Genome Scanning Methods for Comparing Sequences Between Groups, with Application to HIV Vaccine Trials

Peter B. Gilbert,[1,*] Chunyuan Wu,[1] and David V. Jobes[2]

[1] Fred Hutchinson Cancer Research Center and Department of Biostatistics, University of Washington, Seattle, Washington, 98109, U.S.A.

[2] VaxGen, Inc., 1000 Marina Boulevard, Suite 200, Brisbane, California, 94005, U.S.A.

[*] *e-mail:* pgilbert@scharp.org

SUMMARY. Consider a placebo-controlled preventive HIV vaccine efficacy trial. An HIV amino acid sequence is measured from each volunteer who acquires HIV, and these sequences are multiply aligned together with the reference HIV sequence represented in the vaccine. We develop genome scanning methods to identify HIV positions at which the amino acids in sequences from infected vaccine recipients tend to be more divergent from the corresponding reference amino acid than the amino acids in sequences from infected placebo recipients. Such analysis can help guide the sequence design of antigens to include in HIV vaccines. We consider five test statistics for comparing amino acid divergences between groups, based on Euclidean, Mahalanobis, and Kullback-Leibler measures. Weights are incorporated to reflect biological information contained in different amino acid positions and substitutions. Position-wise p-values are obtained by approximating the null distribution of the test statistics either by a permutation procedure or by nonparametric estimation. Modified Bonferroni and false discovery rate procedures that exploit the discrete nature of the genetic data are used to infer statistically significant signature positions. The methods are examined in simulations and are applied to data from the first HIV vaccine efficacy trial.

KEY WORDS: False Discovery Rate; Genetics; High Dimensional Data; Human Immunodeficiency Virus; Kullback-Leibler; Mahalanobis; Multinomial; Sequence Analysis.

1. Introduction

The extensive genetic diversity of the human immunodeficiency virus (HIV) poses a formidable challenge to the development of an efficacious HIV vaccine (Graham, 2002). An HIV vaccine may prevent infections with viruses genetically similar to a virus represented in the vaccine, but fail against genetically dissimilar viruses. Data on the amino acid sequences of the viruses that infect participants in preventive HIV vaccine efficacy trials can be used to assess how the efficacy of the candidate vaccine depends on genetic mismatching of exposing viruses. "Sieve analysis" methods have been developed for this purpose, which are based on comparing the genetic distances (to the vaccine sequence) of the sequences of infected vaccine recipients to the genetic distances of the sequences of infected placebo recipients (Gilbert, Self, and Ashby, 1998). Previously developed sieve analysis methods considered "low dimensional" cases in which viruses are classified exhaustively by a small number of K genotypes (or phenotypes), or are ordered by a scalar continuous summary measure of distance from the virus represented in the vaccine (Gilbert, Lele, and Vardi, 1999). However, there are many thousands of distinct HIV genotypes as defined by amino acid sequence. Consequently, the problem of identifying sequence patterns that differentiate between the two sets of infecting viruses is a high dimensional data problem, in which the number of variables (sequence positions) exceeds the number of observations (infected subjects). In a typical HIV vaccine efficacy trial, 100-400 subjects are infected and hundreds or thousands of sequence positions are studied.

The dataset available from an efficacy trial that we consider is the aligned HIV

amino acid sequences sampled from infected vaccine and placebo recipients, with one sequence per subject. We develop techniques for "genome scanning", whereby a sliding window is used within which the amino acids in the two aligned sequence sets are compared to the amino acid at the corresponding position in the reference vaccine sequence, and the goal is to identify "signature positions" (see Figure 1). A signature position is a position at which vaccinee sequences exhibit significantly greater divergence from the reference amino acid than placebo sequences. Identifying a signature position may suggest that amino acid changes in that position were required in order for HIV to elude the vaccine-induced immune response and hence establish infection. For example, certain N-linked glycosylation positions in the glycoprotein 120 (gp120) region of HIV appear critically important for HIV to evade neutralization (Wei et al., 2003), and the vaccine may fail to protect against viruses with certain mutant amino acids in these positions, due to antibody epitope escape. Finding a signature position could imply the necessity to add multiple different antigens to the vaccine, with amino acid sequences that match contemporary circulating viral strains, in order for the vaccine to elicit broadly protective immune responses. Therefore the results of genome scanning analyses can guide the design of new vaccines.

Consideration of one of the most commonly used methods for studying HIV signature positions, VESPA (Korber and Myers, 1992; http:hiv-web.lanl.govcontenthiv-dbmainpage.html), demonstrates the need for new methodology. VESPA is purely descriptive- it compares the frequency of the most common amino acid at positions between two sequence sets, without weighting the particular amino acids involved, and without using a probabilistic framework to control error rates. Our approach to the scanning analysis divides into three parts:

1. For each position, construct a two-sample test statistic that compares amino acid

divergences or frequencies between the two groups;

2. Approximate the null distribution of the test statistics across the set of studied amino acid positions, and obtain position-specific p-values;

3. Apply a multiple testing adjustment procedure to the set of unadjusted p-values to infer the set of signature positions, controlling for a false positive rate.

For 1., various statistics for evaluating sequence distances have recently been proposed, based on standardized Euclidean and Kullback-Leibler discrepency (Wu, Hsieh, and Li, 2001), and Mahalanobis distance (Kowalski, Pagano, and DeGruttola, 2002). These metrics were developed in different contexts than genome scanning analysis, so that their relative utility for our application is unknown. Accordingly we develop and compare test statistics based on all four of these approaches, generalized to incorporate weight functions that can make amino acid distances more immunologically relevant and thus hopefully more predictive of vaccine efficacy.

The test statistics evaluate the null hypothesis that the amino acids in the two sets of sequences have equal distributions of distance to the reference amino acid. We also consider versions of the statistics that test whether the distribution of amino acid frequencies at the position is equal for the two sets of sequences, irrespective of a reference amino acid.

For 2., we consider two approaches to approximating the null distributions. The first is a standard permutation procedure that only uses information at individual positions. When the number of tests is large compared to the sample size, this position-specific approach can provide too-conservative inference (cf., Pan, 2002). This motivates our second approach, which, following Pan (2003), pools information across all positions and estimates the null distributions of the test statistics directly and nonparametrically.

Efron (2004) also pointed out that a large number of tests presents an opportunity to estimate the null distribution directly as a novel approach to coping with high dimensional data. We apply both the permutation and the nonparametric estimation approaches to obtain unadjusted p-values for each of the positions, which are then subjected to a multiple comparisons adjustment procedure.

For 3., we apply the modified Bonferroni method developed by Tarone (1990) to control the family-wise type I error rate (FWER) across positions, as well as a recent modification of the original false discovery rate (FDR) procedure that also exploits the discrete characteristics of the genetic data to increase power (Gilbert, 2005).

This article is organized as follows. Section 2 develops four new test statistics for identifying signature positions. Section 3 describes the procedures for obtaining p-values and summarizes the multiple comparisons techniques employed, and describes two slightly modified test statistics that are suitable for use with the nonparametric estimation method for deriving p-values. Section 4 compares the performance of the various methods in numerical studies, Section 5 presents an example from the first HIV vaccine efficacy trial, and Section 6 gives concluding remarks.

2. Genome Scanning Methods for Identifying Signature Positions

2.1 *Preliminaries*

The data available for genome scanning analysis are $n_1 + n_2 + 1$ aligned amino acid sequences, one from each infected trial participant (n_1 vaccine arm; n_2 placebo arm), plus a reference sequence represented in the vaccine construct, all of which are p amino acids long. The amino acids compose HIV proteins, and the analysis considers the set of positions that constitute the HIV proteins expressed by the tested vaccine. Current vaccine candidates express proteins spanning $p \sim 500 - 3500$ positions (HVTN, 2005).

For the ith position and the jth sequence in the kth group, $k = 1, 2$, we define

a vector of indicators to represent the 20 amino acids possible at position i, including the possibility of a gap which may have arisen in the alignment. Specifically, let $Y_{kj}(i) = (Y_{kj}(i,1), \cdots, Y_{kj}(i,21))'$, where $Y_{kj}(i,a)$ is 1 if amino acid a is at position i and 0 otherwise, $a = 1, \cdots, 20$ ($a = 1$ represents A, $a = 2$ represents C, and so on), and $a = 21$ represents a gap. Similarly define $Y_r(i) = (Y_r(i,1), \cdots, Y_r(i,21))'$ for the reference sequence , and let $r(i)$ denote the amino acid at position i in the reference sequence. $Y_{kj}(i)$ is a 21-nomial random variable with response probability vector $p_k(i) = (p_k(i,1), \cdots, p_k(i,21))'$. The MLE of $p_k(i)$ is $\widehat{p}_k(i) = (\bar{Y}_k(i,1), \cdots, \bar{Y}_k(i,21))'$, where $\bar{Y}_k(i,a) = n_k^{-1} \sum_{j=1}^{n_k} Y_{kj}(i,a)$.

The biological significance of a difference in two amino acids at a position depends on the particular amino acids being compared (e.g., T vs Y). There is a vast literature on how to weight the $20\times\ 19 = 380$ different amino acid mismatches, by physicochemical or evolutionary properties, and our methods incorporate a weight matrix to reflect such information. Specifically, let M be a 21×21 matrix with nonnegative entries, with $(a,a')^{\text{th}}$ element the weight/score summarizing dissimilarity of amino acids a and a'. For example, the nondiagonal entries of M could be taken to be those from a hydrophobicity/biochemical scoring matrix (George, Barker, and Hunt, 1990) or an HIV-specific point accepted mutation (PAM) matrix (Nickle et al., 2005). The distance between the amino acid at position i in the j^{th} sequence of group k to the amino acid at position i in the reference sequence, $r(i)$, is the appropriate element of M, computed as $d_{kj}(i) = Y_{kj}(i)' M Y_r(i)$. The simplest matrix $M = J - I$, with J the 21 by 21 matrix of ones and I the identity matrix; with this matrix $d_{kj}(i)$ is one if the two amino acids under comparison are different and is zero if they are the same.

2.2 Two-sample Test Statistics

For each position i, test statistics are developed to evaluate $H_0(i) : p_1(i) = p_2(i)$

versus $H_1(i) : p_1(i) \neq p_2(i)$. The choice of weight matrix M determines whether the procedures test for differential amino acid divergence from the reference amino acid or for differential amino acid frequencies, irrespective of any reference. Zeros on the diagonal of M yields tests of the former type, and $M = J$ of the latter type.

For position i, set

$$\widehat{v}^2(i,a) = M(a,r(i)) \left[\frac{(n_1-1)}{(n-2)} \widehat{Var}(\widehat{p}_{11}(i,a) + \frac{(n_2-1)}{(n-2)} \widehat{Var}(\widehat{p}_{21}(i,a)) \right] M(a,r(i)).$$

Let $p^*(i)$ be the number of nonzero components in $\widehat{v}^2(i,a)$, and define

$$Z_E(i) = C_E(i) \sum_{a=1}^{21} \frac{\left(M(a,r(i)) \left[\widehat{p}_1(i,a) - \widehat{p}_2(i,a)\right]\right)^2}{\{\widehat{v}(i,a) + \lambda_1\}^2} I(\widehat{v}(i,a) > 0), \tag{1}$$

where $C_E(i) = (n-2-p^*(i)+1)/(p^*(i)(n-2)) \times ((n_1-1)(n_2-1)/(n-2))$ and λ_1 is a nonnegative constant. Note that $M(a,r(i)) \left[\widehat{p}_1(i,a) - \widehat{p}_2(i,a)\right] = \bar{d}_1(i,a) - \bar{d}_2(i,a)$. The constant λ_1 is added to the denominator of $Z_E(i)$ to stabilize the statistics, which can be very large due to small $\widehat{v}(i,a)$. Efron et al. (2001), Tusher et al. (2001), and Guo et al. (2003) suggested adding a small positive constant to two-sample statistics in microarray applications, and Lönnstedt and Speed (2002) showed that the modified statistics perform better than the usual t-statistic. Following the approach of Tusher et al. (2001), we choose λ_1 to minimize the coefficient of variation of $\{Z_E(i) : i = 1, \cdots, p\}$. An alternative approach would pick λ_1 as the 90^{th} percentile of $\{\widehat{v}(i,a) : i = 1, \cdots, p; a = 1, \cdots, 21\}$ (Efron et al., 2001).

For the second test statistic, Mahalanobis' D^2 statistic for position i is given by

$$D^2(i) = (\widehat{p}_1(i) - \widehat{p}_2(i))' diag(MY_r(i)) \widehat{S}^-_{\lambda_2}(i) diag(MY_r(i)) (\widehat{p}_1(i) - \widehat{p}_2(i)),$$

where $\widehat{S}^-_{\lambda_2}(i)$ is the Moore-Penrose generalized inverse of $\widehat{S}_{\lambda_2}(i) \equiv \widehat{S}(i) + \lambda_2 diag(\underline{1}_{nz}(i))$, with $\widehat{S}(i) = [(n_1-1)\widehat{S}_1(i) + (n_2-1)\widehat{S}_2(i)]/(n-2)$ and λ_2 a nonnegative constant. Here $\widehat{S}_k(i) = \widehat{p}_k(i) I - \widehat{p}_k(i)\widehat{p}_k(i)'$ is the multinomial MLE of $S_k(i) = p_k(i) I - p_k(i) p_k(i)'$, and

$\underline{1}_{nz}(i)$ is the 21-vector of indicators of whether the ath row of $\widehat{S}(i)$ is the zero vector, $a = 1, \cdots, 21$. $S^{-}_{\lambda_2}(i)$ is obtained by computing the Moore-Penrose inverse of the submatrix of $\widehat{S}_{\lambda_2}(i)$ formed by removing the zero-vector rows and columns (corresponding to amino acids never present or always present at position i), and then expanding the inverse to a 21×21 matrix by re-inserting the zero-vector rows and columns. When $M = J$ (so that $diag(MY_r(i)) = I$) and $\lambda_2 = 0$, $D^2(i)$ is the Mahalanobis statistic that has been used extensively (cf., Rao and Chakraborty, 1991).

Let $p^*(i)$ be the rank of $\widehat{S}(i)$, and define the second test statistic $Z_M(i)$ by

$$Z_M(i) = \frac{(n - p^*(i) - 1)}{p^*(i) \times (n-2)} \frac{n_1 n_2}{n} D^2(i).$$

Similarly to $Z_E(i)$, the diagonal matrix $\lambda_2 I$ is added to $\widehat{S}(i)$ to stabilize $Z_M(i)$. The constant λ_2 is selected to minimize the coefficient of variation of the test statistic via the algorithm of Guo et al. (2003, page 1630). An advantage of $Z_M(i)$ compared to the Euclidean statistic $Z_E(i)$ is that it accounts for the correlation structure of the multinomial response vectors, which can increase statistical power, as verified in Section 4.

The third statistic is based on the Kullback-Leibler discrepency, which is relatively easy to compute. For position i, let $Z_{KL}(i) =$

$$\sum_{a=1}^{21} M(a, r(i)) \widehat{p}_1(i,a) \log \left\{ I(\widehat{p}_2(i,a) > 0) \frac{\widehat{p}_1(i,a)}{\widehat{p}_2(i,a)} + I(\widehat{p}_2(i,a) = 0) \frac{(\widehat{p}_1(i,a) + n_1^{-1})}{n_2} \right\}. \quad (2)$$

If in (2) $I(\widehat{p}_2(i,a) > 0)$ is replaced with unity and the second term with $I(\widehat{p}_2(i,a) = 0)$ is deleted, and $M = J$, then $Z_{KL}(i)$ is exactly the Kullback-Leibler discrepency between the 21-nomial empirical densities $\widehat{p}_1(i)$ and $\widehat{p}_2(i)$. The modification, which was suggested by Wu, Hsieh, and Li (2001) prevents $Z_{KL}(i)$ from taking infinite value.

3. Judging Statistical Significance

3.1 Permutation-based Unadjusted p-values

To judge statistical significance of the p tests, first nominal (unadjusted) position-wise p-values are computed. Since analytic p-values based on limiting distributions are not available for the test statistics (except Euclidean and Mahalanobis with $\lambda_1 = \lambda_2 = 0$), and parametric distributional assumptions may be unreliable given the underlying discreteness of the sequence data, we use a permutation procedure to determine p-values. Specifically, B data sets, each of $n = n_1 + n_2$ sequences, are generated by independently permuting the group membership labels of the whole sequences. The p-value for position i is calculated as the fraction of the test statistics computed using the B permuted datasets that equal or exceed the value of the original test statistic.

3.2 Nonparametric Estimated Null Distribution-based Unadjusted p-values

In the second (pooling) approach to computing position-specific p-values, assume that under all $H_0(i)$'s, the test statistics of interest $Z(i)$ have the same distribution for $i = 1, \cdots, p$. We follow Pan's (2003) clever idea for how to directly nonparametrically estimate the null distribution of hundreds of t-statistics. For each group of sequences separately, randomly permute the sequences into two (almost) equally-sized pieces, labeled sets $J_{k1}, J_{k2}, k = 1, 2$. Define $n_{k2} = n_{k1}$ if $n_k = 2n_{k1}$ and $n_{k2} = 2n_{k1} + 1$ otherwise, $k = 1, 2$. Let $w_1(i), i = 1, \cdots, p$ be a vector of nonnegative weights on the positions. Modify (slightly) the test statistic $Z_E(i)$ of (1) to $Z_E^{split}(i) = w_1(i)C_E(i)\times$

$$\sum_{a=1}^{21} \frac{\{M(a, r(i))\left[(\widehat{p}_{11}(i,a) + \widehat{p}_{12}(i,a))/2 - (\widehat{p}_{21}(i,a) + \widehat{p}_{22}(i,a))/2\right]\}^2}{\{\widehat{v}(i,a) + \lambda_1\}^2} I(\widehat{v}(i,a) > 0),$$

where $\widehat{p}_{k1}(i,a) = n_{k1}^{-1} \sum_{j=1}^{n_1} Y_{kj}(i,a) I(j \in J_{k1})$ averages the $Y_{kj}(\cdot)$ in the first permuted half of sample k and $\widehat{p}_{k2}(i,a)$ averages the $Y_{kj}(\cdot)$ in the second permuted half. The statistic $Z_E^{split}(i)$ approximately equals $Z_E(i)$, and motivates a statistic that estimates its null distribution: $z_E^{split}(i) = w_1(i)C_E(i)\times$

$$\sum_{a=1}^{21} \frac{\{M(a, r(i))\left[(\widehat{p}_{11}(i,a) - \widehat{p}_{12}(i,a))/2 + (\widehat{p}_{21}(i,a) - \widehat{p}_{22}(i,a))/2\right]\}^2}{\{\widehat{v}(i,a) + \lambda_1\}^2} I(\widehat{v}(i,a) > 0).$$

Because the numerator of $z_E^{split}(i)$ is the sum of within-sample differences, its mean is zero. Furthermore, the denominators of $Z_E^{split}(i)$ and $z_E^{split}(i)$ are the same, and thus $z_E^{split}(i)$ can be expected to approximate the null distribution of $Z_E^{split}(i)$.

To obtain p-values, once $Z_E^{split}(i)$ is computed, each group of sequences is again separately randomly permuted into two halves, and $z_E^{split}(i)$ is computed. Based on B separate permutations $z_E^{split(b)}(i)$ is computed B times, $b = 1, \cdots, B$. For position i the p-value is then $p_i = N_i/(B \times p)$, where N_i is the number of the test statistics $z_E^{split(b)}(i')$ that equal or exceed $Z_E^{split}(i)$, pooling over $i' = 1, \cdots, p$ and $b = 1, \cdots, B$.

We use a very similar approach to estimate the null distribution of a slightly modified version of $Z_M(i)$, $Z_M^{split}(i)$, defined as

$$Z_M^{split}(i) = w_1(i)\frac{(n - p^*(i) - 1)}{p^*(i) \times (n-2)}\frac{n_1 n_2}{n} D^{2split}(i), \tag{3}$$

$$\begin{aligned} D^{2split}(i) &= \left\{\frac{\widehat{p}_{11}(i) + \widehat{p}_{12}(i)}{2} - \frac{\widehat{p}_{21}(i) + \widehat{p}_{22}(i)}{2}\right\}' diag(MY_r(i))\widehat{S}^-_{\lambda_2}(i) \\ &\quad \times diag(MY_r(i))\left\{\frac{\widehat{p}_{11}(i) + \widehat{p}_{12}(i)}{2} - \frac{\widehat{p}_{21}(i) + \widehat{p}_{22}(i)}{2}\right\}. \end{aligned}$$

The null distribution of $Z_M^{split}(i)$ can be estimated via $z_M^{split}(i)$ defined as in (3) with $(\widehat{p}_{11}(i)+\widehat{p}_{12}(i))/2-(\widehat{p}_{21}(i)+\widehat{p}_{22}(i))/2$ replaced by $(\widehat{p}_{11}(i)-\widehat{p}_{12}(i))/2+(\widehat{p}_{21}(i)-\widehat{p}_{22}(i))/2$. Unadjusted p-values are then obtained in the same way as for $Z_E^{split}(i)$. Note that for $Z_E^{split}(i)$ and $Z_M^{split}(i)$, the inverse variances are not estimated separately for the two halves of the data, to increase the stability of the statistics.

An advantage of the "split" statistics $Z_E^{split}(i)$ and $Z_M^{split}(i)$ is the incorporation of a position-specific weight $w_1(i)$ that can be used to reflect biological information. These weights affect the p-values because the pooling method is used; weights placed in front of the non-split statistics described in Section 2 would not affect the p-values, because they are computed marginally (position-specific).

3.3 Multiple Hypothesis Testing Adjustment

Given the set of p-values, we consider four multiple comparisons adjustment procedures to determine the set of significant signature positions: standard Bonferroni, Tarone's (1990) modified Bonferroni method for discrete data, standard FDR (Benjamini and Hochberg, 1995), and Tarone-modified FDR for discrete data, which we refer to as "Tarone FDR" (Gilbert, 2005). We sketch the latter procedure in two steps: First, compute the integer K and the subset of indices R_K among the p hypotheses as described by Tarone (1990); Second, perform Benjamini and Hochberg's (1995) FDR procedure at level α on the subset of hypotheses R_K. To define K and R_K, for each $k = 1, \cdots, p$, let $m(k)$ be the number of the p positions for which $\alpha_i^* < \alpha/k$, where α_i^* is the minimum achievable significance level for the test for the ith position, computed based on data pooled over the two groups. Then K is the smallest value of k such that $m(k) \leq k$, and R_K is the set of indices satisyfing $\alpha_l^* < \alpha/K$. Because K and R_K are calculated based only on information pooled over the vaccine and placebo groups, this procedure controls the FDR at level α. Due to the complexity of computing the α_i^* for each of the newly proposed test statistics, for the Simulations and Example the α_i^* were computed based on Fisher's exact test.

4. Simulation Study

4.1 Background

The simulation study is designed based on data from the first HIV vaccine efficacy trial (Flynn et al., 2005). Healthy HIV uninfected volunteers were randomized to receive vaccine ($N_v = 3598$) or placebo ($N_p = 1805$) and were tested for HIV infection every 6 months for 36 months. The vaccine was a recombinant envelope gp120 subunit vaccine, and was designed to prevent acquisition of HIV by inducing antibodies that could bind to neutralizing epitopes on HIV gp120 and prevent entry of HIV into host

cells. The vaccine did not prevent HIV infection, with a similar rate of infection in the vaccine (241/3598 = 6.7%) and placebo (127/1805 = 7.0%) arms. For 336 of the 368 infected participants three HIV isolates were sampled at the time of HIV infection detection, and the amino acid sequence of gp120 was determined by direct translation of the DNA sequence for each isolate. Sequences from the same individual were highly similar, and with little loss of information we considered one randomly selected sequence from each subject. The 336 gp120 sequences were aligned together with the two gp120 sequences that were represented in the vaccine construct, named MN and GNE8. Since GNE8 was sampled more recently and was closer genetically to the infecting sequences, it was used as the reference sequence in all analyses. There are $n_1 = 217$ vaccine group sequences and $n_2 = 119$ placebo group sequences, each of length $p = 581$.

For each of the five testing procedures developed above, plus Fisher's exact test for comparison, simulations were carried out to address the following questions: 1) What is the impact of the proportion of positions with a true alternative hypothesis on the performance of the procedures? 2) How much power is there to detect signature positions for vaccine efficacy trials of different sizes? 3) What is the impact of whether the alternative hypotheses are true in conserved or variable positions? 4) How do the position weights $w_1(i)$ influence size and power of the split test statistics? 5) What is the impact of the constants λ_1 and λ_2 in the performance of the procedures? To address these questions, gp120 sequences for the infected placebo group were simulated by randomly sampling with replacement $n_2 = 90$ or 180 whole sequences from the 336 sequences. These sample sizes represent a small and large efficacy trial. Assuming an overall vaccine efficacy of 50%, sequences for the infected vaccine group were generated by sampling with replacement $n_1 = 45$ or 90 whole sequences from the 336 sequences.

To create the alternative hypothesis at a given position i, we used the HIV-specific PAM matrix developed by Nickle et al. (2005) to induce stochastic evolution of the amino acids at i in the vaccinee sequences. Each nondiagonal entry of the PAM matrix corresponds to two different amino acids, and specifies the probability that either of the amino acids mutates into the other one during a certain amount of evolutionary time. We used the PAM-25 matrix, which specifies a total of 25 amino acid interchanges per 100 positions. Thus, at alternative hypothesis positions, on average 25% of the vaccinee sequences have mutations.

Question 1) was addressed by setting 1%, 10% or 25% of the positions to have true alternatives, which amounts to 6, 58, or 145 of the 581 positions. We selected the positions based on previous studies supporting that 39 of the 581 positions are important for HIV neutralization or CD4 co-receptor binding. Specifically, Wyatt et al. (1998) identified the CD4-binding positions 88, 113, 117, 256, 257, 262, 266, 368, 370, 384, 421, 427, 457, 470, 474, 475, 477, 482, 483, 484, the CD4-induced epitope positions 88, 117, 121, 207, 256, 257, 262, 370, 381, 382, 419, 420, 421, 422, 423, 427, 435, 438, 475, and positions 295, 297, 334, 386, 392, 397, which constitute a neutralization epitope defined by the monoclonal antibody 2G12. The positions, here and in the Example, are numbered using the standard HXB2 strain numbering system (Kuiken et al., 2002). In addition, Wei et al. (2003) identified three positions at which amino acid changes can sterically inhibit the accessibility of principal neutralizing epitopes on the virus surface: 245, 274, 309. These comprise 39 unique positions. For the 6 alternative positions, we selected the positions constituting the monoclonal antibody 2G12 neutralization epitope (295, 297, 334, 386, 392, 397); for the 58 alternative positions we selected the 39 key positions listed above plus 19 randomly sampled positions; and for the 145 alternative positions we used these 58 positions plus 87 more randomly sampled positions.

Question 2) was addressed by repeating the simulation experiment for the two sample sizes. Question 3) was addressed by evaluating the power of the testing procedures separately for diverse positions (frequency of modal amino acid < 0.95) and conserved positions (frequency of modal amino acid ≥ 0.95). Question 4) was addressed by running simulations with $w_1(i) = I(H_0(i) \text{ true}) + cI(H_0(i) \text{ false})$ with c set as 2.0 or 0.5, which assess size and power of the split test statistics when the false null hypotheses are upweighted 2-fold (correctly incorporating prior knowledge) or downweighted 2-fold (incorrectly incorporating prior knowledge), respectively. Question 5) was addressed by repeating the simulations for the Euclidean- and Mahalanobis-based tests with λ_1 and λ_2 set to zero.

Amino acid substitutions were weighted equally, by setting $M = J - I$. Except for results reported at the end of Section 4.3, positions in the split statistics were weighted equally ($w_1(i) = 1$). Tests were carried out at 2-sided level $\alpha = 0.05$, using 10,000 permutations to approximate p-values. Empirical false positive rates, false discovery rates, and powers of the testing procedures were computed.

4.3 Simulation Results

Figure 2 shows the estimated sizes and FDRs using the Tarone Bonferroni and Tarone FDR multiple testing adjustment methods, respectively. The tests based on $Z_E(i), Z_M(i), Z_E^{split}(i), Z_M^{split}(i)$ used $\lambda_1 = \lambda_2 = 0$, due to their superior performance as described below. All of the test procedures are conservative under family-wise error (FWER) adjustment. Under FDR adjustment, when 10% or 25% of the null hypotheses are false, all of the procedures except the two Mahalanobis-based tests control the FDR below 0.05 within 2 Monte Carlo standard errors. When only 1% of the null hypotheses are false, all of the proposed procedures have estimated FDRs higher than 0.05. This occurs because in many simulation runs the null hypothesis was only rejected at 1 or

2 positions, in which case a single false rejection makes the FDR very high (e.g., 0.5 for 2 total rejections). This suggests that the FDR method should not be used unless the Tarone Bonferroni method rejects at least one null hypothesis. The tests based on $Z_E(i), Z_M(i), Z_E^{split}(i), Z_M^{split}(i)$ with $\lambda_1 > 0$ and $\lambda_2 > 0$ had very similar FWER and FDR rates.

Figure 3 shows the estimated powers of the procedures (again with $\lambda_1 = \lambda_2 = 0$). We make several observations. First, the Kullback-Leibler and standardized Euclidean statistics are consistently most powerful. Fisher's exact statistic is third most powerful, with $Z_E^{split}(i)$ providing similar power at the larger sample size under FDR adjustment. Second, the statistics that use pooling are generally less powerful the non-pooled counterparts, moreso for lower fractions of false null hypotheses. It appears that the pooling methods perform best when many alternative hypotheses are true (Pan (2003) found greatest advantage of the pooling approach for this setting).

Third, the tests based on $Z_M(i)$ and $Z_M^{split}(i)$ consistently have the lowest power. To explain the poor performance of the Mahalanobis-based statistics, which have inflated false positive rates as well as low power, note that the estimated covariance matrix $\widehat{S}(i)$ is often fairly high-dimensional (e.g., 10×10), which occurs because the gp120 region is highly variable. Consequently there are dozens of covariance terms to estimate, but the sample size is quite limited for doing so. Therefore, we conjecture that the noise in covariance estimation is causing the poor performance. To support this conjecture, we repeated the simulations with all correlation estimates set to zero, in which case the Mahalanobis-based test statistics are very similar to the Euclidean-based statistics. As expected, with this modification these two approaches performed similarly.

Fourth, the tests based on $Z_E(i)$ and $Z_E^{split}(i)$ with $\lambda_1 = 0$ were consistently more powerful than the corresponding tests with $\lambda_1 > 0$ (not shown in figures). For example,

for the smaller sample size and FWER adjustment, power of the $Z_E(i)$ tests with $\lambda_1 = 0$ was 0.45, 0.73, and 0.63 for 1%, 10%, and 25% of $H_0(i)$'s false, respectively, compared to 0.20, 0.59, and 0.51 for the $Z_E(i)$ tests with $\lambda_1 > 0$. Similarly, for the larger sample size and FWER adjustment, power of the $Z_E^{split}(i)$ tests with $\lambda_1 = 0$ was 0.12, 0.42, and 0.27 for 1%, 10%, and 25% of $H_0(i)$'s false, respectively, compared to 0.01, 0.27, and 0.18 for the $Z_E^{split}(i)$ tests with $\lambda_1 > 0$. This result can be explained by the fact that the sum $\sum_{a=1}^{21}$ in $Z_E(i)$ is restricted to contrasts $\widehat{p}_1(i,a) - \widehat{p}_2(i,a)$ for which at least one of $\widehat{Var}(\widehat{p}_{11}(i,a))$ or $\widehat{Var}(\widehat{p}_{21}(i,a))$ is positive, which prevents the denominator from being near 0. A similar explanation attains for $Z_E^{split}(i)$.

Fifth, when 1% of null hypotheses are false, the Kullback-Leibler and standardized Euclidean statistics have much greater power than the other methods. Because only a small number of signature positions are expected in many applications including vaccine trials, these methods are recommended for sparse-signal problems. Fifth, the results for conserved and diverse positions showed that for 10% and 25% true alternatives, power was greater for conserved positions, presumably because there is less background sequence variability (noise). For 1% true alternatives the methods had zero power for conserved positions, due to insufficient sequence variability, whereas the power of the tests for diverse positions was slightly greater than for all positions.

Sixth, the split statistics with true alternative positions upweighted had lower false positive/discovery rates and greater power than the equal-weighted methods; for example with $m_1/m_2 = 45/90$ and 10% of the alternative hypotheses true, the FDR of the $Z_E^{split}(i)$ tests with $\lambda_1 = 0$ was 0.0, compared to 0.032 for the unweighted tests, and power was 0.26/0.64 under Tarone Bonferroni/FDR compared to 0.23/0.56 for the unweighted tests. On the other hand when the true alternative positions were downweighted, the opposite results attained: the $Z_E^{split}(i)$ tests with $\lambda_1 = 0$ gave inflated

FDR = 0.078 and low power 0.10/0.21. These results provide "proof-of-principle" that correct upweighting of positions can improve size and power of the tests based on $Z_E^{split}(i)$, but incorrect weighting can seriously erode performance. This suggests that weighting to incorporate biological knowledge should be done with caution.

5. Example

The matrix M was taken as $J-I$, or as the HIV-specific PAM-250 matrix of Nickle et al. (2005), modified to have zeros on the diagonal and a vector of zeros added to the 21st row and 21st column. This matrix was computed based on thousands of observed among-patient mutations in HIV sequences. Because the previously available amino acid substitution matrices were built using organisms other than HIV, this HIV-specific PAM may yield more accurate estimates of evolutionary distances, as supported with some real data examples in Nickle et al. (2005). Weighting amino acid mismatches by their relative probability of interchange may enhance the biological relevance of the genome scanning analysis.

With $w_1(\cdot) = 1$ and $M = J - I$, Figure 4 shows the -$\log_{10}$-transformed unadjusted p-values based on the six test statistics (setting $\lambda_1 = \lambda_2 = 0$) for the 350 informative positions (i.e., those with enough diversity to possibly reject $H_0(i)$ using the Tarone Bonferroni procedure), and Figure 5 shows histograms of the test statistics. As indicated by the horizontal lines in Figure 4, after multiplicity adjustment the null hypothesis was not rejected for any positions by any of the statistics. The same result was obtained when M was set as the modified PAM matrix. The analysis was repeated for the subset of the 39 biologically-key positions described in the Simulations. Again there were no significant signature positions. The lack of any signatures can be explained by the apparent inability of the tested vaccine to prevent HIV infection. If the vaccine has no effect on susceptibility to acquiring HIV, then the distribution of infecting HIV

sequences should be identical in the infected vaccine and infected placebo groups.

Because there was a suggestion of possible partial vaccine efficacy in non-white participants (infection rates 5.0% and 9.4% in the vaccine and placebo groups, respectively, unadjusted $p = 0.012$) and in participants with high self-reported risk behavior at baseline (infection rates 20.3% and 29.2%, unadjusted $p = 0.032$) (Flynn et al., 2005), the scanning analyses were repeated in these subgroups. Again no significant signatures were found, which supports that the apparent efficacy in these subgroups may not have been real.

To illustrate an application with significant signature positions, 251 gp160 subtype B HIV-1 sequences were downloaded from the Los Alamos HIV Sequence Database (Kuiken et al., 2002), 61 known to be CXCR4 co-receptor utilizing viruses and 192 known to be CCR5 co-receptor utilizing viruses. The sequences were multiply aligned, with common length $p = 857$ amino acid positions. The procedures were applied to the data with $M = J$, to test for positions with different amino acid frequency distributions. Many significant signatures were found by all of the procedures; for example at level $\alpha = 0.05$ the Kullback-Leibler test yielded 13 significant signature positions under Tarone Bonferroni adjustment and 16 under Tarone FDR adjustment, and the pooled-Mahalanobis test yielded 9 and 17 significant signature positions, respectively. In comparison Fisher's exact test provided 5 and 13 significant signature positions, demonstrating greater sensitivity of the new testing procedures.

6. Discussion

We developed and evaluated five new testing procedures for detecting signature positions that distinguish two groups of amino acid sequences. The Kullback-Leibler and standardized Euclidean test statistic (with constant λ_1 in the denominator set to 0) were most powerful and are recommended. The efficiency of the Kullback-Leibler

discrepency likely derives from the fact that it is an expected log likelihood ratio, and has optimality properties closely related to those of likelihood ratio tests, as has been widely studied (Eguchi and Copas, 2002, provide a particularly clear discussion). A related standardized Euclidean statistic was also found to perform well by Wu et al. (2001), and in our setting we conjecture that it provided greater power than the Mahalanobis-based test because it standardized only by the variance estimates, thereby avoiding the noise introduced by estimating the entire covariance matric nonparametrically. Further research into the Mahalanobis-based tests is of interest, for example by using shrinkage covariance matrix estimates (Ledoit and Wolf, 2004).

An advantage of the methods developed here is that the procedures for obtaining position-specific unadjusted p-values and for determining significance of this set of p-values are completely separate; therefore any valid p-value-based multiple testing procedure can be used with any of the test statistics.

Acknowledgements

This research was supported by NIH grant 1 U01 A1054165-01.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate- a new and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57**:289-300.

Efron, B., Tibshirani, R., Storey, J.D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**:1151-1160.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association* **99**:96-104.

Eguchi, S., Copas, J. (2002). Interpreting Kullback-Leibler Divergence with the Neyman-Pearson Lemma. Preprint, available at www.ism.ac.jp/ eguchi/pdf/KLNP.pdf.

George, D.G., Barker, W.C., Hunt, L.T. (1990). Mutation data matrix and its uses. *Methods Enzymol* **183**:333-351.

Gilbert, P.B. (2005). A modified false discovery rate multiple comparisons procedure for discrete data, applied to HIV genetics. *Applied Statistics* **54**:143-158.

Gilbert, P.B., Self, S., Ashby, M. (1998). Statistical methods for assessing differential vaccine protection against human immunodeficiency virus types. *Biometrics* **54**:799-814.

Graham, B.S. (2002). Clinical trials of HIV vaccines. *Annual Review of Medicine* **53**:207-221.

Guo, X., Qi, H., Verfaillie, C.M., and Pan, W. (2003). Statistical significance analysis of longitudinal gene expression data. *Bioinformatics* **19**:1628-1635.

HVTN (2005). The Pipeline Project. Available at: http:www.hvtn.orgscience.

Korber, B. and Myers, G. (1992). Signature pattern analysis: A method for assessing viral sequence relatedness. *AIDS Research and Human Retroviruses* **8**:1549-1560.

Kowalski, J., Pagano, M. and DeGruttola, V. (2002). A nonparametric test of gene region heterogeneity associated with phenotype. *Journal of the American Statistical Association* **97**:398-408.

Kuiken, C., Foley, B., Hahn, B., Marx, P., McCutchan, F., Mellors, J., Wolinsky, S. and

Korber, B. (ed.) (2002). HIV Sequence Compendium 2001. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.

Ledoit, O., Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* **88**:365-411.

Lönnstedt, I. and Speed, T. (2002). Replicated microarray data. *Statistica Sinica* **12**:31-46.

Nickle, D.C., Heath, L., Jensen, M.A., Gilbert, P.B., Kosakovsky Pond, S.L.K., Mullins, J.I. (2005). Amino acid substitution matrices for HIV-1 subtype B. *Journal of Virology*, under revision.

Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* **12**:546-554.

Pan, W. (2003). On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. *Bioinformatics* **19**:1333-1340.

Rao, C.R., Chakraborty, R., Eds. (1991). Handbook of statistics. Volume 8: Statistical methods in biological and medical sciences. Elsevier/North-Holland, New York, New York.

Flynn, N.M., Forthal, D.N., Harro, C.D., Mayer, K.H., Para, M.F.; The rgp120 HIV Vaccine Study Group (2005). Placebo-controlled trial of a recombinant glycoprotein 120 vaccine to prevent HIV infection. *Journal of Infectious Diseases* **191**:654-665.

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

46:515-522.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences USA* **98**:5116-5121.

Wei, X., Decker, J.M., Wang, S., Hui, H., Kappes, J.C., Wu, X., Salazar-Gonzalez, J.F., Salazar, M.G., Kilby, J.M., Saag, M.S., Komarova, N.L., Nowak, M.A., Hahn, B.H., Kwong, P.D. and Shaw, G.M. (2003). Antibody neutralization and escape by HIV-1. *Nature* **422**:307-312.

Wu, T.-J., Hsieh, Y.-C. and Li, L.-A. (2001). Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. *Biometrics* **57**:441-448.

Wyatt, R., Kwong, P.D., Desjardins, E., Sweet, R.W., Robinson, J., Hendrickson, W.A., Sodroski, J.G. (1998). The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature* **393**:705-711.

Figure Legends

Figure 1. Illustration of sequence data available for genome scanning analysis, from 6 randomly selected infected vaccine and placebo recipients in the VaxGen trial, aligned together with the reference HIV vaccine sequence GNE8. The V3 loop region within gp120 is shown, which consists of positions 297-328 of gp160 using the HXB2 strain numbering system. The letters denote amino acids.

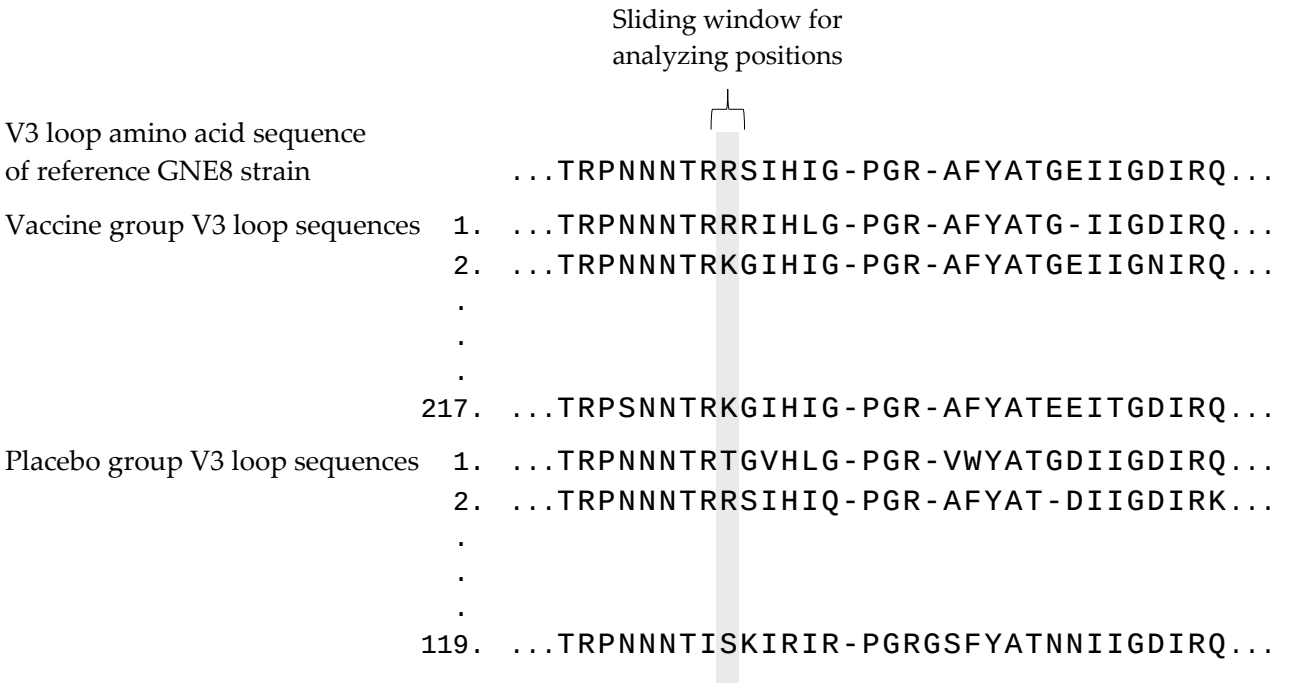
Figure 2. Based on the simulation study, panels (a) and (c) show average false positive rates for the five testing procedures (with $\lambda_1 = \lambda_2 = 0$) using the Tarone Bonferroni multiple testing adjustment procedure, for $n_1/n_2 = 45/90$ and $n_1/n_2 = 90/180$, respectively. Panels (b) and (d) show the corresponding average false discovery rates for the Tarone FDR multiple testing adjustment procedure.

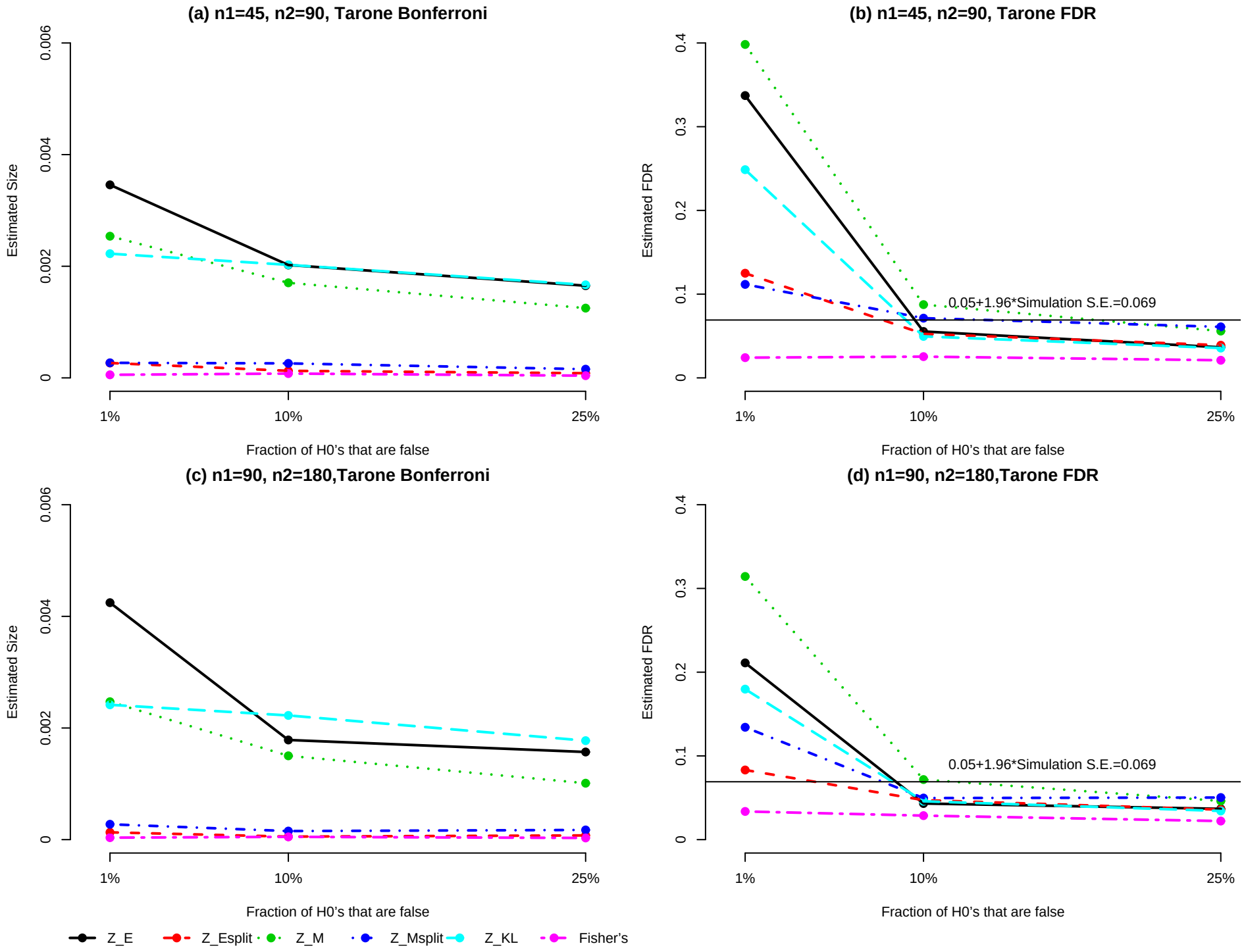
Figure 3. Based on the simulation study, panels (a) and (c) show average true positive rates (powers) for the five testing procedures (with $\lambda_1 = \lambda_2 = 0$) using the Tarone Bonferroni multiple testing adjustment procedure, for $n_1/n_2 = 45/90$ and $n_1/n_2 = 90/180$, respectively. Panels (b) and (d) show the corresponding estimated powers for the Tarone FDR multiple testing adjustment procedure.

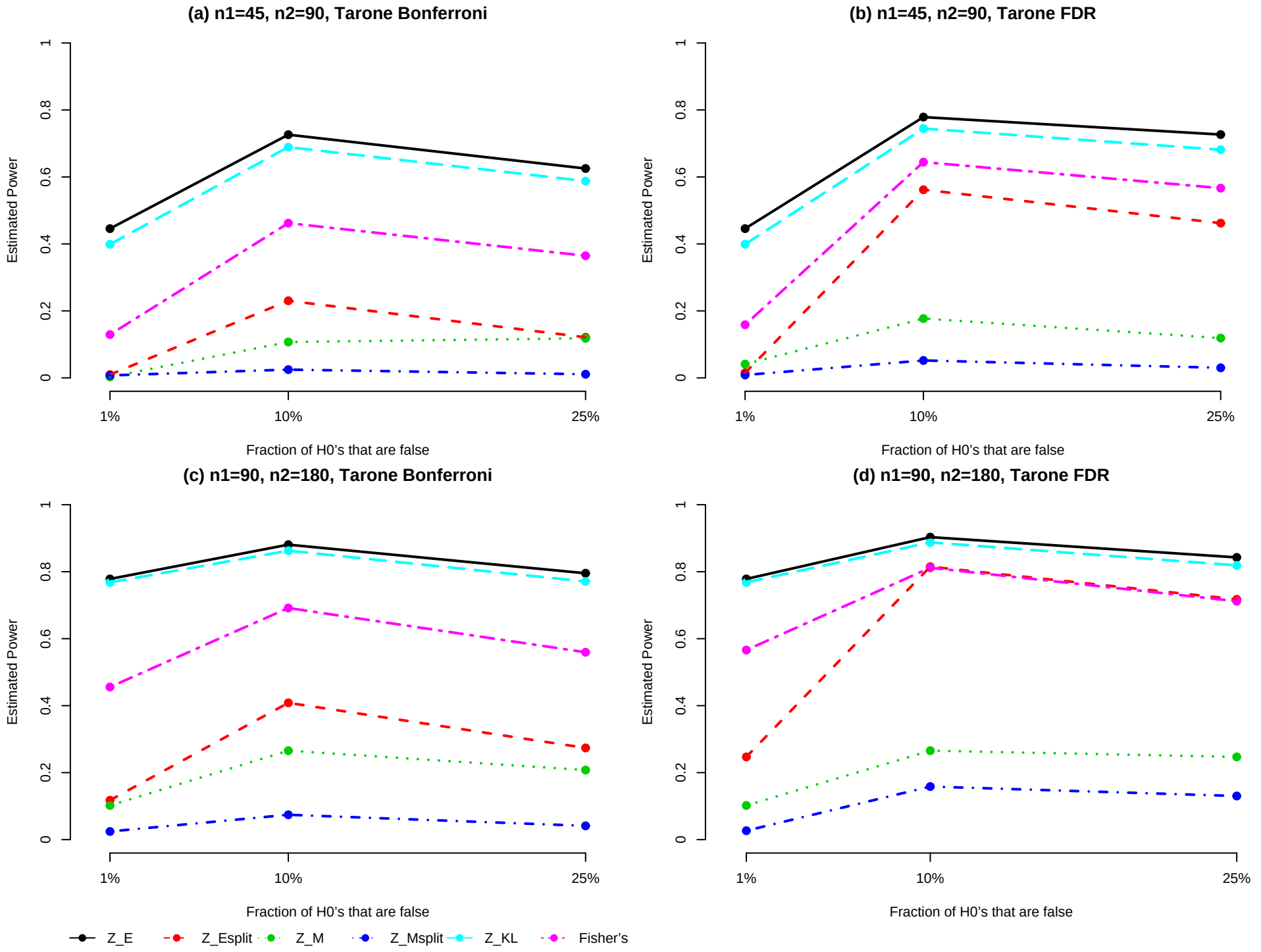
Figure 4. -$\log_{10}$ unadjusted p-values from the five testing procedures (with $\lambda_1 = \lambda_2 = 0$), for the 350 informative positions among the $p = 581$ positions analyzed in the VaxGen trial, with equal weighting of all amino acid substitutions and all positions. The horizontal lines represent cut-off levels of significance after adjustment for multiple testing using four different multiple testing adjustment procedures.

Figure 5. Histograms of the log-transformed test statistics for the 350 informative

positions (positions with some amino acid diversity) among the $p = 581$ positions analyzed in the VaxGen trial, with equal weighting of all amino acid substitutions and all positions.

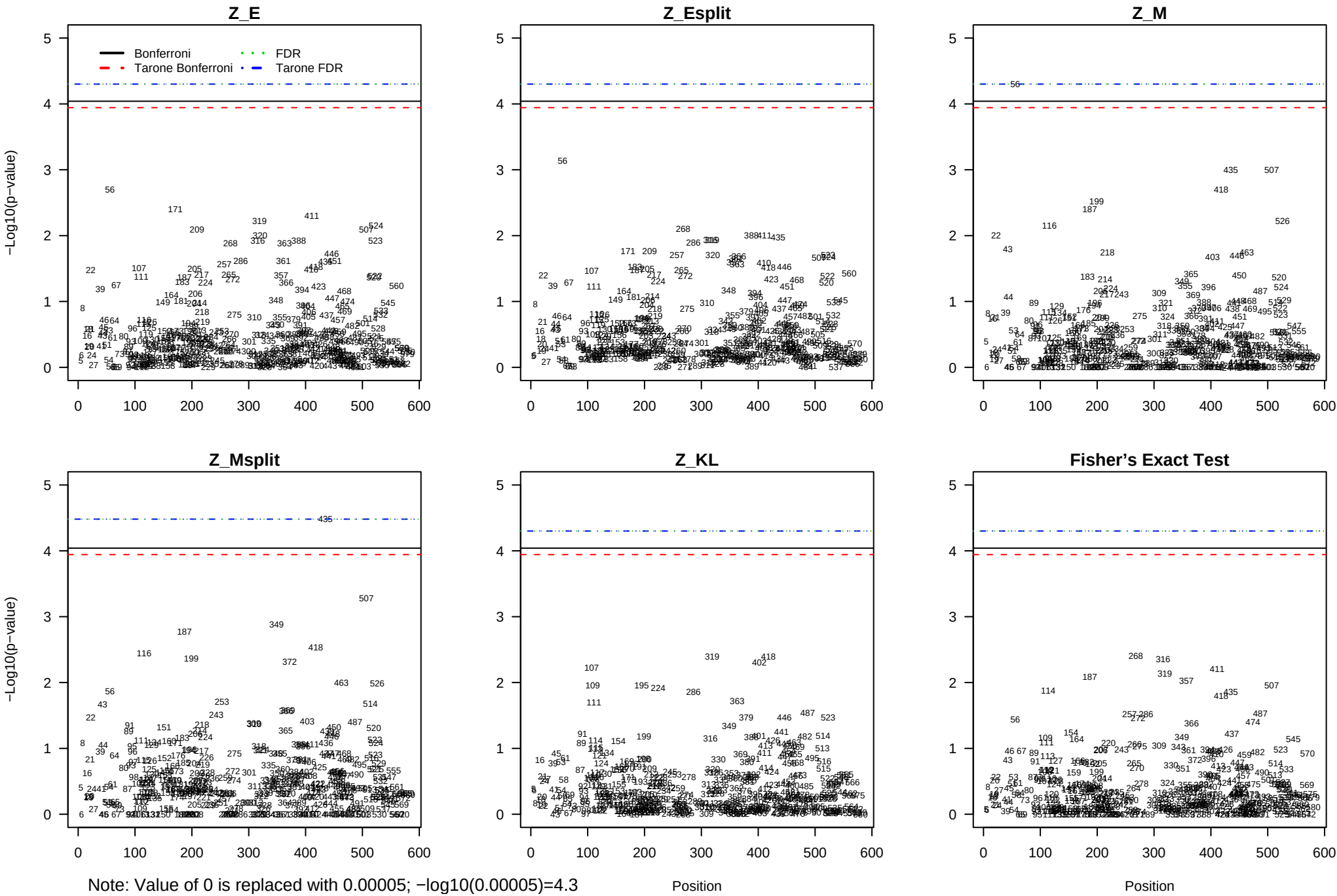

Sliding window for analyzing positions
V3 loop amino acid sequence of reference GNE8 strain
...TRPNNNTRRSIHIG-PGR-AFYATGEIIGDIRQ...
Vaccine group V3 loop sequences
1. ...TRPNNNTRRRIHLG-PGR-AFYATG-IIGDIRQ...
2. ...TRPNNNTRKGIHIG-PGR-AFYATGEIIGNIRQ...
217. ...TRPSNNTRKGIHIG-PGR-AFYATEEITGDIRQ...
Placebo group V3 loop sequences
1. ...TRPNNNTRTGVHLG-PGR-VWYATGDIIGDIRQ...
2. ...TRPNNNTRRSIHIQ-PGR-AFYAT-DIIGDIRK...
119. ...TRPNNNTISKIRIR-PGRGSFYATNNIIGDIRQ...


(a) n1=45, n2=90, Tarone Bonferroni
Estimated Size
0
0.002
0.004
0.006
1%
10%
25%
Fraction of H0's that are false
(b) n1=45, n2=90, Tarone FDR
Estimated FDR
0
0.1
0.2
0.3
0.4
0.05+1.96*Simulation S.E.=0.069
1%
10%
25%
Fraction of H0's that are false
(c) n1=90, n2=180,Tarone Bonferroni
Estimated Size
0
0.002
0.004
0.006
1%
10%
25%
Fraction of H0's that are false
(d) n1=90, n2=180,Tarone FDR
Estimated FDR
0
0.1
0.2
0.3
0.4
0.05+1.96*Simulation S.E.=0.069
1%
10%
25%
Fraction of H0's that are false
Z_E
Z_Esplit
Z_M
Z_Msplit
Z_KL
Fisher's


(a) n1=45, n2=90, Tarone Bonferroni
(b) n1=45, n2=90, Tarone FDR
(c) n1=90, n2=180, Tarone Bonferroni
(d) n1=90, n2=180, Tarone FDR
Estimated Power
0
0.2
0.4
0.6
0.8
1
1%
10%
25%
Fraction of H0's that are false
Z_E
Z_Esplit
Z_M
Z_Msplit
Z_KL
Fisher's

−log10 p−value for comparing Vaccine and Placebo VaxGen gp120 sequences

Equal Weight Matrix



Note: Value of 0 is replaced with 0.00005; −log10(0.00005)=4.3

Histograms of Test Statistics

VAXGEN gp120 vaccine and placebo recipient data set, with Equal Weight Matrix