Chromatin information content landscapes inform transcription factor and DNA interactions

Ricardo D'Oliveira Albanus[1], Yasuhiro Kyono[1,2,3], John Hensley[1], Arushi Varshney[1,2], Peter Orchard[1], Jacob O. Kitzman[1,2], Stephen C. J. Parker[1,2]*

[1]Department of Computational Medicine & Bioinformatics, University of Michigan. Ann Arbor, USA

[2]Department of Human Genetics, University of Michigan. Ann Arbor, USA

[3]Current address: Tempus Labs, Inc. Chicago, IL, Chicago, USA

*Correspondence to: scjp@umich.edu

Abstract

Interactions between transcription factors (TFs) and chromatin are fundamental to genome organization and regulation and, ultimately, cell state. Here, we use information theory to measure signatures of TF-chromatin interactions encoded in the patterns of the accessible genome, which we call chromatin information enrichment (CIE). We calculate CIE for hundreds of TF motifs across human tissues and identify two classes: low and high CIE. The 10-20% of TF motifs with high CIE associate with higher protein-DNA residence time, including different binding sites subclasses of the same TF, increased nucleosome phasing, specific protein domains, and the genetic control of both gene expression and chromatin accessibility. These results show that variations in the information content of chromatin architecture reflect functional biological variation, with implications for cell state dynamics and memory.

Main text

Chromatin is the association between DNA, RNA, and diverse nuclear proteins, including nucleosomes. It enables the ~2-meter human genome to be packaged inside the nucleus while allowing active genes and their corresponding regulatory elements to remain accessible (*2*). Nucleosome positioning is an essential property of chromatin architecture and has been shown to have both passive and active roles in transcription factor (TF) binding (*3–5*). Understanding TF-chromatin interactions is therefore critical to dissect regulatory circuits that lead to differences in transcriptional activity across species, tissues, stimulatory, and genetic contexts. Information theory provides a powerful framework to quantify ordered patterns in data (*6*) and has been successfully used to characterize genome-wide DNA methylation patterns (*7*). Here, we hypothesized that local chromatin architecture encodes rich signatures of TF interactions and developed information-theoretical tools to measure these patterns in human tissues.

We first aimed to quantify patterns of chromatin accessibility around TF-chromatin interactions. We reasoned that TF binding creates a localized impact on chromatin architecture, which may result in TF-specific signatures. To measure chromatin architecture, we focused on the assay for transposase-accessible chromatin using sequencing (ATAC-seq) (*8*), that can simultaneously quantify both TF and nucleosome signatures, which are reflected in the ATAC-seq fragment length patterns. This chromatin architecture can be visualized using V-plots (*9*),

which show the aggregate ATAC-seq fragment midpoints around TF binding sites and can result in a stereotyped "V" pattern of points for bound TFs with well-phased adjacent nucleosomes (Fig. 1A, upper plot). The extent of organization in the V-plot can be measured using Shannon's entropy equations (*6*) to quantify information. We therefore calculated information content (*1*) of the ATAC-seq fragment size distribution around TF binding sites as a way to quantify V-plot organization (Fig. 1A, middle plot). To adjust for potential bias arising from non-uniform ATAC-seq fragment coverage across the V-plot, we devised a metric called chromatin information enrichment (CIE) (*1*) (Fig. 1A, middle and lower plots). We summarized CIE into a single value, named feature V-Plot Information Content Enrichment (f-VICE), which represents the CIE at landmark TF and nucleosomal positions across the V-plot (*1*), which are expected to have high CIE when the nucleosomes are phased around the TF binding site (Fig. 1A, lower plot). Therefore f-VICE quantifies the degree of chromatin architecture organization around a TF.

We initially focused on the GM12878 lymphoblastoid cell line, for which there is high-quality, deeply-sequenced ATAC-seq data (*8*) and 41 TF chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiments that pass our inclusion criteria (Table S1) (*1*, *10*). To increase our ability to detect TF-chromatin interactions, we generated an independent GM12878 ATAC-seq dataset with higher signal-to-noise ratio (Fig. S1). Using these datasets, we created V-plots and calculated f-VICEs centered on bound motif instances for 41 TFs. The ATAC-seq fragment pattern was most ordered around CTCF, a known chromatin organizer (*11*), where we detected clusters of fragments distributed periodically in a "V" pattern indicating nucleosome phasing (Fig. 1B-C, S2). CTCF f-VICE was highest among the 41 TFs (Fig 1D). Other TFs, exemplified by AP-1 and NFKB, had diverse f-VICEs (Fig. 1B-D, S2). These patterns were consistent across independent ATAC-seq libraries, indicating the robustness of the f-VICE metric (Fig. S2). These results indicate extensive differences in TF-chromatin interactions, which are captured in the CIE patterns.

One alternative to determine f-VICEs for TFs without ChIP-seq data is to rely on binding predictions using chromatin accessibility data. This motivated us to first evaluate the performance of current TF binding prediction algorithms. Most algorithms search for footprints, which are regions of low chromatin accessibility embedded within larger accessible regions, thought to be caused by cleavage protection from bound TFs (*12–14*). However, a recent report indicated that ~80% of TFs do not have footprints (*15*). Hence, we developed BMO, an unsupervised method to predict TF binding using negative binomial models of chromatin accessibility (*16–18*) and co-occurring motifs (*19*), without relying on footprints (*1*). We benchmarked BMO and other methods (*12–14*, *20*) using TF ChIP-seq data from GM12878 and HepG2 (Table S1). Overall, the footprint-agnostic methods (BMO, CENTIPEDE, and a custom implementation of CENTIPEDE, called ssCENTIPEDE (*1*, Supplementary Text) outperformed footprint-based methods on most (median of 81% across datasets) tested TFs, particularly on those with lower f-VICEs (Figs. 1E, S3-7; Supplementary Text). These findings indicate that TF binding is more accurately predicted using a simple chromatin accessibility model tuned to each TF motif.

Having determined that our BMO footprint-agnostic method is among the most accurate for predicting TF binding, we proceeded with predictions to estimate f-VICEs for TFs without ChIP-seq data. BMO-predicted f-VICEs were significantly correlated with f-VICEs calculated from TF ChIP-seq data across all datasets (Pearson's $\rho \geq 0.72$, $p \leq 1e\text{-}10$; Fig. S8). We therefore

concluded that BMO can be used to estimate f-VICEs without ChIP-seq data and performed BMO TF binding predictions to calculate f-VICEs for 540 non-redundant (*1*) TF motifs (Tables S2-3). We used high-quality ATAC-seq datasets from four additional human tissues (pancreatic islets (*21*), pancreatic islet sorted alpha and beta cells (*22*), and CD4+ cells (*23*); Table S1), selected by applying a strategy that uses the highly stereotyped chromatin architecture in ubiquitous and conserved CTCF/cohesin binding sites to measure sample quality (Fig. S9) (*1*). We normalized f-VICEs within each sample (*1*) to control for differences in bound motif predictions and overall chromatin accessibility (Fig. S10). Among the 540 motifs, we observed a mixture of two f-VICE distributions and therefore used a mixture of two gaussians to fit the data (*1*). The median percentage of motifs associated with high f-VICEs across datasets was 14% (Figs. 1G, S11), which is comparable to the percentage of motifs associated with DNase footprint protection across datasets (median=19%) from another study (*15*) and supports our conclusion that footprint-based algorithms will not perform well on the majority (median of 86% across datasets) of TFs. Together, these results reinforce the use of footprint-agnostic methods like BMO for accurately calculating f-VICE.

TF residence time, which corresponds to the duration of DNA binding for a TF, is an important biophysical measurement that can influence TF activity (*3*, *24*). Based on the high f-VICEs for CTCF and AP-1 and low f-VICE for NFKB (Fig. 1C-D), which agree with the known residence times for these TFs (Table S4), we hypothesized that CIE correlates with residence time. We correlated BMO-informed f-VICEs with previously measured fluorescence recovery after photobleaching (FRAP) data from mammalian cell lines (Table S4), which provide an upper bound of TF residence time (*25*, *26*). Using a robust linear regression to protect against outlier influence, we found that f-VICE was significantly associated with FRAP recovery times in all samples ($\beta \geq 0.7$, Bonferroni adjusted $p \leq 0.001$; Figs. 2A, S12). This suggests that TFs associated with high CIE have longer residence times.

A recent study found that cohesin has a residence time 10- to 20-fold higher than CTCF (*26*). We reasoned this difference could be reflected in the local chromatin architecture and calculated the CIE of the GM12878 lymphoblastoid cell line CTCF binding sites with and without the presence of cohesin (CTCF/cohesin$^+$ and CTCF/cohesin$^-$), controlling for potential confounding biases (Fig. S13A) (*1*). CTCF/cohesin$^+$ had 1.9-fold higher CIE compared to CTCF/cohesin$^-$ (Figs. 2B, S13B), indicating these distinct CTCF occupancy classes have different CIE signatures. We next compared the nucleosome positioning signals inferred from lymphoblastoid cell line micrococcal nuclease sequencing (MNase-seq) profiles (Table S1). Only the CTCF/cohesin$^+$ class had phased nucleosomes around the binding site (Figs. 2C, S13C), consistent with longer residence times associating with nucleosome phasing. To experimentally validate these results, we generated chromatin accessibility data using a modified ATAC-seq protocol with an additional sonication step (*1*) to disrupt the fragment size information (Fig. S14). There were no detectable nucleosome phasing patterns in the motif-flanking CIE signature of the sonicated sample (Fig. 2D; see vertical arrows). These results are complementary to our residence time results above in that they show our CIE approach can capture differences even when subsets of a single TF have different residence times.

To systematically characterize the association between CIE and nucleosome positioning, we compared GM12878 CIE patterns across TF motifs to lymphoblastoid MNase-seq profiles. First, we used *k*-means clustering (*1*) to divide motifs into broad CIE categories. We found three clusters representing a continuum of CIE at the motif region (Fig. S15A). Clusters one and two

had lower CIE at the motif compared to motif-flanking regions, while cluster three had the highest CIE at the motif (Fig. 2E, S15A-B) and encompassed >95% of the high f-VICE motifs (Fig. S15C-D). Notably, we observed two distinctly anti-correlated MNase-seq signal patterns for the motifs in cluster three (Fig. 2E). These two distinct patterns are consistent with TFs binding at the center of the nucleosome dyad or between phased nucleosomes (*5*, *27*). CIE and MNase signals were anti-correlated at high f-VICE motifs (Fig 2E; yellow-green heatmap), indicating that the highest CIE TFs associate with nucleosome phasing. We quantified nucleosome phasing (*1*) and found that it was significantly correlated with f-VICE in clusters two and three (Spearman's $\rho \geq 0.42$, $p \leq 1e\text{-}7$; Fig. S16). These results suggest that TF-chromatin interaction patterns are driven by TF residence time, resulting in distinct CIE signatures.

Previous reports suggested that a subset of TFs directionally bind DNA, with potential effects on gene regulation (*12*, *28*, *29*). To investigate this further, we extended our information content analyses to quantify CIE asymmetry (*1*). Of the 540 motifs tested, 150 had significantly asymmetric CIE (Bonferroni corrected $p<0.05$; Figs. 2F, S17A). The direction of CIE asymmetry was significantly correlated with the direction of the nearest TSS relative to each motif instance (Spearman's $\rho=0.66$, $p=2e\text{-}16$; Fig. S17B). To determine if asymmetric CIE was an artifact of TSS proximity, we calculated CIE asymmetry separately for TSS-proximal (≤1 kb) and TSS-distal (≥10 kb) motif instances. The TSS-distal and TSS-proximal CIE asymmetry directions agreed significantly more than expected by chance (111/150, binomial test $p=4e\text{-}9$; Fig. S17C-D), suggesting that CIE asymmetry is intrinsic to the TF motif. The magnitude of asymmetry was higher in TSS-proximal motifs (Fig. S17D), suggesting that TSS proximity amplifies TF CIE asymmetry. Accordingly, the correlation between nearest TSS direction and CIE asymmetry was stronger at TSS-proximal motifs (Spearman's $\rho=0.88$, $p=2e\text{-}16$; Fig. 2G). These results indicate that directional binding is an intrinsic property of TF-chromatin interactions.

We next aimed to investigate cross-tissue differences in CIEs. We performed an unsupervised hierarchical clustering of motif f-VICEs and found that it recapitulated the expected tissue grouping (Fig. 3A). A recent study demonstrated that NF-KB (p65) residence time is determined by its DNA-binding domain (DBD) (*30*), which motivated us to ask if DBDs are associated with CIE. We assigned DBDs and protein domains to motifs and designed a permutation-based rank test to calculate domain f-VICE enrichments (*1*). We observed both common and tissue-specific f-VICE enrichments, including IRF and ETS in blood-related samples, PAX in islet-related samples, and HMG/SOX and FOX domains in HepG2 (FDR < 10%; Figs. 3B, S18). Our findings show the landscape of TF-chromatin interactions varies across tissues and reflects protein domain-level TF properties.

The prevalence of tissue-specific differences in CIEs led us to examine the role of high f-VICE TFs in regulating gene expression. We calculated the enrichment of the motifs categorized as high or low f-VICE in GM12878 (Fig. 1F) to overlap lymphoblastoid cis-expression quantitative trait loci (cis-eQTLs) datasets (*31*, *32*), which represent gene expression genetic control regions. High f-VICE motifs had 15-30% higher (median=24%) fold-enrichment in cis-eQTLs compared to low f-VICE motifs (Figs. 3C, S19A), but no differences in eQTL effect sizes (Fig. S19B). These results indicate that high f-VICE TFs are more likely to mediate genetic effects on gene expression, but not their magnitude.

Given that high f-VICE TFs have highly ordered chromatin (Fig. 1), high predicted residence times (Fig. 2A, B, S12), and nucleosome phasing properties (Fig. 2E, S16) , we hypothesized that their regulatory effects (Fig. 3C) could result from acting as or recruiting

pioneer factors that induce chromatin accessibility (*12*, *33*). If true, we would expect increased CIE for single nucleotide polymorphism (SNP) alleles with increased chromatin accessibility (*i.e.* with ATAC-seq allelic imbalance; Fig. 3D). We performed a motif-agnostic approach to calculate the f-VICEs associated with every DNA 6-mer, controlling for differences in chromatin accessibility (*1*). This strategy allows the interrogation of genetic variants by determining the DNA 6-mers formed by each allele and their corresponding f-VICEs. DNA 6-mers have a distribution of f-VICEs (Figs. 3E; S20A) and GC-pure 6-mers had the highest f-VICEs (Fig. S20B), which is consistent with GC-rich sequences driving enhancer activity (*34*) and suggest that high GC-content regions represent anchors of nuclear architecture. Notably, a single base-pair change can lead to large differences in 6-mer f-VICEs (Fig. 3E-F, S20C-E), suggesting that genetic variation impacts CIE. To test this, we determined f-VICEs for 6-mers formed by both alleles at SNPs with significant ATAC-seq allelic imbalance (binomial test $p<0.05$) in GM12878 and pancreatic islets (*1*). The preferred ATAC-seq alleles were significantly biased to form higher f-VICE 6-mers compared to the less favored allele in all samples (permutation test $p<3e-4$; Fig. 3G-H, S21). These findings support a model where TFs with potential pioneer-like properties bookmark regions of the genome to allow binding of other migrant-like TFs (*12*, *33*). Notably, TF motifs that are predictive of binding without any chromatin accessibility data (based solely on the motif match score) have significantly higher f-VICEs in GM12878 and HepG2 (robust linear regression $p\leq0.001$; Fig. S22). This suggests that high f-VICE TFs, particularly CTCF, are more likely to bind any strong motif regardless of its underlying accessibility, while the remaining TFs require motifs located in already accessible regions. Collectively, our results show that application of information theory methods to chromatin profiles captures a dynamic landscape of TF-chromatin interactions, with implications for cell state memory and gene regulation.

References:

1. Materials and methods are available as supplementary materials
2. E. Segal, J. Widom, What controls nucleosome positions? *Trends Genet.* **25**, 335–343 (2009).
3. C. R. Lickwar, F. Mueller, S. E. Hanlon, J. G. McNally, J. D. Lieb, Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature*. **484**, 251–255 (2012).
4. S. Rudnizky, H. Khamis, O. Malik, P. Melamed, A. Kaplan, The base pair-scale diffusion of nucleosomes modulates binding of transcription factors. *Proc. Natl. Acad. Sci.*, 201815424 (2019).
5. F. Zhu, L. Farnung, E. Kaasinen, B. Sahu, Y. Yin, B. Wei, S. O. Dodonova, K. R. Nitta, E. Morgunova, M. Taipale, P. Cramer, J. Taipale, The interaction landscape between transcription factors and the nucleosome. *Nature*. **562**, 76–81 (2018).
6. C. E. Shannon, A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
7. G. Jenkinson, E. Pujadas, J. Goutsias, A. P. Feinberg, Potential energy landscapes identify the information-theoretic nature of the epigenome. *Nat. Genet.* **49**, 719–729 (2017).

8. 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. *Nat. Methods*. **10**, 1213–1218 (2013).

9. J. G. Henikoff, J. A. Belsky, K. Krassovsky, D. M. MacAlpine, S. Henikoff, Epigenome characterization at single base-pair resolution. *Proc. Natl. Acad. Sci.* **108**, 18318–18323 (2011).

10. The ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature*. **489**, 57–74 (2012).

11. Y. Fu, M. Sinha, C. L. Peterson, Z. Weng, The Insulator Binding Protein CTCF Positions 20 Nucleosomes around Its Binding Sites across the Human Genome. *PLoS Genet.* **4**, e1000138 (2008).

12. R. I. Sherwood, T. Hashimoto, C. W. O'Donnell, S. Lewis, A. A. Barkal, J. P. Van Hoff, V. Karun, T. Jaakkola, D. K. Gifford, Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.* **32**, 171–178 (2014).

13. M. H. Sung, M. J. Guertin, S. Baek, G. L. Hager, DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Mol. Cell*. **56**, 275–285 (2014).

14. Z. Li, M. H. Schulz, T. Look, M. Begemann, M. Zenke, I. G. Costa, Identification of transcription factor binding sites using ATAC-seq. *Genome Biol.* **20**, 45 (2019).

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

16. H. H. He, C. A. Meyer, S. S. Hu, M. W. Chen, C. Zang, Y. Liu, P. K. Rao, T. Fei, H. Xu, H. Long, X. S. Liu, M. Brown, Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat. Methods*. **11**, 73–78 (2014).

17. G. Cuellar-Partida, F. A. Buske, R. C. McLeay, T. Whitington, W. S. Noble, T. L. Bailey, Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics*. **28**, 56–62 (2012).

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

19. J. Wang, J. Zhuang, S. Iyer, X. Lin, T. W. Whitfield, M. C. Greven, B. G. Pierce, X. Dong, A. Kundaje, Y. Cheng, O. J. Rando, E. Birney, R. M. Myers, W. S. Noble, M. Snyder, Z. Weng, Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **22**, 1798–1812 (2012).

20. R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, J. K. Pritchard, Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21**, 447–455 (2011).

21. A. Varshney, L. J. Scott, R. P. Welch, M. R. Erdos, P. S. Chines, N. Narisu, R. D. Albanus, P. Orchard, B. N. Wolford, R. Kursawe, S. Vadlamudi, M. E. Cannon, J. P. Didion, J. Hensley, A. Kirilusha, L. L. Bonnycastle, D. L. Taylor, R. Watanabe, K. L.

Mohlke, M. Boehnke, F. S. Collins, S. C. J. Parker, M. L. Stitzel, Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proc. Natl. Acad. Sci.* **114**, 2301–2306 (2017).

22. A. M. Ackermann, Z. Wang, J. Schug, A. Naji, K. H. Kaestner, Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. *Mol. Metab.* **5**, 233–244 (2016).

23. M. R. Corces, A. E. Trevino, E. G. Hamilton, P. G. Greenside, N. A. Sinnott-Armstrong, S. Vesuna, A. T. Satpathy, A. J. Rubin, K. S. Montine, B. Wu, A. Kathiria, S. W. Cho, M. R. Mumbach, A. C. Carter, M. Kasowski, L. A. Orloff, V. I. Risca, A. Kundaje, P. A. Khavari, T. J. Montine, W. J. Greenleaf, H. Y. Chang, An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods*. **14**, 959–962 (2017).

24. A. Loffreda, E. Jacchetti, S. Antunes, P. Rainone, T. Daniele, T. Morisaki, M. E. Bianchi, C. Tacchetti, D. Mazza, Live-cell p53 single-molecule binding is modulated by C-terminal acetylation and correlates with transcriptional activity. *Nat. Commun.* **8** (2017), doi:10.1038/s41467-017-00398-7.

25. F. Mueller, D. Mazza, T. J. Stasevich, J. G. McNally, FRAP and kinetic modeling in the analysis of nuclear protein dynamics: What do we really know? *Curr. Opin. Cell Biol.* **22**, 403–411 (2010).

26. A. S. Hansen, I. Pustova, C. Cattoglio, R. Tjian, X. Darzacq, CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *eLife*. **6**, 1–33 (2017).

27. S. Li, E. B. Zheng, L. Zhao, S. Liu, Nonreciprocal and Conditional Cooperativity Directs the Pioneer Activity of Pluripotency Transcription Factors. *Cell Rep.* **28**, 2689-2703.e4 (2019).

28. A. Kundaje, S. Kyriazopoulou-Panagiotopoulou, M. Libbrecht, C. L. Smith, D. Raha, E. E. Winters, S. M. Johnson, M. Snyder, S. Batzoglou, A. Sidow, Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Res.* **22**, 1735–1747 (2012).

29. S. R. Grossman, J. Engreitz, J. P. Ray, T. H. Nguyen, N. Hacohen, E. S. Lander, Positional specificity of different transcription factor classes within enhancers. *Proc. Natl. Acad. Sci.*, 201804663 (2018).

30. A. Callegari, C. Sieben, A. Benke, D. M. Suter, B. Fierz, D. Mazza, S. Manley, Single-molecule dynamics and genome-wide transcriptomics reveal that NF-kB (p65)-DNA binding times can be decoupled from transcriptional activation. *PLOS Genet.* **15**, e1007891 (2019).

31. GTEx Consortium, Genetic effects on gene expression across human tissues. *Nature*. **550**, 204–213 (2017).

32. T. Lappalainen, M. Sammeth, M. R. Friedländer, P. A. C. ‘t Hoen, J. Monlong, M. A. Rivas, M. Gonzàlez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlöf, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. J. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S.

Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, The Geuvadis Consortium, H. Lehrach, S. Schreiber, R. Sudbrak, Á. Carracedo, S. E. Antonarakis, R. Häsler, A.-C. Syvänen, G.-J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigó, I. G. Gut, X. Estivill, E. T. Dermitzakis, Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*. **501**, 506–511 (2013).

33. K. S. Zaret, J. S. Carroll, Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.* **25**, 2227–2241 (2011).

34. J. O. Yáñez-Cuna, C. D. Arnold, G. Stampfel, L.M. Bory, D. Gerlach, M. Rath, A. Stark, Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. *Genome Res.* **24**, 1147–1156 (2014).

35. J. D. Buenrostro, B. Wu, H. Y. Chang, W. J. Greenleaf, *Curr. Protoc. Mol. Biol.*, in press, doi:10.1002/0471142727.mb2129s109.

36. S. Picelli, A. K. Björklund, B. Reinius, S. Sagasser, G. Winberg, R. Sandberg, Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).

37. N. Rohland, D. Reich, Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939–946 (2012).

38. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. **25**, 1754–1760 (2009).

39. L. J. Scott, M. R. Erdos, J. R. Huyghe, R. P. Welch, A. T. Beck, B. N. Wolford, P. S. Chines, J. P. Didion, N. Narisu, H. M. Stringham, D. L. Taylor, A. U. Jackson, S. Vadlamudi, L. L. Bonnycastle, L. Kinnunen, J. Saramies, J. Sundvall, R. D. Albanus, A. Kiseleva, J. Hensley, G. E. Crawford, H. Jiang, X. Wen, R. M. Watanabe, T. A. Lakka, K. L. Mohlke, M. Laakso, J. Tuomilehto, H. A. Koistinen, M. Boehnke, F. S. Collins, S. C. J. Parker, The genetic regulatory signature of type 2 diabetes in human skeletal muscle. *Nat. Commun.* **7**, 11764 (2016).

40. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079 (2009).

41. Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nussbaum, R. M. Myers, M. Brown, W. Li, X. S. Liu, Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9** (2008), doi:10.1186/gb-2008-9-9-r137.

42. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma. Oxf. Engl.* **26**, 841–842 (2010).

43. J. Köster, S. Rahmann, Snakemake--a scalable bioinformatics workflow engine. *Bioinforma. Oxf. Engl.* **28**, 2520–2522 (2012).

44. 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, A global reference for human genetic variation. *Nature*. **526**, 68–74 (2015).

45. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: scanning for occurrences of a given motif. *Bioinforma. Oxf. Engl.* **27**, 1017–1018 (2011).

46. J. A. Castro-Mondragon, S. Jaeger, D. Thieffry, M. Thomas-Chollier, J. Van Helden, RSAT matrix-clustering: Dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Res.* **45**, 1–13 (2017).

47. J. Hausser, K. Strimmer, Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks. *J. Mach. Learn. Res.* **10**, 1469–1484.

48. A. Chesi, Y. Wagley, M. E. Johnson, E. Manduchi, C. Su, S. Lu, M. E. Leonard, K. M. Hodge, J. A. Pippin, K. D. Hankenson, A. D. Wells, S. F. A. Grant, Genome-scale Capture C promoter interactions implicate effector genes at GWAS loci for bone mineral density. *Nat. Commun.* **10**, 1260 (2019).

49. O. Denas, R. Sandstrom, Y. Cheng, K. Beal, J. Herrero, R. C. Hardison, J. Taylor, Genome-wide comparative analysis reveals human-mouse regulatory landscape and evolution. *BMC Genomics*. **16**, 87 (2015).

50. T. Liptak, On the combination of independent tests. *Magy. Tud Akad Mat Kut. Int Kozl*. **3**, 171–197 (1958).

51. Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependence. *Ann. Stat.* **29**, 1165–1188 (2001).

52. T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*. **10**, 1–21 (2015).

53. J. Davis, M. Goadrich, The relationship between Precision-Recall and ROC curves. *Proc. 23rd Int. Conf. Mach. Learn. - ICML 06*, 233–240 (2006).

54. T. Sing, O. Sander, N. Beerenwinkel, T. Lengauer, ROCR: visualizing classifier performance in R. *Bioinformatics*. **21**, 3940–3941 (2005).

55. J. Grau, I. Grosse, J. Keilwagen, PRROC: Computing and visualizing Precision-recall and receiver operating characteristic curves in R. *Bioinformatics*. **31**, 2595–2597 (2015).

56. T. Benaglia, D. Chauveau, D. R. Hunter, D. S. Young, mixtools: An R Package for Analyzing Mixture Models. *J. Stat. Softw.* **32**, 1–29 (2009).

57. W. N. Venables, B. D. Ripley, *Modern Applied Statistics with S* (Springer-Verlag, New York, ed. 4, 2002; https://www.springer.com/gp/book/9780387954578), *Statistics and Computing, Statistics,Computing Venables,W.N.:Statistics w.S-PLUS*.

58. D. J. Gaffney, G. McVicker, A. A. Pai, Y. N. Fondufe-Mittendorf, N. Lewellen, K. Michelini, J. Widom, Y. Gilad, J. K. Pritchard, Controls of Nucleosome Positioning in the Human Genome. *PLOS Genet.* **8**, e1003036 (2012).

59. M. Garieri, O. Delaneau, F. Santoni, R. J. Fish, D. Mull, P. Carninci, E. T. Dermitzakis, S. E. Antonarakis, A. Fort, The effect of genetic variation on promoter usage and enhancer activity. *Nat. Commun.* **8**, 1–9 (2017).

60. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2013).

61. M. C. Frith, E. Valen, A. Krogh, Y. Hayashizaki, P. Carninci, A. Sandelin, A code for transcription initiation in mammalian genomes. *Genome Res.* **18**, 1–12 (2008).

62. S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, M. T. Weirauch, The Human Transcription Factors. *Cell*. **172**, 650–665 (2018).

63. S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, W. S. Noble, Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).

64. A. L. Mitchell, T. K. Attwood, P. C. Babbitt, M. Blum, P. Bork, A. Bridge, S. D. Brown, H.-Y. Chang, S. El-Gebali, M. I. Fraser, J. Gough, D. R. Haft, H. Huang, I. Letunic, R. Lopez, A. Luciani, F. Madeira, A. Marchler-Bauer, H. Mi, D. A. Natale, M. Necci, G. Nuka, C. Orengo, A. P. Pandurangan, T. Paysan-Lafosse, S. Pesseat, S. C. Potter, M. A. Qureshi, N. D. Rawlings, N. Redaschi, L. J. Richardson, C. Rivoire, G. A. Salazar, A. Sangrador-Vegas, C. J. A. Sigrist, I. Sillitoe, G. G. Sutton, N. Thanki, P. D. Thomas, S. C. E. Tosatto, S.-Y. Yong, R. D. Finn, InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* (2018), doi:10.1093/nar/gky1100.

65. M. T. Weirauch, A. Yang, M. Albu, A. G. Cote, A. Montenegro-Montero, P. Drewe, H. S. Najafabadi, S. A. Lambert, I. Mann, K. Cook, H. Zheng, A. Goity, H. van Bakel, J.-C. Lozano, M. Galli, M. G. Lewsey, E. Huang, T. Mukherjee, X. Chen, J. S. Reece-Hoyes, S. Govindarajan, G. Shaulsky, A. J. M. Walhout, F.-Y. Bouget, G. Ratsch, L. F. Larrondo, J. R. Ecker, T. R. Hughes, Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*. **158**, 1431–1443 (2014).

66. E. M. Schmidt, J. Zhang, W. Zhou, J. Chen, K. L. Mohlke, Y. E. Chen, C. J. Willer, GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics*. **31**, 2601–2606 (2015).

67. O. Delaneau, H. Ongen, A. A. Brown, A. Fort, N. I. Panousis, E. T. Dermitzakis, A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).

68. B. van de Geijn, G. McVicker, Y. Gilad, J. K. Pritchard, WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods*. **12**, 1061–1063 (2015).

69. R. D. Phair, P. Scaffidi, C. Elbi, J. Vecerova, A. Dey, K. Ozato, D. T. Brown, G. Hager, M. Bustin, T. Misteli, Global Nature of Dynamic Protein-Chromatin Interactions In Vivo: Three-Dimensional Genome Scanning and Dynamic Interaction Networks of Chromatin Proteins. *Mol. Cell. Biol.* **24**, 6393–6402 (2004).

70. C. E. Malnou, F. Brockly, C. Favard, G. Moquet-Torcy, M. Piechaczyk, I. Jariel-Encontre, Heterodimerization with different jun proteins controls c-Fos intranuclear dynamics and distribution. *J. Biol. Chem.* **285**, 6552–6562 (2010).

71. B. M. Mayr, E. Guzman, M. Montminy, Glutamine rich and basic region/leucine zipper (bZIP) domains stabilize cAMP-response element-binding protein (CREB) binding to chromatin. *J. Biol. Chem.* **280**, 15103–15110 (2005).

72. H. Nakahashi, K. R. K. Kwon, W. Resch, L. Vian, M. Dose, D. Stavreva, O. Hakim, N. Pruett, S. Nelson, A. Yamane, J. Qian, W. Dubois, S. Welsh, R. D. Phair, B. F. Pugh, V. Lobanenkov, G. L. Hager, R. Casellas, A Genome-wide Map of CTCF Multivalency Redefines the CTCF Code. *Cell Rep.* **3**, 1678–1689 (2013).

73. T. Sekiya, U. M. Muthurajan, K. Luger, A. V. Tulin, K. S. Zaret, Nucleosome-binding affinity as a primary determinant of the nuclear mobility of the pioneer transcription factor FoxA. *Genes Dev.* **23**, 804–809 (2009).

74. D. Bosisio, I. Marazzi, A. Agresti, N. Shimizu, M. E. Bianchi, G. Natoli, A hyper-dynamic equilibrium between promoter-bound and nucleoplasmic dimers controls NF-κB-dependent gene activity. *EMBO J.* **25**, 798–810 (2006).

75. F. L. Groeneweg, M. E. Van Royen, S. Fenz, V. I. P. Keizer, B. Geverts, J. Prins, E. R. De Kloet, A. B. Houtsmuller, T. S. Schmidt, M. J. M. Schaaf, Quantitation of glucocorticoid receptor DNA-binding dynamics by single-molecule microscopy and FRAP. *PLoS ONE*. **9**, 1–12 (2014).

76. M. Tirard, O. F. X. Almeida, P. Hutzler, F. Melchior, T. M. Michaelidis, Sumoylation and proteasomal activity determine the transactivation properties of the mineralocorticoid receptor. *Mol. Cell. Endocrinol.* **268**, 20–29 (2007).

77. P. Hinow, C. E. Rogers, C. E. Barbieri, J. A. Pietenpol, A. K. Kenworthy, E. DiBenedetto, The DNA binding activity of p53 displays reaction-diffusion kinetics. *Biophys. J.* **91**, 330–342 (2006).

Acknowledgments: We thank members of the Parker Lab, L. J. Scott, P. Freddolino, P. Wittkopp, M. Burmeister, G. Higgins, P. Pereira, M. Puthenveedu, and J. Brancho for helpful comments. Sequencing was performed at the UM Sequencing Core Facility. **Funding:** This work was supported by the ADA Pathway to Stop Diabetes Grant 1-14-INI-07 and by the National Institute of Diabetes and Digestive and Kidney Diseases grant R01 DK117960 to SCJP. **Author contributions:** ROA: Analyzed data, designed computational experiments, wrote the manuscript. YK: Generated ATAC-seq datasets. JH: Implemented computational algorithms. AV: analyzed eQTL data. PO: calculated ATAC-seq allelic imbalance. JK: designed ATAC-seq experiments. SCJP: designed experiments, analyzed data, wrote the manuscript, and supervised all aspects of the project. **Competing interests:** Authors declare none. **Data and materials availability:** Code and scripts are available (github.com/ParkerLab/chromatin_information). ATAC-seq data is deposited in GEO (GSE135074).

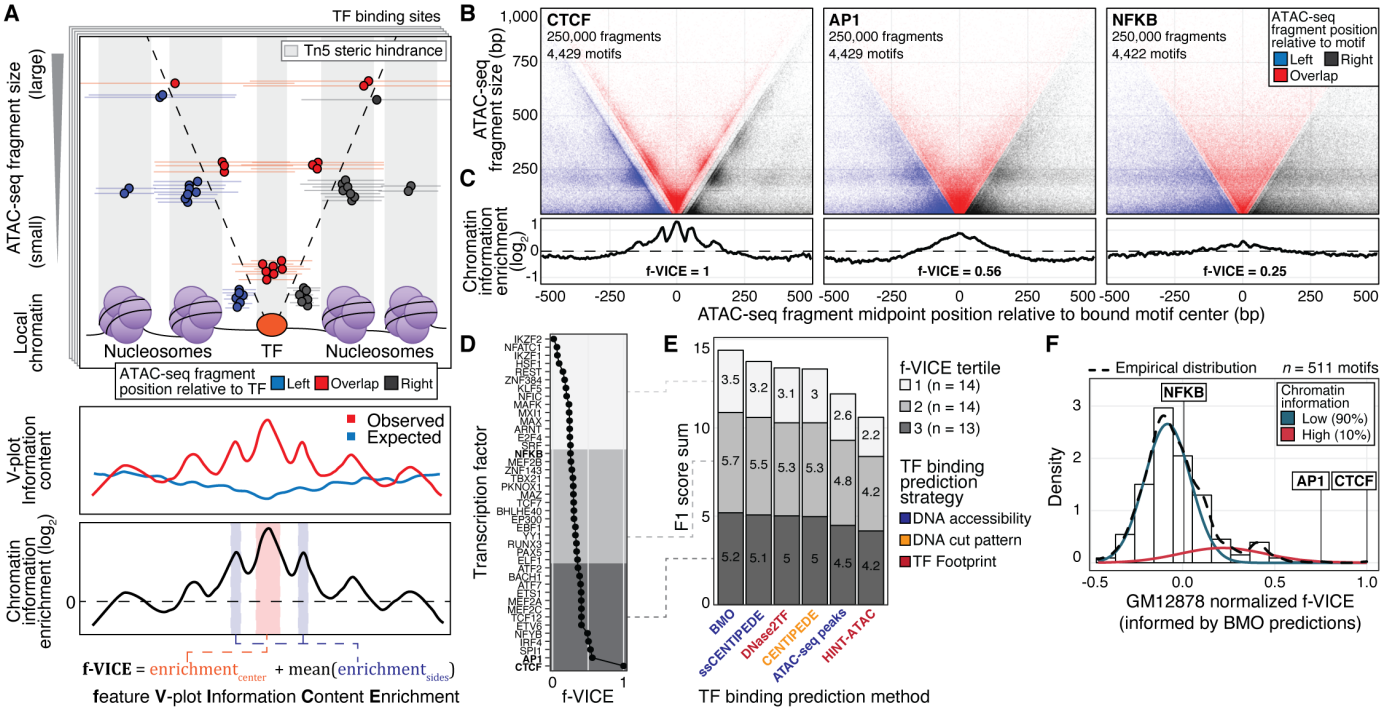

Fig. 1. Information content of TF-chromatin interactions. (**A**) Upper: TF binding impacts the chromatin architecture and the observed ATAC-seq fragment distribution around TF binding sites. Middle and bottom: calculation of CIE and f-VICE. (**B-C**) V-plots and CIEs of CTCF, AP-1, and NFKB (GM12878 ATAC-seq data generated in this study). V-plots were downsampled to highlight differences in chromatin architecture (but not for f-VICE calculation). (**D**) f-VICEs calculated for TFs with GM12878 ChIP-seq data. (**E**) F1 score sum of TF binding prediction algorithms. (**F**) Normalized GM12878 BMO-informed f-VICE distribution.

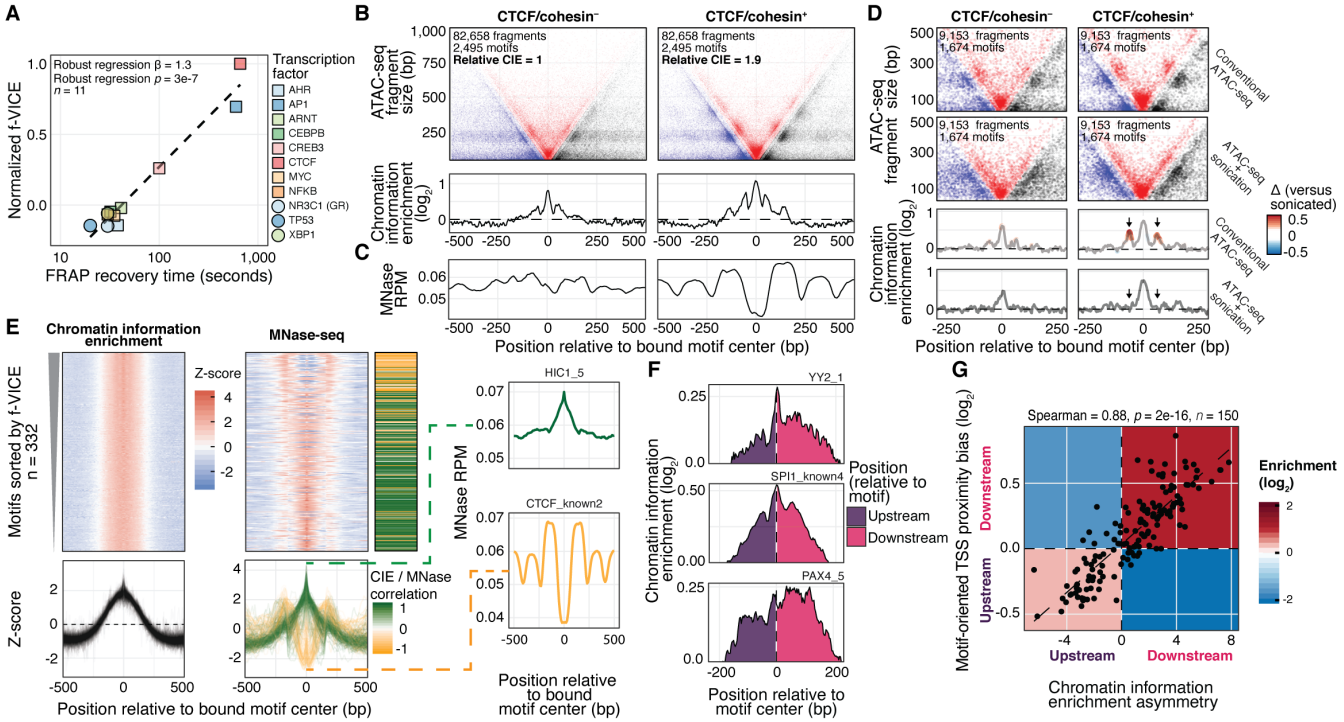

Fig. 2. Chromatin information informs residence times and TF-nucleosome interactions. (**A**) Correlation of FRAP recovery times and GM12878 f-VICEs. Dashed line, linear model fit. (**B**) V-plots and CIEs of CTCF/cohesin$^+$ and CTCF/cohesin$^-$ motifs. (**C**) GM19238 MNase-seq reads per million mapped reads at the same motifs. (**D**) CTCF/cohesin$^+$ and CTCF/cohesin$^-$ motifs in the sonicated and conventional GM12878 ATAC-seq data. Colors, differences relative to sonicated. (**E**) Left: CIE and MNase-seq profiles (*k*-means cluster three). Middle: Heatmap of MNase and CIE Z-score correlations. Right: Example motifs with positive and negative CIE/MNase correlation. (**F**) Top 3 motifs with CIE asymmetry Z-scores in GM12878. (**G**) Scatter plot of motif-oriented TSS position bias and CIE asymmetry in TSS-proximal motifs. Enrichments calculated by permuting the signs of observed values (n=10,000).

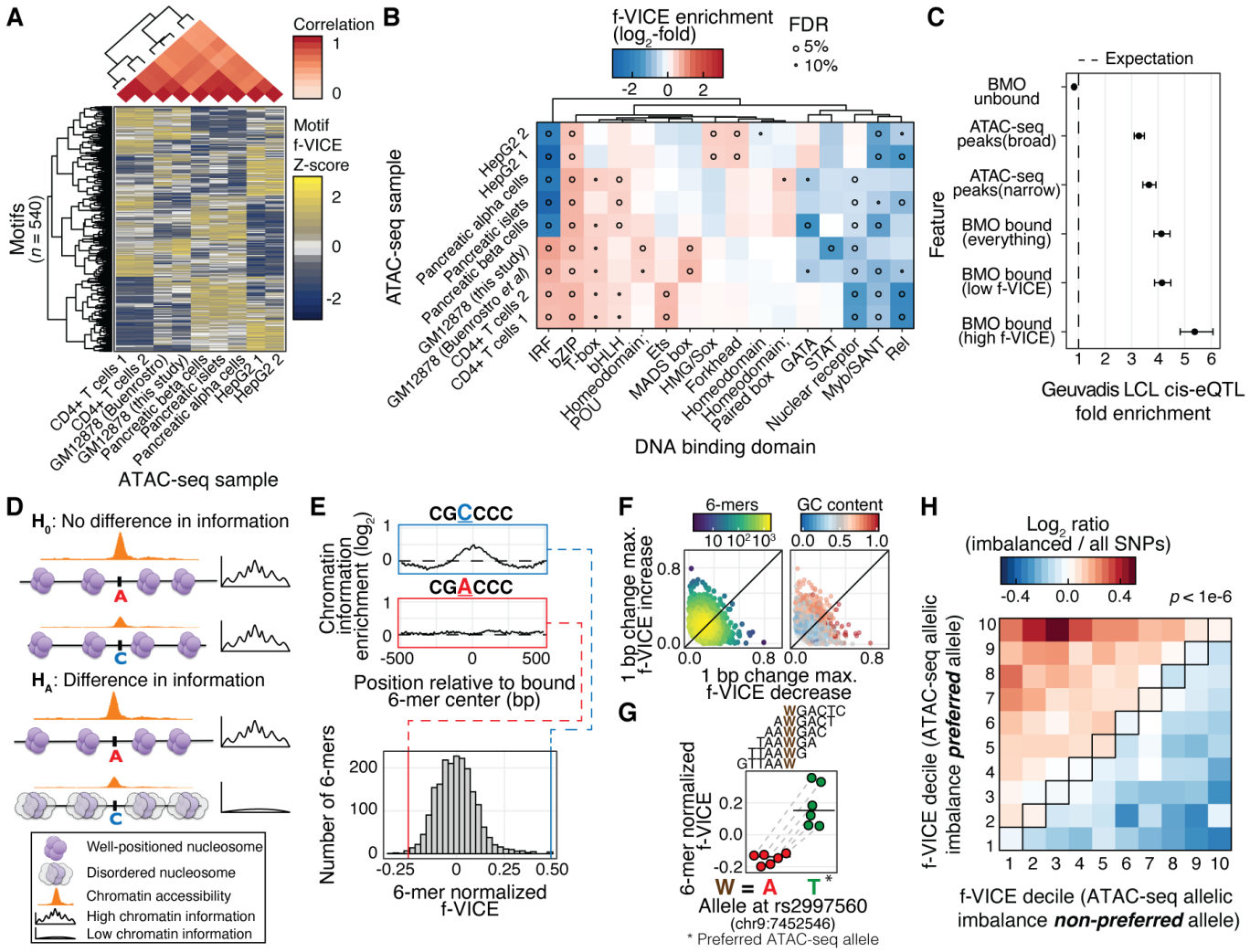

Fig. 3. The chromatin information landscape of human tissues. (**A**) Hierarchical clustering of f-VICE Z-scores. (**B**) f-VICE enrichments across DBDs. (**C**) LCL cis-eQTLs enrichments. Error bars, effect size SD. (**D**) Hypothesis schematic. (**E**) Upper: Two 6-mers with 1bp difference in sequence. Lower: pancreatic islets 6-mer normalized f-VICE distribution. (**F**) Range of f-VICE differences associated with 1-bp difference in 6-mer sequence. (**G**) Predicted f-VICE change associated with rs2997560 in pancreatic islets. Horizontal bars, median. (**H**) Log_2 ratio of f-VICE decile changes associated with the preferred and non-preferred alleles of imbalanced SNPs versus all tested SNPs in pancreatic islets.