Biostatistics (2017), **0**, 0, *pp.* 1–29
doi:10.1093/biostatistics/allele-biostat

An Empirical Bayes Test for Allelic-Imbalance Detection in ChIP-seq

QI ZHANG*

Department of Statistics, University of Nebraska Lincoln, Lincoln, NE

qi.zhang@unl.edu

SÜNDÜZ KELEŞ

Department of Biostatistics and Medical Informatics and Department of Statistics, University of Wisconsin, Madison, WI

Summary

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) has enabled discovery of genomic regions enriched with biological signals such as transcription factor binding sites and histone modifications. Allelic-Imbalance detection is a complementary analysis of ChIP-seq data for associating biological signals with single nucleotide polymorphisms (SNPs) and has been successfully used in elucidating functional roles of non-coding SNPs. Commonly used statistical approaches for Allelic-Imbalance detection are often based on binomial testing and mixture models, both of which rely on strong assumptions on the distribution of the unobserved allelic probability, and have significant practical shortcomings. We propose Non-Parametric Binomial (NPBin) test for Allelic-Imbalance detection and for modeling Binomial data in general.

*To whom correspondence should be addressed.

NPBin models the density of the unobserved allelic probability non-parametrically, and estimates its empirical null distribution via curve fitting. We demonstrate the advantage of NPBin in its interpretability of the estimated density and the accuracy in Allelic-Imbalance detection using simulations and analysis of several ChIP-seq datasets. We also illustrate the generality of our modeling framework beyond Allelic-Imbalance detection by an effect size estimation problem with application to baseball data. This paper has supplementary material online.

Key words: Empirical Bayes; Non-parametric density estimation; Spline; High-throughput sequencing; Allelic-Imbalance.

1. Introduction

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) has been widely used for genome-wide profiling of histone modifications and transcription factor (TF)-DNA interactions. ChIP-seq experiments generate sequencing *reads* from sheared DNA fragments that are enriched at genomic locations with the specific TF-DNA interaction or protein modification of interest. A standard ChIP-seq analysis pipeline maps these reads to the corresponding *reference genome* (i.e., the typical whole genome sequence of the experimental samples), and identifies ChIP-seq peaks (Kharchenko *and others*, 2008). These peaks are the short genomic regions with higher than expected read counts, and are the candidate regions of the signal of interests. When there are two or more experimental groups, it is typical to study the differential enrichment within the peaks across experimental groups (Liang and Keleş, 2012). This pipeline can also handle other high-throughput sequencing experiments such as DNase-seq (Boyle *and others*, 2011) and ATAC-seq (Buenrostro *and others*, 2013) for chromatin accessibility. The main goal of these experiments is similar to that of ChIP-seq, i.e., discovering genomic regions containing of certain biological signals. We will also refer to them as ChIP-seq in this paper.

Standard ChIP-seq data analysis pipelines ignore the differences in DNA sequence among individuals by mapping reads to the reference genome. The diploid genomes consist of one maternal and one paternal copy of the genome. At each genomic position, the paternal and maternal alleles could be different than each other. The majority of these differences are in the form of heterozygous Single-Nucleotide Polymorphism (SNP), which is one letter substitution. Such one letter difference may be associated with significant differences in ChIP-seq signals. Each ChIP-seq peak on the reference genome consists of a pile of reads from the maternal genome, and a pile from the paternal genome. At heterozygous SNPS, we may observe that there are more reads from one allele than the other, which may suggest some biological functional difference between the two alleles (Figure 1(a)). We refer to such differences in ChIP-seq signals at heterozygeous SNPs as ALlelic-Imbalance (ALI) because it characterizes an imbalance between the two alleles in the biological signal level (e.g., read counts) at heterozygeous SNPs. We remark that ALI observed at heterozygeous SNPs are not necessarily caused by this genotype difference. Instead, it could be due to other molecular mechanisms such as imprinting. For the same reason, ALI events could take place at loci where the two alleles are identical. Nevertheless, we focus on ALI at heterozygeous SNPs because ALI signals in ChIP-seq are only observable at such loci. The hetetozygeous SNPs function as markers that differentiate the sequencing reads from the two alleles. Detecting ALI can also be viewed as one type of differential enrichment analysis between the two alleles. Studying ALI is important as it associates the genotype (SNPs) with the epigenetic signal (e.g., ChIP-seq peak) *in vivo* with internal control, and provides potential molecular mechanisms that enable interpretation of GWAS hits. For example, Verlaan *and others* (2009) further dissected a region associated with asthma, type 1 diabetes, primary biliary cirrhosis, and Crohn disease. They showed that a common disease allele with a strong association with asthma was also linked to changes in CTCF binding and nucleosome occupancy leading to altered domain-wide cis-regulation, which brought them one step closer to elucidating a mechanism for genetic susceptibility of ashma.

Detecting ALI includes the following three steps (Figure 1(b)): (1) *Alignment.* Map the reads, and assign each read to an allele; (2) *Read counting.* Count the reads at the SNPs of interests; and (3) *Statistical testing.* Determine the statistical significance and direction of ALI at each SNP and adjust for multiple testing. In this paper, we focus on the multiple testing problem in the last step, i.e., at each heterozygous SNP in a ChIP-seq peak, we are given the read counts from each allele, and need to decide whether it is an ALI SNP. For simplicity and following the standard assumptions in literature, we assume no error in genotyping, no read error at SNP locations, and that the read counts at different SNPs are independent. We propose Non-Parametric Binomial (NPBin) test, an empirical Bayes test for ALI detection. The major innovation of NPBin includes the non-parametric density estimation for the unobserved allele probability, and approximating the ideal empirical null distribution via curve fitting. The rest of this paper is organized as follows. In Section 2, we formulate ALI detection as a multiple testing problem, present NPBin, and draw its connections with the current literature on ALI detection. In Section 3, we compare NPBin with several other ALI detection procedures with simulations and real ChIP-seq data analysis, and further illustrate the generality of NPBin model by applying it to effect size estimation in a baseball example. We conclude with further discussion in Section 4.

2. ALI detection as a multiple testing problem

Detecting ALI at individual SNP positions can be formulated into the following statistical model. For $j = 1, \ldots, M$, let s_j be a phased heterozygous SNP. For ChIP-seq or any other sequencing dataset, let m_j and x_j denote the total number of reads and the number of reads from the maternal allele overlapping SNP s_j, respectively. We are interested in δ_j, an indicator variable denoting ALI status at s_j. We also define a latent variable p_j as the true allelic probability that a read covering s_j originates from the maternal allele. p_j is biologically meaningful because it characterizes the Allelic-Imbalance directly and quantitatively. We use the following hierarchical model for the allelic frequency. For $j = 1, \ldots, M$:

$$\begin{aligned} \delta_j &\sim \text{Bernoulli}(1 - \pi_0), \\ p_j | \delta_j = i &\sim g_i, \quad \text{for } i = 0, 1, \\ x_j | p_j, m_j &\sim \text{Binom}(m_j, p_j), \end{aligned} \tag{2.1}$$

where π_0 is the proportion of SNPs with no ALI and we treat m_j as a nuisance parameter. The marginal density of the unobserved allelic probability p is:

$$g(p) = \pi_0 g_0(p) + (1 - \pi_0) g_1(p).$$

For $i = 0, 1$, let $f_i(x_j; m_j) = \int_0^1 \text{Binom}(x_j; m_j, p) g_i(p) dp$ be the marginal probability of x_j at the SNP without or with ALI. We can also write the overall marginal distribution of x_j as

$$f(x_j; m_j) = \pi_0 f_0(x_j; m_j) + (1 - \pi_0) f_1(x_j; m_j).$$

Similar hierarchical models have been used in the literature of ALI detection (Skelly *and others*, 2011; Zhang *and others*, 2014). Along with other differences, these approaches assume g to be a two-component beta mixture, the disadvantages of which will be discussed later in Section 2.4. In contrast, we propose Non-Parametric Binomial (NPBin) test for ALI detection that estimates g non-parametrically. The formal framework consists of the following steps:

1. Estimate g directly via fitting splines.

2. Estimate g_0 by approximating the ideal empirical null distribution via curve fitting.

3. Control the false discovery rate via thresholding the local false discovery rate.

In what follows, we present these steps (See also Supplementary Notes for the implementation details), and discuss the conceptual advantage of NPBin compared to existing approaches.

2.1 *Non-parametric density estimation for the latent true allelic probability*

Non-parametric density estimation methods such as Poisson regression over histogram has been used in empirical Bayes testing problems when such histogram of the observed data is meaningful (Efron *and others*, 2001; Schwartzman, 2008). When the density of a latent variable needs to be estimated, it is often assumed to be discrete or have a parametric form (Liao *and others*, 2014; Efron, 2016). This is closely connected to the widely studied problem of estimating the mixing density nonparametrically. The works of Laird (1978), Lindsay *and others* (1983), and Efron (2016) were concerned with the case when the mixing density is discrete. Martin and Tokdar (2012), and Mabon (2016) developed deconvolution methods for additive noise. When the observations are i.i.d., Zhang (1995) derived a kernel estimator using fourier method, and Roueff and Rydén (2005), and Rebafka and Roueff (2015) discussed the general theoretical framework for orthogonal series estimators. However, we have not identified a reliable and practical algorithm for the setting considered in this paper where the mixing density is continuous and the non-additive noise depends on a nuisance parameter. This necessitates the development of a new nonparametric estimator.

We propose to estimate g using B-spline density estimation. Let $(2-K)/T, ..., -1/T, 0, 1/T, \ldots, T, 1+1/T, \ldots, 1+(K-2)/T$ be equally spaced knots, and $B_\ell(p;K)$ be the normalized K'th order B-spline defined on interval $[\ell/T, (\ell+K)/T]$ such that $\int_0^1 B_\ell(p;K)dp = 1$. These basis are restricted within $[0,1]$ and $B_{-K+1}(p;K)$ and $B_{T-1}(p;K)$ are discarded. Let $g(p)$ be a smooth density function which we can approximate as:

$$g(p) = \sum_{\ell=-K+2}^{T-2} a_\ell B_\ell(p;K), \tag{2.2}$$

where a_ℓ are non-negative coefficients. Viewing m_j as a nuisance parameter, we apply (2.2) to the joint likelihood of (p_j, x_j) and obtain

$$f(x_j, p_j; m_j) = \sum_{\ell=-K+2}^{T-2} a_\ell \binom{m_j}{x_j} p_j^{x_j} (1-p_j)^{m_j - x_j} B_\ell(p_j;K).$$

Thus, the marginal likelihood of x_j is given by:

$$f(x_j; m_j) = \int_0^1 f(x_j, p; m_j)dp = \sum_{\ell=-K+2}^{T-2} a_\ell \binom{m_j}{x_j} \int_{\ell/T}^{(\ell+K)/T} p^{x_j}(1-p)^{m_j - x_j} B_\ell(p; K)dp.$$

Let $c_{\ell,j} = \binom{m_j}{x_j} \int_{\ell/T}^{(\ell+K)/T} p^{x_j}(1-p)^{m_j-x_j} B_\ell(p;K)dp$. Note that this does not depend on any unknown parameters, and denotes a polynomial integral which can be solved explicitly. Hence,

$$f(x_j; m_j) = \sum_{\ell=-K+2}^{T-2} a_\ell c_{\ell,j}. \tag{2.3}$$

We fix the number of knots and do not add additional smoothness penalty when estimating g because the number of knots already controls the smoothness. Then, we estimate the coefficients a_ℓ's by maximizing the marginal likelihood $\prod_{j=1}^{M} f(x_j; m_j)$ via an Expectation-Maximization (EM) algorithm (Dempster *and others*, 1977). For fixed number of components, (2.3) can be viewed as a mixture distribution, and m_j is a fixed nuisance parameter. For $\ell = -K+2, \cdots, T-2$, define $y_{\ell,j}$ as the indicator that x_j is from component ℓ. Then $P(X_j = x_j | y_{\ell,j} = 1; m_j) = c_{\ell,j}$, and $P(y_{\ell,j} = 1) = a_\ell$. The exact steps of this EM-algorithm are as follows.

1. *Initialization.* Set $a_\ell = a_\ell^{(0)}$ for $\ell = -K+2, \ldots, T-2$ and $L^{(0)} = -\infty$.

2. *E-step.* For current coefficients $\mathbf{a^{(t)}} = (a_{-K+2}^{(t)}, \ldots, a_{T-2}^{(t)})'$,

$$z_{\ell,j}^{(t+1)} \equiv P(y_{\ell,j} = 1 | x_j; \mathbf{a^{(t)}}, m_j) = \frac{a_\ell^{(t)} c_{\ell,j}}{\sum_{k=-K+2}^{T-2} a_k^{(t)} c_{k,j}}.$$

3. *M-step.* Update the coefficients,

$$a_\ell^{(t+1)} = \frac{\sum_{j=1}^{M} z_{\ell,j}^{(t+1)}}{\sum_{k=-K+2}^{T-2} \sum_{j=1}^{M} z_{k,j}^{(t+1)}}.$$

4. *Stopping rule.* Repeat steps 2-3 until the increase in the following estimated marginal likelihood is small in terms of both the absolute and the relative value.

$$L^{(t+1)} = \sum_{j=1}^{M} \log \left(\sum_{\ell=-K+2}^{T-2} a_\ell^{(t+1)} c_{\ell,j} \right).$$

Stop when

$$(L^{(t+1)} - L^{(t)}) \cdot \max(1, 1/|L^{(t+1)}|) < \mathrm{err}_{\max},$$

where $\mathrm{err}_{\max}$ is a pre-specified control parameter.

2.2 *Approximating the ideal empirical null distribution*

Next, we will discuss the estimation of the empirical null. The distribution of the latent variable g is the mixture of the null density g_0 and g_1, the density under the alternative. For the sake of identifiability, it is often assumed in literature that the density of g_1 is zero in the center of the bulk region (e.g., an interval around 0 in the case of z-score, and correspondingly an interval around 0.5 in our setting), which is referred to as *zero assumption* (Efron, 2012). Under this assumption, and when the observed data are i.i.d., empirical null can be estimated using a "central-matching" method (Efron *and others*, 2001; Schwartzman, 2008). However, the above methods are not directly applicable when the random variable that follows g is not observed.

In what follows, we exploit the zero assumption from different angle and derive our estimate of the empirical null from the characterization of the ideal null distribution. When the signal is not too weak,the implicit assumption underneath the zero assumption is that g_0 and g_1 need to be "separable" to a certain degree, so that ALI detection is possible. Then

$$r(p) \equiv \frac{\pi_0 g_0(p)}{g(p)} = \begin{cases} \text{close to 1} & \text{when } p \text{ is close to } 0.5, \\ \text{close to 0} & \text{when } p \text{ is close to 0 or 1,} \end{cases}$$

whose most extreme (and ideal) form is

$$r(p) = 1\{p \in A\}) \text{ where } A \text{ is a sub-interval of } [0,1] \text{ around } 0.5. \tag{2.4}$$

In this case, ALI detection becomes easily tractable. Thus we call g_0 that satisfies (2.4) the "ideal null". We observed that the derivative of $r(p)$ in (2.4) is zero when it can be defined. Motivated by this, we let $g_0(p) = \text{Beta}(p; \alpha_0, \beta_0)$, and approximate the ideal null by minimizing

$$\int_0^1 \left[\frac{d}{dp} \left(\frac{g_0(p)}{g(p)} \right) \right]^2 g(p)dp, \tag{2.5}$$

where we assume that $g(p) > 0$ for $p \in [0, 1]$. This formulation does not require the supports of $g_0(p)$ and $g_1(p)$ to be disjoint, but only that the two densities are separable so that the zero assumption is satisfied. Noting that $\pi_0 g_0(p) \leqslant g(p)$ for any $p \in [0, 1]$, π_0 can be estimated by

$$1/\max_{p\in[0,1]}(g_0(p)/g(p)).$$

In practice, we estimate (2.5) numerically, and set

$$\pi_0 = 1/q_{0.975}(g_0(p)/g(p)), \tag{2.6}$$

for a more robust estimate, where $q_d(b)$ is d-th quantile of b.

2.3 *FDR control by local false discovery rate thresholding*

The local false discovery rate (locfdr) can be defined as

$$locfdr_j = \hat{\pi}_0 \hat{f}_0(x_j; m_j)/\hat{f}(x_j; m_j).$$

where $\hat{f}$, $\hat{f}_0$, and $\hat{\pi}_0$ are the estimated f, f_0, and π_0. The locfdr has the interpretation of $P(H_{0j}|x_j; m_j)$, where H_{0j} is the null hypothesis that $\delta_j = 0$. FDR can be controlled by

$$FDR_J = \frac{1}{J}\sum_{j=1}^{J} locfdr_{(j)},$$

where $locfdr_{(j)}$ is the j'th (increasingly) sorted locfdr value. Similar FDR control methods have been used in other genomics problems (Zhao *and others*, 2013).

2.4 *Connection to Existing ALI Detection Methods*

The binomial test is widely used for count data from next generation sequencing (Rozowsky *and others*, 2011). For ALI detection, it treats g_0 to be a point mass at 0.5, and ignores the natural over-dispersion of p even when there is no ALI. Consequently, it is overpowered when m_j is large. Modeling g as a Beta mixture is an appealing alternative considered by many (Muralidharan, 2010; Skelly *and others*, 2011; Zhang *and others*, 2014), where the null g_0 is one component (potentially with equal shape parameters) and g_1 consists of the other beta components. Although this Beta-Binomial mixture model-based empirical Bayes approach is widely used, the parametric assumptions on the alternative hypothesis can be unrealistic, and may result in inflexibility and may not be able to capture the density of the latent variable p.

A second and perhaps more important issue with the Beta-Binomial mixture model is due to its lack of identifiability and the interpretability of the estimated g. In fact, the estimated g can be visually different from the truth. As a simple exposition, we show the density estimation results from one of our simulation settings ($(\pi_0, \alpha_0, d) = (0.8, 5, 0.3)$; See Section 3.2 for details) where $g = 0.8 \times \text{Beta}(5, 5) + 0.1 \times \text{Beta}(1.25, 5) + 0.1 \times \text{Beta}(5, 1.25)$. We compared the estimated g from the Beta-Binomial mixture model (BBmix; Muralidharan 2010) and our proposed non-parametric method (NPBin). We observed that BBmix led to an inaccurate and spiky estimate even though the true model of x_j is Beta-Binomial mixture. In contrast, NPBin yielded an accurate, and thus more interpretable estimate (Figure 2(a)). This issue was also observed by Muralidharan (2010) who remarked that very different models for g might give nearly identical marginal for f.

A data-integration approach (Skelly *and others*, 2011; Zhang *and others*, 2014) emerged as an alternative in estimating g and g_0. This approach estimates g_0 by fitting a Beta-Binomial model on additional whole genome sequencing (DNA-seq) data of the subjects/cells that the functional assays (e.g., ChIP-seq, RNA-seq) originate from. Assuming this g_0 to be the same as the null distribution for the functional assay, it then estimates the other parameters of a two-component Beta-Binomial mixture from the functional assay for ALI detection. Its critical assumption could be easily violated for two reasons. First, different assays target at different types of genomic regions (e.g., RNA-seq is for transcribed regions and DNase-seq for accessible regions), which may lead to null distributions of allelic probability. Second, many other technical factors may also affect the null distributions. In fact, the estimated g_0's from different technical replicates of the same biological sample could look very different (Figure 2(b)). Additionally, the availability of whole genome DNA-seq data with satisfactory coverage is a bottleneck in practice, e.g., many of the cell lines utilized heavily by the ENCODE project does not have whole genome sequences.

Due to the limitations of the above methods, there is still pressing need for novel statistical tests for ALI detection. Our proposed NPBin relies on fewer assumptions, produces interpretable estimates, and does not require external data. Non-parametric density estimation has been widely used in empirical Bayes testing (Efron *and others*, 2001; Efron, 2016). If the allelic probability p is directly observed, non-parametric estimation of g is relatively easy. However, the allelic probability p is unobserved. The key innovation of NPBin test is the non-parametric estimation of the density g of the latent allelic probability p without requiring external data.

3. Results

We next evaluated our proposed method NPBin from several perspectives. Next three subsections describe the methods included in the comparison, the simulation models, and the ChIP-seq datasets we analyzed. Then, we present comparison results on accuracy and interpretability in density estimation and accuracy in ALI detection using both simulations and the actual ChIP-seq datasets. To demonstrate the broader applicability of NPBin beyond ALI detection, we illustrate the generality and highlight the interpretability of NPBin's density estimate in a baseball dataset.

3.1 *Methods compared in simulations and data analysis*

If p_j's are observed, it is natural to test ALI in a similar fashion to the empirical Bayes tests on z-scores (Efron, 2012). We refer to this method as Empirical Bayes Oracle (**EBO**) because it requires oracle information, i.e., the true values of the latent variable p. EBO first directly estimates

g using splines. Next, EBO estimates g_0 in Beta family by maximum likelihood using the data in the "bulk" region and also accounting for truncation (Efron 2012; See also Supplementary Notes for details). In our simulation studies where we know the true allelic probability, we compared EBO with NPBin. When we do not observe p_j's as is in real data applications, a naive approach is to estimate g and g_0 using $\hat{p}_j = x_j/m_j$ by essentially treating $\hat{p}_j$ as the true value of p_j. We refer this method as Empirical Bayes Estimated (**EBE**). We also included a Beta-Binomial modeling approach for comparisons in simulations, for which the true number of components and the true mean of each component are given. Empirically, we found that these additional oracle information helps avoid the identifiability issues observed by Muralidharan (2010) and us, and yields accurate estimates. This mixture model approach is different from the existing standard approaches as it is enhanced by the oracle information (true means of the components of the simulation model), and we refer it as the Oracle Enhanced Mixture model (**OEMix**). We designed this approach strictly for the comparison purpose in simulations. It is excluded in comparisons using actual ChIP-seq data analysis since such oracle information is not available. For the Binomial test, we used Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) for FDR control. Table 1 summarizes all the methods used in comparisons with simulations.

3.2 *Simulation model*

For $j = 1, \dots, M$, we simulated the ChIP-seq read counts as the following. We first simulated the allele specific binding state u_j of SNP j as

$$u_j \sim \pi_0 \Delta_0 + 0.5(1 - \pi_0)\Delta_1 + 0.5(1 - \pi_0)\Delta_{-1}$$

where Δ_i is a point mass at i for $i \in \{-1, 1, 0\}$. Here $u_j = 1$ represents that at SNP j, the transcription factor only binds to the maternal allele, $u_j = -1$ only to the paternal allele, and $u_j = 0$ refers to binding to both alleles. Since ALI analysis of ChIP-seq are usually done within the pre-defined ChIP-seq peak regions, it is sensible to assume that the factor under consideration binds to at least one allele. Then we simulated the expected ChIP-seq read counts from the maternal and the paternal alleles $(\lambda_{m,j}, \lambda_{p,j})$ as independent Gamma samples

$$\begin{aligned}\lambda_{m,j}|u_j &\sim 1\{u_j \neq -1\}\text{Gamma}(\alpha_0, \theta_0) + 1\{u_j = -1\}\text{Gamma}(\alpha_1, \theta_1) \\ \lambda_{p,j}|u_j &\sim 1\{u_j \neq 1\}\text{Gamma}(\alpha_0, \theta_0) + 1\{u_j = 1\}\text{Gamma}(\alpha_1, \theta_1)\end{aligned} \tag{3.7}$$

In this setup, if the TF binds to one allele, its expected ChIP-seq count follows $\text{Gamma}(\alpha_0, \theta_0)$, and $\text{Gamma}(\alpha_1, \theta_1)$ otherwise. Finally, we simulated the observed counts from the two alleles $(x_j, m_j - x_j)$ as

$$x_j|\lambda_{m,j} \sim \text{Poison}(\lambda_{m,j}) \quad \textit{and} \quad m_j - x_j|\lambda_{p,j} \sim \text{Poison}(\lambda_{p,j})$$

Then the true maternal allele frequency $p_j = \lambda_{m,j}/(\lambda_{m,j} + \lambda_{p,j})$ and there is $x_j|m_j, p_j \sim \text{Binom}(m_j, p_j)$. Let $\theta_0 = \theta_1$, then it is easy to see that the true maternal allele frequencies are actually i.i.d. samples from the following Beta mixture

$$g(p) = \pi_0\text{Beta}(\alpha_0, \alpha_0) + 0.5(1 - \pi_0)\text{Beta}(\alpha_0, \alpha_1) + 0.5(1 - \pi_0)\text{Beta}(\alpha_1, \alpha_0)$$

In particular, the null distribution $(u_j = 0)$ is $g_0(p) = \text{Beta}(\alpha_0, \alpha_0)$.

In this model, π_0 is the proportion of the null, and α_0 controls the over-dispersion level of the null. We further define $d = \frac{\alpha_0 - \alpha_1}{2(\alpha_0 + \alpha_1)}$ representing the strength of the ALI signal, as it is the difference between one component of the alternative (e.g., $\text{Beta}(p; \alpha_0, \alpha_1)$) to 0.5. The simulation parameters are set as follows: $M = 5000$, $\pi_0 = 0.8, 0.9$, $\alpha_0 = 5, 20$ and $d = 0.3, 0.4$. In each setting, we set $\theta_0 = \theta_1 = 15/\alpha_0$, i.e., the average read count from one allele on all SNPs with TF binding is roughly 15, which is reflective of many ChIP-seq experiments. We replaced the SNPs with $m_j < 5$ with new simulations, so the minimal total read count m_j is at least 5. It is easy to check that such manipulation does not influence the distribution of the true allelic frequency. We only present the results for $(\pi_0, \alpha_0, d) = (0.8, 5, 0.4)$ and $(0.9, 20, 3)$. The rest are presented in the Supplementary Materials, and generally consistent with our findings presented here.

3.3 *Pre-processing of ChIP-seq data*

We analyzed many ChIP-seq and ChIP-seq like data from GM12878 cells (Supplementary Table 1), including DNase-seq, ATAC-seq, and ChIP-seq for CCCTC-binding factor (CTCF), one of the most commonly studied transcription factors within consortia projects. GM12878 is a human lymphoblastoid cell line that has been intensively used by large consortia projects such as ENCODE (Consortium *and others*, 2012) and generally in epigenetics. SNPs and diploid genome sequence for this cell line are available through the 1000 Genomes project (Siva, 2008). We used the personalized-genome-based AlleleSeq (Rozowsky *and others*, 2011) to map the reads and obtain (x_j, m_j), the maternal and the overall read counts at each phased heterozygous SNP of GM12878. Then, we compared different statistical tests for ALI detection using the same pre-processed data, among which Binomial test is utilized in the original AlleleSeq. AlleleSeq was originally designed for single-end reads, and we used a modified version from Zhang *and others* (2016) for paired-end reads. For all the experiments, we exclusively focused on the SNPs within the peaks (signal enriched regions) identified for each experiment because these are the candidate regions with biological signal. The peak information was obtained from the publications listed in Supplementary Table 1. We remark that NPBin and many other statistical tests for ALI detection do not depend on the pre-processing methods. As a result, they can be applied as long as the maternal and the overall read counts at each SNP are provided.

3.4 *Estimation accuracy of the overall density and the null density of the latent variable*

We first evaluated the accuracy in estimating g and g_0 with ChIP-seq data and simulation studies. In the analysis of ChIP-seq data, we compared the estimated g and g_0 with the histogram of $\hat{p}_j = x_j/m_j$ for $j = 1, \cdots, M$. The variation in $\hat{p}_j$ has two sources, g and the Binomial model. Thus, the estimated g and g_0 can be considered as being reasonable if they are slightly tighter than the histogram of $\hat{p}_j$, but not too far away, and if g and $\pi_0 g_0$ are close to each other around

0.5. As expected, we found that NPBin yielded reasonable estimates (Figure 3(a)-(c)), while EBE's estimates of g and g_0 were too flat and too dissimilar around 0.5 (Figure 3(d)-(f)).

We next compared the accuracy of different methods in estimating g using $L1$ loss (Table 2 and Supplementary Table 2), and found that NPBin performed almost as well as OEMix and EBO, and significantly better than EBE. Since both OEMix and EBO require oracle information that are not generally available, we concluded that NPBin is much closer to the oracle performance. We also compared the bias and standard error of the estimated null shape parameter α_0 and the proportion of the null π_0 (Table 2 and Supplementary Table 3). Since none of these methods assumes equal shape parameters in the null model, we used the average of the two shape parameters of the estimated null as the estimated α_0. This analysis revealed that all the methods tended to underestimate the shape parameters, and overestimate the proportion of the null. Overall, NPBin led to better accuracy than EBE.

3.5 *Accuracy comparison with simulations*

Next, we assessed the accuracy in ALI detection in two aspects using simulations. First, we evaluated the ranking performance using Precision-Recall Curve (PRC), which is more appropriate for non-balanced data comparing to ROC. Different methods may rank the SNPs differently for ALI evidence. For fixed number of selected top SNPs across all replicates, we calculated the average precision and average recall for each method, and use these averages to draw PRC curves (Figure 4 and Supplementary Figure 2). We found that all empirical Bayes methods performed similarly, and were all much better than the Binomial test. Second, we compared the number of selected loci and empirical FDR at the same nominal FDR control level (Table 2 and Supplementary Tables 4-5). Only NPBin selected reasonable number of loci, and were similar to the numbers from oracle methods. In comparison, Binomial test selected too many, and EBE selected too few. This observation is compatible with the empirical FDR results: The empirical FDR level of NPBin and the oracle-assisted methods were close to the nominal level, while EBE was too conservative, and the Binomial test led to very large empirical FDR.

In order the investigate the impact of misspecification of the null model, we substitute θ_0 in (3.7) with $\theta_{0,j} = \theta_0 v_j$ where $v_j \sim Unif[0.8, 1.2]$, which results in an infinite beta mixture with mean 0.5 as the null model. The accuracy comparison using this model leads to similar results (Supplementary Figure 2 and Supplementary Tables 6-7).

3.6 *ALI detection in ChIP-seq data*

We designed an evaluation criterion for the actual ChIP-seq ALI detection analysis. Specifically, we used `R` package `atSNP` (Zuo *and others*, 2015) to identify the SNPs with significant allelic difference in TF binding motif strength based on the sequence information and the 205 known TF motifs in vertebrates from the JASPAR database (Mathelier *and others*, 2013). Because both information sources are independent of the ChIP-seq data under consideration, the allelic motif strength difference can serve as an external validation criterion (See also Supplementary Notes for more details). Such external validation criterion is only available for 5-10% of SNPs, and the other SNPs simply showed no significant allelic difference in motif strength. Reasons for this include other unknown factors affecting TF binding and chromatin accessibility such as imprinting and the incompleteness of the JASPAR database. For each of the 5-10% SNPs that exhibit significant allelic difference in motif strength, an expected winning allele in ALI detection was assigned to it based on the sequence information. The results were compared with the actual winning alleles in ALI detection from ChIP-seq data. Note that all ALI detection methods will report the same winning allele at the same SNP and from the same ChIP-seq data, e.g., if $x_j > m_j - x_j$, indicating more maternal reads than paternal reads at SNP j, all methods will report the maternal allele as the winning allele. However, their estimated significance could be dramatically different. We label a SNP as a potential True Positive (**TP**) if the expected winning allele based on sequence information and the actual winning allele in ALI detection from ChIP-seq data are the same and a potential False Positive (**FP**) otherwise. When such a potential TP (or FP) SNP has enough statistical significance to be chosen as a ALI SNP from ChIP-seq data, its winning allele will agree (or contradict) with the motif-based benchmark. Thus, it can be viewed as a true positive (or false positive) in the conventional setting. However, potential true negatives or false negatives cannot be defined in the same fashion due to the incompleteness of the databases of known TF motifs, and the limited knowledge on the other factors that may affect ChIP-seq signal.

Since the TP and FP labels were only available for a small proportion of SNPs, quantifying the differences between methods with Precision Recall curves became less relevant (Supplementary Figure 3). As an alternative, we used $\log_2$ (TP/FP) as the measure of accuracy for fixed number of selected ALI SNPs. Since ALI generally only exists at a small proportion of SNPs, we focused on the top 20% SNPs. Using this benchmark, we found that NPBin showed slightly higher accuracy on ATAC-seq and DNase-seq, and all methods performed similarly on CTCF ChIP-seq (Figure 5). The numbers of TP and FP for fixed proportion of selected SNPs are in Supplementary Table 8. Consistent with the simulation results, we found that EBE was often too conservative and the Binomial test was too liberal in ALI detection at fixed nominal FDR (Table 3).

We also compared the ranking self-consistency of these methods across different replicates using Spearman's rank correlation. Specifically, we ran NPBin, EBE and Binom on all individual replicates. For each method, we rank all candidate SNPs by $w_j = (1 - FDR_j) \cdot \mathbf{sgn}(2x_j - m_j)$, where FDR_j is FDR (for the method under consideration) at SNP j. Ranking by w_j also takes into account the direction of ALI signals. For each pair of replicates, we focused on the common candidate SNPs. Then for each method, we calculated the Spearman's correlation of w_j's from the two replicates. Overall, we found that the rankings of EBE are less consistent across replicates (lower correlation) than Binom and NPBin (Supplementary Table 9).

3.7 *Re-analysis of Brown (2008) major league baseball data*

We developed NPBin as an empirical Bayes testing method; however, the non-parametric density estimation procedure of NPBin is applicable to other problems, such as prior and effect size estimation. We illustrated its application in a non-bological context by re-analyzing a Major League Baseball data that has been used in Brown (2008) and Muralidharan (2010). We obtained the data and the code from Muralidharan (2010).

The dataset consists of batting records from the 2005 season. For player j, let m_j be the number of bats, and x_j the number of hits in the first half of the season. It is reasonable to assume $x_j \sim \text{Binomial}(m_j, p_j)$, where p_j is the unknown true batting average of this player. The goal is to predict the batting average of the second half of the season using the posterior mean of p_j from the first half season. The key for this problem is to accurately estimate g, the prior of p_j. Similar to Muralidharan (2010), we focused on the 567 players with at least 11 bats in the first half of the season. Muralidharan (2010) analyzed the data as a whole, and also analyzed the pitchers and non-pitchers separately, because the author argued that better batters bat more which violated the Binomial model, and splitting the players by pitchers and non-pitchers reduced the variation in the number of bats in each group. We followed the same procedure for easy comparison. We compared the Beta-Binomial mixture model (BBmix; Muralidharan 2010), with the non-parametric density estimation component of NPBin.

We first compared the estimated priors by the two methods, with reference to the histogram of $\hat{\mathbf{p}}$, the sample batting average of the first half of the season. The variation in $\hat{\mathbf{p}}$ has two sources, g and the Binomial model. Thus, we expected the estimated priors to be tighter than the histograms, but not too far away. We found that Beta-Binomial mixture model was in favor of very spiky estimates (Figure 6), and sometimes dramatically visually different from the histogram (Figure 6(b)). This again highlighted the identifiability issue and consequently the lack of interpretability of priors estimated by Beta-Binomial mixture models. In contrast, the NPBin estimates were smoother, closer to the histograms, yet still tighter as expected. Thus, they were better and more interpretable representation of the truth.

We next compared the batting average prediction, where the batting records of the second serve as the test set. Brown (2008) and Muralidharan (2010) used the following loss function

$$TSE = \sum_{j=1}^{n} \left(\mu_j - \arcsin\sqrt{\frac{\tilde{x}_j + 1/4}{\tilde{m}_j + 1/2}} \right)^2 - \frac{1}{4\tilde{m}_j},$$

where $(\tilde{m}_j, \tilde{x}_j)$ are the number of bats and hits of player j in the second half of the season, and $\mu_j = E(\arcsin\sqrt{p_j} \mid \mathbf{m}, \mathbf{x})$ where $(\mathbf{m}, \mathbf{x})$ are all the bats and hits in the training data. In addition to this loss function, we also compared the $L2$ loss in batting average prediction given by $\sum_{j=1}^{n}(\tilde{p}_j - \tilde{x}_j/\tilde{m}_j)^2$, where $\tilde{p}_j = E(p_j|\mathbf{m}, \mathbf{x})$. We found that NPBin performed better for the whole data and for the pitcher data, but not for the non-pitcher data (Table 4).

4. Summary and Discussion

In this paper, we studied the problem of detecting Allelic-Imbalance in ChIP-seq signals between the maternal and paternal alleles. We proposed and implemented a Non-Parametric Binomial (NPBin) test for ALI detection and for modeling Binomial data in general. NPBin estimates the overall density of the latent allelic probability nonparametrically, and estimates the empirical null density by curve fitting that approximate the ideal empirical null. It makes minimal assumptions on the data generating model, and does not rely on external data. We illustrated the advantages of NPBin in the accuracy of ALI detection, and the accuracy and interpretability of the estimated density of the latent variable using simulations and real ChIP-seq data analysis. We also illustrated the generality of NPBin by applying it to effect size estimation in the context of baseball.

NPBin takes count data (m_j and x_j) as input, and performs statistical testing for ALI detection. Even though it properly models the variation in allelic probability, it does not correct the bias from haplotype variation, wet lab experiment issues, sequencing error, and read mapping.

Besides improving web lab experiments, bioinformatical proposals for fixing such biases includes incorporating additional DNA-seq data, screening candidate loci, and improving allele-specific mapping (Rozowsky *and others*, 2011; Younesy *and others*, 2013; Bailey *and others*, 2015; Van De Geijn *and others*, 2015). These could result in improved count data for testing using NPBin.

We focused on the problem of ALI detection from ChIP-seq at single SNP level, and from one data set. There have been many other ALI detection methods and software that concern ALI in RNA-seq (Mayba *and others*, 2014), region-level ALI detection (Van De Geijn *and others*, 2015), joint ALI detection from multiple types of functional assays (de Santiago *and others*, 2017; Wei *and others*, 2012), and ALI-informed QTL/eQTL analysis of multiple individuals (Van De Geijn *and others*, 2015; Sun, 2012). The statistical machinery behind these approaches are mostly parametric empirical Bayes models. Our current nonparametric empirical Bayes framework could be potentially applied to these problems. For example, we can combine our SNP level results through a meta-analysis framework similar to MBASED (Mayba *and others*, 2014) and detect region or gene level ALI from ChIP-seq or RNA-seq. In our real data analysis, the accuracy evaluation criterion as based on the binding affinity, which is not the only potential cause of ALI in ChIP-seq. As we have pointed out, other molecular mechanism such as (allele-specific) imprinting may also cause ALI in ChIP-seq, at loci with or without genetic variations. One future study along this line is a joint ALI analysis of ChIP-seq and methylation data.

Supplementary Material

Supplementary material, including R code and sample input data will be available online upon publication of this paper.

Acknowledgments

S.K. was partially supported by NIH BD2K grant U54 AI117924, NIH UO1 grant HG007019, and NIH RO1 grant HG003747. *Conflict of Interest*: None declared.

References

Bailey, Swneke D, Virtanen, Carl, Haibe-Kains, Benjamin and Lupien, Mathieu. (2015). Abc: a tool to identify snvs causing allele-specific transcription factor binding from chip-seq experiments. *Bioinformatics*, btv321.

Benjamini, Yoav and Hochberg, Yosef. (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.

Boyle, Alan P, Song, Lingyun, Lee, Bum-Kyu, London, Darin, Keefe, Damian, Birney, Ewan, Iyer, Vishwanath R, Crawford, Gregory E and Furey, Terrence S. (2011). High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome research* **21**(3), 456–464.

Brown, Lawrence D. (2008). In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *The Annals of Applied Statistics*, 113–152.

Buenrostro, Jason D, Giresi, Paul G, Zaba, Lisa C, Chang, Howard Y and Greenleaf, William J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**(12), 1213–1218.

Consortium, ENCODE Project *and others*. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414), 57–74.

de Santiago, Ines, Liu, Wei, Yuan, Ke, O'Reilly, Martin, Chilamakuri, Chandra Sekhar Reddy, Ponder, Bruce A. J., Meyer, Kerstin B. and Markowetz, Florian. (2017). Baalchip: Bayesian analysis of allele-specific transcription factor binding in cancer genomes. *Genome Biology* **18**(1), 39.

Dempster, Arthur P, Laird, Nan M and Rubin, Donald B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

Efron, Bradley. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, Volume 1. Cambridge University Press.

Efron, Bradley. (2016). Empirical Bayes deconvolution estimates. *Biometrika* **103**(1), 1–20.

Efron, Bradley, Tibshirani, Robert, Storey, John D and Tusher, Virginia. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American statistical association* **96**(456), 1151–1160.

Kharchenko, Peter V, Tolstorukov, Michael Y and Park, Peter J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology* **26**(12), 1351–1359.

Laird, Nan. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* **73**(364), 805–811.

Liang, Kun and Keleş, Sündüz. (2012). Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics* **28**(1), 121–122.

Liao, JG, Mcmurry, Timothy and Berg, Arthur. (2014). Prior robust empirical bayes inference for large-scale data by conditioning on rank with application to microarray data. *Biostatistics* **15**(1), 60–73.

Lindsay, Bruce G *and others*. (1983). The geometry of mixture likelihoods: a general theory. *The annals of statistics* **11**(1), 86–94.

Mabon, Gwennaëlle. (2016). Adaptive deconvolution of linear functionals on the nonnegative real line. *Journal of Statistical Planning and Inference* **178**, 1–23.

Martin, Ryan and Tokdar, Surya. (2012). A nonparametric empirical bayes framework for large-scale multiple testing. *Biostatistics* **13**(3), 427–439.

Mathelier, Anthony, Zhao, Xiaobei, Zhang, Allen W, Parcy, François, Worsley-Hunt, Rebecca, Arenillas, David J, Buchman, Sorana, Chen, Chih-yu, Chou, Alice, Ienasescu, Hans *and others*. (2013). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic acids research*, gkt997.

Mayba, Oleg, Gilbert, Houston N, Liu, Jinfeng, Haverty, Peter M, Jhunjhunwala, Suchit, Jiang, Zhaoshi, Watanabe, Colin and Zhang, Zemin. (2014). Mbased: allele-specific expression detection in cancer tissues and cell lines. *Genome biology* **15**(8), 405.

Muralidharan, Omkar. (2010). An empirical Bayes mixture method for effect size and false discovery rate estimation. *The Annals of Applied Statistics*, 422–438.

Rebafka, Tabea and Roueff, François. (2015). Nonparametric estimation of the mixing density using polynomials. *Mathematical Methods of Statistics* **24**(3), 200–224.

Roueff, Francois and Rydén, Tobias. (2005). Nonparametric estimation of mixing densities for discrete distributions. *Annals of statistics*, 2066–2108.

Rozowsky, Joel, Abyzov, Alexej, Wang, Jing, Alves, Pedro, Raha, Debasish, Harmanci, Arif, Leng, Jing, Bjornson, Robert, Kong, Yong, Kitabayashi, Naoki *and* *others*. (2011). AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular systems biology* **7**(1).

Schwartzman, Armin. (2008). Empirical null and false discovery rate inference for exponential families. *The Annals of Applied Statistics*, 1332–1359.

Siva, Nayanah. (2008). 1000 Genomes project. *Nature biotechnology* **26**(3), 256–256.

Skelly, Daniel A, Johansson, Marnie, Madeoy, Jennifer, Wakefield, Jon and Akey, Joshua M. (2011). A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome research* **21**(10), 1728–1737.

Sun, Wei. (2012). A statistical framework for eQTL mapping using RNA-seq data. *Biometrics* **68**(1), 1–11.

Van De Geijn, Bryce, McVicker, Graham, Gilad, Yoav and Pritchard, Jonathan K. (2015). Wasp: allele-specific software for robust molecular quantitative trait locus discovery. *Nature methods* **12**(11), 1061–1063.

Verlaan, Dominique J, Berlivet, Soizik, Hunninghake, Gary M, Madore, Anne-Marie, Larivière, Mathieu, Moussette, Sanny, Grundberg, Elin, Kwan, Tony, Ouimet, Manon, Ge, Bing *and others*. (2009). Allele-Specific Chromatin Remodeling in the ZPBP2/GSDMB/ORMDL3 Locus Associated with the Risk of Asthma and Autoimmune Disease. *The American Journal of Human Genetics* **85**(3), 377–393.

Wei, Yingying, Li, Xia, Wang, Qian-fei and Ji, Hongkai. (2012). iASeq: integrative analysis of allele-specificity of protein-DNA interactions in multiple ChIP-seq datasets. *BMC genomics* **13**(1), 681.

Younesy, Hamid, Möller, Torsten, Heravi-Moussavi, Alireza, Cheng, Jeffrey B, Costello, Joseph F, Lorincz, Matthew C, Karimi, Mohammad M and Jones, STEVEN JM. (2013). Alea: a toolbox for allele-specific epigenomics analysis. *Bioinformatics*, btt744.

ZHANG, CUN-HUI. (1995). On estimating mixing densities in discrete exponential family models. *The Annals of Statistics*, 929–945.

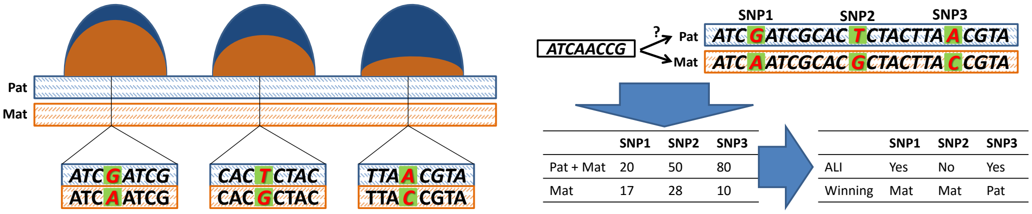
ZHANG, QI, ZENG, XIN, YOUNKIN, SAM, KAWLI, TRUPTI, SNYDER, MICHAEL P AND KELEŞ, SÜNDÜZ. (2016). Systematic evaluation of the impact of ChIP-seq read designs on genome coverage, peak identification, and allele-specific binding detection. *BMC bioinformatics* **17**(1), 1.

ZHANG, SHAOJUN, WANG, FANG, WANG, HONGZHI, ZHANG, FAN, XU, BIN, LI, XIA AND WANG, YADONG. (2014). Genome-wide identification of allele-specific effects on gene expression for single and multiple individuals. *Gene* **533**(1), 366–373.

ZHAO, ZHIGEN, WANG, WEI, WEI, ZHI *and others*. (2013). An empirical bayes testing procedure for detecting variants in analysis of next generation sequencing data. *The Annals of Applied Statistics* **7**(4), 2229–2248.

ZUO, CHANDLER, SHIN, SUNYOUNG AND KELEŞ, SÜNDÜZ. (2015). atSNP: transcription factor binding affinity testing for regulatory SNP detection. *Bioinformatics*, btv328.

[*Received xx, 2017; revised xx, 2017; accepted for publication xx, 201x*]

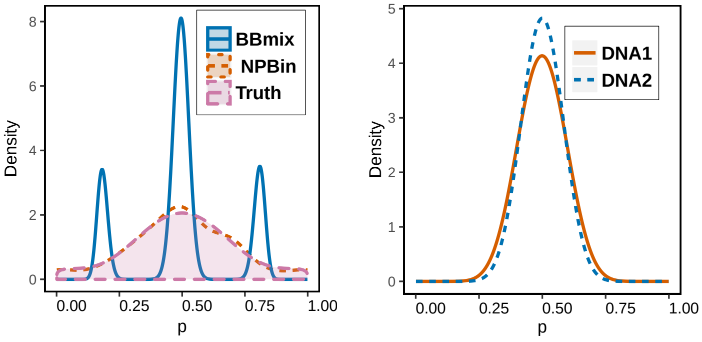


	SNP1	SNP2	SNP3
Pat + Mat	20	50	80
Mat	17	28	10

	SNP1	SNP2	SNP3
ALI	Yes	No	Yes
Winning	Mat	Mat	Pat

(a) (b)

Fig. 1: Definition and detection of Allelic-Imbalance (ALI) in ChIP-seq data. (a) Each ChIP-seq peak with a heterozygous SNP consists of reads from the paternal (Pat) and the maternal (Mat) alleles. On the left is a peak with majority of the reads from Mat; the middle peak has equal numbers of reads from both alleles; while the peak on the right is dominated by the Pat reads. Both the left and right peaks with heterozygeous SNPs exhibit ALI. (b) A typical ALI detection pipeline: 1) *Alignment.* Identifying the genomic origin and the contirbuting allele of each read; 2) *Read counting.* Tallying the total read count (Pat+Mat) and the Mat read count at each SNP; and 3) *Statistical testing.* For each SNP, deciding on the winning allele and whether the difference between Mat and Pat read counts reach a certain statistical significance.



(a) (b)

Fig. 2: (a) Beta-Binomial mixture led to inaccurate and spiky estimates of g. (b) The estimates of g_0 from two separate DNA-seq data of GM12878 cells were different.

Abbr.	Name	Applicable to real data
NPBin	Non-Parametric Binomial	Yes
OEMix	Oracle-Enhanced Mixture model	No
EBO	Empirical Bayes Oracle	No
EBE	Empirical Bayes Estimated	Yes
Binom	Binomial Test	Yes

Table 1: Methods compared in Section 3. Methods marked as not applicable to real data rely on oracle information, and, hence, were used only in simulations for benchmarking purposes.

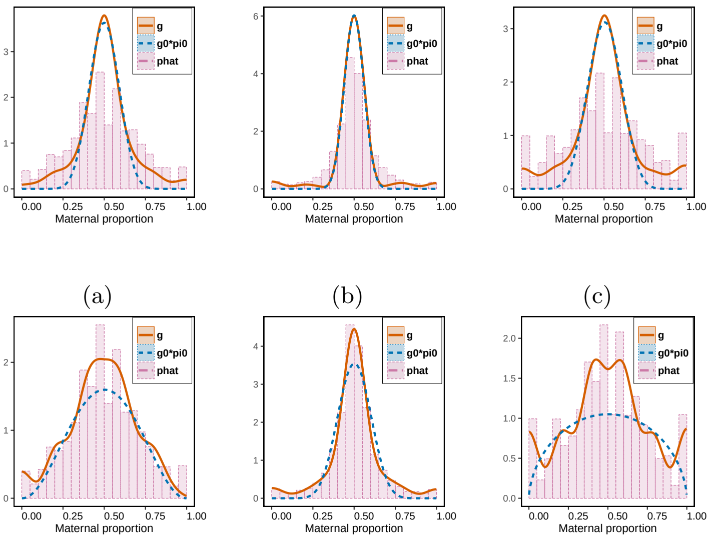
$(\pi_0, \alpha_0, d) = (0.8, 5, 0.4)$

Method	g L1	α_0 error	π_0 error	NS0.05	NS0.10	eFDR0.05	eFDR0.10
NPBin	1.8(0.31)	0.635(0.347)	-0.023(0.027)	436(57)	623(56)	0.063(0.013)	0.098(0.014)
OEMix	0.85(0.14)	-0.039(0.438)	0.27(0.014)	357(51)	533(60)	0.042(0.018)	0.071(0.027)
EBO	1(0.13)	-0.445(0.185)	0.047(0.006)	33(14)	309(27)	0.041(0.01)	0.068(0.013))
EBE	2.26(0.18)	-1.735(0.114)	0.140(0.005)	86(17)	265(25)	0.018(0.013)	0.028(0.008)
Binom	–	–	–	1426(46)	1756(62)	0.515(0.012)	0.565 (0.011)

$(\pi_0, \alpha_0, d) = (0.9, 20, 0.3)$

Method	g L1	α_0 error	π_0 error	NS 0.05	NS 0.10	eFDR 0.05	eFDR 0.10
NPBin	1.92(0.27)	-1.740(1.707)	0.035(0.013)	61(11)	103(18)	0.041(0.03)	0.068(0.035)
OEMix	1.78(0.44)	-1.554(1.271)	0.032(0.007)	62(12)	106(17)	0.045(0.033)	0.07(0.031)
EBO	1(0.17)	0.040(0.779)	0.005(0.012)	81(14)	139(18)	0.069(0.036)	0.114(0.028)
EBE	8.49(0.0.26)	-11.822(0.387)	0.027(0.010)	16(6)	39(8)	0.011(0.027)	0.020(0.023)
Binom	–	–	–	204(32)	319(41)	0.257(0.034)	0.359(0.031)

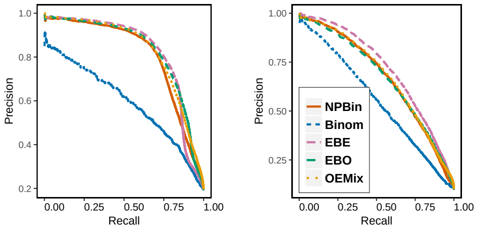
Table 2: Simulation results for two settings: Mean (Standard Error). The column names refer to the following. g L1: L1 loss in estimating g, normalized by the mean of the L1 loss of EBO; α_0 error: error in estimating α_0; π_0 error: error in estimating π_0, the proportion of the null; NS 0.05: number of loci selected for nominal FDR=0.05; NS 0.10: number of loci selected for nominal FDR=0.10; eFDR 0.05: empirical FDR when nominal FDR=0.05; eFDR 0.10: empirical FDR when nominal FDR=0.1



(a) (b) (c)

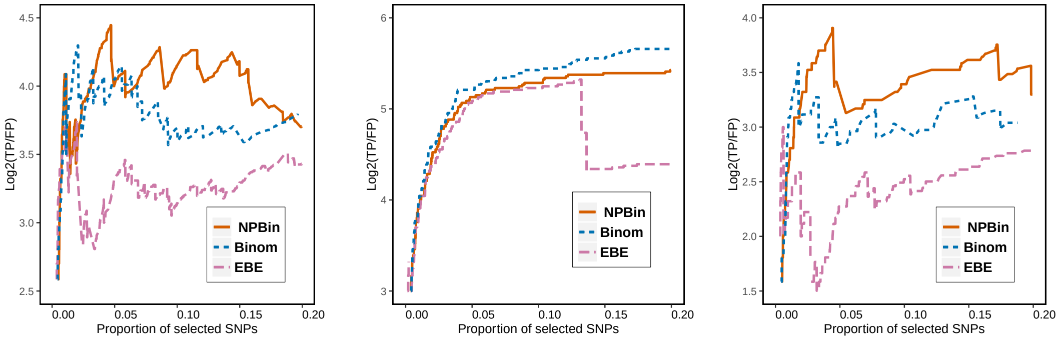
(d) (e) (f)

Fig. 3: Comparison of the estimated overall density g and the null density g_0 with the histogram of $\hat{\mathbf{p}}$. (a) ATAC-seq results by NPBin; (b) CTCF ChIP-seq results by NPBin; (c) DNase-seq results by NPBin; (d) ATAC-seq results by EBE; (e) CTCF ChIP-seq results by EBE; (f) DNase-seq results by EBE.



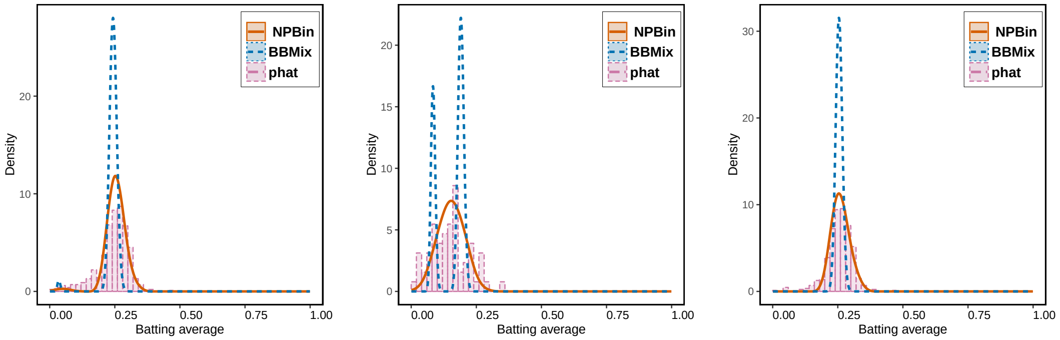
(a) (b)

Fig. 4: The precision-recall plot for two settings of (π_0, α_0, d). (a) (0.8,5,0.4); (b) (0.9,20,0.3)



(a) (b) (c)

Fig. 5: $\log_2$(TP/FP) for fixed proportions of selected SNPs. (a) ATAC-seq; (b) ChIP-seq of CTCF; (c) DNase-seq.



(a) (b) (c)

Fig. 6: Comparison of the estimated density of batting rates (p) with the histogram of $\hat{\mathbf{p}}$. (a) All players; (b) Pitchers; (c) Non-pitchers.

ATAC M=39,659 (987, 89)			
FDR	NPBin	EBE	Binomial
0.05	1,193 (78, 4)	56 (1, 1)	2,412 (129, 7))
0.1	1,827 (106, 4)	98 (6, 1)	3,609 (173, 12)
0.20	3,188 (147, 7)	195 (17, 1)	5,538 (242, 18)
CTCF M=19,782 (112, 5)			
FDR	NPBin	EBE	Binomial
0.05	1,376 (74, 1)	1,085 (67, 1)	2,417 (87, 1)
0.1	1,600 (74, 1)	1315 (72, 1)	2,959 (93, 1)
0.20	1,986 (77, 1)	1,773 (73, 1)	3,840 (100, 1)
DNase M=28,942 (206, 20)			
FDR	NPBin	EBE	Binomial
0.05	729 (22, 1)	0 (0, 0)	1,231 (31, 3)
0.1	1,355 (31, 2)	1 (0, 0)	1,839 (38, 4)
0.20	2,837 (41, 3)	27 (0, 0)	3,611 (57, 6)

Table 3: Numbers of detected Allelic-Imbalance SNPs with numbers of TP and FP ALI SNPs in parentheses at varying nominal FDR levels. The title for each panel indicates the total number of candidate SNPs (M) and the total numbers of potential TP and FP SNPs (when all M candidate SNPs are selected as ALI SNPs) in parentheses for each experiment.

	Overall	Pitchers	Non-pitchers
Training set size	567	81	486
Test set size	499	64	435
BBmix: TSE loss	0.588	0.156	0.314
NPBin: TSE loss	0.576	0.125	0.325
BBmix: L2 loss	1.749	0.335	1.021
NPBin: L2 loss	1.734	0.324	1.033

Table 4: Comparison of BBmix and NPmin on MLB batting average prediction based on loss in effect size estimation.