**Genetics: Published Articles Ahead of Print, published on June 14, 2005 as 10.1534/genetics.104.036699**

# Quantitative Trait Loci Analysis using the False Discovery Rate

Yoav Benjamini [*†] and Daniel Yekutieli [‡]

September 23, 2004

[*]Department of Statistics and Operations Research, Tel Aviv University.

[†]This research was supported by the FIRST foundation of the Israeli Academy of Sciences and Humanities. Yoav Benjamini was also partly supported by a NIH grant. We are especially thankful to two reviewers whose suggestions contributed much to the presentation of this paper.

[‡]Department of Statistics and Operations Research, Tel Aviv University.

*Running head:*

QTL analysis using the FDR.

*Key Words:*

FDR, multiple comparisons, suggestive linkage, significant linkage, confirmed linkage, BH procedure

*Correspondence:*

Daniel Yekutieli,

Department of Statistics and Operations Research,

The Sackler Faculty of Exact Sciences,

Tel Aviv University, Tel Aviv, Israel

Fax (972)-3-6409357, Phone (972)-3-6405386

Email yekutiel@post.tau.ac.il

**Abstract**

False Discovery Rate control has become an essential tool in any study that has a very large multiplicity problem. False Discovery Rate controlling procedures have also been found very effective in QTL analysis: ensuring reproducible results - few falsely discovered linkages, and offering increased power to discover QTL, although their acceptance has been slower than in microarray analysis, for example. The reason is partly because the methodological aspects of applying the False Discovery Rate to QTL mapping are not well developed. Our aim in this work is to lay a solid foundation for the use of the False Discovery Rate in QTL mapping. We review the False Discovery Rate criterion, the appropriate interpretation of the FDR, and alternative formulations of the FDR that appeared in the statistical and genetics literature. We discuss important features of the FDR approach, some stemming from new developments in FDR theory and methodology, which deems it especially useful in linkage analysis. We review False Discovery Rate controlling procedures: the BH, the resampling procedure, and the adaptive two stage procedure, and discuss the validity of these procedures in single and multiple traits QTL mapping. Finally we argue that the control of the False Discovery Rate has an important role in suggesting, indicating the significance and confirming QTLs, and present guidelines for its use.

# 1 Introduction

Overlooking the multiplicity aspects in QTL studies with numerous markers and several quantitative traits can lead to many false discoveries of linkages. Lander and Botstein (1989) were among the first to address the multiplicity problem in QTL mapping, arguing that "Adopting too lax a standard guarantees a burgeoning literature of false positive ... Scientific disciplines erode their credibility when substantial proportion of claims cannot be replicated". The procedure they offered was designed to control the probability of making even one type I error - of making even one false discovery - in a QTL study with a single quantitative trait. This probability is termed the Family Wise Error-Rate (FWE), or in this case the Genome Wise Error-Rate (GWER).

Concerned that controlling the FWE at conventional levels results in very little power to discover QTLs Lander and Kruglyak (1995) set the following widely used terminology. *Suggestive Linkage* - statistical evidence that would be expected to occur one time at random in a genome scan. *significant linkage* - statistical evidence that would be expected to occur at random with probability 0.05. *Confirmed Linkage* - significant linkage confirmed by a further sample with a nominal p-value $< 0.01$.

The FWE controlling procedures employed are designed to control the probability of making one or more false discoveries for a single trait. In studies with multiple traits FWE controlling procedures have to be suitably adjusted by further raising the threshold for significant linkage according to

the number of traits in the study. Consider, for example, a study with 12 quantitative traits. For a linkage to be suggestive, it should pass a threshold which is equivalent to controlling the FWE at level 0.6 (using the Poisson approximation when the average is 1). Findings that would have been reported as significant linkages ($FWE = 0.05$) in a single trait study, should now be reported only as suggestive linkages (if a simple Bonferroni correction was used the single trait threshold should be divided by the number of traits). Such a solution may be acceptable in QTL studies with only a few traits. In studies with many traits the control of the FWE leaves very little power to make discoveries. The problem becomes serious even well before reaching current large problems that combine microarrays with QTL analysis, as in Williams et al. (2002) with its 12422 quantitative traits.

The conflict between the strict control of type I error, as a protection against false discoveries, and the need for increased power led Benjamini and Hochberg (1995) to suggest a new criterion for type I error in multiple testing: let $\mathbf{Q}$ be the proportion of false linkage claims in a genome scan, setting it at zero if no discoveries are made; then the False Discovery Rate (FDR) is the expected value of $\mathbf{Q}$. It thus seems as if Benjamini and Hochberg have taken literarily the warning of Lander and Botstein about the danger in allowing a substantial proportion of false claims, and casted this danger into a well defined statistical criterion.

Benjamini and Hochberg (1995) also introduced a FDR controlling procedure (sometimes called linear step-up or BH procedure), and proved that

it controls the FDR for independent test statistics. Benjamini and Yekutieli (2001) proved that the same holds for some types of positive dependence.

Weller et al. (1998) were the first to apply the FDR approach for QTL mapping. They explained the use of the BH procedure for single markers analysis, and demonstrated the increase in power. While commenting on this work, Zaykin et al. (2000) pointed out that the conditional interpretation to the FDR criterion, i.e. that controlling the FDR at 0.05 means that 95 % of the discoveries are true, does not always hold. Although Weller (2000) eased their concern, the issue is still alive, and has resurfaced in discussions of alternative formulations of the FDR criterion: positive FDR (Storey, 2004), adjusted FDR (Mosig et. al., 2002) and the Proportion of False Positives (PFP - Fernando et. al., 2004).

While the BH procedure received much attention in microarrays analysis, the developments were slower in the context of QTL analysis. Sabatti et. al. (2003) verified that the BH procedure controls the FDR for single trait linkage (see also Yekutieli, 2001) and association genome screens. Lee et al. (2002) addressed interval mapping, discussed the use of the FDR in multiple quantitative traits, and recommended the application of the BH procedure separately for each trait. Sabatti et. al. (2003) further suggested substituting the BH procedure with resampling based FDR procedure capable of capturing the dependency structure of the data to achieve additional power.

What is common to the above researchers, and to the other few users of FDR ideas in QTL analysis, is their enthusiasm about the important role

that these ideas can play in QTL analysis. Since we share this enthusiasm, it seems important to us to lay a solid foundation for the use of FDR in this area, as well as to expand our working methodologies. We present the FDR in precise terms, in conjunction with related formulations of the proportion of false discoveries. We make a clear distinction between the FDR criterion and the BH procedure that controls it under independence. We explain the role that adaptive FDR controlling procedures can have in increasing power while controlling the FDR. We clarify what is known about the performance of FDR controlling procedures under dependency in the context of QTL analysis, and offer more progress on this question. We analyze the performance of the resampling procedure of Yekutieli and Benjamini (1999) in QTL analysis, and we address the important issue of multiple traits. In all of the above, we also make an effort to point at remaining gaps in our knowledge. Finally, we analyze FDR control in independently repeated experiments, and make use of the results to offer a set of FDR guidelines for the significance of findings in QTL studies.

The paper is structured as follows: with a slight danger of repetition, we review the FDR in Section 2. In 2.1 we present the various types of FDR controlling procedures. In 2.2 we discuss what "FDR control" means, its relations to conditional FDR control and the suggestive linkage criterion, and the control of FDR in multi-trait studies. In 2.3 we present a new framework, based on the BH procedure, for FDR-confirmed QTL mapping with an independent study and compare our new approach to Lander and

Kruglyak's *confirmed linkage*. Section 3 is dedicated to simulation studies. We compare the performance of the BH procedure to the performance of the FDR-resampling and FWE-resampling procedures. We examine the FDR control of the BH procedure in cases for which no analytical results are available. We summarize our FDR-guidelines for assessing the significance of the results of QTL experiments in Section 4.

Thus we hope to be able with this paper to remove some of the hurdles placed before the use of FDR in QTL analysis, and help to its wider and faster acceptance.

# 2 The False Discovery Rate

The FDR was defined in Benjamini and Hochberg (1995) as the expected proportion of false discoveries. Formally, if $\mathbf{R}$ is the total number of discoveries, and $\mathbf{V}$ of them are false discoveries, the proportion of false discoveries is $\mathbf{Q} = \mathbf{V}/\mathbf{R}$ if $\mathbf{R} > 0$, and $\mathbf{Q} = 0$ if $\mathbf{R} = 0$. Although the value of $\mathbf{Q}$ is usually not known in a particular experiment, we can discuss, and even control *its expectation* $FDR = E(\mathbf{Q})$.

Benjamini and Hochberg (1995) considered several alternative formulations for the FDR. Two such alternative formulations, passed over by Benjamini and Hochberg, reappeared in the statistical and genetic literature. $E(\mathbf{Q}|\ \mathbf{R} > 0)$ - the positive FDR (pFDR) in Storey (2002); $E(\mathbf{V})/E(\mathbf{R})$ - the Proportion of false positives (PFP) in Fernando et al. 2004. The pFDR and PFP have very desirable properties, however they are not appropriate

criteria for multiple testing: a statistical procedure offering pFDR or PFP control for all configurations of true and false null hypotheses cannot be constructed. Let us articulate this point, if all tested hypotheses are true, say studying a totally non-hereditary trait, any discovery made is a false discovery - no matter what statistical method is being used. The proportion of false discoveries among these discoveries is identically 1. With the exception of the useless "never-reject-anything" rule, the pFDR and the PFP of all procedures are by definition 1. Thus even the procedure suggested in Storey (2002), and is designed to address the pFDR, is not a pFDR controlling testing procedure. It is capable only of estimating the pFDR once a fixed rejection threshold is being used (see 2.1.2) and includes an estimate of $m_0/m$ (see 2.1.3).

In contrast, if all tested hypotheses are true, the FDR equals the traditional FWE. Therefore it also makes sense to use the conventional levels such as 0.05 or 0.01 for FDR control, though in some applications higher values may be justifiable. In particular for QTL analysis Mosig et al. (2001) recommends using 0.1, and Weller et al. (1998) use 0.25.

## 2.1 FDR controlling procedures.

### 2.1.1 The BH procedure

Benjamini and Hochberg (1995) provided a simple stepwise procedure to control the FDR at a desired level $q$. The procedure makes use of the p-values only, so the statistical test itself may be tailored to the problem at

hand, be it a LOD score or a non-parametric test. The procedure runs as following. The individual p-values are sorted from smallest to largest: $P_{(1)} \leq P_{(2)} \ldots \leq P_{(m)}$. Starting from the largest p-value $P_{(m)}$, compare $P_{(i)}$ with $q \cdot i/m$. Continue as long as $P_{(i)} > q \cdot i/m$. Let k be the first time $P_{(i)} \leq q \cdot i/m$, then declare the hypotheses corresponding to the smallest $k$ p-values as significant.

An intuitive explanation of the connection between the BH procedure and FDR control was offered by Benjamini and Hochberg (1995) and later recited by Weller et al.(1998) and Mosig et al.(2001): if we reject the $k$ hypotheses with p-values less than $p_{(k)}$, and treat this $p_{(k)}$ as if it is fixed, then regardless of the joint distribution of the p-values the average number of false discoveries is $p_{(k)} \cdot m_0$, where $m_0$ is the (unknown) number of true null hypotheses. As an upper bound for the FDR we take $p_{(k)} \cdot m/k$. For this to be less than $q$, $p_{(k)}$ has to be less than $q \cdot k/m$. This argument by itself is inappropriate as a proof, mixing expectations with values in the experiment, and ignoring the case when no discovery is made. Hence it cannot be used to imply that the BH procedure always controls the FDR regardless of the test statistic distribution, and in fact in some rare settings the FDR of the BH procedure may get as high as $q(1+1/2+1/3+\ldots 1/m)$. The correct statement, proven precisely in Benjamini and Yekutieli (2001), is as following:

**Theorem 2.1** *If the test statistics are Positively Regression Dependent on each hypothesis from the subset corresponding to true null hypotheses, the BH procedure controls the FDR at level* $q \cdot m_0/m$.

The definition of the Positive Regression Dependent on a Subset (PRDS) condition and a review of important examples where it is satisfied can be found in Benjamini and Yekutieli (2001). The PRDS condition and FDR control of the BH procedure in single trait QTL analysis are discussed in Sabatti et al. (2003). In 2.2.3 we discuss FDR control of the BH procedure in multiple trait QTL analysis.

### 2.1.2 The fixed rejection region approach

Weller et al. (1998) denote $q = P_{(i)} \cdot m/i$ as the false discovery rate when rejecting at $P_{(i)}$. They note the erratic behavior of this $q$ as a function of $i$, and demonstrate in simulations that the mean value of the above expression - for a fixed $i$ - is greater than the mean of the number of tests with p-values $\leq P_{(i)}$ not linked to a simulated QTL divided by $i$.

Such an attempt does not verify the FDR control of the BH procedure. Instead, it shows that if one decides beforehand to reject the $i$ hypotheses with the smallest p-values, the corresponding $P_{(i)} \cdot m/i$ is a conservative estimator of the FDR committed. A more natural choice in this direction is to use the fixed rejection region approach. For example, reject any hypothesis with $LOD > c$, and study the FDR properties of such a rejection rule as $c$ varies. A detailed discussion of this topic, in general settings, can be found in Yekutieli and Benjamini (1999) and Storey (2001).

In the BH procedure $q$ is specified while the LOD threshold and the number of hypotheses rejected varies. If, for each $P_{(i)}$, instead of comparing

$P_{(i)} \cdot m/i$ to $q$ we take the minimum of such terms over all p-values larger than or equal to $P_{(i)}$, we get the BH FDR-adjusted p-values: $P_{(i)}^{BH} = min_{j \geq i}(P_{(j)} \cdot m/j)$ (Westfall, 1997 ; Yekutieli and Benjamini, 1999 ; later called estimated q-values in Storey and Tibshirani, 2003). Obviously, rejecting all hypotheses whose BH FDR-adjusted p-values are smaller than $q$ is equivalent to applying the BH procedure at level $q$. Thus reporting the FDR adjusted p-values $P_{(i)}^{BH}$, or plotting them as a function of $i$, is convenient. Compared to plots of $q = P_{(i)} \cdot m/i$ vs $i$ (or vs $P_{(i)}$ as in Lee et al., 2002), they are monotonic and exhibit less erratic behavior. Plotting the FDR adjusted p-values versus their location on the chromosome is another informative way of displaying results.

### 2.1.3 Two-stage adaptive procedures

When some QTLs exist, implying that the number of markers unlinked to QTL $m_0$ is less than $m$, the BH procedure controls the FDR at too strict a level by a factor of $m_0/m$. If this factor were known, using the BH procedure at level $q^* = q \cdot m/m_0$ instead of $q$, would be more powerful and still achieve FDR control at level $q$. Since $m_0$ is hardly ever known, it seems attractive to first estimate $m_0$ and then use $q^*$ as defined above, with the estimated $m/m_0$ instead of the true quantity. Theorem 2.1 by itself does not guarantee that two stage procedures control the FDR at the desired level, and their properties need separate attention (see Benjamini et al., 2004).

Two stage adaptive FDR controlling procedures were introduced by Ben-

jamini and Hochberg (2000). The study and development of adaptive FDR controlling procedures is a very active area of research: Storey (2002), Storey et al. (2004), Black (2004), to name just a few. In the context of QTL analysis, Mosig et al. (2001) discussed the role of the estimated $m_0$ in the control of FDR, their offering being an adaptive procedure aimed at controlling the FDR (rather than a new criterion). Important modification to the procedure were offered in Fernando et al. (2004).

Benjamini et al. (2004) introduced a two stage procedure in which $m_0$ is estimated by the number of hypotheses *not* rejected by the BH procedure at level $q$, $m - \mathbf{R}$, further multiplied by $1 + q$. At the second stage the BH is used again with the modified $q^*$. They analytically prove FDR control of their new procedure for independent test statistic. Benjamini et al. (2004) further show that with the exception of their procedure, all other two stage procedure (including the two stage procedure of Mosig et al., 2001) fail to control the FDR for some positively regression dependent tests. This does not mean yet that all other procedures are not appropriate for the particular dependency structure encountered in QTL mapping, but it raises doubt as to their appropriateness in their current form.

### 2.1.4 Resampling procedures

Resampling procedures, reshuffling and randomization being close variants, overcome the problem of specifying the marginal distributions of the test statistic at each marker under the null hypothesis of no linkage, and even

more importantly incorporate the dependency between the markers to increase the power of the analysis. It is done by imitating the way the data was generated by sampling, with or without replacement, under the assumption of no linkage. In QTL analysis resampled data can be generated by randomly drawing a with-replacement sample of the phenotype data and then correlate it with the original genotype data.

Yekutieli and Benjamini (1999) introduced a general FDR controlling resampling procedure designed to control the FDR: for $i = 1, 2, \ldots m$ compute the resampling based FDR adjusted p-value $P_{(i)}^{Res}$; reject the null hypotheses corresponding to $P_{(i)}^{Res} < q$. This procedure controls the FDR if the sets of tests statistics corresponding to true and false hypotheses are independent. This condition is satisfied in single trait QTL analysis (see 2.2.3). For multiple traits we bring support for FDR control from the simulation study in Section 3.

## 2.2 Control of the FDR

Level $q$ FDR means that the expected value of $\mathbf{Q}$ is less than or equal to $q$. The actual value of $\mathbf{Q}$ can theoretically vary from 0 to 1, thus FDR control does not generally imply that 95% of the discoveries are true discoveries. Genovese and Wasserman (2004) show that in very large problems with independently distributed test statistics if the data contains proportion of true discoveries then applying the BH procedure does ensure that approximately 95% of the discoveries are true ones, and as $\mathbf{Q}$ hardly varies:

$FDR = pFDR = PFP = 0.05 \cdot m_0/m$. Notice that while these assumptions might be reasonable for microarray data, they do not necessarily apply to QTL mapping.

### 2.2.1 Conditional control of the FDR

Zaykin et al. (2000) argue against control of the FDR as it does not control the expected proportion of false discoveries if interest is restricted to experiments in which some discoveries were made (which is the pFDR), and recommend adhering to control of the FWE. Storey (2002) has been advancing the pFDR as the criterion of interest. Mathematically, $pFDR = E(\mathbf{Q}|\ \mathbf{R} > 0) = FDR/P(\mathbf{R} > 0) > FDR$, thus control of the FDR does not imply pFDR control. We agree with Zaykin et al. (2000) that control of FWE offers more protection than FDR at the same level. However, as we argued before, FWE is not suitable for QTL analysis because at low levels such as 0.05, it leaves too little power to make discoveries of multiple QTLs, and at higher levels, as allowed by suggestive linkage, it does not offer sufficient type I error control (details in 2.2.2). Moreover, in spite of its conservativeness, controlling the FWE still does not control the pFDR.

Actually, the control of the FDR strikes a balance between the usually too conservative control of the FWE and the sometimes impossible control of the pFDR, depending on the data at hand. Studying totally non-hereditary trait with $FDR = q$ is the same as with $FWE = q$, (both being smaller than the pFDR.) On the other hand, when there are QTLs present, as true

discoveries are more likely than false ones, $\Pr(\mathbf{R} > 0) > FWE$. In particular, studies in which many hypotheses are rejected reflect configurations where $\Pr(\mathbf{R} > 0) \approx 1$. Thus $FDR \approx pFDR$, and controlling the FDR in such cases is reasonably close to controlling the pFDR (both being smaller than the FWE), (Weller, 2000).

### 2.2.2 Replacing the suggestive linkage criterion with FDR control

Let us carry the above discussion into the case of suggestive linkage. The threshold is chosen so that there will be one false linkage per genome scan on the average. Using the Poisson approximation, such a threshold is equivalent to controlling the FWE at 0.6. Now when Zaykin et al. (2000) claim that the control of FWE is better than the FDR, they argue: " ... using an FWER (FWE - B.Y.) controlling method, one may claim that all significances obtained in the study are real, gambling upon the occurrence that the given study was not one of the 20 (or whatever FWE level that is used) that will produce a false positive". Consider the above argument applied to the criterion for suggestive linkages: gambling that the given study is not one of the 12 out of 20 that will produce a false positive, is difficult to justify. It is therefore our view that controlling the FWE at 0.6 cannot by itself be trusted to indicate suggestive results. If one reads carefully Lander and Kruglyak (1995) similar skepticism can be sensed, as for example, they do not see a way to confirm suggestive linkages in a second study. We therefore suggest that this criterion be abandoned, and be replaced by FDR control

at lower level. A good choice is $q = 0.1$, as done by Lee et al. (2002). We certainly do not recommend going higher than $q = 0.2$ in published reports.

Controlling the FDR at even lower level, say $q = 0.05$, yields credible results while adapting to the number of traits, their complexity, and their degree of heritability. In studies with a few weak QTLs, such FDR-significant linkage is as conservative as the usual (FWE) significant linkage. In studies with highly heritable complex traits $FDR = 0.05$ control offers much more power than $FWE = 0.05$ control. Thus false discoveries are likely to occur among the FDR-significant linkages, but the false discoveries are expected to be a small proportion of the discoveries made.

### 2.2.3 FDR control in multi-trait studies

Studying multiple traits, a question arises whether there is need to consider all traits jointly when controlling the FDR at level $q$, or each trait can be considered separately. An intuitive reason for the the second approach is that having a fixed proportion of errors in each trait implies the same fixed proportion of errors in the combined study. Lee et al. (2002) considered both options and recommended the latter. However, when each trait is considered separately the actual FDR over the entire study is determined by type of data analyzed. For example, in a study with $k$ independent non heritable traits, the FDR for all traits combined equals $1 - (1 - q)^k$. On the other hand, in a study in which all traits are highly heritable, the combined FDR is approximately $q$.

Fernando et al. (2004) show that if for each trait the PFP is controlled at level $q$ then the PFP of the entire study is less than or equal to $q$. Our ongoing work (Benjamini and Yekutieli, 2002) indicates that controlling the FDR at each trait separately is legitimate as long as the number of markers discovered this way across all traits greatly exceeds the number of traits in the study (e.g., if the number of markers discovered is more than twice the number of traits the combined FDR is $< 2 \cdot q$). The connection between the two results is that in large studies with many discoveries $PFP \approx FDR$. To be on the safe side, we recommend that the test statistics of all the traits in the study have to be tested simultaneously. Unlike in FWE controlling approaches, the resulting decrease in power need not be large (see simulations in 3.1).

Yekutieli (2001) addressed the problem of controlling the FDR when simultaneously testing all test statistics in a multiple trait study. For each trait locus and each quantitative trait the null hypothesis is that the allelic composition at the locus is unassociated with the quantitative trait. The alternative hypothesis depends on the design of the study and can be one sided, but for the most part is two sided: non zero additive effect for the back-cross design; or non zero additive or dominance effects for an inter-cross design. For each trait separately, Yekutieli (2001) proves (a) PRDS dependency between the test statistics in a backcross experiment corresponding to the true null hypotheses. Due to genetic linkage, for each trait all the hypotheses on a chromosome are either true or false depending on the presence of a QTL

on the chromosome. Since test statistics on separate chromosomes are independent: (b) independence between the set of test statistics corresponding to false null hypotheses and those corresponding to true ones. Property (b) is a sufficient condition for the validity of the resampling based FDR controlling procedure. Properties (a) and (b) jointly imply the PRDS condition. Property (b) does not hold for multiple trait - a chromosome might contain QTLs for only a subset of the traits in the study. Traits having negatively correlated environmental components rule out property (a).

## 2.3 Confirming QTL mapping with an independent study

According to the guidelines suggested by Lander and Kruglyak (1995) the most credible linkage is a "confirmed linkage" - a significant linkage from one study found significant at the nominal 0.01 level in an independent study, preferably conducted by other investigators. They argue that each significant linkage can be tested by itself at 0.05 level, since it was chosen in advance. Since typically one further considers a 20 cM neighborhood of markers around the significant linkage, their nominal threshold is lowered to 0.01.

The above approach has two limitations. (1) Using the FWE at level .05 to test the significance of the results in the initial study may leave little power to make any discoveries. (2) If a number of linkages are discovered in the first stage, neglecting to control for multiplicity in the confirmatory stage compromises the strength of the evidence. The latter is especially serious when we consider multiple complex traits, with their many potential QTLs.

We therefore suggest a new FDR controlling testing strategy in which the BH procedure is used in both the initial and the confirmatory studies.

**Definition 2.2** A FDR procedure for confirming QTL mapping in an independent study.

1. Test the m null hypotheses in study 1 using the BH procedure at level $q_1$.

2. Test the $r_1$ hypotheses rejected in study 1 using the BH procedure at level $q_2$.

**Proposition 2.3** *Under the conditions of Theorem 2.1, and for* $q_1$ *and* $q_2$ *fixed in advance, the FDR of procedure 2.2 is less or equal to* $m_0 \cdot q_1 \cdot q_2/m$.

The Proof of Proposition 2.3 is deferred to the Appendix.

If we wish to Adhere to the Lander and Kruglyak (1995) choice of "double 0.05" level, *FDR-confirmed linkages* are significant linkages from the first study which were tested at FDR level of 0.05 in the confirmation study and remained significant. Thus FDR-confirmed linkages enjoy FDR level of 0.0025. This way we address the two limitations of Lander and Kruglyak approach to confirming linkages.

Otherwise, Proposition 2.3 can be used in the design of the experiment: determining the optimal significance level and sample sizes to be used in each study. With the possible goal of decreasing the number of sampling units

(sample size times number of hypotheses), while the combined study still controls the FDR at level 0.05.

# 3 Simulation Study

## 3.1 Use of resampling based procedures in QTL analysis

Sabatti et al. (2003) suggest using resampling based procedures in QTL analysis. To test their suggestion we compared the power of the BH procedure to the power of the resampling FDR procedure and the Churchill and Doerge (1994) FWE controlling resampling procedure.

**Single trait power comparison** This simulated experiment consisted of 550 backcross progeny. The genome of each individual consisted of 20 100cM chromosomes. Markers were situated at 1cM intervals - the dense marker map is meant to work in favor of the resampling based procedures. Quantitative trait values were computed by summing allelic effect at QTLs and noise. Six QTLs with effect sizes of 0.3, 0.2, 0.4, 0.1, 0.4, 0.3 were planted on 4 chromosomes. Each simulation consisted of 500 repetitions. For each outcome variable and marker, the absolute valued Z-score and p-values were computed, and the two FDR procedures and the FWE procedure were applied at level 0.05. In order to conduct the two resampling based methods the data was resampled 3000 times under the complete null hypothesis of no QTL.

A summary of the results of this study are presented in Table 1. The two FDR controlling procedures yield very similar results. Their FDR levels are $0.038 \pm 0.004$ and $0.039 \pm 0.004$, significantly less than $0.04 = 0.05 \times \frac{16}{20}$ (the expected FDR level for independently distributed test statistics). The two FDR procedures enjoy similar power and are superior to the FWE controlling procedure. The FDR procedures have lower mean rejection thresholds (expressed in terms of Z-scores), and more true discoveries ( over 200 for the FDR procedures; 91 for the FWE procedure). The price for the increased power is a few more erroneous linkages, FWE level of 0.63 and 0.64.

—- PLACE TABLE 1 HERE —-

**Multi trait power comparison** In this set of simulation we altered the simulation of the phenotypes. We simulated 1, 2, 4 and 8 traits. QTLs were placed at random. The number of chromosomes having QTLs was sampled from a Poisson distribution with mean 3. The number of QTLs per chromosome was either 1 or 2. Effect sizes were sampled from $U[-0.25, 0.4]$. In this set of simulations an independent random error term was added to each trait.

In Table 2 we present the average threshold for rejecting null hypotheses for the three procedures at level 0.05 (expressed in terms of Z-scores and p-values). The two FDR procedures retain, and even increase, their power

as the number of traits increases. This property is characteristic to FDR procedures when QTLs are present. On the other hand, if no QTLs are present FDR procedures control the FWE and the FDR threshold increases as the number of hypotheses increases. This type of behavior is evident in the performance of the FWE controlling procedure. As the number of traits increases twofold the critical p-value is multiplied by 0.5.

—- PLACE TABLE 2 HERE —-

## 3.2 FDR control of the BH procedure

According to the simulation results in 3.1 the BH and the FDR Resampling procedures are practically identical. In this set of simulations we verify whether the BH procedure controls the FDR in situations for which no analytical results are available.

**FDR control in an intercross experiment** The setting is as in the single trait power comparison, but we altered the design of the experiment from a backcross to an intercross. We therefore added the following 6 dominance effects to the 6 QTLs: 0.15, -0.15, 0.3, -0.06, -0.01, -0.25. The simulation included 1000 replications. The simulated FDR of the BH procedure was $0.0396 \pm 0.001$, indicating that it also controls the FDR at level $q \cdot m_0/m$ in single trait intercross experiments.

**FDR control for negatively correlated multiple traits** The setting is as in the multi trait power comparison with the number of traits set to 8, only this time correlated (rather than independent) errors are added to the traits. This set of simulations consisted of two runs of 1000 replications.

In the first simulation the random error terms added to the phenotypes were a sample from a multivariate normal distribution with 0 mean, unit variance, and 0.36 correlation. FDR value for the BH procedure was $0.042 \pm 0.001$.

In the second simulation the 8 traits were divided into two blocks of 4 traits (traits 1-4 and traits 5-8). The correlation of the error terms within each block remained 0.36, the correlation of the error terms between trait belonging to different blocks was changed to $-0.36$. The FDR value was now $0.044 \pm 0.001$.

While the number of true null hypotheses varied, the expected value of $m_0 \times q/m$ was 0.0425. This reveals that the introduction of negatively correlated error terms may result in FDR exceeding $m_0 \times q/m$. But the striking feature, evident in all the simulations conducted, is that the FDR level of the BH procedure is practically unaffected by correlation: in all simulations the FDR varied from $m_0 \times q/m$ by less than 0.002.

# 4 Summary

The two advantages of the FDR approach, which make it particularly suitable for QTL analysis, are its adaptivity to the amount of information in the data,

and its scalability - controlling the FDR for multiple traits may come with no loss of power. By comparison, the thresholds set by Lander and Kruglyak (1995) are only valid for a single trait. If these thresholds are further modified to achieve FWE control when many traits are studied, the required size of the experiment which is needed to achieve significance may not be feasible.

We establish, via simulations, that the BH procedures can effectively be used to control the FDR in multiple trait studies. Controlling the FDR at the 0.10 level can take the place of the suggestive linkage criterion of Lander and Kruglyak. At the 0.05 level it can assume the role of identifying significant QTLs. Such FDR significant linkages can further be confirmed in an independent study by controlling the FDR again at the 0.05 level, thereby giving rise to FDR-confirmed linkages.

# References


[1] Benjamini Y, Hochberg Y. (1995) "Controlling the False Discovery Rate: a practical and powerful approach to multiple testing", *Journal of the Royal Statistical Society, Series B*; **57** (1): 289-300.

[2] Benjamini Y, Hochberg Y. (2000) "The adaptive control of the false discovery rate in multiple comparison problems", *The Journal of Educational and Behavioral Statistics* **25** (1) 60-83.

[3] Benjamini Y., Krieger A.M., Yekutieli D., (2004) "Adaptive Linear Step-up Proceudres that control the False Discovery Rate",

Biometrika (revised).

[4] Benjamini Y., Yekutieli D.,(2001) "The control of the False Discovery Rate in multiple testing under dependency", *Annals of Statistics*, **29** (4) 1165-1188.

[5] Benjamini Y., Yekutieli D.,(2002) "Hierarchical FDR testing of trees of hypotheses ", Research Paper 02-02 the Department of Statistics and OR, Tel Aviv University.

[6] Churchill G. A., Doerge R. W., (1994) "Empirical Threshold Values for Quantitative Trait Mapping", *Genetics* **138** 963-971.

[7] Fernando R. L., Nettleton D., Southey B. R., Dekkers J. C. M., Rothschild M. F., Soller M., (2004) Controlling the Proportion of False Positives in Multiple Dependent Tests", *Genetics* **166** 611-619.

[8] Genovese C., Wasserman L. (2004) "A Stochastic Process Approach to False Discovery Control", *Annals of Statistics*, **32** (3) 1035-1061.

[9] Haley C. S. and Knott S. A. (1992) "A simple regression method for mapping quantitative trait loci in line crosses using flanking markers", *Heredity* **69** 315-324.

[10] Lander E. S., Botstein D., (1989). "Mapping Mendelian Factors Underlying Quantitative Traits using RFLP Linkage Maps", *Genetics*, **121** 185-190.

[11] Lander E. S., Kruglyak L., (1995)"Genetic Dissection of Complex Traits: Guidelines for Interpreting and Reporting Linkage Results", *Nature Genetics* , **11** 241-247.

[12] Lee H., Dekkers J. C. M., Soller M., Malek M., Fernando R. L., Rothschild M. F., (2002) "Application of the False Discovery Rate to Quantitative Trait Loci Interval Mapping With Multiple Traits", *Genetics*, **161** 905-914.

[13] Mosig M. O., Lipkin E., Khutoreskaya G., Tchourzyna E., Soller M., Friedmann A. A. (2001) "Whole genome scan for quantitative trait loci affecting milk protein percentage e in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion", *Genetics* **157** (4) 1683-1698.

[14] Sabatti C., Service S., Freimer N., (2003) "False Discovery Rate in Linkage and Association Genome Screens for Complex disorders", *Genetics* **164** 829-833.

[15] Storey J. D. (2002) "A direct approach to false discovery rates", *Journal of the Royal Statistical Society: Series B*, **64** 479-498.

[16] Storey J. D., Taylor, J. E., and Siegmund D. (2004) "Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach", *Journal of the Royal Statistical Society: Series B*, **66**, 187-205(19).

[17] Tusher V., Tibshirani R., Chu G. (2001) "Significance Analysis of Microarrays Applied to Transcriptional Responses to Ionizing Radiation", *Proc Natl Acad Sci* **98** 5116-5121.

[18] Weller J. I., Song J. Z., Heyen D. W., Lewin H. A., Ron M. (1998) "A new approach to the problem of multiple comparisons in the genetic dissection of complex traits", *Genetics* **150** (4) 1699-1706.

[19] 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** (4) 1919-1919.

[20] Westfall P. H. (1997) "Multiple testing of general contrasts using logical constraints and correlations", *Journal of the American Statistical Association* **92** (437): 299-306.

[21] Williams R. W, Shou S., Lu L, Qu Y., Wang J., Manly K., Chesler E., Hsu H., Mountz J., Threadgill D. (2002) "Genomic Analysis of Transcriptional Networks: Combining Microarrays with Complex Trait Analysis", abstract in the First Annual CTC meeting, May 2002, *http://www.complextrait.org/*

[22] Yekutieli D.,(2001) "Theoretical results needed for applying the False Discovery Rate in statistical problems", PhD Thesis Department of Statistics and O.R., Tel Aviv University, Tel Aviv.

[23] Yekutieli D., Benjamini Y. (1999) "Resampling based false discovery rate controlling procedure for dependent test statistics", *Journal of Statistical Planning and Inference* **82** (1-2) 171-19.

[24] Zaykin D. V., Young S. S., Westfall P. H. (2000) "Using the False Discovery Rate in the genetic dissection of complex traits: A response to Weller et al.", *Genetics* **154** (4) 1917-1918.

**Appendix: proof of proposition 2.3**

Let $\mathbf{R}_1$, $\mathbf{V}_1$, $\mathbf{R}_2$ and $\mathbf{V}_2$ denote the number of discoveries and false discoveries in the initial and confirmatory studies when applying procedure 2.2. Let

$$\mathbf{Q}_1 = \begin{cases} \mathbf{V}_1/\mathbf{R}_1 & \text{if } \mathbf{R}_1 > 0 \\ 0 & \text{if } \mathbf{R}_1 = 0 \end{cases}, \quad \mathbf{Q}_2 = \begin{cases} \mathbf{V}_2/\mathbf{R}_2 & \text{if } \mathbf{R}_2 > 0 \\ 0 & \text{if } \mathbf{R}_2 = 0 \end{cases},$$

therefore the FDR of Procedure 2.2 is $E(\mathbf{Q}_2)$.

Conditioning on the number of discoveries and the number of false discoveries at the initial study, $\mathbf{R}_1 = r_1$ and $\mathbf{V}_1 = v_1$, and using Theorem 2.1 with $v_1$ and $r_1$ taking the role of $m_0$ and $m$, we get: $E_{R_1=r_1,V_1=v_1}(\mathbf{Q}_2) \leq v_1 \cdot q_2/r_1$ for $r_1 > 0$, and $\mathbf{Q}_2 \equiv 0$ if $r_1 = 0$. To complete the proof we use the independence of the two studies and express the FDR of procedure 2.2,

$$\begin{aligned} E(\mathbf{Q}_2) &\leq q_2 \cdot \{\sum_{r_1=1}^{m} \sum_{v_1=0}^{r_1} \Pr(\mathbf{R}_1 = r_1, \mathbf{V}_1 = v_1) \cdot \frac{v_1}{r_1} + \Pr(\mathbf{R}_1 = 0) \cdot 0\} \\ &= q_2 \cdot E(\mathbf{Q}_1) \\ &\leq q_2 \cdot \frac{m_0 q_1}{m}. \end{aligned}$$

Where the last inequality holds since the BH procedure is also used in the initial stage of Procedure 2.2.

| | Resamp. FDR | BH procedure | Resamp. FWE |
|---|---|---|---|
| Mean Z-score rejection threshold | 2.81 | 2.79 | 3.84 |
| FDR level | 0.039 | 0.038 | 0.002 |
| FWE level | 0.65 | 0.63 | 0.042 |
| Mean number of true discoveries | 207.4 | 205.5 | 90.9 |

Table 1: Comparison of FDR and FWE procedures - single trait

| | Resamp. FDR | BH procedure | Resamp. FWE |
|---|---|---|---|
| Single trait | 3.06 $(2.2 * 10^{-3})$ | 3.08 $(2.1 * 10^{-3})$ | 3.81 $(1.4 * 10^{-4})$ |
| 2 traits | 2.95 $(3.2 * 10^{-3})$ | 2.92 $(3.5 * 10^{-3})$ | 4.01 $(6.0 * 10^{-5})$ |
| 4 traits | 2.96 $(3.1 * 10^{-3})$ | 2.92 $(3.5 * 10^{-3})$ | 4.15 $(3.3 * 10^{-5})$ |
| 8 traits | 2.90 $(3.7 * 10^{-3})$ | 2.89 $(3.8 * 10^{-3})$ | 4.36 $(1.3 * 10^{-5})$ |

Table 2: Mean Z-score (and p-value) rejection thresholds - multiple traits