University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2006 *Paper* 205

Empirical Bayes Approach to Controlling Familywise Error: An Application to HIV Resistance Data

Rhoderick N. Machekano* Alan E. Hubbard†

*Division of Biostatistics, School of Public Health, University of California, Berkeley, rodmach@berkeley.edu

†University of California, Berkeley, hubbard@stat.berkeley.edu

Empirical Bayes Approach to Controlling Familywise Error: An Application to HIV Resistance Data

Rhoderick N. Machekano and Alan E. Hubbard

Abstract

Statistical challenges arise in identifying meaningful patterns and structures from high dimensional genomic data sets. Relating HIV genotype (sequence of amino acids) to phenotypic resistance presents a typical problem. When the HIV virus is under antiretroviral drug pressure, unfavorable mutations of the target genes often lead to greatly increased resistance of the virus to drugs, including drugs the virus has not been exposed to. Identification of mutation combinations and their correlation to drug resistance is critical in guiding efficient prescription of HIV drugs. The identification of a subset of codons associated with drug resistance from a set of several hundreds of codons presents a multiple testing problem. Statistical issues arising from genomic data multiple testing procedures include the choice of the null test-statistic distribution used to define cut-offs. Controlling familywise error rate implies controlling the number of false positives among true nulls. Given the large number of hypotheses to be tested, the number of true nulls is unknown. We apply two multiple testing procedures (MTPs) controlling familywise error rate: an adhoc augmented-Bonferroni method and a Empirical Bayes procedure originally proposed in van der Laan, Birkner and Hubbard(2005). Using simulations, we demonstrate that the proposed MTPs are less conservative than the traditional methods such as Bonferroni and Holm's procedures. We apply the methods to HIV resistance data where we wish to identify mutations in the protease gene associated with Amprenavir resistance.

1 Introduction

New technologies, such as micro-arrays, have revolutionized the genetic study of human disease. Statistical challenges arise in identifying meaningful patterns and structures from high dimensional genomic datasets. One area of statistical research that has generated renewed interest is multiple testing, where several hundreds or thousands of hypotheses are simultaneously tested. This article is motivated by HIV genotype-phenotype association studies, investigating which of hundreds of codon mutations might cause resistance to HIV drugs.

Management of HIV disease requires judicious use of regimens of multiple drugs that target two genes in the HIV genome: the protease (PR) and reverse transcriptase (RT) genes. When the HIV virus is under suboptimal drug pressure, the genes (sequences of amino-acids) often mutate at certain codon positions[1]. Unfavorable mutations lead to greatly increased resistance of the virus to the drugs, including resistance to drugs the virus has not been exposed to, resulting in limited treatment options.

Identification of mutation combinations and their correlation to drug resistance is therefore critical in guiding efficient prescription of HIV drugs in order to suppress viral replication. The RT and PR genes are about 560 codons and 99 codons long respectively. In wild-type virus, codons have a fixed arrangement, with particular codons found at certain positions of the gene. To identify mutant codons in the RT gene associated with drug resistance, 560 hypothesis tests need to be performed, one for each position. With no prior information on the effect of each codon position on drug resistance, the true null hypotheses are unknown. Test statistics generated from these type of data are often correlated in someway because of correlation between codons. The occurrence and effect of mutations at any given position are influenced by the presence of mutations at other positions - engendering potential correlation among positions. These issues motivates the statistical question of how to control the false positive rate when carrying out a large number of tests in which the test statistics maybe correlated and the true null hypotheses are unknown. Since our analyses are exploratory, we desire a testing procedure that is not overly conservative. We want a procedure that helps us identify a small proportion of codons potentially associated with drug resistance for further study.

In the following section, a brief review of the multiple testing framework is given. Section (3) describes two procedures controlling FWER: an adhoc Augmented Bonferroni method (AB) and an Empirical Bayes procedure (EB) proposed by van der Laan, Birkner and Hubbard for the control of the tail probability of the proportion false positives (TPPFP) and adopted here for controlling FWER. In section (4), we

investigate the performance of AB and EB procedures in comparison to the standard Bonferroni and Holm's procedures using simulation studies. In section (5), we apply the procedures to a HIV drug susceptibility dataset where we wish to identify positions in the HIV protease gene associated with resistance to HIV drug amprenavir. Section(6) summarizes our findings.

2 Multiple Testing Framework

For each codon $j = 1 \cdots m$, the hypothesis $H_{0j} : \mu_{1j} = \mu_{0j}$ is tested against $H_{1j} : \mu_{1j} \neq \mu_{0j}$, where μ_{1j} and μ_{0j} is the mean phenotypic resistance in subjects with and without mutation at codon j respectively. Let $T_{n1}, T_{n2}, \cdots, T_{nm}$ be the corresponding test statistics and let $p_1, p_2, \cdots, p_m$ be the associated p-values, where p_i summarizes the strength of evidence against the null hypothesis H_{0j}. In single hypothesis testing, H_{0j} is rejected if $|T_{nj}| > c_j(\alpha)$, or if $p_j \leq \alpha$, for some chosen $\alpha \in (0, 1)$, $j = 1, \cdots, m$. It is often assumed that if H_{0j} is true, T_{nj} comes from some standard theoretical distribution (e.g. normal or t-distribution) from which we get the critical values $c_j(\alpha)$. Define $\mathcal{H} = H_{0j} : j = 1, \cdots, m$ the set of all null hypothesis, $\mathcal{H}_0$ the set of all true null hypotheses and let $\mathcal{H}_1$ represent the set of true non-null hypotheses.

In a multiple testing problem, we would like to accurately estimate the subset $\mathcal{H}_0$, thus its complement $\mathcal{H}_1$, while controlling probabilistically some error rate under an assumed significance level α. Choosing the rejection region, $c_j(\alpha) : j = 1, \cdots, m$ is challenging, because the theoretical joint null distribution for the m test-statistics is not always obvious. Moreover, the difference between the theoretical null and empirical null distributions affects simultaneous inference[2].

2.1 Examples of Type I errors

Several experimentwise type I error rates have been proposed and the choice depends on how many false positives one can tolerate. If a significant test results in expensive

Table 1: *Possible outcomes from testing m hypotheses*

	Accept Null	Reject Null	
Null True	U_n	V_n	m_0
Null False	T_n	S_n	$m - m_0$
	W_n	R_n	m

follow-up testing, then a researcher might want to preclude having (almost) any false positives. If the analysis is more exploratory, perhaps a greater tolerance will be warranted. We review four type I errors that we may wish to control: familywise error rate (FWER), generalized familywise error rate (gFWER), tail probability of the proportion of false positives (TPPFP) and false discovery rate (FDR).

Let $V_n = \sum_{j=1}^{p} I(|T_{n_j}| > c_j | H_{0j} \text{ true})$ be the number of false rejections. Let $R_n = \sum_{j=1}^{p} I(|T_{nj}| > c_j)$ be the total number of rejected hypotheses - Table 1. Then

The familywise error rate (FWER) is defined as the probability of making at least one error by rejecting a true null hypothesis among all tests i.e. $FWER = P(V_n > 0)$. FWER is the most conservative.

The generalized familywise error rate (gFWER) is the probability of making at least k false positives, $P(V_n > k)$, for some $k = 1, \cdots, m$. gFWER generalizes the FWER to tolerate more false positives.

The tail probability of the proportion of false positives (TPPFP) among the total number of rejected hypotheses is defined by $P\left(\frac{V_n}{R_n} > q\right)$ for some set $q \in [0, 1]$. TPPFP controls the proportion of false positives.

The false discovery rate (FDR) is the expectation of the proportion of false positives, $E\left(\frac{V_n}{R_n}\right)$. FDR also controls the proportion of false positives, but does not control the bounds as TPPFP and is the most liberal.

2.2 Background to Multiple Testing Procedures

Traditionally, multiple testing procedures (MTPs) have sought to control the familywise error rate (FWER). Commonly used procedures that achieve control of FWER include the Bonferroni method and its modifications (e.g. Holm's step down procedure, Hochberg, and Hommel)[3–5]. While these procedures reduce the probability of spurious findings, they also severely reduce the probability of identifying real effects[6]. Alternative methods that control less conservative error rates, such as the false discovery rate (FDR) introduced by Benjamin and Hochberg[3,7,8] and its modifications, control the expected proportion of false-positive findings among all rejected hypotheses. The choice of which error to control is often guided by the problem at hand: exploratory versus confirmatory analyses. In exploratory analysis, we want to relax our control on type I error rates, whereas, confirmatory analyses demand strict control of type I error rates. Thus methods controlling FWER have been preferred in confirmatory analyses compared to methods controlling FDR. On the contrary, FDR

procedures have been preferred in exploratory analyses than procedures controlling FWER.

In HIV genomic analyses, we prefer methods that are less conservative because we are searching for new sites of drug resistance. We do not want to rule out those mutations that may induce even the slightest level resistance. Thus, Bonferroni-like procedures might be too conservative because of their strict control of the type I error under all data-generating distributions. Researchers have noted the conservatism of the Bonferroni procedure, particularly when the data (hence the test statistics) are correlated. The Bonferroni procedure provides sharp FWER control only under independence of test statistics. In addition, in the presence of dependence between test statistics, the power of the Bonferroni procedure sharply diminishes with the number of tests[4,9]. Establishing associations between drug resistance and DNA sequences is complicated by the presence of dependence particularly between codons within the same region. The Bonferroni procedure as well as other single-step procedures assume that all null hypothesis are true. Clearly, these assumptions do not hold with genomic data.

Westfall and Young (1993) introduced a resampling based approach in estimating the test statistic null distribution that takes into consideration the dependence between test statistics[5]. Pollard, van der Laan and Dudoit proposed a resampling based multiple testing methodology that controls FWER but does not require the subset pivotality condition, a requirement in Westerfall and Young's resampling method[10,11–14]. Subset pivotality is satisfied when the joint distribution of the test statistics does not dependent on the subset of true null hypothesis. This requires that the covariance of the test statistics under the true data generating distribution is the same as the covariance of the test statistics under a null data generating distribution. On page 42 of their book, Westfall and Young give multiple testing examples where subset pivotaility property is either satisfied (when testing if several means equal zero using t-tests) or violated (when testing if pairwise correlations between several variables equal zero)[5].

Efron and Tibshirani used an Empirical Bayes approach to control the FDR[2] in which an attempt to guess the set of true nulls is made. van der Laan proposed a method for controlling FWER following Efron's Empirical Bayes approach, combining it with Pollard's resampling-based null distribution estimation method[11,15]. Interestingly, Schweder and Spjotvoll, in an earlier paper, demonstrated a graphical method to identify true null hypotheses, and suggested an improved Bonferroni method by dividing α by the estimated number of true nulls[16]. We take a similar approach here for our Augmented Bonferroni approach.

3 Using the Mixture Model to control FWER

First, we take a Bayesian perspective that

1. the indicator of whether a test statistic comes from a null distribution is a random variable, and

2. there is a prior probability that the null hypothesis is true.

As discussed by van der Laan[15] this view of the data-generating mechanism provides sensible control of type I error rates even though the motivation, in reality, is frequentist (whether or not a test statistic is from a null data generating distribution is fixed).

Formally, let $\vec{B} = (B_1, \cdots, B_m)$ be a joint vector of Bernoulli indicator variables $B_j = 1 - I(H_{0j} \quad \text{true})$, where $I(.)$ is the indicator function (i.e. $B_j = 0$ when H_{0j} is true). Assume the marginal distribution of B_j is Bernoulli$(1 - p_0)$, i.e. $P(B_j = 0) = p_0, \forall j = 1, \cdots, m$. Let $\mathcal{S}_0 = \{j : B_j = 0, j = 1, \cdots, m\}$ represent the set of true null hypotheses. Let $\vec{T}_n \equiv (T_{n1}, \cdots, T_{nm})$ be the m vector of test statistics corresponding to testing each of the m null hypotheses $\mathcal{H} = \{H_{0j} : j = 1, \cdots, m\}$ and let $Q_{n|\vec{B}}$ denote the joint conditional distribution of $\vec{T}_n$. Assume that the marginal distributions of $Q_{n|\vec{B}}$ corresponding with the true null hypotheses H_{0j} (i.e. $B_j = 0$) equal a common known distribution $F_{0,n}$, and that the other marginal distributions equal a common unknown distribution $F_{1,n}$ with corresponding density functions $f_{0,n}$ and $f_{1,n}$. We assume the marginal distribution of test statistics are from a mixture of the known null density $f_{0,n}$ (e.g. $\mathcal{N}(0, 1)$) and unknown alternative density $f_{1,n}$ with unknown mixing proportion p_0:

$$T_{nj} \sim f_n \equiv p_0 f_{0,n} + (1 - p_0) f_{1,n}$$

.

Under the mixture model, the posterior probability that T_{nj} came from a true H_{0j} given its observed value T_{nj} is given by:

$$P(B_j = 0 | T_{nj}) \equiv \Phi_n = p_0 \frac{f_{0,n}(T_{nj})}{f_n(T_{nj})} \tag{1}$$

In order to calculate this posterior probability, one needs an estimate of both $f_{0,n}$ and f_n. As stated above, typically one assumes $f_{0,n}$ is known (e.g. $N(0, 1)$). Density estimation procedures (e.g. Kernel smoothing) applied to the test statistics $\{T_{jn} : j = 1, \cdots, m\}$ can be used to estimate f_n. The proportion of true null

hypotheses $p_0 = \frac{|\mathcal{S}_0|}{m}$ is the user's choice, where $p_0 = 1$ is a natural conservative choice.

Below, we discuss two methods that use this posterior distribution to estimate the proportion of true nulls among the set of hypotheses. The idea is to increase the power of typical single-step procedures, which control under a global null, type I error rates by using empirical information to "adjust" the procedures to the actual (estimated) number of true nulls.

3.1 Augmented Bonferroni

The Bonferroni procedure rejects hypothesis H_{0j} when the p-value p_j is less than $\frac{\alpha}{m}$, where α is the preset FWER level and m is the number of null hypotheses. This leads to the Bonferroni single-step adjusted p-value $\tilde{p}_j = \min(mp_j, 1)$[5]. The Bonferroni method controls the FWER under the complete null hypotheses assumption (i.e. all the null hypotheses are true). However, it is desirable to be able to control the FWER regardless of the subset of true null hypotheses (strong control)[4]. Of course, the problem is that we do not know the number true null hypotheses. However, we can simply use equation (1) to estimate the number of true null hypotheses, and propose a modified Bonferroni-like method that attempts to control FWER only among the true nulls. Specifically, the Augmented Bonferroni method estimates adjusted p-value ($\tilde{p}_j$) by multiplying the "raw" p-value (p_j) by the estimated number of true null hypotheses $\hat{m}_0$ (i.e. $\tilde{p}_j = \min(\hat{m}_0 p_j, 1)$) , where

$$\hat{m}_0 = \sum_{j=1}^{m} \hat{P}(B_j = 0|T_{nj}) \tag{2}$$

$$= \sum_{j=1}^{m} \frac{p_0 f_{0,n}(T_{nj})}{\hat{f}(T_{nj})} \tag{3}$$

By rejecting the null hypothesis if the unadjusted p-value for the j^{th} test p_j is less than $\frac{\alpha}{\hat{m}_0}$ or if the adjusted p-value $\tilde{p}_j$ is less than α and failing to reject H_{0j} when otherwise. We speculate that FWER $\leq \alpha$ and this procedure is more powerful than the standard Bonferroni correction.

3.2 Empirical Bayes Procedure

Empirical Bayes procedure goes a step further than Augmented Bonferroni by estimating the test statistic null distribution among the true nulls. The null distribution comes from the distribution of the maximum of the m correlated test statistics or

the minimum of the p-values. As the distribution of these maxima for an unknown correlation structure is unknown, use of a bootstrap procedure seems the reasonable way to estimate the null distribution[17]. Common cut-offs that guarantee control of FWER are chosen from the maximum test statistic distribution.

The method for choosing the cut-off c involves controlling the tail probability of a random variable $\tilde{v}_n(c)$ defined as

$$\tilde{v}_n(c) = \sum_{j=1}^{m} I(\tilde{T}_{n,j} > c, j \in \mathcal{S}_{0n})$$

representing the guessed number of false positives, defined by drawing a random set $\mathcal{S}_{0n}$ which represents a guessed set of true null hypotheses $\mathcal{S}_0$ and a draw $\tilde{T}$ from a null distribution of the test statistic vector. From van der Laan, Birkner and Hubbard[15], the distribution of $\mathcal{S}_{0n}$ and null distribution of $\tilde{T}$ are chosen so that $\tilde{v}_n(c)$ asymptotically dominates in distribution the true number of false positives $V_n(c)$. The result of lemma (1) in appendix 1 justifies this approach to FWER control.

3.2.1 Implementation: Empirical Bayes method

Draw guessed set of null hypothesis $\tilde{\mathcal{B}}_{0n}$

1. Estimate the posterior probability $P(B_j = 0|T_{nj}), \quad \forall j = 1, \cdots, m$ using equation (1)
2. Generate m Bernoulli random variables using probability estimated above
3. Repeat 2 B times to form $m \times B$ matrix

$$\tilde{\mathcal{B}}_{0n} = \begin{array}{c} \overbrace{\qquad\qquad\qquad\qquad}^{\text{B columns}} \\ \begin{array}{ccccc} i & & & & \\ 1 & 1 & 1 & \cdots & 0 \\ 2 & 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ m & 0 & 1 & \cdots & 0 \end{array} \end{array}$$

Test Statistics Null Distribution

1. Sample with replacement from data
2. Calculate re-scaled, null centered test statistics T_{kj} for $j = 1, \cdots, m$
   (a) First calculate the test statistic for each hypothesis

(b) Repeat 1 and 2 B times and generate $m \times B$ matrix

$$T_{Boot} = \overbrace{\begin{array}{ccccc} i & & & & \\ 1 & T_{11} & T_{12} & \cdots & T_{1B} \\ 2 & T_{21} & T_{22} & \cdots & T_{2B} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ m & T_{m1} & T_{m2} & \cdots & T_{mB} \end{array}}^{\text{B columns}}$$

(c) For each row j calculate the mean $\bar{T}_j$ and standard error of the test statistics $SD(T)$.

(d) For each row entry in T, subtract the associated row mean and divide by the standard error, multiply by the "desired" null standard error σ, then add the "desired" null mean μ as follows: $\tilde{T}_{ji} = \frac{(T_{ji} - \bar{T}_j)\sigma}{SD(T_{ji})} + \mu$, for $j = 1, \cdots, m$ and $i = 1, \cdots, B$.

This gives a test statistic null distribution under the complete null hypothesis with mean μ and variance σ^2. However, we need to correct this distribution to take into consideration the fact that a subset of the null hypothesis is false.

Corrected Test Statistic Null Distribution

1. Take the Hadamard product (entry-wise product) of $\tilde{\mathcal{B}}_{0n}$ and $\tilde{T}$

$$\tilde{T} \bullet \tilde{B}_{0n} = T_0 = \overbrace{\begin{array}{ccccc} j & & & & \\ 1 & T_{11} & T_{12} & \cdots & 0 \\ 2 & T_{21} & 0 & \cdots & T_{2B} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ m & 0 & T_{k2} & \cdots & 0 \end{array}}^{\text{B columns}}$$

2. Get maximum in each column $T_1^{\max}, \cdots, T_B^{\max}$ to get the max-T distribution only among the null hypothesis

3. Adjusted p-value for position j, $\tilde{p_j}$, is given by

$$\tilde{p_j} = \frac{1}{B} \sum_{k=1}^{B} I(T_{nj} \geq T_k^{\max}) \tag{4}$$

4. OR, the critical value $c(\alpha)$ is given as the $1 - \frac{\alpha}{2}$ quartile of the max-T distribution. Reject H_{0j} if $|T_{nj}| \geq c(\alpha)$ otherwise do not reject H_{0j}.

The max-T distribution above guarantees asymptotic control of FWER at level less than or equal to α given the distribution of null-centered rescaled T's dominates the null distribution of observed test statistics (see Duboit for proof[18,19]).

4 Simulation studies

We undertake simulation studies to assess the performance of the proposed MT procedures in controlling FWER under a range of test-statistic correlation structures and two codon effect sizes. The simulations are intended to examine the effect of increasing correlation in the data on the relative performance of the test procedures. Three sets of simulations using the following data generating mechanisms

1. Things in common in simulations

   (a) the model for generating mean outcome by covariate group (codon). Given $\mathbf{X}$, we assume a linear model for each independent observation i.e.

   $$Y = \mathbf{X}\beta + \epsilon \tag{5}$$

   where β is a real-valued vector of coefficients representing the effect of each codon on the outcome, $\mathbf{X}$ is a vector of binary indicators for presence or absence of mutation at each codon, and $\epsilon \sim \mathcal{N}(0,1)$, some random error.

   (b) $\beta = 0$ for the nulls (i.e. codons with no effect on outcome) and $\beta = 0.2$ or $\beta = 1$ for codons that have an effect on the outcome because the X's are uncorrelated between group of nulls and alternatives. The codon effect sizes were chosen based on the observed unadjusted codon effects on drug resistance from the Stanford Sequence data.

   (c) In all the three simulations, a vector of 200 codons for each individual $\mathcal{X}$ is generated from a multivariate binary distribution assuming varying correlation structures. In each simulation, the proportion of mutations for each codon (i.e.$P(X_j = 1), j = 1, \cdots, m$) was obtained by randomly sampling a number in the interval $[0.4, 0.7]$. The $\mathcal{X}$ vector was generated using the *bindata* package[20] from the comprehensive R archive network CRAN (www.r-project.org).

   (d) We generated 200 observations in simulation studies.

2. Things that are different between simulations

   (a) In simulation 1, we assumed the 200 codons were independent from each other ($\rho = 0$).

   (b) In simulation 2, we assumed a weak correlation ($\rho = 0.2$) structure between 190 codons in the null group and weak correlation ($\rho = 0.2$) between 10 codons in the alternative group. However, we assumed independence between codons in null and alternative groups.

   (c) In the third simulation, we assumed strong correlation ($\rho = 0.9$) between codons within each of the two groups, but independence of codons between the groups as described in simulation 2.

3. Choice of f_{0n}: we assumed the test-statistic null distribution f_{on} was a t-distribution with $n - 1$ degrees of freedom.

4. Estimation of f: to estimate the marginal distribution of the test statistics f, we used kernel density estimation, assuming a gaussian kernel. We choose the default bandwidth in R (version 2.2.0) which equals 0.9 times the minimum of the standard deviation and the interquartile range divided by 1.34 times the sample size to the negative one-fifth power (this Silverman's "rule of thumb", Silverman (1986, page 48, eqn (3.31)).

5. Choice of the prior proportion of true nulls, p_0: we conservatively assumed all nulls were true i.e. $p_0 = 1$.

6. For each study, treating the entire dataset as the population, we repeatedly random sample with replacement an equal number of observations ($n = 200$) to create 2000 data sets. For each data set, we estimate the number of type I and type II errors for each MTP:

$$\#\text{ type I error} = \sum_{j=1}^{m} I(H_{0j} \quad \text{true}) I(\text{adj. p-value}(j) \leq \alpha)$$

where adj. p-value(j) is the adjusted p-value for the hypothesis j. We estimate the FWER for each procedure by counting the number of times the number of type I errors is greater than or equal to 1 and dividing by the total number of simulations (S=2000).

4.1 Simulation Results

Simulation results are presented in tables 2-4. Table 2 shows the FWERs for the two effect sizes assuming complete independence between codons. All the procedures have comparable FWERs $\approx$ 0.05. The Bonferroni and Holm procedures perform as well as the proposed procedures because the data satisfies independence of test statistics assumption, a cornerstone of the Bonferroni procedure. Under weak codon correlation (table 3), the Bonferroni and Holm procedures (≈ 0.04) are slightly more conservative compared to the Augmented Bonferroni and Empirical Bayes procedures (≈ 0.05 or 0.06). The FWERs are consistent across the two effect sizes. Similarly, when correlation between codons within a group is strong (table 4), the Bonferroni and Holm procedures are appreciably more conservative with FWERs as low as 0.0131 and 0.019. On the contrary, the Augmented Bonferroni procedure is anti-conservative with FWER at 0.065 and 0.0765 for the $\beta = 0.2$ and $\beta = 1$ respectively. The Empirical Bayes procedure (FWERs = 0.0392 and 0.034) is less conservative compared to the Bonferroni and Holm procedures but still gives proper control (≤ 0.05). For all three correlation structures, all four procedures tend to give slightly higher FWERs when there are strong codon effects ($\beta = 1$) compared to when codon effects are weak ($\beta = 0.2$).These results highlight the fact that the Bonferroni and Holm procedures are overly conservative by not accounting for the correlation structure of the test statistics.

5 Application: Identification of mutant codon positions associated with drug resistance

We apply the Augmented Bonferroni and Empirical Bayes MTPs to HIV resistance data from the Stanford HIV Sequence Database, comparing the new Empirical Bayes procedures to the standard

Bonferroni and Holms procedures. The data consists of amino acid sequences from the protease gene and the corresponding fold-resistance as a measure of phenotypic resistance. The aim of the study was to identify mutant codon positions that are significantly associated with viral phenotypic resistance. In this article, we restrict our analysis to identifying protease mutations associated with viral resistance to the HIV drug Amprenavir among 876 patients from the Stanford HIV Sequence Database. The sequence data is represented by 99 binary variables for each codon, taking values 1 when there is a mutation at the codon, and 0 when the codon is wildtype. Each i^{th} patient provides a vector of $X_i = (X_{i,1}, X_{i,2}, \cdots, X_{i,99})$ with $X_{i,j} = 1$ or 0 indicating whether or not a mutation occurred at codon j and a fold-resistance measure Y_i, which we log transform. Before calculating the test statistic, we filtered out conserved positions. We defined conserved positions to be those positions with less than 3% mutation rate. The resulting dataset had forty three positions, and for each position, a test statistic T_{nj} is calculated based on the usual t-test for comparison of two group means:

$$T_{nj} = \frac{\bar{y}_{1j} - \bar{y}_{0j}}{\sqrt{\frac{s^2_{1j}}{n_1 j} + \frac{s^2_{0j}}{n_0 j}}} \tag{6}$$

where $\bar{y}_{1j} = \frac{\sum_{i=1}^{n} X_{i,j} Y_i}{\sum_{i=1}^{n} X_{i,j}}$, $\bar{y}_{0j} = \frac{\sum_{i=1}^{n} (1-X_{i,j}) Y_i}{\sum_{i=1}^{n} (1-X_{i,j})}$, $s^2_{1j} = \frac{\sum_{i=1}^{n} X_{i,j} (Y_i - \bar{y}_{1j})^2}{\sum_{i=1}^{n} X_{i,j} - 1}$, $s^2_{0j} = \frac{\sum_{i=1}^{n} (1-X_{i,j})(Y_i - \bar{y}_{1j})^2}{\sum_{i=1}^{n} (1-X_{i,j}) - 1}$, and $n = 876$.

After eliminating conserved positions, we computed standardized t-statistics T_{nj} comparing mean log fold-resistance between mutant sequences and wild-type sequences and associated p-values p_j for each position j. Figure (5) shows the empirical distribution of 43 test statistics and a kernel density estimate of the distribution. The distribution of the test-statistics suggest that we have a mixture distribution, clear evidence against a global null distribution. Thus Empirical Bayes is an appropriate candidate for multiple testing in this data.

We used kernel density estimation assuming a Gaussian kernel to estimate the marginal density of the test statistics T_{nj} $(f(T_{nj}))$. We first assumed the null distribution of the test-statistics,$f_0(T_{nj})$, was a t-distribution with $(n-1)$ degrees of freedom $t_{(n-1)}$.We then estimated the probability of a true null hypothesis given the test-statistic, $p_{nj} \equiv P(B_j = 0|T_{nj}) \equiv p_0 \frac{f_0(T_{nj})}{f(T_{nj})}$, where we set $p_0 = 1$, and calculated the expected number of true nulls (equation (2)).

Following the estimation outline above, the Augmented Bonferroni and Empirical Bayes adjusted p-values were calculated. Using the *multtest* package in R, Bonferroni-adjusted and Holms-adjusted p-values were obtained. R esults are summarized in table 5 and table 6.

5.1 Data Analysis Results

Table (5) shows the number of rejected null hypotheses by procedure. The unadjusted analysis identified 27 out of 43 positions in the protease gene significantly associated with resistance to Amprenavir at $\alpha = 0.05$. After adjusting for multiple comparison (FWER control), the Bonferroni and Holm's procedures both identify 18 positions associated with amprenavir resistance at $\alpha = 0.05$ significance level. The augmented Bonferroni and Empirical Bayes procedures are less conservative, identifying 2 more positions than regular Bonferroni procedure. Table (6) lists in detail the protease codon positions and the corresponding unadjusted, Bonferroni, Holms, Augmented Bonferroni and Empirical Bayes adjusted p-values.

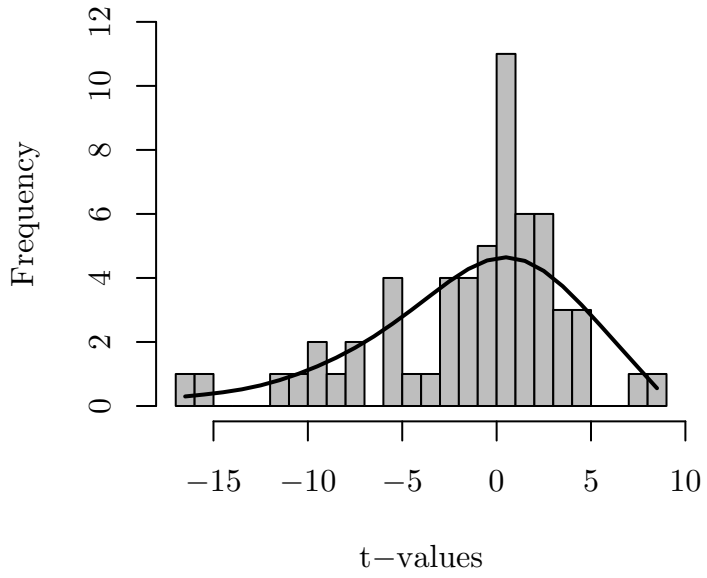

Figure 1: Histogram of 43 t-values from the protease gene mutation-drug resistance analysis

We compared our data analysis results to genotype-drug resistance profiles published at the Stanford HIV Resistance website (http://hivdb.stanford.edu/cgi-bin/PIResiNote.cgi). Our proposed methods confirmed 14 of the the 19 published protease mutations associated with resistance to amprenavir. In addition, we identified 6 new mutations which may have influence on virologic response to treatment with amprenavir. Of these 6 new mutations, position 30 has been associated with increased sensitivity to another drug, Nelfinavir. Our analysis shows that any mutation at this position is also associated with virus sensitivity to amprenavir. Position 89 which is associated with increased resistance to amprenavir is next to position 90 which has been confirmed as highly resistant to amprenavir. The remaining 4 positions (13, 37, 41, and 57) among the new discoveries are associated with increased sensitivity to amprenavir and need to be confirmed through further studies.

6 Discussion

We have presented two recently proposed multiple testing procedures for controlling FWER highlighting their motivation, estimation algorithms and their performance under varying correlation

structures. The procedures follow a Bayesian framework in estimating the number of true nulls (Augmented Bonferroni), and uses resampling methods to estimate the test-statistic null distribution among true nulls (Empirical Bayes). Using simulation studies under different codon position correlation structures we compared their performance in controlling FWER to the classical single-step Bonferroni and step-down Holm approaches.

We assumed a linear model between drug resistance and codon positions. A linear model is probably an oversimplified representation of the relationship between drug resistance and codon mutations because interaction between codon positions are likely to influence drug resistance. In reality, because HIV drugs target certain parts of the gene,"active sites", it is likely that mutation changes occur in groups and may interact between each other to either enhance or reduce resistance. However, this should not undermine the usefulness of the methods at all, besides the fact that adding interaction terms increases the number of tests to be performed.

The performance of the Bonferroni and Holm procedures depend on the strength of correlation between codons, while the Augmented Bonferroni and Empirical Bayes procedures are less influenced by the correlation structure. Under strong correlation, Bonferroni and Holm procedures are extremely conservative with FWERs as low as 0.013. Under weak correlation, Bonferroni and Holm procedures are less conservative with estimated FWERs around 0.04. However, under independence of codon positions, Bonferroni and Holm procedures achieve FWERs consistent with an upper bound on the type I error probability of 0.05. On the contrary, the augmented Bonferroni and Empirical Bayes procedures consistently control FWER around 0.05 irrespective of the correlation of the test statistics. In the data analysis, we discovered two more protease codon mutations associated with amprenavir resistance when we used the Augmented Bonferroni and Empirical Bayes procedures compared to the classical Bonferroni and Holm procedures.

Our study confirms that the Bonferroni procedure is conservative unless test statistics are independent and one can do better in situations where there is some correlation between test-statistics. The Holm procedure is also an attempt to control FWER only among the nulls and is a less conservative procedure compared to the Bonferroni procedure. However, when there is strong correlation between test statistics, Holm procedure can also be conservative. The Augmented Bonferroni procedure is similar to the Holm procedure in that it attempts to control FWER only among nulls, but uses an ad hoc procedure based on posterior probability of the null to identify the 'true' nulls. The Augmented Bonferroni procedure achieves the desired effect i.e. makes the Bonferroni procedure less conservative, but in simulations, is sometimes anti-conservative. Empirical Bayes procedure controls FWER among true nulls and takes into account the dependence structure of the test statistics. Empirical Bayes is our preferred method given similar data structures because it is consistent, with the probability of falsely rejecting any type hypothesis that is a member of the set of true null hypothesis approaching $\alpha = 0.05$). Our proposed methods can be easily implemented and the R code is included in the appendix.

Table 2: Simulation results: independent covariance structure

MTP	FWER	
	$\beta = 0.2$	$\beta = 1$
Bonferroni	0.052	0.047
Holm	0.054	0.047
Augm. Bonf	0.049	0.043
Emp. Bayes	0.057	0.051

Table 3: Simulation results: weak covariance structure

MTP	FWER	
	$\beta = 0.2$	$\beta = 1$
Bonferroni	0.039	0.039
Holm	0.042	0.039
Augm. Bonf	0.061	0.047
Emp. Bayes	0.057	0.049

Table 4: Simulation results: strong covariance structure

MTP	FWER	
	$\beta = 0.2$	$\beta = 1$
Bonferroni	0.013	0.017
Holm	0.013	0.019
Augm. Bonf	0.076	0.065
Emp. Bayes	0.039	0.034

Table 5: Number of rejected null hypotheses out of 43

Method	Rejections
Unadjusted	27
Bonferroni	18
Holms	18
Augmented Bonferroni	20
Empirical Bayes	20

Table 6: Protease Codons associated with resistance to Amprenavir

No.	Codon(j)	T_{jn}	p_j	Bonf.	Holms	Aug. Bonf.	Emp. Bayes
1	84	-16.13	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
2	10	-15.45	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
3	46	-12.53	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
4	54	-10.68	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
5	71	-9.23	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
6	90	-8.70	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
7	82	-7.91	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
8	88	7.40	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
9	33	-7.06	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
10	47	-7.03	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
11	24	-6.38	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
12	32	-5.62	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
13	73	-5.60	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
14	20	-4.06	0.00005	0.00234	0.00163	0.00250	0.00096
15	30	3.84	0.00014	0.00586	0.00395	0.00460	0.00240
16	13	3.69	0.00024	0.01041	0.00678	0.00670	0.00427
17	37	3.54	0.00042	0.01815	0.01139	0.00990	0.00744
18	41	3.41	0.00069	0.02964	0.01792	0.01610	0.01215
19	89	-3.08	0.00217	0.09346	0.05434	0.04010	0.03832
20	57	3.03	0.00256	0.10993	0.06136	0.04450	0.04507
21	64	2.98	0.00302	0.12971	0.06938	0.05120	0.05318
22	63	2.70	0.00719	0.30908	0.15814	0.09870	0.12672
23	35	2.63	0.00868	0.37340	0.18236	0.11550	0.15309
24	12	2.48	0.01325	0.56994	0.26509	0.16000	0.23366
25	14	2.42	0.01567	0.67391	0.29778	0.18140	0.27629
26	53	-2.34	0.01972	0.84782	0.35490	0.22350	0.34758
27	45	2.19	0.02859	1.00000	0.48597	0.30270	0.50395
28	48	-1.96	0.05090	1.00000	0.81435	0.47370	0.89726
29	62	1.93	0.05457	1.00000	0.81852	0.49680	0.96197
30	93	1.88	0.06049	1.00000	0.84679	0.53430	1.00000
31	15	1.65	0.09885	1.00000	1.00000	0.72100	1.00000
32	36	-1.40	0.16070	1.00000	1.00000	0.89390	1.00000
33	69	1.23	0.21964	1.00000	1.00000	0.96060	1.00000
34	70	1.12	0.26237	1.00000	1.00000	0.98230	1.00000
35	61	1.03	0.30315	1.00000	1.00000	0.99230	1.00000
36	50	-0.80	0.42478	1.00000	1.00000	0.99990	1.00000
37	55	-0.67	0.50029	1.00000	1.00000	1.00000	1.00000
38	77	0.67	0.50302	1.00000	1.00000	1.00000	1.00000
39	19	0.33	0.74050	1.00000	1.00000	1.00000	1.00000
40	16	0.22	0.82540	1.00000	1.00000	1.00000	1.00000
41	74	-0.22	0.82815	1.00000	1.00000	1.00000	1.00000
42	60	0.13	0.89902	1.00000	1.00000	1.00000	1.00000
43	72	0.02	0.98764	1.00000	1.00000	1.00000	1.00000

Appendix 1

Lemma (From Dudoit) 1. *Consider the simultaneous test of m null hypothesis $H_{0i}, i = 1, \cdots, m$, based on an m-vector of test statistics $T_n = (T_{ni} : i = 1, \cdots, m)$, with true distribution Q_n. Consider null test statistics $T_0 = (T_{0i} : i = 1, \cdots, m)$ and random guessed sets $\mathcal{H}_{0n} \subseteq 1, \cdots, m$ of true null hypotheses, where, given empirical distribution P_n, T_{0n} and $\mathcal{H}_{0n}$ are independent, with respective conditional distributions Q_{0n} and H_{on}. Given m-dimensional vectors of cut-offs $c = (c_i : i = 1, \cdots, m)$, test statistics T_n and null test statistics T_{0n}, and a random guessed set $\mathcal{H}_{0n}$, define the number of false positives to be $V(c)$. For type I error level $\alpha \in [0, 1]$, the corresponding FWER is*

$$FWER = Pr(V(c, \mathcal{H}_{0n}, T_n) > 0 | P_n) \leq \alpha \tag{7}$$

$$FWER = Pr(V(c, \mathcal{H}_{0n}, T_{0n}) > 0 | P_n) \leq \alpha \tag{8}$$

Proof. For detailed proof, see Dudoit,van der Laan and Birkner □

Appendix 2

R code for Augmented Bonferroni and Emperical Bayes Procedures

Example using t-tests to compare mean virus growth between wild-type and mutant codons

```
# Get t-statistics for all codons
Tn.2<-apply(dat.faux,1,pairedt.teststat)

# Function to get density at each point
fd=function(x,Tn){dg=density(Tn,from=x,to=x,bw='nrd',kernel="gaussian") dg$y[1]}

# Get density of test statistics (f) and under null

Tn.mat.2<-matrix(Tn.2,length(Tn.2),1)
f.Tn.2<-apply(Tn.mat.2,1,fd,Tn=Tn.2)
n<-dim(dat.faux)[2]

# Assume null dist. is t-distribution with n-1 df

f.Tn.0<-dt(Tn.2,df=n-1)

# P(B(j)=0|T_n(j)) pn<-pmin(1,f.Tn.0/f.Tn.2)

# Bootstrap to get joint null dist of test stat
B=5000
p<-dim(dat.faux)[1]
boot.stat.mat<-matrix(0,p,B)
Sn0.mat<-matrix(0,p,B)

# Get a random sample of Sn0 B times as well as the null ## T-stats

for(i in 1:B) {
 cat(" i =
",i,"\n")
  ii<-sample(1:n,n,replace=T)
  dat3<-dat.faux[,ii]

boot.stat.mat[,i]<-apply(dat3,1,pairedt.teststat)
Sn0.mat[,i]<-rbinom(p,1,pn) }

# Center distribution
boot.stat.centered<-boot.stat.mat-matrix(rep(apply(boot.stat.mat,1,mean),B),p,B)
```

```
# Or boot.stat.centered<-boot.stat.mat-matrix(rep(Tn.2,B),p,B)

# Need distribution of maximum Tn's only among nulls

Tn0.boot<-Sn0.mat*boot.stat.centered
abs.max<-function(x) {
    max(abs(x))
    }
    maxT<-apply(Tn0.boot,2,abs.max)

# Adjusted p-values
cdf.F<-ecdf(maxT)
adj.pvalue<-1-cdf.F(abs(Tn.2))

# Adjusted p-value based on using p* = sum(pn) and Bonferroni
pstar<-sum(pn)
adj.bonf<-pmin(pvalue*pstar,1)


# Get number of Type I and Type II errors for different procedures

alpha<-0.05
inv.true<-1-true.Sn
typeI.bonf<-sum(inv.true[oo][res$adj[,"Bonferroni"]<alpha])
typeII.bonf<-sum(true.Sn[oo][res$adj[,"Bonferroni"]>alpha])
typeI.holm<-sum(inv.true[oo][res$adj[,"Holm"]<alpha])
typeII.holm<-sum(true.Sn[oo][res$adj[,"Holm"]>alpha])
typeI.ebayes<-sum(inv.true[adj.pvalue<alpha])
typeII.ebayes<-sum(true.Sn[adj.pvalue > alpha])
typeI.adjbonf<-sum(inv.true[adj.bonf < alpha])
typeII.adjbonf<-sum(true.Sn[adj.bonf > alpha])
results<-cbind(c(typeI.bonf,typeI.holm,typeI.ebayes,typeI.adjbonf),
c(typeII.bonf,typeII.holm,typeII.ebayes,typeII.adjbonf), rep(alpha,4))

rownames(results)<-c("Bonf","Holm","Ebayes","adjbonf")

colnames(results)<-c("type I","type II","alpha") results
```

References

[1] Dulioust A, Paulous S, Guillemot L, Delavalle AM, Boue F, F Clavel. Constrained evolution of human immunodeficiency virus type 1 protease during sequential therapy with two distinct protease inhibitors. Journal of Virology 1999;73(1):850–854.

[2] Efron Bradley. Large-scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. Journal of the American Statistical Association 2004;99(465):96–104.

[3] Benjamini Yoav, Hochberg Yosef. On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics. Journal of Educational and Behavioral Statistics 2000;25(1):60–83.

[4] Hochberg Yosef, Tamhane AjitC. Multiple Comparison Procedures. John Wiley & Sons; 1987.

[5] Westfall PeterH, Young SStanley. Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment. John Wiley & Sons; 1993.

[6] Tadesse MahletG, Ibrahim JosephG, Vannucci Marina, Gentleman Robert. Wavelet Thresholding with Bayesian False Discovery Rate Control. Biometrics 2005;61(1):25–35.

[7] Benjamini Yoav, Yekutieli Daniel. The Control of the False Discovery Rate in Multiple Testing under Dependency. The Annals of Statistics 2001;29(4):1165–1188.

[8] Yekutieli Daniel, Benjamini Yoav. Resampling-based False Discovery Rate Controlling Multiple Test Procedures for Correlated Test Statistics. Journal of Statistical Planning and Inference 1999;82:171–196.

[9] Hochberg Yosef, Benjamini Yoav. More Powerful Procedures for Multiple Significance Testing. Statistics in Medicine 1990;9:811–818.

[10] Pollard KathrineS, van der Laan MarkJ. Resampling-based Multiple Testing: Asymptotic Control of Type I Error and Applications to Gene Expression Data. Division of Biostatistics, School of Public Health, University of California, Berkeley. Berkeley Electronic Press; 2003.

[11] Pollard KathrineS, Birkner MerrillD, van der Laan MarkJ, Dudoit Sandrine. Test Statistics Null Distributions in Multiple Testing: Simulation Studies and Applications to Genomics. Division of Biostatistics, School of Public Health, University of California, Berkeley. Berkeley Electronic Press; 2005.

[12] Dudoit Sandrine, van der Laan MarkJ, Pollard KathrineS. Multiple Testing. Part I. Single-Step Procedures for Control of General Type I Error Rates. Statistical Applications in Genetics and Molecular Biology 2004;3(1):1–71.

[13] van der Laan MarkJ, Dudoit Sandrine, Pollard KathrineS. Multiple Testing. Part II. Step-Down Procedures for Control of the Family-Wise Error Rate. Statistical Applications in Genetics and Molecular Biology 2004;3(1):1–35.

[14] van der Laan MarkJ, Dudoit Sandrine, Pollard KathrineS. Multiple Testing. Part III. Procedures for Control of the Generalized Family-Wise Error Rate and Proportion of False Positives. Division of Biostatistics, School of Public Health, University of California, Berkeley. Berkeley Electronic Press; 2004.

[15] van der Laan MarkJ, Birkner MerrillD, Hubbard AlanE. Resampling-based Multiple Testing: Procedure Controlling Tail Probability of Proportion of False Positives. Division of Biostatistics, School of Public Health, University of California, Berkeley. Berkeley Electronic Press; 2005.

[16] Schweder T, Spjotvoll E. Plots of P-values to evaluate many tests simultaneously. Biometrika 1982;69:493–502.

[17] Reitmeir Peter, Wassmer Gernot. Resampling-based Methods for the Analysis of Multiple Endpoints in Clinical Trials. Statistics in Medicine 1999;18:3453–3462.

[18] Dudoit S, van der Laan MJ. Multiple Testing Procedures and Applications to Genomics. Springer; 2006. (In preparation).

[19] Dudoit Sandrine, van der Laan MarkJ, Birkner MerrillD. Multiple Testing Procedures for Controlling Tail probability Error Rates. Division of Biostatistics, School of Public Health, University of California, Berkeley. Berkeley Electronic Press; 2004.

[20] Leisch Friedrich, Weingessel Andreas, Kurt Hornik. On the Generation of Correlated Artificial Binary Data. Institut for Statiskit und Wahrscheinlichkeitstheorie, Technische Universitat Wien; 1998.