# Note

## False Discovery Rate in Linkage and Association Genome Screens for Complex Disorders

**Chiara Sabatti,*,[1] Susan Service† and Nelson Freimer†**

*Departments of Human Genetics and Statistics, University of California, Los Angeles, California 90095-7088 and
†Center for Neurobehavioral Genetics, Neuropsychiatric Institute, and Departments of Psychiatry and
Human Genetics, University of California, Los Angeles, California 90095-1761

Manuscript received January 9, 2003
Accepted for publication February 21, 2003

ABSTRACT

We explore the implications of the false discovery rate (FDR) controlling procedure in disease gene mapping. With the aid of simulations, we show how, under models commonly used, the simple step-down procedure introduced by Benjamini and Hochberg controls the FDR for the dependent tests on which linkage and association genome screens are based. This adaptive multiple comparison procedure may offer an important tool for mapping susceptibility genes for complex diseases.

RECENT developments in methods for controlling for multiple comparisons in statistical testing may strongly influence strategies for mapping disease genes. The importance of correcting for multiple comparisons in genome screens is well known (Lander and Kruglyak 1995). The current literature is based on controlling the probability of making at least one false rejection [known as the family-wise error rate (FWER)]. The global significance threshold obtainable under such a paradigm with either a sparse map assumption or a continuous map assumption (Lander and Botstein 1989; Lander and Kruglyak 1995), while appropriate to evaluate the locus with the strongest evidence, is, however, too stringent for evaluating multiple loci. With this in mind, conditional and simultaneous searches have been proposed. Unfortunately, conditional and simultaneous linkage analyses, when appropriately correcting for multiple comparison in a FWER framework, have very low power—so that marginal search is still preferable (Lander and Botstein 1986; Dupuis *et al.* 1995).

In 1995, Benjamini and Hochberg introduced the false discovery rate (FDR), a new notion of global error for multiple testing situations. The idea of FDR is to use, as a measure of global error, the expected proportion of false rejections of the null hypothesis among the total number of rejections. The use of the proportion of type I errors among the total number of "significant" results leads to a global cutoff value that is *adaptive* to the data set. That is, if a higher percentage of the null hypotheses tested are truly false, the FDR procedure will identify a lower cutoff level than the universal Bonferroni cutoff. Therefore, FDR defines an *adaptive marginal search*, which is most effective for the identification of loci with secondary effects. On the other hand, if all the null hypotheses are true (none of the analyzed markers is linked with the disease), controlling FDR is equivalent to controling FWER, as in the Lander and Kruglyak (1995) criteria. A number of results now show that the FDR's measure of global error and the multiple comparison procedure it implies are *optimal* in that one can describe statistical problems for which they represent the asymptotic minimax solution (see Abramovich *et al.* 2000; D. Donoho and J. Jin, unpublished results).

The first work to suggest applying the FDR procedure in genetic mapping (Weller *et al.* 1998) recognized its value for investigating multiple quantitative trait loci (QTL). We focus here on qualitative traits, as investigated through either linkage or association studies. Moreover, recent theoretical advances in the statistical theory of FDR allow us to show that a simple thresholding rule originally proposed by Benjamini and Hochberg (1995) controls FDR in the context of dependent tests, which are typical for genome scans.

To illustrate both the FDR approach and the implications of these novel findings, consider a linkage genome screen done under the sparse map assumption and based on $n = 400$ markers. Let $H_i$, $i = 1, \ldots, n$, be the null hypothesis of no linkage with the *i*th marker and let $H_0 = \cap_{i=1}^{n} H_i$. Let $p_1, \ldots p_n$ be the $P$ values associated with each of the test statistics ($n = 400$) and let $p_{(1)} \leq \ldots \leq p_{(n)}$ be their ordered counterpart. According to the Bonferroni rule, one can reject $H_0$ if $p_{(1)} < \alpha/n$,

[1] *Corresponding author:* Department of Human Genetics, UCLA School of Medicine, 695 Charles E. Young Dr. S., Los Angeles, CA 90095-7088. E-mail: csabatti@mednet.ucla.edu

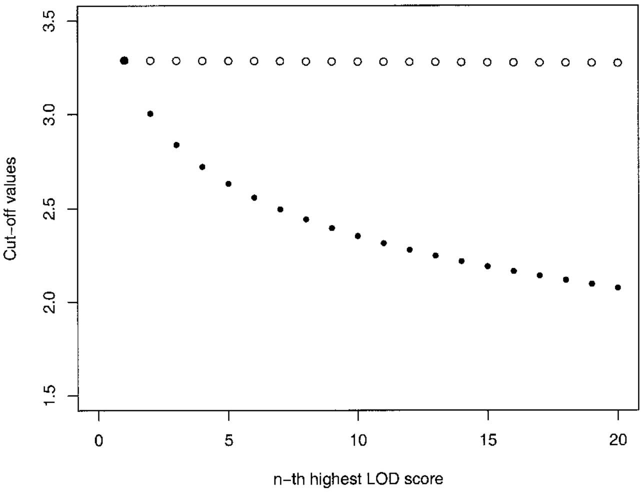

Figure 1.—Significance cutoff of the markers' lod scores. We considered a genome screen involving 400 markers leading to independent tests (sparse map assumption) and we plotted the cutoff values for significance for the highest lod score values in decreasing order. The open circles correspond to a procedure that controls FWER and the solid circles correspond to the BH that controls FDR at the 0.05 level. Note that for FDR the stepwise procedure described in the text has to be followed.

where $\alpha$ is the desired level for the test of $H_0$; for $\alpha = 0.05$, this translates to a lod score of 3.3. Benjamini and Hochberg (BH) proposed the following stepwise procedure: proceed from $i = 1$ to $i = 2, \ldots, n$ until, for the last time, $p_{(i)} \leq i \cdot \alpha/n$. Denote this index by $k$ and reject all $H_{(i)}$ with $i = 1, \ldots, k$. The decreasing cutoff values can be translated in a series of decreasing lod scores, shown in Figure 1: if the locus with the highest lod score has to pass a 3.3 threshold to be significant alone, the second locus is compared with a score of 3, and the third locus with a threshold of 2.8.

While the rule described above (BH) was proposed by Benjamini and Hochberg for independent tests—a sparse map assumption in the context of linkage—recent theoretical developments assure that it can control FDR even in the case of dependent tests, as in the continuous-map assumption. Weller *et al.* (1998) considered the context of QTL analysis, which is based on statistics different from those of the analysis used in genome scans of diseases; they showed with simulations that the described BH procedure controlled FDR in that context. It is now possible to understand more generally why this should be the case and to prove it for the context of linkage genome scans for complex diseases, as well as for genomic association studies—at least as it appears from simulations.

To achieve this, one must consider the result of Benjamini and Yekutieli (2001). They show that BH controls FDR at a level $\alpha \cdot n_0/n$, where $n_0$ is the number of false null hypotheses, if the test statistics are positive-regression dependent on each one from the subset (PRDS) of test statistics corresponding to the true null hypotheses. Technically, the definition of PRDS is as follows. The set $D$ is called increasing if $x \in D$ and $y \geq x$ imply that $y \in D$ as well. The random variables $X_1, \ldots, X_n$ are PRDS on $I_0$ if, for any increasing set $D$, and for each $i \in I_0$, $P(X_1, \ldots, X_n \in D | X_i = x)$ is nondecreasing in $x$. This definition is a specific formal requirement for what we may call "positive dependence," and in the context of genome screens, PRDS can be loosely interpreted as follows: if two markers are linked [or in linkage disequilibrium (LD)] and neither is related to the disease, the $P$ values of the tests conducted at each marker tend to be positively correlated—as one would expect.

In the case of linkage we can prove that the lod score statistic (or a specific approximation of it) satisfies the PRDS requirement. In the case of association studies, we illustrate the meaning of PRDS with respect to a specific model and conduct a simulation study.

**Significance cutoffs for a linkage genome screen under FDR:** Using a specific approximation of the lod score statistics, one can show that they satisfy the PRDS requirement. Consider the Gaussian models for genetic linkage analysis proposed by Feingold *et al.* (1993) in the specific case of grandparent-grandchild pairs. If we restrict our attention to a finite subset of genome locations, the linkage tests have a multivariate Gaussian distribution. The mean value of the test statistic at each unlinked location is 0 and it is positive for linked loci. The covariances between two test statistics are nonnegative and are functions of the recombination fraction across loci. Benjamini and Yekutieli (2001) show that PRDS translates in the following requirement for multivariate normal tests statistics. Let $X \sim N(\mu, \Sigma)$ be a vector of test statistics, each testing the hypothesis $H_i$ that $\mu_i = 0$ against the alternative $\mu_i > 0$, for $i = 1, \ldots, m$. For $i \in I_0$, the true set of null hypotheses, $\mu_i = 0$; otherwise $\mu_i > 0$. If for each $i \in I_0$, and for each $j \neq i$, $\sigma_{ij} \geq 0$, then the distribution of $X$ is PRDS over $I_0$. We can conclude, because the covariances are nonnegative, that the tests are PRDS on $I_0$. Hence the cutoff values illustrated in Figure 1 are guaranteed to control the FDR, even when we relax the independence assumption.

**Significance cutoffs for an association genome screen under FDR:** Depending on the population of origin of the sample and on the distance between markers and their location in the genome, association tests at different markers either may be independent or may display varying degrees of dependence. In the case of independence between markers, we know that the BH rule controls for FDR. In the case of dependent markers, intuition suggests that PRDS should hold and hence the BH rule should also control FDR. In the absence of a general model for dependency among association tests, we illustrate that PRDS should hold by using a simple

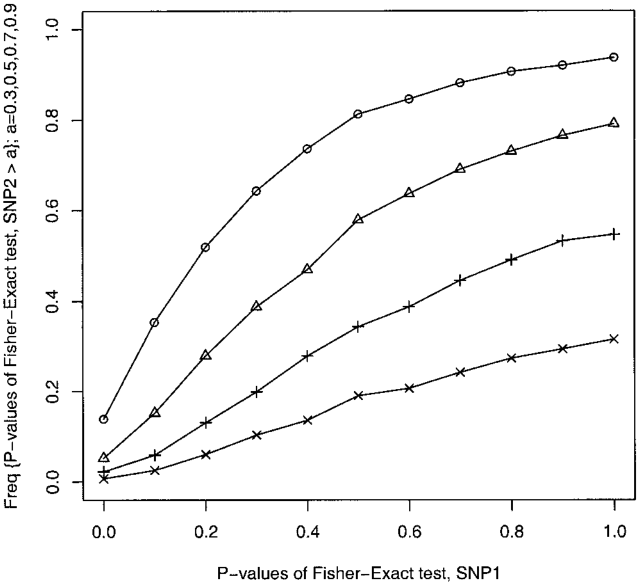

Figure 2.—Positive regression dependency on a subset for the $P$ values of Fisher's exact test between two dependent SNPs in LD and a disease status variable. Let $X_1$ be the total number of allele 0 associated with disease at the first marker and $Y_1$ the total number of allele 0 associated with control at the same marker in the obtained sample. Suppose that there is no association between the SNPs under consideration and the disease. Then $X_1 \sim \text{Binom}(p, N)$ and $Y_1 \sim \text{Binom}(p, N)$, which entirely specifies the distribution of the contingency table $T_1$, collecting allele counts for SNP1 and disease status. To evaluate the distribution of the contingency table $T_2$ for SNP2 and disease status, given $T_1$, it is sufficient to calculate the distribution of $X_2$, $Y_2$ (number of allele 1 in the second SNP associated with disease and control) given $X_1$ and $Y_1$. This is such that both can be viewed as the sum of two independent binomial components: $X_2 \sim \text{Binom}(q + \delta/p, x_1) + \text{Binom}(q - \delta/(1 - p), N - x_1)$ and $Y_2 \sim \text{Binom}(q + \delta/p, y_1) + \text{Binom}(q - \delta/(1 - p), N - y_1)$. As an example, we generated 50,000 tables using two SNPs, each with allele frequency 0.5 and a background linkage disequilibrium parameter $\delta = 0.2(-0.2)$. There was no association between a disease locus and the considered SNPs. We report the empirical equivalent of $\Pr(p_2 > a|p_1 = b)$, for $a = 0.3$ (circles), 0.5 (triangles), 0.7 (+), and 0.9 (x) and $b = 0.1, \ldots, 0.9$, where $p_1$ and $p_2$ are the $P$ values of the tests (values are grouped in bins of size 0.1). Results for $\delta = -0.2$ are comparable.

example and by testing the performance of the BH rule directly with simulations.

Suppose we conduct a screen using $N$ cases and $N$ controls and testing for association at two loci with single-nucleotide polymorphisms (SNPs). Their joint distribution can be represented by the parameters $\Pr(\text{SNP}_1 = 0) = p$, $\Pr(\text{SNP}_2 = 0) = q$, and $\Pr(\text{SNP}_1 = 0 \text{ and } \text{SNP}_2 = 0) = pq + \delta$. If we focus on the distribution of the test of association derived from these two markers, it is possible to see, through computations, that the distribution of their $P$ values ($p_1$, $p_2$) has a property required by PRDS (see Figure 2). While this case of two markers serves as an illustration of the implications of PRDS and why it is reasonable to think that it should hold, a genome screen clearly involves more than two markers. We should investigate PRDS more carefully, for all combinations of multiple statistics—a task clearly impossible in the absence of a simpler model for their dependency. We therefore resorted to a simulation study.

For brevity we report here only on the simulations for haplotype-based tests. The results of single-marker-based tests are totally comparable and are available in a companion technical report (Sabatti *et al.* 2002). We consider a situation where we have three susceptibility genes of equal importance, acting independently and located on three different chromosomes. We evaluated a sample of 200 diseased haploid individuals and 200 controls. One-third of the 200 diseased individuals were carriers of one of the three disease loci. Details of the simulation settings are shown in the Table 1 legend.

Under nearly all simulation conditions reported in Table 1, the BH method achieves control of the FDR: the average estimated FDRs are <0.05. In the high-power, high-dependence setting for the haplotype test, the average FDR is 0.052. Under the high-dependence settings, the variance of the results is considerably greater than that in the other scenarios and indeed, 0.05 is within the 95% bootstrap confidence interval for all the scenarios. The BH FDR method controls only the expected value of the FDR, so that in one replicate the FDR may actually be higher (data not shown). In general, even under high levels of dependency, then, the FDR is controlled at the appropriate level, as suggested by our previous analysis. The column named "no. of false positives" reports the average number of false positives per replicate; while this number is larger using BH than using Bonferroni, it is still ⩽1, indicating that indeed we do not need to be as strict as FWER procedures suggest.

The FDR method leads to a considerable increase in power when compared with FWER. On average, the marginal power estimates of FDR are 25% greater than those of FWER. The increased power of FDR over FWER is even more dramatic when one requires that multiple loci be detected, with the increase in power for identifying all three loci averaging >100%. The power for all methods decreases with increased dependency between markers. This decrease is apparent not only in the power estimates themselves, but also by the fact that both FDR and FWER are usually controlled at a level ⩽0.05 (the cutoff we had specified). This result argues in favor of the necessity of developing adequate resampling-based evaluations of FDR so that the dependence between markers is incorporated to increase the power of the study. This is the goal of a separate investigation.

We have also applied the BH procedure to a recently collected data set from a genome-wide LD study of a complex trait, bipolar disorder (Table 2; Ophoff *et al.* 2002). If we aim at controlling the global error at the standard 0.05 level, FDR accepts two locations as possibly

**TABLE 1**

**Simulation results for the haplotype tests**

| | Power | | | | | | False rejections | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Marginal power | | Power ≥2 | | Power 3 | | No. false positives | | FWER est. | | FDR est. | |
| LD values | Low | High | Low | High | Low | High | Low | High | Low | High | Low | High |
| Low power | | | | | | | | | | | | |
| No MCP | 0.997 | 0.989 | 1 | 0.999 | 0.999 | 0.969 | 53.3 | 52.7 | 1 | 1 | 0.81 | 0.80 |
| Bonferroni | 0.437 | 0.336 | 0.407 | 0.267 | 0.087 | 0.043 | 0.035 | 0.047 | 0.03 | 0.04 | 0.014 | 0.023 |
| BH FDR | 0.563 | 0.441 | 0.603 | 0.452 | 0.245 | 0.149 | 0.223 | 0.239 | 0.16 | 0.12 | 0.044 | 0.028 |
| High power | | | | | | | | | | | | |
| No MCP | 0.999 | 0.997 | 1 | 1 | 0.999 | 0.992 | 53.2 | 53.2 | 1 | 1 | 0.78 | 0.78 |
| Bonferroni | 0.779 | 0.566 | 0.88 | 0.596 | 0.468 | 0.177 | 0.06 | 0.06 | 0.052 | 0.042 | 0.011 | 0.016 |
| BH FDR | 0.904 | 0.715 | 0.961 | 0.788 | 0.757 | 0.412 | 0.44 | 0.43 | 0.30 | 0.22 | 0.048 | 0.052 |

The genome is considered to be 3300 cM long and organized in 22 chromosomes of equal length. One-third of the 200 diseased individuals were carriers of one of the three disease loci. We assumed that 1100 markers, each 3 cM apart, covered the genome. This is approximately the density used in the simplest linkage disequilibrium (LD) screens, those conducted in recently founded population isolates (Ophoff *et al.* 2002, for example). For computational simplicity we considered SNPs rather than microsatellite markers. The distribution of the closest recombination events on the two sides of the disease locus in disease gene-carrying chromosomes was evaluated assuming a founding event 15 generations ago. Outside the conserved region, the disease chromosomes were modeled as control chromosomes with a Markov process of the first order. To cover an interesting range of settings, we considered two degrees of LD among adjacent SNPs and two levels of power. The levels of LD are described by the parameter $\lambda$ as in $\Pr(\mathrm{SNP}_i = 1|\mathrm{SNP}_{i-1} = 1) = \Pr(\mathrm{SNP}_i = 1) + \lambda \Pr(\mathrm{SNP}_i = 2)$ and $\Pr(\mathrm{SNP}_i = 1|\mathrm{SNP}_{i-1} = 2) = \Pr(\mathrm{SNP}_i = 1) - \lambda \Pr(\mathrm{SNP}_i = 2)$. Linkage equilibrium corresponds to $\lambda = 0$. Low LD indicates a scenario where $\lambda$ varies uniformly in $[0, 0.1]$; medium LD is characterized as $\lambda \in [0.2, 0.4]$; and high LD is $\lambda \in [0.8, 1]$. The different power levels are achieved making differential assumptions on the frequencies of the alleles associated with disease loci on the founder chromosomes. (The frequencies of the alleles associated with disease vary uniformly between 0.3 and 0.4 for the high-power case and between 0.4 and 0.5 for the low-power case.) We simulated 1000 samples. We used Fisher's exact *P* values for the association tests. We measured power in three different ways, to emphasize how the gains from FDR are mainly in identifying multiple loci. We also evaluated the average number of false rejections per genome screen and the average FDR and the average FWER across replicates. We considered as true null hypotheses all the $H_0$ relative to markers that are more than three SNPs away from the true disease locus and that are on chromosomes that do not carry any disease locus. The "marginal power" column contains the average percentage of times in which each of the three disease loci was detected (at the 0.05 level after correction by FDR or FWER). The percentage of times in which at least two of the three locations were detected is reported in "power ≥2" and the percentage of times in which all three locations were detected in "power 3."

associated with the disease, while FWER accepts only one possible location. While at this stage in the investigation it is not possible to determine if either of these loci will lead to the identification of a disease gene, considerable circumstantial evidence supports the importance in psychiatric diseases of the locus on chromosome 8, which passes the FDR threshold, but not the FWER one (see the discussion in Ophoff *et al.* 2002). This result suggests that it is immediately useful to implement the FDR for determining significance thresholds of genome-wide LD analysis.

**Is FDR the appropriate measure of global error for disease gene mapping?** The FDR is a powerful, relatively novel measure of global error in multiple testing. The BH controlling strategy is simple and effective in a wide range of circumstances and in particular for disease mapping, as we illustrate. While extensively used in gene expression studies (see, for example, Efron *et al.* 2001), until now, disease mapping efforts have ignored FDR methods; this may be due to the already sobering frequency of "false positives" in the field. While we think this is certainly an issue, we believe that the problem derives more from liberal interpretations of multiple comparison procedures than from the fact that the adopted threshold may be too low; multiple genome screens are, for example, often conducted on the same data, using different hypotheses, to report only the "best" results. Moreover, the BH procedure would affect mainly the way in which we evaluate the significance when multiple loci play a role in the disease. Because of the known lack of power in this case of the Lander and Kruglyak (1995) recommendations, researchers tend to be too liberal in reporting such results. A simple and clear procedure such as BH should streamline the evaluation of the significance of secondary loci and possibly reduce the number of false positives reported. Additionally, determining what measure of global error is appropriate depends crucially on the costs associated with a false positive. While the publicity effect of announcing a spurious finding is constant, the costs of pursuing a wrong

TABLE 2

**Corrected *P* values of a genome screen for bipolar disorder**

| | *P* values | | |
|---|---|---|---|
| Markers | Uncorrected | Corrected FWER | Corrected FDR |
| D2S156 | 1.11*e*-11 | 1.05*e*-09 | 1.05*e*-09 |
| D8S503 | 5.66*e*-05 | 0.0524 | 0.02692 |
| D9S257 | 0.00032 | 0.2667 | 0.07608 |
| D17S1529 | 0.00038 | 0.30189 | 0.07608 |
| D2S325 | 0.0044 | 0.31640 | 0.07608 |
| D8S520 | 0.00089 | 0.57318 | 0.13429 |
| D1S2841 | 0.00099 | 0.60939 | 0.134629 |
| D2S369 | 0.00122 | 0.68602 | 0.14480 |
| D17S788 | 0.0019 | 0.83568 | 0.20066 |
| D2S2241 | 0.0027 | 0.92581 | 0.26011 |

The data (described in the quoted article) were analyzed using the ancestral haplotype reconstruction test (Service *et al.* 1999) on 952 three-marker windows, sliding through the genome. We corrected the *P* values associated with these tests using both the described step-down FWER and FDR controlling procedures. Results are reported for the 10 three-marker windows that gave the most significant results. The effect of the multiple-comparison procedure is illustrated with corrected *P* values: for each three-marker window we report the level of global (genome-wide) significance of the association result, as in Weller *et al.* (1998).

lead with further studies have diminished in recent years. The availability of the complete genome sequence, of a dense map of markers that can be used to further investigate each region, and the possibility of doing interspecies comparison for further studies all reduce the time, efforts, and costs associated with the verification of an initial finding. These considerations should be incorporated in the choice of global error measure, and we believe that they support the use of FDR.

In addition to what we have illustrated in this article, the application of FDR in genome screens has other advantages. It is becoming clearer that to identify the genes responsible for complex disease, a variety of strategies will need to be employed. Multiple phenotypes may be analyzed at the same time; for example, expression levels of candidate genes may be monitored together with the analysis of genotypes at thousands of loci. The BH strategy illustrated in this article can be easily adapted to control for multiple comparisons in these more diverse settings.

C. Sabatti acknowledges support from National Institutes of Health (NIH) grant RO1 MH49499 and S. Service and N. Freimer from NIH grants RO1 MH49499, R01 NS 37484, R01 NS40024, R01 HL66289, and K02 MH01375.

## LITERATURE CITED

Abramovich, F., Y. Benjamini, D. Donoho and I. Johnstone, 2000 Adapting to unknown sparsity by controlling the false discovery rate. Technical Report 2000-19. Department of Statistics, Stanford University, Stanford, CA.

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

Benjamini, Y., and D. Yekutieli, 2001 The control of the false discovery rate in multiple testing under independence. Ann. Stat. **29:** 1165–1188.

Dupuis, J., P. O. Brown and D. O. Siegmund, 1995 Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent. Genetics **140:** 843–856.

Efron, B., R. Tibshirani, J. Storey and V. Tusher, 2001 Empirical Bayes analysis of a microarray experiment. J. Am. Stat. Assoc. **96:** 1151–1160.

Feingold, E., P. O. Brown and D. Siegmund, 1993 Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. Am. J. Hum. Genet. **53:** 234–251.

Lander, E., and D. Botstein, 1986 Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms. Proc. Natl. Acad. Sci. USA **83:** 7353–7357.

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

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

Ophoff, R., M. Escamilla, S. Service, M. Spesny, D. Meshi *et al.*, 2002 Genome-wide linkage disequilibrium mapping of severe bipolar disorder in a population isolate. Am. J. Hum. Genet. **71:** 565–574.

Sabatti, C., S. Service and N. Freimer, 2002 *UCLA Statistical Department Preprint.* University of California, Los Angeles.

Service, S., D. Temple Lang, N. Freimer and L. Sandkuijl, 1999 Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. Am. J. Hum. Genet. **64:** 1728–1738.

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.

Communicating editor: G. Churchill