Controlling the Proportion of False Positives in Multiple Dependent Tests

R. L. Fernando,*,†,[1] D. Nettleton,†,‡ B. R. Southey,§ J. C. M. Dekkers,*,†
M. F. Rothschild*,† and M. Soller**

**Department of Animal Science, †Lawrence H. Baker Center for Bioinformatics and Biological Statistics and ‡Department of Statistics, Iowa State University, Ames, Iowa 50011, §Department of Animal Sciences, University of Illinois, Urbana, Illinois 61801 and **Department of Genetics, Hebrew University, Jerusalem 91904, Israel*

Manuscript received February 8, 2003
Accepted for publication October 1, 2003

ABSTRACT

Genome scan mapping experiments involve multiple tests of significance. Thus, controlling the error rate in such experiments is important. Simple extension of classical concepts results in attempts to control the genomewise error rate (GWER), *i.e.*, the probability of even a single false positive among all tests. This results in very stringent comparisonwise error rates (CWER) and, consequently, low experimental power. We here present an approach based on controlling the proportion of false positives (PFP) among all positive test results. The CWER needed to attain a desired PFP level does not depend on the correlation among the tests or on the number of tests as in other approaches. To estimate the PFP it is necessary to estimate the proportion of true null hypotheses. Here we show how this can be estimated directly from experimental results. The PFP approach is similar to the false discovery rate (FDR) and positive false discovery rate (pFDR) approaches. For a fixed CWER, we have estimated PFP, FDR, pFDR, and GWER through simulation under a variety of models to illustrate practical and philosophical similarities and differences among the methods.

IN recent years a relatively new class of "multiple-test" genetic experiments has come into prominence, in which there is a strong prior assumption that a certain proportion of the tested alternative hypotheses are true. Consider, for example, a genome-wide scan for linkage between a marker and a quantitative trait locus (QTL). In this situation, when heritability analysis shows that QTL are segregating in the population, the large number and close spacing of the markers employed ensures that an appreciable proportion of markers are in linkage to segregating QTL. The challenge is to identify these markers among all of the tested markers. Similarly, prior marker-QTL linkage mapping in a particular population may have identified a set of markers in linkage to segregating QTL. For purposes of marker-assisted selection, it is important to identify individuals heterozygous at these QTL. On Hardy-Weinberg assumptions, over a wide range of QTL allele frequencies one-third to one-half of the QTL will be heterozygous in any given individual. Thus, the experiment to identify the markers in linkage to heterozygous QTL in a particular individual starts with the strong prior assumption that a comparable proportion of the markers tested are indeed in such a state. Again, the challenge is to identify the individual-by-marker combinations for which this is true, among all tested individual-by-marker combinations. In many microarray experiments, treatments that cause physiological changes are administered to experimental units. One main goal of such experiments is to identify which of thousands of genes change expression as a result of treatment. Treatments are often designed to alter the expression of particular genes, so it is reasonable to assume that some measurable changes in gene expression occur.

Clearly in these examples, identification of a marker in linkage to a QTL, identification of an individual-by-marker combination that represents a heterozygous QTL, or identification of differentially expressed genes, there is the possibility of false-positive error. Controlling this error is important scientifically to avoid cluttering the literature with false results and, practically, to avoid expenditure of effort on false leads to genetic improvement or gene cloning.

One of the most widely used approaches to control errors in multiple tests is based on controlling the familywise type I error rate (FWER). The FWER is the probability of rejecting one or more true null hypotheses in a family of tests. In genome scans for QTL, it has been proposed that the family of tests should be defined as the set of all possible tests across the entire genome, thus controlling the genomewise type I error (GWER; Lander and Kruglyak 1995). The drawback of this approach is the drastic loss of power.

An alternative to attempting to avoid all false-positive results is to manage the accumulation of false positives relative to the total number of positive results that appear in the literature. Indeed, this is the approach that

[1] *Corresponding author:* Department of Animal Science, Iowa State University, 225 Kildee Hall, Ames, IA 50011-3150.
E-mail: rohan@iastate.edu

was traditionally taken in human genetics, where it was early realized that for a monogenic trait, if a comparisonwise type I error rate (CWER) of 0.05 is used as the threshold for declaring linkage, a large proportion of declared linkages would be false. Instead, in human linkage analysis error control has been based on controlling the posterior type I error rate (PER), which is the probability of nonlinkage between two loci given that linkage was declared between these two loci (Morton 1955). By definition, this has the above property of controlling the accumulation of false positives relative to the total number of positive results. Although originally defined for the single-test situation, the PER has also been discussed in a multiple-test situation (Risch 1991), where evenly spaced markers spanning the entire genome were sequentially tested for linkage to a single-trait locus. Assuming that the tests were independent, Risch (1991) computed the posterior type I error rate given that linkage was declared after k_s tests. When a constant threshold was used for declaring linkage, the posterior type I error rate decreased as k_s increased (Risch 1991).

In QTL scans, testing does not stop when one of the markers is declared to be linked to a QTL; all markers are tested for linkage to QTL. Further, with the increased availability of closely spaced markers, tests cannot be considered to be independent. Thus, to extend the philosophy underlying the posterior type I error rate to QTL scans, Southey and Fernando (1998) defined the proportion of false positives (PFP) as a generalization of the PER to the genome scan situation. As is shown in subsequent sections of this article, the PFP effectively controls the accumulation of false positives relative to the total number of positive results. In addition, the PFP level for a set of tests does not depend on the number of tests or the correlation structure among the tests. This makes the PFP of particular usefulness in QTL mapping applications that often involve a large number of tests with a complex correlation structure.

Another approach that has been used to control the accumulation of false positives in QTL scans is based on controlling the false discovery rate (FDR; Benjamini and Hochberg 1995; Weller 2000; Mosig *et al.* 2001). Mosig *et al.* (2001) argued intuitively that the FDR as defined by Benjamini and Hochberg (1995) is not appropriate when the experiment has a large number of tests for which the null hypothesis is false; they proposed using an adjusted FDR, which takes this factor into account. Although not considered previously in the QTL mapping context, Storey (2002) defined the positive false discovery rate (pFDR) to be more suitable than FDR as a measure of false discoveries. Differences and similarities of these various methods with respect to PFP are discussed in a subsequent section of this article.

Our development of the PFP is general. However, we use simulations within the QTL mapping application to show how PFP compares to FWER, FDR, and pFDR and to illustrate how the estimated PFP levels compare to true PFP levels.

CONNECTION TO POSTERIOR TYPE I ERROR RATE

The philosophy behind the PFP approach is closely connected to the philosophy of the posterior type I error rate approach developed by Morton (1955) for the case of detecting linkage between a single-marker locus and a monogenic trait locus. In this setting, the PER is the conditional probability that the true status between a randomly selected marker locus and the monogeneic trait locus is one of nonlinkage, given a statistical test result interpreted as declaring linkage (Morton 1955). In technical notation, let the true status of linkage between the two loci be represented by a random variable L that can take one of two values, $L = 1$ if the two loci are linked and $L = 0$ if the two loci are not linked; and let the declared status of linkage between the two loci on the basis of some statistical test be represented by a random variable D that can also take one of two values, $D = 1$ if the two loci are declared linked and $D = 0$ if the two loci are declared not linked. Then the PER is $\Pr(L = 0|D = 1)$. Following Morton (1955), this probability can be written as

$$\begin{aligned}\Pr(L = 0|D = 1) &= \frac{\Pr(L = 0, D = 1)}{\Pr(D = 1)}\\ &= \frac{\Pr(L = 0, D = 1)}{\Pr(L = 0, D = 1) + \Pr(L = 1, D = 1)}.\end{aligned} \tag{1}$$

The probabilities required to compute (1) are

$$\begin{aligned}\Pr(L = 0, D = 1) &= \Pr(D = 1|L = 0)\Pr(L = 0)\\ &= \alpha\Pr(L = 0),\end{aligned} \tag{2}$$

and

$$\begin{aligned}\Pr(L = 1, D = 1) &= \Pr(D = 1|L = 1)\Pr(L = 1)\\ &= \pi\Pr(L = 1),\end{aligned} \tag{3}$$

where α is the CWER and π is the average power of the test for markers for which $L = 1$. Using (2) and (3) in (1) gives

$$\text{PER} = \Pr(L = 0|D = 1) = \frac{\alpha\Pr(L = 0)}{\alpha\Pr(L = 0) + \pi\Pr(L = 1)}. \tag{4}$$

For a monogenic trait in humans, the prior probability that a random marker is within detectable linkage of the trait locus is ~0.02 (Elston and Lange 1975; Ott 1991), so that for a random marker, $\Pr(L = 1) = 0.02$. Using a CWER of 0.05 to represent significance would give a PER of 0.73; *i.e.*, of every 100 declared linkages, ~73 would be false. The traditional LOD score of 3 required to declare linkage corresponds to a CWER between 0.0001 and 0.001 (Elston 1997). Taking 0.001 as the critical CWER to declare linkage, and supposing that average power of the test is 0.90, the PER is

$$\Pr(L = 0|D = 1) = \frac{0.001 \times 0.98}{0.001 \times 0.98 + 0.9 \times 0.02} = 0.05.$$

Thus, using this CWER, of every 100 declared linkage associations, ∼5 would be false. Thus, the PER approach indeed controls the proportion of false positives in the literature as intended.

For the case of a genome scan involving a set of k markers, Southey and Fernando (1998) defined the PFP as

$$\mathrm{PFP} = \frac{\sum_{i=1}^{k}\alpha_i \Pr(H_i)}{\sum_{i=1}^{k}[\alpha_i \Pr(H_i) + (1 - \Pr(H_i))\pi_i]}, \tag{5}$$

where for the ith test, α_i is the CWER, π_i is the power, and $\Pr(H_i)$ is the probability that the null hypothesis is true [if the ith marker is linked to a QTL $\Pr(H_i) = 0$ and if it is not linked to a QTL $\Pr(H_i) = 1$]. Comparing Equations 4 and 5, the correspondence between PER and PFP is evident.

For the general case involving a family of k hypothesis tests, we define

$$\mathrm{PFP} = \frac{E(V)}{E(R)}, \tag{6}$$

where V denotes the number of mistakenly rejected null hypotheses (number of false positives), R denotes the total number of rejected null hypotheses, and $E(V)$ and $E(R)$ denote the mathematical expectations of the random variables V and R, respectively. It is straightforward to show that this general definition of PFP specializes to the definition of PFP given by Southey and Fernando (1998) for the case of a genome scan involving k markers. For an experiment consisting of a single test of linkage between a random marker and a monogenetic disease locus, we have

$$\mathrm{PFP} = \frac{E(V)}{E(R)} = \frac{\Pr(V = 1)}{\Pr(R = 1)} = \frac{\Pr(L = 0, D = 1)}{\Pr(D = 1)}$$
$$= \Pr(L = 0|D = 1) = \mathrm{PER}.$$

Thus PFP simplifies to PER as proposed by Morton (1955) and is a natural extension of PER to the multiple-test setting considered throughout the remainder of the article.

PFP CONTROLS THE PROPORTION OF FALSE POSITIVES ACROSS MANY EXPERIMENTS

In this section we present two useful properties of PFP. Proofs of these properties are presented in the appendix.

Property 1: If the PFP level is equal to γ for each of n sets of tests corresponding to n independent experiments, then the PFP level for the collection of all tests associated with the n experiments is also equal to γ.

Property 2: If the PFP level is equal to γ for each of n sets of tests corresponding to n independent experiments, the observed proportion of false positives out of the total number of rejections across all n experiments converges to γ with probability 1 as the number of experiments increases, provided that the number of tests per experiment does not grow without bound.

Contrast property 1 with the situation encountered in FWER control. If the FWER is controlled at level γ for each of n independent families of tests, the FWER for the family consisting of the union of the n families of tests is $1 - (1 - \gamma)^n$. This quantity may be several times larger than γ for even moderate n. As the number of independent sets of tests increases, it becomes prohibitively difficult to control the probability of one or more false-positive errors.

Rather than attempting to avoid all false positive results, it makes sense to manage the accumulation of false positives relative to the total number of positive results that appear in the literature. The PFP approach provides precisely this type of error management as illustrated by property 2. It is property 2 that suggests "proportion of false positives" as an appropriate name of the error measure $E(V)/E(R)$. We show in a subsequent section of this article that control of other error measures (FWER, FDR, and pFDR) does not necessarily lead to the control of the proportion of false-positive results among all positive results.

PFP DOES NOT DEPEND ON EITHER THE NUMBER OF TESTS OR THE CORRELATION STRUCTURE AMONG THE TESTS

Consider a collection of k tests. Let W_j be 1 or 0 depending on whether or not the jth null hypothesis is falsely rejected. Let S_j be 1 or 0 depending on whether or not the jth null hypothesis is rejected. Suppose the jth test is conducted at CWER α_j, and let π_j denote the probability that the jth null hypothesis is rejected. Let $\mathcal{K}_0$ and $\mathcal{K}_1$ form a partition of the indices $1, \ldots, k$ such that $j \in \mathcal{K}_0$ if the jth null hypothesis is true and $j \in \mathcal{K}_0$ if the jth null hypothesis is false. Then for all $j \in \mathcal{K}_0$, we have $E(W_j) = E(S_j) = \alpha_j$. For all $j \in \mathcal{K}_1$, we have $E(W_j) = 0$ and $E(S_j) = \pi_j$. Now let p_0 denote the proportion of true null hypotheses among all hypotheses tested. Let $\alpha = 1/kp_0\sum_{j\in\mathcal{K}_0}\alpha_j$ denote the average CWER for tests of true null hypotheses. (Typically the same CWER will be used for all tests, in which case $\alpha_j = \alpha$ for all j.) Let $\pi = 1/(k(1 - p_0))\sum_{j\in\mathcal{K}_1}\pi_j$ denote the average power for tests of false null hypotheses. We may write PFP for the collection of k tests as

$$\begin{aligned}\text{PFP} &= \frac{E(V)}{E(R)} = \frac{E(\sum_{j=1}^{k} W_j)}{E(\sum_{j=1}^{k} S_j)} = \frac{\sum_{j=1}^{k} E(W_j)}{\sum_{j=1}^{k} E(S_j)}\\ &= \frac{\sum_{j\in\mathcal{K}_0}\alpha_j}{\sum_{j\in\mathcal{K}_0}\alpha_j + \sum_{j\in\mathcal{K}_1}\pi_j} \qquad (7)\\ &= \frac{\alpha k p_0}{\alpha k p_0 + \pi k(1-p_0)} = \frac{\alpha p_0}{\alpha p_0 + \pi(1-p_0)}. \qquad (8)\end{aligned}$$

From expression (8) we can see that PFP depends only on the average CWER α, the proportion p_0 of true null hypotheses out of all hypotheses tested, and the average power π. Note that, as claimed in the Introduction, the PFP does not depend on either the number of tests or the correlation structure among the tests. These properties are particularly desirable for application of the PFP approach to QTL mapping, where there is a nontrivial correlation structure among a large number of tests.

INTERPRETATION OF PFP FOR A SINGLE EXPERIMENT: THE RELATION OF PFP AND PER

We have shown that PFP = PER for an experiment consisting of a single test of linkage between a random marker and a monogenetic disease locus. In this section we demonstrate a more general result: the level of PER for a test randomly chosen from a family of k tests is equal to the level of PFP for the family of k tests. Let J denote a random index that is equally likely to take each value in $\{1, \ldots, k\}$. Then, using the notation of the previous section,

$$\begin{aligned}\text{PER} &= \Pr(J\in\mathcal{K}_0 \mid S_J = 1) = \frac{\Pr(J\in\mathcal{K}_0,\, S_J = 1)}{\Pr(S_J = 1)}\\ &= \frac{\Pr(S_J = 1\mid J\in\mathcal{K}_0)\Pr(J\in\mathcal{K}_0)}{\Pr(S_J = 1\mid J\in\mathcal{K}_0)\Pr(J\in\mathcal{K}_0) + \Pr(S_J = 1\mid J\in\mathcal{K}_1)\Pr(J\in\mathcal{K}_1)}. \qquad (9)\end{aligned}$$

Now

$$\begin{aligned}\Pr(S_J = 1\mid J\in\mathcal{K}_0) &= \sum_{j\in\mathcal{K}_0}\Pr(S_J = 1, J = j\mid J\in\mathcal{K}_0)\\ &= \sum_{j\in\mathcal{K}_0}\Pr(S_J = 1\mid J = j)\Pr(J = j\mid J\in\mathcal{K}_0)\\ &= \sum_{j\in\mathcal{K}_0}\alpha_j\frac{1}{kp_0} = \alpha. \qquad (10)\end{aligned}$$

Similarly

$$\begin{aligned}\Pr(S_J = 1\mid J\in\mathcal{K}_1) = \pi, \quad \Pr(J\in\mathcal{K}_0) = p_0,\\ \Pr(J\in\mathcal{K}_1) = 1 - p_0. \qquad (11)\end{aligned}$$

Now (9), (10), and (11) imply that PER is equal to (8). Thus PER = PFP.

ESTIMATING PFP FOR A GIVEN EXPERIMENT

For simplicity of notation, we assume henceforth that a single CWER α is used for each of k tests. Consideration of the case where the jth test is conducted at its own CWER α_j is a straightforward generalization. For any given CWER α, (8) indicates that the PFP can be estimated as

$$\widehat{\text{PFP}}_\alpha = \frac{\alpha\hat{p}_0}{\alpha\hat{p}_0 + \hat{\pi}_\alpha(1-\hat{p}_0)}, \qquad (12)$$

where $\hat{p}_0$ and $\hat{\pi}_\alpha$ are estimates of p_0 and π, respectively. Several methods for estimating p_0 are beginning to appear in the literature. Benjamini and Hochberg (2000) described a method for estimating p_0 on the basis of a graphical approach proposed by Schweder and Spjotvoll (1982). Storey (2002) and Storey and Tibshirani (2001) used resampling techniques to approximate p_0. Allison *et al.* (2002) fit a mixture of a uniform distribution and a β distribution to the observed P values. The maximum-likelihood estimate of the mixing proportion corresponding to the uniform distribution serves as an estimate of p_0. Mosig *et al.* (2001) proposed an iterative algorithm for estimating p_0 that uses the number of P values falling into each of several intervals that form a partition of the interval [0, 1]. Their procedure can be considered a nonparametric version of the procedure proposed by Allison *et al.* (2002). Nettleton and Hwang (2003) describe the estimator proposed by Mosig *et al.* (2001) in greater detail and show that the estimator can be computed directly from the observed P values without iteration.

Because $1 - \hat{p}_0$ is an estimate of the proportion of tested null hypotheses that are false (*e.g.*, the proportion of markers linked to QTL), it can be of direct scientific interest. Note, however, that estimating the proportion of null hypotheses that are false is not the same thing as estimating *which* of the null hypotheses are false. Simply identifying the $k(1 - \hat{p}_0)$ tests with the smallest P values as those tests with false null hypotheses will typically result in an unacceptably high PFP (see, for example, Genovese and Wasserman 2002, who considered this issue as part of their thorough investigation of the properties of FDR). Thus it is important to combine estimates of p_0 with estimates of π to approximate PFP.

An estimator of π is given by

$$\hat{\pi}_\alpha = \frac{R_\alpha - \alpha k\hat{p}_0}{k(1-\hat{p}_0)}, \qquad (13)$$

where R_α denotes the observed value of R for the given choice of α. Note that the numerator of (13) is an estimate of the number of true positives while the denominator is an estimate of the number of tests for which the null hypothesis is false. Combining (12) and (13) yields

$$\widehat{\text{PFP}}_\alpha = \frac{\alpha k\hat{p}_0}{R_\alpha}. \qquad (14)$$

When the method of Mosig *et al.* (2001) is used to obtain $\hat{p}_0$, $\widehat{\mathrm{PFP}_\alpha}$ is the estimator that Mosig *et al.* (2001) referred to as "adjusted FDR." In the simulation described in a subsequent section, we use this estimator to produce estimates of PFP ($\widehat{\mathrm{PFP}_\alpha}$) for varying levels of α (Table 2).

COMPARISON OF PFP, FWER, FDR, AND pFDR

Benjamini and Hochberg (1995) defined FDR as

$$\mathrm{FDR} = E\left(\frac{V}{R}\middle| R > 0\right)\Pr(R > 0), \tag{15}$$

where, as defined previously, V represents the number of mistakenly rejected null hypotheses and R denotes the number of rejected null hypotheses. Storey (2002) defined the pFDR as

$$\mathrm{pFDR} = E\left(\frac{V}{R}\middle| R > 0\right) \tag{16}$$

and proposed pFDR as more suitable than FDR as a measure of false discoveries because it more closely matches the type of error control that is desirable in practice. Both FDR and pFDR seem to be gaining in popularity as error measures for multiple-testing problems involving hundreds or thousands of tests. This is especially the case in the analysis of microarray data where thousands of tests are typical. Familywise error rate [FWER = $\Pr(V > 0)$] traditionally has been the most popular error measure for general multiple-testing problems.

We have previously shown that control of PFP across multiple experiments will lead to control of the proportion of false-positive results among all positive results in the long run. We now show by a hypothetical example that the other error measures (FDR, pFDR, and FWER) do not necessarily share this property.

Suppose that for each experiment in a series of independent and identical experiments V/R is 50/100 with probability 0.1, 0/10 with probability 0.5, and 0/0 with probability 0.4. Then

$$\mathrm{PFP} = \frac{50(0.1)}{100(0.1) + 10(0.5)} = \frac{1}{3},$$

which is the proportion of false positives among all positive results that will accrue in the long run over repeated experimentation. On the other hand, the values of FWER, pFDR, and FDR are

$$\mathrm{FWER} = \frac{1}{10}, \quad \mathrm{pFDR} = \left(\frac{50}{100}\right)\left(\frac{0.1}{0.1 + 0.5}\right)$$
$$= \frac{1}{12}, \quad \mathrm{FDR} = \left(\frac{1}{12}\right)(0.6) = \frac{1}{20}.$$

This example shows that control of FDR, pFDR, or FWER will not guarantee control of the accumulation of false-positive results as a proportion of all positive results over multiple experiments. Obviously the example has been artificially constructed to emphasize the differences among the error measures. This example involves independent experiments, which means that the tests in one experiment are independent of tests in another. The tests within any of the experiments, however, are not necessarily independent of each other. Indeed, these tests must be dependent to obtain the behavior described in the example. Note that when a large number of rejections occur, the ratio V/R is high (50/100). On the other hand, when a small number of rejections occur, the ratio V/R is quite low (0/10). Such a situation can arise in the QTL mapping setting. Suppose that a QTL for a trait of interest lies on a chromosome for which few markers are available. Suppose that some other chromosomes have a high density of markers. A high density of markers on a chromosome without the QTL translates into a high positive correlation among tests for which the null hypothesis is true. Because dense markers are positively correlated, a false-positive result at any one of these markers is likely to be accompanied by many other false-positive results at neighboring markers. With few markers on the chromosome containing the QTL, there can never be a large number of true positive results. Thus a large number of rejections will occur only when there are a large number of false positives. It is in such situations that we will see substantial differences between PFP and the other error measures. Such a scenario is created in model 5 of our simulation study described later in this article.

Although the example of this section and model 5 of our simulation show that the error measures can differ substantially, there are many similarities among FDR, pFDR, and PFP. Storey (2002) has shown that when the tests are identically and independently distributed pFDR = PER; *i.e.*, the level pFDR for a set of k tests is equal to the level of PER for a randomly chosen test. Storey (2003) has shown that pFDR = PFP when the tests are independent (Corollary 1 in Storey 2003) and that pFDR and FDR will be approximately equivalent to PER (and thus PFP) as the number of tests in a family grows large as long as the test statistics corresponding to the family of tests satisfy a "weak dependence" condition (Theorem 4 in Storey 2003). We have shown that the equality between PFP and PER holds in general regardless of the dependence structure among the test statistics or the number of tests conducted. A probability interpretation of pFDR that holds even when tests are not independent or identically distributed is given below.

Let A denote the event, "a positive result, randomly selected from all positive results, is a false positive." We have

$$\begin{aligned}\Pr(A \mid R > 0) &= \sum_{r=1}^{k}\sum_{v=0}^{r}\Pr(A,\ V = v,\ R = r \mid R > 0)\\ &= \sum_{r=1}^{k}\sum_{v=0}^{r}\Pr(A \mid V = v,\ R = r,\ R > 0)\Pr(V = v,\ R = r \mid R > 0)\\ &= \sum_{r=1}^{k}\sum_{v=0}^{r}\frac{v}{r}\Pr(V = v,\ R = r \mid R > 0)\\ &= E\left(\frac{V}{R} \mid R > 0\right) = \text{pFDR}.\end{aligned}$$

Thus, even when tests are not independent nor identically distributed, conditional on an experiment having one or more positive test results, pFDR is equal to the probability that a randomly chosen test from among these positive results is a false positive.

It is easiest to understand the somewhat subtle difference between this interpretation of pFDR and the interpretation of PFP as PER by considering the example presented in this section. In the example pFDR is determined as follows. Of the experiments with at least one positive result, about five-sixths of the experiments will have 0 as the probability that a randomly selected positive result will be a false positive while the other one-sixth will have probability 0.5 that a randomly selected positive result is a false positive. Thus pFDR is $(5/6) \cdot 0 + (1/6)(0.5) = 1/12$, which is exactly the probability that a randomly selected positive result will be a false positive, given that the experiment resulted in at least one positive result. Note that this calculation in no way accounts for the fact that there are many more positive results in the less likely experimental outcome $[\Pr(V/R = 50/100) = 0.1]$ than in the more likely outcome $[\Pr(V/R = 0/10) = 0.5]$. On the other hand, PFP = PER is the probability that a randomly selected result is a false positive, given that it is positive. By conditioning on the event that the randomly selected result is positive rather than on the event that the experiment contains at least one positive, PFP accounts for differences in the number of positive results across experimental outcomes because randomly selected events are more likely to be positive in experiments with many positive results. In contrast to pFDR, experimental outcomes V/R are weighted by both their probability of occurrence and the number of rejections R. For our hypothetical example, we can write PFP as a weighted average of the V/R ratios as

$$\text{PFP} = \frac{(0.5)(10)(0/10) + (0.1)(100)(10/100)}{(0.5)(10) + (0.1)(100)} = \frac{1}{3}.$$

A SIMULATION STUDY

A QTL scan with 500 backcross offspring from inbred lines was simulated. The simulation was used to compare PFP with FWER, FDR, and pFDR and to illustrate how the estimated PFP levels compare to true PFP levels. The simulation was repeated for five simple genetic models.

QTL model 1: This model had 10 chromosomes with one QTL at the center of the chromosome; the 10 QTL were of equal effect, so that each accounted for 10% of the genetic variance. The remaining 20 chromosomes had no QTL. The simulated trait was completely additive with a heritability of 0.25 in the F_2 generation. The residuals were normally distributed. Each chromosome was 100 cM long and had 21 equally spaced markers.

QTL model 2: This model was obtained from model 1 by moving the QTL from the center to the left by 25 cM for each of the 10 chromosomes with a QTL.

QTL model 3: This model was obtained from model 1 by increasing the number of chromosomes with a single QTL at the center from 10 to 20 and by decreasing the number of chromosomes with no QTL from 20 to 10. As this model contains 20 QTL of the same effect, each accounted for 5% of the additive genetic variance.

QTL model 4: This model was obtained from model 1 by decreasing the number of chromosomes with a single QTL at the center from 10 to 5 and by increasing the number of chromosomes with no QTL from 20 to 25. As this model contained five QTL of the same size, each accounted for 20% of the additive genetic variance.

QTL model 5: This model with only two chromosomes was constructed to illustrate that PFP can give quite different results from pFDR and FDR. The first chromosome was 100 cM long with one QTL at the center and 11 equally spaced markers. The second chromosome also was 100 cM long with no QTL and 101 equally spaced markers. The heritability for the trait was 0.025.

The scan for QTL was based on testing each marker for linkage to QTL by a *t*-test for comparing the means for the trait between the two marker genotype classes (Soller *et al.* 1976). The null hypothesis of no linkage to a QTL was rejected if the P value for the test was lower than the critical CWER. For each experiment, the numbers of positive (R) and false-positive (V) test results were counted given the critical CWER values of 0.01, 0.001, and 0.0001. For each model, 50,000 replications of the experiment were used to obtain empirical values for PFP, pFDR, FDR, and FWER, which in this context is called the GWER (Lander and Kruglyak 1995). The empirical PFP was obtained as $\overline{V}/\overline{R}$, $\overline{V}$ and $\overline{R}$ being the mean values of V and R over the 50,000 replications of the experiment; empirical pFDR was obtained as the mean value of the ratio V/R over all experiments with $R > 0$; empirical FDR was obtained as empirical pFDR times the proportion of experiments with $R > 0$; and empirical GWER was obtained as the proportion of experiments with $V > 0$. The results for these empirical values are given in Table 1.

Table 1 shows that PFP, pFDR, and FDR were practically identical to each other for model 1 through model 4, while GWER was very different from these. For these four models, using a P-value threshold of 0.001 was sufficient to control PFP, pFDR, or FDR to well below

TABLE 1

Empirical values of PFP, pFDR, FDR, and GWER from 50,000 replicates of a simulated backcross experiment for models 1–5

Model	Critical CWER	PFP	pFDR	FDR	GWER
1	0.01	0.088	0.090	0.090	0.825
	0.001	0.027	0.031	0.031	0.194
	0.0001	0.009	0.012	0.010	0.025
2	0.01	0.094	0.095	0.095	0.824
	0.001	0.028	0.032	0.031	0.192
	0.0001	0.009	0.012	0.010	0.025
3	0.01	0.051	0.053	0.053	0.581
	0.001	0.022	0.026	0.026	0.104
	0.0001	0.010	0.013	0.008	0.012
4	0.01	0.106	0.105	0.105	0.889
	0.001	0.022	0.023	0.023	0.237
	0.0001	0.004	0.005	0.005	0.030
5	0.01	0.275	0.107	0.079	0.111
	0.001	0.081	0.030	0.012	0.015
	0.0001	0.024	0.008	0.002	0.002

TABLE 2

Empirical values of PFP, mean values of PFP estimates, and their mean squared errors from 50,000 replicates of a simulated backcross experiment for models 1–5

Model	Critical *P* value	PFP	Mean ($\widehat{\mathrm{PFP}}$)	MSE ($\widehat{\mathrm{PFP}}$)
1	0.01	0.088	0.115	0.0039
	0.001	0.027	0.049	0.0044
	0.0001	0.009	0.016	0.0002
2	0.01	0.094	0.125	0.0050
	0.001	0.028	0.053	0.0049
	0.0001	0.009	0.016	0.0003
3	0.01	0.051	0.125	0.0133
	0.001	0.022	0.080	0.0123
	0.0001	0.010	0.020	0.0003
4	0.01	0.106	0.115	0.0018
	0.001	0.022	0.028	0.0007
	0.0001	0.004	0.008	0.0008
5	0.01	0.275	0.352	0.101
	0.001	0.081	0.051	0.003
	0.0001	0.024	0.006	0.000

0.05, while with this threshold GWER is well above 0.05. The results for model 5 show that PFP can be quite different from pFDR and FDR and that pFDR can be different from FDR.

DISCUSSION

In linkage analysis, significance testing has not been based on controlling the type I error rate, but on controlling the PER, which is the conditional probability of a false-positive result given a positive test result (Morton 1955; Ott 1991). For QTL scans, which involve multiple tests of linkage, Southey and Fernando (1998) proposed PFP as a natural extension of PER, which was defined for a single test. In this article we provided the mathematical justification for this proposal.

Briefly, the justification is as follows. If the level of PER for a test is γ, then as the number of independent tests increases, the proportion of false positives in the accumulated positive results converges to γ. We have shown here that if the level of PFP in a multiple test experiment is γ, then as the number of such independent experiments increases, the proportion of false positives in the accumulated positive results also converges to γ. Alternatively, we have shown here that when the number k of tests is 1, controlling PER is equivalent to controlling PFP. Further, when $k > 1$, we showed that controlling the PER for a test that is randomly chosen from the set of k tests is equivalent to controlling the PFP defined over all k tests. These results hold for any dependence structure among the k tests in an experiment.

When tests are identically and independently distributed, pFDR = PFP, and thus, in this situation, controlling PER for a randomly chosen test is equivalent to controlling pFDR (Storey 2002). A probability interpretation of pFDR that holds even when tests are not independent nor identically distributed given here is: if an experiment with level γ for pFDR has one or more positive test results, γ is the conditional probability that a randomly sampled result from these positive results is a false positive.

Thus in multiple-test experiments, controlling PFP will result in controlling the proportion of false-positive results in the accumulated positive test results over many experiments, while controlling pFDR will result in controlling the expected proportion of false positives in the positive test results in each experiment. When tests are independently and identically distributed, pFDR = PFP, and, thus, false positives will be controlled to the same level in each experiment and in the accumulated test results over many experiments. The simulation results for models 1–4 show that even when tests are highly dependent, pFDR and PFP can give very similar results. For tests that are identically distributed but dependent, Storey (2003) has given conditions under which pFDR will converge to PER = PFP as the number of tests increases. As demonstrated by the results for model 5, however, it is clear that in some situations controlling PFP is not equivalent to controlling pFDR or FDR.

Like pFDR, PFP = 1 if all the null hypotheses tested are true. Thus neither pFDR nor PFP can be controlled in the same sense that FDR can be controlled (Benjamini and Hochberg 1995). Nonetheless, we believe the more direct interpretations of pFDR and PFP make these error measures worth considering. We have illustrated the connection between the PER and PFP and have shown that, unlike FDR and pFDR, PFP is free of the correlation structure among the tests. Storey and Tibshirani (2001) propose a method for approximating FDR and pFDR under general dependence structures. Using their method requires the ability to draw samples from an approximation to the joint distribution of the test statistics when all null hypotheses are true. This is not a trivial computing exercise and may be very difficult to accomplish in some situations. In contrast, the approach to estimate PFP that we have presented here requires only the P values corresponding to the tests of interest. Such P values can be obtained without simulation in situations where the approximate marginal distributions of the test statistics are known.

In this article we used the method proposed by Mosig *et al.* (2001) to estimate p_0 and π to demonstrate the estimation of PFP (Table 2). In most of the cases our estimates of PFP were conservative. In only two cases with very low heritability and one QTL was the mean of the estimated PFP levels lower than the empirical value. Even in this case, when estimated PFP was ~0.05, the empirical PFP was only slightly higher. Research in methods for estimating p_0 is ongoing, so we believe the estimates of PFP illustrated here can be improved by using improved estimates of p_0. It is also worth noting that in models 1–4, where the number of QTL ranged from 5 through 20, using a critical CWER of 0.001 was sufficient to control PFP <0.05.

The authors thank two anonymous reviewers for useful comments. R. Fernando acknowledges support from the National Research Initiative Competitive Grants Program of the U.S. Department of Agriculture, Award 2002-35205-11546; D. Nettleton acknowledges support from the National Research Initiative Competitive Grants Program of the U.S. Department of Agriculture, Award 1998-35205-10390; M. Soller acknowledges support from the BovMAS project of the European Union FP5 program and the Cotswold Swine Breeding Company.

LITERATURE CITED

Allison, D. B., G. L. Gadbury, M. Heo, J. Fernandez, C.-K. Lee *et al.*, 2002 A mixture model approach for the analysis of microarray gene expression data. Comp. Stat. Data Anal. **39:** 1–20.

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

Benjamini, Y., and Y. Hochberg, 2000 On the adaptive control of the false discovery rate in multiple testing with independent statistics. J. Educ. Behav. Stat. **25:** 60–83.

Elston, R. C., 1997 1996 William Allan award address: algorithms and inferences: the challenges of multifactorial diseases. Am. J. Hum. Genet. **60:** 225–262.

Elston, R. C., and K. Lange, 1975 The prior probability of autosomal linkage. Ann. Hum. Genet. **38:** 341–350.

Genovese, C., and L. Wasserman, 2002 Operator characteristics and extensions of the false discovery rate procedure. J. R. Stat. Soc. Ser. B **64:** 499–517.

Lander, E., and L. Kruglyak, 1995 Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat. Genet. **11:** 241–247.

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

Mosig, M., E. Lipkin, G. Khutoreskaya, E. Tchourzyna, M. Soller *et al.*, 2001 A whole genome scan for QTL affecting milk protein percentage in Israel-Holstein cattle by means of selective milk pooling in a daughter design, using an adjusted false discovery rate criterion. Genetics **157:** 1683–1698.

Nettleton, D., and J. Hwang, 2003 Estimating the number of false null hypotheses when conducting many tests. Preprint Series 2003-09, Technical Report, Department of Statistics, Iowa State University, Ames, IA.

Ott, J., 1991 *Analysis of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore.

Risch, N., 1991 A note on multiple testing procedures in linkage analysis. Am. J. Hum. Genet. **48:** 1058–1064.

Rohatgi, V. K., 1976 *An Introduction to Probability Theory and Mathematical Statistics*. Wiley, New York.

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

Soller, M., T. Brody and A. Genizi, 1976 On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theor. Appl. Genet. **47:** 35–39.

Southey, B. R., and R. L. Fernando, 1998 Controlling the proportion of false positives among significant results in QTL detection. Proceedings of the 6th World Congress on Genetics Applied to Livestock Production, Armidale, Australia, Vol. 26, pp. 221–224.

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

Storey, J. D., 2003 The positive false discovery rate: Bayesian interpretation and the Q-value. Ann. Stat. **31** (6): 2013–2035.

Storey, J. D., and R. Tibshirani, 2001 Estimating false discovery rates under dependence, with applications to DNA microarrays. Technical Report 2001–28, Department of Statistics, Stanford University, Stanford, CA.

Weller, J. I., 2000 Using the false discovery rate approach in the genetic dissection of complex traits: a response to Zaykin *et al.* Genetics **154:** 1919.

Communicating editor: J. B. Walsh

APPENDIX

Proof of Property 1. Let V_i denote the number of rejections of a true null hypothesis in the ith experiment, and let R_i denote the number of rejections of a null hypothesis in the ith experiment. If we have $E(V_i)/E(R_i) = \gamma$ for all $i = 1, \ldots, n$, then the PFP across the set of n independent experiments is given by

$$\frac{E\left(\sum_{i=1}^{n} V_i\right)}{E\left(\sum_{i=1}^{n} R_i\right)} = \frac{\sum_{i=1}^{n} E(V_i)}{\sum_{i=1}^{n} E(R_i)} = \frac{\sum_{i=1}^{n} E(R_i)E(V_i)/E(R_i)}{\sum_{i=1}^{n} E(R_i)}$$

$$= \frac{\sum_{i=1}^{n} E(R_i)\gamma}{\sum_{i=1}^{n} E(R_i)} = \gamma.$$

This proves property 1.

Proof of Property 2. We begin the proof of property 2 by noting that

$$\frac{\sum_{i=1}^{n} V_i}{\sum_{i=1}^{n} R_i} = \frac{\sum_{i=1}^{n}\{V_i - E(V_i)\}/n + \sum_{i=1}^{n} E(V_i)/n}{\sum_{i=1}^{n}\{R_i - E(R_i)\}/n + \sum_{i=1}^{n} E(R_i)/n}. \qquad \text{(A1)}$$

By Corollary 1 to Theorem 6 in Rohatgi (1976), $\Sigma_{i=1}^{n}\{V_i - E(V_i)\}/n$ will converge to 0, in the almost sure sense, as long as $\Sigma_{i=1}^{\infty}\mathrm{Var}(V_i)/i_2 < \infty$. Note that $V_i \leq k_i$, where k_i denotes the number of tests in the ith experiment. There exists $M \geq k_i$ for all i because the number of tests for each experiment does not grow without bound. Thus

$$V_i \leq M, \quad \text{which implies } \mathrm{Var}(V_i) \leq E(V_i^2) \leq M^2 \text{ for all } i.$$

It follows that $\Sigma_{i=1}^{\infty}\mathrm{Var}(V_i)/i^2$ is bounded above by $M^2\Sigma_{i=1}^{\infty}1/i^2$, which is finite. Thus Corollary 1 to Theorem 6 in Rohatgi (1976) implies that

$$\Pr\Big(\lim_{n\to\infty}\sum_{i=1}^{n}\{\mathrm{V}_i - E(V_i)\}/n = 0\Big) = 1.$$

The same basic argument can be used to show that

$$\Pr\Big(\lim_{n\to\infty}\sum_{i=1}^{n}\{R_i - E(R_i)\}/n = 0\Big) = 1.$$

Therefore, using (A1), we have

$$\lim_{n\to\infty}\frac{\sum_{i=1}^{n} V_i}{\sum_{i=1}^{n} R_i} \overset{\text{a.s.}}{=} \lim_{n\to\infty}\frac{\sum_{i=1}^{n} E(V_i)/n}{\sum_{i=1}^{n} E(R_i)/n} = \lim_{n\to\infty}\frac{E(\sum_{i=1}^{n} V_i)}{E(\sum_{i=1}^{n} R_i)} = \gamma,$$

where the last equality follows from property 1 and $\overset{\text{a.s.}}{=}$ denotes equality in the almost sure sense.