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: Institute for Genomics and Systems Biology, University of Chicago, Chicago, USA

Corresponding Author:

Stephen C. J. Parker

Abstract:

The dynamic interactions between transcription factors (TFs) and chromatin are a fundamental component of genome regulation. Their biological principles, however, remain incompletely understood. Here, we develop information theory algorithms to measure signatures of TF-chromatin interactions encoded in patterns of the accessible genome, which we call chromatin information enrichment. We calculate chromatin information enrichment for hundreds of TF motifs across human tissues and find significant associations with TF-DNA residence times and specific DNA binding domains. Most TF motifs have low chromatin information enrichment and have lower performance with footprinting methods, but the subset of TF motifs associated with high chromatin information enrichment are more highly enriched to overlap genetic variants linked to gene expression and associate with increased chromatin accessibility and nucleosome phasing. Our results provide insights into how TF-chromatin interactions shape the regulatory landscape in the cell.

Main text:

Chromatin is the association between DNA, RNA, and diverse nuclear proteins, including nucleosomes. It enables the ~2 m 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 aspect 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 the

regulatory circuits leading to differences in transcriptional activity across diverse species, tissue, 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 hypothesize that chromatin local 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 how chromatin accessibility informs 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 midpoint distribution around TF binding sites and can result in a stereotyped "V" pattern of points for bound TFs that phase nearby nucleosomes (Fig. 1A, upper plot). The extent of organization of data in the V-plot can be quantified using Shannon's entropy (*6*). We calculated the 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 coverage across the V-plot, we devised a metric called chromatin information enrichment, calculated as the $\log_2$ ratio of the observed divided by the expected V-plot information content (*1*). The expected information was obtained by permuting the data in a manner that preserves the observed fragment length distribution in the V-plot (*1*) (Fig. 1A, middle and lower plots). We summarized chromatin information enrichment into a single value, named feature V-Plot Information Content Enrichment (f-VICE), which represents the chromatin information enrichment at landmark TF and nucleosomal positions across the V-plot (Fig. 1A, lower plot) (*1*). These regions are expected to have high chromatin information enrichment when the nucleosomes are phased around the TF binding site. Therefore, our chromatin information theory metric, 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 and 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*). In addition, we generated an independent GM12878 ATAC-seq dataset with higher signal-to-noise ratio, as measured by transcription start site (TSS) enrichment and fraction of reads in peaks (Fig. S1). Using our new data, we created V-plots and calculated f-VICEs centered on bound motif instances for these 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). Accordingly, CTCF f-VICE was highest among the tested TFs (Fig 1D). Other TFs, exemplified by AP-1 and NFKB, had diverse f-VICE (Fig. 1B-D, S2), indicating extensive variation in chromatin architecture between TFs. These results were consistent across independent GM12878 ATAC-seq datasets, indicating the robust nature of the f-VICE metric (Fig. S2). These results indicate that differences in the TF-chromatin interactions are captured in the chromatin information enrichment patterns.

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*, *12*). We noted that CTCF and AP1, which have high experimentally measured residence times (*13*, *14*), had high f-VICEs, while NFKB, which has a low residence time (*15*), had lower f-VICE. We therefore hypothesized that chromatin information enrichment correlates with residence time. However, due to the lack of ChIP-seq data for other TFs with known residence times, we were unable to use f-VICEs calculated from ChIP-seq data to test this hypothesis.

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 (*16*–*18*). However, analyses of footprint depths (accessibility at the center protected region relative to the flanking unprotected regions) across multiple ATAC-seq and DNase-seq datasets suggests that ~80% of TFs do not have footprints (*19*). Further, a TF binding prediction method that relies on the presence of a footprint may bias towards detecting organized chromatin regions. To address these issues, we developed BMO, an unsupervised method to predict TF binding from chromatin accessibility data without using footprints. BMO builds on previous reports that the degree of chromatin accessibility around a motif (*20*–*22*) and the presence of co-occurring motifs (*23*) positively correlate with TF binding, and uses TF-specific negative binomial models of these two signals to estimate the likelihood of a bound instance (*1*). We benchmarked the performance of BMO and other unsupervised TF binding prediction algorithms using ATAC-seq data from the GM12878 and HepG2 (*24*) cell lines and their corresponding TF ChIP-seq data (n=41 and n=59, respectively; Table S1). We compared BMO to three footprinting-based algorithms (HINT-ATAC (*18*), DNase2TF (*17*), PIQ (*16*)), CENTIPEDE, which learns informative DNA cut patterns indicating TF binding (*25*), and a baseline classifier that labels TF motifs within ATAC-seq peaks as bound. BMO outperformed all methods in our high-signal GM12878 dataset (Fig. 1E), whereas BMO and CENTIPEDE had similarly high performance in lower-signal datasets (Figs. S1, S3). PIQ cannot use custom TF motif scans and therefore required separate benchmarking, which revealed lower performance compared to BMO (Fig. S4). These results were consistent across ATAC-seq replicates and cell lines, including downsampled data representing shallower sequencing depths (Fig. S5). Overall, the two

footprinting-agnostic methods (BMO and CENTIPEDE) outperformed footprint-based methods on the majority of tested TFs, indicating that TF binding is more accurately predicted using a simple chromatin accessibility model tuned to each TF motif.

We next sought to determine if the CENTIPEDE approach relied on spatial DNA cut patterns, or if the overall accessibility in the region was sufficient for high performance. We devised an alternative implementation of CENTIPEDE that ignores the DNA cut positions (signal-sum CENTIPEDE; ssCENTIPEDE), thus masking any footprint-like patterns in the motif vicinity (1). This ssCENTIPEDE approach performed almost identically to CENTIPEDE (Figs. 1E, S6), again indicating that footprint patterns in chromatin profiles are not necessary for high prediction performance. One corollary expectation from this conclusion is that footprint-based algorithms should perform comparatively worse when predicting binding for TFs with a low impact on local chromatin. To test this, we compared performance across f-VICE tertiles representing low (tertile one), intermediate (tertile two), and high (tertile three) f-VICEs. Notably, BMO and (ss)CENTIPEDE had relatively higher performance on lower f-VICE tertiles one and two (Fig. 1E, S3). Our findings indicate that footprinting-based methods are more sensitive to the local TF-chromatin architecture.

Having determined that footprint-based methods are less accurate for predicting TF binding, we proceeded with BMO predictions to estimate f-VICEs for TFs without ChIP-seq data. We found that BMO-predicted f-VICEs are significantly correlated with f-VICEs calculated from TF ChIP-seq data in GM12878 and HepG2 (Pearson's $\rho \geq 0.72$, $p \leq 1e\text{-}10$; Fig. S7). We therefore concluded that BMO can be used to estimate f-VICEs without ChIP-seq data and performed BMO TF binding predictions and calculated f-VICEs for 540 non-redundant (*1*) TF motifs (Table S2). We used high-quality ATAC-seq datasets from an additional four human tissues (pancreatic islets (*26*), pancreatic islet sorted alpha and beta cells (*27*), and CD4+ cells (*28*); Table S1). These datasets were selected using a quality control strategy that uses the highly stereotyped chromatin architecture in ubiquitous and conserved CTCF/cohesin binding sites to infer sample quality (Fig. S8) (*1*). We used linear regression to normalize f-VICEs within each sample to control for differences in bound motif predictions and overall chromatin accessibility (Fig. S9). The majority of the 540 TF motifs followed an approximately normal f-VICE distribution, but we observed an upper tail with higher f-VICEs, which motivated us to fit two distributions to the data (*1*) corresponding to low or high f-VICE motifs. The median percentage of motifs associated with high f-VICEs across datasets was 14% (Fig. 1G, S10), which is comparable to the percentage of motifs associated with DNase footprint protection across datasets (median=19%) from another study (*19*). This result supports our conclusion that footprint-based algorithms will not perform well on most TFs. Together, these results reinforce the use of BMO for accurately calculating f-VICE and indicate that a minority of TFs associate with high chromatin information enrichment.

We next asked if chromatin information enrichment reflected TF-DNA residence times. We correlated BMO-informed f-VICEs with previously measured fluorescence recovery after photobleaching (FRAP) data from mammalian cell lines (Table S3), which provide an upper bound of TF residence time (*13*, *29*) . Using a robust linear regression to protect against outlier influence, we found that f-VICE significantly correlated with FRAP recovery times in all samples ($\beta \geq 0.7$, Bonferroni adjusted $p \leq 0.001$), suggesting that TFs associated with high chromatin information enrichment have longer DNA binding times (Fig. 2A, S11). A recent study found that cohesin has a residence time 10- to 20-fold higher than CTCF (*13*). We reasoned this difference could be reflected in the local chromatin architecture and calculated the chromatin information enrichment of GM1278 CTCF binding sites with and without the presence of cohesin (CTCF/cohesin$^+$ and CTCF/cohesin$^-$), informed by CTCF and cohesin subunit Rad21 ChIP-seq experiments. When comparing these two classes of CTCF bound motifs, we performed quantile-based downsampling (*1*) to control for potential confounding biases (Fig. S12A). CTCF/cohesin$^+$ had 1.9-fold higher chromatin information enrichment compared to CTCF/cohesin$^-$ after controlling for potentially confounding variables (Fig. 2B, S12B), indicating these distinct CTCF occupancy classes have different chromatin information enrichment signatures. We next compared the nucleosome positioning signals inferred from lymphoblastoid cell line micrococcal nuclease sequencing (MNase-seq) profiles (*10*, *30*). Only the CTCF/cohesin$^+$ class was associated with phased nucleosomes around the binding site (Fig. 2C), consistent with longer residence times associated 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. S13). There were no detectable nucleosome phasing patterns in the motif-flanking chromatin information enrichment of the sonicated sample (Fig. 2D; see vertical arrows). These results show that chromatin information enrichment signatures of the two classes of CTCF binding result from TF-chromatin interactions instead of differences in chromatin accessibility.

To further characterize the association between chromatin information enrichment and nucleosome positioning, we systematically compared GM12878 chromatin information enrichment patterns across TF motifs to lymphoblastoid MNase-seq profiles. Using *k*-means clustering, we found three clusters representing a continuum of chromatin information enrichment at the motif region (Fig. S14A). Clusters one and two had lower chromatin information enrichment at the motif compared to motif-flanking regions, while cluster three had the highest chromatin information enrichment at the motif (Fig. 2E, S14A-B). Cluster three encompassed >95% of the high f-VICE motifs (Fig. S14C-D). Notably, we observed two distinctly anti-correlated MNase signal patterns for the motifs in cluster three, consistent with TFs binding at the center of the nucleosome dyad or between phased nucleosomes (*5*) (Fig. 2E). The chromatin information enrichment and MNase signals were

anti-correlated at high f-VICE motifs (Fig 2E; right yellow/green heatmap), indicating that the highest chromatin information enrichment TFs are associated with nucleosome phasing. We quantified nucleosome phasing by calculating the $\log_2$ ratio of MNase-seq signal at motif-proximal and motif distal regions (*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. S15). These results suggest that TF-chromatin interaction patterns are driven by TF residence time, resulting in distinct chromatin information enrichment signatures.

Previous reports suggested that a subset of TFs directionally bind DNA, with potential effects on gene regulation (*16*, *31*, *32*). To investigate this further, we extended our information content analyses to quantify chromatin information enrichment asymmetry (*1*). Of the 540 motifs tested, we detected 150 with significantly asymmetric chromatin information enrichment (Bonferroni corrected $p<0.05$; Z-scores ≥ 12.8), including members of the PAX, SPI1 (PU.1), and YY2 families (Fig. 2F, S16A). The direction of chromatin information enrichment asymmetry was positively correlated with the direction of the nearest TSS relative to each motif instance (Spearman's $\rho = 0.66$, $p=2e\text{-}16$; Fig. S16B). To determine if chromatin information enrichment asymmetry was an artifact of TSS proximity, we performed a regression analysis of nearest TSS direction bias and chromatin information enrichment asymmetry, controlling for TSS distance (*1*). Chromatin information enrichment asymmetry remained significant ($p=1e\text{-}15$) when controlling for TSS distance (Fig. S16C), suggesting that chromatin information enrichment asymmetry may be an intrinsic property encoded in the motif and not simply an artifact of TSS proximity. As an independent test, we repeated the analysis calculating asymmetry separately for TSS-proximal (≤1 kb) and TSS-distal (≥10 kb) motif instances. The TSS-distal and TSS-proximal chromatin information enrichment asymmetry directions agreed more than expected by chance (111/150, binomial test $p=4e\text{-}9$; Fig S16D), suggesting that chromatin information enrichment asymmetry is intrinsic to the TF motif. The magnitude of asymmetry was stronger in TSS-proximal motifs (Fig S16D), suggesting that TSS proximity amplifies TF chromatin information enrichment asymmetry. Accordingly, the correlation between nearest TSS direction and chromatin information enrichment asymmetry was more pronounced at TSS-proximal motifs (Spearman's $\rho = 0.88$, $p=2e\text{-}16$; Fig. 2G). These results support that directional binding is a property of TF-chromatin interactions.

We next aimed to investigate cross-tissue differences in chromatin information enrichments. We performed an unsupervised hierarchical clustering of motif f-VICEs and found that it recapitulated the expected tissue grouping, indicating that f-VICEs vary in a tissue-specific manner (Fig. 3A). The motifs driving the clustering patterns included known tissue-specific transcriptional regulators, such as IRF, SPI1, and ETS in blood-related samples, PAX in pancreatic islets, and FOX in HepG2 (Fig. S17), consistent with chromatin information enrichment reflecting TF activity. A recent study demonstrated that NF-KB

(p65) residence time is determined by its DNA-binding domain (DBD) (*33*). We therefore asked if DBDs or other protein domains are associated with chromatin information enrichment. We assigned high-confidence DBDs and protein domains to motifs and designed a permutation-based rank test to calculate the f-VICE enrichment of each domain (*1*). The basic-leucine zipper (bZIP) and beta helix-loop-helix (bHLH) domains were significantly enriched for high f-VICE, while the nuclear receptor, Rel, and Myb/SANT domains were significantly enriched for low f-VICE in the majority of samples (false discovery rate <5%). We observed tissue-specific f-VICE enrichments in the interferon-regulatory factor (IRF) and E26 (ETS) in blood-related samples, paired box (PAX) in islet-related samples, and forkhead (FOX) domains in HepG2 (Fig. 3B, S18). Our findings show the landscape of TF-chromatin interactions varies across tissues and reflects properties of TF biology.

The prevalence of tissue-specific differences in chromatin information enrichments led us to examine the role of high f-VICE TFs in regulating gene expression. To determine this, we calculated the enrichment of the motifs categorized as high or low f-VICE in GM12878 to overlap lymphoblastoid cis-expression quantitative trait loci (cis-eQTLs) datasets (*34*, *35*), which represent gene expression genetic control regions in these cells. High f-VICE motifs had higher (15-30%, median=24%) fold-enrichment in cis-eQTLs compared to low f-VICE motifs (Fig. 3C, S19A), but we did not find differences in effect sizes between eQTL overlapping high and low f-VICE motifs (Fig. S19B). These results indicate that high f-VICE TFs are more likely to mediate genetic effects on gene expression, but not necessarily the magnitude of these effects. Given the highly ordered chromatin (Fig. 1), high predicted residence times (Fig. 2A, B, S11), and nucleosome phasing properties (Fig. 2E, S15) associated with high f-VICE TFs, we hypothesize that their regulatory effects could result from acting as or recruiting pioneer factors (TFs that induce chromatin accessibility) (*16*, *36*). If true, we would expect increased chromatin information enrichment for single nucleotide polymorphism (SNP) alleles with increased chromatin accessibility (*i.e.* with ATAC-seq allelic imbalance; Fig. 3D). We first performed a motif-agnostic approach to calculate the f-VICEs associated with every 6-mer in the human genome, using linear regression to control 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. Similar to TF motifs, DNA 6-mers have a distribution of f-VICEs (Fig. 3E; see histogram). Importantly, 6-mers differing by a single base-pair could have large differences in f-VICE, indicating that genetic variation impacts chromatin information enrichment (Fig. 3E, S20A). We separated the DNA 6-mers into f-VICE deciles and found that they differed in nucleotide composition and enrichment in genomic features, suggesting distinct biological roles (Fig. S20B-C). We calculated the f-VICE of the DNA 6-mers formed by both alleles at SNPs with significant ATAC-seq allelic imbalance (binomial test $p<0.05$) in GM12878 and pancreatic islets. The preferred ATAC-seq alleles

were significantly biased to be associated with higher f-VICE 6-mers compared to the less favored allele (permutation test p<3e-4; Fig. 3F-G, S21). These findings support a model where TFs with distinct properties (pioneers) bookmark regions of the genome to allow binding of other TFs (migrants) (*16*, *36*). Notably, TF motifs that are predictive of binding without any chromatin accessibility data (based solely on the motif match score) are significantly associated with higher f-VICEs in GM12878 and HepG2 (robust linear regression p=0.0002 and 0.002, respectively; Fig. S22). This suggests that high f-VICE TFs, like CTCF, are more likely to bind any strong motif, while the remaining TFs require motifs located in accessible regions. Collectively, our results demonstrate that the chromatin information landscape captures dynamic properties of TF-chromatin interactions, with implications for cell state memory and gene regulation.

Acknowledgements

We thank members of the Parker Lab, L. J. Scott, P. Freddolino, P. Wittkopp, M. Burmeister, and G. Higgins 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.

Competing interests: The authors declare that they have no competing interests.

Author contributions:

ROA: Analyzed data, designed computational experiments, wrote the manuscript. YK: Generated ATAC-seq datasets. JH: Implemented computational algorithms. AV: analyzed eQTL and chromHMM 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.

Data and materials availability: Code and scripts are available at http://github.com/ParkerLab/chromatin_information. ATAC-seq data is deposited in GEO (GSE135074).

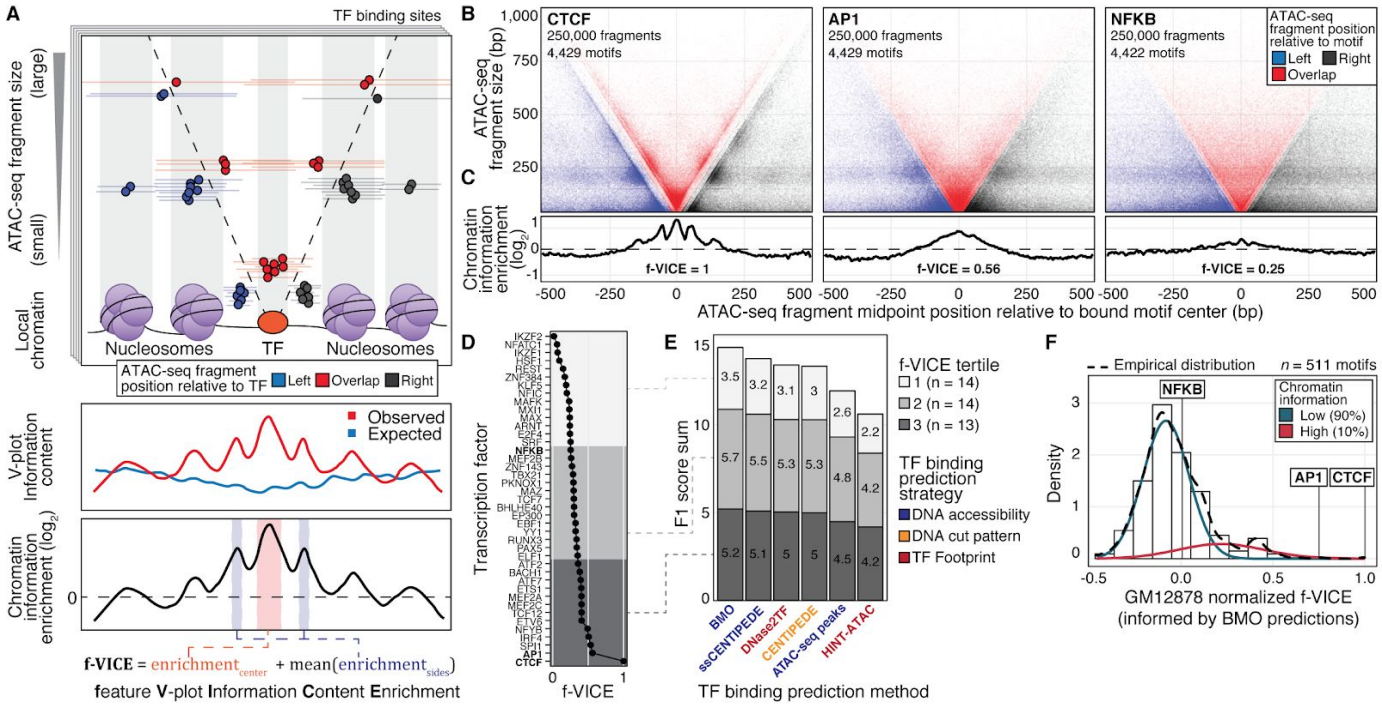

Fig. 1 - Information content of TF-chromatin interactions

A) Upper: The impact of TF binding on the chromatin architecture and on the observed ATAC-seq fragment distribution around TF binding sites. Gray regions represent steric hindrance to Tn5 transposition. Middle and bottom: calculation of chromatin information and f-VICE. **B-C)** V-plots and chromatin information of NFKB, AP-1, and CTCF bound sites in the GM12878 ATAC-seq data generated in this study. V-plots were downsampled to the same number of motifs and ATAC-seq fragments to highlight intrinsic differences in chromatin architecture (no downsampling was used 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 f-VICE distribution based on BMO predictions.

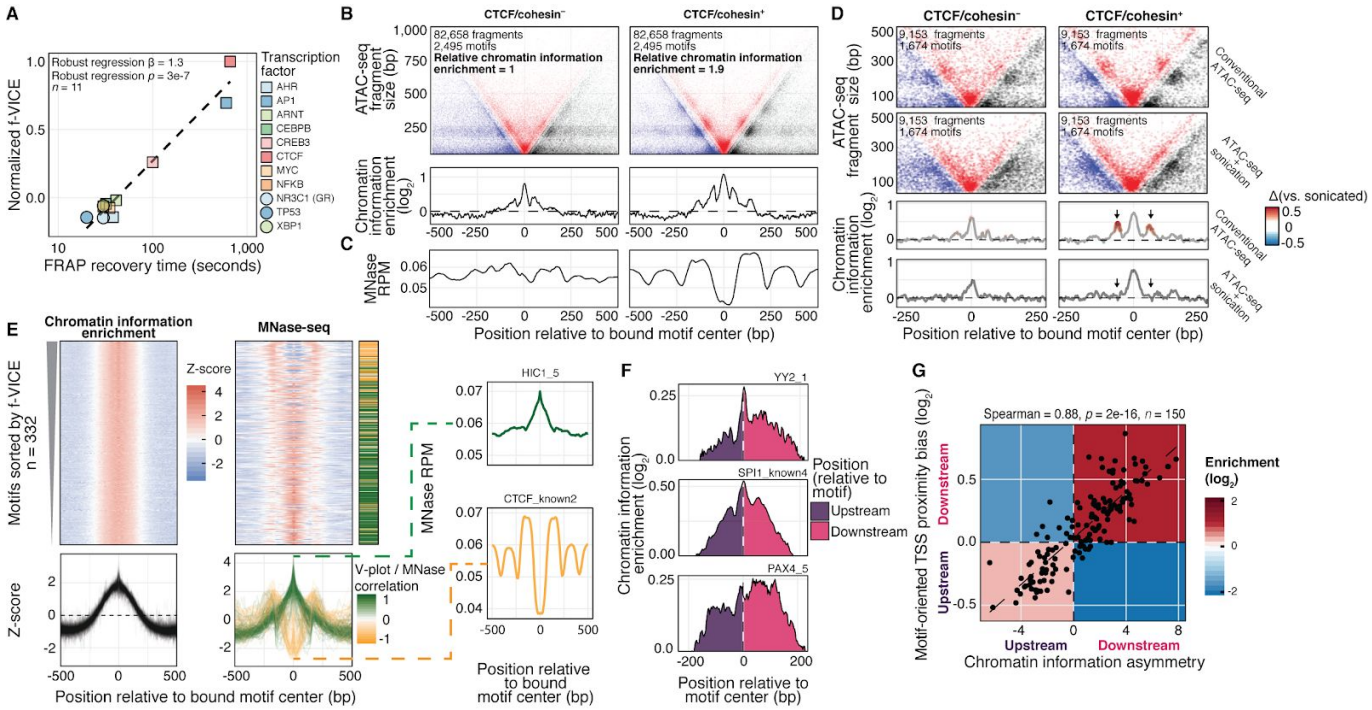

Fig. 2 - Chromatin information informs residence times and TF-nucleosome interactions

A) Correlation of FRAP recovery times and f-VICE in GM12878. **B)** V-plots and chromatin information 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 in the conventional ATAC-seq chromatin information lines indicate differences relative to sonicated. **E)** Left: Chromatin information and MNase-seq profiles of cluster three motifs. Middle: Heatmap of the correlation between MNase and chromatin information Z-scores. Right: Example motifs with positive and negative correlation. **F)** The top 3 motifs with chromatin information asymmetry in GM12878. **G)** Scatter plot of motif-oriented TSS position bias and chromatin information asymmetry in TSS-proximal motifs. Enrichments calculated by permuting the signs of observed values (n=10,000).

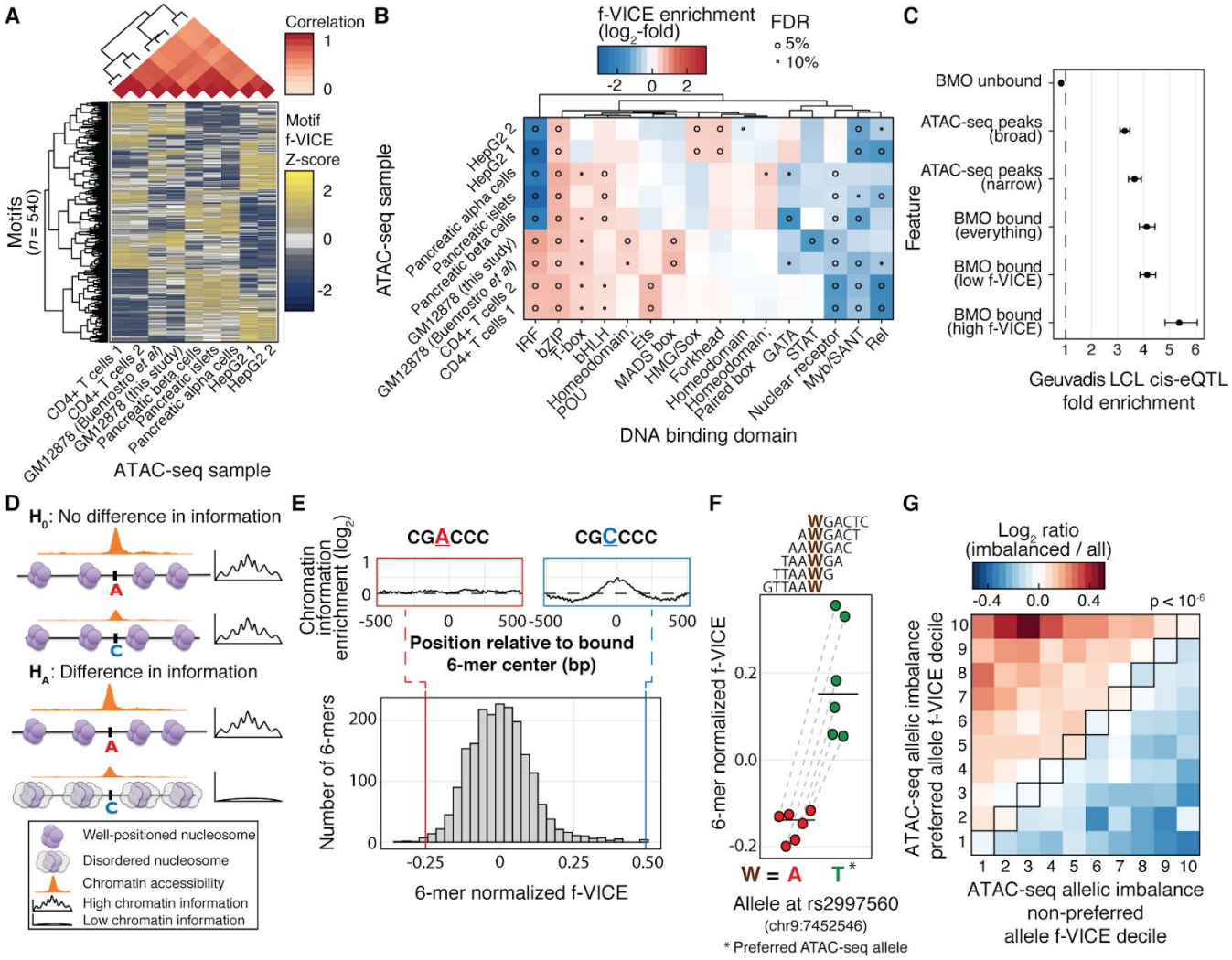

Fig. 3 - The chromatin information landscape of human tissues

A) Hierarchical clustering of f-VICE Z-scores across datasets. **B)** f-VICE enrichments across DBDs. **C)** Enrichment of different features in LCL cis-eQTLs. Error bars represent standard deviation of the effect size. Vertical dashed line represents expectation. **D)** Hypothesis schematic. **E)** Upper: example of two 6-mers with a Hamming distance of 1 and distinct pancreatic islets f-VICEs. Lower: pancreatic islets 6-mer normalized f-VICE distribution. **F)** Example of predicted change in chromatin information associated with SNP rs2997560 in pancreatic islets. Horizontal bars indicate median. **G)** Log_2 ratio of f-VICE decile changes associated with the preferred and non-preferred alleles of imbalanced SNPs versus all tested SNPs in human pancreatic islets.

References

1. Materials and methods are available as supplementary materials.
2. E. Segal, J. Widom, *Trends Genet.* **25**, 335–343 (2009).
3. C. R. Lickwar *et al*, *Nature*. **484**, 251–255 (2012).
4. S. Rudnizky *et al*, *Proc. Natl. Acad. Sci.*, 201815424 (2019).
5. F. Zhu *et al.*, *Nature*. **562**, 76–81 (2018).
6. C. E. Shannon, *Bell Syst. Tech. J.* **27**, 379–423 (1948).
7. G. Jenkinson *et al*, *Nat. Genet.* **49**, 719–729 (2017).
8. J. D. Buenrostro *et al*, *Nat. Methods*. **10**, 1213–1218 (2013).
9. J. G. Henikoff *et al*, *Proc. Natl. Acad. Sci.* **108**, 18318–18323 (2011).
10. The ENCODE Project Consortium, *Nature*. **489**, 57–74 (2012).
11. Y. Fu *et al*, *PLoS Genet.* **4**, e1000138 (2008).
12. A. Loffreda *et al.*, *Nat. Commun.* **8** (2017), doi:10.1038/s41467-017-00398-7.
13. A. S. Hansen *et al*, *eLife*. **6**, 1–33 (2017).
14. C. E. Malnou *et al.*, *J. Biol. Chem.* **285**, 6552–6562 (2010).
15. D. Bosisio *et al.*, *EMBO J.* **25**, 798–810 (2006).
16. R. I. Sherwood *et al.*, *Nat. Biotechnol.* **32**, 171–178 (2014).
17. M. H. Sung *et al*, *Mol. Cell*. **56**, 275–285 (2014).
18. Z. Li *et al.*, *Genome Biol.* **20**, 45 (2019).
19. S. Baek *et al*, *Cell Rep.* **19**, 1710–1722 (2017).
20. H. H. He *et al.*, *Nat. Methods*. **11**, 73–78 (2014).
21. G. Cuellar-Partida *et al.*, *Bioinformatics*. **28**, 56–62 (2012).
22. G. G. Yardımcı *et al*, *Nucleic Acids Res.* **42**, 11865–11878 (2014).
23. J. Wang *et al.*, *Genome Res.* **22**, 1798–1812 (2012).
24. A. Chesi *et al.*, *Nat. Commun.* **10**, 1260 (2019).
25. R. Pique-Regi *et al.*, *Genome Res.* **21**, 447–455 (2011).
26. A. Varshney *et al.*, *Proc. Natl. Acad. Sci.* **114**, 2301–2306 (2017).
27. A. M. Ackermann *et al*, *Mol. Metab.* **5**, 233–244 (2016).
28. M. R. Corces *et al.*, *Nat. Methods*. **14**, 959–962 (2017).
29. F. Mueller, *et al*, *Curr. Opin. Cell Biol.* **22**, 403–411 (2010).
30. D. J. Gaffney *et al.*, *PLOS Genet.* **8**, e1003036 (2012).
31. A. Kundaje *et al.*, *Genome Res.* **22**, 1735–1747 (2012).
32. S. R. Grossman *et al.*, *Proc. Natl. Acad. Sci.*, 201804663 (2018).
33. A. Callegari *et al.*, *PLOS Genet.* **15**, e1007891 (2019).
34. GTEx Consortium, *Nature*. **550**, 204–213 (2017).
35. T. Lappalainen *et al.*, *Nature*. **501**, 506–511 (2013).
36. K. S. Zaret, J. S. Carroll, *Genes Dev.* **25**, 2227–2241 (2011).}