Human transcription factor combinations mapped by 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[1,2,3], 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[1,2,3], 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] *

Affiliations

[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.

* Corresponding author. Email: sunneyxie@biopic.pku.edu.cn

Abstract

An individual's somatic cells have the same genome but exhibit cell-type-specific transcriptome 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 measure human TFs' binding sites genome-wide with single-base resolution by footprinting with deaminase (FOODIE). Single-molecule sequencing reads from thousands of cells after *in situ* deamination yielded site-specific TF binding fractions and the cooperativity among adjacent TFs. In a human lymphoblastoid cell line, we found that, genes in a correlated gene module (CGM) share TF(s) in their *cis*-regulatory elements to facilitate a particular biological function. Finally, single-cell resolved experiments (scFOODIE) allow cell-type-specific TF footprinting in heterogeneous brain tissues.

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. Transcription factors (TFs) bind to *cis*-regulatory elements, including 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, acting as a key to turn on and off an operon, contains 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 (*9-12*). 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 (*13*) and ChIP-seq (*14, 15*), 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 (*16*). 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) (*17*). However, it is not appliable

62 to normal mammalian cells, as the extraneous methyltransferase used was mostly indistinguishable
63 from endogenous methylation.
64 Here we report a novel approach based on double-stranded DNA cytosine deamination, named
65 FOOtprinting with DeamInasE (FOODIE), to detect TFs' binding sequences at single-base
66 resolution along the genome, as the TFs' steric hindrance prevents the conversion of cytosine to
67 uracil. Cost-effective and scalable, FOODIE requires sequencing only thousands of cells,
68 significantly fewer than those required for ChIP-seq or DNase-seq.
69
70 **Genome-wide mapping of TF footprints**
71 FOODIE relies on double-stranded DNA deaminase with high efficiency (Fig. 1A). We
72 initially tested DddA, the first double-stranded DNA deaminase reported (*18*), which only
73 deaminates cytosine preceded by thymine, not capable of footprinting all TFs bound to the genome
74 (Fig. S1A). To overcome this limitation, we characterized a new dsDNA cytosine deaminase from
75 a large group of reported toxins (*19*), which we named DddB (also known as BadTF3 (*20*)) after
76 DddA. Unlike DddA, DddB converts cytosine to uracil largely independent of adjacent bases (Fig.
77 1B), allowing us to footprint almost all TF motif sequences (Fig. S1A). We determined the crystal
78 structure of DddB in complex with its immunity protein at 2.5 Å resolution (Fig. 1C). Despite the
79 overall structural similarity between DddB and DddA, DddB features an intervening helix near the
80 active site, whereas DddA has a loop in this region (Fig. S1B).
81 Similar but one step more than ATAC-seq (*21*), FOODIE uses transposase Tn5 in
82 permeabilized cells to enrich the open chromatin regions, where most TFs bind and exert their
83 regulatory roles. DddB is then added after Tn5 tagmentation to detect protein binding on DNA,
84 whose underlying cytosines are protected from enzymatic deamination. As such, FOODIE is
85 capable of probing the regulatory sites in open chromatin regions in a cost-effective and high-
86 throughput way.
87 TF binding profiles detected by ChIP-seq, FOODIE, and DNase-seq are shown in Fig. 1D.
88 ChIP-seq had a low spatial resolution (200-500bp), not capable of telling whether both or just one
89 of the detected CTCF sequences were occupied. In contrast, FOODIE accurately identified CTCF
90 that the footprint on the right is bound, which was consistent with DNase-seq footprinting. We
91 further compared the number of genome-wide binding sites of CTCF detected by FOODIE with
92 those obtained from ChIP-seq and DNase-seq. The CTCF binding sites detected by FOODIE show

93 strong concordance with those detected by ChIP-seq and largely overlap with those detected by
94 DNase-seq footprinting (Fig. 1E).
95 Next we employed a heat shock experiment to validate the reliability of FOODIE. Upon a
96 temperature increase, the transcription factor HSF is activated to bind to the promoters of target
97 genes to induce expression (*22*). At HSP90AA1 promoter, we observed a decrease in the
98 conversion ratio within HSF motif sequences, which corresponded to the gain of HSF binding
99 upon heat shock (Fig. 1F). The above data suggested that FOODIE can identify precise binding of
100 TFs with high accuracy.
101 To support the interactive analyses of genome-wide footprints identified by FOODIE, we
102 developed a web server (http://foodie.sunneyxielab.org) with a user-friendly interface. This
103 platform includes a genome browser for visualization and a data portal with our FOODIE data.
104 Fig. 2A shows a browser display of the genomic region around the TRNAU1AP promoter in
105 chromosome 1, a chromatin open region between nucleosomes -1 (and -2) and +1 (and +2), with
106 three distinct footprints. The first row is the ATAC track seen by Tn5 tagmentation. The second
107 row is the C to U conversion efficiency showing the footprints. The raw reads of single-molecule
108 sequences below, with the converted cytosines on the Watson strand (red) and the Crick strand
109 (green), show the binding fractions along the genome.
110 We found that the deamination efficiency of DddB is dependent on the neighboring sequences
111 of cytosines (especially from -4 to +2) (Fig. S2), which needs to be normalized when calculating
112 the binding fraction. We measured deamination efficiency along the naked human genome as a
113 positive control for the conversion ratio correction. The binding fraction is one minus conversion
114 ratio.
115 Fig. 2B shows the mean number of footprints present in *cis*-regulatory elements of genes and
116 their relative arrangement (see Material and Methods). The median number of footprints per gene
117 promoter was found to be 5 (median number), compared to 2 in enhancer regions. Notably, the
118 binding fractions of TFs in the promoter regions were higher than those of TFs in enhancer regions.
119 The FOODIE footprints need to be assigned to particular TFs, commonly through motifs
120 tabulated on websites, such as JASPAR (*23*) and HOCOMOCO (*24*), which collected vast amounts
121 of information from *in vitro* data of SELEX (*25*) and PBM's (*26*) as well as *in vivo* data of ChIP-
122 seq (*27*). It turned out that a footprint has many different possible assignments of TFs, which we
123 can not pin down because of the lack of binding affinities from databases and the intracellular

concentrations. While we are working on the methods to fully assign the footprints, in this work we assigned footprints to certain TFs only when they have strong ChIP-seq signal in GM12878 cell line. This resulted in reliable assignments to only seventy-nine known TFs.

As such, the abundance of the assigned TFs is shown in promoter and enhancer regions, we found promoters were particularly enriched for the NRF1, KLF/SP, ZNF143, ETS, and YY1, footprints, which were rarely detected in enhancers (Fig. 2C). Taken together, TF bound to promoters and enhancers differ in average number within open regions, binding fractions and TF types.

Single-molecule sequencing reads revealing binding cooperativity among TFs

While the binding fraction of a particular binding site can be determined by the proportion of single molecules of sequenced DNA with unconverted C to that of the total of converted and unconverted C. The fact that sequencing occurs on individual molecules allows us to determine the cooperativity between TFs bound to two adjacent binding sites.

Single-molecule reads at a CTCF binding site in the genome are shown in Fig. 2D, exhibiting two separate footprints of the bridging structure of the zinc finger hinge (*28*), which are either simultaneously occupied or not. In Fig. 2D, two adjacent TFs RFX and CREB have two distinct footprints (*29*). To evaluate their binding cooperativity, we performed Fisher's exact test for the significance of independence on a pair of cytosines, each selected from the two binding sites. The Φ correlation coefficient is between 0 and 1 for positive cooperativity and between -1 and 0 for negative cooperativity. A red dot indicates significant positive cooperativity ($P<0.001$). A pair of cytosines within the RFX (or CREB) binding sequence are positively cooperative because the binding of RFX to them is simultaneous. Interestingly, cytosine pairs across CREB and RFX binding sequences are also positively cooperative, i.e. RFX and CREB's binding is not independent but facilitates each other (Fig. 2D). We attribute this positive coopertavity to protein-protein interactions between two TFs, which has been previously reported (*29*).

Conversely, negative cooperativity was seen between two NRF1 binding sites within the ACTN4 promoter (Fig. 2D). While the Φ of the two cytosines within each NRF1 binding site is positive, the pairs across different sites are negative (blue: $P<0.001$), meaning their binding is exclusive. Such single-molecule measurements provide unprecedented information about how TFs work together in the genome.

155

156 **Shared TFs among genes within the same CGM**

157 Many protein-coding genes form a module to work together to accomplish a particular
158 biological function. People have been looking for gene modules with the bulk (*12*) and single cells
159 (*9, 10, 30*) by perturbation measurements. By improving the detectability of single-cell
160 transcriptome, we demonstrated the identification of correlated gene modules (CGMs) without
161 perturbation under the steady state condition (*9, 10*). The bursty nature of gene expression has
162 been long established and attributed to the single-molecule nature of DNA in single cells (*31-35*).
163 The coordination of a particular biological function by multiple genes requires their
164 synchronization. Experimentally, such synchronization can be probed by measuring the covariance
165 matrix of any pair of RNA in a single-cell transcriptome experiment (MALBAC-DT) (*9, 10*).

166 Recently we have developed a new method for more accurate identification of CGMs, Triangle
167 Screening of Hierarchical Accumulated-Edges (TriSHA) (*36*). Over Representation Analysis
168 (ORA) reveals that nearly 80% of the modules are functionally annotated with a False Discovery
169 Rate (FDR) below 0.05. Through module connection analyses, we have identified three distinct
170 types of CGMs within the GM12878 cell line based on functional annotation, associated with
171 specific cellular functions: housekeeping, tissue-specific, and cell cycle (Fig. 3A). Supplementary
172 Table S1 lists the CGMs with size larger than 10 that we identified. Gene t-SNE plot of GM12878
173 clearly shows the cluster of genes in the identified CGMs (Fig. 3B).

174 How are the genes in the same module distributed within the 3D genome? Are they clustered
175 together, or spatially dispersed? Our group determined the 3D genome structure of a somatic
176 human cell (*37*). To answer the above question, we employed our latest high-resolution single-cell
177 3D genome structures of GM12878 cells using scMicro-C (*38*). Taking the DNA replication CGM
178 (MD20, cell cycle clade) of 317 genes as an example, Fig. 3C shows that they are not clustered
179 but relatively randomly distributed within the nucleus, with the more quantitative distribution of
180 pair-wise distance shown in Fig. 3D. No spatial clustering of genes in other gene modules was
181 observed either. This observation does not support the idea that the correlated genes are transcribed
182 within certain locations in the nucleus. We hypothesize that the coordination of different genes in
183 the module is achieved through TF presented in the contiguous euchromatin space.

184 With FOODIE, we are in a position to determine how TFs organize the CGMs. We looked at
185 the enrichment of TFs shared among both promoters and enhancers of genes within a module,

186 which is characterized by the hypergeometric test. Out of sixty-five CGMs identified in GM12878
187 (Fig. 3A and Supplementary Table S1), we successfully identified that many TFs are statistically
188 enriched in different gene modules, including for example, E2F factors in the promoters of genes
189 in MD16 cell cycle-related CGM (*39*) and RELB in the enhancers of genes in MD47 regulation of
190 immune system CGM (*40*) (Fig. 4A). As a control, when we randomly selected tens of genes (size
191 from 30 to 150) out of the ~4000 expressed genes, no TFs were significantly enriched.

192 Significantly we found enrichments of specific TFs are clearly different for the three types of
193 CGMs. For cell cycle and housekeeping CGMs, the TF enrichment is primarily in promoters. For
194 example, the NF-Y factor is enriched in the genes of MD14 cholesterol metabolism CGM (Fig.
195 4A). The number of enriched TFs ($P<0.05$) were 31 and 59 for cell cycle modules and house
196 keeping modules, respectively. As a control, if the cell cycle and house keeping modules were
197 grouped together two large groups, the number enriched TFs ($P<0.05$) were reduced to 9 and 30,
198 respectively. This lends further support on module-specific TF enrichment.

199 For tissue-specific CGMs, the TF enrichment is primarily in enhancers (Fig. 4A). For example,
200 STAT3 enrichment is present in MD 48 immune response CGM. This suggests that there are two
201 distinct modes of CGM regulation: genes with essential cellular functions rely on TFs' binding at
202 promoters for robustness, while tissue-specific genes depend more on enhancers for larger
203 tunability.

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

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

216

217 **Footprinting after cell typing by scFOODIE in heterogeneous tissues**

218 Our goal is to characterize different human tissues with FOODIE and determine their genome-
219 wide TF footprints. However, human tissues are heterogeneous with many different cell types, so
220 it is necessary to determine each cell's identity. We note that each gene only has one maternal
221 allele and one paternal allele in an individual cell, which is not sufficient to provide reliable
222 statistics for TF footprints. However, thousands of single cells identified for a particular cell type
223 would be sufficient to carry out footprinting.

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

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

238

239 **Outlook**

240 In conclusion, we have shown that FOODIE is a high-precision method for genome-wide *in*
241 *situ* mapping of TF footprints inside the nucleus with single-base resolution. FOODIE is
242 experimentally simple, high-throughput, and cost-effective. ScFOODIE allows for cell typing in a
243 heterogeneous tissue sample and TF footprinting in any cell type within a complex tissue. A
244 systematic FOODIE database of different human tissues is underway. We anticipate that such
245 information about TF combinations of different cell types and correlated gene modules will lead
246 to the fundamental understanding of transcriptional regulation and cellular functions in eukaryotic

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. **Data and materials availability:** Raw sequencing data and processed data have been deposited at http://foodie.sunneyxielab.org.

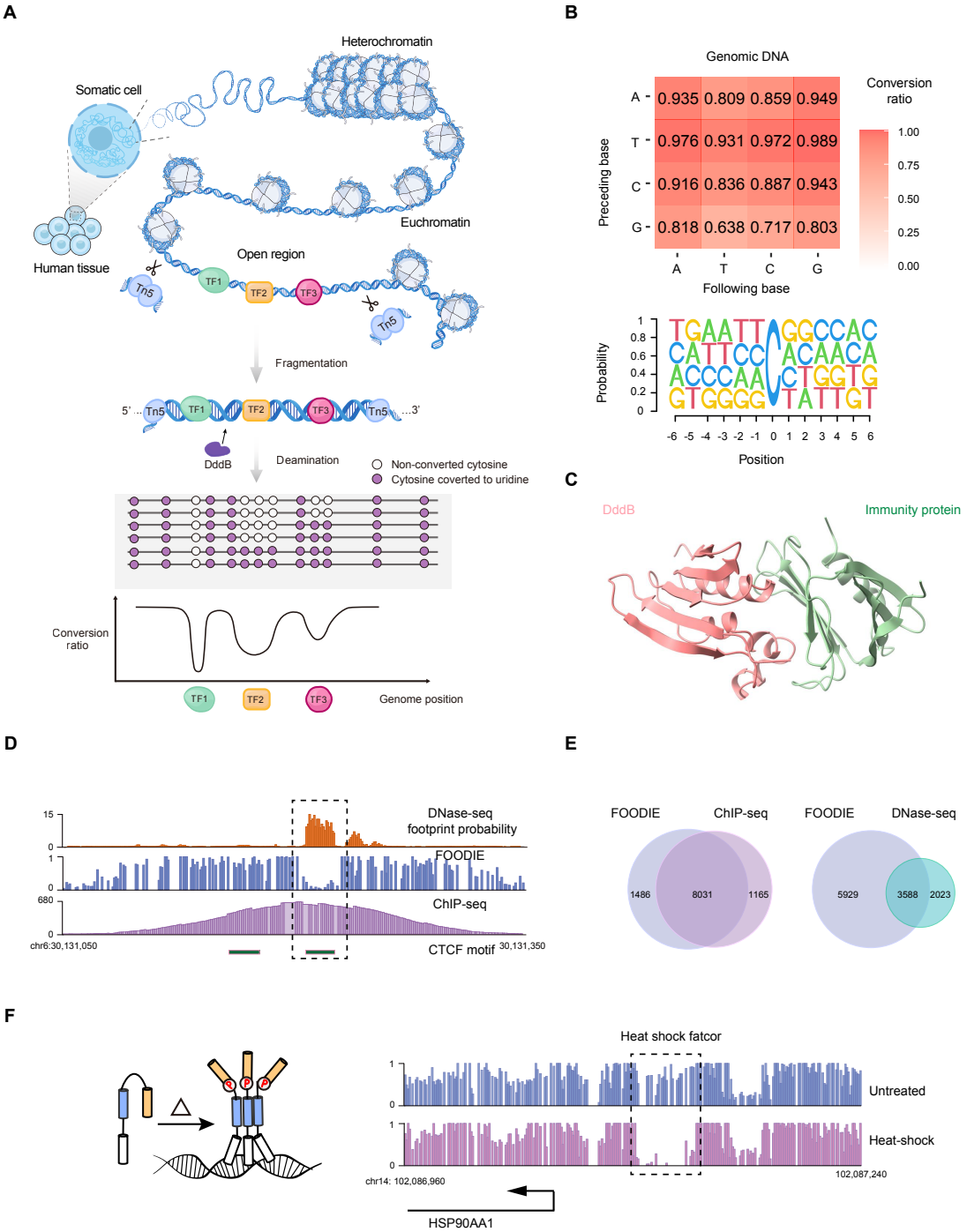


Fig. 1. Genome-wide mapping of TF binding sites by FOODIE. (A) Workflow of FOODIE. Cells were permeabilized for Tn5 tagmentation and DddB deamination. Amplicons of DNA undergone *in situ* enzymatic deamination were sequenced to identify TF footprints. (B) (Top) DddB C to U conversion efficiency of naked 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.

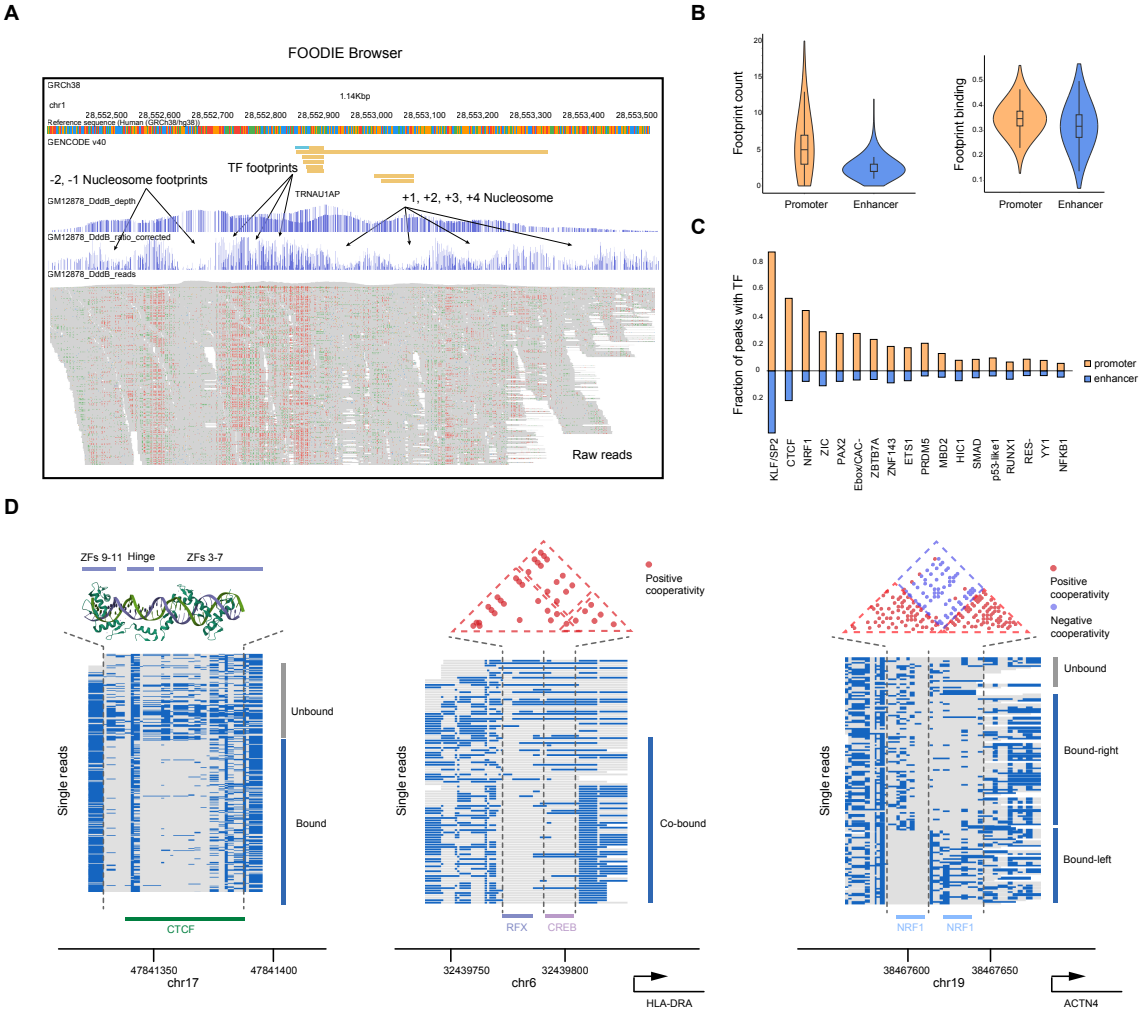


Fig. 2. TF binding fractions and cooperativity among TFs detected by FOODIE. (A) FOODIE browser display at the promoter region of TRNAU1AP, with two nucleosomes at each side flanking the open chromatin region. The FOODIE sequencing depth is similar to ATAC-seq, whereas the FOODIE conversion ratio shows 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.

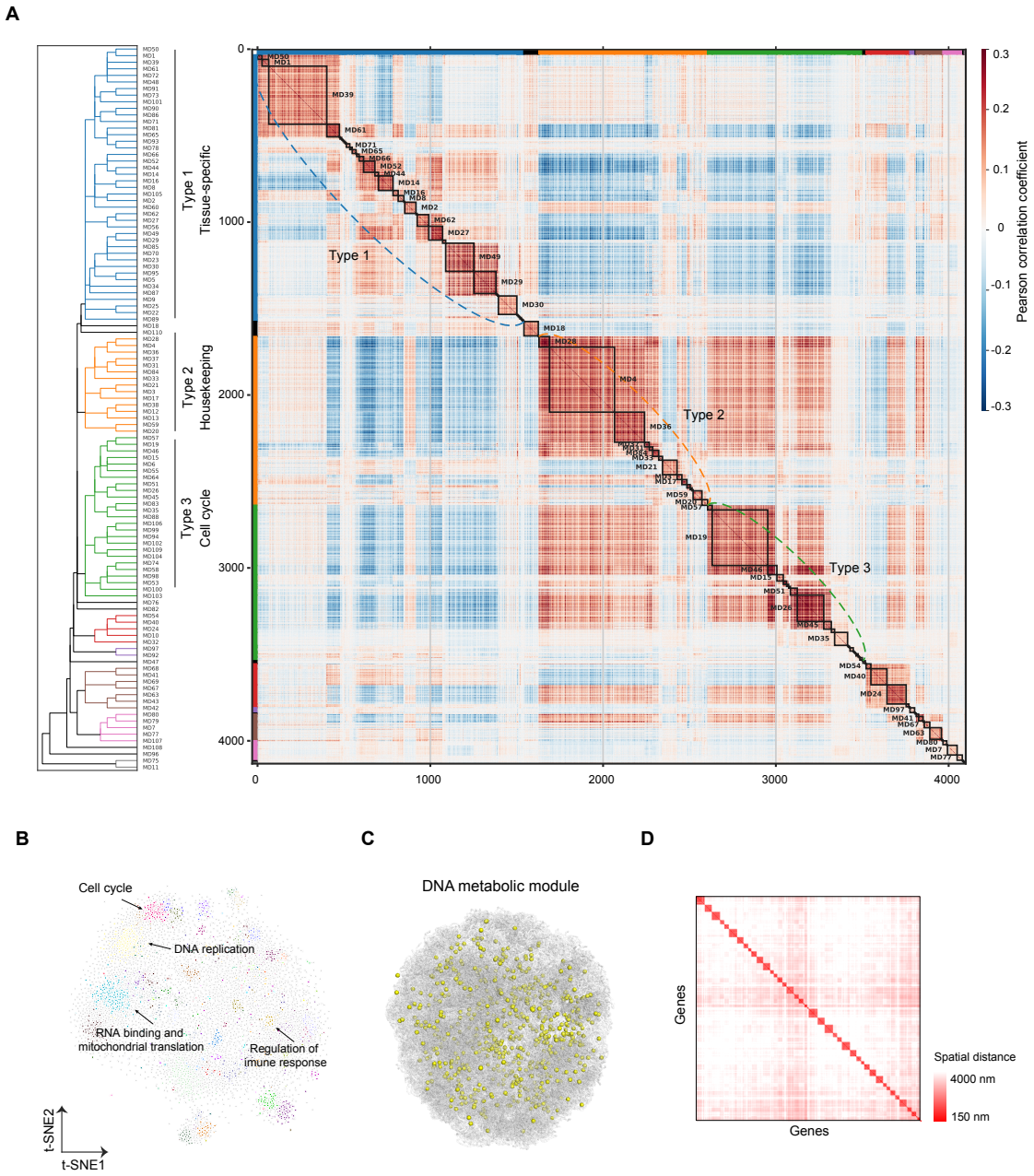


Fig. 3. Genes in a CGM for a biological function not spatially clustered in nucleus. (A) Pair-wise Pearson correlation coefficients for ~ 4000 detected mRNA levels of GM12878 single cells under steady-state conditions. CGMs were identified by the improved algorithm TriSHA (*36*), and indicated by enclosed squares along with the names of major ones. Shown on the left was the

hierarchical clustering of CGMs with three major clades: cell cycle, 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 (*38*). 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.

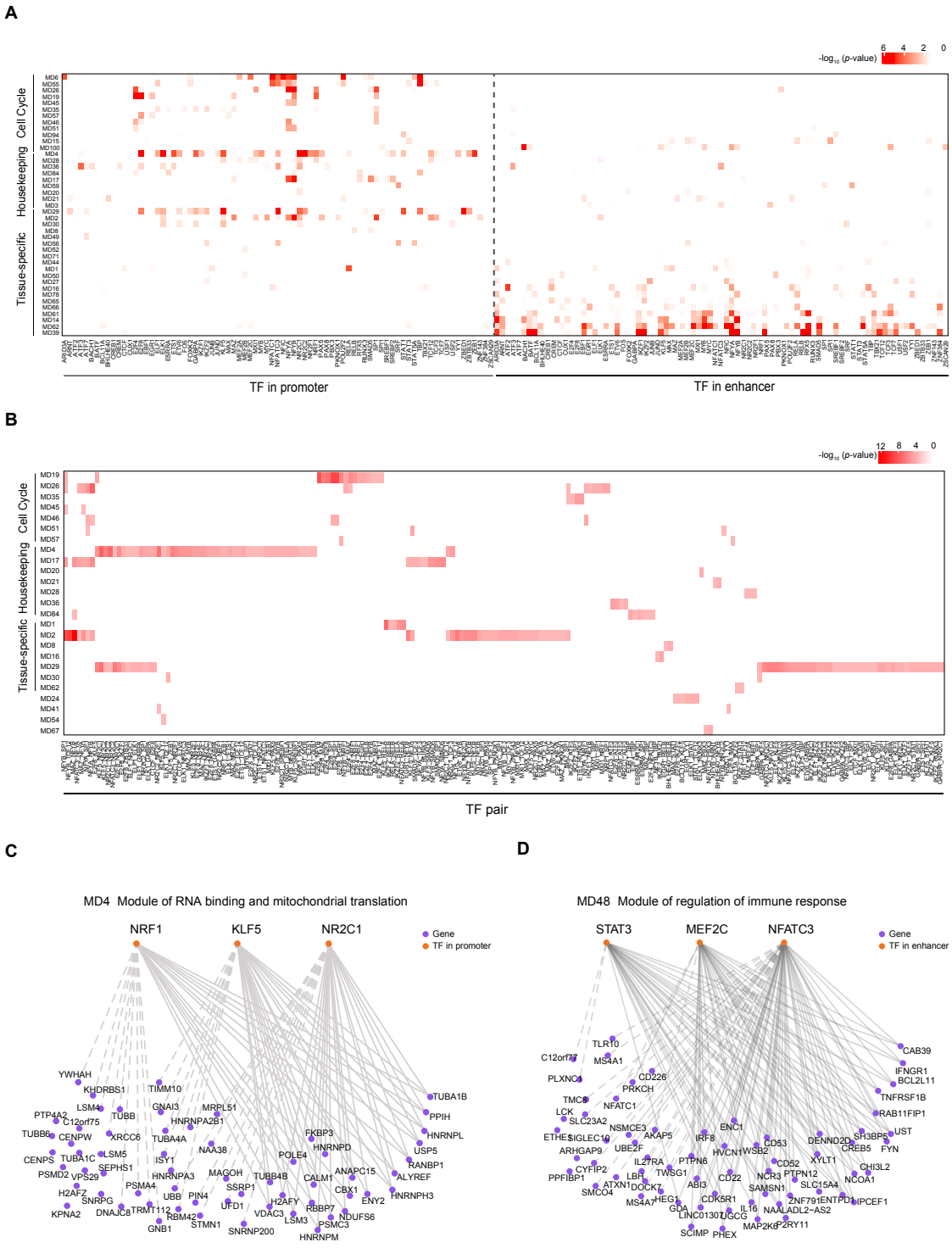


Fig. 4. Genes in a CGM sharing TF(s) in their promoters and enhancers. (A) Heatmap showing the TF enrichment (-log *p*-value) within the major CGMs. Each row indicates a CGM

375 and each column indicates TFs assigned by ChIP-seq to the measured footprints. Color density
376 denotes the -log *p*-values calculated by the hypergeometric test. (B) Heatmap showing the TF pairs
377 enrichment (-log *p*-value) within the major CGMs. Each row indicates a CGM and each column
378 indicates TF pairs. Similarly, the -log *p*-values were calculated by the hypergeometric test. (C)
379 NRF1-KLF5-NR2C1 TF combination at promoters regulating the genes in the M4 RNA binding
380 and mitochondrial translation CGM. The solids lines are between a gene that connects more than
381 one TFs, whereas the dash lines are between a gene with only one TF specified (and other un-
382 assigned TFs). (D) STAT3-MEF2C-NFATC3 TF combination at enhancers regulating the genes
383 in the MD48 regulation of immune response CGM.
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400

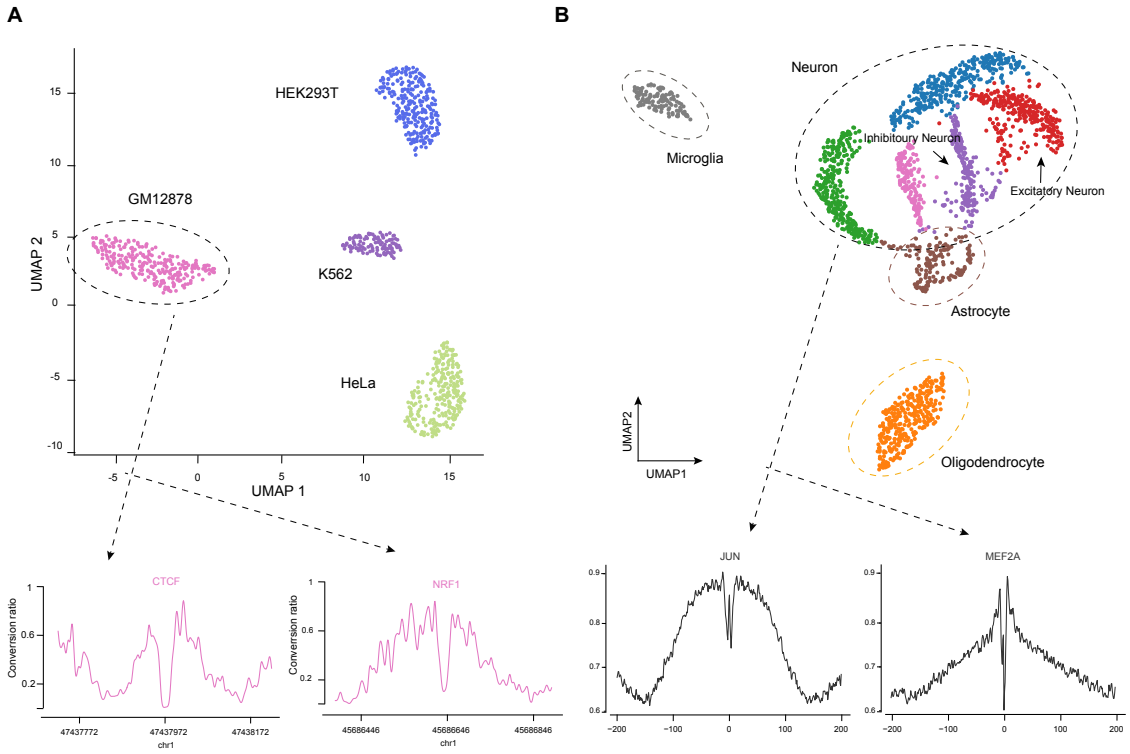


Fig. 5. Cell-type-specific TF footprinting in heterogeneous tissues by scFOODIE. (A) Cells from a mixture of four human cell lines are classified into four different cell types based on their barcoded scFOODIE data. Unsupervised clustering was performed using UMAP. CTCF and NRF1 footprints at two genome 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 scFOODIE 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.

417 **Materials and Methods:**

418 Plasmid construction

419 Expression constructs for the DddA toxin domain (DddAtox) and DddB were obtained from
420 Genescript through a gene synthesis service. To generate a pETDuet-1-based expression construct
421 for DddAtox, the DddAtox encoding sequence:
422 ATGGGCAGCAGCCATCACCATCATCACCACAGCCAGGATCCGGGTAGCTATGCGCT
423 GGGTCCGTATCAGATCTCTGCTCCGCAGCTGCCGGCATATAACGGTCAGACTGTTGG
424 TACTTTCTATTATGTTAACGATGCTGGCGGTTTAGAAAGCAAAGTTTTCAGCTCTGGT
425 GGTCCGACCCCGTATCCGAACTATGCTAACGCTGGTCACGTTGAAGGTCAGTCTGCT
426 CTGTTCATGCGTGATAACGGTATCTCTGAAGGTCTGGTTTTCCATAACAACCCGGAA
427 GGTACCTGTGGTTTTTGTGTTAACATGACCGAAACCCTGCTGCCGGAAAACGCTAAA
428 ATGACCGTTGTTCCGCCGGAAGGTGCGATTCCGGTTAAACGTGGTGCTACCGGTGAA
429 ACCAAAGTTTTCACCGGTAACTCTAACTCTCCGAAATCTCCGACCAAAGGTGGTTGC
430 TAA (of which the translated protein sequence is:
431 MGSSHHHHHHSQDPGSYALGPYQISAPQLPAYNGQTVGTFYYVNDAGGLESKVFSSGG
432 PTPYPNYANAGHVEGQSALFMRDNGISEGLVFHNNPEGTCGFCVNMTETLLPENAKMT
433 VVPPEGAIPVKRGATGETKVFTGNSNSPKSPTKGGC.), is synthesized and cloned into the
434 MCS-1 (NcoI and HindIII sites, retaining an N-terminal hexahistidine tag), and the immunity
435 protein DddAI encoding sequence:
436 ATGTATGCGGATGACTTTGACGGGGAAATTGAGATTGATGAAGTTGATAGCCTAGTT
437 GAGTTTCTGAGCCGTCGTCCGGCGTTCGATGCGAACAACTTCGTTCTGACCTTCGAA
438 GAAAGCGGCTTCCCGCAGCTGAACATCTTCGCGAAAAACGATATCGCGGTTGTTTAC
439 TACATGGATATCGGCGAAAACTTCGTTAGCAAAGGCAACAGCGCGAGCGGCGGCAC
440 CGAAAAATTCTACGAAAACAAACTGGGCGGCGAAGTTGATCTGAGCAAAGATTGCG
441 TTGTTAGCAAAGAACAGATGATCGAAGCGGCGAAACAGTTCTTCGCGACCAAACAG
442 CGTCCGGAACAGCTGACCTGGAGCGAACTGTAA (of which the translated protein
443 sequence is:
444 MYADDFDGEIEIDEVDSLVEFLSRRPAFDANNFVLTFEESGFPQLNIFAKNDIAVVYYMD
445 IGENFVSKGNSASGGTEKFYENKLGGEVDLSKDCVVSKEQMIEAAKQFFATKQRPEQLT
446 WSEL.), is synthesized and cloned into the MCS-2 (NdeI and XhoI sites, removing an C-terminal
447 S-tag). To generate a pETDuet-1-based expression construct for DddB, the DddB encoding

448 sequence:
449 ATGGGCAGCAGCCATCACCATCATCACCACAGCCAGGATCCTGGTTGGAAATTTTCT
450 AACGGTAAACGCCGTCCGCCGCACAAAGCAACGGTAACTGTGACCGATAAAAACGG
451 TGTCGTTAAACACAAAAGCAACCTGGTTTCTGGCAACATGACTGAAGCCGAAAAGA
452 AACTGGGCTTCCCGAACAACTCCCTGGCGACCCACACCGAAAACCGTGCTACCCGC
453 CTGATCGATCTGAACCAAGGTGATACTATGCTGATCGAGGGCCAATACCGTCCGTGT
454 CCACGTTGTAAAGGTGCAATGCGCGTGAAAGCGGAGGAATCCGGTGCGAAAGTGAT
455 CTACACCTGGCCAGAAGATGGTGACCTGAAAAAACGTGAATGGGAAGGCACTCCGT
456 GCGACAAAAAATAA (of which the translated protein sequence is:
457 MGSSHHHHHHSQDPGWKFSNGKRRPPHKATVTVTDKNGVVKHKSNLVSGNMTEAEK
458 KLGFPNNSLATHTENRATRLIDLNQGDTMLIEGQYRPCPRCKGAMRVKAEESGAKVIYT
459 WPEDGDLKKREWEGTPCDKK.), is synthesized and cloned into the MCS-1 (NcoI and HindIII
460 sites, retaining an N-terminal hexahistidine tag), and the immunity protein DddBI encoding
461 sequence:
462 ATGACCAAATCTAAAATGCTGAGCAACATCGTCATCCAGGAGGTCAAATTTGCGATC
463 GAAGATTACTGCGCTATTCTGAGCTTCGCTTCTGACTCTTATGAAGTGCCGGAGCAG
464 TATTTTATCATTACCCGTTCTACCACCGAACGTTCTGGCGGTATTCCGGAGGGCGAC
465 ATCTACCTGGAATCTAACCTGTTTCTGGATTTTAACCCGTACGGCCTGAGCGGTTACC
466 TGCTGTCTGAGCCGAACTGCGTAGATCTGCTGATCGAACCGAACAACTACGTTCGTC
467 TGCGTCTGATCGAAAAAATCGATATCCTGGAAGTGGAAAACCACCTGAAATTTCTGT
468 TCGACAACTAA (of which the translated protein sequence is:
469 MTKSKMLSNIVIQEVKFAIEDYCAILSFASDSYEVPEQYFIITRSTTERSGGIPEGDIYLESN
470 LFLDFNPYGLSGYLLSEPNCVDLLIEPNNYVRLRLIEKIDILEVENHLKFLFDN.), is
471 synthesized and cloned into the MCS-2 (NdeI and KpnI sites, removing an C-terminal S-tag). The
472 pETDuet-1::dddAtox + dddAI vector and pETDuet-1::DddBtox + DddBI were transformed into
473 *E. coli* strains DH5α BL21 and stored at -20 degrees.
474
475 <u>Bacterial strains and culture conditions</u>
476 Unless otherwise noted, all Escherichia coli (E. coli) strains used in this study were grown in
477 Lysogeny Broth (LB) at 37 °C or on LB medium solidified with agar (LBA, 1.5% w/v). The media

478 was supplemented with ampicillin (100 μg per ml) or IPTG (0.5 mM) if it is required. E. coli
479 strains DH5α and BL21 were used for plasmid maintenance and protein expression, respectively.
480
481 Purification of DddA enzyme
482 The purification of the DddA toxin domain (DddAtox) protein has been reported previously
483 (Beverly et al. 2020, Nature). Briefly, to purify the his-tagged DddAtox in complex with DddAI,
484 *E. coli* BL21 (pETDuet-1::dddAtox + dddAI) was used to inoculate 2 L of LB broth in a 1:100
485 dilution and cultured overnight. After the culture reached an OD600 of approximately 0.6, 0.5 mM
486 IPTG was added, and the culture was incubated with shaking for 16 hours at 18˚C. Bacterial cell
487 pellets were collected by centrifugation at 4000 g for 30 minutes and then resuspended in 50 ml of
488 lysis buffer (50 mM Tris-HCl pH 8.0, 500 mM NaCl, 10 mM imidazole, 1 mg/mL lysozyme, and
489 a protease inhibitor cocktail). The cells were lysed by sonication (five pulses, 10 seconds each),
490 and the supernatant was separated from the debris by centrifugation at 25,000 g for 30 minutes.
491 The his-tagged DddAtox–DddAI complex was purified from the supernatant using a Nickel
492 column and eluted with elution buffer (50 mM Tris-HCl pH 7.5, 300 mM imidazole, 500 mM
493 NaCl, 30 mM imidazole, and 1 mM DTT). The eluted DddAtox–DddAI complex was denatured
494 by adding 50 ml of 8 M urea denaturing buffer (50 mM Tris-HCl pH 7.5, 300 mM NaCl, and 1
495 mM DTT) and incubated for 16 hours at 4˚C. The denatured proteins in the 8 M urea denaturing
496 buffer were reloaded onto the Nickel column, which was then washed with 50 ml of the same
497 buffer to remove any remaining DddAI. Sequential washes were performed with denaturing buffer
498 containing decreasing concentrations of urea (6 M, 4 M, 2 M, 1 M), followed by a final wash with
499 buffer without urea. DddAtox bound to the column was eluted with 5 ml of elution buffer. The
500 eluted DddAtox was subjected to size-exclusion chromatography using fast protein liquid
501 chromatography (FPLC) on a Superdex200 column (GE Healthcare) with sizing buffer (20 mM
502 Tris-HCl pH 7.5, 200 mM NaCl, 1 mM DTT, and 5% (w/v) glycerol). The purity of the eluted
503 DddAtox was assessed by SDS-PAGE gel stained with Coomassie blue, and the protein was stored
504 at -80˚C.
505
506 Purification of DddB enzyme
507 To purify the his-tagged DddB toxin domain (DddBtox) in complex with DddBI, *E. coli* BL21
508 (pETDuet-1::dddBtox + dddBI) was used to inoculate 2 L of LB broth in a 1:100 dilution and

509 cultured overnight. After the culture reached an OD600 of approximately 0.6, 0.5 mM IPTG was
510 added, and the culture was incubated with shaking for 16 hours at 18˚C. Bacterial cell pellets were
511 collected by centrifugation at 4000 g for 30 minutes and then resuspended in 50 ml of lysis buffer
512 (50 mM Tris-HCl pH 8.0, 500 mM NaCl, 10 mM imidazole, 1 mg/ml lysozyme, and a protease
513 inhibitor cocktail). The cells were lysed by sonication (five pulses, 10 seconds each), and the
514 supernatant was separated from the debris by centrifugation at 25,000 g for 30 minutes. The his-
515 tagged DddBtox–DddBI complex was purified from the supernatant using a Nickel column and
516 eluted with elution buffer (50 mM Tris-HCl pH 7.5, 300 mM NaCl, 30 mM imidazole, and 1 mM
517 DTT). The eluted DddBtox–DddBI complex was denatured by adding 50 ml of 8 M urea
518 denaturing buffer (50 mM Tris-HCl pH 7.5, 300 mM NaCl, and 1 mM DTT) and incubated for 16
519 hours at 4˚C. The denatured proteins in the 8 M urea denaturing buffer were reloaded onto the
520 Nickel column, which was then washed with 50 ml of the same buffer to remove any remaining
521 DddBI. Sequential washes were performed with denaturing buffer containing decreasing
522 concentrations of urea (6 M, 4 M, 2 M, 1 M), followed by a final wash with buffer without urea.
523 DddBtox bound to the column was eluted with 5 ml of elution buffer. The eluted DddBtox was
524 subjected to size-exclusion chromatography using fast protein liquid chromatography (FPLC) on
525 a Superdex200 column (GE Healthcare) with sizing buffer (20 mM Tris-HCl pH 7.5, 200 mM
526 NaCl, 1 mM DTT, and 5% (w/v) glycerol). The purity of the eluted DddBtox was assessed by
527 SDS-PAGE gel stained with Coomassie blue, and the protein was stored at -80˚C.
528
529 Crystallization and structure determination
530 Crystals of the DddB toxin-immunity protein complex were obtained at 5~12 mg/ml in
531 crystallization buffer (25 mM Tris pH 7.5, 200 mM NaCl, 1.0 mM DTT), mixed 1:1 with
532 crystallization solution containing 10% (w/v) PEG 8000, 0.1 M Potassium phosphate monobasic/
533 Sodium phosphate dibasic pH 6.2, 200 mM NaCl2. Rectangular crystals grew to 400 × 200 × 100
534 μm over 5 days. DddB toxin-immunity protein crystals displayed the symmetry of space group
535 C2221 (a = 84.1 Å, b = 130.6 Å, c = 61.6 Å, α = β = γ = 90°). Before data collection, crystals were
536 cryoprotected in a crystallization solution of 20% glycerol, 10% (w/v) PEG 8000, 0.1 M Potassium
537 phosphate monobasic/ Sodium phosphate dibasic pH 6.2, 200 mM NaCl2. The X-ray diffraction
538 data were collected using a Rigaku XTALAB SYNERGY CUSTOM diffractometer with a Cu
539 target and were processed with CrysAlis(Pro) software.

540 The DddB toxin-Immunity protein complex structure was solved by molecular replacement
541 using Phaser in Phenix (*41*), with the Alphafold2 predicted structure as the search model (*42*).
542 Phenix (Adams et al., 2010) and Coot (*43*) were used for the standard refinement. The structure
543 validation was completed with MolProbity (*44*). The data processing and refinement statistics are
544 summarized in Table S2.
545
546 DNA deamination activity assays
547 Lambda DNA or genomic DNA extracted from K562 or GM12878 cell lines was used to
548 evaluate the double-stranded DNA deamination activity. Reactions were performed in 10 µl of
549 deamination buffer consisting of 20 mM Tris-HCl pH 7.4, 100 mM NaCl, 1 mM DTT, 50ng DNA
550 substrate, and deaminase. Due to batch effects in enzyme purification, it is recommended to
551 perform titration tests to reach the saturation of deamination. Reactions were incubated for 1 hr or
552 the indicated time course at 37˚C, followed by DNA purification using Zymo DNA Clean &
553 Concentrator-5 kit. Purified DNA was subjected to library preparation using TruePrep DNA
554 Library Prep Kit V2 for Illumina (Vazyme) following the manual, except replacing the polymerase
555 mix with 1× Q5U PCR master mix plus Bst 3.0 polymerase (0.08 U per µl). The uracil conversion
556 ratio at each cytosine site was calculated and averaged to evaluate the deamination activity in each
557 possible sequence context.
558
559 Cell preparation
560 Human cell line K562 and GM12878 were grown with RPMI 1640 medium (cat. no.
561 11875093, Gibco) containing 10% FBS (cat. no. 10099141C, Gibco) in the 5% CO2 incubator at
562 37°C. If necessary, cells were cultured in the 5% CO2 incubator at 42°C for 30min for heat shock
563 treatment. When needed, 0.2 million cells were crosslinked in 0.2 ml of freshly prepared 1% PFA
564 fixation buffer at room temperature (RT) for 10 min. To quench the fixation, the cells were rinsed
565 once with 1 ml of DPBS (cat. no. 14040117, Thermofisher)-1% BSA (cat. no. A1933, Sigma), and
566 further incubated with DPBS-1% BSA for 5 min.
567
568 FOODIE for bulk samples
569 1. Nuclei preparation. To prepare nuclei, fresh or fixed cells were spun down at 450g for 5
570 min, followed by washing using an equal volume of cold 1× PBS and centrifugation at 500g for 5

571 min at 4 °C. A total of 30,000 cells were permeabilized using cold permeabilization buffer (10 mM
572 Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% IGEPAL CA-630, 0.1% Tween-20, 0.1%
573 digitonin). Immediately after permeabilization, nuclei were spun down at 550g for 5 min at 4 °C.
574 Supernatant was carefully removed from the pellet after centrifugation.

575 2. Tagmentation reaction. Prepare 2X TD buffer (20mM TAPS ph8.5, 10mM MgCl2, 20%
576 DMF). The nuclei pellet was immediately resuspended in the transposase reaction mix (25 µL 2×
577 TD buffer, 1 µL transposase (Vazyme, S111), and 24 µL PBS with 0.1% digitonin). The
578 transposition reaction was carried out for 30 min at 37 °C on a thermomixer at 800 rpm. For fixed
579 cells, the transposition reaction can be extended to 1.5 h. After that, 100 µl ice-cold RSB was added
580 to the mixer, which was followed by centrifugation at 550g for 5 min at 4 °C. Supernatant was
581 carefully removed from the pellet after centrifugation.

582 3. Deamination reaction. Prepare 2X DRB buffer (20 mM Tris-HCl pH 7.5, 20 mM NaCl, 2
583 mM DTT). The nuclei pellet was immediately resuspended in the deaminase reaction mix. For
584 DddAtox, the deaminase reaction mix contains 12 µl DRB and 18 µl DddA enzyme (50 uM, stored
585 in sizing buffer). For DddB, the deaminase reaction mix contains 5 µl DRB and 5 µl DddB enzyme
586 (50 uM, stored in sizing buffer). Due to batch effects in enzyme purification, it is recommended to
587 perform titration tests to achieve optimal deamination performance. The nuclei mix was incubated
588 at 37°C for 10min. For fixed cells, the deamination reaction can be extended to 45 min, followed
589 by a proteinase K overnight digestion for de-crosslinking. The DNA was purified using a Zymo
590 DNA Clean & Concentrator-5 kit.

591 4. Library amplification. After DNA purification, library fragments were amplified using 1×
592 Q5U PCR master mix, Bst 3.0 polymerase (0.08 U per µl) and 1.25 µM of Nextera PCR primers
593 (forward and reverse), using the following PCR conditions: 65 °C for 5 min; 80 °C for 12 min;
594 98 °C for 2 min; and thermocycling at 98 °C for 15 s, 60 °C for 30 s and 72 °C for 1 min. It is
595 recommended to monitor the PCR reaction using qPCR to stop amplification before saturation and
596 avoid GC and size bias in PCR. After a pre-amplification of full libraries for five cycles, an aliquot
597 of the PCR reaction was collected for qPCR. A total of 20 cycles of qPCR was performed to
598 determine the additional number of cycles needed for the remaining PCR reaction. Generally, a
599 total of 10–12 cycles of amplification yields high-quality libraries. The libraries were purified
600 using a Zymo Select-a-Size DNA Clean & Concentrator Kit to collect DNA with fragment size
601 above 200bp. Size-selected libraries are ready for sequencing.

602

603 FOODIE for single cells

604 Single-cell FOODIE in four cell lines was performed separately in the same experiment on the
605 same day. Cultured cells with more than 95% viability were collected freshly washed with ice-
606 cold PBS twice and counted with cell counters (Invitrogen). 50, 000 cells of each cell line were
607 used following the above bulk FOODIE procedure, except the incubation temperature was reduced
608 to 24°C. After the STOP step, the nuclei pellet was diluted into 1 mL ice-cold PBS. DAPI (Sigma)
609 was incubated in 1:1000 with the nuclei for 10 minutes in the dark on ice. Then the stained single
610 nuclei were sorted into single wells of a 96-well plate containing 2ul lysis buffer (30mM Tris-HCL
611 pH7.5, 20mM NaCl, 0.5% SDS, 1mM DTT, 0.4ug/ml QP) by FACS machine (BD Aria II). After
612 the lysis step at 50C for 30min, distinct cell barcodes were added to different wells through primers,
613 and a 10 µL volume of PCR reaction was conducted using 22 cycles. More cycles should be
614 considered when dealing with other cell types. 3µL of the PCR product of each well in one plate
615 was pooled together and purified using 1.1X Ampure beads twice for size selection.

616 The C57BL/6N mouse strain is utilized in the scFOODIE project. The procedure for isolating
617 mouse hippocampal nuclei follows that outlined in reference (*45*), while the beginning steps for
618 dissociated mouse brain cells were performed as previously described in bulk FOODIE. Post
619 deaminase incubation, add 0.2 µl of 1 mg/ml DAPI per 100 µl reaction for 5 minutes for staining
620 on ice. Subsequently, sorting single cell to a 96-well plate containing 4 µl lysis buffer (20 mM
621 Tris-HCl pH 7.5, 20 mM NaCl, 0.4% SDS) using flow cytometry. Incubate the 96-well plate at
622 65℃ for 15 minutes for nuclear lysis. Finally, add 8 µl of PCR mix (2×Q5U master mix 6µl, 10%
623 Tween20 2µl), and amplify as follows: 72 °C for 5 min; 98 °C for 1 min; and cycling for 22 cycles
624 at 98 °C for 10 s, 63 °C for 30 s and 72 °C for 1 min. After PCR, wells from the same plate are
625 pooled and purified using 1×Ampure beads for sequencing.

626

627 FOODIE data preprocessing

628 Raw paired-end reads were adapter-trimmed using Trim Galore (version 0.6.10) (*46*) and then
629 mapped to the hg38 genome and deduplicated using Bismark Bisulfite Mapper (version 0.24.2)
630 (*47*). We then used custom Python scripts to extract C-to-T conversion states for deduplicated
631 reads and generate a conversion ratio for every genome position. For correction of the binding
632 fraction, we measured deamination efficiency along the naked human genome as a positive control,

633 from which we obtained conversion probability for all 4^6 (-4 to +2 bases adjacent to cytosine)
634 possible sequences. For *in vivo* FOODIE data, conversion ratios of any cytosine sites were
635 corrected by normalization with the positive control.
636
637 Single-cell FOODIE data processing
638 For single-cell FOODIE data analysis, adapter sequences are first trimmed using Trimmomatic
639 (version 0.39) (*48*), followed by mapping to the mouse reference genome mm10 with Bismark
640 (version 0.24.0) (*47*). Duplicated reads are subsequently removed using the "deduplicate_bismark"
641 method within Bismark. For the single-cell data, the cytosines on the deduplicated reads are then
642 extracted for calculating the "non-conversion ratio" of detected genes. The resulting non-
643 conversion ratio matrix undergoes filtering for cell clustering analysis using Scanpy (*49*), requiring
644 a minimum of 600 genes per cell with a non-zero non-conversion ratio and each gene is detected
645 in at least 5 cells. Further analysis includes Principal Component Analysis using "scanpy.tl.pca,"
646 construction of a neighborhood graph with "scanpy.pp.neighbors," and dimension reduction via
647 "scanpy.tl.umap" for UMAP visualization. Subsequently, clusters are identified using the
648 "scanpy.tl.leiden" algorithm. Cluster-specific marker genes are analyzed using
649 "scanpy.tl.rank_genes_groups," with the top 20 ranked genes selected for each cluster for cell type
650 annotation. Cell type annotation is performed by matching the marker genes of each cluster against
651 the CellMarker 2.0 database (*50*). The cell type with the highest number of matched genes is
652 assigned to the respective cluster.
653
654 *De novo* footprinting
655 We used a custom-modified method to de novo call TF binding footprints (original code from
656 https://noble.gs.washington.edu/proj/footprinting). In brief, we used a set of sliding windows with
657 different sizes to determine footprint positions. Windows with a conversion ratio significantly
658 lower than flanking regions are considered footprints.
659
660 Assignment of TF recognition sequences to FOODIE footprints
661 To assign TF recognition sequences to FOODIE footprints, *de novo* called footprints were
662 overlapped with TF recognition sequences. TF recognition sequences were downloaded from
663 JASPAR (*23*) website (https://frigg.uio.no/JASPAR/JASPAR_TFBSs/2024). A motif that

overlaps with a footprint of more than 50% length of the motif was reported for further analysis. When needed, ChIP-seq data was downloaded from the ENCODE Project (*27*) for TF assignments, using a principle that TF recognition sequences fully covered by the corresponding ChIP-seq peaks were used for further analysis.

Statistic analyses

To investigate the potential cooperativity among transcription factors (TFs), we analyzed patterns of cytosine conversion at both sites. For each cytosine pair, we constructed 2x2 contingency tables to capture the bivariate frequency distribution of their conversion events. We then performed Fisher's exact test on each table, testing the null hypothesis that the conversion events between cytosines are independent. The type of cooperativity (positive or negative) was determined based on the Φ correlation coefficients derived from these tables. Specifically, a positive Φ coefficient indicates positive cooperativity, while a negative Φ coefficient suggests negative cooperativity. We retained cytosine pairs with significant *P*-values ($P < 0.001$) for constructing a pairwise correlation map. In these maps, individual TFs exhibited strong self-correlation, depicted by a densely populated triangular area with positive Φ values. The cooperativity between two TFs was represented by the cross-section of two individual binding sites, displaying either positive or negative patterns.

Given the binding sites of a TF, we can identify the most frequently occurring modules. Since some modules are larger than others, they tend to show up more frequently for any TF. Thus, it is valuable to identify the statistically significantly enriched modules. For a specific TF and a specific module, if all analyzed modules have N total genes, among which M are involved in the module, and the TF binds to the promoters (or enhancers) of n genes, among which m are involved in the same module, the *p* value is calculated as follows:

$$p = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Then we performed Benjamini-Hochberg FDR correction with default cutoff of 0.05.

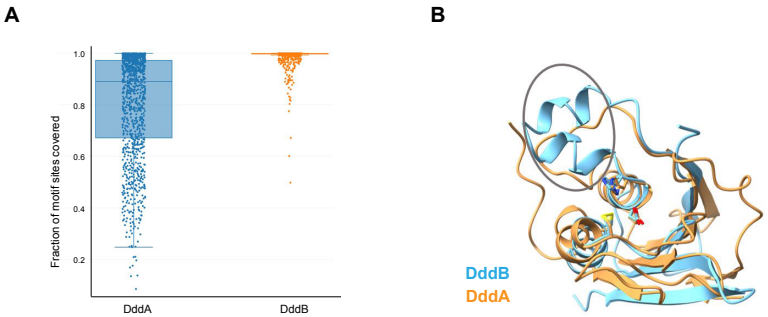
694 **References:**

695 1. S. A. Lambert *et al.*, The Human Transcription Factors. *Cell* **172**, 650–665 (2018).
696 2. F. Jacob, J. Monod, Genetic regulatory mechanisms in the synthesis of proteins. *Journal*
697 *of molecular biology* **3**, 318-356 (1961).
698 3. M. Ptashne, ISOLATION OF THE lambda PHAGE REPRESSOR. *Proc Natl Acad Sci U*
699 *S A* **57**, 306-313 (1967).
700 4. J. Elf, G.-W. Li, X. S. Xie, Probing transcription factor dynamics at the single-molecule
701 level in a living cell. *Science* **316**, 1191-1194 (2007).
702 5. J. Yu, J. Xiao, X. Ren, K. Lao, X. S. Xie, Probing gene expression in live cells, one
703 protein molecule at a time. *Science* **311**, 1600-1603 (2006).
704 6. P. J. Mitchell, R. Tjian, Transcriptional regulation in mammalian cells by sequence-
705 specific DNA binding proteins. *Science* **245**, 371-378 (1989).
706 7. J. C. M. Gebhardt *et al.*, Single-molecule imaging of transcription factor binding to DNA
707 in live mammalian cells. *Nature methods* **10**, 421-426 (2013).
708 8. C. Nüsslein-Volhard, E. Wieschaus, Mutations affecting segment number and polarity in
709 Drosophila. *Nature* **287**, 795-801 (1980).
710 9. A. R. Chapman *et al.*, Correlated gene modules uncovered by high-precision single-cell
711 transcriptomics. *Proceedings of the National Academy of Sciences* **119**, e2206938119
712 (2022).
713 10. M. Hu *et al.*, Correlated Protein Modules Revealing Functional Coordination of
714 Interacting Proteins Are Detected by Single-Cell Proteomics. *The Journal of Physical*
715 *Chemistry B* **127**, 6006-6014 (2023).
716 11. J. M. Replogle *et al.*, Mapping information-rich genotype-phenotype landscapes with
717 genome-scale Perturb-seq. *Cell* **185**, 2559-2575. e2528 (2022).
718 12. P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network
719 analysis. *BMC Bioinformatics* **9**, 559 (2008).
720 13. B. Ren *et al.*, Genome-Wide Location and Function of DNA Binding Proteins. *Science*
721 **290**, 2306-2309 (2000).
722 14. G. Robertson *et al.*, Genome-wide profiles of STAT1 DNA association using chromatin
723 immunoprecipitation and massively parallel sequencing. *Nature Methods* **4**, 651-657
724 (2007).
725 15. D. S. Johnson, A. Mortazavi, R. M. Myers, B. Wold, Genome-Wide Mapping of in Vivo
726 Protein-DNA Interactions. *Science* **316**, 1497-1502 (2007).
727 16. J. R. Hesselberth *et al.*, Global mapping of protein-DNA interactions in vivo by digital
728 genomic footprinting. *Nature Methods* **6**, 283-289 (2009).
729 17. C. Sönmezer *et al.*, Molecular Co-occupancy Identifies Transcription Factor Binding
730 Cooperativity In Vivo. *Molecular Cell*, (2020).
731 18. B. Y. Mok *et al.*, A bacterial cytidine deaminase toxin enables CRISPR-free
732 mitochondrial base editing. *Nature* **583**, 631-637 (2020).
733 19. L. M. Iyer, D. Zhang, I. B. Rogozin, L. Aravind, Evolution of the deaminase fold and
734 multiple origins of eukaryotic editing and mutagenic nucleic acid deaminases from
735 bacterial toxin systems. *Nucleic acids research* **39**, 9473-9497 (2011).
736 20. M. H. de Moraes *et al.*, An interbacterial DNA deaminase toxin directly mutagenizes
737 surviving target populations. *eLife* **10**, e62967 (2021).

21. 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).
22. 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).
23. 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).
24. 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).
25. A. Jolma *et al.*, Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research* **20**, 861-873 (2010).
26. S. Mukherjee *et al.*, Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature Genetics* **36**, 1331-1339 (2004).
27. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636-640 (2004).
28. M. J. MacPherson, P. D. Sadowski, The CTCF insulator protein forms an unusual DNA structure. *BMC Molecular Biology* **11**, 101 (2010).
29. 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).
30. 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).
31. D. M. Suter *et al.*, Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics. *Science* **332**, 472–474 (2011).
32. A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, S. Tyagi, Stochastic mRNA Synthesis in Mammalian Cells. *PLOS Biology* **4**, e309 (2006).
33. J. R. Chubb, T. Trcek, S. M. Shenoy, R. H. Singer, Transcriptional Pulsing of a Developmental Gene. *Current Biology* **16**, 1018–1025 (2006).
34. S. Chong, C. Chen, H. Ge, X. S. Xie, Mechanism of Transcriptional Bursting in Bacteria. *Cell* **158**, 314–326 (2014).
35. D. E. Nelson *et al.*, Oscillations in NF-kappaB signaling control the dynamics of gene expression. *Science* **306**, 704-708 (2004).
36. W. Ma, X. S. Xie, TriSHA: investigating modularity in high-precision single-cell transcriptomic data. *In preparation*, (2024).
37. 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).
38. 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).
39. 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).
40. K. Taniguchi, M. Karin, NF-κB, inflammation, immunity and cancer: coming of age. *Nature Reviews Immunology* **18**, 309-324 (2018).
41. A. J. McCoy *et al.*, Phaser crystallographic software. *J Appl Crystallogr* **40**, 658-674 (2007).
42. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

43. P. D. Adams *et al.*, PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213-221 (2010).
44. C. J. Williams *et al.*, MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science* **27**, 293-315 (2018).
45. L. Tan *et al.*, Changes in genome architecture and transcriptional dynamics progress independently of sensory experience during post-natal brain development. *Cell* **184**, 741-758 e717 (2021).
46. F. Krueger, T. Galore, A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. *Cambridge, UK: Babraham Institute*, (2015).
47. F. Krueger, S. R. Andrews, Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *bioinformatics* **27**, 1571-1572 (2011).
48. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
49. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19**, 15 (2018).
50. C. Hu *et al.*, CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res* **51**, D870-D876 (2023).

820 **Supplementary Materials**



821

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

825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841

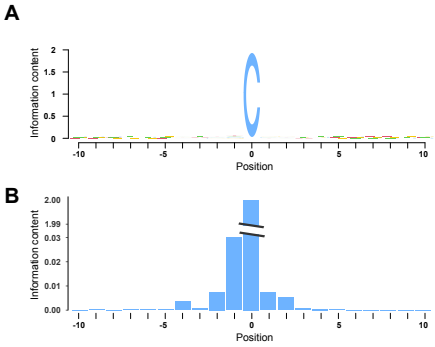


Figure S2. (A) Sequence logo of DddB conversion yields for a cytosine on naked DNA *in vitro*. (B) Bases from -4 to +2 positions of cytosine show relatively large bias on conversion.

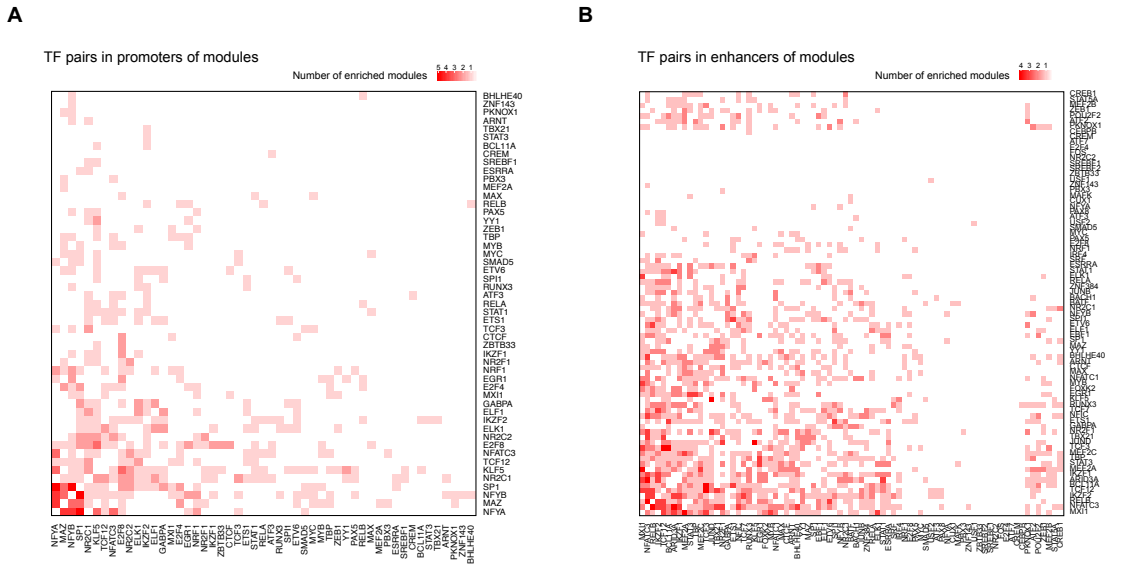


Figure S3. Enhancers harbor more enriched TF pairs than promoters. (A) Heatmap showing the number of modules, in which the TF pairs in a promoter are enriched. Color bar on the top denotes the number of enriched modules. (B) Heatmap showing the number of modules, in which the TF pairs in a enhancer are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