A Q-Q plot aids interpretation of the false discovery rate

Nicholas W. Galwey

Research Statistics Group, GlaxoSmithKline Medicines Research Centre, Stevenage, Hertfordshire, UK

Correspondence
Nicholas W. Galwey, Research Statistics Group, GlaxoSmithKline Medicines Research Centre, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, UK. Email: nicholas.w.galwey@gsk.com

This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract
False discovery rates are routinely controlled by application of the Benjamini–Hochberg step-up procedure to a set of p-values. A method is demonstrated for representing the values so obtained (the BH-FDRs) on a quantile–quantile (Q-Q) plot of the p-values transformed to the negative-logarithmic scale. Recognition of this connection between the BH-FDR and the Q-Q plot facilitates both understanding of the meaning of the BH-FDR and interpretation of the BH-FDR in a particular data set.

KEYWORDS
bioinformatics, computer intensive methods, data mining, multivariate data, statistical genetics

1 | INTRODUCTION

Quantile–quantile (Q-Q) plots and false discovery rates (FDRs) are both routinely used in the interpretation of data analyses that produce large sets of p-values, such as high-throughput screens of multiple response variables (e.g., gene expression values potentially influenced by a single cause (Shedden et al., 2005)) or multiple explanatory variables (e.g., genetic variants potentially associated with a single response (The Wellcome Trust Case Control Consortium, 2007; Kathiresan et al., 2009)). However, there is an important connection between these two statistical tools that does not seem to be widely known. A Q-Q plot, transformed to the negative-logarithmic scale, is valuable both for understanding the meaning of the FDR, and for interpreting the FDRs obtained from a particular set of m significance tests.

2 | THE METHOD AND ITS INTERPRETATION

2.1 | The Benjamini–Hochberg (BH) step-up procedure for FDR control

When a set of m significance tests have been performed, it is often desirable to select a subset of the hypotheses tested for further investigation. Conventional assessment of the p-values obtained against a significance threshold \( \alpha \), typically 0.05, may not provide an adequate basis for selection, because when \( m > 1 \) tests have been performed, the probability that some of them will give \( p < \alpha \) is greater than \( \alpha \), even if the null hypothesis (\( H_0 \)) is true for all tests. Indeed, if \( H_0 \) is true in a specified proportion of cases, then as \( m \to \infty \), the probability that at least some tests will give such a false-positive result approaches certainty. Various approaches have been developed to address this issue, notably the Bonferroni correction (Bonferroni, 1936), which applies a much more stringent individual-test significance threshold \( \alpha / m \), thereby ensuring that even if \( H_0 \) is true for every test, the probability that one or more tests will produce a significant result is still less than...
However, this defense against false-positive results comes at a price: many tests for which the alternative hypothesis \( H_1 \) is true, and deserves further research, may fail to give significant results. The proportion of false-negative results is increased and the power of the significance test is reduced.

An attractive alternative approach is control of the FDR, formally described by Benjamini and Hochberg (1995). This approach assumes that a procedure is available that decides, for each test, on the basis of its \( p \)-value, whether \( H_0 \) should be rejected, and hence \( H_1 \) accepted, and a discovery announced. The approach recognizes that the tests for which \( H_1 \) is accepted will fall into two categories:

- cases in which \( H_1 \) is true: true discoveries (true positives), and
- cases in which \( H_1 \), though accepted, is false and \( H_0 \) is true: false discoveries (false positives).

Let

\[ S = \text{number of true positives} \]

\[ V = \text{number of false positives} \]

among the \( m \) tests, both \( S \) and \( V \) being unknown. The aim is then to control the ratio \( V / (V + S) \), keeping its expected value below a specified level. The method proposed by Benjamini and Hochberg is as follows. First rank the \( p \)-values, \( p_i, i = 1 \ldots m \), in ascending order, \( p_1 \leq p_2 \leq \ldots \leq p_m \). (Note the use of brackets in the subscript to distinguish the labeling of the ranked \( p \)-values from that of the \( p \)-values in their original order.) Then specify a rate \( q^* \) at which the FDR is to be controlled \((0 \leq q^* \leq 1)\). Then find the largest \( p \)-value for which

\[ p_i \leq \frac{i}{m} q^*. \]

Specify this as the \( k \)th \( p \)-value, and interpret it as the largest significant \( p \)-value: that is, reject \( H_0 \) for this test and for all tests giving smaller \( p \)-values. It can then be proved that if the \( p \)-values are mutually independent in those cases where \( H_0 \) is true, the expected (i.e., average) proportion of false discoveries among the test results declared significant is

\[ E \left( \frac{V}{V + S} \right) \leq q^*. \]

(If no value \( p_i \) meets criterion (1), then \( H_0 \) is not rejected for any test and no discoveries are announced.)

This method is the Benjamini-Hochberg (BH) step-up procedure, and \( q^* \) is the BH-FDR. (The term “step-up” refers to the fact that the search proceeds by stepping from smaller to larger values of \( p_i \). There are other FDR control methods, delivering different subsets of the \( m \) tests: these are well reviewed by Goeman and Solari (2014).)

For each \( p \)-value, \( p_i, i = 1 \ldots m \), one may note the smallest value \( q^* \) that will cause \( p_i \) to be included in the subset declared significant, and designate this \( q_i \). That is,

\[ q_i = \min \left[ q^* \mid p_i \leq \frac{i}{m} q^* \right]. \]

If no value \( q^* \) meets criterion (1), then \( q_i = 1 \). This distinction between \( q^* \) and \( q_i \) is important and helpful: \( q^* \) is a prespecified value to be applied to a set of \( p \)-values, whereas the \( q_i, i = 1 \ldots m \), are functions of the set of \( p \)-values. The FDR values produced by standard statistical software are obtained on this basis—for example, the “adjusted \( p \)-values” produced by the R function \( \text{p.adjust()} \) with either of the synonymous option settings ‘method’ = ‘BH’ or ‘method’ = ‘fdr’ (R Core Team, 2020). Thus, rejection of \( H_0 \) on the basis of such an “adjusted \( p \)-value” is equivalent to the BH procedure. This is a convenient and useful rejection criterion, and avoids the excessive stringency of Bonferroni-type corrections. However, the term “adjusted \( p \)-value” is potentially misleading, because a \( p \)-value in the usual sense is the probability of announcing a discovery conditional on \( H_0 \), whereas an FDR is the probability of \( H_0 \) conditional on the announcement of a discovery.

In practice, the \( m \) \( p \)-values in a multiple-testing situation are rarely independent: they usually test related hypotheses, for example, due to correlations among the explanatory or response variables. However, the majority of such correlations between \( p \)-values are usually positive, and it can be shown that positive correlations between \( p \)-values make the FDR conservative (Benjamini & Yekutieli, 2001, Theorem 1.2): that is, they tend to reduce \( k \), the number of tests declared
significant, and hence, increase the stringency of the significance threshold, biasing $V/(V + S)$ upward and making it an overestimate of the FDR.

### 2.2 Illustration of the BH-FDR on a quantile–quantile plot

An intuitive understanding of how the BH step-up procedure controls the FDR at the specified rate can be gained by considering two sets of simulated $p$-values with contrasting characteristics. Figure 1 shows a plot of $p(i)$ versus $i/m$ for the first of these sets. (The corresponding plot for the second set of simulated $p$-values is shown in the Appendix, in Figure A1.)

This is a Q-Q plot, that is, it compares the quantiles of one probability distribution with those of another. The vertical coordinates of the plotted points are the quantiles of the observed distribution of $p$-values; their horizontal coordinates are a set of $m$ evenly spaced quantiles from a Uniform(0, 1) distribution. If $H_0$ is true for all tests, then the observed $p$-values are a random sample from this distribution, and the quantiles on the horizontal axis are closely related to, but not quite the same as, the expected quantiles of such an ordered sample, namely, $i/(m + 1)$, $i = 1 \ldots m$. The slightly larger values given by $i/m$ are required to obtain a valid upper bound for the BH-FDR, and the difference between corresponding values, $i[1/m - 1/(m + 1)]$, approaches zero as $m \to \infty$. $i/(m + 1)$, $i/m$, and other choices for the quantiles on the horizontal axis are considered by Thode (2002, Section 2.2.2). (Thode states that the use of $i/m$ prevents the use of the largest observed value in the plot; however, this reservation relates to the Normal distribution, in which the quantile $m/m$ is located at $+\infty$, and does not apply to the Uniform(0, 1) distribution, in which this quantile is located at 1.)

The line of unit slope passing through the origin indicates the distribution of points that would be expected as $m \to \infty$ if $H_0$ were true for all tests, and will be referred to here as the $H_0$ line—though there would be stochastic variation, above and below this line, in the observed values. When a line with slope $q^*$ passing through the origin is drawn on this plot, the point with the largest $p$-value lying below this line represents $p(k)$: this value, and all the smaller $p$-values represented by points lying to its left, are declared significant. The horizontal coordinate of this point indicates the proportion of the $p$-values that are declared significant, both true positives and false positives; the vertical coordinate indicates the proportion expected to be declared significant at level $p(k)$ if $H_0$ were true for all tests—the false positives. Thus,

$$p(k) \left/ \frac{k}{m} \right.$$  \hspace{1cm} (4)

is an estimate of the BH-FDR, and the plot shows that this value, the slope of a line connecting the point representing $p(k)$ to the origin, is less than $q^*$. 

**FIGURE 1**  Quantile–quantile (Q-Q) plot of the first set of simulated $p$-values, with annotations to illustrate the use of the Benjamini–Hochberg step-up procedure to achieve a specified false discovery rate (the BH-FDR)
FIGURE 2  Quantile–quantile (Q-Q) plots transformed to the negative-logarithmic scale, with conventions and annotations to indicate the BH-FDR: (a) plot for the first set of simulated p-values, in which the minimum q(i) is attained at the smallest value of p(i) and (b) plot for the second set of simulated p-values, in which the minimum q(i) is attained at an intermediate value of p(i).

2.3  Negative-logarithmic transformation of the Q-Q plot

In this Q-Q plot, the points representing small p-values—the ones of greatest interest—are bunched together near the origin. However, transformation of the plot from the p scale to the -$\log_{10}(p)$ scale (Figure 2a) spreads these points out clearly in the upper right-hand region, and also provides further insight into the interpretation of the BH-FDR. A specified BH-FDR ($q^*$) can then be represented by a line parallel to the $H_0$ line. The intercept of this BH-FDR line is -$\log_{10}(q^*)$, and the largest p-value to be declared significant, according to the BH step-up procedure, is the first one above the line, reading from the left. The significance threshold ($\alpha$) implied by the choice of $q^*$ can be obtained by reading off the coordinate of the largest significant p-value from the vertical scale, and the proportion of the full set of p-values that are declared significant can be obtained by reading off the coordinate from the horizontal scale, backtransforming from -$\log_{10}()$ in both cases. Thus, in Figure 2(a), if $q^* = 0.3$ is specified (N.B., an FDR that would be considered excessively generous in most applications), then $\alpha = 10^{-1.243} = 0.057$, and a proportion $10^{-0.585} = 0.26$ of the tests (26%, 13 out of $m = 50$) are declared significant.

When each point on the negative-log transformed Q-Q plot is color-coded according to the corresponding value of q(i), the plot acquires further power as a tool for interpretation. In the present data set, the smaller p-values lie further and further from the $H_0$ line, and are associated with smaller and smaller values of q(i): if the relatively stringent value $q^* = 0.1$ is specified (a BH-FDR boundary commonly used in practice), 12% of them (6 out of 50) are still declared significant, and these are readily identified by their deep-red plotting symbol. The p-values for points colored light red will be declared significant only if the BH-FDR boundary is relaxed to $q^* = 0.3$. The p-values declared significant when $q^* = 0.1$ are also in the larger subset identified when $q^* = 0.3$, and are again declared significant.

From Figure 2(a), it might appear that the relationship between $p_{(i)}$ and $q_{(i)}$ is strictly monotonic, but the corresponding plot for the second set of p-values, which have a different pattern, shows that this is not the case (Figure 2b). If $q^* = 0.3$ is specified for this data set, then $\alpha = 10^{-0.701} = 0.199$, and a proportion $10^{-0.167} = 0.68$ of the tests (68%, 136 out of $m = 200$) are declared significant. All p-values to the right of this are declared significant conditional on the specified value of $q^*$, whether or not they are above the BH-FDR line (parallel to the $H_0$ line and with intercept $\log_{10}(0.3) = 0.523$): this is due to the “step-up” nature of the BH procedure. The minimum attainable FDR in this data set is $q^* = 0.250$: if a lower value of $q^*$ is specified, none of the p-values are declared significant. This value of $q^*$ results in a generous individual-test significance threshold of $\alpha = 0.089$: 35% of the p-values are smaller than this, but the sequence of smaller p-values curves back toward the $H_0$ line (the most extreme ones lying below it), so none of them achieves a lower FDR—and this is indicated by their light-red plotting symbol. This representation of the p-values on the negative-logarithmic scale color-coded by the corresponding values of $q_{(i)}$ makes clear the relationship between the p-values declared significant and the full set—and one important message that this representation conveys is that the BH-FDR $q^*$ identifies a set of p-values, and is not a property of any individual value.
Patterns like that in Figure 2(b) commonly occur in practice, and result in a not-strictly-monotonic relationship between $p_{(i)}$ and $q_{(i)}$: there are regions of the ordered sequence of $p$-values in which $q_{(i)}$ does not change. This may be puzzling until the reason is elucidated by the corresponding Q-Q plot and criterion (3). Such a pattern is often an indication of positive correlations among the significance tests. The effective number of tests (a concept reviewed by Li et al., 2012) is then less than $m$, the small $p$-values are not as small as would be expected on the basis of $m$ independent tests, and the BH-FDR is conservative as noted earlier.

2.4 Alternative visualizations

There are other ways in which a set of $p$-values and the associated FDRs can be visualized to assist in understanding the relationship between them and in their interpretation. Several boundary values $q^*$ may be represented on the untransformed Q-Q plot (Figure 3a). The “BH-adjusted $p$-values” $q_{(i)}$ may then be plotted against $i/m$, with the same $q^*$ values marked. The values of $q_{(i)}$ and $q^*$ may be negative-log transformed, while the values of $i/m$ are left untransformed (a semi-
FIGURE A2  Semi-log plot of the second set of simulated $p$-values

log plot—Figure 3b, for the first set of simulated $p$-values: the corresponding plot for the second set of simulated $p$-values is shown in the Appendix, Figure A2). This plot indicates clearly which subset of $p$-values meets each BH-FDR criterion, and makes it easy to determine the minimum attainable FDR. However, it does not, on its own, demonstrate “why” each subset meets the corresponding criterion, as the Q-Q plot with both axes negative-log transformed does (Figure 2a): the semi-log plot must be interpreted in conjunction with the untransformed Q-Q plot in order to achieve this.

3  |  CONCLUSION

A Q-Q plot of a set of $m$ ranked $p$-values, $p(i), i = 1...m$, against the corresponding values $i/m$, both axes transformed to the negative-logarithmic scale, with the conventions and annotations described here, can be of great value both to the theoretician seeking to understand the FDR obtained by the Benjamini–Hochberg step-up procedure (the BH-FDR), and to the practitioner seeking to explain the signals in the data to an audience.

Source code to reproduce the results in this paper is available as Supporting Information on the journal’s web page.

ACKNOWLEDGMENTS
The author is grateful to the members of GlaxoSmithKline’s Research Statistics Group for helpful comments on the presentation of this method, and to an anonymous reviewer for important observations concerning the relationship of the BH-FDR to the expectation on $H_0$ and alternative visualizations.

CONFLICT OF INTEREST
The author has declared no conflict of interest.

FUNDING
The author is a full-time employee of GlaxoSmithKline.

DATA AVAILABILITY STATEMENT
The data that supports the findings of this study are available in the supplementary material of this article.
This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID
Nicholas W. Galwey https://orcid.org/0000-0002-2110-3194

REFERENCES
Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57, 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29, 1165–1188.

Bonferroni, C. E. (1936). Teorìa statistìca delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commercialì di Firenze*.

Goeman, J. J., & Solari, A. (2014). Multiple hypothesis testing in genomics. *Statistics in Medicine*, 33, 1946–1978. https://doi.org/10.1002/sim.6082

Kathiresan, S., Willer, C. J., Peloso, G. M., Demissie, S., Musunuru, K., Schadt, E. E., Kaplan, L., Bennett, D., Li, Y., Tanaka, T., Voight, B. F., Bonnycastle, L. L., Jackson, A. U., Crawford, G., Surti, A., Guiducci, C., Burtt, N. P., Parish, S., Clarke, R.,… Cupples, L. A. (2009). Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature Genetics*, 41, 56–65. https://doi.org/10.1038/ng.291

Li, M.-X., Yeung, J. M. Y., Cherny, S. C., & Sham, P. C. (2012). Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Human Genetics*, 131, 747–756. https://doi.org/10.1007/s00439-011-1118-2

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Shedden, K., Chen, W., Kuick, R., Ghosh, D., Macdonald, J., Cho, K. R., Girodano, T. J., Gruber, S. B., Fearon, E. R., Taylor, J. M. G., & Hanash, S. (2005). Comparison of seven methods for producing Affymetrix expression scores based on false discovery rates in disease profiling data. *Bmc Bioinformatics [Electronic Resource]*, 6, 26. https://doi.org/10.1186/1471-2105-6-26

The Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447, 661–678. https://doi.org/10.1038/nature05911

Thode, H. C. (2002). *Testing for normality*. Marcel Dekker. ISBN 08247-9613-6.

SUPPORTING INFORMATION
Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Galwey, N. W. (2023). A Q-Q plot aids interpretation of the false discovery rate. *Biometrical Journal*, 65, 2100309. https://doi.org/10.1002/bimj.202100309