Human transcription factor combinations mapped by genome-wide footprinting with deaminase

Runsheng He[1,2,*], Wenyang Dong[1,2,4,*], Wenping Ma[2,*], Zhi Wang[1,2,4,*], Long Gao[2,*], Chen Xie[1,2,*], Dubai Li[2,3], Ke Shen[1,2,4], Fanchong Jian[1,2,5], Jiankun Zhang[1,2,4], Yuan Yuan[1,2,5], Xinyao Wang[1,3], Yuxuan Pang[1,3], Zhen Zhang[2], Yinghui Zheng[1], Shuang Liu[2], Cheng Luo[3], Xiaoran Chai[1], Jun Ren[2], Zhanxing Zhu[2], Xiaoliang Sunney Xie[1,2,#]

1. Beijing Advanced Innovation Center for Genomics (ICG) & Biomedical Pioneering Innovation Center (BIOPIC), Peking University, Beijing 100871, China
2. Changping Laboratory, Beijing, 102206, China
3. Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, China
4. School of Life Sciences, Peking University, Beijing, 100871, China
5. College of Chemistry and Molecular Engineering, Peking University, Beijing, 100871, China
* Authors contributed equally.
# Correspondence: sunneyxie@biopic.pku.edu.cn

Abstract

An individual's somatic cells have the same genome but exhibit cell-type-specific gene expression regulated by a combination of transcription factors (TFs) for each gene. Mapping of TF sites on the human genome is critically important for understanding functional genomics. Here we report a novel technique to map TFs' binding sites along the human genome with single-base resolution by footprinting with deaminase (FOODIE). Single-molecule amplicons from the *in situ* deamination yielded TF binding fractions and the cooperativity between adjacent TFs. Interestingly, we found that, in a human lymphoblastoid cell line, genes in a correlated gene module (CGM) for a particular biological function share TF(s) in their promoters and enhancers. Finally, we show that single-cell FOODIE allows cell-type specific TF footprinting in heterogeneous mouse brain tissue.

Introduction

Every somatic cell in the human body has essentially the same genome but carries out distinct functions in specific tissues with different gene expression patterns. Transcription factors (TFs) bind to promoters and enhancers of genes on the genome, to control gene expression and regulation, cell differentiation, and development, in response to environmental signals through regulatory networks (*1*). Our knowledge about gene expression and regulation first came from the study of bacteria, where one TF, acts as a key to turn on and off an operon, containing multiple protein-coding genes that carry out a particular biological function (*2, 3*). Such TF activities have been directly monitored and quantitatively described on a single-cell and single-molecule basis (*4, 5*).

In contrast, TFs in eukaryotes, ever since their first isolation (*6*), seem to operate differently in that each binds to many sites on the genome with lower binding affinities (*7*), and gene regulation is achieved through the binding of a combination of several TFs at promoters and enhancers (*8*). The fact that several TFs, acting together as a set of keys, control one protein-coding gene makes the total number of ~1600 TFs sufficient to control ~22000 genes in human cells. More importantly, we hypothesize, that this provides a way to coordinate multiple genes in a gene module to accomplish a particular biological function, such as protein synthesis or glucose metabolism (*9, 10*). However, due to the lack of experimental techniques, little is known about how TF combinations are organized to coordinate regulatory networks, which remains the central question in eukaryotic biology.

Most TF binding sites in eukaryotes have been identified by chromatin immunoprecipitation ChIP-on-chip (*11*) and ChIP-seq (*12, 13*), each analyzing one TF at a time, through antibody pulling down followed by DNA sequencing, with a low spatial resolution of a few hundred base pairs. A common method for determining cellular binding motifs is footprinting, i.e. detecting DNA sequences that are protected from enzymatic treatment, most notably by DNase-seq, which detects the endpoints of DNA fragments cleaved by DNase I (*14*). However, both ChIP-seq and DNase-seq require bulk samples with millions of cells, usually of cell lines, and are incapable of detection on a single DNA.

To overcome these limitations, another footprinting method by cytosine methylation had been previously attempted with Single Molecule Footprinting (SMF) (*15*). However, it can not be applied to normal mammalian cells, as the extraneous methyltransferase used was mostly indistinguishable from endogenous methylation.

63 Here we report a double-stranded DNA cytosine deamination-based approach, which was
64 named FOOtprinting with DeamInasE (FOODIE). FOODIE can detect TFs' binding sequences
65 along the genome, as the TFs' steric hindrance prevents the conversion of cytosine to uracil.
66 FOODIE requires sequencing samples of only thousands of cells, significantly fewer than those
67 required for ChIP-seq or DNase-seq.
68
69 **Genome-wide mapping of TF footprints**
70 FOODIE relies on double-stranded DNA deaminase with high efficiency (Fig. 1A). We
71 initially tested DddA, the first double-stranded DNA deaminase reported (*16*), which only
72 deaminates cytosine preceded by thymine, not capable of footprinting all TFs bound to the genome
73 (Fig. S1A). To overcome this limitation, we characterized a new dsDNA cytosine deaminase from
74 a large group of reported toxins (*17*), which we named DddB (also known as BadTF3 (*18*)) after
75 DddA. Unlike DddA, DddB converts cytosine to uracil largely independent of adjacent bases (Fig.
76 1B), allowing us to footprint almost all TF motif sequences (Fig. S1A). We determined the crystal
77 structure of DddB in complex with its immunity protein at 2.5 Å resolution (Fig. 1C). Despite the
78 overall structural similarity between DddB and DddA, DddB features an intervening helix near the
79 active site, whereas DddA has a loop in this region (Fig. S1B).
80 In the FOODIE workflow (Fig. 1A), similar way to ATAC-seq (*19*), transposase Tn5 was
81 used in permeabilized cells to enrich the open chromatin regions, where most TFs bind and exert
82 their regulatory roles. DddB is then added after Tn5 tagmentation to detect protein binding on
83 DNA, whose underlying cytosines are protected from enzymatic deamination. By performing one
84 more step after ATAC, FOODIE is capable of probing the regulatory sites in open chromatin
85 regions in a cost-effective and high-throughput way.
86 The TF binding detection results by ChIP-seq, FOODIE, and DNase-seq are shown in Fig.
87 1D. ChIP-seq had a low spatial resolution (200-500bp), not capable of telling whether both or just
88 one of the detected CTCF sequences were occupied. In contrast, FOODIE accurately identified
89 CTCF that the footprint on the right is bound, which was consistent with DNase-seq footprinting.
90 We further compared the number of genome-wide binding sites of CTCF detected by
91 FOODIE with those obtained from ChIP-seq and DNase-seq. The CTCF binding sites detected by
92 FOODIE show strong concordance with those detected by ChIP-seq and largely overlap with those
93 detected by DNase-seq footprinting (Fig. 1E).

94 Next we employed a heat shock experiment to validate the reliability of FOODIE. As the
95 temperature increases, the transcription factor HSF is activated to bind to the promoters of target
96 genes to induce expression (*20*). At HSP90AA1 promoter, we observed a decrease in the
97 conversion ratio within HSF motif sequences, which corresponded to the gain of HSF binding
98 upon heat shock (Fig. 1F). Those data suggested that FOODIE can identify precise binding of TFs
99 with high accuracy.

100 To support the interactive analyses of genome-wide footprints identified by FOODIE, we
101 developed a web server with a user-friendly interface. This platform includes a genome browser
102 for visualization and a data portal with FOODIE data in this study. Fig. 2A shows a browser
103 display of the genomic region around the TRNAU1AP promoter in chromosome 1, a chromatin
104 open region between nucleosomes -1 (and -2) and +1 (and +2), with three distinct footprints. The
105 first row is the ATAC track seen by Tn5 tagmentation. The second row is the C to U conversion
106 efficiency showing the footprints. The raw reads of single-molecule sequences below, with the
107 converted cytosines on the Watson strand (red) and the Crick strand (green), show the binding
108 fractions along the genome.

109 We found that the deamination efficiency of DddB is dependent on the neighboring sequences
110 of cytosines (especially from -4 to +2) (Fig. S2), which needs to be normalized when calculating
111 the binding fraction. We measured deamination efficiency along the naked human genome as a
112 positive control for the conversion ratio correction.

113 We calculated the mean number of footprints present in cis-regulatory elements of genes and
114 their relative arrangement (see Material and Methods). The median number of footprints per gene
115 promoter was found to be 5 (median number), compared to 2 in enhancer regions (Fig. 2B).
116 Notably, the binding fractions of TFs in the promoter regions were higher than those of TFs in
117 enhancer regions (Fig. 2B).

118 FOODIE footprints need to be assigned to particular TFs, commonly through motifs tabulated
119 on websites, such as JASPAR (*21*) and HOCOMOCO (*22*), which collected vast amounts of
120 information from *in vitro* data of SELEX (*23*) and PBM's (*24*) as well as *in vivo* data of ChIP-seq
121 (*25*). It turned out that a footprint has many different possible assignments of TFs, which we can
122 not pin down because of the lack of binding affinities from databases and the intracellular
123 concentrations. While we are working on the methods to fully assign the footprints, in this work

124 we assigned footprints to certain TFs only when they have strong ChIP-seq signal in GM12878
125 cell line. This resulted in reliable assignments to only seventy-nine known TFs.
126 As such, the abundance of the assigned TFs is shown in promoter and enhancer regions, we
127 found promoters were particularly enriched for the NRF1, KLF/SP, ZNF143, ETS, and YY1,
128 footprints, which were rarely detected in enhancers (Fig. 2C). Taken together, TF bound to
129 promoters and enhancers differ in average number within open regions, binding fractions and TF
130 types.
131
132 **Single-molecule reads revealing positive and negative binding cooperativity between TFs**
133 While the binding fraction of a particular binding site can be determined by the proportion of
134 single DNA molecules with unconverted C to U, the fact that sequencing occurs on individual
135 molecules allows us to determine the cooperativity between TFs bound to two adjacent binding
136 sites.
137 The single-molecule reads at a CTCF binding site in the genome are shown in Fig. 2D,
138 exhibiting two separate footprints of the bridging structure of the zinc finger hinge (*26*), which are
139 either simultaneously occupied or not. In Fig. 2D, two adjacent TFs RFX and CREB have two
140 distinct footprints (*27*). To evaluate their cooperativity, we performed Fisher's exact test for the
141 significance of independence on a pair of cytosines, each selected from the two binding sites. The
142 φ correlation coefficient is between 0 and 1 for positive cooperativity and between -1 and 0 for
143 negative cooperativity. A red dot indicates significant positive cooperativity ($P<0.001$). The pairs
144 within RFX and CREB sites are positively cooperative because the binding of cytosines is
145 simultaneous. Interestingly, the pairs across different sites are also positively cooperative, i.e. the
146 binding of RFX and CREB are not independent but facilitate each other (Fig. 2D). We attribute
147 this mechanism to protein-protein interactions between two TFs, which has been previously
148 reported (*27*).
149 Conversely, negative cooperativity was seen between two NRF1 binding sites within the
150 ACTN4 promoter (Fig. 2D). While the φ of the two cytosines within each NRF1 binding site is
151 positive, the pairs across different sites are negative (blue: $P<0.001$), meaning their binding is
152 exclusive. Such single-molecule measurements provide valuable information about how TFs work
153 together in the genome.
154

155 **Shared TFs within the same correlated gene module**

156 Many protein-coding genes form a module to work together to accomplish a particular
157 biological function. People have been looking for gene modules with the bulk (*28*) and single cells
158 (*9, 10, 29*) by perturbation measurements. By improving the detectability of single-cell
159 transcriptome, we demonstrated the identification of correlated gene modules (CGMs) without
160 perturbation under the steady state condition (*9, 10*). The bursty nature of gene expression has
161 been long established and attributed to the single-molecule nature of DNA in single cells (*30-34*).
162 The coordination of a particular biological function by multiple genes requires their
163 synchronization. Experimentally, such synchronization can be probed by measuring the covariance
164 matrix of any pair of RNA in a single-cell transcriptome experiment (MALBAC-DT) (*9, 10*).

165 Recently we have developed a new method for more accurate identification of CGMs, Triangle
166 Screening of Hierarchical Accumulated-Edges (TriSHA) (*35*). Over Representation Analysis
167 (ORA) reveals that nearly 80% of the modules are functionally annotated with a False Discovery
168 Rate (FDR) below 0.05. Through module connection analyses, we have identified three distinct
169 types of CGMs within the GM12878 cell line based on functional annotation, associated with
170 specific cellular functions: housekeeping, tissue-specific, and cell cycle (Fig. 3A). Supplementary
171 Table S1 lists the CGMs with size larger than 10 that we identified. Gene t-SNE plot of GM12878
172 clearly shows the cluster of genes in the identified CGMs (Fig. 3B).

173 How are the genes in the same module distributed within the 3D genome? Are they clustered
174 together, or spatially dispersed? Our group determined the 3D genome structure of a somatic
175 human cell (*36*). To answer the above question, we employed our latest high-resolution single-cell
176 3D genome structures of GM12878 cells using scMicro-C (*37*). Taking the DNA replication CGM
177 (MD20, cell cycle clade) of 317 genes as an example, Fig. 3C shows that they are not clustered
178 but relatively randomly distributed within the nucleus, with the more quantitative distribution of
179 pair-wise distance shown in Fig. 3D. No spatial clustering of genes in other gene modules was
180 observed either. This observation does not support the idea that the correlated genes are transcribed
181 within certain locations in the nucleus. We hypothesize that the coordination of different genes in
182 the module is achieved through TF presented in the contiguous euchromatin space.

183 With FOODIE, we are in a position to determine how TFs organize the CGMs. We looked at
184 the enrichment of TFs shared among both promoters and enhancers of genes within a module,
185 which is characterized by the hypergeometric test. Out of sixty-five CGMs identified in GM12878

186 (Fig. 3A and Supplementary Table S1), we successfully identified that many TFs are statistically
187 enriched in different gene modules, including for example, E2F factors in the promoters of genes
188 in MD16 cell cycle-related CGM (*38*) and RELB in the enhancers of genes in MD47 regulation of
189 immune system CGM (*39*) (Fig. 4A).

190 Interestingly, we found enrichments of specific TFs are clearly different for the three types of
191 CGMs. For cell cycle and housekeeping CGMs, the TF enrichment is primarily in promoters. For
192 example, the NF-Y factor is enriched in the genes of MD14 cholesterol metabolism CGM (Fig.
193 4A). In contrast, for tissue-specific CGMs, the TF enrichment is primarily in enhancers. For
194 example, STAT3 enrichment is present in MD 48 immune response CGM. This suggests that there
195 are two distinct modes of CGM regulation: genes with essential cellular functions rely on TFs'
196 binding at promoters for robustness, while tissue-specific genes depend more on enhancers for
197 larger tunability.

198 We further observed certain TF pairs are shared among genes in both promoters and enhancers
199 within the same CGM (Fig. 4B) and evaluated the enrichment of TF using the hypergeometric test.
200 TF pairs are highly abundant in GCMs, and there are more TF pairs in enhancer regions (Fig. S2A).
201 Around one-third of the CGMs have enriched TF pairs, either at the promoters (Fig. 4B), or at the
202 enhancers (Fig. S2B). A higher fraction is expected when TF footprints are fully assigned. This
203 suggests the coordination of different genes in a module for a particular biology function through
204 TF combinations is ubiquitous and by no means random. Concrete examples of TF pairs shared
205 by multiple genes in a CGM are shown in Fig. 4C and 4D for promoters and enhancers.

206 As such, we are able to use FOODIE to probe how regulatory networks are orchestrated by a
207 combination of TFs. Although we do not yet have all the TF footprints assigned, even an
208 incomplete set (seventy-nine assigned TFs), already uncovered TFs pairs previously unknown in
209 order to coordinate gene modules.

210

211 **Single-cell FOODIE for cell typing in tissues**

212 Our goal is to characterize different human tissues with FOODIE and determine their genome-
213 wide TF footprints. However, human tissues are heterogeneous with different types of cells mixed,
214 so it is necessary to determine each cell's identity. We note that each gene only has one maternal
215 allele and one paternal allele in an individual cell, which is not sufficient to provide reliable

statistics for TF footprints. However, thousands of single cells identified for a particular cell type would be sufficient to carry out FOODIE.

Similar to cell typing by single-cell transcriptome or single-cell ATAC-seq, we use FOODIE signal along the genome to do unsupervised clustering of a population of cells. We evaluated the "non-conversion ratio", i. e. the ratio of unconverted cytosines to the total number of detected cytosines within the gene body ± 1000 bp regions for each gene. Using four well-characterized human cell lines (GM12878, K562, HEK293T, and HeLa) to form a mix of single-cell FOODIE data, different cell types were clearly identified (Fig. 5A). After merging single-cell data of a particular cell type (GM12878 for example) as a pure bulk, we observed precise TF footprints of CTCF and NRF1 at the specific genome locations (Fig. 5A).

We further conducted single-cell FOODIE for ~1500 cells from the mouse hippocampus. Four main cell types were identified, including oligodendrocytes, microglia, astrocytes, and neurons (both inhibitory and excitatory), which were annotated by matched marker genes (Fig. 5B). Using merged single-cell FOODIE data of neurons, we observed footprints of JUN and MEF2B, two TFs critical for neural cellular functions (Fig. 5B). Experiments like this are expected to provide new information of neuronal gene regulation in response to either physical or chemical stimulates in the brain.

Outlook

In conclusion, we have shown that FOODIE is a high-precision method for genome-wide *in situ* mapping of TF footprints inside the nucleus with single-base resolution. Unlike ChIP-seq, FOODIE is experimentally simple, high-throughput, and cost-effective. Single-cell FOODIE would allow for cell typing in a heterogeneous tissue sample, enabling TF mapping in any cell type within a complex tissue. A systematic TF footprinting database of different human tissues is underway. Combining with machine learning, we anticipate that such a database of TF combinations of different cell types and correlated gene modules will lead to the fundamental understanding of transcriptional regulation and cellular functions in eukaryotic organism biology, especially for human biology, which might hold the potential for understanding many human diseases.

Acknowledgments

We thank Prof. Ge Gao for the critical reading of the manuscript. We thank Qiuwan Liu for helping with the figures, and Wenzhong Zhang for helping with the web server. **Funding**: This project is financially supported by Changping Laboratory (2021C040201) and the Ministry of Science and Technology of China. The initiation work was made possible by the support from Beijing Advanced Innovation Center for Genomics at Peking University. W. D. was supported in part by the Postdoctoral Fellowship of Peking-Tsinghua Center for Life Sciences. X. W. was supported in part by Peking University Boya Postdoctoral Fellowship. **Author contributions**: X. S. X., R. H., W. D., Z. W., and W. M. designed the experiments. R. H., W. D., W. M., Z. W., Y. Y., Y. Z., S. L., C. L., and J. R. collected the experimental data. L. G., C. X., D. L., K. S., F. J., J. Z., X. W., Y. P., Z. Zhang, X. C. and Z. Zhu performed the data analyses. X. S. X. supervised the project and wrote the manuscript with the help of all authors. **Competing interests**: W. D., R. H., Z. W., and X. S. X. are coinventors on patent application WO/2024/065721 that includes discoveries described in this manuscript. The remaining authors declare no competing interests.

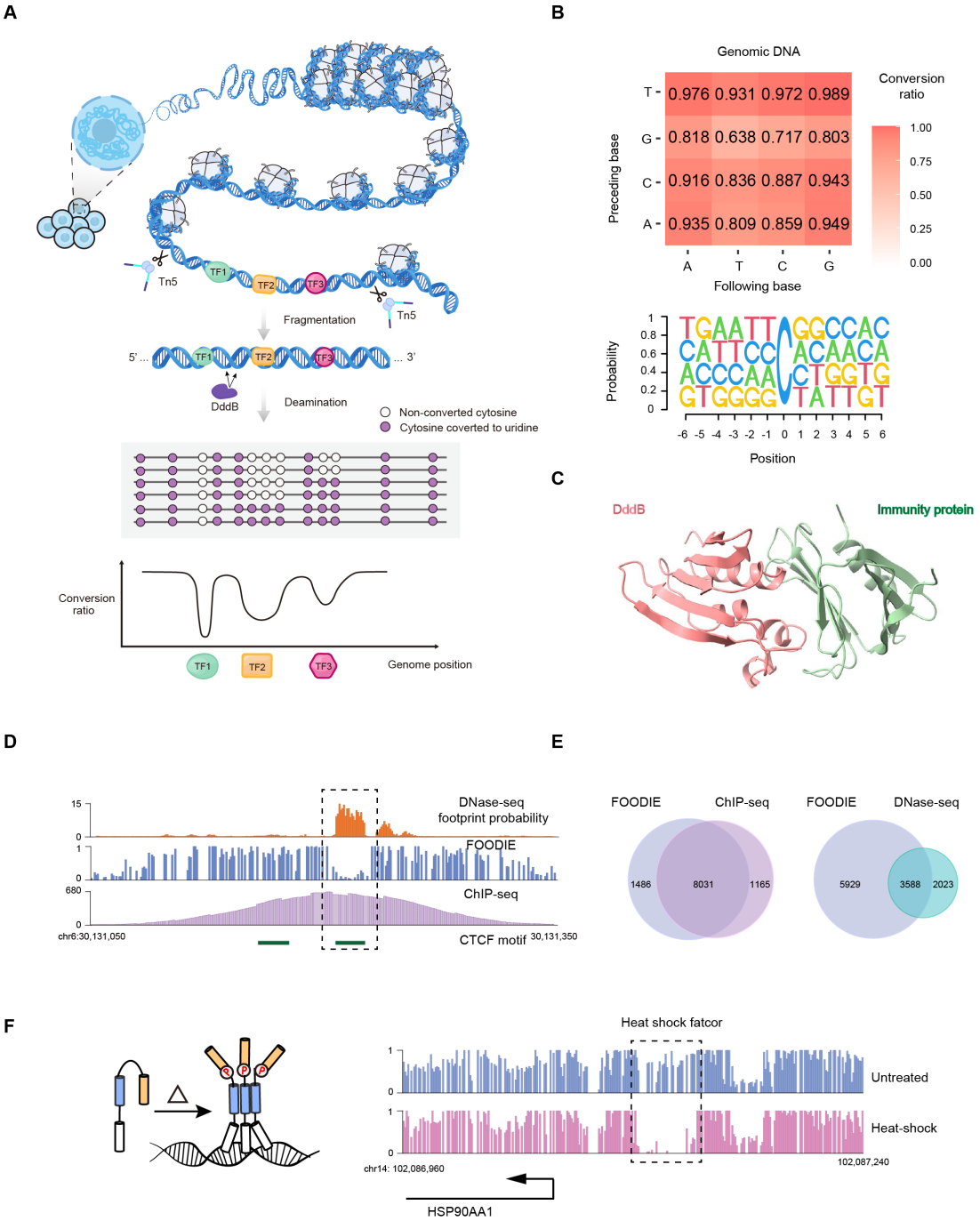


Figure 1. (A) Workflow of FOODIE. Cells were permeabilized for Tn5 tagmentation and DddB deamination. Amplicons of DNA undergone *in situ* enzymatic deamination were sequenced to identify TF footprints. (B) (Top) DddB C to U conversion efficiency of naked DNA *in vitro*. The x-axis is the base following, and the y-axis base proceeding the cytidine. (Bottom) Sequence logo

of the above data. (C) The crystal structure of DddB (red) complexed with its immunity protein (green). (D) Human genome tracks of DNase-seq footprint probability, FOODIE, and ChIP-seq in GM12878 cell line. The horizontal green bars represent the two possible CTCF motif sequences. Only the right one is bound by CTCF. (E) Venn diagrams of FOODIE with ChIP-seq or DNase-seq for the numbers of CTCF binding sites in open chromatin regions. (F) Heat shock factor (HSF) responding to heat shock. Upon heat shock, HSF monomers trimerize, bind to the Hsp90AA1 promoter, and leave a TF footprint on DNA.

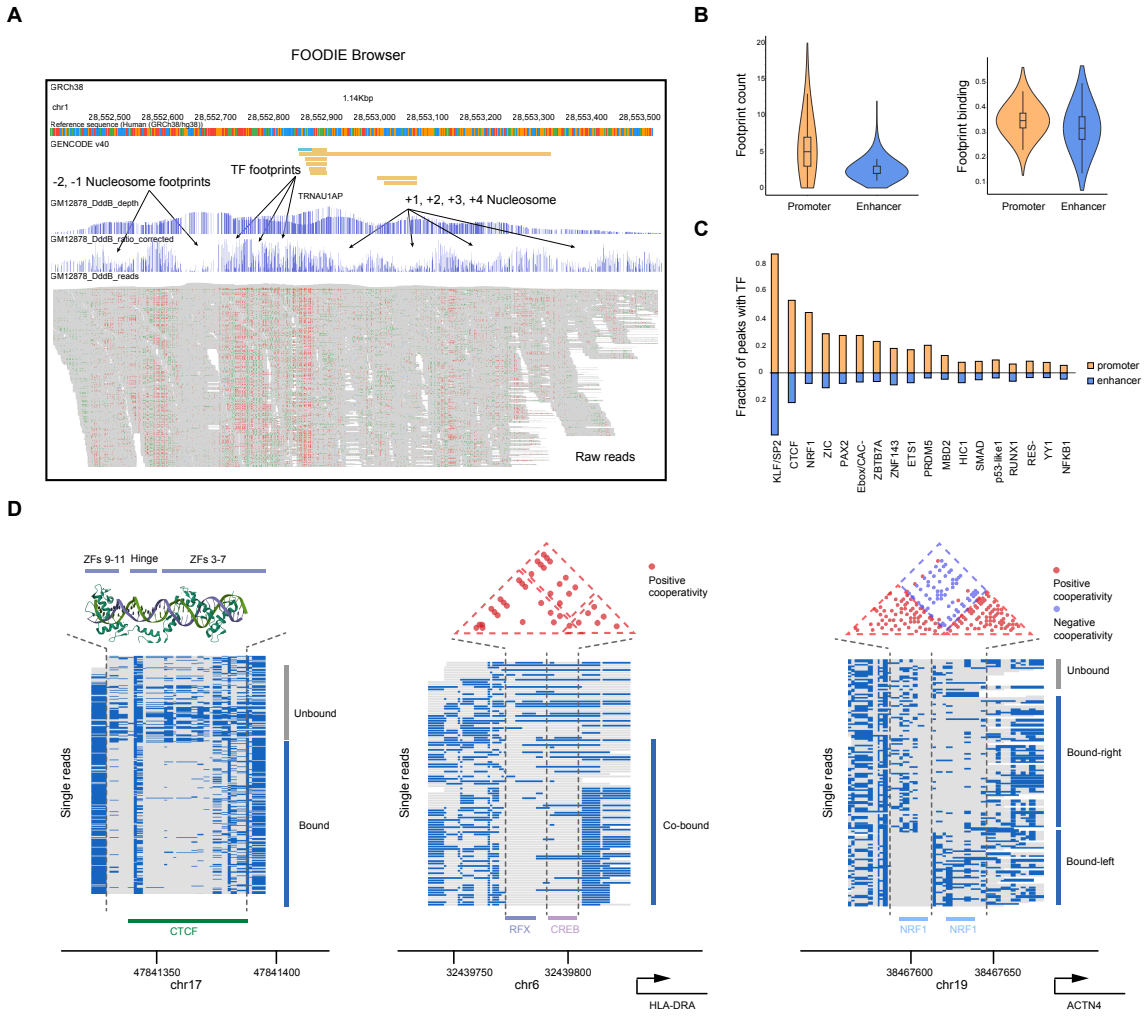
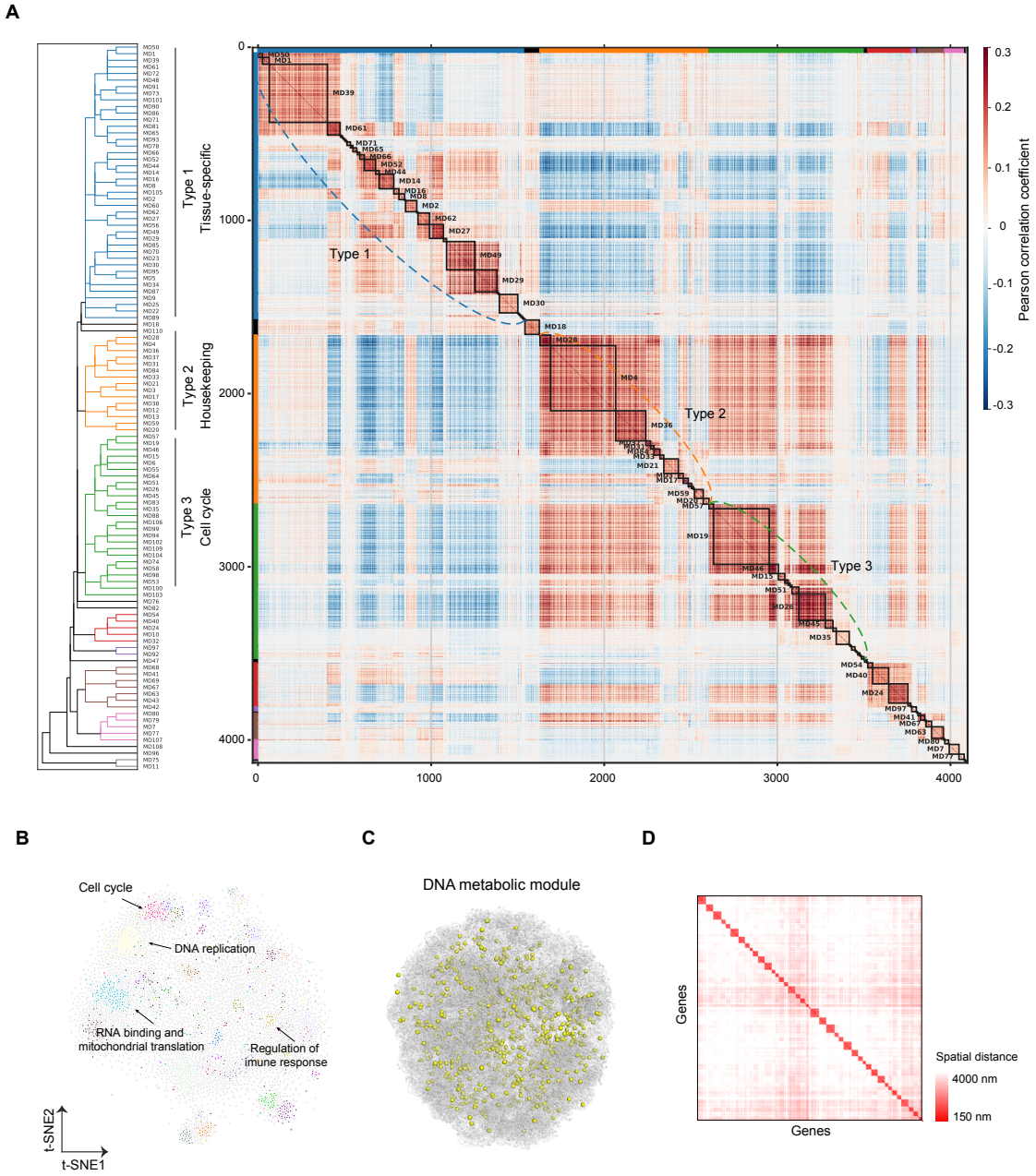


Figure 2. (A) FOODIE browser display at the promoter region of TRNAU1AP, with two nucleosomes at each side flanking the open chromatin region. The FOODIE sequencing depth is similar to ATAC-seq, whereas the FOODIE conversion ratio shows footprints of nucleosomes -1 & -2 and +1 & +2, and three distinct TFs in between. Below are single-molecule sequencing raw reads with the converted cytosines on the Watson strand (red) and the Crick strand (green). (B) (Left) Distributions for numbers of TF footprints in the promoters (median=5) and enhancers (median=2). (Right) Distributions for binding fractions of TF footprints in the promoters (median=0.348) and enhancers (median=0.317). (C) Abundance of those TF footprints with ChIP-seq-confirmed assignments on the promoters and enhancers. (D) (Left) Structure of CTCF (ZFs 3-11) and the single-molecule sequencing reads of FOODIE footprints on a CTCF binding site.

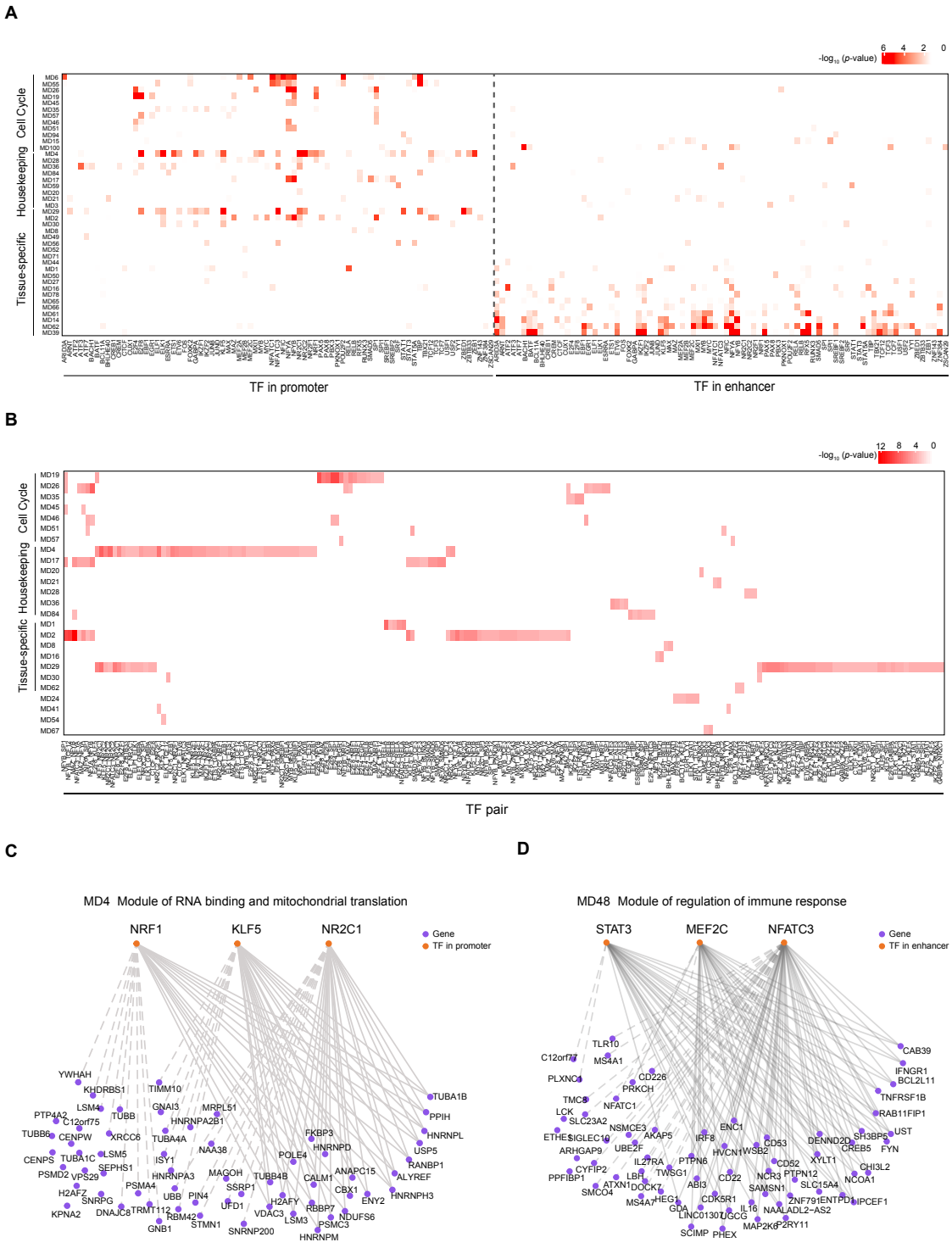
(Middle) Single-molecule sequencing reads showing RFX and CREB footprints at the HLA-DRA promoter. The positive binding cooperativity between two TFs is evident from the positive φ correlation coefficient (dots colored in red) between a pair of cytosines, each from both TF binding sites. (Right) Single-molecule sequencing reads with two NRF1 binding sites at the ACTN4 promoter have negative binding cooperativity, which is evident from the negative φ correlation coefficient (dots colored in blue) between a pair of cytosines, each from both TF binding sites.



333

334 Figure 3. (A) Pair-wise Pearson correlation coefficients for ~ 4000 detected mRNA levels of
335 GM12878 single cells under steady-state conditions. CGMs were identified by the improved
336 algorithm TriSHA (*35*), and indicated by enclosed squares along with the names of major ones.
337 Shown on the left was the hierarchical clustering of CGMs with three major clades: cell cycle,
338 housekeeping, and tissue-specific functions, respectively. (B) Gene t-sne plot of the CGMs along

with names of the representative ones. Each dot is an expressed gene, and genes in one CGM are labeled with a particular color. (C) 3D genome structure of a GM12878 cell experimentally determined by scMicro-C (*37*). Yellow points represent the promoter positions of the genes in the DNA replication CGM. (D) Pairwise spatial distance matrix for the genes in the DNA replication CGM, averaged over all scMicro-C structures, suggesting the genes in a CGM are far apart from each other rather than clustered.



364

365 Figure 4. (A) Heatmap showing the TF enrichment (-log *p*-value) within the major CGMs. Each

366 row indicates a CGM and each column indicates TFs assigned by ChIP-seq to the measured

367 footprints. Color density denotes the -log *p*-values calculated by the hypergeometric test. (B)
368 Heatmap showing the TF pairs enrichment (-log *p*-value) within the major CGMs. Each row
369 indicates a CGM and each column indicates TF pairs. Similarly, the -log *p*-values were calculated
370 by the hypergeometric test. (C) NRF1-KLF5-NR2C1 TF combination at promoters regulating the
371 genes in the M4 RNA binding and mitochondrial translation CGM. The solids lines are between a
372 gene that connects more than one TFs, whereas the dash lines are between a gene with only one
373 TF specified (and other un-assigned TFs). (D) STAT3-MEF2C-NFATC3 TF combination at
374 enhancers regulating the genes in the MD48 regulation of immune response CGM.
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391

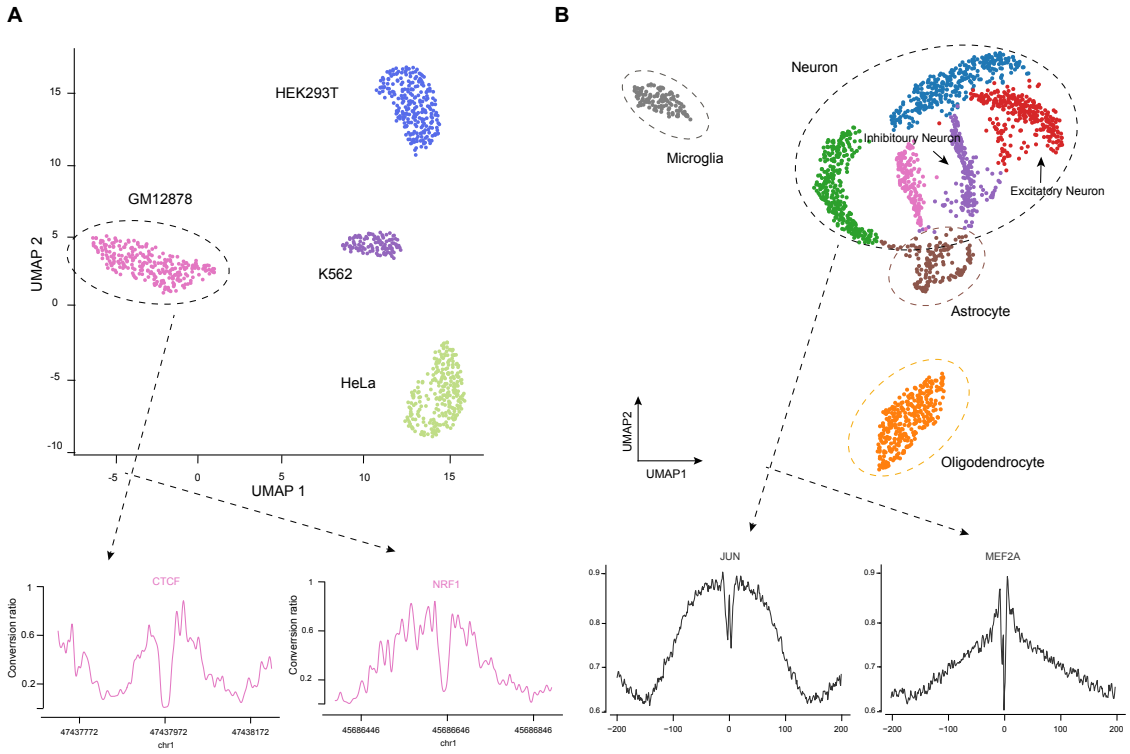


Figure 5. (A) Cells from a mixture of four human cell lines are classified into four different cell types based on their barcoded single-cell FOODIE data. Unsupervised clustering was performed using UMAP. CTCF and NRF1 footprints at two genome locations from pooled cells of the GM12878 cell type are shown in the inset. (B) Cell typing of ~1500 mouse hippocampus cells based on their single-cell FOODIE data. Different colors indicate distinct cell types, astrocytes, microglia, oligodendrocytes, and neurons (including inhibitory and excitatory neurons), with annotations based on the marker genes. Footprints of JUN and MEF2B in neurons at merged genome-wide binding sites are shown in the inset.

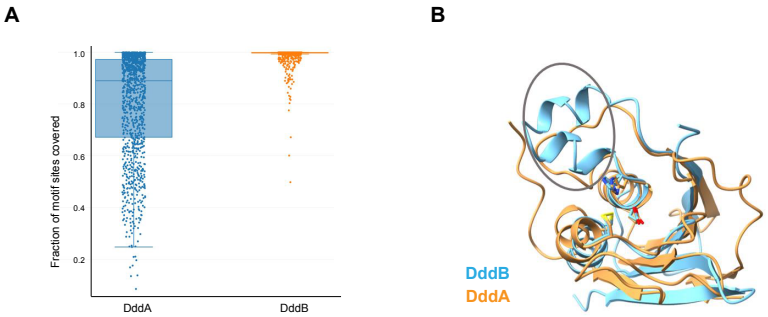


Figure S1. (A) Percentage of TF motifs on the human genome that can be assessed by DddA (probing every TC) and DddB (probing every C) enzymes. (B) Comparsion of cystal sturcutes of DddA and DddB.

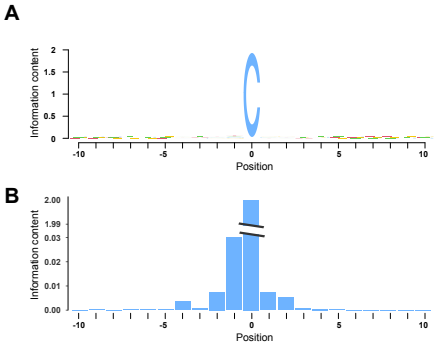


Figure S2. (A) Sequence logo of DddB conversion yields for a cytosine on naked DNA *in vitro*. (B) Bases from -4 to +2 positions of cytosine show relatively large bias on conversion.

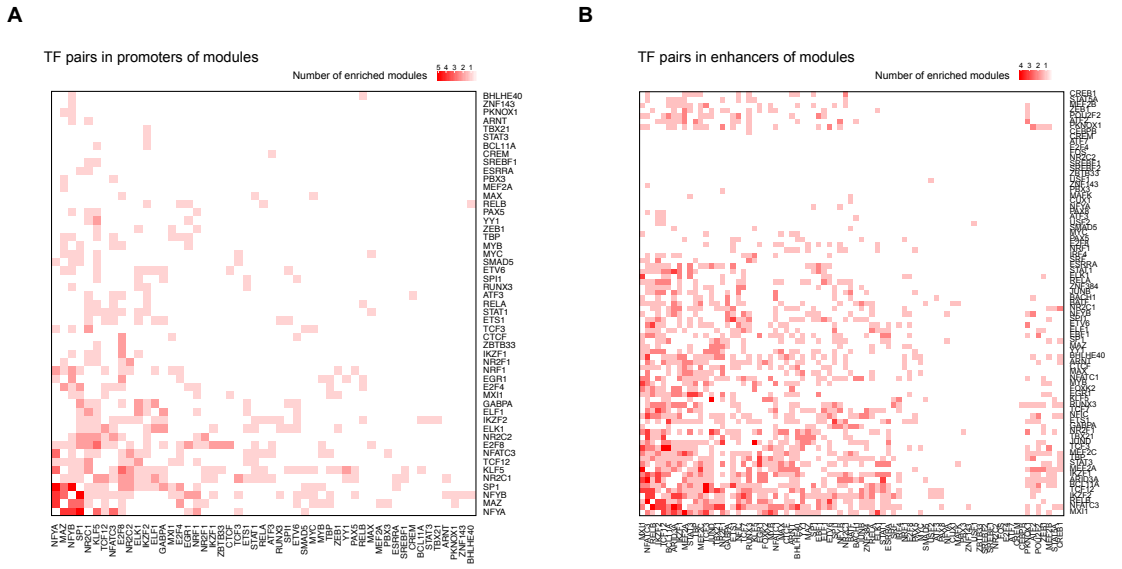


Figure S3. Enhancers harbor more enriched TF pairs than promoters. (A) Heatmap showing the number of modules, in which the TF pairs in a promoter are enriched. Color bar on the top denotes the number of enriched modules. (B) Heatmap showing the number of modules, in which the TF pairs in a enhancer are enriched. Color bar on the top denotes the number of enriched modules.

Material and Methods

Material and methods are placed in the supplementery document.

References:

1. S. A. Lambert *et al.*, The Human Transcription Factors. *Cell* **172**, 650–665 (2018).
2. F. Jacob, J. Monod, Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology* **3**, 318-356 (1961).
3. M. Ptashne, ISOLATION OF THE lambda PHAGE REPRESSOR. *Proc Natl Acad Sci U S A* **57**, 306-313 (1967).
4. J. Elf, G.-W. Li, X. S. Xie, Probing transcription factor dynamics at the single-molecule level in a living cell. *Science* **316**, 1191-1194 (2007).
5. J. Yu, J. Xiao, X. Ren, K. Lao, X. S. Xie, Probing gene expression in live cells, one protein molecule at a time. *Science* **311**, 1600-1603 (2006).
6. P. J. Mitchell, R. Tjian, Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245**, 371-378 (1989).
7. J. C. M. Gebhardt *et al.*, Single-molecule imaging of transcription factor binding to DNA in live mammalian cells. *Nature methods* **10**, 421-426 (2013).
8. C. Nüsslein-Volhard, E. Wieschaus, Mutations affecting segment number and polarity in Drosophila. *Nature* **287**, 795-801 (1980).
9. A. R. Chapman *et al.*, Correlated gene modules uncovered by high-precision single-cell transcriptomics. *Proceedings of the National Academy of Sciences* **119**, e2206938119 (2022).
10. M. Hu *et al.*, Correlated Protein Modules Revealing Functional Coordination of Interacting Proteins Are Detected by Single-Cell Proteomics. *The Journal of Physical Chemistry B* **127**, 6006-6014 (2023).
11. B. Ren *et al.*, Genome-Wide Location and Function of DNA Binding Proteins. *Science* **290**, 2306-2309 (2000).
12. G. Robertson *et al.*, Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods* **4**, 651-657 (2007).
13. D. S. Johnson, A. Mortazavi, R. M. Myers, B. Wold, Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* **316**, 1497-1502 (2007).
14. J. R. Hesselberth *et al.*, Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods* **6**, 283-289 (2009).
15. C. Sönmezer *et al.*, Molecular Co-occupancy Identifies Transcription Factor Binding Cooperativity In Vivo. *Molecular Cell*, (2020).
16. B. Y. Mok *et al.*, A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. *Nature* **583**, 631-637 (2020).
17. L. M. Iyer, D. Zhang, I. B. Rogozin, L. Aravind, Evolution of the deaminase fold and multiple origins of eukaryotic editing and mutagenic nucleic acid deaminases from bacterial toxin systems. *Nucleic acids research* **39**, 9473-9497 (2011).
18. M. H. de Moraes *et al.*, An interbacterial DNA deaminase toxin directly mutagenizes surviving target populations. *eLife* **10**, e62967 (2021).

19. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, 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).
20. M. Åkerfelt, R. I. Morimoto, L. Sistonen, Heat shock factors: integrators of cell stress, development and lifespan. *Nature Reviews Molecular Cell Biology* **11**, 545-555 (2010).
21. I. Rauluseviciute *et al.*, JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **52**, D174-D182 (2024).
22. I. E. Vorontsov *et al.*, HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors. *Nucleic Acids Research* **52**, D154-D163 (2024).
23. A. Jolma *et al.*, Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research* **20**, 861-873 (2010).
24. S. Mukherjee *et al.*, Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature Genetics* **36**, 1331-1339 (2004).
25. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636-640 (2004).
26. M. J. MacPherson, P. D. Sadowski, The CTCF insulator protein forms an unusual DNA structure. *BMC Molecular Biology* **11**, 101 (2010).
27. W. Reith, S. LeibundGut-Landmann, J.-M. Waldburger, Regulation of MHC class II gene expression by the class II transactivator. *Nature Reviews Immunology* **5**, 793-806 (2005).
28. P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
29. A. Dixit *et al.*, Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866.e1817 (2016).
30. D. M. Suter *et al.*, Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics. *Science* **332**, 472–474 (2011).
31. A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, S. Tyagi, Stochastic mRNA Synthesis in Mammalian Cells. *PLOS Biology* **4**, e309 (2006).
32. J. R. Chubb, T. Trcek, S. M. Shenoy, R. H. Singer, Transcriptional Pulsing of a Developmental Gene. *Current Biology* **16**, 1018–1025 (2006).
33. S. Chong, C. Chen, H. Ge, X. S. Xie, Mechanism of Transcriptional Bursting in Bacteria. *Cell* **158**, 314–326 (2014).
34. D. E. Nelson *et al.*, Oscillations in NF-kappaB signaling control the dynamics of gene expression. *Science* **306**, 704-708 (2004).
35. W. Ma, X. S. Xie, TriSHA: investigating modularity in high-precision single-cell transcriptomic data. *In preparation*, (2024).
36. L. Tan, D. Xing, C.-H. Chang, H. Li, X. S. Xie, Three-dimensional genome structures of single diploid human cells. *Science* **361**, 924-928 (2018).
37. H. Wu, J. Zhang, L. Tan, X. S. Xie, Extruding transcription elongation loops observed in high-resolution single-cell 3D genomes. *bioRxiv*, 2023.2002. 2018.529096 (2023).
38. W. Zhu, P. H. Giangrande, J. R. Nevins, E2Fs link the control of G1/S and G2/M transcription. *The EMBO Journal* **23**, 4615-4626-4626 (2004).
39. K. Taniguchi, M. Karin, NF-κB, inflammation, immunity and cancer: coming of age. *Nature Reviews Immunology* **18**, 309-324 (2018).