Bayesian Analysis (2023) **TBA**, Number TBA, pp. 1–30

Nonparametric Bayes Differential Analysis of Multigroup DNA Methylation Data*

Chiyu Gu†, Veerabhadran Baladandayuthapani‡, and Subharup Guha§

Abstract. DNA methylation datasets in cancer studies are comprised of measurements on a large number of genomic locations called cytosine-phosphate-guanine (CpG) sites with complex correlation structures. A fundamental goal of these studies is the development of statistical techniques that can identify disease genomic signatures across multiple patient groups defined by different experimental or biological conditions. We propose *BayesDiff*, a nonparametric Bayesian approach for differential analysis relying on a novel class of first order mixture models called the Sticky Pitman-Yor process or two-restaurant two-cuisine franchise (2R2CF). The BayesDiff methodology flexibly utilizes information from all CpG sites or biomarker probes, adaptively accommodates any serial dependence due to the widely varying inter-probe distances, and makes posterior inferences about the differential genomic signature of patient groups. Using simulation studies, we demonstrate the effectiveness of the BayesDiff procedure relative to existing statistical techniques for differential DNA methylation. The methodology is applied to analyze a gastrointestinal (GI) cancer dataset exhibiting serial correlation and complex interaction patterns. The results support and complement known aspects of DNA methylation and gene association in upper GI cancers.

Keywords: 2R2CF, first order models, mixture models, sticky Pitman-Yor process, two-restaurant two-cuisine franchise.

1 Introduction

Recent advances in array-based and next-generation sequencing (NGS) technologies have revolutionized biomedical research, especially in cancer. The rapid decline in the cost of genome technologies has facilitated the availability of datasets involving intrinsically different sizes and scales of high-throughput data and provided genome-wide, high resolution information about the biology of cancer. A common analytical goal is the identification of differential genomic signatures between groups of samples corresponding to different treatments or biological conditions, e.g., treatment arms, response to adjuvant chemotherapy, tumor subtypes, or cancer stages. Challenges include the high dimensionality of genomic biomarkers or probes, usually in the hundreds of thousands,

*This work was supported by the National Science Foundation and National Institutes of Health under Grants DMS-1854003, R01 CA269398, and U01 CA209414 to SG, and Grants DMS-1463233 and R01 CA160736 to VB. The authors thank the anonymous Editor, Associate Editor, and two referees for many insightful comments that improved the content and presentation of the paper.

†Formerly at the University of Missouri. Currently employed at Bayer Crop Science, 700 Chesterfield Pkwy W, Chesterfield, MO 63017, guchiyu@gmail.com

‡Department of Biostatistics, University of Michigan, veerab@umich.edu

§Department of Biostatistics, University of Florida, s.guha@ufl.edu

 https://doi.org/10.1214/23-BA1407

and the relatively small number of patient samples, usually no more than a few hundred. This "small n, large p" setting results in unstable inferences due to collinearity. Further, the data exhibit complex interaction patterns, such as signaling or functional pathway-based interactions, and location-based serial correlation for high-throughput sequencing data. These characteristics challenge statistical techniques for detecting differential signatures.

Differential DNA Methylation in Cancer Studies DNA methylation is an epigenetic mechanism that involves the addition of a methyl (CH_3) group to DNA, resulting in the modification of gene functions. It typically occurs at genomic locations called cytosine-phosphate-guanine (CpG) sites. Alterations in DNA methylation, e.g., hypomethylation of oncogenes and hypermethylation of tumor suppressor genes, are often associated with the development and progression of cancer (Feinberg and Tycko, 2004). It was previously believed that these alterations occur almost exclusively at promoter regions known as CpG islands, i.e., chromosomal regions with high concentrations of CpG sites. However, with the advent of high-throughput technologies, it has been shown that a significant proportion of cancer-related alterations do not occur in promoters or CpG islands (Irizarry et al., 2009), prompting higher resolution, epigenome-wide investigations.

Gastrointestinal (GI) cancer, the most common form of cancer in the U.S. (Siegel et al., 2017), refers to malignant conditions affecting the digestive system associated with epigenetic alterations (Vedeld et al., 2017). Molecular characterization of different cancer types, facilitated by the identification of differentially methylated CpG sites, is therefore key to better understanding GI cancer. In the motivating application, we analyze methylation profiles publicly available from The Cancer Genome Atlas (TCGA) project and comprising 1,224 tumor samples belonging to four GI cancers of the upper digestive tract: stomach adenocarcinoma (STAD), liver hepatocellular carcinoma (LIHC), esophageal carcinoma (ESCA) and pancreatic adenocarcinoma (PAAD). For 485,577 probes, where each probe is mapped to a CpG site, DNA methylation levels or Beta-values ranging from 0 (no methylation) to 1 (full methylation) were measured using the Illumina Human Methylation 450 platform.

Figure 1 displays the methylation levels for CpG sites near TP53, a tumor suppressor gene located on chromosome 17. A random subset of the tumor samples was chosen to facilitate an informal visual evaluation. Each plotted point represents the methylation level of a tumor sample at a CpG site. As indicated in the figure legend, the four sets of colors and shapes of the points represent the four upper GI cancers. The vertical dashed lines indicate the boundaries of the TP53 gene. Although differential methylation is clearly visible at some CpG sites, the differences are generally subtle, demonstrating the need for sophisticated statistical analyses. An obvious feature is the correlation of the apparent methylation statuses of nearby CpG sites (Eckhardt et al., 2006; Irizarry et al., 2008; Leek et al., 2010). The dependence of proximal CpG sites is also seen in Figure 1 of Supplementary Material (Gu et al., 2023) with highly significant tests for serial correlations. Furthermore, the highly variable inter-probe spacings in Figure 1 suggests the need to model distance-based dependencies.

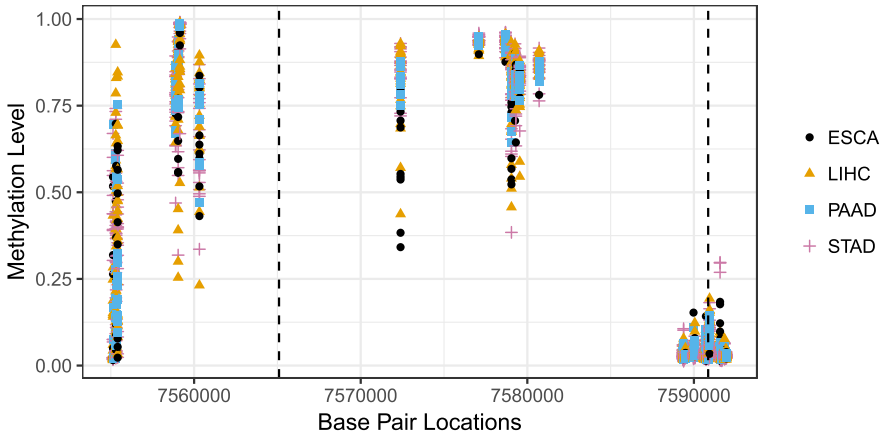


Figure 1: Methylation levels of CpG sites near gene TP53 for randomly chosen tumor samples of the TCGA upper GI dataset. Each plotted point represents the methylation level at a CpG site, with the shapes and colors corresponding to the different GI cancers indicated in the legend. The vertical dashed lines demarcate the TP53 gene boundaries.

Existing Statistical Approaches for Differential DNA Methylation and Limitations Numerous frequentist and Bayesian methods have been developed for differential DNA methylation, and can be broadly classified into four categories: *(i) Testing-based methods*, such as Illumina Methylation Analyzer (IMA) (Wang et al., 2012), City of Hope CpG Island Analysis Pipeline (COHCAP) (Warden et al., 2013), and BSmooth (Hansen et al., 2012). These methods rely on two-sample or multiple-sample tests for the mean group differences at each CpG site. *(ii) Regression based models*, such as Methylkit (Akalin et al., 2012), bump hunting (Jaffe et al., 2012), Biseq (Hebestreit et al., 2013), and RADMeth (Dolzhenko and Smith, 2014). After applying smoothing or other adjustments, these methods fit individual regression models for each CpG site and test for significance. *(iii) Beta-binomial model-based methods*, such as MOABS (Sun et al., 2014), DSS (Feng et al., 2014), and methylSig (Park et al., 2014). These methods fit separate models to each CpG site. *(iv) Hidden Markov models (HMMs)*, such as MethPipe (Song et al., 2013), Bisulfighter (Saito et al., 2014), and HMM-DM (Yu and Sun, 2016). These methods detect differentially methylated sites using inferred hidden states.

The aforementioned methods have several deficiencies. Because they fit separate models to each probe, most methods ignore the strong correlations between neighboring probes, reducing detection power. Additionally, beta-binomial, HMM, and most testing-based methods are able to accommodate only two treatments and rely on inefficient adjustments for multiple treatments. The methods that account for serial dependence (e.g., HMMs) do not adjust for the widely varying inter-probe distances, and instead, assume uniform inter-probe dependencies. The few methods that account for inter-probe distances (e.g., Hansen et al., 2012; Jaffe et al., 2012; Hebestreit et al., 2013) rely on ad hoc parameter-tuning procedures that do not adjust for the data characteristics.

Motivated by these challenges, we propose general and flexible methodology for differential analysis in DNA methylation data, referred to as *BayesDiff*. Rather than fitting a separate model to each CpG site or probe, BayesDiff relies on a common analytical framework for simultaneous inferences that adapts to the unique data attributes. To diminish collinearity effects and achieve dimension reduction, the probes are allocated to a smaller, unknown number of latent clusters based on the similarities of probe-specific multivariate parameters. Posterior inferences are made on differential state variables to delineate the disease genomic signature of multiple treatments.

For realistically modeling the probe-cluster allocation mechanism of DNA methylation profiles, we devise an extension of Pitman-Yor processes (PYPs) (Perman et al., 1992) called the *Sticky PYP* (equivalently, the *two-restaurant two-cuisine franchise*). In addition to accounting for long-range biological interactions, this nonparametric process accommodates distance-based serial dependencies. Separately for the differential and non-differential probes, the data flexibly direct the choice between PYPs, and their special case, Dirichlet processes, in finding the best-fitting allocation schemes.

We implement an inferential procedure for Sticky PYPs using a Markov chain Monte Carlo (MCMC) algorithm specifically designed for posterior inferences in the typically large methylation datasets. Simulation results show that our approach significantly outperforms existing methods for multigroup comparisons in data with or without serial correlation. For the motivating TCGA dataset, in addition to confirming known features of DNA methylation and disease-gene associations, the analysis reveals interesting aspects of the biological mechanisms of upper GI cancers.

The rest of the paper is organized as follows. Section 2 describes the BayesDiff approach, with Section 2.1 introducing the Sticky PYP or two-restaurant two-cuisine franchise (2R2CF) for differential DNA methylation. Section 3 outlines an effective inference procedure for detecting differential probes. Section 4 uses artificial datasets with varying noise and correlation levels to assess the accuracy of BayesDiff in detecting disease genomic signatures and compares BayesDiff with established techniques for DNA methylation data. The motivating upper GI dataset is analyzed in Section 5. Finally, conclusions and future work are discussed in Section 6.

2 The BayesDiff Model

Sequencing technologies measure DNA methylation levels of p biomarkers represented by CpG sites ("probes") and n matched patient or tissue samples ("individuals"). Usually, p is much larger than n. The methylation levels, which belong to the interval $[0, 1]$, are arranged in an $n \times p$ matrix of proportions, $\mathbf{X} = ((x_{ij}))$, for individuals i and probes j, with the probes sequentially indexed by their genomic locations. The distances between adjacent probes are denoted by $e_1, \ldots, e_{p-1}$, and typically exhibit high variability. For instance, in the upper GI TCGA dataset, the inter-probe distances range from 2 base pairs to a million base pairs; a base pair is a unit of DNA length consisting of two nucleobases bound to each other by hydrogen bonds (e.g., Baker et al., 2008).

Each individual i is associated with a known experimental or biological condition

("treatment") denoted by t_i and taking values in $\{1, \ldots, T\}$ with $T \geq 2$. In the motivating TCGA data, there are $T = 4$ upper GI cancer types. We model the logit transformation of the methylation levels, $z_{ij} = \log(x_{ij}/(1 - x_{ij}))$, as follows:

$$z_{ij} \sim N\left(\xi_i + \chi_j + \theta_{t_i j}, \sigma^2\right), \tag{1}$$

where ξ_i represents the ith subject's random effect, χ_j represents the jth probe's random effect, and θ_{tj} is the treatment t–probe j interaction random effect; refer to the directed acyclic graph (DAG) in Figure 6 of Supplementary Material (Gu et al., 2023). Logit methylation levels differ from M-values (Du et al., 2010), commonly used in differential analyses, by $1 - \log(2) = 0.306$; however, the key results are identical for both transformations.

The main inferential goal is the detection of differential probes, i.e., probes j, for which $\boldsymbol{\theta}_j = (\theta_{1j}, \ldots, \theta_{Tj})'$ does not have all identical elements. Consequently, we define a binary *differential state variable*, s_j, with $s_j = 1$ indicating that probe j is not differential and $s_j = 2$ indicating that it is differential:

$$s_j = \begin{cases} 1 & \text{if } \theta_{1j} = \theta_{2j} = \cdots = \theta_{Tj}, \\ 2 & \text{otherwise,} \end{cases} \tag{2}$$

for $j = 1, \ldots, p$. The parameters of interest are $s_1, \ldots, s_p$, with the differential genomic signature consisting of probes with $s_j = 2$. Figure 6 of Supplementary Material (Gu et al., 2023) positions the differential state variables in the BayesDiff model hierarchy. Motivated by the distance-dependent correlations of DNA methylation data and the deficiencies of existing statistical approaches, this paper fosters a Bayesian nonparametric framework for random effects $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p$ underlying the differential state variables.

Modeling Probe Clusters

In addition to high-dimensionality, the analytical challenges include pervasive collinearity caused by dependencies between physically proximal probes. Additionally, nonadjacent probes may have long-range dependencies due to biological interactions, e.g., signaling or functional pathways. To accommodate these complex dependence structures and extract information from the large number of probes, we allocate the p probes to a much smaller number, q, of latent clusters based on the similarities of their random effects $\boldsymbol{\theta}_j$. We favor clustering to dimension reduction methods such as principal components analysis (PCA); each PC being a linear combination of all p biomarkers, PCA is less useful because it is unable to select sparse features, i.e., probes. By contrast, the proposed approach facilities biological interpretations by identifying CpG sites relevant to the differential genomic signatures between multiple treatments.

Suppose an *allocation variable*, c_j, assigns probe j to one of q latent clusters, where q is unknown. The event $[c_j = k]$ indicates that the j^{th} probe is assigned to the k^{th} latent cluster, $k = 1, \ldots, q$. We assume that the q clusters are associated with *latent vectors*, $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_q$, where the probe-specific random effects and cluster-specific latent vectors have the relation:

$$\boldsymbol{\theta}_j = \boldsymbol{\lambda}_k \quad \text{if } c_j = k. \tag{3}$$

That is, all probes in a cluster are assumed to have identical random effects equal to that cluster's latent vector. The differential state variables, defined in equation (2), then become a shared attribute of their parent cluster, and clusters as a whole are either differentially or non-differentially methylated. Further, if probe j belongs to cluster k (i.e., $c_j = k$), then the condition $\theta_{1j} = \theta_{2j} = \cdots = \theta_{Tj}$ in equation (2) is equivalent to $\lambda_{1k} = \lambda_{2k} = \cdots = \lambda_{Tk}$, and the differential cluster indexes comprise

$$\mathcal{D} = \{k : \lambda_{tk} \neq \lambda_{t'k}, \text{ for some } t \neq t', k = 1, \ldots, q\}. \tag{4}$$

Mixture Models for Allocation Bayesian infinite mixture models are a natural choice for allocating p probes to a smaller, unknown number of latent clusters. Dirichlet processes (Ferguson, 1973) are arguably the most frequently used infinite mixture models; see Müller and Mitra (2013, chap. 4) for a comprehensive review. The use of Dirichlet processes to achieve dimension reduction has precedence in the literature, albeit in unrelated applications (see Medvedovic et al., 2004; Kim et al., 2006; Dunson et al., 2008; Dunson and Park, 2008; Guha and Baladandayuthapani, 2016). Lijoi, Mena, and Prünster (2007a) advocated the use of Gibbs-type priors (Gnedin and Pitman, 2006) for accommodating more flexible clustering mechanisms and demonstrated the utility of Pitman-Yor processes (PYPs) in genomic applications. An overview of Gibbs-type priors and characterization of the learning mechanism is provided by De Blasi et al. (2015). Formally, the PYP (Perman et al., 1992) relies on a discount parameter $d \in [0, 1)$, positive mass parameter α, and T-variate base distribution W, and is denoted by $\mathcal{W}(d, \alpha, W)$. The value $d = 0$ yields a Dirichlet process with mass parameter α and base distribution W. Suppose $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p$ are distributed as $\mathcal{W}(d, \alpha, W)$. The *stick-breaking representation* of $\mathcal{W}(d, \alpha, W)$ is $\boldsymbol{\theta}_j \overset{\text{i.i.d.}}{\sim} \mathcal{P}$, where random distribution $\mathcal{P}$ is the discrete mixture $\sum_{v=1}^{\infty} \pi_v \delta_{\boldsymbol{\phi}_v}$, with $\delta_{\boldsymbol{\phi}_v}$ denoting a point mass located at the atom $\boldsymbol{\phi}_v \overset{\text{i.i.d.}}{\sim} W$. The random stick-breaking probabilities have the form $\pi_1 = V_1$, and $\pi_h = V_h \prod_{v=1}^{h-1}(1 - V_v)$ for $h > 1$, where $V_h \overset{\text{indep}}{\sim} \text{beta}(1 - d, \alpha + hd)$. Guha and Baladandayuthapani (2016) introduced VariScan, a technique that utilizes PYPs and Dirichlet processes for clustering, variable selection, and prediction in high-dimensional regression problems in general, and in gene expression datasets in particular. They also demonstrated that PYPs are overwhelmingly favored over Dirichlet processes in gene expression datasets, which typically exhibit no serial correlation.

Limitations of Existing Mixture Models Although the aforementioned mixture models achieve dimension reduction and account for long-range biological interactions between non-adjacent probes, a potential drawback is their implicit assumption of a priori exchangeability of the probes. Consequently, these techniques cannot accommodate serial correlation in methylation data. Infinite HMMs, such as the hierarchical Dirichlet process hidden Markov model (HDP-HMM) (Teh et al., 2006) and Sticky HDP-HMM (Fox et al., 2011), could be utilized to fill this gap. Although these models are a step in the right direction, they have several undesirable features for differential analysis. *First*, the degree of first order dependence is uniform irrespective of the inter-probe distances. This is unrealistic in methylation datasets where the correlation typically

decreases with inter-probe distance (Hansen et al., 2012; Jaffe et al., 2012; Hebestreit et al., 2013). *Second*, an ad hoc exploratory analysis of the GI cancer dataset reveals that the serial correlation in the treatment-probe effects is weaker than the serial dependence between the differential state variables in equation (2). Although there may not be a biological explanation for this phenomenon, this makes sense from a statistical perspective because the differential states are binary functions of the treatment-probe interactions; the differential states are more sensitive in detecting first order dependence even when the higher-dimensional (and noisier) treatment-probe interactions show negligible correlation. This suggests that a hypothetical two-group Markov model, rather than an infinite-group Markov model such as HDP-HMM or Sticky HDP-HMM, would provide a better fit for the data. *Third*, the range of allocation patterns supported by infinite HMMs is relatively limited. In particular, realistic allocation patterns, such as power law decays in the cluster sizes and large numbers of small-sized clusters, a common feature of cancer datasets (Lijoi et al., 2007b), are assigned relatively small prior probabilities by infinite HMMs.

2.1 Sticky PYP: A Two-restaurant, Two-cuisine Franchise (2R2CF) for Differential Analysis

The proposed Sticky PYP comprises a cohort of regular PYPs producing the probe-specific random effects by switching the generative PYP at random locations along the probe sequence. Alternatively, the Chinese restaurant franchise (CRF) metaphor for HDP-HMMs and Sticky HDP-HMMs can be generalized to the *two-restaurant two-cuisine franchise* (2R2CF) to give an equivalent representation of Sticky PYPs appropriate for differential analysis. We first present a descriptive overview of 2R2CF.

Imagine a franchise with two restaurants labeled 1 and 2. Each restaurant consists of two sections, labeled section 1 and 2. Each section serves a single cuisine and the section-cuisine menu consists of infinite dishes. Section 1 of both restaurants exclusively serves cuisine 1. The cuisine 1 menus of restaurant 1 and 2, along with the selection probabilities of the dishes, are identical. Similarly, section 2 of both restaurants exclusively serves cuisine 2, and the cuisine 2 menus of the two restaurants are identical.

A succession of p customers, representing the CpG sites or probes, arrive at the franchise. The waiting times between successive customers correspond to the inter-probe distances, $e_1, \ldots, e_{p-1}$. Each customer first selects a restaurant and then a section (equivalently, cuisine) in that restaurant. Each restaurant section has an infinite number of tables, and a customer either sits at a table already occupied by the previous customers or sits at a new table. All customers at a table are served the same dish chosen from the section-cuisine menu by the first customer who sat at that table. The first customer at a table independently picks a dish from the infinite cuisine menu with a cuisine-specific probability associated with each dish. As a result, multiple tables at a restaurant section may serve the same dish.

Restaurant 1 specializes in cuisine 1. Consequently, section 1 is more popular with the restaurant 1 patrons. Similarly, restaurant 2 specializes in cuisine 2, and so, restaurant 2 customers tend to favor section 2 over section 1. By design, if a customer has eaten a

cuisine 1 (2) dish, then the next customer is more likely to visit restaurant 1 (2), where cuisine 1 (2) is more popular. In this manner, each customer tends to select the same cuisine as the previous customer.

In the metaphor, cuisine 1 symbolizes the non-differential state and cuisine 2 symbolizes the differential state. The dish that franchise customer j eats represents the probe-specific random effect, $\boldsymbol{\theta}_j$. Since cuisine 1 represents the non-differential state, its dishes are characterized by T-variate random vectors with all equal elements; see equation (2). In contrast, cuisine 2 (differential state) dishes are characterized by T-variate random vectors with at least two unequal elements.

The dependence in the restaurant and cuisine choices of consecutive customers account for the long runs of differential or non-differential states seen in DNA methylation data. However, a customer's influence on the next customer diminishes as the time interval between the two customers increases; the differential statuses of two adjacent probes are statistically independent in the limit as the inter-probe distance grows.

The 2R2CF process is illustrated in Figure 2 and discussed below in greater detail. The following specification conditions on G, an unknown distribution in $\mathcal{R}$ that is assigned a Dirichlet process prior with mass parameter $\beta > 0$ and univariate normal base distribution, $G_0 = N\left(\mu_G, \tau_G^2\right)$. The stick-breaking representation of the Dirichlet process implies that distribution G is almost surely discrete because it has the infinite mixture distribution:

$$G \overset{d}{=} \sum_{v=1}^{\infty} \varpi_v \delta_{\zeta_v}, \text{ where } \sum_{v=1}^{\infty} \varpi_v = 1 \text{ and } \zeta_v \overset{\text{i.i.d.}}{\sim} G_0. \tag{5}$$

The distribution of the random probabilities, ϖ_v, which depend on mass parameter β, was derived in Sethuraman (1994); see also Ishwaran and James (2003) and Lijoi and Prünster (2010). In the sequel, we condition on distribution G; equivalently on the probabilities, ϖ_v, and univariate atoms, ζ_v, for $v \in \mathcal{N}$, the natural numbers.

Cuisine 1 Menu Recall that cuisine 1 represents the non-differential state, for which the T-variate random vectors (i.e., dishes in the metaphor) has all equal elements. Cuisine 1 menu, with its countably infinite dishes and their associated probabilities, is modeled as a discrete *menu distribution*, W_1, in $\mathcal{R}^T$. With $\mathbf{1}_T$ denoting the column vector of T ones, let

$$\boldsymbol{\vartheta} \mid \psi = \psi \mathbf{1}_T \text{ where } \psi \sim G. \tag{6}$$

Cuisine 1 menu distribution W_1 is defined as the law of random vector $\boldsymbol{\vartheta}$. Then $\mathcal{S}_1 = \{\zeta_v \mathbf{1}_T : v \in \mathcal{N}\}$ represents the available cuisine 1 dishes and the support of W_1. The selection probability associated with dish $\zeta_v \mathbf{1}_T$ is ϖ_v.

The continuity of base distribution G_0 in equation (5) guarantees that the menu W_1 dishes are unique. On the other hand, the discreteness of distribution G has practical implications for 2R2CF: *(a)* cuisine 1 consists of discrete dishes, as required, rather than a continuous spectrum, and *(b)* since section 1 at both restaurants serve the same menu, two section 1 customers may eat the same dish even if they select different restaurants.

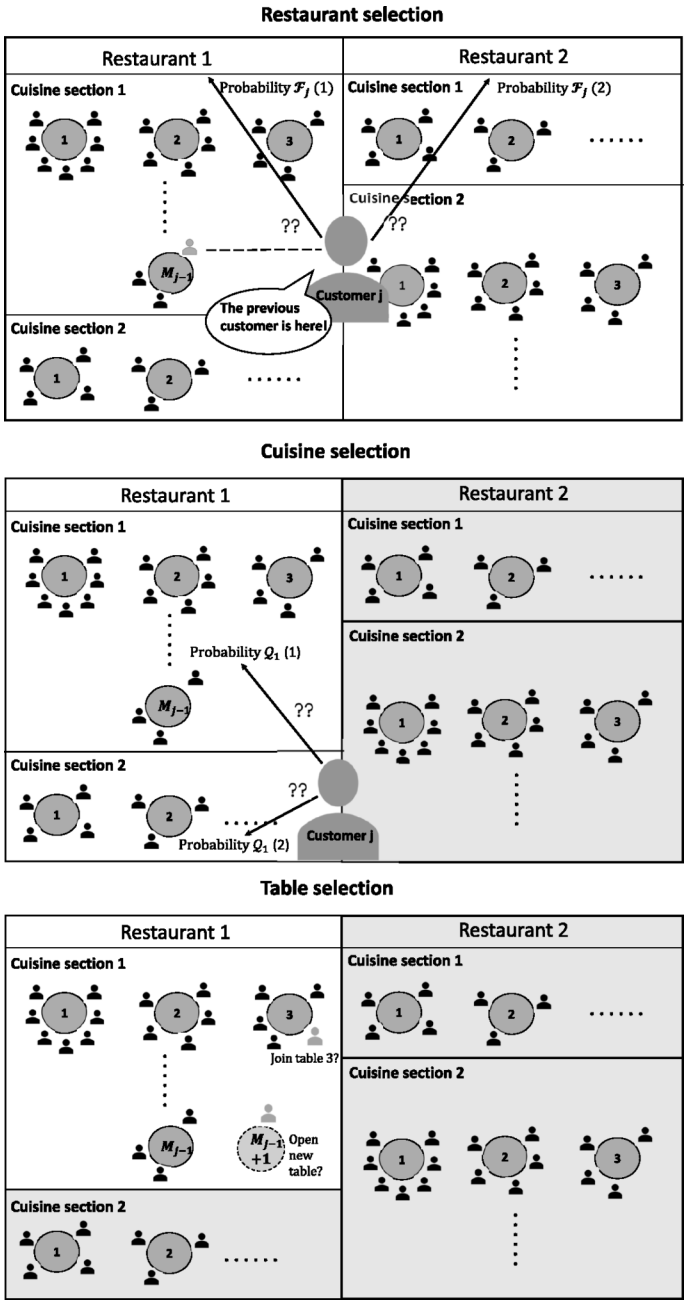


Figure 2: Cartoon representation of the two-restaurant two-cuisine franchise for differential analysis, showing the progressive choice of restaurant, cuisine section, and table by customer j, for $j > 1$. The numbered circles represent table numbers. See the text in Section 2.1 for a detailed description of the 2R2CF process.

Cuisine 2 Menu As mentioned, cuisine 2 depicts the differential state and its dishes represent T-variate random vectors with at least two unequal elements. Its menu comprises countably infinite cuisine 2 dishes along with associated probabilities. The menu is therefore modeled by a T-variate distribution, W_2, satisfying two conditions: *(i)* it has a countably infinite support, and *(ii)* each T-variate atom of W_2 has at least two unequal elements. For any given $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_T)' \in \mathcal{R}^T$, a probability mass function for W_2 that satisfies these two conditions is

$$W_2(\boldsymbol{\phi}) = \begin{cases} \prod_{t=1}^{T} G(\phi_t) / \left(1 - \sum_{v=1}^{\infty} \varpi_v^T\right) & \text{if } \phi_t \neq \phi_{t'} \text{ for some } t \neq t', \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where $G(\phi)$ denotes the mass function of distribution G evaluated at $\phi \in \mathcal{R}$. For a graphical summary of the parameters in equations (5)–(7), refer to the "Cuisine menus" block of Figure 6 of Supplementary Material (Gu et al., 2023). In line 1 of expression (7), normalizing constant $(1 - \sum_{v=1}^{\infty} \varpi_v^T)$ is the total probability that a T-variate random vector whose elements are i.i.d. G has at least two distinct elements. Referring back to the atoms, ζ_v, of distribution G in (5), $\mathcal{S}_2 = \{\boldsymbol{\phi} : \phi_t = \zeta_v \text{ for some } v \in \mathcal{N}, \text{ but not all } \phi_t \text{ are identical for } t = 1, \ldots, T\}$ is the list of cuisine 2 dishes and the support of W_2. The selection probability associated with dish $\boldsymbol{\phi}$ is $W_2(\boldsymbol{\phi})$.

Restaurant, Section, Table, and Dish Choices of the 2R2CF Customers

Let the restaurant chosen by franchise customer j be denoted by g_j and the chosen cuisine (i.e., section) be denoted by s_j. Suppose he or she sits at table v_j in that restaurant section and eats dish $\boldsymbol{\theta}_j$.

Customer 1 At time 0, suppose the first customer selects restaurant $g_1 = 1$ with probability $\rho_1 > 0$ and selects restaurant $g_1 = 2$ with probability $\rho_2 = 1 - \rho_1 > 0$. For reasons that will become clear, we refer to ρ_1 as the *baseline non-differential proportion* and ρ_2 as the *baseline differential proportion*. Typically, the differential state is less frequent, and so $\rho_2 < \rho_1$ (i.e., $\rho_2 < 1/2$). Proportion ρ_1 is given a uniform prior on the interval $(1/2, 1)$.

Choice of cuisine s_1 Next, customer 1 selects a section in restaurant g_1. Since each restaurant specializes in its namesake cuisine, the cuisine is more popular with the restaurant customers. This is modeled as follows. Within restaurant g_j (with $j = 1$ for customer 1), customer j selects cuisine 1 with probability

$$\mathcal{Q}_{g_j}(1) = \begin{cases} \rho_1 + \rho_2\gamma & \text{if } g_j = 1, \\ \rho_1 - \rho_1\gamma & \text{if } g_j = 2, \end{cases} \tag{8}$$

for a *speciality cuisine popularity parameter*, $\gamma \in (0, 1)$, determining the degree to which a restaurant's patrons favor its namesake cuisine. For instance, if γ is nearly 1, then a restaurant 1 (2) customer almost always (never) chooses cuisine 1. At the other extreme, if γ is nearly 0, then the customer chooses cuisine 1 with approximate probability ρ_1 irrespective of the restaurant. Parameter γ is assigned an independent uniform prior

on the unit interval. The probability that a restaurant g_j customer chooses cuisine 2 is then $\mathcal{Q}_{g_j}(2) = 1 - \mathcal{Q}_{g_j}(1)$.

Choice of table v_1 ***and dish*** $\boldsymbol{\theta}_1$ Within section s_1 of restaurant g_1, we assume without loss of generality (since the table identifiers are arbitrary) that customer 1 sits at table $v_1 = 1$. At table 1, customer 1 randomly orders a cuisine s_1 dish from menu distribution W_{s_1}. The dish he or she eats represents the random effect of the first probe. That is, $\boldsymbol{\theta}_1 \mid s_1 \sim W_{s_1}$. As the 2R2CF process evolves as more patrons arrive, the tables in a restaurant's section are sequentially assigned new labels as they are occupied.

Customer j**, for** $j > 1$ The *restaurant* choice of a subsequent customer is influenced by the previous customer's *cuisine* and waiting time. Suppose customer j arrives at the franchise after a time interval of e_{j-1} following the $(j-1)$th customer. Without loss of generality, $e_1, \ldots, e_{p-1}$ can be scaled so that their total equals 1. Since the probes in differential analysis typically represent CpG sites on a chromosome or gene, it has a scaled length of 1.

To model the dependencies between the franchise customers, we define a non-negative ***dependence parameter*** η that transforms waiting time e_{j-1} to an ***affinity*** measure between customer $(j-1)$ and customer j:

$$r_j = \exp(-e_{j-1}/\eta), \qquad j > 1. \tag{9}$$

The affinity measure belongs to the unit interval when $\eta > 0$. If $\eta = 0$, r_j is defined as 0 irrespective of waiting time e_{j-1}. The affinity influences the restaurant choice through assumption (10) below.

Choice of restaurant g_j The cuisine s_{j-1} of the $(j-1)$th customer influences the restaurant choice of the jth customer through affinity r_j and popularity parameter γ:

$$g_j \,\big|\, s_{j-1}, \rho_1, \eta, \gamma \sim \mathcal{F}_j.$$

Specifically, the probability that customer j selects restaurant 1 is assumed to be

$$\mathcal{F}_j(1) \stackrel{def}{=} P\big(g_j = 1 \mid s_{j-1}, \rho_1, \eta, \gamma\big) = \begin{cases} \rho_1 + \rho_2 r_j/\gamma & \text{if } s_{j-1} = 1, \\ \rho_1 - \rho_1 r_j/\gamma & \text{if } s_{j-1} = 2, \end{cases} \tag{10}$$

and $\mathcal{F}_j(2) = 1 - \mathcal{F}_j(1)$. The idea is illustrated in the top panel of Figure 2, where customer j chooses restaurant 1 with probability $\mathcal{F}_j(1)$ and restaurant 2 with probability $\mathcal{F}_j(2)$. If dependence parameter $\eta = 0$, then the restaurant choices of the customers are independent; specifically, $\mathcal{F}_j(1) = \rho_1$ irrespective of the cuisine s_{j-1}.

It can be verified that $\mathcal{F}_j$ is a probability mass function if and only if $r_j/\gamma < 1$. Since the scaled waiting times are bounded above by 1, a globally sufficient condition is $\eta < -1/\log\gamma$. We therefore assume a mixture prior for dependence parameter η:

$$\eta \mid \gamma \sim \frac{1}{2}\delta_0 + \frac{1}{2}\mathcal{H} \cdot \mathcal{I}(\eta < -1/\log\gamma), \tag{11}$$

where the second mixture component involves a continuous distribution, $\mathcal{H}$, restricted to the interval $[0, -1/\log\gamma)$, enforcing the globally sufficient condition. In our experience, posterior inferences on η are relatively robust to the continuous prior $\mathcal{H}$ provided the prior is not highly concentrated on a small part of interval $[0, -1/\log\gamma)$. Refer to the "Restaurant and cuisine of customer j" block of Figure 6 of Supplementary Material (Gu et al., 2023) for a DAG depicting the relationships of these parameters.

When $\eta = 0$, we have a *zero-order Sticky PYP*; when $\eta > 0$, we obtain a *first order Sticky PYP*. Some interesting consequences of specification (10) are:

1. **Zero-order Sticky PYP**: When $\eta = 0$, each customer independently chooses restaurant 1 (or 2) with a baseline probability of ρ_1 (or ρ_2). The p customers act identically.
2. **First order Sticky PYP with e_{j-1}/η large**: At large relative distances, customer j acts approximately independently of the history. Somewhat similarly to customer 1, customer j chooses restaurant 1 (2) with a probability approximately, but not exactly, equal to ρ_1 (ρ_2).
3. **First order Sticky PYP with e_{j-1}/η small**: In the limit as $e_{j-1}/\eta \to 0$ (e.g., for a small inter-probe distance e_{j-1}), the restaurant choice of customer j follows a hidden Markov model.

Since it drives the dependence characteristics of DNA methylation data, parameter η is of interest. Prior specification (11) allows the data to direct the model order through posterior probability, $P[\eta = 0 \mid \mathbf{X}]$, an MCMC estimate of which is readily available; see Section 3.

Choice of cuisine s_j In restaurant g_j, customer j selects cuisine-section s_j with distribution $\mathcal{Q}_{g_j}$ defined in expression (8). For bookkeeping purposes, among franchise customers $1, \ldots, j$, let $p_{gs}^{(j)}$ be the number of customers that choose section s in restaurant g; that is, $p_{gs}^{(j)} = \sum_{l=1}^{j} \mathcal{I}(g_l = g, s_l = s)$ for $g, s = 1, 2$.

For a graphical depiction of cuisine selection by the jth customer, see the middle panel of Figure 2, where $g_j = 1$. That is, customer j, having already chosen restaurant 1, now chooses a cuisine-section. Restaurant 2 has been greyed out because it is no longer accessible to this customer. In the lower panel of Figure 2, we find that the customer picked cuisine-section 1, and so $s_j = 1$.

Choice of table v_j Applying the above notation, among customers $1, \ldots, (j-1)$, there are $p_{g_j s_j}^{(j-1)}$ customers in the same restaurant and section as the jth customer. Suppose these customers have occupied tables $1, \ldots, M_{g_j s_j}^{(j-1)}$, and that there are $p_{g_j s_j k}^{(j-1)}$ customers seated at the kth table. Let $\mathcal{M}_{j-1} = \{p_{g_j s_j k}^{(j-1)} : k = 1, \ldots, M_{g_j s_j}^{(j-1)}\}$ comprise these aggregated table occupancies.

Recall that the newly arrived jth customer may sit at any of the $M_{g_j s_j}^{(j-1)}$ occupied tables or a new $(M_{g_j s_j}^{(j-1)} + 1)$th table. Two possibilities are illustrated in the lower panel

of Figure 2. For a PYP with mass parameter α_{s_j} and cuisine-specific discount parameter $d_{s_j} \in [0, 1)$, the predictive distribution of table v_j of customer j is related to the table occupancies as follows:

$$P\Big(v_j = k \mid \mathcal{M}_{j-1}\Big) \propto \begin{cases} p^{(j-1)}_{g_j s_j k} - d_{s_j} & \text{if } k = 1, \ldots, M^{(j-1)}_{g_j s_j}, \\ \alpha_{s_j} + M^{(j-1)}_{g_j s_j} d_{s_j} & \text{if } k = (M^{(j-1)}_{g_j s_j} + 1), \end{cases} \tag{12}$$

where the second line corresponds to customer j sitting at a new table, in which case the new number of occupied tables is $M^{(j)}_{g_j s_j} = M^{(j-1)}_{g_j s_j} + 1$ and table index $v_j = M^{(j)}_{g_j s_j}$. Otherwise, if customer j sits at a previously occupied table, then table index $v_j \leq M^{(j-1)}_{g_j s_j}$ and the number of occupied tables remains unchanged: $M^{(j)}_{g_j s_j} = M^{(j-1)}_{g_j s_j}$. See the DAG in Figure 6 of Supplementary Material (Gu et al., 2023).

The above predictive distribution implies that customer j is more likely to choose tables with several occupants, positively reinforcing that table's popularity for future customers. The number of occupied tables stochastically increases with the PYP mass and discount parameter.

For section s, if the PYP discount parameter $d_s = 0$, we obtain the well-known Pòlya urn scheme for Dirichlet processes (Ferguson, 1973). PYPs act as effective dimension reduction devices because the random number of occupied tables is much smaller than the number of customers. In general, as the number of patrons in section s of restaurant g grows as more customers arrive at the franchise, that is, as $p^{(j)}_{gs} \to \infty$, the number of occupied tables, $M^{(j)}_{gs}$, is asymptotically equivalent to

$$\begin{cases} \alpha_s \log p^{(j)}_{gs} & \text{if } d_s = 0, \\ T_{d_s \alpha_s} (p^{(j)}_{gs})^{d_s} & \text{if } 0 < d_s < 1, \end{cases} \tag{13}$$

for a positive random variable $T_{d_s \alpha_s}$ (Lijoi and Prünster, 2010). The asymptotic order of the number of occupied tables increases with discount parameter d_s.

Choice of dish $\boldsymbol{\theta}_j$ As discussed, all customers seated at a given table of section s_j are served the same dish, chosen from the cuisine s_j menu by the first customer to sit at that table. Let $\boldsymbol{\phi}_{g_j s_j k}$ denote the common dish eaten by customers at the kth table, $k = 1, \ldots, M^{(j-1)}_{g_j s_j}$. The dish that customer j eats represents the probe-specific random effect $\boldsymbol{\theta}_j$, and

$$\boldsymbol{\theta}_j \begin{cases} = \boldsymbol{\phi}_{g_j s_j v_j} & \text{if } v_j = 1, \ldots, M^{(j-1)}_{g_j s_j}, \\ \sim W_{s_j} & \text{if } v_j = (M^{(j-1)}_{g_j s_j} + 1). \end{cases} \tag{14}$$

In the latter case (line 2), the dish $\boldsymbol{\theta}_j$ randomly selected by customer j is registered as $\boldsymbol{\phi}_{g_j s_j M^{(j)}_{g_j s_j}}$, where $M^{(j)}_{g_j s_j} = M^{(j-1)}_{g_j s_j} + 1$, and is served to all future customers who sit at table $M^{(j)}_{g_j s_j}$. Assumptions (12) and (14) imply that although the restaurants serve the same menus, the overall relative popularity of each dish is restaurant-specific. Refer to the block entitled "Cuisine menus" in Figure 6 of Supplementary Material (Gu et al., 2023).

The aforementioned process continues for the remaining 2R2CF customers. Expressions (8) and (10) guarantee that a cuisine is more popular at its namesake restaurant and the cuisine selected by a customer influences the restaurant choice of the next customer, making the next customer likely to select the same cuisine. This accounts for the lengthy runs of differential or non-differential probes seen in methylation data. In addition to achieving dimension reduction, the proposed Sticky PYP models the serial dependencies of adjacent probes as a decreasing function of the inter-probe distances.

Latent Clusters and Their Differential States

Latent clusters, introduced earlier in expression (3), comprise probes with identical random effects and form the basis of the dimension reduction strategy. Returning to the 2R2CF metaphor, we identify a cluster as the set of customers who eat the same dish. However, in addition to the customers seated at a table, multiple tables in both restaurants may serve the same dish because of the shared cuisine menu. Therefore, irrespective of the restaurant, aggregating customers eating the same dishes, we obtain the probe-cluster allocation variables $c_1, \ldots, c_p$, and the number of latent clusters, q. The collection of customers eating the same cuisine 2 (differential state) dishes corresponds to a distinct differential cluster in $\mathcal{D}$, defined in equation (4).

From expression (13), we expect the number of occupied tables in the franchise to be much smaller than the number of customers, p. Furthermore, since multiple tables may serve the same dish, we expect the number of latent clusters, q, to be smaller than the number of occupied tables. With high probability, this implies that q is much smaller than p.

PYP Discount Parameter d_2 Consider the differential state cuisine menu, W_2, defined in (7). It can be shown that as the number of treatments, T, and the number of probes, p, increase, the differential clusters are not only asymptotically identifiable but consistently detectable in the posterior; refer to Section 4 of Guha and Baladandayuthapani (2016) for a detailed discussion of this remarkable phenomenon in standard PYP settings. Since the differential clusters can be inferred with high accuracy when T and p are large, discount parameter d_2 is given the mixture prior:

$$d_2 \sim \frac{1}{2}\delta_0 + \frac{1}{2}U(0,1) \tag{15}$$

where $d_2 = 0$ corresponds to a Dirichlet process. This provides 2R2CF the posterior flexibility to choose between a Dirichlet process and a more general PYP for a suitable clustering pattern of the differential probes. An allocation pattern typical of Dirichlet processes, such as exponentially decaying cluster sizes dominated by a few large clusters, results in a high posterior probability that d_2 equals 0. By contrast, an allocation pattern characteristic of non-Dirichlet PYPs, such as slower-than-exponential power law decays in the cluster sizes and relatively large numbers of smaller-sized clusters, causes the posterior of discount parameter d_2 to concentrate near 1 and exclude 0. A proof of the intrinsically different cluster patterns of Dirichlet processes and PYPs is given in Theorem 2.1 of Guha and Baladandayuthapani (2016).

Since distribution G is discrete, all atoms of T-variate distribution W_2 may not be unique. Indeed, this is common for $T = 2$ treatments. However, as T grows, and provided the number of probes, p, grows at a slower-than-exponential rate as T, the probability that two atoms allocated to the probes are identical rapidly decays to 0. In regression problems unrelated to differential analysis, Section 2.3 of Guha and Baladandayuthapani (2016) derived a similar result for a simpler zero-order stochastic process. We have verified this phenomenon in simulation studies on differential analysis datasets. In several hundred artificial datasets generated from the Sticky PYP, for $p = 1{,}500$ probes and T as small as four, no two allocated atoms of W_2 were identical.

PYP Discount Parameter d_1 Consider again the (non-differential) cuisine menu W_1 defined in (6). In general, the flexibility provided by PYP allocation patterns is not necessary for non-differential probes. This is because the allocation patterns of W_1 are driven by univariate parameter ψ in (6) and mixture allocations of univariate objects are unidentifiable (e.g., Frühwirth-Schnatter, 2006). Consequently, we set PYP discount parameter $d_1 = 0$, reducing the two PYPs associated with the non-differential state (i.e., section 1 in both restaurants) to Dirichlet processes.

Other Model Parameters

Depending on the specifics of the application, an appropriate model is assumed for the subject-specific parameters $\xi_1, \ldots, \xi_n$. For example, we may assume $\xi_i \overset{\text{i.i.d.}}{\sim} N(0, \tau_\epsilon^2)$. In other applications, it may be more appropriate to assume non-zero means and flexible error distributions: $\xi_i = b_i + \epsilon_i$, where b_i represents lane or batch effects in methylation data, and the i.i.d. ϵ_i follow a random distribution with a Dirichlet process prior. Similarly, appropriate models for the probe-specific parameters $\chi_1, \ldots, \chi_p$ may include i.i.d. zero-mean normal distributions, and finite mixtures or HMMs with state-specific normal distributions. Inverse-gamma priors are assigned to σ^2 and τ_ϵ^2. Suitable priors are assumed for mass parameters β, α_1, and α_2 in expressions (6) and (12). Mean μ_G and variance τ_G^2 of base distribution G_0 in expression (6) are given a joint normal-inverse gamma prior. The DAG in Figure 6 of Supplementary Material (Gu et al., 2023) summarizes the complex relationships between the different model parameters.

3 Posterior Inference

Due to the analytical intractability of the BayesDiff model, we rely on MCMC methods for posterior inference and detection of differential probes.

3.1 MCMC Strategy

The model parameters are initialized using naïve estimation techniques and iteratively updated by MCMC techniques until the chain converges. We split the MCMC updates into three blocks. An outline of the MCMC procedure is as follows. Further details can be found in Section 1 of Supplementary Material (Gu et al., 2023).

1. **Restaurant-cuisine-table-dish $(g_j, s_j, v_j, \boldsymbol{\theta}_j)$ of customer j:** For each probe $j = 1, \ldots, p$, we sample vector $(g_j, s_j, v_j, \boldsymbol{\theta}_j)$ given the vectors of the other $(p-1)$ probes. This is achieved by proposing a new value of $(g_j, s_j, v_j, \boldsymbol{\theta}_j)$ from a carefully constructed approximation to its full conditional, and by accepting or rejecting the move in a Metropolis-Hastings step. As discussed in Section 2.1, probe-cluster allocations $c_1, \ldots, c_p$ are immediately available from the restaurant-cuisine-table allocations (g_j, s_j, v_j) of the p probes, as are available the q latent clusters with their allocated probes and the set of differential clusters, $\mathcal{D}$.

2. **Latent vectors $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_q$:** The Tq latent vector elements are not necessarily distinct because of the Dirichlet process prior on distribution G. Although the latent vector elements are known from the aforementioned block 1 updates, MCMC mixing is considerably improved by updating the latent vector elements conditional on the p probe-cluster allocations. As the calculation in Supplementary Material (Gu et al., 2023) shows, this is achieved by Gibbs sampling.

3. **Remaining model parameters**: Generated by standard MCMC techniques.

We discarded an initial burn-in of 10,000 MCMC samples and used the subsequent 50,000 draws for posterior inferences. Convergence was informally assessed by trace plots of various hyperparameters to validate the MCMC sample sizes. For the proposed moves (in discrete parameter space) described in Step 1, the average Metropolis-Hastings acceptance rate exceeded 90% in all our analyses.

3.2 Detection of Differential Probes with FDR Control

Post-processing the MCMC sample, a Bayesian approach for controlling the false discovery rate (FDR) (Newton et al., 2004) is applied to detect the probes j with differential state $s_j = 2$. Specifically, let q_0 be the nominal FDR level and ω_j be the posterior probability that probe j is differential, so that $\omega_j = P[s_j = 2 \mid \mathbf{X}]$. An empirical average estimate, $\hat{\omega}_j$, is available from the MCMC sample. To achieve the desired FDR level in calling the differential probes, we first rank all the probes in decreasing order of $\hat{\omega}_j$. Let $\hat{\omega}_{(1)} > \hat{\omega}_{(2)} > \cdots > \hat{\omega}_{(p)}$ denote the ordered posterior probability estimates. For each $b = 1, \ldots, p$, we evaluate the posterior expected FDR of calling differential the first b probes in the sorted sequence:

$$\widehat{\mathrm{FDR}}_b = \frac{\sum_{j=1}^{p} (1 - \hat{\omega}_j)\, \mathcal{I}\left(\hat{\omega}_j \geq \hat{\omega}_{(b)}\right)}{\sum_{j=1}^{p} \mathcal{I}\left(\hat{\omega}_j \geq \hat{\omega}_{(b)}\right)} = \frac{\sum_{j=1}^{b} \left(1 - \hat{\omega}_{(j)}\right)}{b}, \tag{16}$$

where the simplification occurs because the $\hat{\omega}_{(j)}$ are sorted. Finally, we pick the largest value of b, denoted by b^*, for which $\widehat{\mathrm{FDR}}_b < q_0$. A nominal FDR level of q_0 is achieved by labeling the first b^* probes, arranged in decreasing order of $\hat{\omega}_j$, as differential.

4 Simulation Studies

Using artificial datasets with $T = 5$ treatments, we analyzed the accuracy of BayesDiff in detecting differentially methylated probes. We compared the results with established differential methylation procedures and general statistical techniques for multiple treatment comparisons. We also evaluated the ability of the BayesDiff procedure in discovering the complex dependence structures of DNA methylation data.

Generation Strategy Proportions representing DNA methylation data were generated using the logit transformation as in equation (1). The inter-probe distances were the actual distances from the motivating TCGA dataset, scaled to add to 1. In order to capture the complexity of methylation data, such as the existence of multiple latent methylation states (e.g,. CpG islands and shores), different read depths across CpGs, and the incomplete conversion of bisulphite sequencing, the generation strategy was partly based on techniques implemented in WGBSSuite, a flexible stochastic simulation tool for generating single-base resolution methylation data (Rackham et al., 2015). However, the generation procedure differed from WGBSSuite in some respects. Specifically, it allowed more than two treatments ($T = 5$). Additionally, as in actual methylation datasets, the generation procedure incorporated serial dependence not only in the methylation levels but also in the differential states of the probes.

The probe-specific read depths were generated as $n_j \overset{\text{i.i.d.}}{\sim} \text{Poisson}(50)$. Unlike assumption (1), there were no subject-specific random effects in the generation mechanism. Instead, the normal means incorporated additive probe-specific random effects, $\chi_1^{(0)}, \ldots, \chi_p^{(0)}$, that were generated using the following steps:

1. Generate the true methylation status of the probes, denoted by $h_1^{(0)}, \ldots, h_p^{(0)}$, using the 4-state *distance-based* HMM of Rackham et al. (2015), with the states respectively representing the methylated, first transit, demethylated, and second transit states.

2. Set the baseline methylation levels for the methylated, (first or second) transit, and demethylated states as $p_{\text{methylated}} = 0.8$, $p_{\text{transit}} = 0.5$, and $p_{\text{un-methylated}} = 0.2$.

3. For $j = 1, \ldots, p$, compute the probe-specific means:

$$\tilde{\chi}_j^{(0)} = \begin{cases} \log\left(\frac{p_{\text{methylated}}}{(1-p_{\text{methylated}}}\right) & \text{if } h_j = 1 \text{ (i.e., methylated state)}, \\ \log\left(\frac{p_{\text{transit}}}{1-p_{\text{transit}}}\right) & \text{if } h_j = 2, 4 \text{ (first or second transit state)}, \\ \log\left(\frac{p_{\text{demethylated}}}{1-p_{\text{demethylated}}}\right) & \text{if } h_j = 3 \text{ (i.e., demethylated state)}. \end{cases}$$

4. Generate $\chi_j^{(0)} \overset{\text{indep}}{\sim} N\left(\tilde{\chi}_j^{(0)}, \tau_\chi^2\right)$ for $j = 1, \ldots, p$.

Noise and Dependence Levels We investigated four scenarios corresponding to the combinations of two noise levels and two dependence levels. For each scenario, 20 datasets were independently generated, with each dataset consisting of $p = 500$ probes

α_1	α_2	d_2	β	γ	ρ_2	μ_G	τ_G^2	τ_χ^2
20	20	0.33	20	0.9	0.1	0	1	0.1225

Table 1: True parameter values used to generate the artificial datasets.

and four samples associated with each of $T = 5$ treatments, i.e., a total to $n = 20$ samples. The low noise setting corresponded to true variance parameter $\sigma_0^2 = 0.36$; equivalently, to a signal-to-noise of $R_0^2 \approx 70\%$. The high noise setting corresponded to $\sigma_0^2 = 1$ or $R_0^2 \approx 40\%$. The true between-probe dependencies comprised two levels: no serial correlation (i.e., a zero-order Sticky PYP) with $\eta_0 = 0$, and positive serial correlation (i.e., a first order Sticky PYP) with $\eta_0 = 0.004$. Although $\eta_0 = 0.004$ may appear to be small, its value is calibrated to the scaled inter-probe distances and represents fairly high inter-probe dependence. For example, when the distance between two adjacent probes is equal to the standardized average distance of $\bar{e} = 1/(p-1) = 1/499$, $\eta_0 = 0.004$ gives an affiliation of $r_0 = 0.6$ in equation (9). For convenience, we will refer to the two dependence levels as "no-correlation" and "high correlation." True values of the other model parameters were common to the four scenarios and are displayed in Table 1. Setting a true baseline differential proportion of $\rho_2 = 0.1$ resulted in approximately 10% true differentially methylated CpGs in each dataset.

Posterior Inferences Assuming all model parameters to be unknown, each artificial dataset was analyzed using a BayesDiff model that differed in key respects from the true generation mechanism. For example, unlike the 4-state HMM of the generation strategy, the probe-specific random effects χ_j were analyzed using a BayesDiff model that ignored the first order dependence, and instead, relied on a 3-state finite mixture model representing the methylated, transit, and unmethylated states. Additionally, in contrast to the zeroed-out subject-specific random effects during data generation, BayesDiff assumed i.i.d. normal random effects with zero means.

To assess BayesDiff's accuracy in detecting the absence or presence of inter-probe serial correlation, in the no-correlation ($\eta_0 = 0$) situation, we evaluated $\log\left(\frac{P[\eta=0|\mathbf{X}]}{P[\eta>0|\mathbf{X}]}\right)$, the log-Bayes factor comparing zero order to first order Sticky PYPs. In the high correlation ($\eta_0 = 004$) situation, we evaluated $\log\left(\frac{P[\eta>0|\mathbf{X}]}{P[\eta=0|\mathbf{X}]}\right)$, the log-Bayes factor comparing first order to zero order Sticky PYPs. Thus, in any scenario, a large positive value of this measure provides strong evidence that BayesDiff detects the correct model order.

Although conceptually straightforward, the estimation of Bayes factors requires multiple MCMC runs even for relatively simple parametric models (Chib, 1995). Basu and Chib (2003) extended the estimation strategy to infinite dimensional models such as Dirichlet processes. However, the computational costs are prohibitively high for big datasets, and multiple MCMC runs stretch present-day computational resources beyond their limits. Faced with these challenges, we relied on an alternative strategy for estimating the *lower bounds* of log-Bayes factors as by-products of the Section 3.1 algorithm. As it turns out, this is often sufficient to infer Sticky PYP model orders. Let Θ^- denote all BayesDiff model parameters except η. In the high correlation situation, applying Jensen's inequality, a lower bound for the corresponding log-Bayes factor is

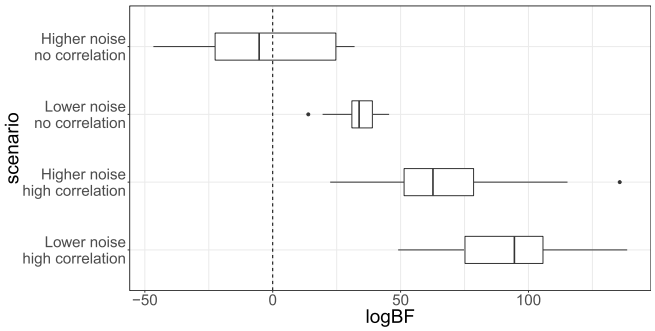


Figure 3: Simulation study box plots for estimated lower bounds of the log-Bayes factors in favor of the true model order.

$E\left[\log\left(\frac{P[\eta>0|\mathbf{X},\Theta^-]}{P[\eta=0|\mathbf{X},\Theta^-]}\right) \mid \mathbf{X}\right]$. Unlike log-Bayes factors, this lower bound can be easily estimated using a single MCMC run. In the no-correlation situation, a lower bound for the log-Bayes factor, $\log\left(\frac{P[\eta=0|\mathbf{X}]}{P[\eta>0|\mathbf{X}]}\right)$, is similarly estimated.

In the four simulation scenarios, box plots of these estimated lower bounds for the 20 datasets are presented in Figure 3. Except for the high noise–no-correlation scenario, for which the results were inconclusive, the estimated lower bounds of the log-Bayes factors in favor of the true correlation structure were all positive and large. In the low noise–no-correlation scenario, BayesDiff decisively favored zero-order models, and the smallest lower bound among the 20 datasets was 13.9, corresponding to Bayes factors exceeding $e^{13.9} = 1{,}088{,}161$. The 25^{th} percentile of these lower bounds was 30.9, corresponding to Bayes factors exceeding $e^{30.9} = 2.63 \times 10^{13}$. This is strong evidence that the BayesDiff approach is reliable in this scenario. For the high-correlation scenarios, the estimated lower bounds were even higher, indicating that BayesDiff overwhelmingly favors first order models when the data are serially correlated.

Comparisons with Other Methods We evaluated the success of the BayesDiff procedure in detecting disease genomic signatures relative to six well-known statistical methods. These included generic multiple comparison techniques, namely, one-way analysis of variance (ANOVA) and the Kruskal-Wallis test. Also included were specially developed methods for detecting differential methylation in more than two treatments: COHCAP (Warden et al., 2013), methylKit (Akalin et al., 2012), BiSeq (Hebestreit et al., 2013), and RADMeth (Dolzhenko and Smith, 2014). The ANOVA and Kruskal-Wallis test procedures were performed separately on each probe after applying the inverse-logit transform to the data. The COHCAP method was directly applied to the synthetic data. The remaining three methods are designed for bisulfite sequencing, which consists of total methylation reads for each CpG site. For these methods, the methylation reads were calculated by multiplying the methylation proportions by the total reads. As recommended, the bandwidth smoothing parameter of the BiSeq method was tuned to optimize overall detection. For all six competing methods, probe-specific p-values were

evaluated and adjusted for multiplicity using the FDR control procedure of Benjamini and Hochberg (1995).

Like most nonparametric Bayes techniques, the computational times of BayesDiff are considerably higher than frequentist methods, but negligible compared to the time frames over which the experimental data are collected. Furthermore, as we demonstrate, the substantially greater accuracy of BayesDiff more than compensates for its computational costs. On a personal computer with an Intel Core i7-4770 processor with 3.40 GHz frequency and 8 GB RAM, the average run time for the Section 3.1 MCMC algorithm, applied to the synthetic datasets with $n = 20$ samples, $T = 5$ treatments, and $p = 500$ probes, was 0.60 seconds per iteration. However, the computational times are greatly reduced by running the datasets in parallel across multiple cores of a research computing cluster. Analyzing datasets of various sizes, we found that the computational cost is $O(Tp^2)$ but does not appreciably depend on n/T. This is reasonable because the mixture model primarily focuses on $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p \in \mathcal{R}^T$. Due to the intensive nature of the one-parameter-at-a-time Gibbs sampling updates in Block 2, the Metropolis-Hastings algorithm of Guha (2010) can be applied to significantly speed up the updates and make the calculations more scalable. As part of ongoing work developing a fast R package, we find that ten- to hundred-fold speedups are possible with this fast MCMC strategy, which can also accelerate the block 1 parameter updates of Section 3.1.

We computed the receiver operating characteristic (ROC) curves for differential probe detection for all seven methods. For a quantitative assessment, we calculated the area under curve (AUC), declaring the method with the largest AUC as the most reliable in each scenario. The ROC curves, averaged over the 20 datasets under each simulation scenario, are shown with the AUCs in Figure 2 of Supplementary Material (Gu et al., 2023). In all except the high-noise–no-correlation scenario, BayesDiff uniformly outperformed the other methods. Even in the high-noise–no-correlation scenario, BayesDiff performed better in the low FPR region. As expected, all seven methods had lower accuracies for higher noise levels. BayesDiff did significantly better than the competing methods in the high correlation scenarios, suggesting that the incorporation of between-probe dependencies improves its accuracy in situations typical of DNA methylation data.

Since researchers typically focus on small false positive rates (FPRs), that is, small significance levels, we also calculated the measures, AUC_{20} and AUC_{10}. AUC_{20} (AUC_{10}) is defined as the area under the ROC curve multiplied by 5 (10) when the FPR does not exceed 0.2 (0.1). The multiplicative factors ensure that the areas potentially vary between 0 and 1. The three versions of AUC are presented in Table 2 in Supplementary Material (Gu et al., 2023). As also seen in Figure 2, Table 2 reveals that in three of the four scenarios, BayesDiff had the largest overall AUC. Furthermore, BayesDiff had vastly improved reliability for low FPRs. For example, consider the low noise–high correlation scenario. The overall AUC for BayesDiff was 0.035 greater than that for ANOVA. In contrast, the gains for BayesDiff, relative to ANOVA, were +0.107 for AUC_{20} and +0.146 for AUC_{10}. The advantages of BayesDiff were even greater relative to the other competing methods. In the high noise–low-correlation scenario, BayesDiff had a relatively low AUC, as mentioned. However, even in this scenario, it had the

greatest AUC_{20} and AUC_{10} among all the methods. Additionally, for a nominal FDR of $q_0 = 0.05$, the achieved FDR of BayesDiff was between 0 and 0.03 in every dataset and simulation scenario. These results demonstrate the ability of BayesDiff to accurately detect differential probes, even in challenging situations in which the FPR is small.

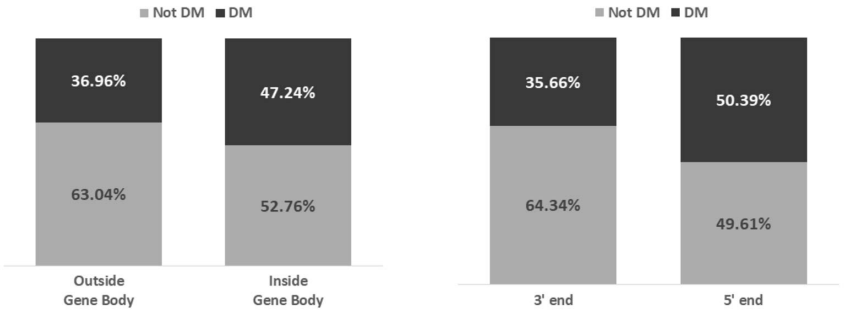
5 Data Analysis

We returned to the motivating DNA methylation data consisting of the upper GI cancers: stomach adenocarcinoma (STAD), liver hepatocellular carcinoma (LIHC), esophageal carcinoma (ESCA), and pancreatic adenocarcinoma (PAAD). Applying the BayesDiff procedure, we detected the differentially methylated CpG loci among the cancer types.

Data Processing The dataset was obtained from The Cancer Genome Atlas project, publicly available through The Genomic Data Commons (GDC) portal (Grossman et al., 2016). The measurements on 485,577 probes located at CpG sites were made using the Illumina Human Methylation 450 platform. At the time of analysis, the dataset consisted of 1,224 tumor samples. We analyzed the data on a gene-by-gene basis, selecting 443 genes with mutations in at least 5% of the samples. To ensure that all CpG sites potentially linked to a gene were included in the analysis, we selected sites located within 50K base pairs outside the gene body, upstream from the 5' end as well as downstream from the 3' end. The number of gene-specific CpG sites ranged from 1 to 769, and are displayed in Figure 3(a) of Supplementary Material (Gu et al., 2023). As a final preprocessing step, since the methylation patterns of short genes are less informative in cancer investigations, we eliminated the 25 genes mapped to 20 or fewer CpG sites.

Inference Procedure The data were analyzed using the proposed BayesDiff approach. Exploratory analyses indicated that eliminating the probe-specific random effects χ_j in expression (1) produces a satisfactory model fit. Since experimental batch information is not available in the TCGA dataset, we assumed that the parameters $\xi_1, \ldots, \xi_n$ in (1) are i.i.d. with a random distribution having a Dirichlet process prior. The MCMC procedure of Section 3.1 was applied to obtain posterior samples for each gene. For detecting differential CpG sites, we applied the Section 3.2 procedure with a nominal FDR of $q_0 = 0.05$.

Results Among the differentially methylated CpG sites detected by our approach, approximately 40.6% of the sites were located outside the gene body. Figure 4 displays the associations between detected methylation status and position of the CpG sites. We defined "near the 5' (3') end" as CpG sites located within one-fourth length of the gene body, either inside or outside the gene boundary, and closer to the transcription start (termination) site. Our results indicate that the proportion of differential methylation is higher for CpG sites inside the gene body and most differentially methylated loci are situated within the gene body, as is well known from numerous previous studies.



(a) Contingency table of detected methylation status and location of CpG site with respect to gene body

(b) Contingency table of detected methylation status and proximity of CpG site to chromosomal end

Figure 4: Associations of detected methylation status and position of CpG sites.

However, our analysis also revealed significant amounts of differential methylation outside the gene body. Despite the common belief that DNA methylation analysis should focus on the 5' end region, we found that CpG sites near the 3' ends also displayed considerable differential methylation. These findings support the recommendations of Irizarry et al. (2009) that studies of DNA methylation alteration should be conducted on a higher resolution, epigenome-wide basis.

Among the differentially methylated sites detected by BayesDiff, we estimated the pairwise differences between random effects associated with the four cancer types. Site-wise summaries of the largest pairwise differences of the cancer-specific effects are displayed in Figure 5. None of the four cancer types displayed consistent hypermethylation or hypomethylation across all genes or over entire chromosomes. However, we found that LIHC is frequently differentially methylated relative to one of the other cancer types, implying that it is the most volatile disease with respect to DNA methylation.

For each gene, Figure 3(b) of Supplementary Material (Gu et al., 2023) displays 95% credible intervals for the lower bounds of log-Bayes factors of a first versus zero-order 2R2CF model, i.e., $\eta = 0$ versus $\eta > 0$ in expression (9). Models with first order dependence are overwhelmingly favored for a majority of the genes, suggesting that statistical techniques that fail to account for dependence between neighboring CpG sites are less effective for these data. Figure 4 of Supplementary Material (Gu et al., 2023) displays the detailed differential methylation pattern for the top two mutated genes, TP53 and TTN. An obvious feature of both genes is that the differential methylation patterns are strongly serially correlated. For gene TP53, there are almost no differentially methylated loci within the gene body. The 3' end region outside the gene body has a cluster of differentially methylated loci, for which cancer type STAD is mostly hypermethylated. The results for gene TTN tell a quite different story: most of the differentially methylated loci are inside the gene body and near the 5' end. Cancer type LIHC is hypomethylated compared to PAAD around the 5' end region, but it is hypermethylated compared to STAD near the 3' end. Genes with at least 90% differentially methylated

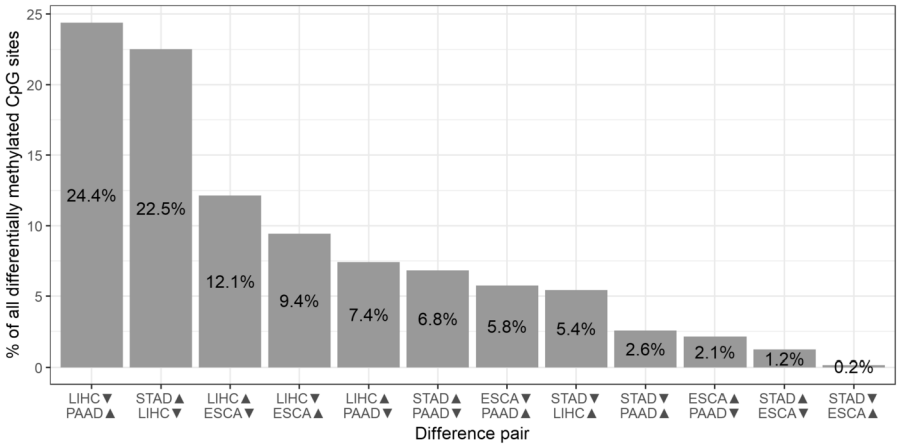


Figure 5: Site-wise summary of the largest pairwise differences of differentially methylated loci among the four upper GI cancer types.

sites detected by BayesDiff are listed in Table 3 of Supplementary Material (Gu et al., 2023), along with the largest pairwise difference between the four cancer types among the differentially methylated loci. The number of CpG sites within each segment is listed in Table 4 of Supplementary Material (Gu et al., 2023).

Existing medical literature both supports and complements our findings. For example, hypermethylation of the EDNRB and SLIT2 genes have been found in STAD (Tao et al., 2012). Gene FBN2 was hypermethylated in ESCA (Tsunoda et al., 2009). While several studies have found that the gene and protein expressions of ABC transporter genes, such as ABCC9, are useful for understanding the prognosis of esophageal cancer (Vrana et al., 2018), we find that hypermethylation of ABCC9 is a major difference between cancer types ESCA and LIHC. Gene FLRT2 is a potential tumor suppressor that is hypermethylated and downregulated in breast cancer (Bae et al., 2017). Our results indicate that this gene is also hypermethylated in cancer type STAD versus LIHC. Mutations in SPTA1 gene has been linked with PAAD (Murphy et al., 2013); our results indicate that hypermethylation of this gene distinguishes PAAD from LIHC.

Finally, we compared our findings with those of ANOVA for multiple treatment comparisons. Table 5 of Supplementary Material (Gu et al., 2023) lists the common set of genes with at least 90% differentially methylated sites identified by both BayesDiff *and* ANOVA. Table 6 displays the genes identified by *only* ANOVA, whereas Table 7 displays the large number of genes detected by *only* BayesDiff. Cross-referencing with the medical literature, we find that genes FLRT2 and FBN2 were detected by both methods. However, genes EDNRB, SLIT2, ABCC9, and SPTA1 were only identified by BayesDiff, revealing the benefits of the proposed Bayesian nonparametric method.

Accounting for Data Characteristics To avoid making misleading biological interpretations, a statistical model must account for the observed biomarker means and

variances, especially in multiple-testing approaches where the first two sample moments are important (Subramaniam and Hsiao, 2012). From this perspective, certain aspects of the BayesDiff model, such as variance σ^2 being a priori unrelated to the mean in expression (1), may appear to be unduly restrictive. However, even though it was not specifically designed to match data summaries such as sample moments, in practice, the nonparametric nature of the Sticky PYP allows the posterior to flexibly adapt to unique data characteristics, such as sample moments, and account for mean-variance relationships in a robust manner. For example, consider again the top mutated genes, TP53 and TTN, discussed in Figure 4 of Supplementary Material (Gu et al., 2023). The ability of BayesDiff to match the sample moments of the gene-specific probes can be demonstrated as follows. Given the inter-probe distances, the joint posterior of the BayesDiff parameters induces predictive distributions on the n measurements for each probe. Functionals of these predictive distributions, such as probe-specific sample moments, are easily estimated by post-processing the MCMC sample. For these two genes, Figure 5 of Supplementary Material (Gu et al., 2023) reveals that the sample moments predicted by BayesDiff are a close match to the actual first and second sample moments with correlations exceeding 99% in each plot. Similar results were observed in other datasets.

6 Discussion

DNA methylation data exhibit complex structures due to biological mechanisms and distance-dependent correlations between adjacent CpG sites or probes. The identification of the differential signatures of multiple sets of tumor samples is crucial for developing targeted treatments for disease. This paper formulates a flexible approach applicable to multiple treatments called BayesDiff. The technique relies on a novel Bayesian mixture model called the Sticky PYP or the two-restaurant two-cuisine franchise. In addition to facilitating simultaneous inferences on the probes, the model accommodates distance-based serial dependence and accounts for the complex interaction patterns commonly observed in cancer data. An effective MCMC strategy for detecting the differential probes is developed. The success of the BayesDiff procedure in differential DNA methylation, relative to well-established methodologies, is exhibited via simulation studies. The new technique is applied to the motivating TCGA dataset to detect the differential genomic signatures of four upper GI cancers. The results both support and complement known facts about epigenomic differences between these cancer types, while identifying genes with high proportions of differentially methylated CpG sites.

In addition to providing a good fit to the data, a statistical model must be able to account for features such as sample moments. The success of the BayesDiff model in this regard is demonstrated in Section 5 using the upper GI dataset. It must be emphasized, however, that BayesDiff may be less successful in accounting for the characteristics of some other datasets, possibly due to slow asymptotic convergence of the posterior to the underlying generative process. In such situations, more flexible global transformations (Li et al., 2016) or variance-stabilizing transformations (Durbin et al., 2002) may be utilized. Alternatively, local Laplace approximations of exponential family likelihoods

through link functions (Zeger and Karim, 1991; Chib and Winkelmann, 2001) may extend the BayesDiff model to better explain the data characteristics.

Like most Bayesian models comprising several latent parameters, the proposed 2R2CF may be marginalized over different parameter sets to obtain equivalent versions of the same model. For example, we could marginalize over *restaurants* to obtain an equivalent "sticky cuisine" version in which there is just one restaurant with two cuisine-sections and a customer more likely to favor the *cuisine* selected by the previous customer. Alternatively, we could marginalize over *sections* to obtain an equivalent "sticky restaurant franchise" in which each restaurant comprises a single section with restaurant-specific probabilities ensuring that Cuisine 1 or 2 dishes are more popular at their namesake restaurant; a customer is then more likely to favor the *restaurant* selected by the previous customer. In all equivalent versions, however, a probe's differential state is determined by the customer's dish in the metaphor.

The 2R2CF perspective offers the twin advantages of parameter interpretability and generalizability. Section 3 of Supplementary Material (Gu et al., 2023) presents the generalized form of the Sticky PYP, revealing the full potential of the proposed method in analyzing not only DNA methylation datasets, but other types of omics datasets, such as gene expression, RNASeq, and copy-number alteration data. Beyond biomedical applications, the generalized formulation offers a diverse palette of parametric and nonparametric models for capturing the distinctive features of datasets. These Bayesian mixture models are special cases of Sticky PYPs for particular choices of a countable group parameter (e.g., two "restaurants" in the 2R2CF metaphor for differential methylation problems) and countable state parameter (e.g., two "cuisines" in 2R2CF) with the state of a customer influencing the group of the next customer. In addition to extending PYPs to discrete time series-type data, the range of models includes Dirichlet processes, PYPs, infinite HMMs, hierarchical Dirichlet processes (Teh et al., 2006), hierarchical Pitman-Yor processes (Teh et al., 2006; Camerlenghi et al., 2019), finite HMMs, nested Chinese restaurant processes (Blei and Jordan, 2005), nested Dirichlet processes (Rodriguez et al., 2008), and analysis of densities models (Tomlinson and Escobar, 2003).

Ongoing work involves extending the correlation structure to model more sophisticated forms of inter-probe dependence in DNA methylation data. Commented R code implementing the BayesDiff method is available on GitHub at `https://github.com/cgz59/BayesDiff`. Using high-performance Rcpp subroutines, we are developing a fast R package for detecting differential genomic signatures in a wide variety of omics datasets. Initial results indicate that order-of-magnitude speedups will allow the fast analyses of high-dimensional datasets on personal computers.

Supplementary Material

MCMC algorithm and Generalized Sticky PYP (DOI: 10.1214/23-BA1407SUPP; .pdf). Detailed description of MCMC procedure; generalized form of the Sticky PYP; additional graphs and tables.

References

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., and Mason, C. E. (2012). "methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles." *Genome Biology*, 13(10): R87. 3, 19

Bae, H., Kim, B., Lee, H., Lee, S., Kang, H.-S., and Kim, S. J. (2017). "Epigenetically regulated fibronectin leucine rich transmembrane protein 2 (FLRT2) shows tumor suppressor activity in breast cancer cells." *Scientific Reports*, 7(1): 272. 23

Baker, T. A., Bell, S. P., Gann, A., Levine, M., Losick, R., and Inglis, C. (2008). *Molecular Biology of the Gene*. San Francisco, CA, USA: Pearson/Benjamin Cummings. 4

Basu, S. and Chib, S. (2003). "Marginal likelihood and Bayes factors for Dirichlet process mixture models." *Journal of the American Statistical Association*, 98(461): 224–235. MR1965688. doi: https://doi.org/10.1198/01621450338861947. 18

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)*, 289–300. MR1325392. 20

Blei, D. M. and Jordan, M. I. (2005). "Variational inference for Dirichlet process mixtures." *Bayesian Analysis*, 1: 1–23. MR2227367. doi: https://doi.org/10.1214/06-BA104. 25

Camerlenghi, F., Lijoi, A., Orbanz, P., and Prünster, I. (2019). "Distribution theory for hierarchical processes." *The Annals of Statistics*, 47(1): 67–92. MR3909927. doi: https://doi.org/10.1214/17-AOS1678. 25

Chib, S. (1995). "Marginal likelihood from the Gibbs output." *Journal of the American Statistical Association*, 90(432): 1313–1321. MR1379473. 18

Chib, S. and Winkelmann, R. (2001). "Markov chain Monte Carlo analysis of correlated count data." *Journal of Business & Economic Statistics*, 19(4): 428–435. MR1947077. doi: https://doi.org/10.1198/07350010152596673. 25

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). "Are Gibbs-type priors the most natural generalization of the Dirichlet process?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2): 212–229. 6

Dolzhenko, E. and Smith, A. D. (2014). "Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments." *BMC Bioinformatics*, 15(1): 215. MR3167142. 3, 19

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., and Lin, S. M. (2010). "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis." *BMC Bioinformatics*, 11. 5

Dunson, D. B., Herring, A. H., and Engel, S. M. (2008). "Bayesian selection and clustering of polymorphisms in functionally-related genes." *Journal of the American*

Statistical Association, 103: 534–546. MR2523991. doi: https://doi.org/10.1198/016214507000000554. 6

Dunson, D. B. and Park, J.-H. (2008). "Kernel stick-breaking processes." *Biometrika*, 95: 307–323. MR2521586. doi: https://doi.org/10.1093/biomet/asn012. 6

Durbin, B., Hardin, J., Hawkins, D., and Rocke, D. (2002). "A variance-stabilizing transformation for gene-expression microarray data." *Bioinformatics*, 18: S105–S110. MR2704889. 24

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., et al. (2006). "DNA methylation profiling of human chromosomes 6, 20 and 22." *Nature Genetics*, 38(12): 1378. 2

Feinberg, A. P. and Tycko, B. (2004). "The history of cancer epigenetics." *Nature Reviews Cancer*, 4(2): 143. 2

Feng, H., Conneely, K. N., and Wu, H. (2014). "A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data." *Nucleic Acids Research*, 42(8): e69–e69. 3

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 1: 209–230. MR0350949. 6, 13

Fox, E., Sudderth, E., Jordan, M., and Willsky, A. (2011). "The sticky HDP-HMM: Bayesian nonparametric hidden Markov models with persistent states." *Annals of Applied Statistics*, 5: 1020–1056. MR2840185. doi: https://doi.org/10.1214/10-AOAS395. 6

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer. MR2265601. 15

Gnedin, A. and Pitman, J. (2006). "Exchangeable Gibbs partitions and Stirling triangles." *Journal of Mathematical Sciences*, 138: 5674–5685. MR2160320. doi: https://doi.org/10.1007/s10958-006-0335-z. 6

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., and Staudt, L. M. (2016). "Toward a shared vision for cancer genomic data." *New England Journal of Medicine*, 375(12): 1109–1112. 21

Gu, C., Baladandayuthapani, V., and Guha, S. (2023). "Supplementary Material for "Nonparametric Bayes differential analysis of multigroup DNA methylation data"." *Bayesian Analysis*. doi: https://doi.org/10.1214/23-BA1407SUPP. 2, 5, 10, 12, 13, 15, 16, 20, 21, 22, 23, 24, 25

Guha, S. (2010). "Posterior simulation in countable mixture models for large datasets." *Journal of the American Statistical Association*, 105(490): 775–786. MR2724860. doi: https://doi.org/10.1198/jasa.2010.tm09340. 20

Guha, S. and Baladandayuthapani, V. (2016). "A nonparametric Bayesian technique for high-dimensional regression." *Electronic Journal of Statistics*, 10: 3374–3424. MR3572854. doi: https://doi.org/10.1214/16-EJS1184. 6, 14, 15

Hansen, K. D., Langmead, B., and Irizarry, R. A. (2012). "BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions." *Genome Biology*, 13(10): R83. 3, 7

Hebestreit, K., Dugas, M., and Klein, H.-U. (2013). "Detection of significantly differentially methylated regions in targeted bisulfite sequencing data." *Bioinformatics*, 29(13): 1647–1653. 3, 7, 19

Irizarry, R. A., Ladd-Acosta, C., Carvalho, B., Wu, H., Brandenburg, S. A., Jeddeloh, J. A., Wen, B., and Feinberg, A. P. (2008). "Comprehensive high-throughput arrays for relative methylation (CHARM)." *Genome Research*, 18(5): 780–790. 2

Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., et al. (2009). "Genome-wide methylation analysis of human colon cancer reveals similar hypo-and hypermethylation at conserved tissue-specific CpG island shores." *Nature Genetics*, 41(2): 178. 2, 22

Ishwaran, H. and James, L. F. (2003). "Generalized weighted Chinese restaurant processes for species sampling mixture models." *Statistica Sinica*, 13: 1211–1235. MR2026070. 8

Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P., and Irizarry, R. A. (2012). "Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies." *International Journal of Epidemiology*, 41(1): 200–209. 3, 7

Kim, S., Tadesse, M. G., and Vannucci, M. (2006). "Variable selection in clustering via Dirichlet process mixture models." *Biometrika*, 93: 877–893. MR2285077. doi: https://doi.org/10.1093/biomet/93.4.877. 6

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010). "Tackling the widespread and critical impact of batch effects in high-throughput data." *Nature Reviews Genetics*, 11(10). 2

Li, D., Wang, X., Lin, L., and Dey, D. K. (2016). "Flexible link functions in nonparametric binary regression with Gaussian process priors." *Biometrics*, 72(3): 707–719. MR3545664. doi: https://doi.org/10.1111/biom.12462. 24

Lijoi, A., Mena, R., and Prünster, I. (2007a). "Bayesian nonparametric estimation of the probability of discovering new species." *Biometrika*, 94: 769–786. MR2416792. doi: https://doi.org/10.1093/biomet/asm061. 6

Lijoi, A., Mena, R., and Prünster, I. (2007b). "Controlling the reinforcement in Bayesian nonparametric mixture models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69: 715–740. 7

Lijoi, A. and Prünster, I. (2010). *Models Beyond the Dirichlet Process*, 80–136. Cambridge Series in Statistical and Probabilistic Mathematics. MR2730661. 8, 13

Medvedovic, M., Yeung, K. Y., and Bumgarner, R. E. (2004). "Bayesian mixture model based clustering of replicated microarray data." *Bioinformatics*, 20: 1222–1232. 6

Müller, P. and Mitra, R. (2013). "Bayesian nonparametric inference – why and how." *Bayesian Analysis (Online)*, 8(2). MR3066939. doi: https://doi.org/10.1214/13-BA811. 6

Murphy, S. J., Hart, S. N., Lima, J. F., Kipp, B. R., Klebig, M., Winters, J. L., Szabo, C., Zhang, L., Eckloff, B. W., Petersen, G. M., et al. (2013). "Genetic alterations associated with progression from pancreatic intraepithelial neoplasia to invasive pancreatic tumor." *Gastroenterology*, 145(5): 1098–1109. 23

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). "Detecting differential gene expression with a semiparametric hierarchical mixture method." *Biostatistics*, 5(2): 155–176. 16

Park, Y., Figueroa, M. E., Rozek, L. S., and Sartor, M. A. (2014). "MethylSig: a whole genome DNA methylation analysis pipeline." *Bioinformatics*, 30(17): 2414–2422. 3

Perman, M., Pitman, J., and Yor, M. (1992). "Size-biased sampling of Poisson point processes and excursions." *Probability Theory and Related Fields*, 92(1): 21–39. MR1156448. doi: https://doi.org/10.1007/BF01205234. 4, 6

Rackham, O. J., Dellaportas, P., Petretto, E., and Bottolo, L. (2015). "WGBSSuite: simulating whole-genome bisulphite sequencing data and benchmarking differential DNA methylation analysis tools." *Bioinformatics*, 31(14): 2371–2373. 17

Rodriguez, A., B., D. D., and Gelfand, A. E. (2008). "The nested Dirichlet process (with discussion)." *Journal of the American Statistical Association*, 103: 1131–1144. MR2528831. doi: https://doi.org/10.1198/016214508000000553. 25

Saito, Y., Tsuji, J., and Mituyama, T. (2014). "Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions." *Nucleic Acids Research*, gkt1373. 3

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 639–650. MR1309433. 8

Siegel, R. L., Miller, K. D., and Jemal, A. (2017). "Cancer statistics, 2017." *CA: A Cancer Journal for Clinicians*, 67(1): 7–30. 2

Song, Q., Decato, B., Hong, E. E., Zhou, M., Fang, F., Qu, J., Garvin, T., Kessler, M., Zhou, J., and Smith, A. D. (2013). "A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics." *PloS One*, 8(12): e81148. 3

Subramaniam, S. and Hsiao, G. (2012). "Gene-expression measurement: variance-modeling considerations for robust data analysis." *Nature Immunology*, 13(3): 199–203. 24

Sun, D., Xi, Y., Rodriguez, B., Park, H. J., Tong, P., Meong, M., Goodell, M. A., and Li, W. (2014). "MOABS: model based analysis of bisulfite sequencing data." *Genome Biology*, 15(2): R38. 3

Tao, K., Wu, C., Wu, K., Li, W., Han, G., Shuai, X., and Wang, G. (2012). "Quantitative

analysis of promoter methylation of the EDNRB gene in gastric cancer." *Medical Oncology*, 29(1): 107–112. 23

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). "Hierarchical Dirichlet processes." *Journal of the American Statistical Association*, 101: 1566–1581. MR2279480. doi: https://doi.org/10.1198/016214506000000302. 6, 25

Tomlinson, G. and Escobar, M. (2003). "Analysis of densities." *Talk given at the Joint Statistical Meeting*, 103: 1131–1144. 25

Tsunoda, S., Smith, E., De Young, N. J., Wang, X., Tian, Z.-Q., Liu, J.-F., Jamieson, G. G., and Drew, P. A. (2009). "Methylation of CLDN6, FBN2, RBP1, RBP4, TFPI2, and TMEFF2 in esophageal squamous cell carcinoma." *Oncology Reports*, 21(4): 1067–1073. 23

Vedeld, H. M., Goel, A., and Lind, G. E. (2017). "Epigenetic biomarkers in gastrointestinal cancers: The current state and clinical perspectives." In *Seminars in Cancer Biology*. Elsevier. 2

Vrana, D., Hlavac, V., Brynychova, V., Vaclavikova, R., Neoral, C., Vrba, J., Aujesky, R., Matzenauer, M., Melichar, B., and Soucek, P. (2018). "ABC transporters and their role in the neoadjuvant treatment of esophageal cancer." *International Journal of Molecular Sciences*, 19(3): 868. 23

Wang, D., Yan, L., Hu, Q., Sucheston, L. E., Higgins, M. J., Ambrosone, C. B., Johnson, C. S., Smiraglia, D. J., and Liu, S. (2012). "IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data." *Bioinformatics*, 28(5): 729–730. 3

Warden, C. D., Lee, H., Tompkins, J. D., Li, X., Wang, C., Riggs, A. D., Yu, H., Jove, R., and Yuan, Y.-C. (2013). "COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis." *Nucleic Acids Research*, 41(11): e117–e117. 3, 19

Yu, X. and Sun, S. (2016). "HMM-DM: identifying differentially methylated regions using a hidden Markov model." *Statistical Applications in Genetics and Molecular Biology*, 15(1): 69–81. MR3464011. doi: https://doi.org/10.1515/sagmb-2015-0077. 3

Zeger, S. L. and Karim, M. R. (1991). "Generalized linear models with random effects: A Gibbs sampling approach." *Journal of the American Statistical Association*, 86: 79–86. MR1137101. 25