Accurate estimation of intrinsic biases for improved analysis of chromatin accessibility sequencing data using SELMA

Shengen Shawn Hu¹, Lin Liu², Qi Li³, Wenjing Ma^{1,4}, Michael J. Guertin⁵, Clifford A. Meyer⁶, Ke Deng³, Tingting Zhang⁷, Chongzhi Zang^{1,8,9*}

¹Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22908, USA

²Institute of Natural Sciences, MOE-LSC, School of Mathematical Sciences, and SJTU-Yale Joint Center for Biostatistics and Data Science, Shanghai Jiao Tong University, Shanghai 200240, China

³Center for Statistical Science, Tsinghua University, Beijing 100084, China

⁴Department of Computer Science, Emory University, Atlanta, GA 30322, USA

⁵Center for Cell Analysis and Modeling, Department of Genetics and Genome Sciences, University of Connecticut, Farmington, CT 06030, USA

⁶Department of Data Science, Dana-Farber Cancer Institute, Boston, MA 02215, USA

⁷Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260, USA

⁸Department of Public Health Sciences, University of Virginia, Charlottesville, VA 22908, USA

⁹UVA Cancer Center, University of Virginia, Charlottesville, VA 22908, USA

*Correspondence should be addressed to C.Z.: zang@virginia.edu

Abstract

Genome-wide profiling of chromatin accessibility by DNase-seq or ATAC-seq has been widely used to identify regulatory DNA elements and transcription factor binding sites. However, enzymatic DNA cleavage exhibits intrinsic sequence biases that confound chromatin accessibility profiling data analysis. Existing computational tools are limited in their ability to account for such intrinsic biases. Here, we present Simplex Encoded Linear Model for Accessible Chromatin (SELMA), a computational method for systematic estimation of intrinsic cleavage biases from genomic chromatin accessibility profiling data. We demonstrate that SELMA yields accurate and robust bias estimation from both bulk and single-cell DNase-seq and ATAC-seq data. We show that transcription factor binding inference from DNase footprints can be improved by incorporating estimated biases using SELMA. We also demonstrate improved cell clustering of single-cell ATAC-seq data by considering the SELMA-estimated bias effect. SELMA can be applied to existing bioinformatics tools to improve the analysis of chromatin accessibility sequencing data.

Introduction

Cis-regulatory elements in the genome play a critical role in transcription regulation by interaction with protein molecules such as transcription factors (TFs). These DNA elements are usually unwrapped from packed nucleosomes and are accessible in the chromatin structure^{1,2}. Genome-wide profiles of chromatin accessibility are a means to measure the global landscapes of active regulatory elements in different cell types. DNaseI hypersensitivity sequencing (DNase-seq) and the assay for transposase-accessible chromatin using sequencing (ATAC-seq) have become widely used for the genomic profiling of chromatin structure and accessibility^{3,4}. Signal enrichments, or “peaks”, from DNase-seq or ATAC-seq data are considered to represent accessible chromatin regions and can be used for inferring regulatory elements or TF binding sites. In addition, DNase-seq and ATAC-seq data also exhibit footprint patterns, which are fine structures in the accessible chromatin where high-affinity protein-DNA interactions protect the DNA from DNaseI or Tn5-transposase cleavages^{4,5}. DNase/ATAC-seq footprint detection has been implicated as an effective approach for identifying accurate TF binding sites at base-pair resolution^{6,7}. A few computational tools have been developed for detecting footprints from DNase-seq (RepFootprint⁸, Wellington⁹, PIQ¹⁰) or ATAC-seq data (HINT-ATAC¹¹, ToBIAS¹²). A recent study integrated 243 DNase-seq samples from different human cell and tissue types and reported approximately 4.5 million DNaseI consensus footprints associated with TF occupancy across the human genome as one of the largest maps of human regulatory DNA⁷.

The premise of using DNase-seq and ATAC-seq data to profile chromatin accessibility is that enzymatic DNA cleavages represented by sequence reads reflect local chromatin openness only. However, it has been shown that both DNaseI and Tn5 transposase exhibit sequence selection biases in high-throughput sequencing data^{13–16}. Such intrinsic enzymatic biases in

sequencing data can potentially confound observed cleavage patterns and footprint detection. The characterization and correction of intrinsic cleavage biases are essential to DNase/ATAC-seq data analysis. To characterize intrinsic cleavage biases, most studies use a k-mer model, in which the k-mer DNA sequence centered at a cleavage site of DNaseI/Tn5 is used as the signature of this cleavage^{14,15,17}. The sequence bias can be estimated by counting the occurrences of cleavages with each k-mer in one dataset relative to the genome-wide occurrences of this k-mer. A naïve k-mer model assumes that k-mers are independent of each other, resulting in an exponential increase in the degree of freedom when k increases. Therefore, a naïve k-mer model becomes less feasible in practice with a large k, especially with samples having insufficient sequencing depth. Although most studies use 6-mers with $4^6 = 4096$ parameters^{8,10,14,15,17,18}, it is unclear whether a different model with a larger k-mer size and more feasible parameter estimation can achieve better performance.

Several studies have used various computational approaches for intrinsic cleavage bias estimation^{8,15,16,19} and footprint detection with bias correction^{15,18–20} for bulk DNase/ATAC-seq data. Recently, single-cell ATAC-seq (scATAC-seq) has enabled chromatin accessibility profiling in thousands of individual cells at one time^{21–24}. Due to the high sparsity of single-cell data and because most chromatin accessibility regions contain only one read in one cell, any potential bias can be substantial in scATAC-seq data compared to bulk data, creating additional challenges in computational analysis. Incorporating intrinsic cleavage bias effects for improved scATAC-seq analysis also requires comprehensive assessment and development of innovative methods.

Here, we present Simplex Encoded Linear Model for Accessible Chromatin (SELMA), a computational framework for the accurate estimation of intrinsic cleavage biases and improved analysis of DNase/ATAC-seq data for both bulk and single-cell experiments. We demonstrate

that SELMA generates more accurate and robust bias estimation from bulk DNase/ATAC-seq data than the naïve k-mer model and that SELMA can utilize mitochondrial DNA (mtDNA) for bias estimation instead of requiring a separate naked DNA sample. We show an improved TF occupancy inference on ENCODE consensus footprints by including SELMA-estimated biases for each footprint. Finally, we show that SELMA-estimated biases can be incorporated with existing scATAC-seq computational tools to generate more accurate cell clustering analysis.

Results

SELMA improves cleavage bias estimation in DNase-seq and ATAC-seq data

We developed SELMA for an accurate and robust estimation of intrinsic cleavage biases from chromatin accessibility sequencing data. In SELMA, we start with a naïve k-mer model to calculate the frequency of observed cleavage sites at each k-mer relative to the total occurrences of this k-mer (Fig. 1a). We further encode each k-mer as a vector in the Hadamard Matrix H_{16} , derived from a simplex encoding model, in which the k-mer sequences are encoded as the vertices of a regular 0-centered simplex^{25,26}. Intuitively, a k-mer can be represented as k mononucleotides and $k - 1$ adjacent dinucleotides. Each mononucleotide is encoded as the 3-dimensional vector of one of the four tetrahedral vertices of the cube of side 2 centered at the origin. Each dinucleotide is encoded as the outer product of the 2 vectors representing the associated nucleotides (Fig. 1b). This k-mer simplex encoding model has $1 + 3 \times k + 9 \times (k - 1) = 12k - 8$ parameters, much fewer than the naïve k-mer model (4^k). We use a linear model to fit these $12k - 8$ parameters from the naïve k-mer biases, and use the fitted values as the SELMA-modeled cleavage biases (Fig. 1c).

As an intrinsic property of the enzyme (DNaseI/Tn5 transposase), the cleavage biases are expected to be invariant across cell types and independent of chromatin states (Fig. S1a). Comparing data from two cell types using different 8-mer models as an example, we found that the biases estimated using SELMA have a higher correlation than those estimated using the naïve k-mer model, for both DNase-seq (Fig. 1d,e) and ATAC-seq (Fig. 1f,g). Using sequence reads from genomic regions with different chromatin accessibility levels, the naïve k-mer model was not able to generate very consistent bias estimations (Fig. S1b,c), but the k-mer biases estimated by SELMA using the same data were highly consistent (Fig. S1b-e). The sequence preferences of DNaseI or Tn5 cleavage can be better reflected when the enzymes are applied to deproteinized naked genomic DNA^{15,17}. We found that the k-mer cleavage biases in naked DNA DNase/ATAC-seq data estimated with the naïve k-mer model can still be improved by SELMA, obtaining more consistent bias scores between different cell systems (Fig. 1h-k). These data demonstrated that SELMA can improve the accuracy of estimating k-mer cleavage biases in DNase-seq and ATAC-seq data.

With fewer parameters, SELMA enabled us to assess the effect of k-mer size on the performance of bias estimation. Using a “bias expected cleavage” approach^{8,14,15,19}, we compared the bias estimation performances measured as the correlation coefficient between the genome-wide observed cleavages and bias expected cleavages estimated using SELMA with different k. A higher correlation coefficient indicates a more accurate bias estimation to calculate the expected cleavages. For both DNase-seq and ATAC-seq data from two different cell lines, we found that 10-mer outperforms any other k-mer (Fig. 1l-o). We then applied this analysis to more DNase-seq and ATAC-seq data from a variety of human tissues generated by ENCODE and found that 10-mer always exhibited the best performance (Fig. S2). The above empirical evidence suggested that 10-mer is the optimal choice for intrinsic cleavage bias estimation for both DNaseI (DNase-seq) and Tn5 (ATAC-seq) cleavages.

SELMA improves ATAC-seq bias estimation by considering dimeric Tn5 cleavages

Many studies processed DNase-seq and ATAC-seq data in a similar way, treating individual DNA cleavage sites directly as indications of accessible chromatin^{8,16}. However, the mechanisms of enzymatic DNA cleavage are different between DNaseI and Tn5 transposases. Unlike DNaseI, the Tn5 transposase binds DNA as a dimer and inserts adaptors on the two strands separated by 9bp^{4,27} (Fig. 2a). As a result, the presence of each observed Tn5 cleavage in ATAC-seq data should be the consequence of two insertion events induced by the same Tn5 dimer, and the bias estimation of a Tn5 cleavage site should consider both the observed cleavage and the coupling cleavage 9bp downstream on the reverse strand. Therefore, SELMA estimated the bias on an ATAC-seq cleavage site as the geometric mean of the bias scores of the 10-mers at the 5' observed cleavage and at the 3' cleavage 9bp downstream on the reverse strand (Fig. 2a). To show the dimeric Tn5 cleavage effect, we calculated the cross-correlation between the genome-wide profiles of the plus strand cleavages and the minus strand cleavages. As expected, we observed a peak at 9bp of the cross-correlation curve specifically in the ATAC-seq data but not the DNase-seq data (Fig. 2b,c). Similarly, we observed perfectly matching aggregate cleavage patterns on TF motif consensus sites between plus strand and minus strand cleavages shifted by 9bp (Fig. S3a,b).

Using the similar “observed-expected correlation” approach described above, we found that SELMA considering dimeric cleavages outperformed models considering only 5' cleavage in generating a more accurate bias expected cleavage pattern for ATAC-seq data (Fig. 2d,e). We confirmed that this observation was specific to ATAC-seq, as similar analyses for DNase-seq from the same cell lines did not show a similar level of improvement (Fig. S3c,d). We also compared SELMA with several existing bias estimation approaches¹⁸ and found that SELMA's performance was the best for ATAC-seq data from several cell lines as well as different human

tissue types from ENCODE (Fig. 2d,e, Fig. S3e-j). Several new Tn5-based techniques have recently been developed for improved chromatin accessibility profiling, including THS-seq²⁸, Fast-ATAC-seq²⁹, and Omni-ATAC-seq³⁰. Data from these technologies showed intrinsic cleavage biases similar to those of conventional ATAC-seq (Fig. S3k-n). We found that SELMA also outperformed other approaches in bias estimation (Fig. 2f-i). Collectively, these data suggested that SELMA can most accurately estimate intrinsic cleavage biases in data from ATAC-seq and other Tn5-based techniques.

SELMA enables accurate bias estimation by utilizing mitochondrial DNA (mtDNA) reads

Accurate estimation of enzymatic cleavage biases independent of chromatin usually requires a control sample of deproteinized naked DNA digested by the same enzyme. Biases estimated from the naked DNA control data can be used to correct the chromatin accessibility profiling data^{14–18} (Fig. 3a). In conventional DNase/ATAC-seq data analyses, sequence reads mapped to mitochondrial DNA (mtDNA) are usually discarded⁴. Lacking histones responsible for the chromatin structure, mtDNA is similar to deproteinized naked DNA^{31–33}. Therefore, we sought to use the mtDNA reads from DNase/ATAC-seq data for cleavage bias estimation as an alternative to using a naked DNA control sample (Fig. 3b). Likely because of relatively low read counts and lack of sequence diversity (e.g., human mtDNA is only < 20kb long), the naïve k-mer model exhibited poor performance on bias estimation from mtDNA reads, using bias scores estimated from naked DNA as a reference (Fig. 3c). In contrast, SELMA generated a more accurate bias estimation from the same mtDNA reads, which was highly consistent with the bias scores estimated from the naked DNA data (Fig. 3d), demonstrating the power of SELMA to use less input to make accurate bias estimations. We applied this approach to a series of ATAC-seq and DNase-seq datasets for different human tissues from ENCODE and found that SELMA was better than the naïve k-mer model in yielding a consistently higher correlation coefficient (>0.9) between mtDNA-estimated bias and naked DNA-estimated bias for every ATAC-seq and

DNase-seq sample tested (Fig. S4a,b). Many optimized ATAC-seq protocols aimed to reduce mtDNA reads to increase the fraction of chromatin DNA reads for chromatin accessibility signal yield^{29,30}. We sampled down mtDNA reads to test the performance of SELMA in making robust bias estimations and found that SELMA could accurately estimate the bias with as few as 50,000 mtDNA reads (Fig. 3e, Fig. S4c), which is approximately 0.2% of the sequencing depth of a regular ATAC-seq sample and lower than the fraction of mtDNA reads in any existing ATAC-seq experiment^{29,30}. These data demonstrated that by applying SELMA, mtDNA reads can be utilized to substitute naked DNA control samples for accurate bias estimation.

Considering SELMA-estimated bias improves TF binding inference on ENCODE DNaseI footprint regions

With an accurate bias estimation model developed, we next sought to examine the potential effect of intrinsic biases on TF binding footprints in chromatin accessibility profiling data. Focusing on the 4,465,728 human DNaseI consensus footprints across the human genome, we plotted the DNaseI cleavages from different human cell lines and observed similar classic DNaseI cleavage protection (“footprint”) patterns across these footprint regions (Fig. 4a, Fig. S5a). Interestingly, we also observed a recurrent pattern of the SELMA-estimated DNaseI cleavage bias that is well aligned with the DNaseI cleavage pattern across these footprint regions (Fig. 4b, Fig. S5b). In the aggregate view of footprint regions of different lengths, the DNaseI cleavage signals exhibited a clear “cliff-bound valley”-shaped footprint pattern (Fig. S5c). The DNaseI cleavage bias scores exhibited a pattern of two spikes located inside the footprint coordinates (Fig. S5d). After normalizing various footprint lengths, we plotted the aggregate DNaseI cleavage patterns across all consensus footprint regions and found that the overall “footprint pattern” clearly included the double spike pattern of cleavage biases (Fig. 4c), which aligned well with the two spikes in the aggregate bias score pattern (Fig. 4d), indicating a possible contribution of intrinsic biases to the DNaseI footprinting.

To assess the interference of intrinsic biases with DNaseI cleavage patterns at TF binding footprints, we aligned the 4.5 million consensus footprints with more than 10,000 human TF ChIP-seq datasets from the Cistrome Data Browser database^{34,35} and collected two sets of footprint regions: “TF binding hotspots”, the footprint regions overlapping with TF binding sites detected from more than 3,000 ChIP-seq datasets, and “TF binding deserts”, the footprint regions that do not overlap with any TF binding sites from any ChIP-seq dataset or any human TF motif sequence from the HOCOMOCO database³⁶ (Fig. S5e). We compared the aggregate DNaseI cleavage patterns and the bias score patterns in these two sets of consensus footprint regions and found that the DNaseI cleavage pattern in TF binding hotspot regions was dominated by the classic DNaseI footprint pattern with little interference from the bias (Fig. 4e,f), while in the TF binding desert regions, the entire cleavage “footprint” pattern was apparently determined by the two spikes from the intrinsic bias (Fig. 4g,h). These results suggested that although the overall DNaseI cleavage pattern is indicative of TF binding, the intrinsic cleavage bias may interfere with the real footprint pattern, and the effect on those footprint regions with rare TF binding events can be drastic. These observations were reproducible in DNase-seq data from different cell and tissue types (Fig. S6a-b).

To quantify the level of intrinsic cleavage biases in a consensus footprint region, we designed a footprint bias score (FBS), namely, the relative SELMA-estimated bias score at the two spikes compared to the average SELMA bias score across the rest of the region for each footprint. Consistent with what we observed in the aggregate view, the FBSs for TF binding deserts were significantly higher while the FBSs for TF binding hotspots were significantly lower than the background of all consensus footprints (Fig. 4i, $p < 10^{-5}$, by Wilcoxon test), indicating that FBS might be an informative feature of the consensus footprint regions and might help separate true TF binding footprints from false-positive DNaseI footprint patterns. Next, we used a model-

based approach to assess the potential power of SELMA-derived FBS in boosting the performance of TF binding inference from DNase-seq signal patterns on consensus footprint regions containing the TF motif sequence. For every TF with both the motif and ChIP-seq data available, we employed a multinomial logistic regression model using different sets of features, each of which may include DNase-seq read count, DNaseI footprint score produced from an existing method, and an optional FBS, to infer the TF binding occupancy (determined by peak occurrence in ChIP-seq data) in the motif-containing footprint regions and compared the inference performances. We included our previous footprint method¹⁴ and several representative methods that outperform others, Wellington⁹ and HINT^{11,19} (with and without bias correction mode) in this comparison. For each footprint method, we compared the TF binding inference performance of the original method (read count + footprint score), the original method plus a randomized naïve k-mer bias score as a control, and the original method plus the SELMA bias score (FBS), using an integrated rank score strategy. For a total of 48 TFs in the human cell line GM12878 and 45 TFs in K562, the models with SELMA FBS significantly outperformed those without SELMA FBS, regardless of the approach used originally to calculate the footprint score (Fig. 4j, Fig. S6). These results suggested that considering the intrinsic cleavage bias could always improve the performance of existing footprint computational methods for TF binding inference.

To assess whether the intrinsic bias has different levels of interference with the footprint regions for different TFs, we selected two subsets of sequence motif-containing footprint regions for each TF as the top 10% with the highest FBS or the bottom 10% with the lowest FBS, and compared the frequencies of observing actual TF binding (overlapping with a ChIP-seq peak) in these subsets of footprints. We found that different TFs had various preferences for binding at low-FBS footprints or high-FBS footprints. For most TF samples in the two cell lines, the footprint regions with a low FBS showed a significantly higher chance of observing actual TF

binding than those with a high FBS (Fig. 4k). For example, HLH family TFs such as MYC and MAX¹⁷ showed a higher fraction of TF binding occupancies in low-FBS footprint regions than at high-FBS regions. Zinc finger family TFs including CTCF, REST, and NRF1^{6,17}, showed a higher chance of TF binding in high-FBS footprint regions than in low-FBS regions (Fig. S6d). This is possibly because of the different DNA binding affinities and DNA residence times between different TFs¹⁵. TFs with stronger DNA-binding affinity or longer residence time might be less affected by intrinsic cleavage biases, while TFs with shorter residence time or weaker binding affinity might be more severely affected by intrinsic biases. These results suggested that intrinsic cleavage biases might affect different TFs at various levels in divergent directions, and considering intrinsic bias should improve the footprint-based TF binding inference for most TFs.

SELMA improves the accuracy of cell clustering from single-cell ATAC-seq data

Single-cell ATAC-seq (scATAC-seq) technologies enable the detection of accessible chromatin regions at single-cell resolution in thousands of cells at a time²¹⁻²³. Due to the scarcity of cleavage events in an individual cell, most chromatin accessibility regions in a single cell have only one aligned fragment, making the potential influence of intrinsic cleavage biases more substantial in scATAC-seq data than in bulk ATAC-seq data. We collected three scATAC-seq datasets generated using two different platforms, human hematopoietic cells³⁷, human cell lines²¹, and a mouse primitive gut tube sample, and found that the scATAC-seq data contained similar intrinsic cleavage biases to bulk data with highly correlated bias scores estimated by SELMA (Fig. S7a-c). We estimated the average cleavage bias for each individual cell (cell bias score, CBS) and found that the distribution of CBS was different across cell cluster patterns, batches, and annotated cell types (Fig. 5a-c, Fig. S7d-i). These observations indicated that, regardless of experimental platforms and biological systems, the Tn5 intrinsic cleavage biases can contribute to cellular heterogeneity observed from scATAC-seq data and can affect cell clustering analysis.

We next assessed whether considering intrinsic cleavage biases can improve cell clustering based on scATAC-seq data. We used the actual cell-type labels as the clustering ground truth for the human hematopoietic cell sample and the mixed human cell line sample and used scRNA-seq data-projected cell labels as a “pseudo” ground truth for the mouse primitive gut tube sample. We used the adjusted Rand index³⁸ to quantify the accuracy of an unsupervised clustering result. As scATAC-seq-based cell clustering can be performed on all chromatin accessibility regions (ATAC-seq peaks), we sought to address whether removing peaks with high intrinsic biases can increase the clustering accuracy. We first tested a simple clustering approach that involved dimensionality reduction using principal component analysis (PCA) followed by K-means clustering. For all three scATAC-seq datasets, cell clustering using SELMA-debiased peaks only (excluding 1%-50% of peaks with the highest PBS) is significantly more accurate than clustering using all peaks under the background of randomly excluding the same number of peaks (Fig. 5d-f). Then, we tested several representative scATAC-seq analysis tools for cell clustering, including APEC³⁹, Seurat⁴⁰, scran⁴¹, and snapATAC⁴², and found that using SELMA-debiased peaks can significantly improve the clustering results in most cases (Fig. 5g-j, Fig. S7j-r). These results suggested that scATAC-seq data contain intrinsic cleavage biases that can be reduced by excluding SELMA-identified high-bias peaks from the cell clustering analysis.

Discussion

The existence of enzymatic cleavage biases in DNase-seq and ATAC-seq experiments has been widely acknowledged in the field, but to what extent such intrinsic biases affect data analysis has not been systematically assessed. SELMA provides a quantitative approach for the

accurate and robust estimation of intrinsic cleavage biases in both bulk and single-cell chromatin accessibility sequencing data and requires fewer sequence reads than the naïve k-mer model. Taking Tn5 dimerization into consideration, SELMA yields more accurate bias estimation specifically for ATAC-seq data by including k-mer sequences at two Tn5 cleavages 9bp apart. SELMA can improve functional analysis and interpretation of chromatin accessibility profiles. On bulk DNase-seq data, we showed that considering SELMA-estimated biases can improve TF binding inference from ENCODE DNaseI consensus footprints, with superior performances compared to existing tools, including some that already considered “biases”. On single-cell ATAC-seq data, we demonstrated that selecting chromatin accessibility regions based on the SELMA-estimated bias level can greatly increase the accuracy of cell clustering by several existing tools and hence can help researchers obtain more biological insights from the data.

SELMA is built on top of the widely used k-mer model by combining simplex encoding and a linear model. Simplex encoding has the unique ability to capture the pairwise symmetry and orthogonality between mononucleotides and interactions within each dinucleotide. It significantly reduces the degrees of freedom without losing any variance information compared to the naïve k-mer model. These properties enable SELMA to make robust estimations from fewer reads or smaller datasets than are required by other approaches, hence enabling *de novo* bias estimation from a smaller DNA molecule, such as mtDNA in a DNase/ATAC-seq sample, without having to generate a separate genomic naked DNA sample. However, SELMA still relies on sufficient read counts for each k-mer for an accurate estimation. Although the performance of SELMA may increase with k, this effect is not unlimited. The performance using 12-mer is not as good as using 10-mer possibly because there are not enough reads in the dataset for many 12-mers. Nevertheless, SELMA works well for most existing DNase/ATAC-seq datasets tested in our study. In addition, the feasibility and superior performance of simplex encoding have also

been demonstrated for TF motif characterization⁴³. Following similar encoding strategies, SELMA can potentially be applied to any k-mer-based model or to any high-throughput sequencing data for robust sequence bias estimation and pattern recognition.

The k-mer biases estimated by SELMA are consistent across species and cell types, as expected based on the assumption that the cleavage biases are intrinsic features of the enzymatic assay. Therefore, one can directly use the SELMA-estimated k-mer biases from DNaseI or Tn5-digested naked DNA data as universal intrinsic biases for DNase-seq and ATAC-seq, respectively, and incorporate these precalculated bias scores into the data analysis. However, this bias dataset is not guaranteed to remain accurate in many species that have not been profiled, as there might be unknown biases that have not been captured. Although we are confident that the SELMA-estimated results should still be largely consistent, one can always use the SELMA package for *de novo* estimation of cleavage biases from one's own datasets.

When applying SELMA-estimated FBS to correct biases in DNase footprints, our data were limited to ENCODE DNaseI consensus footprints as a proof-of-principle study. While this ENCODE dataset is thus far the largest DNase footprint repertoire, users might be interested in *de novo* detection of footprints from their customized DNase/ATAC-seq data. As many bioinformatics tools are already available for such tasks using various computational models²⁰, SELMA or SELMA-generated bias information can be incorporated into each of those models for improved performance for more accurate regulatory DNA identification from footprints. Similarly, while we show significant improvement of scATAC-seq cell clustering by removing peaks with high PBS under the framework of most existing tools, performance still varies across these tools due to the different statistical and computational models used in these methods. Further optimization of existing methods or development of new methods is expected for the integration of SELMA for the systematic removal of intrinsic biases.

Methods

High-throughput sequencing data collection and processing

Data collection. Publicly available ATAC-seq, single-cell ATAC-seq, DNase-seq and ChIP-seq data used in this study were collected from Gene Expression Omnibus (GEO) and the ENCODE consortium. The metadata and accession numbers are listed in [Table S1](#).

Processing of bulk ATAC-seq and DNase-seq data. Raw sequencing reads were aligned to the GRCh38 (hg38) reference genome with bowtie2 (v2.2.9) (-X 2000 for paired end data)⁴⁴. Low-quality reads (MAPQ < 30) were discarded. For paired-end sequencing data, reads with two ends aligned to different chromosomes (chimeric reads) were also discarded. For paired-end data, reads with identical 5' end positions for both ends were regarded as redundant reads and discarded. The nonredundant reads were separated into chromosomal DNA (chromatin reads) and mitochondrial DNA (mtDNA reads) based on their genomic location. Peak detection was performed on the nonredundant chromatin reads using MACS2 (v2.1.2) (-q 0.01, --extsize 50) and +/- 200bp centered on the peak summits was collected as the genome-wide chromatin accessible regions. The accessible regions in each dataset were separated into 5 groups from highest accessibility to lowest accessibility based on the read count on each peak (for [Fig. S1](#)). The 5' end nucleotides of each read were piled up to generate the genome-wide observed cleavage profile.

Processing of single-cell ATAC-seq data. For the human HSC sample, raw sequencing reads were aligned to the GRCh38 (hg38) reference genome with bowtie2 (-X 2000). Low quality reads (MAPQ < 30), chimeric reads and duplicate reads for each individual cell were discarded.

For the mouse gut tube sample, scATAC-seq data from the 10x Genomics platform were preprocessed with Cell Ranger ATAC with the default parameters to generate fragments for each individual cell. The fragment file from the standard Cell Ranger ATAC output was used as input for the subsequent analysis. Because the Cell Ranger ATAC pipeline shifted from the Tn5 cleavage sites to +4/-5 bp in generating the fragment file, the coordinates were shifted back to represent the actual cleavage loci. For both datasets, cells with more than 10,000 reads were retained for analysis.

Processing of ChIP-seq data. ChIP-seq data were collected and processed as described in a previous study⁴⁵. In brief, based on the quality measurements from the Cistrome data browser (CistromeDB)^{34,35}, we included only the ChIP-seq datasets that pass at least four out of the first five QC measures (i.e., sequence quality, mapping quality, library complexity, ChIP enrichment, and signal-to-noise ratio). We also excluded the ChIP-seq datasets with fewer than 1000 peaks identified by MACS2⁴⁶ with an FDR cutoff of 0.01. For TFs with multiple datasets available in the same cell type, we kept only one dataset with the best quality based on passing most of the 6 CistromeDB QC measures, including sequence quality, mapping quality, library complexity, ChIP enrichment, signal-to-noise ratio, and regulatory region enrichment. The ChIP-seq peak files (detected by MACS2) for the selected samples were downloaded from CistromeDB for the following analysis.

TF motif analysis. The motifs of human TFs were collected from the HOCOMOCO database (v11)³⁶, and the genome-wide motif sites of TFs were detected by FIMO (v4.12.0) in the MEME package⁴⁷. Motif sites located outside of the genome-wide 36bp unique mappable regions were excluded from the analysis. In total, 25,027,116 motif sites for 66 TF motifs from the human genome were included in the analysis.

SELMA model

Naïve k-mer bias model for intrinsic cleavage bias estimation. The naïve k-mer biases were calculated as described in a previous study¹⁴. In brief, a naïve k-mer bias was estimated as the observed frequency of the cleavages with the k-mer relative to the frequency of that k-mer in the background. For each mapped sequence read in a DNase-seq or ATAC-seq dataset, the enzymatic cleavage site was between genomic positions i and $i-1$ for the plus (+) strand reads and between i and $i+1$ for the minus (−) strand reads, where i represents the genomic position of the 5' nucleotide of the reads. The associated k-mer sequence was thus assigned as $\left[i - \frac{k}{2}, i + \frac{k}{2} - 1\right]$ for the plus strand reads and $\left[i - \frac{k}{2} + 1, i + \frac{k}{2}\right]$ for the minus strand reads. The naïve k-mer bias score for k-mer j is defined as the number of all observed cleavages with that k-mer divided by the occurrences of that k-mer in the background:

$$S_j = \frac{N_j}{M_j}$$

where N_j is the count of cleavages with k-mer j , and M_j is the total count of occurrences of k-mer j in the background (Fig. 1a). In the case of chromatin DNA, this background included 400 bp from each chromatin accessible region centered at the peak summit detected by MACS2. The background in the naked DNA samples included genome-wide 36bp unique mappable regions. The background in mtDNA included the mitochondrial DNA sequence. The median was further subtracted from the bias scores for visualization in the scatter plots (e.g., Fig. 1d-k).

In the naïve k-mer model, the bias score for each k-mer is independent and empirically determined from the data. The number of independent variables in the model is the total number of k-mers, i.e., 4^k .

SELMA model for intrinsic cleavage bias estimation. In SELMA, a simplex encoding model followed by a linear model was employed on top of the naïve k-mer model to better estimate the

intrinsic cleavage biases for each k-mer. To efficiently encode a k-mer sequence considering their intrinsic similarities, a simplex encoding model was adopted from previous studies^{25,26}, in which the DNA nucleotides were encoded as vectors representing the four tetrahedral vertices of a regular 0-centered simplex (Fig. 1b).

$$\begin{aligned} A &= \begin{bmatrix} 1 & -1 & -1 & -1 \end{bmatrix} \\ C &= \begin{bmatrix} -1 & 1 & -1 & -1 \end{bmatrix} \\ G &= \begin{bmatrix} -1 & -1 & 1 & -1 \end{bmatrix} \\ T &= \begin{bmatrix} -1 & -1 & -1 & 1 \end{bmatrix} \end{aligned}$$

In the simplex encoding, the vectors representing the four nucleotides are of equal length, mutually orthogonal, and equidistant from each other. To account for interactions between adjacent nucleotides, a dinucleotide was additionally encoded as the outer product of the two vectors associated with the two nucleotides:

$$\begin{aligned} AA &= \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 \end{bmatrix} \\ AC &= \begin{bmatrix} -1 & 1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 \end{bmatrix} \\ AG &= \begin{bmatrix} -1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix} \\ AT &= \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 \end{bmatrix} \\ CA &= \begin{bmatrix} -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix} \\ CC &= \begin{bmatrix} 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \end{bmatrix} \\ CG &= \begin{bmatrix} 1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \end{bmatrix} \\ CT &= \begin{bmatrix} -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 \end{bmatrix} \\ GA &= \begin{bmatrix} -1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 \end{bmatrix} \\ GC &= \begin{bmatrix} 1 & -1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 \end{bmatrix} \\ GG &= \begin{bmatrix} 1 & 1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 \end{bmatrix} \\ GT &= \begin{bmatrix} -1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 \end{bmatrix} \\ TA &= \begin{bmatrix} 1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 \end{bmatrix} \\ TC &= \begin{bmatrix} -1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 \end{bmatrix} \\ TG &= \begin{bmatrix} -1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 \end{bmatrix} \\ TT &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \end{aligned}$$

One can show that these vectors for dinucleotide interactions are also of equal length, mutually orthogonal, and equidistant from each other. In fact, in the simplex encoding, mononucleotides

and dinucleotides were encoded as rows in a Hadamard matrix of order 4 and a Hadamard matrix of order 16, respectively.

Considering both mononucleotides and dinucleotide interactions, a k -mer can then be encoded as k mononucleotides and $k - 1$ dinucleotides. Therefore, the dimensionality of a k -mer simplex encoding is

$$p(k) = 1 + 3k + 9(k - 1) = 12k - 8$$

In SELMA, the intrinsic k -mer biases were then estimated in a linear model with these $12k - 8$ parameters using the observed naïve k -mer bias scores (Fig. 1c). In detail, we had the formula:

$$y \sim \sum_{i=1}^{12k-8} a_i x_i \quad (1)$$

where each observation y is the naïve k -mer bias score; $x_i \in \{1, -1\}$, is the independent variable in the simplex encoding vectors; and a_i is the parameter to be estimated. After linear regression, the model-fitted value $\hat{y}$ was defined as the SELMA bias score for each k -mer.

Genome-wide cross-correlation analysis. For the analysis presented in Fig. 2b,c, reads from plus (+) and minus (-) strands on chromatin accessible regions (peaks) were collected separately to generate plus strand observed cleavage profile and minus strand observed cleavage profile, respectively. The Pearson correlation coefficient between the plus strand signal and the k -bp shifted minus strand signal is:

$$\rho_k = \frac{\sum_i (P_i - \bar{P})(M_{ik} - \bar{M}_k)}{\sqrt{\sum_i (P_i - \bar{P})^2 \sum_i (M_{ik} - \bar{M}_k)^2}} \quad (2)$$

where P_i is the log-scaled plus strand cleavage count at genomic position i ($\log_2(\text{cleavage}+1)$), M_{ik} is the log-scaled minus strand cleavage count at genomic position i with a k -bp shift, $\bar{P}$ is the mean of all the P_i , and $\bar{M}_k$ is the mean of M_{ik} for all i . All genomic positions on the genome-wide DNase/ATAC-seq peaks were included in the analysis. k is chosen from 1 to 20 (x-axis in Fig. 2b,c).

Comparison of different bias estimation methods. We use a_i^+ and a_i^- to denote the “5’ only” intrinsic sequence bias score at genomic position i on the plus strand and minus strand, respectively. We use $a_i'^+$ and $a_i'^-$ to denote the bias score from other bias estimation methods in this section. Different bias estimation methods used in Fig. 2 and its associated sections are listed below:

- For the “5’ only” method, a_i^+ was calculated based on the k-mer ratio associated with the k-mer spanning positions $[i - k/2, i + k/2 - 1]$ on the plus strand, and a_i^- was calculated in a similar way based on the nucleotides spanning genomic positions $[i - k/2 + 1, i + k/2]$ on the minus strand (reverse complement of the DNA sequence on the corresponding plus strand). This method was applied to both DNase-seq and ATAC-seq, while the other methods were applied only to ATAC-seq as they were specifically designed for ATAC-seq.
- For SELMA (Fig. 2a), the bias score was calculated as the geometric mean of the “5’ only” bias score at the given position and the “5’ only” bias score at 9bp downstream of the other strand, i.e., $a_i'^+ = \sqrt{a_i^+ \times a_{i+9}^-}$, and $a_i'^- = \sqrt{a_i^- \times a_{i-9}^+}$.
- For the model in Martins et al¹⁶, the bias score at genomic position i was calculated in a similar way to the “5’ only” method but using a gapped 11-mer model. The model could be represented as XXXXXXNNNXNXCXXNNXNNNNXNX, in which position i was represented by C; positions that were ignored were represented by X and informative positions were represented by N.
- For the model in Baek et al¹⁸, the bias score at genomic position i was calculated in a similar way to the “5’ only” method but the cleavages were shifted +4/-5 bp for +/- strand cleavages. In practice, following the description in the “bagfoot” package, the bias score at position i was calculated as $a_i'^+ = a_{i+5}^+$, and $a_i'^- = a_{i-5}^+$.

- For the model in Calviello et al⁸, the bias score at genomic position i was calculated in a similar way to the “5’ only” method but using the 6-mer bias table provided in the study.

Calculation of observed and bias-expected cleavage. The 1bp at 5’ end positions for DNase-seq or ATAC-seq reads were piled up to generate the genome-wide observed cleavage profile. To generate the bias-expected cleavage profile, for a 50-bp window centered on nucleotide i , we placed the same number of observed cleavages following the sequence bias contribution in this window. Let $\widehat{N}_i^s$ represent the bias expected cleavage at position i on strand $s \in \{+, -\}$, N_i^s represent the observed cleavage at position i on strand s , and a_i^s denote the intrinsic cleavage bias (estimated with any of the above methods) at position i on strand s . Then we have

$$\widehat{N}_i^s = N_i^s y_i^s \quad (3)$$

$$y_i^s = \frac{2^{a_i^s}}{\sum_{j=i-25}^{i+24} 2^{a_j^s}} \quad (4)$$

We used the Pearson correlation coefficient to compare the observed cleavage profile and the bias-predicted cleavage profile (Fig. 2). We only considered positions within the accessible regions to ensure that all positions had sufficient reads in the 50-bp window for accurate estimation.

DNase footprint analysis

Data processing. The genome-wide DNaseI consensus footprint regions were downloaded from Reference⁷

(<https://resources.altius.org/~jvierstra/projects/footprinting.2020/consensus.index/>). The

observed DNaseI cleavage profile from a DNase-seq dataset and DNaseI SELMA bias scores across +/- 50bp centered on the footprint region were plotted as heatmaps (Fig. 4a,b). The footprint regions were ordered by the footprint lengths, and each 1000 footprint regions with

similar lengths were compressed as one row in the heatmap for better visualization. The +/- strand signals were plotted separately. We aligned the footprint regions based on the two observed bias spikes in each footprint region (located 7bp to the right of the left boundary and 7bp to the left of the right boundary of the footprint, labeled as -0 and +0 in Fig. 4c-h). The center regions between the bias spikes were scaled to 4 bins to align footprint regions with different lengths.

Footprint bias score. The footprint bias score (FBS, in Fig. 4i-k) was defined as the difference between “spike bias” and “center bias”. The “spike bias” was calculated as the average of the two SELMA bias scores at the spike positions, while the “center bias” was calculated as the median SELMA bias score at the rest of the positions in the footprint region. Let FBS_j , b_j , and c_j denote the footprint bias score, spike bias and center bias of footprint j , respectively. The FBS, spike bias, and center bias are given by:

$$FBS_j = b_j - c_j \quad (5)$$

$$b_j = \overline{a_{i \in B}^{s \in \{+, -\}}} \quad (6)$$

$$c_j = \widetilde{a_{i \in C}^{s \in \{+, -\}}} \quad (7)$$

where a_i^s represents the SELMA bias score at genomic position i on the strand $s \in \{+, -\}$, overbar represents the average, tilde represents the median, and B and C represent spike positions (within the flanking 1bp of the bias spikes) and the remaining positions of the footprint j , respectively. To calculate the randomly shuffled FBS (random k-mer bias), the FBS was calculated in the same way, but the bias score a_i^s for each position was randomly selected from the SELMA bias score table.

Calculation of footprint score with different methods. The following methods were used to calculate footprint scores for comparison:

- Raw footprint: The raw footprint score was calculated following a previous study¹⁴, using the formula $f = -\left(\log \frac{n_C+1}{n_R+1} + \log \frac{n_C+1}{n_L+1}\right)$, where n_C , n_R and n_L denote the DNase cleavage count in the motif region, and the flanking regions to the right and left of the motif, respectively. The flanks are both the same length as the motif.
- Wellington footprint: We used Wellington⁹ (v0.2.0) with default parameters to detect genome-wide footprints and selected significant footprint regions with p-values < 1e-10. The output footprint score was assigned to the TF motif overlapping with the footprint region.
- HINT footprint: We used HINT^{11,19} (v0.12.3) with default parameters to detect genome-wide footprints. For bias correction mode, we used an additional parameter: --bias-correction. The footprint score of each footprint region was assigned to the overlapping TF motifs.

Inference of TF binding with different features. For each TF in Fig. 4j, the TF motifs overlapping with consensus footprint regions were collected as target regions. DNase-seq read count in the footprint region (“reads”), footprint score from an existing method, and bias score were used as features in a multinomial logistic regression model to infer TF occupancy at footprint-overlapping motif regions. For each available method, model training with cross-validation and predictions were performed using different combination of features: “original method” refers to reads + footprint score. An additional feature of either SELMA FBS or a random k-mer bias was added for different models. We used a performance measure integration approach¹¹ to evaluate different models’ prediction power. For each model, we calculated the area under the ROC curve (AUROC) at 100%, 10%, and 1% false positive rate (FPR). We also calculated the area under the precision-recall curve (AUPRC) at 100%, 10%,

and 1% recall. We then combined these six performance measures as a rank score S , defined as

$$S = \frac{1}{6} \sum_i -\log \frac{r_i}{N+1}$$

where $i = 1, \dots, 6$ denotes the 6 performance measures, r_i is the rank of a model for performance measure i , and N is the total number of models.

To calculate the random k-mer bias, we randomly permuted the SELMA k-mer bias table and generated the k-mer bias table for a “simulated enzyme”. We used this “simulated” bias table to calculate the FBS for each footprint region and performed TF inference modeling similar to what we did for the DNaseI SELMA FBS. This permutation was performed 100 times to generate 100 performance rank scores for random k-mer bias used as controls.

Motif prediction power of TF binding sites. In Fig. 4k, we collected the genome-wide motif sites overlapping with consensus footprint regions and the ChIP-seq peaks for each TF with HOCOMOCO³⁶ motif and ChIP-seq data available in GM12878 or K562 (Table S1). As a result, we collected genome-wide motif sites for 65 TFs (48 in GM12878 and 45 in K562), and 35,721,822 motif sites in total. For each TF, we selected two subgroups of its motif sites based on the FBS of their overlapped footprint regions: the top 10% of motif sites with the highest FBS, defined as “sites with high-bias footprint”; and the bottom 10% of motif sites with the lowest FBS, defined as “sites with low-bias footprint”. We calculated the proportion of motif sites overlapping with TF ChIP-seq peaks for each of the two subgroups and plotted on a scatter plot (Fig. 4k). To test whether low-bias footprints tend to have more TF binding than high-bias footprints for most TFs, we conducted a t-test comparing the distribution of TF binding log likelihood ratios of low-bias over high-bias footprints to the standard normal distribution, and the test p-value is labeled in Fig. 4k.

Single-cell ATAC-seq clustering analysis

Gold standard of single-cell clustering. For single-cell ATAC-seq data in the human hematopoietic cells and human cell line samples, the cell-type information for each individual cell was used as the ground truth, or the gold standard. For single-cell ATAC-seq data in the mouse gut tube sample, the cell-type information was assigned based on label transfer⁴⁰ from the single-cell RNA-seq dataset in the same system⁴⁸ (GSE136689), as the “pseudo” ground truth, or the silver standard. In detail, we integrated scRNA-seq and scATAC-seq data using the ArchR package⁴⁹ (v1.0.1). Individual cells with a high RNA integration score (unconstrained predicted score < 0.56) were collected as high-quality cells for the analysis. The cutoff of the RNA integration score (0.56) was determined by maximizing the interclass variance in the RNA integration score using Otsu’s method⁵⁰. Cell types represented by fewer than 10 cells were discarded.

Average bias for individual cells. The average bias for each individual cell (median SELMA cell bias score, median CBS, used in Fig. 5) was calculated as the median of the bias for all the fragments in the individual cell. The bias for each paired-end fragment was calculated as the mean of the SELMA bias scores for the 5’ end and the 3’ end.

Peak detection and peak bias score calculation. We first combined all the single-cell ATAC-seq reads in the dataset and performed MACS2⁴⁶ peak calling with a q-value cutoff of 0.1 to include all the potential accessible regions in the human hematopoietic cell and human cell line systems (100,456 peaks detected for human hematopoietic cells and 83,318 peaks for human cell lines). For the 10X scATAC-seq data in mouse gut tube, the q-value cutoff of 0.01 was used for peak calling (146,098 peaks detected). Potential accessible regions with fewer than 10 covered reads or more than 4,000 covered reads were discarded in the subsequent analysis. To

consider the effect of intrinsic cleavage bias in scATAC-seq data, we summarized the bias for each potential accessible region by calculating the peak bias score (PBS), defined as the median SELMA bias score of all reads (from all individual cells) located in the peak region.

Cell clustering with scATAC-seq data. In the scATAC-seq clustering analysis, as a naïve method, we performed PCA dimensionality reduction on the accessible regions by individual cells matrix of normalized read count and kept 60 PCs, followed by k-means clustering. The number of clusters (k) in k-means clustering was determined as the actual number of cell types in the dataset. To evaluate the performance of cell clustering, we compared the clustering results with the gold standard using adjusted rand index (ARI)³⁸. We further repeated the k-means clustering with 100 different random seeds to ensure the stability of the clustering results and the average ARI of the 100 runs was used to estimate the clustering accuracy. We also applied several published methods, including APEC³⁹, Seurat⁴⁰, scran⁴¹, and snapATAC⁴² with default parameters, on the same dataset. We selected the top 60 dimensions for those methods at the dimensional reduction step (PCA, Seurat, scran and snapATAC) to keep more variance potentially introduced by intrinsic biases. We used the ArchR package to practice the Seurat and scran clustering methods. To explore the effect of intrinsic cleavage bias on scATAC-seq analysis, we selected different percentages (from 50% to 99%, with a 1% increment) of peaks with the lowest PBS (SELMA debiased peaks in Fig. 5) as input to perform cell clustering. We also randomly selected the same percentage of peaks as a control to estimate the relative rank of the clustering performance from using all peaks and debiased peaks. In detail, for each percentage from 50% to 99%, we randomly sampled peaks 100 times and defined the relative rank as the number of ARIs from random samples that were less than the ARI from the same percentage of debiased peaks or all peaks. The relative ranks for different percentages from 50% to 99% are summarized and plotted as boxplots in Fig. 5.

Code and data availability

The SELMA package is available at <https://github.com/zang-lab/SELMA>. User instructions and example data files can be found in the README document. Essential annotation data, analysis results, and scripts are also available in the repository. The mouse gut tube scATAC-seq dataset has been deposited in the Gene Expression Omnibus (GEO) with accession number GSE168373. Publicly available data used in this study can be found in the GEO and the ENCODE project. Accession numbers for all data used in the study are listed in Additional file 2: Table S1.

Acknowledgements

The authors thank Dr. Tae-Hee Kim for sharing the mouse primitive gut tube scATAC-seq data and Dr. Xiaole Shirley Liu for helpful discussions. This work was supported by US National Institutes of Health (NIH) grants R35GM133712 and K22CA204439 (to C.Z.), R35GM128635 (to M.J.G.), National Science Foundation grant NSF-2048991 (to T.Z.), UVA Cancer Center Farrow Fellowship and the NCI Cancer Center Support Grant P30 CA44579 (to S.S.H.).

Author contributions

C.Z. conceived and directed the research. S.S.H., C.A.M., and C.Z. designed the method. L.L., M.J.G., K.D., and T.Z. contributed to the method design. S.S.H. implemented the method and analyzed the data. Q.L. and W.M. helped the data analysis. S.S.H. and C.Z. wrote the manuscript. All authors approved the manuscript.

Competing interests

The authors declare that they have no competing interests.

References

1. McGhee, J. D., Wood, W. I., Dolan, M., Engel, J. D. & Felsenfeld, G. A 200 base pair region at the 5' end of the chicken adult β -globin gene is accessible to nuclease digestion. *Cell* **27**, 45–55 (1981).
2. Keene, M. A., Corces, V., Lowenhaupt, K. & Elgin, S. C. DNase I hypersensitive sites in *Drosophila* chromatin occur at the 5' ends of regions of transcription. *Proc National Acad Sci* **78**, 143–146 (1981).
3. Boyle, A. P. *et al.* High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* **132**, 311–322 (2008).
4. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10**, 1213–1218 (2013).
5. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods* **6**, 283–289 (2009).
6. Nepf, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).
7. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020).
8. Calviello, A. K., Hirsekorn, A., Wurmus, R., Yusuf, D. & Ohler, U. Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling. *Genome Biol* **20**, 42 (2019).
9. Piper, J. *et al.* Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Research* **41**, e201–e201 (2013).
10. Sherwood, R. I. *et al.* Discovery of directional and nondirectional pioneer transcription

factors by modeling DNase profile magnitude and shape. *Nat Biotechnol* **32**, 171–178 (2014).

11. Li, Z. et al. Identification of transcription factor binding sites using ATAC-seq. *Genome Biol* **20**, 45 (2019).

12. Bentsen, M. et al. ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat Commun* **11**, 4267 (2020).

13. Meyer, C. A. & Liu, X. S. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat Rev Genet* **15**, 709–721 (2014).

14. He, H. H. et al. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat Methods* **11**, 73–78 (2014).

15. Sung, M.-H., Guertin, M. J., Baek, S. & Hager, G. L. DNase Footprint Signatures Are Dictated by Factor Dynamics and DNA Sequence. *Molecular Cell* **56**, 275–285 (2014).

16. Martins, A. L., Walavalkar, N. M., Anderson, W. D., Zang, C. & Guertin, M. J. Universal correction of enzymatic sequence bias reveals molecular signatures of protein/DNA interactions. *Nucleic Acids Res* **46**, gkx1053- (2017).

17. Yardımcı, G. G., Frank, C. L., Crawford, G. E. & Ohler, U. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic Acids Research* **42**, 11865–11878 (2014).

18. Baek, S., Goldstein, I. & Hager, G. L. Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity. *CellReports* **19**, 1710–1722 (2017).

19. Gusmao, E. G., Allhoff, M., Zenke, M. & Costa, I. G. Analysis of computational footprinting methods for DNase sequencing experiments. *Nature Methods* **13**, 303–309 (2016).

20. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker’s guide to ATAC-seq data analysis. *Genome Biol* **21**, 22 (2020).

21. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

22. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by

combinatorial cellular indexing. *Science* **348**, 910–914 (2015).

23. Lake, B. B. et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nature Biotechnology* **36**, 70–80 (2017).

24. Preissl, S. et al. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat Neurosci* **21**, 432–439 (2018).

25. Zhang, C.-T. & Zhang, R. Analysis of distribution of bases in the coding sequences by a digrammatic technique. *Nucleic Acids Res* **19**, 6313–6317 (1991).

26. Stormo, G. D. Maximally Efficient Modeling of DNA Sequence Motifs at All Levels of Complexity. *Genetics* **187**, 1219–1224 (2011).

27. Adey, A. et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol* **11**, R119 (2010).

28. Sos, B. C. et al. Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. *Genome Biol* **17**, 20 (2016).

29. Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics* **48**, 1193–1203 (2016).

30. Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature Methods* **14**, 959–962 (2017).

31. Kornberg, R. D. Chromatin structure: a repeating unit of histones and DNA. *Science* **184**, 868–871 (1974).

32. Caron, F., Jacq, C. & Rouvière-Yaniv, J. Characterization of a histone-like protein extracted from yeast mitochondria. *Proceedings of the National Academy of Sciences* **76**, 4265–4269 (1979).

33. DeFrancesco, L. & Attardi, G. In situ photochemical crosslinking of HeLa cell mitochondrial DNA by a psoralen derivative reveals a protected region near the origin of replication. *Nucleic Acids Research* **9**, 6017–6030 (1981).

34. Mei, S. et al. Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility

770 data in human and mouse. *Nucleic Acids Research* 45, D658–D662 (2017).

771 35. Zheng, R. et al. Cistrome Data Browser: expanded datasets and new tools for gene
772 regulatory analysis. *Nucleic Acids Research* **47**, D729–D735 (2018).

773 36. Kulakovskiy, I. V. et al. HOCOMOCO: towards a complete collection of transcription factor
774 binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res* 46,
775 gkx1106- (2017).

776 37. Buenrostro, J. D. et al. Integrated Single-Cell Analysis Maps the Continuous Regulatory
777 Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548.e16 (2018).

778 38. Rand, W. M. Objective Criteria for the Evaluation of Clustering Methods. *J Am Stat Assoc*
779 **66**, 846 (1971).

780 39. Li, B. et al. APEC: an accession-based method for single-cell chromatin accessibility
781 analysis. *Genome Biol* 21, 116 (2020).

782 40. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21
783 (2019).

784 41. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis
785 of single-cell RNA-seq data with Bioconductor. *F1000research* **5**, 2122 (2016).

786 42. Fang, R. et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat*
787 *Commun* 12, 1337 (2021).

788 43. Zhao, Y., Ruan, S., Pandey, M. & Stormo, G. D. Improved Models for Transcription Factor
789 Binding Site Identification Using Nonindependent Interactions. *Genetics* **191**, 781–790 (2012).

790 44. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods*
791 **9**, 357–359 (2012).

792 45. Hu, S. et al. ncHMR detector: a computational framework to systematically reveal non-
793 classical functions of histone modification regulators. *Genome biology* **21**, 48 (2020).

794 46. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *9*, R137-9 (2008).

795 47. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif.

796 *Bioinformatics* **27**, 1017–1018 (2011).

797 48. Han, L. et al. Single cell transcriptomics identifies a signaling network coordinating
798 endoderm and mesoderm diversification during foregut organogenesis. *Nat Commun* **11**, 4158
799 (2020).

800 49. Granja, J. M. et al. ArchR is a scalable software package for integrative single-cell chromatin
801 accessibility analysis. *Nat Genet* **53**, 403–411 (2021).

802 50. Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *Ieee Transactions Syst*
803 *Man Cybern* **9**, 62–66 (1979).

804

805

Figure legends

Figure 1. SELMA framework for cleavage bias estimation.

(a) Schematic of a naïve k-mer model for cleavage bias estimation.

(b) Simplex encoding model. The coordinates of the 4 tetrahedral vertices of the cube encode the 4 nucleotides. Each dinucleotide is encoded as the outer product of the 2 mononucleotides.

p represents the number of parameters in a k-mer simplex encoding model.

(c) SELMA uses k-mer simplex encoding followed by a linear regression for k-mer cleavage bias estimation.

(d-k) Comparison between the naïve k-mer model (d,f,h,j) and SELMA (e,g,i,k) on k-mer cleavage bias scores estimated from DNase-seq data (d,e,h,i) and ATAC-seq data (f,g,j,k).

Each dot in a scatter plot represents an 8-mer, with its estimated bias score from different datasets represented in the x- and y-axes. Chromatin DNA (d-g) and naked DNA (h-k) from different human cell lines (d-i) or different species (j,k) were compared as labeled. R represents the Pearson correlation coefficient.

(l-o) Intrinsic cleavage bias estimation accuracy measured by correlation between genome-wide observed (OBS) and bias-expected (EXP) cleavages with different k-mers. A higher Pearson correlation coefficient (R) indicates a better prediction of the observed cleavage profile using the estimated biases.

Figure 2. SELMA consideration of dimeric Tn5 cleavages for ATAC-seq.

(a) Schematic of SELMA consideration of dimeric Tn5 cleavages.

(b,c) Cross-correlation between 5' cleavages and 3' cleavages across genome-wide accessible chromatin regions in the human GM12878 cell line from ATAC-seq (b) and DNase-seq (c) experiments. The x-axis represents the shift distance (in bp) between 5' and 3' cleavages.

(d-i) Comparison of ATAC-seq intrinsic cleavage bias estimation accuracy measured by

Pearson correlation coefficients (R) between genome-wide observed (OBS) and bias-expected (EXP) cleavages for human GM12878 (**d**) and K562 (**e**) cell lines, as well as several modified Tn5-based assays (**f-i**). Different bars represent different estimation approaches: gray for considering the 10-mer at the observed cut only (5' only); orange for SELMA; and blue for several published approaches. Modified Tn5-based assays include (**f**) THS-seq with standard Tn5; (**g**) THS-seq with mutated Tn5; (**h**) fast-ATAC-seq; and (**i**) omni-ATAC-seq.

Figure 3. SELMA bias estimation using mtDNA reads.

(**a**) Conventional approach to intrinsic cleavage bias estimation, in which a naked DNA sample is required for bias estimation, and mtDNA reads from the chromatin DNA ATAC-seq experiment are discarded.

(**b**) SELMA approach of intrinsic cleavage bias estimation using mtDNA reads. Chromatin reads and mtDNA reads from the same ATAC-seq dataset were separated for Tn5 cleavage profiling and bias estimation with SELMA, respectively.

(**c-d**) Scatter plots demonstrating the robustness of 10-mer bias scores estimated from mtDNA (mtDNA reads from chromatin ATAC-seq data, x-axis) and genomic DNA (reads from naked DNA ATAC-seq data, y-axis), comparing the naïve k-mer model (**c**) and SELMA (**d**). SELMA shows a more robust estimation using mtDNA reads with higher correlation than the naïve k-mer model.

(**e**) Comparison of the correlation coefficients between biases estimated from mtDNA reads and from naked DNA with different mtDNA read counts.

Figure 4. TF binding inference on DNaseI footprint regions is affected by intrinsic cleavage biases and improved by SELMA.

(**a,b**) Heatmaps of DNaseI cleavage patterns in the GM12878 cell line (**a**) and SELMA-estimated bias scores (**b**) around ENCODE DNaseI consensus footprint regions. 5' and 3'

patterns were plotted separately. Footprint regions were consistently ranked by footprint length, and signals for every 1000 regions with similar lengths were averaged as one row in the heatmap.

(c-h) Aggregate plots of DNaseI cleavage patterns (**c,e,g**) and SELMA-estimated bias scores (**d,f,h**) at all ENCODE DNaseI consensus footprint regions (**c,d**), footprint regions overlapping with TF binding hotspots (**e,f**), and footprint regions overlapping with TF binding deserts (**g,h**). DNaseI cleavages are from GM12878 for (**c & e**) and merged from GM12878, K562 and ENCODE tissues (**g**). Dashed lines represent estimated bias spikes in the footprint regions.

(i) SELMA-estimated footprint bias scores (FBS) for different groups of footprint regions. P-values were calculated by the Wilcoxon signed-rank test. The centerline, bounds of box, top line and bottom line represent the median, 25th to 75th percentile range, 25th percentile - 1.5*interquartile range (IQR) and 75th percentile + 1.5*IQR, respectively.

(j) Performance rank scores for transcription factor binding prediction from DNaseI footprint using various methods for the human GM12878 cell line. Boxplots with different colors represent different approaches for prediction as indicated in the legends. Different sets of boxplots represent footprint scores calculated by different published methods as labeled. P-values were calculated by the Wilcoxon signed-rank test.

(k) Scatter plot showing the heterogeneous effect of intrinsic cleavage biases on different TF motifs. Each data point represents motif sites for one TF in one cell type. Fractions of the TF-bound motif sites overlapping with high-FBS footprint regions and with low-FBS footprint regions are plotted on the x- and y-axes, respectively. More motifs located above the diagonal line indicate that more TFs are more likely to bind at low-FBS sites than at high-FBS sites. The p-value was calculated by t-test comparing the distribution of log likelihood ratios (y-axis/x-axis) to the standard normal distribution.

Figure 5. Intrinsic cleavage biases affect scATAC-seq data analysis.

884 **(a,b)** UMAP visualization of single cells from scATAC-seq data in human hematopoietic cells.
885 Cells are colored by cell type **(a)** or by cell bias score (CBS) **(b)**.
886 **(c)** CBS distributions of cells from different cell types and batches. Boxes are colored by cell
887 type using the same color palette as **(a)**.
888 **(d-f)** Comparisons of cell clustering accuracy before and after considering the peak bias score
889 (PBS) on scATAC-seq data for human hematopoietic cells **(d)**, human cell lines **(e)**, and the
890 mouse primitive gut tube **(f)**. K-means clustering was performed after PCA dimensionality
891 reduction. Blue, clustering using all ATAC-seq peaks; orange, using debiased peaks, i.e., 50%
892 to 99% of peaks with the lowest peak bias score (PBS). For each percentage from 50% to 99%
893 with a 1% increment, 100 randomly sampled subsets of peaks were used as the background for
894 determining the relative ranks of all peaks or debiased peaks. The relative ranks of the adjusted
895 Rand index (ARI), defined as the rank relative to the 100 randomly sampled subdatasets, for the
896 50 cases from 50% to 99%, were plotted.
897 **(g-j)** Comparisons of cell clustering accuracy before and after (orange) considering the peak
898 bias score (PBS) generated by existing methods on scATAC-seq data for human hematopoietic
899 cells. Each panel represents the result from a clustering approach: APEC **(g)**, Seurat **(h)**, scan
900 **(i)**, and snapATAC **(j)**. P-values were calculated by the Wilcoxon signed-rank test.

Figure 1

Figure 2

Figure 3

a Conventional approach

b SELMA approach

Figure 5