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

Shengen Shawn Hu[1], Lin Liu[2], Qi Li[3], Wenjing Ma[1,4], Michael J. Guertin[5], Clifford A. Meyer[6], Ke Deng[3], Tingting Zhang[7], Chongzhi Zang[1,8,9]*

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

[2]Institute 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

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

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

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

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

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

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

[9]UVA 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 (RepFootprint[8], Wellington[9], PIQ[10]) or ATAC-seq data (HINT-ATAC[11], ToBIAS[12]). 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 DNA[7].

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

65 sequencing data can potentially confound observed cleavage patterns and footprint detection.
66 The characterization and correction of intrinsic cleavage biases are essential to DNase/ATAC-
67 seq data analysis. To characterize intrinsic cleavage biases, most studies use a k-mer model, in
68 which the k-mer DNA sequence centered at a cleavage site of DNaseI/Tn5 is used as the
69 signature of this cleavage[14,15,17]. The sequence bias can be estimated by counting the
70 occurrences of cleavages with each k-mer in one dataset relative to the genome-wide
71 occurrences of this k-mer. A naïve k-mer model assumes that k-mers are independent of each
72 other, resulting in an exponential increase in the degree of freedom when k increases.
73 Therefore, a naïve k-mer model becomes less feasible in practice with a large k, especially with
74 samples having insufficient sequencing depth. Although most studies use 6-mers with $4^6 = 4096$
75 parameters[8,10,14,15,17,18], it is unclear whether a different model with a larger k-mer size and more
76 feasible parameter estimation can achieve better performance.
77
78 Several studies have used various computational approaches for intrinsic cleavage bias
79 estimation[8,15,16,19] and footprint detection with bias correction[15,18–20] for bulk DNase/ATAC-seq
80 data. Recently, single-cell ATAC-seq (scATAC-seq) has enabled chromatin accessibility
81 profiling in thousands of individual cells at one time[21–24]. Due to the high sparsity of single-cell
82 data and because most chromatin accessibility regions contain only one read in one cell, any
83 potential bias can be substantial in scATAC-seq data compared to bulk data, creating additional
84 challenges in computational analysis. Incorporating intrinsic cleavage bias effects for improved
85 scATAC-seq analysis also requires comprehensive assessment and development of innovative
86 methods.
87
88 Here, we present Simplex Encoded Linear Model for Accessible Chromatin (SELMA), a
89 computational framework for the accurate estimation of intrinsic cleavage biases and improved
90 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).

116 As an intrinsic property of the enzyme (DNaseI/Tn5 transposase), the cleavage biases are
117 expected to be invariant across cell types and independent of chromatin states (Fig. S1a).
118 Comparing data from two cell types using different 8-mer models as an example, we found that
119 the biases estimated using SELMA have a higher correlation than those estimated using the
120 naïve k-mer model, for both DNase-seq (Fig. 1d,e) and ATAC-seq (Fig. 1f,g). Using sequence
121 reads from genomic regions with different chromatin accessibility levels, the naïve k-mer model
122 was not able to generate very consistent bias estimations (Fig. S1b,c), but the k-mer biases
123 estimated by SELMA using the same data were highly consistent (Fig. S1b-e). The sequence
124 preferences of DNaseI or Tn5 cleavage can be better reflected when the enzymes are applied
125 to deproteinized naked genomic DNA[15,17]. We found that the k-mer cleavage biases in naked
126 DNA DNase/ATAC-seq data estimated with the naïve k-mer model can still be improved by
127 SELMA, obtaining more consistent bias scores between different cell systems (Fig. 1h-k). These
128 data demonstrated that SELMA can improve the accuracy of estimating k-mer cleavage biases
129 in DNase-seq and ATAC-seq data.
130
131 With fewer parameters, SELMA enabled us to assess the effect of k-mer size on the
132 performance of bias estimation. Using a “bias expected cleavage” approach[8,14,15,19], we
133 compared the bias estimation performances measured as the correlation coefficient between
134 the genome-wide observed cleavages and bias expected cleavages estimated using SELMA
135 with different k. A higher correlation coefficient indicates a more accurate bias estimation to
136 calculate the expected cleavages. For both DNase-seq and ATAC-seq data from two different
137 cell lines, we found that 10-mer outperforms any other k-mer (Fig. 1l-o). We then applied this
138 analysis to more DNase-seq and ATAC-seq data from a variety of human tissues generated by
139 ENCODE and found that 10-mer always exhibited the best performance (Fig. S2). The above
140 empirical evidence suggested that 10-mer is the optimal choice for intrinsic cleavage bias
141 estimation for both DNaseI (DNase-seq) and Tn5 (ATAC-seq) cleavages.

142

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

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

160

161 Using the similar "observed-expected correlation" approach described above, we found that
162 SELMA considering dimeric cleavages outperformed models considering only 5' cleavage in
163 generating a more accurate bias expected cleavage pattern for ATAC-seq data (Fig. 2d,e). We
164 confirmed that this observation was specific to ATAC-seq, as similar analyses for DNase-seq
165 from the same cell lines did not show a similar level of improvement (Fig. S3c,d). We also
166 compared SELMA with several existing bias estimation approaches[18] and found that SELMA's
167 performance was the best for ATAC-seq data from several cell lines as well as different human

168 tissue types from ENCODE (Fig. 2d,e, Fig. S3e-j). Several new Tn5-based techniques have
169 recently been developed for improved chromatin accessibility profiling, including THS-seq[28],
170 Fast-ATAC-seq[29], and Omni-ATAC-seq[30]. Data from these technologies showed intrinsic
171 cleavage biases similar to those of conventional ATAC-seq (Fig. S3k-n). We found that SELMA
172 also outperformed other approaches in bias estimation (Fig. 2f-i). Collectively, these data
173 suggested that SELMA can most accurately estimate intrinsic cleavage biases in data from
174 ATAC-seq and other Tn5-based techniques.
175
176 **SELMA enables accurate bias estimation by utilizing mitochondrial DNA (mtDNA) reads**
177 Accurate estimation of enzymatic cleavage biases independent of chromatin usually requires a
178 control sample of deproteinized naked DNA digested by the same enzyme. Biases estimated
179 from the naked DNA control data can be used to correct the chromatin accessibility profiling
180 data[14–18] (Fig. 3a). In conventional DNase/ATAC-seq data analyses, sequence reads mapped to
181 mitochondrial DNA (mtDNA) are usually discarded[4]. Lacking histones responsible for the
182 chromatin structure, mtDNA is similar to deproteinized naked DNA[31–33]. Therefore, we sought to
183 use the mtDNA reads from DNase/ATAC-seq data for cleavage bias estimation as an
184 alternative to using a naked DNA control sample (Fig. 3b). Likely because of relatively low read
185 counts and lack of sequence diversity (e.g., human mtDNA is only < 20kb long), the naïve k-mer
186 model exhibited poor performance on bias estimation from mtDNA reads, using bias scores
187 estimated from naked DNA as a reference (Fig. 3c). In contrast, SELMA generated a more
188 accurate bias estimation from the same mtDNA reads, which was highly consistent with the bias
189 scores estimated from the naked DNA data (Fig. 3d), demonstrating the power of SELMA to use
190 less input to make accurate bias estimations. We applied this approach to a series of ATAC-seq
191 and DNase-seq datasets for different human tissues from ENCODE and found that SELMA was
192 better than the naïve k-mer model in yielding a consistently higher correlation coefficient (>0.9)
193 between mtDNA-estimated bias and naked DNA-estimated bias for every ATAC-seq and

194 DNase-seq sample tested (Fig. S4a,b). Many optimized ATAC-seq protocols aimed to reduce
195 mtDNA reads to increase the fraction of chromatin DNA reads for chromatin accessibility signal
196 yield[29,30]. We sampled down mtDNA reads to test the performance of SELMA in making robust
197 bias estimations and found that SELMA could accurately estimate the bias with as few as
198 50,000 mtDNA reads (Fig. 3e, Fig. S4c), which is approximately 0.2% of the sequencing depth
199 of a regular ATAC-seq sample and lower than the fraction of mtDNA reads in any existing
200 ATAC-seq experiment[29,30]. These data demonstrated that by applying SELMA, mtDNA reads
201 can be utilized to substitute naked DNA control samples for accurate bias estimation.
202

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

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

220

221 To assess the interference of intrinsic biases with DNaseI cleavage patterns at TF binding
222 footprints, we aligned the 4.5 million consensus footprints with more than 10,000 human TF
223 ChIP-seq datasets from the Cistrome Data Browser database[34,35] and collected two sets of
224 footprint regions: "TF binding hotspots", the footprint regions overlapping with TF binding sites
225 detected from more than 3,000 ChIP-seq datasets, and "TF binding deserts", the footprint
226 regions that do not overlap with any TF binding sites from any ChIP-seq dataset or any human
227 TF motif sequence from the HOCOMOCO database[36] (Fig. S5e). We compared the aggregate
228 DNaseI cleavage patterns and the bias score patterns in these two sets of consensus footprint
229 regions and found that the DNaseI cleavage pattern in TF binding hotspot regions was
230 dominated by the classic DNaseI footprint pattern with little interference from the bias (Fig. 4e,f),
231 while in the TF binding desert regions, the entire cleavage "footprint" pattern was apparently
232 determined by the two spikes from the intrinsic bias (Fig. 4g,h). These results suggested that
233 although the overall DNaseI cleavage pattern is indicative of TF binding, the intrinsic cleavage
234 bias may interfere with the real footprint pattern, and the effect on those footprint regions with
235 rare TF binding events can be drastic. These observations were reproducible in DNase-seq
236 data from different cell and tissue types (Fig. S6a-b).

237

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

246 based approach to assess the potential power of SELMA-derived FBS in boosting the
247 performance of TF binding inference from DNase-seq signal patterns on consensus footprint
248 regions containing the TF motif sequence. For every TF with both the motif and ChIP-seq data
249 available, we employed a multinomial logistic regression model using different sets of features,
250 each of which may include DNase-seq read count, DNaseI footprint score produced from an
251 existing method, and an optional FBS, to infer the TF binding occupancy (determined by peak
252 occurrence in ChIP-seq data) in the motif-containing footprint regions and compared the
253 inference performances. We included our previous footprint method[14] and several
254 representative methods that outperform others, Wellington[9] and HINT[11,19] (with and without bias
255 correction mode) in this comparison. For each footprint method, we compared the TF binding
256 inference performance of the original method (read count + footprint score), the original method
257 plus a randomized naïve k-mer bias score as a control, and the original method plus the SELMA
258 bias score (FBS), using an integrated rank score strategy. For a total of 48 TFs in the human
259 cell line GM12878 and 45 TFs in K562, the models with SELMA FBS significantly outperformed
260 those without SELMA FBS, regardless of the approach used originally to calculate the footprint
261 score (Fig. 4j, Fig. S6). These results suggested that considering the intrinsic cleavage bias
262 could always improve the performance of existing footprint computational methods for TF
263 binding inference.
264
265 To assess whether the intrinsic bias has different levels of interference with the footprint regions
266 for different TFs, we selected two subsets of sequence motif-containing footprint regions for
267 each TF as the top 10% with the highest FBS or the bottom 10% with the lowest FBS, and
268 compared the frequencies of observing actual TF binding (overlapping with a ChIP-seq peak) in
269 these subsets of footprints. We found that different TFs had various preferences for binding at
270 low-FBS footprints or high-FBS footprints. For most TF samples in the two cell lines, the
271 footprint regions with a low FBS showed a significantly higher chance of observing actual TF

272 binding than those with a high FBS (Fig. 4k). For example, HLH family TFs such as MYC and
273 MAX[17] showed a higher fraction of TF binding occupancies in low-FBS footprint regions than at
274 high-FBS regions. Zinc finger family TFs including CTCF, REST, and NRF1[6,17], showed a higher
275 chance of TF binding in high-FBS footprint regions than in low-FBS regions (Fig. S6d). This is
276 possibly because of the different DNA binding affinities and DNA residence times between
277 different TFs[15]. TFs with stronger DNA-binding affinity or longer residence time might be less
278 affected by intrinsic cleavage biases, while TFs with shorter residence time or weaker binding
279 affinity might be more severely affected by intrinsic biases. These results suggested that
280 intrinsic cleavage biases might affect different TFs at various levels in divergent directions, and
281 considering intrinsic bias should improve the footprint-based TF binding inference for most TFs.
282

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

284 Single-cell ATAC-seq (scATAC-seq) technologies enable the detection of accessible chromatin
285 regions at single-cell resolution in thousands of cells at a time[21–23]. Due to the scarcity of
286 cleavage events in an individual cell, most chromatin accessibility regions in a single cell have
287 only one aligned fragment, making the potential influence of intrinsic cleavage biases more
288 substantial in scATAC-seq data than in bulk ATAC-seq data. We collected three scATAC-seq
289 datasets generated using two different platforms, human hematopoietic cells[37], human cell
290 lines[21], and a mouse primitive gut tube sample, and found that the scATAC-seq data contained
291 similar intrinsic cleavage biases to bulk data with highly correlated bias scores estimated by
292 SELMA (Fig. S7a-c). We estimated the average cleavage bias for each individual cell (cell bias
293 score, CBS) and found that the distribution of CBS was different across cell cluster patterns,
294 batches, and annotated cell types (Fig. 5a-c, Fig. S7d-i). These observations indicated that,
295 regardless of experimental platforms and biological systems, the Tn5 intrinsic cleavage biases
296 can contribute to cellular heterogeneity observed from scATAC-seq data and can affect cell
297 clustering analysis.

298

299 We next assessed whether considering intrinsic cleavage biases can improve cell clustering
300 based on scATAC-seq data. We used the actual cell-type labels as the clustering ground truth
301 for the human hematopoietic cell sample and the mixed human cell line sample and used
302 scRNA-seq data-projected cell labels as a “pseudo” ground truth for the mouse primitive gut
303 tube sample. We used the adjusted Rand index[38] to quantify the accuracy of an unsupervised
304 clustering result. As scATAC-seq-based cell clustering can be performed on all chromatin
305 accessibility regions (ATAC-seq peaks), we sought to address whether removing peaks with
306 high intrinsic biases can increase the clustering accuracy. We first tested a simple clustering
307 approach that involved dimensionality reduction using principal component analysis (PCA)
308 followed by K-means clustering. For all three scATAC-seq datasets, cell clustering using
309 SELMA-debiased peaks only (excluding 1%-50% of peaks with the highest PBS) is significantly
310 more accurate than clustering using all peaks under the background of randomly excluding the
311 same number of peaks (Fig. 5d-f). Then, we tested several representative scATAC-seq analysis
312 tools for cell clustering, including APEC[39], Seurat[40], scran[41], and snapATAC[42], and found that
313 using SELMA-debiased peaks can significantly improve the clustering results in most cases
314 (Fig. 5g-j, Fig. S7j-r). These results suggested that scATAC-seq data contain intrinsic cleavage
315 biases that can be reduced by excluding SELMA-identified high-bias peaks from the cell
316 clustering analysis.

317

318

319 **Discussion**

320

321 The existence of enzymatic cleavage biases in DNase-seq and ATAC-seq experiments has
322 been widely acknowledged in the field, but to what extent such intrinsic biases affect data
323 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

350 been demonstrated for TF motif characterization[43]. Following similar encoding strategies,
351 SELMA can potentially be applied to any k-mer-based model or to any high-throughput
352 sequencing data for robust sequence bias estimation and pattern recognition.
353
354 The k-mer biases estimated by SELMA are consistent across species and cell types, as
355 expected based on the assumption that the cleavage biases are intrinsic features of the
356 enzymatic assay. Therefore, one can directly use the SELMA-estimated k-mer biases from
357 DNaseI or Tn5-digested naked DNA data as universal intrinsic biases for DNase-seq and
358 ATAC-seq, respectively, and incorporate these precalculated bias scores into the data analysis.
359 However, this bias dataset is not guaranteed to remain accurate in many species that have not
360 been profiled, as there might be unknown biases that have not been captured. Although we are
361 confident that the SELMA-estimated results should still be largely consistent, one can always
362 use the SELMA package for *de novo* estimation of cleavage biases from one's own datasets.
363
364 When applying SELMA-estimated FBS to correct biases in DNase footprints, our data were
365 limited to ENCODE DNaseI consensus footprints as a proof-of-principle study. While this
366 ENCODE dataset is thus far the largest DNase footprint repertoire, users might be interested in
367 *de novo* detection of footprints from their customized DNase/ATAC-seq data. As many
368 bioinformatics tools are already available for such tasks using various computational models[20],
369 SELMA or SELMA-generated bias information can be incorporated into each of those models
370 for improved performance for more accurate regulatory DNA identification from footprints.
371 Similarly, while we show significant improvement of scATAC-seq cell clustering by removing
372 peaks with high PBS under the framework of most existing tools, performance still varies across
373 these tools due to the different statistical and computational models used in these methods.
374 Further optimization of existing methods or development of new methods is expected for the
375 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-esq 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.

402 For the mouse gut tube sample, scATAC-seq data from the 10x Genomics platform were
403 preprocessed with Cell Ranger ATAC with the default parameters to generate fragments for
404 each individual cell. The fragment file from the standard Cell Ranger ATAC output was used as
405 input for the subsequent analysis. Because the Cell Ranger ATAC pipeline shifted from the Tn5
406 cleavage sites to +4/-5 bp in generating the fragment file, the coordinates were shifted back to
407 represent the actual cleavage loci. For both datasets, cells with more than 10,000 reads were
408 retained for analysis.
409
410 **Processing of ChIP-seq data.** ChIP-seq data were collected and processed as described in a
411 previous study[45]. In brief, based on the quality measurements from the Cistrome data browser
412 (CistromeDB)[34,35], we included only the ChIP-seq datasets that pass at least four out of the first
413 five QC measures (i.e., sequence quality, mapping quality, library complexity, ChIP enrichment,
414 and signal-to-noise ratio). We also excluded the ChIP-seq datasets with fewer than 1000 peaks
415 identified by MACS2[46] with an FDR cutoff of 0.01. For TFs with multiple datasets available in the
416 same cell type, we kept only one dataset with the best quality based on passing most of the 6
417 CistromeDB QC measures, including sequence quality, mapping quality, library complexity,
418 ChIP enrichment, signal-to-noise ratio, and regulatory region enrichment. The ChIP-seq peak
419 files (detected by MACS2) for the selected samples were downloaded from CistromeDB for the
420 following analysis.
421
422 **TF motif analysis.** The motifs of human TFs were collected from the HOCOMOCO database
423 (v11) [36], and the genome-wide motif sites of TFs were detected by FIMO (v4.12.0) in the MEME
424 package[47]. Motif sites located outside of the genome-wide 36bp unique mappable regions were
425 excluded from the analysis. In total, 25,027,116 motif sites for 66 TF motifs from the human
426 genome were included in the analysis.
427

428 **SELMA model**

429 **Naïve k-mer bias model for intrinsic cleavage bias estimation.** The naïve k-mer biases were
430 calculated as described in a previous study[14]. In brief, a naïve k-mer bias was estimated as the
431 observed frequency of the cleavages with the k-mer relative to the frequency of that k-mer in the
432 background. For each mapped sequence read in a DNase-seq or ATAC-seq dataset, the
433 enzymatic cleavage site was between genomic positions i and i-1 for the plus (+) strand reads
434 and between i and i+1 for the minus (–) strand reads, where i represents the genomic position
435 of the 5' nucleotide of the reads. The associated k-mer sequence was thus assigned as
436 $\left[i-\frac{k}{2}, i+\frac{k}{2}-1\right]$ for the plus strand reads and $\left[i-\frac{k}{2}+1, i+\frac{k}{2}\right]$ for the minus strand reads. The
437 naïve k-mer bias score for k-mer *j* is defined as the number of all observed cleavages with that
438 k-mer divided by the occurrences of that k-mer in the background:

439 $$S_j = \frac{N_j}{M_j}$$

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

446

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

450

451 **SELMA model for intrinsic cleavage bias estimation.** In SELMA, a simplex encoding model
452 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).

$$
\begin{aligned}
A &= [\ 1 \quad -1 \quad -1\ ] \\
C &= [\ -1 \quad 1 \quad -1\ ] \\
G &= [\ -1 \quad -1 \quad 1\ ] \\
T &= [\ 1 \quad 1 \quad 1\ ]
\end{aligned}
$$

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:

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

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

467 and dinucleotides were encoded as rows in a Hadamard matrix of order 4 and a Hadamard
468 matrix of order 16, respectively.
469
470 Considering both mononucleotides and dinucleotide interactions, a k-mer can then be encoded
471 as k mononucleotides and $k-1$ dinucleotides. Therefore, the dimensionality of a k-mer simplex
472 encoding is

473 $$p(k) = 1 + 3k + 9(k-1) = 12k - 8$$

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

476 $$y \sim \sum_{i=1}^{12k-8} a_i x_i \quad (1)$$

477 where each observation y is the naïve k-mer bias score; $x_i \in \{1, -1\}$, is the independent
478 variable in the simplex encoding vectors; and a_i is the parameter to be estimated. After linear
479 regression, the model-fitted value $\hat{y}$ was defined as the SELMA bias score for each k-mer.
480
481 **Genome-wide cross-correlation analysis.** For the analysis presented in Fig. 2b,c, reads from
482 plus (+) and minus (-) strands on chromatin accessible regions (peaks) were collected
483 separately to generate plus strand observed cleavage profile and minus strand observed
484 cleavage profile, respectively. The Pearson correlation coefficient between the plus strand
485 signal and the k-bp shifted minus strand signal is:

486 $$\rho_k = \frac{\sum_i (P_i - \bar{P})(M_{ik} - \overline{M_k})}{\sqrt{\sum_i (P_i - \bar{P})^2 \sum_i (M_{ik} - \overline{M_k})^2}} \quad (2)$$

487 where P_i is the log-scaled plus strand cleavage count at genomic position *i* (log2(cleavage+1)),
488 M_{ik} is the log-scaled minus strand cleavage count at genomic position *i* with a k-bp shift, $\underline{P}$ is
489 the mean of all the P_i, and $\overline{M_k}$ is the mean of M_{ik} for all *i*. All genomic positions on the genome-
490 wide DNase/ATAC-seq peaks were included in the analysis. *k* is chosen from 1 to 20 (x-axis in
491 Fig. 2b,c).

492

493 **Comparison of different bias estimation methods.** We use a_i^+ and a_i^- to denote the "5' only"
494 intrinsic sequence bias score at genomic position i on the plus strand and minus strand,
495 respectively. We use a'^+_i and a'^-_i to denote the bias score from other bias estimation methods in
496 this section. Different bias estimation methods used in Fig. 2 and its associated sections are
497 listed below:

498 - For the "5' only" method, a_i^+ was calculated based on the k-mer ratio associated with the
499 k-mer spanning positions $[i - k/2,\ i + k/2 - 1]$ on the plus strand, and a_i^- was
500 calculated in a similar way based on the nucleotides spanning genomic positions $[i -$
501 $k/2 + 1,\ i + k/2]$ on the minus strand (reverse complement of the DNA sequence on the
502 corresponding plus strand). This method was applied to both DNase-seq and ATAC-seq,
503 while the other methods were applied only to ATAC-seq as they were specifically
504 designed for ATAC-seq.
505 - For SELMA (Fig. 2a), the bias score was calculated as the geometric mean of the "5'
506 only" bias score at the given position and the "5' only" bias score at 9bp downstream of
507 the other strand, i.e., $a'^+_i = \sqrt{a_i^+ \times a_{i+9}^-}$, and $a'^-_i = \sqrt{a_i^- \times a_{i-9}^+}$
508 - For the model in Martins et al[16], the bias score at genomic position i was calculated in a
509 similar way to the "5' only" method but using a gapped 11-mer model. The model could
510 be represented as $XXXXXXNNNXNXCXXNNXNNNNXNX$, in which position i was
511 represented by C; positions that were ignored were represented by X and informative
512 positions were represented by N.
513 - For the model in Baek et al[18], the bias score at genomic position i was calculated in a
514 similar way to the "5' only" method but the cleavages were shifted +4/-5 bp for +/- strand
515 cleavages. In practice, following the description in the "bagfoot" package, the bias score
516 at position i was calculated as $a'^+_i = a_{i+5}^+$, and $a'^-_i = a_{i-5}^+$.

517 - For the model in Calviello et al[8], the bias score at genomic position i was calculated in a
518 similar way to the "5' only" method but using the 6-mer bias table provided in the study.
519

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

527 $$\widehat{N_i^s} = N_i^s y_i^s \quad (3)$$

528 $$y_i^s = \frac{2^{a_i^s}}{\sum_{j=i-25}^{i+24} 2^{a_j^s}} \quad (4)$$

529 We used the Pearson correlation coefficient to compare the observed cleavage profile and the
530 bias-predicted cleavage profile (Fig. 2). We only considered positions within the accessible
531 regions to ensure that all positions had sufficient reads in the 50-bp window for accurate
532 estimation.
533

534 **DNaseI footprint analysis**

535 **Data processing.** The genome-wide DNaseI consensus footprint regions were downloaded
536 from Reference[7]
537 (https://resources.altius.org/~jvierstra/projects/footprinting.2020/consensus.index/). The
538 observed DNaseI cleavage profile from a DNase-seq dataset and DNaseI SELMA bias scores
539 across +/- 50bp centered on the footprint region were plotted as heatmaps (Fig. 4a,b). The
540 footprint regions were ordered by the footprint lengths, and each 1000 footprint regions with

541 similar lengths were compressed as one row in the heatmap for better visualization. The +/-
542 strand signals were plotted separately. We aligned the footprint regions based on the two
543 observed bias spikes in each footprint region (located 7bp to the right of the left boundary and
544 7bp to the left of the right boundary of the footprint, labeled as -0 and +0 in Fig. 4c-h). The
545 center regions between the bias spikes were scaled to 4 bins to align footprint regions with
546 different lengths.
547
548 **Footprint bias score.** The footprint bias score (FBS, in Fig. 4i-k) was defined as the difference
549 between "spike bias" and "center bias". The "spike bias" was calculated as the average of the
550 two SELMA bias scores at the spike positions, while the "center bias" was calculated as the
551 median SELMA bias score at the rest of the positions in the footprint region. Let FBS_j, b_j, and c_j
552 denote the footprint bias score, spike bias and center bias of footprint j, respectively. The FBS,
553 spike bias, and center bias are given by:

554 $$FBS_j = b_j - c_j \tag{5}$$

555 $$b_j = \overline{a_{i \in B}^{s \in \{+,-\}}} \tag{6}$$

556 $$c_j = \widetilde{a_{i \in C}^{s \in \{+,-\}}} \tag{7}$$

557 where a_i^s represents the SELMA bias score at genomic position i on the strand $s \in \{+,-\}$,
558 overbar represents the average, tilde represents the median, and B and C represent spike
559 positions (within the flanking 1bp of the bias spikes) and the remaining positions of the footprint
560 j, respectively. To calculate the randomly shuffled FBS (random k-mer bias), the FBS was
561 calculated in the same way, but the bias score a_i^s for each position was randomly selected from
562 the SELMA bias score table.
563
564 **Calculation of footprint score with different methods.** The following methods were used to
565 calculate footprint scores for comparison:

566 - Raw footprint: The raw footprint score was calculated following a previous study[14], using
567 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
568 cleavage count in the motif region, and the flanking regions to the right and left of the
569 motif, respectively. The flanks are both the same length as the motif.
570 - Wellington footprint: We used Wellington[9] (v0.2.0) with default parameters to detect
571 genome-wide footprints and selected significant footprint regions with p-values < 1e-10.
572 The output footprint score was assigned to the TF motif overlapping with the footprint
573 region.
574 - HINT footprint: We used HINT[11,19] (v0.12.3) with default parameters to detect genome-
575 wide footprints. For bias correction mode, we used an additional parameter: --bias-
576 correction. The footprint score of each footprint region was assigned to the overlapping
577 TF motifs.
578
579 **Inference of TF binding with different features.** For each TF in Fig. 4j, the TF motifs
580 overlapping with consensus footprint regions were collected as target regions. DNase-seq read
581 count in the footprint region ("reads"), footprint score from an existing method, and bias score
582 were used as features in a multinomial logistic regression model to infer TF occupancy at
583 footprint-overlapping motif regions. For each available method, model training with cross-
584 validation and predictions were performed using different combination of features: "original
585 method" refers to reads + footprint score. An additional feature of either SELMA FBS or a
586 random k-mer bias was added for different models. We used a performance measure
587 integration approach[11] to evaluate different models' prediction power. For each model, we
588 calculated the area under the ROC curve (AUROC) at 100%, 10%, and 1% false positive rate
589 (FPR). We also calculated the area under the precision-recall curve (AUPRC) at 100%, 10%,

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

591 as

592 $$S = \frac{1}{6}\sum_i -\log\frac{r_i}{N+1}$$

593 where $i = 1, \ldots, 6$ denotes the 6 performance measures, r_i is the rank of a model for

594 performance measure i, and N is the total number of models.

595

596 To calculate the random k-mer bias, we randomly permuted the SELMA k-mer bias table and

597 generated the k-mer bias table for a "simulated enzyme". We used this "simulated" bias table to

598 calculate the FBS for each footprint region and performed TF inference modeling similar to what

599 we did for the DNaseI SELMA FBS. This permutation was performed 100 times to generate 100

600 performance rank scores for random k-mer bias used as controls.

601

602 **Motif prediction power of TF binding sites.** In Fig. 4k, we collected the genome-wide motif

603 sites overlapping with consensus footprint regions and the ChIP-seq peaks for each TF with

604 HOCOMOCO[36] motif and ChIP-seq data available in GM12878 or K562 (Table S1). As a result,

605 we collected genome-wide motif sites for 65 TFs (48 in GM12878 and 45 in K562), and

606 35,721,822 motif sites in total. For each TF, we selected two subgroups of its motif sites based

607 on the FBS of their overlapped footprint regions: the top 10% of motif sites with the highest FBS,

608 defined as "sites with high-bias footprint"; and the bottom 10% of motif sites with the lowest

609 FBS, defined as "sites with low-bias footprint". We calculated the proportion of motif sites

610 overlapping with TF ChIP-seq peaks for each of the two subgroups and plotted on a scatter plot

611 (Fig. 4k). To test whether low-bias footprints tend to have more TF binding than high-bias

612 footprints for most TFs, we conducted a t-test comparing the distribution of TF binding log

613 likelihood ratios of low-bias over high-bias footprints to the standard normal distribution, and the

614 test p-value is labeled in Fig. 4k.

615

616 **Single-cell ATAC-seq clustering analysis**

617 **Gold standard of single-cell clustering.** For single-cell ATAC-seq data in the human
618 hematopoietic cells and human cell line samples, the cell-type information for each individual
619 cell was used as the ground truth, or the gold standard. For single-cell ATAC-seq data in the
620 mouse gut tube sample, the cell-type information was assigned based on label transfer[40] from
621 the single-cell RNA-seq dataset in the same system[48] (GSE136689), as the "pseudo" ground
622 truth, or the silver standard. In detail, we integrated scRNA-seq and scATAC-seq data using the
623 ArchR package[49] (v1.0.1). Individual cells with a high RNA integration score (unconstrained
624 predicted score < 0.56) were collected as high-quality cells for the analysis. The cutoff of the
625 RNA integration score (0.56) was determined by maximizing the interclass variance in the RNA
626 integration score using Otsu's method[50]. Cell types represented by fewer than 10 cells were
627 discarded.

628

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

633

634 **Peak detection and peak bias score calculation.** We first combined all the single-cell ATAC-
635 seq reads in the dataset and performed MACS2[46] peak calling with a q-value cutoff of 0.1 to
636 include all the potential accessible regions in the human hematopoietic cell and human cell line
637 systems (100,456 peaks detected for human hematopoietic cells and 83,318 peaks for human
638 cell lines). For the 10X scATAC-seq data in mouse gut tube, the q-value cutoff of 0.01 was used
639 for peak calling (146,098 peaks detected). Potential accessible regions with fewer than 10
640 covered reads or more than 4,000 covered reads were discarded in the subsequent analysis. To

641 consider the effect of intrinsic cleavage bias in scATAC-seq data, we summarized the bias for
642 each potential accessible region by calculating the peak bias score (PBS), defined as the
643 median SELMA bias score of all reads (from all individual cells) located in the peak region.
644
645 **Cell clustering with scATAC-seq data.** In the scATAC-seq clustering analysis, as a naïve
646 method, we performed PCA dimensionality reduction on the accessible regions by individual
647 cells matrix of normalized read count and kept 60 PCs, followed by k-means clustering. The
648 number of clusters (k) in k-means clustering was determined as the actual number of cell types
649 in the dataset. To evaluate the performance of cell clustering, we compared the clustering
650 results with the gold standard using adjusted rand index (ARI)[38]. We further repeated the k-
651 means clustering with 100 different random seeds to ensure the stability of the clustering results
652 and the average ARI of the 100 runs was used to estimate the clustering accuracy. We also
653 applied several published methods, including APEC[39], Seurat[40], scran[41], and snapATAC[42] with
654 default parameters, on the same dataset. We selected the top 60 dimensions for those methods
655 at the dimensional reduction step (PCA, Seurat, scran and snapATAC) to keep more variance
656 potentially introduced by intrinsic biases. We used the ArchR package to practice the Seurat
657 and scran clustering methods. To explore the effect of intrinsic cleavage bias on scATAC-seq
658 analysis, we selected different percentages (from 50% to 99%, with a 1% increment) of peaks
659 with the lowest PBS (SELMA debiased peaks in Fig. 5) as input to perform cell clustering. We
660 also randomly selected the same percentage of peaks as a control to estimate the relative rank
661 of the clustering performance from using all peaks and debiased peaks. In detail, for each
662 percentage from 50% to 99%, we randomly sampled peaks 100 times and defined the relative
663 rank as the number of ARIs from random samples that were less than the ARI from the same
664 percentage of debiased peaks or all peaks. The relative ranks for different percentages from
665 50% to 99% are summarized and plotted as boxplots in Fig. 5.
666

667

668 **Code and data availability**

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

676

677 **Acknowledgements**

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

683

684 **Author contributions**

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

689

690 **Competing interests**

691 The authors declare that they have no competing interests.

692

693

694 **References**

695 1. McGhee, J. D., Wood, W. I., Dolan, M., Engel, J. D. & Felsenfeld, G. A 200 base pair region
696 at the 5′ end of the chicken adult β-globin gene is accessible to nuclease digestion. *Cell* **27**, 45–
697 55 (1981).
698 2. Keene, M. A., Corces, V., Lowenhaupt, K. & Elgin, S. C. DNase I hypersensitive sites in
699 Drosophila chromatin occur at the 5' ends of regions of transcription. *Proc National Acad Sci* **78**,
700 143–146 (1981).
701 3. Boyle, A. P. *et al.* High-Resolution Mapping and Characterization of Open Chromatin across
702 the Genome. *Cell* **132**, 311–322 (2008).
703 4. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of
704 native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding
705 proteins and nucleosome position. *Nature Methods* **10**, 12**13**–1218 (2013).
706 5. Hesselberth, J. R. et al. Global mapping of protein-DNA interactions in vivo by digital genomic
707 footprinting. Nature Methods **6**, 283–28**9** (2009).
708 6. Neph, S. et al. An expansive human regulatory lexicon encoded in transcription factor
709 footprints. Nature **489**, 83–90 (2012).
710 7. Vierstra, J. et al. Global reference mapping of human transcription factor footprints. Nature
711 **583**, 729–736 (2020).
712 8. Calviello, A. K., Hirsekorn, A., Wurmus, R., Yusuf, D. & Ohler, U. Reproducible inference of
713 transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias
714 modeling. *Genome Biol* **20**, 42 (2019).
715 9. Piper, J. et al. Wellington: a novel method for the accurate identification of digital genomic
716 footprints from DNase-seq data. *Nucleic Acids Research* **41**, e201–e201 (2013).
717 10. Sherwood, R. I. et al. Discovery of directional and nondirectional pioneer transcription

718 factors by modeling DNase profile magnitude and shape. *Nat Biotechnol* **32**, **17**1–178 (20**14**).
719 **11**. Li, Z. et al. Identification of transcription factor binding sites using ATAC-seq. Genome Biol
720 20, **45** (20**19**).
721 12. Bentsen, M. et al. ATAC-seq footprinting unravels kinetics of transcription factor binding
722 during zygotic genome activation. *Nat Commun* **11**, **42**67 (2020).
723 13. Meyer, C. A. & Liu, X. S. Identifying and mitigating bias in next-generation sequencing
724 methods for chromatin biology. *Nat Rev Genet* **15**, 709–7**21** (2014).
725 14. He, H. H. et al. Refined DNase-seq protocol and data analysis reveals intrinsic bias in
726 transcription factor footprint identification. *Nat Methods* **11**, 73–78 (2014).
727 15. Sung, M.-H., Guertin, M. J., Baek, S. & Hager, G. L. DNase Footprint Signatures Are
728 Dictated by Factor Dynamics and DNA Sequence. *Molecular Cell* **56**, 275–285 (2014).
729 16. Martins, A. L., Walavalkar, N. M., Anderson, W. D., Zang, C. & Guertin, M. J. Universal
730 correction of enzymatic sequence bias reveals molecular signatures of protein/DNA interactions.
731 Nucleic Acids Res **46**, gkx1053- (2017).
732 17. Yardımcı, G. G., Frank, C. L., Crawford, G. E. & Ohler, U. Explicit DNase sequence bias
733 modeling enables high-resolution transcription factor footprint detection. Nucleic Acids Research
734 42, 11865–11878 (2014).
735 18. Baek, S., Goldstein, I. & Hager, G. L. Bivariate Genomic Footprinting Detects Changes in
736 Transcription Factor Activity. *CellReports* **19**, 1710–1722 (2017).
737 19. Gusmao, E. G., Allhoff, M., Zenke, M. & Costa, I. G. Analysis of computational footprinting
738 methods for DNase sequencing experiments. Nature Methods 13, 303–309 (2016).
739 20. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide
740 to ATAC-seq data analysis. Genome Biol 21, 22 (2020).
741 21. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory
742 variation. Nature **523**, 486–490 (2015).
743 22. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by

744 combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
745 23. Lake, B. B. et al. Integrative single-cell analysis of transcriptional and epigenetic states in
746 the human adult brain. *Nature Biotechnology* **36**, 70–80 (2017).
747 24. Preissl, S. et al. Single-nucleus analysis of accessible chromatin in developing mouse
748 forebrain reveals cell-type-specific transcriptional regulation. *Nat Neurosci* **21**, 432–439 (2018).
749 25. Zhang, C.-T. & Zhang, R. Analysis of distribution of bases in the coding sequences by a
750 digrammatic technique. Nucleic Acids Res 19, 6313–6317 (1991).
751 26. Stormo, G. D. Maximally Efficient Modeling of DNA Sequence Motifs at All Levels of
752 Complexity. *Genetics* **187**, **12**19–1224 (2011).
753 27. Adey, A. et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-
754 density in vitro transposition. Genome Biol 11, R119 (2010).
755 28. Sos, B. C. et al. Characterization of chromatin accessibility with a transposome
756 hypersensitive sites sequencing (THS-seq) assay. Genome Biol 17, 20 (2016).
757 29. Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human
758 hematopoiesis and leukemia evolution. *Nature Genetics* **48**, 1193–1203 (2016).
759 30. Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables
760 interrogation of frozen tissues. Nature Methods 14, 959–962 (2017).
761 31. Kornberg, R. D. Chromatin structure: a repeating unit of histones and DNA. Science **184**,
762 868–871 (1974).
763 32. Caron, F., Jacq, C. & Rouvière-Yaniv, J. Characterization of a histone-like protein extracted
764 from yeast mitochondria. *Proceedings of the National Academy of Sciences* **76**, 4265–4269
765 (1979).
766 33. DeFrancesco, L. & Attardi, G. In situ photochemical crosslinking of HeLa cell mitochondrial
767 DNA by a psoralen derivative reveals a protected region near the origin of replication. Nucleic
768 Acids Research 9, 6017–6030 (1981).
769 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 accesson-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

806 **Figure legends**

807

808 **Figure 1.** SELMA framework for cleavage bias estimation.

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

810 (**b**) Simplex encoding model. The coordinates of the 4 tetrahedral vertices of the cube encode

811 the 4 nucleotides. Each dinucleotide is encoded as the outer product of the 2 mononucleotides.

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

813 (**c**) SELMA uses k-mer simplex encoding followed by a linear regression for k-mer cleavage

814 bias estimation.

815 (**d-k**) Comparison between the naïve k-mer model (**d,f,h,j**) and SELMA (**e,g,l,k**) on k-mer

816 cleavage bias scores estimated from DNase-seq data (**d,e,h,i**) and ATAC-seq data (**f,g,j,k**).

817 Each dot in a scatter plot represents an 8-mer, with its estimated bias score from different

818 datasets represented in the x- and y-axes. Chromatin DNA (**d-g**) and naked DNA (**h-k**) from

819 different human cell lines (**d-i**) or different species (**j,k**) were compared as labeled. R represents

820 the Pearson correlation coefficient.

821 (**l-o**) Intrinsic cleavage bias estimation accuracy measured by correlation between genome-wide

822 observed (OBS) and bias-expected (EXP) cleavages with different k-mers. A higher Pearson

823 correlation coefficient (R) indicates a better prediction of the observed cleavage profile using the

824 estimated biases.

825

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

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

828 (**b**,**c**) Cross-correlation between 5' cleavages and 3' cleavages across genome-wide accessible

829 chromatin regions in the human GM12878 cell line from ATAC-seq (**b**) and DNase-seq (**c**)

830 experiments. The x-axis represents the shift distance (in bp) between 5' and 3' cleavages.

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

832 Pearson correlation coefficients (R) between genome-wide observed (OBS) and bias-expected
833 (EXP) cleavages for human GM12878 (**d**) and K562 (**e**) cell lines, as well as several modified
834 Tn5-based assays (**f-i**). Different bars represent different estimation approaches: gray for
835 considering the 10-mer at the observed cut only (5' only); orange for SELMA; and blue for
836 several published approaches. Modified Tn5-based assays include (**f**) THS-seq with standard
837 Tn5; (**g**) THS-seq with mutated Tn5; (**h**) fast-ATAC-seq; and (**i**) omni-ATAC-seq.
838
839 **Figure 3.** SELMA bias estimation using mtDNA reads.
840 (**a**) Conventional approach to intrinsic cleavage bias estimation, in which a naked DNA sample
841 is required for bias estimation, and mtDNA reads from the chromatin DNA ATAC-seq
842 experiment are discarded.
843 (**b**) SELMA approach of intrinsic cleavage bias estimation using mtDNA reads. Chromatin reads
844 and mtDNA reads from the same ATAC-seq dataset were separated for Tn5 cleavage profiling
845 and bias estimation with SELMA, respectively.
846 (**c-d**) Scatter plots demonstrating the robustness of 10-mer bias scores estimated from mtDNA
847 (mtDNA reads from chromatin ATAC-seq data, x-axis) and genomic DNA (reads from naked
848 DNA ATAC-seq data, y-axis), comparing the naïve k-mer model (**c**) and SELMA (**d**). SELMA
849 shows a more robust estimation using mtDNA reads with higher correlation than the naïve k-
850 mer model.
851 (**e**) Comparison of the correlation coefficients between biases estimated from mtDNA reads and
852 from naked DNA with different mtDNA read counts.
853
854 **Figure 4.** TF binding inference on DNaseI footprint regions is affected by intrinsic cleavage
855 biases and improved by SELMA.
856 (**a,b**) Heatmaps of DNaseI cleavage patterns in the GM12878 cell line (**a**) and SELMA-
857 estimated bias scores (**b**) around ENCODE DNaseI consensus footprint regions. 5' and 3'

858 patterns were plotted separately. Footprint regions were consistently ranked by footprint length,
859 and signals for every 1000 regions with similar lengths were averaged as one row in the
860 heatmap.
861 (**c-h**) Aggregate plots of DNaseI cleavage patterns (**c,e,g**) and SELMA-estimated bias scores
862 (**d,f,h**) at all ENCODE DNaseI consensus footprint regions (**c,d**), footprint regions overlapping
863 with TF binding hotspots (**e,f**), and footprint regions overlapping with TF binding deserts (**g,h**).
864 DNaseI cleavages are from GM12878 for (**c** & **e**) and merged from GM12878, K562 and
865 ENCODE tissues (**g**). Dashed lines represent estimated bias spikes in the footprint regions.
866 (**i**) SELMA-estimated footprint bias scores (FBS) for different groups of footprint regions. P-
867 values were calculated by the Wilcoxon signed-rank test. The centerline, bounds of box, top line
868 and bottom line represent the median, 25th to 75th percentile range, 25th percentile -
869 1.5*interquartile range (IQR) and 75th percentile + 1.5*IQR, respectively.
870 (**j**) Performance rank scores for transcription factor binding prediction from DNaseI footprint
871 using various methods for the human GM12878 cell line. Boxplots with different colors represent
872 different approaches for prediction as indicated in the legends. Different sets of boxplots
873 represent footprint scores calculated by different published methods as labeled. P-values were
874 calculated by the Wilcoxon signed-rank test.
875 (**k**) Scatter plot showing the heterogeneous effect of intrinsic cleavage biases on different TF
876 motifs. Each data point represents motif sites for one TF in one cell type. Fractions of the TF-
877 bound motif sites overlapping with high-FBS footprint regions and with low-FBS footprint regions
878 are plotted on the x- and y-axes, respectively. More motifs located above the diagonal line
879 indicate that more TFs are more likely to bind at low-FBS sites than at high-FBS sites. The p-
880 value was calculated by t-test comparing the distribution of log likelihood ratios (y-axis/x-axis) to
881 the standard normal distribution.
882
883 **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**), scran
900 (**i**), and snapATAC (**j**). P-values were calculated by the Wilcoxon signed-rank test.

Figure 1

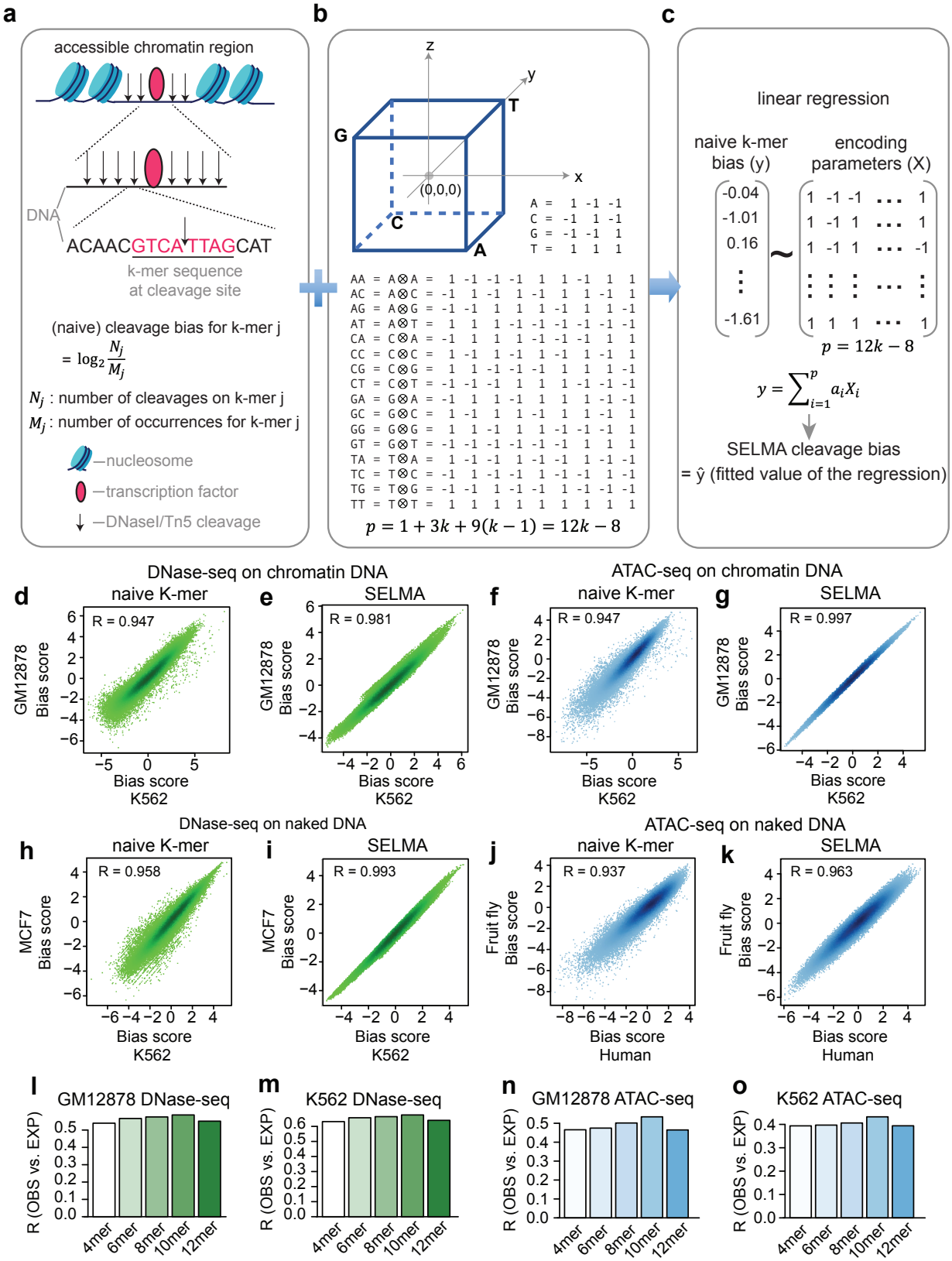

a
accessible chromatin region
DNA
ACAACGTCATTAGCAT
k-mer sequence at cleavage site
(naive) cleavage bias for k-mer j
$= \log_2 \frac{N_j}{M_j}$
N_j : number of cleavages on k-mer j
M_j : number of occurrences for k-mer j
—nucleosome
—transcription factor
—DNaseI/Tn5 cleavage
b
z
y
x
T
G
C
A
(0,0,0)
A = 1 -1 -1
C = -1 1 -1
G = -1 -1 1
T = 1 1 1
AA = A⊗A = 1 -1 -1 -1 1 1 -1 1 1
AC = A⊗C = -1 1 -1 1 -1 1 1 -1 1
AG = A⊗G = -1 -1 1 1 1 -1 1 1 -1
AT = A⊗T = 1 1 1 -1 -1 -1 -1 -1 -1
CA = C⊗A = -1 1 1 1 -1 -1 -1 1 1
CC = C⊗C = 1 -1 1 -1 1 -1 1 -1 1
CG = C⊗G = 1 1 -1 -1 -1 1 1 1 -1
CT = C⊗T = -1 -1 -1 1 1 1 -1 -1 -1
GA = G⊗A = -1 1 1 -1 1 1 1 -1 -1
GC = G⊗C = 1 -1 1 1 -1 1 1 -1 1
GG = G⊗G = 1 1 -1 1 1 -1 -1 -1 1
GT = G⊗T = -1 -1 -1 -1 -1 -1 1 1 1
TA = T⊗A = 1 -1 -1 1 -1 -1 1 -1 -1
TC = T⊗C = -1 1 -1 -1 1 -1 -1 1 -1
TG = T⊗G = -1 -1 1 -1 -1 1 -1 -1 -1
TT = T⊗T = 1 1 1 1 1 1 1 1 1
$p = 1 + 3k + 9(k-1) = 12k - 8$
c
linear regression
naive k-mer bias (y)
encoding parameters (X)
-0.04
-1.01
0.16
-1.61
$p = 12k - 8$
$y = \sum_{i=1}^{p} a_i X_i$
SELMA cleavage bias
= ŷ (fitted value of the regression)
DNase-seq on chromatin DNA
ATAC-seq on chromatin DNA
d naive K-mer R = 0.947
e SELMA R = 0.981
f naive K-mer R = 0.947
g SELMA R = 0.997
GM12878 Bias score
Bias score K562
DNase-seq on naked DNA
ATAC-seq on naked DNA
h naive K-mer R = 0.958
i SELMA R = 0.993
j naive K-mer R = 0.937
k SELMA R = 0.963
MCF7 Bias score
Fruit fly Bias score
Bias score Human
l GM12878 DNase-seq
m K562 DNase-seq
n GM12878 ATAC-seq
o K562 ATAC-seq
R (OBS vs. EXP)
4mer
6mer
8mer
10mer
12mer

Figure 2

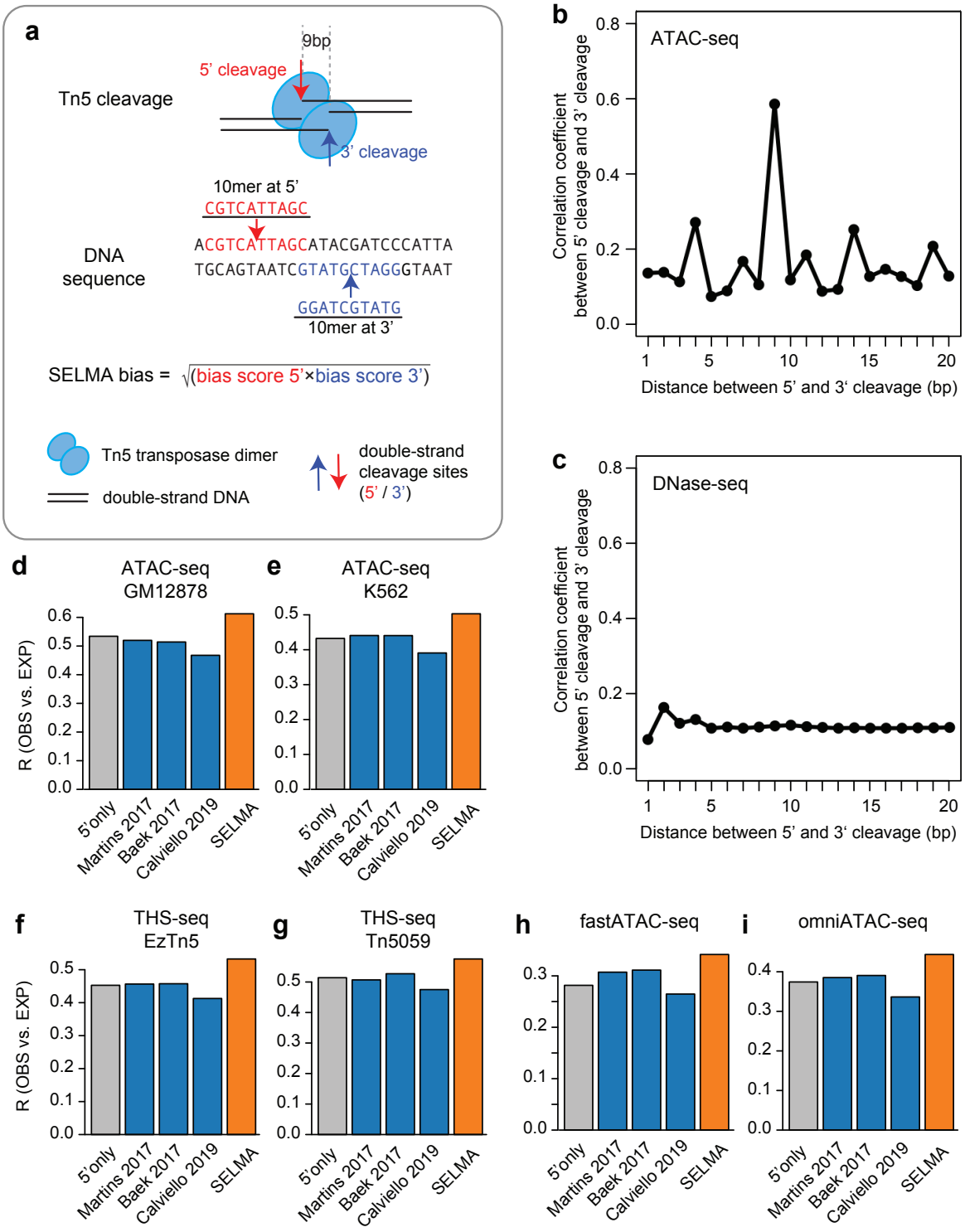

a
Tn5 cleavage
9bp
5' cleavage
3' cleavage
10mer at 5'
CGTCATTAGC
DNA sequence
ACGTCATTAGCATACGATCCCATTA
TGCAGTAATCGTATGCTAGGGTAAT
GGATCGTATG
10mer at 3'
SELMA bias = √(bias score 5'×bias score 3')
Tn5 transposase dimer
double-strand DNA
double-strand cleavage sites (5' / 3')
b
ATAC-seq
Correlation coefficient between 5' cleavage and 3' cleavage
Distance between 5' and 3' cleavage (bp)
c
DNase-seq
Correlation coefficient between 5' cleavage and 3' cleavage
Distance between 5' and 3' cleavage (bp)
d
ATAC-seq GM12878
e
ATAC-seq K562
f
THS-seq EzTn5
g
THS-seq Tn5059
h
fastATAC-seq
i
omniATAC-seq
R (OBS vs. EXP)
5'only
Martins 2017
Baek 2017
Calviello 2019
SELMA

Figure 3

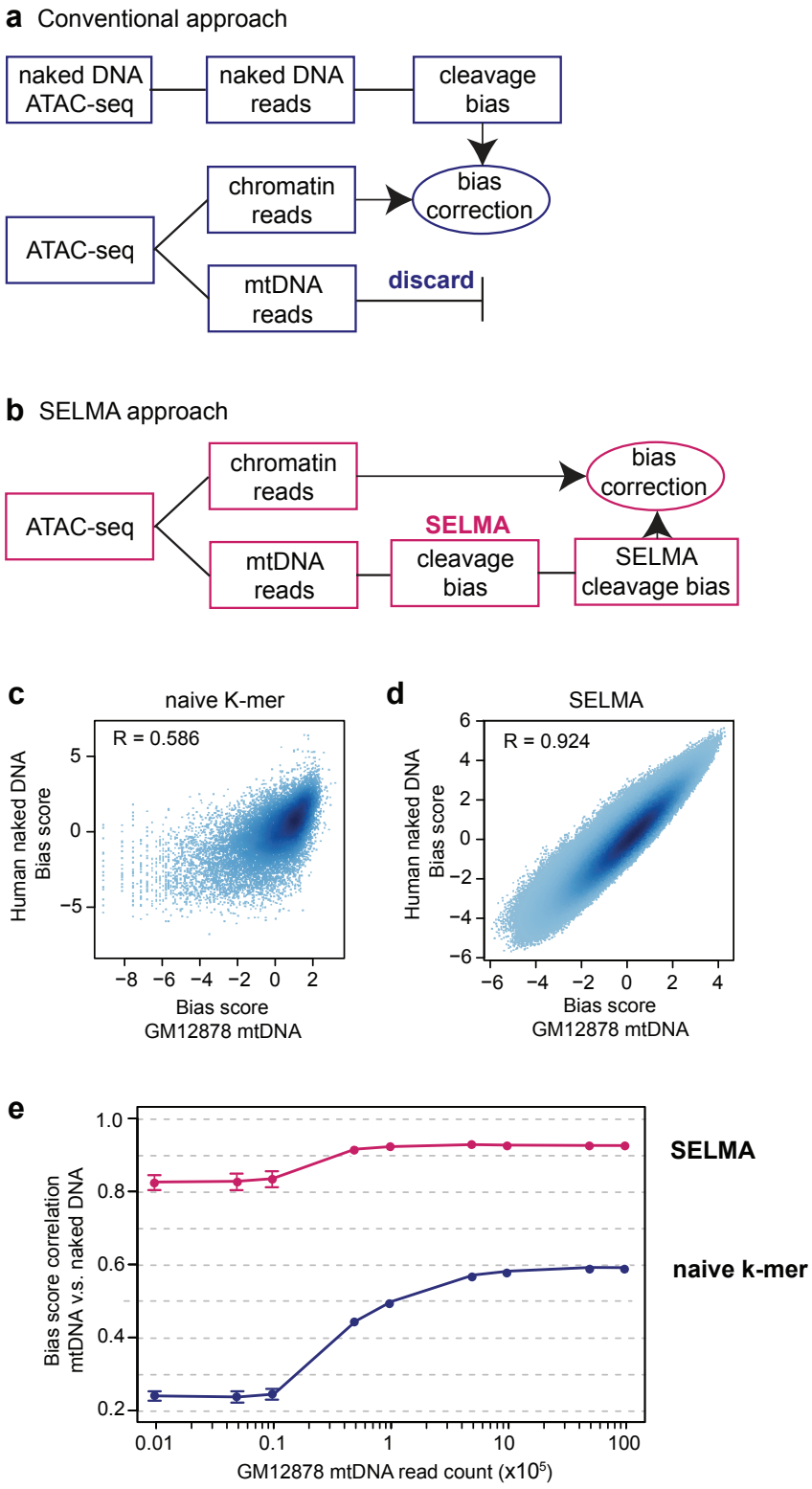

a Conventional approach
naked DNA ATAC-seq
naked DNA reads
cleavage bias
chromatin reads
bias correction
ATAC-seq
mtDNA reads
discard
b SELMA approach
chromatin reads
bias correction
ATAC-seq
SELMA
mtDNA reads
cleavage bias
SELMA cleavage bias
c
naive K-mer
R = 0.586
Human naked DNA Bias score
Bias score GM12878 mtDNA
d
SELMA
R = 0.924
Human naked DNA Bias score
Bias score GM12878 mtDNA
e
Bias score correlation mtDNA v.s. naked DNA
SELMA
naive k-mer
GM12878 mtDNA read count (x10^5)

Figure 4

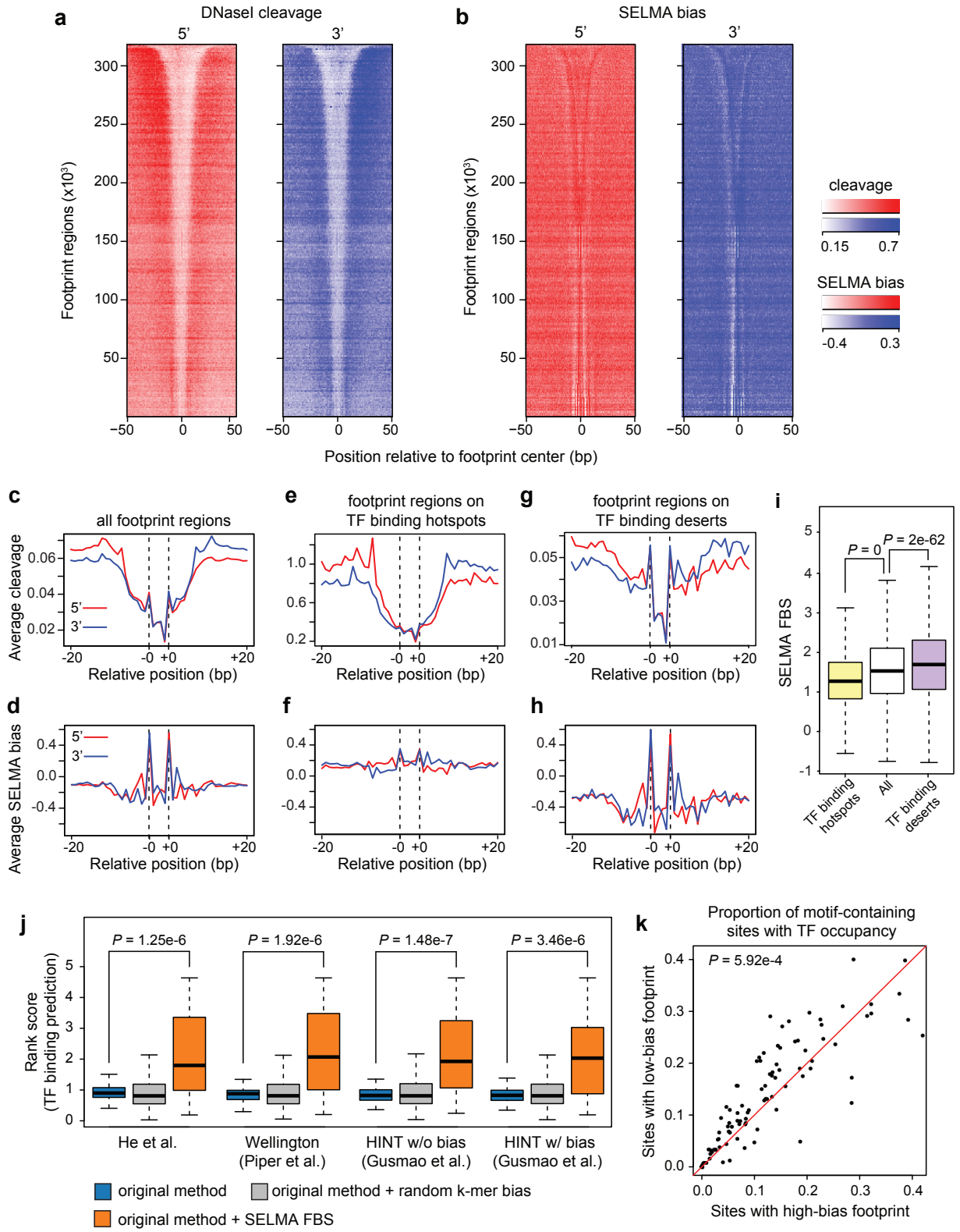
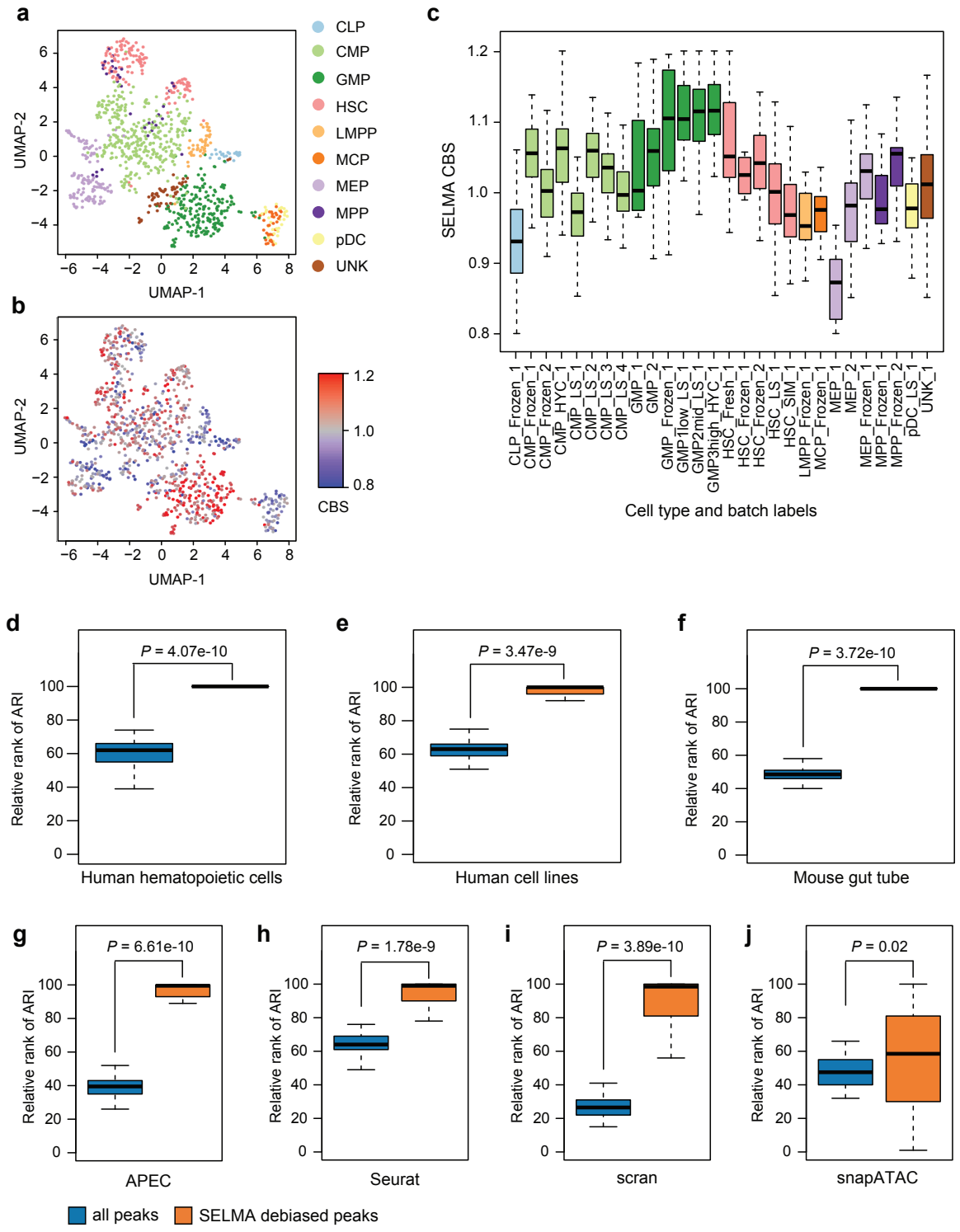

a
DNaseI cleavage
5'
3'
Footprint regions (x10³)
b
SELMA bias
cleavage
0.15
0.7
SELMA bias
-0.4
0.3
Position relative to footprint center (bp)
c
all footprint regions
Average cleavage
Relative position (bp)
e
footprint regions on TF binding hotspots
g
footprint regions on TF binding deserts
d
Average SELMA bias
f
h
i
SELMA FBS
P = 0
P = 2e-62
TF binding hotspots
All
TF binding deserts
j
Rank score (TF binding prediction)
P = 1.25e-6
P = 1.92e-6
P = 1.48e-7
P = 3.46e-6
He et al.
Wellington (Piper et al.)
HINT w/o bias (Gusmao et al.)
HINT w/ bias (Gusmao et al.)
original method
original method + random k-mer bias
original method + SELMA FBS
k
Proportion of motif-containing sites with TF occupancy
P = 5.92e-4
Sites with low-bias footprint
Sites with high-bias footprint

Figure 5


a
UMAP-2
UMAP-1
CLP
CMP
GMP
HSC
LMPP
MCP
MEP
MPP
pDC
UNK
b
CBS
c
SELMA CBS
Cell type and batch labels
d
P = 4.07e-10
Relative rank of ARI
Human hematopoietic cells
e
P = 3.47e-9
Human cell lines
f
P = 3.72e-10
Mouse gut tube
g
P = 6.61e-10
APEC
h
P = 1.78e-9
Seurat
i
P = 3.89e-10
scran
j
P = 0.02
snapATAC
all peaks
SELMA debiased peaks