Appl. Statist. (2005)
54, *Part* 1, *pp.* 143–158

A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics

Peter B. Gilbert

Fred Hutchinson Cancer Research Center and University of Washington, Seattle, USA

[Received February 2003. Revised January 2004]

Summary. To help to design vaccines for acquired immune deficiency syndrome that protect broadly against many genetic variants of the human immunodeficiency virus, the mutation rates at 118 positions in HIV amino-acid sequences of subtype C *versus* those of subtype B were compared. The false discovery rate (FDR) multiple-comparisons procedure can be used to determine statistical significance. When the test statistics have discrete distributions, the FDR procedure can be made more powerful by a simple modification. The paper develops a modified FDR procedure for discrete data and applies it to the human immunodeficiency virus data. The new procedure detects 15 positions with significantly different mutation rates compared with 11 that are detected by the original FDR method. Simulations delineate conditions under which the modified FDR procedure confers large gains in power over the original technique. In general FDR adjustment methods can be improved for discrete data by incorporating the modification proposed.

Keywords: Bonferroni; False discovery rate; Genetics data; High dimensional data; Human immunodeficiency virus vaccine trial; Hypothesis testing; Simultaneous inference

1. Introduction

Consider the following problem in genetics. Data are available on two sets of amino-acid sequences, aligned such that all the sequences have the same number of amino-acids. (Amino-acids are the basic building-blocks of proteins; there are 20 amino-acids, denoted by capital letters.) For each sequence set, the degree of polymorphism at each position in the sequences can be measured by the frequency of non-consensus amino-acids at the position. (The consensus amino-acid is the modal amino-acid for the sequence set.) The problem that is addressed here is how to identify the 'differentially polymorphic' positions, i.e. the positions at which the probability of a non-consensus amino-acid differs between the two sequence sets. For example, position 6 is differentially polymorphic if the consensus amino-acid has frequency 0.87 for the first set and 0.52 for the second set. This two-sample problem occurs in many genetics studies; here we focus on the application of developing a vaccine for acquired immune deficiency syndrome.

The development of an efficacious preventive vaccine for the human immunodeficiency virus (HIV) is challenged by the extensive genetic heterogeneity of HIV (World Health Organization–UNAIDS, 2001; Gaschen *et al.*, 2002). Within a phase III trial of a HIV vaccine, an important objective is to assess how the level of efficacy of the vaccine to prevent infection depends on genetic characteristics of the exposing HIV (such an analysis has been named 'sieve analysis';

Address for correspondence: Peter B. Gilbert, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, PO Box 19024, Seattle, WA 98109-1024, USA.
E-mail: pgilbert@scharp.org

 0035–9254/05/54143

Berman *et al.*, 1997). Gilbert *et al.* (1999, 2000) developed statistical methods for sieve analysis, in which, for each HIV-infected participant, a genetic distance is computed between the infecting virus and the HIV represented in the vaccine, and the distributions of these distances are compared between the vaccine and placebo arms. The distance between two HIV sequences can be defined as the percentage mismatch of amino-acids with appropriate weighting of positions for their immunological significance, and an important research area is how to select the weights (see De Groot *et al.* (2002)). In this paper, we consider the analysis of a data set that informs this choice of weighting. Specifically, several vaccines are under development that are based on a subtype C strain of HIV that includes the gag p24 protein (Graham, 2002), and there are plans to test these vaccines in regions where both subtype C and subtype B HIVs circulate (subtype C is the globally predominant HIV subtype and subtype B predominates in North America and Europe; Papathanasopoulos *et al.* (2003)). For such vaccines, a genetic distance can be defined on the basis of the amino-acids in the gag p24 sequence, and one weighting strategy upweights positions that have a different degree of polymorphism among subtype C and B viruses. Differential polymorphism at a position may reflect an important immunobiological difference in the subtypes at this location and suggests that a mismatch of the amino-acid at this position in the exposing HIV sequence compared with that in the vaccine sequence might be key for causing failure of a vaccine. Upweighting such positions can improve the power of sieve analyses. In addition, for forthcoming phase III HIV vaccine trial data sets, a useful sieve analysis would compare the polymorphism of positions between the vaccine and placebo sequences. Positions with greater polymorphism in vaccine than in placebo sequences are flagged as positions at which 'vaccine resistance' mutations may have occurred, i.e. amino-acid mutations that allowed HIV to break through the immune responses induced by vaccination. Identifying the differentially polymorphic positions would aid the design of a vaccine by suggesting the positions at which multiple different amino-acids should be added to the vaccine formulation to broaden the protection that is conferred by the vaccine.

To study differential polymorphism in gag p24, we analyse a data set of 146 gag p24 amino-acid sequences, with half the sequences sampled from Southern Africans infected with a subtype C HIV (group 1, $n_1 = 73$ individuals) and half the sequences sampled from North Americans infected with a subtype B HIV (group 2, $n_2 = 73$ individuals) (Novitsky *et al.*, 2002a; Kuiken *et al.*, 2002) (Fig. 1). In addition to informing the weighting of amino-acid metrics, this assessment is useful for designing 'cytotoxic T-lymphocyte (CTL) epitope cocktail' vaccines, a leading vaccine approach (De Groot *et al.*, 1997, 2002; Novitsky *et al.*, 2002b). Such vaccines contain many HIV epitopes (i.e. contiguous strips of 8–11 HIV amino-acids that induce CTL immune responses, e.g. PIVQNLQGQ). At epitope regions with positions with different mutation rates among subtype C and B HIVs, it is important to include multiple different epitope sequences in the vaccine, to maximize the number of HIVs that the vaccine can protect against. Thus, identifying differentially polymorphic positions guides the design of CTL epitope cocktail vaccines by informing where to place multiple HIV epitope sequences.

The 146 amino-acid sequences were aligned to have all the same length (231 positions). The alignment was constructed by using ClustalX version 1.81 (Thompson *et al.*, 1997) and manually edited by using BioEdit (Hall, 1999). The alignment is correct with high probability because most of the positions in gag p24 sequences are conserved. To attempt to identify the differentially polymorphic positions, 231 hypothesis tests can be conducted, one for each position. The investigator then faces the question of how to identify the set of significant results while controlling the false positive error rate. This problem motivates the statistical issue that is addressed in this paper, which has broad application to genetics and other fields: how to control the false positive rate when carrying out a large number of hypothesis tests for data with discrete distributions?

Multiple-comparisons procedures that control the familywise error rate FWER, such as the widely used Bonferroni method, provide stringent type I error control but often allow a high rate of type II errors, and the power sharply diminishes with the number of tests (Hochberg and Tamhane, 1987; Westfall and Young, 1993; Hsu, 1996). For many practical applications, it is preferable to apply a multiple-comparisons procedure that controls the risks of type I and II errors more evenly. This objective is addressed by sequential *P*-value methods that control the false discovery rate (FDR), i.e. the expected proportion of rejected hypotheses that are false rejections. Such methods have been used successfully in many applications including wavelet analysis of signal processing (Abramovich and Benjamini, 1996), genome scans for locating traits (Weller *et al.*, 1998; Storey and Tibshirani, 2003), psychiatric research (Mallet *et al.*, 1998), educational research (Williams *et al.*, 1999) and deoxyribonucleic acid microarray experiments (Efron *et al.*, 2001).

Benjamini and Hochberg (1995) developed the first FDR controlling method and showed that it provides large power gains over FWER controlling methods that increase with the number of tests. Despite its power advantages, the original FDR procedure is often quite conservative, with the true FDR well below the prespecified FDR. Several modifications of the original FDR method have been proposed that can provide improved power (e.g. Benjamini and Liu (1999) and Benjamini and Hochberg (2000)). However, when the test statistics have discrete distributions, apparently none of the published FDR controlling methods explicitly account for the discrete characteristics, which can be exploited to improve the power considerably.

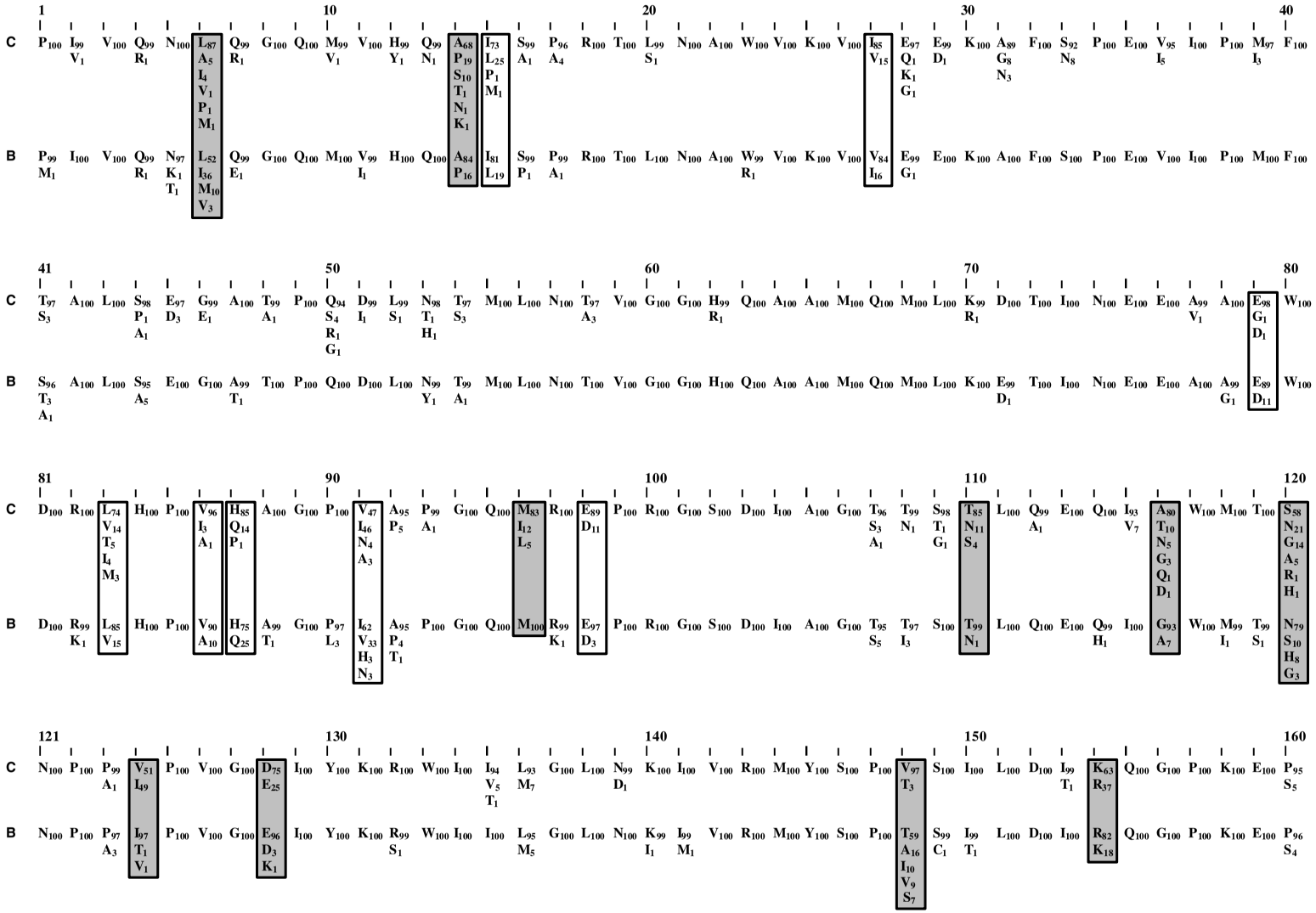
For FWER controlling multiple-comparisons procedures, Westfall and Wolfinger (1997) and others have shown that accounting for the discreteness of the data can yield large gains in power. In particular, Tarone (1990) developed a modified Bonferroni method for discrete data. In this paper, we show that Tarone's (1990) modification also applies to FDR controlling procedures, with a similar capacity to improve the power. The newly proposed procedure is applied to the HIV sequence data set in Section 2. Section 3 provides background on the method of Benjamini and Hochberg (1995) and Tarone's (1990) method, and Section 4 describes the modification of Benjamini and Hochberg's FDR procedure. The new and existing procedures are evaluated in a simulation study in Section 5. Section 6 provides a discussion of several recently developed FDR controlling procedures and makes the point that these techniques can be improved for discrete data by incorporating the modification proposed. The data and the programs that were used to analyse them can be obtained from

`http://www.blackwellpublishing.com/rss`

2. Example

The goal of the HIV sequence problem is to identify the positions in gag p24 amino-acid sequences at which the probability of a non-consensus amino-acid differs between the sets of subtype C and B sequences. Of the 231 positions in HIV gag p24, 113 have the modal amino-acid in all 146 sequences, i.e. are perfectly conserved. Thus, $m = 118$ positions contribute comparative information. Let p_{1i} and p_{2i} respectively be the probabilities of a non-consensus amino-acid at position i for group 1 and group 2 sequences; the goal is to test simultaneously $H_i : p_{1i} = p_{2i}, i = 1, \ldots, 118$.

Fisher's exact test was used to compute 118 unadjusted *P*-values. An alternative approach to this problem would compare the probabilities of the 19 possible amino-acid substitutions from the consensus amino-acid between the sets of sequences, using Fisher's exact test generalized for 20-category variables. However, the data set has few positions with more than two


1 10 20 30 40
C P100 I99 V1 V100 Q99 R1 N100 L87 A5 I4 V1 P1 M1 Q99 R1 G100 Q100 M99 V1 V100 H99 Y1 Q99 N1 A68 P19 S10 T1 N1 K1 I73 L25 P1 M1 S99 A1 P96 A4 R100 T100 L99 S1 N100 A100 W100 V100 K100 V100 I85 V15 E97 Q1 K1 G1 E99 D1 K100 A89 G8 N3 F100 S92 N8 P100 E100 V95 I5 I100 P100 M97 I3 F100
B P99 M1 I100 V100 Q99 R1 N97 K1 T1 L52 I36 M10 V3 Q99 E1 G100 Q100 M100 V99 I1 H100 Q100 A84 P16 I81 L19 S99 P1 P99 A1 R100 T100 L100 N100 A100 W99 R1 V100 K100 V100 V84 I16 E99 G1 E100 K100 A100 F100 S100 P100 E100 V100 I100 P100 M100 F100
41 50 60 70 80
C T97 S3 A100 L100 S98 P1 A1 E97 D3 G99 E1 A100 T99 A1 P100 Q94 S4 R1 G1 D99 I1 L99 S1 N98 T1 H1 T97 S3 M100 L100 N100 T97 A3 V100 G100 G100 H99 R1 Q100 A100 A100 M100 Q100 M100 L100 K99 R1 D100 T100 I100 N100 E100 E100 A99 V1 A100 E98 G1 D1 W100
B S96 T3 A1 A100 L100 S95 A5 E100 G100 A99 T1 T100 P100 Q100 D100 L100 N99 Y1 T99 A1 M100 L100 N100 T100 V100 G100 G100 H100 Q100 A100 A100 M100 Q100 M100 L100 K100 E99 D1 T100 I100 N100 E100 E100 A100 A99 G1 E89 D11 W100
81 90 100 110 120
C D100 R100 L74 V14 T5 I4 M3 H100 P100 V96 I3 A1 H85 Q14 P1 A100 G100 P100 V47 I46 N4 A3 A95 P5 P99 A1 G100 Q100 M83 I12 L5 R100 E89 D11 P100 R100 G100 S100 D100 I100 A100 G100 T96 S3 A1 T99 N1 S98 T1 G1 T85 N11 S4 L100 Q99 A1 E100 Q100 I93 V7 A80 T10 N5 G3 Q1 D1 W100 M100 T100 S58 N21 G14 A5 R1 H1
B D100 R99 K1 L85 V15 H100 P100 V90 A10 H75 Q25 A99 T1 G100 P97 L3 I62 V33 H3 N3 A95 P4 T1 P100 G100 Q100 M100 R99 K1 E97 D3 P100 R100 G100 S100 D100 I100 A100 G100 T95 S5 T97 I3 S100 T99 N1 L100 Q100 E100 Q99 H1 I100 G93 A7 W100 M99 I1 T99 S1 N79 S10 H8 G3
121 130 140 150 160
C N100 P100 P99 A1 V51 I49 P100 V100 G100 D75 E25 I100 Y100 K100 R100 W100 I100 I94 V5 T1 L93 M7 G100 L100 N99 D1 K100 I100 V100 R100 M100 Y100 S100 P100 V97 T3 S100 I100 L100 D100 I99 T1 K63 R37 Q100 G100 P100 K100 E100 P95 S5
B N100 P100 P97 A3 I97 T1 V1 P100 V100 G100 E96 D3 K1 I100 Y100 K100 R99 S1 W100 I100 I100 L95 M5 G100 L100 N100 K99 I1 I99 M1 V100 R100 M100 Y100 S100 P100 T59 A16 I10 V9 S7 S99 C1 I99 T1 L100 D100 I100 R82 K18 Q100 G100 P100 K100 E100 P96 S4

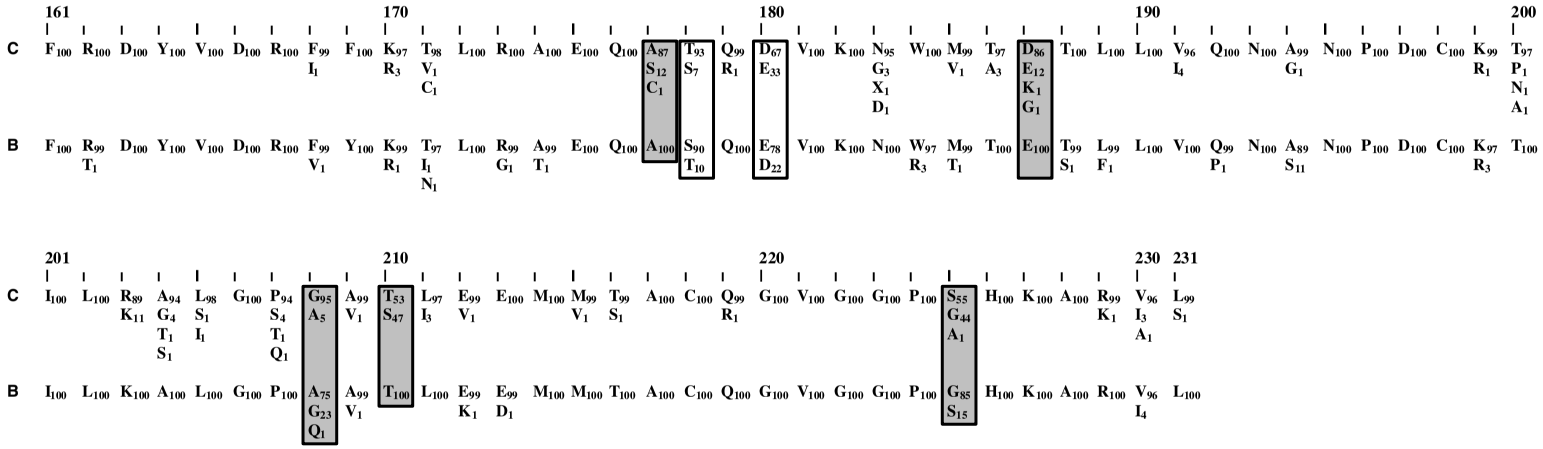


Fig. 1. Amino-acid frequencies at each of the 231 positions in gag p24 for the 73 subtype C HIV gag p24 sequences sampled from Southern Africans and for the 73 subtype B HIV gag p24 sequences sampled from North Americans: I_{99} indicates that amino-acid I was present in 99% of the 73 sequences; the $m(K) = 25$ positions that have enough variability possibly to reject $H_i : p_{1i} = p_{2i}$ using the Tarone modification are highlighted with boxes; shaded boxes indicate the 15 positions that were inferred to have significantly differential polymorphism among the C- and B-sequences according to the Tarone modified FDR procedure with adjustment; positions 14, 116, 120 and 154 were not significant by Benjamini and Hochberg's (1995) FDR procedure

amino-acids represented appreciably in either set of sequences, so measuring polymorphism by binary response variables captures most of the available information (see Fig. 1).

With $\alpha = 0.05$, all P-values from Fisher's exact test less than $0.05/118 = 0.000\,424$ are significant by the ordinary Bonferroni method. Applying Tarone's (1990) modification of the Bonferroni procedure that is described in Section 3.1, we compute the sequence $m(1), m(2), \ldots,$ and stop at the first value $K = k$ such that $m(k) \leqslant k$. We obtain $m(K) = K = 25$, and thus R_K contains the 25 indices with $\alpha_i^* < 0.05/25 = 0.002$ (these positions are highlighted with boxes in Fig. 1). Because $\eta_\cdot = \Sigma_{i \in R_K} \eta_i = 0.035\,11$ is less than $\alpha = 0.05$, the significance cut-off value 0.002 can be further improved. Implementing the approach that is described in the last paragraph of Section 3.2, the Tarone Bonferroni procedure was applied to a sequence of α-values ranging from 0.05 to 0.10 in increments of 0.001. The largest α such that $\eta_\cdot \leqslant 0.05$ was 0.081, for which $m(K) = K$ was 27, yielding an improved cut-off value $0.081/27 = 0.003$, which we refer to as the Tarone Bonferroni cut-off with adjustment. Using the three Bonferroni procedures, 6, 9 and 10 of the positions have significant P-values after multiple-comparisons adjustment.

Next, we apply Benjamini and Hochberg's (1995) FDR procedure to the 118 positions. The largest ordered P-value $P_{(i)}$ with $P_{(i)} \leqslant (i/118)0.05$ is 0.004 52, so all positions with $P_i \leqslant 0.004\,52$ are significant, yielding 11 positive results. With the new Tarone-modified FDR procedure that is described in Section 4.1, the largest $P_{(i)}$ in R_K with $P_{(i)} \leqslant (i/25)0.05$ is 0.015 26. The increase in cut-off value from 0.004 52 to 0.015 26 represents an improvement, yielding 13 compared with 11 positive results. Finally, the modified FDR procedure with adjustment was carried out, in which the modified FDR procedure was repeated for several values of $\alpha \geqslant 0.05$. The largest $\alpha^* > \alpha$ such that

$$\mathrm{FDR}^* = \frac{1}{m(K)^*} \sum_{i=1}^{m(K)^*} \eta_{lim(K)^*} \leqslant 0.05$$

is 0.058, for which $m(K)^* = K^* = 25$ and $R_K^* = R_K$, and the Tarone-modified FDR procedure conducted at level 0.058 (instead of 0.05) yields a cut-off value of 0.0338, and 15 positive results. In sum, of the six multiple-comparisons procedures that were used, the five improvements all successively make more discoveries (6, 9, 10, 11, 13 and 15).

Because the FDR procedures are only guaranteed to control FDR $\leqslant 0.05$ if the test statistics are independent or satisfy certain conditions such as positive regression dependence (see Section 4.2), it is possible (but unlikely) that the FDR procedures were liberal. To account for this possibility, Benjamini and Hochberg's (1995) FDR procedure was rerun at level $0.05/\Sigma_{i=1}^{118}(1/i) = 0.0093$ and the Tarone-modified FDR procedures were rerun at level $0.05/\Sigma_{i=1}^{25}(1/i) = 0.0131$, to provide upper bound conservative procedures that are guaranteed to control FDR $\leqslant 0.05$ for completely general test statistics. Six, 10 and 10 hypotheses are rejected by Benjamini and Hochberg's (1995) FDR procedure and the Tarone-modified FDR procedure without and with adjustment respectively.

In conclusion, the newly proposed FDR procedure identified 15 positions in gag p24 that are differentially polymorphic between subtypes C and B (which are highlighted with shaded boxes in Fig. 1). These positions can be upweighted in HIV metrics that are used in sieve analyses of forthcoming efficacy trials of subtype C HIV vaccines that are tested either in a geographic region where both subtype C and subtype B HIVs circulate (e.g. China) or at multiple sites, some with subtype C HIVs and others with subtype B HIVs. In addition, multiple HIV epitopes with different amino-acid sequences can be included in CTL epitope cocktail HIV vaccine constructs that are under development (De Groot *et al.*, 1997, 2003; Novitsky *et al.*, 2002b) at the regions containing the 15 identified positions.

3. Background on existing multiple-comparisons procedures

3.1. Original false discovery rate procedure

Consider the problem of testing simultaneously m null hypotheses $H_1, \ldots, H_m$. Define random variables $\mathbf{R}$ and $\mathbf{V}$ as the number of null hypotheses and the number of true null hypotheses that are rejected by a testing procedure respectively. The FDR is defined as $E(\mathbf{V}/\mathbf{R})\, I(\mathbf{R} > 0)$, i.e. the expected proportion of the rejected null hypotheses that are erroneously rejected, with $\text{FDR} \equiv 0$ if no hypotheses are rejected. Order the m testwise P-values $P_{(1)} \leqslant \ldots \leqslant P_{(m)}$, with $H_{(1)}, \ldots, H_{(m)}$ the corresponding null hypotheses. Fix $\alpha \in (0, 1)$, and let k be the largest i for which $P_{(i)} \leqslant (i/m)\alpha$. Then, Benjamini and Hochberg's (1995) FDR procedure specifies rejecting $H_{(i)}$ for $i = 1, \ldots, k$. For independent, continuous test statistics Benjamini and Hochberg proved that this procedure controls the FDR at α ($\text{FDR} \leqslant \alpha$), and Benjamini and Yekutieli (2001) proved this result for independent discrete test statistics. Furthermore, Benjamini and Yekutieli (2001) proved this result for dependent statistics satisfying certain dependence structures and argued on the basis of simulation experiments and other means that it is expected to hold quite generally. Main advantages of the FDR procedure include that it is often much more powerful than FWER controlling methods and it is simple to use.

3.2. Tarone's (1990) modified Bonferroni procedure

Given P-values $P_1, \ldots, P_m$ for testing m null hypotheses $H_1, \ldots, H_m$, the original Bonferroni procedure rejects all H_i with $P_i \leqslant \alpha/m$. The main virtues of this procedure are that it controls the FWER for any data set, and it is very simple to use. Tarone (1990) proposed a modification to the Bonferroni procedure to make it more powerful for problems with discrete test statistics. To help to describe this procedure as well as the new procedure in Section 4.1, consider the example data set. Let x_{1i} and x_{2i} be the number of observed non-consensus amino-acids in the group 1 and group 2 sample respectively, and set $x_{\cdot i} = x_{1i} + x_{2i}, i = 1, \ldots, 118$. By conditioning on the denominators $n_{1i} = n_{2i} = n_1 = n_2 = 73$ sequences, Fisher's exact test can be applied to obtain unadjusted two-sided P-values $P_1, \ldots, P_{118}$. Let α_i^* be the minimum achievable significance level at position i, which for this example equals

$$\alpha_i^* = 2\binom{n_{1i}}{x_{\cdot i}} \Big/ \binom{n_{1i}+n_{2i}}{x_{\cdot i}} = 2 \times \frac{73!}{(73-x_{\cdot i})!} \Big/ \frac{146!}{(146-x_{\cdot i})!}.$$

Table 1 lists the number of positions i with observed values of x_{1i} and x_{2i} satisfying $x_{\cdot i} = 1, 2, \ldots, 12$ and $x_{\cdot i} > 12$, and the corresponding values of α_i^*. Note that many positions are fairly conserved, with few non-consensus amino-acids, in which cases α_i^* is relatively large. For example, 83 of the 118 positions have fewer than six non-consensus amino-acids, which implies that $\alpha_i^* > 0.05$. As illustrated in the example in Section 2, the large fraction of positions with sizable α_i^* can be exploited to enlarge the number of significant discoveries substantially.

Tarone's (1990) modified Bonferroni method works as follows. For each $k = 1, \ldots, m$, let $m(k)$ be the number of the m tests for which $\alpha_i^* < \alpha/k$, and let K be the smallest value of k such that $m(k) \leqslant k$. Let R_K be the set of indices satisfying $\alpha_i^* < \alpha/K$, which contains $m(K)$ indices. Then, the ith test is deemed significant if i is in R_K and the nominal (testwise) significance level $P_i < \alpha/K$. This procedure always controls the FWER rate at level α, with the probability of at least one false rejection (the FWER) bounded by $m(K)\alpha/K \leqslant \alpha$.

Tarone (1990) showed that the Bonferroni method can be improved further by considering values η_i, $i \in R_K$, with η_i the largest achievable significance level such that $\eta_i \leqslant \alpha/K$. Because $\text{FWER} = \Sigma_{i \in R_K} \eta_i \equiv \eta_\cdot$, it follows that, if $\eta_\cdot < \alpha$, then it may be possible to expand the rejection region. Tarone suggested a systematic approach whereby the tail outcome of smallest probability

Table 1. Observed frequencies of non-consensus (i.e. non-modal) amino-acids for the two groups of HIV amino-acid sequences of length 118 positions

Number of positions	x_i/n_i *for group with smaller* x_{1i}, x_{2i}	x_i/n_i *for other group*	$x_.$	α_i^*	η_i†	P_i
50	0/73	1/73	1	1.00	ND	1.00
7	0/73	2/73	2	0.50	ND	0.50
7	1/73	1/73	2	0.50	ND	1.00
2	0/73	3/73	3	0.24	ND	0.24
6	1/73	2/73	3	0.24	ND	1.00
2	0/73	4/73	4	0.12	ND	0.12
2	1/73	3/73	4	0.12	ND	0.62
1	2/73	2/73	4	0.12	ND	1.00
5	0/73	5/73	5	0.058	ND	0.058
1	2/73	3/73	5	0.058	ND	1.00
1	0/73	6/73	6	0.028	ND	0.028
1	2/73	4/73	6	0.028	ND	0.68
1	3/73	3/73	6	0.028	ND	1.00
2	3/73	4/73	7	0.013	ND	1.00
2	0/73	8/73	8	0.0064	ND	0.0064
1	4/73	4/73	8	0.0064	ND	1.00
1	1/73	8/73	9	0.0030	ND	0.033
1	4/73	5/73	9	0.0030	ND	1.00
1	0/73	10/73	10	0.0014	0.0014	0.0014
2	2/73	8/73	10	0.0014	0.0014	0.097
1	3/73	7/73	10	0.0014	0.0014	0.33
1	0/73	11/73	11	0.00065	0.00065	0.00065
1	0/73	12/73	12	0.00030	0.00030	0.00030
1	1/73	11/73	12	0.00030	0.00030	0.0045
18			>12	<0.00030	$\leqslant 0.002$	

†ND, not defined.

not included in the current rejection region is sequentially added to make the final rejection region as near as possible to α. This improvement is relatively complicated to implement, and in practice it may be preferable to use a simpler adjustment, in which the unmodified Tarone procedure is repeated for a grid of values $\alpha^* \geqslant \alpha$. The rejection decisions are based on the procedure that is performed using the value of α^* for which $\eta_.$ is closest to α without exceeding it.

4. Modification of the original false discovery rate procedure

4.1. Combining false discovery rate and Tarone's modification

Benjamini and Yekutieli's (2001) theorem 5.1 and its proof illuminate why the original Benjamini and Hochberg (1995) FDR procedure can be made more powerful for discrete data. For independent test statistics, theorem 5.1 states that the Benjamini and Hochberg procedure conducted at level α controls the FDR at exactly $(m_0/m)\alpha$ for continuous test statistics, and at level less than or equal to $(m_0/m)\alpha$ for general test statistics. Equality holds for continuous test statistics because the m P-values are uniformly distributed under the null hypotheses in this case, which implies that $\Pr\{P_i \leqslant (k/m)\alpha\} = (k/m)\alpha$ for all $i, k = 1, \ldots, m$, which is a key step in the proof. For discrete statistics, $\Pr\{P_i \leqslant (k/m)\alpha\}$ may be less than $(k/m)\alpha$ and, the greater the gaps between these terms, the greater the opportunity to improve the power.

The newly proposed modified FDR procedure is a simple two-step combination of the Tarone and BH procedures: first, compute the integer K and the subset of $m(K)$ indices R_K among the m

hypotheses as described in Section 3.2; second, perform Benjamini and Hochberg's (1995) FDR procedure at level α on the subset of hypotheses R_K. Because K, $m(K)$ and R_K are calculated on the basis of marginal information only (pooled over the two groups), it follows that the FDR procedure conducted on the subset of indices in R_K controls the FDR at level less than or equal to $\{m_0(K)/m(K)\}\alpha \leqslant \alpha$, where $m_0(K)$ is the number of true null hypotheses among the $m(K)$ tests. The modified FDR procedure simplifies to the original Benjamini and Hochberg FDR procedure if R_K consists of all m positions ($K = m(K) = m$), and usually has a larger rejection region otherwise, with the gain in power generally increasing with $m - m(K)$.

The improvement of the Tarone Bonferroni procedure that was described at the tail end of Section 3.2 can also be applied to the FDR method to improve it further for some data sets. For each i and $k = 1, \ldots, m(K)$, let η_{ik} be the largest achievable significance level that is less than or equal to $\{k/m(K)\}\alpha$. Then, on the basis of equation (19) in Benjamini and Yekutieli (2001), it follows that, when the above modified FDR procedure is carried out at level α, the FDR is bounded above by

$$\sum_{i=1}^{m(K)} \sum_{k=1}^{m(K)} \frac{1}{k} \eta_{ik} w_{ik},$$

where $\Sigma_{k=1}^{m(K)} w_{ik} = 1$. To derive a simple practical procedure from this inequality that usually controls the FDR, set $w_{ik} = I\{k = m(K)\}$, so that for each i all weight is placed on the largest η_{ik}, $\eta_{im(K)}$. Under this approximation the FDR is bounded by

$$\mathrm{FDR}^* = \frac{1}{m(K)} \sum_{i=1}^{m(K)} \eta_{im(K)},$$

and FDR^* may be substantially less than α. This fact leads to the following procedure: conduct the modified FDR procedure for a grid of values $\alpha^* \geqslant \alpha$ for which $\mathrm{FDR}^* \leqslant \alpha$, and keep the results from the iteration for which FDR^* is as near as possible to α without exceeding it.

4.2. Dependent test statistics

When Benjamini and Hochberg's (1995) FDR procedure controls the FDR, so does the modified FDR procedure. When their procedure is not guaranteed to control the FDR, nor is the modified procedure. To help to explain the effect of dependences in the test statistics on the modified procedure, a summary is given on the known operating characteristics of Benjamini and Hochberg's FDR procedure applied to dependent data. First, a general result establishing that their procedure confers FDR control for dependent test statistics is not available. Benjamini and Yekutieli (2001) proved that Benjamini and Hochberg's procedure controls the FDR for statistics with positive regression dependence, a condition that is satisfied for some statistics of interest such as multivariate normal statistics with non-negative correlations between statistics for true null hypotheses and all the other statistics. Simulation studies by Benjamini *et al.* (1997) showed that Benjamini and Hochberg's procedure controlled the FDR for equally positively correlated normally distributed (possibly Studentized) test statistics. Fisher's exact test statistics that were used in the example have a limiting multivariate normal distribution, with non-negative correlations more plausible than negative correlations, supporting that the FDR procedures are unlikely to be liberal.

Benjamini and Yekutieli (2001) also proved that a simple modification of Benjamini and Hochberg's (1995) procedure controls the FDR for all forms of dependence: conducting the FDR procedure at level $\alpha/\Sigma_{i=1}^{m}(1/i)$ guarantees control of the FDR $\leqslant \alpha$. This modification

also ensures FDR control by the newly proposed procedure for general test statistics: first $m(K)$ and R_K are computed and then Benjamini and Hochberg's FDR procedure is run at level $\alpha/\Sigma_{i=1}^{m(K)}(1/i)$ on the indices in R_K. In addition, Yekutieli and Benjamini (1999) developed a resampling-based FDR controlling procedure that uses information on the dependence structure of the test statistics, and they showed that it can substantially improve the power, especially if many null hypotheses are true and the P-values for testing the true null hypotheses are highly positively correlated. This approach could be incorporated into the new procedure; however, the resampling method is not guaranteed to control the FDR, and more work is needed to characterize its utility in practice.

5. Simulation study

We conduct a simulation experiment to compare the false positive error control and the power of the new Tarone FDR procedure with those of existing multiple-comparisons procedures. Consider a two-sample problem in which a vector of m independent binary responses is observed for each of 100 individuals in each group, and the goal is to test simultaneously the m hypotheses $H_i : p_{1i} = p_{2i}$, $i = 1, \ldots, m$, where p_{1i} and p_{2i} are the success probabilities for the ith binary response in group 1 and group 2 respectively. To study power for high dimensional data sets, we use several large values of m: 100, 200, 400, 800, 1600 and 3200. With $m_1 + m_2 + m_3 = m$, data are generated so that the response is Bernoulli(0.01) at m_1 positions for both groups, is Bernoulli(0.10) at m_2 positions for both groups and is Bernoulli(0.10) at m_3 positions for group 1 and Bernoulli(0.30) at m_3 positions for group 2. The small success probability (0.01) for the m_1 positions reflects the fact that they are quite conserved, with only two successes (non-consensus amino-acids) expected for the two groups combined. The high degree of conservation implies that many of the minimum significance levels α_i^* for these m_1 positions will be large, which provides an opportunity for improving the power of a multiple-comparisons procedure by accounting for the α_i^*. For each of the m_2 positions, there is substantial diversity, so the α_i^* will tend to be too small (less than 0.001) to be useful for improving a procedure. Whereas the null hypothesis is true for the m_1 and m_2 positions, the alternative hypothesis is true for the m_3 positions, for which there is also substantial diversity and little opportunity for the Tarone modification to confer improvement.

Three sets of simulations are run, with m_1 set to be 20%, 50% or 80% of the value m. These cases reflect 'lightly conserved', 'moderately conserved' and 'heavily conserved' data sets, with increasing opportunity for gains in power via the Tarone modification. Given m_1, m_2 is set to be $f = 5\%$, 25%, 50%, 75% of the remaining $m_2 + m_3$ positions. The percentage f represents the proportion of the $m_2 + m_3$ hypotheses for which the null hypothesis is true; in total $f_{\text{null}} = m_2/(m_1 + m_2 f^{-1})\%$ of the m null hypotheses are true. For each of the 72 parameter configurations specified by m, m_1 and f, 5000 data sets are generated. For each data set an unadjusted P-value from Fisher's exact test is computed for each of the m positions at which there is at least one success in the pooled data set, and the six multiple-comparisons procedures that are described in the example are applied. For each parameter configuration the 'size' of each FDR controlling procedure is evaluated by computing the true FDR as the fraction of the rejected hypotheses that are truly null, averaged over the 5000 simulations. In addition, the power of each procedure is evaluated by computing the empirical expectation of the fraction of the m_3 false null hypotheses that are rejected.

Fig. 2 shows the FDRs of the three FDR controlling procedures performed at level 0.05. For all parameter configurations, each procedure is conservative, with the unmodified procedure most conservative and the modified procedure with adjustment least conservative. The

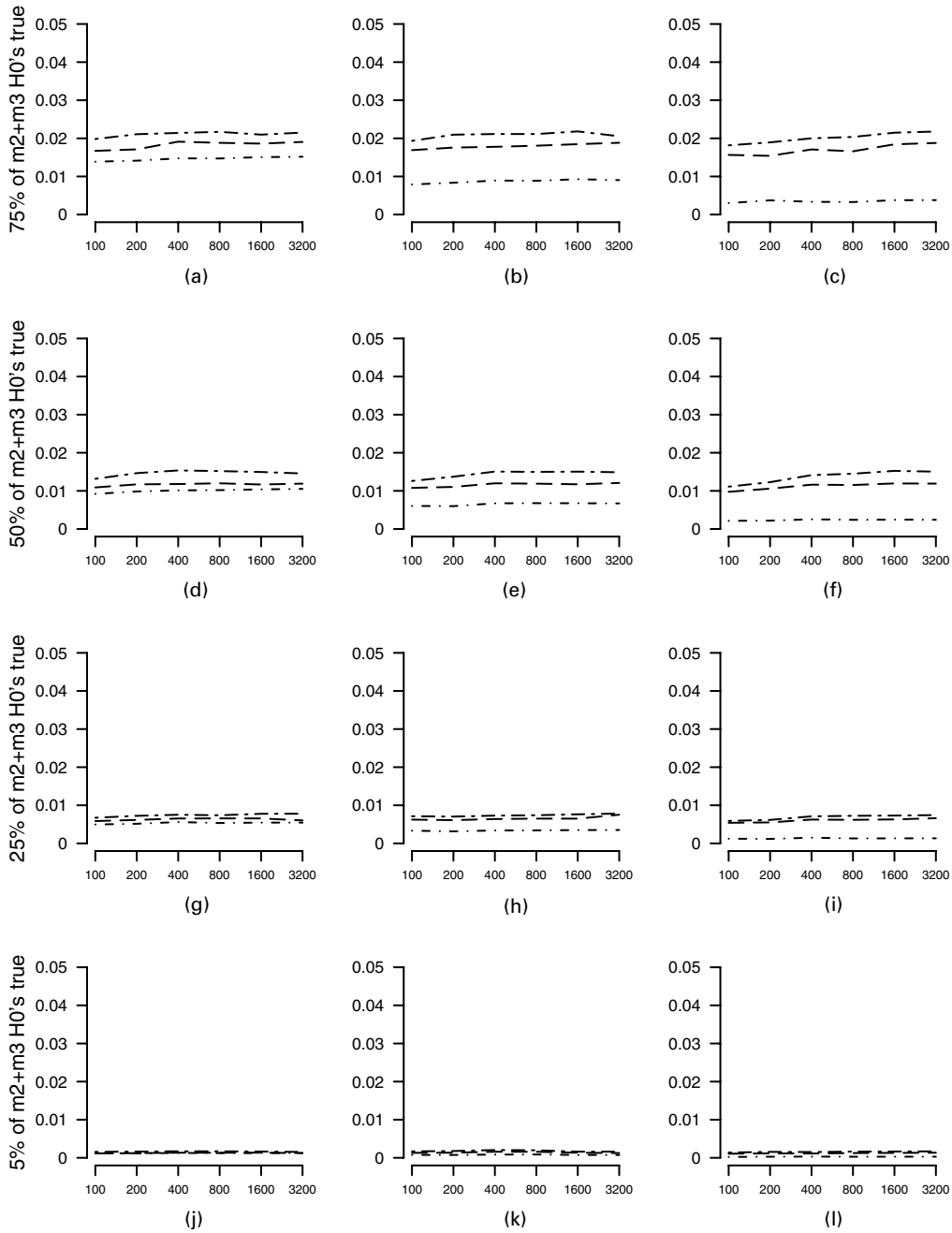


Fig. 2. Based on the simulation study, the figure shows the FDR (the fraction of rejected null hypotheses that are truly null averaged over the 5000 simulations) for Benjamini and Hochberg's (1995) FDR method (–·–·–), the Tarone-modified FDR method (— —) and the Tarone-modified FDR method with adjustment (- — -), for testing simultaneously $m = 100, 200, 400, 800, 1600, 3200$ hypotheses: (a) 60% of m H_0s true; (b) 37.5% of m H_0s true; (c) 15% of m H_0s true; (d) 40% of m H_0s true; (e) 25% of m H_0s true; (f) 10% of m H_0s true; (g) 20% of m H_0s true; (h) 13.5% of m H_0s true; (i) 5% of m H_0s true; (j) 4% of m H_0s true; (k) 2.5% of m H_0s true; (l) 1% of m H_0s true; (a), (d), (g), (j) $m_1/m = 0.2$; (b), (e), (h), (k) $m_1/m = 0.5$; (c), (f), (i), (l) $m_1/m = 0.8$

new methods are consistently closer to the nominal 0.05 level than the original FDR method, with the gap widening with the degree of conservancy of the data set (moving from left to right across the panels). Further note that, although the relative improvement in the modified FDR methods is substantial for all parameter configurations, all three FDR methods become more conservative with the fraction of false null hypotheses and are highly conservative when 95% of the $m_2 + m_3$ null hypotheses are false (Fig. 2(l)).

Fig. 3 shows the average power for the six procedures. We summarize the results. First, the power of the Bonferroni-based methods decreases with the number of tests m, whereas the power of the FDR-based methods is steady with m. Second, the ordinary Bonferroni method always has the lowest power, and the two modified FDR methods always have the highest power. Third, the extent of power gained by the modified FDR methods compared with the original FDR method

(a) does not depend on the number of tests m,
(b) increases slightly with the percentage of true null hypotheses and
(c) increases substantially with the degree of conservatism (i.e. moving from left to right across the panels).

Fourth, when $m \leqslant 200$, the data set is heavily conserved and 75% of the $m_2 + m_3$ null hypotheses are true (Fig. 3(c)), the Tarone modification has a greater effect on the power than the use of the FDR *versus* Bonferroni procedure. Since the majority of null hypotheses are expected to be true in many practical applications, this point highlights that Tarone's idea can be as helpful as the FDR idea for improving the power of a multiple-comparisons testing procedure when the data are discrete and quite 'conserved'. Fifth, the powers of the modified FDR procedures with and without adjustment are comparable; this occurs because for some data sets the adjustment enhances the power and for others it diminishes the power. To see why, note that the unadjusted and adjusted methods carry out Benjamini and Hochberg's (1995) FDR procedure using sequential cut-off levels $i\alpha/m(K)$ and $i\alpha^*/m(K)^*$ respectively, where α^*, $m(K)$ and $m(K)^*$ depend on the particular data set and $\alpha^* \geqslant \alpha$ and $m(K)^* \geqslant m(K)$. These inequalities imply that rejection regions for the adjusted method are sometimes greater than those for the unadjusted method and sometimes smaller, depending on the relative size of the ratios $\alpha/m(K)$ and $\alpha^*/m(K)^*$. In practice, these two ratios can be computed (on the basis of pooled data only) and, if $\alpha^*/m(K)^* > \alpha/m(K)$, then the adjusted method is expected to be more powerful.

Additional simulations were conducted for low dimensional data sets (with $m = 10$), and the Tarone modification was found to provide similar gains in power over the original FDR procedure. Simulations were also performed that were equivalent to the above simulations except that the probabilities $p_{1i} = p_{2i} = 0.1$ for the m_2 null positions were replaced with $p_{1i} = p_{2i} = 0.3$, and the probabilities $p_{1i} = 0.1$, and $p_{2i} = 0.3$ for the m_3 alternative positions were replaced with $p_{1i} = 0.01$ and $p_{2i} = 0.1$. This case was chosen to illustrate that the original FDR procedure sometimes beats the modified procedures. The results on size were similar to those which are reported in Fig. 2, but for lightly and moderately conserved data sets the original FDR procedure consistently had the greatest power, with the gain in power increasing with the number of hypothesis tests. This is explained by the fact that the true success probability is small in group 1 (0.01), which implies that zero successes occur in group 1 for an expected 36.6% of all positions. For these positions, P_i from Fisher's exact test is tiny and equals α_i^*, and many of these indices are excluded from R_K, and therefore these truly false null hypotheses cannot be rejected by the modified FDR methods. The original FDR procedure does have the opportunity to reject these hypotheses, however, and by virtue of the large number of such indices the power of the original procedure can exceed that of the modified procedures.

Finally, to assess the effect of dependences of the test statistics across amino-acid positions

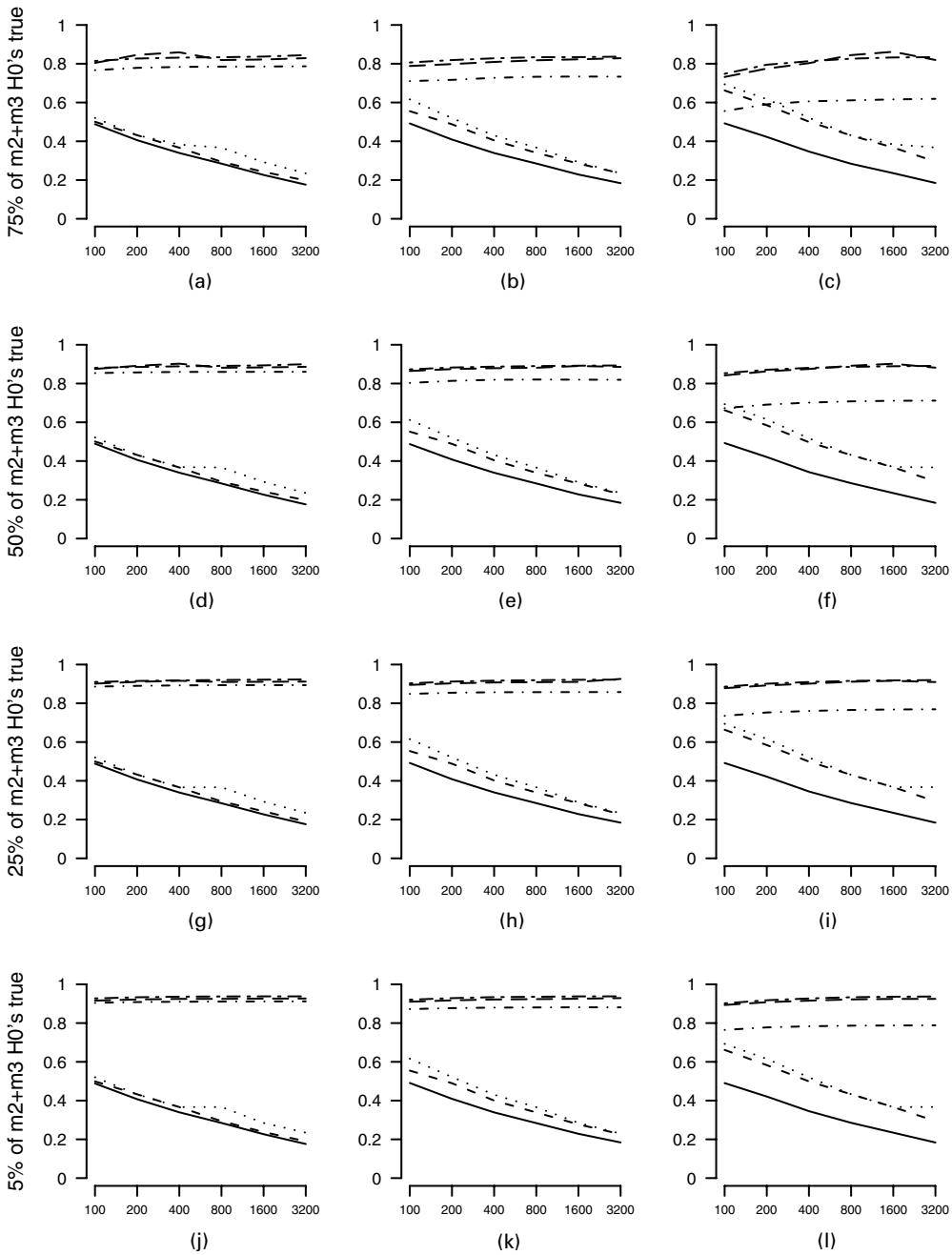


Fig. 3. Based on the simulation study, the figure shows the average power (the fraction of false null hypotheses rejected averaged over the 5000 simulations) for the Bonferroni method (———), the Tarone-modified Bonferroni method (– – –), the Tarone-modified Bonferroni method with adjustment (·····), Benjamini and Hochberg's (1995) FDR method (– · – ·), the Tarone-modified FDR method (— —) and the Tarone-modified FDR method with adjustment (- — -): (a) 60% of m H_0s true; (b) 37.5% of m H_0s true; (c) 15% of m H_0s true; (d) 40% of m H_0s true; (e) 25% of m H_0s true; (f) 10% of m H_0s true; (g) 20% of m H_0s true; (h) 13.5% of m H_0s true; (i) 5% of m H_0s true; (j) 4% of m H_0s true; (k) 2.5% of m H_0s true; (l) 1% of m H_0s true; (a), (d), (g), (j) $m_1/m = 0.2$; (b), (e), (h), (k) $m_1/m = 0.5$; (c), (f), (i), (l) $m_1/m = 0.8$

in the example data on the size of the procedures, simulations (with $\alpha = 0.05$) were conducted in which two sets of 73 sequences were randomly sampled with replacement from the pooled set of 146 HIV sequences, and the true FWERs and FDRs were estimated on the basis of 5000 resampled data sets. Resampling whole sequences for individuals preserves the dependence structure of the data. The FDRs of all three methods were bounded by 0.01, with the original FDR procedure most conservative, demonstrating that for the example the new procedures are improbably liberal.

6. Discussion

Benjamini and Hochberg's (1995) original FDR multiple-comparisons procedure has broad application, including to clinical trials, genomics and microarray experiments. This paper has developed a modified FDR procedure for discrete data, on the basis of Tarone's (1990) modification of the Bonferroni procedure. The new procedure is simple to apply and was shown in simulations and an example to improve the power substantially over Benjamini and Hochberg's FDR procedure for certain applications. The extent of the gain in power depends on the degree of variability in the outcome of interest, with greatest gains for highly conserved data sets (e.g. genetic sequence data sets) and negligible gains for data sets that have substantial variability in the outcome for all hypothesis tests. The improvement that is conferred by the new method is consistent across low dimensional and high dimensional problems, for which the number of tests is small or large compared with the sample size respectively. In general, FDR controlling procedures have their greatest advantage over FWER controlling procedures when the number of hypothesis tests is large; thus the modified FDR procedure that was developed here is expected to be particularly useful for the analysis of high dimensional data sets. For the example data set with 146 subjects and 118 tests, the new procedure identified 15 positions in the gag p24 sequence of HIV that are differentially polymorphic between subtypes C and B, and this information can be applied to weight HIV metrics that are used in sieve analyses of forthcoming HIV vaccine efficacy trials, and to inform the on-going construction of CTL epitope cocktail HIV vaccines.

In addition to the original FDR procedure, the modification that is described here can be applied to improve several recently developed FDR controlling procedures when the data are discrete. For example, the adaptive method of Benjamini and Hochberg (2000) proceeds as follows: an allowable FDR α is prespecified, the number of true hypotheses m_0 is estimated with an upwardly biased estimator $\hat{m}_0$ and Benjamini and Hochberg's (1995) procedure is conducted with m replaced by $\hat{m}_0$. Similarly, Storey (2002) proposed an upwardly biased estimator $\hat{\pi}_0$ for the probability that a null hypothesis is true and noted that, for independent test statistics, performing Benjamini and Hochberg's procedure with prespecified FDR $= \alpha/\hat{\pi}_0$ controls the FDR at level α. Both of these adaptive FDR procedures are substantially more powerful than the original FDR procedure when a large fraction of the m null hypotheses are false, and the Tarone modification for discrete data can be implemented directly as described in Section 4.1.

Development of FDR controlling multiple-comparisons techniques is an active area of research, and we expect that many of the newly developed procedures can be modified straightforwardly by using the idea that is described here, to improve their power when the data are discrete.

Acknowledgements

This work was supported by National Institutes of Health grants 1 RO1 AI054165-01 and 1 UO1 AI46703-01.

References

Abramovich, F. and Benjamini, Y. (1996) Adaptive thresholding of wavelet coefficients. *Comput. Statist. Data Anal.*, **22**, 351–361.

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

Benjamini, Y. and Hochberg, Y. (2000) On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Statist.*, **25**, 60–83.

Benjamini, Y., Hochberg, Y. and Kling, Y. (1997) False discovery rate control in multiple hypotheses testing using dependent test statistics. *Research Paper 97-1*. Department of Statistics and Operational Research, Tel Aviv University, Tel Aviv.

Benjamini, Y. and Liu, W. (1999) A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *J. Statist. Planng Inf.*, **82**, 163–170.

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

Berman, P. W., Gray, A. M., Wrin, T., Vennari, J. C., Eastman, D. J., Nakamura, G. R., Francis, D. P., Gorse, G. and Schwartz, D. H. (1997) Genetic and immunologic characterization of viruses infecting MN-rgp120-vaccinated volunteers. *J. Infect. Dis.*, **176**, 384–397.

De Groot, A. S., Jesdale, B., Martin, W., Saint Aubin, C., Sbai, H., Bosma, A., Lieberman, J., Skowron, G., Mansourati, F. and Mayer, K. H. (2003) Mapping cross-clade HIV-1 vaccine epitopes using a bioinformatics approach. *Vaccine*, **21**, 4486–4504.

De Groot, A. S., Jesdale, B. M., Szu, E. and Schafer, J. R. (1997) An interactive web site providing MHC ligand predictions: application to HIV research. *AIDS Res. Hum. Retrovir.*, **13**, 539–541.

De Groot, A. S., Sbai, H., Aubin, C. S., McMurray, J. and Martin, W. (2002) Immuno-informatics: mining genomes for vaccine components. *Immunol. Cell Biol.*, **80**, 255–269.

Efron, B., Tibshirani, R., Storey, J. and Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Statist. Ass.*, **96**, 1151–1160.

Gaschen, B., Taylor, J., Yusim, K., Foley, B., Gao, F., Lang, D., Novitsky, V., Haynes, B., Hahn, B. H., Bhattacharya, T. and Korber, B. (2002) Diversity considerations in HIV-1 vaccine selection. *Science*, **296**, 2354–2360.

Gilbert, P. (2000) Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. *Ann. Statist.*, **28**, 151–194.

Gilbert, P., Lele, S. and Vardi, Y. (1999) Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika*, **86**, 27–43.

Graham, B. S. (2002) Clinical trials of HIV vaccines. *A. Rev. Med.*, **53**, 207–221.

Hall, T. A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.*, **41**, 95–98.

Hochberg, Y. and Tamhane, A. C. (1987) *Multiple Comparison Procedures*. New York: Wiley.

Hsu, J. (1996) *Multiple Comparisons Procedures*. London: Chapman and Hall.

Kuiken, C., Foley, B., Hahn, B., Marx, P., McCutchan, F., Mellors, J., Wolinsky, S. and Korber, B. (eds) (2002) HIV sequence compendium 2001. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos.

Mallet, L., Mazoyer, B. and Martinot, J. L. (1998) Functional connectivity in depressive, obsessive-compulsive, and schizophrenic disorders: an explorative correlational analysis of regional cerebral metabolism. *Psychiatr. Res. Neurimgng*, **82**, 83–93.

Novitsky, V., Cao, H., Rybak, N., Gilbert, P., McLane, M. F., Gaolekwe, S., Peter, T., Thior, I., Ndung'u, T., Marlink, R., Lee, T. H. and Essex, M. (2002a) Magnitude and frequency of cytotoxic T-lymphocyte responses: identification of immunodominant regions of human immunodeficiency virus type 1 Subtype C. *J. Virol.*, **76**, 10155–10168.

Novitsky, V., Smith, U. R., Gilbert, P., McLane, M. F., Chigwedere, P., Williamson, C., Ndung'u, T., Klein, I., Chang, S. Y., Peter, T., Thior, I., Foley, B. T., Gaolekwe, S., Rybak, N., Gaseitsiwe, S., Vannberg, F., Marlink, R., Lee, T. H. and Essex, M. (2002b) Human Immunodeficiency Virus Type 1 Subtype C molecular phylogeny: consensus sequence for an AIDS vaccine design? *J. Virol.*, **76**, 5435–5451.

Papathanasopoulos, M. A., Hunt, G. M. and Tiernessen, C. T. (2003) Evolution and diversity of HIV-1 in Africa—a review. *Virus Genes*, **26**, 151–163.

Storey, J. D. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc.* B, **64**, 479–498.

Storey, J. D. and Tibshirani, R. (2001) Estimating false discovery rates under dependence, with applications to DNA microarrays. *Technical Report 2001-28*. Department of Statistics, Stanford University, Stanford.

Storey, J. D. and Tibshirani, R. (2003) SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In *The Analysis of Gene Expression Data: Methods and Software* (eds G. Parmigiani, E. S. Garrett, R. A. Irizarry and S. L. Zeger). New York: Springer.

Tarone, R. E. (1990) A modified Bonferroni method for discrete data. *Biometrics*, **46**, 515–522.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. and Higgins, D. G. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acid Res.*, **25**, 4876–4882.

Weller, J. I., Song, J. Z., Heyen, D. W., Lewin, H. A. and Ron, M. (1998) A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics*, **150**, 1699–1706.
Westfall, P. H. and Wolfinger, R. D. (1997) Multiple tests with discrete distributions. *Am. Statistn*, **51**, 3–8.
Westfall, P. H. and Young, S. S. (1993) *Resampling Based Multiple Testing*. New York: Wiley.
Williams, V. S. L., Jones, L. V. and Tukey, J. W. (1999) Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *J. Statist. Planng Inf.*, **24**, 42–69.
World Health Organization–UNAIDS (2001) Report from a meeting of the WHO-UNAIDS Vaccine Advisory Committee Geneva, 21-23 February 2000. Approaches to the development of broadly protective HIV vaccines: challenges posed by genetic, biological and antigenic variability of HIV-1. *AIDS*, **15**, W1–W25.
Yekutieli, D. and Benjamini, Y. (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Statist. Planng Inf.*, **82**, 171–196.