False Discovery Rate and Correction for Multiple Comparisons in Linkage Disequilibrium Genome Screens

Chiara Sabatti, Susan Service, and Nelson Freimer

C. Sabatti is with the Human Genetics and Statistics Departments at UCLA, 695 Charles Young Drive South, Los Angeles, CA 90095-7088, USA. Phone: (310) 794-9567. Fax: (310) 794-5446. E-mail: csabatti@mednet.ucla.edu.

S. Service is with the Center for Neurobehavioral Genetics, UCLA

N. Freimer is with the Center for Neurobehavioral Genetics and the Human Genetics Department, UCLA.

Abstract

Population based linkage disequilibrium genome screens represent one of the most recent approaches for the localization of genes responsible for complex diseases. One open problem in this context is represented by the definition of an appropriate significance threshold that takes into account the multiple comparison problem. We explore the conceptual and practical implications of the multiple testing procedure known as False Discovery Rate (FDR). We argue that controlling the FDR better represents the interest of researcher in this area than more traditional approaches. We then explore the applicability of the Benjamini-Hochberg (BH) FDR controlling procedure in the specific context of association mapping from case-control data. We analyze the nature of dependency between the test statistics with analytic work and simulations and we conclude that the BH rule effectively controls FDR in our context of interest. The dependency between test statistics translates into a decrease of power, which highlights the necessity of developing resampling based rules to control FDR.

Keywords

Significance cut-off, linkage mapping, linkage disequilibrium, association mapping, Bonferroni, Family Wise Error Rate.

I. Introduction

Association studies have been proposed multiple times as a viable mapping strategy for complex diseases (see for ex. Risch and Merikangas, 1996). Some of the most recent work describing the patters of recombination in chromosomes (Daly et al. 2001) suggest that there are blocks of highly conserved haplotypes. A mapping strategy for the future could be based on studies designed to analyze the association between disease traits and SNPs representative of each of these blocks. These collection of SNPs would be on the order of hundred of thousands. Even before such high-resolution genotyping becomes possible, it is clear that a successful association mapping procedure has to be able to deal with the problem of multiple comparisons across thousands of tests. Our first goal is to analyze the potential of relatively new paradigm for multiple testing in the context of genemoscreen association studies. In 1995 Benjamini and Hochberg proposed to control, when correcting for multiple comparisons, the proportion of wrongly rejected null hypothesis over all the rejections. They named this quantity False Discovery Rate and they gave a simple step-wise rule that selects significant results, guaranteeing a FDR of a specified level, among independent tests. In 1998 Weller et al. proposed the application of this procedure in a mapping context with reference to the QTL methodology and simultaneous investigation of multiple traits. In the most recent years the theoretical understanding of the FDR rule described by Benjamini and Hochberg has considerably improved. In particular, it has become apparent that the same rule controls FDR when the test statistics are dependent. This makes it applicable to a much wider set of problems, including complex disease mapping based on association tests—which will be the focus of our investigation.

In section 1 we introduce the problem of multiple comparison in the context of gene localization, with specific reference to the well studied case of linkage mapping. In section 2 we give some background on general statistical approaches to multiple comparison; we formally introduce the notion of FDR; we describe Benjamini and Hochberg (BH) step-wise rule; and we discuss why we believe this approach to multiple comparison to be particularly suitable for the mapping of complex diseases. In section 3 we analyze the applicability of the BH procedure in the context of genome screens, and in particular population based association studies. We study to the nature of dependence between the various test-statistics and its implications. Section 4 describes the results of a simulation study that confirms our results of section 3 and allows us to compare the power of different correction procedures. We conclude that the BH procedure effectively controls FDR in association studies. It results in a significant power increase with respect to procedures that control family-wise error rates. However, it is also apparent that FDR rules constructed taking into account the existing dependency would result in further a increase of power.

II. THE PROBLEM OF SIGNIFICANCE CUT-OFFS IN GENE MAPPING STUDIES

In order to understand the parameters of our problem, it is useful to review the approaches to establish significant cut-off values in the well-studied linkage analysis for monogenic diseases. As the type of data-sets available for mapping have evolved, so have the criteria used to assess the significance level.

Morton popularized the the cut-off value of 3 for the lod-score in a world where few markers were available. In such a context, a stringent cut-off was required in view of the considerably small chance that one of the few markers available could turn out to be close to the locus under investigation. Formally, assuming a prior probability of 0.2 for an analyzed marker to be actually linked to the disease under study, Morton concluded that one could stop collecting families where the lod score reached 3. The same Bayesian argument can be applied to a non sequential procedure and translates, again, to a cut of significance level of 3.

As the number of markers available for analysis increased substantially, the perspective changed. Since a considerable number of markers are spread around the genome, the prior probability that one of them is linked to the disease is high. However, conducting multiple tests makes it necessary to worry about multiple comparisons. Even if the disease was not genetic and all the null hypotheses of no linkage to each of the considered markers were true, the simple fact of looking at many markers increases the probability of finding one that shows a significant pattern. To address this issue, two simplifying hypotheses on the marker structure have been made, corresponding to the idea of a “sparse map” and a “dense map” (Lander and Botstein, 1989). In the first case, markers are assumed to be independent, in the second case, markers are assumed to cover the entire genome so that the lod scores observed are actually a continuous process. In both these frameworks, lod-score values of 3-3.5 appear to produce good evidence for linkage. In the sparse map assumption, consider a typical genome screen with 400 markers. Then one can apply Bonferroni correction and obtain that to have an overall level of significance of 0.05, one need an individual p-value lower than 0.0001, corresponding to a lod-score of 3.3. This approximation is not valid as we further increase the number of markers, as the assumption of independence between the markers becomes increasingly unrealistic and would lead to an unnecessary loss of power. In this context, the dense map approximation is useful as the tests for linkage can be shown to follow a Ornstein-Uhlenbeck process and extremal probabilities from this process can used to define the level of significance (Feingold et al, 1993), and obtaining results comparable to the 3.3 cut-off value.

A further parameter that has been introduced over time in the discussion on the appropriate cut-off values for a mapping study is the increased interest in complex diseases, where more than one locus is expected to be involved. In the context of linkage studies, Lander and Botstein (1986) and Depuis, Brown and Siegmund (1995) compared the performances of marginal search (focusing on one locus at the time), and simultaneous and conditional search that are carried out under the explicit assumption that more than one locus should be involved. From a practical standpoint, the conclusion of these studies has been that even if conditional and simultaneous search are potentially more powerful, they require such high levels of corrections for multiple comparison that they are often not worth pursuing. Certainly marginal search plays the leading role in practical applications.

On this background, the following statements can be made about the specific characteristics of Linkage Disequilibrium (LD) mapping.

- The sparse map assumption of independent markers appears not realistic. LD studies are based on thousand of markers and the distance between then is such that often they are in linkage disequilibrium in a random sample of individuals (see Service et al. 2001)
- So far a continuous process approximation of LD tests has not been obtained and it seems difficult to do so, given the indirect nature of LD tests (which are not based on counting recombination events or sharing), the variability of the LD patterns across the genome, and the dependence of the test statistics on auxiliary variables—as the number of alleles at a given marker or their distribution.
- LD studies are almost exclusively considered in the context of complex diseases, where more than one locus is expected to play a role.

In reference to the first two points, we are going to consider the discrete model that studies n tests corresponding to n locations in the genome. Let then $T_1, \dots, T_n$ be a set of test statistics for testing the hypotheses $\{H_1, \dots, H_n\}$, where H_i is true if marker i is not linked to a disease locus. We are not going to assume that T_i are independent. In order to develop a procedure that corrects for multiple comparison but is also reasonably sensitive to signal associated to different markers we propose to adopt the FDR viewpoint. We devote the following section to a description of the issues involved in selecting a multiple comparison procedure and the specific characteristics of FDR. In order to provide a vivid illustration of the various methodologies reviewed we will often refer to their implications in linkage mapping.

III. Multiple comparison procedures and control of the False Discovery Rate

Let then $T_1, \dots, T_n$ be a set of test statistics for testing the hypotheses $\{H_1, \dots, H_n\}$, where H_i is true if marker i is not linked to a disease locus. Let H_0 be the hypothesis that corresponds to each of the H_i being true $H_0 = \cap_{i=1}^n H_i$. When we conduct tests of these hypotheses, we try to answer two types of questions: (1) can H_0 be rejected? (2) If H_0 is rejected, which of the H_i should be rejected? The statistical technique used to answer the first question is called global test, while the one addressing the second is named multiple test procedure. Note that in the case of a disease whose genetic component is well established, the goal of a global test is really to determine if the sample size and the available genotypes are enough to resolve the location of the susceptibility genes. When the disease is monogenic, a positive answer to this question coincides with the identification of the disease locus. When, instead, multiple genes are known to play a role, addressing question (2) is a really different problem. The well known Bonferroni procedure offers an easy answer to both problems. Let $p_1, \ldots p_n$ be the p-values associated with each of the tests statistics and let $p_{(1)}, \ldots, p_{(n)}$ be their ordered counterpart. According to Bonferroni, one can reject H_0 if $p_{(1)} < \alpha/n$, where α is the desired level for the test of H_0. As for the second question, the hypothesis H_i for which $p_i < \alpha/n$ will be rejected. This, as illustrated in the previous section, is equivalent to rejecting H_i for which the lod-score be larger than 3.3. The idea behind the Bonferroni correction is very simple. Suppose each hypothesis is tested at the α/n level

$$Pr(\text{Reject } H_0 | (\text{when } H_0 \text{ is true}) = Pr(p_{(1)} \leq \alpha/n | H_0) \leq \sum_{i=1}^{n} Pr(p_i \leq \alpha/n | H_0),$$

which, assuming that the p_i are uniform under H_0–which is true when the T are continuous, p_i are exact and can be very far from true when they are approximated–, is equal to α. It is clear that the procedure is conservative, as based on an inequality. It is difficult to improve on a general level such procedure, and there is a vast literature in statistics on this topic that we cannot consider here. However, there are three aspects that are important for our problem.

(1) If the test statistics are positively correlated, the Bonferroni procedure is much too conservative. This is best exemplified by looking at the extreme case where all the tests are the same

$$Pr(\text{Reject } H_0 | (\text{when } H_0 \text{ is true}) = Pr(p_{(1)} \leq \alpha/n | H_0) = Pr(p_1 \leq \alpha/n | H_0) = \alpha/n.$$

The actual level of the test is now α/n; in other words, we did not need any correction. In general, it is known that if the tests statistics are positively dependent, the correction proposed by Bonferroni is too strong. It is however very difficult to construct rules that take into account a general form of dependence. In most cases, the literature resorts to re-sampling techniques to estimate the distribution of $p_{(1)}$ given the particular structure of the data.

(2) The other characteristic of the Bonferroni procedure that reduces its power is that it is a single step method, that is all the p-values from the various statistics are compared to the same benchmark value. In contrast to this, step-wise methods generally work on ordered set of p-values and have a different cut-off value for each $p_{(i)}$. One of the first procedures of this kind was developed by

Holm. The idea behind these methods is that once the H_i corresponding to $p_{(1)}$ has been rejected, we should believe that it is false and then we are now fishing among only $n - 1$ hypothesis, so that the appropriate cutoff for the second significant result should be $\alpha/(n - 1)$. Indeed, the following procedure is an acceptable alternative to Bonferroni:

(FWER) Start with $i = 1$. If $p_{(i)} > \alpha/(n-i+1)$ accept $H_{(i)}, \ldots, H_{(n)}$ and stop. Otherwise, reject $H_{(i)}$ and continue.

If we applied this criteria to the sparse map approximation for the linkage genome screen, we would obtain that the first significant locus has to give a lod-score of 3.2868, the second a lod score of 3.2860, etc, as illustrated in figure 1. Clearly, this is not a terribly interesting modification of the procedure, as the variation of the cut-off values in the range of interest is practically nonexistent. A much different picture can be obtained if we consider the following rule:

(BH) Proceed from $i = n$ to $i = n - 1$ et cetera, until, for the first time, $p_{(i)} \leq i\alpha/n$. Denote that i by k and reject all $H_{(i)}$ with $i = 1, \ldots, k$.

If, once again, we report this to the sparse map assumption, and monitor the implied cut-off values in term of lod-score we get the situation illustrated in Figure 1.

[Figure 1 about here.]

Clearly, now, we have a significant departure from the request of a uniform cut-off of 3. This procedure has appeared a number of times in the multiple comparison literature with out becoming really successful, as it does not control the same error measure that is controlled by Bonferroni. In particular, it behaves as Bonferroni in terms of global test, but not as a multiple comparison procedure.

(3) The Family Wise Error Rate (FWER) is defined as the probability to wrongly reject at least one H_i. The idea to control the FWER has been, for a long time, the dominating paradigm in statistics for analyzing the problem of multiple comparisons. There are two precise definitions of FWER: as the probability of rejecting at least one hypothesis H_i when they are all true, or the probability or rejecting at least one of the true hypotheses, regardless which these are. A method that controls the FWER in the first definition is said to control it in a weak sense and a method that controls the FWER as defined in the second way is said to control it in the strong sense. The FWER is a somewhat natural extension of the significance level of a single test to the context of multiple

tests. In some sense, it represents the "p-value of the p-value": it gives the probability of observing a p-value as low as the minimum one when all the null hypothesis are true. Certainly this similarity with the familiar p-value concept has contributed to the popularity of such a conservative criteria, at least on a theoretical level. From the practical standpoint, however, its short-comings have been noted a number of times (a nice discussion can be found in Benjamini and Yekutieli, 2001): Lander and Kruglyak (1995), for example, resort to the mythological analogy of "choice between Shylla and Charybdis" to describe the difficult balance between FWER and power. FWER is an appropriate measure of error when we strongly desire not to make any wrong rejections. However, it does not necessarily reflect the attitude of researchers, who often are interested in measuring the overall error of multiple tests, in a framework that is similar to a classification procedure, where of main relevance is the percentage of errors, rather that the presence of at least one of them. So, even if a wrong rejection of the null hypothesis is the error that one wants to control, as in classical testing, what matters is the fraction of wrongly rejected hypotheses rather than on their absolute number. In general, it is perceived as a less serious problem to falsely reject one H_i if we correctly rejected 100, than if it is the only rejected one. Specifically to model this behavior, Benjamini and Hochberg in 1995 introduced a new paradigm for approaching the problem of multiple comparisons, which is now receiving a considerable amount of attention in the statistics and data mining literature. They suggest shifting the attention from the probability of wrongly rejecting at least one null hypothesis to the expected fraction of mistakes among the rejected hypothesis, which they call False Discovery Rate. If all the H_i are true, the two approaches coincide, so that FDR and FWER lead to similar global test conclusions. Their difference is really as multiple comparison procedures. In general, if one controls FWER, FDR is also controlled, but not viceversa. As FDR is a less stringent criteria, it is also more powerful. In particular, the step-down procedure leading to the lod-score cutoffs of figure 1 controls FDR. We believe that this paradigm is particularly relevant in the case of complex disease mapping. It is becoming clearer that genome-wide studies have to be considered precisely as screening tools, rather than experiments that will immediately lead to the identification of the disease gene. Because of the complex nature of the disease, we expect that more than one locus in the genome may be implicated, so that we are effectively interested in multiple hits. In this context, one wants to be able to follow all the good leads that may result in the identification of a disease locus. While too many wrong clues are to be avoided as costly, what really matters is the proportion of these over the total number of clues that are warranted further investigation.

Before turning to consideration of the association genome screens, we want to point out an other application of the FDR principle in the context of linkage study that illustrates its potential. The (BH) rule we presented applies to independent tests, so that its immediate application in linkage screens is under the sparse map assumption. However, one can also adopt the FDR approach under the continuous map hypothesis. Consider the Grandparent-Grandchild model of Feingold et al. (1993). The Ornstein-Uhlenbeck process offers a good approximation of the dependence structure between test statistics on one chromosome. Markers on different chromosomes are independent. Indeed, in Feingold et al. (1993) the cut-off value is obtained by setting the per-chromosome significance level to 0.05/25 (they are looking at the mythical unicorn that has 25 chromosomes of equal length). If we follow FDR approach, while using the same extremal probabilities to correct for multiple comparisons within a chromosome, we can increase our power on the independent comparisons. The simplest way is to consider the 25 hypothesis of linkage to any locus in each of the chromosomes. We can calculate the p-value for each of these using the Ornstein-Uhlenbeck based approximation of Feingold et al. We would then reject the hypothesis of no linkage using the FDR threshold that we have repeatedly presented: this would lead to cut-off values for subsequent chromosomes ordered in terms of decreasing evidence for linkage reported in Figure 2.

[Figure 2 about here.]

To then determine which are the locations that are associated with linkage in the previous chromosomes, we would look at the locations where the maximum of the lod-scores is higher than the lod-score associated with p^*, where p^* is the p-value of the largest i for which $p_i < \alpha 25/i$. This is an example of exploiting the specific dependence-independence structure to construct a multiple comparison procedure that is based on the FDR approach and increases the power of traditional constructions.

IV. Controlling FDR in population based LD genome-screens

We are now going to explore how the FDR control can be applied in the case of population based LD genome screens for the mapping of complex disease. The clear interest of the FDR controlling strategy is in its increased power in circumstances where more than one of the null hypothesis is false, which is what one expects if there is more than one gene influencing the disease. The step-down procedure that we have described in the previous section controls FDR if the test statistics $T_1, \ldots T_n$ are independent, which is what is assumed in the sparse-map approximation. However, the data collected so far on the levels of background LD suggest that this is not a realistic hypothesis in the case of tests of association between a disease and a set of finely spaced markers. There are two implications coming from the departure from independence. The first one that the step-down procedure may not control FDR for tests with generic dependence. The second one is a consideration similar to the one already discussed for the Bonferroni procedure: the cut-off value of α/n for $p_{(1)}$ may be excessively conservative if the tests are positively associated. We have already commented on the second implication of dependence among the test statistics, while the first may appear contradictory. Indeed, loss of power is generally associated with *positive* dependence, while increased error rates with *negative* dependence. For illustration, one may want to consider the example of Hochberg and Rom (1996): two normally distributed test statistics with negative correlation and the BH rule lead to FEWR larger than α.

The rest of the paper will deal to the investigation of how these two issues can be dealt with in the context of LD mapping.

Controlling FDR under dependency: Recent work of Benjamini and Yekutieli shows that under some forms of dependence (Positive Regression Dependency on each one from a Subset - PRDS), the procedure described in (BH) controls FDR. If the tests statistic under association satisfy this requirement, then, we can use the presented step-down procedure and be reassured that it will control the overall FDR. 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 non-decreasing in x. Benjamini and Yekutieli (2001) were able to prove that the procedure illustrated for independent tests, also control the FDR at a level $n_0/n\alpha$, where n_0 is the number of false null hypotheses, if the joint distribution of the tests statistics are PDRS on the subset of test statistics corresponding to the true null hypothesis. The definition of PRDS may seem rather arcane. However, one illustration with reference to linkage should serve to clarify the nature of this hypothesis and illustrate its adaptability to the mapping context. Benjamini and Yekutieli (2001) show that PDRS translates in the following requirement for multivariate normal tests statistics. Consider $X \sim N(\mu, \Sigma)$, 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 hypothesis, $\mu_i = 0$; 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. If we now consider the Gaussian models for genetic linkage analysis proposed by Feingold et al. (1993), it is easy to see that they satisfy this condition. Consider the model proposed for Grandparent-Grandchild pairs. If we restrict our attention to a finite subset of genome locations we get a multivariate Gaussian. The mean values of the test statistics at each un-linked location is 0 and it is positive for linked loci. The covariance between two test-statistics are non negative and a function of the recombination fraction across loci. Because the covariances are non negative, we can conclude that the tests are PRDS on I_0 and hence the cutoff values illustrated in Figure 1 are actually guaranteed to control the FDR, even when we relax the independence assumption.

It should be clear by know that PRDS is a property that is likely to hold also for linkage disequilibrium test statistics in the sense that if two markers are in LD and one happens to show random association to the disease, the other marker in LD would have increased chances of showing association higher than a given threshold. Because we do not have a general model for the dependency between tests of association, nor there is likely to be a realistic one in the near future, it is difficult to translate this intuitive idea in a precise general statement. However, we can consider some specific cases that capture the essence of the dependence between test of association involving markers in LD. The detail study of these simplified cases will shed some light on general patterns and illustrate the meaning of the PRDS condition.

Consider the situation in which one examines the association between a series of SNP and a disease in a case-control study, where N disease and N control haplotypes are sampled. If the SNPs are in linkage equilibrium, the tests statistic will be independent, while if they are in LD they are not. In the following analysis, we will focus on the dependence between the test of association originated by two SNPs, between which there is a specified amount of LD. This is representative of the joint distribution of any set of SNPs when one assumes a Markovian structure of the first order for the LD between SNPs.

Let's then consider two SNPs, their joint distribution can be represented as follows:

$M1 \setminus M2$	1	2	
1	$pq+\delta$	$p(1-q)-\delta$	p
2	$(1-p)q-\delta$	$(1-p)(1-q)+\delta$	$1-p$
	q	$1-q$	1

,

where the parameters p and q represent the population frequencies of the 1 alleles in the two SNPs and δ the amount of association existing between them. Let X_1 be the total number of allele 1 associated with disease in the first marker, and Y_1 the total number of allele 1 associated with control in the same marker. Suppose there is no association between the SNPs under consideration and the disease. Then $X_1 \sim$ Binom(p, N) and $Y_1 \sim$ Binom(q, N), which entirely specifies the distribution of the contingency table T_1, collecting allele counts for SNP 1 and disease status.

$T_1 =$

$Dis. \setminus M1$	1	2	
D	X_1	$N-X_1$	N
ND	Y_1	$N-Y_1$	N
	X_1+Y_1	$2N-X_1-Y_1$	$2N$

$T_2 =$

$Dis. \setminus M2$	1	2	
D	X_2	$N-X_2$	N
ND	Y_2	$N-Y_2$	N
	X_2+Y_2	$2N-X_2-Y_2$	$2N$

In order to evaluate the distribution of $T_2|T_1$, 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, is such that both can be viewed as the sum of two independent binomial components: $X_2 \sim$ Binom$(q + \delta/p, x_1) +$ Binom$(q - \delta/p, N - x_1)$ and $Y_2 \sim$ Binom$(q + \delta/p, y_1) +$ Binom$(q - \delta/p, N - y_1)$. Using this expression for the joint distribution of T_1 and T_2, we can evaluate some properties of the dependence between the p-values of tests of association that are relevant to establish if they are PRDS. As an example, we generated 5000 tables using two SNPs, each with allele frequency .5 and a background linkage disequilibrium parameter $\delta = .2(-.2)$. There was no association between disease locus and the considered SNPs. Figures 3 and 4 report the observations for X_1 and X_2, the scatter plot of the p-values for association tests based on tables T_1 and T_2 (p_1 and p_2) and the empirical equivalent of $\Pr(p_2 > a|p_1 = b)$, for $a = 0.3, 0.5, 0.7, 0.9$ and $b = 0.1, \ldots, 0.9$, calculated binning the values of p_1 and p_2 in equally spaced intervals between 0 and 1.

[Figure 3 about here.]

[Figure 4 about here.]

The latest probability is at the base of the definition of PRDS. In order to check that the p-values of association tests satisfy the PRDS on the collection of true null hypothesis, we should verify that non-decreasing relationships as the ones in figures 3(c) and 4(c) hold for any collection of tests conditional on any p-value corresponding to a true null hypothesis. This task cannot, obviously, be carried trough by systematic enumeration. What we have done is simply to verify numerically the one-dimensional implications of the PRDS property. Additionally, we can observe that if the markers corresponding to the p-values whose probability is evaluated are in linkage equilibrium with the table corresponding to the conditional p-value, the relation will necessarily be true. Moreover, one can restrict one,s attention to the p-values of true null hypothesis, as the p-values corresponding to H_0^c, should also be, by definition, independent from any p-value corresponding to H_0.

Another source of dependency that plays an important role in determining the joint distribution of p-values from association studies, is the fact that the same markers may be involved in the definition of different haplotypes, whose distribution in cases and controls may be the precise object of comparison. We consider two models to describe the nature of this dependence. In order to distinguish it clearly from the one studied in the previous example, we assume that different markers are in linkage equilibrium. The first model is really based on an approximation of haplotype and uses χ^2 tests. The second model relies again on Fisher's exact test. Again suppose that each marker is a SNP. To mimic the effect of considering haplotypes, one can think of adding the χ^2 statistics across the SNPs in a sliding window. This is less stringent than considering the chi-square table deriving from the haplotype table, but mimics the consideration of historical recombination and mutation that are incorporated in much of the haplotype methods. Consider the case where we add the chi-square statistics from 3 adjacent SNPs. Assume that the χ^2s are independent—this is of interest only because we have already explored the effect of LD in the previous example. Let X_i be the counts of allele 1 on disease chromosomes at marker i. Let $Y = X_1 + X_2 + X_3$ and $W = X_2 + X_3 + X_4$. It is easy to check that W is positively regression dependent on Y, as long as Y satisfies the null hypothesis. We have

$$\begin{aligned} P(W \geq w|Y \geq y) &= P(X_2 + X_3 + X_4 \geq w|Y = y) = \\ &\quad P(\frac{X_2 + X_3}{Y} Y + X_4 \geq w|Y = y) = P(Ry + X_4 \geq w|Y = y), \end{aligned}$$

where R is distributed as a Beta of parameters 1/2 and 1, and X_4 is independent from Y. This probability is clearly increasing in y. Again, to fully check the PDRS condition we have to look at joint conditional distributions.

An other—more direct—way of looking at the distribution of haplotype tests is to study the case of 2 overlapping SNPs haplotypes. We assume that there is no association with the disease and that the markers 2 and 3 are in the same dependence relation above. The table of counts for the association between disease and haplotypes of the first two markers is as follows:

$$T_1 = \begin{array}{c|c|c|c|c|c} Dis. \setminus H12 & 11 & 12 & 21 & 22 & \\ \hline D & X_1 & X_2 & X_3 & X_4 & N \\ \hline ND & Y_1 & Y_2 & Y_3 & Y_4 & N \\ \hline & X_1 + Y_1 & X_2 + Y_2 & X_3 + Y_3 & X_4 + Y_4 & \end{array},$$

where (X_1, X_2, X_3, X_4) and $(Y_1, \ldots, Y_4)$ are multinomials of parameters depending on the haplotype frequencies. Sliding down of one marker, the table counts will have the following shape:

$$T_2 = \begin{array}{c|c|c|c|c|c} Dis. \setminus H23 & 11 & 12 & 21 & 22 & \\ \hline D & X_1' & X_2' & X_3' & X_4' & N \\ \hline ND & Y_1' & Y_2' & Y_3' & Y_4' & N \\ \hline & X_1' + Y_1' & X_2' + Y_2' & X_3' + Y_3' & X_4' + Y_4' & \end{array},$$

with conditional distribution described by $X_1' \sim$ Binom$(q + \delta/p, X_1 + X_3)$, $X_2' = X_1 + X_3 - X_1'$, $X_3' \sim$ Binom$(q - \delta/p, X_2 + X_4)$,$X_4' = X_2 + X_4 - X_3'$. Again, we used this joint distribution of T_1 and T_2 to analyze the validity of the PRDS requirements on one dimensional distributions. For example, we considered again 5000 tables, generated with SNPs with allele frequency 0.5. We assumed independence across SNPs and with the disease and evaluated the p-value of association tests between two overlapping haplotypes and the disease locus. The results are in figure 5: again, PRDS seems a likely structure for this problem.

[Figure 5 about here.]

The results in this section suggest that positive regression dependency from the subset of the null hypothesis test statistics is likely to hold in the context of LD studies. However, we did not prove theoretically this to be the case and we have made a series of simplifying assumptions. The simulation study carried out in the following will contribute to clarify the scope of these results.

Power of BH procedure under dependency: If all the tests are perfectly dependent, we should not correct for multiple comparisons: this is true for FDR as much as for FWER. The best way to investigate the extent of the loss of power under these circumstances is the use of simulations, which we will illustrate in the next section. The most effective way of correcting this problem relies on developing a precise model for the dependency and incorporating it in the definition of an FDR controlling procedure. This does not appear as an easily attainable goal, as the collected evidence on distribution of LD across genome and populations shows a significant, unexplained, variability. A more practical strategy relies, instead, on permutation-based evaluations of FDR within the sample of interest. While there has been some work describing appropriate permutation methods (Yakutieli and Benjamini and Storey and Tibshirani), the design of an efficient resampling scheme that is practical for genome screen purposes is still needed. On this background, the BH rules remains very attractive as computationally inexpensive. We hence decided to study its performance in presence of the dependence structure that characterizes the association studies with a set of simulations.

V. Simulation Study: empirical FDR and power

Our simulation study had two goals: (1) to validate the conjectures of our previous section that the BH rule control FDR in the case of association tests and (2) to assess the extent of the loss of power that is due to dependency among the test statistics.

Because the multiple comparison procedures considered differ only when more than one null hypothesis is false, as in the case of multiple genes influencing the diseases, we considered a situation where we have three susceptibility genes of equal importance and acting independently. These genes are located on three different chromosomes. The total number of chromosomes is 22; they are assumed of equal length for a total genome length of 3300 cM. The data analyzed to investigate association consist on a sample of 200 diseased aneuploid individuals and 200 control ones: that is, for each individual, we analyzed 22 chromosomes. One third of the 200 diseased individuals was a carrier of each of one of the three disease loci. To mimic a typical setting for association genome screens, we assumed that 1100 SNPs, each 3cM a part, covered the genome. The haplotypes of susceptibility genes carrying chromosomes was generated assuming one founding event (for that location), 15 generations old. This determined the distribution of the closest recombination events on the two sides of the disease locus. Outside the conserved region, the disease chromosome were modeled as the control ones with a Markov process of the first order.

In order to cover an interesting range of settings, we considered three degree of LD among adjacent SNPs and two levels of powers. The levels of LD are described by the parameter λ as in

$$\begin{aligned} Pr(SNP_i = 1|SNP_{i-1} = 1) &= Pr(SNP_i = 1) + (1-\lambda)Pr(SNP_i = 2) \\ Pr(SNP_i = 1|SNP_{i-1} = 2) &= Pr(SNP_i = 1) - (1-\lambda)Pr(SNP_i = 2) \end{aligned}$$

Linkage equilibrium corresponds to $\lambda = 0$; we indicated as low LD a scenario where λ varies uniformly in [0,0.1]; medium LD is characterized as $\lambda \in [0.2, .4]$; and high LD to $\lambda \in [0.8, 1]$.

The different power levels are achieved making differential assumptions on the frequencies of the alleles associated with disease locus on the founder chromosomes. In analyzing the results we considered both a single marker test and an haplotype test. Because of the intrinsic power differences in these approaches, it was instructive to use different parameter settings in the two cases in order to achieve comparable power levels. In detail, for the single marker case, the high power scenario corresponds to a frequency of the associated alleles on the ancestral disease chromosomes that varies uniformly between 0.2 and 0.3; in the low power case such frequency varies uniformly between 0.3 and 0.4. For the haplotype test, the frequency of the associated alleles on the ancestral disease chromosomes varied uniformly between 0.3 and 0.4 for the high power case and 0.4 and 0.5 for the low power case.

The results of genome-scans for 200 diseased and two hundred controls individuals were simulated 1000 times for each of the described scenarios. Both in the single marker case and in the two-markers haplotype test we used Fisher's exact p-values for the association tests. Tables 1 and

2 report three measures of power, the average FDR and the average FWER across replicates.

[Table 1 about here.]

[Table 2 about here.]

Here the false discovery rate is defined as

$$FDR = \begin{cases} \frac{V}{V+S} & \text{if} \quad V+S > 0 \\ o & \text{otherwise} \end{cases} ,$$

where V are false rejections and S correct rejections of the null hypothesis. We considered as true null hypothesis all the H_0 relative to markers that are more than 3 SNPs away from the true disease locus and that are on chromosomes that do not carry any disease locus. Attention has to be paid to the fact that the variance of the results for high LD is considerably greater than in the other scenarios, due to the high levels of dependency.

The first evident conclusion from the table is that the BH method achieves control of the FDR: the average estimated FDR are below 0.05. (notice, incidentally, that the method only controls expected FDR, so that in one replicate the FDR may actually be higher).

Secondly, it is evident that FDR leads to an increased power with respect to FWER. We measured power in different ways. For each of the three disease locations we calculated the percentage of times in which they are detected. Since these three locations are essentially identical in terms of simulation parameters, we averaged their results in what is called “marginal power” column. To emphasize the fact that the increased power of FDR is due to an increased power of detecting more than one locus, we then also evaluated the percentage of time in which at least two locations were detected and in which all three locations where detected with the three methods. The increase in power ranges from 25% in the marginal power to over 100% for the three-hit power.

A third aspect worth noting is that the FDR (as FWER) seems to be controlled at a level lower than 0.05, which is the cut-off we set. This is due to the dependency structure and results in a loss of power. To quantify the extent of this power loss, for the single marker test, we considered the power of an genome screen identical to the one of our simulations, but where all the marker are independent—which is reported under the column of no LD. It can be seen that the loss of power due to unaccounted positive dependence between test statistics is indeed significant for both FWER and FDR controlling methods, ranging from 10% to over 100%, in the extreme case of low power and very high dependence. This result argues in favor of the necessity of developing adequate resampling-based evaluation 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.

VI. CONCLUSIONS

We have illustrated conceptually and with numerous references to linkage studies how controlling the False Discovery Rate is a most satisfying procedure for correcting for multiple comparisons and how it translates in substantial increase of power. We have then analyzed in particular the case of genome screens with case-control data and the performance of the Benjaminin Hochberg procedure in this setting.

The simple analytic models and the simulations results documented in this study suggest that the BH methodology can be effectively used to correct for multiple comparison in the case of LD genome screens. Additionally, it is evident that the development of other procedures that explicitly take into account the dependency between tests would result in an increase of power. Further work is needed to investigate the applicability in this context of re-sampling-estimates of FDR and the development of novels multiple comparison procedures.

REFERENCES

Benjamini, Y. and 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. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under independence, The Annals of Statistics, to appear.

Daly, M. Rioux, J., Schaffner, S., Hudson, T. and Lander, E. (2001) High-resolution haplotype structure in the human genome, Nature genetics 29:229-232.

Efron, B, Tibshirani, R., Storey, J. and Tusher, V. (2001) Empirical bayes analysis of a microarray experiment, to appear in JASA.

Holm, S (1979) A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6: 65-70

Lander, E.and Botstein (1989) Mapping mandelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185-190

Lander, E and Kruglyak, L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nature Genetics 11:241-247.

Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. Science 273:1516-1517.

Service SK, Ophoff RA, Freimer NB. (2001) The genome-wide distribution of background linkage disequilibrium in a population isolate.Human Molecular Genetics 10: 545–51.

Weller, Song, Heyen, Lewin, Ron (1998) A new approach to the problem of multiple comparisons in the genetic dissection of complex traits, Genetics 150:1699–1706

Yekutieli, D. and Benjamini, Y. (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics, Journal of Statistical Planning and Inference 82:171-196

LIST OF FIGURES

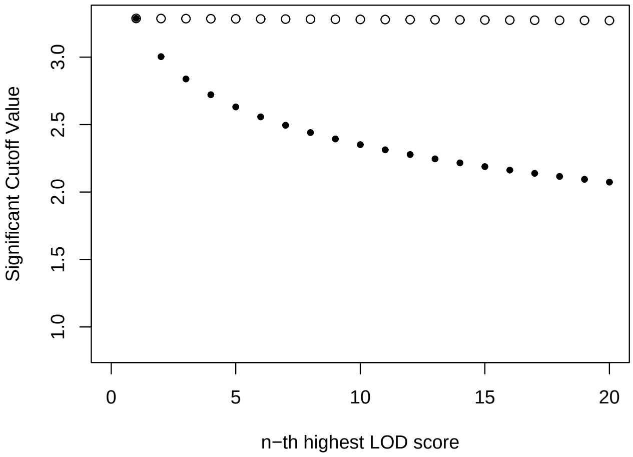

Fig. 1. Significance Cut-off of the LOD-Score based on a genome-screen with 400 independent markers

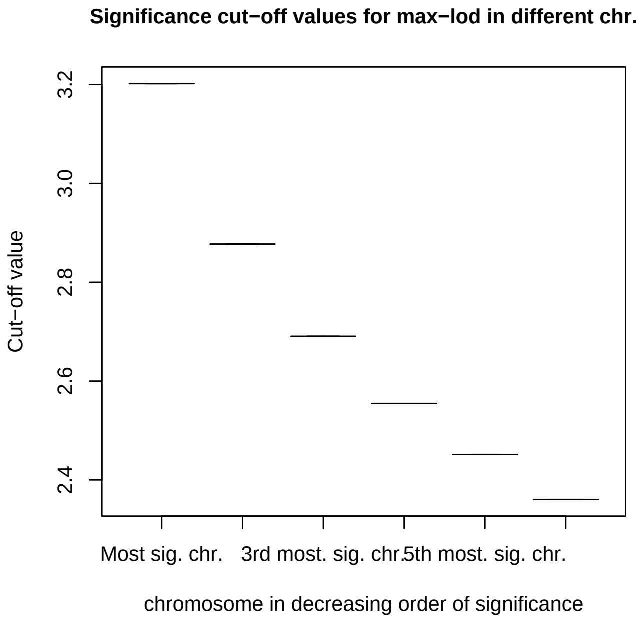

Fig. 2. Cut-off values for lodscore on sequential chromosomes

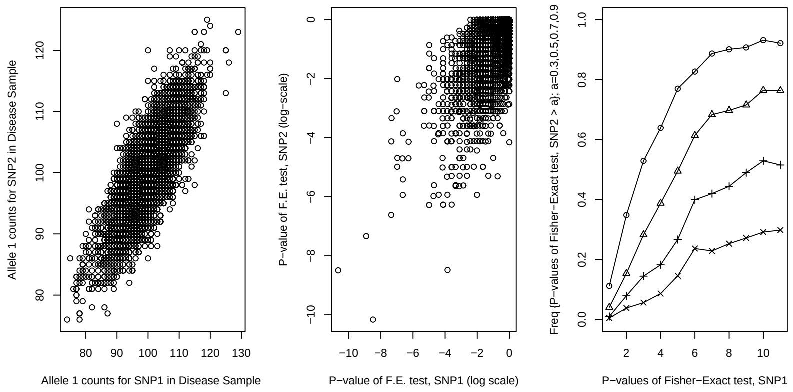

Fig. 3. Dependency between the p-values of two tables

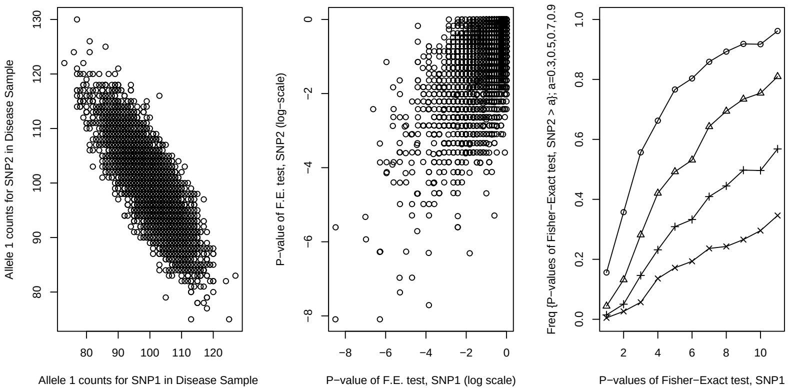

Fig. 4. Dependency between the p-values of two tables

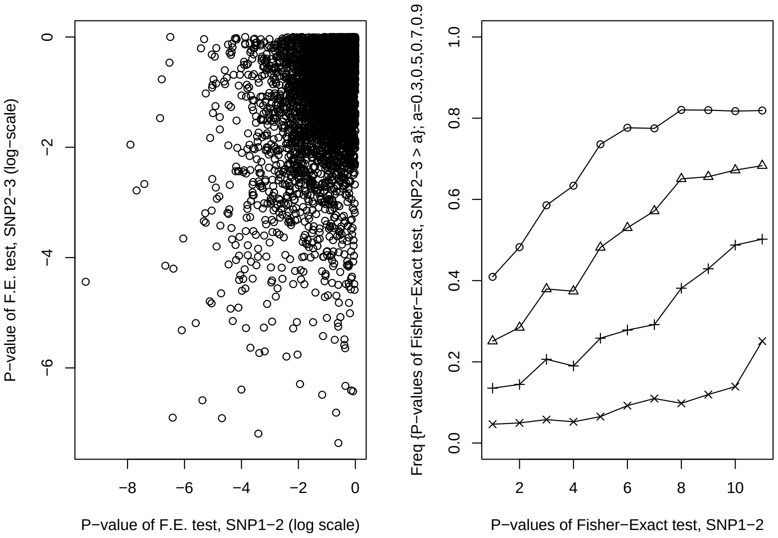

Fig. 5. dependency between the p-values of sliding haplotypes tests

LIST OF TABLES

Low Power

	False Positives			Marginal Power			Power ≥ 2			Power 3			FWER est.			FDR est.		
LD values	No	Med	High	No	Med	High	No	Med	High	No	Med	High	No	Med	High	No	Med	High
No MCP	42.96	43.60	43.09	99.9	99.7	97.3	100	100	99.7	99.7	99.3	92.2	1	1	1	0.8	0.8	0.787
Bonferroni	.045	0.032	0.04	47.3	41.3	32.3	45.1	38.2	24.2	11.8	6.3	4.2	0.04	0.03	0.025	0.02	0.015	0.011
Step FWER	.046	0.032	0.04	47.3	41.3	32.3	45.1	38.2	24.2	11.8	6.3	4.2	0.045	0.03	0.025	0.02	0.015	0.011
B-H FDR	.17	0.147	0.272	59.4	52.7	40.6	63.5	55.1	39	27.4	21	11.7	0.153	0.124	0.09	0.045	0.035	0.034

High Power

	False Positives			Marginal Power			Power ≥ 2			Power 3			FWER est.			FDR est.		
LD values	No	Med	High	No	Med	High	No	Med	High	No	Med	High	No	Med	High	No	Med	High
No MCP	43.65	43.58	44.43	100	100	99.6	100	100	100	100	100	99	1	1	1	0.77	0.77	0.76
Bonferroni	0.041	0.042	0.029	78.3	78.3	63.2	87.5	88.6	70.5	48.3	47	24.2	0.04	0.04	0.02	0.01	0.009	0.0052
Step FWER	0.041	0.042	0.031	78.3	78.3	63.2	87.5	88.6	70.5	48.3	47	24.2	0.04	0.04	0.02	0.010	0.009	0.0054
B-H FDR	0.247	0.289	0.362	90.9	89.8	77.7	96.6	95.9	85.7	76.5	73.8	51.5	0.203	0.24	.13	0.037	0.041	0.034

TABLE I

SIMULATION RESULTS. SINGLE MARKER TESTS.

Low Power

	False Positives		Marginal Power		Power ≥ 2		Power 3		FWER est.		FDR est.	
LD values	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High
No MCP	53.3	52.7	99.7	98.9	100	99.9	99.9	96.9	1	1	0.81	0.80
Bonferroni	0.035	0.047	43.7	33.6	40.7	26.7	8.7	4.3	0.03	0.04	0.014	0.023
Step FWER	0.035	0.047	43.7	33.6	40.7	26.7	8.7	4.3	0.03	0.04	0.014	0.023
B-H FDR	0.223	0.239	56.3	44.1	60.3	45.2	24.5	14.9	0.16	0.12	0.044	0.028

High Power

	False Positives		Marginal Power		Power ≥ 2		Power 3		FWER est.		FDR est.	
LD values	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High
No MCP	53.2	53.2	99.9	99.7	100	100	99.9	99.2	1	1	0.78	0.78
Bonferroni	0.06	.06	77.9	56.6	88	59.6	46.8	17.7	0.052	0.042	0.011	0.016
Step FWER	0.06	0.06	77.9	56.7	88	59.6	46.8	17.8	0.052	0.042	0.011	0.016
B-H FDR	0.44	0.43	90.4	71.5	96.1	78.8	75.7	41.2	0.30	0.22	0.048	0.052

TABLE II

SIMULATION RESULTS. TWO HAPLOTYPE TESTS.