Genomics, Prior Probability, and Statistical Tests of Multiple Hypotheses

Kenneth F. Manly, Dan Nettleton and J.T. Gene Hwang

Genome Res. 2004 14: 997-1001
Access the most recent version at doi:10.1101/gr.2156804

Genomics, Prior Probability, and Statistical Tests of Multiple Hypotheses

Kenneth F. Manly,[1,4] Dan Nettleton,[2] and J.T. Gene Hwang[3]

[1]Department of Molecular and Cellular Biology, Roswell Park Cancer Institute, Buffalo, New York 14263, USA; [2]Department of Statistics, Iowa State University, Ames, Iowa 50011, USA; [3]Departments of Mathematics and Statistical Science, Cornell University, Ithaca, New York 14853, USA

Genomic methods have made statistical multiple-test methods important to geneticists and molecular biologists. These tests apply to identification of quantitative trait loci and measurement of changes in RNA or DNA abundance by microarray methods. Recently developed multiple-test methods provide more statistical power when many of the tested null hypotheses are false. At the same time, these methods can provide stringent control of errors in cases when most or all of the tested null hypotheses are true. These methods control errors in a different way from previous hypothesis tests, controlling or estimating quantities called the posterior error rate (PER), false discovery rate (FDR), or proportion of false positives (PFP), rather than the type I error. In this study, we attempt to clarify the relationships among these methods and demonstrate how the proportion of true null hypotheses among all tested hypotheses plays an important role.

Genomic methods, those that evaluate many genes or many genomic locations for some property, often require testing a large set of statistical hypotheses, called a family of hypotheses. Such a family may include thousands of hypotheses. For example, detection of quantitative trait loci involves testing a statistical association between trait values and genotypes at several hundred marker loci (Lander and Botstein 1989). Microarray analysis of RNA expression may involve looking for changes among thousands of RNA species (Lockhart et al. 1996). Combining the two techniques (Jansen and Nap 2001; Brem et al. 2002; Schadt et al. 2003), tests pairwise associations between thousands of RNA expression patterns and genotypes at hundreds of marker loci.

Naive application of standard hypothesis tests with no adjustment for multiple testing will yield large numbers of nonreproducible positive results or false discoveries (Soriç 1989). On the other hand, using multiple testing methods to control the familywise type I error rate (FWER, see below) can greatly reduce the power to detect discoveries in families of tests where many such cases should be detected. In this study, we examine the relationship between the traditional type I error and the other criteria that seem more useful for genetics hypothesis testing. The issue is essentially that faced by Morton (1955) when he proposed that an LOD score of 3.0 be required to declare linkage of genetic loci in humans.

Formal statistical hypothesis tests provide a standard method for interpreting experimental data. They contrast a null hypothesis, H_0, and an alternative hypothesis, H_1. The null hypothesis H_0 is chosen so that the probability of any experimental outcome can be calculated assuming H_0 to be true. If under H_0 the probability of observing results as extreme or more extreme than the observed results from an experiment is less than some desired value α, H_0 is rejected and H_1 is accepted. The value α is the desired type I error rate, or the probability of rejecting H_0 when H_0 is true. This value may also be called the comparisonwise type I error rate (CWER) when it refers to the rate for a single test in a family of tests. Suppose now that there is a family of m tests and that the null hypothesis is true for m_o of the tests and false for $m_1 = m - m_0$ of them. Table 1 summarizes the possible outcomes for this family of m tests. Each test yields a value x for a statistic X and a p-value, which is $P(X \geq x|H_0)$, the probability that X would match or exceed the observed value x under the assumption that the null hypothesis is true. For m_0 of the tests, those for which the null hypothesis is true, the p-values are uniformly distributed. For the other m_1 tests, the distribution of p-values will be stochastically smaller than a uniform distribution (i.e., $P(p \leq x|H_1) \geq x = P(p \leq x|H_0)$ for any x between 0 and 1). The ratios $\pi_o = m_o/m$ and $\pi_1 = m_1/m$, when m is large, can be interpreted as approximate Bayesian prior probabilities of the null and alternative hypotheses, respectively.

The terms "discovery" (Soriç 1989) or "positive result" are often used to refer to a hypothesis test in which the null hypothesis is rejected. The terms "false discovery" or "false positive" are commonly used to describe the case in which the null hypothesis is rejected, although it is, in fact, true. The power of a test is the probability that a positive result will be obtained when the null hypothesis is false. Tests with high power will tend to produce high values of S in Table 1, whereas tests with low power may produce high values of T in Table 1.

The familywise error rate (FWER), also known as the overall type I error rate, is the probability of one or more type I errors in a family of tests. In terms of the definitions in Table 1, the FWER is simply P(V > 0). Much of the past research in the area of multiple testing has focused on the development of methods that control this probability. Strong control of the FWER at level α is achieved if the FWER is less than or equal to α, regardless of the number of false null hypotheses (m_1). Weak control is obtained if the FWER is less than or equal to α whenever all tested null hypotheses are true ($m_0 = m$) and not necessarily less than or equal to α when some of the null hypotheses are false ($m_1 > 0$).

Controlling the FWER is important for tests in which the family is being tested as a unit, and the rejection of any null hypothesis affects the whole family. More generally, control of the FWER is important whenever it is necessary for an analysis to produce no false positives with high probability. When a family contains many tests, producing no false positives with high probability may require a substantial degree of conservativeness that could lead to many type II errors (i.e., large values of T in Table 1). Suppose, for example, that a method used to test each of the m genes for differential expression in a microarray experiment correctly rejects 99 false null hypotheses and incorrectly rejects one true null hypothesis (S = 99, V = 1, R = 100). From the standpoint of FWER error control, the performance of the method on this one data set would be considered in error because of the one false positive result. Such an error would be allowed to occur in only 5% of the experiments to which the method would be

[4]Corresponding author.
E-MAIL Kmanly@Tennessee.edu; FAX (901) 448-7193.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.2156804.

Table 1. Classification of Outcomes Among a Family of *m* Statistical Tests, for m_0 of Which the Null Hypothesis Is True

	Accept null No discovery Negative result	Reject null Declare discovery Positive result	Total
True null hypothesis	U	V	m_0
False null hypothesis	T	S	m_1
Total	W	R	m

Column headings give three alternative phrases with the same meaning. The letters R through W represent the number of test that fall into each category. For example, R is the number of tests that are declared significant, rejecting the null hypothesis. The V tests that reject the null when it is true commit type I errors. The T tests that accept the null when it is false make Type II errors.

applied if control of the FWER at the 5% level was required. However, rather than considering such a performance a failure to be avoided 95% of the time, a researcher may be quite happy with a method that routinely produced a list of 100 differentially expressed genes that contained only one false positive and, perhaps, would be willing to tolerate even more false positives in such analyses if this tolerance would permit more true discoveries while maintaining a low ratio of false positive to positive results (V/R).

Alternative Error Measures

In this study, we consider alternatives to control the FWER that will tolerate more false positive results in exchange for greater levels of true discovery. The error measures corresponding to these alternative methods are closely related to the concept of posterior error rate (PER) proposed by Morton (1955). Morton defined the PER for a single test as the probability of the null hypothesis being true, given that the test resulted in the rejection of the null hypothesis. The PER depends on the prior probability that the null hypothesis is true. The prior probability of the null hypothesis being true is simply m_0/m if a null hypothesis is randomly selected from a family of m null hypotheses, of which m_0 are true. Fernando et al. (2004) defined the proportion of false positives (PFP) as an error measure that is equivalent to the PER in the sense that the PFP for a family of m tests is equal to the PER for a test that is randomly selected from a family of m tests. Storey (2002; 2003) defined the positive false discovery rate (pFDR) and described situations in which the pFDR is equivalent to the PER. Storey's work is closely related to Benjamini and Hochberg's (1995) landmark paper on the false discovery rate (FDR). This error measure, along with the pFDR, PER, and PFP are defined formally in Table 2.

When a test yields a discovery, an experimental scientist would like to know that the discovery is repeatable; that is, that it is not a false discovery. Standard hypothesis testing controls the probability of a type I error, but type I error control may not lead to a suitably low PER, a situation known as the "screening paradox". For example, suppose we screen for some condition in a population for which the frequency of the condition is 1 in 10,000, screening with a test that yields 1% false positives and a negligible number of false negatives. That is, in a single test, the test has a type I error rate of 0.01 and a type II error rate near 0. Using this test on 10,000 individuals, we would make, on average, around 100 false discoveries and probably one true discovery. The PER would be >99%; that is, almost all discoveries would not be repeatable. Although the true and false discoveries can be distinguished by repeated testing, a test with a type I error rate of 1% is only useful in populations where the condition to be detected is itself much more frequent than 1%.

Users of statistical tests often assume that the type I error rate of a test and the PER are the same, or at least, that a low type I error rate implies a low PER. That is, they assume that if a result is declared significant at a type I error rate of 5%, then the PER is about 5% and there is a 95% chance that the result is repeatable. In general, this assumption is false, as demonstrated in the previous paragraph. But if the probability of a discovery π_1 is high enough, it is almost true. Standard hypothesis tests provide an acceptably low PER, because scientists intuitively choose experiments that are likely to "work", that is, experiments that have a moderately high probability of the null hypothesis being false (K.F. Manly is indebted to N.J. Schork for this insight). For these experiments, a low type I error rate implies a low PER.

For a single test, we obtain below the relationship between PER and π_1 by combining parameters of standard hypothesis testing with Bayes' theorem. A standard hypothesis test is characterized by two parameters, the type I error rate α, and the type II error rate, β, which is the probability failing to reject the null hypothesis when it is false. If these parameters are combined with Bayes' theorem, we obtain a relationship comparable to that described by Morton (Morton 1955; Kurhekar et al. 2002).

$$PER = \frac{1}{1 + \frac{(1-\beta)\pi_1}{\alpha(1-\pi_1)}}$$

If we plot this relationship for typical values of α (0.05 or 0.001) and β (0.2), we obtain the relationship shown in Figure 1. This figure shows PER as a function of the π_1/α ratio, because this ratio almost completely determines the PER when $\pi_1 \ll 1$ and $\beta \ll 1$.

From this, we can see that a standard hypothesis test has an acceptably low PER only if π_1 is well above α. Specifically, the PER will be acceptable (<20%) only if α is chosen to be smaller than approximately $\pi_1/5$. If we observe a discovery at the usual significance level of $\alpha = 0.05$, we can count on that result being repeatable only if the result was already moderately likely before the experiment. The minimum factor of π_1/α needed to achieve PER <20% is $4/(1 - \beta + 4\alpha)$. This factor is not greatly sensitive to changes in α and β; for $0.001<\alpha<0.05$ and $0.1<\beta<0.4$, the minimum factor π_1/α varies between 3.6 and 6.6.

How can we assure that α is sufficiently small relative to π_1 to obtain low PER when π_1 is unknown? In many cases, π_1 is not completely unknown. Theory or prior knowledge often provides some clue as to the magnitude of π_1. For example, if a trait differs in two lines of inbred mice and segregates in crosses as a Mendelian factor, it is quite likely that we can identify some gene that controls a part of that trait. Because the trait must be linked to some gene, we can calculate the probability (≈0.05; Morton 1955) that the gene will be linked to an arbitrarily chosen marker locus.

Table 2. Definitions of Error Rates Related to False Discovery Rate and Posterior Error Rate

Abbreviation	Name	Definition	Reference
FDR	False discovery rate	$E\left(\frac{V}{R} \mid R>0\right)P(R>0)$	(Benjamini and Hochberg 1995)
pFDR	Positive false discovery rate	$E\left(\frac{V}{R} \mid R>0\right)$	(Storey 2002)
PER	Posterior error rate	$P(V=1 \mid R=m=1)$	(Morton 1955)
PFP	Proportion of false positives	$\frac{E(V)}{E(R)}$	(Fernando et al. 2004)

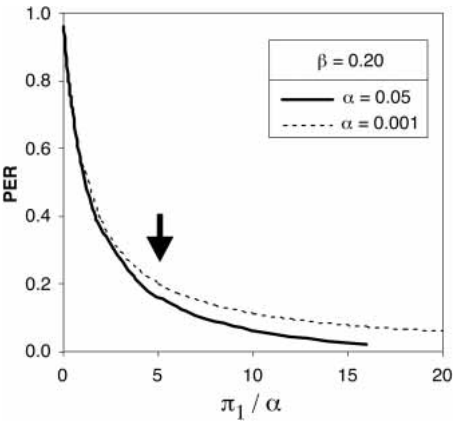

Figure 1 Posterior error rate for an experiment in which the null hypothesis is rejected, as a function of the ratio of π_1 to significance level α, for $\beta = 0.2$, and two significance levels as shown. As discussed in the text, π_1 is the fraction of tests in a family for which the null hypothesis is actually false. The arrow shows the approximate minimum ratio to achieve a PER of no more than 0.2.

Even without prior information, Benjamini and Hochberg (2000), Mosig et al. (2001), Storey (2002), Storey and Tibshirani (2001), and Allison et al. (2002) describe methods for obtaining information about π_1 from the observed data.

Multiple-Test Methods for Controlling the FWER

As already mentioned, several tests have been developed for situations in which a family of hypotheses are to be tested. One of the oldest and best known is the Bonferroni correction. This correction modifies a standard hypothesis test by controlling the false positive rate more stringently. For a family of m tests, Bonferroni specifies controlling the CWER for each test at α/m. This results in FWER $\leq \alpha$ in the strong sense. This method is effective even for situations in which π_1 is relatively low, that is, for families of tests in which almost no discoveries are expected. The Bonferroni correction will assure a satisfactory PER under these conditions. By extension of the argument presented above, it will assure a satisfactory PER, even if π_1 is as low as $4\alpha/[m(1-\beta)+4\alpha]$ for the family of tests. The Bonferroni correction is also useful in cases where even one false positive would be troublesome. However, the Bonferroni correction is not well suited for cases in which π_1 is high enough that several discoveries can be expected, and in which a minority of false discoveries can be tolerated. In these cases, the Bonferroni correction suffers because it has low power, that is, because it results in a large number of type II errors.

Methods for Controlling the FDR

Several methods have been described as improvements to the Bonferroni correction (Holm 1979; Simes 1986; Hochberg 1988; Hommel 1988). These methods, although they differ in detail, are all sequential procedures in which the hypotheses are tested in ascending or descending order of p-value. Details are explained in the Appendix.

Like the Bonferroni procedure, the methods of Holm, Hochberg, and Hommel tend to result in satisfactory PER. However, despite their enhanced power relative to the Bonferroni procedure, these methods often still suffer from a lack of power when π_1 is high. Benjamini and Hochberg (1995) were the first to develop a powerful method with reasonably good PER properties across a wide range of conditions. The Benjamini and Hochberg procedure rejects the hypotheses corresponding to the smallest k p-values ($p_{(1)}, \ldots, p_{(k)}$) whenever $mp_{(k)}/k \leq \alpha$. Simes (1986) had previously shown that this procedure provides weak control of the FWER under general conditions and suggested that the criterion might be used as an exploratory tool. Benjamini and Hochberg were the first to prove that this procedure controls the FDR at level α.

Note that the numerator left of the inequality $mp_{(k)}/k \leq \alpha$ is simply an estimate of the expected number of false discoveries that would result if all null hypotheses are true and a null is rejected whenever its p-value is no larger than $p_{(k)}$. The denominator left of the inequality is simply the actual number of discoveries that result for the observed data when a null is rejected whenever its p-value is no larger than $p_{(k)}$. Thus, $mp_{(k)}/k$ is an estimate of the proportion of false discoveries among all discoveries, and the Benjamini and Hochberg method attempts to reject as many hypotheses as possible, subject to the constraint that this estimated false discovery rate is no larger than α. By insisting that this estimated false discovery rate be no larger than α, the Benjamini and Hochberg procedure results in a testing procedure with generally low PER.

Note that the quantity $mp_{(k)}$ in the numerator of the Benjamini and Hochberg criterion will tend to overestimate the expected number of false discoveries when rejecting null hypotheses with p-values less than or equal to $p_{(k)}$, unless all null hypotheses are true ($m = m_0$). Benjamini and Hochberg (1995) recognized this and proved that their procedure is conservative in the sense that it controls FDR at (m_0/m) times the nominal rate. In general, it would be more appropriate to use $m_0 p_{(k)}$ in the numerator of the Benjamini and Hochberg criterion, but the challenge, of course, is that m_0 is unknown. The multiple testing procedures proposed by Benjamini and Hochberg (2000), Mosig et al. (2001), Storey and Tibshirani (2003), and Fernando et al. (2004) reject the null hypotheses corresponding to the k smallest p-values, where k is the largest i, such that $\hat{m}_{0P(i)}/i \leq \alpha$ and $\hat{m}_0$ is an estimate of m_0 obtained from the empirical distribution of observed p-values. Thus, these methods differ from the original FDR controlling procedure proposed by Benjamini and Hochberg (1995), only in that m is replaced by an estimate of m_0. The procedures differ from each other only in the method used to estimate m_0. By extracting information about m_0, or equivalently, π_1 from the observed data, these procedures achieve greater power than the original procedure developed by Benjamini and Hochberg.

Illustration of the Methods

Let us suppose that we test a family of 500 cases for which the null hypothesis is always true ($\pi_1 = 0$). The p-values for these cases will be uniformly distributed between 0 and 1. If these p-values are sorted and each plotted against its rank in the sort (Schweder and Spjøtvoll 1982), the p-values will tend to form a straight line as shown in Figure 2A. If, on the other hand, we test a family of 500 cases for which 80 null hypotheses are false ($\pi_1 = 0.16$), this family will produce 420 uniformly distributed p-values mixed with 80 that tend to be smaller. When these cases are plotted in the same way, the line formed by the points will be curved or bent, as shown in Figure 2B.

The multiple-test methods mentioned in the previous sections each establish a criterion for significance that can be represented by a threshold line or curve like those in Figure 2. Points that fall below each depicted threshold define a group of cases that can be declared significant. The methods differ in detail as to the position and shape of the threshold as well as how the group is defined. According to the Hochberg (1988) test, for example, the rightmost p-value below the curve, and all smaller p-values can be declared significant at a specified FWER. Figure 2 shows the Holm, Hochberg, and Hommel thresholds for FWER control at 20%, along with the Benjamini and Hochberg threshold and the Storey and Tibshirani (2003) threshold for FDR control at 20%. When all 500 null hypotheses were true, none of the methods rejected any null hypotheses, because all points fell above the significance thresholds (not discernable at the scale of Figure 2A). When 80 null hypotheses were false and 420 null hypotheses were true (Fig. 2B,C,D), the Holm, Hochberg, and Hommel methods each declared 11 discoveries, the Benjamini and Hochberg method rejected declared 58 discoveries, and the Storey and Tibshirani method declared 65 discoveries. The Holm, Hochberg, and Hommel methods made no type I errors in this case, but the conservativeness required for FWER control led to 69 type II errors. The Benjamini and Hochberg procedure committed seven type I errors for an observed false discovery rate of ~12%. The Storey and Tibshirani procedure committed

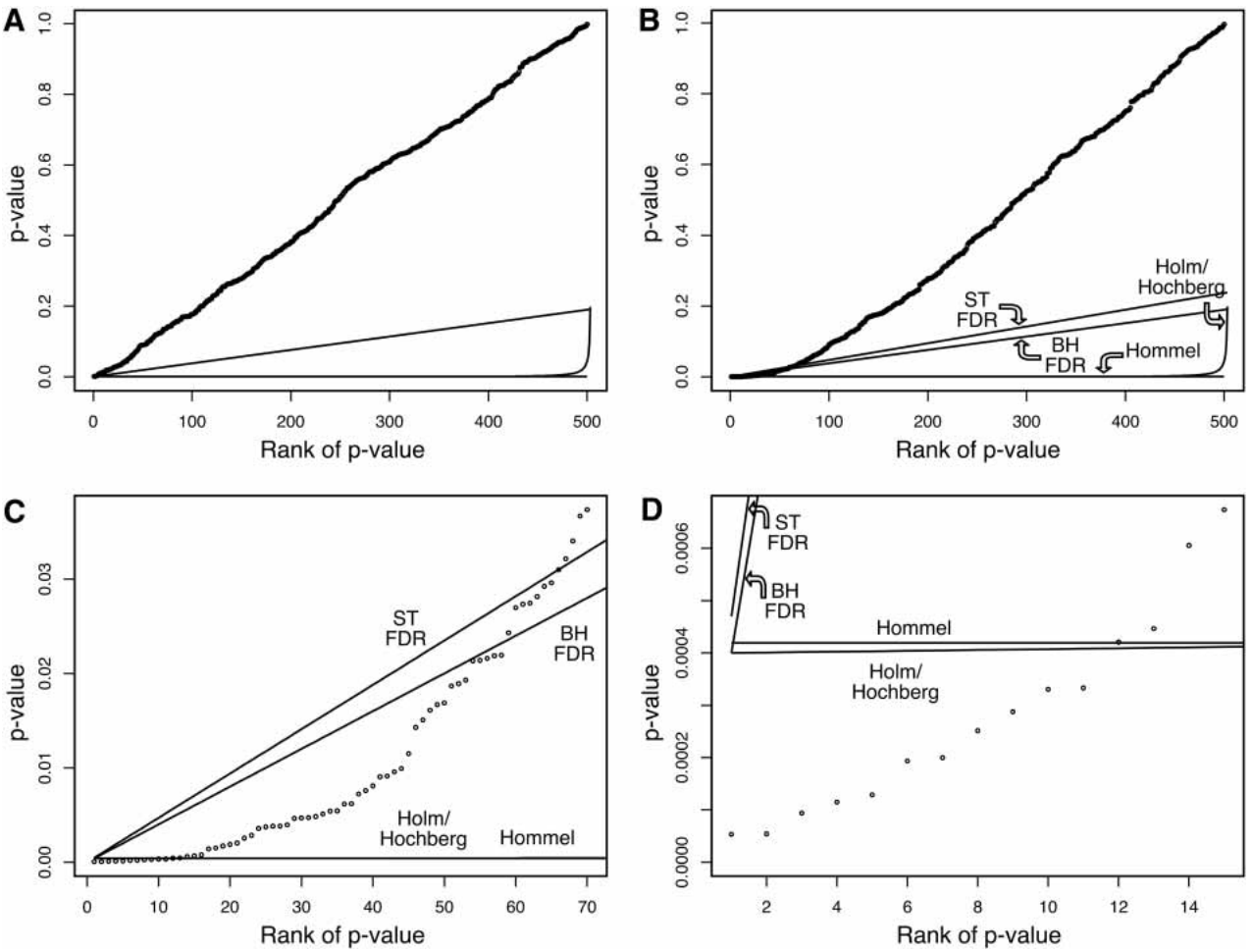

Figure 2 Comparison of multiple-test methods. (*A*) *p*-values from family of tests in which all null hypotheses are true. The Benjamini and Hochberg and Storey and Tibshirani thresholds for FDR control at 20% coincide and are represented by the line with positive slope. The Holm/Hochberg significance threshold for FWER control begins below and then rises dramatically above the Hommel threshold for FWER control at 20% (difference only discernable on the far *right* of the plot). (*B*) *p*-values from a family of tests in which 80 null hypotheses are false; (*C*) data from *B* at expanded scale; (*D*) data from *B* and *C* at expanded scale.

nine type I errors for an observed false discovery rate of ~14%, well below the nominal 20%.

Performance of the Bonferroni, Holm, Hochberg, Hommel, Benjamini and Hochberg, and Storey and Tibshirani methods was tracked over 1000 replications of the scenario depicted in Figure 2B,C,D. The mean number of type I and type II errors along with the mean observed FDR are reported in Table 3. Also included are estimates of the FWER, given by the proportion of the 1000 replications, in which one or more type I errors were committed, and estimates of the PER, given by the proportion of false positives among all positive results observed over all 500,000 tests. The standard error of each estimated mean or proportion is provided after the ± symbol.

DISCUSSION

The example and small simulation study in the previous section illustrate some general concepts regarding multiple testing methods. First, Bonferroni and related methods that attempt to control FWER will tend to achieve low PER due to the small CWER used for individual tests. The low PER, however, comes at the cost of power; the FWER-controlling procedures will tend to make many type II errors when the number of tests m is large and the proportion of true alternative hypotheses π_1 is high. Although the modifications to the Bonferroni procedure do provide some improvements in power, the performance of the modified methods will often be quite similar to the performance of the Bonferroni method. When the number of tests m is large and the proportion of true alternative hypotheses π_1 is high, the FDR-controlling methods of Benjamini and Hochberg (2000) and Storey and Tibshirani (2003) will tend to achieve PER values near their target FDR levels, while at the same time permitting far greater discovery than the FWER-controlling procedures. Thus, FDR-controlling methods

Table 3. Comparison of Multiple-Test Methods by Simulation

Method	Type I Errors	Type II Errors	Observed FDR	Observed PER	Observed FWER
Bonferroni	0.171 ± 0.01	68.4 ± 0.1	0.014 ± 0.001	0.012 ± 0.001	0.154 ± 0.01
Holm	0.174 ± 0.01	68.2 ± 0.1	0.014 ± 0.001	0.012 ± 0.001	0.157 ± 0.01
Hochberg	0.174 ± 0.01	68.2 ± 0.1	0.014 ± 0.001	0.012 ± 0.001	0.157 ± 0.01
Hommel	0.179 ± 0.01	68.0 ± 0.1	0.014 ± 0.001	0.012 ± 0.001	0.162 ± 0.01
Benjamin and Hochberg	12.1 ± 0.14	21.0 ± 0.1	0.167 ± 0.002	0.174 ± 0.001	1.000
Storey and Tibshirani	16.3 ± 0.20	17.9 ± 0.1	0.202 ± 0.002	0.215 ± 0.001	1.000

are recommended for exploratory genomics experiments in which a specified proportion of false positive results among all positive results can be tolerated. In applications where no type I errors can be tolerated, an FWER-controlling method should be used, because the FDR-controlling procedures may result in one or more false positive results with high probability. (Note that the FWER for the FDR-controlling methods was estimated to be 1.0 for the small simulation study of the previous section.)

We presented results for 500 tests with FWER and FDR control at 20%, primarily to make Figure 2 easier to read. Many genomics experiments will involve far more than 500 tests, and control of FWER or FDR at lower error rates is often desired. The general performance characteristics illustrated in our example and simulation carry over to smaller error rates and larger numbers of tests. The choice of magnitude of the error rate is best left to individual researchers to determine on the basis of the cost of false positives and false negatives in the situation at hand.

We focused on the FDR-controlling procedures of Benjamini and Hochberg (1995) and Storey and Tibshirani (2003) because these procedures have received a more thorough treatment in the statistics literature than other procedures designed to control or estimate quantities related to FDR. Storey, in particular, has published several results on the relationship between these two leading methods for FDR-control and has established several formal properties of his FDR-controlling procedures (see Storey 2002, 2003; Storey et al. 2004). Storey (2002) presents simulations showing that his procedure can achieve power of up to eight times that of the Benjamini and Hochberg (1995) procedure, but this extreme advantage occurs only when π_1 is well above 0.5.

Summary

In a single test, PER is affected by the prior probability π_1 of a discovery, and if that probability is low compared with the type I error rate, the PER will be unacceptably high. Because the PER provides an indication of whether an experimental result will be repeatable, it is often as important to an experimental scientist as the type I error rate. Past work on the problem of multiple testing has focused on control of the FWER. Although methods that control the FWER will have generally low PER, the procedures are more conservative than necessary for exploratory genomics studies. New multiple testing procedures that attempt to control or estimate error measures PER, PFP, FDR, or pFDR tend to achieve reasonable PER levels without unduly sacrificing the power to discover.

APPENDIX

Holm's procedure (Holm 1979) rejects the null hypotheses corresponding to the k smallest p-values $(p_{(1)},\ldots_{(k)})$ if $p_{(i)} \leq \alpha/(m-i+1)$ for all $i \leq k$. Note that the smallest p-value is evaluated with a stringency equivalent to that of the Bonferroni correction, but larger p-values are tested with less stringent criteria. Holm's method provides strong control of the FWER in all circumstances and, due to its enhanced power when multiple null hypotheses are false, should be preferred to the Bonferroni procedure in all situations.

The method of Hochberg (1988) rejects the null hypotheses corresponding to the k smallest p-values as long as $p_{(k)} \leq \alpha/(m-k+1)$. The Hochberg procedure is clearly more powerful than Holm's procedure, because the hypothesis corresponding to $p_{(i)}$ can be rejected even if $p_{(i)} > \alpha/(m-i+1)$ as long as $p_{(k)} \leq \alpha/(m-k+1)$ for some $k > i$. Strong control of the FWER, however, is guaranteed for Hochberg (1988) only when the test statistics used to test the family of hypotheses are independent (although simulations in Simes [1986] suggest the strong control holds for Hochberg [1988] in more general situations). The same basic error controlling properties are shared by the Hommel procedure (Hommel 1988), which rejects all hypotheses whose p-values are $\leq \alpha/k$, where k is defined as the largest value of i satisfying $p_{(n-i+j)} > j\alpha/i$ for all $j = 1,\ldots/$. If no such value of i exists, all null hypotheses in the family are rejected. Although the Hommel procedure is more complex than Hochberg's procedure, Hommel (1989) showed that it is also more powerful.

ACKNOWLEDGMENTS

This work was supported by grants P41-HG01656 from the National Human Genome Research Institute and P20-MH62009 from the Human Brain Project (funded jointly by the National Institute of Mental Health, the National Institute on Drug Abuse, and the National Science Foundation), and the Cooperative State Research, Education, and Extension Service, U.S. Department of Agriculture, under Agreement No. 2002-35300-12619.

REFERENCES

Allison, D.B., Gadbury, G.L., Moonseong, H., Fernandez, J.R., Lee, C.-K., Prolla, T.A., and Weindruch, R. 2002. A mixture model approach for the analysis of microarray gene expression data. *Computat. Stat. Data Anal.* **39:** 1–20.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Royal Stat. Soc. B* **57:** 289–300.

———. 2000. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educa. Behav. Stat.* **25:** 60–83.

Brem, R.B., Yvert, G., Clinton, R., and Kruglyak, L. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296:** 752–755.

Fernando, R.L., Nettleton, D., Southey, B.R., Dekkers, J.C.M., Rothschild, M.F., and Soller, M. 2004. Controlling the proportion of false positives (PFP) in multiple dependent tests. *Genetics* **166:** 611–619.

Hochberg, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75:** 800–802.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6:** 65–70.

Hommel, G. 1988. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75:** 383–386.

———1989. A comparison of two modified Bonferonni procedures. *Biometrika* **76:** 624–625.

Jansen, R.C. and Nap, J.P. 2001. Genetical genomics: The added value from segregation. *Trends Genet.* **17:** 388–391.

Kurhekar, M.P., Adak, S., Jhunjhunwala, S., and Raghupathy, K. 2002. Genome-wide pathway analysis and visualization using gene expression data. *Pac. Symp. Biocomput.* 462–473.

Lander, E.S. and Botstein, D. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121:** 185–199.

Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., et al. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14:** 1675–1680.

Morton, N.E. 1955. Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* **7:** 277–318.

Mosig, M.O., Lipkin, E., Khutoreskaya, G., Tchourzyna, E., Soller, M., and Friedmann, A. 2001. A whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion. *Genetics* **157:** 1683–1698.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., et al. 2003. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422:** 297–302.

Schweder, T. and Spjøtvoll, E. 1982. Plots of P-values to evaluate many tests simultaneously. *Biometrika* **69:** 493–502.

Simes, R. 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73:** 751–754.

Soriç, B. 1989. Statistical "discoveries" and effect size estimation. *J. Am. Stat. Assoc.* **84:** 608–610.

Storey, J.D. 2002. A direct approach to false discovery rates. *J. Royal Stat. Soc. B* **64:** 479–498.

———. 2003. The positive false discovery rate: A Bayesian interpretation and the Q-value. *Ann. Statistics* **31:** 2013–2035.

Storey, J.D. and Tibshirani, R. 2001. *Estimating false discovery rates under dependence, with application to DNA microarrays.* Department of Statistics, Stanford University, Stanford, CT.

———. 2003. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* **100:** 9440–9445.

Storey, J.D., Taylor, J., and Siegmund, D. 2004. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *J. Royal Stat. Soc. Series B* **66:** 184–205.