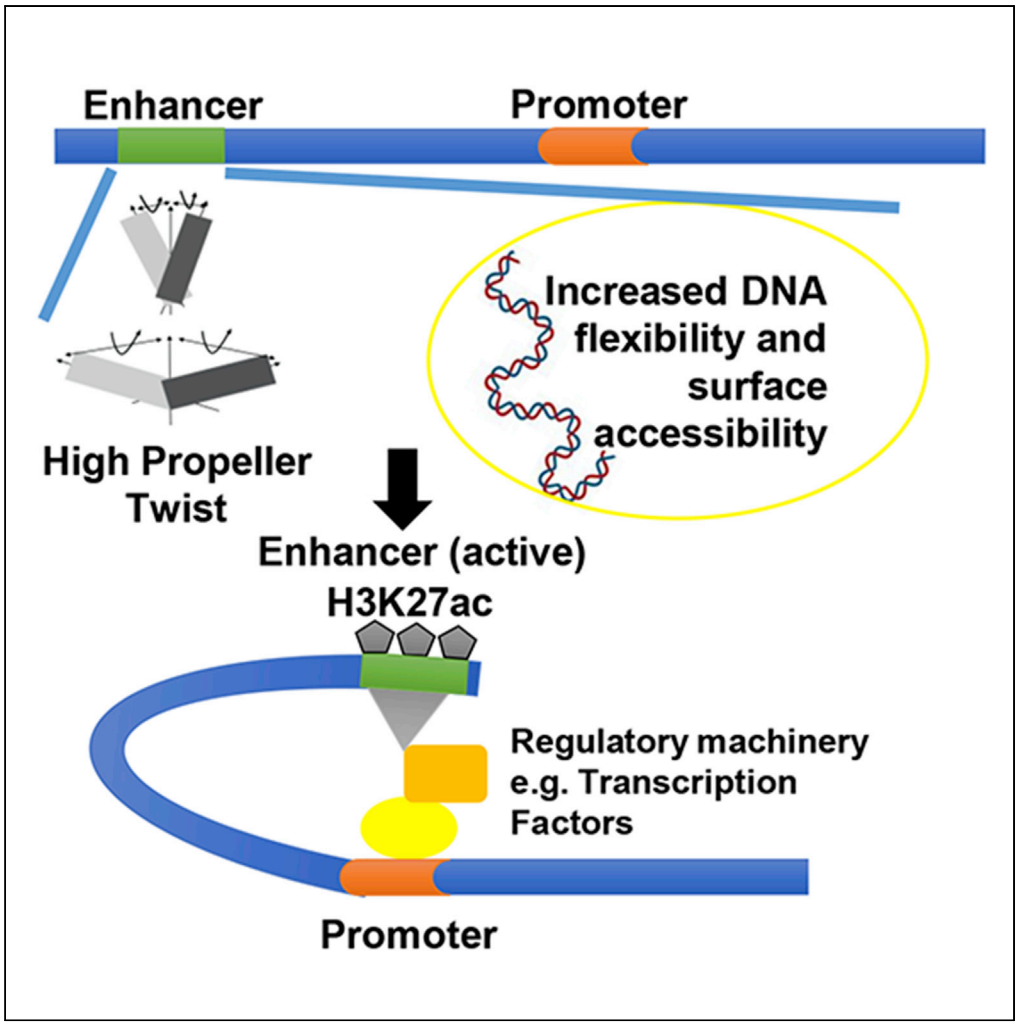
iScience

CellPress

Article

Deciphering the Gene Regulatory Landscape Encoded in DNA Biophysical Features



Abhijeet Pataskar, Willem Vanderlinden, Johannes Emmerig, Aditi Singh, Jan Lipfert, Vijay K. Tiwari

v.tiwari@qub.ac.uk

HIGHLIGHTS

DNA shape features encode genomic surface accessibility and flexibility

High ProT is a deterministic feature of enhancers

ProT levels correlate with nuclear organization of epigenetic states

Cell-fate switches involve a transient usage of low ProT regulatory elements

Pataskar et al., iScience *21*, 638–649
November 22, 2019
https://doi.org/10.1016/j.isci.2019.10.055

iScience

CellPress

Article

Deciphering the Gene Regulatory Landscape Encoded in DNA Biophysical Features

Abhijeet Pataskar,[1,4] Willem Vanderlinden,[2] Johannes Emmerig,[2] Aditi Singh,[3] Jan Lipfert,[2] and Vijay K. Tiwari[3,4,5,*]

SUMMARY

Gene regulation in higher organisms involves a sophisticated interplay between genetic and epigenetic mechanisms. Despite advances, the logic in selective usage of certain genomic regions as regulatory elements remains unclear. Here we show that the inherent biophysical properties of the DNA encode epigenetic state and the underlying regulatory potential. We find that the propeller twist (ProT) level is indicative of genomic location of the regulatory elements, their strength, the affinity landscape of transcription factors, and distribution in the nuclear 3D space. We experimentally show that ProT levels confer increased DNA flexibility and surface accessibility, and thus potentially primes usage of high ProT regions as regulatory elements. ProT levels also correlate with occurrence and phenotypic consequences of mutations. Interestingly, cell-fate switches involve a transient usage of low ProT regulatory elements. Altogether, our work provides unprecedented insights into the gene regulatory landscape encoded in the DNA biophysical features.

INTRODUCTION

The genome consists of multiple gene regulatory units comprising of proximal and distal regulatory elements. Recent studies have shown that the function and utilization of these elements during cellular differentiation and in response to intracellular and extracellular cues relies on a dynamic control by epigenetic machinery in concert with transcription factors (TFs) (Li et al., 2007; Long et al., 2016; Margueron and Reinberg, 2010; Moris et al., 2016). Enhancers are known to be critical for setting up transcriptome underlying cell-type identity and function from far away distances on their target promoters (Heinz et al., 2015; Long et al., 2016; Rickels and Shilatifard, 2018; Sakabe et al., 2012; Schoenfelder and Fraser, 2019; Shlyueva et al., 2014; Spitz and Furlong, 2012). Importantly, mutations in these gene regulatory elements are known to disrupt their function affecting gene expression and ultimately cell identity and hence underlie several diseases (Li et al., 2019; Rickels and Shilatifard, 2018; Sakabe et al., 2012; Weinhold et al., 2014). Typically, a range of methods are employed to identify and validate such distal regulatory elements including quantifying certain histone modifications and DNase hypersensitivity assays (Ong and Corces, 2011; Pradeepa et al., 2016; Shlyueva et al., 2014; Zentner and Henikoff, 2013). These methods have their own limitations and a number of alternate assays and histone modifications have recently been used to discover enhancers (Arnold et al., 2013; Pradeepa et al., 2016; Vanhille et al., 2015). Thus, our current approach to reveal regulatory elements in entirety is highly limited and vouches to search for conserved features that can explain enhancer evolution and function.

Interestingly, sequences from regulatory loci are able to recapitulate endogenous TF binding pattern, chromatin state, and cell-type-specific activity when placed at an exogenous genomic site or tested in isolation (Lienert et al., 2011; Yanez-Cuna et al., 2013). In addition, computational analysis has further shown that the occurrence of certain sequences at genomic loci is predictive of their regulatory potential (Colbran et al., 2017; Yanez-Cuna et al., 2014; Yang et al., 2017a). These lines of evidence strongly suggest the existence of inherent gene regulatory potential of these genomic loci at the sequence level. Despite these advances, we lack understanding of the evolutionary constrains in the selection of certain genomic DNA elements for their gene regulatory function (Pennacchio et al., 2013). It is thus important to decode the power of sequence features in determining the gene regulatory potential and differential usage in cell-type specification. In addition, it is important to catalog novel regulatory elements, an effort that is limited by insufficient knowledge of existing features of these elements.

Several laboratories have attempted to employ computational approaches to predict enhancers based on sequence information (Kleftogiannis et al., 2016; Lee et al., 2011; Rusk, 2014). Although these methods were

[1]Netherlands Cancer Institute, Amsterdam, the Netherlands

[2]Department of Physics and Center for NanoScience, LMU Munich, 80799 Munich, Germany

[3]Wellcome-Wolfson Institute for Experimental Medicine, School of Medicine, Dentistry & Biomedical Science, Queens University Belfast, Belfast BT9 7BL, UK

[4]Former Address: Institute of Molecular Biology, 55128 Mainz, Germany

[5]Lead Contact

*Correspondence: v.tiwari@qub.ac.uk

https://doi.org/10.1016/j.isci.2019.10.055

able to predict enhancers to a certain degree, they were unable to decipher the underlying code that drives enhancer selection and strength (Pennacchio et al., 2013). A previous study suggested that the local DNA topography differs at functional noncoding regions of the genome including enhancers (Parker et al., 2009). Interestingly, DNA shape features such as propeller twist (ProT), major grove width, and helical twisting determine different local geometries, which in turn contribute to the control of transcription factor binding and gene regulation (Greenbaum et al., 2007; Ma et al., 2017; Mathelier et al., 2016; Zhou et al., 2013). Overall, the existing evidences suggest a genetic feature code beyond simple sequence that may dictate selection of enhancers and their strength of function. Here we show that DNA shape features are highly informative of the gene regulatory potential of genomic loci. We discover that the ProT levels can reveal the location of enhancers, their strength, the affinity landscape of transcription factors, and distribution in the nuclear 3D space with high accuracy. Using experimental assays including single-molecule AFM imaging measurements, we show that indeed high ProT levels cause increased DNA flexibility and surface accessibility and may potentially explain their usage as regulatory elements. Furthermore, ProT levels also determine the effectivity landscape of the genome to tolerate mutations. Altogether, this work reveals the gene regulatory landscape encoded in the basic genetic sequence features and provides a significant advance in unfolding the mysteries of genetic code.

RESULTS

Genomic Surface Accessibility and Flexibility Are Encoded in DNA Shape Features

The ability for genomic regions to function as gene regulatory elements is thought to be significantly influenced by their inherent accessibility for DNA-binding proteins such as TFs (Bell et al., 2011). To probe accessibility, we began by investigating whether the surface accessibility of DNA is influenced by its biophysical features. We used hydroxyl radical cleavage maps as a proxy for solvent accessible surface area of the DNA (Greenbaum et al., 2007) and correlated this with various DNA shape features such as ProT, major groove width (MGW), helix turn (HelT), and roll predicted by an established tool—DNAshape (Zhou et al., 2013). We found that ProT, defined as the angle of twisting of two neighboring nucleotides from the axis of their geometrical center, highly correlates with the DNA surface accessibility (Pearson correlation coefficient = 0.967, pval <0.001) (Figure 1A). The other features do not show as strong a correlation with hydroxyl radical cleavage maps and with each other (Figures S1A–S1C). This analysis established ProT as a proxy to measure inherent surface accessibility of DNA.

Next, to directly test how increased levels of ProT affect the mechanical properties of DNA segments, we decided to carry out high-resolution AFM imaging experiments of ~1 kbp long DNA sequences with different ProT levels (Data S1). Toward this, we first used PCRs to generate different linear sequences predicted to be either genomic "random" or "high" in terms of ProT levels. Subsequently, atomic force microscopy (AFM) images of the DNA molecules were measured and analyzed by tracing the DNA paths (Figure 1B). An analysis of the mean-squared separation of pairs of points located at different distances along the contour length confirmed that the DNA molecules were equilibrated at the surface (Figure S1D) and allowed determination of the bending persistence length. We found bending persistence length values $P \approx 56$ nm, in good agreement with previous measurements under similar conditions (Mazur and Maaloum, 2014; Rivetti et al., 1996; Wiggins et al., 2006). The data did not reveal significant differences in the persistence lengths from control ($P = 56.1 \pm 0.2$ nm) and high ProT sequences ($P = 56.7 \pm 0.6$ nm), suggesting that the bending stiffness at longer length scales is similar for different levels of ProT.

Therefore, to probe the local flexibility of the DNA sequences, we analyzed the distribution of bend angles between points separated by 5 nm along the contour. Taking the negative logarithm of the histogram of bend angles directly gives the effective bending energy (Figure 1C). For both random and high ProT sequences, the data for angles up to $\theta \sim 1$ rad are well described by a simple elastic model, the so-called worm-like chain, whereas for bending angles $\theta > 1$ rad clear deviations from the elastic model are apparent, as have been observed previously (Wiggins et al., 2006). Interestingly, the high ProT sequences exhibited larger deviations from the elastic model and a significantly higher fraction of medium ($\theta > 0.8$ rad; $p = 7.5 \cdot 10{-}6$) and large bends ($\theta > 1.1$ rad; $p = 0.0013$) compared with the control sequences (Figure 1D). In contrast, different control sequences and different high ProT sequences gave the same fractions of medium and large bends, respectively, within experimental error. Taken together, the AFM imaging analysis suggested that on short length scales (~5 nm), high ProT sequences exhibit enhanced bendability compared to random sequences.

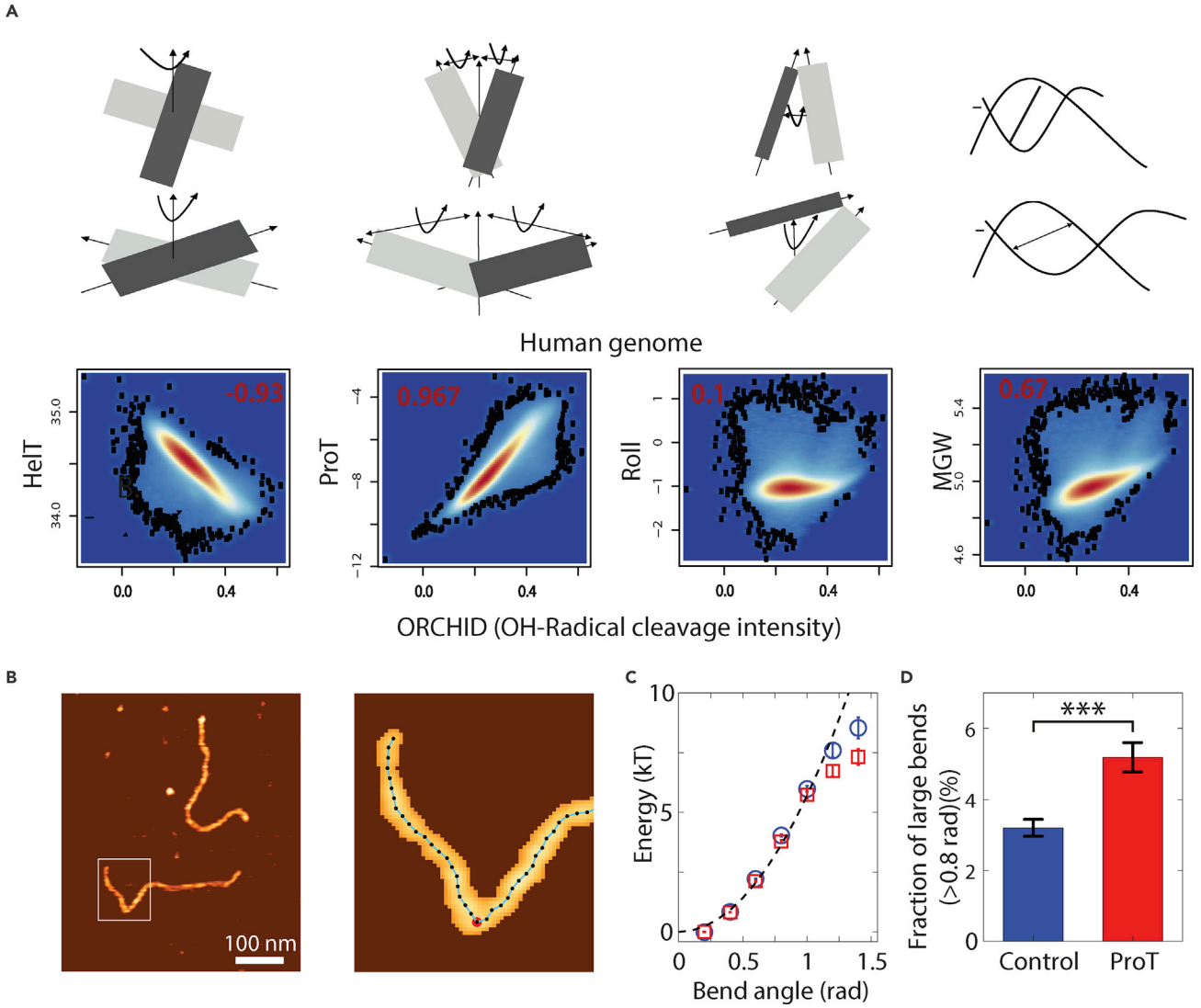


Figure 1. DNA Surface Accessibility and Flexibility Are Encoded in Its Biophysical Features

(A) Human genome-wide correlations of predicted values of DNAshape features; helix turn (HelT), propeller twist (ProT), roll (Roll), and major groove width (MGW) with surface accessibility of the DNA as measured by OH-radical cleavage intensity predictions.

(B) Typical AFM topographic image depicting two linear DNA molecules (left). Enlarged view of the boxed area (left) superimposed with the traced contour (right). The red point highlights the location of a large bend (>0.8 rad).

(C) Energy landscape for bend angles reconstructed from the bend angle distribution for pooled control ($N_{molecules}$ = 801) and pooled ProT ($N_{molecules}$ = 425) sequences. In total, we traced 87,168 bend angles from 1,226 imaged DNA molecules. The broken line depicts the energy landscape expected for a worm-like chain with persistence length P = 55 nm.

(D) Fraction of large bends (>0.8 rad) for pooled control ($N_{large\ bends}$ = 182; $N_{bends,\ total}$ = 56,874) and high ProT ($N_{large\ bends}$ = 157; $N_{bends,\ total}$ = 30,294) sequences. The fraction of large bends is significantly higher for the ProT versus control sequences ($p = 7.5 \times 10^{-6}$). The error bar is the standard deviation from counting statistics, i.e. the square root of the counts divided by the number of total counts.

See also Figure S1.

Propeller Twist Levels Correlate with 3D Nuclear Positioning of Distinct Chromatin States

Eukaryotic genomes are compartmentalized into distinct domains marked by active (eu-) and inactive (hetero-) chromatin. Inspired by the observation that ProT highly correlates with the inherent surface accessibility and bendability of DNA, we hypothesized that these regions could potentially mark open active chromatin regions that are also known to be more fluid in nature. A previous study employed single cell Hi-C assays to reconstruct the 3D genome of mouse embryonic stem (mES) cells at a high resolution (Stevens et al., 2017). We processed these data and overlaid with histone modifications indicative of

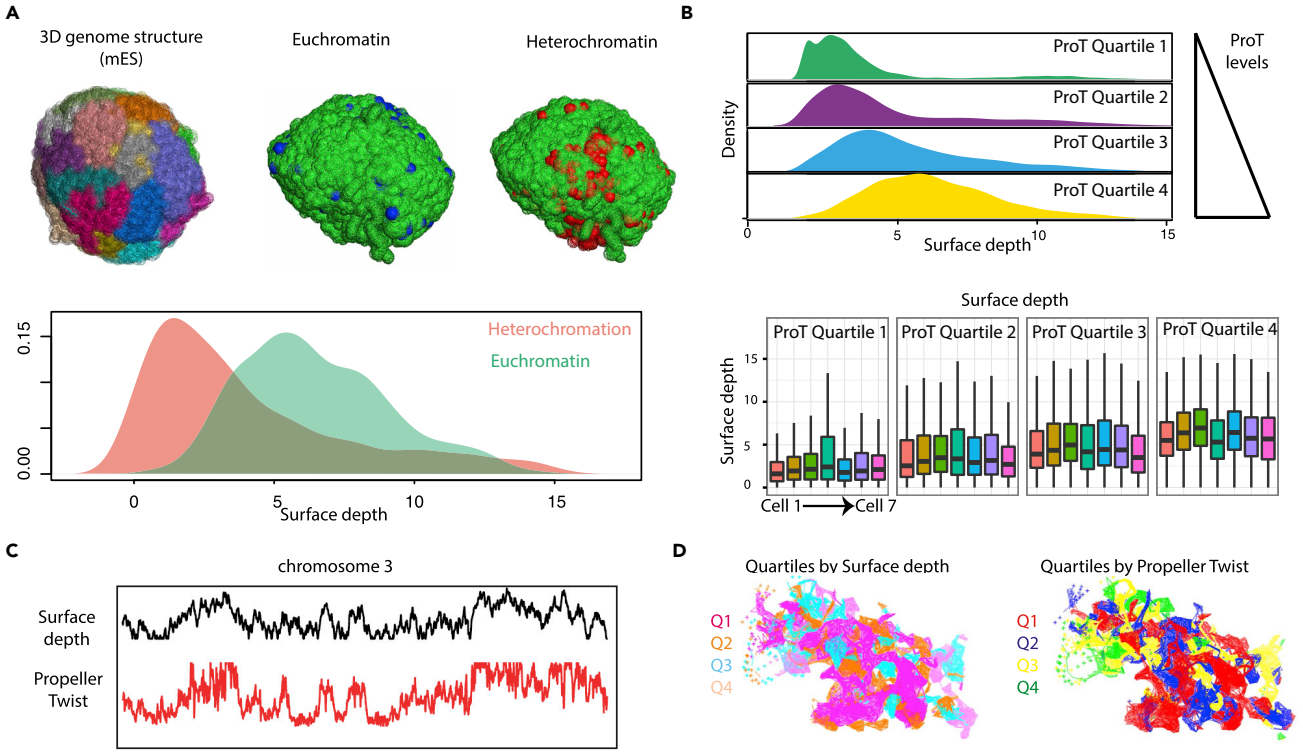


Figure 2. Propeller Twist Predicts 3D Nuclear Positioning Hallmark of Distinct Chromatin States

(A) Upper panel: (Left) chromosomes overlaid in different colors in reconstructed single cell genome structure with the resolution of 1MB from single cell HiC studies in mouse ES cells (Stevens et al., 2017). (Mid and Right) Genomic subunits each of 1MB highlighted in blue and red depending of enrichment of euchromatin (blue) and heterochromatin (red) features. Lower panel: density plot depicting surface depth in the reconstructed genome structures of euchromatin feature and heterochromatin features.

(B) Density plot depicting enrichment of surface depth from reconstructed genome structure of mES cell into genome subunits characterized into four quartiles in increasing amount of median Propeller Twist (ProT) values arranged from top to bottom. Lower panel: boxplots depicting surface depth values for every quartile of ProT values in all seven different studies single cell genome structures.

(C) Line plot depicting profiles for ProT (red) and surface depth (Cell 1, black) aimed toward displaying linear correlation of these two features across chromosome 3.

(D) Reconstructed 3D structure of chromosome 3 color-overlaid with quartiles of Surface depth (left) and ProT levels (right).

See also Figure S2.

euchromatin and heterochromatin, H3K27ac and H3K9me3, respectively. Interestingly, euchromatin was found to have a higher surface depth (as defined in Stevens et al., 2017) as compared with heterochromatin (Figures 2A and S2A). These findings are also in line with the local enrichment of heterochromatic lamina-associated domains (LADs) at the nuclear periphery (van Steensel and Belmont, 2017).

We next analyzed the radial distribution of sequences with different ProT levels within the Hi-C data derived from 3D nuclear positioning. Interestingly, ProT levels were found to correlate well with the nuclear distribution, where "high ProT" sequences occupy an internal position, whereas "low ProT" sequences are localized at the periphery (Figure 2B). Importantly, while the surface depths of genomic loci at the single cell level across various ES cells is variable, the overall radial positioning of differential ProT regions in the genome is highly consistent (Figures S2B, S2C, and 2B). Simultaneous visualization of chromosome-wide surface depth and ProT profiles also showed a clear correlation between these two features (Figures 2C, 2D, and S2D). Collectively, these results suggested a potential contribution of ProT in influencing the nuclear positioning and its association with distinct epigenetic states.

Propeller Twist Encodes the Regulatory Potential of Genetic Elements

Intrigued by the above findings, we next attempted to perform a detailed characterization of high ProT regions. In line with our previous findings, we find that "high ProT" sequences are prevalent at regions

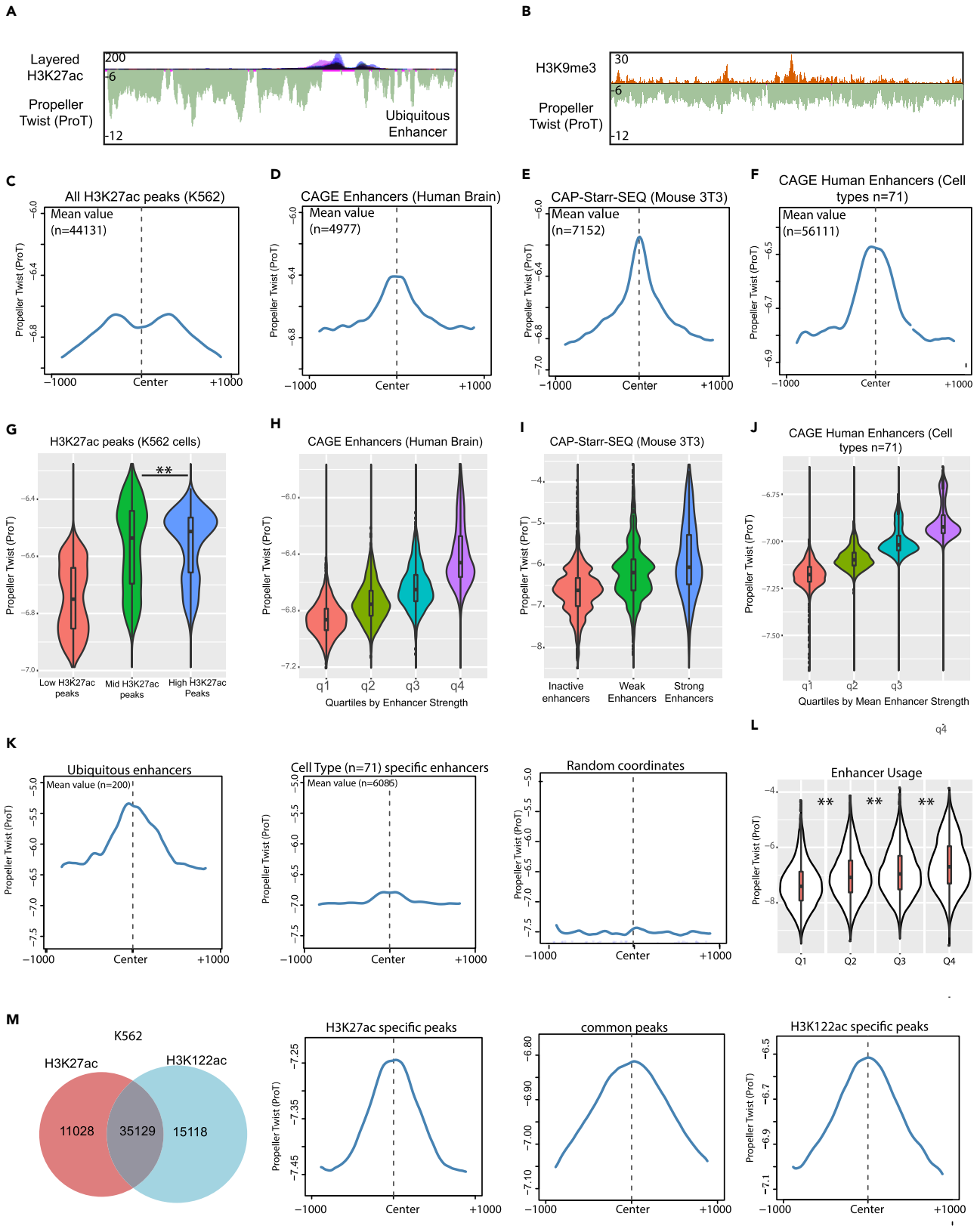

A
Layered H3K27ac
Propeller Twist (ProT)
200
-6
-12
Ubiquitous Enhancer
B
H3K9me3
Propeller Twist (ProT)
30
-6
-12
C
All H3K27ac peaks (K562)
Mean value (n=44131)
Propeller Twist (ProT)
−1000
Center
+1000
D
CAGE Enhancers (Human Brain)
Mean value (n=4977)
E
CAP-Starr-SEQ (Mouse 3T3)
Mean value (n=7152)
F
CAGE Human Enhancers (Cell types n=71)
Mean value (n=56111)
G
H3K27ac peaks (K562 cells)
**
Low H3K27ac peaks
Mid H3K27ac peaks
High H3K27ac Peaks
H
CAGE Enhancers (Human Brain)
q1
q2
q3
q4
Quartiles by Enhancer Strength
I
CAP-Starr-SEQ (Mouse 3T3)
Inactive enhancers
Weak Enhancers
Strong Enhancers
J
CAGE Human Enhancers (Cell types n=71)
Quartiles by Mean Enhancer Strength
K
Ubiquitous enhancers
Mean value (n=200)
Cell Type (n=71) specific enhancers
Mean value (n=6085)
Random coordinates
L
Enhancer Usage
**
Q1
Q2
Q3
Q4
M
K562
H3K27ac
H3K122ac
11028
35129
15118
H3K27ac specific peaks
common peaks
H3K122ac specific peaks

Figure 3. Propeller Twist Encodes the Regulatory Potential of Genetic Elements

(A) UCSC genome browser track displaying layered H3K27ac tracks and Propeller Twist (ProT) showing higher ProT values at H3K27ac positive regions.

(B) UCSC genome browser track displaying H3K9me3 (K562) tracks and ProT showing lower ProT values at H3K27me3 positive regions.

(C) Density plot depicting ProT at H3K27ac peaks identified in K562 cells.

(D–F) Same as (C) but as expressed enhancers in Human Brain identified in CAGE experiments (D), STARRseq identified enhancers in mouse NIH3T3cells (E) and all CAGE defined enhancers across 71 cell types from FANTOM5 atlas (F).

(G) Violin-boxplot depicting ProT levels at H3K27ac peaks clustered into three categories based on enrichment (low, mid, and high).

(H) Same as G but at four groups obtained from increasing quartile expression of Human Brain enhancers identified from CAGE experiments.

(I) Violin-boxplot depicting ProT levels at capSTARR-seq defined enhancers classified into three classes as inactive, weak, and strong enhancers.

(J) Same as H but at quartiles defined by mean expression across 71 cell types.

(K) Density plot depicting ProT profiles at genomic coordinates marked by ubiquitous enhancers (left), cell-type-specific enhancers (mid) and random coordinates (right).

(L) Violin-boxplot depicting ProT levels at enhancers ranked into four quartile groups into the increasing order of enhancer usage (cross-cell type usage of enhancers).

(M) Venn-diagram depicting overlap of peaks from H3K27ac and H3K122ac ChIP-seq study in Human K562 cells (left). ProT profile plotted as density plot across H3K27ac specific peaks identified from this comparison (second from left), common peaks (second from right), and H3K122ac (right).

See also Figure S3.

enriched with H3K27ac, a marker of active promoter and enhancer regions, whereas they are depleted at regions enriched in the repressive epigenetic mark H3K9me3 (Figures 3A and 3B). To further validate these findings, we segmented the epigenome of human K562 myeloma cells into 15 different chromatin states using ChromHMM (Ernst and Kellis, 2012) and determined their ProT levels. Consistent with the previous observations, we found generally higher ProT levels at genomic regions marked by active chromatin marks as compared with repressive ones (Figures S3A–S3C). An interesting exception was H3K27me3, a repressive mark, which correlates with higher ProT levels (Figure S3C). This may be explained by the fact that H3K27me3 marks certain genomic regions that permit enhancer activity under certain physiological conditions (Taberlay et al., 2011). Furthermore, this mark is also known to be present at "poised" promoters that represent a transcription ready state (Bernstein et al., 2006).

We next extracted experimentally validated regulatory regions from a variety of cell types and analyzed their ProT profiles. Strikingly, we noticed that the regulatory elements defined by H3K27ac mark show a highly characteristic distribution of ProT levels where ProT peaks appear symmetrically next to the center of H3K27ac peaks (Figure 3C). Furthermore, ProT peaks overlap with the centers of regulatory elements identified by CAGE or STARR-seq experiments, suggesting that ProT is an intrinsic property of regulatory regions (Figures 3D and 3E). Extended analysis of CAGE-defined enhancers across 71 cell types further supports these findings (Andersson et al., 2014) (Figures 3F and S3D).

Next, we sought to monitor the correlation of ProT levels with enhancer activity in a quantitative manner. Because H3K27ac levels at enhancers are known to correlate with gene expression levels, we used this as a proxy for enhancer usage (Karlic et al., 2010). A comparison of H3K27ac enrichment with ProT levels demonstrated a clear relationship (Figure 3G), which was also true with CAGE- or STARR-seq-determined enhancer strength across multiple cell types (Figures 3H–3J). Based on these observations we also hypothesized whether ProT levels could also help discriminate enhancer usage across cell types. Interestingly, we indeed observed that higher ProT level-containing regions tend to be ubiquitous enhancers, whereas those showing lower ProT level were enhancers of cell-type specific genes (Figure 3K). This may relate to an easy activatable state of housekeeping genes versus those of cell-type specific genes that generally require distinct machinery and program to induce their expression. Further, we observed that the ProT levels correlate with expression levels (Figures 3L and S3E), suggesting that the transcriptional competence is potentially orchestrated at the genetic level by DNA shape features.

The current repertoire of histone modifications does not seem sufficient to define all genomic regulatory elements, and efforts are continuously being made to uncover new chromatin features that allow mapping all enhancers. In line with this, H3K122ac modification was shown to mark enhancers that do not exhibit any H3K27ac mark (Pradeepa et al., 2016). Further corroborating our previous observations, H3K122ac positive and H3K27ac negative enhancers show a characteristic high ProT profile (Figure 3M). Thus, high ProT levels constitute a common feature of enhancers irrespective of the chromatin mark defining these regions. These findings argue that high ProT levels constitute a common feature of

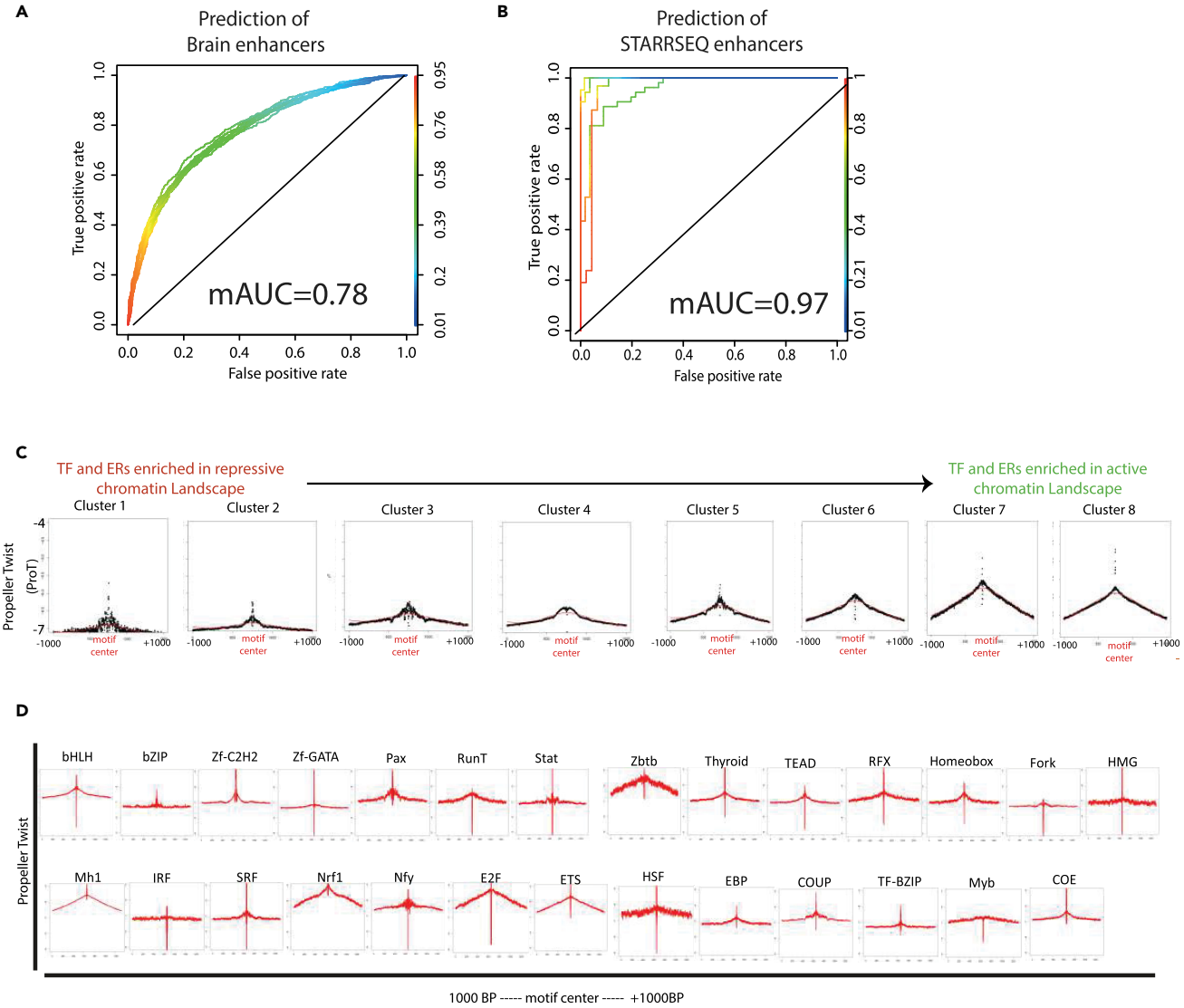


Figure 4. Propeller Twist Is Deterministic of Regulatory Potential of Genetic Elements

(A) Receiver operator curve (ROC) depicting prediction accuracy of SVM models (n = 9) trained to classify between enhancer sequences in brain and random genome loci using 2000BP single nucleotide ProT predictions.

(B) Same as A but SVM models (n = 5) trained at capSTARR-seq defined enhances in mouse NIH3T3cells.

(C) ProT density plots over each of the clusters (identified in Figure S4A) arranged from left to right in the increasing order of enrichment in repressive to active chromatin landscape.

(D) ProT density plots as median over all factors across particular transcription factor family.

See also Figure S4.

enhancers, which overrides limitations of other modes of predicting enhancers including those based on the epigenetic state.

ProT Profile Is a Deterministic Feature of Enhancers

To further establish predictive nature of ProT levels in priming genomic regions for a gene regulatory function we developed SVM (Support Vector Machine) models to classify between random genomic loci and CAGE-defined brain enhancers using single nucleotide ProT values across a 2000 bp window. The resulting nine models could classify the location of enhancers at randomly chosen genomic sites with very high accuracy (mAUC = 0.78) (Figure 4A). Next, we trained five models in a similar manner to identify STARR-seq defined enhancers from mouse NIH3T3cells. Again, our SVM models closely predicted enhancer locations (AUC = 0.96) (Figure 4B).

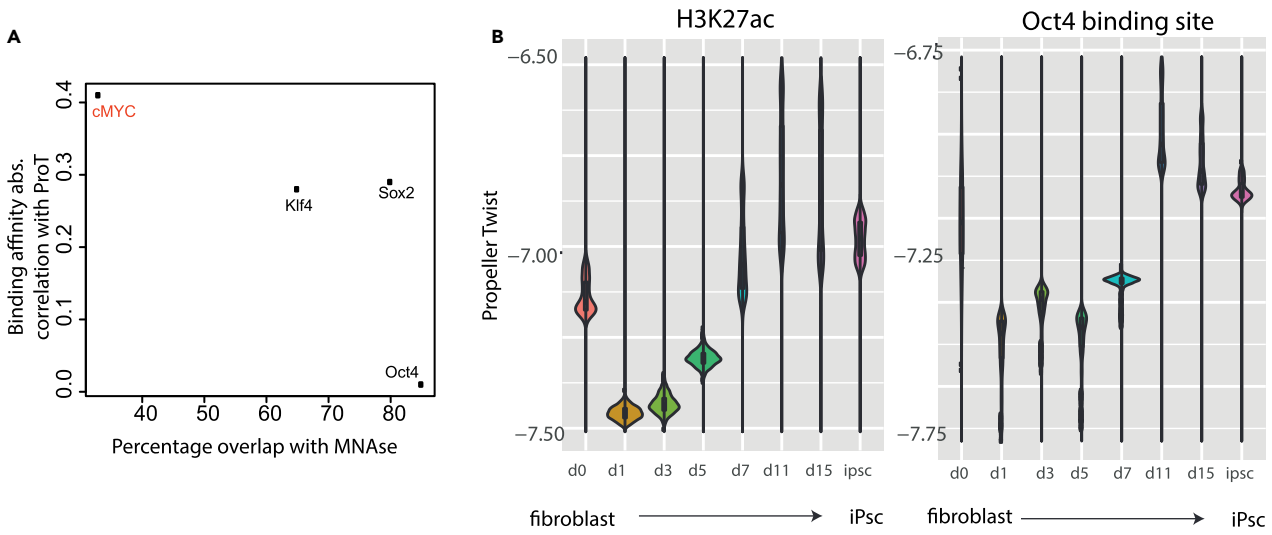


Figure 5. Switch between Distinct Cell-Fates Involves Transient Usage of Low ProT Regulatory Elements

(A) Scatter plot with enrichment of reprogramming factors on nucleosome (stage; fibroblasts) on X axis as determined by the study by Soufi et al. (Soufi et al., 2015) and binding affinity (stage: 48 h of reprogramming) correlation with ProT on Y axis.

(B) Violin-boxplots with ProT levels at H3K27ac peaks (left) and Oct4 binding sites (right) at various stages of reprogramming on mouse fibroblasts into iPSC cells.

Enhancers are known to contain multiple TF (Transcription Factor) binding sites, and given a strong relationship between ProT and enhancer occurrence and usage, we next had a closer look at TF motifs at ProT profiles. Here we clustered TF-bound motifs, as derived from actual ChIP-seq assays for these TFs, into eight different clusters based on various histone modification patterns at these sites as hallmark of euchromatin and heterochromatin (Figures S4A–S4C). Interestingly, an analysis of ProT levels at the center of motif at these TF bound sites for each of these clusters showed that the activator and repressor TF motifs can be clearly delineated by ProT levels. The TFs that function primarily as activators preferably target motifs embedded in high ProT environments, whereas TFs acting mainly as repressors bind motifs within lower ProT environments (Figures 4C and 4D). These findings argue that ProT profile is a deterministic feature of enhancers and can predict the active distal gene regulatory landscape with high accuracy.

Cell-Fate Switches Involve a Transient Usage of Low ProT Regulatory Elements

We next explored the ProT dependency landscape of different TFs to reveal the possible impact of DNA structure on the function of general vs cell-fate-determining TFs. Toward this, we looked for systems that involve dynamic reprogramming of cell-fate using defined TFs. Somatic cells can be efficiently reprogrammed into an embryonic stem cell state, i.e. induced pluripotent stem cells (iPSCs), using a distinct set of TFs, namely Oct4, Sox2, Klf4, and c-Myc (Takahashi and Yamanaka, 2006). Therefore, using datasets from a previous study we analyzed binding of these four TFs during fibroblast-to-iPSC reprogramming and assessed its relation to ProT levels and nucleosome occupancy in pre-induced human fibroblasts, as measured by MNase sequencing (MNase-seq) (Chronis et al., 2017). Interestingly, c-Myc showed lesser affinity for nucleosome bound regions, whereas Oct4, Sox2, and Klf4 preferentially targeted nucleosome-enriched sites (Figure 5A). These data are consistent with a previous finding that Oct4, Sox2, and Klf4, but not c-Myc, could function as pioneer TFs during reprogramming by virtue of their ability to target "closed" chromatin sites (Soufi et al., 2015). Furthermore, while c-Myc, Klf4 and Sox2 ChIP-seq enrichment relied on levels of ProT, this was less so in the case of Oct4.

Intrigued by these findings, we analyzed the rewiring of the active chromatin landscape using H3K27ac mark and its relationship with Oct4 binding dynamics during distinct stages of reprogramming. Strikingly, although H3K27ac sites in either fibroblasts or iPSCs show stable ProT levels, those occurring during any of the transient states during reprogramming show a significant drop in their ProT levels (Figure 5B). This pattern was closely mimicked by genomic regions targeted by Oct4 at distinct stages of reprogramming (Figure 5B). These results suggest that while Oct4 binding and enhancer activation occurs at "low ProT"

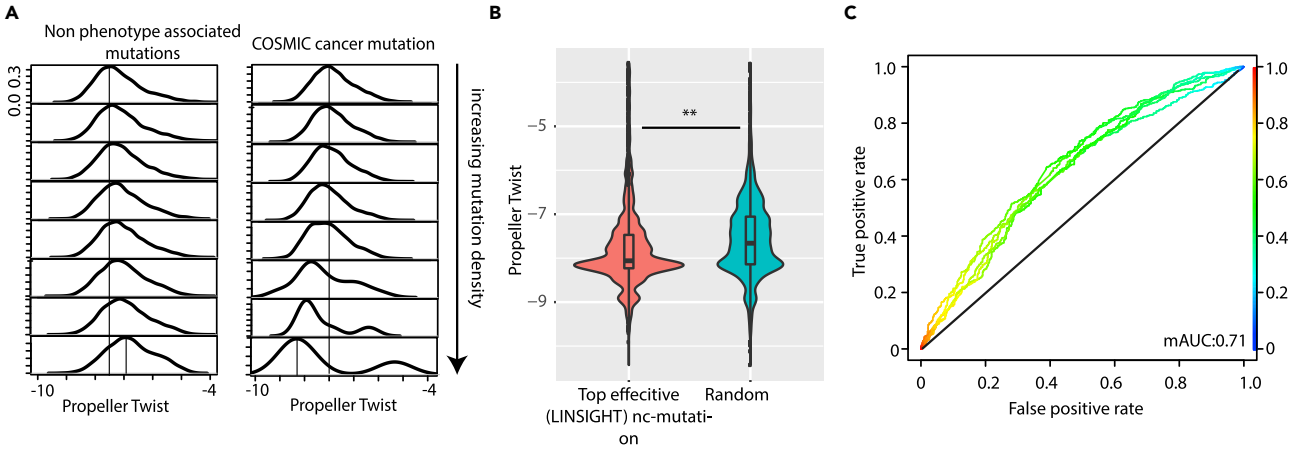


Figure 6. Functional Implications of Differential ProT Levels in the Genome

(A) Left: density plot depicting ProT density across eight categories of genomic loci (1000 bp) classified into increasing mutation density of non-phenotype associated mutation class. Right: same as Left, but for mutations associated to cancer in COSMIC database (Forbes et al., 2017).

(B) Violin-boxplot for effectivity of genomic loci to be effective in terms of phenotype as determined by LINSIGHT database (Huang et al., 2017) in high and low ProT classes.

(C) SVM models accuracy plot as ROC curve depicting efficient classification of genomic classes as effective or otherwise as determined LINSIGHT.

See also Figure S5.

regions in transient cell states occurring during reprogramming, the acquisition of a fully reprogrammed cell-fate involves utilization of "average ProT" sites as enhancers (Figure 5B). Altogether, these findings imply that although high ProT sites are hallmark of enhancers in defined cell types, relatively lower ProT sites may play a crucial role during setting up of these cell-fates during reprogramming and potentially in development.

ProT Levels Correlate with Occurrence and Phenotypic Consequences of Mutations

Given our findings of a deterministic role of ProT in the regulatory potential of distinct genomic sites, we next assessed the differential sensitivity of ProT sites to tolerate mutations. Strikingly, our analysis revealed a higher occurrence of random mutations (i.e. non-phenotype associated) at "high ProT" regions of the genome (Figures 6A and S5). This implies that inherent higher DNA accessibility plays a critical role in enhancing mutability, possibly because mutagenic agents or machinery have an easier access to such sites. In contrast, the occurrence of cancer-associated mutations (i.e. phenotype associated) is higher in genomic loci characterized by a lower ProT (Figures 6A and S5). It is likely that these low ProT sites offer reduced access to DNA repair machineries and consequently more likely to result in phenotypic consequences.

We further employed LINSIGHT predictions to determine effectivity of SNPs in causing a detectable phenotype (Huang et al., 2017). We analyzed genomic segments in 1 kb bins and found that ProT levels are lower at phenotype-associated SNPs than at random genomic loci, in line with our previous observations (Figure 6B). We further implemented SVM models trained on ProT information for 2 kbp genomic segments to classify effective versus random genomic loci in terms of phenotypic association. Strikingly, the generated models were able to predict whether mutation at a specific locus could be phenotypic with a considerably high accuracy (Figure 6C). Taken together, we show that ProT levels are correlated with occurrence and phenotypic consequences of mutations.

DISCUSSION

The DNA sequence composition is known to influence local DNA shape across the genome (Parker et al., 2009). However, at a broader scale, the DNA sequence and structures appear as independent and deconvolved features (Abe et al., 2015). Previous studies have hinted upon a role of DNA sequence as well as topography in determining certain epigenetic features (Arnold et al., 2017; Parker et al., 2009; Rusk, 2014; Wang and Moazed, 2017). The genetic features were also shown to be important in determining TF access on the DNA (Yang et al., 2017b; Zhou et al., 2015). However, whether the local DNA biophysical features play any role in a chromatin context and in gene regulation is unknown. This study provides

unprecedented insights into the deterministic role of DNA biophysical features in governing epigenetic and gene regulatory landscape underlying cell identity and function.

Our study has discovered ProT as a novel proxy for measuring inherent surface accessibility as determined by OH-radical cleavage mapping and local DNA flexibility as probed by atomic force microscopy. Further assessment of distribution of different ProT sequences within the 3D nuclear space revealed that higher ProT regions are enriched in euchromatic domains and are more interiorly located in 3D genome structure. In contrast, lower ProT regions are enriched in heterochromatic domains and are more exterior in their location within the 3D genome structure. These intriguing results implied a novel role for ProT levels of DNA sequences in guiding DNA surface accessibility and flexibility, epigenetic state, and ultimately the nuclear organization of distinct chromatin domains. As LADs are known to be AT-rich, future studies should attempt to dissect any contribution of DNA sequence versus DNA shape to our observations (van Steensel and Belmont, 2017). Importantly further, although our analysis found ProT to positively correlate with the DNA surface accessibility, HelT showed negative correlation. Further investigation is required to uncover the relevance of this anti-correlation, in particular towards the gene regulatory landscape.

A large scale, systematic analysis showed specific enrichment of ProT at the regulatory elements that were previously identified by a number of independent experimental measures in multiple cell-types and across species. Importantly, further, the regulatory potential of the genomic regions showed a strong correlation with ProT levels. Additional analysis revealed that ProT is a deterministic feature of the enhancers irrespective of the chromatin mark used for their identification. Strikingly, the predictive power of ProT to identify sequence-intrinsic enhancer features, as experimentally measured by STARR-seq (Arnold et al., 2013; Muerdter et al., 2015), was very high, suggesting that ProT predictions are able to decode sequence logic with confidence in such experiments and offers an alternative to heavy experimentation-based analysis. Collectively, our findings argue that the local DNA biophysical features hold the potential to prime a genomic region for particular epigenetic state and gene regulatory potential, thus revealing an underestimated role of DNA structure in guiding genome function. This is further in line with previous studies that have shown that the DNAshape algorithm works better than certain k-mer (2,3) combinations for some biological functions (Abe et al., 2015; Mathelier et al., 2016).

ProT, along with other DNA shape features, have been shown to contribute to TF access DNA (Yang et al., 2017b; Zhou et al., 2015). Our analysis shows that the activator and repressor TFs have an inverse binding affinity with ProT levels. Importantly, further, in contrast to general TFs, pioneer TFs have lesser dependency on these DNA features for their binding. A number of efficient reprogramming TFs are known to be pioneer TFs (Pataskar et al., 2016; Soufi et al., 2015). Interestingly, employing pioneer TF binding and H3K27ac data during distinct stages of cellular reprogramming of a differentiated to a pluripotent state, we find that the switch between cell-fates involves a transient usage of low ProT as regulatory elements. In further support of this highly influential role of ProT in guiding gene regulation, DNA flexibility, and surface accessibility, we found a higher occurrence of random or non-phenotype-associated mutations at "high ProT" regions of the genome, whereas the occurrence of phenotype-associated mutation is higher at lower ProT genomic regions, which we identified to mark cell-identity enhancers. This implies that high ProT-imposed increased DNA accessibility plays a critical role in enhancing mutability in the genome, potentially because mutagenic agents or machinery have an easier access to such sites. A previous study has suggested that the GC-rich sequences tend to have less pronounced ProT values, whereas AT-rich sequences tend to have more negative ProT values (Hancock et al., 2013). It is thus also likely that the observed relationship to the mutation rate is also influenced to certain extent by the sequence composition and warrants further investigation.

Altogether, our work provides unprecedented insights into the gene regulatory landscape encoded in the DNA biophysical features. Our findings open a new area of investigation that was previously underestimated for its relevance and vouches for the necessity to include DNA shape features while studying epigenetic gene regulatory mechanisms in various contexts. Our study hypothesizes that DNA sequences have evolved in a highly orchestral manner, wherein genomic DNA is segmented into compartments of different inherent biophysical states, which are then chosen to be in different nuclear chromatin compartments of different regulatory potential. Follow-up studies should aim to investigate how epigenetic machineries such as DNA modifying enzymes alter the DNA structure at specific sites. Furthermore, it will also be important to determine how combinatorial TF binding influences DNA flexibility and if this crosstalk is relevant for

cell-fate decisions during development. In addition, it will be interesting to investigate how genomic sites with various ProT levels are utilized in the 3D nuclear space in response to external cues during development and in various diseases. Such investigations will ultimately decode the relationship between genetic and epigenetic mechanisms, which is key for a comprehensive understanding of genome function in health and disease.

Limitations of the Study

DNAshape algorithm predicts shape from DNA sequence, and hence it will show high correlation to the sequence. Given this close dependency, it is almost impossible to fully dissect the contribution of DNA sequence versus shape. This limitation is valid for several previous publications using DNAshape algorithm as well as our own study.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2019.10.055.

ACKNOWLEDGMENTS

We would like to especially thank Dr. Michael Stadler (FMI, Switzerland) and Dr. Attila Nemeth for assistance and critical feedback to this study. We would also like to thank the members of the Tiwari lab for their cooperation and critical feedback during this study. We also thank Tim Liedl and Joachim Rädler for access to the AFM imaging lab. This study was supported by the Deutsche Forschungsgemeinschaft TI 799/1-3 to V.K.T. and Deutsche Forschungsgemeinschaft SFB 863, project A11 to J.L..

AUTHOR CONTRIBUTIONS

A.P. analyzed data and wrote the manuscript. W.V and J.E. acquired and analyzed A.F.M. images. A.S. helped during revision of the manuscript. J.L. provided suggestions for A.F.M. experiments and wrote the manuscript. V.T. designed the study, analyzed data, and wrote the manuscript. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing financial interests.

Received: August 11, 2019
Revised: October 20, 2019
Accepted: October 24, 2019
Published: November 22, 2019

REFERENCES

Abe, N., Dror, I., Yang, L., Slattery, M., Zhou, T., Bussemaker, H.J., Rohs, R., and Mann, R.S. (2015). Deconvolving the recognition of DNA shape from sequence. Cell *161*, 307–318.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

Arnold, C.D., Gerlach, D., Stelzer, C., Boryn, L.M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science *339*, 1074–1077.

Arnold, C.D., Zabidi, M.A., Pagani, M., Rath, M., Schernhuber, K., Kazmar, T., and Stark, A. (2017). Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. Nat. Biotechnol. *35*, 136–144.

Bell, O., Tiwari, V.K., Thoma, N.H., and Schubeler, D. (2011). Determinants and dynamics of genome accessibility. Nat. Rev. Genet. *12*, 554–564.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315–326.

Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J., and Plath, K. (2017). Cooperative binding of transcription factors orchestrates reprogramming. Cell *168*, 442–459.e20.

Colbran, L.L., Chen, L., and Capra, J.A. (2017). Short DNA sequence patterns accurately identify broadly active human enhancers. BMC Genomics *18*, 536.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods *9*, 215–216.

Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L., et al. (2017). COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. *45* (D1), D777–D783.

Greenbaum, J.A., Pang, B., and Tullius, T.D. (2007). Construction of a genome-scale structural map at single-nucleotide resolution. Genome Res. *17*, 947–953.

Hancock, S.P., Ghane, T., Cascio, D., Rohs, R., Di Felice, R., and Johnson, R.C. (2013). Control of DNA minor groove width and Fis protein binding by the purine 2-amino group. Nucleic Acids Res. *41*, 6750–6760.

Heinz, S., Romanoski, C.E., Benner, C., and Glass, C.K. (2015). The selection and function of cell type-specific enhancers. Nat. Rev. Mol. Cell Biol. *16*, 144–154.

Huang, Y.F., Gulko, B., and Siepel, A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat. Genet. *49*, 618–624.

Karlic, R., Chung, H.R., Lasserre, J., Vlahovicek, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. Proc. Natl. Acad. Sci. U S A *107*, 2926–2931.

Kleftogiannis, D., Kalnis, P., and Bajic, V.B. (2016). Progress and challenges in bioinformatics approaches for enhancer identification. Brief. Bioinform. *17*, 967–979.

Lee, D., Karchin, R., and Beer, M.A. (2011). Discriminative prediction of mammalian enhancers from DNA sequence. Genome Res. *21*, 2167–2180.

Li, B., Carey, M., and Workman, J.L. (2007). The role of chromatin during transcription. Cell *128*, 707–719.

Li, P., Marshall, L., Oh, G., Jakubowski, J.L., Groot, D., He, Y., Wang, T., Petronis, A., and Labrie, V. (2019). Epigenetic dysregulation of enhancers in neurons is associated with Alzheimer's disease pathology and cognitive symptoms. Nat. Commun. *10*, 2246.

Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F., and Schubeler, D. (2011). Identification of genetic elements that autonomously determine DNA methylation states. Nat. Genet. *43*, 1091–1097.

Long, H.K., Prescott, S.L., and Wysocka, J. (2016). Ever-changing landscapes: transcriptional enhancers in development and evolution. Cell *167*, 1170–1187.

Ma, W., Yang, L., Rohs, R., and Noble, W.S. (2017). DNA sequence+shape kernel enables alignment-free modeling of transcription factor binding. Bioinformatics *33*, 3003–3010.

Margueron, R., and Reinberg, D. (2010). Chromatin structure and the inheritance of epigenetic information. Nat. Rev. Genet. *11*, 285–296.

Mathelier, A., Xin, B., Chiu, T.P., Yang, L., Rohs, R., and Wasserman, W.W. (2016). DNA shape features improve transcription factor binding site predictions in vivo. Cell Syst. *3*, 278–286.e4.

Mazur, A.K., and Maaloum, M. (2014). Atomic force microscopy study of DNA flexibility on short length scales: smooth bending versus kinking. Nucleic Acids Res. *42*, 14006–14012.

Moris, N., Pina, C., and Arias, A.M. (2016). Transition states and cell fate decisions in epigenetic landscapes. Nat. Rev. Genet. *17*, 693–703.

Muerdter, F., Boryn, L.M., and Arnold, C.D. (2015). STARR-seq - principles and applications. Genomics *106*, 145–150.

Ong, C.T., and Corces, V.G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat. Rev. Genet. *12*, 283–293.

Parker, S.C., Hansen, L., Abaan, H.O., Tullius, T.D., and Margulies, E.H. (2009). Local DNA topography correlates with functional noncoding regions of the human genome. Science *324*, 389–392.

Pataskar, A., Jung, J., Smialowski, P., Noack, F., Calegari, F., Straub, T., and Tiwari, V.K. (2016). NeuroD1 reprograms chromatin and transcription factor landscapes to induce the neuronal program. EMBO J. *35*, 24–45.

Pennacchio, L.A., Bickmore, W., Dean, A., Nobrega, M.A., and Bejerano, G. (2013). Enhancers: five essential questions. Nat. Rev. Genet. *14*, 288–295.

Pradeepa, M.M., Grimes, G.R., Kumar, Y., Olley, G., Taylor, G.C., Schneider, R., and Bickmore, W.A. (2016). Histone H3 globular domain acetylation identifies a new class of enhancers. Nat. Genet. *48*, 681–686.

Rickels, R., and Shilatifard, A. (2018). Enhancer logic and mechanics in development and disease. Trends Cell Biol. *28*, 608–630.

Rivetti, C., Guthold, M., and Bustamante, C. (1996). Scanning force microscopy of DNA deposited onto mica: equilibration versus kinetic trapping studied by statistical polymer chain analysis. J. Mol. Biol. *264*, 919–932.

Rusk, N. (2014). Capturing promoter-enhancer interactions in high throughput. Nat. Methods *11*, 231.

Sakabe, N.J., Savic, D., and Nobrega, M.A. (2012). Transcriptional enhancers in development and disease. Genome Biol. *13*, 238.

Schoenfelder, S., and Fraser, P. (2019). Long-range enhancer-promoter contacts in gene expression control. Nat. Rev. Genet. *20*, 437–455.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nat. Rev. Genet. *15*, 272–286.

Soufi, A., Garcia, M.F., Jaroszewicz, A., Osman, N., Pellegrini, M., and Zaret, K.S. (2015). Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. Cell *161*, 555–568.

Spitz, F., and Furlong, E.E. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. *13*, 613–626.

Stevens, T.J., Lando, D., Basu, S., Atkinson, L.P., Cao, Y., Lee, S.F., Leeb, M., Wohlfahrt, K.J., Boucher, W., O'Shaughnessy-Kirwan, A., et al. (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature *544*, 59–64.

Taberlay, P.C., Kelly, T.K., Liu, C.C., You, J.S., De Carvalho, D.D., Miranda, T.B., Zhou, X.J., Liang, G., and Jones, P.A. (2011). Polycomb-repressed genes have permissive enhancers that initiate reprogramming. Cell *147*, 1283–1294.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663–676.

van Steensel, B., and Belmont, A.S. (2017). Lamina-associated domains: links with chromosome architecture, heterochromatin, and gene repression. Cell *169*, 780–791.

Vanhille, L., Griffon, A., Maqbool, M.A., Zacarias-Cabeza, J., Dao, L.T., Fernandez, N., Ballester, B., Andrau, J.C., and Spicuglia, S. (2015). High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. Nat. Commun. *6*, 6905.

Wang, X., and Moazed, D. (2017). DNA sequence-dependent epigenetic inheritance of gene silencing and histone H3K9 methylation. Science *356*, 88–91.

Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. Nat. Genet. *46*, 1160–1165.

Wiggins, P.A., van der Heijden, T., Moreno-Herrero, F., Spakowitz, A., Phillips, R., Widom, J., Dekker, C., and Nelson, P.C. (2006). High flexibility of DNA on short length scales probed by atomic force microscopy. Nat. Nanotechnol. *1*, 137–141.

Yanez-Cuna, J.O., Kvon, E.Z., and Stark, A. (2013). Deciphering the transcriptional cis-regulatory code. Trends Genet. *29*, 11–22.

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

Yang, B., Liu, F., Ren, C., Ouyang, Z., Xie, Z., Bo, X., and Shu, W. (2017a). BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone. Bioinformatics *33*, 1930–1936.

Yang, L., Orenstein, Y., Jolma, A., Yin, Y., Taipale, J., Shamir, R., and Rohs, R. (2017b). Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. Mol. Syst. Biol. *13*, 910.

Zentner, G.E., and Henikoff, S. (2013). Regulation of nucleosome dynamics by histone modifications. Nat. Struct. Mol. Biol. *20*, 259–266.

Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A.C., Ghane, T., Di Felice, R., and Rohs, R. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res. *41*, W56–W62.

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R.S., Bussemaker, H.J., Gordan, R., and Rohs, R. (2015). Quantitative modeling of transcription factor binding specificities using DNA shape. Proc. Natl. Acad. Sci. U S A *112*, 4654–4659.

ISCI, Volume *21*

Supplemental Information

Deciphering the Gene Regulatory Landscape Encoded in DNA Biophysical Features

Abhijeet Pataskar, Willem Vanderlinden, Johannes Emmerig, Aditi Singh, Jan Lipfert, and Vijay K. Tiwari

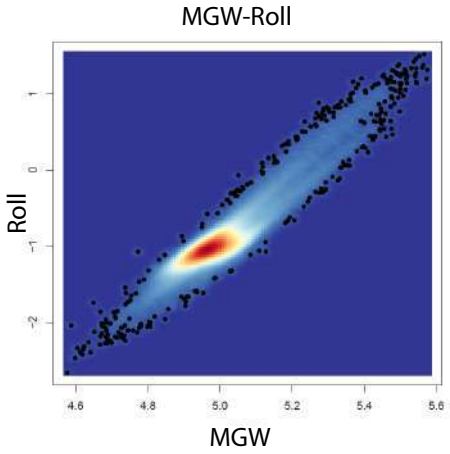
SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1: Related to Figure 1.

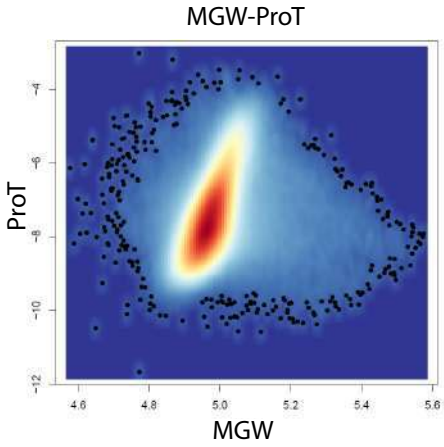
(A-C) Genome-wide correlations of DNA shape features; Major Groove Width (MGW) and Roll (A), MGW and Propeller Twist (ProT,B) and ProT and Roll (C) (D) Left: Measurements of Contour length distributions; schema (top) and Fitted regression line of contour length distributions (below) Mid: Density plot of contour length distribution in low ProT control (above) and high ProT construct (below). Right: Schema for measurement of contour length with length scales (5nm).

Figure S1

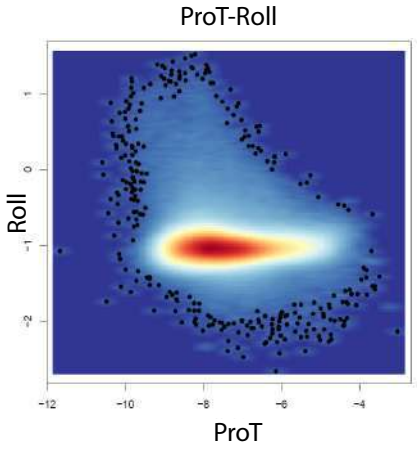
A



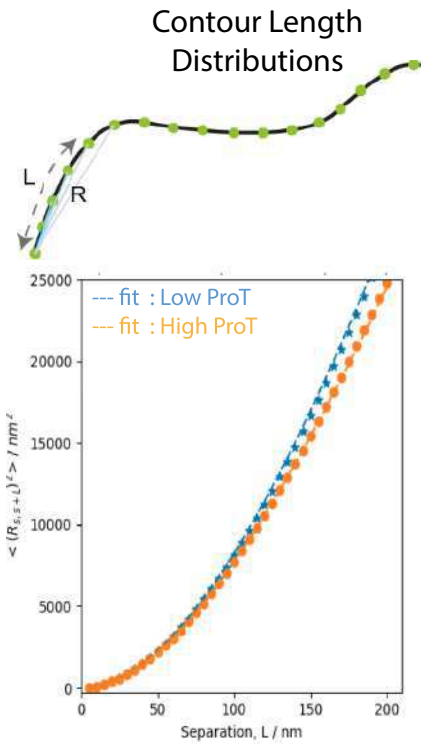
B

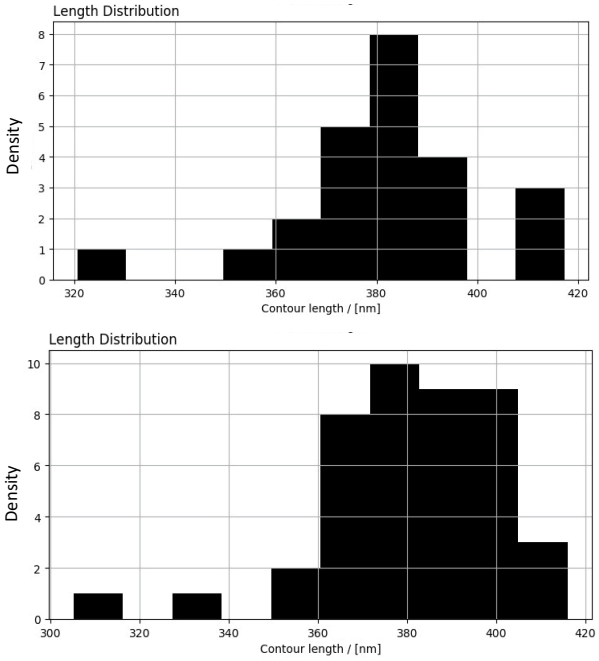


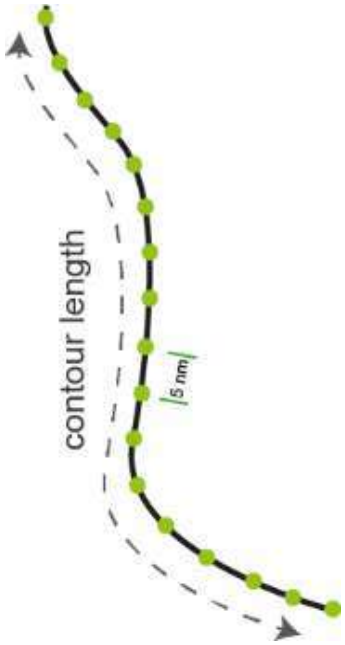
C



D





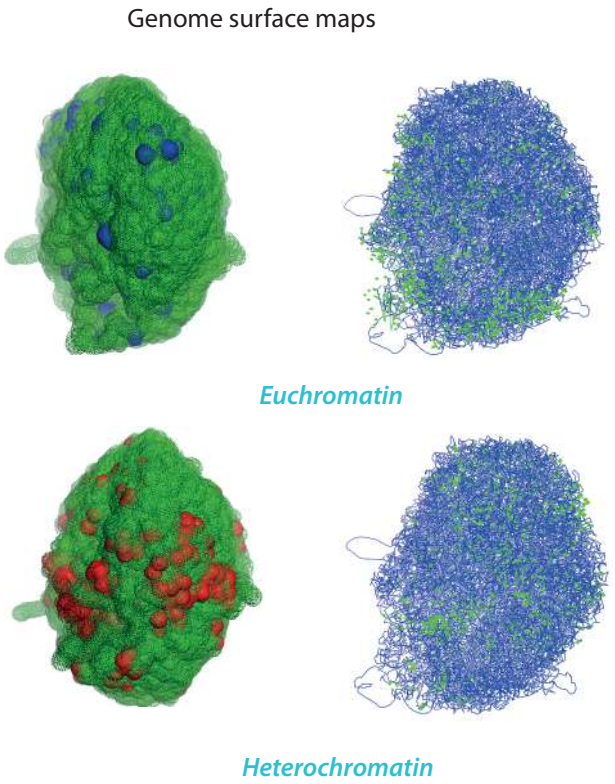


Supplementary Figure 2: Related to Figure 2.

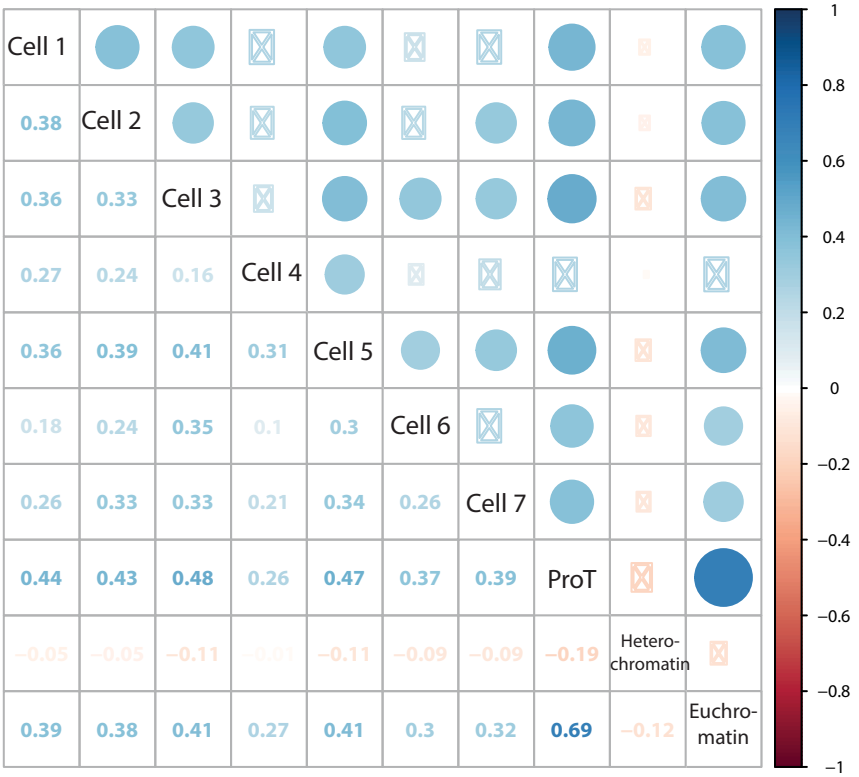
(A) Genome Surface maps highlighted with Euchromatin in blue (above) and Heterochromatin in red (below) derived from the reconstructed genome structure from the single cell HiC experiments in mES cells. (B) Correlation scatter matrix with correlation coefficients (numbers) derived genome-wide from the information of surface depths in all analyzed 7 cells, ProT, Euchromatin and Heterochromatin density. (C) Cross correlation scatter plot from the surface depths of genomic loci from seven different reconstructed single cell genome structures of mES cells (D) Line plot depicting linear profile of surface depths (black) and ProT (red) across lengths of each chromosomes in mouse ES cells.

Figure S2

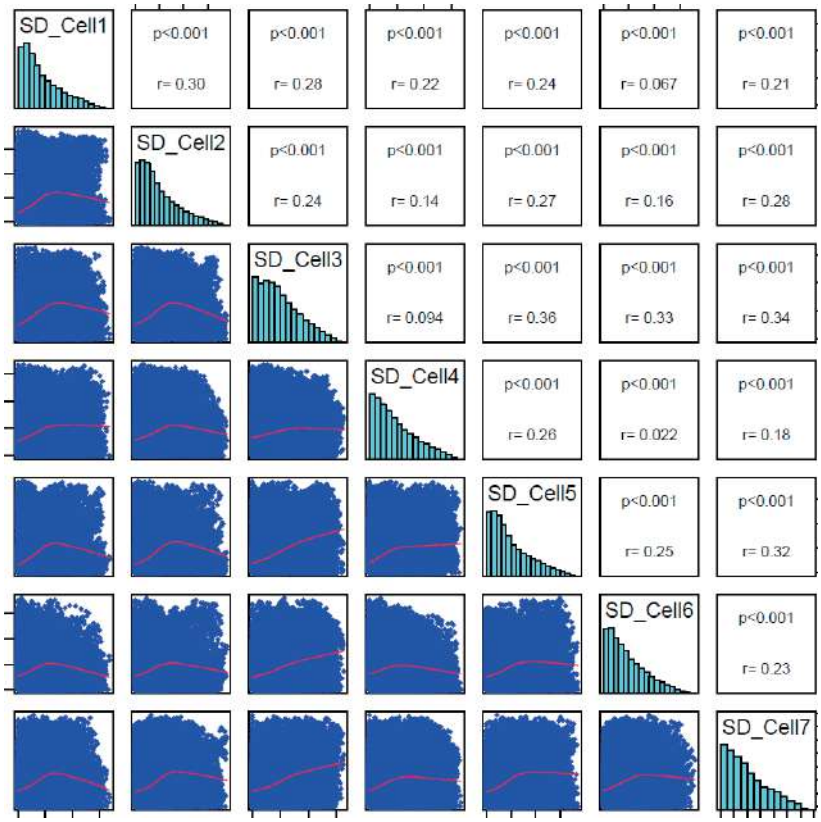
A



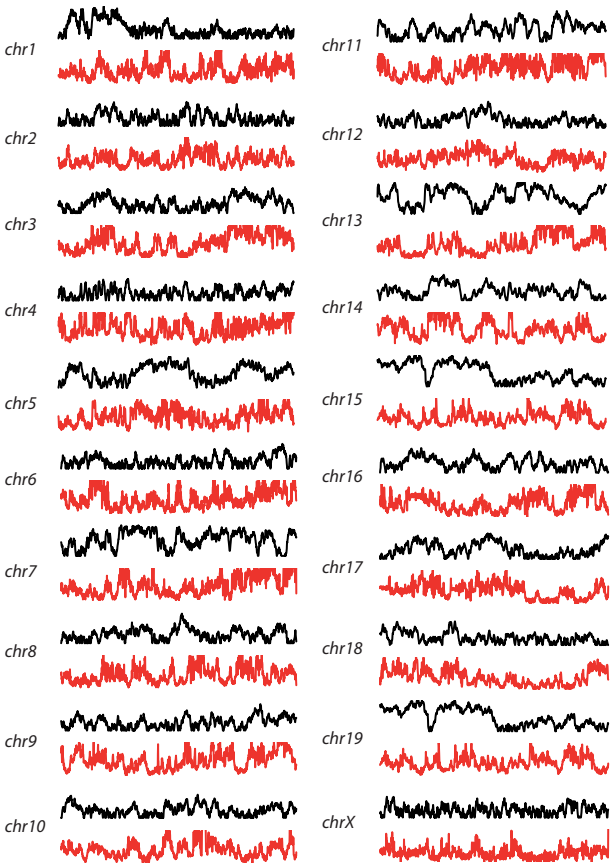
B



C



D

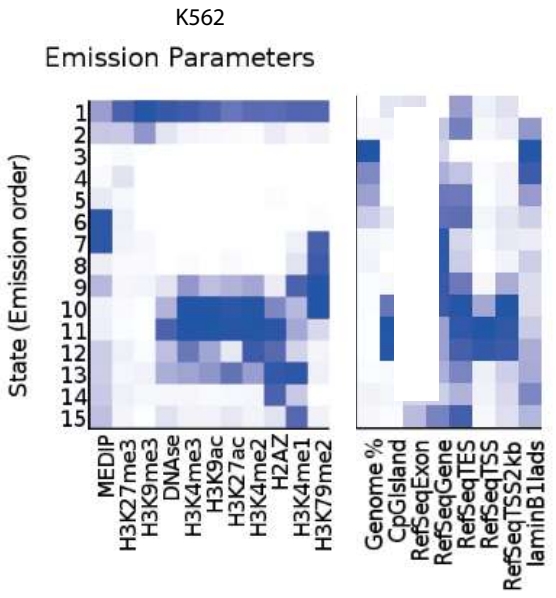


Supplementary Figure 3: Related to Figure 3.

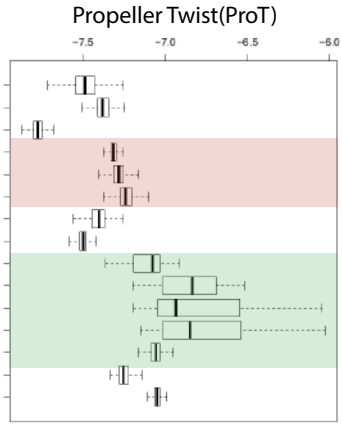
(A) ChromHMM based clustering of genome into segments with 15 different chromatin states. (B) Propeller Twist (ProT) levels depicted as box-plots in each of these 15 different chromatin states. Highlights: red- heterochromatin, green-euchromatin. (C) Line plot depicting genome-wide correlation coefficients of ProT with different chromatin marks. (D) Above: Example scatter plots from two cell types with enhancer strength as measured by CAGE experiment on X-axis and ProT on y-axis. Below: Same information depicted as regression line plot for ProT when enhancers are arranged in increasing order of strength. (E) ProT levels plotted as density plot to -50 KB to +50 KB of TSS and TES respectively of genes classified into four quartiles by expression value in Human myeloma K562 cells.

Figure S3

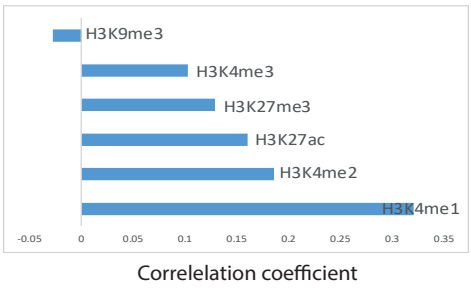
A



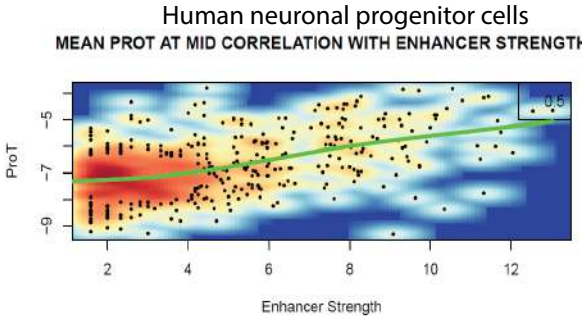
B

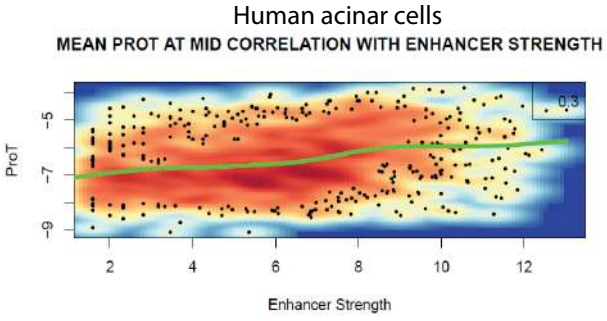


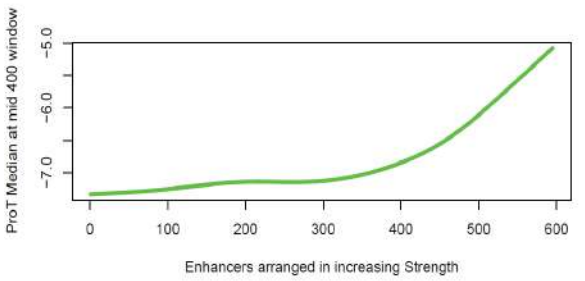
C

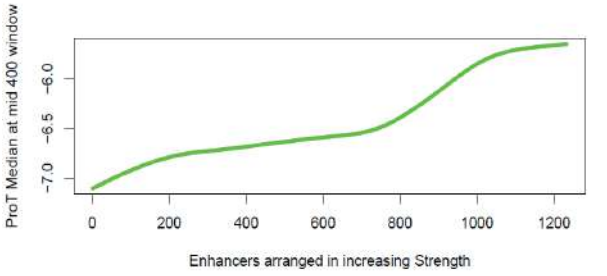


D

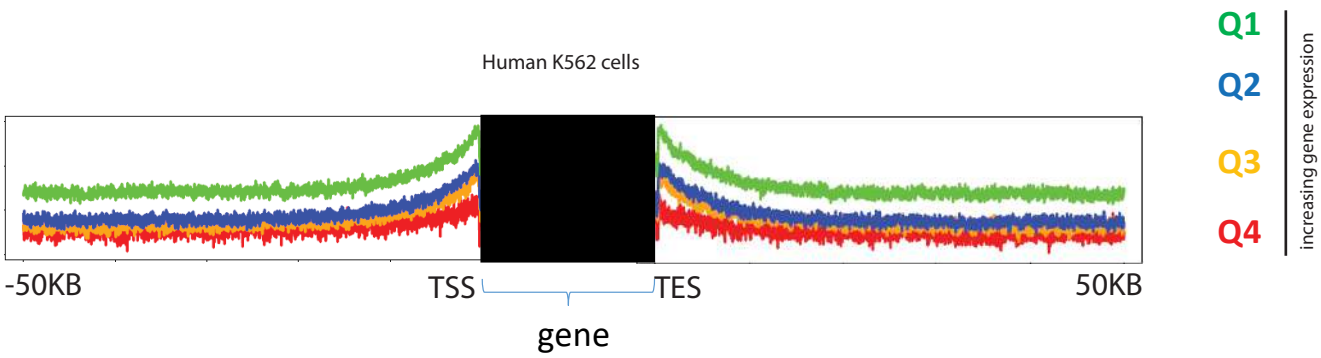








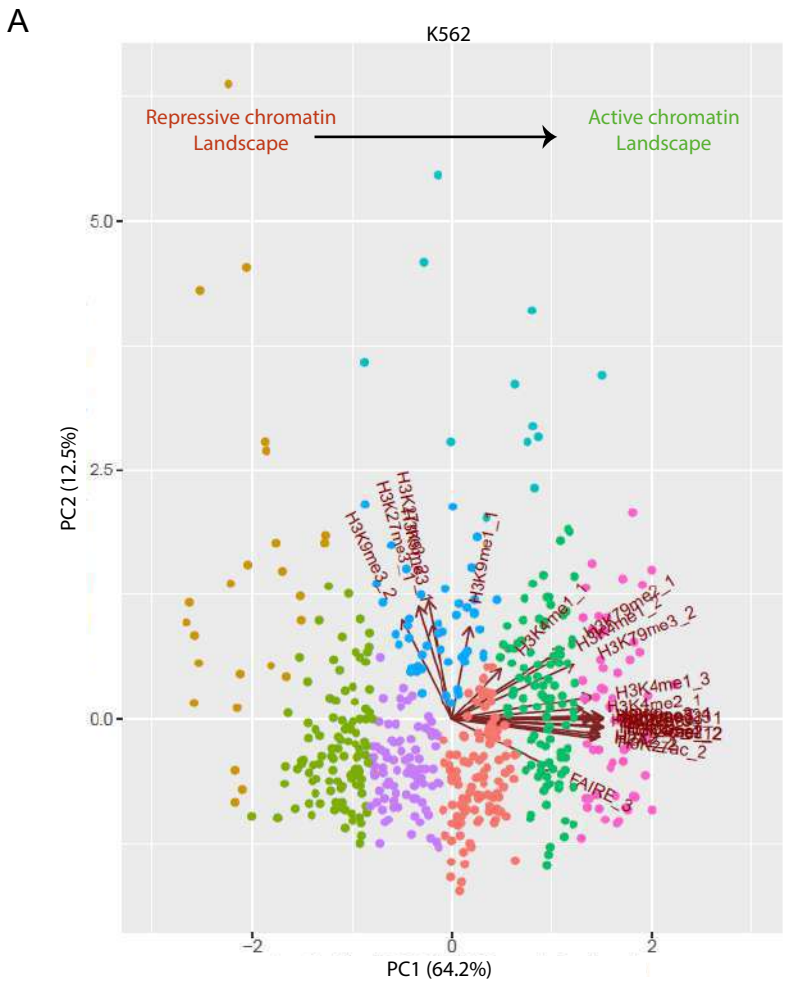
E

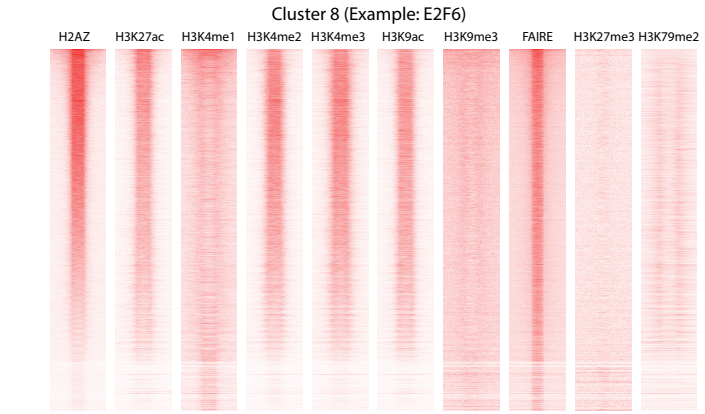


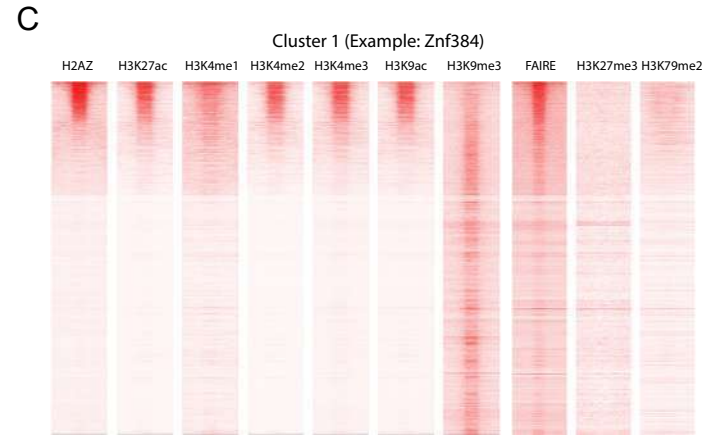
Supplementary Figure 4: Related to Figure 4.

(A) PCA plot depicted distribution of Transcription factor binding enrichment over chromatin marks, thereby clustered into 8 different groups colored differently. Vectors represent the contributing features for the PCA distribution. PCA overall classifies transcription factors on PC1 with enrichment of factors on repressive chromatin landscape on X-axis and on active chromatin landscape on Y-axis. Data acquired from K562 cells. (B-C) Example of cluster 8 (B) and cluster 1 (C) transcription factors where enrichment of different chromatin marks.

Figure S4


A
K562
Repressive chromatin Landscape
Active chromatin Landscape
PC2 (12.5%)
PC1 (64.2%)
5.0
2.5
0.0
−2
0
2
H3K9me1_1
H3K4me1_1
H3K79me3_2
H3K4me1_3
FAIRE_3

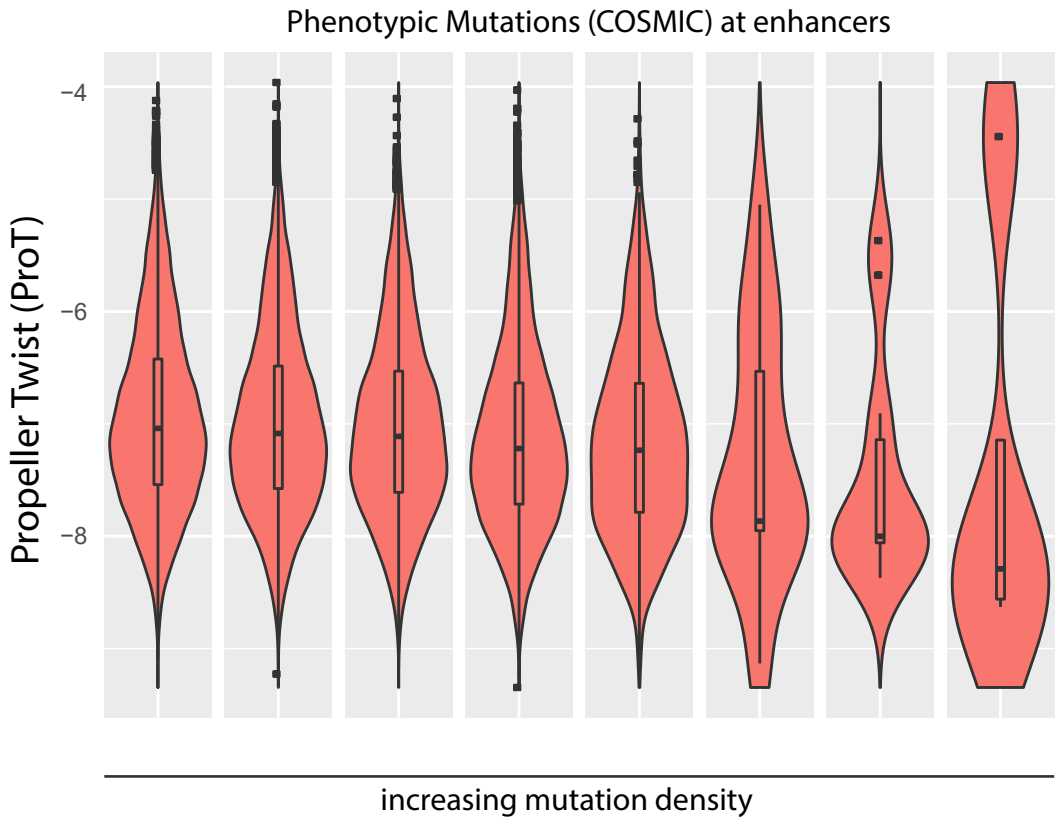

B
Cluster 8 (Example: E2F6)
H2AZ
H3K27ac
H3K4me1
H3K4me2
H3K4me3
H3K9ac
H3K9me3
FAIRE
H3K27me3
H3K79me2


C
Cluster 1 (Example: Znf384)
H2AZ
H3K27ac
H3K4me1
H3K4me2
H3K4me3
H3K9ac
H3K9me3
FAIRE
H3K27me3
H3K79me2

Supplementary Figure 5: Related to Figure 6.

Violin-box plots depicting ProT levels at distinct enhancer regions arranged in the increasing order of occurrence of phenotypic (COSMIC) mutations.

Figure S5


Phenotypic Mutations (COSMIC) at enhancers
Propeller Twist (ProT)
−4
−6
−8
increasing mutation density

TRANSPARENT METHODS

Preparation of DNA constructs for AFM imaging

DNA sequences (1167 bp long) were generated by polymerase chain reaction using Phusion master mix (2x Phusion High-Fidelity PCR Master Mix, Thermo Fisher Scientific Inc., Waltham, MA, USA) using plasmid pUC57 with sequences of interest introduced at HindIII site (biocat) as a template and using the sequences CCC AGT CAC GAC GTT GTA AAA CG and AGC GGA TAA CAA TTT CAC ACA GG as forward and reverse primers respectively. PCR products were purified using a PCR cleanupCleanup kit and stored in Tris-EDTA buffer (10 mM Tris- HCl, pH = 8.0; 1 mM EDTA) at 4° C. To verify successful amplification of the desired sequence, the product was analyzed on a 1 % (w/v) agarose gel.

AFM imaging

A buffer solution (12 mM $MgCl_2$; 10 mM Tris-HCl, pH = 8.0) containing 1167 bp linear DNA construct at 0.5 ng/ L was deposed onto freshly cleaved mica by drop-casting for 30 seconds, followed by gentle rinsing with milliQ water (25 mL) and drying under N^2. AFM imaging was performed on a Multimode AFM equipped with a Nanoscope III controller (Digital Instruments) and a type E scanner (Bruker). Images were recorded on dried samples, under ambient conditions, and using silicon cantilevers (Nanosensor; SSS-NCHR; resonance frequency ≈ 300 kHz). Typical scans were recorded at 1–3 Hz line frequency, with optimized feedback parameters and 512 × 512 pixels of 1.9 nm each.

AFM data analysis

AFM topography images were loaded in Scanning Probe Imaging Processor software (v6.4) and background corrected using global fitting with a second order polynomial and in a line-by-line fashion using the histogram alignment routine. The processed images were saved in ASCII format and analyzed using custom-written code implemented in Python framework. Image analysis involves finding (x,y) coordinates that define the DNA curvilinear length by tracing the molecular contour with a step-length l = 5 nm, following the routine introduced by Wiggins et al. (Wiggins et al., 2006). Bend angles are defined as the deviation from linearity between a certain set of tangent vectors separated by l. In total, we traced 87168 bend angles from 1226 imaged DNA molecules. The energy landscape for bending was reconstructed from the corresponding bend angle distribution by taking the negative logarithm. The energy required to introduce a bend increases with the bend angle θ. Up to θ ~ 1 rad the increase is approximately quadratic and in agreement with the prediction of the worm-like chain model $E_{WLC}(\theta) = 1/2 \cdot k_B T \cdot (P/l) \cdot \theta^2$, with P ~ 55 nm (Figure 1C, dashed line), where k_B is the Boltzmann constant, T the absolute temperature (295 K), and l the segment

length (5 nm in our analysis). For bending angles θ > 1 rad, the energy to bend the DNA grows more slowly -in other words large bends occur more frequently- than predicted by the WLC model, as has been reported previously (Wiggins et al., 2006). While both the control and high ProT sequences deviate from the WLC prediction, the energy required to introduce a large bend for the high ProT sequences is lower and, therefore, deviates more strongly from the WLC prediction compared to the control sequences.

Further, the (x,y) coordinates are used to quantify the end-to-end distances *R* between any two traced positions as a function of their separation *L* along the contour. In turn, the relation between the mean squared end-to-end distance $\langle R^2 \rangle$ and *L* can be used to examine surface equilibration using the formula $\langle R^2 \rangle = 4\,P\,L\,(1 - 2\,P/L\,(1 - \exp(-L\,/\,2P)))$ with *P* the bending persistence length.

DNAshape and OH-radical cleavage prediction

DNAshapeR package (Chiu et al., 2016) was used to generate and plot DNA shape feature predictions, namely Propeller Twist, Major Groove Width, Helix Turn and Roll. Density plots of these features were plotted using DNAShapeR (Chiu et al., 2016). OH-radical cleavage predictions were obtained from ORCHID2 (Greenbaum et al., 2007) server. For genome-wide predictions of DNAshape features, bigwig files (mm9 and hg19 for mouse and human respectively) were downloaded from GBShape (Chiu et al., 2015). Correlation within DNAshape (Chiu et al., 2015) features and with OH-radical cleavage intensity (ORCHID) were done by using Deeptools (Ramirez et al., 2014) at the resolution of 1KB.

Reconstruction of 3D Genome Structures and overlay with Chromatin features and Propeller Twist predictions

HDF5 Datafiles containing genome structure features reconstructed from single-cell HiC experiments of seven mouse ES cells were downloaded from supplementary information provided in study from Stevens et al (Stevens et al., 2017). The surface depths with resolution of 1MB were analyzed from these HDF5 files. Euchromatin segments were defined as those 1 MB segments that shows more than 10 peaks of H3K27ac chromatin marks, while heterochromatin marks showed enrichment of more than 10 peaks of H3K9me3. Propeller Twist predictions at resolution of 1MB in mouse genome (mm9) was analyzed from BigWig files downloaded from GBShape (Chiu et al., 2015) using Deeptools (Ramirez et al., 2014). For visualization of 3D genome and chromosome structures, PDB files were downloaded from supplementary information provided in study from Stevens et al.(Stevens et al., 2017). The PDB files visualization and programming was done using Pymol. Overlay of chromatin states, surface depth and propeller Twist quartiles was done by

processing PDB files in R and visualization in Pymol. Scripts used for processing are available upon request.

Analysis of CAGE and STARR-seq enhancers

capSTARR-seq data was downloaded from the supplementary data published in the study from Vanhille et al(Vanhille et al., 2015). CAGE Enhancer analysis was downloaded from FANTOM5 Atlas (Andersson et al., 2014). Propeller Twist profiles were overlaid using mean across BED file, centered at mid, using DNAshapeR (Chiu et al., 2016).

Histone modification ChIPseq analysis

Following ChIP-seq were processed in this manner: mES H3K27ac, mES H3K9me3, K562 H3K27ac. The ChIP-sequencing output in FASTQ format was subjected to a quality check using FASTQC v2.6.14 (Andrews). Bowtie v0.12.9 (Langmead, 2010) was used to align the reads uniquely, i.e., each read was maximally aligned to one position, to mm9 genome with UCSC annotations (Rosenbloom et al., 2015). The alignment output files from two biological replicates were merged together after checking for correlations across the replicates using the SAMTOOLS v0.1.19 (Li et al., 2009) merge function. SAMTOOLS v0.1.19 (Li et al., 2009) was used for the alignment file format conversions and sorting of alignment output files. The WIGGLE files for the alignment files were generated using QuasR package (Gaidatzis et al., 2015). The peaks were computed without providing input with MACS v2.0.10.20120913 (Zhang et al., 2008) using the default parameters. The enrichment was calculated by QuasR (Gaidatzis et al., 2015) using the following formulae:

$$Enrichment = log2((\frac{ns}{Ns} * \min(ns, nb) + p)/(\frac{nb}{Nb} * \min(ns, nb) + \mathrm{p}))$$

where ns is the total number of reads that align at the genome level in the ChIP-seq Sample, Ns is the number of reads that aligned the entire ChIP-seq sample, nb is the number of reads that aligned at the genomic level in the input, and Nb is the total number of reads that aligned in the input. P is the pseudocount, which is used to correct the enrichment values at the genomic features with low read counts and was set to 8. For ChIPseq of H3K27ac during time-points of reprogramming, data was downloaded from supplementary information provided in the study by Chen et al.(Chen et al., 2016) All other ChIPseq data was downloaded pre-analyzed from ENCODE (de Souza, 2012).

Transcription factor ChIPseq analysis

Transcription factor ChIPseq data for K562 cells was obtained from ENCODE (de Souza, 2012). ChIPseq data was analyzed in the same way as Histone Modification data given below with specifically restricting to narrow peak calling algorithm from MACS (Zhang et al.,

2008) Transcription factor motifs were obtained from TRANSFAC (Wingender et al., 2000) database. Only those Transcription factors were further analyzed for which motif information were available. Motif enrichment analysis was performed using HOMER (Heinz et al., 2010). Binding sites with ChIPseq peaks centered accross motifs were retained for further DNAshape analysis using DNAshapeR (Chiu et al., 2016). The transcription factors were classified according to families from AnimalTF Database (Zhang et al., 2015). For ChIPseq of Oct4 during time-points of reprogramming, data was downloaded from supplementary information provided in the study by Chen et al. (Chen et al., 2016).

Histone modification ChIP-seq in K562 analysis

Histone modification ChIP-seq data in K562 was obtained from ENCODE (de Souza, 2012) as BAM and peak files. Heatmap and density plot enrichment of Histone modifications was generated using ngs.plot.r (Loh and Shen, 2016; Shen et al., 2014). For the PCA plot of Transcription factor distribution on PCA dimensions as a function of enrichments across various chromatin states was undertaken with data of overlap of Transcription factor peaks with histone modification peaks.

Support Vector Machine and linear and logistic regression analysis

Support vector machine was implemented using e1071 R package 80% of data was used for training of SVM model, while rest 20% of rest was used for testing unless specified. Multiple models as independently mentioned were trained for the reducing chances of overfitting. ROC curves were plotted using ROCR package (Sing et al., 2005). Regression analysis was done using R with similar strategy.

Generating of features for Support vector machine analysis

For classification of enhancer positive and random genomic loci, the corresponding sequences of 2000BP centered on the peak-mid were extracted using BEDTOOLS (Quinlan, 2014)and DNAshape analysis was undertaken using DNAshapeR (Chiu et al., 2016). Propeller Twist values for each base pair step was used as feature set for SVM analysis. For classification of cell-type-cluster specific classification, the feature set included PHASTCons (Felsenstein and Churchill, 1996) derived conservation scores of enhancer sequences from each clusters, TRANSFAC (Wingender et al., 1996) transcription factor motifs enrichments for each enhancer sets, PhastCons derived Conservation scores of the transcription factor motif gene were used.

Conservation and Mutation analysis

For mutation analysis, COSMIC database (Forbes et al., 2017) was used to download

phenotypic and non-phenotypic mutations. The genomic classes in increasing order of mutation density was used by scanning number of mutations occurring in each non-overlapping 2000 BP regions, and stratification by the quartile analysis of the counts. Conservation scores were obtained from PhastCons (Felsenstein and Churchill, 1996)

Programming and scripting

Scripts for data analysis is written in R and PERL and is available upon request.

SUPPLEMENTAL REFERENCES

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T.*, et al.* (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455-461.

Chen, J., Chen, X., Li, M., Liu, X., Gao, Y., Kou, X., Zhao, Y., Zheng, W., Zhang, X., Huo, Y.*, et al.* (2016). Hierarchical Oct4 Binding in Concert with Primed Epigenetic Rearrangements during Somatic Cell Reprogramming. Cell reports *14*, 1540-1554.

Chiu, T.P., Comoglio, F., Zhou, T., Yang, L., Paro, R., and Rohs, R. (2016). DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. Bioinformatics *32*, 1211-1213.

Chiu, T.P., Yang, L., Zhou, T., Main, B.J., Parker, S.C., Nuzhdin, S.V., Tullius, T.D., and Rohs, R. (2015). GBshape: a genome browser database for DNA shape annotations. Nucleic acids research *43*, D103-109.

de Souza, N. (2012). The ENCODE project. Nature methods *9*, 1046.

Felsenstein, J., and Churchill, G.A. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. Molecular biology and evolution *13*, 93-104.

Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L.*, et al.* (2017). COSMIC: somatic cancer genetics at high-resolution. Nucleic acids research *45*, D777-D783.

Gaidatzis, D., Lerch, A., Hahne, F., and Stadler, M.B. (2015). QuasR: quantification and annotation of short reads in R. Bioinformatics *31*, 1130-1132.

Greenbaum, J.A., Pang, B., and Tullius, T.D. (2007). Construction of a genome-scale structural map at single-nucleotide resolution. Genome research *17*, 947-953.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell *38*, 576-589.

Langmead, B. (2010). Aligning short sequencing reads with Bowtie. Current protocols in bioinformatics *Chapter 11*, Unit 11 17.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Loh, Y.H., and Shen, L. (2016). Analysis and Visualization of ChIP-Seq and RNA-Seq Sequence Alignments Using ngs.plot. Methods in molecular biology *1415*, 371-383.

Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Current protocols in bioinformatics *47*, 11 12 11-34.

Ramirez, F., Dundar, F., Diehl, S., Gruning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic acids research *42*, W187-191.

Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M.*, et al.* (2015). The UCSC Genome Browser database: 2015 update. Nucleic acids research *43*, D670-681.

Shen, L., Shao, N., Liu, X., and Nestler, E. (2014). ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. BMC genomics *15*, 284.

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

Stevens, T.J., Lando, D., Basu, S., Atkinson, L.P., Cao, Y., Lee, S.F., Leeb, M., Wohlfahrt, K.J., Boucher, W., O'Shaughnessy-Kirwan, A.*, et al.* (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature *544*, 59-64.

Vanhille, L., Griffon, A., Maqbool, M.A., Zacarias-Cabeza, J., Dao, L.T., Fernandez, N., Ballester, B., Andrau, J.C., and Spicuglia, S. (2015). High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. Nature communications *6*, 6905.

Wiggins, P.A., van der Heijden, T., Moreno-Herrero, F., Spakowitz, A., Phillips, R., Widom,

J., Dekker, C., and Nelson, P.C. (2006). High flexibility of DNA on short length scales probed by atomic force microscopy. Nat Nanotechnol *1*, 137-141.
Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Pruss, M., Reuter, I., and Schacherer, F. (2000). TRANSFAC: an integrated system for gene expression regulation. Nucleic acids research *28*, 316-319.
Wingender, E., Dietze, P., Karas, H., and Knuppel, R. (1996). TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic acids research *24*, 238-241.
Zhang, H.M., Liu, T., Liu, C.J., Song, S., Zhang, X., Liu, W., Jia, H., Xue, Y., and Guo, A.Y. (2015). AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. Nucleic acids research *43*, D76-81.
Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W*., et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome biology *9*, R137.