The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3528–3548
https://doi.org/10.1214/24-AOAS1950

SCALABLE TEST OF STATISTICAL SIGNIFICANCE FOR PROTEIN-DNA BINDING CHANGES WITH INSERTION AND DELETION OF BASES IN THE GENOME

BY QINYI ZHOU[1,a], CHANDLER ZUO[1,b], YUANNYU ZHANG[2,d], MIN CHEN[1,c], JIAN XU[2,e] AND SUNYOUNG SHIN[3,f]

[1]*Department of Mathematical Sciences, University of Texas at Dallas,* [a]*zhouqy0531@aliyun.com,* [b]*chandler.c.zuo@gmail.com,* [c]*yuannyu.zhang@stjude.org*

[2]*Center of Excellence for Leukemia Studies, Department of Pathology, St. Jude Children's Research Hospital,* [d]*mchen@utdallas.edu,* [e]*jian.xu@stjude.org*

[3]*Department of Mathematics, Pohang University of Science and Technology,* [f]*sunyoungshin@postech.ac.kr*

Mutations in the noncoding DNA, which represents approximately 99% of the human genome, have been crucial to understanding disease mechanisms through dysregulation of disease-associated genes. One key element in gene regulation that noncoding mutations mediate is the binding of proteins to DNA sequences. Insertion and deletion of bases (InDels) are the second most common type of mutations, following single nucleotide polymorphisms, that may impact protein-DNA binding. However, no existing methods can estimate and test the effects of InDels on the process of protein-DNA binding. We develop a novel test of statistical significance, namely, the binding change test (BC test), using a Markov model to evaluate the impact and identify InDels altering protein-DNA binding. The test predicts binding changer InDels of regulatory significance with an efficient importance sampling algorithm generating background sequences in favor of large binding affinity changes. Simulation studies demonstrate its excellent performance. The application to human leukemia data uncovers, in critical cis-regulatory elements, candidate pathological InDels on modulating TF binding in leukemic patients. We develop an R package atIndel, which is available on GitHub.

1. Introduction. An increasing number of disease-associated polymorphisms and mutations have been identified in the human genome with the advances in high-throughput sequencing technologies. While mutations in protein-coding regions, due to their direct effects on gene expression or cellular functions, have been well studied, it remains challenging to elucidate the pathologic roles of genetic mutations outside of the coding regions. The transcription of disease-associated genes may be regulated by noncoding mutations through regulatory elements such as binding of proteins to DNA sequences, DNA methylation, histone modification, and chromatin structure (The ENCODE Project Consortium et al. (2020a)). Several large research consortia, such as The Encyclopedia of DNA Elements (ENCODE) and Roadmap Epigenomics Mapping Consortium, have devoted significant efforts to generating regulatory genomic data and discovering regulatory roles of noncoding genomic sequences in gene transcription and translation and, in general, their biological functions (The ENCODE Project Consortium (2012), The ENCODE Project Consortium et al. (2020b), Fredriksson et al. (2014), Kundaje et al. (2015)).

Various genome annotation tools leveraging large scale genetic data have been developed for prioritization of functional effects of noncoding genetic variants. ANNOVAR (ANNOtate VARiation) calculates functional importance scores and finds variants in conserved elements

Received March 2024; revised August 2024.
Key words and phrases. Noncoding mutations, p-value based test statistic, importance sampling, sequence-based models, transcription factor binding, test of significance.

(Wang, Li and Hakonarson (2010)). The Ensembl Variant Effect Predictor (VEP) predicts the effects of genetic variants on functional status of resultant proteins and reports variants in noncoding RNAs and regulatory regions (McLaren et al. (2016)). SnpEff annotates variants based on their genomic locations and predicts coding effects (Cingolani et al. (2012)).

The modification caused by genetic variants in the binding of transcription factors (TFs) to DNA targets may determine regulatory consequences of the variants (Jensen and Liu (2008), Li and Zhang (2010), Li et al. (2020)). A p-value-based scoring system is established to evaluate the computational identification of TF binding sites (Huang et al. (2004)). Computational prediction of changes in protein-DNA binding, owing to mutations, has been facilitated by a growing number of libraries of TF-binding motifs, which are DNA sequence patterns that specifically bind to TFs (Jensen and Liu (2008), Li and Zhang (2010)). A TF-binding motif is mathematically expressed with a position probability matrix (PPM) in which each column contains the probabilities of the four nucleotides in the corresponding position of the motif (Stormo et al. (1982)). In silico, motif-based prediction is extremely useful for screening TFs with regulatory power from a large number of candidate TFs since in vitro experiments, such as ChIP-Seq (chromatin immunoprecipitation followed by sequencing) and ChIP-Exo, would be prohibitively expensive if used for genome-wide scans over a large number of TFs.

Insertions and deletions (InDels), characterized by addition and deletion of nucleotides in the genomic DNA, are the second most common type of genetic variations in the human genome, only next to single nucleotide polymorphisms (SNP) (Mills et al. (2006), Mullaney et al. (2010)). InDels may account for 15–20% of human polymorphisms (Mullaney et al. (2010)) and, in particular, those in noncoding regions have been implicated in many complex diseases (Sehn (2015)). InDels may have a substantial impact on the gene regulation and downstream biological processes if they alter the landscape of protein-DNA binding (Lin et al. (2017)). Therefore, it is crucial to catalogue noncoding InDels that may affect transcriptional regulation through computational prediction and assessment of disruption of an existing or creation of a novel protein-DNA binding.

However, there have been no statistical testing approaches to evaluate the effects of InDels in switching on and off a TF binding site. Existing computational and statistical methods, including SNP-MAPPER (Riva (2012)), motifbreakR (Coetzee, Coetzee and Hazelett (2015)), is-rSNP (Macintyre et al. (2010)), and atSNP (Zuo, Shin and Keleş (2015)), are all designed specifically for detecting SNPs, not suitable for InDels. Unlike SNPs, InDels change the length of the mutant DNA sequence. An insertion produces a longer sequence than the original one, while a deletion generates a shorter sequence.

We develop a novel test, namely, protein-DNA binding change test (BC test), to identify InDels that can cause the switch between on and off status of a protein-DNA binding. BC test first performs two statistical testings, one for the longer sequence and another for the shorter one, under the null hypothesis that they do not change the TF-binding affinity. To establish the null distributions for calculating empirical p-values, we employ a Markov model to generate background DNA sequences. Then it computes a binding-change score from the two p-values to compare the difference in the binding affinity between the original and mutant sequences. Markov models have been found to successfully predict DNA sequences (Avery (1987), Cowan (1991)) and thus have been used in computational algorithms to predict TF bindings (Zhou and Liu (2004), Bailey et al. (2006)).

To address the computational challenges, we further propose an importance sampling scheme to accelerate the calculation of empirical p-values. The importance sampling has been quite useful in Monte Carlo sampling for reducing computational cost (Liang (2002), Gupta and Ibrahim (2007), Chan and Zhang (2007), Chan, Zhang and Chen (2010), Cappello and Palacios (2020), Retkute et al. (2021), Li, Ko and Byon (2021)). We derive a general importance distribution that efficiently generates random sequence pairs for any given InDels

and binding motifs. The tilting parameter for the importance distribution is estimated such that the algorithm generates many sequences of large binding change scores. The algorithm can predict and prioritize the regulatory consequences of millions of InDels with thousands of binding motifs.

The rest of the paper is organized as follows. Section 2 presents a brief review on atSNP framework. In Section 3 we propose the methodology of the BC test for InDels along with Markov chain modelling for the null model. Section 4 develops the algorithm for computing p-values of the BC test based on the importance sampling method. In Section 5 we demonstrate statistical validity of the BC test with simulations under the null model and rigorously selected alternative models. Section 6 shows computational cost of the proposed BC test. Section 7 presents large-scale data analysis of primary leukemia samples with the BC test (Li et al. (2020)). Section 8 concludes with future directions and remarks.

2. TF binding affinity test for SNP. atSNP (Zuo, Shin and Keleş (2015)) considers a given motif of L positions and a reference sequence of length $2L-1$, denoted as $\mathbf{x} = (x_1, x_2, \ldots, x_{2L-1})$, where $x_i \in \{A, C, G, T\}$, $i = 1, \ldots, 2L-1$. For the remainder of the paper, we simply use $\{1, 2, 3, 4\}$, instead of the four nucleotides, $\{A, C, G, T\}$. The nucleotide in the middle, x_L, is the SNP position, where the nucleotide is substituted. Under the null hypothesis $\mathcal{H}_0$, $\mathbf{x}$ follows an irreducible first-order Markov model with prior probabilities $\pi_0(k) = P(y_l = k)$, $k = 1, \ldots, 4$, and transition probabilities $a_0(k, n) = P(y_{l+1} = n | y_l = k)$, $k, n = 1, \ldots, 4$

$$f_{\mathcal{H}_0}(\mathbf{x}) = \pi_0(x_1) \prod_{l=1}^{2L-2} a_0(x_l, x_{l+1}). \tag{1}$$

The SNP sequence, denoted by $\mathbf{x}^a = (x_1, \ldots, x_{L-1}, x_L^a, x_{L+1}, \ldots, x_{2L-1})$, is different from the reference sequence $\mathbf{x}$ only by x_L. The probabilities of having one of three nucleotides other than x_L as a substituted nucleotide are assumed to be equal each other, and consequently, the joint distribution of $\mathbf{x}$ and x_L^a under the null hypothesis is given by

$$f_{\mathcal{H}_0}(\mathbf{x}, x_L^a) = \frac{1\{x_L^a \neq x_L\}}{3} \pi_0(x_1) \prod_{l=1}^{2L-2} a_0(x_l, x_{l+1}).$$

The null sequences are specified jointly by the independent multinomial distribution and Markov chain model.

Denote the PPM of the motif by a $4 \times L$ matrix W, each column of which contains the four probabilities $W(\cdot, l)$, $l = 1 \ldots, L$ such that $\sum_{k=1}^{4} W(k, l) = 1$, $\forall l$. Zuo, Shin and Keleş (2015) define the binding score for a subsequence of the sequence $\mathbf{x}$, which starts at position s with a fixed length of L,

$$C(\mathbf{x}, s) = \sum_{l=1}^{L} \log W(x_{l+s-1}, l), \tag{2}$$

where $s \in \{1, \ldots, L\}$ is the protein binding start position. The PPM W of the motif is used as a scoring scheme for any input subsequences of length L. The binding score of the sequence $\mathbf{x}$ is defined as

$$C(\mathbf{y}) = \max_{s \in \{1, \ldots, L\}} \{C(T(\mathbf{y}), s) : T \in \{I, R\}\}, \tag{3}$$

where I and R are two strand operators with $I(\mathbf{y}) = \mathbf{y}$ and $R(\mathbf{y}) = (5 - y_{2L-1}, 5 - y_{2L-2}, \ldots, 5 - y_1)$, which is the reverse complement sequence. We have $C(\mathbf{y})$ as the maximum binding score from all forward and reverse subsequences with the L start positions.

Similarly to $\mathbf{y}$, the binding score of the mutated sequence $\mathbf{y^a}$ is defined as

$$C^a(\mathbf{y^a}) = \max_{s\in\{1,\dots,L\}} \{C(T(\mathbf{y^a}), s) : T \in \{I, R\}\}. \tag{4}$$

More details on the testing procedure and implementation can be found in the Supplementary Material of Zuo, Shin and Keleş (2015).

3. TF binding change test for InDels. The protein-DNA binding change test for InDels (BC test) detects and quantifies possible binding changes caused by InDels. The null hypothesis is there is no binding change with the InDel.

3.1. *Markov chain background model.* InDel mutations may promote or impair TF binding to the DNA sequence. The proposed BC test evaluates the influence of InDel mutations on binding of a TF to DNA using the corresponding motif and the pair of sequences. We compare protein bindings to the reference and the InDel mutated sequences, the lengths of which are different. Given an insertion mutation, the shorter sequence is the reference sequence, and the longer one is the mutated sequence consisted of both the reference and inserted nucleotides. Given a deletion mutation, the longer sequence is the reference sequence, and the shorter one is the mutated sequence constructed by removing some nucleotides from the reference. In both cases the key component is the stretch of added or deleted nucleotides, referred to as a contrasting sequence hereafter. Without loss of generality, consider the PPM of a given motif of L positions, W, as in Section 2, and a contrasting sequence of m nucleotides in the forward strand, denoted by $\mathbf{y^c} = (y_L, \dots, y_{L+m-1})$, where $y_i \in \{1, 2, 3, 4\}$, $i = L, \dots, L+m-1$. Here m and L are fixed numbers. Given m and L, the lengths of the two sequences are set to be $2L+m-2$ and $2L-2$, respectively, such that a potential binding site covers at least part of the contrasting sequence, and hence the InDels may cause a change in the binding affinity. Let $\mathbf{y} = (y_1, \dots, y_{L-1}, y_L, y_{L+1}, \dots, y_{L+m-1}, y_{L+m}, \dots, y_{2L+m-2})$ denote the longer sequence in the forward strand, where $y_i \in \{1, 2, 3, 4\}$, $i = 1, \dots, 2L+m-2$. Let $\mathbf{y^a} = (y_1, \dots, y_{L-1}, y_{L+m}, \cdots, y_{2L+m-2})$ denote the shorter sequence in the forward strand. With the contrasting sequence $\mathbf{y^c}$, $\mathbf{y} = (\mathbf{y^a}_1, \mathbf{y^c}, \mathbf{y^a}_2)$, and $\mathbf{y^a} = (\mathbf{y^a}_1, \mathbf{y^a}_2)$, where $\mathbf{y^a}_1 = (y_1, \dots, y_{L-1})$ and $\mathbf{y^a}_2 = (y_{L+m}, \dots, y_{2L+m-2})$.

The main idea of the BC test is to assess the difference of the TF binding affinity between the longer sequence $\mathbf{y}$ and the shorter sequence $\mathbf{y^a}$. It is sufficient to model the longer sequence $\mathbf{y} = (\mathbf{y^a}_1, \mathbf{y^c}, \mathbf{y^a}_2)$ since the shorter sequence $\mathbf{y^a} = (\mathbf{y^a}_1, \mathbf{y^a}_2)$ is just a subvector of $\mathbf{y}$. The null model for the longer sequences is an irreducible first-order Markov model with prior probabilities $\pi_0(k) = P(y_l = k)$, $k = 1, \dots, 4$, and transition probabilities $a_0(k, n) = P(y_{l+1} = n | y_l = k)$, $k, n = 1, \dots, 4$,

$$f_{\mathcal{H}_0}(\mathbf{y}) = \pi_0(y_1) \prod_{l\in\{1,\dots,2L+m-3\}} a_0(y_l, y_{l+1}). \tag{5}$$

Background DNA sequences are usually fitted by the Markov chain model as the occurrence of a nucleotide at a given position depends on the previous nucleotides in the sequence (Avery and Henderson (1999), Menéndez et al. (2011), Reinert, Schbath and Waterman (2000)). Using the Markov chain for modelling the background sequences was successfully supported by comprehensive numerical experiments (Huang et al. (2004), Menéndez et al. (2011)).

3.2. *Test statistic and p-value.* We define the binding score for a subsequence of the longer sequence $\mathbf{y}$, which starts at position s with a fixed length of L,

$$C(\mathbf{y}, s) = \sum_{l=1}^{L} \log W(y_{l+s-1}, l), \tag{6}$$

where $s \in \{1, \ldots, L+m-1\}$ is the protein binding start position. The binding score of the sequence $\mathbf{y}$ is defined as

$$C(\mathbf{y}) = \max_{s\in\{1,\ldots,L+m-1\}} \{C(T(\mathbf{y}), s) : T \in \{I, R\}\}, \tag{7}$$

where I and R are two strand operators that are defined in Section 2. The binding score of $\mathbf{y}$, $C(\mathbf{y})$, is the maximum binding score of the forward and reverse subsequences with the $(L+m-1)$ start positions. Similarly to $\mathbf{y}$, the binding score of the shorter sequence $\mathbf{y^a}$ is

$$C^a(\mathbf{y^a}) = \max_{s\in\{1,\ldots,L-1\}} \{C(T(\mathbf{y^a}), s) : T \in \{I, R\}\}. \tag{8}$$

We use the maximum value of the $2\cdot(L-1)$ subsequence binding scores as the binding score of $\mathbf{y^a}$. Here $\mathbf{y}$ has m more subsequences than $\mathbf{y^a}$, unlike SNPs where the lengths of both $\mathbf{y}$ and $\mathbf{y^a}$ are equal.

The binding changes, due to InDels, are tested by comparing binding significance on $\mathbf{y}$ to that on $\mathbf{y^a}$. Given a longer sequence $\mathbf{y_0}$, the TF binding p-value is obtained as the probability that binding score of a sequence from the null model is at least as large as $C(\mathbf{y_0})$,

$$p_l(\mathbf{y_0}) = P\{C(\mathbf{y}) \geq C(\mathbf{y_0}) | \mathbf{y} \sim f_{\mathcal{H}_0}\}. \tag{9}$$

Similarly, for a shorter sequence $\mathbf{y_0^a}$, we obtain the p-value based on the observed binding score $C^a(\mathbf{y_0^a})$ as

$$p_s(\mathbf{y_0^a}) = P\{C^a(\mathbf{y^a}) \geq C^a(\mathbf{y_0^a}) | \mathbf{y} \sim f_{\mathcal{H}_0}\}. \tag{10}$$

The binding significance of a sequence is free of the sequence length; therefore, it is viable to make direct comparison between $p_l(\mathbf{y_0})$ and $p_s(\mathbf{y_0^a})$.

Our BC test statistic for the pair $(\mathbf{y}, \mathbf{y^a})$, named "binding change score," is the difference between the logarithm of the binding p-values to $\mathbf{y}$ and $\mathbf{y^a}$,

$$T \equiv T(\mathbf{y}, \mathbf{y^a}) = \log\{p_s(\mathbf{y^a})\} - \log\{p_l(\mathbf{y})\}, \tag{11}$$

where $p_l(\cdot)$ and $p_s(\cdot)$ are in (9) and (10), respectively. Under the null hypothesis that there is no biding change, the binding p-values are comparable regardless of the difference between $\mathbf{y}$ and $\mathbf{y^a}$. The binding p-values for longer and shorter sequences, $p_l(\mathbf{y_0})$ and $p_s(\mathbf{y_0^a})$, are evaluated under $f_{\mathcal{H}_0}$. The magnitude of T is expected to be large if there is a large difference between the binding affinity between $\mathbf{y}$ and $\mathbf{y^a}$. We can determine whether or not the binding is enhanced or disrupted from the sign of the binding change score. For example, a positive test statistic value for insertion mutations corresponds to binding creation. On the other hand, the binding disruption due to insertion mutations, where $p_s(\mathbf{y^a})$ is smaller than $p_l(\mathbf{y})$, produces a negative test statistic value. The test statistic is computed based on p-values, which are also functions of the sample data. Such p-value-based test statistics have been considered in the literature, for example, the higher criticism test statistic, proposed by Donoho and Jin (2004), and its variation that Mukherjee, Pillai and Lin (2015) introduced for binary regression.

Similar to $(\mathbf{y}, \mathbf{y^a})$, in the observed sequence pair $(\mathbf{y_0}, \mathbf{y_0^a})$, $\mathbf{y_0^a}$ is a subsequence of $\mathbf{y_0}$. We define p-value to assess whether or not the change in the protein binding affinity with $(\mathbf{y_0}, \mathbf{y_0^a})$ is statistically different from what would be expected by chance under the null model,

$$p\text{-value} \equiv p(\mathbf{y_0}, \mathbf{y_0^a}) = 2\cdot\min\{P(T \geq t_0 | \mathbf{y} \sim f_{\mathcal{H}_0}), P(T \leq t_0 | \mathbf{y} \sim f_{\mathcal{H}_0})\}, \tag{12}$$

where $t_0 \equiv T(\mathbf{y_0}, \mathbf{y_0^a}) = \log\{p_s(\mathbf{y_0^a})\} - \log\{p_l(\mathbf{y_0})\}$. The p-value we obtain in (12) doubles the smaller tail probability to address the two-sided BC test. The null distribution of the binding change score T is obtained with sequence pairs from $f_{\mathcal{H}_0}$.

An alternative test to detect the binding change is binding score difference test whose test statistics is the naive score difference based on the difference between the binding scores normalized by the sequence lengths,

$$T_d \equiv T_d(\mathbf{y}, \mathbf{y^a}) = C(\mathbf{y})/(2L+m-2) - C^a(\mathbf{y^a})/(2L-2)\}, \tag{13}$$

where $C(\cdot)$ and $C^a(\cdot)$ are from (7)–(8). The binding score difference test assesses how far the observed binding score difference is distant from the expected score difference under the null model in a probabilistic sense. Similarly to the BC test, the random longer sequences are generated under the first-order Markov model while the corresponding shorter sequences are obtained by removing the contrasting sequences from the longer sequences. For a sequence pair $(\mathbf{y}, \mathbf{y^a})$, $C(\mathbf{y})$ is more likely to be larger than $C^a(\mathbf{y^a})$, since $C(\mathbf{y})$ is the maximum binding score evaluated on $\mathbf{y}$ that has more subsequences than $\mathbf{y^a}$. Thus, the scores are normalized by the corresponding sequence lengths. We obtain the p-value for the binding score difference from the best subsequence matches of both sequences as follows:

$$p_d(\mathbf{y_0}, \mathbf{y_0^a}) = \min\Bigg(2P\left\{\frac{C(\mathbf{y})}{2L+m-2} - \frac{C^a(\mathbf{y^a})}{2L-2} \geq \frac{C(\mathbf{y_0})}{2L+m-2} - \frac{C^a(\mathbf{y_0^a})}{2L-2}, \mathbf{y} \sim f_{\mathcal{H}_0}\right\},$$
$$2P\left\{\frac{C(\mathbf{y})}{2L+m-2} - \frac{C^a(\mathbf{y^a})}{2L-2} \leq \frac{C(\mathbf{y_0})}{2L+m-2} - \frac{C^a(\mathbf{y_0^a})}{2L-2}, \mathbf{y} \sim f_{\mathcal{H}_0}\right\}\Bigg).$$

In Sections 5.1–5.2, we conduct comprehensive numerical studies to compare the BC test to the binding score difference test.

4. General importance sampling algorithm. A practical obstacle of conducting the BC test is the theoretical calculation of the null distribution of the test statistic T for distinct values of m, L, and W that are determined by InDel mutations and motifs. We develop an efficient algorithm that computes empirical p-value of the BC test for any values of m, L, and W based on the importance sampling technique, requiring a much smaller number of pairs to be simulated. In the BC test, the null model rarely generates sequence pairs with large binding change scores. The importance sampling speeds up the Monte Carlo procedure by generating many random samples from the importance distribution that has higher density in the importance region of rare events (Kahn and Marshall (1953), Chan and Zhang (2007), Chan, Zhang and Chen (2010)). The technique approximates p-value, which can be expressed as an expectation with respect to the null distribution, with a weighted average for the random draws from the importance distribution (Chan, Zhang and Chen (2010), Zuo, Shin and Keleş (2015)). The algorithm that we have coded in `R` and `C++` is available in R package *atIndel* and is scalable for the BC tests on hundreds of thousands InDel mutations against thousands of binding motifs.

4.1. *Importance distribution.* We design the importance distribution for the BC test, denoted by $h_\theta(\mathbf{y})$, to sample sequence pairs with large score differences driven by contrasting sequences. The challenge for constructing the importance distribution is that m can be an arbitrary integer and the length of the overlapping part of the contrasting sequence with the protein binding site can be also arbitrary between 1 and m, which atSNP does not suffer from.

The p-value for the observed sequence pair $(\mathbf{y_0}, \mathbf{y_0^a})$ can be rewritten with $h_\theta(\mathbf{y})$,

$$\begin{aligned} p(\mathbf{y_0}, \mathbf{y_0^a}) &= 2 \cdot \min\big[E\big[1\{T \geq T(\mathbf{y_0}, \mathbf{y_0^a})\}|\mathbf{y} \sim f_{\mathcal{H}_0}(\mathbf{y})\big], \\ &\qquad E\big[1\{T \leq T(\mathbf{y_0}, \mathbf{y_0^a})\}|\mathbf{y} \sim f_{\mathcal{H}_0}(\mathbf{y})\big]\big] \\ &= 2 \cdot \min\left[E\left[1\{T \geq T(\mathbf{y_0}, \mathbf{y_0^a})\} \cdot \frac{f_{\mathcal{H}_0}(\mathbf{y})}{h_\theta(\mathbf{y})}|\mathbf{y} \sim h_\theta(\mathbf{y})\right],\right. \\ &\qquad \left. E\left[1\{T \leq T(\mathbf{y_0}, \mathbf{y_0^a})\} \cdot \frac{f_{\mathcal{H}_0}(\mathbf{y})}{h_\theta(\mathbf{y})}|\mathbf{y} \sim h_\theta(\mathbf{y})\right]\right], \end{aligned} \tag{14}$$

where $1(\cdot)$ is the indicator function.

We consider the conditional importance distribution of $\mathbf{y}$, given the protein binding start position s, $1 \le s \le L+m-1$, as follows:

$$
\begin{aligned}
h_\theta(\mathbf{y}|s) = {} & \frac{1}{M_s(\theta)} f(y_1,\ldots,y_{s-1})^{1(s\ge 2)} \Bigg[\prod_{l=s}^{L-1} IW(y_l, l-s+1)\Bigg]^{1(1\le s\le L-1)} \\
& \times \Bigg[\prod_{c=\max(L,s)}^{\min(s,m)+L-1} D(y_c, c-s+1)^\theta\Bigg] \\
& \times \Bigg[\prod_{l=L+m}^{L+s-1} IW(y_l, l-s+1)\Bigg]^{1(m+1\le s\le L+m-1)} \\
& \times f(y_{L+s},\ldots,y_{2L+m-2})^{1(s\le L+m-2)},
\end{aligned} \tag{15}
$$

where θ is a tilting parameter and $M_s(\theta)$ is the normalizing constant. Here $IW(\cdot,l) = \{W(\cdot,l)+1/4\}/2$ and the $4\times L$ matrix D has the following entries:

$$
D(k,l) = \exp\Bigg\{\sum_{j=1}^{4} \pi_0(j)\big(\log W(k,l) - \log W(j,l)\big)\Bigg\}.
$$

The function $f(\cdot)$ is an irreducible first-order Markov chain with parameters π_0 and a_0. The sequence $\mathbf{y}$ is modelled by three parts: a part of the protein binding site overlapping with the contrasting sequence with $D(\cdot,l)^\theta$, another part of the protein binding site outside of the contrasting sequence with $IW(\cdot,l)$, and the protein unbinding site with $f(\cdot)$. The intuition behind $D(k,l)$ is to compare the nucleotide k at the lth position of the binding subsequence of length L with a random nucleotide that follows a zero-order Markov chain with parameters π_0. In replacing the nucleotide k at the lth position with the random nucleotide, $\log D(k,l)$ is the expected score change from the random nucleotide to the nucleotide k. The part of the contrasting sequence overlapped with the protein binding site follows the weighted distribution $D(\cdot,l)^\theta$ so that the binding subsequence within the contrasting sequence generates a large expected change in binding scores. Next, the other part of the protein binding site is modelled to generate sequences following $IW(\cdot,l)$, a transformation of $W(\cdot,l)$, to prevent from always generating $\mathbf{y}$ exactly following the PPM. The conditional distribution $h_\theta(\mathbf{y}|s)$ encompasses all types of overlaps between the protein binding site and the longer sequence such as partial overlaps and complete overlaps, regardless of $m<L$ or $m\ge L$. An example of the overlaps is illustrated in Section 1.1 of the Supplementary Material (Zhou et al. (2024)).

By integrating $\mathbf{y}$ out on both sides of (15), we have an explicit form of $M_s(\theta)$,

$$
\begin{aligned}
M_s(\theta) &= \prod_{l=L+m-s+1}^{L}\Bigg\{\sum_{i=1}^{4} IW(i,l)\Bigg\}\prod_{l=1}^{L-s}\Bigg\{\sum_{i=1}^{4} IW(i,l)\Bigg\}\prod_{j=\max(L,s)-s+1}^{\min(s,m)+L-s}\Bigg\{\sum_{i=1}^{4} D(i,j)^\theta\Bigg\} \\
&= \prod_{j=\max(L,s)-s+1}^{\min(s,m)+L-s}\Bigg\{\sum_{i=1}^{4} D(i,j)^\theta\Bigg\}.
\end{aligned}
$$

We set the distribution of the binding start position s to be proportional to the corresponding normalizing constant,

$$
h_\theta(s) = \frac{M_s(\theta)}{M(\theta)}, \quad s = 1,\ldots,L+m-1,
$$

where $M(\theta) = \sum_{s=1}^{L+m-1} M_s(\theta)$.

Consequently, we obtain the importance joint distribution of $(\mathbf{y}, s)$,

$$
\begin{aligned}
h_\theta(\mathbf{y}, s) = {} & \frac{1}{M(\theta)} f(y_1, \ldots, y_{s-1})^{1(s\geq 2)} \left[\prod_{l=s}^{L-1} IW(y_l, l-s+1)\right]^{1(1\leq s\leq L-1)} \\
& \times \left[\prod_{c=\max(L,s)}^{\min(s,m)+L-1} D(y_c, c-s+1)^\theta\right] \\
& \times \left[\prod_{l=L+m}^{L+s-1} IW(y_l, l-s+1)\right]^{1(m+1\leq s\leq L+m-1)} \\
& \times f(y_{L+s}, \ldots, y_{2L+m-2})^{1(s\leq L+m-2)}.
\end{aligned} \tag{16}
$$

By taking sum of the joint distribution over s from 1 to $L+m-1$, we have the following importance distribution of $\mathbf{y}$:

$$
h_\theta(\mathbf{y}) = \sum_{s=1}^{L+m-1} h_\theta(\mathbf{y}, s). \tag{17}
$$

Further, we have the importance distribution of $\mathbf{y^a}$ by marginalization,

$$
\begin{aligned}
h_\theta^a(\mathbf{y^a}) = {} & \sum_{\{y_L, \ldots, y_{L+m-1}\}\in\{1,2,3,4\}^m} h_\theta(\mathbf{y}) \\
= {} & \sum_{s=1}^{L+m-1} \sum_{\{y_L, \ldots, y_{L+m-1}\}\in\{1,2,3,4\}^m} h_\theta(\mathbf{y}|s) h_\theta(s) \\
= {} & \sum_{s=1}^{L+m-1} \frac{1}{M(\theta)} f(y_1, \ldots, y_{\min(s-1,L-1)})^{1(s\geq 2)} \\
& \times \left[\prod_{l=s}^{L-1} [IW(y_l, l-s+1)^{1(1\leq s\leq L-1)}]\right] \\
& \times \left[\prod_{c=\max(L,s)}^{\min(s,m)+L-1} \left[\sum_{i=1}^{4} D(i, c-s+1)^\theta\right]\right] \\
& \times \left[\prod_{l=L+m}^{L+s-1} [IW(y_l, l-s+1)^{1(m+1\leq s\leq L+m-1)}]\right] \\
& \times f(y_{\max(L+s,L+m)} \cdots, y_{2L+m-2})^{1(s\leq L+m-2)}.
\end{aligned}
$$

Our algorithm for the BC test employs the importance distribution in (17) for computing p-value of $(\mathbf{y_0}, \mathbf{y_0^a})$ in (14).

4.2. *Choice of tilting parameter.* We now discuss the estimation of the tilting parameter θ that determines the importance distribution. It is desirable to choose the tilting parameter for each BC test for an InDel mutation against a given motif such that the importance distribution has a suitable weight on the importance region. With the computed tilting parameter, we obtain random sequence pairs from the importance distribution and estimate the p-value.

To estimate the tilting parameter θ, we set the expected difference in binding scores between subsequences of length L under the importance distribution and background subsequences equal to the observed binding score change. Proposition 1 computes the expected

score change between the binding subsequence and the random sequence of length L from the zero-order Markov model with π_0.

PROPOSITION 1. *Suppose that a random vector of the sequence and the binding start position* $(\mathbf{y}, s)$ *follows the importance distribution* $h_\theta(\mathbf{y}, s)$ *in* (16). *Further, suppose that* $\mathbf{y}' = (y'_s, \ldots, y'_{s+L-1})$ *is independent of* $(\mathbf{y}, s)$ *and follows zero-order Markov model with* $\pi(\cdot)$. *The expected binding score difference between the binding subsequence* $(y_s, \ldots, y_{s+L-1})$ *and the subsequence* $\mathbf{y}'$ *is as follows*:

$$
\begin{aligned}
&E_{\mathbf{y},\mathbf{y}',s}\left\{\sum_{j=1}^{L}\big(\log W(y_{j+s-1}, j) - \log W\big(y'_{j+s-1}, j\big)\big)\right\} \\
&\quad= \sum_{s=1}^{L+m-1} \frac{M_s(\theta)}{M(\theta)} \\
&\quad\quad\times\left[\sum_{i<\max(L,s) \text{ or } i\geq\min(s,m+L)}\left\{\sum_{k=1}^{4}\big(IW(k, i-s+1) - \pi(k)\big)\log W(k, i-s+1)\right\}\right. \\
&\quad\quad\left.+\sum_{i=\max(s,L)}^{\min(s,m+L-1)}\frac{\sum_k D(k, i-s+1)^\theta \log D(k, i-s+1)}{\sum_k D(k, i-s+1)^\theta}\right].
\end{aligned}
\tag{18}
$$

The proof is in Section 1.2 of the Supplementary Material (Zhou et al. (2024)). We consider changing each nucleotide of the binding subsequence to the random nucleotide y'_i. From the result we set the observed binding score change $C(\mathbf{y_0}) - C^a(\mathbf{y_0^a})$ to be equal to the expected score difference and solve for θ to acquire the optimal tilting parameter. Under the importance sampling distribution with the optimal tilting parameter, it is not rare to obtain the observed score difference.

4.3. *Importance sampling Monte Carlo algorithm.* We sample the nucleotides of a random sequence following the joint importance distribution $h_\theta(\mathbf{y}, s)$ in (16) in a sequential manner. The start position is first simulated from its marginal importance distribution in (4.1), and the nucleotides are simulated one by one as follows:

$$
h_\theta(y_l | 2 \leq s \leq L+m-2) = \pi_0(y_l), \quad \text{for } l = 1, L+s, \tag{19}
$$

$$
\begin{aligned}
&h_\theta(y_l | y_{l-1}, 2 \leq s \leq L+m-2) = a_0(y_{l-1}, y_l), \\
&\quad \text{for } l = 2, \ldots, s-1, L+s+1, \ldots, 2L+m-2,
\end{aligned}
\tag{20}
$$

$$
h_\theta(y_l | 1 \leq s \leq L-1) = IW(y_l, l-s+1), \quad \text{for } l = s, \ldots, L-1 \tag{21}
$$

$$
\begin{aligned}
&h_\theta(y_l | m+1 \leq s \leq L+m-1) = IW(y_l, l-s+1), \\
&\quad \text{for } l = L+m, \ldots, L+s-1
\end{aligned}
\tag{22}
$$

$$
h_\theta(y_l) = \frac{D(y_l, l-s+1)^\theta}{\sum_{i=1}^{4} D(i, l-s+1)^\theta}, \quad \text{for } l = \max(L,s), \ldots, \min(s,m)+L-1. \tag{23}
$$

The key to generating the random sequence pairs is to decompose the conditional distribution (15) into (19)–(23). We construct the distribution for the complement of the binding subsequences, that is, $f(y_1, \ldots, y_{s-1})$ and $f(y_{L+s}, \ldots, y_{2L+m-2})$ with (19) and (20). The product of the probabilities in (21) simply forms $\prod_{l=s}^{L-1}[IW(y_l, l-s+1)^{1(1\leq s\leq L-1)}]$ and so does the probability product in (22). Similarly, the probability product in (23) is $\prod_{c=\max(L,s)}^{\min(s,m)+L-1} D(y_c, c-s+1)^\theta$. Thus, according to the sampling procedure, we can simulate random sequences from the importance distribution.

4.4. *Empirical p-value for binding change test.* With the importance sampling algorithm, we generate N independent random longer sequences $\{\mathbb{Y}_t : t = 1, \ldots, N\}$, under $h_\theta(\mathbf{y})$, and obtain corresponding shorter sequences $\{\mathbb{Y}_t^a : t = 1, \ldots, N\}$ by removing the contrasting sequences. Conventionally, the empirical p-value for a random longer sequence $\mathbb{Y}_t$ is given as $\frac{1}{N}\sum_{s=1}^{N} 1\{C(\mathbb{Y}_s) \geq C(\mathbb{Y}_t)\}\frac{f_{\mathcal{H}_0}(\mathbb{Y}_s)}{h_\theta(\mathbb{Y}_s)}$. However, the naive estimator is biased due to the fact that $\mathbb{Y}_t \in \{\mathbb{Y}_1, \ldots, \mathbb{Y}_N\}$,

$$E\left[\frac{1}{N}\sum_{s\neq t} 1\{C(\mathbb{Y}_s) \geq C(\mathbb{Y}_t)\}\frac{f_{\mathcal{H}_0}(\mathbb{Y}_s)}{h_\theta(\mathbb{Y}_s)}\right] + \frac{f_{\mathcal{H}_0}(\mathbb{Y}_t)}{Nh_\theta(\mathbb{Y}_t)} = \frac{N-1}{N}p_l(\mathbb{Y}_t) + \frac{f_{\mathcal{H}_0}(\mathbb{Y}_t)}{Nh_\theta(\mathbb{Y}_t)}.$$

To remove the finite-sample bias, we use the empirical estimator of p_l based on all but $\mathbb{Y}_t$ as follows:

$$\widehat{p_l}(\mathbb{Y}_t) = \frac{1}{N-1}\sum_{s\neq t} 1\{C(\mathbb{Y}_s) \geq C(\mathbb{Y}_t)\}\frac{f_{\mathcal{H}_0}(\mathbb{Y}_s)}{h_\theta(\mathbb{Y}_s)}.$$

The distribution of the empirical p-values has been found to follow the uniform distribution well. In a similar manner, the empirical p-value for the corresponding shorter sequence Y_t^a is adjusted,

$$\widehat{p_s}(\mathbb{Y}_t^a) = \frac{1}{N-1}\sum_{s\neq t} 1\{C(\mathbb{Y}_s^a) \geq C(\mathbb{Y}_t^a)\}\frac{\tilde{f}_{\mathcal{H}_0}^a(\mathbb{Y}_s^a)}{h_\theta^a(\mathbb{Y}_s^a)},$$

where $\tilde{f}_{\mathcal{H}_0}^a$ is chosen to be an irreducible first-order Markov model with π_0 and a_0,

$$\tilde{f}_{\mathcal{H}_0}^a(\mathbf{y^a}) = \pi_0(y_1) \prod_{l\in\{1,\ldots,L-2,L+m,\ldots,2L+m-3\}} a_0(y_l, y_{l+1}) \cdot a_0(y_{L-1}, y_{L+m}). \tag{24}$$

It is computationally expensive to obtain an output value of the marginal probability density function for $\mathbb{Y}_s^a$ from the null model, $f_{\mathcal{H}_0}^a$. To lower computational burden, we use a simpler first-order Markov model for the shorter sequence $\tilde{f}_{\mathcal{H}_0}^a(\mathbf{y^a})$ in (24). We compute the empirical p-values of the observed sequences $(\mathbf{y_0}, \mathbf{y_0^a})$ that are unbiased as follows:

$$\widehat{p_l}(\mathbf{y_0}) = \frac{1}{N}\sum_{s=1}^{N} 1\{C(\mathbb{Y}_s) \geq C(\mathbf{y_0})\}\frac{f_{\mathcal{H}_0}(\mathbb{Y}_s)}{h_\theta(\mathbb{Y}_s)}, \tag{25}$$

$$\widehat{p_s}(\mathbf{y_0^a}) = \frac{1}{N}\sum_{s=1}^{N} 1\{C^a(\mathbb{Y}_s^a) \geq C^a(\mathbf{y_0^a})\}\frac{\tilde{f}_{\mathcal{H}_0}^a(\mathbb{Y}_s^a)}{h_\theta^a(\mathbb{Y}_s^a)}. \tag{26}$$

The total computational complexity for (25) and (26) is $N(L+m)(3L+m)(1+o(1))$. Detailed analysis on the complexity is found in Section 1.3 of the Supplementary Material (Zhou et al. (2024)).

Now, we have N binding change test statistics corresponding to the simulated sequences,

$$\mathbb{T}_t \equiv \log\{\widehat{p_s}(\mathbb{Y}_t^a)\} - \log\{\widehat{p_l}(\mathbb{Y}_t)\}.$$

The observed test statistic for $(\mathbf{y_0}, \mathbf{y_0^a})$ is given as

$$T_0 \equiv \log\{\widehat{p_s}(\mathbf{y_0^a})\} - \log\{\widehat{p_l}(\mathbf{y_0})\}.$$

In conclusion, with the importance samples $\{\mathbb{Y}_t\}$ and $\{\mathbb{Y}_t^a\}$, we obtain our target p-value of the observed sequence pair $(\mathbf{y_0}, \mathbf{y_0^a})$ as follows:

$$\widehat{p}(\mathbf{y_0}, \mathbf{y_0^a}) = 2 \cdot \min\left[\frac{1}{N}\sum_{t=1}^{N} 1\{\mathbb{T}_t \geq T_0\}\frac{f_{\mathcal{H}_0}(\mathbb{Y}_t)}{h_\theta(\mathbb{Y}_t)}, \frac{1}{N}\sum_{t=1}^{N} 1\{\mathbb{T}_t \leq T_0\}\frac{\tilde{f}_{\mathcal{H}_0}(\mathbb{Y}_t)}{h_\theta(\mathbb{Y}_t)}\right].$$

In practice, the sequence sample size between 2000 and 10,000 shows a quite satisfactory performance.

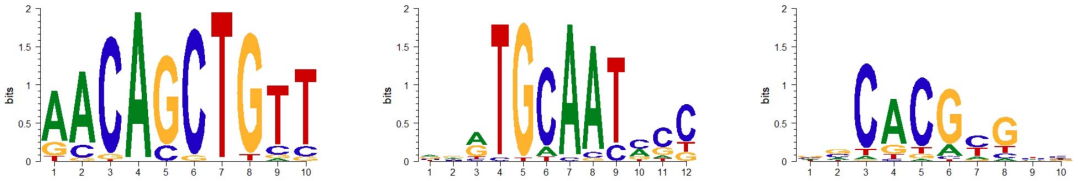


FIG. 1. *Three JASPAR motifs used for simulation studies.*

5. Simulation studies. We investigate the performance of the BC test with InDels that are simulated from the background and alternative sequence pair models. From JASPAR 2018 vertebrate motif library (Portales-Casamar et al. (2010), Khan et al. (2017)), we select three motifs with various levels of information content (IC). IC measures the motif conservation; higher IC corresponds to greater certainty of the binding sequences. (Hertz and Stormo (1999)). The median IC values of the JASPAR motifs vary from 0.133 to 2, and their mean IC values are from 0.492 to 1.857. Figure 1 shows the selected motifs MSC, Ddit3::Cebpa, and Hes1. Denote the motif lengths by $L_{\text{MSC}} = 10$, $L_{\text{DC}} = 12$, and $L_{\text{Hes1}} = 10$. The median IC values of the three motifs are 1.522, 0.995, 0.458, and their mean IC values are 1.486, 0.971, 0.543, respectively. For each of the given motifs, binding affinity changes driven by the InDels are evaluated with the BC test.

We set the null model parameters of the BC test π_0 and a_0 for $k, n = 1, \ldots, 4$ based on the human reference genome version GRCh37 (hg19). We first randomly selected 100,000 sequences of length 100 from the reference genome excluding each chromosome's first and last 10,000 bases that contain unknown bases. Next, we fitted the sequences with a first-order Markov model and obtained the model parameters in Tables S1–S2 in Section 2.1 of the Supplementary Material (Zhou et al. (2024)). Since there is no existing statistical test or scoring scheme for binding affinity changes with InDels, for the comparison we use the score difference test illustrated in Section 3.2. Both test algorithms are available in R package *atIndel*.

5.1. *Simulations under the null model.* We examine if our method successfully controls the Type I error under the null model. We generated a random sample of 10,000 longer sequences from the first-order Markov model with the computed model parameters. The length of the sample longer sequences we chose is 28 and the contrasting sequence length is 6 ($L = 12$, $m = 6$) such that the sample shorter sequences were obtained by dropping the middle six bases from the longer sequences. Empirical rejection probabilities were calculated at different levels of significance $p = 0.01$, 0.05, and 0.1. The Monte Carlo sample size for the importance algorithm was set as 2000.

TABLE 1
Empirical rejection probabilities of the binding change test

	Empirical rejection probability		
p-value	MSC	Ddit3::Cebpa	Hes1
0.01	0.0138	0.0176	0.0094
0.05	0.0527	0.0684	0.0502
0.10	0.1038	0.1164	0.1008

FIG. 2. *Q-Q plots of p-values from binding change tests under the null model.*

In Table 1 the empirical rejection probabilities of the BC test are around the nominal significance level. For Ddit::Cebpa, the p-value approximation tends to overestimate, which may be attenuated with a larger Monte Carlo sample size. The QQ plots in Figure 2 show the distributions of p-values from the binding change tests of the three motifs under the null model are approximately uniform. The results of the score difference test are presented in Table 2. Overall, the score difference test is more conservative than the BC test.

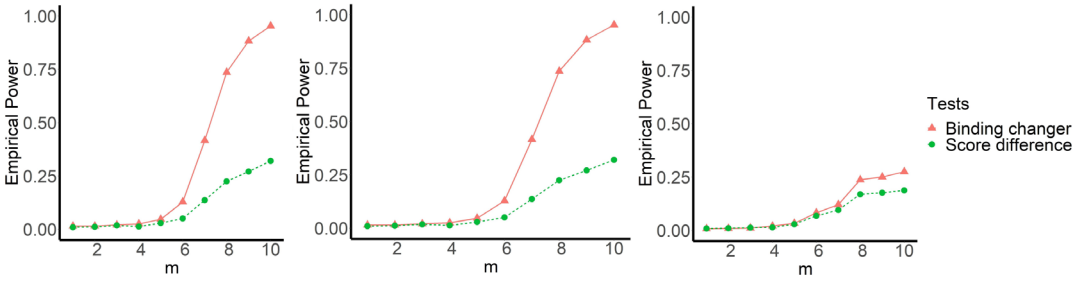
5.2. *Simulations under a defined set of alternatives.* To evaluate the power of the BC test, we consider the alternative model that introduces high binding affinity of a motif driven by the contrasting sequences which have the same length as the motif ($m = L$) such that the InDels may create or disrupt binding. The probability mass function of the full alternative model is as follows:

$$
\begin{aligned}
f_{\mathcal{H}_1}(\mathbf{y}) &= \sum_{\{y_L,\ldots,y_{2L-1}\}\in\{1,2,3,4\}^L} f_{\mathcal{H}_0}(\mathbf{y}) \prod_{l\in\{L,\ldots,2L-1\}} W(y_l, l-L+1) \\
&= \sum_{\{y_L,\ldots,y_{2L-1}\}\in\{1,2,3,4\}^L} \pi_0(y_1) \prod_{l\in\{1,\ldots,3L-3\}} a_0(y_l, y_{l+1}) \\
&\quad \times \prod_{l\in\{L,\ldots,2L-1\}} W(y_l, l-L+1).
\end{aligned} \tag{27}
$$

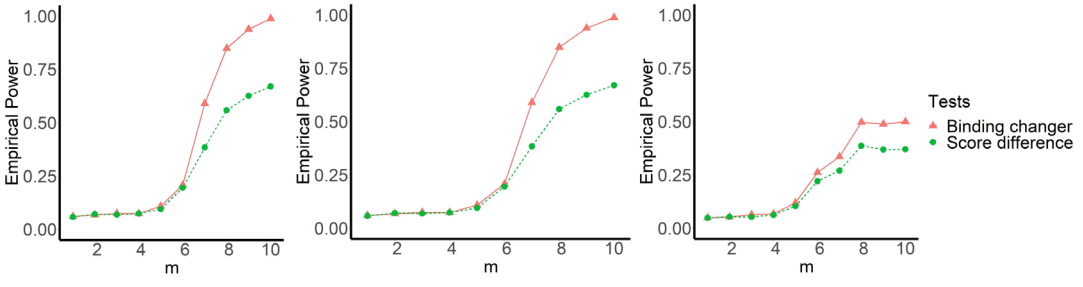
The contrasting sequence of length L $\{y_L, \ldots, y_{2L-1}\}$ is obtained from the PPM $W(\cdot, l - L + 1)$ such that the occurrence of one nucleotide in y_l follows a multinomial distribution of the $(l - L + 1)$th column of W, $l = L, \ldots, 2L - 1$. The shorter sequence of length $2L - 2$ is obtained by concatenating $\{y_1, \ldots, y_{L-1}\}$ and $\{y_{2L}, \ldots, y_{3L-2}\}$. Under the full alternative model, the longer sequences may have the nucleotide pattern of the motif while the shorter sequences lack the pattern.

TABLE 2
Empirical rejection probabilities of the score difference test

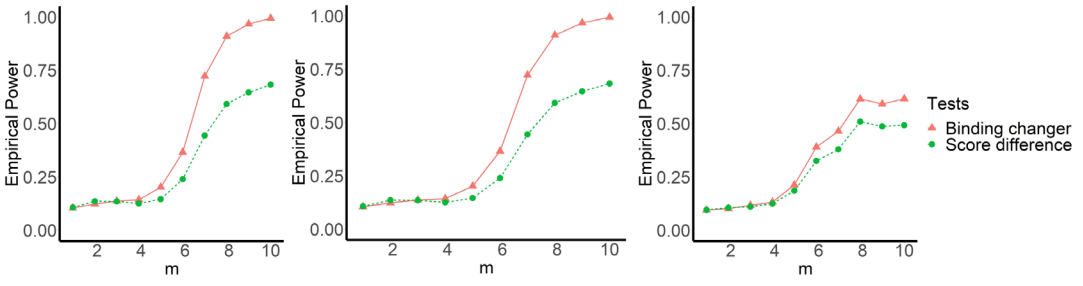
	Empirical rejection probability		
p-value	MSC	Ddit3::Cebpa	Hes1
0.01	0.0111	0.0114	0.0097
0.05	0.0521	0.0512	0.0476
0.10	0.1038	0.1024	0.1029



(a) Significance level 0.01



(b) Significance level 0.05



(c) Significance level 0.1

FIG. 3. *Power curves evaluated with the three motifs.*

Further, we consider local alternative models $f_{\mathcal{H}_{1\Delta}}$,

$$f_{\mathcal{H}_{1\Delta}}(\mathbf{y}) = \sum_{\{y_L,\ldots,y_{L+\Delta-1}\}\in\{1,2,3,4\}^{\Delta}} \pi_0(y_1) \prod_{l\in\{1,\ldots,3L-3\}} a_0(y_l, y_{l+1}) \times \prod_{l\in\{L,\ldots,L+\Delta-1\}} W(y_l, l-L+1),$$

where $\Delta \in \{1, \ldots L-1\}$ is the number of bases in the contrasting sequence following W. From the local alternative models, the contrasting sequences do not completely follow W but may contain a partial nucleotide pattern of the motif. The full alternative model can be viewed as $f_{\mathcal{H}_{1\Delta}}(\mathbf{y})$ with $\Delta = L$. We generated 2000 random sequence pairs for each of the alternative models and set the Monte Carlo sample size to 2000.

The empirical power curves of the various alternative models for the BC test and the binding score difference test are plotted at significance level 0.01, 0.05, and 0.1 in Figure 3. We clearly see that the empirical powers of the BC test are higher than those of the binding score

TABLE 3
Empirical rejection probability for the independent multinomial distribution

	Empirical rejection probability		
p-value	MSC	Ddit3::Cebpa	Hes1
0.01	0.0135	0.0172	0.0093
0.05	0.0523	0.0691	0.0504
0.10	0.1027	0.1218	0.0996

difference test. The poor performance of the score difference test may be attributable to a limited magnitude of changes to the score differences.

With the increase of Δ, we observe a dramatically increased rejection probability. In the case of motif MSC, the empirical power of the BC tests on InDels with $\Delta = L$ is close to one at all significance levels. At significance levels of 0.05 and 0.1, the empirical power of the BC tests on InDels following motif Ddit3::Cebpa is close to one, and the empirical power of the motif Hes1, which has low level of IC, is about 0.5. This shows that the IC is an important factor that determines the power of the test. For motif Hes1 there is little difference in the empirical rejection probabilities evaluated between $\Delta = 8$ and $\Delta = 10$ since the last two positions of the motif have low IC. Similarly, there is little difference in the empirical rejection probabilities of Ddit3::Cebpa evaluated between $\Delta = 9$ and $\Delta = 11$. Table S3 in Section 2.2 of the Supplementary Material (Zhou et al. (2024)) summarizes the power evaluated at $f_{\mathcal{H}_1}(\mathbf{y})$, where $\Delta = L$. In Section 2.3 of the Supplementary Material (Zhou et al. (2024)), we consider another set of alternatives where the protein binding start position varies. Similarly to the results above, the BC test outperforms the score difference test.

5.3. *Simulation under misspecification*. FIMO considers the independent multinomial distribution as the null model (Grant, Bailey and Nobel (2011)). Delcher et al. (1999), Borodovsky and McIninch (1993), Borodovsky et al. (1995) discussed that the fifth-order Markov model is an effective model for gene prediction. We examine if the first-order Markov chain model we employ is robust such that the Type I error is well controlled, even when the null model is specified to independent multinomial distribution or the fifth-order Markov model.

First, we calculate the independent multinomial distributions from the selected 100,000 sequences from the GRCh37/hg19. The probabilities for nucleotide A, C, G, T are estimated as 0.2918, 0.2073, 0.2080, and 0.2929, respectively. We used the estimated model parameters to generate a random sample of 10,000 sequence pairs. The rest of the settings follow Section 5.1. We calculate p-values for the binding affinity change of the random sequences with each motif. As shown in Table 3, the empirical rejection probabilities of the simulated reference and InDel sequences are close to the nominal significance levels. The BC test performs successfully under the independent multinomial model.

Next, we fitted the selected 100,000 sequences from hg19 with a fifth-order Markov model. Based on the $4^5 \times 4$ transition matrix estimated and the four prior probabilities of 0.2918, 0.2073, 0.2080, and 0.2929, we generated 10,000 random sequence pairs. The empirical rejection probabilities for p-value of 0.1 are around 0.1 in Table 4. There are minor discrepancies when the p-values are 0.01 or 0.05; thus, our test may be used for the random sequence pairs from the fifth-order Markov model.

6. Computation time. The BC test equipped with parallelization can significantly reduce computational time for the analysis of large-scale InDel data. We tested the running

TABLE 4
Empirical rejection probability for the fifth-order Markov model

	Empirical rejection probability		
p-value	MSC	Ddit3::Cebpa	Hes1
0.01	0.0185	0.0170	0.0078
0.05	0.0602	0.0635	0.0390
0.10	0.1135	0.1106	0.0843

time with a MacBook Pro (2.3 GHz Intel®Core i5), a Windows PC (2.4 GHz Intel®Core i5-1135G7), and a Linux server (Intel®Xeon Gold 5320 @ 2.20 GHz). The results shown in Tables 5–7 confirm the BC method is easily scalable to large genomic data with computational efficiency. In terms of computational time, both BC test and the score difference test are similar since they use the same set of samples.

7. Application to leukemia mutation data. We apply the BC test to 5737 somatic InDels in samples of primary acute myeloid leukemia (AML), which Li et al. (2020) discovered by targeted resequencing in 22,262 blood-cell associated cis-regulatory elements (CREs) using H3K27ac ChIP-seq peaks. Strelka (Saunders et al. (2012)) and Scalpel (Fang et al. (2016)) are used for mutation calling. Table 8 shows the distribution of the contrasting sequence lengths (m) from the somatic AML InDels. The insertions or deletions of a single base pair ($m = 1$) are most common, but the values of m can be as large as 57.

We use the JASPAR 2018 vertebrate library containing 579 motifs (Portales-Casamar et al. (2010), Khan et al. (2017)). To identify TF binding created or disrupted by the InDels, 304 motifs whose average IC is higher than 1.2 are considered. The cutoff value is the median IC value of the JASPAR motifs. We conducted 1,744,048 BC tests for the 5737 InDels against the 304 motifs with our R package *atIndel*. We use the null Markov chain model parameters from Section 5 based on the randomly selected 100,000 sequences of length 100 from the reference genome. To improve the test accuracy, Monte Carlo sample size was set to 10,000. To successfully control Type I error of the multiple tests, we compute adjusted p-values using Benjamini–Hochberg method per motif (Benjamini and Hochberg (1995)).

We select binding changer InDels that have strong TF binding to either the reference or the mutated sequences with the following criteria: (i) adjusted p-value less than 0.10 and (ii) at least one of the two binding p-values less than 0.05 (p_l, p_s). The binding p-value computation is also available in R package *atIndel*.

We nominate 28 binding changer InDels in the blood-cell associated CREs that lead to the gain or loss of 10 or more TF binding motifs. Table 9 presents the coordinates and the reference and mutant alleles of the InDels. Further, we mark 14 InDels as critical CREs (cCREs) by ENCODE SCREEN, a web interface searching the registry of candidate CREs (cCREs)

TABLE 5
Running time on Mac

Test	# of motifs	# of mutations	MC size	# of cores	time (minutes)
Binding change	10	100	2000	1	6.19
Score difference	10	100	2000	1	5.83
Binding change	10	100	2000	4	2.22
Score difference	10	100	2000	4	1.89

TABLE 6
Running time on Windows

Test	# of motifs	# of mutations	MC size	# of cores	time (minutes)
Binding change	10	100	2000	1	5.43
Score difference	10	100	2000	1	5.21
Binding change	10	100	2000	4	2.66
Score difference	10	100	2000	4	2.60

(The ENCODE Project Consortium et al. (2020a)). The human SCREEN registry is based on genome version GRCh38 (hg38); thus, we obtain cCREs on hg19 with liftover. We further annotate the binding changer InDels with their nearest neighbor genes using GREAT (Genomic Regions Enrichment of Annotations Tool) (McLean et al. (2010)) in Table 10. The distance from the middle coordinate of the InDel to the transcription start site (TSS) of the gene is given in the parenthesis. A positive direction "+" indicates that the InDel is downstream of the TSS and a negative direction "-" indicates that the InDel is upstream of the TSS. A deletion mutation, chr12:25,359,464-25,359,465, numbered 22nd, is approximately 44 kb downstream of the KRAS gene, which is a proto-oncogene that can become cancerous due to mutations (Li et al. (2020)). Braun et al. (2004), Van Meter et al. (2007) suggest KRAS coding mutations may be detrimental to normal hematopoietic stem cells. According to Li et al. (2020), the KRAS dosage regulation through altered noncoding CREs may attribute AML development to the genetic alternations. We further examine if any nearest genes are differently expressed in AML compared to normal cells using GEPIA2 (Tang et al. 2019), a web service for gene expression analysis based on TCGA tumor and normal samples (Tomczak, Czerwińska and Wiznerowicz (2015)). The nearest genes, TRAF3IP3, IQCG, SERINC5, FKBP9, TTC26, EZH2, KDM4C, ZNF484, PSMD5, TUFM, ATF5, ID1, MAGEA1, PNMA6A, and RUNX1 are differentially expressed with strong statistical significance. The 15 genes could be related to AML, and the corresponding InDels might need further investigation. Table S4 in Section 3 of the Supplementary Material (Zhou et al. (2024)) shows TF binding motifs gained or lost by each of the binding changer InDels. By the score difference test, only two deletion mutations, chr5: 40,798,695–40,798,696 and chr20: 30,198,508–30,198,509 are identified as changing TF binding.

8. Concluding remarks. We developed the test for statistical significance and built R package *atIndel* on the changes in TF binding affinity driven by InDels. To efficiently analyze large scale mutation data, we devised the general importance distribution suitable for

TABLE 7
Running time on server

Test	# of motifs	# of mutations	MC size	# of cores	time (minutes)
Binding change	10	100	2000	1	5.80
Score difference	10	100	2000	1	5.32
Binding change	10	100	2000	4	1.46
Score difference	10	100	2000	4	1.33
Binding change	500	20	2000	12	6.02
Score difference	500	20	2000	12	5.61
Binding change	500	20	10,000	12	36.76
Score difference	500	20	10,000	12	26.35

TABLE 8
Distribution of contrasting sequence lengths (m)

m	1	2	3	4	5	6–10	11–20	21-57	Total
#	4061	858	237	202	69	200	92	18	5737
%	70.8	15.0	4.1	3.5	1.2	3.5	1.6	0.3	100

any given InDels and motifs, which generates more sequence pairs of large score differences in the background sequence sampling. To improve computational efficiency, the tilting parameter value for the importance distribution was calculated such that the expected binding score difference is matched to the observed score difference. Simulation studies showed that the test controls Type I error and has high power under the alternative models. It showed high performance to test for motifs with high IC, where the nucleotide distribution of the positions is certain. We also applied our test to somatic InDels identified in primary leukemia samples against 304 motifs to uncover potential AML driver InDels. Eight InDels, whose nearest genes differentially expressed between AML tumor and normal samples, are found in critical CREs by ENCODE search. In computing the empirical p-value for the shorter sequence, one may attain high estimation accuracy at the cost of heavy computations by using the exact null distribution of the shorter sequences $f^a_{\mathcal{H}_0}$ instead of $\tilde{f}^a_{\mathcal{H}_0}$ in (24). As an effort to better

TABLE 9
Annotation information of leukemia binding changer InDels

No	Chr	Start	End	Ref	Mut	Critical CRE
1	chr1	17,221,878	17,221,879	TC	T	Yes
2	chr1	65,299,551	65,299,551	G	GAATT	
3	chr1	200,271,192	200,271,192	T	TA	
4	chr1	200,271,192	200,271,192	T	TAA	
5	chr1	202,779,915	202,779,916	GT	G	Yes
6	chr1	209,957,953	209,957,954	TC	T	Yes
7	chr2	216,979,898	216,979,900	TTG	T	
8	chr3	121,265,599	121,265,599	T	TAA	Yes
9	chr3	197,686,538	197,686,539	TA	T	
10	chr5	40,798,695	40,798,696	GC	G	
11	chr5	79,543,049	79,543,050	TA	T	Yes
12	chr6	88,032,984	88,032,988	TTTTG	T	
13	chr6	106,442,246	106,442,248	TAA	T	Yes
14	chr7	32,981,874	32,981,874	A	AC	
15	chr7	138,804,207	138,804,208	TA	T	
16	chr7	148,508,938	148,508,939	GT	G	
17	chr9	6,704,250	6,704,251	TG	T	Yes
18	chr9	66,494,029	66,494,029	C	CTG	Yes
19	chr9	95,639,846	95,639,850	CTAAT	C	
20	chr9	123,605,711	123,605,727	CGAAGGCGTGAGTAATA	C	Yes
21	chr9	135,995,961	135,995,962	CA	C	
22	chr12	25,359,464	25,359,465	TA	T	
23	chr12	29,542,090	29,542,091	AC	A	
24	chr16	28,836,192	28,836,195	TTAA	T	Yes
25	chr19	49,126,927	49,126,928	GT	G	Yes
26	chr19	50,400,737	50,400,738	CT	C	Yes
27	chr20	30,198,508	30,198,509	GA	G	Yes
28	chrX	152,419,769	152,419,770	AG	A	Yes

TABLE 10
Nearest genes of leukemia binding changer InDels

No	Nearest Genes	DE p-value
1	NBPF1 (−281,896), CROCC (−26,567)	
2	RAVER2 (+88,773), JAK1 (+132,636)	
3	ZNF281 (+107,992), NR5A2 (+274,462)	
4	ZNF281 (+107,992), NR5A2 (+274,462)	
5	KDM5B (−1317)	
6	IRF6 (+21,436), TRAF3IP3 (+26,110)	TRAF3IP3 (<0.0001)
7	XRCC5 (+7712), MARCH4 (+256,851)	
8	POLQ (−746)	
9	LMLN (−555), IQCG (+352)	IQCG (<0.0001)
10	PRKAA1 (−219)	
11	SERINC5 (+8849), THBS4 (+211,878)	SERINC5 (<0.0001)
12	SMIM8 (+680)	
13	PREP (−591,288), PRDM1 (−91,948)	
14	KBTBD2 (−50,365), FKBP9 (−15,143)	FKBP9 (0.0042)
15	TTC26 (−14,317), ZC3HAV1 (−9813)	TTC26 (<0.0001)
16	EZH2 (+72,432), CUL1 (+112,934)	EZH2 (<0.0001)
17	GLDC (−58,600), KDM4C (−53,406)	KDM4C (0.0007)
18	SPATA31A7 (−984,419)	
19	ZNF484 (+370)	ZNF484 (0.0047)
20	PSMD5 (−457)	PSMD5 (<0.0001)
21	RALGDS (+602)	
22	LYRM5 (+11,314), KRAS (+44,273)	
23	ERGIC2 (−8004), OVCH1 (+108,529)	
24	ATXN2L (+1779), TUFM (+21,536)	TUFM (<0.0001)
25	RPL18 (−4134)	
26	ATF5 (−31,663), AKT1S1 (−19,452)	ATF5 (<0.0001)
27	COX4I2 (−27,183), ID1 (+5422)	ID1 (<0.0001)
28	MAGEA1 (+66,346), PNMA6A (+81,468)	MAGEA1 (<0.0001), PNMA6A (<0.0001)

evaluate the Type I error of the test, we may consider the use of sequences of no TF binding inferred from the reference genome.

In the future one may consider a model other than the Markov chain by leveraging information about special patterns of the deletion/insertion sequences. Also, a collection of mutations that reside in a broader site may be investigated together with the BC testing results. This will boost the power of detecting regulatory regions in the genome and discovering the associated gene regulatory pathways. Further, we may study multiple bindings with different TFs such as super-enhancers, since multiple TF binding sites may have a larger impact on transcriptional regulation (Huang et al. (2016), Liu et al. (2017, 2020)). The BC test for DNA sequence may also be extended to RNA sequence.

Acknowledgments. The authors are grateful to Dr. Michael Q. Zhang and Dr. Zhenyu Xuan at University of Texas at Dallas for helpful discussions.

Funding. Shin was supported in part by U.S. NSF Grant DMS-2113674, Korean NRF grant funded by the Korea government (MSIT) (RS-2023-00243012, RS-2023-00219980), POSTECH Basic Science Research Institute Fund (NRF-2021R1A6A1A10042944), and POSCO HOLDINGS grant 2023Q033.

Xu is a Scholar of The Leukemia & Lymphoma Society (LLS) and an American Society of Hematology (ASH) Scholar.

SUPPLEMENTARY MATERIAL

Supplement to "Scalable test of statistical significance for protein-DNA binding changes with insertion and deletion of bases in the genome" (DOI: 10.1214/24-AOAS1950SUPPA; .pdf). We provide additional materials to support the results in the paper. Section 1 includes more details on the importance algorithm for the binding changer test. Section 1.1 shows how we construct the importance distribution; Section 1.2 includes a proof of Proposition 1; Section 1.3 shows the computational complexity of the algorithm. Section 2 includes more details on the simulation studies. Section 2.1 presents the model parameters in the simulation; Sections 2.2 and 2.3 provide the test performance under alternatives; Section 2.4 shows the test performance under misspecification. Section 3 presents the results of the leukemia mutation data analysis.

Source code to "Scalable test of statistical significance for protein-DNA binding changes with insertion and deletion of bases in the genome" (DOI: 10.1214/24-AOAS1950SUPPB; .zip). We provide R code for implementing the simulation studies and the application study. The R code for the binding change test is available at https://github.com/sunyoungshin/atIndel. The leukemia mutation data used in Section 7 were originally collected in Li et al. (2020). The data are available by request to Jian Xu at jian.xu@stjude.org.

REFERENCES

AVERY, P. J. (1987). The analysis of intron data and their use in the detection of short signals. *J. Mol. Evol.* **26** 335–340. https://doi.org/10.1007/BF02101152

AVERY, P. J. and HENDERSON, D. A. (1999). Detecting a changed segment in DNA sequences. *J. R. Stat. Soc., Ser. C* **48** 489–503. MR1721441 https://doi.org/10.1111/1467-9876.00167

BAILEY, T. L., WILLIAMS, N., MISLEH, C. and LI, W. W. (2006). MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **34** W369–W373.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BORODOVSKY, M. and MCININCH, J. (1993). GENMARK: Parallel gene recognition for both DNA strands. *Comput. Chem.* **17** 123–133.

BORODOVSKY, M., MCLNINCH, J. D., KOONIN, E. V., RUDD, K. E., MÉDIGUE, C. and DANCHIN, A. (1995). Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucleic Acids Res.* **23** 3554–3562.

BRAUN, B. S., TUVESON, D. A., KONG, N., LE, D. T., KOGAN, S. C., ROZMUS, J., LE BEAU, M. M., JACKS, T. E. and SHANNON, K. M. (2004). Somatic activation of oncogenic Kras in hematopoietic cells initiates a rapidly fatal myeloproliferative disorder. *Proc. Natl. Acad. Sci. USA* **101** 597–602.

CAPPELLO, L. and PALACIOS, J. A. (2020). Sequential importance sampling for multiresolution Kingman-Tajima coalescent counting. *Ann. Appl. Stat.* **14** 727–751. MR4117827 https://doi.org/10.1214/19-AOAS1313

CHAN, H. P. and ZHANG, N. R. (2007). Scan statistics with weighted observations. *J. Amer. Statist. Assoc.* **102** 595–602. MR2370856 https://doi.org/10.1198/016214506000001392

CHAN, H. P., ZHANG, N. R. and CHEN, L. H. Y. (2010). Importance sampling of word patterns in DNA and protein sequences. *J. Comput. Biol.* **17** 1697–1709. MR2749757 https://doi.org/10.1089/cmb.2008.0233

CINGOLANI, P., PLATTS, A., WANG, L. L., COON, M., NGUYEN, T., WANG, L., LAND, S. J., LU, X. and RUDEN, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* (*Austin*) **6** 80–92. https://doi.org/10.4161/fly.19695

COETZEE, S. G., COETZEE, G. A. and HAZELETT, D. J. (2015). motifbreakR: An R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31** 3847–3849.

THE ENCODE PROJECT CONSORTIUM (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489** 57–74. https://doi.org/10.1038/nature11247

THE ENCODE PROJECT CONSORTIUM, MOORE, J. E., PURCARO, M. J., PRATT, H. E., EPSTEIN, C. B., SHORESH, N., ADRIAN, J., KAWLI, T., DAVIS, C. A. et al. (2020a). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583** 699–710.

THE ENCODE PROJECT CONSORTIUM, SNYDER, M. P., GINGERAS, T. R., MOORE, J. E., WENG, Z., GERSTEIN, M. B., REN, B., HARDISON, R. C., STAMATOYANNOPOULOS, J. A. et al. (2020b). Perspectives on ENCODE. *Nature* **583** 693–698.

COWAN, R. (1991). Expected frequencies of DNA patterns using Whittle's formula. *J. Appl. Probab.* **28** 886–892. MR1133796 https://doi.org/10.2307/3214691

HUANG, H., KAO, M.-C. J., ZHOU, X., LIU, J. S. and WONG, W. H. (2004). Determination of local statistical significance of patterns in Markov sequences with application to promoter element identification. *J. Comput. Biol.* **11** 1–14.

DELCHER, A. L., HARMON, D., KASIF, S., WHITE, O. and SALZBERG, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27** 4636–4641. https://doi.org/10.1093/nar/27.23.4636

DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195 https://doi.org/10.1214/009053604000000265

FANG, H., BERGMANN, E. A., ARORA, K., VACIC, V., ZODY, M. C., IOSSIFOV, I., O'RAWE, J. A., WU, Y., BARRON, L. T. J. et al. (2016). Indel variant analysis of short-read sequencing data with Scalpel. *Nat. Protoc.* **11** 2529–2548.

FREDRIKSSON, N. J., NY, L., NILSSON, J. A. and LARSSON, E. (2014). Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat. Genet.* **46** 1258–1263.

GRANT, C. E., BAILEY, T. L. and NOBEL, W. S. (2011). FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **7** 1017. https://doi.org/10.1093/bioinformatics/btr064

GUPTA, M. and IBRAHIM, J. G. (2007). Variable selection in regression mixture modeling for the discovery of gene regulatory networks. *J. Amer. Statist. Assoc.* **102** 867–880. MR2411650 https://doi.org/10.1198/016214507000000068

HERTZ, G. Z. and STORMO, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15** 563–577. https://doi.org/10.1093/bioinformatics/15.7.563

HUANG, J., LIU, X., LI, D., SHAO, Z., CAO, H., ZHANG, Y., TROMPOUKI, E., BOWMAN, T. V., ZON, L. I. et al. (2016). Dynamic control of enhancer repertoires drives lineage and stage-specific transcription during hematopoiesis. *Dev. Cell* **36** 9–23. https://doi.org/10.1016/j.devcel.2015.12.014

JENSEN, S. T. and LIU, J. S. (2008). Bayesian clustering of transcription factor binding motifs. *J. Amer. Statist. Assoc.* **103** 188–200. MR2420226 https://doi.org/10.1198/016214507000000365

KHAN, A., FORNES, O., STIGLIANI, A., GHEORGHE, M., CASTRO-MONDRAGON, J. A., VAN DER LEE, R., BESSY, A., CHÈNEBY, J., KULKARNI, S. R. et al. (2017). JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46** D260–D266. https://doi.org/10.1093/nar/gkx1126

KUNDAJE, A., MEULEMAN, W., ERNST, J., BILENKY, M., YEN, A., HERAVI-MOUSSAVI, A., KHERADPOUR, P., ZHANG, Z., WANG, J. et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* **518** 317–330.

LI, F. and ZHANG, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *J. Amer. Statist. Assoc.* **105** 1202–1214. MR2752615 https://doi.org/10.1198/jasa.2010.tm08177

LI, K., ZHANG, Y., LIU, X., LIU, Y., GU, Z., CAO, H., DICKERSON, K. E., CHEN, M., CHEN, W. et al. (2020). Noncoding variants connect enhancer dysregulation with nuclear receptor signaling in hematopoietic malignancies. *Cancer Discov.* **10** 724–745. https://doi.org/10.1158/2159-8290.CD-19-1128

LI, S., KO, Y. M. and BYON, E. (2021). Nonparametric importance sampling for wind turbine reliability analysis with stochastic computer models. *Ann. Appl. Stat.* **15** 1850–1871. MR4355079 https://doi.org/10.1214/21-aoas1490

LIANG, F. (2002). Dynamically weighted importance sampling in Monte Carlo computation. *J. Amer. Statist. Assoc.* **97** 807–821. MR1941411 https://doi.org/10.1198/016214502388618618

LIN, M., WHITMIRE, S., CHEN, J., FARREL, A., SHI, X. and GUO, J.-T. (2017). Effects of short indels on protein structure and function in human genomes. *Sci. Rep.* **7** 1–9.

LIU, X., CHEN, Y., ZHANG, Y., LIU, Y., LIU, N., BOTTEN, G. A., CAO, H., ORKIN, S. H., ZHANG, M. Q. et al. (2020). Multiplexed capture of spatial configuration and temporal dynamics of locus-specific 3D chromatin by biotinylated dCas9. *Genome Biol.* **21** 1–20.

LIU, X., ZHANG, Y., CHEN, Y., LI, M., ZHOU, F., LI, K., CAO, H., NI, M., LIU, Y. et al. (2017). In situ capture of chromatin interactions by biotinylated DCas9. *Cell* **170** 1028–1043.e19. https://doi.org/10.1016/j.cell.2017.08.003

MACINTYRE, G., BAILEY, J., HAVIV, I. and KOWALCZYK, A. (2010). is-rSNP: A novel technique for in silico regulatory SNP detection. *Bioinformatics* **26** 524–530.

KAHN, H. MARSHALL, A. W. (1953). Methods of reducing sample size in Monte Carlo computations. *J. Oper. Res. Soc. Am.* **1** 263–278.

MCLAREN, W., GIL, L., HUNT, S. E., RIAT, H. S., RITCHIE, G. R., THORMANN, A., FLICEK, P. and CUNNINGHAM, F. (2016). The ensembl variant effect predictor. *Genome Biol.* **17** 1–14.

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M. and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28** 495–501. https://doi.org/10.1038/nbt.1630

Menéndez, M. L., Pardo, L., Pardo, M. C. and Zografos, K. (2011). Testing the order of Markov dependence in DNA sequences. *Methodol. Comput. Appl. Probab.* **13** 59–74. MR2755132 https://doi.org/10.1007/s11009-008-9107-1

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S. and Devine, S. E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* **16** 1182–1190. https://doi.org/10.1101/gr.4565806

Mukherjee, R., Pillai, N. S. and Lin, X. (2015). Hypothesis testing for high-dimensional sparse binary regression. *Ann. Statist.* **43** 352–381. MR3311863 https://doi.org/10.1214/14-AOS1279

Mullaney, J. M., Mills, R. E., Pittard, W. S. and Devine, S. E. (2010). Small insertions and deletions (INDELs) in human genomes. *Hum. Mol. Genet.* **19** R131–R136.

Portales-Casamar, E., Thongjuea, S., Kwon, A. T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W. W. et al. (2010). JASPAR 2010: The greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **38** D105–D110. https://doi.org/10.1093/nar/gkp950

Reinert, G., Schbath, S. and Waterman, M. S. (2000). Probabilistic and statistical properties of words: An overview. *J. Comput. Biol.* **7** 1–46. https://doi.org/10.1089/10665270050081360

Retkute, R., Touloupou, P., Basáñez, M.-G., Hollingsworth, T. D. and Spencer, S. E. F. (2021). Integrating geostatistical maps and infectious disease transmission models using adaptive multiple importance sampling. *Ann. Appl. Stat.* **15** 1980–1998. MR4355085 https://doi.org/10.1214/21-aoas1486

Riva, A. (2012). Large-scale computational identification of regulatory SNPs with rSNP-MAPPER. *BMC Genomics* **13** S7. https://doi.org/10.1186/1471-2164-13-S4-S7

Saunders, C. T., Wong, W. S. W., Swamy, S., Becq, J., Murray, L. J. and Cheetham, R. K. (2012). Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28** 1811–1817. https://doi.org/10.1093/bioinformatics/bts271

Sehn, J. K. (2015). Chapter 9—insertions and deletions (indels). In *Clinical Genomics* (S. Kulkarni and J. Pfeifer, eds.) 129–150. Academic Press, Boston, MA. https://doi.org/10.1016/B978-0-12-404748-8.00009-5

Stormo, G. D., Shneider, T. D., Gold, L. and Ehrenfeucht, A. (1982). Use of 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res.* **10** 2997–3010.

Tang, Z., Kang, B., Li, C., Chen, T. and Zhang, Z. (2019). GEPIA2: An enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res.* **47** W556–W560.

Tomczak, K., Czerwińska, P. and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol.* **19** A68–A77.

Van Meter, M. E., Díaz-Flores, E., Archard, J. A., Passegué, E., Irish, J. M., Kotecha, N., Nolan, G. P., Shannon, K. and Braun, B. S. (2007). K-RasG12D expression induces hyperproliferation and aberrant signaling in primary hematopoietic stem/progenitor cells. *Blood* **109** 3945–3952.

Wang, K., Li, M. and Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38** e164–e164.

Zhou, Q. and Liu, J. S. (2004). Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics* **20** 909–916.

Zhou, Q., Zuo, C., Zhang, Y., Chen, M., Xu, J. and Shin, S. (2024). Supplement to "Scalable test of statistical significance for protein-DNA binding changes with insertion and deletion of bases in the genome." https://doi.org/10.1214/24-AOAS1950SUPPA, https://doi.org/10.1214/24-AOAS1950SUPPB

Zuo, C., Shin, S. and Keleş, S. (2015). atSNP: Transcription factor binding affinity testing for regulatory SNP detection. *Bioinformatics* **31** 3353–3355. https://doi.org/10.1093/bioinformatics/btv328