Empirical Null Estimation Using Zero-Inflated Discrete Mixture Distributions and its Application to Protein Domain Data

Iris Ivy M. Gauran, Junyong Park, Johan Lim, DoHwan Park, John Zylstra, Thomas Peterson, Maricel Kann, and John L. Spouge

1 Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, Maryland 21250, U.S.A.
2 Department of Statistics, Seoul National University, Seoul, 08826, Republic of Korea
3 Department of Biological Sciences, University of Maryland, Baltimore County, Baltimore, Maryland 21250, U.S.A.
4 National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, U.S.A.
5 School of Statistics, University of the Philippines Diliman, Quezon City, 1101, Philippines

*email: jumpark@umbc.edu

Summary. In recent mutation studies, analyses based on protein domain positions are gaining popularity over gene-centric approaches since the latter have limitations in considering the functional context that the position of the mutation provides. This presents a large-scale simultaneous inference problem, with hundreds of hypothesis tests to consider at the same time. This article aims to select significant mutation counts while controlling a given level of Type I error via False Discovery Rate (FDR) procedures. One main assumption is that the mutation counts follow a zero-inflated model in order to account for the true zeros in the count model and the excess zeros. The class of models considered is the Zero-inflated Generalized Poisson (ZIGP) distribution. Furthermore, we assumed that there exists a cut-off value such that smaller counts than this value are generated from the null distribution. We present several data-dependent methods to determine the cut-off value. We also consider a two-stage procedure based on screening process so that the number of mutations exceeding a certain value should be considered as significant mutations. Simulated and protein domain data sets are used to illustrate this procedure in estimation of the empirical null using a mixture of discrete distributions. Overall, while maintaining control of the FDR, the proposed two-stage testing procedure has superior empirical power.

Key words: Local false discovery rate; Protein domain; Zero-in ated generalized poisson

1. Introduction

Interest towards multiple testing procedures has been growing rapidly in the advent of the so-called genomic age. With the breakthrough in large-scale methods to purify, identify, and characterize DNA, RNA, proteins, and other molecules, researchers are becoming increasingly reliant on statistical methods for determining the significance of biological findings (Pollard et al., 2005). Gene-based analyses of cancer data are classic examples of studies which present thousands of genes for simultaneous hypothesis testing. However, Nehrt et al. (2012) reported that gene-centric cancer studies are limited since the functional context that the position of the mutation provides is not considered. In lieu of this, Nehrt et al. (2012) have shown that protein domain level analyses of cancer somatic variants can identify functionally relevant somatic mutations where traditional gene-centric methods fail by focusing on protein domain regions within genes, leveraging the modularity and polyfunctionality of genes.

In protein domain-centric studies, somatic mutations from sequenced tumor samples are mapped from their genomic positions to positions within protein domains, enabling the comparison of distant genomic regions that share similar structure and amino acid composition (Peterson et al., 2013). In the analysis of sequenced tumor samples, it is assumed that the mutational distribution will consist of many “passenger” mutations, which are non-functional randomly distributed background mutations, in addition to rare functional “driver” mutations that reoccur at specific sites within the domain and contribute to the initiation or progression of cancer (Parmigiani et al., 2007). The major interest is in a single domain, how to identify the highly mutated positions compared to the background where the number of positions in a domain can be as large as several hundreds.

Motivated by the aforementioned domain-level analyses, we propose a methodology for identifying significant mutation counts while controlling the rate of false rejections. Dudoit et al. (2003) reported that much of the statistics microarray literature is focused on controlling the probability of a Type I error, a “false discovery.” A traditional approach is to control the family-wise error rate (FWER), the probability of
Empirical Null Estimation Using Zero-Inflated Discrete Mixture Distributions

2. Multiple Testing Procedures Controlling FDR

In this section, we briefly discuss the motivating example and review the existing procedures for analysis. The collection of the original dataset is \( a = (a_1, a_2, \ldots, a_N) \), where \( a_i \) is the number of mutations in the \( i \)th position of the specific domain with \( N \) positions. We define \( \mathcal{A} \) as the set of the unique values of \( a \), \( K = \max(a) \), and \( L \) is the cardinality of \( \mathcal{A} \) where \( L \leq K + 1 \). Some relevant features of \( a \) follow. A large proportion of positions do not have any mutation, \( a_i = 0 \). Also, \( L \) is relatively small compared to \( N \), which means that the number of mutations in many positions are tied. Since our goal is to identify the positions with extra disease mutation counts, it is only reasonable to have the same conclusion for positions wherein the number of mutations are tied. Therefore, we transform the data into the observed “histograph” of positions over “mutation counts.” We define \( n_j = |\{ i : a_i = j \}| \), as the number of positions with \( j \) mutations, \( j \in \mathcal{A} \), where \( \mathcal{A} = \{ j : j \geq 0, n_j > 0 \} \) and \( \sum_{j \in \mathcal{A}} n_j = N \). The ordered data \( x_N \) can be represented as a partition of the unique values of \( a \), that is,

\[
x_N = (x_{K_0}', x_1', \ldots, x_K') = (0, 0, \ldots, 0, 1, 1, \ldots, K, K, \ldots, K)
\]

where \( x_N \) is the column vector containing \( n_j \) of \( j \).

For any single domain of interest, a total of \( L \) mutation counts can be decomposed into two groups, \( A_0 \) and \( A_1 \), where \( A_0 \) is the collection of small number of mutation counts which is considered to be non-significant and \( A_1 \) is the set of large number of mutation counts which consists of significantly mutated positions. Let the prior probabilities of the two groups be \( \pi_0 \) and \( \pi_1 = 1 - \pi_0 \), and assume corresponding densities, \( f_0 \) and \( f_1 \). Denote \( f_0 \) to be the null distribution and \( f_1 \) to be the alternative distribution. Therefore, we consider the problem of testing \( L \) null hypotheses simultaneously, \( H_{0j} \) is true for \( j \in A \). \( |A| = L \), where \( H_{0j} \) is stated as the number of mutations \( j \) is generated from \( f_0 \).

For a given position, the number of mutations follow one of the two distributions \( f_0 \) or \( f_1 \), so the probability density function of the mixture distribution can be represented as

\[
f(a) = \pi_0 f_0(a) + (1 - \pi_0) f_1(a)
\]

and our goal is to identify the positions which have significantly different patterns from the null. For continuous data, Efron (2005) introduced the idea of “zero assumption” where observations around the central peak of the distribution consists mainly of null cases. Using this assumption, \( f_0 \) is estimated using Gaussian quadrature which is based on derivative at the mode. However, such a procedure is not applicable to discrete data. In our problem on discrete data, we introduce the following assumption on the null distribution which plays a key role throughout this article.

Assumption on \( f_0 \):

\[
f(a) = \pi_0 f_0(a) \text{ for } a_i \leq C \text{ for some integer } C
\]

From the assumption, \( a_i \leq C \) are guaranteed to be from \( f_0 \) and \( a_i > C \) are generated from the mixture of \( f_0 \) and \( f_1 \). We will discuss more details about how to choose the value of \( C \) in the next section. We do not have the issue of identifiability in the mixture model (1) since \( f_1 \) has a different support from that of \( f_0 \) from the assumption (2), so any parametric model for \( f_0 \) considered in our article is uniquely identified.

Benjamini and Hochberg (1995) employed a sequential p-value method to determine \( r \) that tells us to reject \( p_1^{(r)}, p_2^{(r)}, \ldots, p_C^{(r)} \), where \( p_1^{(r)}, p_2^{(r)}, \ldots, p_K^{(r)} \) are the ordered observed p-values. Storey (2002) improved the Benjamini-Hochberg (BH) procedure with the inclusion of the estimator of the null proportion, \( \pi_0 \), which indicates that we reject
parameter in (4) to characterize the distribution. Then significantly higher frequency, we can include a zero-inflation modeling 1. If the model, it is overdispersed with variance to mean ratio exceeding 1. A zero-inflated integer-valued random variable where relative to Poisson distributions. Moreover, if the true null distribution is ZIP, it should still be able to capture properties of the zero-inflated variations of NB and GP distributions, such as probability mass and skewness, while keeping the first two moments fixed. Using this result, it is worthwhile to consider ZIGP rather than ZINB given that the mutation count data exhibited both features.

3. Methodology

3.1. Model Specification

Depending on the application, we assume that the mutation counts follow a zero-inflated model in order to account for the true zeros in the count model and the excess zeros. For example, the zero-inflation observed in protein domain data is due to the negative selection of mutations that can not occur in either a healthy cell or a cancer cell. Biologically, zero-inflation is expected due to the many protein positions that are crucial for proper functioning, often leading to a cell that can not perform the functions necessary for life. The class of models considered is the Generalized Poisson (GP) distribution introduced by Consul and Jain (1970), with an additional zero-inflation parameter. Let $T$ be a nonnegative integer-valued random variable where relative to Poisson model, it is overdispersed with variance to mean ratio exceeding 1. If $T \sim GP(\lambda, \theta)$, then the probability mass function can be written as

$$P(T = t) = g(t) = \frac{\lambda(t + \theta)^{t-1}}{t!} e^{-\lambda - \theta} \quad (4)$$

where $0 \leq \theta < 1$ and $\lambda > 0$. If zero is observed with a significantly higher frequency, we can include a zero-inflation parameter in (4) to characterize the distribution. Then $X \sim ZIGP(\eta, \lambda, \theta)$ and the probability that $X = j$, denoted by $f_0(j)$, is

$$f_0(j) = \eta I_{\{0\}}(j) + (1 - \eta) g(j) I_{\{0, 1, 2, \ldots\}}(j)$$

where $j$ is a nonnegative integer, $0 \leq \eta < 1$, $0 \leq \theta < 1$ and $\lambda > 0$. The indicator function $I_S(j)$ is equal to 1 if $j \in S$ and 0 otherwise. Recently, ZIGP models have been found useful for the analysis of heavy-tailed count data with a large proportion of zeros (Famoye and Singh (2006)). The ZIGP model reduces to Zero-Inflated Poisson (ZIP) distribution when $\theta = 0$, Generalized Poisson distribution (GP) when $\eta = 0$ and Poisson distribution when $\eta = 0$ and $\theta = 0$. The ZIP model, first introduced by Lambert (1992), is applied when the count data possess the equality of mean and variance property while taking into consideration the structural zeros and zeros which exist by chance. Meanwhile, the Zero-Inflated Negative Binomial (ZINB) model is widely used for handling data with population heterogeneity which may be caused by the occurrence of excess zeros and the overdispersion due to unobserved heterogeneity (Phang and Loh, 2013). The rationale for choosing ZIGP as a model for $f_0$ was provided by Joe and Zhu (2005). They showed that ZIGP provides a better fit than ZINB when there is a large fraction of zeros and the data is heavily right-skewed. They compared the probabilistic properties of the zero-inflated variations of NB and GP distributions, such as probability mass and skewness, while keeping the first two moments fixed. Using this result, it is worthwhile to consider ZIGP rather than ZINB given that the mutation count data exhibited both features.

3.2. Estimation of $f_0$, $\beta$, and $\pi_0$

From equation (3), the local FDR formulation consists of unknown quantities $f_0$, $\beta$, and $\pi_0$ which must be estimated accordingly. We follow the idea of “zero assumption” in Efron (2005) where $f_0$ is assumed to be normally distributed and Park et al. (2011) which modeled $f_0$ as a mixture of normal distributions.

However, since $f_0$ is unknown in practice, four count models will be compared in order to come up with estimates for the parameters of the null distribution. These models belong to the class of ZIGP distribution, namely, (i) ZIGP (ii) ZIP (iii) Generalized Poisson, and (iv) Poisson distribution. If the true $f_0$ is ZIGP and the model used to estimate $f_0$ is ZIGP then we expect superior results compared to the other three distributions. Moreover, if the true null distribution is ZIP, then we expect better results for ZIP and ZIGP distribution compared to GP and Poisson distribution. This suggests that since ZIGP can characterize overdispersion, even if there is none such as the case of ZIP, it should still be able to capture the behavior of $f_0$ accurately.

To estimate the parameters of $f_0$ for any of these four count models, the EM Algorithm proposed by McLachlan and Jones (1988) will be utilized. For truncated data sets described in (2), fitting the model using EM algorithm is not straightforward as when all data points are available. The details of the EM Algorithm are provided in Web Appendix A.

Moreover, it is straightforward to estimate $f(j)$ by using relative frequency given by $f(j) = n_j/N$. Using the assumption on $f_0$, for $j \in C$, $f(j)$ from (1) reduces to $\pi_0 f_0(j)$. Hence, $\sum_{j \in C} \pi_0 f_0(j) = \sum_{j \in C} f(j)$. Finally, the estimate of $\pi_0$ is

$$\pi_0 = \frac{\sum_{j \in C} \hat{f}(j)}{\sum_{j \in C} f_0(j)}$$

where $\hat{f}(j) = n_j/N$ and the estimate of $f_0$ after plugging in $\hat{\beta}$ resulting from the EM algorithm.
3.3. Choice of the Cut-off C

In our model, we assume that we can identify a cut-off C, wherein positions with number of mutations greater than C contain more mutations than what would be expected in the null model. The choice of the cut-off C is of paramount importance since the estimation of \( f_0 \) and \( \pi_0 \) depend on C. It is more realistic to assume that C is unknown, so such a pre-determined C may affect the result of local FDR procedure seriously. In particular, if C is predetermined and is chosen to be larger than the true value, the null distribution is estimated based on observations from alternative hypothesis as well as null hypothesis, so the estimated null distribution is contaminated by the alternative distribution. This will cause insensitivity of local FDR procedure in detecting the alternative hypothesis. On the other hand, if C is chosen to be smaller, then the null distribution is estimated only based on small values, so the estimation of the null distribution especially at the tail part is less reliable. Empirically, the FDR procedure yields liberal results in that there are too many rejections resulting in failure in controlling a given level of FDR.

The estimation of the cut-off C has been formulated using the likelihood function. Define the index sets \( \mathcal{A} = \{ j : j \geq 0, n_j > 0 \}, \mathcal{A}(C) = \{ j : 0 \leq j < C, n_j > 0, f_C(j) = 0 \}. \) Note that \( \mathcal{A}(C_1) \subset \mathcal{A}(C_2) \) for \( C_1 < C_2 \) and \( f(j) = \pi_0 f_0(j) \) when \( j \in \mathcal{A}(C) \). The likelihood function of \((0,n_0), \ldots, (K,n_K)\) for a given \( \mathcal{A}(v) \) is

\[
L(\Theta^*, f) = \prod_{j \in K} f(j)^{n_j} = \prod_{j \in \mathcal{A}(v)} (\pi_0 f_0(j))^{n_j} \prod_{j \notin \mathcal{A}(v)} f(j)^{n_j}
\]

where \( \pi_0 f_0 \) depends on \( \Theta^* = (\pi_0, \eta, \theta, \lambda) = [\pi_0] \cup \Theta \). The log likelihood is also

\[
\log L(\Theta^*, f) = \ell_v(\Theta^*, f) = \sum_{j \in \mathcal{A}(v)} n_j \log(\pi_0 f_0(j)) + \sum_{j \notin \mathcal{A}(v)} n_j \log f(j)
\]

since \( f(j) = \pi_0 f_0(j) \) for \( j \in \mathcal{A}(v) \). This leads to

\[
\ell_v(\Theta^*, f) = \sum_{j \in \mathcal{A}(v)} n_j \log \frac{\pi_0 f_0(j)}{f(j)} + \sum_{j \notin \mathcal{A}(v)} n_j \log f(j)
\]

We adopt the idea of sequential testing to detect the change point in which the observations are generated from the mixture distribution \( f \). More specifically, suppose we observed \((0,n_0),(1,n_2), \ldots, (K,n_K)\) sequentially from \( f_0(0), f_0(1), \ldots, f_0(C), f(C+1), \ldots, f(K) \) where distribution is changed from \( f_0 \) to \( f \) at \( C+1 \). Our goal is to detect the change point \( C \) based on assuming that we observe \( 0, 1, 2, \ldots, K \) sequentially. For a given \( v \), we define \( S_v(\Theta, f) \) as

\[
S_v(\Theta, f) = \sum_{j \leq v} n_j \log f_0(j) + \sum_{j \geq v+1} n_j \log f(j)
\]

Maximizing \( S_v(\Theta, f) \) is equivalent to the CUSUM(cumulative sum) \( \sum_{j \leq v} n_j \log \frac{f_0(j)}{f(j)} \). Since the parameters \( \Theta \) is estimated from EM algorithm and \( \bar{f}(j) = \frac{n_j}{N} \), our procedure is

\[
\hat{C} = \arg \max_{v=1,2,\ldots,K} S_v(\hat{\Theta}_v)
\]

where \( \hat{\Theta}_v \) is the estimator from the EM algorithm with the value of \( C \) set to \( v \). One may consider the full likelihood of all observations and find out some connection between \( S_v \) and the full likelihood presented as follows:

\[
S_v(\Theta^*, f) = \ell_v(\Theta^*, f) - N_v \log \pi_0 - \sum_{j \in \mathcal{A}(v)} n_j \log f_0(j) - \sum_{j \notin \mathcal{A}(v)} n_j \log f(j)
\]

where \( N_v = \sum_{j \leq v} n_j \) and \( \ell_v(\Theta^*, f) = \log \pi_0 f_0(j) - f(j) \).

Sheetlin et al. (2011) offer an objective-change point method that can replace the subjective approaches performed by eye-balling the data. Their proposed method resembles the change-point regression and robust regression but it is tailored to estimate the change point from a transient to an asymptotic regime. Given a tuning parameter \( c \) and a criterion function \( \rho \), depending on \( \beta \), the estimator for the change point \( k^* \) is defined as

\[
k^* = \arg \min_{k=0,1,\ldots,n} \left( \min_{\beta} \sum_{i=k+1}^n \left( \rho(\epsilon_i) - c \right) \right)
\]

where \( \rho(\epsilon) \) is the estimated least-squares normalized residual. In (5), there is a tuning parameter \( c \) which should be given ahead. The value of \( c \) plays the role of penalty for adding terms \( \rho(\epsilon) \) in (5), so the predetermined value of \( c \) affects \( k^* \) arbitrarily. We see that our proposed estimation of \( C \) is related to the form (5). We estimate \( C \) via

\[
\hat{C}_1 = \arg \min_{v=1,2,\ldots,k} \sum_{j \notin \mathcal{A}(C)} n_j (\rho(\hat{\Theta}_v^*, \hat{f})(j) - \bar{c}_v)
\]

where \( \hat{\Theta}_v^* = (\hat{\pi}_0, \hat{\eta}, \hat{\Theta}_v, \hat{\lambda}) \) is obtained from the EM algorithm discussed in the previous section, \( \bar{f}(j) = \frac{n_j}{N} \), \( \rho_j(\Theta^*, f) = -\ell_j(\Theta^*, f) \) and \( \bar{c}_v = -\log \hat{\pi}_0 \). In (5), \( c \) is a pre-determined value, however we don’t need to predetermine any parameter in (5). The proposed criterion (6) is related to the penalized model selection such as AIC and BIC. When we use the information that \( n = \sum_{j \leq v} n_j \) observed values are generated from \( f_0, -\sum_{j \leq v} n_j \ell_j(\Theta^*, f) \) is increasing in \( v \), there is a compromise term \( c = -\log \pi_0 \) for each observation to compensate adding additional terms. There is a total of \( N_v \) positions, so when we use the assumption \( v = C \), we consider \( N_v \log \pi_0 \)
penalty to the log likelihood function $\ell_*$. Most of well known model selection criteria have similar forms where the penalty terms are related to penalize the complexity of models. In our context, the term $-\log \pi_0$ gives penalty to using the information that $j$ for $j \leq v$ are generated from $f_0$. For a small value of $\pi_0$, the corresponding penalty ($-\log \pi_0$) is large since a large penalty should be given to a low chance of $f_0$. On the other hand, if $\pi_0$ is close to 1, there becomes small risk from assuming observations are from the null hypothesis.

For the second method, we consider the extension of the methodology proposed by Efron (2007) which explicitly uses the zero assumption. This stipulates that the non-null density $f_1$ is supported outside some set $\{0, 1, \ldots, C\}$. The likelihood function for $x_n = (x_0, \ldots, x_C)$ is defined as

$$L(\hat{\Theta}_n^* | x_n) = \xi^n (1 - \xi)^{\sum_{j \geq v} f_0(j)}$$

where $\xi = \sum_{j \geq v} f_0(j)$. The cut-off can be computed as

$$\hat{C}_2 = \arg\min_{i=1, 2, \ldots, K} (\log L(\hat{\Theta}_n^* | x_n))$$

### 3.4. Modification of Local FDR by Truncation

In practice, if a given domain position has a large number of mutations, then these mutations are expected to be significant. In many cases, there are relatively few positions in a protein domain where large values of mutations can be observed. This indicates that for large values of $j$, estimation of $f$ based on relative frequency is not accurate due to the sparse data in the tail part. Consequently, the estimated local FDR is not reliable since it depends on the estimator of $f$. Rather than testing significance based on inaccurate local FDRs from large mutation counts, we consider a screening process so that the number of mutations exceeding a certain value should be considered as significant mutations. Such a critical value will be decided depending on the estimated null distribution. When we have observations $a_i$ for $1 \leq i \leq N$ generated from the null distribution, we are interested in figuring out $D_N$ such that

$$P_{H_0} \left( \max_{1 \leq i \leq N} a_i < D_N \right) \rightarrow 1$$

as $N \rightarrow \infty$. Once a sequence $D_N$ is identified, $a_i(\geq D_N)$ is hardly observed under the null hypothesis, so the corresponding null hypothesis is rejected directly rather than making decision based on local FDR procedure. There are many choices of $D_N$, but a smaller sequence of $D_N$ satisfying (8) is of our interest since any sequence $B_N$ satisfying $B_N > D_N$ also satisfies the property. The details of the calculation of $D_N$ are provided in Web Appendix B.

The calculation of $D_N$ when $f_0$ is modeled using ZIP or Generalized Poisson is similar since the derivation will eventually yield leading terms which does not involve $\eta$. Likewise, the calculation of $D_N$ when $f_0$ is modeled using either ZIP or Poisson is the same. On the other hand, since the true values

of the parameters $\lambda$ and $\theta$ are unknown, we calculate $D_N$ using the estimates $\hat{\lambda}$ and $\hat{\theta}$. Hence, $D_N$ can be calculated as

$$D_N = [\max(D_1, D_2)]$$

where $[x]$ is the smallest integer greater than or equal to $x(x > 0)$. $D_1$ and $D_2$ are presented in Section 3.3 for the choice of cut-off. The true null distribution account for the excessive number of zeros which is a characteristic of the mutation count data. Following the key assumption on $f_0$, the support of $f_1$ does not contain values in $[0, C]$. Hence, $f_1$ can be expressed as $f_1 = C + 1 + W$ where $W$ follows another count model. For the model specification of $W$, Geometric($p = 0.08$) and Binomial($n = 250, p = 0.20$) distribution are utilized. They exhibit the pattern of the mutation count observed in the real data set.
Using these model specifications in terms of the true $f_0$ and $f_1$, there are 15 mixture models considered for data generation as presented in Web Table S1. For each of the specification of $f_0$, $\hat{C}$ is calculated and the corresponding set of parameter estimates $\Theta_k^* = (\tau_k, \lambda_k, \theta_k, \tau_0, \hat{\alpha}_k)$ from EM Algorithm are obtained. Results show that regardless of the true null distribution, if $f_0$ is modeled using the ZIGP distribution, then $\hat{C}$ produces the most accurate estimate for $\hat{C}$ in terms of the bias and standard error. This validates the robustness of ZIGP as a model for $f_0$, that is, even if the true null distribution is ZIP, the most accurate estimate of $\hat{C}$ can still be observed when $f_0$ is modeled using ZIGP. The bias of the parameter estimates for each of these specifications are presented in Web Tables S2–S5 in the Supplementary Materials.

A total of $L$ hypotheses tests were performed for independent random variables $n_j$ over 1000 replications. For each replication, the proportion of $n_j$ from the null distribution is set to be $\pi_0$ and the total number of positions $N$ is specified to be 1000. To calculate the False Discovery Rate, $\hat{\text{FDR}}$, for the $k$th generated data, $k = 1, 2, \ldots, 1000$, we compute the False Discovery Proportion (FDP) which is defined by

$$ FDP_k = \frac{V_k}{R_k} I(R_k > 0) $$

where $V_k$ and $R_k$ are the number of falsely rejected hypotheses (false discoveries) and the total number of rejected hypotheses in the $k$th generated data, respectively. FDR is the expected value of the false discovery proportion and can be computed empirically as

$$ \hat{\text{FDR}} = \frac{1}{1000} \sum_{k=1}^{1000} \frac{V_k}{R_k} I(R_k > 0) $$

In our simulations, the decision rule is to reject the null $H_0(j)$ if $\hat{\text{FDR}}(j) = \pi_0 f_0(j)/\hat{f}(j) \leq \alpha$. Throughout the simulations, we consider the level of significance $\alpha = 0.05$. The True Positive Rate, $\hat{\text{TPR}}$ is computed empirically as

$$ \hat{\text{TPR}} = \frac{1}{1000} \sum_{k=1}^{1000} \frac{S_k}{S_k + T_k} $$

where $S_k$ and $T_k$ are the number of correctly rejected hypotheses (true discoveries) and the number of falsely accepted hypotheses (false non-discoveries) in the $k$th generated data, respectively. Three procedures are compared in terms of controlling $\hat{\text{FDR}}$ and the value of $\hat{\text{TPR}}$, namely the one-stage local FDR procedure, the proposed two-stage procedure and Storey’s procedure. The results which yields the superior $\hat{\text{TPR}}$ while controlling FDR are highlighted using bold face quantities.

As displayed in Figure 1a–c, the non-null distribution is Geometric, the proportion of null cases is 0.80 and the fraction of zeros is 0.80. The degree to which the null model is mixed with the non-null model is described using the three cases. ZIP1 represents the well-separated case, ZIGP1 is the moderately mixed case while ZIGP2 can be described as the heavily mixed case. The corresponding numerical comparison is shown in Table 1.

Based upon the results of Table 1, since null and non-null distribution is moderately mixed for ZIGP1, the resulting $\hat{\text{TPR}}$ for all three procedures is substantially higher than the $\hat{\text{TPR}}$ for ZIGP2, regardless of the model used for the estimation of $f_0$. Given that $\hat{\text{FDR}}$ is controlled in all procedures, if the model for $f_0$ is ZIGP, then the Two-Stage procedure yields the highest $\hat{\text{TPR}}$.

This suggests that the proposed procedure is better than the other existing procedures. Meanwhile, due to the “well-separation” if the true null is ZIP1, then the $\hat{\text{TPR}}$ for ZIP1 is slightly higher than the $\hat{\text{TPR}}$ for ZIGP1. Moreover, the $\hat{\text{FDR}}$ for all three procedures for ZIP1 are noticeably lower than the $\hat{\text{FDR}}$ for ZIGP1. This means that the number of rejections for ZIGP1 and ZIP1 are almost the same but there are more erroneous rejections for ZIGP1. This result can be explained by the presence of overdispersion in ZIGP1, thereby suggesting that the presence of overdispersion in the data could lead to erroneous rejections. Overall, there are more rejections using $\hat{C}_1$ as a cut-off compared to $\hat{C}_2$. Also, even if $\hat{C}_1$ yields more rejections, it still controls the value of FDR demonstrating the superiority of $\hat{C}_1$ as a cut-off method.

Figure 1d–f also presents the histograms when the non-null distribution is Binomial, the proportion of null cases is 0.35 and the fraction of zeros is 0.40. Unlike the parametrization of the Geometric non-null distribution which appears to be skewed to the right, this $f_1$ exhibits near symmetry. In terms of the mixing of the null and non-null distribution, ZIGP2 represents the well-separated case, ZIGP3 is the moderately mixed case while ZIGP4 can be described as the heavily mixed case. The numerical comparison is shown in Table 2.

According to Table 2, when $f_0$ is modeled using ZIGP, using $\hat{C}_1$ as a cut-off yielded many more rejections, regardless of the procedure used. This suggests that the extension of Efron’s method is conservative and would miss significant positions. The difference between $\hat{C}_1$ and $\hat{C}_2$ is further highlighted for ZIGP4, where the true null distribution is heavily mixed with the non-null distribution and overdispersion is present. For scenarios where $\hat{\text{FDR}}$ is controlled for ZIP2 and ZIGP3, using Two-Stage procedure leads to the highest $\hat{\text{TPR}}$. However, for ZIGP4 where overdispersion is evident and $f_0$ is heavily mixed with $f_1$, the value of $\hat{\text{TPR}}$ is substantially higher using Storey’s procedure, while keeping the $\hat{\text{FDR}}$ controlled.

Another scenario considered is when the true non-null distribution is Geometric, the proportion of null cases is 0.85 but the fraction of zeros is 0.40. Unlike the scenario presented in Table 1 and Figure 1, the specified proportion of zeros is reduced to half. The interest is to determine whether there would be a change in pattern should there be a significant decrease in the number of positions without a mutation. The histograms and the corresponding numerical comparison are presented in Web Figure S1 and Web Table S6 in the Supplementary section. It can be noted that regardless of the magnitude of the fraction of zeros, a similar pattern can be observed in terms of the superiority of $\hat{C}_1$ as a method for choosing $C$. Among the procedures, the $\hat{\text{TPR}}$ is consistently highest for the Two-Stage procedure, given that the $\hat{\text{FDR}}$ is
Figure 1. Histogram of different scenarios for comparison when $C = 5$. (a) The first panel in the first row: $f_0$ is $\text{ZIP}_1(\eta = 0.80, \lambda = 1.5)$ which represents the well-separated case, $f_1$ is shifted Geometric($p = 0.08$), $\pi_0 = 0.80$ (b) The second panel in the first row: $f_0$ is $\text{ZIGP}_1(\eta = 0.80, \lambda = 1.5, \theta = 0.3)$ which represents the moderately mixed case, $f_1$ is shifted Geometric($p = 0.08$), $\pi_0 = 0.80$ (c) The first panel in the third row: $f_0$ is $\text{ZIGP}_2(\eta = 0.80, \lambda = 3, \theta = 0.3)$ which represents the heavily mixed case, $f_1$ is shifted Geometric($p = 0.08$), $\pi_0 = 0.80$ (d) The second panel in the second row: $f_0$ is $\text{ZIP}_2(\eta = 0.40, \lambda = 1.5)$ which represents the well-separated case, $f_1$ is shifted Binomial($n = 250, p = 0.20$), $\pi_0 = 0.35$ (e) The first panel in the third row: $f_0$ is $\text{ZIGP}_3(\eta = 0.40, \lambda = 1, \theta = 0.20)$ which represents the moderately mixed case, $f_1$ is shifted Binomial($n = 250, p = 0.20$), $\pi_0 = 0.35$ (f) The second panel in the third row: $f_0$ is $\text{ZIGP}_4(\eta = 0.40, \lambda = 3, \theta = 0.20)$ which represents the heavily mixed case, $f_1$ is shifted Binomial($n = 250, p = 0.20$), $\pi_0 = 0.35$. 
Table 1
Numerical Comparison when \( f_1 \) is shifted Geometric \((p = 0.08), \pi_0 = 0.80, \) and \( C = 5. \) ZIGP \((\eta = 0.80, \lambda = 1.5) \) represents the well-separated case, \( \text{ZIP} \) \((\eta = 0.80, \lambda = 0.3) \) represents the moderately mixed case, and \( \text{ZIGP} \) \((\eta = 0.80, \lambda = 3, \theta = 0.3) \) represents the heavily mixed case. The number in ( ) corresponds to the standard error. The results which yields the superior TPR while controlling FDR are highlighted using bold face quantities.

| True Model \( f_0 \) | Choice of Model \( C \) for \( f_0 \) | Two-stage procedure | One-stage procedure | Storey’s FDR |
|----------------------|-----------------------------------------------|-------------------|-------------------|--------------|
|                      | \( R \) | \( \hat{FDR} \) | \( \hat{TPR} \) | \( R \) | \( \hat{FDR} \) | \( \hat{TPR} \) | \( R \) | \( \hat{FDR} \) | \( \hat{TPR} \) |
| \( \text{ZIP}_1 \)  | \( \hat{C}_1 \) ZIGP | 200.91 | 0.0042 | 1.000 | 187.88 | 0.0009 | 0.938 | 182.91 | 0.0005 | 0.914 |
|                     | (12.86) | (0.0064) | (0.000) | (13.58) | (0.0023) | (0.025) | (13.90) | (0.0017) | (0.037) |
| \( \text{ZIP} \)    | \( \hat{C}_2 \) ZIGP | 198.27 | 0.0027 | 0.988 | 198.25 | 0.0027 | 0.988 | 191.89 | 0.0017 | 0.957 |
|                     | (14.20) | (0.0040) | (0.024) | (14.23) | (0.0040) | (0.024) | (14.65) | (0.0035) | (0.037) |
| \( \text{GP} \)     | \( \hat{C}_1 \) ZIGP | 0.16  | 0.0000 | 0.001 | 1.14  | 0.0000 | 0.001 | 0.14  | 0.0000 | 0.001 |
|                     | (4.93)  | (0.0000) | (0.032) | (4.46)  | (0.0000) | (0.029) | (4.46)  | (0.0000) | (0.029) |
| \( \text{P} \)      | \( \hat{C}_2 \) ZIGP | 236.75 | 0.1513 | 1.000 | 236.75 | 0.1513 | 1.000 | 219.30 | 0.0864 | 1.000 |
|                     | (20.21) | (0.0623) | (0.000) | (20.21) | (0.0623) | (0.000) | (13.99) | (0.0486) | (0.000) |

| \( \text{ZIP}_1 \)  | \( \hat{C}_1 \) ZIGP | 194.00 | 0.0420 | 0.928 | 179.77 | 0.0278 | 0.878 | 170.23 | 0.0202 | 0.834 |
|                     | (39.02) | (0.0207) | (0.175) | (36.31) | (0.0153) | (0.164) | (31.40) | (0.0125) | (0.144) |
| \( \text{ZIP} \)    | \( \hat{C}_2 \) ZIGP | 208.30 | 0.0523 | 0.986 | 208.12 | 0.0521 | 0.986 | 198.94 | 0.0416 | 0.952 |
|                     | (15.30) | (0.0211) | (0.027) | (15.58) | (0.0213) | (0.028) | (16.93) | (0.0200) | (0.044) |
| \( \text{GP} \)     | \( \hat{C}_1 \) ZIGP | 0.00  | 0.0000 | 0.000 | 0.00  | 0.0000 | 0.000 | 0.00  | 0.0000 | 0.000 |
|                     | (0.00)  | (0.0000) | (0.000) | (0.00)  | (0.0000) | (0.000) | (0.00)  | (0.0000) | (0.000) |
| \( \text{P} \)      | \( \hat{C}_2 \) ZIGP | 258.86 | 0.2255 | 1.000 | 258.86 | 0.2255 | 1.000 | 242.28 | 0.1728 | 1.000 |
|                     | (17.02) | (0.0472) | (0.000) | (17.02) | (0.0472) | (0.000) | (16.01) | (0.0463) | (0.000) |

| \( \text{ZIP}_2 \)  | \( \hat{C}_1 \) ZIGP | 50.36  | 0.0184 | 0.237 | 50.23  | 0.0182 | 0.237 | 58.05  | 0.0148 | 0.279 |
|                     | (72.57) | (0.0341) | (0.035) | (72.40) | (0.0339) | (0.035) | (62.16) | (0.0274) | (0.070) |
| \( \text{ZIP} \)    | \( \hat{C}_2 \) ZIGP | 227.11 | 0.1581 | 0.954 | 211.90 | 0.1340 | 0.916 | 192.28 | 0.1032 | 0.858 |
|                     | (19.67) | (0.0361) | (0.032) | (32.17) | (0.0532) | (0.072) | (26.79) | (0.0428) | (0.073) |
| \( \text{GP} \)     | \( \hat{C}_1 \) ZIGP | 19.35  | 0.3140 | 0.68  | 16.45  | 0.1946 | 0.068 | 14.35  | 0.1483 | 0.063 |
|                     | (71.88) | (0.0401) | (0.252) | (61.04) | (0.0264) | (0.252) | (53.25) | (0.0228) | (0.232) |
| \( \text{P} \)      | \( \hat{C}_2 \) ZIGP | 313.13 | 0.3606 | 1.000 | 313.14 | 0.3605 | 1.000 | 310.41 | 0.3553 | 1.000 |
|                     | (16.02) | (0.0349) | (0.000) | (16.00) | (0.0349) | (0.000) | (15.11) | (0.0296) | (0.000) |
Table 2
Numerical Comparison when $f_1$ is shifted Binomial($n = 250, p = 0.20$), $\pi_0 = 0.35$ and $C = 5$. ZIGP$_2(\eta = 0.40, \lambda = 1.5)$ is the well-separated case, ZIGP$_2(\eta = 0.40, \lambda = 1.0, \theta = 0.30)$ represents the moderately mixed case, and ZIGP$_4(\eta = 0.40, \lambda = 4, \theta = 0.30)$ represents the heavily mixed case. The number in (·) represents the standard error. The results which yields the superior TPR while controlling FDR are highlighted using bold face quantities.

| True Choice of C for $f_0$ | Two-stage procedure | One-stage procedure | Storey’s FDR |
|---------------------------|---------------------|---------------------|--------------|
|                           | $R$ | $\hat{FDR}$ | $\hat{TPR}$ | $R$ | $\hat{FDR}$ | $\hat{TPR}$ | $R$ | $\hat{FDR}$ | $\hat{TPR}$ |
| ZIP$_2$ ZIGP $\hat{C}_1$  | 643.15 | **0.0015** | **1.000** | 631.41 | 0.0003 | 0.983 | 632.95 | 0.0003 | 0.985 |
|                           | (15.32) | (0.0016) | (0.001) | (16.20) | (0.0007) | (0.007) | (16.00) | (0.0008) | (0.007) |
| ZIP 635.66 | 0.0006 | 0.989 | 635.62 | 0.0006 | 0.989 | 638.45 | 0.0009 | 0.993 |
| GP 1.85 | 0.0000 | 0.003 | 1.85 | 0.0000 | 0.003 | 123.76 | 0.0000 | 0.192 |
| P 644.35 | 0.0033 | 1.000 | 644.35 | 0.0033 | 1.000 | 646.02 | 0.0058 | 1.000 |
|                           | (15.21) | (0.0046) | (0.000) | (15.21) | (0.0046) | (0.000) | (15.23) | (0.0039) | (0.000) |
| ZIP$_3$ ZIGP $\hat{C}_2$  | 620.41 | 0.0018 | 0.964 | 608.78 | 0.0002 | 0.948 | 614.49 | 0.0003 | 0.956 |
|                           | (99.01) | (0.0022) | (0.152) | (96.45) | (0.0006) | (0.148) | (78.24) | (0.0007) | (0.119) |
| ZIP 636.53 | 0.0007 | 0.990 | 635.35 | 0.0007 | 0.988 | 637.63 | 0.0009 | 0.992 |
| GP 0.37 | 0.0000 | 0.001 | 0.37 | 0.0000 | 0.001 | 106.13 | 0.0000 | 0.165 |
| P 659.04 | 0.0259 | 1.000 | 659.04 | 0.0259 | 1.000 | 657.95 | 0.0243 | 1.000 |
|                           | (15.23) | (0.0049) | (0.000) | (15.23) | (0.0049) | (0.000) | (15.22) | (0.0039) | (0.000) |
| ZIP$_4$ ZIGP $\hat{C}_2$  | 506.01 | 0.0064 | 0.782 | 497.86 | 0.0002 | 0.787 | 509.59 | 0.0037 | 0.790 |
|                           | (185.12) | (0.0061) | (0.283) | (178.91) | (0.0037) | (0.276) | (158.92) | (0.0040) | (0.244) |
| ZIP 648.00 | 0.0110 | 0.998 | 647.92 | 0.0110 | 0.998 | 649.10 | 0.0125 | 0.998 |
| GP 0.00 | 0.0000 | 0.000 | 0.00 | 0.0000 | 0.000 | 68.95 | 0.0000 | 0.107 |
| P 659.07 | 0.0259 | 1.000 | 659.07 | 0.0259 | 1.000 | 658.00 | 0.0244 | 1.000 |
|                           | (15.49) | (0.0111) | (0.000) | (15.49) | (0.0111) | (0.000) | (15.60) | (0.0098) | (0.000) |
| ZIP$_4$ ZIGP $\hat{C}_2$  | 256.82 | 0.0086 | 0.390 | 256.80 | 0.0086 | 0.390 | 352.65 | **0.0092** | **0.538** |
|                           | (230.14) | (0.0157) | (0.345) | (230.15) | (0.0157) | (0.345) | (163.87) | (0.0155) | (0.241) |
| ZIP 671.07 | 0.0724 | 0.964 | 582.86 | 0.0257 | 0.877 | 588.60 | 0.0271 | 0.888 |
| GP 0.04 | 0.0000 | 0.000 | 0.04 | 0.0000 | 0.000 | 90.70 | 0.0001 | 0.140 |
| P 794.02 | 0.1876 | 1.000 | 794.02 | 0.1876 | 1.000 | 772.11 | 0.1645 | 1.000 |
|                           | (15.34) | (0.0184) | (0.000) | (15.34) | (0.0184) | (0.000) | (16.62) | (0.0179) | (0.000) |
controlled. Also, even if the true model is ZIP, using ZIGP to model $f_0$ produced better results in terms of $\hat{TPR}$, while keeping $\hat{FDR}$ controlled.

Finally, the last scenario considered is when the non-null distribution is Binomial, the fraction of zeros is still 0.40 but the proportion of null cases is increased to 0.80. When the true $f_0$ exhibits near symmetry, the goal is to determine whether there would be a change in pattern should there be a significant increase in the number of positions without a mutation. The histograms and the corresponding numerical comparison are presented in Web Figure S2 and Web Table S7 in the Supplementary section. Results revealed that the difference between $\hat{C}_1$ and $\hat{C}_2$ is apparent when there is overdispersion and $f_0$ is heavily mixed with $f_1$. If ZIGP is the model used for the estimation of $f_0$, the value of $\hat{TPR}$ is substantially higher using $\hat{C}_1$, while keeping the $\hat{FDR}$ controlled. Moreover, the resulting $\hat{TPR}$ for the moderately mixed case is substantially higher than the $\hat{TPR}$ for the heavily mixed case, regardless of the model used for the estimation of $f_0$ and the procedure employed. However, given that $\hat{FDR}$ is controlled by specifying either of the two models, using ZIGP leads to a higher $\hat{TPR}$ than when the true model ZIP is specified. This result implies using ZIGP would yield satisfactory results even under model misspecification.

Taking everything into account, for the well-separated and moderately mixed case, if the null model is correctly specified and $\hat{FDR}$ is controlled, using the Two-Stage procedure yields $\hat{FDR}$ closest to the nominal level $\alpha$. Consequently, the Two-Stage procedure is superior in terms of $\hat{TPR}$ in most cases. If the true null model is ZIP and the null model is correctly specified, $\hat{FDR}$ is controlled in all procedures. However, the Two-Stage procedure is better than the local FDR procedure and Storey’s procedure in terms of $\hat{TPR}$. It can also be noted that if the true model is ZIP and ZIP is used to model the null distribution, then the Two-Stage Procedure still yields the closest $\hat{FDR}$ to $\alpha$ and leads to higher $\hat{TPR}$ as compared to the other procedures. This implies using the Two-Stage Procedure when the null model is misspecified would still produce satisfactory results. Moreover, regardless of the shape of the non-null distribution, the Two-Stage Procedure yields better results than the other procedures. Additionally, the condition (2) can be weakened so that mutation counts belonging in $[0, C]$ can be contaminated by $f_1$ with probability close to zero. In fact, if there are alternative observations below $C$, then the true $C$ is changed to be a smaller value (say $C'$) where all observations below $C'$ are actually from $f_0$. Intuitively, we can posit that this may affect the estimates of null parameters and the TPR. To address this concern, we perform additional simulation studies using $C = 1$ which represents that small values of data can be observed from the alternative. The bias and standard error of the parameter estimates are provided in Web Tables S8–S11. Results show that the estimates of null parameters tend to have large biases and standard errors compared to those for $C = 5$ in Web Tables S2–S7. Furthermore, the corresponding numerical comparisons of FDR and TPR were also provided in Web Tables S12–S15. When there are alternative observations for a small value of $C$, the situations become more heavily mixed case, so this results in lower TPR while FDRs are still controlled.

4.2. Application to Protein Domain Data

In the study of cancer at the molecular level, it is important to understand which somatic mutations contribute to tumor initiation or progression in order to develop new treatments or to identify patients for which a given treatment will be effective. In this field, some researchers focus on the patterns of somatic variants on protein domains due to the well-defined function and structure of these units, which can lead to a better understanding of how these functions are disrupted in cancer (Nehrt et al., 2012; Peterson et al., 2010). Here, one interesting issue is identifying “protein domain hotspots”, or positions within domains that are found to be mutated frequently (Peterson et al., 2012). It is among a fixed number of positions in a single domain, which ones are significantly different from the majority.

In application to cancer, Peterson et al. (2012) discovered that known cancer variants tended to cluster at specific positions more than variants involved with unrelated diseases, suggesting that protein domain hotspots could be useful in understanding the molecular mechanisms involved with cancer. It is a novel solution for the application of protein domain hotspots to somatic variants in sequenced tumor samples, which has the potential to distinguish between driver variants that contribute to cancer and passenger variants that do not contribute and are assumed to be distributed randomly.

As an example, we analyze the mutation data obtained from the tumors of 5848 patients from The Cancer Genome Atlas (TCGA) data portal (http://tcga-data.nci.nih.gov/tcga/, Collins and Barker, 2007). These were mapped to specific positions within protein domain models to identify clusters. TCGA MAF files were obtained for 20 cancer types.

Among several hundreds of domains, we focus on six functionally well-known domains to identify the hotspots of somatic variants in TCGA sequenced tumor samples. We start with the hotspots on growth factors (cd00031), which are known to harbor reoccurring somatic mutations involved with clonal expansion, invasion across tissue barriers, and colonization of distant niches (Jeanes et al., 2008). Furthermore, protein kinases (cd00180) and RAS-Like GTPase family of genes (cd00882) are well-known for their role in regulating pathways important to cancer (Tsatsanis and Spanidios, 2000). Genes with kinases or RAS-like GTPases are expected to harbor driver mutations that reoccur at specific sites since they are classic examples of proto-oncogenes that mutate into oncogenes, contributing to cancer (Anderson et al., 1992). Additionally, we identify hotspots on ankyrin domains (cd00204), which play a role in mediating protein-protein interactions important in cancer (Imaoka et al., 2014). Also, we find hotspots on transmembrane domains of proteins that are known to be involved with signal transduction, which is relevant in controlling processes involved with cancer (Sever and Brugge, 2015) and experimental evidence confirms the important regulatory role played by membrane proteins in cancer Neuhaus et al. (2009).

Since the mutation counts are discrete, we apply our proposed method based on various discrete models, such as ZIGP, ZIP, Generalized Poisson and Poisson distribution for $f_0$. The estimated parameters based on those models are reported in Web Table S16 in the Supplementary Section. Figure 2 shows
the distribution of each protein domain and its total number of positions. In order to assess the general goodness-of-fit using these different parametric models, we estimate the probabilities for $a_i \leq \hat{C}$ and compared these estimated values with relative frequencies. The results are presented in Web Figures S7–S12 in the Web Supplementary Materials.

The identified number of positions which are mutated differently from expected are in Table 3. For example, when assume that $f_0$ follows ZIGP, the results show that the identified hotspots on growth factor domain (cd00031) based on One-stage and Two-stage procedures are 143 positions among a total of 366 positions, using $\hat{C}_2$. On the other hand, the
Empirical Null Estimation Using Zero-Inflated Discrete Mixture Distributions

Table 3

Comparison of Number of Rejections for Protein Domain Data

| Data    | Method | One-stage procedure | Two-stage procedure | Storey’s FDR |
|---------|--------|---------------------|---------------------|--------------|
|         |        | ZIGP | ZIP | GP | P | ZIGP | ZIP | GP | P | ZIGP | ZIP | GP | P |
| cd00031 | $\hat{C}_1$ | 191 | 212 | 140 | 212 | | | | | | | 201 | 212 | 141 | 212 | 200 | 211 | 154 | 211 |
|         | $\hat{C}_2$ | 143 | 205 | 16 | 212 | | | | | | | | 143 | 205 | 17 | 212 | 162 | 204 | 85 | 211 |
| cd00054 | $\hat{C}_1$ | 45 | 45 | 42 | 45 | | | | | | | | 45 | 45 | 42 | 45 | 25 | 25 | 25 | 25 |
|         | $\hat{C}_2$ | 45 | 45 | 26 | 45 | | | | | | | | 45 | 45 | 26 | 45 | 25 | 25 | 20 | 25 |
| cd00180 | $\hat{C}_1$ | 254 | 288 | 0 | 304 | | | | | | | | 254 | 270 | 5 | 284 | |
|         | $\hat{C}_2$ | 63 | 288 | 0 | 304 | | | | | | | | 63 | 288 | 1 | 304 | 122 | 284 | 0 | 284 |
| cd00240 | $\hat{C}_1$ | 129 | 130 | 19 | 130 | | | | | | | | 129 | 130 | 20 | 130 | 125 | 128 | 60 | 128 |
|         | $\hat{C}_2$ | 129 | 130 | 12 | 130 | | | | | | | | 130 | 130 | 13 | 130 | 125 | 128 | 52 | 128 |
| cd00882 | $\hat{C}_1$ | 148 | 155 | 148 | 169 | | | | | | | | 155 | 155 | 155 | 169 | 147 | 154 | 147 | 160 |
|         | $\hat{C}_2$ | 148 | 155 | 148 | 169 | | | | | | | | 161 | 155 | 155 | 169 | 147 | 154 | 147 | 160 |
| pfam00001 | $\hat{C}_1$ | 255 | 340 | 253 | 341 | | | | | | | | 255 | 405 | 253 | 403 | 242 | 206 | 240 | 204 |
|         | $\hat{C}_2$ | 55 | 340 | 57 | 341 | | | | | | | | 55 | 405 | 57 | 403 | 177 | 206 | 174 | 204 |

local FDR with $\hat{C}_1$ identifies more hotspots for Two-stage procedure (201) than One-stage (191) and Storey’s procedure (200). The rest of the domains can be analyzed in the similar manner.

Results from Table 3 revealed that using $\hat{C}_1$ yields more rejections. This suggests that the data analysis for the real data shows the same pattern as the simulation results presented previously. Moreover, two domains can be highlighted in terms of the difference in the number of rejections, namely, cd00180 and pfam00001. The number of rejections using $\hat{C}_1$ is almost four times higher if the model used for $f_0$ is ZIGP and the procedure employed is either local FDR or Two-stage method. For both of these data sets, the number of rejections using $\hat{C}_1$ is almost twice compared to $\hat{C}_2$ given that the model for $f_0$ is ZIGP and using Storey’s FDR. Overall, the results for the real data analysis is consistent with the simulation studies.

These results recapitulate much of what is known about how protein domain families contribute to the initiation or progression of cancer and are useful to biologists studying the molecular mechanisms underlying cancer. For instance, hotspots identified on the calcium-binding domain of epidermal growth factors (cd00054) are mapped to the 1EMN structure in Figure 3. Here, all three residues known to bind to calcium are also identified as hotspots, confirming the known importance of this binding pocket in cancer. In addition, we identify 42 other residues on the domain as hotspots that do not bind with calcium but tend to occur around the binding pocket and are present in many patients, suggesting their importance in cancer.

Figure 3. Patterns of protein domain hotspots on the calcium-binding domain of epidermal growth factors (cd00054). This figure appears in color in the electronic version of this article.
5. Conclusion
In this article, our primary interest is to select significant mutation counts for a specific protein domain, while controlling a given level of Type I error via False Discovery Rate (FDR) procedures. For the rth position, the number of mutations \( a_i \) follow one of the two distributions \( f_0 \) or \( f_1 \), namely, the null or alternative distribution, respectively. With \( \pi_0 \) and \( 1 - \pi_0 \) representing the prior probabilities of the two groups, the probability density function of the mixture distribution can be represented as \( f(a_i) = \pi_0 f_0(a_i) + (1 - \pi_0) f_1(a_i) \). We assume that if the number of mutations \( a_i \leq C \), then \( a_i \) is guaranteed to be from the null model, for some positive integer \( C \). We propose a method for identify a cut-off \( C \) and show that this is superior to the cut-off developed by extending Efron’s proposal. In addition, after the selection of this cut-off, we consider a screening process so that the number of mutations exceeding a certain value \( D_N (C) \) should be considered as significant mutations.

The proposed two-stage procedure in the selection of \( C \) and \( D_N \) yielded a testing procedure which is superior in terms TPR in most cases. For the well-separated and moderately mixed case, if the null model is correctly specified, then using the Two-Stage procedure yields FDR closest to the nominal level \( \alpha \) and the highest TPR. It can be noted that if the true model is ZIP and ZIGP is used to model \( f_0 \), then the Two-Stage Procedure still yields the closest FDR to \( \alpha \) and leads to highest TPR. This means that even when the null model is misspecified, the Two-Stage procedure would still produce satisfactory results. Also, regardless of the shape of the alternative distribution, the Two-Stage Procedure yields better results than the other procedures. Furthermore, results from six chosen protein domain data sets revealed that using the proposed cut-off for \( C \) yielded more rejections. In general, the results for the real data analysis is consistent with the simulation studies. These results sum up how protein domain families contribute to the initiation or progression of cancer and are useful to biologists studying the molecular mechanisms underlying cancer.

6. Supplementary Materials
Web Appendix A, referenced in Section 3.2 and Web Appendix B which is referenced in 3.4, is available with this article at the Biometric website on Wiley Online Library. The R codes used in this article are available at the Biometrics website on Wiley Online Library.

Acknowledgements
This research was supported in part by the Intramural Research Program of the NIH, National Library of Medicine.

References
Anderson, M. W., Reynolds, S. H., You, M., and Maronpot, R. M. (1992). Role of proto-oncogene activation in carcinogenesis. Environmental Health Perspectives 98, 13–24.
Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B (Methodological) 57, 289–300.
Conseil, P. C. and Jain, G. C. (1970). On the generalization of Poisson distribution. The Annals of Mathematical Statistics 41, 1387.
Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. Statistical Science 18, 71–103.
Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. Journal of the American Statistical Association 96, 1151–1160.
Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. Journal of the American Statistical Association 99, 465.
Efron, B. (2005). Local false discovery rates. Division of Biostatistics, Stanford, CA: Stanford University. http://statweb.stanford.edu/ckirby/brad/papers/2005LocalFDR.pdf
Efron, B. (2007). Doing thousands of hypothesis tests at the same time. Metron-International Journal of Statistics 65, 3–21.
Famula, T. F. and Singh, K. P. (2006). Zero-inflated generalized Poisson regression model with an application to domestic violence data. Journal of Data Science 4, 117–130.
Imaoka, T., Okutani, T., Daino, K., Iizuka, D., Nishimura, M., and Shimada, Y. (2014). Overexpression of NOTCH-regulated ankyrin repeat protein is associated with breast cancer cell proliferation. Anticancer Research 34, 2165–2171.
Janes, A., Gottardi, C. J., and Yap, A. S. (2008). Cadherins and Cancer: How does cadherin dysfunction promote tumor progression and quest. Oncogene 27, 6920–6929.
Joe, H. and Zhu, R. (2005). Generalized poisson distribution: The property of mixture of Poisson and comparison with negative binomial distribution. Biometrical Journal 47, 219–229.
Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34, 1–14.
McLachlan, G. J. and Jones, P. N. (1988). Fitting mixture models to grouped and truncated data via the EM algorithm. Biometrics 44, 571–578.
Nehrt, N. L., Peterson, T. A., Park, D., and Kann, M. G. (2012). Domain landscapes of somatic mutations in cancer. BMC Genomics 13, 1.
Neuhaus, E. M., Zhang, W., Gels, L., Deng, Y., Nokhush, J., and Hatt, H. (2009). Activation of an olfactory receptor inhibits proliferation of prostate cancer cells. Journal of Biological Chemistry 284, 16218–16225.
Park, D., Park, J., Zhong, X., and Sadelain, M. (2011). Estimation of empirical null using a mixture of normals and its use in local false discovery rate. Computational Statistics and Data Analysis 55, 2421–2432.
Parmigiani, G., Lin, J., Boca, S., Sjolblom, T., Kinzler, K. W., Velculescu, V. E., et al. (2007). Statistical methods for the analysis of cancer genome sequencing data. Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 126. http://biostats.bepress.com/jhubiostat/paper126
Peterson, T. A., Adadey, A., Santana-Cruz, I., Sun, Y., Winder, A., and Kann, M. G. (2010). DMDM: Domain mapping of disease mutations. Bioinformatics 26, 2458–2459.
Peterson, T. A., Nehrt, N. L., Park, D., and Kann, M. G. (2012). Incorporating molecular and functional context into the analysis and prioritization of human variants associated with cancer. Journal of the American Medical Informatics Association 19, 275–283.
Peterson, T. A., Park, D., and Kann, M. G. (2013). A protein domain-centric approach for the comparative analysis of human and yeast phenotypically relevant mutations. *BMC Genomics* **14**, 1.

Phang, Y. N. and Loh, E. F. (2013). Zero inflated models for overdispersed count data. *International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering* **7**, 817–819.

Pollard, K. S., Birkner, M. D., Van Der Laan, M. J., and Dudoit, S. (2005). Test statistics null distributions in multiple testing: Simulation studies and applications to Genomics. *Journal de la société française de statistique* **146**, 77–115.

Sever, R. and Brugge, J. S. (2015). Signal transduction in cancer. *Cold Spring Harbor Perspectives in Medicine* **5**, a006098.

Sheetlin, S., Park, Y., and Spouge, J. L. (2011). Objective method for estimating asymptotic parameters, with an application to sequence alignment. *Physical Review E* **84**, 031914.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 479–498.

Tsatsanis, C. and Spandidos, D. A. (2000). The role of oncogenic kinases in human cancer (Review). *International journal of molecular medicine* **5**, 583–590.

Received October 2016. Revised August 2017. Accepted August 2017.