Application of the False Discovery Rate to Quantitative Trait Loci Interval Mapping With Multiple Traits

Hakkyo Lee,* Jack C. M. Dekkers,[†,1] M. Soller,[‡] Massoud Malek,[†] Rohan L. Fernando[†] and Max F. Rothschild[†]

**Hankyong National University, Ansung-si, Kyonggi, 456-749, Korea, [†]Department of Animal Science, Iowa State University, Ames, Iowa 50011 and [‡]Department of Genetics, Hebrew University of Jerusalem, Jerusalem, 91904 Israel*

Manuscript received February 6, 2001
Accepted for publication March 13, 2002

ABSTRACT

Controlling the false discovery rate (FDR) has been proposed as an alternative to controlling the genome-wise error rate (GWER) for detecting quantitative trait loci (QTL) in genome scans. The objective here was to implement FDR in the context of regression interval mapping for multiple traits. Data on five traits from an F_2 swine breed cross were used. FDR was implemented using tests at every 1 cM (FDR1) and using tests with the highest test statistic for each marker interval (FDRm). For the latter, a method was developed to predict comparison-wise error rates. At low error rates, FDR1 behaved erratically; FDRm was more stable but gave similar significance thresholds and number of QTL detected. At the same error rate, methods to control FDR gave less stringent significance thresholds and more QTL detected than methods to control GWER. Although testing across traits had limited impact on FDR, single-trait testing was recommended because there is no theoretical reason to pool tests across traits for FDR. FDR based on FDRm was recommended for QTL detection in interval mapping because it provides significance tests that are meaningful, yet not overly stringent, such that a more complete picture of QTL is revealed.

DUE to availability of large numbers of polymorphic markers, it is now possible to scan a complete genome for loci affecting quantitative traits of interest, so-called quantitative trait loci (QTL). Because of the large number and correlated statistical tests conducted and associated concerns about a flood of false-positive claims for QTL if comparison-wise type I error rates (CWER) are not properly controlled, methods to set CWER thresholds for declaring the presence of a QTL have received much attention over the past decade. The most common approach is to set CWER so as to control the genome-wise type I error rate (GWER). To achieve this, Churchill and Doerge (1994) proposed an empirical permutation test method (referred to as CD) that provides CWER thresholds controlling GWER for the set of markers that is included in the experiment, while Lander and Schork (1994) and Lander and Kruglyak (1995) presented a method (referred to as LK) that provides CWER thresholds that control GWER as though based on a high-density marker map for the genome under analysis.

More recently, Weller *et al.* (1998) proposed controlling the comparison-wise false discovery rate (FDR), as developed by Benjamini and Hochberg (1995), as an alternative to controlling GWER in genome scans. The FDR was defined as the expected proportion of false positives among tests that are declared significant (Weller *et al.* 1998). Controlling FDR is intuitively attractive because it enables a more reasoned calculation of the tradeoffs involved in conducting follow-up research or in investing selection effort in a putative QTL against the possibility that the result is a false positive.

Interval mapping based on least squares (Haley and Knott 1992) or maximum likelihood (Lander and Botstein 1989) is the common statistical method used to detect QTL. This involves conducting a statistical test at every position for a putative QTL (typically every 1 cM). In their implementation of FDR, Weller *et al.* (1998) considered only tests at individual markers and not tests conducted at each possible QTL position, as in interval mapping. Interval mapping provides additional power to detect QTL because markers that flank an interval provide partially independent information to detect QTL. Furthermore, single-marker analyses do not allow separate estimation of QTL effect and position. Thus, implementation of FDR for interval mapping is warranted (Spelman 1998).

The methods for hypothesis testing described above have generally been applied to single traits. Yet, QTL mapping experiments typically involve several to many traits and this must be taken into account when setting significance thresholds. Technically, this can be readily achieved in the CD and FDR approaches by grouping tests across traits, as though they were generated by a single analysis; for the LK approach single-trait thresholds can be adjusted by a Bonferroni correction (Lander and Kruglyak 1995; Spelman *et al.* 1996).

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

The objective of this study was to implement the FDR approach for least-squares regression interval mapping of single and multiple traits. A secondary objective was to compare CWER thresholds and power for QTL detection using FDR to those from the CD and LK approaches. Data from an F_2 cross of outbred lines in pigs were used to address these objectives but methods and results have a more general application.

MATERIALS AND METHODS

Data and QTL analyses: Data used were from a complete genome scan based on 125 microsatellite markers in 525 F_2 progeny from a cross between two breeds of swine, Berkshire and Yorkshire. Full details are in Malek *et al.* (2001a,b). Data on five meat quality traits were used: carcass weight, last rib back fat thickness, loin eye area, and cholesterol content and marbling score of the loin eye.

The least-squares regression interval mapping procedure and program of Haley *et al.* (1994) for a cross between outbred lines was used for QTL analysis. A statistical model was fitted at each 1-cM position k on the chromosome, to phenotypic records y,

$$y = \text{fixed effects} + b_{a,k}c_{a,k} + b_{d,k}c_{d,k} + \text{residual},$$

where $b_{a,k}$ and $b_{d,k}$ are regression coefficients that estimate the additive and dominance effects for the putative QTL at position k, and $c_{a,k}$ and $c_{d,k}$ are the additive and dominance "breed-origin" coefficients at that position. Breed-origin coefficients were based on breed-origin probabilities for alleles at the putative position. Breed-origin probabilities were derived using all available marker data following Haley *et al.* (1994). The statistic for testing the presence of a QTL at a particular position was derived as an F-statistic following Haley *et al.* (1994). For a single test, this statistic has 2 and 517 d.f. for our data and model.

False discovery rate (FDR): CWER thresholds to control FDR to a level α_F, as suggested by Weller *et al.* (1998), were derived by first ranking all tests on the basis of the CWER of the F-statistic. The FDR for the ith ranked test can then be computed as $FDR_i = N \times CWER_i/i$, where N is the total number of tests and $CWER_i$ is the CWER for the ith ranked test. Note that $N \times CWER_i$ is the expected number of tests declared significant if no QTL were present and the CWER threshold was set at $\alpha_C = CWER_i$, while i is the number of tests that are actually declared significant at that level in the current experiment. Significance thresholds to control FDR at a level α_F were then determined as the CWER corresponding to the largest i for which FDR_i was below the desired level α_F. For multiple-trait thresholds, tests were ranked across traits.

Initially, FDR were derived on the basis of all tests conducted, *i.e.*, at every 1-cM position, referred to as FDR1. The CWER for individual tests were obtained from the standard F-distribution. FDR1 included 2050 tests per trait and 10,250 tests across the five traits. In a second approach, referred to as FDRm, only the highest F-statistic within each marker interval was included, as suggested in Spelman (1998). FDRm included 106 tests per trait and 530 tests across the five traits. For FDRm, a standard F distribution cannot be used to determine the type I error rate for a given test because each test represents the largest test within a marker interval and is already the result of multiple testing. To account for this, instead of the CWER, the interval-wise error rate (IWER) was used to compute the expected number of false positives for FDRm. The IWER represents the type I error rate for the null hypothesis that QTL are not present for any of the 1-cM tests conducted in a given interval.

The IWER for a given marker interval was determined by the distribution of the maximum F-statistic in that interval under the null hypothesis, which can be derived by data permutation. Because densities are required for low values of IWER (<0.001), this would require a very large number of permutations to be conducted for every marker interval. To provide an alternative requiring much less computation, a prediction equation was derived that allowed prediction of IWER on the basis of the CWER for the observed maximum F-value in the interval and the degree of dependence of tests conducted in that interval. The dependence of tests at two positions k and l on the chromosome can be quantified by the correlations of the breed-origin coefficients at these positions, *i.e.*, the correlation of $c_{a,k}$ with $c_{a,l}$ and the correlation of $c_{d,k}$ with $c_{d,l}$. Correlations between breed-origin coefficients at the flanking markers were computed across the F_2 individuals for each interval, separately for additive and dominance coefficients. The average of the two correlations was used to quantify the dependence of tests conducted within the interval. The rationale for using correlations between flanking markers is that all information to map a QTL in an interval is present at the markers that flank the interval (Whittaker *et al.* 1996; Kadarmideen and Dekkers 1999).

Data from 13 marker intervals (6 on chromosome 1 and 7 on chromosome 2) were used to derive the prediction equation for IWER. For each interval, the distribution of the maximum F-statistic under the null hypothesis of no QTL was derived by data permutation (10,000). Threshold F-values were obtained for a range of IWER and used to derive the relationship of IWER with CWER and the average correlation between breed-origin coefficients at the flanking markers. The resulting prediction equation was used to derive IWER for all tests included in FDRm.

Approaches to control GWER: The CD and LK methods were used to derive CWER thresholds that controlled GWER at 0.10, 0.05, and 0.01. The CD method was implemented as in Churchill and Doerge (1994) with 10,000 permutations. For multiple-trait thresholds, the maximum F-statistic for all tests conducted across the five traits was recorded for each permuted data set. Analytical thresholds to control GWER (LK) were derived following Lander and Schork (1994) and Lander and Kruglyak (1995), using $C = 19$ chromosomes, a dependence coefficient between tests of $\rho = 1.5$, and $d = 2$ d.f. for each test. For controlling GWER across the five traits, a Bonferroni adjustment was made, on the realistic assumption that the traits are independent (Table 1). The single-trait GWER ($GWER_{ST}$) required to control the multiple trait GWER at $GWER_{MT}$ was then derived from $GWER_{MT} = 1 - (1 - GWER_{ST})^5$.

RESULTS

Population parameters for the five traits that were included in the analyses are in Table 1. Traits were chosen because of their independence and range of heritabilities. Traits were approximately independent, as indicated by close to zero phenotypic correlations.

Prediction of IWER: Table 2 shows characteristics of the 13 marker intervals that were used to develop the prediction equation for IWER. They represented a range of marker distances and information contents. Correlations between breed-origin coefficients were lower for intervals that were longer and that had higher

TABLE 1

Means, standard deviations, heritabilities, and phenotypic correlations of the five traits analyzed in the F_2 population

Trait	Mean	Standard deviation	Heritability	Phenotypic correlation			
				Last rib	Loin eye	Marbling	Cholesterol
Carcass weight (kg)	87.1	5.7	0.18[a]	0.26	0.17	0.09	0.06
Last rib back fat (cm)	3.16	0.61	0.36		−0.25	0.14	0.12
Loin eye area (cm^2)	35.6	5.7	0.48			−0.25	−0.07
Marbling score (1–5)	3.8	0.73	0.13				0.09
Cholesterol (mg/100 g)	57.7	8.3	0.31				

Heritability estimates are from Goodwin and Burroughs (1995).
[a] Heritability of dressing percentage.

information content. Correlations between dominance coefficients were consistently lower than correlations between additive coefficients. Data on all of the 106 marker intervals showed a high correlation (0.97) between the two correlation coefficients.

Thresholds of the *F*-statistic for IWER were obtained by data permutation for each of the 13 intervals of Table 2 and plotted against their corresponding CWER. Figure 1 illustrates the relationship between IWER and CWER for intervals with a low and a high correlation between QTL coefficients (intervals 2 and 4 on chromosome 2). For these intervals, 50,000 permutations were run, such that thresholds for IWER as low as 0.0005 could be derived.

The IWER and CWER were linearly related on the logarithmic scale for IWER < 0.4 (Figure 1), which is the IWER region of interest. For the interval with the high correlation, the relationship between IWER and CWER was close to equality (IWER = CWER), which is equivalent to conducting a single test across the interval. For the interval with the low correlation, IWER was substantially greater than CWER, except for CWER close to 1. Thus, the CWER required for a given IWER decreased with magnitude of the correlation.

The following prediction equation was derived on the basis of CWER data points corresponding to IWER equal to 0.01, 0.05, 0.1, 0.2, 0.3, and 0.4 for the 13 intervals,

$$\log[\mathrm{IWER}_j] = 0.313 + 0.855\log(\mathrm{CWER}_j) - 0.256\overline{\mathrm{Corr}}_j + 0.100[\log(\mathrm{CWER}_j) \times \overline{\mathrm{Corr}}_j],$$

where IWER_j and CWER_j are the IWER and corresponding CWER for interval j, and $\overline{\mathrm{Corr}}_j$ is the average of the correlations for the additive and dominance breed-origin coefficients at the flanking markers (Table 2). The model R square was 0.998, which indicates a very good

TABLE 2

Characteristics of the 13 marker intervals on chromosomes 1 and 2

Chromosome	Marker interval	Interval length (cM)	Average marker information content[a]	Correlation between breed origin coefficients at flanking markers		
				Additive	Dominance	Average
1	1	18	0.61	0.79	0.62	0.70
	2	10	0.58	0.87	0.78	0.82
	3	12	0.92	0.75	0.56	0.66
	4	13	0.95	0.77	0.55	0.66
	5	40	0.86	0.41	0.18	0.29
	6	16	0.94	0.68	0.46	0.57
2	1	27	0.90	0.54	0.28	0.41
	2	43	0.81	0.41	0.12	0.27
	3	14	0.81	0.73	0.58	0.65
	4	4	0.67	0.97	0.94	0.95
	5	19	0.58	0.68	0.51	0.60
	6	25	0.79	0.62	0.34	0.48
	7	5	0.92	0.89	0.84	0.86

[a] Information content was based on the ability to determine breed origin of marker alleles averaged over all F_2 progeny, using information from that marker only (Malek *et al.* 2001a).

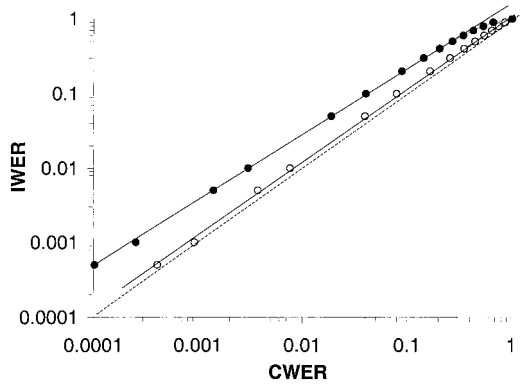

FIGURE 1.—Relationship of interval-wise error rate (IWER) with comparison-wise error rate (CWER) for intervals with low (solid symbols; interval 2 on chromosome 2) and high (open symbols; interval 4 on chromosome 2) correlations between QTL coefficients at flanking markers. The dashed line corresponds to IWER = CWER.

fit. For IWER = 0.01, the average absolute error was 0.0007 or 7% and the maximum absolute error was 25%. Data for IWER < 0.01 were not used to develop the prediction equation because the number of permutations was limited to 10,000. Results displayed in Figure 1, which are based on 50,000 replicates, however, show that the linear prediction can readily be extended to IWER < 0.01.

False discovery rate (FDR): An example of the calculation of FDRm is in Table 3. For each interval, the IWER corresponding to the maximum F value was derived on the basis of the prediction equation. Tests were ranked by IWER and the 20 lowest tests are shown in Table 3. Although FDR generally increased with IWER, a steplike pattern was occasionally seen, where FDR decreased with error rate. This behavior is caused by disproportionate changes in the numerator and denominator of FDR when ranked tests differ little in IWER. When this occurs, the numerator of the expression for FDR, $N \times \text{IWER}_i$, remains the same, while the denominator, i, increases, leading to a reduction in FDR. For example, in going from rank 4 to 5 (Table 3), IWER increased from 0.00143 to 0.00150, while FDR decreased from 0.038 to 0.032.

The stepwise behavior of FDR is very apparent in Figures 2 and 3, which show FDR1 and FDRm, respectively, for last rib back fat and across the five traits. Steps were more pronounced for FDR1. For low CWER values, FDR1 increased dramatically with decreasing CWER. For example, FDR1 was 0.05 for CWER = 0.00012 and 0.19 for CWER = 0.00009 (Figure 2). This behavior is caused by the large number of tests included, combined with the small differences in CWER among the top ranking tests, which tend to originate from the same marker interval.

For a given CWER or IWER, FDR tended to be higher when based on all traits than on tests for last rib back fat alone (Figures 2 and 3). Single-trait results for marbling, loin eye area, and carcass weight (data not shown) were similar to those for back fat. For cholesterol content, both FDR1 and FDRm behaved erratically and never reached FDR levels <0.8.

Comparison of significance testing methods: *Single-trait analyses:* The CWER thresholds required to control GWER or FDR at the 0.10, 0.05, and 0.01 levels for different approaches are in Table 4. For FDRm, both IWER thresholds and the CWER for the associated tests are shown for completeness.

By definition, LK thresholds were the same for all traits (Table 4). Thresholds based on CD differed slightly by trait due to differences in phenotypic distributions and sampling, the latter in particular for the 0.01 GWER level. Thresholds for FDR varied considerably by trait and could not be found for all significance levels for some traits. This variability is caused by the specific CWER values obtained for the set of tests included in the analysis. Part of this variability may be due to the number of segregating QTL. Inability to obtain the target FDR level for a particular trait indicates that none of the tests were significant at that level.

The LK approach required the most stringent CWER thresholds, followed by CD and FDR (Table 4). The CWER thresholds for FDR1 and FDRm were generally similar but varied relative to each other. This variability is caused by the specific tests included in the analyses and by the dependence of CWER thresholds for FDRm on interval characteristics.

The CWER thresholds decreased with decreasing GWER or FDR levels for all methods (Table 4). Decreases in thresholds were relatively small in going from GWER = 0.10 to GWER = 0.05 and greater in going from GWER = 0.05 to GWER = 0.01. Thresholds for FDR decreased markedly in going from FDR = 0.05 to FDR = 0.01, coming close to those for CD. However, at this level of FDR, only two traits had tests that met the target FDR level.

Multiple-trait analyses: Testing for multiple traits decreased CWER thresholds five- to sixfold for both LK and CD (Table 4). Multiple-trait thresholds were also reduced for FDR1 and FDRm, when compared to the average CWER or IWER of single-trait thresholds, but less than for LK or CD. At the 0.10 level, multiple-trait thresholds were reduced only by a factor of 2.3 for FDRm and FDR1, compared to the fivefold reductions observed for LK and CD. At the 0.05 level, multiple-trait thresholds were reduced by a factor of 4 for FDRm and 10 for FDR1. The 0.01 level was not reached for the FDR approaches.

Number of QTL detected: The number of QTL declared significant on the basis of the various thresholds reported in Table 4 are listed in Table 5. Graphs of the test statistic are shown in MALEK *et al.* (2001a,b). In most cases peaks of the test statistic that exceeded significance thresholds spanned more than one consecutive marker

TABLE 3

Example computation of false discovery rate (FDRm) for determining significance thresholds for carcass weight on the basis of tests with the maximum *F*-statistic within each marker interval

Rank	Chromosome no.	Marker interval no.	Average correlation between breed origin coefficients	*F*-statistic (F_i)	Comparison-wise error rate ($CWER_i$)	Interval-wise error rate ($IWER_i$)	False discovery rate $FDRm_i$ ($N \times IWER_i / i$)
1	4	7	0.58	11.8	0.00001[a]	0.00004	0.004
2	4	6	0.83	8.6	0.00021	0.00045	0.024
3	7	5	0.80	7.7	0.00051	0.00107	0.038
4	4	5	0.72	7.5	0.00062	0.00143	0.038
5	7	6	0.62	7.6	0.00056	0.00150	0.032
6	8	2	0.55	7.3	0.00075	0.00212	0.038
7	8	3	0.72	6.9	0.00110	0.00244	0.037
8	4	4	0.68	6.7	0.00134	0.00310	0.041
9	7	1	0.46	6.9	0.00110[b]	0.00340	0.040
10	13	5	0.74	5.5	0.00433	0.00846	0.090
11	14	4	0.73	5.5	0.00433	0.00862	0.083
12	14	5	0.70	5.5	0.00433	0.00891	0.079
13	13	4	0.69	5.5	0.00433[c]	0.00900	0.073
14	3	3	0.51	4.9	0.00780	0.01887	0.143
15	3	5	0.65	4.7	0.00949	0.01937	0.137
16	7	4	0.60	4.6	0.01047	0.02238	0.148
17	3	4	0.96	4.1	0.01712	0.02437	0.152
18	13	6	0.60	4.5	0.01155	0.02452	0.144
19	3	6	0.37	4.5	0.01155	0.03092	0.172
20	12	1	0.32	4.5	0.01155	0.03263	0.173

[a] CWER threshold for FDR < 0.01.
[b] CWER threshold for FDR < 0.05.
[c] CWER threshold for FDR < 0.1.

interval on a chromosome. These cases were, however, recorded as evidence for a single QTL.

As anticipated, the number of detected QTL tended to be in proportion to the required CWER threshold. At the very stringent thresholds required by LK, only 3 QTL were uncovered with the single-trait analyses at the 0.10 GWER level, and 1 and 0 at the 0.05 and 0.01 levels, respectively. The CD approach allowed more QTL to be detected than did LK. Both FDRm and FDR1 performed distinctly better than either LK or CD. The FDRm uncovered a total of 17, 9, and 2 QTL at the 0.10, 0.05, and 0.01 FDR levels, respectively, for the single-trait analyses. FDR1 resulted in very similar numbers of QTL detected as FDRm.

When computed across traits, both LK and CD lost much of their power to detect QTL (Table 5). The FDR method maintained relatively high power at an FDR of 0.10 but not at the 0.05 and 0.01 levels. For FDR1 for marbling at the 0.10 level, more QTL were detected by the multiple-trait test (1) than on the basis of the single-trait test (0).

DISCUSSION

Implementation of FDR for interval mapping: With interval mapping, a test for presence of a QTL is typically conducted at each 1-cM position on the genome. This results in a large number of tests with very high correlations among tests at adjacent positions. Although FDR does not require independence of tests (Benjamini and Hochberg 1995), Figure 2 illustrates that including a large number of highly correlated tests in the analysis results in erratic and stepwise behavior of FDR, in particular for tests with the lowest CWER values. As suggested by Spelman (1998) this problem can be overcome in part by including only the highest test per marker interval (FDRm), as illustrated in Figure 3. Indeed, although FDRm still exhibited a noticeable stepwise pattern, the specific erratic behavior found for low CWER values with FDR1 was not present.

Despite their somewhat different behaviors, FDRm and FDR1 resulted in very similar CWER thresholds and numbers of QTL detected. This similarity is consistent with the theoretical argument that the proportion of false positives is independent of the number of tests included, provided prior probabilities of a true test and statistical power are unaffected (Southey and Fernando 1998). The small discrepancies that were observed between FDR1 and FDRm are caused by the dependence of FDR on the actual tests included, which are subject to some random noise. Thus, either FDR1 or FDRm can be used to implement FDR, although

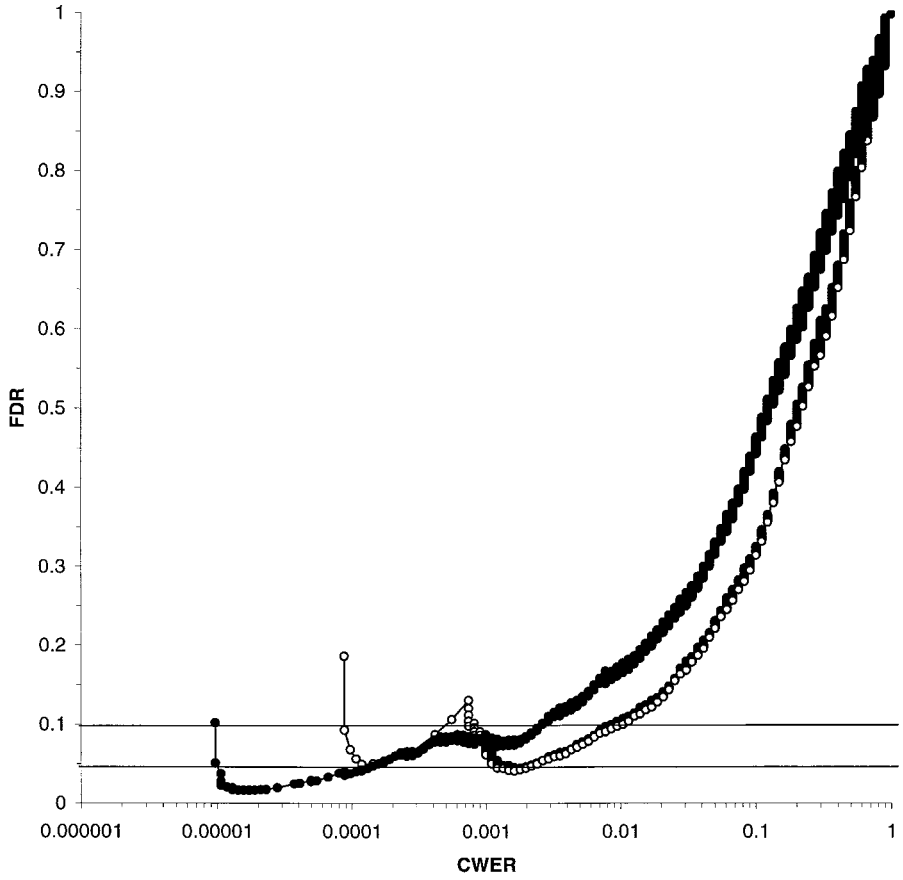

FIGURE 2.—False discovery rate based on a test for every 1-cM position (FDR1) plotted against the comparison-wise error rate (CWER) for last rib back fat (open symbols) and across five traits (solid symbols). Horizontal lines represent significance thresholds at the 5 and 10% levels.

FDRm is marginally preferred because of the observed erratic behavior of FDR1.

SPELMAN (1998) argued that FDRm would return to the same behavior as FDR1 as the number of markers included in genome scans increases and marker intervals become smaller. However, in efficiently designed experiments, large numbers of closely spaced markers would be used only if the number of meioses included

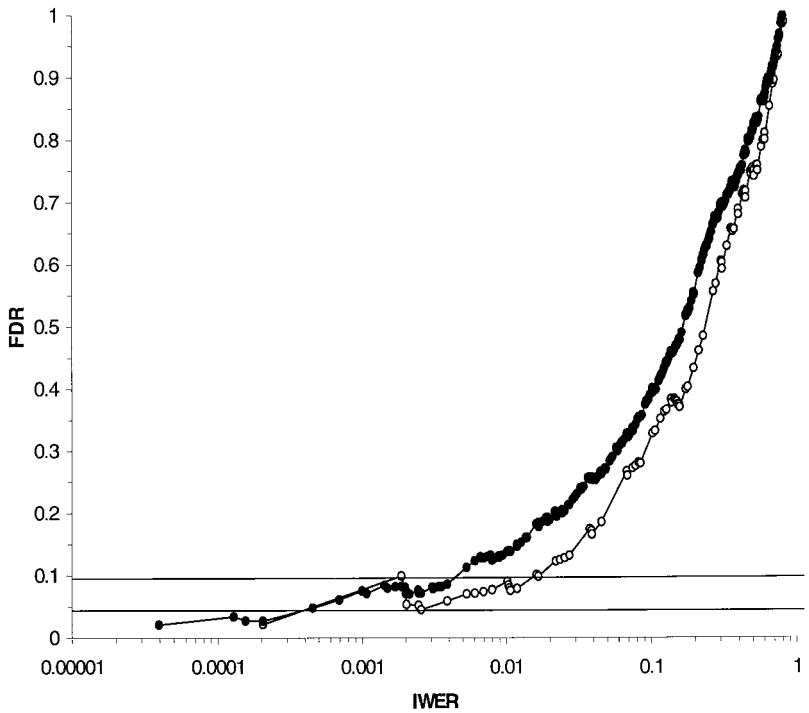

FIGURE 3.—False discovery rate based on a single maximum test per marker interval (FDRm) plotted against the interval-wise error rate (IWER; corrected for multiple testing in the interval) for last rib back fat (open symbols) and across five traits (solid symbols). Horizontal lines represent significance thresholds at the 5 and 10% levels.

TABLE 4

Comparison-wise error rate (CWER) thresholds to control genome-wise type I error rates (α_G) or false discovery rates (α_F) at 10, 5, or 1% for five traits separately or across the five traits based on four approaches

	$\alpha_G = \alpha_F = 10\%$				$\alpha_G = \alpha_F = 5\%$				$\alpha_G = \alpha_F = 1\%$			
Trait	LK	CD	FDRm	FDR1	LK	CD	FDRm	FDR1	LK	CD	FDRm	FDR1
Carcass wt.	0.000089	0.00068	0.00433[a] 0.00900[b]	0.00581	0.000034	0.00031	0.00110 0.00340	0.00180	0.000006	0.000055	0.000010 0.000039	0.000081
Back fat	0.000089	0.00062	0.00707 0.01666	0.00860	0.000034	0.00028	0.00100 0.00256	0.00218	0.000006	0.000050	—	—
Loin eye area	0.000089	0.00068	0.00163 0.00306	0.00148	0.000034	0.00031	0.00042 0.00100	0.00019	0.000006	0.000060	0.000067 0.000160	—
Cholesterol	0.000089	0.00062	—	—	0.000034	0.00028	—	—	0.000006	0.000055	—	—
Marbling	0.000089	0.00068	0.00068 0.00170	—	0.000034	0.00034	—	—	0.000006	0.000067	—	—
Across traits	0.000014	0.00012	0.00148 0.00389	0.00241	0.000007	0.00006	0.00026 0.00045	0.00014	0.000001	0.000011	—	—

LK, Lander and Kruglyak approach; CD, Churchill and Doerge permutation test; FDRm, false discovery rate based on the highest test per marker interval; FDR1, false discovery rate based on tests at every 1-cM position.

[a] Comparison-wise error rate (CWER) for the maximum test in the interval, based on the standard *F*-distribution for a single test, ignoring the effective number of tests in the interval.

[b] Interval-wise error rate (IWER) for the maximum test in the interval, corrected for effective number of tests in the interval.

TABLE 5

Number of QTL detected on the basis of significance thresholds at the genome-wise level based on different approaches using single- (ST) and multiple-trait (MT) testing procedures

		10%				5%				1%			
Trait		LK	CD	FDRm	FDR1	LK	CD	FDRm	FDR1	LK	CD	FDRm	FDR1
Carcass weight	ST	1	2	6	6	1	1	4	4	0	1	1	1
	MT	1	1	4	4	0	1	1	1	0	1	0	0
Last rib back fat	ST	1	2	8	7	0	1	3	4	0	0	0	0
	MT	0	1	4	4	0	0	1	1	0	0	0	0
Loin eye area	ST	1	2	2	2	0	1	2	1	0	1	1	0
	MT	0	1	2	2	0	1	1	1	0	0	0	0
Cholesterol	ST	0	0	0	0	0	0	0	0	0	0	0	0
	MT	0	0	0	0	0	0	0	0	0	0	0	0
Marbling score	ST	0	1	1	0	0	1	0	0	0	0	0	0
	MT	0	0	1	1	0	0	0	0	0	0	0	0
Total	ST	3	7	17	16	1	4	9	9	0	2	2	1
	MT	1	3	11	11	0	2	3	3	0	1	0	0

LK, Lander and Kruglyak approach; CD, Churchill and Doerge permutation test; FDRm, false discovery rate based the highest test per marker interval; FDR1, false discovery rate based on tests at every 1-cM position.

in the design is sufficient to allow high mapping resolution. In such experiments, correlations between tests in adjacent intervals would not be excessively high, even if the intervals are small, because sufficient recombinants would be present, and FDRm would behave as presented here. For experiments for which marker density is high relative to mapping resolution, maximum tests computed across several adjacent intervals could be included in FDRm to avoid excessive correlations.

Implementation of FDRm requires adjustment of the CWER for the multiple tests that are conducted within that interval. The IWER was introduced for these purposes. As demonstrated here, IWER can be derived with high accuracy from (i) a linear relationship between the logarithms of CWER and IWER for IWER < 0.4 and (ii) the dependence of the parameters of this linear relationship on the correlation between breed-origin coefficients at the flanking markers. Further work is needed to confirm these relationships for other designs.

Further development of FDR also requires accommodating the concerns of Zaykin *et al.* (2000) that the FDR is defined in an unconditional manner and cannot be used to control FDR conditional upon having declared one or more tests significant. Weller (2000) argued that the difference between the conditional and unconditional proportion of false positives will be minor if the probability of at least one significant test is high. In the present study, however, QTL were not detected for two of the five traits analyzed at FDR ≤ 0.1 (Table 5). Mosig *et al.* (2001) recently showed that the information required to control the conditional proportion of false positives can be obtained from the tests that are included in the analysis.

Because FDR has not yet been used widely for hypotheses testing, there is no consensus as to the appropriate levels of declaring significance of QTL. A limited number of studies have examined the impact of type I and type II errors on the efficiency of marker-assisted selection (Kashi *et al.* 1990; Moreau *et al.* 1997). The general conclusion from these studies is that in some circumstances increasing power to detect QTL is more important than reducing type I errors for maximizing response to marker-assisted selection. At present an FDR of 0.1 would appear conservative for marker-assisted selection. A more stringent FDR will be appropriate when QTL mapping is aimed at providing a platform for gene identification and positional cloning. In contrast, controlling GWER at levels of 0.05 to 0.01 always requires a very low CWER, irrespective of circumstances. This reduces the statistical power of the experiment and the potential response from marker-assisted selection.

Comparison of significance testing approaches: The main conclusion to be drawn from the results presented with regard to comparison of significance testing methods is that CWER significance thresholds at the same GWER or FDR levels differ substantially between methods (Table 4), leading to different numbers of QTL detected (Table 5). Specifically, FDR resulted in less stringent significance thresholds (Table 4) and in more QTL detected (Table 5), as compared to the GWER controlling methods. Compared to LK, the CD method resulted in less stringent thresholds (Table 4) and in more QTL detected (Table 5). Although these results may depend on the specific data set used, they illustrate several conceptual differences between approaches, as is discussed below.

Conceptual differences: Controlling type I error rate on the basis of a null hypothesis of zero effect is a well-accepted principle in statistical testing of scientific hypotheses. The GWER controlling methods of CD and LK attempt to extend this principle to multiple testing in a QTL scan by taking the null hypothesis of no QTL as valid for all tests conducted across the genome. This null hypothesis is, however, by definition false for traits that have been shown by prior biometrical analyses to have nonzero heritabilities. Instead, the statistical problem is to identify regions that harbor QTL *vs.* those that do not. The FDR approach deals directly and quantitatively with this challenge by controlling the proportion of false positives among all significant results. The GWER approaches deal with this only qualitatively, by controlling the probability that significant results include no more than one false positive.

The CD and LK approaches differ conceptually in the use of only tests based on the set of markers being analyzed in the given experiment for CD and consideration of all tests that would be conducted in a high-density marker map in the LK approach. This results in more stringent thresholds and fewer QTL detected for LK, as illustrated in Tables 4 and 5. The implications of this conceptual difference have been discussed previously (Lander and Kruglyak 1995, 1996; Witte *et al.* 1996). Thresholds for the FDR approach are in principle not affected by the number of markers included in the analysis. Thus, FDR thresholds derived for the current set of markers control the false discovery rate regardless of whether additional tests are conducted in the future.

Multiple-trait testing: Consideration of multiple traits leads to even more stringent significance thresholds based on GWER and further reduces the power to detect QTL, as demonstrated in Tables 4 and 5 for CD and LK. This is not necessarily the case for the FDR approach, provided the proportion of false positives among significant results is not affected by the number of tests. This relies on the condition that adding tests does not affect the prior probability of a true test or the average statistical power across tests (Southey and Fernando 1998). In this regard there is an important difference under FDR (but not under GWER) between adding markers or tests for a single trait *vs.* adding tests on other traits. Added tests for the same trait can be considered to represent a random sample from the same infinite pool and do not change the basic probabilities of a false discovery. In contrast, added traits can have very different QTL structures and heritabilities, ranging from traits with many detectable QTL to traits without detectable QTL. The potential impact of adding tests from other traits can be illustrated by considering the FDR column in Table 3. Increasing the number of traits will increase N in proportion, but if the added traits do not bring with them additional high F-values, CWER thresholds for a given FDR will decrease. Thus, compared to single-trait thresholds, consideration of multiple traits will result in more stringent thresholds and fewer QTL detected for traits with many detectable QTL. This is clearly shown in Table 4, where the nonsignificant tests for cholesterol and marbling increased the stringency of thresholds for the other three traits when considered in a multiple-trait scenario. As a result, in the multiple-trait test, the number of QTL detected for carcass weight and last rib back fat at the 10% level was reduced from 14 to 8 (Table 5). Paradoxically, as pointed out by Spelman (1998), when grouped with traits having many detectable QTL, tests for the traits having few or no QTL will be pushed down to a high rank number. This will tend to produce less stringent CWER thresholds for given FDR level and hence more QTL detected for these traits than when analyzed alone. This is seen in Table 5 for marbling at the FDR 0.10 level.

In principle, GWER controlling methods require pooling of traits in a single analysis, since they all share the same null hypothesis of zero QTL. This is not the case for FDR, since there is no prior assumption that traits have the same number of QTL. Furthermore, there is no advantage to losing power for a trait with many QTL from including tests for a trait with few QTL. Thus, for maximum power, FDR should be implemented for each trait separately.

Spelman (1998) extended this argument, proposing that chromosomes also should be analyzed separately, so that tests for chromosomes with few QTL do not dilute the power for tests on chromosomes with many detectable QTL. The situation for chromosomes, however, differs from that for traits because there is no *a priori* reason for the number of detectable QTL per map distance to differ by chromosome, whereas there are *a priori* differences between traits in terms of heritability. Thus, unless previous results indicate a preponderance of QTL on specific chromosomes, chromosomes should be analyzed jointly.

CONCLUSIONS

Our general conclusion is that FDR allows detection of more QTL and provides a more appropriate strategy for setting significance thresholds for QTL mapping than controlling GWER because it allows a means for controlling the proportion of true results among all those declared significant. From a conceptual point of view, this appears to be the most crucial error rate for follow-up studies or application, although further work is needed to clarify the impact of different types of errors and to address the concerns of Zaykin *et al.* (2000). Furthermore, although testing across traits is expected to have less impact on stringency of tests based on FDR, as compared to tests based on GWER, there is no theoretical reason for combining tests across traits with FDR.

Thus, with FDR traits can be analyzed separately, which will maximize power.

The authors are grateful to Daniel Nettleton, Dirk-Jan De Koning, and an anonymous reviewer for their input and critical review. This work was supported in part by a consortium of the National Pork Producers Council, Iowa Pork Producers Association, Iowa Purebred Swine Council, Babcock Swine, Danbred USA, DEKALB, PIC, Seghersgenetics USA, and Shamrock Swine Breeders. Additional support was from the Cooperative State Research, Education, and Extension Service, U.S. Department of Agriculture, under Agreement no. 00-52100-9610; the Iowa Agriculture and Home Economics Experimental Station, Ames, paper no. J-19082, project no. 3600; as well as Hatch and State of Iowa funds. Part of this work was completed while Dr. Soller was on leave at Iowa State University, supported by a visiting scientist fellowship provided by Cotswold Inc.

LITERATURE CITED

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

Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138:** 963–971.

Goodwin, R., and S. Burroughs, 1995 Genetic evaluation terminal line program results. National Pork Producers Council, Des Moines, IA.

Haley, C. S., and S. A. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69:** 315–324.

Haley, C. S., S. A. Knott and J. M. Elsen, 1994 Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics **136:** 1195–1207.

Kadarmideen, H. N., and J. C. M. Dekkers, 1999 Regression on markers with uncertain allele transmission for QTL mapping in half-sib designs. Genet. Sel. Evol. **31:** 437–455.

Kashi, Y., E. Hallerman and M. Soller, 1990 Marker-assisted selection of candidate bulls for progeny testing programmes. Anim. Prod. **51:** 63–74.

Lander, E. S., and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–190.

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

Lander, E. S., and L. Kruglyak, 1996 Genetic dissection of complex traits—correspondence. Nat. Genet. **11:** 241–247.

Lander, E. S., and N. S. Schork, 1994 Genetic dissection of complex traits. Science **265:** 2037–2048.

Malek, M., J. C. M. Dekkers, H. K. Lee, T. J. Baas and M. F. Rothschild, 2001a A molecular genome scan analysis to identify chromosomal regions influencing economic traits in the pig. I. Growth and body composition. Mamm. Genome **12:** 630–636.

Malek, M., J. C. M. Dekkers, H. K. Lee, T. J. Baas and M. F. Rothschild, 2001b A molecular genome scan analysis to identify chromosomal regions influencing economic traits in the pig. II. Meat and muscle composition. Mamm. Genome **12:** 637–645.

Moreau, L., A. Charcosset, F. Hospital and A. Gallais, 1997 Marker-assisted selection efficiency in populations of finite size. Genetics **148:** 1353–1365.

Mosig, M. O., E. Lipkin, G. Khutoreskaya, E. Tchouryzna, E. Ezra *et al.*, 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.

Southey, B., 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.

Spelman, R. J., 1998 Detection and utilisation of quantitative trait loci in dairy cattle. Ph.D. Thesis, Wageningen Agricultural University, Wageningen, The Netherlands.

Spelman, R. J., W. Coppieters, L. Karim, J. A. M. van Arendonk and H. Bovenhuis, 1996 Quantitative trait loci analysis for five milk production traits on chromosome six in the Holstein-Friesian population. Genetics **144:** 1799–1808.

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.

Weller, J. I., J. Z. Song, D. W. Heyen, H. A. Lewin and M. Ron, 1998 A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. Genetics **150:** 1699–1706.

Whittaker, J. R., R. Thompson and P. M. Visscher, 1996 On the mapping of QTL by regression of phenotype on marker-type. Heredity **77:** 23–32.

Witte, J. S., R. C. Elston and N. J. Schork, 1996 Genetic dissection of complex traits—correspondence. Nat. Genet. **12:** 365–366.

Zaykin, D. V., S. S. Young and P. H. Westfall, 2000 Using the false discovery rate approach in the genetic dissection of complex traits: a response to Weller *et al.* Genetics **154:** 1917–1918.

Communicating editor: J. A. M. van Arendonk