

# Genome Scanning Tests for Comparing Amino Acid Sequences Between Groups

Peter B. Gilbert,<sup>1,\*</sup> Chunyuan Wu,<sup>1</sup> and David V. Jobes<sup>2</sup>

<sup>1</sup> Fred Hutchinson Cancer Research Center and Department of Biostatistics,  
University of Washington, Seattle, Washington, 98109, U.S.A.

<sup>2</sup> VaxGen, Inc., 1000 Marina Boulevard, Suite 200, Brisbane, California, 94005,  
U.S.A.

\**e-mail*: pgilbert@ssharp.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 aligned together with the reference HIV sequence represented in the vaccine. We develop genome scanning methods to identify positions at which the amino acids in infected vaccine recipient sequences either (A) are more divergent from the reference amino acid than the amino acids in infected placebo recipient sequences; or (B) have a different frequency distribution than the placebo sequences, irrespective of a reference amino acid. We consider t-test-type statistics for problem A and Euclidean, Mahalanobis, and Kullback-Leibler-type statistics for problem B. The test statistics incorporate weights to reflect biological information contained in different amino acid positions and mismatches. Position-specific  $p$ -values are obtained by approximating the null distribution of the statistics either by a permutation procedure or by nonparametric estimation. A permutation method is used to estimate a cut-off  $p$ -value to control the per comparison error rate at a pre-specified level. The methods are examined in simulations and are applied to two HIV examples. The methods for problem B address the general problem of comparing discrete frequency distributions between groups in a high-dimensional data setting.

KEY WORDS: Genetics; High Dimensional Data; Hypothesis Testing; Kullback-Leibler; Mahalanobis; Multinomial; Sequence Analysis; Signature Position; Vaccine Trial.

## 1. Introduction

The extensive genetic diversity of the human immunodeficiency virus (HIV) poses a formidable challenge to the development of an efficacious preventive HIV vaccine (HVTN, 2006). 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/phenotypes, or are ordered by  $K$  scalar summary measures of distance. However, there are many thousands of distinct HIV genotypes as defined by amino acid sequence. Consequently, the problem of identifying sequence positions that distinguish 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 efficacy trial, 100-400 subjects are infected and 500-3300 sequence positions are studied.

The data set 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 at each position, the amino acids in the two aligned sequence sets are compared to the amino acid at the corresponding position in the reference sequence, and the goal is to identify “signature positions” (see Figure 1). A signature position is a position at which vaccine 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 (gp120 is composed of a protein and a carbohydrate and is exposed on the surface of the HIV envelope), 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. Finding a signature position could imply the necessity to insert multiple different HIV strains into the vaccine, with amino acid sequences that match contemporary circulating HIV strains, in order for the vaccine to protect broadly. Therefore the results of genome scanning analyses can guide the design of new vaccines.

A “signature position” may alternatively be defined as a position at which the amino acid frequency distributions differ among the two sequence sets, irrespective of any reference amino acid. We develop methodology for detecting both types of signature positions. Henceforth we refer to signature positions involving (not involving) a reference amino acid as type A (type B) signatures.

The data set we analyze derives 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,

designed to prevent acquisition of HIV by inducing antibodies that could bind to neutralizing epitopes on HIV gp120 and destroy the virus before it infects host cells. The vaccine did not prevent HIV infection, with a similar rate of infection in the vaccine ( $241/3598 = 6.7\%$ ) and placebo ( $127/1805 = 7.0\%$ ) groups. 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 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. The alignment was constructed using ClustalX v.1.81 (Thompson et al., 1997) and manually edited. 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  $p = 581$ .

Consideration of one of the most commonly used methods for studying HIV signature positions, VESPA (Korber and Myers, 1992; <http://hiv-web.lanl.gov/content/hiv-db/mainpage.html>), demonstrates the need for new methodology. VESPA is purely descriptive- it evaluates potential type B signatures by comparing the frequency of the most common amino acid at positions between two sequence sets, without considering 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 (type A) or frequencies (type B) between the two sequence sets;
2. Approximate the null distribution of the test statistics across the set of studied amino acid positions, and obtain position-specific  $p$ -values;

3. Determine the set of signature positions as those with  $p$ -value less than a cut-off  $p_{cut}$ , estimated to control a false positive error rate at a pre-specified level.

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

For 2., we consider two approaches to approximating the null distributions. The first is a permutation procedure that only uses information at individual positions. The second approach, 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 an approach to coping with high dimensional data. We apply both approaches to obtain unadjusted  $p$ -values for each of the positions. For 3., we apply a permutation method to estimate the cut-off  $p$ -value  $p_{cut}$ .

This article is organized as follows. Section 2 develops four new test statistics for identifying type B signature positions and two new test statistics for identifying type A signature positions. Section 3 describes the procedures for obtaining  $p$ -values and the method for estimating the cut-off  $p$ -value, and describes four 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 tests in numerical

studies, Section 5 presents two examples, 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  $n_1 + n_2$  aligned amino acid sequences, one from each infected trial participant ( $n_1$  vaccine arm;  $n_2$  placebo arm), all of which are  $p$  amino acids long. For problem A the alignment also includes the reference sequence, which is the HIV sequence represented in the vaccine construct. 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  $p \sim 500 - 2600$  positions (HVTN, 2006).

For the  $i$ th position and the  $j$ th sequence in the  $k$ th 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  $Y_{kj}(i) = (Y_{kj}(i, 1), \dots, Y_{kj}(i, 21))^T$ , where  $Y_{kj}(i, a)$  is 1 if amino acid  $a$  is at position  $i$  and 0 otherwise,  $a = 1, \dots, 20$  ( $a = 1$  represents A, Alanine;  $a = 2$  represents C, Cysteine; and so on in the standard order), and  $a = 21$  represents a gap. Similarly define  $Y_{\text{ref}}(i) = (Y_{\text{ref}}(i, 1), \dots, Y_{\text{ref}}(i, 21))^T$  for the reference sequence, and let  $r(i)$  denote the amino acid at position  $i$  in the reference sequence. The vector  $Y_{kj}(i)$  is a 21-nomial random variable with response probability vector  $p_k(i) = (p_k(i, 1), \dots, p_k(i, 21))^T$ . The MLE of  $p_k(i)$  is  $\hat{p}_k(i) = (\bar{Y}_k(i, 1), \dots, \bar{Y}_k(i, 21))^T$ , 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 physico-chemical or evolutionary properties, and for problem A our methods incorporate a weight matrix to reflect such information. Let  $M$  be a  $21 \times 21$  matrix with nonnegative

entries, with  $(a, a')$ <sup>th</sup> element the weight/score summarizing dissimilarity of amino acids  $a$  and  $a'$ . The distance between the amino acid at position  $i$  in the  $j$ <sup>th</sup> sequence of group  $k$  to the amino acid at position  $i$  in the reference sequence is the appropriate element of  $M$ , computed as  $d_{kj}(i) = Y_{kj}(i)^T M Y_{\text{ref}}(i)$ . The simplest matrix  $M = J - I$ , with  $J$  the 21 by 21 matrix of ones and  $I$  the identity matrix; with this matrix  $d_{kj}(i) = 0$  (1) if the two amino acids under comparison are the same (different).

## 2.2 Two-sample Test Statistics for Problem B (No Reference Sequence)

For each position  $i$ , test statistics are developed to evaluate  $H_0^B(i) : p_1(i) = p_2(i)$  versus  $H_1^B(i) : p_1(i) \neq p_2(i)$ . Testing  $H_0^B(i)$  is equivalent to the well-known problem of testing for independence in a two-way ( $2 \times 21$ ) contingency table. Fisher’s exact test applies to this problem. However, it may not be most powerful for sequence data sets collected in practice, and the simulations verify that some of the new tests provide greater power than Fisher’s exact test.

Our first three proposed test statistics are based on summing weighted differences  $\{\hat{p}_1(i, a) - \hat{p}_2(i, a)\}^2$  over  $a = 1, \dots, 21$ , with three different approaches to standardizing/studentizing the summands. The first two statistics are related to Wu, Hsieh, and Li’s (2001) Euclidean-distance based statistics; the first is unstandardized (a numerator statistic) and the second divides each summand by its estimated variance. For high-dimensional data sets with small sample sizes the noise in variance estimation can erode power, potentially rendering the simpler numerator statistic more powerful (c.f., Pollard and van der Laan, 2003). Related to Wu, Hsieh, and Li’s (2001) Mahalanobis-distance based statistic, the third “fully standardized” statistic standardizes using the inverse of a nonparametric estimate of the  $21 \times 21$  covariance matrix of  $\hat{p}_1(i) - \hat{p}_2(i)$ . Heuristically these three statistics incorporate a hierarchy of degrees of regularization to dampen noise due to variance-covariance estimation: the first statistic employs full

regularization (no variance estimation), the third statistic employs no regularization (estimate the entire variance-covariance matrix), and the second statistic employs intermediate regularization (estimate the variances but set all covariances to zero). For HIV sequence data sets it is unknown which test performs best, and accordingly our simulations are designed to address this question.

In addition, we consider a test statistic based on Kullback-Leibler discrepancy, which is approximately an expected weighted log likelihood ratio comparing  $\hat{p}_1(i)$  and  $\hat{p}_2(i)$ . The Kullback-Leibler discrepancy has been widely studied and has well-known optimality properties closely related to those of likelihood ratio tests (c.f., Eguchi and Copas, 2002), which raises the conjecture that it will provide relatively high power.

For problem B our Euclidian-type statistics are defined by

$$\begin{aligned} Z_{E1}^B(i) &\equiv \sum_{a=1}^{21} \{\hat{p}_1(i, a) - \hat{p}_2(i, a)\}^2 I(\hat{v}(i, a) > 0), \\ Z_{E2}^B(i) &\equiv \sum_{a=1}^{21} \{(\hat{p}_1(i, a) - \hat{p}_2(i, a))/\hat{v}(i, a)\}^2 I(\hat{v}(i, a) > 0), \end{aligned} \quad (1)$$

where  $\hat{v}(i, a)^2$  estimates  $Var(\hat{p}_1(i, a) - \hat{p}_2(i, a))$ :

$$\hat{v}(i, a)^2 = \frac{(n_1 - 1)}{(n - 2)} \widehat{Var}(\hat{p}_1(i, a)) + \frac{(n_2 - 1)}{(n - 2)} \widehat{Var}(\hat{p}_2(i, a)),$$

with  $\widehat{Var}(\hat{p}_k(i, a)) = \hat{p}_k(i, a)(1 - \hat{p}_k(i, a))/n_k$ ,  $k = 1, 2$ .

The third statistic is given by

$$Z_M^B(i) \equiv (\hat{p}_1(i) - \hat{p}_2(i))^T \widehat{S}^{-}(i) (\hat{p}_1(i) - \hat{p}_2(i)), \quad (2)$$

where  $\widehat{S}^{-}(i)$  is the Moore-Penrose generalized inverse of  $\widehat{S}(i) = [(n_1 - 1)\widehat{S}_1(i) + (n_2 - 1)\widehat{S}_2(i)]/(n - 2)$ . Here  $\widehat{S}_k(i) = \text{diag}(\hat{p}_k(i)) - \hat{p}_k(i)\hat{p}_k(i)^T$  is the multinomial MLE of the  $21 \times 21$  covariance matrix  $S_k(i) = \text{diag}(p_k(i)) - p_k(i)p_k(i)^T$ . The matrix  $\widehat{S}^{-}(i)$  can be computed by the following steps: (1) Calculate the singular value decomposition



of  $\widehat{S}(i)$ ,  $\widehat{S}(i) = U \text{diag}(d) V^T$ , where  $U$  and  $V$  are orthogonal matrices and  $\text{diag}(d)$  is a diagonal matrix with diagonal vector  $d$ ; (2) Set  $d^*(a) = I(d(a) > 0)/d(a)$ ,  $a = 1, \dots, 21$ ; (3) Set  $\widehat{S}^-(i) = V \text{diag}(d^*) U^T$ . The statistic  $Z_M^B(i)$  is the Mahalanobis statistic that has been used extensively in many applications, although more commonly for quantitative data, not multinomial data (cf., Rao and Chakraborty, 1991).

The fourth statistic, based on Kullback-Leibler discrepancy, is relatively easy to compute. For position  $i$ , let

$$\begin{aligned} Z_{KL}^B(i) &\equiv \sum_{a=1}^{21} I(\widehat{p}_1(i, a)\widehat{p}_2(i, a) > 0) \widehat{p}_1(i, a) \log \left\{ \frac{\widehat{p}_1(i, a)}{\widehat{p}_2(i, a)} \right\} \\ &+ \sum_{a=1}^{21} I(\widehat{p}_1(i, a)\widehat{p}_2(i, a) = 0) \left( \widehat{p}_1(i, a) + n_1^{-1} \right) \log \left\{ \frac{\widehat{p}_1(i, a) + n_1^{-1}}{\widehat{p}_2(i, a) + n_2^{-1}} \right\}. \end{aligned} \quad (3)$$

Note that the standard Kullback-Leibler discrepancy for comparing  $\widehat{p}_1(i)$  and  $\widehat{p}_2(i)$  is  $\sum_{a=1}^{21} \widehat{p}_1(i, a) \log \{ \widehat{p}_1(i, a) / \widehat{p}_2(i, a) \}$ . If all possible amino acids and the gap character are not represented in group 2 sequences at position  $i$ , then this statistic equals infinity. Following the suggestion of Wu, Hsieh, and Li (2001), our statistic  $Z_{KL}^B(i)$  defined in (3) is modified to keep it finite.

### 2.3 Two-sample Test Statistics for Problem A (With a Reference Sequence)

To evaluate a type A signature at position  $i$ , we develop tests for  $H_0^A(i) : p_1(i, r(i)) = p_2(i, r(i))$  versus  $H_1^A(i) : p_1(i, r(i)) \neq p_2(i, r(i))$ , which assesses equal frequencies of the reference amino acid at position  $i$  in the two sequence sets. We base tests of  $H_0^A(i)$  on a comparison of average distances  $d_{kj}(i) = Y_{kj}(i)^T M Y_{ref}(i)$  (defined at the end of Section 2.1) between groups  $k = 1$  and  $2$ , with  $\text{diag}(M) = 0$ . These averages can be written as  $\bar{d}_k(i) = n_k^{-1} \sum_{j=1}^{n_k} d_{kj}(i) = \sum_{a=1}^{21} M(a, r(i)) \widehat{p}_k(i, a)$ . Parallel to the type B statistics  $Z_{E1}^B(i)$  and  $Z_{E2}^B(i)$ , we consider unstandardized and standardized statistics,

$$\begin{aligned} Z_1^A(i) &\equiv \bar{d}_1(i) - \bar{d}_2(i), \\ Z_2^A(i) &\equiv \left\{ \bar{d}_1(i) - \bar{d}_2(i) \right\} / s(i), \end{aligned} \quad (4)$$

where  $s(i) = [\{(n_1 - 1)/(n - 2)\}s_1^2(i) + \{(n_2 - 1)/(n - 2)\}s_2^2(i)]^{1/2}$ , with  $s_k^2(i)$  the sample variance of  $d_{kj}(i)$ ,  $j = 1, \dots, n_k$ , for  $k = 1, 2$ .

### 3. Judging Statistical Significance

#### 3.1 Permutation-based Unadjusted $p$ -values (Marginal- No Pooling)

To judge statistical significance of the  $p$  tests, first nominal (unadjusted) position-specific  $p$ -values are computed. Although analytic  $p$ -values can be computed for most of the test statistics, to avoid the requirement of large sample sizes and to create a uniform approach for the different statistics, we use a permutation procedure to determine  $p$ -values (except for Fisher’s exact test for which we use analytic  $p$ -values). Specifically,  $B^{perm}$  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  $B^{perm}$  permuted data sets that equal or exceed the value of the original test statistic.

#### 3.2 Nonparametric Estimated Null Distribution-based Unadjusted $p$ -values (Pooling)

In the second (pooling) approach to computing position-specific  $p$ -values, slightly modified versions of  $Z_{Em}^B(i)$  and  $Z_m^A(i)$  are needed,  $m = 1, 2$ , as described below. These modified statistics incorporate a position-specific weight  $w_1(i)$ ,  $i = 1, \dots, p$ , which can be used to reflect biological information. For example, positions could be weighted by their conservancy (a position is relatively conserved if most sequences contain the same amino acid at the position), since conserved positions may be more functionally or structurally important than variable positions. For exploratory analyses, where the aim is to generate hypotheses about positions that warrant further biological examination, equal weights  $w_1(i) = 1$  may be recommended, because they prevent subjective biases from influencing the results, and they may be agreed upon broadly among investigators. For these reasons equal weights are used in the Examples.

To develop the pooling approach, we follow Pan's (2003) clever idea for how to directly nonparametrically estimate the null distribution of hundreds of t-statistics. Assume that under all  $H_0(i)$ 's, the test statistics of interest  $Z(i)$  have the same distribution for  $i = 1, \dots, p$ . 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  $n_{k2} = n_{k1} + 1$  otherwise,  $k = 1, 2$ . To evaluate type B signatures, the test statistic  $Z_{E1}^B(i)$  of (1) is modified (slightly) to  $Z_{E1}^{Bsplit}(i) = w_1(i) \times$

$$\sum_{a=1}^{21} \{(\widehat{p}_{11}(i, a) + \widehat{p}_{12}(i, a))/2 - (\widehat{p}_{21}(i, a) + \widehat{p}_{22}(i, a))/2\}^2 I(\widehat{v}(i, a) > 0),$$

where  $\widehat{p}_{k1}(i, a) = n_{k1}^{-1} \sum_{j=1}^{n_k} Y_{kj}(i, a) I(j \in J_{k1})$  averages the  $Y_{kj}(\cdot)$  in the first permuted half of sample  $k$  and similarly  $\widehat{p}_{k2}(i, a)$  averages the  $Y_{kj}(\cdot)$  in the second permuted half. The statistic  $Z_{E1}^{Bsplit}(i)$  approximately equals  $Z_{E1}^B(i)$ , and motivates a statistic that estimates its null distribution:  $z_{E1}^{Bsplit}(i) \equiv w_1(i) \times$

$$\sum_{a=1}^{21} \{(\widehat{p}_{11}(i, a) - \widehat{p}_{12}(i, a))/2 + (\widehat{p}_{21}(i, a) - \widehat{p}_{22}(i, a))/2\}^2 I(\widehat{v}(i, a) > 0).$$

Because the numerator of  $z_{E1}^{Bsplit}(i)$  is the sum of within-sample differences, its mean is zero, and  $z_{E1}^{Bsplit}(i)$  can be expected to approximate the null distribution of  $Z_{E1}^{Bsplit}(i)$ . Split statistics  $Z_{E2}^{Bsplit}(i)$  and  $z_{E2}^{Bsplit}(i)$  are formed in the same way, with  $\widehat{v}_1(i, a)^2 + \widehat{v}_2(i, a)^2$  added to the denominator of the summand of each statistic, where

$$\widehat{v}_k(i, a)^2 = \frac{(n_{k1} - 1)}{(n_k - 2)} \widehat{Var}(\widehat{p}_{k1}(i, a)) + \frac{(n_{k2} - 1)}{(n_k - 2)} \widehat{Var}(\widehat{p}_{k2}(i, a)),$$

with  $\widehat{Var}(\widehat{p}_{kl}(i, a)) = \widehat{p}_{kl}(i, a)(1 - \widehat{p}_{kl}(i, a))/n_{kl}$ ,  $k = 1, 2; l = 1, 2$ .

To obtain  $p$ -values, once  $Z_{Em}^{Bsplit}(i)$  is computed, each group of sequences is again separately randomly permuted into two halves, and  $z_{Em}^{Bsplit}(i)$  is computed. Based on  $B_{split}^{perm}$  separate permutations  $z_{Em}^{Bsplit(b)}(i)$  is computed  $B_{split}^{perm}$  times,  $b = 1, \dots, B_{split}^{perm}$ . For position  $i$  the  $p$ -value is then  $p_i = N_i / (B_{split}^{perm} \times p)$ , where  $N_i$  is the number of the

test statistics  $z_{Em}^{Bsplit(b)}(i')$  that equal or exceed  $Z_{Em}^{Bsplit}(i)$ , pooling over  $i' = 1, \dots, p$  and  $b = 1, \dots, B_{split}^{perm}$ . We also considered a split-version of  $Z_M^B(i)$ ; however since it performed poorly in simulations we do not discuss it further. A computational advantage of the split test statistics is that setting  $B_{split}^{perm} = 5$  achieves good performance, as verified in the simulations. A small number of permutations suffices because of the pooling of information across positions.

For developing split tests of type A signatures, set  $\bar{d}_{k1}(i) = n_{k1}^{-1} \sum_{j=1}^{n_k} d_{kj}(i)$   $I(j \in J_{k1})$  and  $\bar{d}_{k2}(i) = n_{k2}^{-1} \sum_{j=1}^{n_k} d_{kj}(i)I(j \in J_{k2})$ ,  $k = 1, 2$ . Define the test statistic

$$Z_1^{Asplit}(i) \equiv w_1(i) \left\{ (\bar{d}_{11}(i) + \bar{d}_{12}(i))/2 - (\bar{d}_{21}(i) + \bar{d}_{22}(i))/2 \right\}.$$

The null distribution of  $Z_1^{Asplit}(i)$  can be approximated by

$$z_1^{Asplit}(i) \equiv w_1(i) \left\{ (\bar{d}_{11}(i) - \bar{d}_{12}(i))/2 + (\bar{d}_{21}(i) - \bar{d}_{22}(i))/2 \right\}.$$

Similar statistics  $Z_2^{Asplit}(i)$  and  $z_2^{Asplit}(i)$  are formed by placing  $(\hat{\tau}_1(i)^2 + \hat{\tau}_2(i)^2)^{1/2}$  in the denominator of each statistic, where  $\hat{\tau}_k(i)^2 = \{(n_{k1} - 1)/(n_k - 2)\}s_{k1}^2(i) + \{(n_{k2} - 1)/(n_k - 2)\}s_{k2}^2(i)$ , with  $s_{kl}^2(i)$  the sample variance of  $\{d_{kj}(i) : j \in J_{kl}\}$ , for  $k = 1, 2; l = 1, 2$ .

Note that the position weights  $w_1(i)$  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 permutation-based  $p$ -values, because they are computed marginally.

We also studied modified versions of the statistics  $Z_{E2}^B(i)$ ,  $Z_M^B(i)$ , and  $Z_{E2}^{Bsplit}(i)$  that incorporate a small positive constant in the denominator to stabilize the statistic (see Web Appendix A). In simulations these tests had lower power than the tests described above, and therefore are not considered further.

### 3.3 Permutation-Based Control of the Per Comparison Error Rate

We use a permutation-based method that requires no distributional assumptions to estimate the number of false positive rejections  $V$  at some pre-fixed number. This

allows estimating the per comparison error rate (PCER) by  $\widehat{V}/p$ , and the false discovery rate by  $I(R > 0)\widehat{V}/R$ , where  $R$  is the number of rejections. The estimate  $\widehat{V}$  is computed with the following steps: (1) Construct a permuted data set that satisfies  $H_0(i)$  for all  $i$ , by permuting whole sequences as described in Section 3.1; (2) Analyze the permuted data set in the same way as the real data set, yielding unadjusted p-values for this data set; (3) For a given fixed p-value threshold  $p_{cut}$ , count the number of rejections, which estimates  $V$ ; (4) Fine tune the choice of  $p_{cut}$  such that  $\widehat{V} = V$ ; (5) Reject  $H_0(i)$  if  $p_i < p_{cut}$ . The parameter  $V$  can be estimated more precisely by generating  $N^{null}$  permuted data sets in Step (1) and estimating  $V$  in Step (3) by the average number of rejections over the  $N^{null}$  data sets. In the simulations  $N^{null} = 1$  gave good performance.

### 3.4 Screening Out Highly Conserved Positions

There is little or no power to detect signatures at positions with very limited amino acid variability. Therefore highly conserved positions are pre-screened out, based on Tarone’s (1990) technique for improving power of the Bonferroni correction for discrete data. Tarone’s (1990) procedure first screens out hypothesis tests using a simple algorithm, leaving  $K \leq p$  hypotheses to test, and second rejects the  $i$ th hypothesis if the unadjusted p-value  $p_i < \alpha/K$ . If  $K < p$  this method can provide greater power than the Bonferroni method. The procedure involves computing a minimum achievable significance level  $\alpha_i^*$  for each test, calculated from data pooled over the two groups. Due to the complexity of computing the  $\alpha_i^*$  for each of the new test statistics, for the Simulations and Examples the  $\alpha_i^*$  were computed based on Fisher’s exact test.

## 4. Simulation Study

### 4.1 Design of the Simulation Study

The simulation study is designed based on data from the first HIV vaccine efficacy trial (Flynn et al., 2005) as described in the introduction. For each of the 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) How do the position weights  $w_1(i)$  influence performance of the split test statistics? 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. Assuming half as many vaccine recipients got infected as placebo recipients (i.e., vaccine efficacy = 50%),  $n_1 = 45$  or 90 sequences were generated for the infected vaccine group. These sequences were generated in two steps. First, sequences were sampled in the same way as the placebo sequences. Second, the HIV-specific Point Accepted Mutation (PAM) matrix developed by Nickle et al. (2005) was used to create amino acid mutations in some of the vaccine group sequences at the positions  $i$  where the alternative hypothesis is true. The PAM matrix is  $20 \times 20$ , with the 20 amino acids running down the rows and across the columns (see Web Table 1). Each nondiagonal entry of the PAM matrix corresponds to two different amino acids, and equals the estimated probability that either of the amino acids mutates into the other one, given a specified probability of any mutation at all. The estimated probabilities of amino acid interchange were computed based on thousands of observed mutations in HIV sequences (see Web Appendix B). We used the PAM–25 matrix, which specifies a 25% chance that the amino acid at position  $i$  in a vaccine recipients’ sequence will be mutated. Independently for each alternative hypothesis position and each vaccine group sequence, the amino acid was mutated to one of the 19 other amino acids according to the probabilities in the PAM-25 matrix.

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 36 positions that are involved with CD4-binding, are in CD4-induced epitopes, or that constitute a neutralization epitope defined by the monoclonal antibody 2G12. 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. The positions, here and in the Example, are numbered using the standard HXB2 strain numbering system (Kuiken et al., 2002). 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 considered 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 small ( $n_1/n_2 = 45/90$  infections) and large ( $n_1/n_2 = 90/180$  infections) efficacy trials. Question 3) 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 evaluate the split test statistics when the true alternative hypotheses are upweighted 2-fold (correctly incorporating prior knowledge) or downweighted 2-fold (incorrectly incorporating prior knowledge), respectively.

Except for results reported at the end of Section 4.2, positions in the split statistics were weighted equally ( $w_1(i) = 1$ ). Tests were carried out at pre-specified PCER = 0.01, 0.05, and 0.10, using  $B^{perm} = 1000$  permutations to approximate unadjusted p-values for the non-split statistics and  $B_{split}^{perm} = 5$  permutations for the split statistics. In Step (1) of the algorithm described in Section 3.3  $B^{null} = 1$  data set is generated under

the complete null hypothesis. The PCER, the false positive rate (FPR, the percentage of true null positions rejected), and the true positive rate (TPR, the percentage of true alternative hypotheses rejected) were estimated by averaging over 250 simulated vaccine trials. The performance of the tests can be compared by plotting the estimated TPRs versus FPRs. We also evaluated the PCER because this error rate can be controlled at a fixed level in applications, whereas precise control of the FPR is difficult to achieve.

#### 4.2 Simulation Results

For the type B tests, Figures 2-4 show the estimated TPRs versus FPRs, for the scenarios where 1%, 10%, and 25% of the alternative hypotheses are true, respectively. We make several observations. First, the Kullback-Leibler ( $Z_{KL}^B(i)$ ) and standardized Euclidean ( $Z_{E2}^B(i)$ ) statistics are consistently most powerful. Their power advantage is greatest at low FPRs. Second, when  $FPR \geq 0.05$ , the power of  $Z_{KL}^B(i)$  and  $Z_{E2}^B(i)$  is almost matched by that of  $Z_{E1}^B(i)$  and  $Z_{E2}^{Bsplit}(i)$ . Third, the test based on  $Z_M^B(i)$  has relatively low power, especially when 1% of the alternative hypotheses are true. To explain the poor performance of the Mahalanobis-based statistic, note that the rank of the estimated covariance matrix  $\hat{S}(i)$  is often fairly high, which occurs because the gp120 region of HIV 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 covariance estimates set to zero, in which case the Mahalanobis-based statistic is very similar to the Euclidean-based statistic  $Z_{E2}^B(i)$ . With this modification these two approaches performed similarly.

Fourth, the split statistics with true alternative positions upweighted have greater TPRs and smaller FPRs than the equal-weighted methods; for example with  $m_1/m_2 = 45/90$ ,  $PCER = 0.05$  and 10% of the alternative hypotheses true, the TPR/FPR of



the  $Z_{E2}^{Bsplit}(i)$  tests is 0.94/0.015 compared to 0.90/0.051 for the unweighted tests. On the other hand when the true alternative positions were downweighted, the opposite results attains, with TPR/FPR of 0.74/0.068. These results provide preliminary “proof of principle” that correct upweighting of positions can improve performance of the split test statistics, but incorrect weighting can erode performance. This suggests that weighting to incorporate biological knowledge should be done with caution. Fifth, most of the estimated PCERs are close to their pre-specified values, showing that the procedures correctly control the PCER (results not shown).

The four tests for type A signatures were evaluated using the same simulated data sets. The estimated PCERs are close to their pre-specified values (results not shown). In addition the tests have comparable TPRs, although the split tests sometimes outperform or underperform the non-split tests (Web Figure 1). The comparable powers may be explained by the fact that the type A tests are all variants of t-statistics.

## 5. Examples

We now consider the evaluation of type A signature positions for the data from the efficacy trial described in the introduction. The matrix  $M$  was taken as  $J - I$ , or as the reciprocal of 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. Because the previously available amino acid substitution matrices were built using organisms other than HIV, this PAM may yield more accurate rates of HIV amino acid interchanges. Taking the reciprocal upweights rare amino mismatches, which may have greater biological significance.

The tests were performed using  $B^{perm} = 10,000$  or  $B_{split}^{perm} = 1000$  permutations and  $N^{null} = 1$  null data set generated in Step (1) of the algorithm described in Section 3.3. Tarone’s (1990) procedure screened out 232 of the 581 amino acid positions. With

$w_1(\cdot) = 1$ ,  $M = J - I$ , and PCER fixed at 0.01, the cut-off  $p$ -value  $p_{cut}$  ranged from 0.0009 to 0.0085 for the 4 type A test statistics. The statistics rejected 1, 1, 6, and 6 hypotheses (Figure 5). Because 349 positions were analyzed, 3.5 false positives per test statistic are expected. Therefore the results are consistent with all of the null hypotheses being true. This conclusion is supported by the observation that only one test across all the positions and test statistics is significant under the Holm-Bonferroni adjustment procedure applied to control the family-wise error rate at level 0.10 (with cut-off  $p$ -value of 0.00023). Specifically position 268 had unadjusted  $p = 0.0027, 0.0023, 0.0057, < .0001$  for  $Z_1^A(i), Z_2^A(i), Z_1^{Asplit}(i), Z_2^{Asplit}(i)$ . Similar null results were obtained when  $M$  was set as the reciprocal PAM matrix. The result of no signature positions can be explained by the inability of the tested vaccine to prevent HIV infection. If the vaccine does not impact susceptibility to HIV acquisition, then the distribution of infecting sequences should be identical in the vaccine and placebo groups.

To illustrate the tests for evaluating type B 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. Many significant signatures are found by all of the procedures; for example with PCER = 0.01,  $Z_{E2}^B(i)$  and  $Z_{KL}^B(i)$  yield 67 and 71 significant signature positions out of 448 screened-in positions, compared to 35 for Fisher's exact test.

## 6. Discussion

For comparing two sets of amino acid sequences, we developed and evaluated four new testing procedures for detecting type A signature positions (positions where amino acids have a different probability of mismatch relative to a reference amino acid) and six

new testing procedures for detecting type B signature positions (positions where amino acids have a different frequency distribution, irrespective of a reference amino acid). For evaluating type B signatures the Kullback-Leibler statistic  $Z_{KL}^B(i)$  and the standardized Euclidean statistic  $Z_{E2}^B(i)$  were most powerful and are recommended. The split statistic  $Z_{E2}^{Bsplit}(i)$  may also be recommended, based on its fairly good performance and its computational speed, only requiring  $\approx 5$  permutations. A statistic similar to our  $Z_{E2}^B(i)$  was found to perform well by Wu et al. (2001). We conjecture that the Euclidean-based statistics provide greater power than the Mahalanobis-based statistic  $Z_M^B(i)$  because the latter statistic includes a nonparametric estimate of a large covariance matrix, introducing considerable noise. The efficiency of the Kullback-Leibler test likely derives from its similarity to a likelihood ratio test.

The four t-type statistics developed for evaluating type A signatures performed comparably in the simulation models considered. These tests differ from previously developed t-type tests for large-scale significance testing in that the data (amino acid distances) are discrete, and the test statistics advantageously incorporate a weight matrix specifying dissimilarity values for all pairs of different amino acids. Furthermore, the type A and B split-statistic methods advantageously incorporate weights on amino acid positions. These weights allow the techniques to flexibly reflect biological knowledge about sequences, and their tailoring to different applications.

#### SUPPLEMENTARY MATERIALS

Web Appendices, Tables and Figures referenced in Sections 3.2 and 4.1 are available under the Paper Information link at the Biometrics website <http://www.tibs.org/biometrics>.

#### ACKNOWLEDGEMENTS

We are grateful to the Associate Editor for many suggestions that led to corrections and improvements. This research was supported by NIH grant 2 R01 AI54165-04.

## REFERENCES

- 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](http://www.ism.ac.jp/~eguchi/pdf/KLNP.pdf).
- Flynn, N.M., Forthal, D.N., Harro, C.D., Mayer, K.H.; 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.
- Gilbert, P.B., Self, S. and Ashby, M. (1998). Statistical methods for assessing differential vaccine protection against human immunodeficiency virus types. *Biometrics* **54**:799-814.
- HVTN (2006). The Pipeline Project. Available at: <http://www.hvtn.org/science>.
- 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.
- 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. *Technical Re-*

port, University of Washington.

- Pan, W. (2003). On the use of permutation in and the performance of a class of non-parametric methods to detect differential gene expression. *Bioinformatics* **19**:1333-1340.
- Pollard, K.S. and van der Laan, M.J. (2003). Multiple testing for gene expression data: an investigation of null distributions with consequences for the permutation test. *Proceedings of the 2003 International MultiConference in Computer Science and Engineering, METMBS 2003 Conference*:3-9.
- Rao, C.R., Chakraborty, R., Eds. (1991). Handbook of statistics. Volume 8: Statistical methods in biological and medical sciences. Elsevier, New York, New York.
- 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 Research* **25**:4876-4882.
- Wei, X., Decker, J.M., Wang, S., et al. (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., et al. (1998). The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature* **393**:705-711.

## Figure Legends

**Figure 1.** Illustration of amino acid sequence data available for genome scanning analysis, from 6 randomly selected vaccine and placebo recipients who got HIV infected during the VaxGen trial, aligned together with the reference HIV vaccine sequence GNE8. Each capital letter denotes an amino acid, which is a basic building block of proteins. A - denotes a gap that arose in the alignment; gaps occur because the lengths of HIV sequences differ. The V3 loop region within the HIV protein gp120 is shown, which consists of positions 297-329 of gp160 using the HXB2 strain numbering system (Kuiken et al., 2002).

**Figure 2.** Average true positive rates (TPRs) versus average false positive rates (FPRs) for evaluating type B signatures with the alternative hypothesis true for 1% of positions. (a) and (b) are for trials with 45/90 and 90/180 vaccine/placebo sequences.

**Figure 3.** Average true positive rates (TPRs) versus average false positive rates (FPRs) for evaluating type B signatures with the alternative hypothesis true for 10% of positions.

**Figure 4.** Average true positive rates (TPRs) versus average false positive rates (FPRs) for evaluating type B signatures with the alternative hypothesis true for 25% of positions.

**Figure 5.** Histograms of the 4 type A test statistics for the 349 screened-in positions among the  $p = 581$  HIV gp120 positions sequenced in the VaxGen trial, with equal weighting of all positions and amino acid mismatches. For  $PCER = 0.01$  the statistics  $Z_1^A(i)$ ,  $Z_2^A(i)$ ,  $Z_1^{Asplit}(i)$ , and  $Z_2^{Asplit}(i)$  rejected 1, 1, 6, and 6 hypotheses, respectively.

Examine each amino acid position as a potential 'signature position'


V3 loop amino acid sequence of reference GNE8 strain      ...TRPNNNTRRSIHIG-PGR-AFYATGEIIGDIRQ...

Vaccine group V3 loop sequences

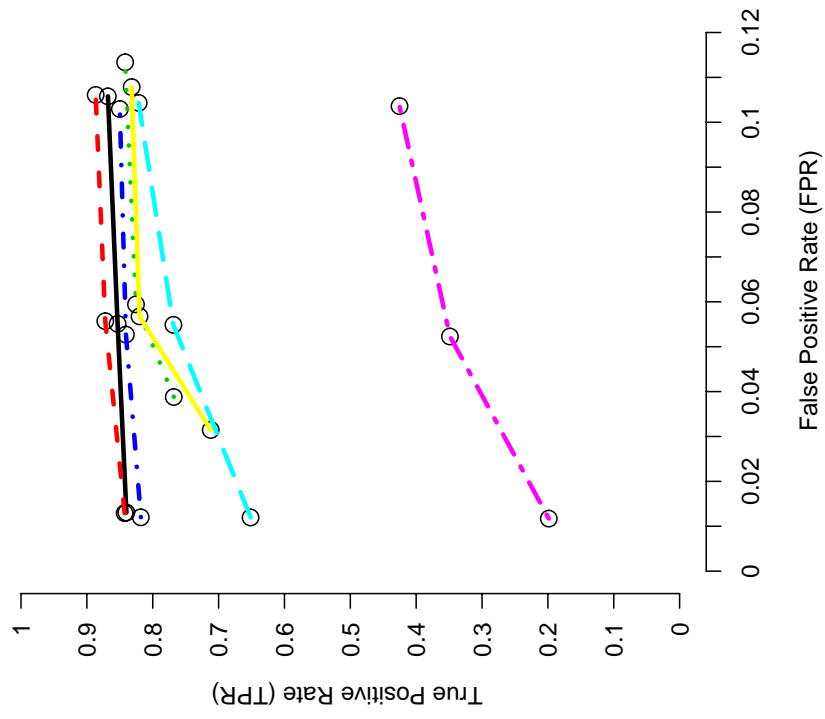
1. ...TRPNNNTRRRIH LG-PGR-AFYATG-IIGDIRQ...
2. ...TRPNNNTRKGIHIG-PGR-AFYATGEIIGNIRQ...
- .
- .
- .
217. ...TRPSNNTRKGIHIG-PGR-AFYATEEITGDIRQ...

Placebo group V3 loop sequences

1. ...TRPNNNTRTG VHLG-PGR-VWYATGDIIGDIRQ...
2. ...TRPNNNTRRSIHIQ-PGR-AFYAT-DIIGDIRK...
- .
- .
- .
119. ...TRPNNNTISKIRIR-PGRGSFYATNNIIGDIRQ...



(a) 45/90 Vaccine/Placebo Sequences



(b) 90/180 Vaccine/Placebo Sequences

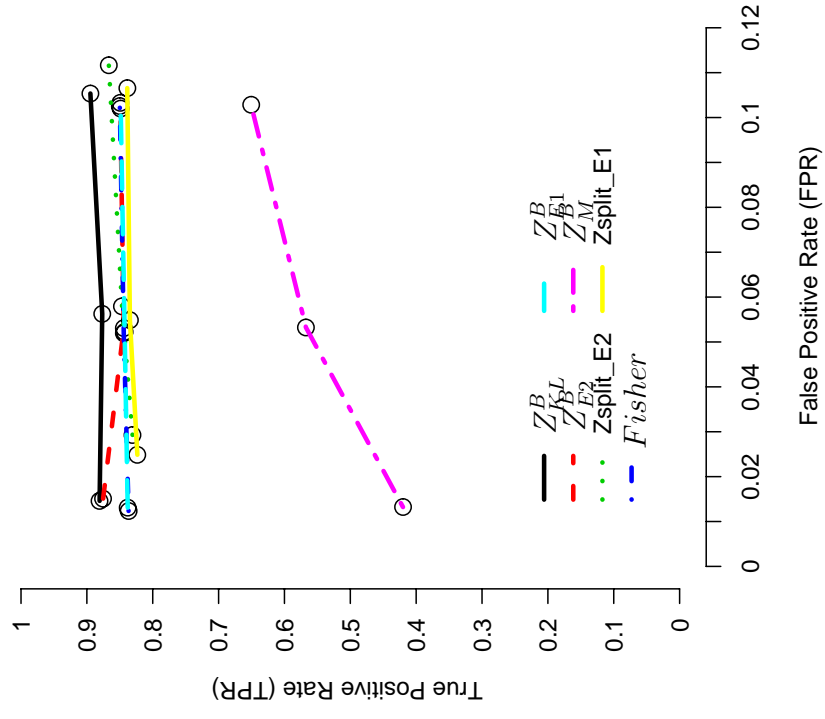
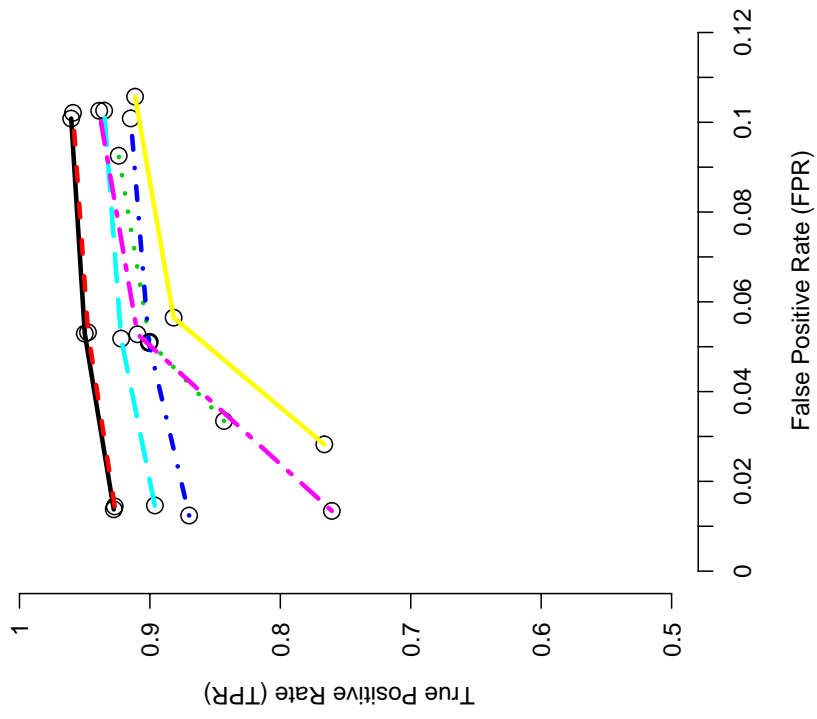


Figure 2. Alternative Hypothesis True for 1% of Positions



(a) 45/90 Vaccine/Placebo Sequences



(b) 90/180 Vaccine/Placebo Sequences

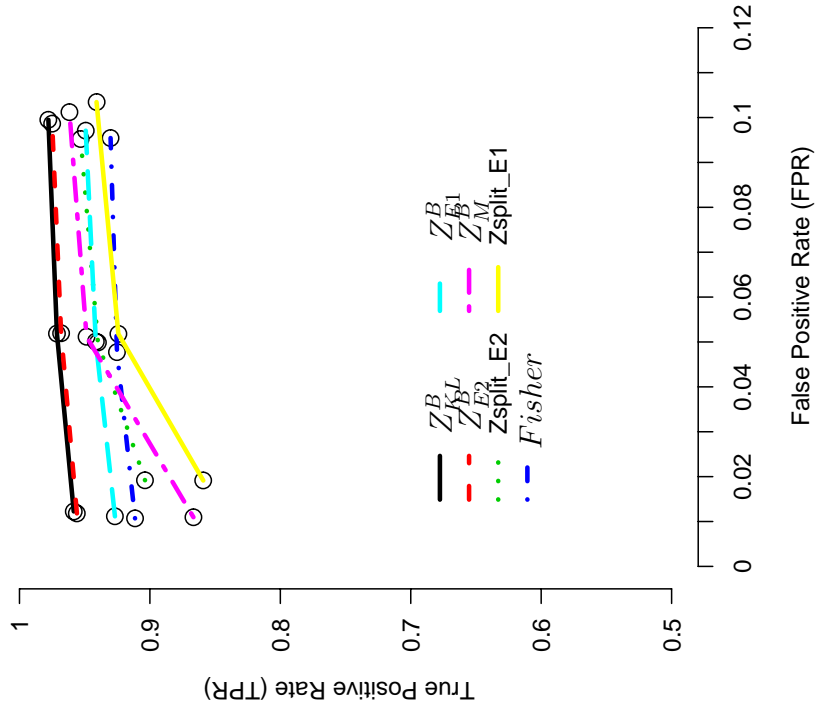
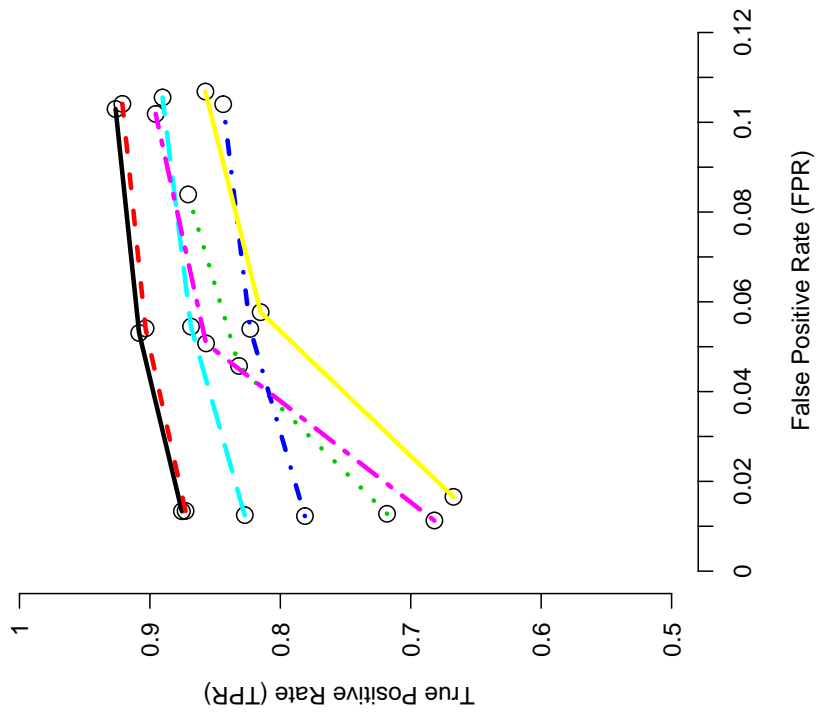


Figure 3. Alternative Hypothesis True for 10% of Positions

(a) 45/90 Vaccine/Placebo Sequences



(b) 90/180 Vaccine/Placebo Sequences

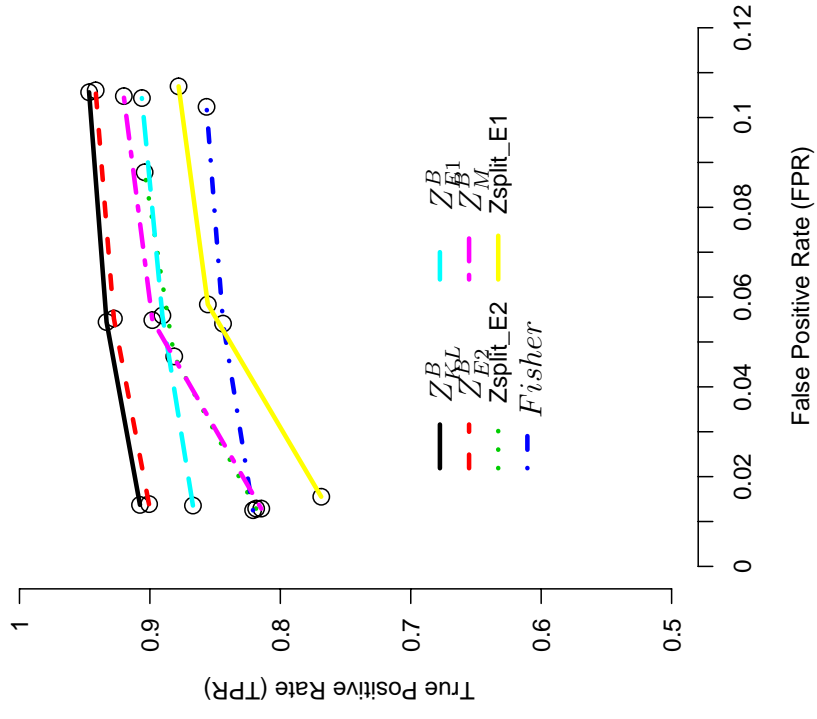


Figure 4. Alternative Hypothesis True for 25% of Positions

