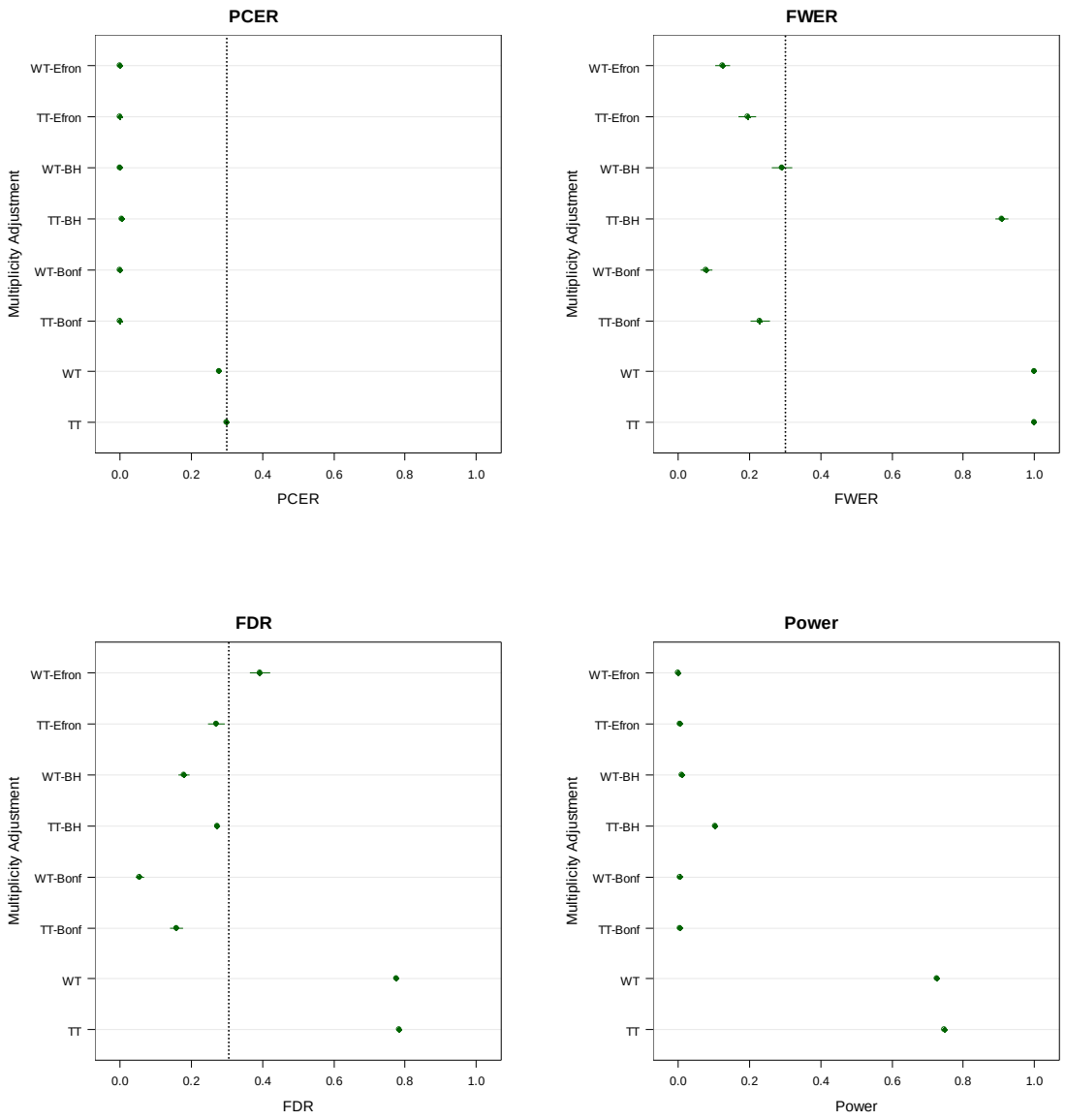

Figure G-14. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

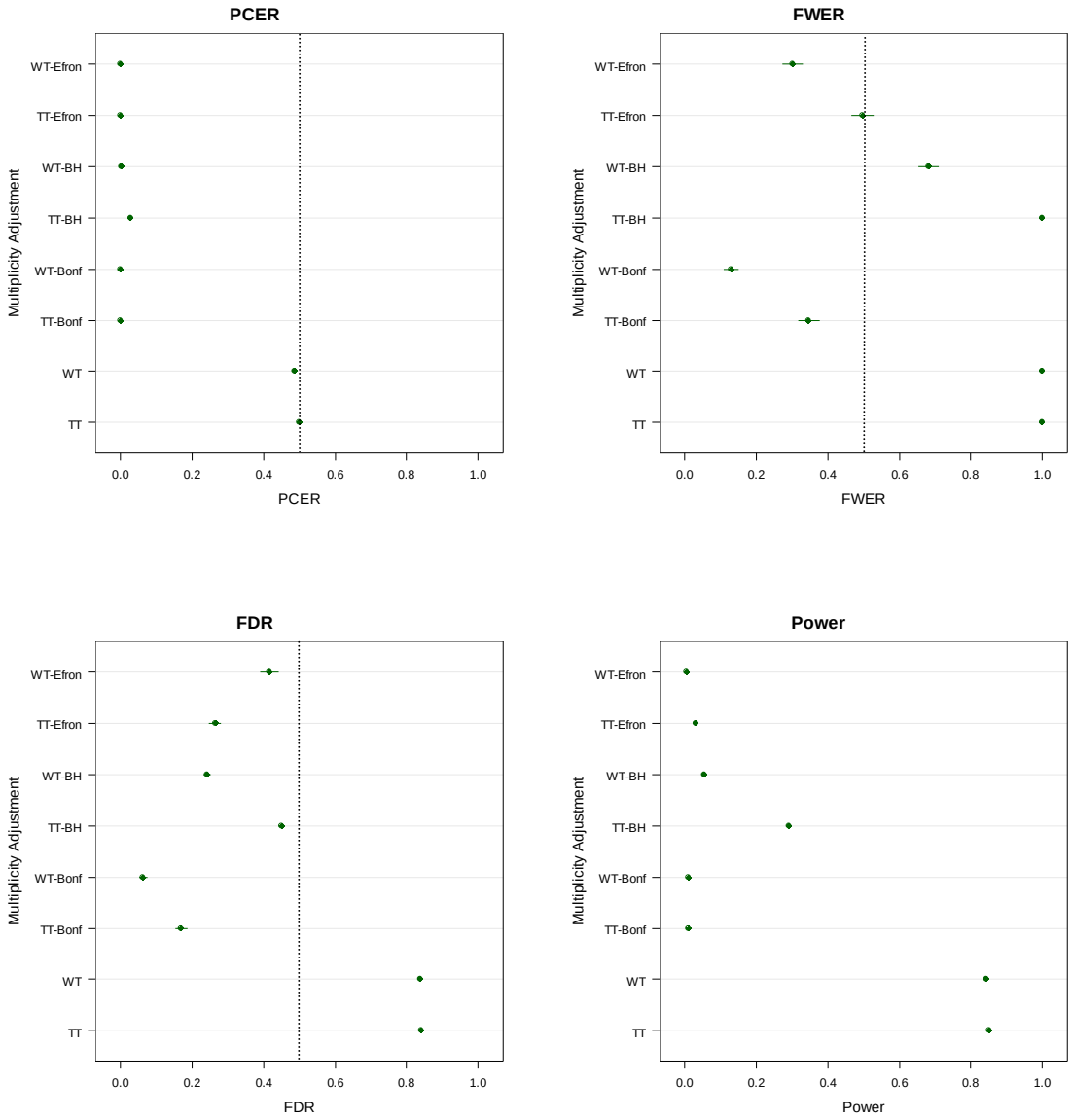

Figure G-15. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 3 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

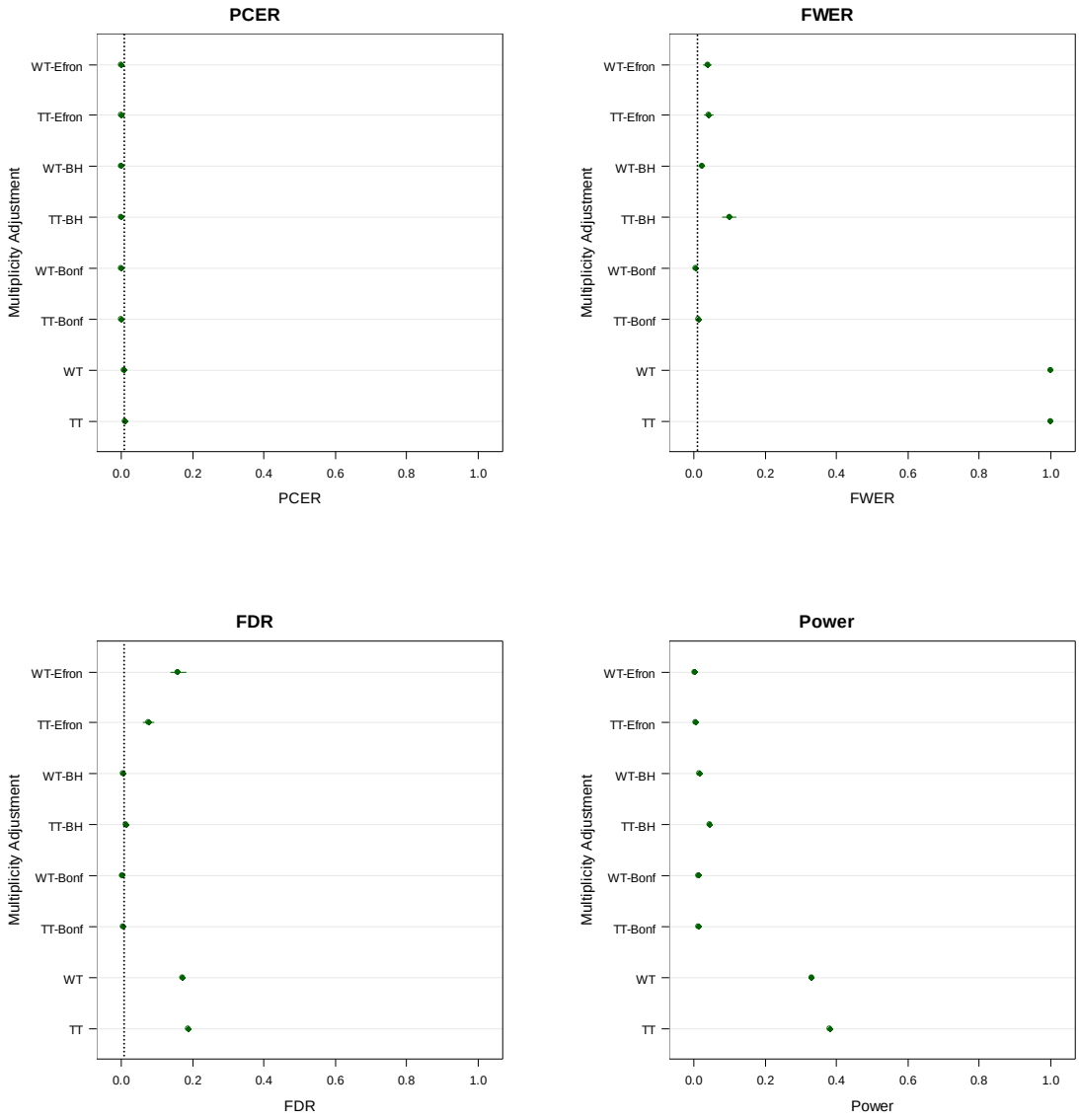

Figure G-16. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

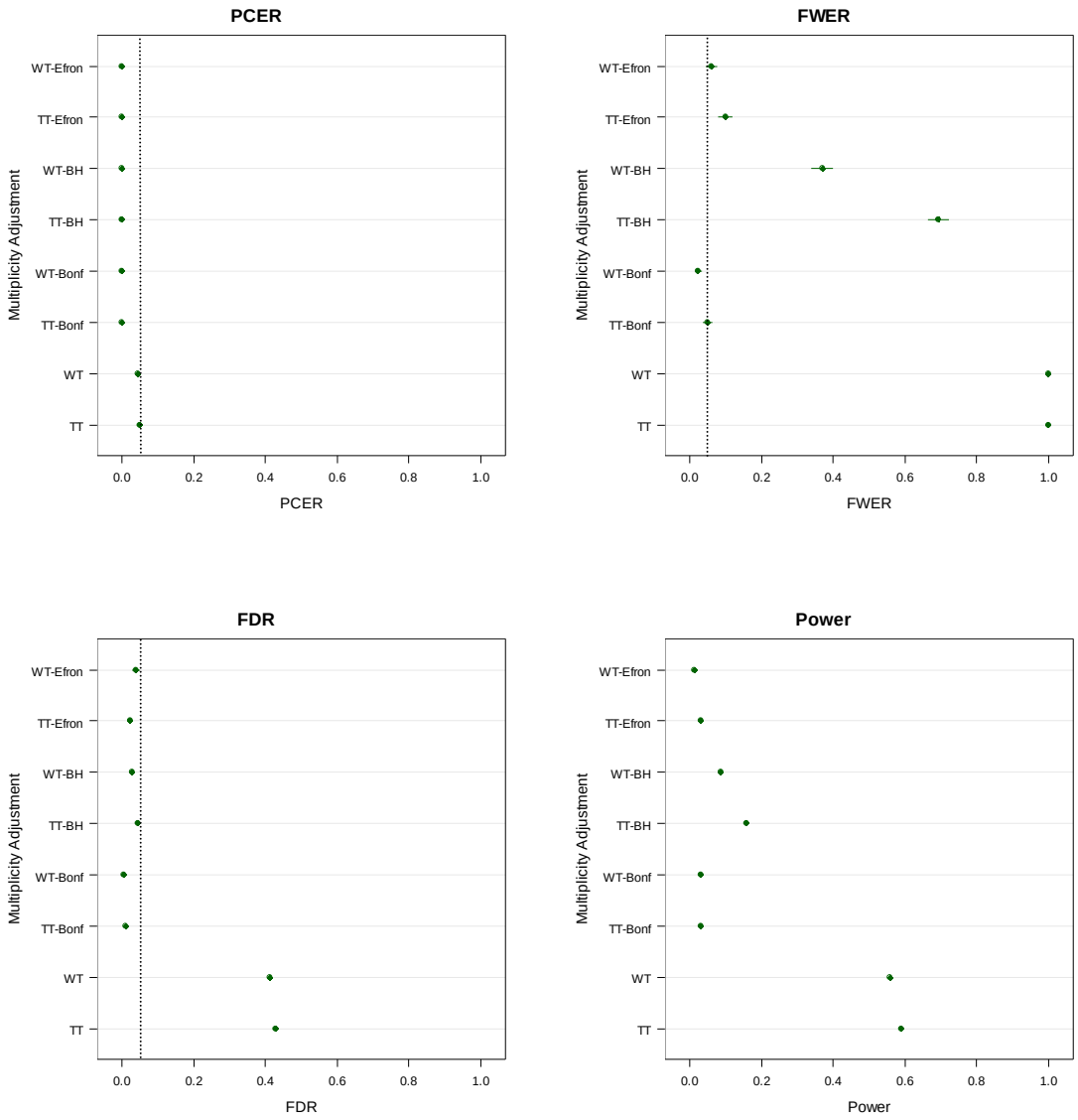

Figure G-17. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

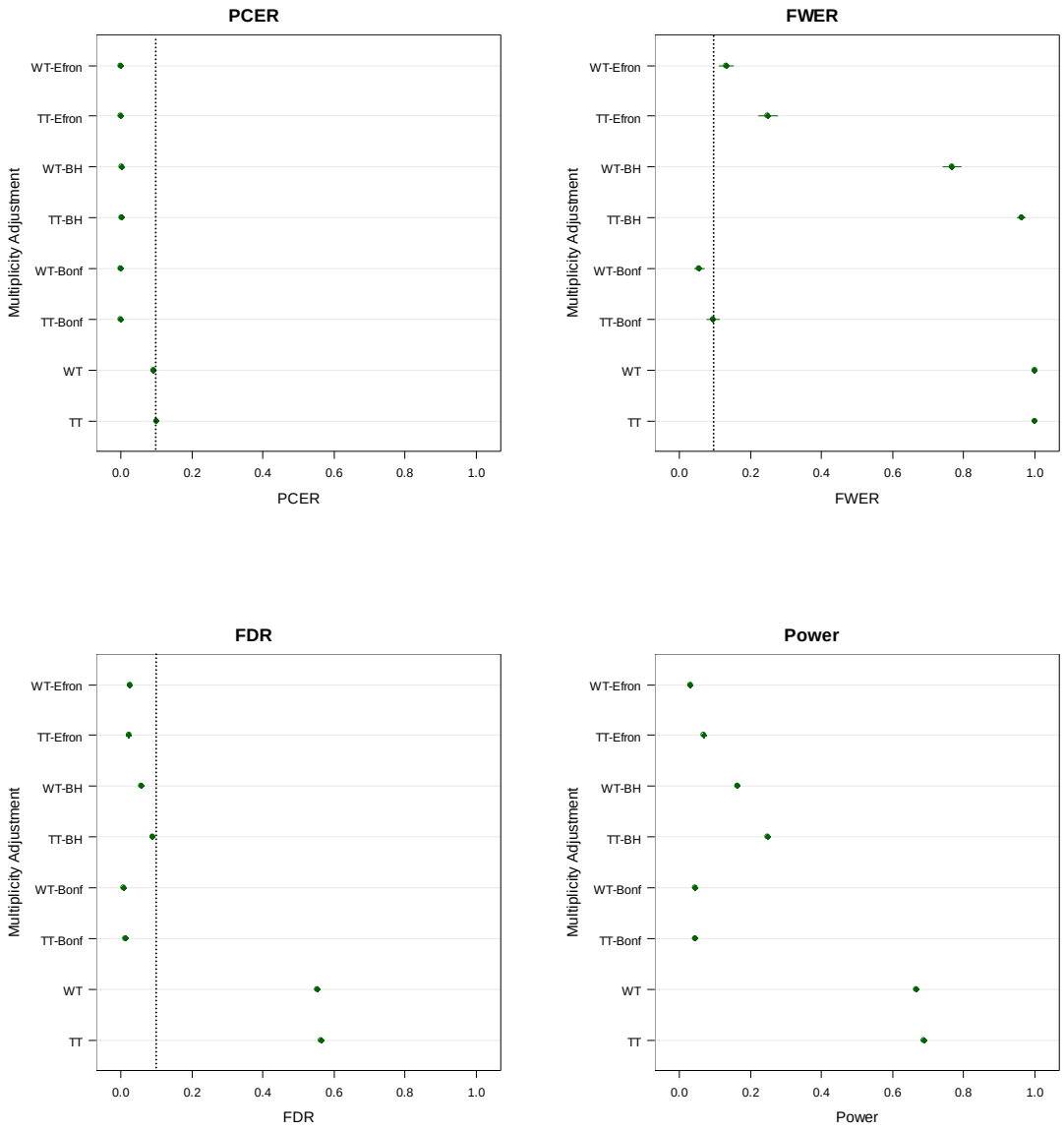

Figure G-18. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

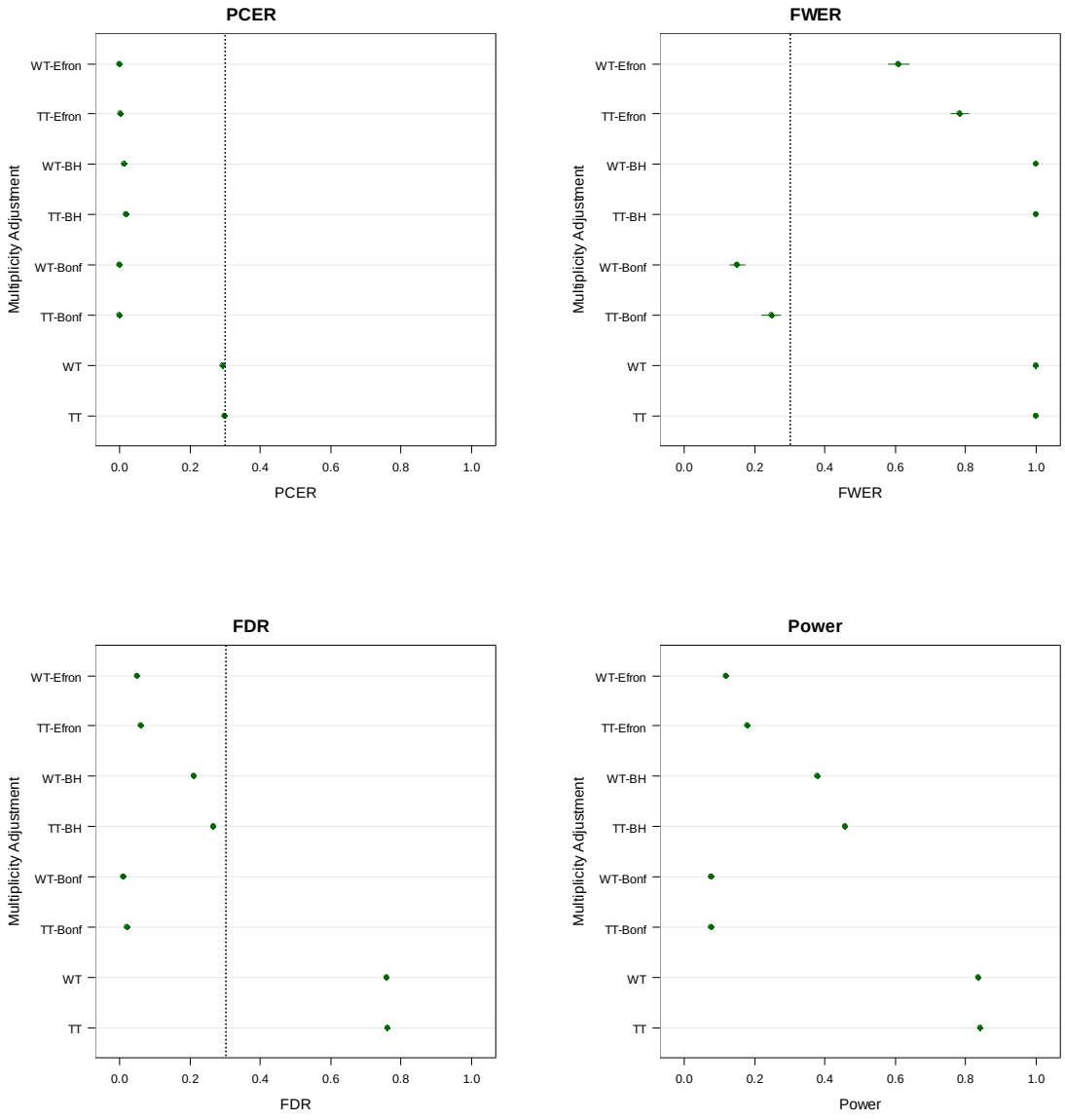

Figure G-19. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

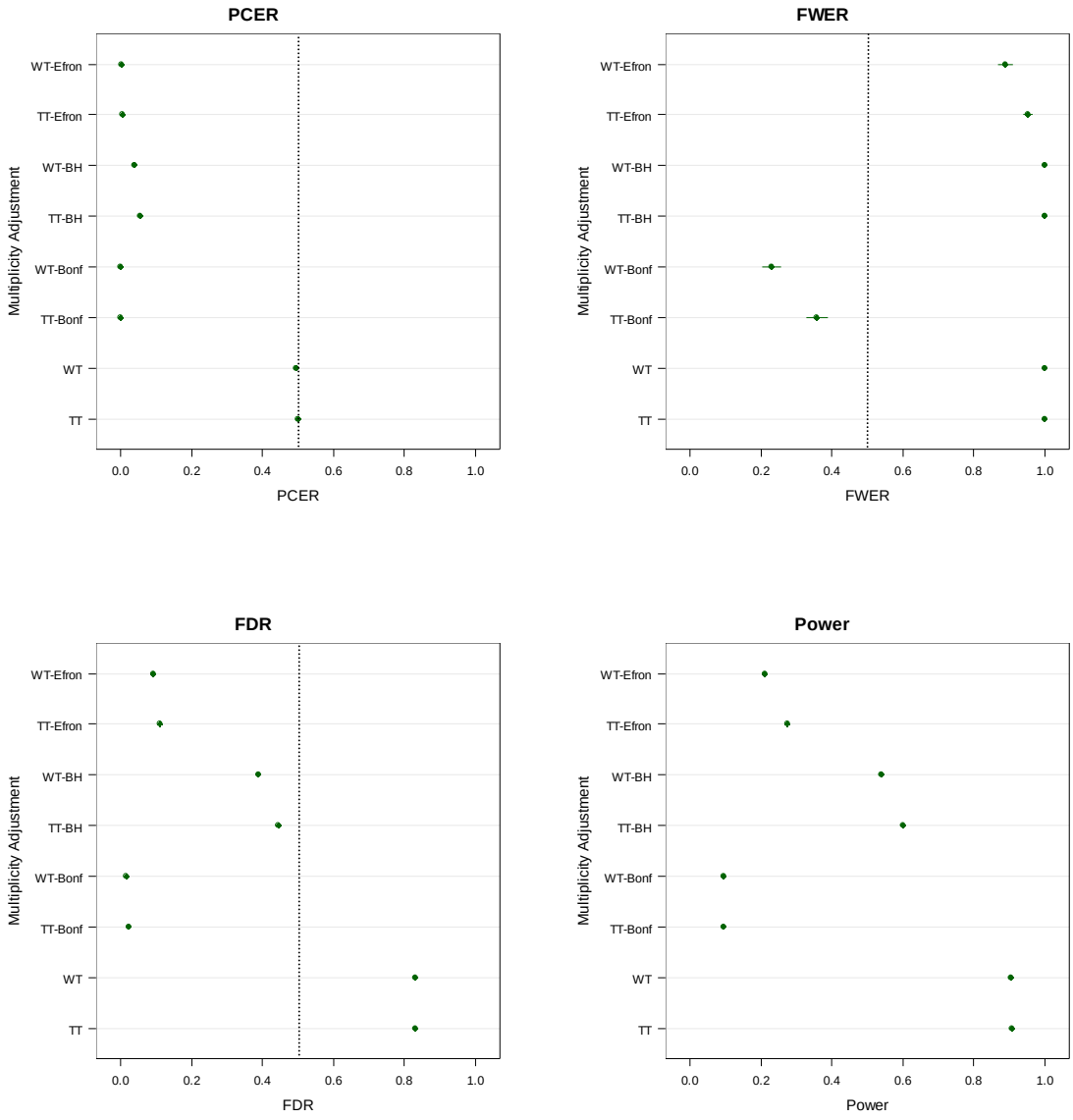

Figure G-20. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 5 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

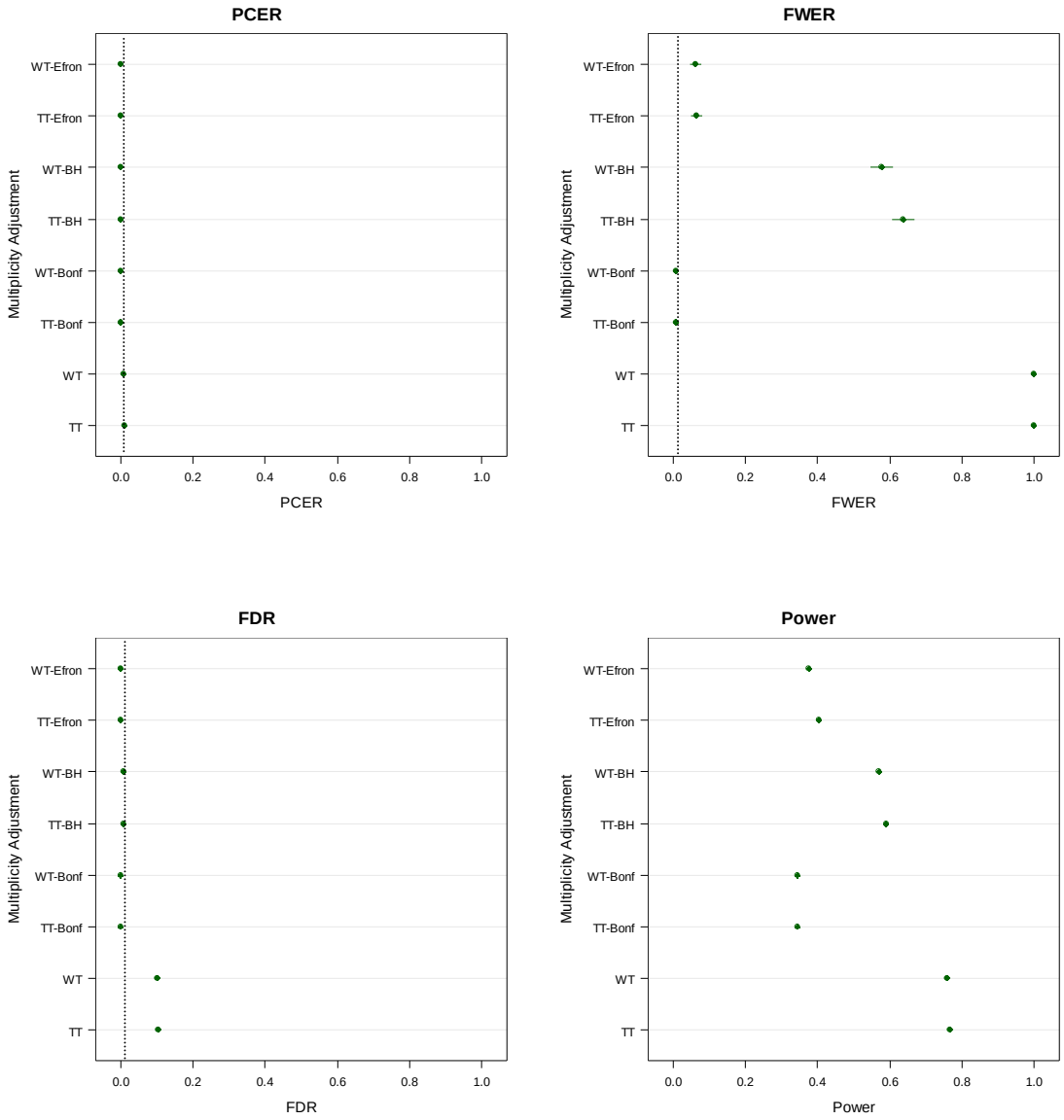

Figure G-21. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

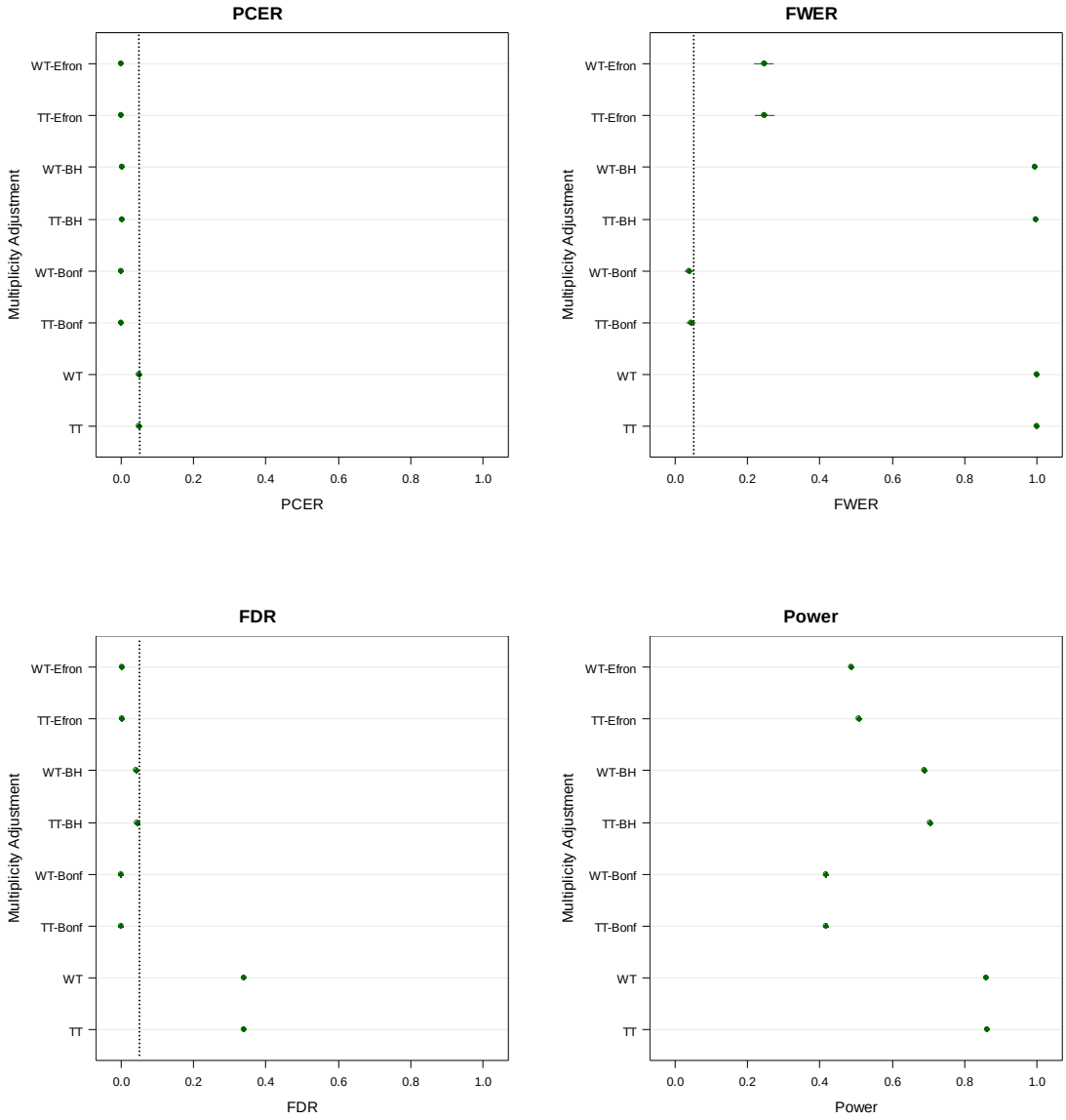

Figure G-22. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

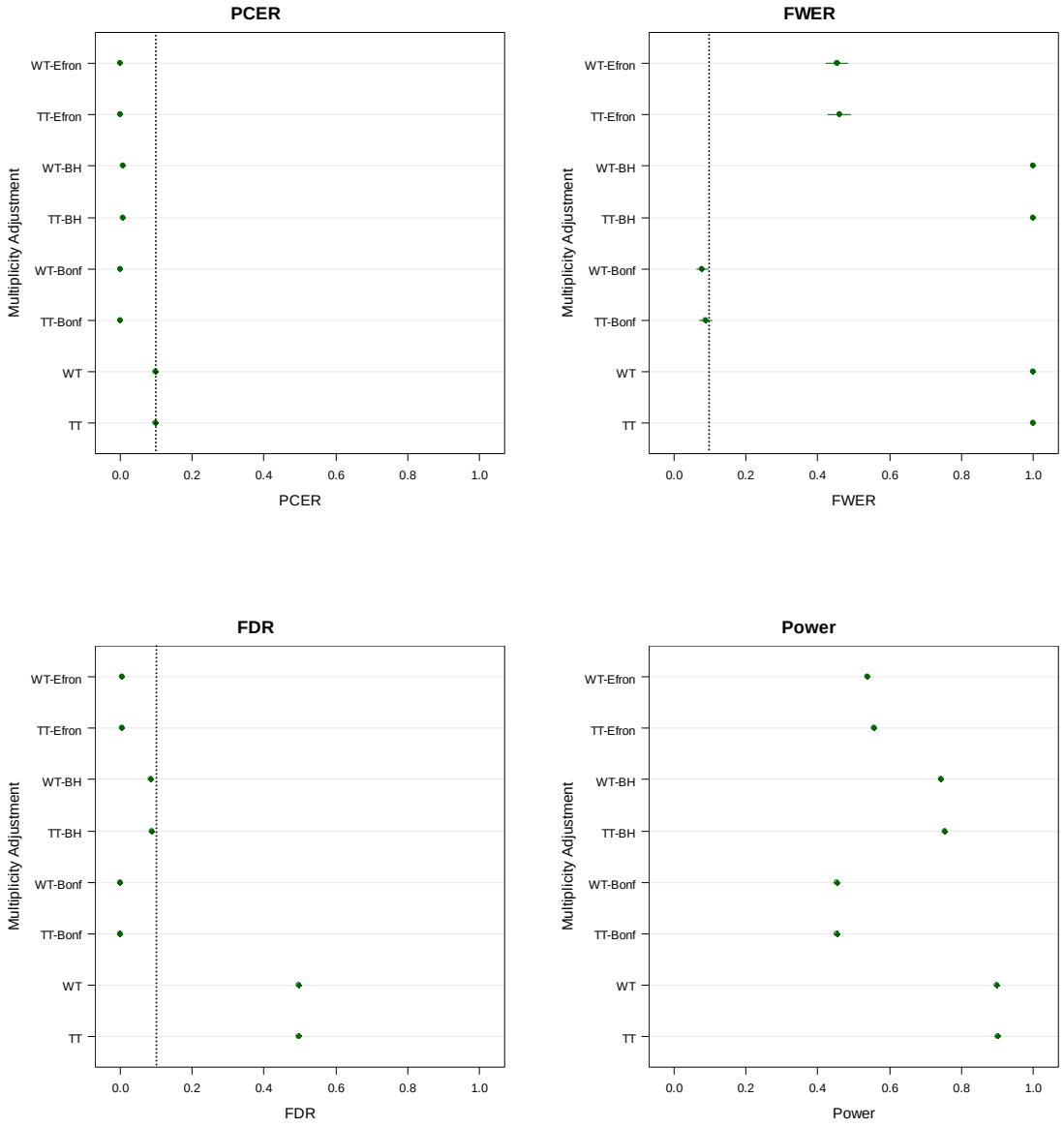

Figure G-23. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

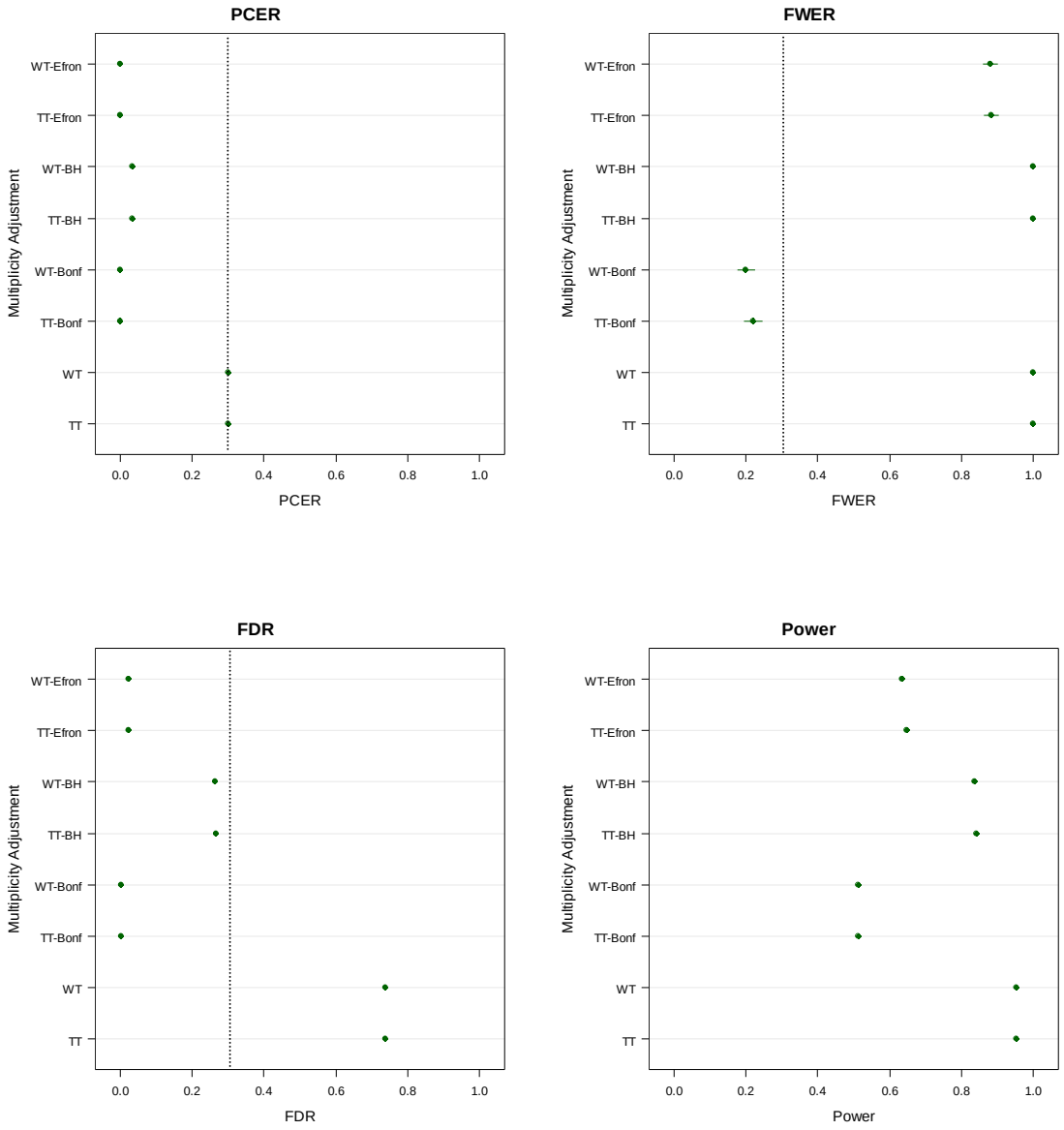

Figure G-24. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

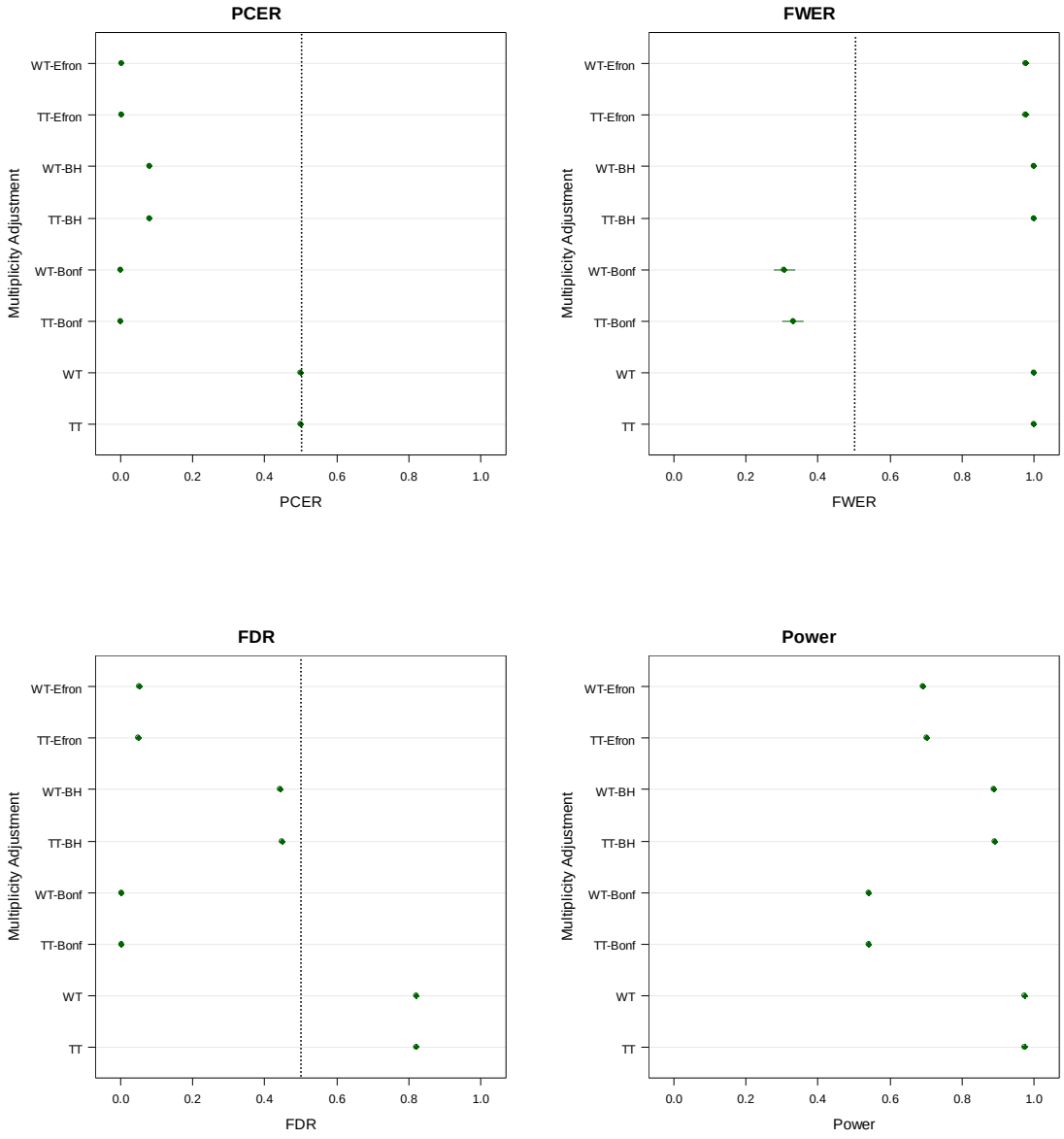

Figure G-25. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 15 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

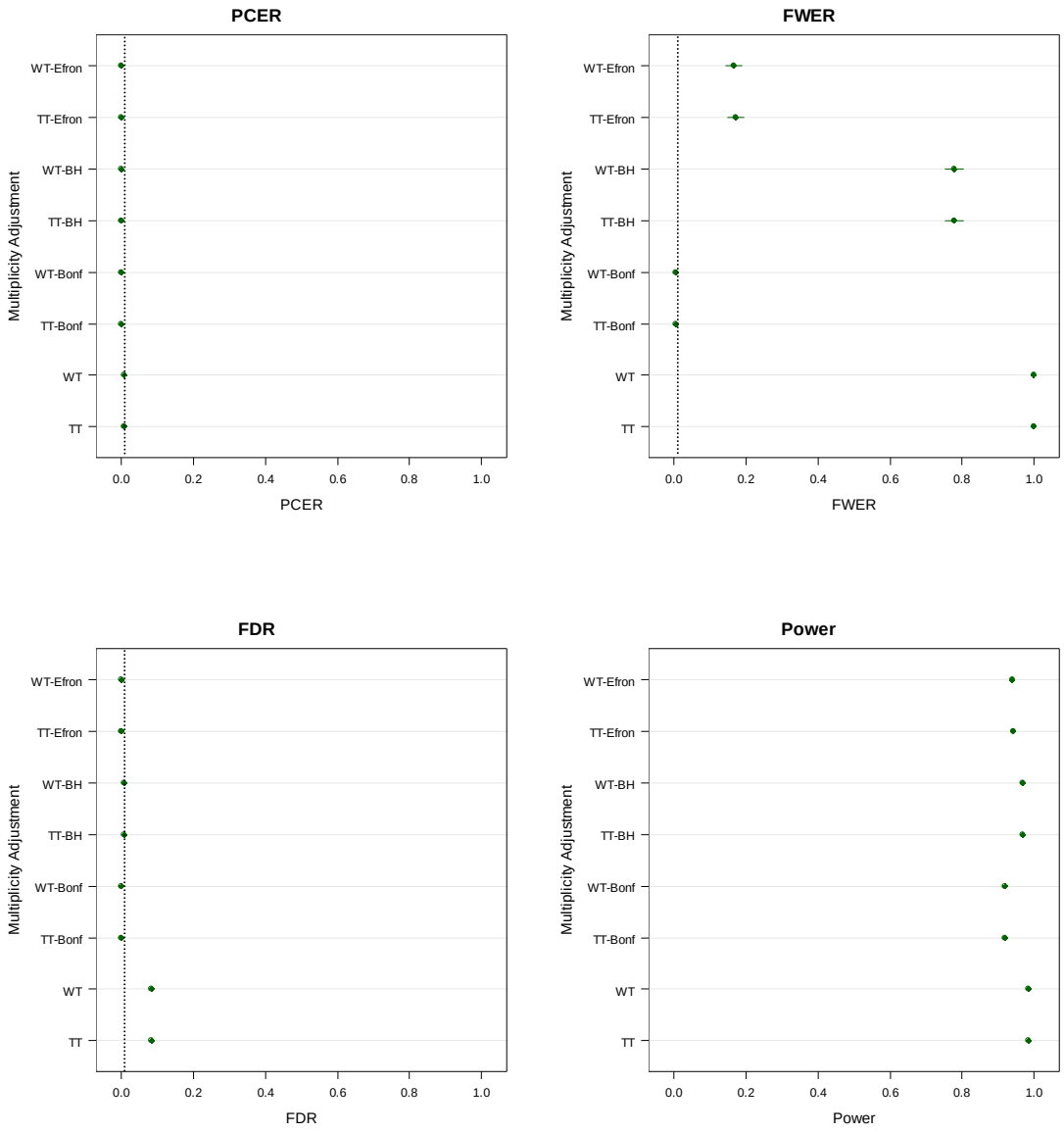

Figure G-26. Simulation Results (0.01); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.01.

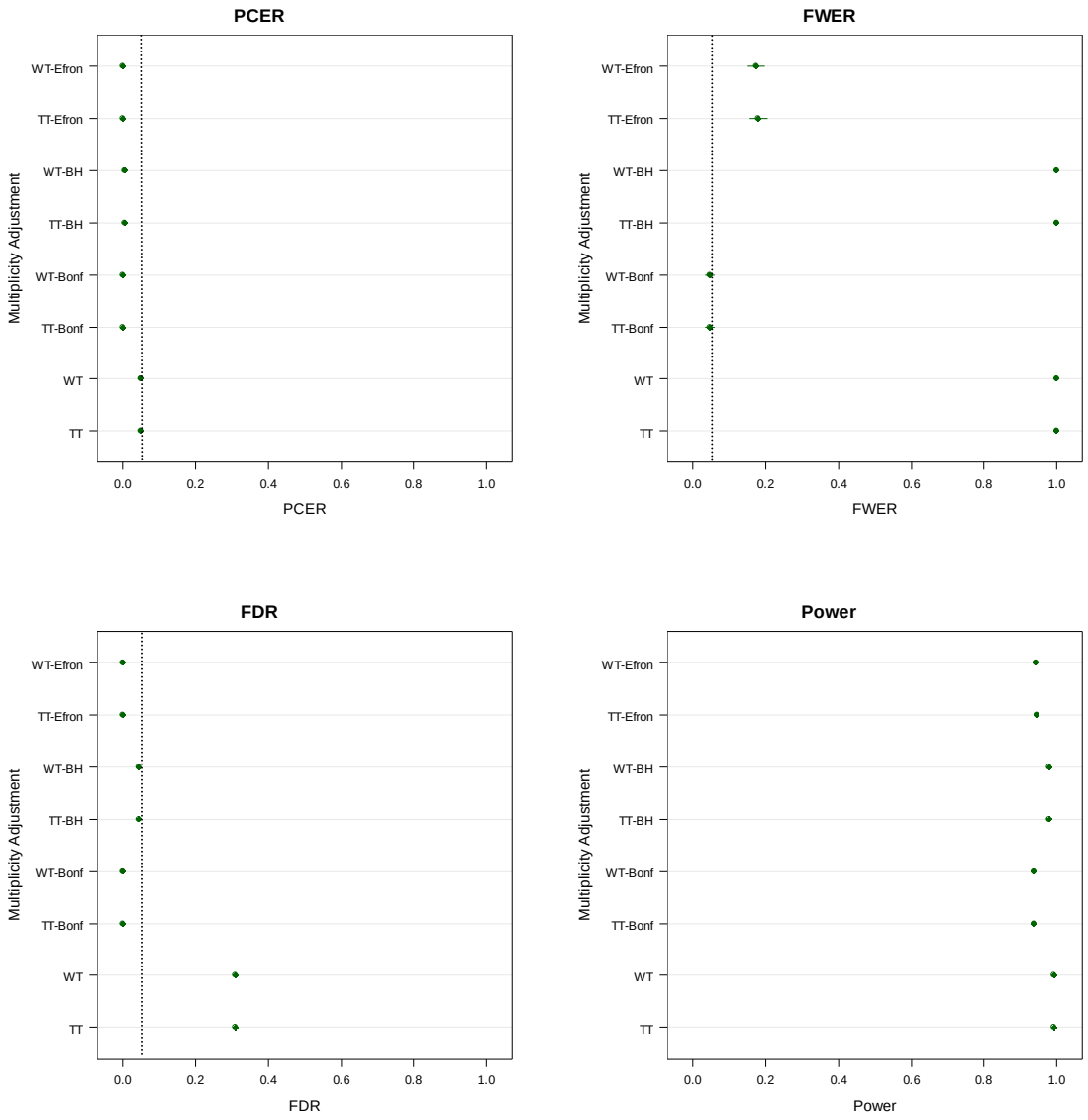

Figure G-27. Simulation Results (0.05); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.05.

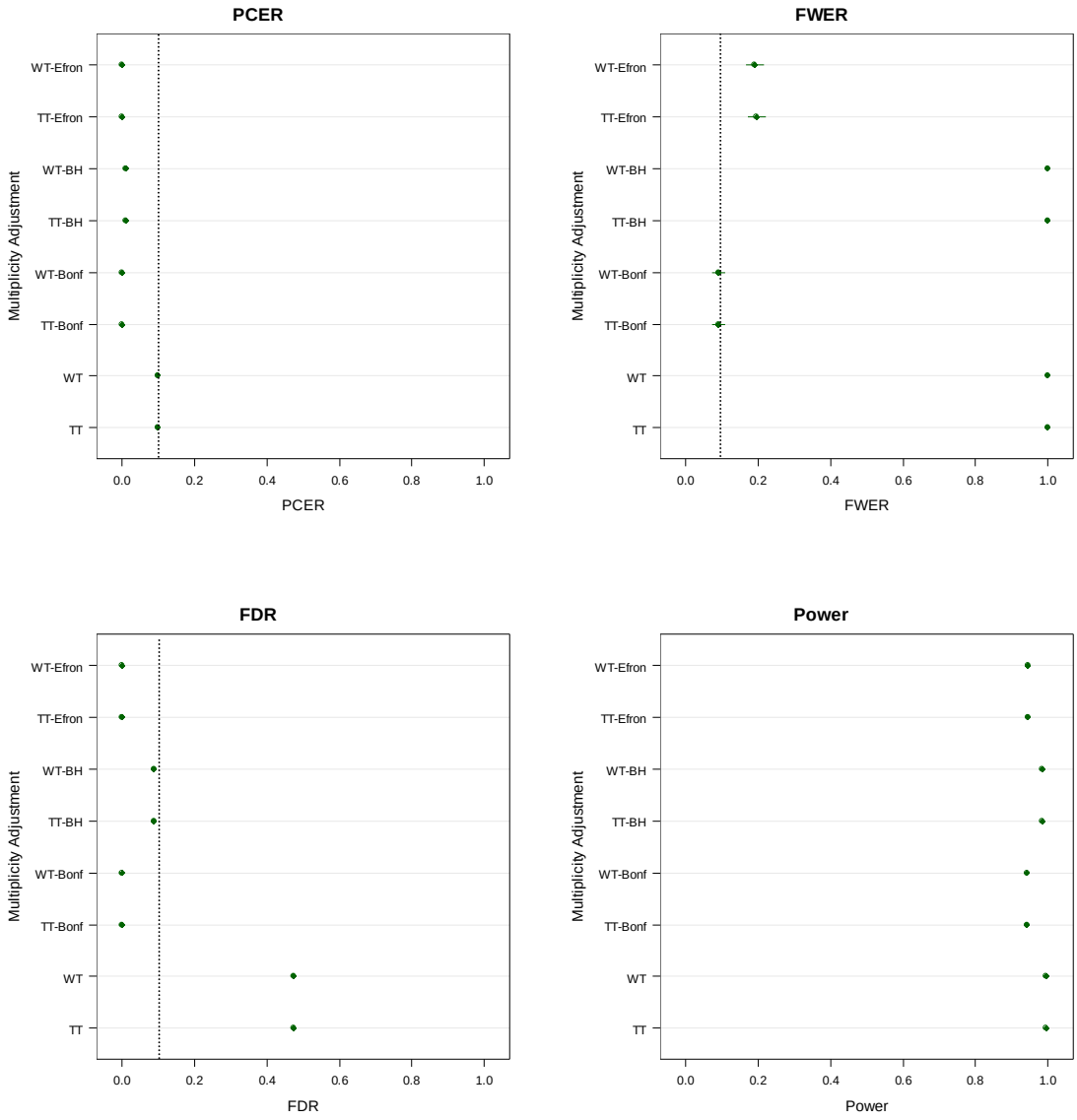

Figure G-28. Simulation Results (0.10); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.10.

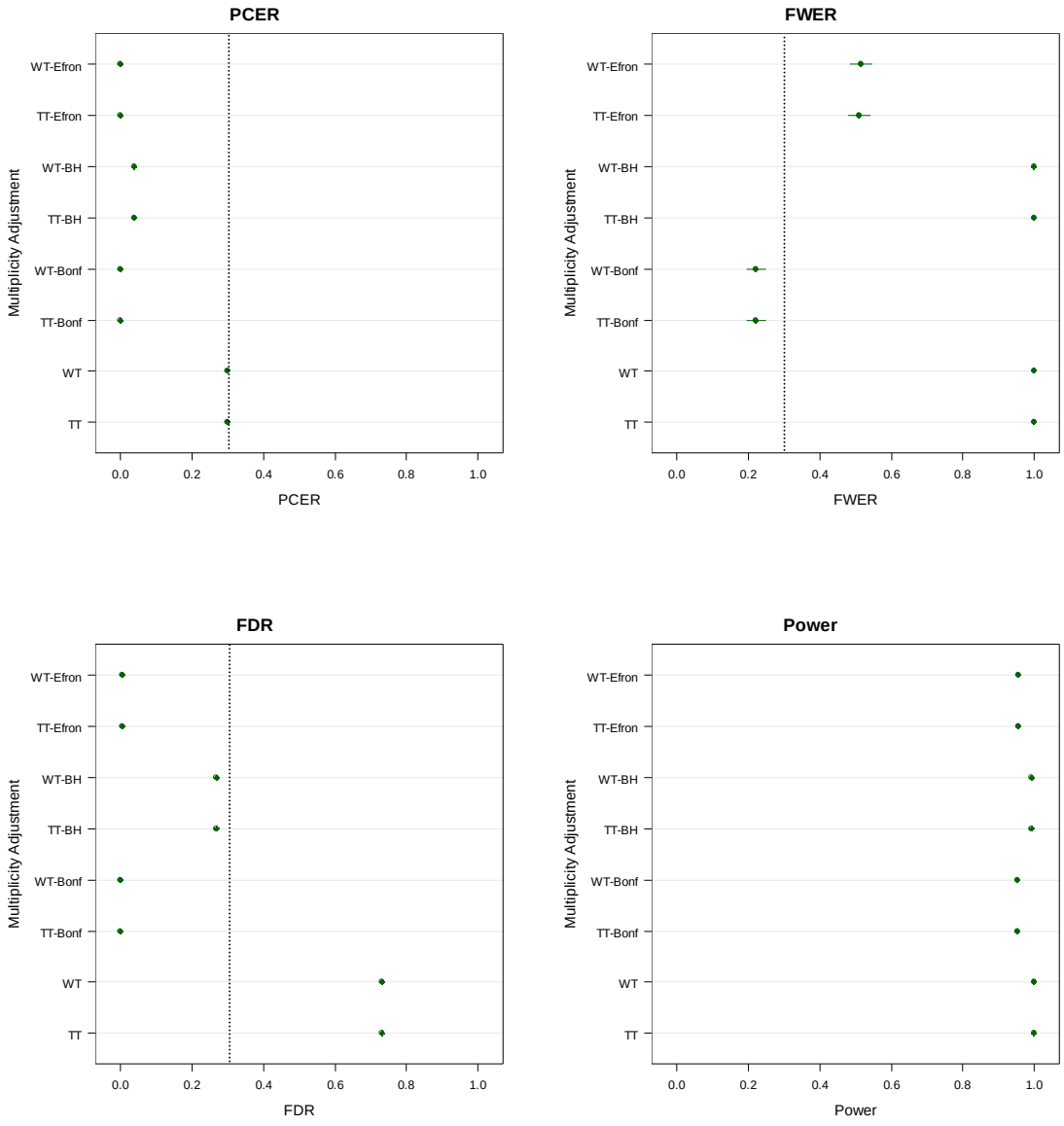

Figure G-29. Simulation Results (0.30); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.30.

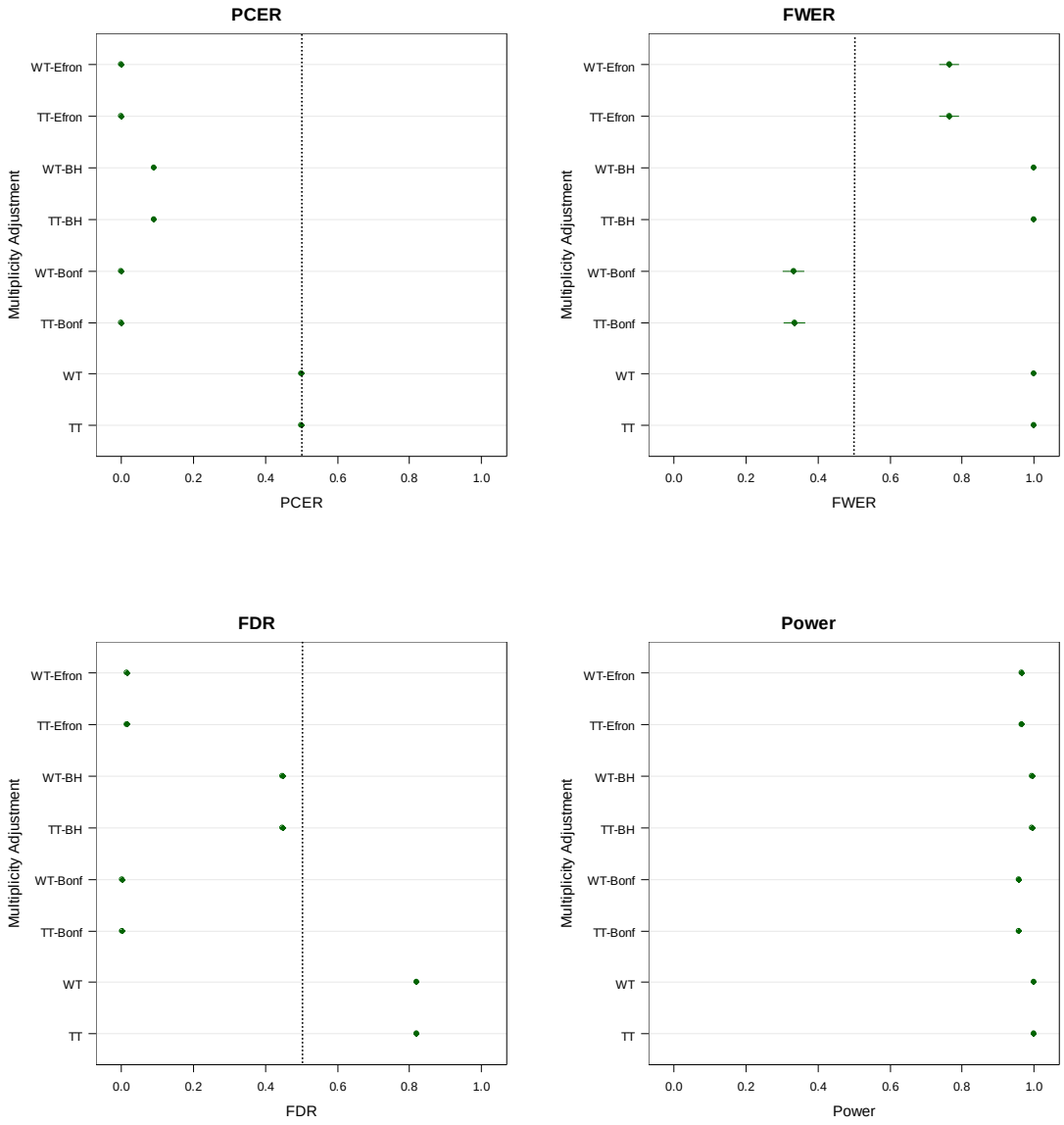

Figure G-30. Simulation Results (0.50); Error and Power Summary Comparing Eight Methods of Adjustment for Multiplicity. Scenario: 2,000 simulated genes (1/2 were correlated in groups of 10), 10% differentially expressed, 100 individuals in each group. In this summary, genes were declared differentially expressed if the adjusted P-value (or ASL) was below 0.50.

VITA

Eric Poole Hintze
329 North 1000 East, Kaysville, Utah, 84037

Educational Background

Ph.D., Statistics, Texas A&M University, College Station, Texas, 2005

M.S., Statistics, Brigham Young University, Provo, Utah, 2000

B.S., Mathematics, Brigham Young University, Provo, Utah, 1997