HIV VACCINE TRIALS NETWORK

Genome scanning methods for comparing HIV sequences between two groups

Biostat 578A: Lecture 9

A manuscript corresponding to this talk is posted on the course webpage (GilbertWuJobes.2006.pdf)

Peter Gilbert and Chunyuan Wu

HIV VACCINE TRIALS NETWORK

Outline

- Introduction/Scientific objectives
- Statistical methods
- Simulations
- Examples
- Discussion

HIV VACCINE TRIALS NETWORK

Introduction

- The extensive genetic diversity of HIV poses a formidable challenge to the development of an efficacious HIV vaccine
- An HIV vaccine may only prevent infections with exposing viruses that are genetically identical or highly similar in certain positions or regions of HIV
  - An amino acid mismatch in one or more key positions may disallow protective efficacy

HIV VACCINE TRIALS NETWORK

Introduction

- In an HIV vaccine efficacy trial, the HIVs that infect participants are sequenced (nucleotides and amino acids)
- Comparison of the sequences between infected vaccine recipients and infected placebo recipients forms the basis for assessing how vaccine efficacy depends on genetic mismatching of exposing HIVs
- Statistical methods have been developed for the case that sequence differences are summarized by 1 or a few numbers
  - Low-dimensional case: $n = 150 - 400,\ p = 1 - 4$

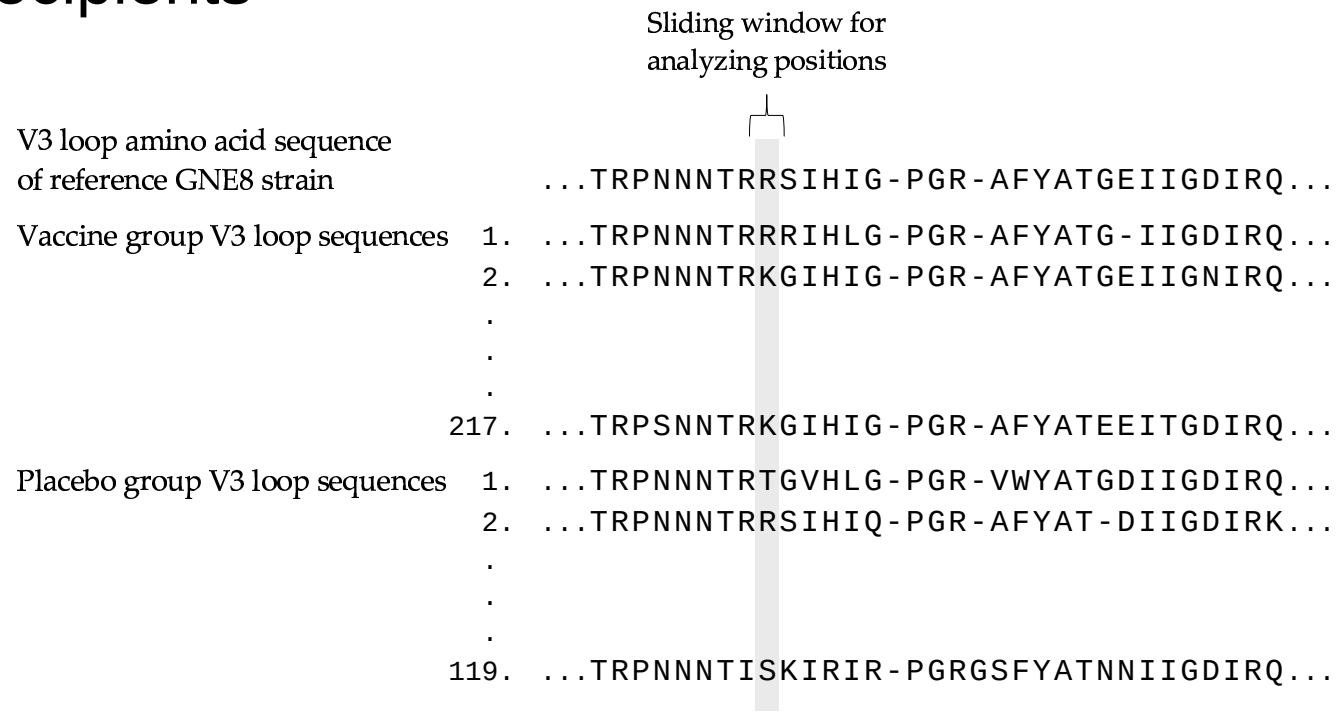
HIV VACCINE TRIALS NETWORK

Introduction

- This work considers the amino acid positions as the variables
  - High-dimensional case: $n = 150 - 400$, $p = 35 - 3000$
  - $p >> n$
- Techniques are developed for “genome scanning"
  - A sliding window is used position-by-position across the multiple alignment of amino acids to search for “signature positions"

Goal of Genome Scanning

- **Goal:** Identify positions at which the amino acids in sequences from infected vaccinees tend to be more divergent from the corresponding amino acid in the reference sequence (vaccine prototype) than the amino acids in sequences from infected placebo recipients



HIV VACCINE TRIALS NETWORK

Goal of Genome Scanning

- Finding a signature position is useful because:
  - It provides knowledge of a position that is key for neutralization, informing on the link between genotype and serotype
  - It suggests adding immunogens to the vaccine that represent multiple different amino acids at the position, to protect against a broader spectrum of HIV strains
- The purpose of genome scanning analysis is to guide the iterative (re)-design of HIV vaccines

Methods: Approach to Genome Scanning

(i) Define the dissimilarity between amino acids

(ii) For each position, construct a two-sample test statistic that compares amino acid divergences or frequencies between the two groups

(iii) Approximate the null distribution of the test statistics across the set of studied amino acid positions, and obtain position-specific p-values

(iv) 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

(i) Dissimilarity Between Two Amino Acids

- Simplest distance: 0 if match; 1 if mismatch (VESPA, Korber and Myers, 1992)
- Of interest to generalize to weight the different kinds of amino acid mismatches (A vs C, G vs Y, etc.)
  - Weight by physical/chemical/biological properties relevant to neutralizing antibodies
    - Hydrophilicity (Hopp and Woods)
    - Surface accessibility based on 3-D structure (Hopp)
    - Antigenicity scale based on known continuous antibody epitopes (Welling et al.)

HIV VACCINE TRIALS NETWORK

Preliminaries/Notation

- Data are $n_1 + n_2 + 1$ aligned amino acid sequences each with length p
- For the ith position and the jth sequence in the kth group, let

$$Y_{kj}(i) = (Y_{kj}(i,1), \cdots, Y_{kj}(i,21))'$$

  be the 21-vector with a 1 at the observed amino acid and zeros elsewhere
- For the ith position in the reference sequence: $Y_r(i) = (Y_r(i,1), \cdots, Y_r(i,21))'$
  - Let $r(i)$ denote the amino acid at position i in the reference sequence

HIV VACCINE TRIALS NETWORK

Preliminaries/Notation

- Vector of response probabilities

$$p_k(i) = (p_k(i,1), \cdots, p_k(i,21))'$$

- MLE of $p_k(i)$:

$$\widehat{p}_k(i) = (\overline{Y}_k(i,1), \cdots, \overline{Y}_k(i,21))'$$

where $\overline{Y}_k(i,a) = n_k^{-1} \sum_{j=1}^{n_k} Y_{kj}(i,a)$

HIV VACCINE TRIALS NETWORK

Preliminaries/Notation

- Let M be a 21×21 weight matrix with nonnegative entries, with $(a, a')^{\text{th}}$ element the weight/score indicating the dissimilarity of amino acids a and a'
  - $M(a, r(i))$ is the divergence between amino acid a and the amino acid $r(i)$ in the reference sequence
- For the j^{th} sequence in group k at position i: Distance is

$$d_{kj}(i) = Y_{kj}(i)' M Y_r(i)$$

Preliminaries/Notation

- Simplest amino acid weight matrix:

$$M = J - I$$

with J the 21 by 21 matrix of ones and I the identity matrix

  - $d_{kj}(i) = 0$ if the amino acids match; 1 if mismatch
  - Equal weighting of all mismatches

HIV VACCINE TRIALS NETWORK

(ii) Two-Sample Test Statistics

- For a position i, test statistics are developed to evaluate

$$H_0(i) : p_1(i) = p_2(i) \quad \text{vs} \quad H_1(i) : p_1(i) \neq p_2(i)$$

- Two types of tests:
  - Tests for differential amino acid divergence from the reference amino acid
    - Specified by zeros on the diagonal of M
  - Tests for differential amino acid frequencies, irrespective of any reference
    - Specified by positive elements on the diagonal of M

(ii) Previous Approaches to Summarizing Amino Acid Divergence

- Standardized Euclidean (Wu et al., 2001)
- Mahalanobis (Kowalski et al., 2002)
- Kullback-Leibler (Wu et al., 2001)

- This work generalizes each of these approaches

(ii) Standardized Euclidean Test Statistic

- For position i, set $\widehat{v}^2(i,a) =$

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

- Define

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

where $C_E(i)$ is a leading constant depending on sample size and λ_1 is a nonnegative constant

(ii) Standardized Euclidean Test Statistic

- λ_1 is added to the denominator to stabilize the statistics
- Several authors including 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
  - Lonnstedt and Speed (2002) showed that the modified statistics perform better than the usual t-statistic
- Following Tusher et al. (2001), λ_1 is chosen to minimize the coefficient of variation of the $Z_E(i)$
- Alternatively, λ_1 could be chosen as the 90^{th} percentile of $\{\widehat{v}(i,a) : i = 1, \cdots, p; a = 1, \cdots, 21\}$ (Efron et al., 2001)

(ii) Standardized Euclidean Test Statistic

- If $\lambda_1 = 0$, then in large samples, under $H_0(i)$, $Z_E(i)$ (with $M = J$) has an F distribution with 1 and $n - p^*(i) - 1$ degrees of freedom
- $Z_E(i)$ makes a standardized comparison of the weighted distances between the two groups

HIV VACCINE TRIALS NETWORK

(ii) Mahalanobis Test Statistic

- Mahalanobis' D^2 statistic for position i:

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

$\widehat{S}^{-}_{\lambda_2}(i)$ is the Moore-Penrose generalized inverse of

$$\widehat{S}_{\lambda_2}(i) \equiv \widehat{S}(i) + \lambda_2 diag(\underline{1}_{nz}(i))$$

- $\widehat{S}(i) = [(n_1 - 1)\widehat{S}_1(i) + (n_2 - 1)\widehat{S}_2(i)]/(n-2)$
- $\widehat{S}_k(i) = \widehat{p}_k(i)I - \widehat{p}_k(i)\widehat{p}_k(i)'$ is the multinomial MLE of $S_k(i) = p_k(i)I - p_k(i)p_k(i)'$
- λ_2 is a nonnegative constant
- $\underline{1}_{nz}(i)$ is the 21-vector of indicators of whether the ath row of $\widehat{S}(i)$ is not the zero vector

(ii) Computing the Moore-Penrose Generalized Inverse $\widehat{S}_{\lambda_2}^{-}(i)$

- **Step 1**: Compute 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)
- **Step 2**: Expand the resulting generalized inverse to a 21×21 matrix by re-inserting the zero-vector rows and columns

(ii) Mahalanobis Test Statistic

- When $M = J$ 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 $\widehat{S}(i)$
- Mahalanobis Test Statistic:

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

- If $\lambda_2 = 0$, under $H_0(i)$, asymptotically $Z_M(i)$ (with $M = J$) has an F distribution with $p^*(i)$ and $n - p^*(i) - 1$ degrees of freedom (see Johnson and Wichern, 2002, page 285)

HIV VACCINE TRIALS NETWORK

(ii) Mahalanobis Test Statistic

- Similarly to $Z_E(i)$, the diagonal matrix $\lambda_2 diag(\underline{1}_{nz})$ 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)
- Potential advantage of $Z_M(i)$ compared to the Euclidean statistic $Z_E(i)$: It accounts for the correlation structure of the multinomial response vectors
  - Potentially increases statistical power

HIV VACCINE TRIALS NETWORK

(ii) Kullback-Leibler Distance

- For random variables $X \sim f$ and $Y \sim g$, the "Kullback-Leibler distance" between f and g is

$$KL(f,g) = E_X \left\{ log\left(\frac{f(X)}{g(Y)}\right)\right\} = \sum_{i=1}^{n} log\left(\frac{f(X_i)}{g(Y_i)}\right) f(X_i)$$

- Properties of Kullback-Leibler distance
  - $KL(f,g) \geq 0$, and equals 0 if and only if $f = g$
  - $KL(f,g)$ is a log-likelihood ratio, and has optimality properties associated with likelihood ratio tests and MLEs
  - $KL(f,g) \neq KL(g,f)$, i.e., not symmetric, so that KL is not a distance ("discrepency" is more accurate)

(ii) Kullback-Leibler Distance [Extend Wu, Hsieh, and Li (2001, Biometrics)]

- For position i, let $Z_{KL}(i) = \sum_{a=1}^{21} M(a, r(i)) \widehat{p}_1(i,a) \times$

$$\log \left\{ I(\widehat{p}_2(i,a) > 0) \frac{\widehat{p}_1(i,a)}{\widehat{p}_2(i,a)} + I(\widehat{p}_2(i,a) = 0) \frac{(\widehat{p}_1(i,a) + n_1^{-1})}{n_2} \right\}$$

- $I(\widehat{p}_2(i,a) > 0)$ prevents the statistic $Z_{KL}(i)$ from taking infinite value
- With $I(\widehat{p}_2(i,a) > 0)$ replaced with 1 and the second term removed, and $M = J$, $Z_{KL}(i) =$ Kullback-Leibler discrepency between the 21-nomial empirical densities $\widehat{p}_1(i)$ and $\widehat{p}_2(i)$

HIV VACCINE TRIALS NETWORK

(iii) Obtaining Position-Wise p-values

- Compute nominal position-wise p-values by a standard permutation method
  - B data sets of $n = n_1 + n_2$ sequences each are generated by independently permuting group membership indices on whole sequences within Groups 1 and 2
  - For each position i compute test statistics on permuted data
  - 2-sided p-values computed as empirical rejection fractions

(iii) Obtaining Position-Wise p-values

- Alternative approach for Euclidean and Mahalanobis statistics
  - Follow Pan's (2003) idea to directly nonparametrically estimate the null distribution of hundreds of t-statistics
  - Assume that under H_0, the statistics for all of the positions have the same distribution
    - A *pooling* approach
  - Can apply weights to the positions: Upweighting biologically important positions based on prior knowledge can increase statistical power

HIV VACCINE TRIALS NETWORK

(iii) Weighting Positions

- Some possible reasons for upweighting a position
  - It is within a known antibody epitope
  - It is under diversifying selection
  - It covarys with a position known to be critical
- Certain gp120 positions have been found to be important for:
  - Antibody binding and neutralization (Wyatt et al., 1998; Wei et al., 2003)
  - Key steps of HIV entry into host T cells such as CD4 co-receptor binding (Wyatt et al., 1998)
  - Evasion of host immune responses, e.g., through an evolving glycan shield (Wei et al., 2003)

HIV VACCINE TRIALS NETWORK

(iii) Weighting Positions

- Wyatt et al. (1998)
  - 6 positions contained in a neutralization epitope defined by the monoclonal antibody 2G12 [295, 297, 334, 386, 392, 397]
  - 20 CD4-binding positions [88, 113, 117, 256, 257, 262, 266, 368, 370, 384, 421, 427, 457, 470, 474, 475, 477, 482, 483, 484]
  - 19 CD4-induced epitope positions [88, 117, 121, 207, 256, 257, 262, 370, 381, 382, 419, 420, 421, 422, 423, 427, 435, 438, 475]
- Wei et al. (2003)
  - 3 positions at which changes can sterically inhibit the accessibility of principal neutralizing epitopes on the virus surface [201, 240, 268]
- 39 total positions to possibly upweight

(iii) Pan's (2003) Approach for Euclidean Statistics

- Modify (slightly) the test statistic $Z_E(i)$

$$Z_E^{split}(i) = w_1(i)C_E(i) \times$$
$$\sum_{a=1}^{21} \frac{\left\{M(a,r(i))\left[\frac{\widehat{p}_{11}(i,a)+\widehat{p}_{12}(i,a)}{2} - \frac{\widehat{p}_{21}(i,a)+\widehat{p}_{22}(i,a)}{2}\right]\right\}^2}{\{\widehat{v}(i,a)+\lambda_1\}^2} I(\widehat{v}(i,a) > 0)$$

  - $\widehat{p}_{k1}(i,a)$ averages the $Y_{kj}(\cdot)$ in the first permuted half of sample k
  - $\widehat{p}_{k2}(i,a)$ averages the $Y_{kj}(\cdot)$ in the second permuted half

(iii) Pan's (2003) Approach for Euclidean Statistics

- A statistic $z_E^{split}(i)$ can be used to estimate the null distribution of $Z_E^{split}(i)$:

$$z_E^{split}(i) = w_1(i) C_E(i) \times$$
$$\sum_{a=1}^{21} \frac{\left\{ M(a, r(i)) \left[\frac{\hat{p}_{11}(i,a) - \hat{p}_{12}(i,a)}{2} + \frac{\hat{p}_{21}(i,a) - \hat{p}_{22}(i,a)}{2} \right] \right\}^2}{\{\hat{v}(i,a) + \lambda_1\}^2} I(\hat{v}(i,a) > 0)$$

(iii) Pan's (2003) Approach for Euclidean Statistics

- 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 $Z_E^{split}(i)$

(iii) Pan's (2003) Approach for Euclidean Statistics

- To obtain p-values, once $Z_E^{split}(i)$ is computed, each sample is again separately randomly permuted into two halves, and the statistic $z_E^{split}(i)$ is computed
- Based on B separate permutations the statistic $z_E^{split(b)}(i)$ is computed B times, $b = 1, \cdots, B$
- For position i the 2-sided p-value is then obtained as $p_i = N_i/(B * p)$, 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 $b = 1, \cdots, B$

(iii) Pan's (2003) Approach for Mahalanobis Statistics

- Slightly modified version of $Z_M(i)$:

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

where $D^{2split}(i) =$

$$\left\{\frac{\widehat{p}_{11}(i)+\widehat{p}_{12}(i)}{2}-\frac{\widehat{p}_{21}(i)+\widehat{p}_{22}(i)}{2}\right\}' diag(MY_r(i))\widehat{S}^{-}_{\lambda_2}(i)$$

$$\times diag(MY_r(i))\left\{\frac{\widehat{p}_{11}(i)+\widehat{p}_{12}(i)}{2}-\frac{\widehat{p}_{21}(i)+\widehat{p}_{22}(i)}{2}\right\}$$

- Apply a similar data-splitting procedure as used for the Euclidean statistics

(iv) Multiple Hypothesis Testing Adjustment

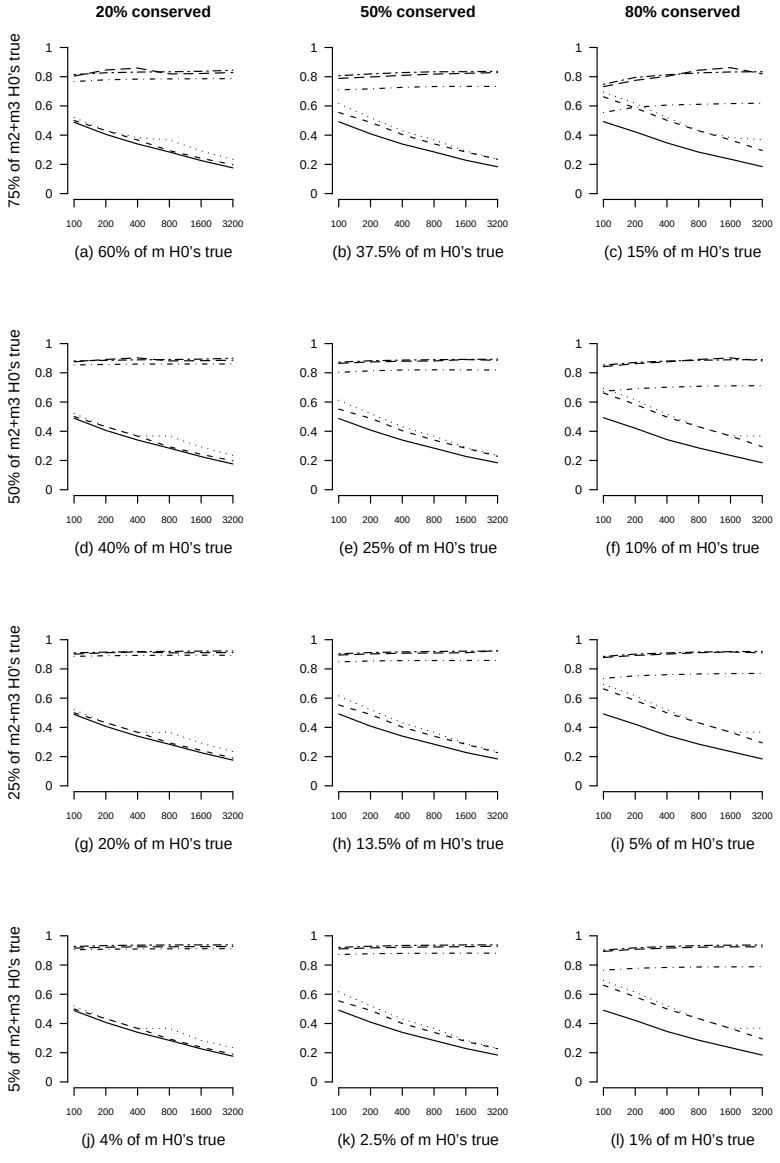
- Four different multiple comparisons adjustment procedures to determine the set of significant signature positions:
  - Standard Bonferroni
  - Tarone's (1990) modified Bonferroni method for discrete data
  - Standard FDR (Benjamini and Hochberg, 1995)
  - Tarone-modified FDR for discrete data (Gilbert, 2004)
    - For conserved regions can increase power 10-50%
    - For diverse regions absent-to-slight power gains

HIV VACCINE TRIALS NETWORK

(iv) Tarone-modified FDR

- **Step 1**: Screen out the conserved positions, for which it would not be possible to reject $H_0(i)$, based on the minimum achievable significance level α_i^*
  - Let $m(k)$ be the number of positions for which $\alpha_i^* < \alpha/k$
  - K is the smallest value of k such that $m(k) \leq k$, and R_K is the set of indices satisyfing $\alpha_i^* < \alpha/K$
- **Step 2**: Perform Benjamini and Hochberg's (1995) FDR procedure at level α on the subset of hypotheses R_K

Power Gains from Tarone-FDR Method



HIV VACCINE TRIALS NETWORK

Simulation Study

- Questions addressed:
  - How much power is there to detect signature positions for vaccine efficacy trials of different sizes?
  - What is the impact of the proportion of positions with a true alternative hypothesis on the performance of the procedures?
  - What is the influence of the small positive constants λ_1 and λ_2 in the denominators of the Euclidean- and Mahalanobis-based test statistics?

HIV VACCINE TRIALS NETWORK

Simulation Study Set-Up

- Simulations based on the VAX004 gp120 sequence data
  - $p = 581$ positions
- **Infected placebo group**: gp120 sequences simulated by randomly sampling with replacement $n_2 = 90$ or 180 whole sequences from the 336 Vax004 sequences
- **Infected vaccine group**: Assuming $VE = 50\%$, gp120 sequences simulated by sampling with replacement $n_1 = 45$ or 90 whole sequences from the 336 sequences

HIV VACCINE TRIALS NETWORK

Simulation Study Set-Up

- To create an alternative hypothesis at a position i, the HIV-1B-specific PAM matrix developed by Nickle et al. (2005) is used to induce stochastic evolution of the vaccinees' amino acids at position i
- Entries of the PAM matrix are probabilities that one amino acid mutates into another during a certain amount of evolutionary time
  - A PAM-25 matrix is used, which specifies a total of 25 amino acid interchanges per 100 positions
    - At each true alternative position, on average 1/4 of the vaccinee sequences have a mutation

HIV-specific PAM matrix with all probabilities multiplied by 100

	Ala	Cys	Asp	Glu	Phe	Gly	His	Ileu	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
Ala	99																			
Cys	0.01	99.9																		
Asp	0.03	0	98.7																	
Glu	0.14	0	0.46	98.8																
Phe	0	0.01	0	0	99.6															
Gly	0.11	0.01	0.1	0.53	0	99.3														
His	0.01	0.01	0	0	0	0	98.8													
Ileu	0.01	0	0.01	0.03	0.09	0.01	0.02	98.9												
Lys	0.03	0	0.09	0.59	0.01	0.11	0	0.05	98.3											
Leu	0.01	0	0	0	0.26	0.01	0.08	0.44	0.02	99.4										
Met	0.01	0	0.01	0.02	0	0.03	0	0.61	0.09	0.36	98.7									
Asn	0.02	0	1.16	0.07	0	0.04	0.25	0.05	0.52	0.01	0	98.8								
Pro	0.03	0	0	0	0	0	0.13	0	0.01	0.11	0	0.02	99.6							
Gln	0.02	0	0.02	0.2	0	0.03	0.53	0.01	0.6	0.06	0.01	0.1	0.2	98.9						
Arg	0.04	0.01	0.01	0.03	0	0.28	0.1	0.09	1.32	0.01	0.17	0.08	0.01	0.28	98.5					
Ser	0.13	0.05	0.06	0.01	0.05	0.12	0.03	0.08	0.1	0.05	0.03	0.5	0.14	0	0.28	98.8				
Thr	0.58	0	0.05	0.06	0	0.04	0.07	0.41	0.33	0.02	0.24	0.25	0.09	0.04	0.13	0.54	98.8			
Val	0.49	0	0.03	0.16	0.06	0.04	0.02	0.83	0.04	0.07	0.18	0.01	0.01	0	0.02	0.01	0.1	99.1		
Trp	0	0.04	0	0	0.06	0.13	0.02	0	0	0.13	0.01	0	0.01	0	0.05	0	0	0.01	99.7	
Tyr	0	0.16	0.01	0.01	0.25	0	0.43	0.01	0.01	0.02	0	0.14	0.01	0.02	0.01	0.04	0	0.01	0.03	99.4

HIV VACCINE TRIALS NETWORK

Simulation Study Set-Up

- Question 1) was addressed by carrying out the simulation experiment for the two sample sizes
  - $n_1/n_2 = 45/90$: Small Phase 3 trial
  - $n_1/n_2 = 90/180$: Large Phase 3 trial

HIV VACCINE TRIALS NETWORK

Simulation Study Set-Up

- Question 2) 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 (Wyatt et al., 1998; Wei et al., 2003)
  - 6 alternative positions: Those constituting the monoclonal antibody 2G12 neutralization epitope (295, 297, 334, 386, 392, 397)
  - 58 alternative positions: The 39 plus 19 randomly sampled positions
  - 145 alternative positions: Same as the 58 plus 87 more randomly sampled positions

Simulation Study Set-Up

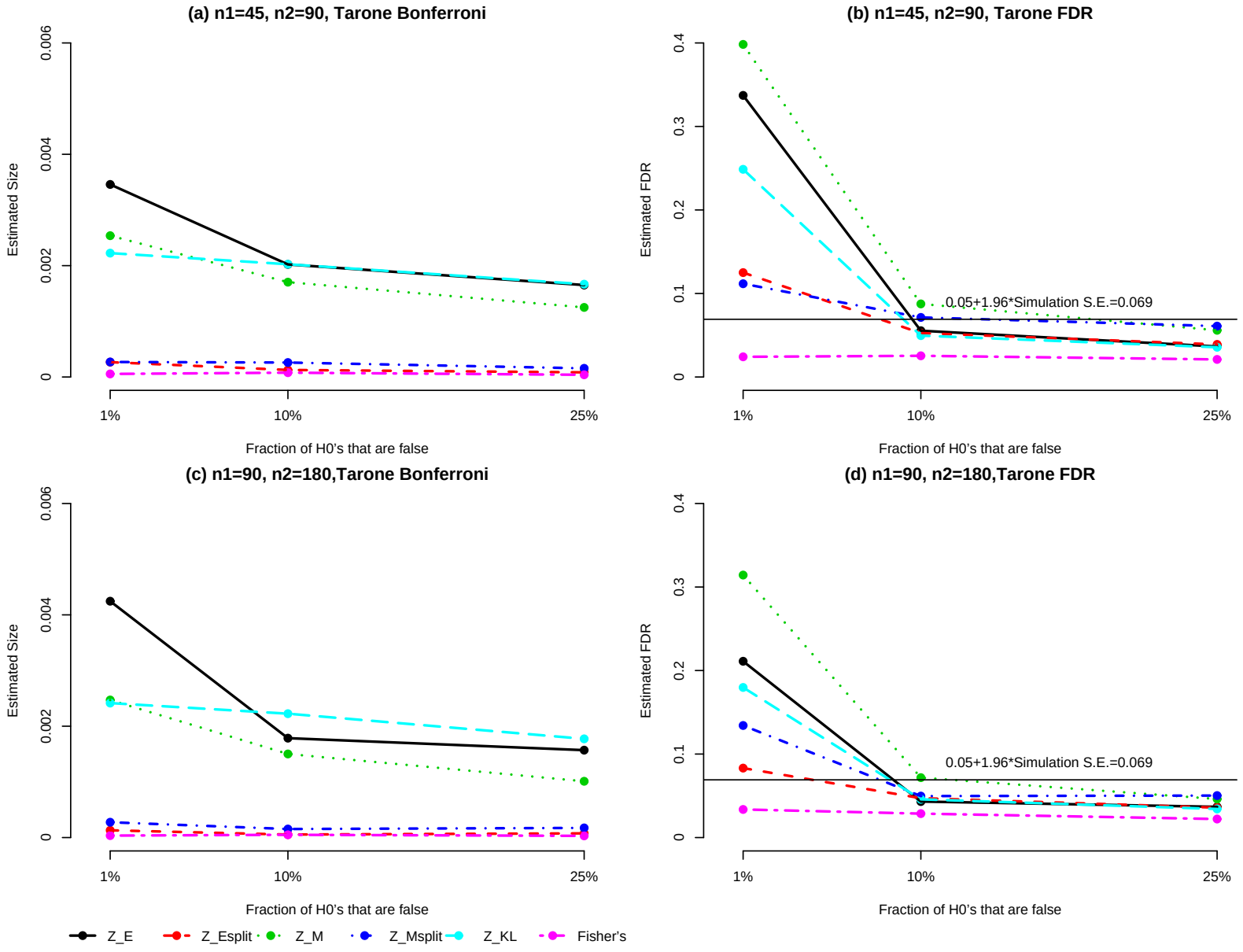
- Question 3) was addressed by repeating the simulations setting λ_1 and λ_2 in the denominators of $Z_E, Z_M, Z_E^{split}, Z_M^{split}$ equal to 0

HIV VACCINE TRIALS NETWORK

Simulation Study Set-Up

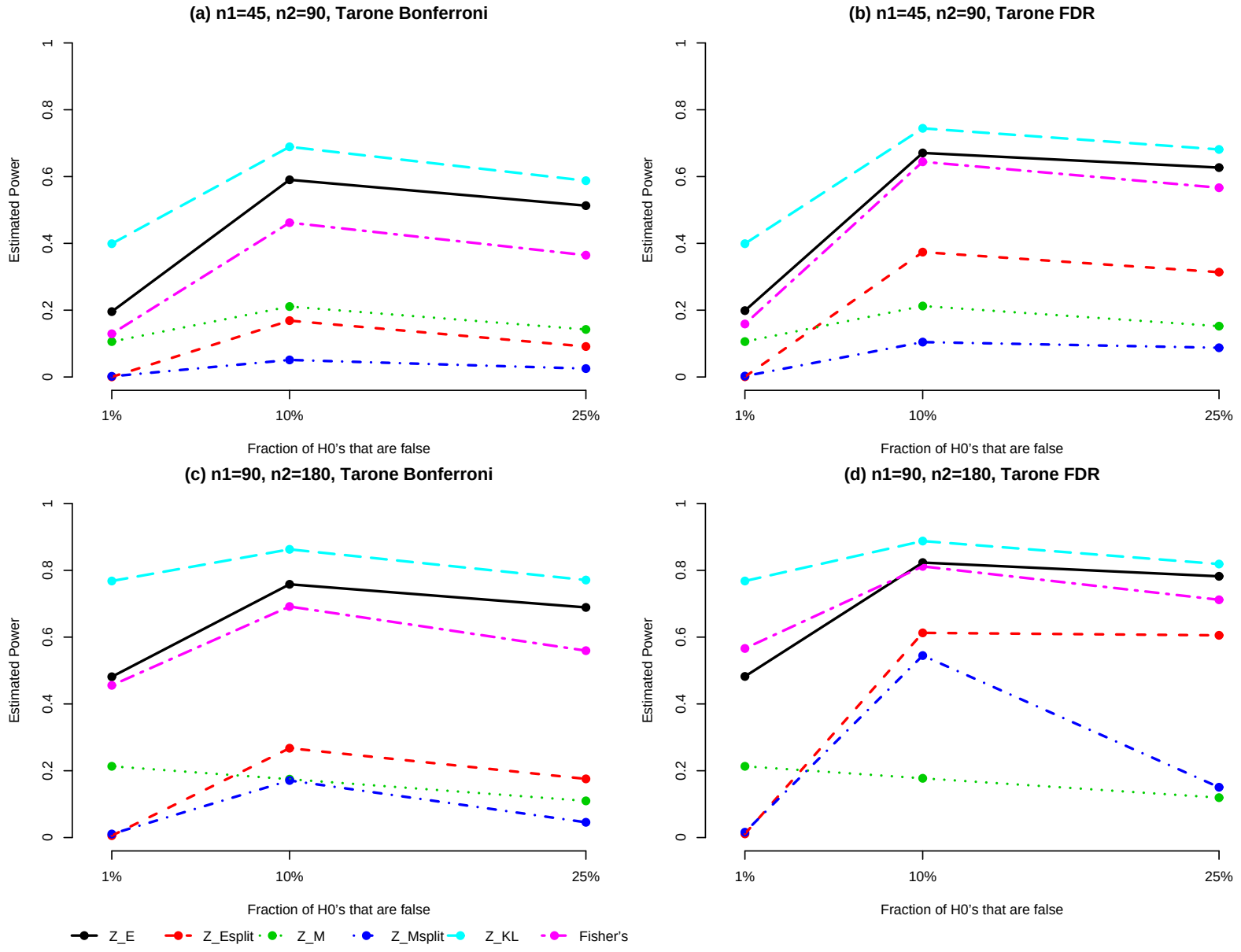
- For each of 500 simulated trials, all of the developed tests, plus Fisher's exact test for comparison, were carried out
  - at 2-sided level $\alpha = 0.05$
  - using 250 permutations to compute p-values
- Empirical false positive rates, false discovery rates, and powers of the testing procedures were computed
  - Tarone Bonferroni
  - Tarone FDR

Simulation Results: False Positive Rates

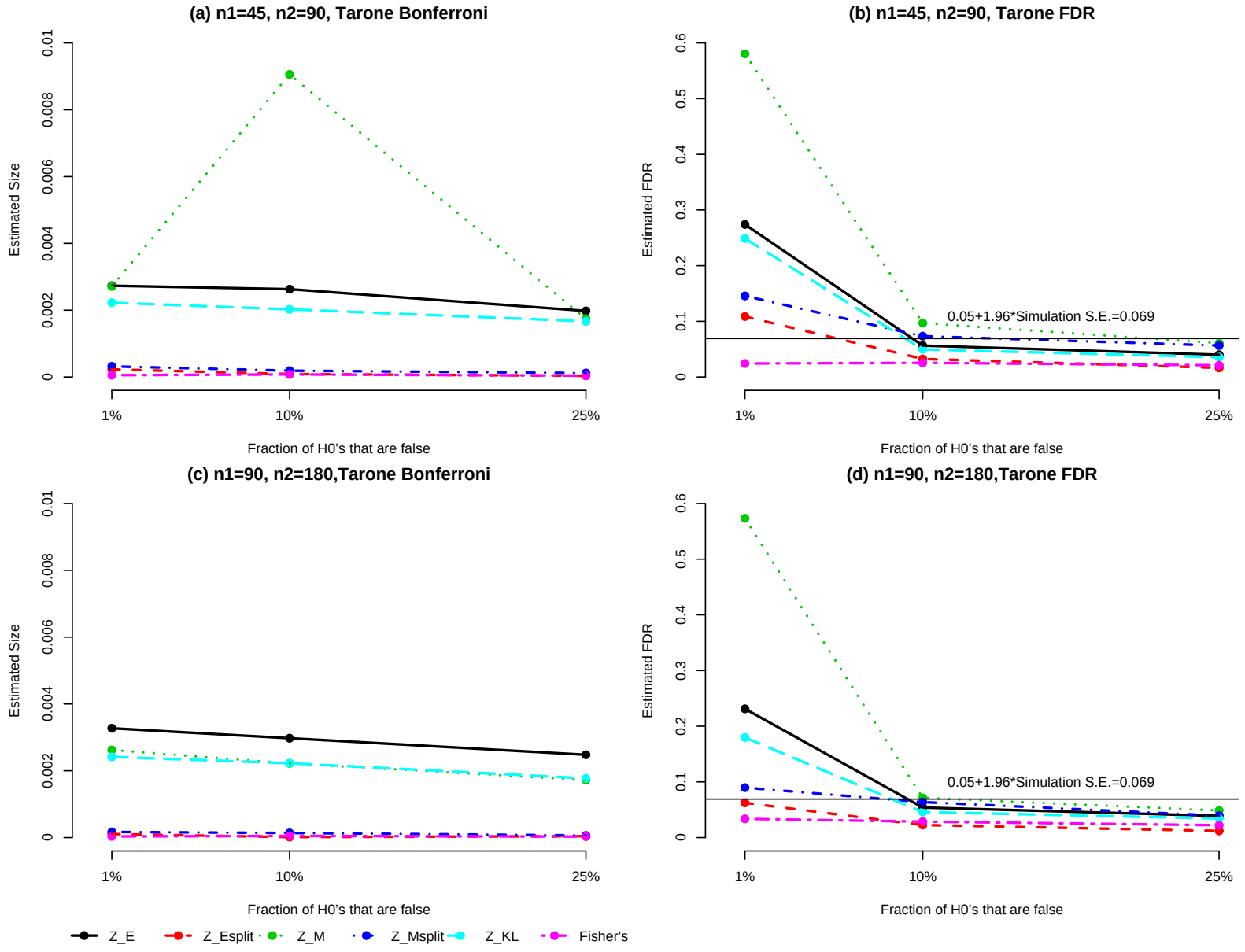


HIV VACCINE TRIALS NETWORK

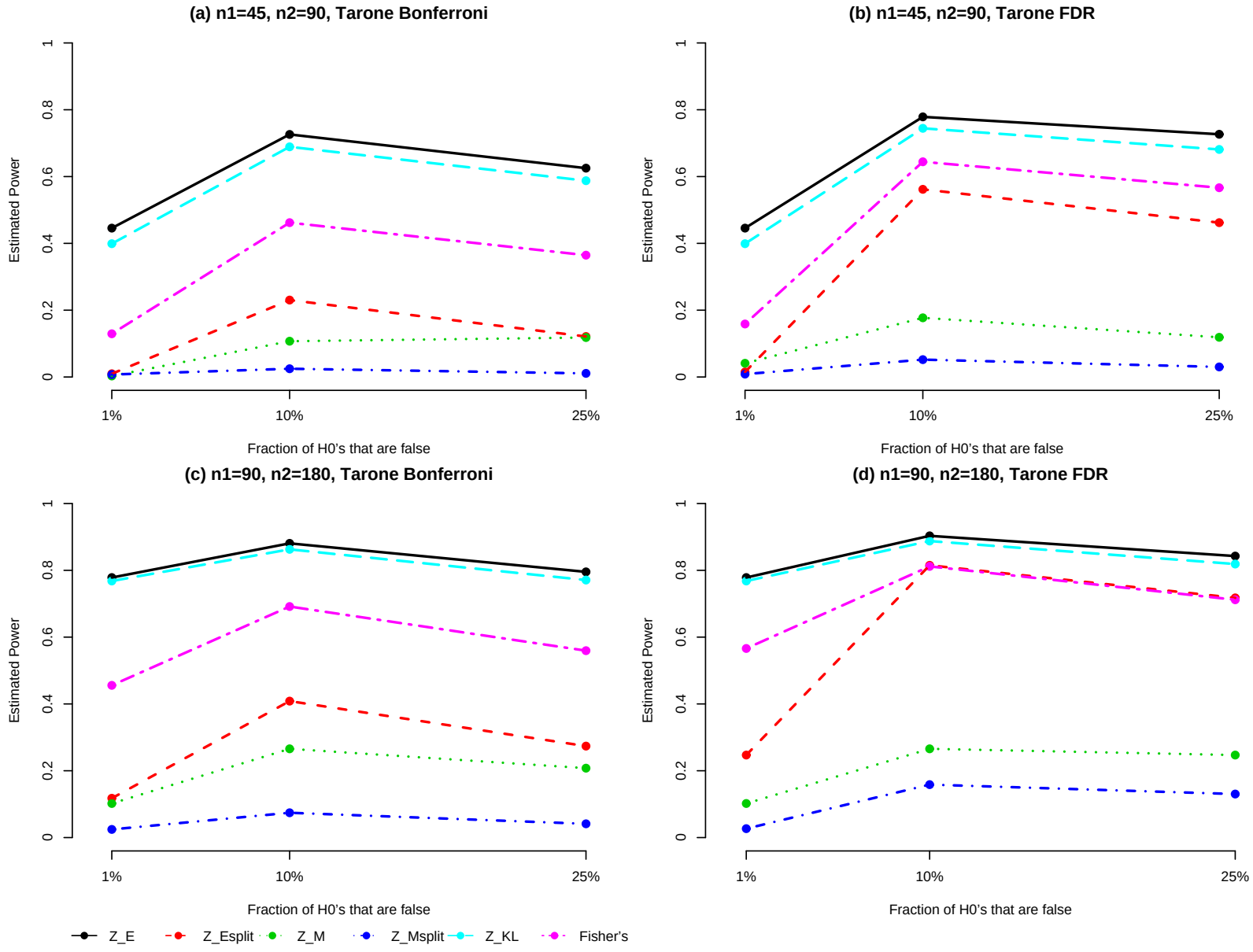
Simulation Results: Power



Simulation Results with $\lambda_1 = \lambda_2 = 0$: False Positive Rates



Simulation Results with $\lambda_1 = \lambda_2 = 0$: Power



HIV VACCINE TRIALS NETWORK

Examples

- In all examples 10,000 permutations were used to approximate p-values
- The Bonferroni, Tarone Bonferroni, FDR, and Tarone FDR procedures were conducted with $\alpha = 0.05$

Example 1: VAX004 Trial

- First preventive HIV vaccine efficacy trial completed in February 2003
- Tested vaccine AIDSVAX, a recombinant gp120 vaccine based on two patient isolates [MN and GNE8]
- Trial conducted in U.S./Netherlands/Canada/Carribean, $n = 5403$, 2:1 randomization to vaccine:placebo
- Volunteers tested for HIV infection every 6 months for 36 months
- For HIV infected subjects, the gp120 region of HIV was sequenced

VAX004: No Vaccine Efficacy To Prevent HIV Infection

- **Primary analysis:**

	Number Randomized	Number Infected	Percent Infected
Vaccine	3598	241	6.7%
Placebo	1805	127	7.0%

$$\widehat{VE} = 5.7\%, \quad 95\% \ CI \ \ (-17.0\%, 24.0\%), p = 0.59$$

HIV VACCINE TRIALS NETWORK

Example 1: VAX004 Trial

- The 336 infecting HIV gp120 sequences were aligned together with the two gp120 sequences that were represented in the vaccine construct (MN and GNE8)
- GNE8 was used as the reference sequence because sampled more recently
- $n_1 = 217$ vaccine sequences and $n_2 = 119$ placebo sequences, each of length $p = 581$

HIV VACCINE TRIALS NETWORK

Example 1: VAX004 Trial

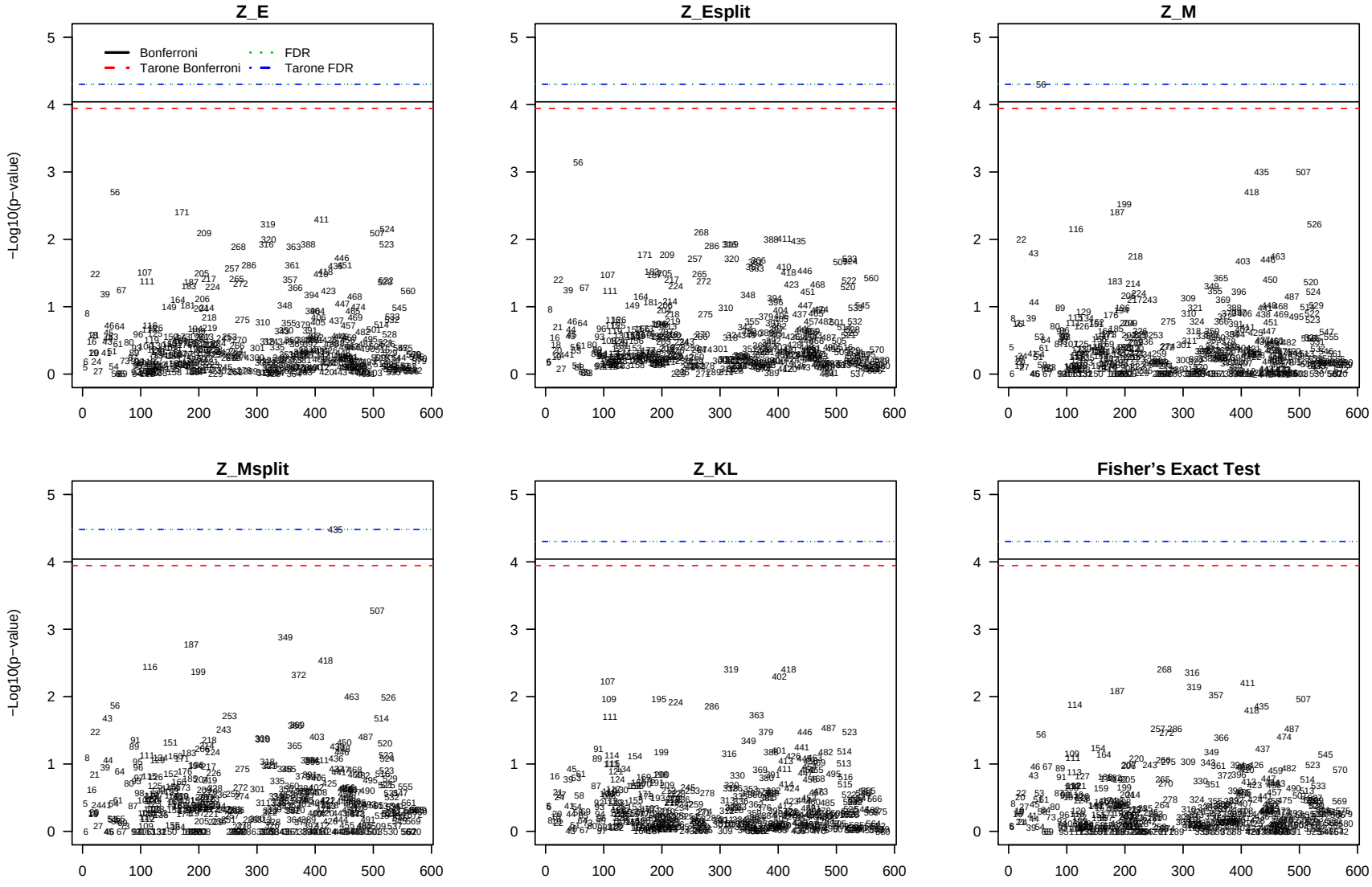
- Set $M = J - I$, so that all amino acid mismatches with the reference sequence are weighted equally
- Of the 581 positions, 348 have enough diversity (by the Tarone screen) to evaluate

VAX004: Equal-weight AA Substitution Matrix

HIV VACCINE TRIALS NETWORK

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

Equal Weight Matrix

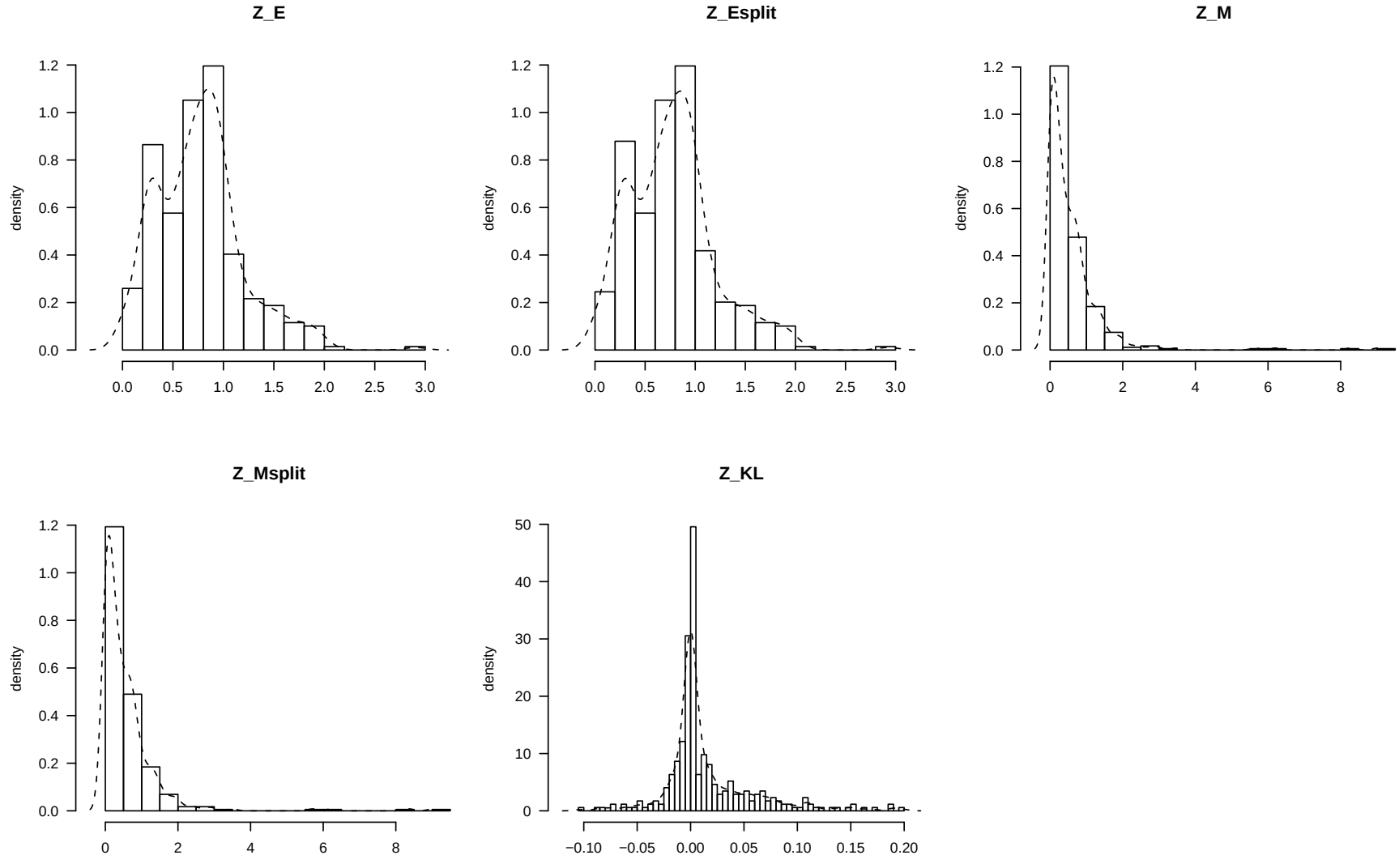


Note: Value of 0 is replaced with 0.00005; −log10(0.00005)=4.3

VAX004: Equal-weight AA Substitution Matrix

Histograms of Test Statistics

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



HIV VACCINE TRIALS NETWORK

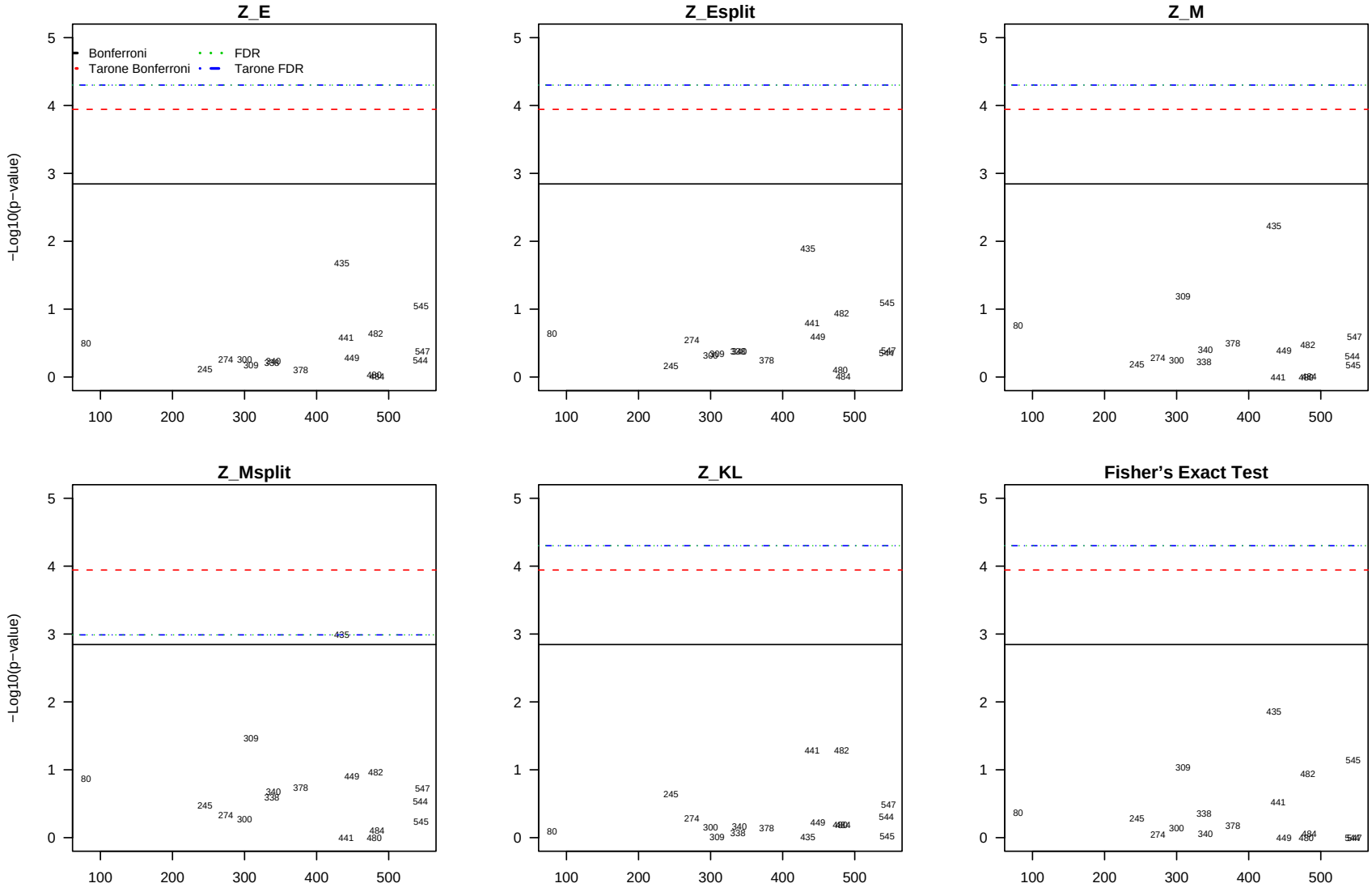
Example 1: VAX004 Trial

- Repeat the analysis for the 39 key positions
- Of these, 17 have enough diversity (by the Tarone screen) to evaluate

VAX004: Equal-weight AA Substitution Matrix; Key Positions

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

39 key positions

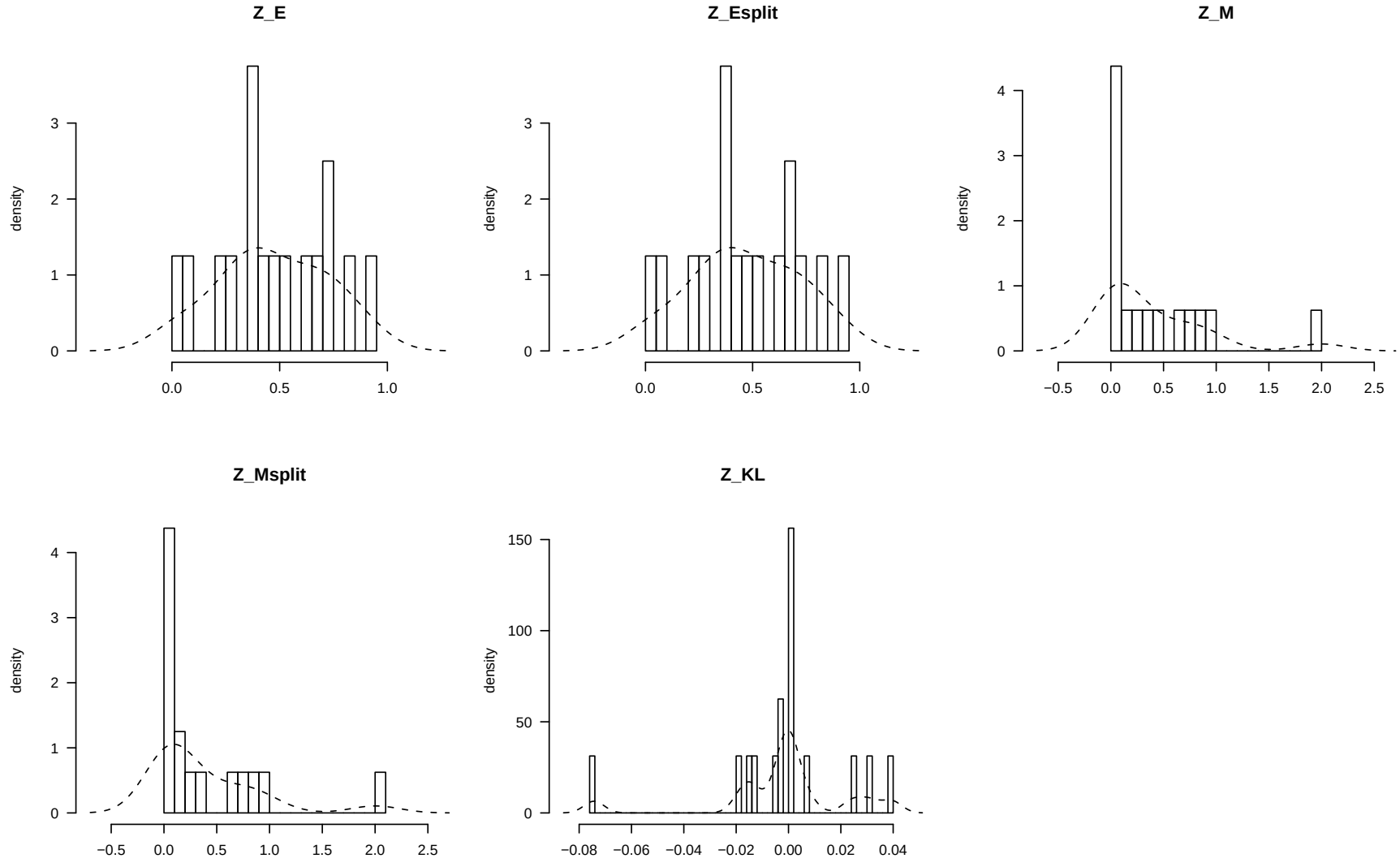


Note: Value of 0 is replaced with 0.00005; −log10(0.00005)=4.3

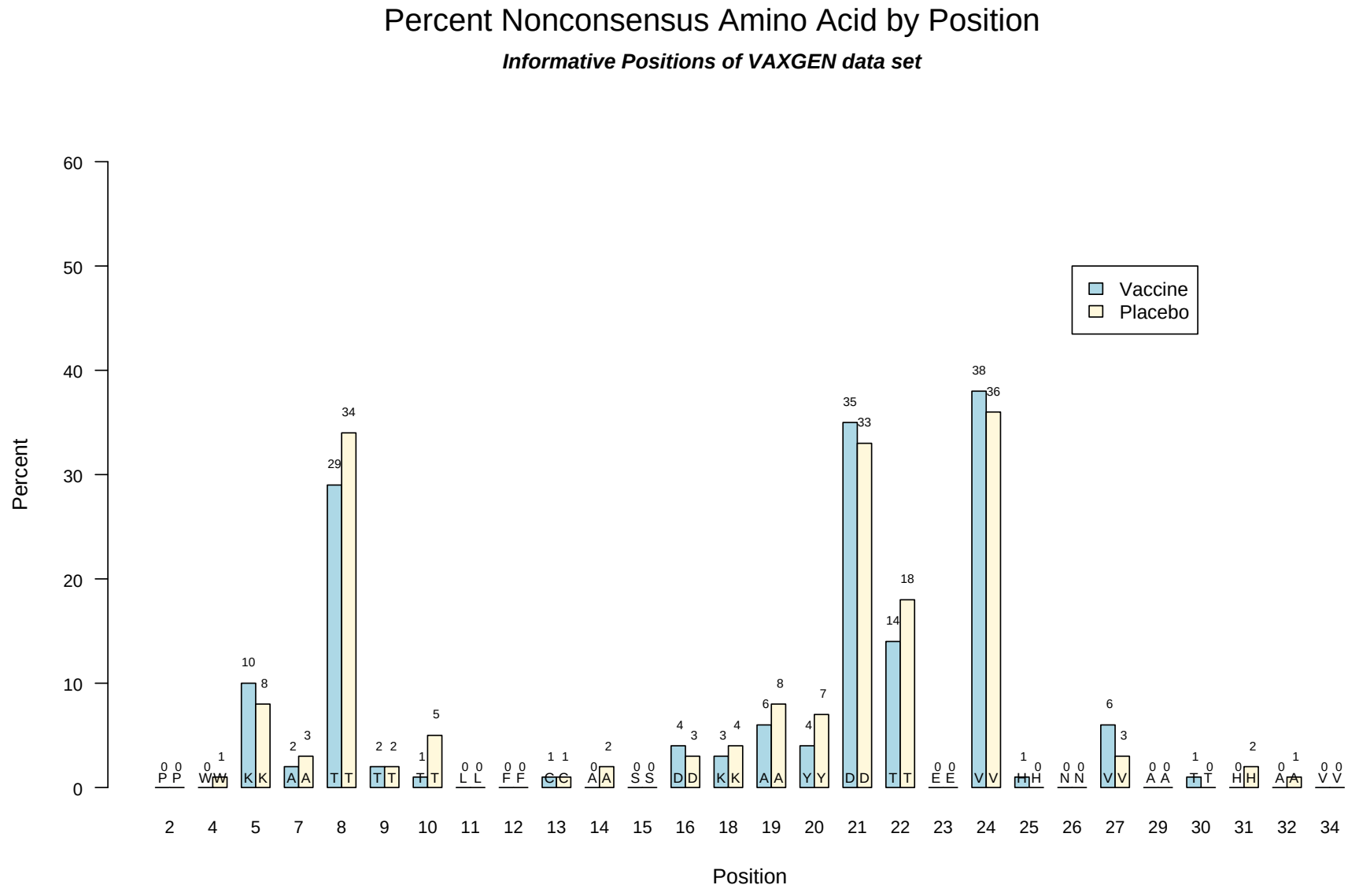
VAX004: Equal-weight AA Substitution Matrix; Key Positions

Histograms of Test Statistics

VAXGEN gp120 vaccine and placebo recipient data set, p=39 key positions



VAX004: Equal-weight AA Substitution Matrix; Key Positions



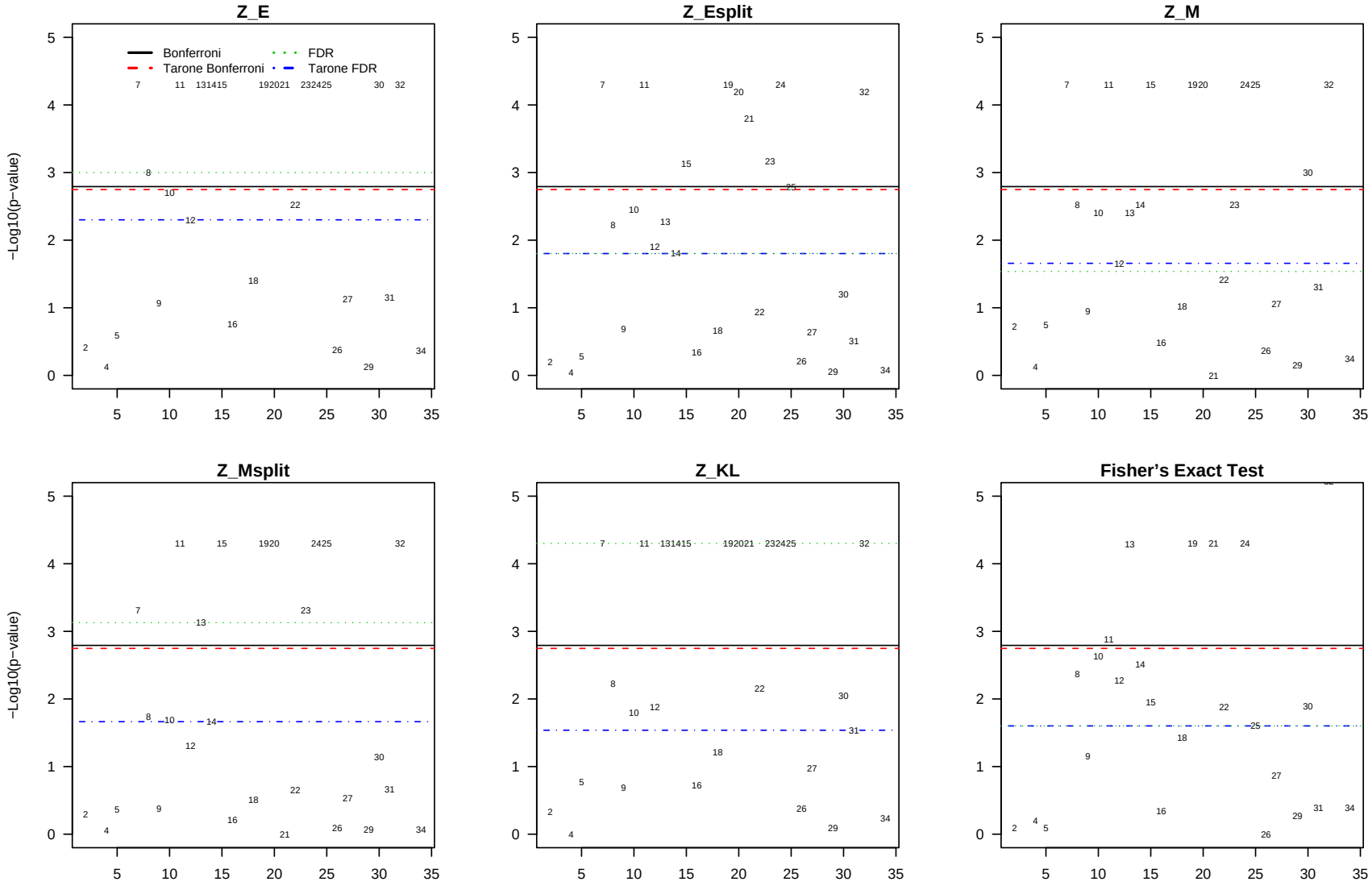
The number in each bar denotes the percent noncnsensus.

Example 2: HIV-1B X4 versus R5 Viruses

- The CD4 co-receptor usage phenotypes of HIV (X4, R5) are distinguishable by V3 loop amino acid sequences
- Algorithms are published to predict phenotype based on V3 loop amino acid sequence
- Fusheng downloaded a dataset from the Los Alamos database of $n_1 = 56$ X4 and $n_2 = 176$ R5 viruses
- The genome scanning methods are applied with $M = J$ (test for differential amino acid frequencies)
- Of 35 positions, 28 have enough diversity (by the Tarone screen) to evaluate

$n_1 = 56$ ***X4 versus*** $n_2 = 176$ ***R5 V3 Loop AA Sequences***

−log10 p−value for comparing X4 and R5 virus sequences



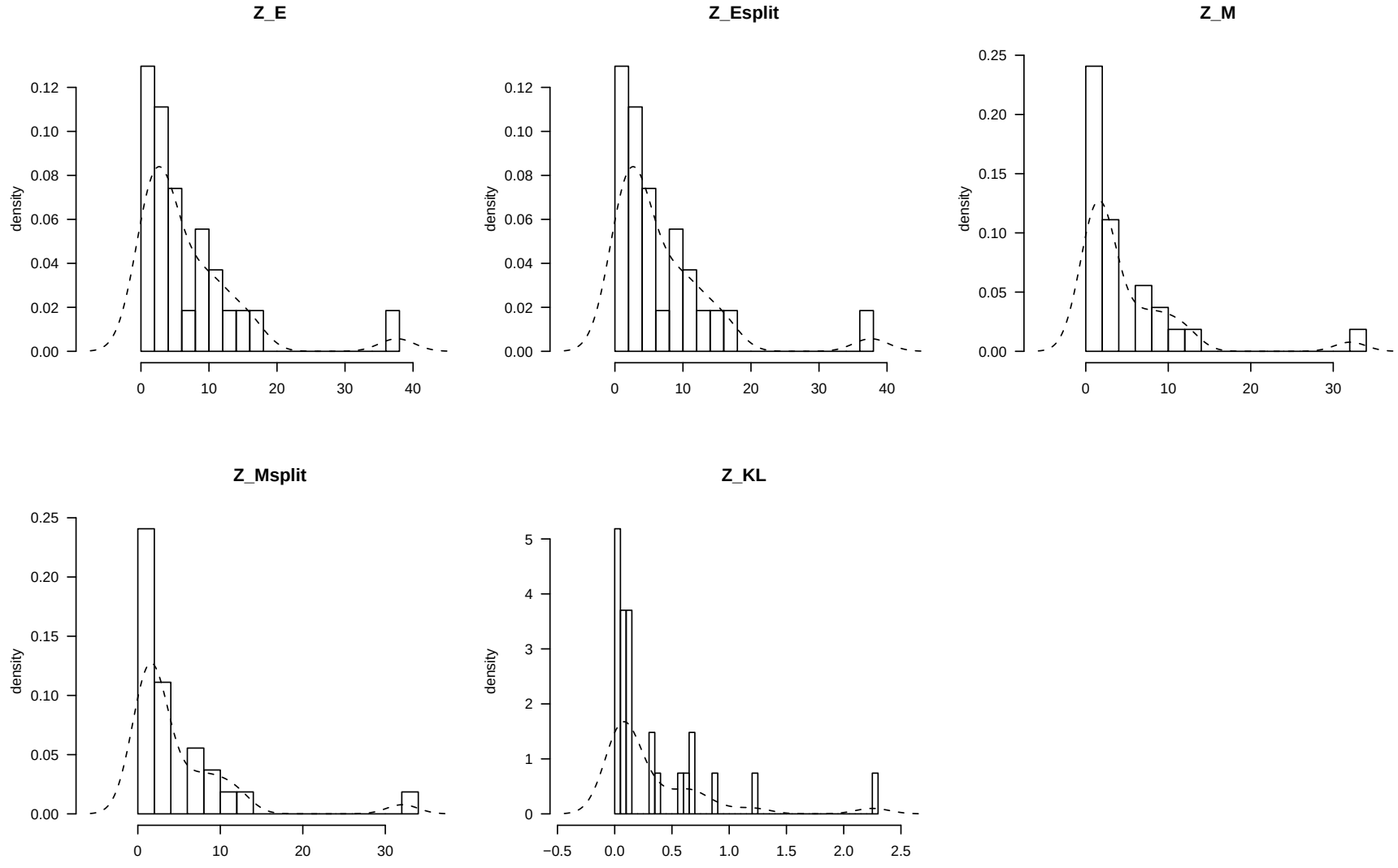
Note: Value of 0 is replaced with 0.00005; −log10(0.00005)=4.3

$n_1 = 56$ ***X4 versus*** $n_2 = 176$ ***R5 V3 Loop HIV-1B AA Sequences***

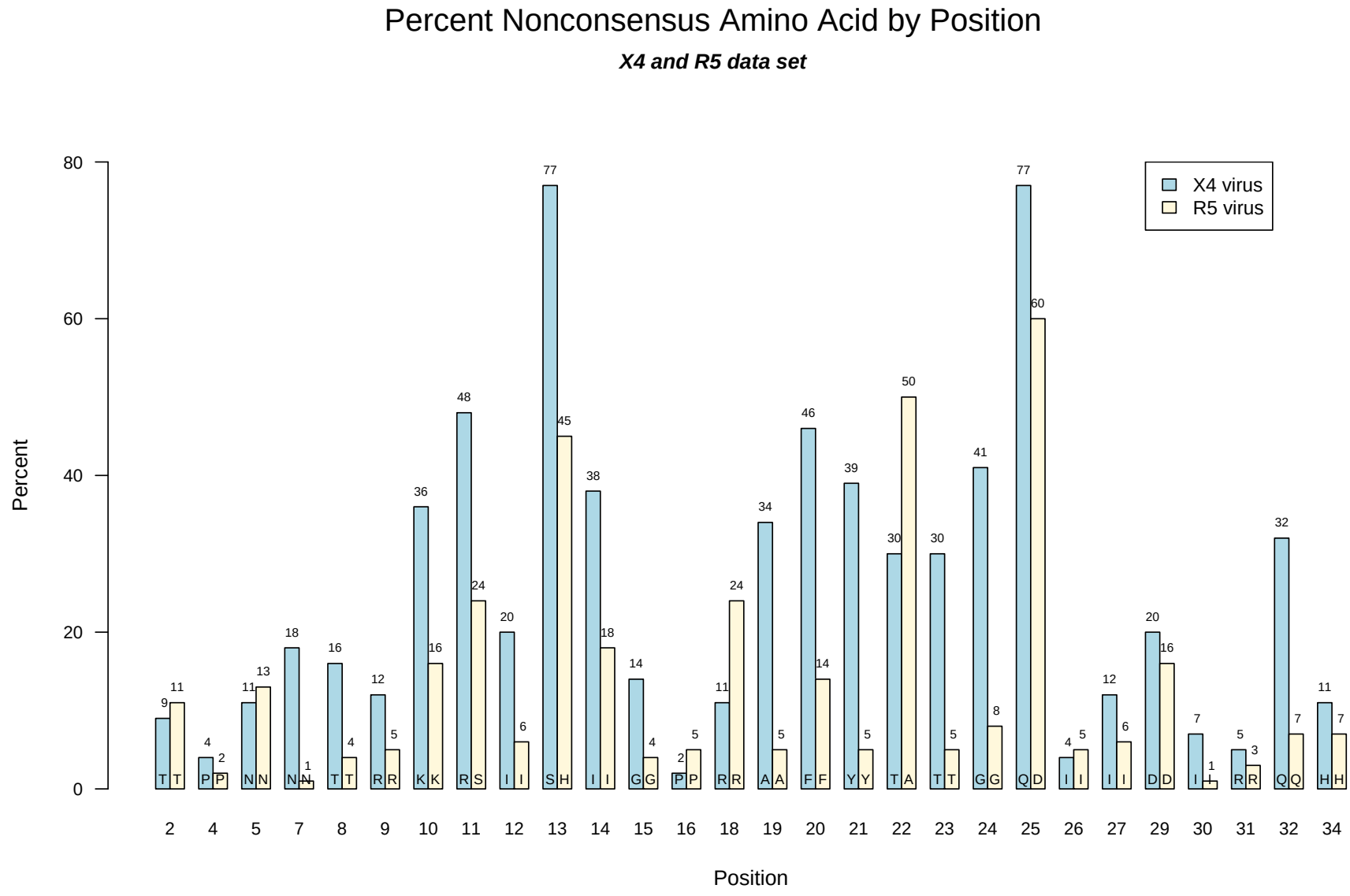
HIV VACCINE TRIALS NETWORK

Histograms of Test Statistics

X4 and R5 virus sequences



$n_1 = 56$ ***X4 versus*** $n_2 = 176$ ***R5 V3 Loop HIV-1B AA Sequences***



The number in each bar denotes the percent noncnsensus.

Example 3: CTL Non-responders versus CTL Responders

- Botswana-Harvard Patnership studied CD8+ T cell ELISpot responses of HIV infected blood donors in Botswana
- Targets were overlapping 15-mer CONSENSUS HIV-1C peptides across the entire HIV genome
- CTL response to Gag p24 (summed over peptides) was inversely correlated with RNA plasma and DNA pro viral load

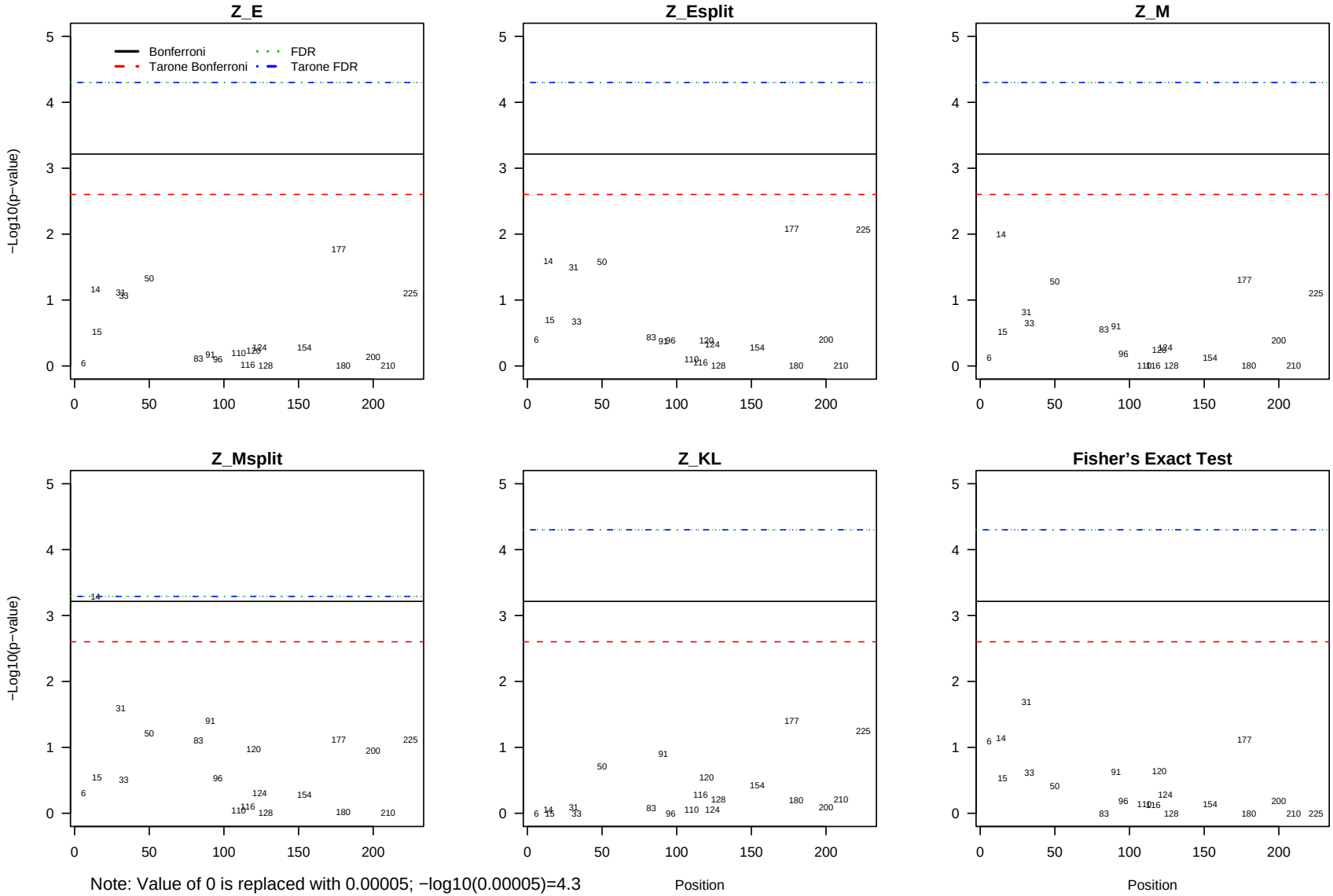
Example 3: CTL Non-responders versus CTL Responders

- Relative to the consensus sequence, are there signature positions in Gag p24 amino acids that distinguish the $n_1 = 17$ subjects who had no response to Gag p24 versus the $n_2 = 34$ subjects who responded to Gag p24?
- Of 231 positions, 20 have enough diversity (by the Tarone screen) to evaluate
  - Note that Tarone FDR provides substantial power gains here

CTL Non-responders vs Responders

HIV-1C Gag p24 AA Sequences

−log10 p−value for comparing p24 responder and non−responder sequences



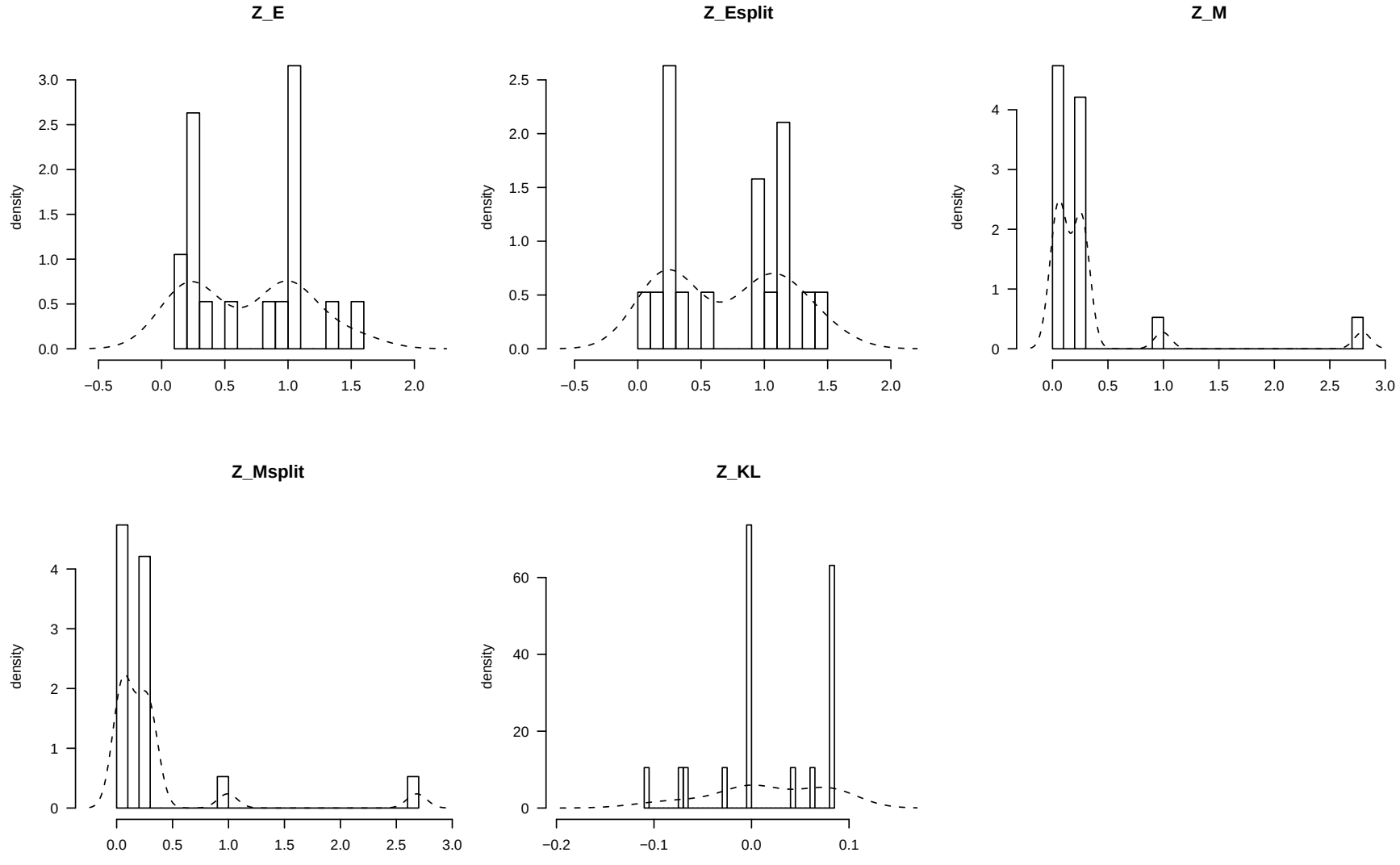
Note: Value of 0 is replaced with 0.00005; −log10(0.00005)=4.3

CTL Non-responders vs Responders

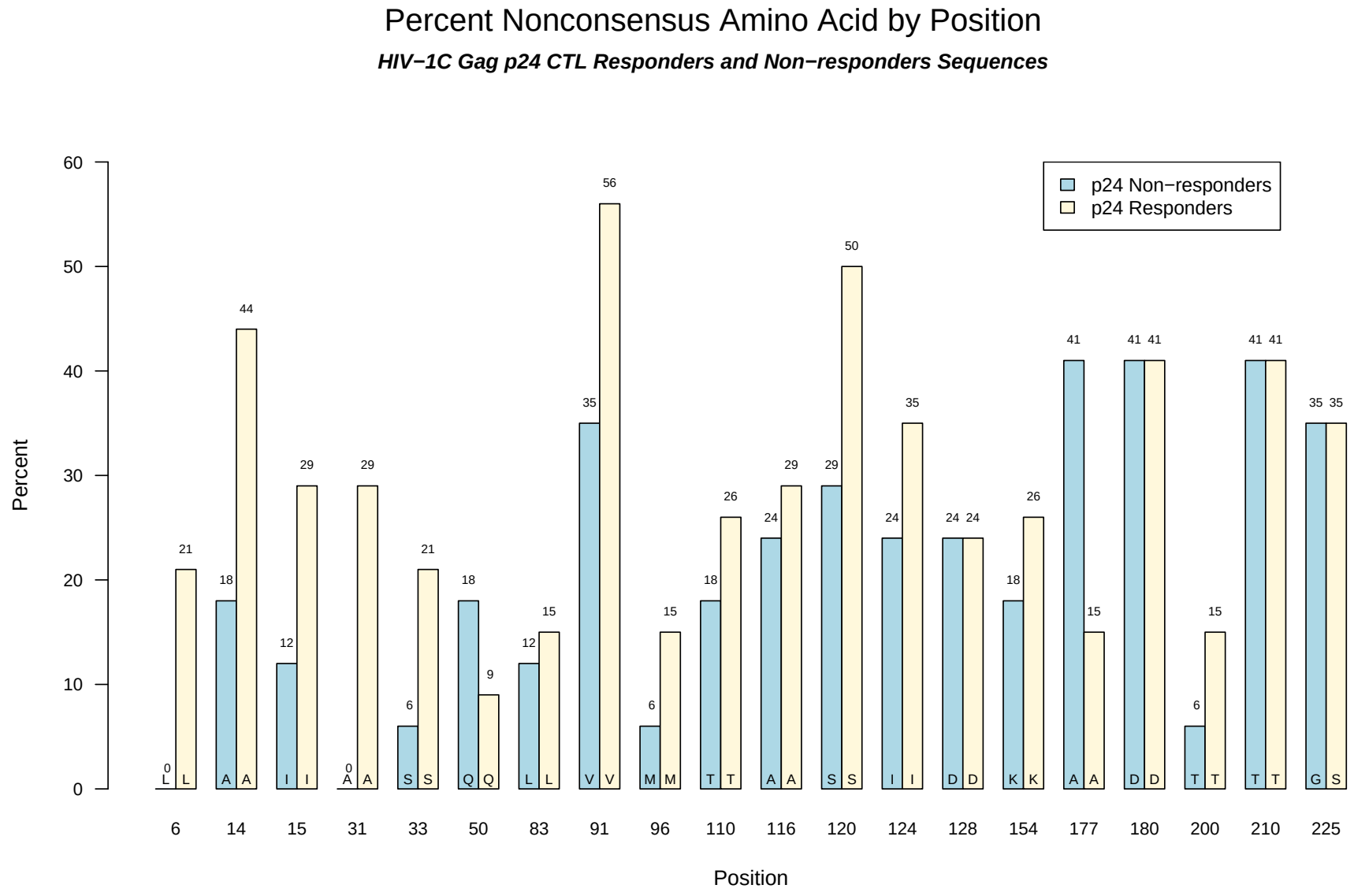
HIV-1C Gag p24 AA Sequences

Histograms of Test Statistics

p24 responder and non−responder sequences



CTL Non-responders vs Responders
HIV-1C Gag p24 AA Sequences



The number in each bar denotes the percent noncnsensus.

HIV VACCINE TRIALS NETWORK

Discussion

- Kullback-Leibler test and standardized Euclidean test (with $\lambda_1 = 0$) the most powerful
  - Both are computationally simple and fast
  - Both control the false positive rate (comparably), and have similar power
  - Recommend either method
- Recommend using a multiplicity adjustment method with the Tarone-modification
- In problems with few expected signature positions, Tarone-Bonferroni safer than Tarone-FDR for avoiding false positives

HIV VACCINE TRIALS NETWORK

Discussion

- Pooling methods appear less powerful than position-wise methods
  - May still be useful when up-weighting certain positions is justified
  - Additional simulations showed that:
    - Correctly upweighting positions that are true alternatives lowers false positive rates and increases power
    - Incorrectly upweighting positions that are true nulls raises false positive rates and decreases power

    $\Rightarrow$ **Weight positions with caution**

HIV VACCINE TRIALS NETWORK

Discussion

- Study design and sampling strongly affect interpretation of results
  - Complicated interpretation for epidemiological studies
    - Geographic differences vs biological differences
  - Clearer interpretation for randomized studies

HIV VACCINE TRIALS NETWORK

Discussion: Further Research

- Extend Kullback-Leibler method to peptide regions (Masters thesis of U of W student Allan DeCamp)
- Multiple sequences per subject (U-statistic approach)
- Continuous outcome (e.g., study relationship between AAs at positions and viral load)
- Include covariates (regularized regression such as threshold gradient descent regularization [TGDR])