NEWS AND COMMENTARY

Statistical Genetics

False discovery or missed discovery?

B Devlin, K Roeder and L Wasserman

Heredity (2003) **91**, 537–538, advance online publication, 22 October 2003;
 doi:10.1038/sj.hdy.6800370

The Human Genome Project ushered the era of big science into human genetics. For some time, scientists have scanned the genome for evidence of linkage between widely spaced markers and disease phenotypes, but now tens of thousands of loci are being queried for their association with phenotypes related to complex disease, and thousands of genes are evaluated for differential expressions in different tissues or types of tumors. In the genomics arena, hypothesis tests to identify widespread genomic features are commonplace. In a recent article in *Proc Natl Acad Sci USA*, Storey and Tibshirani (2003) point to a unifying principle in many of these investigations, namely a large number of features are tested, often independently, against a simple null hypothesis. When confronted with such a vast number of hypothesis tests and the potential for numerous false positives, the traditional statistical approach is to impose a penalty to account for multiple testing, such as the Bonferroni correction. As Storey and Tibshirani point out, that penalty can be far too severe, especially so when it is likely that many of the alternatives are true.

In response to the new scientific environment, a competing approach to multiple testing has come to light that is based on controlling the false discovery rate (FDR). As noted by Storey and Tibshirani, this approach is appropriate when the goal of the experiment is 'to identify as many significant features in the genome as possible, while incurring a relatively low proportion of false positives'.

The FDR approach is due to Benjamini and Hochberg (1995, 2000). Hypothesis testing is concerned with deciding between two hypotheses: a null hypothesis (the *status quo*) and the alternative hypothesis. The data analyst evaluates a test statistic T and then finds the P -value, the probability of getting a test statistic as or more extreme than T under the assumption that the null hypothesis is true. Small P -values are taken as evidence that the null hypothesis

is false. When testing multiple hypotheses, the situation is much more complicated.

There are several ways to measure the error rate (Table 1). Let m be the number of tests, F be the number of false positives, T be the number of true positives, and $S = F + T$ be the total number of discoveries, or rejected null hypotheses. The common error measures are

PCER = $E(F)/m$	per comparison error rate
FWER = $P(F \geq 1)$	family wise error rate
FDR = $E(F/S)$	false discovery rate

in which P means probability and E means average (or expected) value.

Traditionally, FWER has been the quantity of interest. However, when

thousands of hypotheses are tested, a noteworthy fraction of which are expected to be true, it makes more sense to control the FDR. The Benjamini–Hochberg (BH) method for controlling FDR works as follows. First, order the m P -values from the tests: $P_1 \leq P_2 \leq \dots \leq P_m$. Now determine a threshold value for rejection by finding the largest integer j such that $P_j \leq j\alpha/m$. Finally, reject any hypothesis whose P -value is smaller than or equal to P_j (Figure 1). Remarkably, Benjamini and Hochberg (1995) prove that this procedure is guaranteed to force $FDR \leq \alpha$.

False discovery rates have received much recent attention in the statistics literature (Benjamini and Hochberg, 1995, 2000; Yekutieli and Benjamini, 1999; Abramovich *et al.*, 2000; Benjamini and Yekutieli, 2001; Genovese and Wasserman, 2001, 2002, 2003; Storey, 2002, 2003; Storey and Tibshirani, 2001, 2003; Finner and Roters, 2002; Storey *et al.*, 2003; Sarkar, 2002). Storey (2003) pioneered a new way to look at the FDR procedure that continues to control the false discovery rate, but attains somewhat more power to detect true discoveries. In keeping with the spirit of the P -value, Storey (2003) proposed a

Table 1 Outcomes from m hypothesis tests

	Called significant	Called not significant	Total
Null true	F	$m_0 - F$	m_0
Alternative true	T	$m_1 - T$	m_1
Total	S	$m - S$	m

Figure 1 Illustration of the BH procedure. The line labeled α is the prespecified threshold of significance for a single test, the line labeled T_{BH} is the threshold for FDR, and the line labeled α/m is the threshold for FWER.

quantity known as the q -value, which can be calculated for each feature under investigation. The q -value for a particular feature is the proportion of false positives incurred, on average, when that feature defines the threshold value. So, for example, if the investigator decides to perform follow-up investigations on 100 genetic features yielding the smallest P -values, and the q -value for the largest of these discoveries is 0.03, then, on average, she can anticipate that three of these features will be false positives.

Storey and Tibshirani (2003) introduce and explore the use of the q -value in a broad spectrum of problems, showing how these truly big problems in genetics are handled much more coherently by the q -value than by more traditional or even *ad hoc* P -value approaches. Storey and Tibshirani's most notable result pertains to the differential expression of genes in tumors induced by mutations in BRCA1 versus BRCA2. By examining their data on expression of 5361 genes from BRCA1- and BRCA2-positive tumors, Hedenfalk *et al* (2001) argue that nine and 11 genes show distinct expression in BRCA1- and BRCA2-positive tumors, and they identify 176 genes as being the most important discriminators of the tumor types. By contrast, Storey and Tibshirani's analysis using the q -value suggests that over a thousand

of those genes are differentially expressed.

In other developments pertaining to genetic modeling, Devlin *et al* (2003) show that FDR methods can be successfully implemented in a model selection framework. Specifically, in the linear model $Y = \beta_0 + \sum_j \beta_j X_j + \text{error}$, model selection is concerned with deciding which covariates X_j should be retained in the model. The most common model selection strategy is stepwise regression. This strategy can lead to a model with good predictive ability, but it does not control the number of false terms included in the model. Devlin *et al* show that FDR can be applied by computing a P -value for each term in the model, even interactions. This approach controls the rate of falsely included terms and, as shown by M. Wegkamp and F. Bunea (personal communication), will lead to the discovery of the true model for sufficiently large data sets. Sabatti *et al* (2003) explore the behavior of FDR for genetic linkage analysis, and find the tools of Benjamini and Yekutieli (2001) perfect for the job.

Model selection problems figure prominently in the dissection of the genetic basis of complex disease, and massive data explorations figure prominently in genomic analysis. Thus, tools related to the FDR concept promise to play a critical role in the future of genetic research.

This work was supported by MH57881. For software to compute q -values, see <http://genomine.org/qvalue/>.

Bernie Devlin is at the Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA; and Kathryn Roeder and Larry Wasserman are at the Department of Statistics, Carnegie Mellon University, USA

e-mail: roeder@stat.cmu.edu

- Abramovich F *et al* (2000). Technical Report 2000-19. Department of Statistics, Stanford University.
- Benjamini Y, Hochberg Y (1995). *J R Statist Soc B* **57**: 289–300.
- Benjamini Y, Hochberg Y (2000). *J Educ Behav Statist* **25**: 60–83.
- Benjamini Y, Yekutieli D (2001). *Ann Statist* **29**: 1165–1188.
- Devlin B *et al* (2003). *Genet Epidemiol* **25**: 36–47.
- Finner H, Roters M (2002). *Ann Statist* **30**: 220–238.
- Genovese CR, Wasserman L (2001). Technical Report. Department of Statistics, Carnegie Mellon University.
- Genovese CR, Wasserman L (2002). *J R Statist Soc B* **64**: 499–518.
- Genovese CR, Wasserman L (2003). *Ann Statist*. in press.
- Hedenfalk I *et al* (2001). *N Engl J Med* **344**: 539–548.
- Sabatti C *et al* (2003). *Genetics* **164**: 829–833.
- Sarkar SK (2002). *Ann Statist* **30**: 239–257.
- Storey JD (2003). *Ann Statist*. in press.
- Storey JD (2002). *J R Statist Soc B* **6**: 479–498.
- Storey JD, Tibshirani R (2001). Technical Report 2001-28. Department of Statistics, Stanford University.
- Storey JD, Tibshirani R (2003). *Proc Natl Acad Sci USA* **100**: 9440–9445.
- Storey JD *et al* (2003). *J R Statist Soc B*. in press.
- Yekutieli D, Benjamini Y (1999). *J Statist Planning Inference* **82**: 171–196.