Characterizing and Correcting Tn5 Sequence Bias in ATAC-seq Data

Jacob Brandon Wolpe
Springfield, Virginia

B.S. Integrated Science and Technology James Madison University, 2012
Concentration: Biotechnology

A Dissertation presented to the Graduate Faculty of the University of Virginia
in Candidacy for the Degree of Doctor of Philosophy

Department of Cell Biology

University of Virginia
May 2023

Abstract

Chromatin accessibility assays enable base pair resolution measurements of the chromatin landscape and facilitate inferences about the nature of regulatory state. These assays determine which genomic regions are actively regulated by quantifying enzymatic digestion of DNA in a given region. However, the enzymes which conduct this digestion are biased with respect to their propensity to cleave specific sequences. Enzymatic sequence bias can introduce artifacts into data, leading to misinterpretation of downstream analysis. Previous work to address this bias relied on calculating enzymatic bias for specific k-mers and correcting these values to their genomic frequency, known as k-mer scaling. K-mer scaling was successful in correcting nuclease bias, but not the bias of Tn5 transposase, the enzyme used in ATAC-seq. This dissertation illustrates that the breadth and complexity of Tn5 bias hinders the use of k-mer scaling for sequence bias correction. Comparison of Tn5 bias with nucleases highlights why k-mer scaling is ineffective: Tn5 sequence bias is based on a region greater than 20 bp. K-mer scaling can only be applied to k-mers which have many instances in the data set and genome, an impossibility with k-mer sizes greater than 9, due to the number of reads contained in most ATAC-seq experimental data sets. To model this large bias window, we used a machine learning approach, rule ensemble, which integrates information from many input k-mers into a computational bias correction. We created a workflow using this approach, seqOutATACBias, in order to promote bias correction in other studies. We applied seqOutATACBias to naked DNA and found that it effectively diminishes both local sequence bias in addition to correcting a previously unreported Tn5 regional bias for high GC content. Correction of enzymatic sequence bias is of utmost importance for determining ground truth of chromatin accessibility assays.

Contents

Chapter 1

Introduction

1.1 Chromatin accessibility regulates gene expression

Regulation of gene expression is a necessary component of multicellular life. As each cell contains theoretically identical genetic material, it is the regulated expression of these genes which creates cellular types as diverse as red blood cells and neurons (H. Wu and Sun 2006). Gene expression regulation occurs through many mechanisms from DNA to functional protein, frequently in unique interactions for each gene. Of these mechanisms, regulation of transcription by DNA-binding proteins is thought of as the most important, in part because it is the most energetically efficient form of expression regulation (Pulverer 2005). Regardless of relative importance, transcriptional regulation is the initial regulatory step of protein function, and is universal as it applies to every protein. This process is conducted by about 1500 DNA binding proteins which recognize specific sequence motifs, known as transcription factors (TFs) (Dynan and Tjian 1983; Ignatieva, Levitsky, and Kolchanov 2015). TFs bind to specific DNA sequences, and often require DNA in an uncondensed, accessible state to regulate transcription of a given gene (Cremer et al. 2015; Felsenfeld et al. 1996). DNA accessibility is itself a highly regulated process. DNA accessibility is mediated through a host of chromatin remodeling proteins and histones. Interactions between chromatin remodeling proteins and histones determine nucleosome (a complex of DNA and histones) positioning and local availability for TFs and RNA polymerase (Clapier and Cairns 2009; Blossey and Schiessel 2018).

TFs and nucleosome remodeling proteins are known to regulate transcription and chromatin accessibility, respectively. A deterministic understanding as to how interactions between the two alter regulation of either process is currently a topic of investigation (M. Li et al. 2015; Ren et al. 2019; Hansen, Loell, and Cohen 2022). However, it is well known that genomic regions with high transcriptional rates have proportionally less nucleosome occupancy (Lee et al. 2004). Therefore, accessibility of regions in the genome can be measured to infer the degree of regulatory activity acting on the resident genes (Felsenfeld et al. 1996). Previous studies have leveraged this principle to show alterations in accessibility due to developmental changes or between cell types (John et al. 2011; Thurman et al. 2012). Accessible regions in the genome occur dynamically in response to environmental or developmental stimuli, account for the majority of TF-bound locations, and define cellular ability to differentiate (John et al. 2011; C. Wu, Bingham, et al. 1979; C. Wu, Wong, and Elgin 1979; Di Stefano et al. 2016).

Early measurements of chromatin accessibility used DNase to digest genetic material, as the nuclease preferentially cuts accessible regions of the genome (Weintraub and Groudine 1976). These initial DNase digestions were followed by hybridization with a labeled complementary probe (Southern blot) to determine the sensitivity to digestion of the areas around the sequence of interest (C. Wu, Bingham, et al. 1979; C. Wu, Wong, and Elgin 1979). DNase digestions produce fragments of DNA when two independent cutting events occur in close enough proximity. Because cutting events happen more frequently in accessible, actively regulated regions, fragments from these hypersensitive sites are more abundant than those from condensed DNA (Elgin 1988; Gross and Garrard 1988). Initial work using these methods generally examined only a few sequences of interest, as each sequence required its own probe. Subsequent advancements in sequencing technology allowed for each fragment produced by DNase digestion to be sequenced and aligned to a reference genome, a technique known as DNase-seq (Boyle et al. 2008; Song and Crawford 2010). DNase-seq produces genome-wide information about chromatin accessibility. Analysis of DNase-seq chromatin accessibility data identified genome-wide footprints of transcription factor binding as a reduction of signal within regions of hypersensitivity (Hesselberth et al. 2009). These footprints had previously been reported, as initial *in vitro* DNase digestions of DNA were shown to be protected by interaction with DNA-binding proteins (Galas and Schmitz 1978). However, the genome-wide scope of DNase-seq allowed for alignment of sequences on a common motif and analysis of signal at all of these positions combined. Analysis of this combined signal was thought to show a statistical average of DNA-protein interaction among the population of surveyed cells, as some studies observed an inverse correlation between footprint signal and information content of a TF motif (Hesselberth et al. 2009). Although these footprint signals allowed researchers to generate new insights into TF-DNA interactions, they are also susceptible to enzymatic sequence bias. This footprint susceptibility to bias is a consequence of determining signal given aggregation of similar sequences and thus similar biases. In order to correct these biases seqOutBias, a command line interface, was developed in order to apply an algorithmic bias correction to individual aligned reads in a data set (Martins et al. 2018).

1.2 seqOutBias uses direct k-mer scaling to correct enzyme bias

Further investigation into the relationship between DNase hypersensitivity and footprinting yielded remarkable links between DNA-TF crystal structures and observed DNase-seq signal at aggregated TF motifs (Neph et al. 2012). In short, it was observed that genetic positions protected by a bound TF had reduced signal, while nearby positions exposed to DNase digestion had increased signal. Unfortunately, these findings were partially subsumed by studies which examined the sequence bias of DNase. Two independent studies, published within a year of each other, recapitulated many of the DNase-seq TF footprinting patterns observed in previous studies, using naked DNA (Sung et al. 2014; He et al. 2014). Because naked DNA lacks proteins bound to the DNA, the signal observed at aggregated TF motifs could not be due to interactions between TFs and DNA. Instead, these studies demonstrated that using the DNase cut frequency of each k-mer, DNase sequence bias could be modeled reproducibly across experiments and cell types. This modeling strategy was also shown to apply to the nucleases Benzonase and Cyanase, indicating that nuclease preference for specific sequences was a reproducible phenomenon. Intuitively, if one can imagine on a molecular scale, nuclease contact with its DNA substrate involves chemical interactions between the two, some genetic sequences may favor or disfavor interaction over others.

Furthering the modeling of enzymatic sequence bias, `seqOutBias` was published as a software package which corrects this bias using a direct k-mer scaling approach (Martins et al. 2018). `seqOutBias` leverages the observation that enzymatic sequence preferences are consistent across species and chromatin state. This obviates the need for a naked DNA baseline of the applied nuclease's k-mer cutting preferences, and instead allows this correction to be derived from and directly applied to experimental data regardless of experimental parameters.

In order to apply direct k-mer scaling to a given data set, the `seqOutBias` algorithm executes several computations. First, each k-mer specified at a relative position in both the data set and reference genome are tallied. These k-mer counts are then used to determine percent frequency in both the reference genome and data set for each possible -mer of length k. A given k-mer's percent frequency in the reference genome is then divided by its percent frequency in the data set, creating a 'scale factor' for each possible k-mer. Each scale factor is the ratio of genomic occurrence of a k-mer to its occurrence in the data set. The final computation multiplies each read by the scale factor for its respective k-mer in order to scale the occurrence of all k-mers in the data set to their observed frequency in the reference genome. Thus, `seqOutBias` corrects the observed k-mer bias for the specified position relative to each read.

As read values are only scaled based on the k-mer specified by a position relative to each read, sequence bias is only corrected for these corresponding positions. Consequently, to truly correct the sequence bias for a given enzyme, one must scale reads based on the positions which bias cutting. Intuitively, these locations can be thought of as the bases which interact with the enzyme and create biased DNA sequence-enzyme interactions with respect to cutting propensity. A simple solution to this problem would be to specify a large window around a cut site so as to incorporate all possibly relevant positions. Unfortunately, aside from being computationally expensive (there are 4^k possible k-mers), direct k-mer scaling necessitates several instances of each k-mer for bias correction. This limitation becomes apparent in a recommended data set size of 50 million reads using a 9-mer. This would result in the average k-mer having 191 instances, and many k-mers would have few instances based on a normal distribution (Buenrostro, B. Wu, Chang, et al. 2015). The size of the human genome (3,099,706,404 bases in hg38) further limits k-mer size to smaller than 12, as the average k-mer would have 185 instances.

Application of the `seqOutBias` algorithm was shown to correct the surveyed nucleases' (DNase, Cyanase, Benzonase, MNase) sequence bias by scaling reads to the k-mer several positions around a cutsite (Martins et al. 2018). These masks were determined using an enzyme cutsite model. However, this method did not sufficiently correct Tn5 transposase bias. Instead, the best performing Tn5 bias correction was produced using a hill climbing optimization method (Martins et al. 2018). This method sought to reduce the sum of standard deviations between positions in a composite plot of signal at a set of a PSWMs. Hill climbing optimization of the positions which bias cutting in Tn5 data specified a wide range of spaced positions around a cutsite as the most relevant for bias correction. Unfortunately, these positions also did not fully smooth the bias observed in composite plots.

1.3 ATAC-seq uses Tn5 transposase to measure chromatin accessibility

ATAC-seq (Assay for Transposase Accessible Chromatin with high-throughput sequencing), first described in 2013, is a technique that uses a modified Tn5 transposase to fragment DNA and measure chromatin accessibility (Buenrostro, Giresi, et al. 2013). The annual number of published ATAC-seq data sets surpassed those of other chromatin accessibility assays within four years of its introduction. This rapid increase in publications is due to its comparative ease of use, the broad applicability of the data generated, and the smaller sample size required (Yan et al. 2020; Buenrostro, B. Wu, Chang, et al. 2015). These factors have facilitated the refinement and application of ATAC-seq to single-cell measurements of chromatin accessibility (Buenrostro, B. Wu, Litzenburger, et al. 2015). Tn5 transposase's rapid adoption for these purposes was facilitated by the previous knowledge gained and modifications made to it as an early model system.

Tn5 transposase, a dimer, is the key component of ATAC-seq. It was initially identified for its ability to transfer a suite of antibiotic resistance genes from *Escherichia coli* to an infecting phage's genome (Berg et al. 1975). This led to its use as a model system for transposon studies due to its spontaneous transposition, relative simplicity, and mechanistic similarities to other biologically relevant processes such as HIV-DNA integration (Reznikoff 1993). However, one initial challenge in studying Tn5 as a model transposase was its low rate of endogenous activity. To address this, a hyperactive variant of the complex was created through numerous modifications to both its protein structure and target DNA sequence (Reznikoff 2008; Igor Yu Goryshin and Reznikoff 1998). This variant was able to fragment DNA at a reasonable rate, enabling the measurement of sequence preferences (Igor Y Goryshin et al. 1998). Further analysis revealed that regions of the dimer interact with DNA sequences outside of the 9bp active site, contributing to sequence bias for insertions (Ason and Reznikoff 2004; Gradman et al. 2008). These findings suggest that using a direct k-mer scaling approach to correct Tn5's sequence bias would require a k-mer significantly larger than 9bp. As previously explained, a direct k-mer scaling approach using a k-mer this large is not feasible for most data sets. Consequently, a modeling strategy would need to incorporate information from many smaller direct k-mer scaling approaches to successfully model the full Tn5 sequence bias. One such modeling strategy is a rule ensemble, using as input the output from many different direct k-mer scaling approaches.

1.4 Rule ensemble combination of multiple direct k-mer scaling inputs models enzyme bias

The degree to which a model corrects chromatin accessibility assay enzymatic sequence bias is best measured using data from deproteinated, naked DNA. Theoretically unbiased naked DNA digestion would result in an average signal of read depth divided by positions within the reference genome, containing variations within noise. Unbiased signal from naked DNA chromatin accessibility studies should be evenly distributed throughout the genome, regardless of sequence context. Variations on this even distribution can be easily visualized using a composite plot of sequences similar to a chosen motif. If any sequences from this motif are not uniformly preferred by the digesting enzyme, bias will be visualized as either an increase or depletion of signal in the plot. Hence, the degree to which a model corrects enzymatic sequence bias is measured by its ability to return composite plot signal to the known ground truth.

Composite plots of a set of sites which conform to a TF motif are an average of signal at known coordinates, and thus their sequences are known. Using these known sequences, one can partition each position into the k-length sequences surrounding it. These k-length sequences can be used to determine the frequency of each possible k-mer at each position within the plot. Multiplying k-mer frequency values by the inverse of `seqOutBias` scale factors for a k-mer at a given relative position yields the predicted bias for that k-mer position and composite plot. Predicted bias is the value which `seqOutBias` aims to correct by multiplying reads by non-inverted scale factors. Therefore, predicted bias values from different k-mer sizes and positions can be used as unique training input into the predictive modeling technique of the user's choice to estimate the observed bias in a composite plot of raw data. Using the trained modeling technique's formula, `seqOutBias` output values can be integrated in an identical fashion which combines corrections for each respective input k-mer size and position.

Statistical inference uses training input to create a model which predicts output, given new data. Rule ensemble (RE) is a statistical inference modeling method which has predictive accuracy comparable to boosted tree ensembles and random forests (Fokkema 2017). The accuracy of RE arises from combining a linear basis function (similar to a linear regression model) with decision trees (rules) (Friedman and Popescu 2008). This combination helps to alleviate the deficiency of decision trees in predicting linear relationships and conversely, allows nonlinear interactions to be captured by rules. Models are trained based on the importance sampled learning ensemble methodology and terms are scaled using the lasso penalty (Friedman, Popescu, et al. 2003). A unique advantage to REs is the interpretability of predictions. This interpretability can be assessed at the level of the individual prediction, insofar as the linear combination of inputs and each rule can be evaluated for its contribution to the prediction. Single point interpretability enables an assessment of the prediction and can reveal which nonlinear interactions between inputs are necessary for the estimate. More generally, REs produce importance values for each input, quantifying both the linear and rule-based contributions each makes.

Chapter 2

Results

2.1 Correction of transposase sequence bias in ATAC-seq data with rule ensemble modeling

2.1.1 Abstract

Chromatin accessibility assays have revolutionized the field of transcription regulation by providing single-nucleotide resolution measurements of regulatory features such as promoters and transcription factor binding sites. ATAC-seq directly measures how well the Tn5 transpose accesses chromatinized DNA. Tn5 has a complex sequence bias that is not effectively scaled with traditional bias-correction methods. We model this complex bias using a rule ensemble machine learning approach that integrates information from many input k-mers proximal to the ATAC sequence reads. We effectively characterize and correct single-nucleotide sequence biases and regional sequence biases of the Tn5 enzyme. Correction of enzymatic sequence bias is an important step in interpreting chromatin accessibility assays that aim to infer transcription factor binding and regulatory activity of elements in the genome.

2.1.2 Introduction

Chromatin accessibility assays measure the relative frequency that exogenous enzymes access DNA. Chromatin accessibility is not a direct measure of molecular features of chromatin such as transcription factor occupancy or histone modification status. However, accessibility is considered a proxy measurement of regulatory element activity irrespective of the constellation of factors bound to the DNA (C. Wu, Bingham, et al. 1979; C. Wu, Wong, and Elgin 1979). Accessible regions are enriched for transcription factor binding and histone modifications that are hallmarks of functional *cis*-regulatory elements (Moore et al. 2020; Thurman et al. 2012; Tewari et al. 2012; Guertin et al. 2012; Boyle et al. 2008). Deconvolution of genome-wide chromatin hypersensitivity data to infer individual transcription factor binding events remains a challenge (H. Li, Quang, and Guan 2019).

ATAC-seq revolutionized the chromatin accessibility field by providing a straightforward method that requires fewer than 5000 cells (Buenrostro, B. Wu, Chang, et al. 2015). ATAC-seq leverages a hyperactive Tn5 transposase that directly inserts high throughput sequencing adapters into accessible DNA to create sequencing libraries. The analysis of ATAC-seq data requires additional considerations beyond traditional hypersensitivity assays because the molecular biology of transposase function is distinct compared to DNase and other enzymes that are used to measure chromatin accessibility (Buenrostro, Giresi, et al. 2013; J. P. Smith et al. 2021).

Although using Tn5 to measure chromatin accessibility is experimentally easier than using DNase, the dimeric Tn5 enzyme recognizes a wider region when interacting with DNA and this is reflected in the sequence bias of Tn5 (Z. Li et al. 2019). Characterizing and correcting enzymatic sequence bias is an essential step for accurate interpretation of chromatin accessibility data (Koohy, Down, and Hubbard 2013; He et al. 2014; Sung et al. 2014). Strategies to correct biases use a variety of sequence inputs to build models, including k-mer instances, position weight matrices, and long stretches of DNA. DNase bias can be directly scaled based on the 6-mer sequence centered on the DNase cleavage site (Martins et al. 2018; Yardımcı et al. 2014; Schwessinger et al. 2017; J. R. Wang, Quach, and Furey 2017). The advantage of direct k-mer scaling is the simplicity: reads are scaled by the expected / observed k-mer count ratio. If a k-mer is found less often than expected by chance, then the read is scaled to a higher value. However, direct k-mer scaling does not effectively correct Tn5 biases from ATAC-seq data (Martins et al. 2018; Karabacak Calviello et al. 2019; Schwessinger et al. 2017), so other methods were developed to correct Tn5 bias. `ATACorrect` scales individual ATAC-seq reads based upon a dinucleotide weight matrix representing Tn5 bias and introduces simulated reads to offset observed biases (Bentsen et al. 2020). Similar to weight matrix scaling, `HINT-ATAC` employs position dependency models, which account for interactions between weight matrix positions to model ATAC-seq bias (Z. Li et al. 2019). `SELMA` encodes k-mer data into a simplex vector, which is incorporated into a Hadamard matrix for all mono and dinucleotide combinations. These data are input into a linear regression model to capture and correct both DNase and Tn5 bias (Hu et al. 2022). More sophisticated methods that correct ATAC-seq data are often less interpretable than k-mer and weight matrix scaling. `Seqbias` trains a Bayesian network to encode nucleotide preferences using a 40 base window centered on the Tn5 recognition site (Vinayak, Vinay, and Shiv 2019). `MsCentipede` uses naked DNA cleavage to train a Bayesian multi-scale model of a Poisson distribution of reads to account for sequence bias (Raj et al. 2015). Another method employs a convolutional neural network to account for intrinsic sequence biases (Ansari, Fischer, and Theis 2020).

Among bias correction software packages, `seqOutBias` is the only existing method that scales each individual input read and does not couple the bias correction to downstream processing steps.

We previously found that direct k-mer scaling corrects the majority of nuclease sequence bias. We developed `seqOutBias` to implement k-mer scaling for molecular genomics data. The `seqOutBias` package is stand-alone software that specializes in sequence bias correction for a range of k-mer lengths and gapped k-mers. `SeqOutBias` output files can be piped into programs that specialize in peak calling and footprint inference (Gaspar 2018), but `seqOutBias` performs poorly with ATAC-seq data. Here, we expanded the `seqOutBias` package to accommodate ATAC-seq data by coupling `seqOutBias` output to a rule ensemble modeling framework that effectively scales individual ATAC-seq reads to correct Tn5 bias. This modeling approach captures complex interactions between k-mers and quantifies the importance of individual positions that contribute to overall Tn5 bias. Moreover, the importance of k-mers and positions are intrepretable locally for each position in the genome, because the rules and terms applied to each aligned read are explicit. This reproducible workflow efficiently corrects single-nucleotide resolution Tn5 sequence bias and addresses regional baseline bias determined by GC content. To facilitate bias correction of chromatin accessibility data, we developed a workflow that can be applied to existing high-throughput sequencing analysis pipelines. The maintained and updated repository for this work will be within the lab's GitHub repository: `https://github.com/guertinlab/Tn5bias`. The archived Zenodo link for the analysis at the time of submission will be stable at the following link: `https://doi.org/10.5281/zenodo.7757436`.

2.1.3 Results

2.1.3.1 Tn5 sequence bias is more extensive than other nuclease biases

ATAC-seq measures how well Tn5 transposase accesses DNA in chromatin, which is a proxy measurement for regulatory element activity. Although chromatin condensation is the main contributor that influences Tn5-mediated DNA cleavage, the local sequence content biases transposition activity (Buenrostro, Giresi, et al. 2013). A dimeric Tn5 complex nicks opposite strands of DNA 9 base pairs apart, and each monomer inserts a sequencing-compatible adapter at each nick position (Figure 2.1A) (Reznikoff 2008; Reznikoff 2003). The interaction of Tn5 with a precise genomic region results in reads aligning to two converging genomic coordinates 9 bases apart on opposite strands. Therefore the field shifts the reads aligning to either reference genome strand so that a common position anchors the Tn5 recognition site regardless of the orientation of the adapter. We refer to the center nucleotide of this 9 base insertion site as the **central** Tn5 recognition **base**, in contrast to a traditional nuclease which cleaves a single position between two adjacent bases. Shifting ATAC-seq reads results in base-pair resolution data that precisely identifies the center of the Tn5 recognition site.

Detection of a single Tn5-inserted adapter necessitates an independent Tn5 insertion within approximately 500 bases. Only half of the Tn5 insertion events are capable of amplification, even when an independent insertion occurs within 500 bases (Figure 2.1B,C,D). To specify this central base, we shift the forward-strand aligned reads downstream by 4 bases and we shift the reverse-strand aligned reads upstream by 4 bases (Figure 2.1E). Most current ATAC-seq analysis workflows shift the reverse-strand aligned reads upstream by 5 bases so that the data looks more similar to the chromatin accessibility predecessor DNase-seq. However, this approach requires an independent shift of one base for reverse-aligned sequences to ensure that a single DNA cleavage event is specified by a single genomic coordinate (Public Twitter correspondence: `https://twitter.com/jeffvierstra/status/1396900282634625025`).

All enzymes that recognize nucleic acids as substrates have sequence biases. These sequence biases are quantified as the observed frequency of each base at individual positions relative to a cleavage site (or the centrally recognized base in the case of transposases) compared to the expectation of random genomic cleavage. Enzyme bias motifs are represented by a position probability matrix that reports the fraction of each base observed at each position relative to the cut site, which is most easily visualized as a seqLogo (Schneider and Stephens 1990; Gavin E. Crooks and Brenner 2004). We confirm, as others have shown, that Tn5 favors CG-rich DNA and the seqLogo is a reverse complement palindrome, which is because Tn5 recognizes DNA as a dimer (Figure 2.2A) (Buenrostro, Giresi, et al. 2013; Vinayak, Vinay, and Shiv 2019; Z. Li et al. 2019; Bentsen et al. 2020). The nucleases Benzonase, Cyanase, DNase, and MNase have distinct biases that span fewer positions compared to Tn5 (Figure 2.2A) (Grøntved et al. 2012; Lazarovici et al. 2013; Iwata-Otsubo et al. 2017). In addition to Tn5 having a wider bias window, the cumulative information content within the window is highest for Tn5 compared to the nucleases (Figure 2.2A). Importantly, we consider background nucleotide frequency when calculating information content. If we did not background-correct, then random cleavage would be interpreted as an AT bias because A/T bases account for 59% of the genome (Figure 2.3A). Previous methods that accurately correct enzyme sequence biases use k-mer scaling (Martins et al. 2018). However, k-mer scaling does not

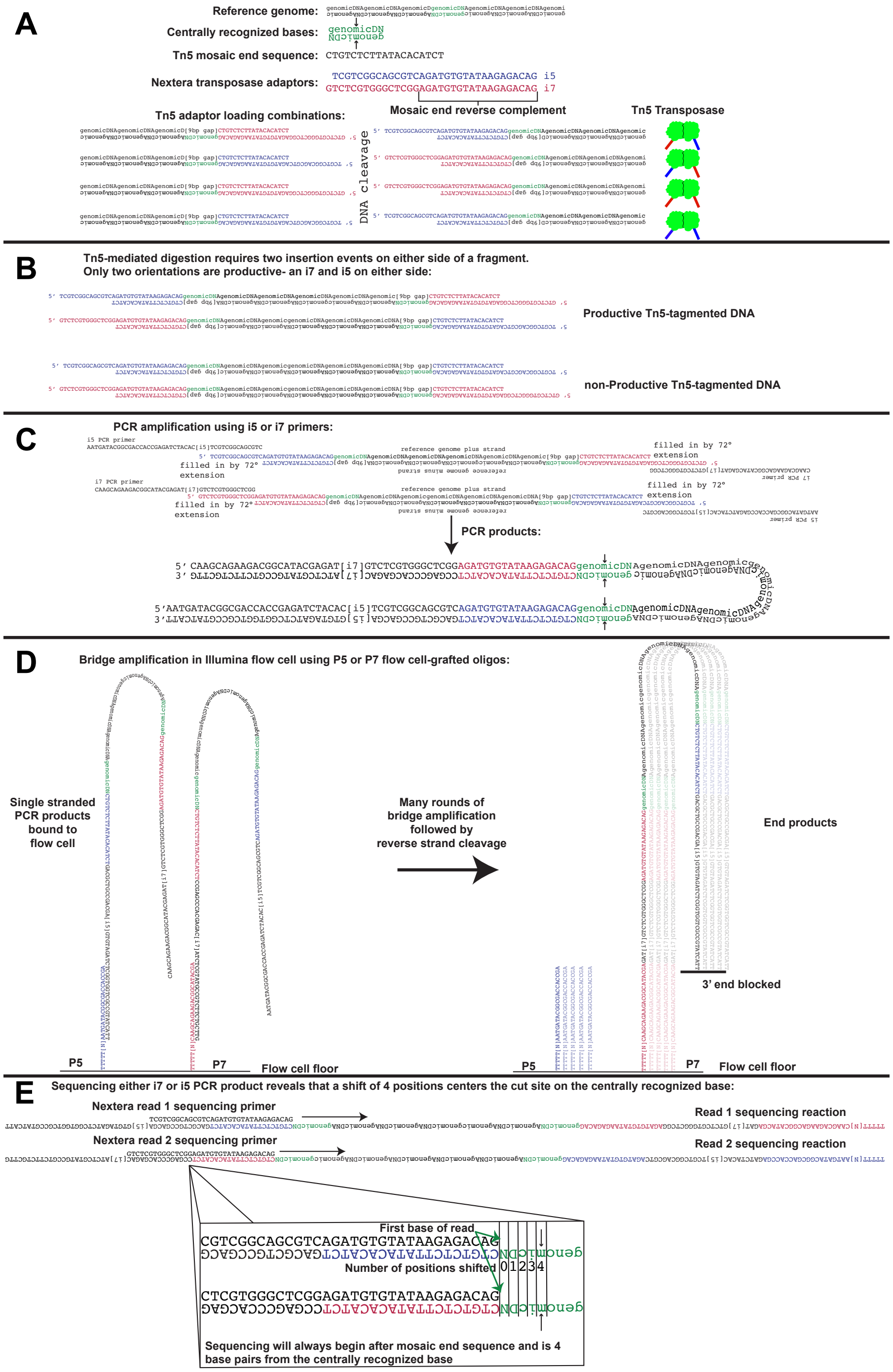

A
Reference genome:
Centrally recognized bases:
genomicDN
Tn5 mosaic end sequence:
CTGTCTCTTATACACATCT
Nextera transposase adaptors:
TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG i5
GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG i7
Mosaic end reverse complement
Tn5 adaptor loading combinations:
Tn5 Transposase
DNA cleavage
B
Tn5-mediated digestion requires two insertion events on either side of a fragment.
Only two orientations are productive- an i7 and i5 on either side:
Productive Tn5-tagmented DNA
non-Productive Tn5-tagmented DNA
C
PCR amplification using i5 or i7 primers:
i5 PCR primer
AATGATACGGCGACCACCGAGATCTACAC[i5]TCGTCGGCAGCGTC
i7 PCR primer
CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG
reference genome plus strand
reference genome minus strand
filled in by 72° extension
PCR products:
5' CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGgenomicDNAgenomicDNAgenomic
5'AATGATACGGCGACCACCGAGATCTACAC[i5]TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGgenomicDNAgenomicDNAgenomicDNA
D
Bridge amplification in Illumina flow cell using P5 or P7 flow cell-grafted oligos:
Single stranded PCR products bound to flow cell
Many rounds of bridge amplification followed by reverse strand cleavage
End products
3' end blocked
P5
P7
Flow cell floor
P5
P7
Flow cell floor
E
Sequencing either i7 or i5 PCR product reveals that a shift of 4 positions centers the cut site on the centrally recognized base:
Nextera read 1 sequencing primer
TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
Read 1 sequencing reaction
Nextera read 2 sequencing primer
GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG
Read 2 sequencing reaction
First base of read
CGTCGGCAGCGTCAGATGTGTATAAGAGACAG
genomicDN
Number of positions shifted
0 1 2 3 4
CTCGTGGGCTCGGAGATGTGTATAAGAGACAG
genomicDN
Sequencing will always begin after mosaic end sequence and is 4 base pairs from the centrally recognized base

Figure 2.1: Caption on next page

Figure 2.1: **Shifting forward and reverse aligned ATAC-seq data reads by 4 bases captures the center of the Tn5 duplication.** A) Each monomer of the Tn5 transposase dimer is loaded with one of two Nextera adapter sequences. During an insertion event, each adapter is attached to one side of the duplicated 9 base pairs. Each adapter sequence is a concatenation of the Tn5 hyperactive mosaic end sequence and either an i7 or i5 sequencing primer. Each Tn5 dimer is capable of four possible DNA insertion events, based on which adapters are loaded into each Tn5 monomer. B) For a fragment to be sequenced, it must have an i7 and i5 primer on either side. Two possible combinations result in non-productive fragments. C) i5 and i7 overhangs are first filled in by a 72° extension step. This filled in sequence is then used as a template for PCR amplification of fragments. D) To begin bridge amplification, PCR products are first denatured into single strands. Next, these single strands are flowed over a lawn of oligos grafted to the flow cell floor which have a complementary sequence that anneals to the P5 or P7 sequences of the index 1 (i7) or index 2 (i5) PCR primers. Once annealed, complementary strands are then polymerized and the original binding strand washed away after being denatured. The single stranded, newly polymerized sequence is then able to bind to a complementary oligo grafted to the flow cell, and the complementary strand is polymerized and subsequently denatured. This process repeats many times to create clonal colonies of each read. Finally, the reverse strands are cleaved and washed away and $3'$ ends of bound reads are blocked. The initial binding and end product stages of bridge amplification are depicted in the panel. E) Sequencing primers are used to determine the sequence downstream of the insertion event. Regardless of orientation, the first base of every read (the end of the adapter sequence) will be 4 positions away from the centrally recognized base.

effectively correct Tn5 bias because of the wide bias window and the limitation that large k-mers occur rarely in both the observed and expected k-mer counts.

If nucleotide sequence far from the Tn5 recognition site center influences cleavage and insertion, then we would expect that k-mers found at positions distal from Tn5's centrally recognized base would deviate substantially from random expectation. We quantified the influence of distal and proximal k-mers on cleavage bias by plotting the $\log_2$ ratio of observed 5-mer frequency to expected 5-mer frequency for all 1024 5-mers; expected frequency is based on the genomic frequency of each 5-mer. The $\log_2$(observed/expected) ratio is zero if a 5-mer is observed at the expected frequency. We find that the distribution of this ratio becomes more tightly centered around zero as we query more distal 5-mers (Figure 2.2B). The distribution does not stabilize for Tn5 until we exceed 11 bases from the Tn5 recognition site, while the other four enzymes stabilize within 4 bases of the cleavage event. This analysis confirms that distal sequences more strongly influence Tn5-mediated cleavage compared to the other enzymes, and indicates that a k-mer spanning at least 21 base pairs is required to model Tn5 bias effectively.

The background corrected sequence motif (Figure 2.2A) for Tn5 is a reverse palindrome, which is common for homodimeric DNA binding factors (Welboren et al. 2009; Sasse et al. 2017). The consensus Tn5 recognition sequence contains CAG trimers present on each strand of the DNA, each CAG is displaced from the adjacent reverse complemented CAG by 5 bases. This unique feature suggested that Tn5 dimers may interact with this sequence motif in multiple orientations to direct the transposition of adapters with higher frequencies. We plotted the read-depth normalized composite profile of each enzyme's bias motif to compare their relative strength of directing DNA cleavage (Figure 2.2C). The Tn5 recognition site directs a central peak that is substantially higher than the other enzymes. The composite profiles also highlight peaks 5 and 10 base pairs downstream and upstream of the centrally recognized base (Figure 2.2C). These flanking peaks are not observed in the other nucleases, although they are comparable in intensity as the central peak of other enzymes. Although not definitive, this pattern is consistent with Tn5 dimers interacting with opposite sides of the DNA helix, displaced by five bases. Taken together, the length of the Tn5 bias motif, the strong influence of distal k-mers, the stronger cutting preference, and 5-base spaced cutting pattern all preclude conventional k-mer based approaches for correcting Tn5 sequence bias.

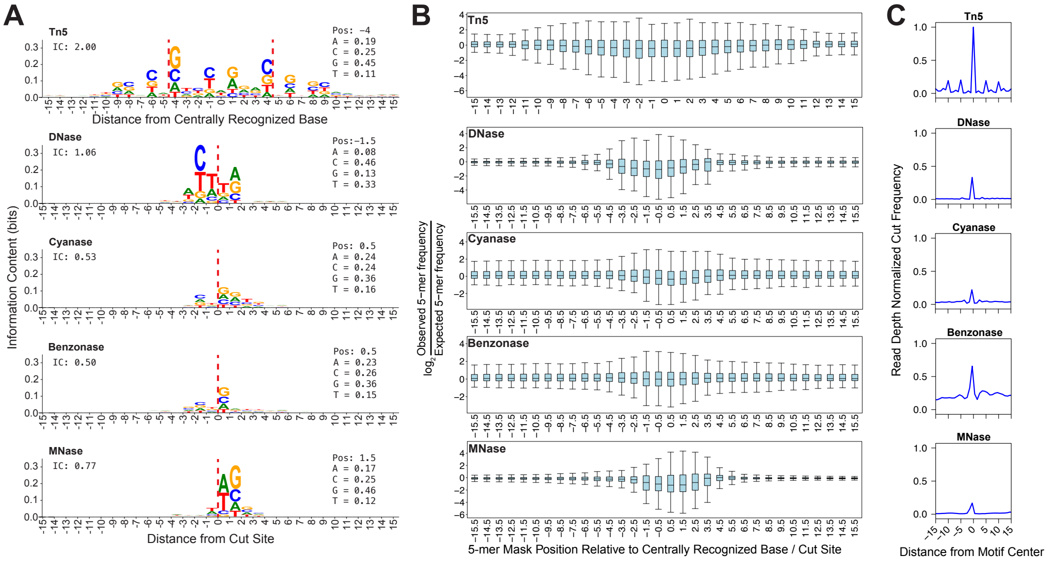


Figure 2.2: **Tn5 sequence bias is more complex than nuclease sequence bias.** A) The seqLogo sequence bias motifs are corrected for background nucleotide content and they illustrate that Tn5 bias is wider and more complex than other commonly used chromatin accessibility enzymes. Nucleotide frequencies listed in the inset to the right correspond to the highest information content position; for instance, G is found at position -4 45% of the instances that Tn5 inserts into DNA. Nuclease cleavage sites are indicated by dashed red line. Total information content (IC) from positions -10 to 10 is listed in the inset to the upper left. B) We plotted the $\log_2$(observed/expected 5-mer frequency) for all positions surrounding DNA cleavage sites or Tn5 recognition sites as box and whisker plots. C) We plotted read depth normalized composite signal from various molecular genomics assays for the top 400,000 sites that conform most stringently to the respective enzyme's bias motif.

2.1.3.2 Tn5 sequence bias is complex and not modeled well by previous methods

A key assumption of representing sequence biases as probability matrices or seqLogos is that any nucleotide observed at a position within the motif does not influence the probability of observing nucleotides in other positions, this property is referred to as positional independence (Guertin et al. 2012; Sharon, Lubliner, and Segal 2008). If each position is independent from the others, observed k-mer

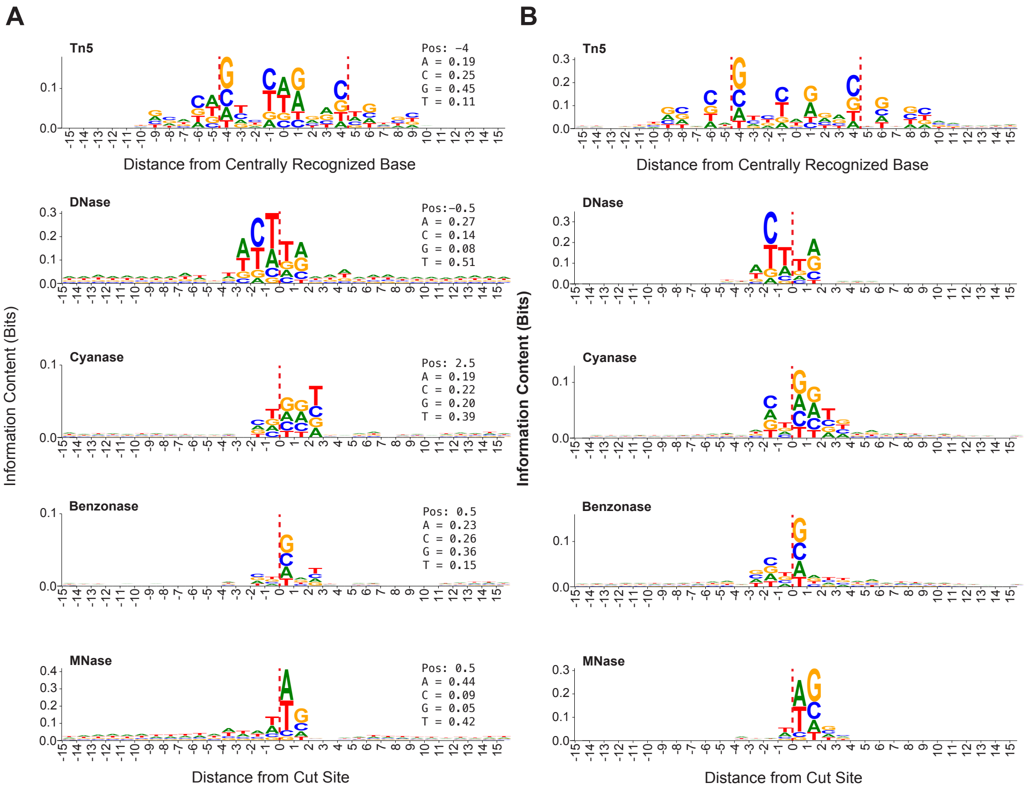


Figure 2.3: **Tn5 sequence bias motif is more complex than nucleases.** A) These sequence logos (Schneider and Stephens 1990; Gavin E. Crooks and Brenner 2004) representing enzyme biases were generated with equiprobable background nucleotide frequencies, which does not incorporate information about genomic AT/CG content. Nucleotide frequencies listed in the inset to the right correspond to the highest information content position (Pos). Nuclease cleavage sites are indicated by the dashed red line. B) Plots of enzymatic sequence bias motifs as seen in (Figure 2.2A), with the y-axis limit scaled to the position with highest information content.

frequencies should be equal to those expected based on nucleotide frequency in the position probability matrix. To test Tn5 sequence bias position independence, we determined the observed frequency of each 3-mer found in position -6 to -4 relative to the central Tn5 recognition base and divided by the expected frequency (Figure 2.4A). We find that 3-mers are observed at frequencies that deviate from the calculated expectation. We can infer basic features describing which 3-mers are disfavored (NAA) and preferred (NCA), but characterizing a comprehensive set of basic rules for many k-mer combinations and k-mer positions is not straightforward. However, since k-mer scaling does consider dependencies between positions, k-mer scaling remains an attractive starting point for developing more complex bias correction models.

Next we tested the feasibility of using rationally spaced k-mers to systematically reduce Tn5 bias. The 3-mer ”CAG” is overrepresented in three positions in the Tn5 bias and each CAG bias is displaced by 5 bases (Figure 2.2A). We measured the composite ATAC-seq signal in the window centered on 400,000 random CAG instances in the hg38 genome assembly. As expected, we observe that the CAG 3-mer directs cutting in three peaks, each displaced from one another by 5 bases (Figure 2.4B). Conventional k-mer scaling is performed with any reasonably short k-mer size (<9 bases) and the k-mer can be positioned at any distance from the observed Tn5 recognition site. In an effort to determine whether k-mer scaling would effectively flatten these three peaks, we used `seqOutBias` to scale reads using 3-mers that are located at the Tn5 recognition site and up/down stream by 5 bases. Recall that each peak corresponds to a peak of Tn5 recognition and that each peak is relative to the CAG 3-mer that anchors the plot at position x = 0. Therefore, we expect that scaling based on the k-mer located 5 bases upstream will flatten the +5 peak because the CAG directing this downstream peak is located 5 bases upstream. Likewise, the central k-mer scaling would abolish the central peak and downstream k-mer scaling would ablate the upstream peak. This exercise indicates that we can rationally design k-mer scaling to correct known biases, but it also highlights the fact that multiple k-mers are needed to correct biases even within the context of this simple example. Therefore, we pursued a rule ensemble modeling approach to integrate multiple k-mer sizes from a range of positions relative to the Tn5 recognition site as input data and determine whether interactions between these inputs contribute to Tn5 insertion bias.

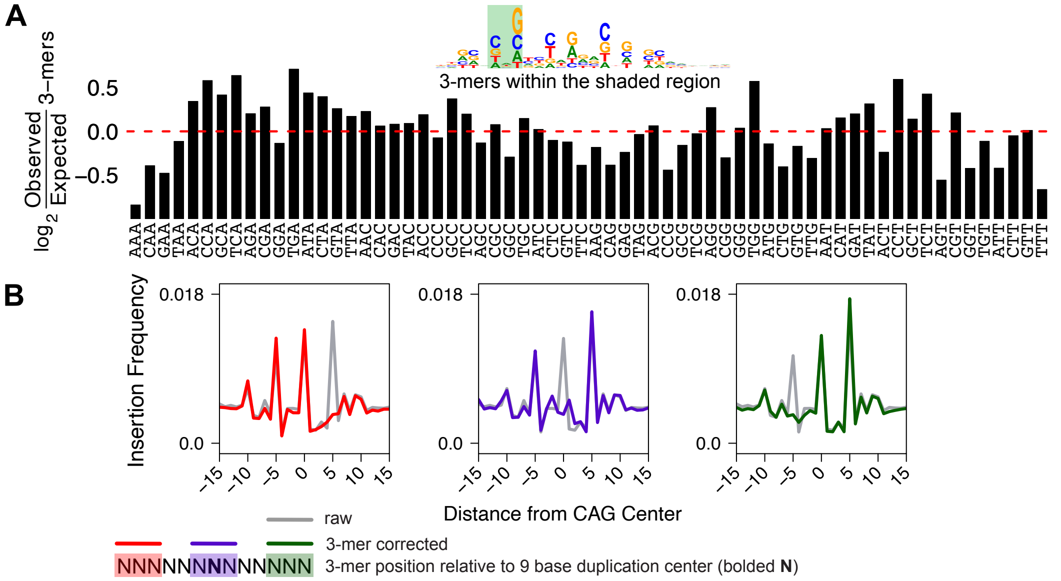
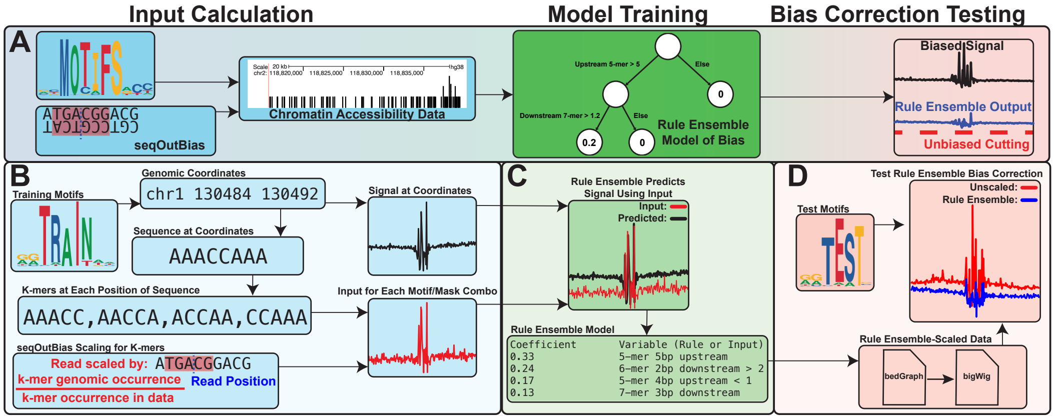


Figure 2.4: **K-mers capture more bias complexity than weight matrices and provide a basis for which to build more sophisticated bias-correction models.** A) We quantified the frequency that all 64 3-mers occur at positions -4 to -6 from the center of a Tn5 recognition site and compared observed frequency to the position weight matrix prediction. A bar chart of $\log_2$ observed divided by expected (from k-mer prevalence in the bias position probability matrix) k-mer frequency highlights the interdependency between sequences are different positions. B) We plotted the signal from 400,000 randomly selected ”CAG” instances in the genome. Overlaid on top of the unscaled signal, each plot shows that the individual peaks from composite profile can be corrected by rationally designed spaced k-mer scaling, based on the position of the scaling k-mer and the peak.

2.1.3.3 Rule ensemble modeling leverages interaction terms and k-mer scaling to correct sequence biases

We previously developed `seqOutBias`, which directly scales chromatin accessibility data by expected/observed k-mer frequency. Although we specify `seqOutBias` to scale based on a 6-mer centered on the DNA cleavage event for DNase, one can specify any k-mer length and relative k-mer position. Our data indicate that Tn5 bias is too broad to scale with a single k-mer and this suggests that bias correction would need to incorporate information from several k-mer positions. Therefore, we chose to use `seqOutBias` scaling outputs with different invocations of k-mer positions and lengths as the input for the rule ensemble model, and to test its capacity to correct Tn5 bias (Figure 2.5A). Sequence biases were first characterized by quantifying enzymatic cut frequency in average composite profiles that are centered on transcription factor binding motifs (He et al. 2014; Sung et al. 2014). We previously developed a model (seqToSign) that reproduced these composite profiles by scaling the genome-wide cut frequency of each k-mer found in a given position in the composite by the number of occurrences of the k-mer at that position (Sung et al. 2014). We attempted a similar approach to predict biased composites by using the inverse of `seqOutBias` scale factors as an input, since this value represents the genomic cut frequency for a k-mer. For each k-mer we multiplied its frequency at each position in the composite by the inverse of that k-mer's corresponding `seqOutBias` scaling factor. We summed these values for all 4^k k-mers at each position in the composite profile (Figure 2.5B). Each of these values at each position in the composite is a covariate input for the rule ensemble model and we repeated the process for many k-mer length and relative k-mer position combinations (Figure 2.6A). We trained a rule ensemble model using these covariate inputs to predict multiple transcription factor composite profiles. The output of this trained model is the formulaic combination of k-mer positions using linear regression coefficients and decision trees (Figure 2.5C). To test the efficacy of the model, we first combined the scaled read values from each k-mer covariate according to the output formula. This generated a single rule ensemble-scaled read file which we tested for its ability to correct test group composite profiles to resemble theoretically unbiased genomic cleavage (Figure 2.5D).



```
Coefficient        Variable (Rule or Input)
0.33               5-mer 5bp upstream
0.24               6-mer 2bp downstream > 2
0.17               5-mer 4bp upstream < 1
0.13               7-mer 3bp downstream
```

Figure 2.5: **Rule ensemble modeling of enzymatic bias combines k-mer scaling approaches to enhance bias correction.** A) We generate the rule ensemble input by combining seqOutBias scaling output with k-mer frequencies observed at transcription factor motifs found throughout the genome. This input is then used to train a rule ensemble model to predict the enzymatic sequence bias. Finally, the model is tested for its ability to correct this bias. B) For each training motif, we identify the 400,000 occurrences in the genome that conform most stringently to the weight matrix. We extract sequences (200 bases) in the region for each identified motif and calculate the frequency of each k-mer at each position relative to the motif center. We perform independent runs of seqOutBias to calculate scale factors for each k-mer in each position within a defined window from the center of the motif. K-mer frequency for each position is then multiplied by the inverse scale factor for every frame of reference. These are the rule ensemble input values for each motif/k-mer (size and position) pairing. C) We train a rule ensemble model to predict the biased signal measured at each of these region sets in a deproteinized data set. D) Output from seqOutBias corresponding to all input frames of reference are then combined according to the rule ensemble model to generate a single BED or bigWig file with bias-corrected values for each sequence read. Successful bias correction is evaluated using a held-out test set of TF motifs.

2.1.3.4 Choosing training and test transcription factor motifs

Transcription factors recognize a diversity of DNA sequences that vary by length, degeneracy, and nucleotide content. We chose training and test data for

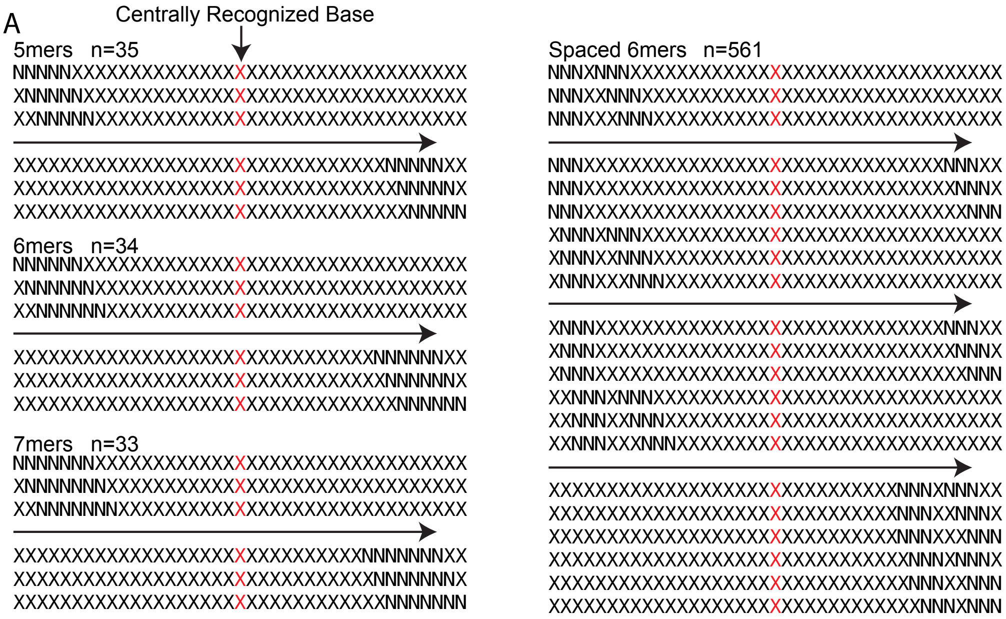

A
Centrally Recognized Base
5mers n=35
NNNNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XNNNNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXNNNNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNNNNNXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNNNNNX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNNNNN
6mers n=34
NNNNNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XNNNNNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXNNNNNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNNNNNNXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNNNNNNX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNNNNNN
7mers n=33
NNNNNNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XNNNNNNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXNNNNNNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNNNNNNNXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNNNNNNNX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNNNNNNN
Spaced 6mers n=561
NNNXNNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
NNNXXNNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
NNNXXXNNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
NNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNNNXX
NNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNNNX
NNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNNN
XNNNXNNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XNNNXXNNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XNNNXXXNNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XNNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNNNXX
XNNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNNNX
XNNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNNN
XXNNNXNNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXNNNXXNNNXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXNNNXXXNNNXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNNNXNNNXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNNNXXNNNX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNNNXXXNNN
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNNNXNNNX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNNNXXNNN
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXNNNXNNN

Figure 2.6: **Many 5/6/7-mer and spaced 6-mer combinations were inputs to rule ensemble modeling of Tn5 bias.** A) This graphic depicts direct k-mer scaling positions used as input for the Tn5 rule ensemble model. In this example, the positions marked 'X' represent unmodeled positions, while the ones marked 'N' are the positions which each read is scaled by. The 'n' in the upper left corner is the number of permutations per k-mer size. Each graphic for each contiguous k-mer size represents the first three and last three input positions. The graphic for spaced 6-mer positions shows how the leading 3-mer is moved across the span, followed by a single base pair movement of the trailing 3-mer, to sample all possible combinations of spacing between the two.

the model by selecting sequence motifs based on these features (Figure 2.7A). In a sequence logo representation of DNA binding preference, the information content is a function of sequence length and degeneracy. We used hierarchical clustering of these values for 43 input motifs that represent a wide range of information content and GC content, encompassing well-characterized binding motifs (Figure 2.7B). Groupings consisted of either doublets or triplets, with a single singlet. Within a group, a single motif was assigned as testing data and the others were assigned as training data.

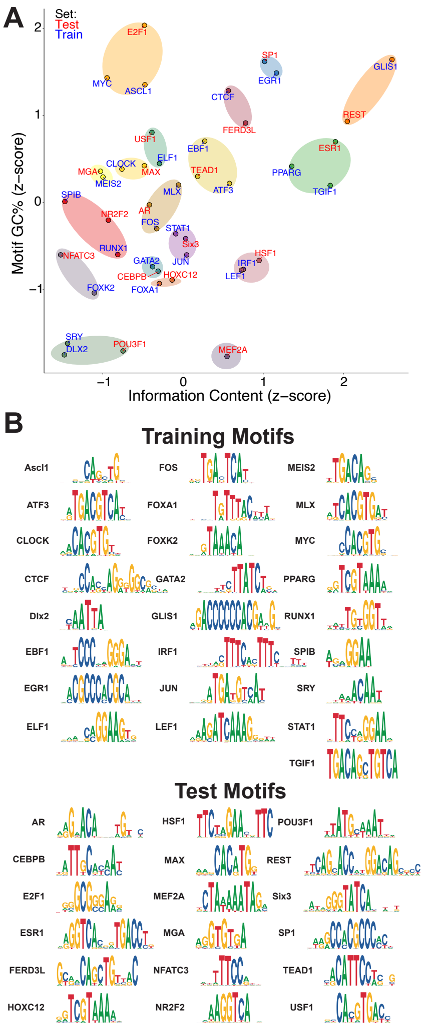


Figure 2.7: **Clustering of transcription factor motifs creates diverse test and training sets.** A) We clustered transcription factor motifs into test and training sets based on motif information content and GC content. We indicate clusters with transparent ovals. Training and test sets are indicated as red or blue text that indicates the motif name. B) The seqLogos of training and test transcription factor motifs illustrate their diversity.

2.1.3.5 Rule ensemble modeling corrects DNase sequence bias

We chose rule ensemble modeling because the importance of covariate inputs are easily interpreted in the overall model and for any individual model prediction. Since DNase-seq biases are well-characterized, we have an expectation of the importance of positions that lead to DNase specificity. We chose to test the feasibility of this model using naked DNA DNase-seq data. We invoked `seqOutBias` with staggered 5-mers within 9.5 base pairs of the cut site as input for the model. The DNase rule ensemble model confirmed that the 3 bases on either side of the cleavage event most strongly influence DNase cleavage (Figure 2.8A).

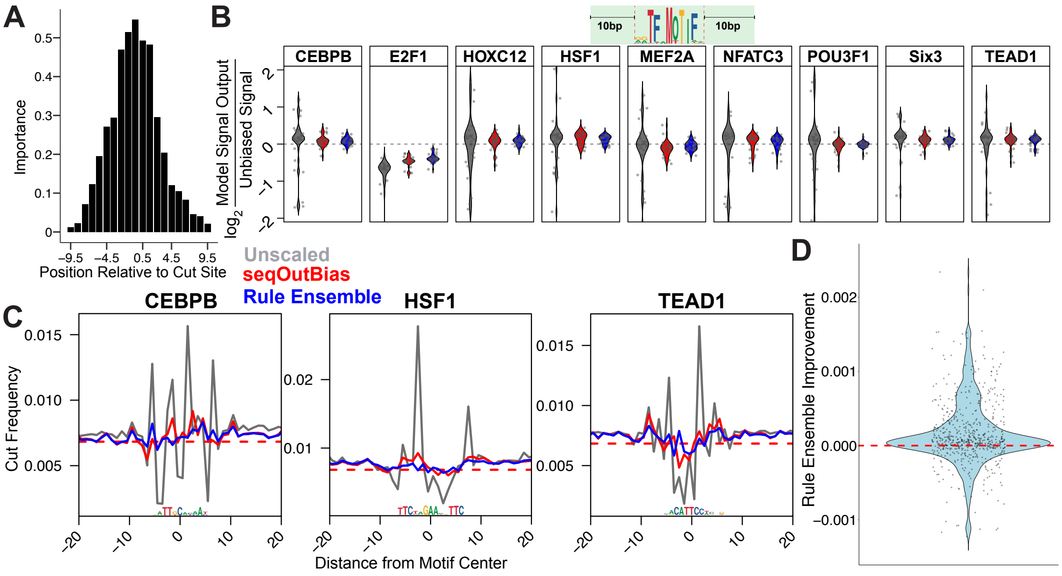


Figure 2.8: **Rule ensemble modeling effectively corrects DNase bias.** A) Rule ensemble modeling provides information about the importance of variable in the model. The positions proximal to DNase cleavage are most important. B) Violin plots quantify the $\log_2$ ratio of unbiased signal to output for each transcription factor's composite profile given the method of k-mer scaling. Regions measured in each plot are within ±10 base pairs of each motif, as indicated by the graphic above the figure panel. C) We plotted composite profiles of transcription factors from rule ensemble output compared with direct k-mer scaling of the 5-mer that encompasses the five most influential positions in panel A. D) We visualized improvement of rule ensemble modeling over seqOutBias by using a violin plot of distance from calculated random cut frequency for rule ensemble output subtracted from seqOutBias. Positions which rule ensemble outperforms `seqOutBias` will be above 0 (dashed red line); 68% of positions in this plot were improved.

We visualized the correction of single nucleotide bias by calculating the $\log_2$ ratio of theoretically unbiased signal to either unscaled, direct k-mer scaling (se-

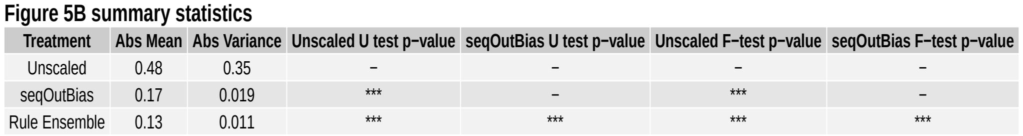


Figure 5B summary statistics

Treatment	Abs Mean	Abs Variance	Unscaled U test p−value	seqOutBias U test p−value	Unscaled F−test p−value	seqOutBias F−test p−value
Unscaled	0.48	0.35	–	–	–	–
seqOutBias	0.17	0.019	***	–	***	–
Rule Ensemble	0.13	0.011	***	***	***	***

Table 2.1: **Statistical tests for Figure 5B and Figure S4A highlight the improvement rule ensemble correction provides.** The absolute value mean and variance were determined for all data of a given treatment of DNase single nucleotide bias correction, as displayed in Figure 5B and S4A. The smaller the absolute $\log_2$ values of variance and mean are, the better bias correction is performing. A value of 0 would indicate perfect bias correction. Q-Q plots indicated that these data are not normally distributed, therefore Mann-Whitney U tests were used to compare differences in mean. F-tests were used to show differences in variation between the values output from the scaling methods. P-values for U and F-tests are indicated by asterisks: ***:p<0.0005, **:p<0.005, *:p<0.05

qOutBias), or rule ensemble output within the region of ±10 base pairs for each motif (Figure 2.8B & Figure 2.9A). Our results indicate that direct k-mer scaling corrects DNase bias well, but rule ensemble modeling outperforms k-mer scaling (Table 2.1). This is most easily illustrated by observing individual composite profiles (Figure 2.8C & Figure 2.9B). Although improvement is comparable between direct k-mer scaling and rule ensemble modeling, we sought to directly compare the two more rigorously. We calculated the absolute value of the difference between each corrected value and random cleavage for each point within 10 base pairs of each motif for both bias correction methods. This metric represents how much the values deviate from perfect bias correction for each method (`seqOutBias` and rule ensemble). For each position, we subtracted the rule ensemble deviation from unbiased cleavage value from the `seqOutBias` deviation from unbiased cleavage and plotted the difference (Figure 2.8D). Each point represents the difference in improvement between `seqOutBias` and rule ensemble modeling. For each point greater than 0, the rule ensemble modeling outperforms `seqOutBias` (Figure 2.8D).

We find that 68% of all points showed improvement, indicating that rule ensemble modeling generally outperforms direct k-mer scaling for correcting DNase bias.

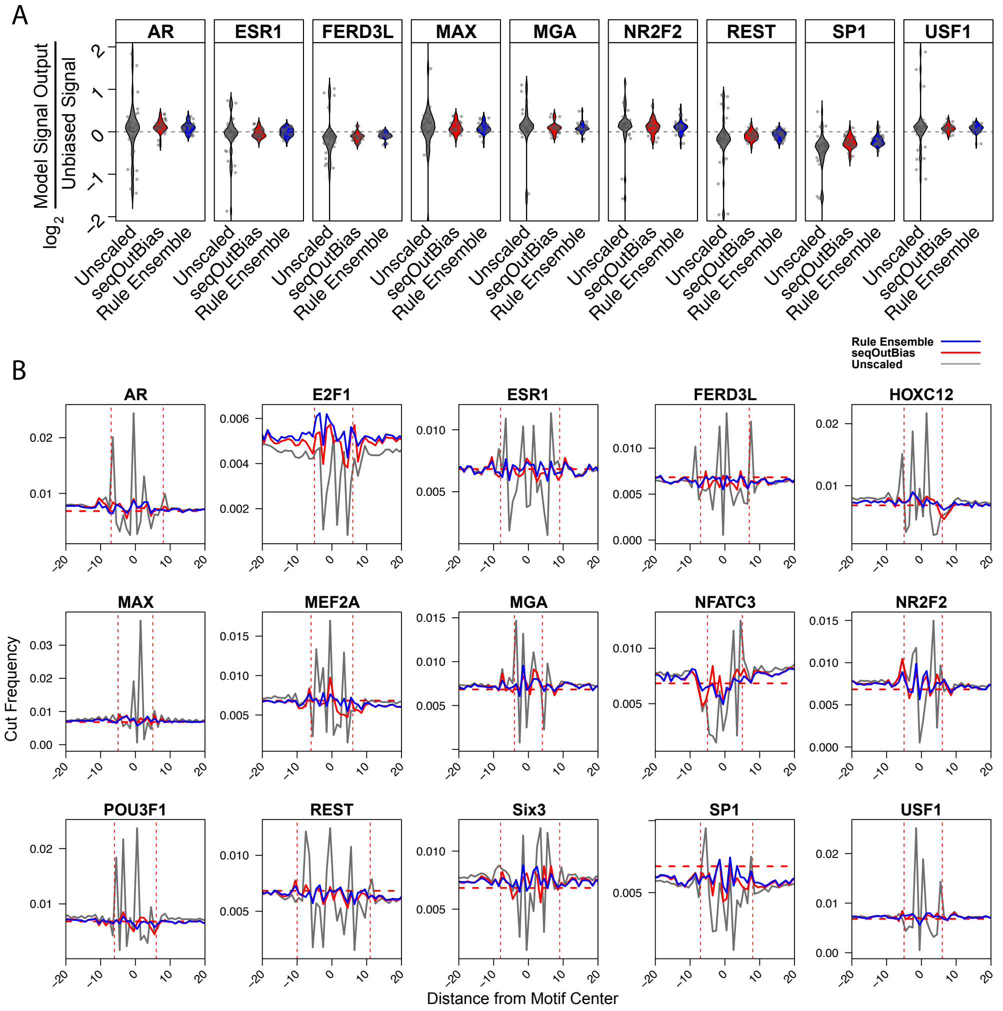


Figure 2.9: **Rule ensemble modeling corrects DNase bias.** A) The remaining violin plots which compare the $\log_2$ of output divided by unbiased signal for each test transcription factor's composite profile. B) Composite profiles of the test set transcription factors show correction of DNase sequence bias. Each plot shows rule ensemble output compared with direct k-mer scaling of the 5-mer that encompasses the five most influential positions in Figure 2.8A. The horizontal dashed red lines indicate random cleavage, while the vertical dashed red lines indicate the beginning and end of the motif.

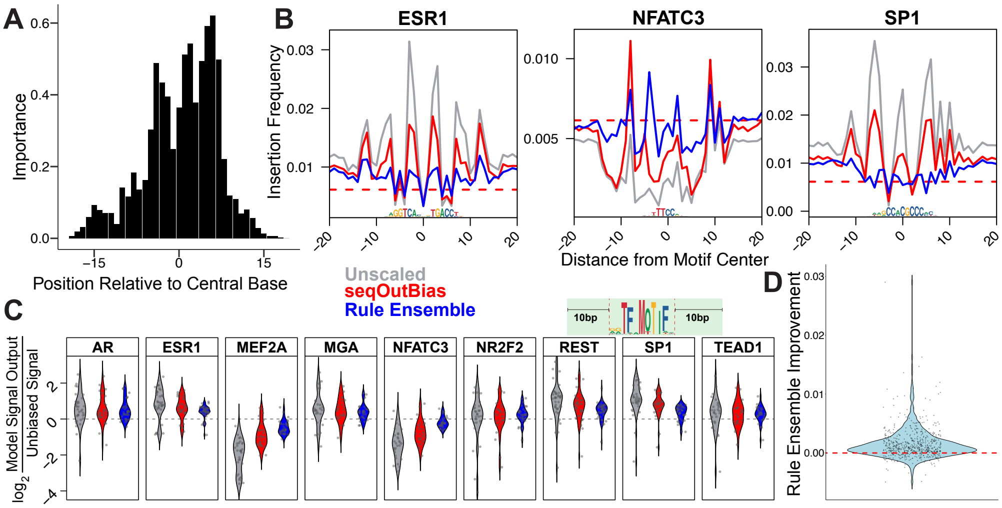


Figure 2.10: **Rule ensemble modeling corrects Tn5 transposon single nucleotide bias in ATAC-seq data.** A) The importance values for each position relative to Tn5's recognition site mirror the information content values of positions in the seqLogo representation of Tn5 bias. B) We directly compare bias correction of Tn5 data by `seqOutBias` and rule ensemble modeling by visualizing the composite corrected signals and comparing to the unscaled composite traces. C) We quantify the single nucleotide correction for each transcription factor's composite profile by measuring divergence from unbiased signal. Regions measured are within ±10 base pairs of each motif, as indicated by the graphic above the figure panel. D) We visualized improvement of rule ensemble modeling over `seqOutBias` by using a violin plot of distance from calculated random cut frequency for rule ensemble output subtracted from `seqOutBias`. Positions which rule ensemble outperforms `seqOutBias` will be above 0 (dashed red line); 78% of positions in this plot were improved and those that are not improved have lower magnitude deviations from the expectation of random cleavage. E) Composite profile overlays of test motifs highlight the improvement of rule ensemble modeling compared to `seqOutBias` correction. The red dashed line indicates the calculated random cleavage frequency.

2.1.3.6 Rule ensemble modeling corrects local Tn5 transposition bias

Since this modeling approach corrected DNase bias, we developed a rule ensemble model to correct bias from deproteinated ATAC-seq data. Since Tn5 bias is more complex than DNase, we ran `seqOutBias` for all contiguous 5-mers, 6-mers, 7-mers, and spaced 6-mers within 19 bases of the DNA cleavage/insertion site for the model input. In total, we included 663 distinct k-mer combinations as inputs (Figure 2.6A). The number of input variables was much higher compared to DNase, so we prioritized the input covariates by first modeling the data using linear regression to determine the most influential k-mers (formalized calculations are included in the Methods), then we incorporated the most influential 10% of these variables into a full rule ensemble model to predict the biased signal for each training motif (Figure 2.5C). The relative importance values for each nucleotide position revealed three central peaks that are separated by 5 bases (Figure 2.10A). Comparison of individual composite profile traces highlights the improvements (Figure 2.10B & Figure 2.11A&B). We systematically measured how bias-corrected signal deviated from random DNA cleavage at sequence motifs (± 10 bases) and compared to unscaled and k-mer scaled correction (Figure 2.10C). The rule ensemble bias correction outperforms k-mer scaling with respect to reducing variance and scaling reads to more accurately approximate random DNA cleavage (Table 2.2). For each individual position, the rule ensemble model outperforms k-mer scaling 78% of the time (Figure 2.10D). Examination of the 22% of positions with worse performance reveals the magnitude by which the model is outperformed is modest compared to the magnitude of improvement for the other 78% of instances. This can be visualized by comparing the distribution above zero (improvement) versus below zero (outperformed by k-mer) in Figure 2.10D. Visual inspection of the composite traces of the 18 test motif composite profiles illustrates that rule ensemble models effectively flatten these profiles (Figure 2.10B & Figure 2.11B). The traces approach the theoretical random cleavage red dotted line more closely than k-mer scaling alone (Figure 2.10B & Figure 2.11B). Therefore, the local sequence bias, which we define as the region within 10 bases of sequence motifs, is more effectively corrected with rule ensemble modeling.

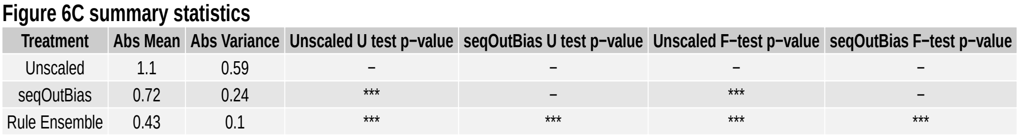


Figure 6C summary statistics

Treatment	Abs Mean	Abs Variance	Unscaled U test p−value	seqOutBias U test p−value	Unscaled F−test p−value	seqOutBias F−test p−value
Unscaled	1.1	0.59	–	–	–	–
seqOutBias	0.72	0.24	***	–	***	–
Rule Ensemble	0.43	0.1	***	***	***	***

Table 2.2: **Statistical analysis of the metrics in Figure 6C and Figure S5A highlight the improvement that rule ensemble modeling provides.** For all motifs and positions within the motifs (±10bp), we $\log_2$ transformed the scaled values, converted to their absolute values, and reported the resultant means and variances of each scaling method from the data in Figure 6C and Figure S5A. The closer values are to 0 for each group, the more effective that treatment is at correcting Tn5 single nucleotide bias. The distribution of values for each method was not normal, so we used the Mann-Whitney U test for comparison. We compared the variances using an F-test. P-values for Mann-Whitney U and F-tests are indicated by asterisks. ***:p<0.0005, **:p<0.005, *:p<0.05

2.1.3.7 Rule ensemble modeling corrects regional sequence biases caused by GC-content

We developed these models and measured correction of enzyme biases using these composite motif profiles because this visual representation is an intuitive way to observe enzyme biases in genomic assays. As we move away from the sequence ”anchor” in DNase-seq composite profiles, the traces flatten and approach the expectation of random cleavage because the sequence content in these regions is more random (Figure 2.12A). Unlike DNase-seq, the composite profiles for ATAC-seq do not approach random expectation, even at distances 100 bases from the sequence motif center. The Tn5 recognition site is GC-rich (Figure 2.2A) and it is known that AT/GC richness is clustered throughout the genome (International Human Genome Sequencing Consortium 2001). We hypothesized that Tn5 preference for GC-rich regions would lead to generally elevated signal within these regions, accompanied by depleted signal in AT-rich regions. Therefore, we determined whether there is

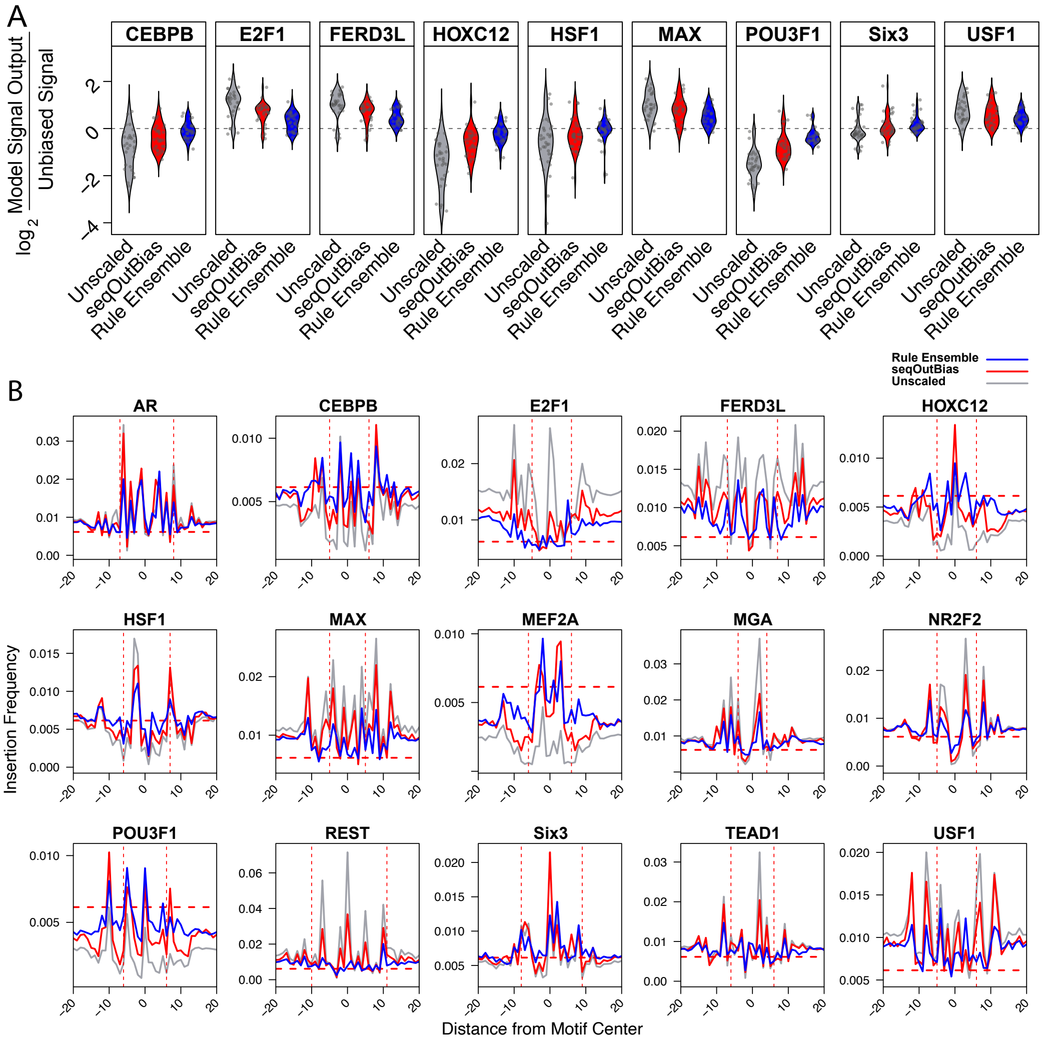


Figure 2.11: **Rule ensemble modeling corrects Tn5 bias.** A) The single nucleotide bias correction comparison of the $\log_2$ ratio of model output to unbiased signal for each transcription factor indicate that rule ensemble outperforms direct k-mer scaling. These motifs represent the remaining test set motifs that are not in Figure 2.10. B) The composite profiles of transcription factors from rule ensemble output compared with direct k-mer scaling highlight the improvement of rule ensemble modeling. Dashed horizontal red lines denote random cleavage and the dashed vertical lines bookend each transcription factor motif.

a relationship between motif GC content and this regional ”baseline signal”. Motif GC content linearly correlates with baseline signal (Figure 2.12B&C). Rule en-

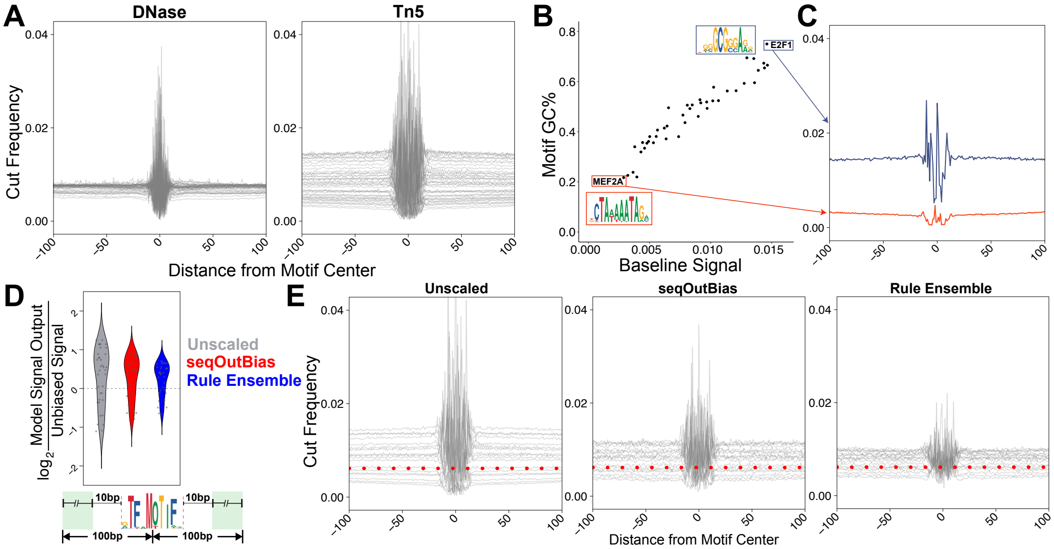


Figure 2.12: **Rule ensemble modeling corrects regional sequence biases caused by GC-content.** A) We illustrate the range of baselines for each enzyme by plotting unscaled overlays of DNase and Tn5 composite profiles for the 43 transcription factors surveyed. B) A plot of motif GC content vs. baseline signal indicates a linear and correlated relationship between the variables. C) The E2F1 and MEF2A motifs have very different GC content and this is reflected in their baseline signal within the composite traces. D) Baseline correction is measured by the ratio of unscaled, seqOutBias, and rule ensemble scaled reads to random cleavage. Two points for each motif are plotted; each point is the average of upstream or downstream signal in the window ±10 base pairs from the edge of the motif spanning 100 bases from the motif center (green box).

semble modeling outperforms `seqOutBias` for 88% of the positions when regional GC bias correction (Figure 2.12D&E) and scaling within the motif (Figure 2.10C) are considered together. We find that rule ensemble corrections improve scaling compared to seqOutBias in terms of both reducing baseline mean and variance (Table 2.3). Rule ensemble modeling effectively corrects Tn5's regional and local sequence biases.

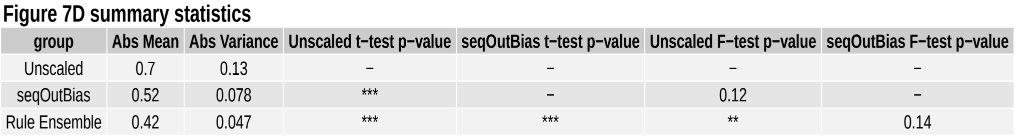


Figure 7D summary statistics

group	Abs Mean	Abs Variance	Unscaled t−test p−value	seqOutBias t−test p−value	Unscaled F−test p−value	seqOutBias F−test p−value
Unscaled	0.7	0.13	−	−	−	−
seqOutBias	0.52	0.078	***	−	0.12	−
Rule Ensemble	0.42	0.047	***	***	**	0.14

Table 2.3: **Statistical analysis of the data in Figure 7D indicates that rule ensemble modeling provides statistically significant improvement of regional sequence biases.** Mean and variance are reported for the absolute values of regional $\log_2$ Tn5 composite data displayed in Figure 7D. Perfect bias correction in this comparison is a value of 0 for both measurements. The data is normally distributed, so we performed t-tests to compare differences between means and F-tests were used to measure differences between variance. P-values for F- and t-test are indicated by asterisks. ***:p<0.0005, **:p<0.005, *:p<0.05

2.1.4 Discussion

Transcription factor binding sites and promoters are the most interesting regions of the genome with respect to gene regulation. ATAC-seq signal correlates with the regulatory potential of these regions. ATAC-seq is a simple experimental assay, but analysis of the data requires dedicated pipelines, specialized software, and unique considerations (J. P. Smith et al. 2021). Importantly, Tn5 sequence bias was described in the first ATAC-seq paper (Buenrostro, Giresi, et al. 2013) and many groups have developed methods to characterize and correct these biases. Approaches often combine bias correction and footprinting or propose models that cannot be easily interpreted. Here, we provide a workflow that directly scales individual ATAC-seq reads to correct Tn5 bias and provides common output files that can be used for peak calling or footprinting. The rule ensemble modeling approach that we employed is interpretable globally, in that we identify the positions relative to Tn5 recognition that influence the model. Moreover, for any individual scaled read the model is interpretable locally. If a rule does not apply to the local sequence read, then the rule drops out of the model and does not contribute to that individual instance of read scaling.

Enzymatic sequence preferences must be considered when interpreting molecular genomics data, particularly when the data is single-nucleotide resolution. We introduce this rule ensemble framework as a novel method that efficiently scales individual sequence reads to correct Tn5 bias.

2.2 Materials & Methods

2.2.1 Reproducible analysis and plotting code

We provide a detailed and reproducible vignette to reproduce all analyses and figure panels directly from raw data here: `https://github.com/guertinlab/Tn5bias/tree/master/Manuscript_Vignette`. We provide a workflow vignette using only chr21 ATAC-seq reads, the chr21 reference genome, and estrogen receptor motifs as a quick-start reference document: `https://github.com/guertinlab/Tn5bias/tree/master/seqOutATACBias_workflow_Vignette`

2.2.2 Chromatin accessibility data preprocessing and read alignment

We downloaded the hg38 and mm39 reference genomes from the UCSC genome browser (International Human Genome Sequencing Consortium 2001; Consortium 2002). We retrieved data sets for the respective nucleases and Tn5 from the NCBI SRA (sequence read archive), in fastq format, using `fasterq-dump` (Leinonen et al. 2010). The following SRA accession numbers were used for mouse (mm39) liver data: SRR535737, SRR535738, SRR535739, SRR535740, SRR535741, SRR535742, SRR535743, SRR535744 (Grøntved et al. 2012), SRR5723785 (Iwata-Otsubo et al. 2017). The following SRA accession numbers were used for human (hg38) DNase (lung fibroblast) SRR769954 (Lazarovici et al. 2013) and Tn5 (T lymphoblast) SRR5123141 (Martins et al. 2018). Reads from each data set were aligned to their respective reference genomes using `bowtie2` (Langmead and Salzberg 2012), then sorted and converted to BAM files with samtools (Danecek et al. 2021).

2.2.3 Determining sequences around each cleavage site

All aligned reads (plus, minus, and unseparated) were used individually as input for unscaled seqOutBias (Martins et al. 2018) runs, which generated bigWig and bed format output files. We deconvolved independent reads that aligned to the same genomic coordinate into separate bed file entries, then converted to sequence files using fastaFromBed (Quinlan and I. M. Hall 2010). Sequences corresponding to minus-aligned reads were reverse complemented before concatenating them with the sequences corresponding to plus aligned reads.

2.2.4 Counting positional nucleotide frequencies for each enzyme

We determined nucleotide counts at each position relative to the start of a read by loading all the sequences into R, separating sequences into columns, and tallying nucleotide bases at each position. Results were output in TRANSFAC format and input into Weblogo (Gavin E. Crooks and Brenner 2004).

2.2.5 Plotting background nucleotide frequency-corrected sequence logos

We desired background corrected information content values for our sequence logos generated from `Weblogo` (Gavin E. Crooks and Brenner 2004). We retrieved these values by modifying the source code of the `Weblogo` command line interface. A step-by-step guide on the modifications we made to retrieve these values is included in the vignette on Github. As a coherence check, we calculated the background corrected information content values for each position. First, we calculated the Shannon entropy:

$$H(l) = -\sum_{b=a}^{t} = f(b,l)log_2 f(b,l)$$

Here, H(l) is the entropy at any given position, and f(b,l) is the frequency of a base (b) at this position (l). Subtracting this value from 2 is the classic calculation for information content. We next calculated the background entropy:

$$H(background) = -\sum_{b=a}^{t} = f(b,l)log_2(\frac{0.25}{f(background,b)})$$

Where H(background) is the correction for background nucleotide frequency. And f(background,b) is the background frequency (background) for a base (b) Thus, for any position the background corrected information content is:

$$Information\ Content = 2 - (H(l) + H(background))$$

These values were plotted using the `Weblogo` command line interface.

2.2.6 Determining enzyme bias motif genomic coordinates

Starting with the TRANSFAC format nucleotide count files, we used `transfac2meme`

(Bailey et al. 2015) to convert to the meme file format. We then used these files as input into `FIMO` to generate region sets for each bias motif using the appropriate refererence genome (Grant, Bailey, and William Stafford Noble 2011). The highest scoring 400,000 regions were used for each composite profile plot.

2.2.7 Plotting composite signal of genomic coordinates

We visualized the average signal at genomic coordinates by aligning plus and minus locations on the same position and retrieving the signal in the plotted interval from a given data set's bigWig file, using the `bigWig` R package (`https://github.com/andrelmartins/bigWig`). We then divided all signal in this interval by the number of genomic coordinates within the plot to calculate the average cut or insertion (for Tn5) frequency at this position relative to the motif.

2.2.8 Determining background nucleotide frequencies of reference genomes

Background nucleotide frequencies for reference genomes were calculated by using the grep command to count the occurrences for a given nucleotide.

2.2.9 Calculating scale factors for k-mer position

Genomic and data set k-mer frequencies were determined by using the `seqOutBias` `table` (Martins et al. 2018) command to tally each k-mer in the reference genome and those in the data sets for each k-mer size and position. We then used these k-mer counts to calculate scale factors for each k-mer by dividing the expected k-mer frequency in the genome by the observed k-mer frequency in the data set (Martins et al. 2018).

2.2.10 Calculating observed and expected upstream k-mers

Starting with the sequences flanking each insertion in the Tn5 data set in fasta format, we counted each 3-mer in relevant positions from the cut site to determine observed k-mer occurrence. To determine expected k-mer occurrence, nucleotide frequency from the sequence logo motif at each position was multiplied together, in all possible combinations, to construct each possible k-mer. This value is the expected frequency for each k-mer; we plotted the $\log_2$ ratio of observed to expected k-mers.

2.2.11 CAG peak direction

We generated a list of all 3-mer locations in the genome by invoking the `seqOutBias dump` (Martins et al. 2018) command using a 3-mer .tbl file (output from previous invocations of `seqOutBias table`) for input. From this file, we determined the genomic location of all CAG instances in the reference genome. We then plotted ATAC signal at 400,000 random CAG instances from this bed file. Rationally designed masks were implemented using the `seqOutBias` (Martins et al. 2018) software with the noted k-mer masks.

2.2.12 Transcription factor motif genomic interval determinations

Motifs for each transcription factor included in the test and training sets were downloaded from the JASPAR (Castro-Mondragon et al. 2022) database in meme format. These motifs were then used as input, along with the hg38 reference genome, into FIMO to output genomic regions conforming best to the motif. We took the top 400,000 highest scoring genomic instances for each motif for both the plus and minus strands and used them as input for the rule ensemble model.

2.2.13 Rule ensemble target input

We calculated the target values by plotting the composite signal at each transcription factor's genomic coordinates using unscaled data produced from `seqOutBias` (Martins et al. 2018), using the `--no-scale` option. `bigWig` files were accessed and plotted using the `bigWig` R package (`https://github.com/andrelmartins/bigWig`). These plotted values were normalized using a common factor, to preserve variation between motifs and allow for accurate rule ensemble prediction.

2.2.14 K-mer frequency calculation

We determined k-mer frequency using the 400,000 genomic locations for each strand of each transcription factor motif previously generated using `FIMO`. From these locations, we retrieved the sequences using `fastaFromBed` (Quinlan and I. M. Hall 2010). We split each sequence into k-sized slices for each k-mer. For example, using a 5-mer, the sequence "AAACCAAA" would be split into: AAACC, AACCA, ACCAA, CCAAA. Each of these sections are a position within the original sequence: position 1-AAACC; position 2-AACCA; etc. For each position, we then determined the frequency of each k-mer among the original 400,000 input sequences.

2.2.15 Rule ensemble independent variable input

We computed the rule ensemble input for each combination of transcription factor motif region set and k-mer size/position. As we previously determined k-mer frequency at each position of a composite profile, we multiply these frequencies by the inverse of the scale factor for each matching k-mer, which was also described above. Finally, we sum these values for all possible k-mers at a position for the input value. This process was repeated for each modeled position in the composite corresponding to each included k-mer size and location relative to the cut site. These values are the `seqOutBias` predicted values output for a k-mer size and position, and modeled genomic regions.

2.2.16 Rule ensemble modeling

After calculating the independent variable input, a rule ensemble model was trained to predict the training set biased target values using the `prediction rule ensemble` package in R (Fokkema 2017). This package implements the original rule ensemble modeling framework (Friedman and Popescu 2008). The output rules and scaling coefficients were recorded as a single equation for later implementation.

2.2.17 Rule ensemble model implementation

We implemented the rule ensemble model by combining scaled `seqOutBias` (Martins et al. 2018) output as defined by the model. We first aggregated the required `seqOutBias` output by combining output bed format files using `unionbedg` software (Quinlan and I. M. Hall 2010). These output values were then combined by applying the modeling equation at every read to generate a rule ensemble-scaled bed file. This bed file was then scaled to the same total read depth as the original, unscaled data. Finally, we converted the bed file to bigWig format for subsequent analysis using the `bedGraphToBigWig` command line interface (Kent et al. 2010).

2.2.18 k-mer importance formalization

Calculations for the importance of a given input variable are conducted by the PRE package in R. The formalized method is imported with little modification from the accompanying paper (Fokkema 2017; Friedman and Popescu 2008).

In order to calculate input k-mer position importance, the importance of individual rules and linear terms in which it appears must be determined. We can calculate linear term importance using the formula:

$$I_j = |\hat{b}_j| \cdot \frac{sd(l_j(x_j))}{sd(y)}$$

Here, $l_j(x_j)$ indicates the linear term for predictor variable x_j (k-mer frequencies of a given size and position relative to cut site multiplied by inverse scale factors for the same k-mer size and position), *sd* signifies standard deviation, and *y* represents output, or bias predicted by the model. The term $|\hat{b}_j|$ is the absolute value of the k-mer's coefficient in the trained model.

To calculate the global importance of a given rule, we use the equation:

$$I_v = |\hat{a}_v| \cdot \frac{\sqrt{s_v(1 - s_v)}}{sd(y)}$$

Similar to a linear term's importance, $\sqrt{s_v(1 - s_v)}$ evaluates to the rule's sample standard deviation. This value is measured by first determining s_v, or the support for rule *v*, which is the fraction of training data for which the rule's conditions are satisfied.

We can calculate the support for a given rule (r_v) using the expression:

$$s_v = \frac{1}{N} \sum_{i=1}^{N} r_v(x_i)$$

To evaluate the total importance of a predictor k-mer, we sum all rules in which it applies, in addition to its linear term. This is calculated in the following expression:

$$J_j = I_j + \sum_{x_j \in r_v} \frac{I_v}{c_v}$$

In which the variable c_v is equal to the number of requirements for rule *v*. This results in the importance of any given rule being equally divided between the various k-mers included in the rule.

2.2.19 Calculating positional importances from rule ensemble models

We calculated variable importances, for each position within our range of inputs to visualize how the rule ensemble model combined input positions. Each k-mer's importance was derived from the model and we calculated the importance of each position by summing the values of each position for all k-mers. We performed this calculation for DNase, however Tn5 k-mers were not equally distributed at each input position and this caluclation was not appropriate. We scaled the normalized Tn5 positional importance by dividing the sum of values by the number of inputs that included the respective position in the input masks.

2.2.20 Calculating improvement of rule ensemble modeling compared to seqOutBias

We directly compared the improvement of rule ensemble modeling over `seqOutBias` by first determining the absolute value difference between scaled output from either method and the calculated unbiased output. This value was determined for each position in our test set composite profiles. Each of these values is the difference between either method and perfect bias correction at a given position—a value of 0 means the method perfectly corrected bias at this position. For each position we subtracted the rule ensemble output from the `seqOutBias`; for every value above 0, rule ensemble modeling outperforms `seqOutBias` correction.

2.3 Hill climbing optimization of seqOutBias masks

One approach we initially explored to correct Tn5 sequence bias was hill climbing optimization of seqOutBias masks (positions used to scale reads), similarly to the method reported in Martins et al. 2018. The rationale of this optimization algorithm is that the positions which are most beneficial to correcting Tn5 bias may not be contiguous, but rather spaced across many positions relative to the cut site. These spaced positions can be identified by their ability to reduce enzyme bias to a greater extent than other positions. To visualize enzyme bias, we plotted the reads scaled by each position onto several TF aggregate plots. We then define enzyme bias as the sum of standard deviations for the points on these aggregate plots. Consequently, the plots with the minimum sum of standard deviations between positions will have the lowest bias.

Hill climbing optimization of seqOutBias masks was carried out using a window of 40 base pairs around the cut site to determine the best positions for correcting bias in PE1 plus naked Tn5 reads. Each iteration of the optimization quantified the sum of standard deviations of the aggregate plots, for each possible position within the 40 base pair window. The position whose aggregate plots had the lowest sum of standard deviations was included as a scaling position for subsequent rounds of optimization. This process was allowed to continue until the top 12 positions had been determined (Figure 2.13).

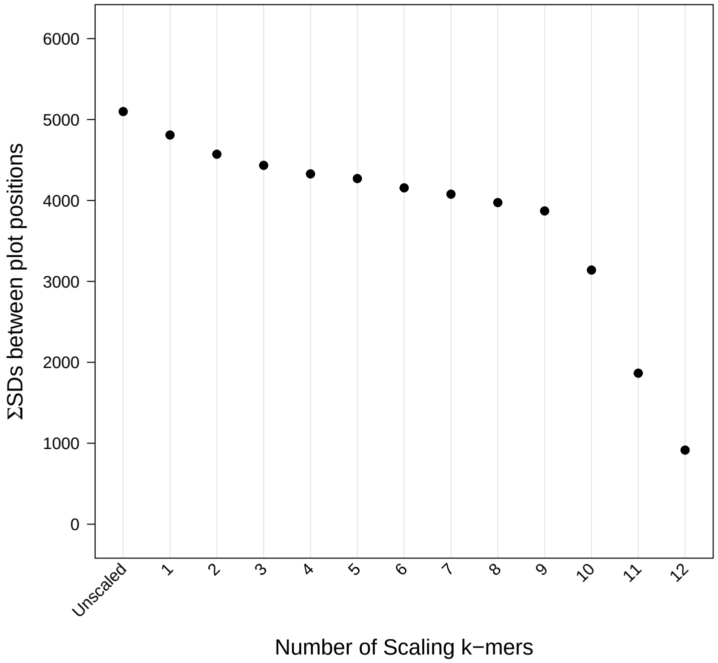


Figure 2.13: **Hill climbing optimization of naked ATAC-seq data reduces variance in aggregate plots.** For each position determined to best correct Tn5 bias, the sum of standard deviations is plotted. Each k-mer includes all previous best positions as scaling positions, and uses the new best position as an additional scaling position.

Visualization of hill climbing optimization showed that it successfully reduced the variance between positions in the aggregate plots. Unfortunately, these reductions in variance were greatest with large k-mers, which are not experimentally feasible. Further inspection of the aggregate plots showed that hill climbing optimization preferentially selected positions which depressed the overall signal, rather than selectively correcting enzyme bias (Figure 2.14). This results in the spaced 12-mer read depth being equal to roughly a quarter of the unscaled read depth, indicating a loss of signal rather than bias correction.

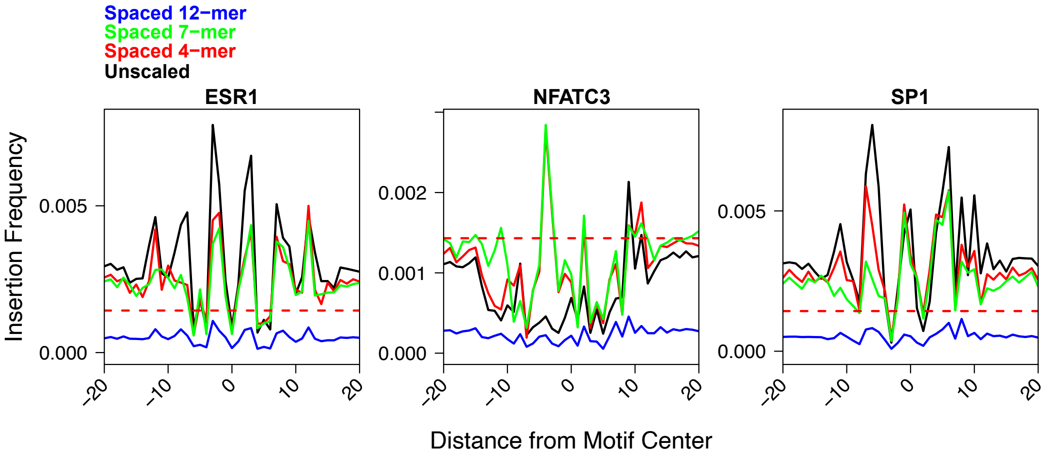


Figure 2.14: **Hill climbing optimization reduces total signal in aggregate plots.** Aggregate plots of the hill climbing-derived indicated best scaling k-mers. The red dashed line indicates theoretically random cleavage.

2.4 Tn5 regional bias and genomic GC organization

Early work on the human genome project revealed that GC content (the frequency of guanine or cytosine in a region) was clustered throughout the genome (International Human Genome Sequencing Consortium 2001). Although this knowledge informs our view of genomic organization, it does not visualize or expand on how clustering in the genome takes place (e.g. is this clustering stable on the order of kilo bases, mega bases; what is the range of possible GC percents etc.). In order to more directly illustrate this clustering, we first determined the sequence of the 100 bases surrounding every position in the human genome. Using these sequences, we next calculated the surrounding percent GC at all positions in the genome. We finally plotted these values for the first 6Mb in chromosome one. This plot illustrates that the GC content of the human genome has a wide range, from nearly 100%

at some positions, to 0% in others (Figure 2.15). While the extremes of this range are due to local outliers, the overall GC percent observed is relatively stable, and gradually changes from 30% to 60% on the order of mega bases.

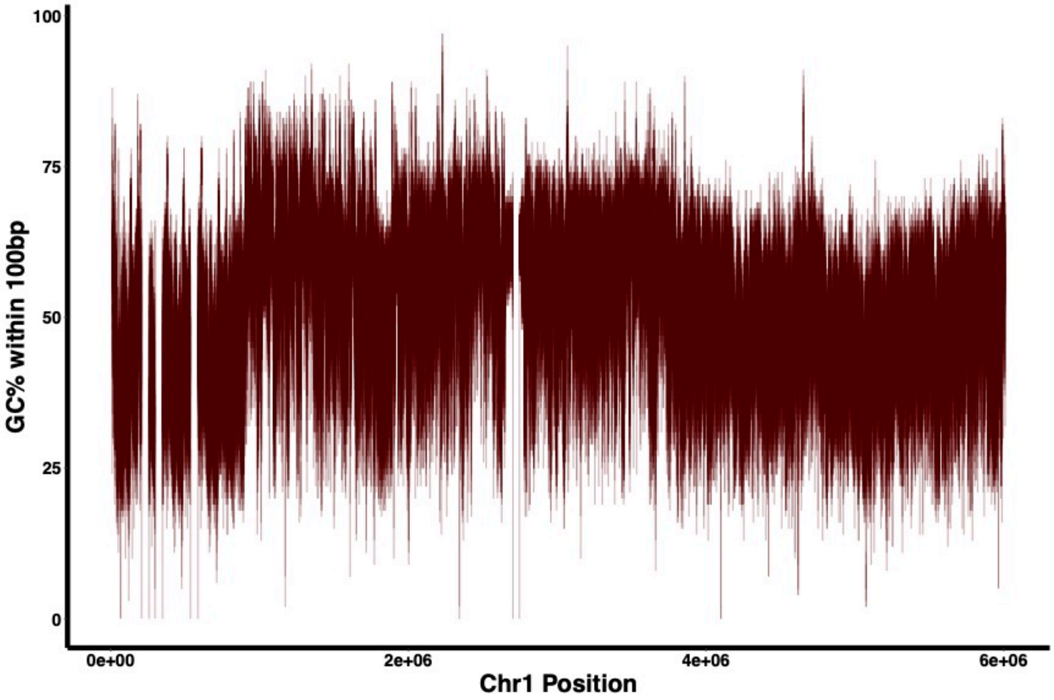


Figure 2.15: **GC content of 6Mb in chromosome 1 shows regional consistency.** Percent GC of each position was calculated using a 100 base pair window. Every possible position within the first 6Mb of chromosome 1 in hg38 is plotted. Image resolution was reduced for size purposes, as the original plot contains 6 million plotted points.

The stability of GC content in the genome led us to next question whether or not the GC content in baselines (defined as ±10 base pairs outside of the motif) of aggregate plots reflected the GC content of the aggregate motif. To determine if a correlation exists, we first retrieved the known sequences which make up aggregate plots. We then determined the GC content for a 5 base pair window around each position within the plot, and averaged these values for each TF. Plotting this average baseline GC content against each motif's GC content demonstrated a strong correlation between the two (Figure 2.16). This result is consistent with the organization of GC content in the genome (Figure 2.15). Taken together, these results indicate an aggregate of similar sequences (a PSWM in this case) should have comparable GC content to the motif used to aggregate, up to several kilobases from the central base pair.

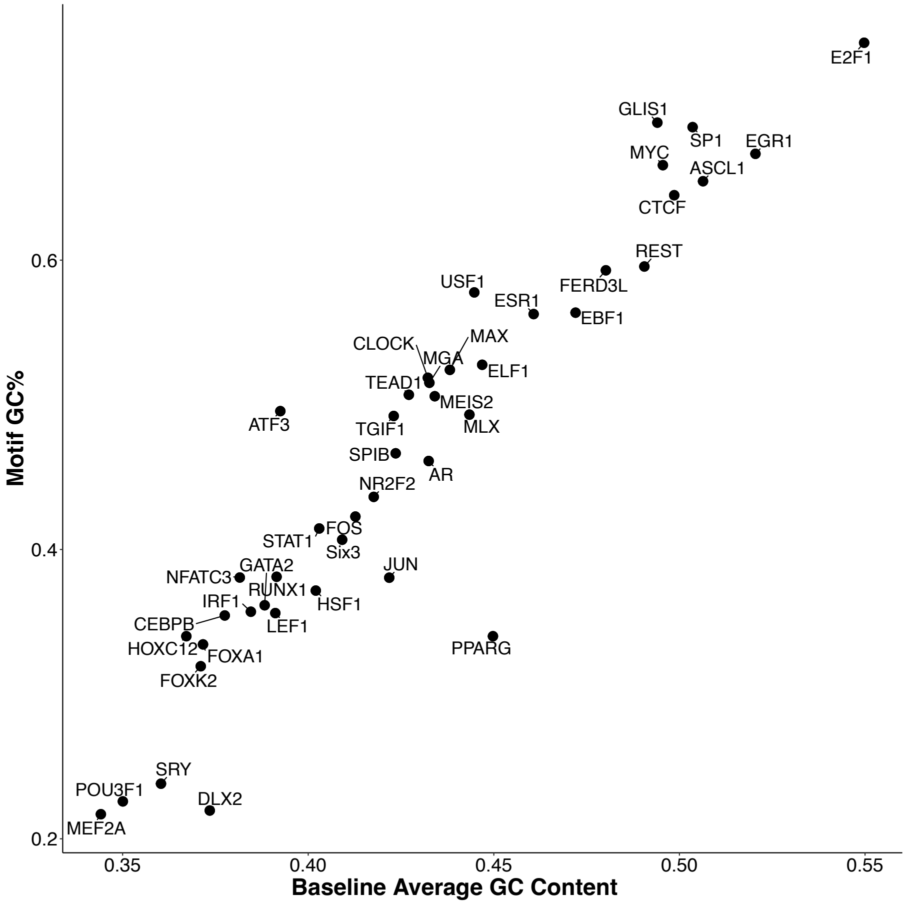


Figure 2.16: **Aggregate plot baseline GC content is correlated with motif GC content.** Baseline GC content for each TF was determined by using a single step, 5 base pair window. These values were then averaged to give a single baseline GC content. Motif GC content is calculated based on PSWM information content and nucleotide identity.

Encouraged by these results, we used the previously calculated GC content as additional input into a rule ensemble model, alongside the original k-mer frequency multiplied by inverse scale factor input. Comparison of bias correction between the rule ensemble model which included GC content, and models which did not showed no improvement from the inclusion of this information (data not shown). We speculate that this lack of improvement is due to the original k-mer frequency multiplied by inverse scale factor input already containing GC content information, at the k-mer level.

Chapter 3

Conclusion

3.1 Summary of major conclusions

Past investigations have shown that interactions with DNA, whether chemical or physical, have biases for specific nucleotide sequences (Meyer and Liu 2014; H. Zhang et al. 2021). Failure to correct these biases can lead to errant conclusions about the ground truth of studied biological systems (Neph et al. 2012; Sung et al. 2014; Teng and Irizarry 2017). This sequence bias also applies to the cleavage of DNA, the foundation of chromatin accessibility assays (Sung et al. 2014; He et al. 2014). Among chromatin accessibility assays, ATAC-seq is the most widely used (Yan et al. 2020). Tn5 transposase, the enzyme used in ATAC-seq, has a sequence bias which is not satisfactorily corrected by direct k-mer scaling, a method which eliminates the majority of observed bias from other commonly used enzymes (Martins et al. 2018). Therefore, it is important to address and correct for these biases before analysis of ATAC-seq data.

Previous studies which sought to correct Tn5 enzymatic sequence bias in ATAC-seq data displayed the PSWM of the Tn5 bias motif as a characterization of its bias (Vinayak, Vinay, and Shiv 2019; Z. Li et al. 2019; Bentsen et al. 2020). This inspired us to not only begin our study by characterizing Tn5 bias, but to quantitatively compare it with other enzymes used for chromatin accessibility assays. In this initial comparison of enzymatic sequence biases, we determined that Tn5 had nearly twice the information content and breadth in its bias motif as was seen in the second widest and highest information content motif (DNase). We advanced this comparison of enzymatic sequence bias by determining the distribution of 5-mer scale factors for each position relative to the cut site or centrally recognized base. Scale factor distributions for the other enzymes surveyed showed a slim window, within 4bp, of positions necessary to model to effectively correct their biases. This is in contrast to the 21bp window observed for Tn5, which confirms direct k-mer scaling is not capable of modeling the transposase's bias. We next sought to directly compare the strength of enzymatic sequence bias. In order to visualize this comparison, we plotted the signal at sites of greatest bias for each enzyme and normalized these values to the read depth of each data set. The resultant plots showed that Tn5 bias was stronger in magnitude for preferred sequences than all other enzymes surveyed.

The stronger sequence bias of Tn5 and its GC-rich bias motif likely also contribute to the observed regional GC preference. When regions with high or low GC content were compared, it was clear that baseline Tn5 signal was elevated in high GC content regions. We termed this 'regional bias' as it acted on the genomic GC content organization in which some regions contained elevated, while others contained depleted, GC content. This regional bias indicated that aside from single nucleotide bias, Tn5 interaction with DNA was influenced by regional factors as well. These findings corroborated previous work, which showed that Tn5 could form scaffolds on DNA with which it had interacted, suggesting interactions between long stretches of DNA and many Tn5 molecules (Adey et al. 2014). Further examination of Tn5 bias led us to test the assumption of positional independence in probability matrices, such as those used to depict the sequence bias motif. Here, we saw that even in a limited 3bp window of the bias motif (positions -4 to -6 from the centrally recognized base) sequence identity at one position influenced cutting at the other positions, violating positional independence. This again suggested that more advanced modeling strategies were necessary to fully account for the observed Tn5 sequence bias.

As we devised methods to correct Tn5 sequence bias, we first attempted to optimize a direct k-mer scaling approach by determining the best performing spaced k-mer to correct the bias. This attempt involved using a hill climbing algorithm to find the local minima of standard deviation between plotted positions for each k-mer position modeled. The best performing 12-mer determined using this method simply suppressed all signal, rather than specifically correcting biased signal and returning composite signal to calculated random cleavage. Because of these results, this method was abandoned. We next explored using statistical inference and modeling to predict and correct Tn5 bias.

3.2 Rule ensemble combination of direct k-mer scaling input corrects Tn5 bias

Before modeling Tn5 bias could be attempted, we needed to devise how to prepare input for the model. A key component of the input was that it needed to represent Tn5 bias across many different genomic intervals and environments. We accomplished this by first selecting a group of transcription factor motifs representative of a wide range of information and GC content. Using these motifs, we determined composite signal at the top 400,000 sites which conformed to each motif, for both strands. Next, enzymatic insertion frequency for each possible k-mer at a given location relative to the centrally recognized base was used together with k-mer frequency at each position to predict these composite signals. This formed the input for the RE model.

Once the RE model was trained, we first measured the importance of each position relative to the centrally recognized base, to determine the degree to which it contributed to the model's prediction. Positional importance values showed that the RE model used input k-mers in a pattern similar to the previously determined Tn5 bias motif's information content distribution. We next tested the RE output by determining how close corrected values were to calculated random cleavage and comparing this with unscaled and seqOutBias (direct k-mer scaling) output. These results revealed that at a single nucleotide level, the biased composite signal was greatly reduced by RE scaling in comparison with direct k-mer scaling. This led us to examine how RE scaling affected the observed regional bias of Tn5. Similarly, testing RE regional bias scaling against seqOutBias output showed that Tn5 regional bias was more greatly reduced by our RE model. Quantification of this improvement showed that 88% of positions were improved by RE modeling over seqOutBias scaling. These findings led us to conclude that RE modeling successfully corrects Tn5 sequence bias.

3.3 Future Directions

Although RE modeling largely corrects observed Tn5 transposase bias, alterations to the method may yield improved results. One way to likely improve the scaling method involves incorporating larger k-mers. When first implementing the RE model for Tn5 data, we reduced the number of input variables by selecting the 10% most important and removing the rest. An unreported aspect of this input variable refinement was that the model heavily preferred larger k-mers over smaller ones. While we desired to incorporate larger k-mers into the model, this was a mathematical impossibility. As we previously mentioned, effective direct k-mer scaling requires many instances of both observed (input data) and genomic k-mers. This means that the human genome can theoretically support k-mers on the scale of at least 11-mers for direct k-mer scaling. Mathematically this would mean the average 11-mer would have 763 occurrences (3.2 billion base pairs divided by 4^{11}), and as k-mer occurrence is a distribution, those at the lowest end would still have several instances. However, because most ATAC-seq experiments do not produce this many reads, incorporating larger k-mers is limited by the number of reads in a data set. Therefore, if one wished to create a high fidelity, ultra low bias data set, a new RE could be implemented in which larger k-mers could be incorporated and used for training. This would require that the experimental protocol be modified to accommodate these larger k-mers by increasing the read count, which has been proposed previously for footprinting (Buenrostro, B. Wu, Chang, et al. 2015). While this would add new experimental constraints, it would also likely reduce the bias further than the reduction observed using 7-mers as we did in this research project.

Another addition which could be made to our RE model is the ability to correct Tn5 sequence bias in single-cell ATAC-seq data. Several previous studies which examined the downstream effects of bias correction in ATAC-seq data also implemented single-cell bias correction with promising results which showed enhanced cell type clustering (Hu et al. 2022; Z. Li et al. 2019). Characterization of the Tn5 sequence bias observed in single-cell data sets showed that the observed bias was highly correlated with the bias observed in bulk ATAC-seq data. This is to be expected as the basic enzyme-DNA interaction is the same. Therefore, applying the

RE correction to single-cell ATAC-seq data would likely not require a specially trained model. Nonetheless, applying the RE workflow to single-cell data would require some accommodations. Namely, the current RE workflow only includes scaled read's genomic coordinates and signal values, and is lacking any information about cell barcodes. Barcodes would need to be included in this new workflow, as combining single-cell reads would be necessary to create enough observed k-mers for bias correction to take place. Hence, the current workflow would need to be modified to include this information, as it is necessary for proper analysis of single-cell data.

The RE model could be further improved by inclusion of DNA shape as a variable. One study investigating Tn5 sequence bias attempted to classify Tn5 insertion sites and random genomic regions using the DNA motif at any given site (H. Zhang et al. 2021). This investigation showed that the accuracy of this classification could be improved in deproteinated ATAC-seq data by including local DNA shape as a variable. Interestingly, this same study later used seqOutBias and direct k-mer scaling to show that peak calling fidelity was increased by bias correction, but did not include DNA shape as a variable for bias correction. Because our method integrates several runs of seqOutBias (and was shown to be capable of incorporating GC content) to effect bias correction, DNA shape could be incorporated into our bias correction model. Similarly to GC content, whether or not DNA shape contributes to bias correction could both be determined by analysis of output, and regression coefficient after the Lasso penalty is applied.

3.4 Final Thoughts

Meaningful scientific understanding requires a pure, artifact-free grasp of ground truth. As ATAC-seq has become the dominant assay used to probe chromatin accessibility, the correction of Tn5 sequence bias has become a priority for accurate analysis of many downstream applications. Failure to account for Tn5 sequence bias has, and will continue to lead to incorrect interpretations of experimental results. This work defines a method to apply an easily implementable, machine learning bias correction to ATAC-seq data. We designed this method to be easily implementable insofar as it accepts commonly used file types as input and generates output which is readily incorporated into field-standard pipelines. Further, this method is interpretable at several levels, giving a rare understanding into the mechanisms by which a machine learning approach achieves its modeling accuracy. We additionally add to the literature and findings which define the nature of Tn5 transposase bias in comparison with nucleases. These findings give reasoning behind the difficulty of modeling Tn5 bias and clarify why previously successful approaches for other enzymes did not apply well to this system. Integration of our RE modeling approach enables others to gain a true understanding of their ATAC-seq output, free from sequence bias artifacts.

References

Adey, Andrew et al. (2014). “In vitro, long-range sequence information for de novo genome assembly via transposase contiguity”. In: *Genome research* 24.12, pp. 2041–2049.

Ansari, Meshal, David S Fischer, and Fabian J Theis (2020). “Learning Tn5 Sequence Bias from ATAC-seq on Naked Chromatin”. In: *International Conference on Artificial Neural Networks*. Springer, pp. 105–114.

Ason, Brandon and William S Reznikoff (2004). “DNA sequence bias during Tn5 transposition”. In: *Journal of molecular biology* 335.5, pp. 1213–1225.

Bailey, Timothy L et al. (2015). “The MEME suite”. In: *Nucleic acids research* 43.W1, W39–W49.

Bentsen, Mette et al. (2020). “ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation”. In: *Nature communications* 11.1, pp. 1–11.

Berg, Douglas E et al. (1975). “Transposition of R factor genes to bacteriophage lambda.” In: *Proceedings of the National Academy of Sciences* 72.9, pp. 3628–3632.

Blossey, Ralf and Helmut Schiessel (2018). “The latest twists in chromatin remodeling”. In: *Biophysical Journal* 114.10, pp. 2255–2261.

Boyle, Alan P et al. (2008). “High-resolution mapping and characterization of open chromatin across the genome”. In: *Cell* 132.2, pp. 311–322.

Buenrostro, Jason D, Paul G Giresi, et al. (2013). “Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position”. In: *Nature methods* 10.12, pp. 1213–1218.

Buenrostro, Jason D, Beijing Wu, Howard Y Chang, et al. (2015). “ATAC-seq: a method for assaying chromatin accessibility genome-wide”. In: *Current protocols in molecular biology* 109.1, pp. 21–29.

Buenrostro, Jason D, Beijing Wu, Ulrike M Litzenburger, et al. (2015). “Single-cell chromatin accessibility reveals principles of regulatory variation”. In: *Nature* 523.7561, pp. 486–490.

Castro-Mondragon, Jaime A et al. (2022). “JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles”. In: *Nucleic acids research* 50.D1, pp. D165–D173.

Clapier, Cedric R and Bradley R Cairns (2009). “The biology of chromatin remodeling complexes”. In: *Annual review of biochemistry* 78.1, pp. 273–304.

Consortium, Mouse Genome Sequencing (2002). “Initial sequencing and comparative analysis of the mouse genome”. In: *Nature* 420.6915, pp. 520–562.

Cremer, Thomas et al. (2015). “The 4D nucleome: Evidence for a dynamic nuclear landscape based on co-aligned active and inactive nuclear compartments”. In: *FEBS letters* 589.20, pp. 2931–2943.

Danecek, Petr et al. (2021). “Twelve years of SAMtools and BCFtools”. In: *Gigascience* 10.2, giab008.

Di Stefano, Bruno et al. (2016). “C/EBPα creates elite cells for iPSC reprogramming by upregulating Klf4 and increasing the levels of Lsd1 and Brd4”. In: *Nature cell biology* 18.4, pp. 371–381.

Dynan, William S and Robert Tjian (1983). “The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter”. In: *Cell* 35.1, pp. 79–87.

Elgin, Sarah CR (1988). “The formation and function of DNase I hypersensitive sites in the process of gene activation”. In.

Felsenfeld, Gary et al. (1996). “Chromatin structure and gene expression.” In: *Proceedings of the National Academy of Sciences* 93.18, pp. 9384–9388.

Fokkema, Marjolein (2017). “Fitting prediction rule ensembles with R package pre”. In: *arXiv preprint arXiv:1707.07149*.

Friedman, Jerome H, Bogdan E Popescu, et al. (2003). “Importance sampled learning ensembles”. In: *Journal of Machine Learning Research* 94305, pp. 1–32.

Friedman, Jerome H and Bogdan E Popescu (2008). “Predictive learning via rule ensembles”. In: *The annals of applied statistics*, pp. 916–954.

Galas, David J and Albert Schmitz (1978). “DNAase footprinting a simple method for the detection of protein-DNA binding specificity”. In: *Nucleic acids research* 5.9, pp. 3157–3170.

Gaspar, John M (2018). “Improved peak-calling with MACS2”. In: *BioRxiv*, p. 496521.

Gavin E. Crooks Gary Hon, John-Marc Chandonia and Steven E. Brenner (Jan. 2004). “WebLogo: A Sequence Logo Generator”. In: *Genome Research* 14.6, pp. 1188–1190. DOI: `10.1101/gr.849004`. eprint: `https://pubmed.ncbi.nlm.nih.gov/15173120/`. URL: `https://genome.cshlp.org/content/14/6/1188.full`.

Goryshin, Igor Y et al. (1998). “Tn 5/IS 50 target recognition”. In: *Proceedings of the National Academy of Sciences* 95.18, pp. 10716–10721.

Goryshin, Igor Yu and William S Reznikoff (1998). “Tn5 in vitro transposition”. In: *Journal of Biological Chemistry* 273.13, pp. 7367–7374.

Gradman, Richard J et al. (2008). “A bifunctional DNA binding region in Tn5 transposase”. In: *Molecular microbiology* 67.3, pp. 528–540.

Grant, Charles E, Timothy L Bailey, and William Stafford Noble (2011). “FIMO: scanning for occurrences of a given motif”. In: *Bioinformatics* 27.7, pp. 1017–1018.

Grøntved, Lars et al. (2012). "Rapid genome-scale mapping of chromatin accessibility in tissue". In: *Epigenetics & Chromatin* 5.10. DOI: https://doi.org/10.1186/1756-8935-5-10. URL: https://epigeneticsandchromatin.biomedcentral.com/articles/10.1186/1756-8935-5-10#citeas.

Gross, David S and William T Garrard (1988). "Nuclease hypersensitive sites in chromatin". In: *Annual review of biochemistry* 57.1, pp. 159–197.

Guertin, Michael J et al. (2012). "Accurate prediction of inducible transcription factor binding intensities in vivo". In: *PLoS genetics* 8.3, e1002610.

Hansen, Jeffrey L, Kaiser J Loell, and Barak A Cohen (2022). "A test of the pioneer factor hypothesis using ectopic liver gene activation". In.

He, Housheng Hansen et al. (2014). "Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification". In: *Nature Methods* 11, pp. 73–78. ISSN: 1548-7105. DOI: https://doi.org/10.1038/nmeth.2762. URL: https://www.nature.com/articles/nmeth.2762.

Hesselberth, Jay R et al. (2009). "Global mapping of protein-DNA interactions in vivo by digital genomic footprinting". In: *Nature methods* 6.4, pp. 283–289.

Hu, Shengen Shawn et al. (2022). "Intrinsic bias estimation for improved analysis of bulk and single-cell chromatin accessibility profiles using SELMA". In: *Nature Communications* 13.1, pp. 1–17.

Ignatieva, Elena V, Victor G Levitsky, and Nikolay A Kolchanov (2015). "Human genes encoding transcription factors and chromatin-modifying proteins have low levels of promoter polymorphism: a study of 1000 genomes project data". In: *International journal of genomics* 2015.

International Human Genome Sequencing Consortium (2001). "Initial sequencing and analysis of the human genome". In: *nature* 409.6822, pp. 860–921.

Iwata-Otsubo, Aiko et al. (2017). “Expanded Satellite Repeats Amplify a Discrete CENP-A Nucleosome Assembly Site on Chromosomes that Drive in Female Meiosis”. In: *Current Biology* 27.15, 2365–2373.e8. ISSN: 0960-9822. DOI: https://doi.org/10.1016/j.cub.2017.06.069. URL: https://www.sciencedirect.com/science/article/pii/S0960982217308060.

John, Sam et al. (2011). “Chromatin accessibility pre-determines glucocorticoid receptor binding patterns”. In: *Nature genetics* 43.3, pp. 264–268.

Karabacak Calviello, Aslıhan et al. (2019). “Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling”. In: *Genome biology* 20.1, pp. 1–13.

Kent, W James et al. (2010). “BigWig and BigBed: enabling browsing of large distributed datasets”. In: *Bioinformatics* 26.17, pp. 2204–2207.

Koohy, Hashem, Thomas A Down, and Tim J Hubbard (2013). “Chromatin accessibility data sets show bias due to sequence specificity of the DNase I enzyme”. In: *PloS one* 8.7, e69853.

Langmead, Ben and Steven L Salzberg (2012). “Fast gapped-read alignment with Bowtie 2”. In: *Nature methods* 9.4, pp. 357–359.

Lazarovici, Allan et al. (2013). “Probing DNA shape and methylation state on a genomic scale with DNase I”. In: *Proceedings of the National Academy of Sciences* 110.16, pp. 6376–6381. DOI: 10.1073/pnas.1216822110. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1216822110.

Lee, Cheol-Koo et al. (2004). “Evidence for nucleosome depletion at active regulatory regions genome-wide”. In: *Nature genetics* 36.8, pp. 900–905.

Leinonen, Rasko et al. (2010). “The sequence read archive”. In: *Nucleic acids research* 39.suppl_1, pp. D19–D21.

Li, Hongyang, Daniel Quang, and Yuanfang Guan (2019). “Anchor: trans-cell type prediction of transcription factor binding sites”. In: *Genome research* 29.2, pp. 281–292.

Li, Ming et al. (2015). “Dynamic regulation of transcription factors by nucleosome remodeling”. In: *Elife* 4, e06249.

Li, Zhijian et al. (2019). “Identification of transcription factor binding sites using ATAC-seq”. In: *Genome biology* 20.1, pp. 1–21.

Martins, André L et al. (2018). “Universal correction of enzymatic sequence bias reveals molecular signatures of protein/DNA interactions”. In: *Nucleic acids research* 46.2, e9–e9.

Meyer, Clifford A and X Shirley Liu (2014). “Identifying and mitigating bias in next-generation sequencing methods for chromatin biology”. In: *Nature Reviews Genetics* 15.11, pp. 709–721.

Moore, Jill E et al. (2020). “Expanded encyclopaedias of DNA elements in the human and mouse genomes”. In: *Nature* 583.7818, pp. 699–710.

Neph, Shane et al. (2012). “An expansive human regulatory lexicon encoded in transcription factor footprints”. In: *Nature* 489.7414, pp. 83–90.

Pulverer, Bernd (2005). “Getting specific”. In: *Nature Reviews Molecular Cell Biology* 6.1, S12–S12.

Quinlan, Aaron R and Ira M Hall (2010). “BEDTools: a flexible suite of utilities for comparing genomic features”. In: *Bioinformatics* 26.6, pp. 841–842.

Raj, Anil et al. (2015). “msCentipede: modeling heterogeneity across genomic sites and replicates improves accuracy in the inference of transcription factor binding”. In: *PloS one* 10.9, e0138030.

Ren, Jianke et al. (2019). “The chromatin remodeling protein Lsh alters nucleosome occupancy at putative enhancers and modulates binding of lineage specific transcription factors”. In: *Epigenetics* 14.3, pp. 277–293.

Reznikoff, William S (1993). “The TN5 transposon”. In: *Annual review of microbiology* 47.1, pp. 945–963.

— (2003). “Tn5 as a model for understanding DNA transposition”. In: *Molecular microbiology* 47.5, pp. 1199–1206.

— (2008). “Transposon tn 5”. In: *Annual review of genetics* 42, pp. 269–286.

Sasse, Sarah K et al. (2017). “Glucocorticoid receptor ChIP-seq identifies PLCD1 as a KLF15 target that represses airway smooth muscle hypertrophy”. In: *American journal of respiratory cell and molecular biology* 57.2, pp. 226–237.

Schneider, Thomas D. and R.Michael Stephens (Oct. 1990). “Sequence logos: a new way to display consensus sequences”. In: *Nucleic Acids Research* 18.20, pp. 6097–6100. ISSN: 0305-1048. DOI: `10.1093/nar/18.20.6097`. eprint: `https://academic.oup.com/nar/article-pdf/18/20/6097/3939026/18-20-6097.pdf`. URL: `https://doi.org/10.1093/nar/18.20.6097`.

Schwessinger, Ron et al. (2017). “Sasquatch: predicting the impact of regulatory SNPs on transcription factor binding from cell-and tissue-specific DNase footprints”. In: *Genome research* 27.10, pp. 1730–1742.

Sharon, Eilon, Shai Lubliner, and Eran Segal (2008). “A feature-based approach to modeling protein–DNA interactions”. In: *PLoS computational biology* 4.8, e1000154.

Smith, Jason P et al. (2021). “PEPATAC: an optimized pipeline for ATAC-seq data analysis with serial alignments”. In: *NAR genomics and bioinformatics* 3.4, lqab101.

Song, Lingyun and Gregory E Crawford (2010). “DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells”. In: *Cold Spring Harbor Protocols* 2010.2, pdb–prot5384.

Sung, Myong-Hee et al. (2014). “DNase Footprint Signatures Are Dictated by Factor Dynamics and DNA Sequence”. In: *Molecular Cell* 56.2, pp. 275–285. ISSN: 1097-2765. DOI: `https://doi.org/10.1016/j.molcel.2014.08.016`. URL: `https://www.sciencedirect.com/science/article/pii/S1097276514006716`.

Teng, Mingxiang and Rafael A Irizarry (2017). “Accounting for GC-content bias reduces systematic errors and batch effects in ChIP-seq data”. In: *Genome research* 27.11, pp. 1930–1938.

Tewari, Alok K et al. (2012). “Chromatin accessibility reveals insights into androgen receptor activation and transcriptional specificity”. In: *Genome biology* 13.10, pp. 1–17.

Thurman, Robert E et al. (2012). “The accessible chromatin landscape of the human genome”. In: *Nature* 489.7414, pp. 75–82.

Vinayak, Viswanadham, Mahajan Vinay, and Pillai Shiv (2019). “A Bayesian approach for correcting Tn5 transposition bias in ATAC-seq footprinting”. In: *bioRxiv*, p. 525808.

Wang, Jeremy R, Bryan Quach, and Terrence S Furey (2017). “Correcting nucleotide-specific biases in high-throughput sequencing data”. In: *BMC bioinformatics* 18.1, pp. 1–10.

Weintraub, Harold and Mark Groudine (1976). “Chromosomal Subunits in Active Genes Have an Altered Conformation: Globin genes are digested by deoxyribonuclease I in red blood cell nuclei but not in fibroblast nuclei.” In: *Science* 193.4256, pp. 848–856.

Welboren, Willem-Jan et al. (2009). “ChIP-Seq of ERα and RNA polymerase II defines genes differentially responding to ligands”. In: *The EMBO journal* 28.10, pp. 1418–1428.

Wu, Carl, Paul M Bingham, et al. (1979). “The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence”. In: *Cell* 16.4, pp. 797–806.

Wu, Carl, Yuk-Chor Wong, and Sarah CR Elgin (1979). “The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity”. In: *Cell* 16.4, pp. 807–814.

Wu, Hao and Yi Eve Sun (2006). “Epigenetic regulation of stem cell differentiation”. In: *Pediatric research* 59.4, pp. 21–25.

Yan, Feng et al. (2020). “From reads to insight: a hitchhiker’s guide to ATAC-seq data analysis”. In: *Genome biology* 21.1, pp. 1–16.

Yardımcı, Galip Gürkan et al. (2014). “Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection”. In: *Nucleic acids research* 42.19, pp. 11865–11878.

Zhang, Houyu et al. (2021). “Comprehensive understanding of Tn5 insertion preference improves transcription regulatory element identification”. In: *NAR genomics and bioinformatics* 3.4, lqab094.