BAYESIAN LOCAL FALSE DISCOVERY RATE FOR SPARSE COUNT DATA WITH APPLICATION TO THE DISCOVERY OF HOTSPOTS IN PROTEIN DOMAINS

BY IRIS IVY M. GAURAN *, JUNYONG PARK †, ILIA RATTSEV ‡, THOMAS A PETERSON §, MARICEL G KANN ‡, DOHWAN PARK ¶

Biostatistics Group, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia *, Department of Statistics, Seoul National University, Korea †, Department of Biological Sciences, University of Maryland, Baltimore County, MD, USA ‡, University of California, San Francisco, Institute for Computational Health Science, CA, USA §, Department of Mathematics and Statistics, University of Maryland Baltimore County, MD, USA ¶

In cancer research at the molecular level, it is critical to understand which somatic mutations play an important role in the initiation or progression of cancer. Recently, studying cancer somatic variants at the protein domain level is an important area for uncovering functionally related somatic mutations. The main issue is to find the protein domain hot spots which have significantly high frequency of mutations. Multiple testing procedures are commonly used to identify hotspots, however when data is not large enough, existing methods produce unreliable results with failure in controlling a given Type I error rate. We propose multiple testing procedures based on Bayesian local false discovery rate for sparse count data and apply it in the identification of clusters of somatic mutations across entire gene families using protein domain models. In multiple testing for count data, it is not clear what kind of the null distribution should be admitted. In our proposed algorithms, we implement the zero assumption in the context of Bayesian methods to identify the null distribution for count data rather than using any theoretical null distribution. Furthermore, we also address different types of modeling of alternative distributions. The proposed fully Bayesian models are efficient when the number of count data is small \(50 \leq N < 200\) while the local false discovery rate procedures based on the empirical Bayes is desirable for a large number of data \(N > 800\). We provide numerical studies to show that the proposed fully Bayesian methods can control a given level of false discovery rate for small number of positions while existing approaches based on nonparametric empirical Bayes fail in controlling a false discovery rate. In addition, we present real data examples of protein domain data to select hotspots in protein domain data.

1. Introduction. DNA microarray technology propelled researchers into the “post-genomics” era, as it allowed the advancement from structural to

Keywords and phrases: Bayesian Local False Discovery Rate, Sparse Count Data, Zero Inflated Generalized Poisson, Protein Domains
functional analysis of genomes. It is now crucial to decipher which genetic variations are consequential and which are merely random noise. Motivated by the need to analyze genomic data, there is a plethora of scientific literature about Bayesian approach to multiple testing (see Newton et al. (2004); Müller et al. (2004); Scott and Berger (2006); Cipolli III et al. (2016); Gutiérrez et al. (2019), for examples). In particular, Ibrahim et al. (2002) asserts that the Bayesian paradigm is well-suited for examining microarray data because the desired inferences, such as the calculation of quantiles, standard deviations, credible sets, and predictions, only require the computation of the posterior distribution.

Recently, several computational tools for analyzing genetic changes in cancer based on Bayesian approach have been developed. SAMtool in (Li, 2011), SOAPsnp in Li et al. (2008) and OncoSNP in Yau (2013) utilized a Bayesian approach for the analysis of single nucleotide polymorphism (SNP) microarray data. In addition, SomaticSniper in Larson et al. (2011), Strelka in Saunders et al. (2012) and Seurat in Christoforides et al. (2013) also employed a Bayesian-based analysis of sequenced genome pairs for the identification of somatic mutations in cancer. Likewise, Shiraishi et al. (2013) developed a method called Empirical Bayesian mutation calling (EBCall) which discriminates somatic mutations from sequencing errors by estimating the model parameters from multiple non-paired normal samples Shiraishi et al. (2013).

However, detection and evaluation of genetic aberrations including somatic mutations remains to be an unsolved problem (Ding et al., 2010). In contrast to the above mentioned gene-based methods for detecting somatic variants, we are interested in identifying protein domain hotspots. According to Chothia (1992), nature often brings several domains together to form multi-domain and multi-functional proteins with numerous possibilities. Protein domains can serve as modules for building up large assemblies such as virus particles or muscle fibers. Moreover, they can provide specific catalytic or binding sites as found in enzymes or regulatory proteins. Among a fixed number of positions in a single domain, hotspots are positions which are found to be mutated frequently and considered to be significantly different from the majority Peterson et al. (2013, 2017).

To detect hotspots among different positions in a protein domain, statistical testing with multiplicity correction are commonly implemented to control some types of error rates. One commonly used procedure in multiple testing is based on the local false discovery rate proposed by Efron and Tibshirani (2002). The local false discovery rate was originally developed by Efron and Tibshirani (2002) based on Nonparametric Empirical
Bayes (NPEB) method for continuous gene expression data. In the local false discovery rate procedure for continuous data, it is a crucial issue to identify the null distribution empirically rather than using a theoretical distribution. Furthermore, Efron and Tibshirani (2002) introduced the idea of “zero assumption” where values around the central peak of the distribution are considered to be null observations. Under this assumption, the parameters were estimated using Gaussian quadrature at the mode or conditional maximum likelihood estimation.

Compared to continuous data, there are two key differences in the context of count data: (i) what null distribution to specify, and (ii) how to determine the corresponding parameters. Gauran et al. (2018) defined the zero assumption for count data from protein domain and proposed multiple testing procedures based on the local false discovery rate using zero inflated discrete distributions. However, Gauran et al. (2018) used the idea of NPEB as in Efron et al. (2001) which is efficient when a large number of data are available. Furthermore, the zero assumption used in Gauran et al. (2018) is restrictive in that an interval consisting of all data from the null distribution is not realistic since small numbers of mutation counts may be generated from the alternative distribution. Protein domain data usually have sparse count data at the right tail, so procedures based on the NPEB for sparse count data are not reliable in the sense that a given false discovery rate may not be controlled quite often.

This paper is organized as follows. In Section 2 we provide a rationale for the inflated FDR and an explanation of why the frequentist paradigm fails. Bayesian False Discovery Rate is defined in Section 3. The model specification, choice of prior distributions and the posterior distribution are discussed in Section 3. The performance of the new procedure is studied via simulations in Section 5. The results for real data sets are shown in this section as well. Lastly, concluding remarks are presented in Section 6.

2. Multiple testing in Protein Domain Data. In this section, we present a specific application which motivated the development of our proposed fully Bayesian methods. This includes a discussion of the data structure of a given protein domain consisting of multiple positions with mutation counts. Also, a brief introduction of some existing methods are included in this section.

2.1. Application to Protein domain data. This study utilized protein domain data available in the Domain Mapping of Disease Mutations database (DMDM) (http://bioinf.umbc.edu/DMDM/). In this database, after mapping all human coding mutations on the protein domain, the number of dis-
ease mutations and polymorphisms in each domain position are displayed vis-à-vis some relevant functional information. Compared to gene-centric visualization tools, this unique protein domain view highlights molecular relationships among mutations from various diseases.

From at least ten thousand protein domains available, we categorized the number of positions in a protein domain into four groups. If the number of positions is less than 200, then we characterize these protein domain data as having small \( N \) while protein domains with \( 200 \leq N < 500 \) is classified as moderately small. In addition, if \( 500 \leq N < 800 \) then \( N \) is categorized as moderately large while protein domains with large number of positions refer to domains with at least 800 positions. The rationale for characterizing the different values of \( N \) into groups is to allow for a pragmatic consideration on the framework that we propose in the next sections. Moreover, this enables us to describe numerically what we meant by small number of positions. This is also consistent with the benchmark procedure proposed in Gauran et al. (2018) where the number of positions considered to be large is \( N = 1,000 \). However, out of 11,244 protein domains considered, only 5.27\% of these have number of positions at least 800 as to compared to 40.15\% domains with small \( N \) and 42.87\% with moderately small \( N \). This prompted the need to develop procedures that can handle sparse count data when the number of positions is considered to be small or moderately small.

To identify hotspots which could be insightful in studying the molecular mechanisms involved with cancer, we selected six domains with small number of positions. The distribution of each protein domain and its total number of positions are displayed in Figure 1. We start by probing into the Fibronectins (FN), which encompasses a variety of major roles in cell adhesion, migration, differentiation and proliferation (Hynes, 2012). Specifically, we are interested in the FN Type III domain (cd00063), an extradomain of which is only expressed during embryogenesis, wound healing and tumorigenesis (Bencharit et al., 2007). Motivated by this, Blumson (2017) mentioned that there is the potential to develop therapeutic agents based on the FN Type III connecting segment structure.

Likewise, calcium binding to EGF-like domains (cd00054, smart00179) are posited to be crucial for numerous protein-protein interactions involving EGF-like domains in coagulation factors, plasma proteins, and membrane protein (Selander-Sunnerhagen et al., 1992). Another domain we delved into is the helix-loop-helix (HLH) family of transcriptional regulatory proteins (cd00083) which are involved in several developmental processes including neurogenesis, myogenesis, hematopoiesis, sex determination and gut development (Massari and Murre, 2000).
In addition, we considered the RNA recognition motif (RRM) also known as the RNA binding domain (cd00590). RNA binding proteins are key players in RNA processing, post-transcriptional regulation of gene expression (Biamonti and Riva, 1994) and are involved in several steps at which mRNA biogenesis, stability, translation and decay is exerted (Cassola et al., 2010). Lastly, we looked into the hotspots of Thrombospondin type 1 repeats (smart00209). According to Atanasova et al. (2019), this protein play a role in wound repair and healing, platelet aggregation, angiogenesis, and tumorigenesis.

The raw data illustrated in Figure 1 can be viewed as a vector of mutation counts among $N$ positions in a given protein domain where the number of positions is small. Formally, let $x_N = (x_1, \ldots, x_N)$, where $x_i$ is the number of mutations in the $i$th position, $i = 1, 2, \ldots, N$. $K$ is the observed maximum number of mutation counts indicating that $0 \leq x_i \leq K$. For each mutation count $x_i$, there are two possibilities:

$H_{0i} : x_i \sim f_0$ vs. $H_{1i} : x_i \sim f_1$

where $f_0$ and $f_1$ are the null and alternative distributions representing the background mutations and disease–based mutations. These hypotheses can
be represented as $f = \pi_0 f_0 + (1 - \pi_0) f_1$ where $\pi_0 = P(H_0)$ is the prior probability that the null hypothesis is true. We are interested in detecting hotspots in Peterson et al. (2017) representing positions with unusually high mutation counts compared to background mutation counts.

For count data, Gauran et al. (2018) used a class of zero inflated discrete distributions known as the Zero Inflated Generalized Poisson (ZIGP) distribution for $f_0$ and estimated the corresponding parameters using the idea of zero assumption used in Efron et al. (2001). Efron et al. (2001) and Gauran et al. (2018) are based on Nonparametric Empirical Bayes (NPEB) approach which is efficient when the number of positions in the protein domain data $(N)$ is fairly large.

2.2. Nonparametric Empirical Bayes (NPEB) False Discovery Rate. Extending Efron et al. (2001)’s zero assumption to discrete data, we define the key assumption in this study as

\begin{equation}
 f(x_i) = \pi_0 f_0(x_i), \text{if } x_i \leq C \text{ for some unknown } C
\end{equation}

wherein small number of mutation counts are generated from $f_0$.

The NPEB approach consists of maximum likelihood estimation for the parameters in $f_0$ and estimation of $f$ which depend only on the data. However, when $N$ is not large enough, the NPEB method may suffer from inaccurate estimators of $f_0$ and $f$ leading to unstable performance of the local false discovery rate. Gauran et al. (2018) proposed NPEB based multiple testing procedures controlling FDR for protein domain data and they focus on the cases of large numbers of positions in protein domain data such as $N = 1,000$ in their numerical studies. On the other hand, when the number of positions $N$ is small, the NPEB based procedures may fail in controlling a given level of FDR.

An alternative method to remedy this issue is to introduce the fully Bayesian models implementing the zero assumption to estimate the parameters in $f_0$. Angers and Biswas (2003) highlighted that most of the studies involving zero inflated count data were carried out from the frequentist viewpoint. They pointed out the limitations of these approaches by citing scenarios wherein classical statistical inference procedures such as maximum likelihood estimation and large sample approximation of a confidence interval may not be suitable for making inferences about the parameters. Meanwhile, an attractive feature of the Bayesian paradigm in multiple testing is the smoothing property which is crucial in the sparse count data scenario.

The motivation of this study is that fully Bayesian methods are expected to be more reliable in controlling a given level of FDR. The following sections
present the models we propose and the corresponding MCMC algorithms incorporating the zero assumption. We will demonstrate that the proposed Bayesian approaches have more reliable results compared to NPEB based approaches via numerical studies when the number of positions in a protein domain is small.

3. Proposed Bayesian models for the calculation of FDR. In this section, we propose three different frameworks to construct the fully Bayesian models and explore the choice of prior distributions. In order to investigate mutation counts in the protein domain data, there are two key points for the calculation of local false discovery rate: (i) we consider different approaches in modeling $f$ or $f_1$, and (ii) we incorporate the zero assumption in our proposed Bayesian models. We need to identify three components, $(\pi_0, f_0, f)$, to calculate the local false discovery rate.

3.1. Model Specification. At the onset, the available information to us are the histograms of the protein domain data presented in Figure 1. The choice of the null distribution and its corresponding parameters is a result of mimicking the characteristics displayed by these histograms. As a choice of $f_0$, we use the Zero Inflated Generalized Poisson ZIGP($\eta, \lambda, \theta$) distribution because it accounts for both zero-inflation and overdispersion which are features of the mutation counts. The probability distribution is given by

$$f_0(x|\phi_0) = \eta I_0(x) + (1-\eta)g(x)$$

where $\phi_0 = (\eta, \lambda, \theta)$, $0 \leq \eta \leq 1$, $\lambda > 0$ and $0 \leq \theta < 1$ and

$$g(x) = \frac{\lambda^{(\lambda + 1)x-1}}{x!}e^{-\lambda}$$

wherein $g(x)$ is the probability function of Generalized Poisson GP($\lambda, \theta$) distribution. The rationale for choosing ZIGP instead of ZINB as a model for $f_0$ was provided by Joe and Zhu (2005). They highlighted that ZIGP provides a better fit than ZINB when there is a large fraction of zeros and the data is heavily right-skewed by comparing the probabilistic properties of the zero-inflated variations of NB and GP, 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.

Meanwhile, there are several frameworks to model $f$. Similar to the NPEB approach, $f$ is estimated through the nonparametric approach. On the other hand, one can estimate $f$ by modeling the alternative distribution $f_1$ and
incorporating it with $f_0$ and $\pi_0$. In this study, we explored three models of $f$ as follows:

1. **Nonparametric Model**: $f(x_i \mid \phi) \sim D(\beta)$ where $f(\cdot \mid \beta, \phi_0, \pi_0, C)$ is marginally distributed as Dirichlet with concentration parameter $\beta$.

2. **Semiparametric Model**: $f(x_i \mid \phi) = \pi_0 f_0(x_i \mid \phi_0) + (1 - \pi_0) f_1(x_i \mid \beta)$, where $f_1(x_i \mid \beta) \sim D(\beta)$ is the Dirichlet distribution with concentration parameter $\beta$.

3. **Parametric Model**: $f(x_i \mid \phi) = \pi_0 f_0(x_i \mid \phi_0) + (1 - \pi_0) f_1(x_i \mid \phi_1)$, where $f_1(x_i \mid \phi_1)$ is a known parametric discrete distribution which accounts for overdispersion such as the Generalized Poisson distribution.

Another key consideration in this study is the implementation of the zero assumption which plays an important role in the context of the empirical Bayes false discovery rate as in Gauran et al. (2018) and Efron et al. (2001). This is carried out in the likelihood presented in Section 3.2 and further highlighted in the proposed methods in Section 4. Based on the zero assumption, one can obtain reliable estimates of the empirical null distribution, i.e., all parameters in $f_0$.

Formally, we define the overall parameter vector $\phi = (\phi_0, \phi_1, \pi_0, C)$ where $\phi_0$ is the vector of the null distribution parameters, $\phi_1$ is the vector of alternative distribution parameters, $\pi_0$ is the proportion of observations from the null distribution and $C$ is the cut-off for the implementation of the zero assumption. If $\phi_0$ has dimension $\ell_0$ and $\phi_1$ has dimension $\ell_1$, then we are interested in obtaining draws for the $(\ell_0 + \ell_1 + 2)$-dimensional vector $\phi$.

### 3.2. Likelihood

Since $x_i$ is generated from the mixture of $f_0$ and $f_1$, we introduce the latent variables for $x_i$, say $z_N = (z_1, z_2, \ldots, z_N)$ where $z_i$ indicates the membership of $x_i$, either 1 or 0 for $f_0$ and $f_1$, respectively. In particular, the zero assumption states that, for a given $C$, $z_i = 1$ with probability 1 since $x_i \leq C$ for a given $C$ ensures that $x_i$ is generated from $f_0$. The likelihood function for $x_N$ is

$$L(\phi \mid x_N) = \prod_{i=1}^{N} f(x_i \mid \phi) = \prod_{i: x_i \leq C} \pi_0 f_0(x_i \mid \phi_0) \prod_{i: x_i > C} f(x_i \mid \phi)$$

$$= \prod_{j \leq C} (\pi_0 f_0(j \mid \phi_0))^n_j \prod_{j > C} f(j \mid \phi)^n_j$$

-1}
where \( n_{0j} = \sum_{i=1}^{N} z_i I(x_i = j) \), \( n_{1j} = \sum_{i=1}^{N} (1 - z_i) I(x_i = j) \) and \( n_j = n_{0j} + n_{1j} \).

The full likelihood function for \((x_N, z_N)\) is

\[
L(\phi \mid x_N, z_N) = \prod_{i=1}^{N} \left[ \pi_0 f_0(x_i \mid \phi_0) \right]^{z_i} \left[ (1 - \pi_0) f_1(x_i \mid \phi_1) \right]^{1-z_i}
\]

\[
= \pi_0^* \prod_{j \leq C} f_0(j \mid \phi_0)^{n_{0j}} \prod_{j > C} f_0(j \mid \phi_0)^{n_{0j}} f_1(j \mid \phi_1)^{n_{1j}}
\]

where \( \pi_0^* = \pi_0 \sum_{i=1}^{N} z_i \) \( \pi_0 \sum_{i=1}^{N} (1 - z_i) \). The inclusion of the latent variables \( z_N \) allowed us to rewrite the likelihood function in (3.3) into (3.4)

\[
n_j = \begin{cases} 
  n_{0j} & \text{if } j \leq C \\
  n_{0j} + n_{1j} & \text{if } j > C.
\end{cases}
\]

### 3.3. Choice of Prior Distributions

A crucial issue in Bayesian inference is the sensitivity of inferences to prior specification. To address this, we choose priors with minimal influence on the inference.

Given that \( C \) is unknown for a certain protein domain data, the cutoff value \( C \) is assumed to have a non-degenerating prior distribution rather than a fixed unknown constant assumed in Gauran et al. (2018). This enables the proposed fully Bayesian models to offer more flexibility as compared to NPEB procedures in Gauran et al. (2018). We use Poisson distribution as our prior distribution for \( C \) and consider the following hierarchical model:

\[
C \mid \tau \sim \text{Poisson}(\tau)
\]

\[
\tau \mid \kappa_\tau, \vartheta_\tau \sim \text{Gamma}(\kappa_\tau, \vartheta_\tau)
\]

where \( \kappa_\tau \) and \( \vartheta_\tau \) are predetermined hyperparameters. In fact, this hierarchical model for \( C \) is the negative binomial setup. We will demonstrate how the zero assumption based on \( C \) affect Bayesian sampling procedures in Section 5.1. Under the ZIGP null distribution specification, we also choose non-informative prior distributions for \( \eta, \theta \) and \( \pi_0 \) which are

\[
\eta \sim \mathcal{U}(0, 1), \quad \theta \sim \mathcal{U}(0, 1), \quad \pi_0 \sim \mathcal{U}(0, 1)
\]

and the prior for \( \lambda \) is

\[
\lambda \sim g(\lambda) \propto \lambda^{-0.5}
\]
which is the Jeffreys’ prior for Poisson distribution. Moreover, the upper limit for the observed maximum mutation count \( K \) is specified to be \( P \). Given that \( K \) is an observed maximum, it is possible that the true maximum \( P \) could be higher. To incorporate this in our proposed methods, we have specified \( P = (1 + \kappa)K \), \( 0 < \kappa < 1 \) as the length of the Dirichlet prior distribution to assign non-zero probabilities for values \( K + 1 \) to \( P \). The imposed restriction on the value of \( \kappa \) is necessary because choosing a value of \( \kappa >> 1 \) severely affects the prior distribution. This will lead to very sparse probabilities and it defeats the purpose of implementing the Bayesian paradigm because the smoothing property will not be utilized. On the other hand, choosing \( \kappa \to 0 \), meant that we are assigning 0 probability to values which can potentially be the true maximum. Hence, we settled for \( \kappa = 0.5 \) in our proposed methods.

Summarizing the three different models described in the previous section and incorporating the zero assumption given \( C \), we have the following cases:

1. **Nonparametric Case of \( f \):** The probability vector of the marginal variable, denoted by \( f \equiv (f(0), f(1), \ldots, f(K), \ldots, f(P)) \), follows

   \[
   f \sim D(\beta = \beta 1_P)
   \]

   for a given \( P > K = \max 1 \leq x_i \leq N \) where \( D(\beta = \beta 1_P) \) is the Dirichlet distribution for with concentration parameter \((\beta, \beta, \ldots, \beta), \beta > 0\).

2. **Semiparametric Case:** The probability vector of \( f_1 \), denoted by \( f_1 \equiv (f_1(0), f_1(1), \ldots, f_1(P)) \)

   \[
   f_1 \sim D(\beta)
   \]

   given \( C \) where \( \beta = (\epsilon, \epsilon, \ldots, \epsilon, \gamma, \gamma, \ldots, \gamma) \) for some given \( \epsilon > 0 \), \( \gamma > 0 \), \( \epsilon << \gamma \).

3. **Parametric Case of \( f_1 \):** The shifted discrete distribution can be expressed as \( x_i|H_{1i}, C \sim W + C + 1 \) where \( W \sim GP(\delta, \nu) \) with \( g(\delta) \propto \delta^{-0.5} \) and \( \nu \sim U(0,1) \). \( GP(\delta, \nu) \) is defined in (3.2).

**Remark:** For the nonparametric case, we consider five values of \( \beta \): (i) \( \beta = 1/P \) which is the reference prior in Berger et al. (2015) (ii) \( \beta = 0.5 \), which is the Jeffreys’ prior, (iii) \( \beta = 1 \), which is the continuous uniform scenario, (iv) \( \beta = 1.5 \), which exhibits near continuous uniform scenario, and (v) \( \beta = 3 \), for the informative prior scenario.
3.4. Posterior distributions. From the previous sections, we presented the likelihood and the summary of the three different models capturing all necessary prior distributions as well as incorporating the zero assumption given $C$. We derive the conditional posterior distributions of $\phi_0 = (\eta, \lambda, \theta)$, $\phi_1$, $\pi_0$, $C$, $\tau$ and $z_N$ because these are necessary for our proposed MCMC algorithm. The details involved in the derivations are presented in the Supplementary Material.

4. Proposed Methods. In this section, we present MCMC algorithms for three different models described in Section 3.3. Using the specification of the conditional posterior of all the parameters related to $f_0$ and $\pi_0$, we implement the Adaptive Metropolis-Hastings within Gibbs sampling algorithms based from the different models for $f$ or $f_1$. The main emphasis in our proposed sampling algorithms is that we take into account the zero assumption in 2.2. First, we define the following notations:

\[ f = (f(0), f(1), \ldots, f(C), f(C + 1), \ldots, f(P)) \]
\[ f_1 = (f_1(0), f_1(1), \ldots, f_1(C), f_1(C + 1), \ldots, f_1(P)) \]

where $P$ is a given number larger than $K = \max_{1 \leq i \leq K} x_i$. The zero assumption implies that, for a given $C$,

\[
f|C = (f(0), f(1), \ldots, f(C), f(C + 1), \ldots, f(P))
= (\pi_0 f_0(0), \pi_0 f_0(1), \ldots, \pi_0 f_0(C), f(C + 1), \ldots, f(P))
\equiv (\psi_0, \ldots, \psi_C, \psi_{C+1}, \ldots, \psi_P)
\equiv (\Psi_0, \ldots, \Psi_C, \Psi_{C+1}, \ldots, \Psi_P)
\equiv (\Psi_0, \Psi_1) \equiv \Psi
\]

(4.1)

and

\[
f_1|C = (f_1(0), f_1(1), \ldots, f_1(C), f_1(C + 1), \ldots, f_1(P))
\equiv (\Upsilon_0, \ldots, \Upsilon_C, \Upsilon_{C+1}, \ldots, \Upsilon_P)
\equiv (\Upsilon_0, \ldots, \Upsilon_1)
\equiv \Upsilon
\]

(4.2)

The nonparametric model generates $f$ directly while the semiparametric and parametric models generate $f$ by combining $\pi_0 f_0(j|\phi_0)$ and $f_1$.

4.1. Nonparametric Model for Bayesian False Discovery Rate. For a given $C$, $\Psi(0)$ is the same as $\pi_0 f_0(j|\phi_0)$, so we use $\pi_0 f_0(j|\phi_0)$ for $\Psi(0)$ based on
samples of \((\pi_0, \phi_0)\). Since \(\Psi\) is assumed to follow the Dirichlet distribution, the posterior distribution of \(\Psi(1)\) given \(\Psi(0)\) is

\[
\frac{1}{1 - \alpha_0} \Psi_{(1)} \left| (\Psi_{(0)}, x_N, z_N, \beta) \sim D(\beta_{C}^*) \right.
\]

where \(\alpha_0 = \sum_{j=0}^{C} \pi_0 f_0(j|\phi_0)\) and \(\beta_{C}^* = (\beta_{C+1}^*, \ldots, \beta_{P}^*)\) with

\[
\beta_{j}^* = \beta_j + n_{0j} + n_{1j}
\]

for \(j \geq C + 1\). Based on the previous computations on the posterior distribution of \(\pi_0\) and \(\phi_0\) in \(f_0\), we propose the following algorithm for sampling in the nonparametric model.

Algorithm for Nonparametric Model:

1. **Initialization:**
   (a) **Time instants:** Set \(t = 0\) and choose the values \(T_{\text{start}} < T_{\text{stop}} < T_{\text{total}}\) where \(T_{\text{start}}\) is the iteration to begin adaptation, \(T_{\text{stop}}\) is the iteration to end adaptation and \(T_{\text{total}}\) is the total number of iterations of the chain.
   (b) **Proposal:** Choose the initial settings for \(\phi_{0(0)}\), \(\pi_{0(0)}\), \(\Psi^{(0)}\), \(\tau^{(0)}\), \(z_{N(0)}^{(0)}\) and \(\Sigma^{(0)}\).

2. **Gibbs step for \(C\):** Update \(C^{(t)}\) by sampling from (A.2) in Supplementary Material.

3. **Gibbs step for \(\tau\):** Update \(\tau^{(t)}\) by sampling from (A.4) in Supplementary Material.

4. **Gibbs step for \(z_N\):** Update \(z_i^{(t)}\) by sampling from (A.5) in Supplementary Material for \(i = 1, 2, \ldots, N\).

5. **Gibbs step for \(\pi_0\):** Update \(\pi_0^{(t)}\) by sampling from (A.8) in Supplementary Material.

6. **Metropolis-Hastings Steps:**
   Let \(h(\phi_0) = h((\lambda, \eta, \theta)) = (\log \lambda, \log \frac{\eta}{1-\eta}, \log \frac{\theta}{1-\theta})\).
   (a) Randomly generate \(w_t\) from \(\ell_0\)-variate Standard Normal and let

   \[
   \phi_{0(t)}^{(t)} = \left(\Sigma^{(t)}\right)^{1/2} w_t + \phi_{0(t)}^{(t)}.
   \]

   (b) Accept \(\phi_{0(t+1)} = h^{-1}(\phi_{0(t)}^{(t)})\) with probability defined in (A.9) in Supplementary Material. Otherwise, set \(\phi_{0(t+1)} = \phi_{0(t)}^{(t)}\).
7. **Updating:** Suppose $T_{\text{thin}}$ is the frequency with which updating occurs and $T_{\text{prop}}$ is the proportion of previous states to include when updating. If $T_{\text{start}} < t < T_{\text{stop}}$ and $t \equiv 0 \pmod{\text{mod } T_{\text{thin}}}$, identify the set of indices $I$ to be used for updating.

$$I = \{\lfloor t \cdot T_{\text{prop}} \rfloor, \lfloor t \cdot T_{\text{prop}} \rfloor + 1, \ldots, t\}$$

Update the parameters of the proposal covariance matrix as follows:

$$\Sigma^{(t + 1)} = \frac{1}{|I|} \sum_{i \in I} \left( \phi_0^{(i)} - \bar{\phi}_0 \right) \left( \phi_0^{(i)} - \bar{\phi}_0 \right)^T$$

where $\bar{\phi}_0 = \frac{1}{|I|} \sum_{i \in I} \phi_0^{(i)}$. If $t < T_{\text{total}}$, repeat from Step 6.

8. **Gibbs step for $\Psi$:** Compute $\Psi_{(0)}^{(t)} = (\pi_0^{(t)} f_0^{(t)}(0 | \phi_0^{(t)}), \ldots, \pi_0^{(t)} f_0^{(t)}(C | \phi_0^{(t)}))$ and generate $\Psi_{(1)}^{(t)}$ from (4.3).

9. Repeat Steps (2) to (8) for $t = 1, 2, \ldots, T$.

Following this method, we use the marginal posterior distribution to calculate the local false discovery rate

$$fdr(j | x_N) = E_{x_N, \phi | x_N} [ fdr(j | \phi, x_N, z_N) ]$$

$$\approx \frac{1}{T} \sum_{t=1}^{T} fdr(j | \phi^{(t)}, x_N, z_N^{(t)})$$

for $fdr(j | \phi, x_N, z_N) = \frac{\pi_0 f_0(j | \phi, x_N, z_N)}{\sum_{j} \pi_0 f_0(j | \phi, x_N, z_N)}$ and mutation counts $j = 0, 1, \ldots, K$. We reject the null hypothesis if

$$fdr(j | x_N) \leq \alpha = 0.05.$$  

4.2. **Semiparametric Model for Bayesian False Discovery Rate.** In contrast to the scenario described in Section 4.1, we consider a nonparametric distribution for $f_1$ instead of $f$. The prior distribution of $\Upsilon$ is given by $D(\beta)$ where $\beta = (\epsilon, \epsilon, \ldots, \epsilon, \gamma, \gamma, \ldots, \gamma) = (\epsilon \cdot 1_{C+1}, \gamma \cdot 1_{P-C-1})$. To reflect the zero assumption, we assign $\epsilon \approx 0$ so that $\Upsilon_0, \ldots, \Upsilon_C$ are generated to be almost 0 for a given $C$. Note that $z_i = 0$ corresponding to $x_i \leq C$ for a given $C$ from the zero assumption. This leads to $n_{0j} = 0$ for $j \leq C$. The details of the proposed implementation and sampling algorithm of the Semiparametric Model is presented in the Supplementary Materials.
4.3. **Parametric Model for Bayesian False Discovery Rate.** Instead of a nonparametric distribution for \( f_1 \), we consider a parametric distribution in this framework. Specifically, we use the shifted Generalized Poisson distribution to reflect the zero assumption such that for a given \( C \), \( x \sim f_1(x) \) where \( X = W + C \) and \( W \) is the given parametric distribution for \( W \geq 0 \). More information on the proposed implementation and sampling algorithm of the Parametric Model is available in the Supplementary Materials.

5. **Numerical Studies.** To arrive at sound conclusions and gather meaningful insights regarding the identification of hotspots, we performed exhaustive simulation studies. We incorporate the multidimensional settings such as \( N, \pi_0 \), choice of \( f_0 \) parameters, and choice of prior distribution hyperparameters, among others, which could potentially impact the False Discovery Rate (FDR) and empirical power.

5.1. **Simulation Studies.** To simulate the data, \( \lceil N \pi_0 \rceil \) of the observations were generated using ZIGP(\( \eta, \lambda, \theta \)) as the true null distribution \( f_0 \). The remaining observations were generated from a discrete non-null distribution \( f_1 \). The support of \( f_1 \) does not contain values in \([0, C]\) following the key assumption on \( f_0 \). Hence, \( f_1 \) can be expressed as \( f_1 = C+1+U \) where \( U \) follows another count model. For the model specification of \( U \), Geometric(\( p = 0.08 \)) and Binomial(\( n = 250, p = 0.20 \)) distribution are utilized. They were chosen because the resulting distribution of the observations exhibit the pattern of the mutation count observed in the real data set (Gauran et al., 2018).

Each randomly generated vector of observations mimicking the properties of the protein domain data is inputted in the adaptive Gibbs sampler. After removing a burn-in of 1000 draws, a total of 5000 draws without thinning were used to compute the false discovery proportion. To compute the FDR, 1000 replications were carried out and the average of the false discovery proportion is reported. The empirical False Discovery Rate and True Positive Rate are computed as

\[
\begin{align*}
\hat{\text{FDR}} &= \frac{1}{1000} \sum_{d=1}^{1000} \frac{V_d}{R_d} I(R_d > 0), \\
\hat{\text{TPR}} &= \frac{1}{1000} \sum_{d=1}^{1000} \left( \frac{S_d}{S_d + T_d} \right)
\end{align*}
\]

where \( V_d \) and \( R_d \) are the number of false discoveries and the total number of discoveries and \( S_d \) and \( T_d \) are the number of true discoveries and the number of false non-discoveries in the \( d \)th generated data.

We provide a numerical comparison among the three proposed fully Bayesian methods versus the two different Empirical Bayes (EB) methods developed by Gauran et al. (2018) as well as the modified Storey’s procedure when the
numbers of positions in a protein domain $N$ are varied from small to large. Following the categories described in Section 2.1, we specify small $N$ to be 50 and 100, $N = 200$ is considered moderately small, $N = 500$ is moderately large while $N = 1000$ is described as large. We also compare the results in terms of the value of $C$ wherein $C = 10$ and $C = 5$ reflect a moderately mixed and heavily mixed null and non-null distributions, respectively. The results which yield the superior $\hat{TPR}$ while controlling $\hat{FDR}$ are highlighted using bold face quantities.

Table 1 results show that all fully Bayesian methods control the $\hat{FDR}$ for a given $\alpha = 0.05$, regardless of the choice of $N$ or $C$. In contrast, we can see that all nonparametric empirical Bayes (NPEB) methods failed in controlling FDR as shown by the $\hat{FDR}$ values ranging between 0.1 to 0.2 which exceed the nominal level when $N$ is small or moderately small. Hence, we recommend Nonparametric Model for Bayesian False Discovery Rate when the number of positions in a protein domain is small, i.e. $N$ is less than 200. The simulation studies show that there is an advantage for the full Bayesian methods when the sample size is small in terms of controlling the FDR. When $N$ is moderately small and the null and non-null distributions are moderately mixed, One-stage NPEB method marginally control the FDR and obtain more TPR than the full Bayesian methods. However, we still propose the use of fully Bayesian methods when $N$ is moderately small and we have a heavily mixed scenario.

When $N = 500$ and $C = 10$, the $\hat{FDR}$ for the One-stage NPEB method is controlled and it yielded the highest $\hat{TPR}$. Under the moderately mixed scenario, as $N$ increases, the EB methods show improvement in controlling FDR and obtain more TPR than the full Bayesian methods. On the other hand, under the heavily mixed scenario, the One-stage NPEB method shows marginal control in $\hat{FDR}$ only when is large, i.e. $N = 1000$. Both Two-stage and modified Storey’s procedure failed to control FDR for any $N$ because of the strong overdispersion and the heavy mixing between $f_0$ and $f_1$. Even though the FDR among fully Bayesian methods are controlled, we see a decrease in the $\hat{TPR}$. As an overall guideline, we propose the use of fully Bayesian methods when the $N$ is small, i.e. when $N$ is less than 200.

In Section 3.3, we specify objective or noninformative priors to minimally influence the inference and results. The prior sensitivity analysis is presented in Table 2. The numerical comparison among different sample sizes show results that are similar to each other for most of cases. We propose using the uniform prior for convenience and simplicity of interpretation/justification. Given that studies involving mutation counts in protein data are fairly recent, it is hard to reflect solid scientific and biological knowledge.
Table 1

Numerical Comparison when \( f_0 \) is ZIGP(\( \eta = 0.4, \lambda = 2, \theta = 0.3 \)), \( f_1 \) is shifted Binomial, and \( \pi_0 = 0.80 \). \( R \) is the average number of rejections while \( \hat{FDR} \) and \( \hat{TPR} \) correspond to the calculated False Discovery Rate and True Positive Rate, respectively. The number in ( ) corresponds to the standard error.

<table>
<thead>
<tr>
<th>( N )</th>
<th>Method</th>
<th>( C = 10 )</th>
<th>( \frac{R}{FDR} )</th>
<th>( \frac{R}{TPR} )</th>
<th>( C = 5 )</th>
<th>( \frac{R}{FDR} )</th>
<th>( \frac{R}{TPR} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>Nonparametric</td>
<td>9.33</td>
<td>0.0141</td>
<td>0.9404</td>
<td>7.40</td>
<td>0.0178</td>
<td>0.7374</td>
</tr>
<tr>
<td></td>
<td>Parametric</td>
<td>8.36</td>
<td>0.0035</td>
<td>0.8502</td>
<td>7.07</td>
<td>0.0102</td>
<td>0.7119</td>
</tr>
<tr>
<td></td>
<td>Semiparametric</td>
<td>8.84</td>
<td>0.0067</td>
<td>0.8929</td>
<td>6.06</td>
<td>0.0068</td>
<td>0.6092</td>
</tr>
<tr>
<td></td>
<td>Two-Stage</td>
<td>10.60</td>
<td>0.1582</td>
<td>0.9009</td>
<td>10.31</td>
<td>0.1618</td>
<td>0.8663</td>
</tr>
<tr>
<td></td>
<td>One-Stage</td>
<td>9.73</td>
<td>0.0959</td>
<td>0.9096</td>
<td>9.33</td>
<td>0.1011</td>
<td>0.8541</td>
</tr>
<tr>
<td></td>
<td>Storey</td>
<td>10.83</td>
<td>0.1513</td>
<td>0.9393</td>
<td>10.62</td>
<td>0.1568</td>
<td>0.9112</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(3.32)</td>
<td>(0.1375)</td>
<td>(0.1379)</td>
<td>(3.63)</td>
<td>(0.1386)</td>
<td>(0.1745)</td>
</tr>
<tr>
<td>100</td>
<td>Nonparametric</td>
<td>18.18</td>
<td>0.0069</td>
<td>0.9221</td>
<td>11.79</td>
<td>0.0666</td>
<td>0.5986</td>
</tr>
<tr>
<td></td>
<td>Parametric</td>
<td>17.14</td>
<td>0.0037</td>
<td>0.8709</td>
<td>12.99</td>
<td>0.0108</td>
<td>0.6493</td>
</tr>
<tr>
<td></td>
<td>Semiparametric</td>
<td>17.07</td>
<td>0.0039</td>
<td>0.8653</td>
<td>7.53</td>
<td>0.0030</td>
<td>0.3836</td>
</tr>
<tr>
<td></td>
<td>Two-Stage</td>
<td>20.77</td>
<td>0.1361</td>
<td>0.9046</td>
<td>20.20</td>
<td>0.1379</td>
<td>0.8749</td>
</tr>
<tr>
<td></td>
<td>One-Stage</td>
<td>18.99</td>
<td>0.0693</td>
<td>0.9039</td>
<td>18.16</td>
<td>0.1072</td>
<td>0.8526</td>
</tr>
<tr>
<td></td>
<td>Storey</td>
<td>20.93</td>
<td>0.1180</td>
<td>0.9388</td>
<td>20.33</td>
<td>0.1209</td>
<td>0.9069</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(5.59)</td>
<td>(0.0925)</td>
<td>(0.1566)</td>
<td>(6.06)</td>
<td>(0.0925)</td>
<td>(0.1854)</td>
</tr>
<tr>
<td>200</td>
<td>Nonparametric</td>
<td>35.24</td>
<td>0.0058</td>
<td>0.8958</td>
<td>16.02</td>
<td>0.0025</td>
<td>0.4060</td>
</tr>
<tr>
<td></td>
<td>Parametric</td>
<td>33.54</td>
<td>0.0022</td>
<td>0.8564</td>
<td>21.07</td>
<td>0.0030</td>
<td>0.5558</td>
</tr>
<tr>
<td></td>
<td>Semiparametric</td>
<td>31.93</td>
<td>0.0029</td>
<td>0.8131</td>
<td>7.85</td>
<td>0.0032</td>
<td>0.1734</td>
</tr>
<tr>
<td></td>
<td>Two-Stage</td>
<td>42.58</td>
<td>0.1306</td>
<td>0.9349</td>
<td>41.06</td>
<td>0.1323</td>
<td>0.8979</td>
</tr>
<tr>
<td></td>
<td>One-Stage</td>
<td>38.88</td>
<td>0.0588</td>
<td>0.9320</td>
<td>36.76</td>
<td>0.0692</td>
<td>0.8699</td>
</tr>
<tr>
<td></td>
<td>Storey</td>
<td>41.90</td>
<td>0.1016</td>
<td>0.9579</td>
<td>40.09</td>
<td>0.1041</td>
<td>0.9124</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(9.24)</td>
<td>(0.0652)</td>
<td>(0.1444)</td>
<td>(10.02)</td>
<td>(0.0667)</td>
<td>(0.1723)</td>
</tr>
<tr>
<td>500</td>
<td>Nonparametric</td>
<td>86.56</td>
<td>0.0039</td>
<td>0.8815</td>
<td>33.38</td>
<td>0.0080</td>
<td>0.3385</td>
</tr>
<tr>
<td></td>
<td>Parametric</td>
<td>78.32</td>
<td>0.0012</td>
<td>0.8004</td>
<td>39.06</td>
<td>0.0006</td>
<td>0.3980</td>
</tr>
<tr>
<td></td>
<td>Semiparametric</td>
<td>69.49</td>
<td>0.0012</td>
<td>0.7083</td>
<td>4.65</td>
<td>0.0001</td>
<td>0.0468</td>
</tr>
<tr>
<td></td>
<td>Two-Stage</td>
<td>110.43</td>
<td>0.1281</td>
<td>0.9813</td>
<td>107.68</td>
<td>0.1310</td>
<td>0.9530</td>
</tr>
<tr>
<td></td>
<td>One-Stage</td>
<td>100.17</td>
<td>0.0451</td>
<td>0.9784</td>
<td>95.76</td>
<td>0.0607</td>
<td>0.9192</td>
</tr>
<tr>
<td></td>
<td>Storey</td>
<td>104.97</td>
<td>0.0798</td>
<td>0.9878</td>
<td>100.09</td>
<td>0.0835</td>
<td>0.9375</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(13.39)</td>
<td>(0.0832)</td>
<td>(0.0838)</td>
<td>(14.29)</td>
<td>(0.0838)</td>
<td>(0.0954)</td>
</tr>
<tr>
<td>1000</td>
<td>Nonparametric</td>
<td>176.49</td>
<td>0.0038</td>
<td>0.9014</td>
<td>80.01</td>
<td>0.0009</td>
<td>0.4091</td>
</tr>
<tr>
<td></td>
<td>Parametric</td>
<td>152.97</td>
<td>0.0008</td>
<td>0.7838</td>
<td>71.32</td>
<td>0.0004</td>
<td>0.3654</td>
</tr>
<tr>
<td></td>
<td>Semiparametric</td>
<td>127.03</td>
<td>0.0004</td>
<td>0.6502</td>
<td>8.40</td>
<td>0.0000</td>
<td>0.0427</td>
</tr>
<tr>
<td></td>
<td>Two-Stage</td>
<td>221.74</td>
<td>0.1222</td>
<td>0.9960</td>
<td>216.06</td>
<td>0.1250</td>
<td>0.9670</td>
</tr>
<tr>
<td></td>
<td>One-Stage</td>
<td>201.67</td>
<td>0.0385</td>
<td>0.9940</td>
<td>192.56</td>
<td>0.0572</td>
<td>0.9303</td>
</tr>
<tr>
<td></td>
<td>Storey</td>
<td>208.02</td>
<td>0.0649</td>
<td>0.9973</td>
<td>196.18</td>
<td>0.0682</td>
<td>0.9368</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(16.46)</td>
<td>(0.0246)</td>
<td>(0.0412)</td>
<td>(17.99)</td>
<td>(0.0257)</td>
<td>(0.0528)</td>
</tr>
</tbody>
</table>
We believe that the fully Bayesian approaches may gain some advantage by incorporating informative priors, especially if these information are justified by biological knowledge. If biological investigations are available in the near future, this scientific knowledge can provide us with some guidelines on choosing the prior distributions and their associated hyperparameters.

5.2. Protein Domain Data Analysis. The number of positions considered to be significantly mutated among six selected protein domains are presented in Table 3. Results reveal that the Empirical Bayes (EB) procedure yields the most number of rejections overall while the nonparametric method has the most number of rejections among the fully Bayesian procedures.

Consequently, the most pronounced difference between the Empirical Bayes and fully Bayesian methods is for cd00083. This helix loop helix (HLH) domain is characterized by very few observations toward the tail. As shown Table 3, the nonparametric and parametric methods identified one position (red) whereas the Empirical Bayes method identified seven positions. It is important to investigate the HLH domain further because the discovery of their diverse functions in the cell cycle, cell-lineage development and tumorigenesis (Murre et al., 1994) made it a potential target for new drug therapies for conditions including heart disease and cancer (Jones, 2004). In particular, cd00083 is a small protein domain, present in 91 human genes (171 proteins), the small number of positions and high prevalence in human genes would result in a high number of somatic mutations regardless of biological significance. More importantly, cd00083 is a DNA binding protein domain with mutations known to be associated with prostate cancer, and thus likely to be of cancer relevance.

As displayed in Figure 2, the proposed fully Bayesian methods are more stringent but still identifies cd00083 as an oncodomain, due to one significant position (Position 15) which is an important position since it is part of the DNA binding site. These results suggest that the proposed method successfully eliminated redundant hotspot positions that might lead to lower specificity of the method when applied to small-sized domains without failing to identify oncodomains with putative high cancer relevance. Furthermore, the results obtained for cd00083 and other small-sized domains, demonstrate the ability of the proposed fully Bayesian method introduced in this paper in solving an important statistical issue concerning the correct identification of oncodomains.

In Section 5, we have seen that for data sets with small number of positions which display sparsity in the right-tail part, EB methods fail to control FDR. Overall simulation results suggest that lesser number of rejections
Table 2
Prior Sensitivity Analysis when $f_0$ is ZIGP($\eta = 0.4$, $\lambda = 2$, $\theta = 0.3$), $f_1$ is shifted Binomial, $C = 10$, and $\pi_0 = 0.80$. $R$ is the average number of rejections while $\hat{FDR}$ and $\hat{TPR}$ correspond to the calculated False Discovery Rate and True Positive Rate, respectively. The number in ( ) corresponds to the standard error.

<table>
<thead>
<tr>
<th>$N$</th>
<th>Prior</th>
<th>Nonparametric</th>
<th></th>
<th></th>
<th>Semiparametric</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$R$</td>
<td>$\hat{FDR}$</td>
<td>$\hat{TPR}$</td>
<td>$R$</td>
<td>$\hat{FDR}$</td>
<td>$\hat{TPR}$</td>
</tr>
<tr>
<td>50</td>
<td>Objective</td>
<td>10.70</td>
<td>0.0718</td>
<td>0.9623</td>
<td>7.59</td>
<td>0.0631</td>
<td>0.7699</td>
</tr>
<tr>
<td></td>
<td>Jeffreys’</td>
<td><strong>9.43</strong></td>
<td><strong>0.0159</strong></td>
<td><strong>0.9497</strong></td>
<td>(3.29)</td>
<td>(0.0248)</td>
<td>(0.1906)</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>9.33</td>
<td>0.0141</td>
<td>0.9404</td>
<td>8.84</td>
<td>0.0667</td>
<td>0.8929</td>
</tr>
<tr>
<td></td>
<td>Near Uniform</td>
<td>9.26</td>
<td>0.0115</td>
<td>0.9364</td>
<td>8.85</td>
<td>0.0068</td>
<td>0.8943</td>
</tr>
<tr>
<td></td>
<td>Informative</td>
<td>9.18</td>
<td>0.0097</td>
<td>0.9318</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>Objective</td>
<td><strong>19.05</strong></td>
<td><strong>0.0219</strong></td>
<td><strong>0.9391</strong></td>
<td>12.72</td>
<td>0.0015</td>
<td>0.6477</td>
</tr>
<tr>
<td></td>
<td>Jeffreys’</td>
<td>18.38</td>
<td>0.0083</td>
<td>0.9312</td>
<td>16.80</td>
<td>0.0040</td>
<td>0.8518</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>18.18</td>
<td>0.0069</td>
<td>0.9221</td>
<td>17.07</td>
<td>0.0039</td>
<td>0.8653</td>
</tr>
<tr>
<td></td>
<td>Near Uniform</td>
<td>18.08</td>
<td>0.0065</td>
<td>0.9172</td>
<td>17.14</td>
<td>0.0040</td>
<td>0.8685</td>
</tr>
<tr>
<td></td>
<td>Informative</td>
<td>17.92</td>
<td>0.0055</td>
<td>0.9098</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>Objective</td>
<td>35.50</td>
<td>0.0170</td>
<td>0.8852</td>
<td>16.79</td>
<td>0.0006</td>
<td>0.4260</td>
</tr>
<tr>
<td></td>
<td>Jeffreys’</td>
<td><strong>35.81</strong></td>
<td><strong>0.0069</strong></td>
<td><strong>0.9096</strong></td>
<td>31.07</td>
<td>0.0028</td>
<td>0.7907</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>35.24</td>
<td>0.0058</td>
<td>0.8958</td>
<td>31.93</td>
<td>0.0029</td>
<td>0.8131</td>
</tr>
<tr>
<td></td>
<td>Near Uniform</td>
<td>34.84</td>
<td>0.0050</td>
<td>0.8862</td>
<td>32.14</td>
<td>0.0029</td>
<td>0.8181</td>
</tr>
<tr>
<td></td>
<td>Informative</td>
<td>34.36</td>
<td>0.0043</td>
<td>0.8747</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>Objective</td>
<td>88.87</td>
<td>0.0972</td>
<td>0.7328</td>
<td>43.80</td>
<td>0.0002</td>
<td>0.4443</td>
</tr>
<tr>
<td></td>
<td>Jeffreys’</td>
<td><strong>88.34</strong></td>
<td><strong>0.0049</strong></td>
<td><strong>0.8992</strong></td>
<td>(18.30)</td>
<td>(0.0022)</td>
<td>(0.1714)</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>86.56</td>
<td>0.0039</td>
<td>0.8815</td>
<td>69.49</td>
<td>0.0112</td>
<td>0.7803</td>
</tr>
<tr>
<td></td>
<td>Near Uniform</td>
<td>85.26</td>
<td>0.0034</td>
<td>0.8686</td>
<td>(19.29)</td>
<td>(0.0038)</td>
<td>(0.1760)</td>
</tr>
<tr>
<td></td>
<td>Informative</td>
<td>82.99</td>
<td>0.0030</td>
<td>0.8455</td>
<td>70.02</td>
<td>0.0015</td>
<td>0.7128</td>
</tr>
<tr>
<td>1000</td>
<td>Objective</td>
<td>173.07</td>
<td>0.0596</td>
<td>0.7863</td>
<td>98.65</td>
<td>0.0001</td>
<td>0.5041</td>
</tr>
<tr>
<td></td>
<td>Jeffreys’</td>
<td><strong>178.24</strong></td>
<td><strong>0.0042</strong></td>
<td><strong>0.9101</strong></td>
<td>(22.64)</td>
<td>(0.0010)</td>
<td>(0.1024)</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>176.49</td>
<td>0.0038</td>
<td>0.9014</td>
<td>123.44</td>
<td>0.0003</td>
<td>0.6318</td>
</tr>
<tr>
<td></td>
<td>Near Uniform</td>
<td>174.55</td>
<td>0.0034</td>
<td>0.8918</td>
<td><strong>128.13</strong></td>
<td><strong>0.0005</strong></td>
<td><strong>0.6558</strong></td>
</tr>
<tr>
<td></td>
<td>Informative</td>
<td>170.1</td>
<td>0.0027</td>
<td>0.8695</td>
<td>126.51</td>
<td>0.0006</td>
<td>0.6471</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(20.42)</td>
<td>(0.0021)</td>
<td>(0.1646)</td>
</tr>
</tbody>
</table>
Table 3

Comparison of Number of Rejections among the Protein Domain Data

<table>
<thead>
<tr>
<th>Data</th>
<th>N</th>
<th>EB</th>
<th>NP</th>
<th>SP</th>
<th>GP</th>
<th>Data</th>
<th>N</th>
<th>EB</th>
<th>NP</th>
<th>SP</th>
<th>GP</th>
</tr>
</thead>
<tbody>
<tr>
<td>cd00054</td>
<td>90</td>
<td>45</td>
<td>43</td>
<td>42</td>
<td>40</td>
<td>cd00063</td>
<td>193</td>
<td>100</td>
<td>95</td>
<td>89</td>
<td>89</td>
</tr>
<tr>
<td>cd00083</td>
<td>93</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>cd00590</td>
<td>175</td>
<td>74</td>
<td>69</td>
<td>19</td>
<td>66</td>
</tr>
<tr>
<td>smart00179</td>
<td>182</td>
<td>44</td>
<td>43</td>
<td>38</td>
<td>38</td>
<td>smart00209</td>
<td>188</td>
<td>54</td>
<td>48</td>
<td>14</td>
<td>32</td>
</tr>
</tbody>
</table>

is preferable because the Empirical Bayes method fails to control $\hat{FDR}$. Hence, the results of the real data analysis are consistent with the simulation studies.

6. Conclusion and Future Work. In this paper, we are interested in identifying significant mutation counts for protein domains with small number of positions and which also display sparsity towards the tail part. In Gauran et al. (2018), the NPEB approaches were proposed where the local FDR is estimated from using empirical null distribution. When the number of data is not large enough in protein domain data, we employ fully Bayesian analysis and developed methods which can control a given level of FDR more reliably compared to the NPEB approaches. The proposed Bayesian approaches implement: (i) zero assumption and (ii) different types of modeling of the marginal or alternative distributions, named parametric, semiparametric and nonparametric methods. Based on the simulation studies, there is no single superior fully Bayesian method. However, since some of these fully Bayesian methods have the ability to control $\hat{FDR}$ when the NPEB method fails, it is still worthy of study.

Overall simulation results suggest that lesser number of rejections is preferable when the number of data is relatively small such as around a hundred. The number of identified hotspots in the real data analysis such as cd00083 are consistent with the simulation studies, and therefore, it is expected that the proposed fully Bayesian methods can identify oncodomains with a given level of FDR.

As future work, we consider the following issues. The primary interest in our analysis is to capture the significant mutations within a given protein domain while implementing multiplicity correction. Biologically, we are interested in the protein domains instead of the gene-centric approach because of the functional information these domains carry. However, because each protein domain perform varying functions and are implicated in different diseases including cancer, we cannot simply merge different protein domains. Hence, by the very nature of the data, we analyze the domain one
Fig 2: Results of the EB method (blue) and Nonparametric method (red) to identify hotspots of somatic mutations in cancer patients are projected into the representative protein structure for the DNA-binding domain, the helix-loop-helix (HLH) domain (cd00083) using Cn3D. Position 15 of the HLH domain (red), was identified by both methods and is a DNA binding position.

Moreover, as shown in the figure for the helix-loop-helix domain cd00083, we can observe a spatial structure among the mutation counts. To incorporate the spatial dependence, we can consider Markov Random Field (MRF) model with a new definition of zero assumption. The implementation of this approach will be completely different from the frameworks presented in our
manuscript. In our revision, we specify this as a possible direction in future researches.

7. Supplementary Materials. The Supplementary Materials are organized as follows. Appendix A includes the derivations of the posterior distributions in Section 3.4. The existing literature and description of the Adaptive Metropolis-Hastings within Gibbs Sampling is presented in Appendix B. The details of the proposed Semiparametric and Parametric Model for Bayesian False Discovery Rate in Section 4 are discussed in Appendix C. Lastly, the results of the exhaustive simulation studies presented as figures are available in Appendix D.

References.


Yuichi Shiraishi, Yusuke Sato, Kenichi Chiba, Yusuke Okuno, Yasunobu Nagata, Kenichi Yoshida, Norio Shiba, Yasuhide Hayashi, Haruki Kume, Yukio Homma, et al. An empir-
Bayesian False Discovery Rate for Sparse Count Data


Biostatistics Group, CEMSE Division, King Abdullah University of Science and Technology, Thuwal, KSA

Christopher Yau

Iris Ivy M. Gauran

Department of Statistics, Seoul National University, Seoul, Korea

Junyong Park

Department of Biological Sciences, University of Maryland Baltimore County, USA

Ilia Rattsev

University of California, Institute for Computational Health Science, USA

Thomas A. Peterson

Department of Biological Sciences, University of Maryland Baltimore County, USA

Maricel G. Kann

Department of Mathematics and Statistics, University of Maryland Baltimore County, USA

DoHwan Park

E-mail: dhpark@umbc.edu