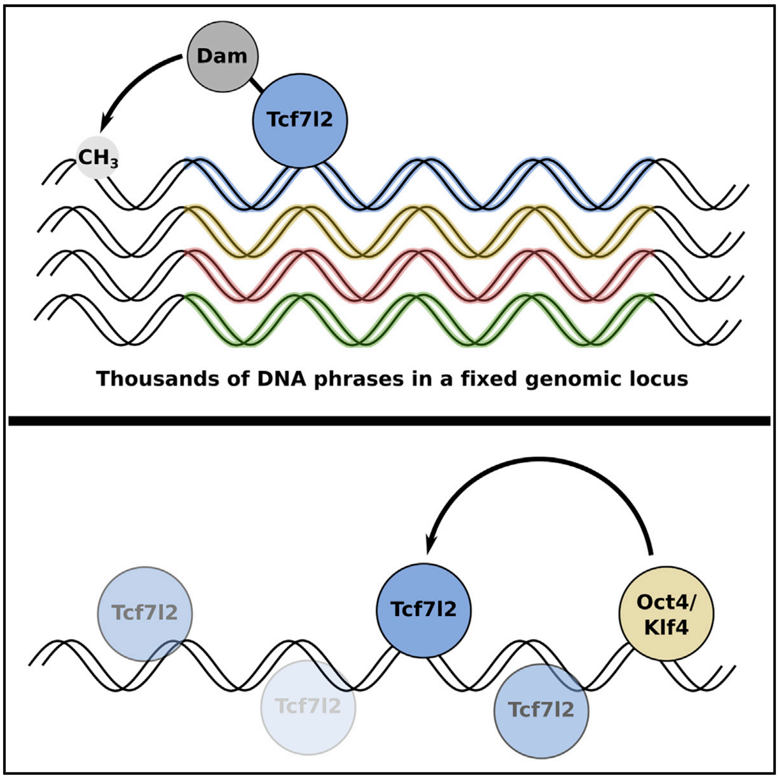
Cell Systems

Report

A High-Throughput Genome-Integrated Assay Reveals Spatial Dependencies Governing Tcf7l2 Binding

Graphical Abstract



Highlights

- Interactions between transcription factors regulate their genomic binding
- A genomically integrated high-throughput screen for transcription factor binding
- Measures sequence determinants of Tcf7l2 binding
- Tcf7l2 binding depends on nearby and in-phase Oct4 and Klf4 motifs

Authors

Tomasz Szczesnik, Lendy Chu,
Joshua W.K. Ho, Richard I. Sherwood

Correspondence

rsherwood@rics.bwh.harvard.edu

In Brief

Transcription factor binding in cells depends on interactions with other proteins. We have developed a high-throughput screening platform to study transcription factor binding strength at thousands of variable sequences in a fixed genomic locus. We find that binding of the Wnt effector Tcf7l2 in mouse embryonic stem cells depends on proximity and phasing that matches the turn of the DNA helix relative to its cofactors Oct4 and Klf4.

Szczesnik et al., 2020, Cell Systems *11*, 315–327
September 23, 2020 © 2020 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.cels.2020.08.004

Cell Systems

CellPress
OPEN ACCESS

Report

A High-Throughput Genome-Integrated Assay Reveals Spatial Dependencies Governing Tcf7l2 Binding

Tomasz Szczesnik,[1,2,3] Lendy Chu,[4] Joshua W.K. Ho,[1,2,5] and Richard I. Sherwood[4,6,7,*]
[1]Victor Chang Cardiac Research Institute, Darlinghurst, NSW 2010, Australia
[2]St Vincent's Clinical School, University of New South Wales, Darlinghurst, NSW 2010, Australia
[3]Department of Biosystems Science and Engineering, ETH Zürich, Mattenstrasse 26, Basel 4058, Switzerland
[4]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA
[5]School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong SAR, China
[6]Hubrecht Institute, 3584 CT Utrecht, the Netherlands
[7]Lead Contact
*Correspondence: rsherwood@rics.bwh.harvard.edu
https://doi.org/10.1016/j.cels.2020.08.004

SUMMARY

Predicting where transcription factors bind in the genome from their *in vitro* DNA-binding affinity is confounded by the large number of possible interactions with nearby transcription factors. To characterize the *in vivo* binding logic for the Wnt effector Tcf7l2, we developed a high-throughput screening platform in which thousands of synthesized DNA phrases are inserted into a specific genomic locus, followed by measurement of Tcf7l2 binding by DamID. Using this platform at two genomic loci in mouse embryonic stem cells, we show that while the binding of Tcf7l2 closely follows the *in vitro* motif-binding strength and is influenced by local chromatin accessibility, it is also strongly affected by the surrounding 99 bp of sequence. Through controlled sequence perturbation, we show that Oct4 and Klf4 motifs promote Tcf7l2 binding, particularly in the adjacent ~50 bp and oscillating with a 10.8-bp phasing relative to these cofactor motifs, which matches the turn of a DNA helix.

INTRODUCTION

Transcription factors recognize and bind to short DNA sequences, motifs, which can be measured directly through *in vitro* binding assays or discovered as enriched at sites bound across the genome. Such motifs, however, are insufficient to accurately predict where in the genome a transcription factor is bound, as most transcription factors bind to fewer than 10% of their strong motifs in any given cell type (ENCODE-DREAM Consortium 2017). Moreover, transcription factors exhibit cell-type-specific binding patterns, despite no change in the DNA-binding motif or genomic sequence. It is known that transcription factors influence each other's binding, either through direct interactions, competition for binding sites, or indirectly by altering DNA organization and accessibility. In different cell types it is then the set of transcription factors expressed that shapes their individual binding profiles. These interactions should be reflected in a "grammar": a logic in how the organization of individual transcription factor-binding motifs shapes the higher order interactions.

The non-random distribution of sequences in the genome, however, makes it difficult to draw further inferences from features enriched at transcription factor-binding sites. Two transcription factors could bind together because they control a similar set of genes, rather than because they stabilize each other's binding. On the other hand, particularly strong arrangements of transcription factors could cause ectopic activation and be selected against, resulting in native enhancers being comprised weaker than possible arrangements of motifs as they provide a sharper response to external signals (Farley et al., 2016). As such the most informative arrangements of motifs for detecting interaction effects are likely under-represented in the genome. Furthermore, it is not clear how much of the binding of a transcription factor at a specific location is due to the sequence immediately surrounding it. Each site is a unique position in the genome and could be influenced by the chromatin organization in the region, looping and interactions with distal regions, and impact of transcription in the area. For example, chromatin immunoprecipitation sequencing (ChIP-seq) experiments performed on livers of mouse F1 crosses have detected the impacts of genomic variants up to 10 kb away from binding sites (Wong et al., 2017)). Additionally, detecting binding sites is usually done through chromatin immunoprecipitation, which uses the same cross linking step as for detecting 3D interactions between distal genomic segments (looping). Without careful titration of this reaction one cannot be sure that it is only direct transcription factor with DNA interactions, and not some larger complex, that one extracts (Teytelman et al., 2013). This

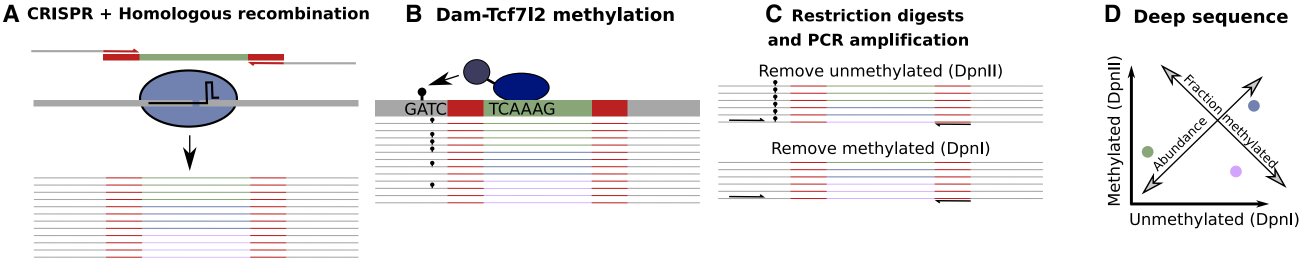


Figure 1. High-Throughput Locus-Specific DamID for Assaying Transcription Factor Binding to Variations in a Specific Genomic Locus
(A) The phrase library (12,000 or 2,000 oligos) is integrated using CRISPR-Cas9 and homologous recombination into a specific locus.
(B) Binding of Dam-Tcf7l2 to a phrase results in methylation of a GATC adjacent to the site of integration.
(C) Following genomic DNA extraction, two pools of completely methylated or unmethylated phrases are created by digestion with methylation specific restriction enzymes followed by PCR amplification with one locus-specific primer and one library-specific primer (black arrows).
(D) The amplified phrases are then deep sequenced and mapped back to the initial library. The relative enrichment of a phrase in the two pools indicates the level of Tcf7l2 binding. The sum of reads in both digests (abundance) and the fraction that is methylated (DpnII/Abundance) are used in later plots.

unavoidable combination of complex genomic features confounding measurement, removal of distal positional information, and biased sequence distribution means that one is limited in the ability to determine transcription factor-binding logic from computational analysis of genomic binding patterns.

The transcription factor Tcf7l2 provides a case study in inadequate prediction of cell-type-specific binding patterns. Tcf7l2 is part of the Tcf/Lef transcription factor family (Tcf7, Tcf7l1, Tcf7l2, and Lef1) (Arce et al., 2006), which all bind DNA through a conserved high mobility group (HMG) domain that prefers the sequence SCTTTGWWS. This recognition occurs through the DNA minor groove (Wetering et al., 1991; Wetering and Clevers, 1992), opening it up and creating a bend of 90–127 degrees (Love et al., 1995; Giese et al., 1995). Tcfs act primarily as effectors of the Wnt signaling pathway, binding the transcriptional activator β-catenin, which upon Wnt signaling ceases to be constitutively degraded (Nelson and Nusse, 2004).

As part of a conserved developmental pathway, Tcfs regulate various functions throughout different cell types in response to Wnt signaling. In mouse embryonic stem cells (mESCs) Tcfs appear necessary to reduce expression of other transcription factors that maintain pluripotency (notably Nanog) in order to allow differentiation (Pereira et al., 2006). In the intestine, Tcf7l2 helps maintain a constant proliferation of adult stem cells that support tissue renewal; a lack of dominant-negative isoforms of Tcf7l2 and mutations in APC—part of the complex that enables GSK-3β to cause degradation of β-catenin—is linked to colorectal cancers (Korinek et al., 1997). Tcf7l2-knockout mice show problems with endoderm development and maintenance of intestinal stem cell populations (Korinek et al., 1998). In liver and pancreatic tissues, Tcf7l2 underpins glucose homeostasis (Norton et al., 2014), with intronic mutations that reduce expression of Tcf7l2 being associated with type 2 diabetes (Grant et al., 2006). One mechanism by which Tcf7l2 achieves cell-type-specific effects is by binding at different genomic locations, hence, regulating a different set of target genes. Across 6 human cell lines (5 endodermal and 1 epithelial) Tcf7l2 bound a largely disparate set of sites, with only 1,800 out of 116,000 total Tcf7l2-binding sites shared between the 6 (Frietze et al., 2012). Supporting the idea of a grammar, different cofactor motifs tend to be enriched at cell-type-specific binding sites. Similarly, in Frietze et al. (2012), the Foxa2 and Hnf4α motifs are enriched in a hepatocyte cell line (Hnf4α appears to function with Tcfs in hepatocytes; Norton et al., 2014), while in an adenocarcinoma cell line it appears that the Gata3 motif helps bind Tcf7l2 when its own motif is absent.

To assess Tcf7l2-binding logic while avoiding the complexity of inferring from sites bound across the genome, we have developed an approach to measure Tcf7l2 binding to thousands of phrases of 99-bp variable DNA sequence transplanted into fixed, defined genomic loci using a quantitative DamID assay (Vogel et al., 2007; Szczesnik et al., 2019). This strategy allows us to detect differences in binding induced by minimal, designed sequence alterations while controlling for effects of the surrounding DNA sequence, thus, enabling us to determine causal relationships between DNA sequence and Tcf7l2 binding. Importantly, our assay is performed in a native cellular chromatin context, allowing us to account for effects of chromatin organization and interactions with other proteins that are missing in *in vitro* binding assays.

Using our approach, we find that while *in vivo* Tcf7l2 binding is dependent on the presence and match of its *in vitro* motif at individual binding sites, Tcf7l2 binding also varies dramatically based on the sequence surrounding it, and cell-type-specific Tcf7l2 binding at genomic loci can be partially recapitulated by the local surrounding sequence. Particularly, the presence of Oct4 and Klf4 motifs favors Tcf7l2 binding in mouse embryonic stem cells (mESCs), and this effect is strongest when occurring within an adjacent ~20 to 50-bp region and oscillates approximately every 10.8-bp shift in distance between the Tcf motif and cofactors. This effect is strongest surrounding when the Oct4 motif occurs as a part of the joint Sox2-Oct4 motif (Chen et al., 2008) and particularly helps promote binding in inaccessible chromatin, which is otherwise refractory to Tcf7l2 binding. This high-throughput DamID assay provides a powerful platform to determine local DNA-sequence grammars that causally influence transcription factor binding.

RESULTS

DamID for Locus Integrated Phrase Library

We developed an assay for measuring Tcf7l2 binding to thousands of pre-determined DNA "phrases" at a specific genomic locus (Figure 1). A library of synthetic oligos containing a 99-bp variable region (phrase) flanked by short constant sequences used as primers is integrated into specific genomic locations in mESCs by CRISPR-Cas9-based homology-directed repair. We find site-specific integration of one of the variable phrases in 20%–40% of alleles through quantitative PCR, confirming previous work (Hashimoto et al., 2016; Rajagopal et al., 2016).

Binding of Tcf7l2 to each integrated phrase is measured by DamID (Vogel et al., 2007). We use a mESC line that enables Cre-LoxP-mediated single genomic integration of a doxycycline-inducible transgene at a fixed site (Iacovino et al., 2011). Prior to phrase library integration, we use Cre/LoxP to integrate a fusion protein of Tcf7l2 and the N126A mutant of Dam, which we have recently shown to allow accurate measurement of Tcf7l2 binding genome wide with reduced off-target methylation as compared with the wild-type Dam enzyme (Szczesnik et al., 2019). As a control, we use a cell line with identical single-copy integration of unfused Dam N126A, which removes variability stemming from random integration and copy numbers. Due to its smaller size, Dam N126A expresses and diffuses more efficiently, resulting in more total methylation than the Dam-Tcf7l2 fusion. A Dam methylation site (GATC) is located directly adjacent to the site of phrase library integration, Dam will methylate phrases in proportion to Tcf7l2-binding strength within the phrase. After this 24-h induction period, genomic DNA from >2 × 10^7 phrase library-integrated cells is separately digested with restriction enzymes specifically recognizing unmethylated GATC (DpnII) or methylated GATC (DpnI). Undigested phrases are then PCR amplified with primers designed to specifically amplify genomically integrated phrases to avoid contamination with unintegrated phrases. These two pools of methylated and unmethylated phrases are then sequenced by Illumina next-generation sequencing (NGS), and the relative abundance of a particular phrase between the two pools is used to estimate the level of Dam methylation, and hence of Tcf7l2 binding, to the integrated phrase. Following DpnII digestion of uninduced cells there was no methylation at the adjacent GATC detected by qPCR, while after 24 h of expression wild-type Dam (which methylates almost the entire genome) over 90% was methylated, indicating that digestion and PCR amplification at this GATC can detect a large range of methylation.

We initially designed and screened a library of 12,000 phrases using this Tcf7l2 DamID approach, split between phrases comprising native genomic sequences and designed sequences. Half of this library was comprised of 99-bp genomic phrases sampled from ChIP-seq peaks that show variable Tcf binding. Cohorts of phrases were sampled from ChIP-seq peaks bound in either Tcf7l1 ChIP-seq in mESCs or Tcf7l2 ChIP-seq in intestinal endoderm (IE) cells, or in both (1,200 phrases each; see STAR Methods and Figure S1). If present, any Tcf motif was included in the sampled 99-bp phrase, and we specifically included ChIP-seq peaks lacking any clear Tcf motif (half of each group). The final cohort of genomic phrases contained unbound Tcf motifs (>10 kb from any Tcf ChIP-seq peak) in regions of open chromatin near marks of active enhancers (H3K27ac ChIP-seq peaks) (2,400 phrases). The other half of the library (6,000 phrases) was generated *de novo* from different arrangements of binding motifs for a set of transcription factors we deemed likely to influence the binding of Tcf7l2, based on published protein-protein interactions or motif enrichment adjacent to Tcf binding sites (see STAR Methods for details) (Frietze et al., 2012; Norton et al., 2014; Cole et al., 2008). This library was first integrated in 2 biological replicates into the inert Rosa26 locus, which resides in natively accessible chromatin (Zambrowicz et al., 1997) (see Figure S9 for DNase-seq signal at this locus). DamID on each of the replicates was done with both Dam-Tcf7l2 and unfused Dam, which has been shown to vary with chromatin accessibility (Kladde and Simpson, 1992) and thus provides a control for differences in Dam methylation rates between phrases independent of Tcf7l2 binding.

Statistical Processing of Read Counts

In order to draw valid comparisons between phrases we need to account for differences in the integration efficiency and sequencing coverage across the phrase library and different conditions. In particular, we noticed that the read counts for a large number of phrases exhibited significant dropout in either the DpnI or DpnII digested samples (seen in the histogram for DpnII counts in Figure 2B), which would confound analysis based on the fraction of methylated counts for each phrase. These dropouts suggest a bottleneck in unique methylation events prior to PCR amplification and sequencing. To estimate the (unobserved) number of methylated or unmethylated alleles for each phrase from the observed sequencing read count, we developed a statistical modeling pipeline.

Briefly, methylated and unmethylated allele counts are modeled as a negative-binomial distribution to capture our assumption that a large variable phrase library should contain a smooth, unimodal distribution over the frequency of methylation. Observed sequencing reads are modeled as a Poisson distribution stemming from a linear amplification of these allele counts, which recreates the observed dropouts (0 genomic counts) and staggering at the lower end of the observed read counts (1, 2… genomic counts) (see STAR Methods for details). The relation between unobserved allele counts ("normalized") to the observed sequencing counts ("raw") captured by the model is shown in Figure 2A.

To calculate allele counts for each experimental replicate, the negative binoimial distribution parameters and amplification rate are tuned to best match the distributions in the observed data (see Supplemental Information for parameter values). The amplification rate is then used to estimate the initial number of allele counts for each phrase in methylated and unmethylated samples (Figure 2B). Following normalization the observed difference between replicates (Figure 2C) follows the expected binomial sampling distribution (Figure 2D), indicating that the majority of difference between replicates stems from sampling variability, and not from a technical or biological source of variability.

Having normalized the data to obtain allele counts, we observed that replicates show a high degree of heteroscedasticity: high abundance phrases have low variability between replicates, low abundance phrases have high variability (Figure 2C). While it is expected that phrases with fewer unique alleles will have more variable measurement, we must account for this issue to performed balanced statistical analysis of library data. Thus, in order to use information from the entire phrase library we need to quantify the uncertainty in our estimate of the Dam methylation fraction for each phrase.

For this task we use a beta-binomial empirical Bayes model, which models the distribution of methylation across the whole

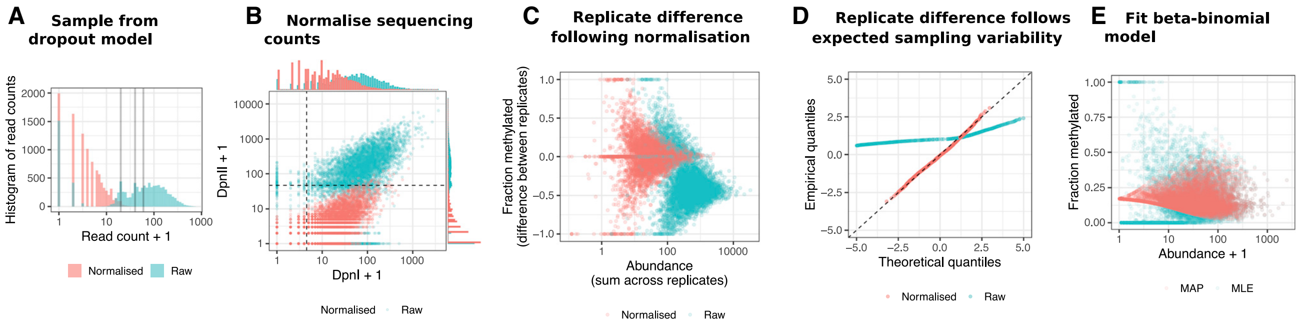


Figure 2. Statistical Modeling Reveals that the Expected Sampling Variability Explains the Majority of Differences between Replicates

(A) Example counts from the negative-binomial model to show the expected distribution of read counts ("raw" in blue) from over-sequencing fewer genomic counts ("normalized" in red). Vertical lines show the position of peaks corresponding to 1, 2, and 3 original genomic phrase counts (referred to as normalized).

(B) Estimated amplification rates (dashed line) for the phrase library in the accessible chromatin locus following Dam-Tcf7l2 expression are used to calculate the underlying genomic counts ("normalized") from the observed sequencing counts ("raw").

(C) Normalization reveals that high abundance phrases have high concordance between replicates, which decreases as the abundance decreases.

(D) Variability between replicates closely follows the expected binomial sampling distribution. Due to this, counts are pooled across replicates for further analysis.

(E) A beta-binomial empirical Bayes distribution is fit to each sample, which reduces the effect of heteroscedasticity by biasing low coverage samples toward the mean fraction methylated. MAP, maximum a posteriori estimate of beta distribution; MLE, maximum likelihood estimate or binomial distribution.

library with a beta distribution, which captures the unimodal spread between 0 and 1. This beta distribution is then used as the prior for the binomial methylation of each phrase, generating a posterior beta distribution that gives the credible interval for the frequency of each phrase's methylation. In practice this adds a few pseudocounts (1 to 14, see Supplemental Information for values) to every sample, biasing low coverage phrases toward the overall population mean but not affecting the high coverage phrases (compare the maximum a posterior MAP estimate from the beta distribution, to the maximum likelihood estimate MLE of the initial binomial distribution in Figure 2E). Overall, this computational pipeline allows us to estimate Tcf7l2 binding to each of a 12,000-phrase library integrated into a fixed genomic context while accurately quantifying the uncertainty in the frequency of each phrase's methylation for subsequent statistical analysis.

Phrase Library with Genomically Sampled and Synthetically Generated Phrases

We next proceeded to assess features governing Tcf7l2 binding, by comparing the amount of methylation seen with Dam-Tcf7l2 (Figure 3A) with that of unfused Dam (Figure 3B). To measure the effects of a set of features of each phrase (motif presence, Tcf binding, and histone methylation status of the genomic loci from which a phrase was derived) on its methylation frequency, we used logistic regression with lasso penalty and cross-validation (Friedman et al., 2010). Effect size is used to refer to how much the fraction of methylated shifts along a logistic function when a feature is present in a phrase; at 0.5 fraction methylated this is a linear effect and becomes asymptotic toward 0 and 1.

For the 6,000 phrases derived from native genomic sites, cell-type-specific binding of Tcf7l2 tends to be retained when these phrases are transplanted to an open chromatin site in mESCs, with phrases derived from mESC ChIP-seq peaks bound more strongly than those derived from intestinal endoderm ChIP-seq peaks or those without ChIP-seq binding in either cell type (0.23 versus 0.046 effect sizes, unbound phrases are at 0; Figure 3C). Presence of a Tcf motif is a strong predictor of Tcf7l2 binding in our assay, as phrases originating from Tcf ChIP-seq peaks but lacking a Tcf motif tend to show weak or absent Tcf7l2 binding, as compared with Tcf motif containing sites within the same ChIP-seq binding profile (ESC 0.36, IE 0.27, both 0.27 effect sizes). While phrases with ChIP-seq binding only in intestinal endoderm in general tend not to acquire binding when transplanted into the Rosa26 locus, those with a Tcf motif are more likely to exhibit Tcf7l2 binding than those without one (0.046 versus 0.27 effect sizes). The positive effects of Tcf motif presence among phrases derived from intestinal endoderm ChIP-seq peaks are largely attenuated when the phrase originates from a site that contains enhancer marks in mESCs (IE with Tcf motif and enhancer marks = 0.046 + 0.27 − 0.24 = 0.076 effect size). This may indicate that the regulation of Tcf7l2 binding between mESCs and intestinal endoderm is in some case through broader changes in the chromatin organization and in others by local sequence features that are permissive in only one cell type. Overall, our analysis of 99-bp phrases transplanted to the Rosa26 locus from native genomic regions finds that Tcf7l2 binding is strongest when the native regions contain Tcf motifs and derive from regions with native mESC Tcf7l2 binding. We conclude that features present within the sequence immediately surrounding a Tcf motif are strongly responsible for regulating the cell-type-specific binding of Tcf7l2 *in vivo*.

For the 6,000 phrases containing different motif arrangements, we calculated the effect of the presence of each motif on Tcf7l2 binding. We found that the presence of the Tcf motif had the strongest effect on Tcf7l2 binding (0.49 effect size), and other putative cofactors had weaker but still positive effects on Tcf7l2 binding (∼0.1 to 0.2 effect size; Figure 3D). To distinguish effects on Tcf7l2 binding driven by protein-protein interaction as compared with those driven indirectly by induction of chromatin accessibility adjacent to the Tcf motif, we integrated this 12,000-phrase library into a genomic locus with minimal native chromatin accessibility in mESCs (upstream of the T-cell-specific CD8 gene: uCD8; see Figure S8 for DNase-seq signal at this locus). Consistent with chromatin accessibility affecting both Dam methylation and Tcf7l2 binding, the overall methylation was reduced in this inaccessible locus (library beta-binomial mean:

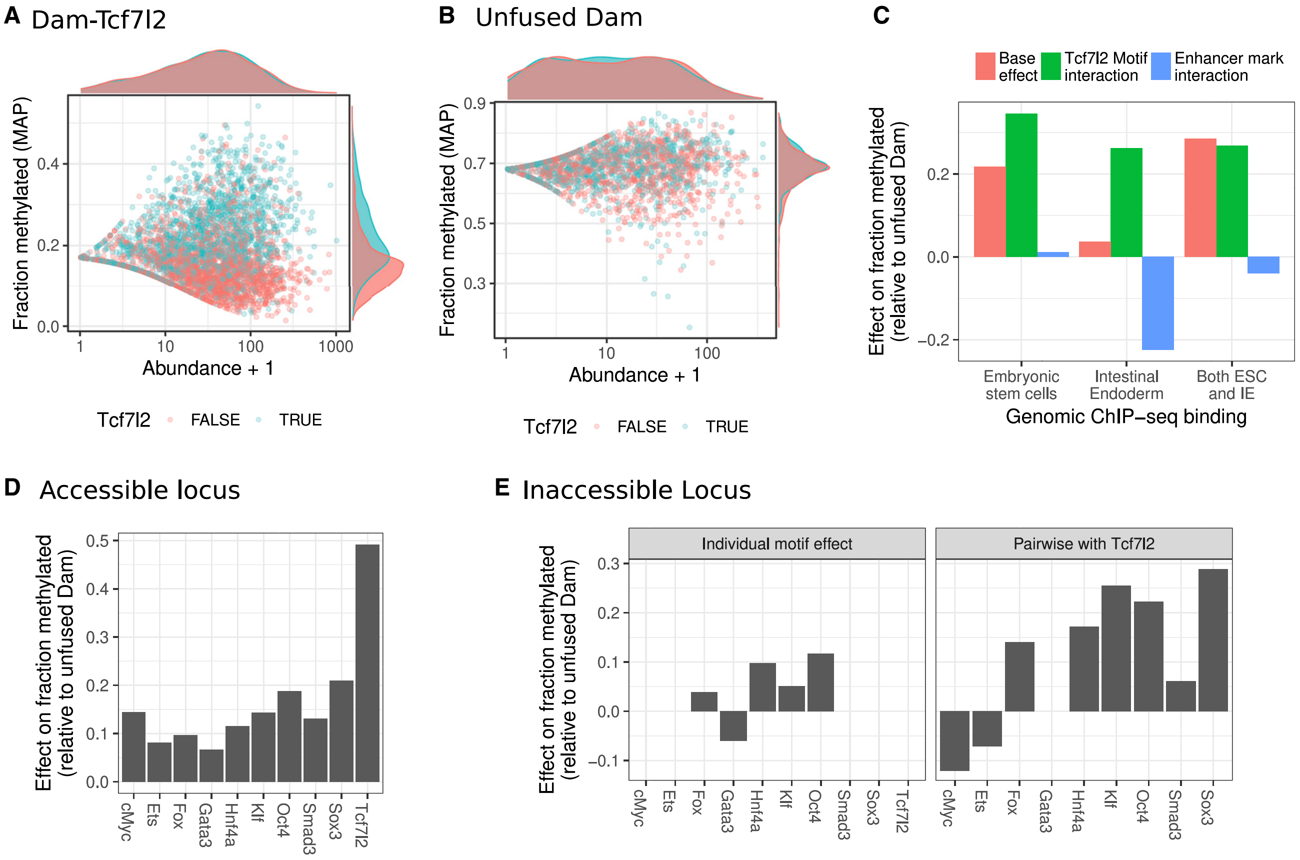


Figure 3. 99 bp of Local Sequence Regulates Binding of Tcf7l2 to Its Motif across Different Genomic Sites

(A and B) (A) Dam-Tcf7l2 and (B) Dam methylation of phrases in the accessible locus Rosa26, split by the presence of a Tcf7l2 motif.

(C) Logistic regression effect size of genomic features on phrases transplanted to the accessible locus Rosa26. Note the base effect is additive: a phrase that was bound in intestinal endoderm, which a Tcf motif, and an mESC enhancer marks has methylation as the sum the red, green, and blue bars for intestinal endoderm.

(D) Logistic regression effect size for the presence of cofactor motifs after integration of the phrase library in the accessible locus.

(E) Logistic regression effect size for the presence of cofactor motifs, either individually or as pairwise interaction with the presence of the Tcf motif, after integration of the phrase library in the inaccessible locus uCD8. This specific sample was measured with wild-type Dam, instead of the N126A variant used elsewhere. Non-significant features are shrunk to zero (e.g., cMyc, Ets,…).

Dam-Tcf7l2 from 0.17 accessible to 0.06 inaccessible locus, unfused Dam from 0.68 to 0.15). Due to the low signal detected with Dam-N126A for this library in this locus, we used the wild-type version of the Dam enzyme which gives stronger signal, however, at the cost of being strongly confounded by changes in chromatin accessibility. We found that, in this natively inaccessible locus, Tcf7l2 binding required the pairwise interaction of its motif with that of specific cofactors, particularly Oct4, Klf4, or Sox2 (0.25, 0.22, and 0.30 effect size; Figure 3E). These stronger pairwise effects between the Tcf and cofactor motifs suggest that this effect is driven more by cooperative interactions rather than independent changes in chromatin. Thus, in this controlled assay, Tcf7l2 binding is impacted by adjacent motifs, and these motifs become more important when phrases are integrated into a locus without surrounding chromatin accessibility.

Tiled Tcf Motif Phrase Library

A deeper analysis of interactions between Tcf7l2 and its cofactors would require single phrase-resolution data of this library, which we were unable to obtain due to insufficient data coverage. Out of the 12,000 phrases, we observed integration of only 3,000–4,000, and only a few hundred (9%) had sufficient coverage for accurate estimates of their true methylation frequency (Figure 4A), limiting analysis to population trends and preventing the detection of sparser levels of binding in the inaccessible locus. This limited integration is largely due to the incomplete efficiency of CRISPR-Cas9-based homology-directed repair and limitations on total cell number.

To investigate the adjacent motifs and spatial determinants governing Tcf7l2 binding at higher resolution, we designed a 2,000-phrase library that systematically varied the position of the Tcf motif across a set of 59 backbone phrases. By reducing the number of unique phrases from 12,000 to 2,000, we posited that we would increase coverage of each phrase and thus enable phrase-resolution analysis. Backbone phrases were chosen from both the native genomic and synthetically generated phrases in the initial library so as to span a range of Dam-Tcf7l2 methylation rates across both accessible and

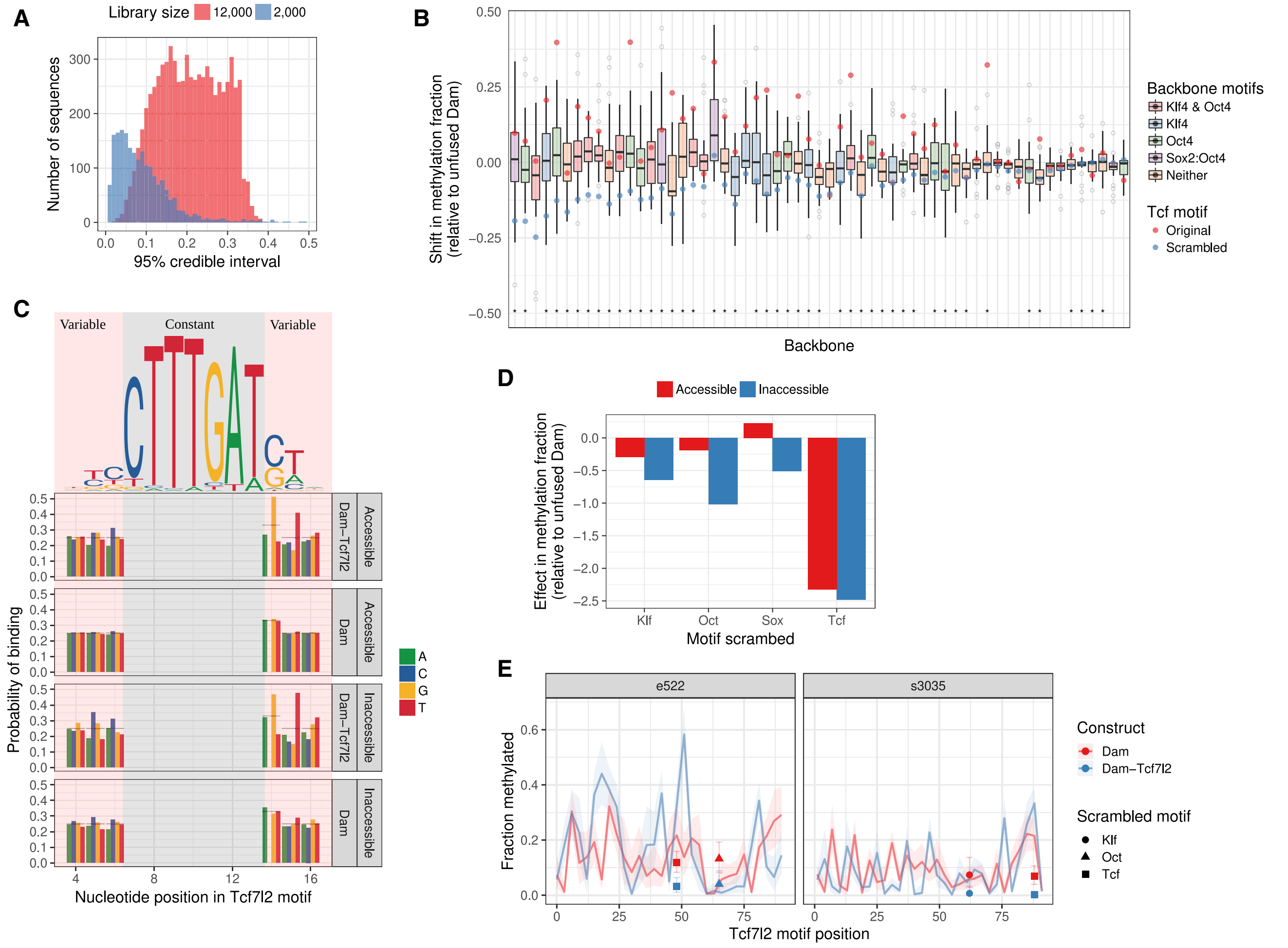


Figure 4. Individual Phrase-Level Resolution of Dam-Tcf7l2 Binding Shows Influence of Motif Score and Cofactor Motif Presence

(A) Comparison of 95% posterior credible interval of fraction methylated for the 12,000 and 2,000 phrase libraries with Dam-Tcf7l2 in the accessible locus.

(B) Spread of Tcf7l2 binding (Dam-Tcf7l2 relative to unfused Dam) as the Tcf motif is tiled across each backbone phrase. Scrambling the Tcf motif (blue dot) substantially decreases binding relative to original Tcf motif (red dot) (*p < 0.05). Gray dots are outliers from the box plot.

(C) The two 5′ and 3′ nucleotides flanking the Tcf motif consistently explain some of a variability in binding and are consistent with lower informative bases in the estimated motif for Tcf7l2 from protein binding microarrays.

(D) Logistic regression effect of scrambling each motif across the whole library. For raw values see Figure S2.

(E) Example of footprint for the tiled Tcf motif disrupting the Oct4 (e522) or Klf4 motif (s3035) to the same level as scrambling the motif. Shaded region/error bars show 95% posterior credible interval.

inaccessible chromatin loci (see STAR Methods for details). For each backbone phrase, phrases were designed with a scrambled version of the initial Tcf motif, and a set of phrases was designed in which the most informative, core part of the Tcf motif (CTTTGAT) was tiled across each backbone phrase in 3-bp increments, replacing the sequence that had been in that position. We also included phrases for each backbone phrase in which we scrambled motifs for Oct4, Klf4, and Sox2, which were identified from the initial library as likely influencing Tcf7l2 binding (see STAR Methods for scrambled sequences). We integrated this 2,000-phrase library into the Rosa26 (accessible chromatin) and uCD8 (inaccessible chromatin) genomic loci in two biological replicates in mESCs and performed DamID with Dam-Tcf7l2 and unfused Dam as for the previous library. In this experiment, a methylation site (GATC) was included on both sides of the integrated phrase in order to reduce possible confounding from the variable distance between Dam-Tcf7l2 and the GATC methylation site.

Integrating this smaller phrase library vastly reduced the uncertainty in estimated methylation rate for individual phrases (Figure 4A), providing much higher-resolution data (9% to 65% increase in the number of phrases with a 95% credible interval of the spread in methylation less than 0.1; >99% phrases were recovered). This was due to sampling many more unique genomic instances of each phrase, and since the variability between replicates remained consistent with the beta-binomial model (at a 0.05 cutoff, 0.935 of the methylation rates of the second replicate fell within the posterior distribution of the first replicate) the concordance between replicates increased proportionally (Figure S7B). Thus, we were able to perform individual phrase-level analysis of Tcf7l2 binding for the majority of the 2,000 phrases.

We first looked at Tcf7l2 binding in the accessible locus. Consistent with our findings in the initial library, scrambling the Tcf motif led to significantly decreased Tcf7l2 binding in most backbone phrases (46/59 at $p < 0.05$; Figure 4B). When comparing Tcf7l2 binding in phrases from the same backbone phrase, we found substantial variation as the Tcf motif is tiled across each backbone phrase (standard deviation range 0.029–0.16; mean of 0.084; Figure 4B). This intra-backbone variation is significantly larger than the standard deviation in the estimate of each phrase's methylation rate for 98.5% of phrases, indicating that tiling the Tcf motif across a phrase leads to robust changes in Tcf7l2 binding. In fact, on an average, Tcf7l2 binding in the phrase with the most unfavorable Tcf motif position within a backbone phrase was equivalently low to the phrase with scrambled Tcf motif −0/59 had significantly higher methylation, while 4/59 backbone phrases had significantly lower methylation ($p < 0.05$, with adjustment for multiple hypothesis testing)—indicating that the sequence context is a strong determinant of binding (Figure 4B). The few backbone phrases that did not have a significant drop in methylation upon scrambling the Tcf motif also had a lower variability in Tcf7l2 binding as the Tcf motif was tiled (mean standard deviation of 0.06 versus 0.09; t test $p = 0.01$), suggesting that other sequence features are required to promote Tcf7l2 binding. We conclude that the location of a core Tcf motif relative to surrounding local sequence plays an important role in determining Tcf7l2 binding strength. We thus turned to investigating patterns in these Tcf motif tiling experiments to determine the key local sequence features underlying Tcf7l2-binding logic.

We find that one major cause of the variability in Tcf7l2 binding as the motif is tiled across each backbone phrase is the change in nucleotides flanking the core Tcf motif. While we tiled the 7 nucleotide core Tcf motif, the position weight matrix that best explains Tcf7l2 in protein binding experiments (Badis et al., 2009) contains contributions from two nucleotides on either side of the core motif. By calculating the average Tcf7l2 binding across all possible base identities in the flanking positions, we find that Tcf7l2 binding strength in our assay correlates with the optimal nucleotide identities of the full Tcf7l2 position weight matrix—a 3′ guanine followed by thymine, and a slight 5′ cytosine or thymine preference (Figure 4C). We note that 3′ cytosines are not present in our phrases as they were replaced with G to avoid the creation of an extra GATC methylation site. A logistic model of the contribution of flanking nucleotide position to Tcf7l2 binding found that including the effects of di- or tri- nucleotides reduces the model log-likelihood, indicating that independent contributions of single nucleotides are sufficient to explain variation in binding strength from flanking nucleotides. This finding rules out effects of new motifs being reproducibly formed from tiling the Tcf motif, as they would result in a model that prefers base-pair dependencies and would likely differ from the *in vitro* binding preference. The changing affinity for the full Tcf motif as it is tiled along a backbone phrase, however, does not explain any differences in average binding affinity between different backbone phrases. This is because the nucleotides that happen to flank the Tcf motif at each tiled position are more-or-less random and are not correlated within a backbone phrase nor between them (this would not be the case if the backbone phrases were strongly enriched in a specific nucleotide or dinucleotide).

Much of the remaining variation in positional Tcf7l2 binding appears dependent on the presence of neighboring motifs, as backbone phrases with the highest binding rate contained either a Klf4 or Oct4 motif (Figure 4B), and shuffling these motifs also tended to reduce binding (Figure 4D). Plotting the Tcf7l2 binding strength as the Tcf motif is tiled along a backbone phrase, we identify striking patterns of reduced Tcf7l2 binding resembling ‘‘footprints’’ coinciding with phrases in which the Tcf motif disrupts an underlying Oct4 or Klf4 motif, with a similar decrease in Tcf7l2 binding as scrambling these motifs (Figure 4E). Since the Tcf motif is tiled by 3 bp, any motif longer than 3 bp should be detected. Nonetheless, we observed no robust loss of Tcf7l2 binding for across adjacent tiles occurring for any other known motifs. We cannot rule out that other motifs would show this effect if we had tiled the Tcf motif across a larger cohort of backbone phrases. Even within the set of backbone phrases containing Oct4 or Klf4 motifs, we observed substantial backbone phrase-specific variation in the magnitude of Tcf7l2 binding and the loss of such binding upon disruption or scrambling of these cofactor motifs (Figure S2). The strongest effects were localized around backbone phrases containing an Oct4 motif as part of a joint Sox2-Oct4 motif, which hints that much of this variability stems from differences in binding affinities of these cofactors between backbone phrases. However, because of the low numbers of instances of either Oct4 or Klf4 motifs in the 59 backbone phrases and the fact that we did not vary the strengths of these motifs in a controlled way, we cannot make strong conclusions about the role of cofactor motif strength.

Gaussian Process Model for Spatial Effects on Tcf Binding

Having identified significant roles of the extended Tcf motif and the presence of adjacent Oct4 and Klf4 motifs in modulating Tcf7l2 binding, we examined whether there are spatial constraints on the positioning of the Tcf motif relative to the cofactor motifs. In order to measure such spatial effects, we use a Gaussian process model, a non-linear regression technique, to model how the position of the Tcf motif within the backbone phrase affects the binding of Tcf7l2. Importantly, within the Gaussian process framework we can define classes of non-linear functions that vary smoothly with the position of the Tcf motif within the backbone phrase, allowing the model to generalize across several Tcf motif positions. This is preferable to treating each spatial position as independent, which results in an overly flexible model lacking in statistical power, or pooling across several adjacent positions, which would blur the underlying spatial effect. As a result, Gaussian process modeling should allow us to more easily identify reproducible cofactor interactions that lack fixed spacing, such as those that slowly change in strength over a region or occur at repeating positions. Gaussian process models also allow us to account for the confounding effects of Tcf motif strength and locus-specific effects to generalize spatial trends across multiple phrases.

We use two classes of non-linear functions to capture different ways in which the changing Tcf motif position could affect Tcf7l2 binding. The first class of functions are designed to identify contiguous stretches of higher or lower Tcf7l2 binding, for example due to a cofactor motif promoting binding nearby. These functions are modeled by a radial basis kernel, which fits smoothly varying functions and is parametrized by a length scale that controls how quickly the function varies. The second class of functions are similar in nature but oscillate periodically, for example due to effects caused by the regular turning of the DNA helix. These functions are also encoded by a periodically repeating radial basis kernel and thus are parametrized by both a length scale and periodicity. To separate out possible spatial effects in the locus, for example due to the position of a nearby nucleosome, each of these functions is present in 3 forms: as backbone phrase-invariant across all backbone phrases, as backbone phrase-invariant across backbone phrases sharing the same Tcf motif orientation, and as spatial effects unique to each backbone phrase. Finally, in order to reliably detect these spatial interactions, we also need to account for the confounding effects within the data. As such the effects of nucleotides flanking the Tcf motif, the average binding to each backbone phrase, and the uncertainty in estimating each phrase's methylation are included as linear effects in the model.

Fitting this Gaussian process model to the observed Tcf7l2 DamID data is done in empirical Bayesian fashion: the internal parameters are integrated out and hence averaging over the various spatial functions consistent with the data, while the model hyperparameters—the length scale, periodicity, and linear weights—are tuned to optimize the model likelihood. Following this we find that the model that best fits the observed Tcf7l2 DamID data in both accessible and inaccessible chromatin loci contains inputs from multiple distinct components (Figure 5A). The most salient individual component influencing intra-backbone variation in Tcf7l2 binding is Tcf motif score as determined by the identity of the four nucleotides adjacent to the Tcf7l2 core motif, which was tiled, which explains 35% and 40% of Dam-Tcf7l2 methylation variability in accessible and inaccessible loci, respectively. The various spatial effects of Tcf motif position explain a further 45% of the variation in binding across both loci. Within these, the backbone phrase-invariant spatial effects—those caused by general features of the two genomic loci used for integration—explain 20% and 25% of variable methylation in accessible and inaccessible loci. Backbone-phrase-specific spatial effects—those dependent on cofactors or other sequence features—explain a further 25% and 20% of variable methylation in accessible and inaccessible loci.

Thus, a Gaussian process regression model with a minimal set of features is capable of explaining the majority of variation (80%–85%) in Tcf7l2 binding across the 2,000 phrases, suggesting that these features of Tcf motif strength and smoothly and periodically varying spatial constraints with cofactors explain much of Tcf7l2 binding strength. The residual variability of 20% and 15% in the accessible and inaccessible loci indicates that either there is little information left to extract from this dataset or that further inference would require improving the coverage of each phrase in order to extract out more subtle features. These could, for example, be due to the effect of a rigid spacing requirement, formation of some unique sequence overlapping the Tcf motif, or confounders of the Dam methylation rate.

Since we are optimizing the hyperparameters of the Gaussian process model likelihood, it is important to ensure that the model is not so explicit as to overfit the data. However, we do not find no evidence of overfitting: the length and periodicity is estimated as almost the same between different loci, the component weights are similar between the two different loci, features with no explanatory power in the unfused Dam samples are automatically discarded, and the presence of residual variability indicates that the model is not fitting directly up to the limit imposed by sampling variability (Figure 5A). As such we think the model is accurately capturing general trends in the underlying dataset.

Unexpectedly, the overall position of the Tcf motif in the locus has an effect on Tcf7l2 binding. The spatial effect of the Tcf motif position across all backbone phrases (irrespective of orientation) shows a similar pattern in both accessible and inaccessible loci: the length scale optimizes to ~8-bp, and extracting the estimated function (Figure 5B) shows that Dam-Tcf7l2 methylation is highest when the motif is located within the middle of a backbone phrase or toward either end with dips in intervening regions. Since this effect is so similar for Dam-Tcf7l2 methylation across both loci, we posit that a likely cause is steric constraints on how well Dam-Tcf7l2 can methylate the GATC sites located adjacent to the ends of each integrated phrase, rather than changes in Tcf7l2 binding. Unfused Dam in the inaccessible locus also shows an effect of the overall Tcf motif position (Figure 5B); however, this mimics an overall downward trend of Dam-Tcf7l2 and not the middle and end peaks, presumably because unfused Dam is not binding to the Tcf7l2 motif. This also rules out a differential accessibility between the Tcf motif and the GATC—for example, if a nucleosome was laterally displaced from the Tcf motif and covered the GATC—as this would create a negative correlation in the backbone phrase-invariant effects between Dam-Tcf7l2 and unfused Dam. Thus, we do identify one feature that is best explained as an artifact of the DamID method—Tcf7l2 is on an average more adept at methylating GATCs with particular distance constraints.

When backbone phrase-invariant effects are calculated only across backbone phrases that share the same Tcf motif orientation, the model identifies a periodic function only in the accessible locus (Figure 5B). The periodicity parameter optimizes to every 10.8 bases, which is close to the estimate of a DNA helix rotation (10.4–10.6) (Wang, 1979; Rhodes and Klug, 1980; Klug and Lutter, 1981). Interestingly, the periodic component for the two-Tcf motif orientations are completely out of phase. Since Tcf7l2 binding introduces a large bend in the DNA (90–127 degrees) (Love et al., 1995; Giese et al., 1995), a possible explanation is that DNA bending in the Rosa26 accessible locus is more energetically favorable in one direction. This would favor Tcf7l2-binding sites that are in-phase with one another with respect to the rotation of a DNA helix, since these would all bend in the same direction. Alternatively, it could indicate an interaction with a transcription factor or nucleosome at a specific position nearby the site of integration.

Lastly, we investigated the backbone phrase-specific spatial effects. The optimal model includes input from both smoothly varying and periodic backbone phrase-specific components with similar hyperparameters as the global and orientation effects (~8 rbf length scale and 10.8 periodicity), in both the accessible and inaccessible loci. The smoothly varying, periodic, and constant backbone phrase-specific effects are extracted and summed up for a set of representative backbone phrases (Figure 5C).

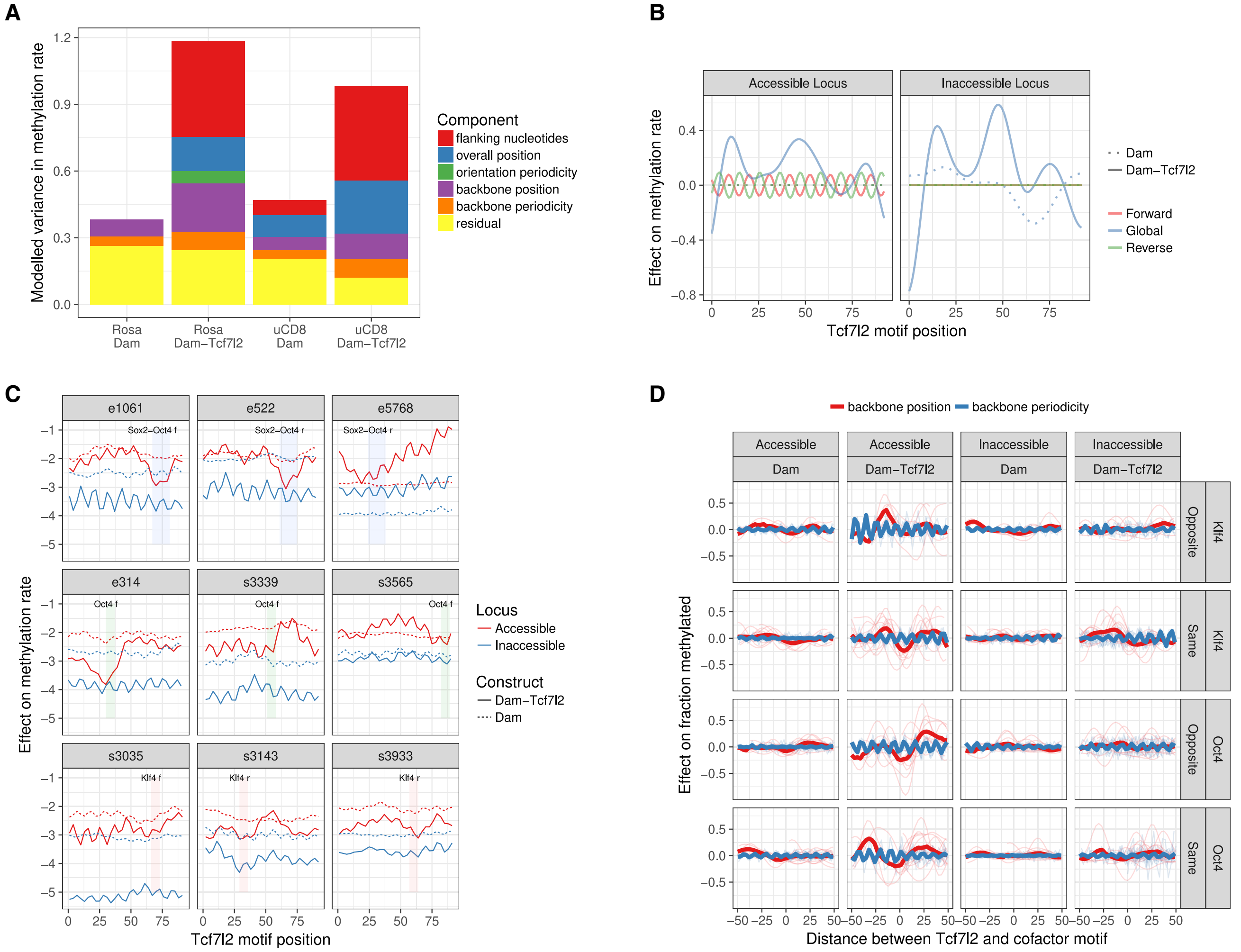


Figure 5. Tcf7l2 Binding Depends on Locus and Cofactor-Dependent Spatial Interactions

(A) Backbone phrase-specific variance in methylation rate explained by Gaussian process components across both accessible and inaccessible loci, and Dam-Tcf7l2 and unfused Dam.

(B) Estimated effect of the backbone phrase-invariant smooth and orientation specific periodic effects.

(C) Observed methylation components extracted from the Gaussian process for a representative set of backbone phrases. Shown is the sum of the backbone-phrase-specific components: constant, smooth, and periodic.

(D) Aggregate profiles of the backbone phrase-specific smooth and periodic components for Dam-Tcf7l2 or unfused Dam methylation, grouped by cofactors present. Phrases are centered at midpoint of the respective cofactor motif and calculated to the midpoint of the Tcf motif for each of the possible orientations. Sox2-Oct4 joint motifs are excluded as they would exaggerate the average signal compared with other Oct4 sites.

The smooth component captures the footprints of reduced binding as an underlying cofactor motif (Oct4 or Klf4) is disrupted (compare backbone phrases e522 and s3035 in Figure 5C with Figure 4E). Additionally, the smooth component captures regions of higher Tcf7l2 binding near the Oct4 and Klf4 cofactor motifs (2nd and 3rd row in Figure 5C), which tend to spread over an ∼20- to 50-bp region adjacent to the cofactor motifs. It is possible that another motif is being disrupted after these stretches of higher binding; however, since we could not reliably identify any other motifs at these locations we believe that it represents the effect of Oct4 or Klf4 promoting adjacent Tcf7l2 binding within this optimal window of proximity.

The backbone phrase-specific periodicity captures a reproducible oscillation in binding strength estimated to occur every 10.8 bp. We hypothesize that backbone phrase-specific periodicity in Tcf7l2 binding arises from interaction between Tcf7l2 and cofactors such as Oct4 that is strengthened when both factors reside on the same side of the DNA helix. The estimated periodic effects are similar for Dam-Tcf7l2 across both loci (0.35 correlation, versus 0.075–0.14 to unfused Dam; Figure S3), indicating that it is detecting oscillating patterns specifically promoting adjacent Tcf7l2 binding within this window.

The backbone phrase-specific smooth and periodic effects across different backbone phrases tend to align when aggregated across different backbone phrases based on the relative position between the Tcf and the Oct4 or Klf4 cofactor motif (Figure 5D). There is a large variability in the strength of these backbone phrase-specific effects across different backbone phrases—the three backbone phrases that contain a joint Sox2-Oct4 motif possess the strongest such effects (first row of Figure 5C). Since there are only 3 such backbone phrases, they are excluded from the aggregate profiles in Figure 5D to avoid exaggerating them, but the estimated positional effects overlap for the shared part of the orientation and gap spacing (Figure S5) These three Sox2-Oct4 motif containing backbone phrases show strong dips when the Sox2-Oct4 motif is disrupted, longer than average stretches of improved Tcf7l2 binding nearby, and strong backbone phrase-specific periodic effects, with maximal Tcf7l2 binding when Tcf7l2 is separated from Oct4 by 27–29, 37–39, 48–50, and 61 bp (relative to the midpoint of the Tcf7l2 and Oct4 motifs and similar across all four orientations of motifs). Since we were unable to accurately model the strength of the cofactor binding from their sequence motifs, this periodic effect had to be estimated for each backbone phrase specifically. As such, information is pooled across all positions of the tiled Tcf motif, resulting in the periodic effect being estimated to also pass through the cofactor motif. Thus, aligning backbone phrases by the distance between the Tcf7l2 motif and Oct4/Klf4 cofactor motif reveals that the smooth and oscillating effects are consistent and are accentuated in the three Sox2-Oct4 motif containing backbone phrases where cofactor binding strength is likely to be strongest.

The variability in Dam-Tcf7l2 methylation is considerably higher than that of unfused Dam seen when tiling the Tcf motif (4× higher in the accessible locus, 2× higher in the inaccessible locus, Figure 5A). Unfused Dam methylation has much higher proportion of unexplained residuals (accessible 75%, inaccessible 50%), suggesting that it is being influenced by some feature not utilized in the model—for example the specific sequence being overwritten by the tiled Tcf motif. The unfused Dam model identifies minor contributions from backbone phrase-specific effects (accessible 25%, inaccessible 20%) and an effect of overall position of the Tcf motif only in the inaccessible locus (15% of variability). The inaccessible locus also has ~10% of variance explained by contribution of flanking nucleotides; however, in this case it does not recapitulate the *in vitro* Tcf position weight matrix and instead appears to resemble a putative weak Sox2-Oct4 motif: 5′ Cs and 3′ A-T-G bias (Figure 4C). These results suggest that the unfused Dam signal in inaccessible chromatin is capturing subtle changes in chromatin accessibility arising from the creation of weak alternative motifs while tiling the Tcf motif—the Tcf, Sox2, and Oct4 motifs all share a core TTTG stretch, making it difficult to definitively assign the most likely binding factors to sites with weak position matrix weight matches. The effect of cofactors influencing accessibility can also be seen in the smooth backbone phrase-specific effect with unfused Dam, which while weaker is correlated with Dam-Tcf7l2 (accessible: 0.35, inaccessible: 0.18; Figure S3). In sum, the Gaussian process regression model is less effective at determining the causes of variation in unfused Dam methylation, likely because the input features have been tailored to predicting Tcf7l2 binding variation. This finding reinforces that the model is learning features specific to Tcf7l2 binding and not to confounders introduced by the DamID method.

In summary, in-depth analysis of this collection of 2,000 phrases that examine Tcf7l2-binding logic at 59-Tcf motif containing backbone phrases reveals that Tcf7l2 binding is dependent on the binding of Oct4 and Klf4 motifs in a spatially dependent manner. Strongest Tcf7l2 binding occurs when the motifs are separated by 20–50 bp with ~10.8-bp oscillatory strength that matches the turn of a DNA helix. The strength of these effects appears dependent on the strength of the cofactor motif: it is strongest at joint Sox2-Oct4 sites, moderate at most other Oct4 and Klf4 sites, and weak at a few Oct4 and Klf4 motifs. Since Tcf7l2 binding is strongly dependent on sequence context, it is likely that similar effects exist that govern the binding strength of Oct4 and Klf4. Further deconvolution all of these variables will require a substantially larger set of backbone phrases that systematically vary in Oct4/Klf4 motif strength as well as relative Tcf7l2 position. The lack of other observed motifs in any observed footprints nor in the flanking nucleotides suggests that there are unlikely to be many other motifs that substantively influence mESC Tcf7l2 binding in the backbone phrases analyzed.

DISCUSSION

The binding motif for Tcf7l2 has been well characterized *in vitro*; however, it is a poor predictor of the *in vivo* binding of Tcf7l2 and lacks the ability to explain differences in cell-type-specific binding (Frietze et al., 2012). Here, we use a combination of site-specific integration and *in vivo* transcription factor-binding measurement to show that a large contribution to the specificity of Tcf7l2 binding in mESCs is contained within the 99 bp of sequence surrounding a genomic Tcf motif.

In particular, by systematically varying the position of a core Tcf7l2 motif relative to cofactor motifs, we find that the presence of Oct4 and Klf4 motifs in an adjacent window of ~20–50 bp with spacing that places these motifs on the same side of the DNA helix as Tcf7l2 produces optimal mESC Tcf7l2 binding. We also show by inserting the same cohort of phrases into an intrinsically accessible and inaccessible genomic locus that Tcf7l2 binding is strongly determined by the local chromatin accessibility, as Tcf7l2 gains the ability to bind in mESCs at some sites usually only bound by Tcf7l2 in intestinal endoderm when these sites are transplanted to accessible chromatin. Thus, our results provide strong supporting evidence with an earlier classification of Tcf7l2 as a ''migrant'' transcription factor that is dependent on both local chromatin accessibility and on interactions with cofactors for binding (Sherwood et al., 2014).

A similar helical-dependent enhancing of co-binding as shown here for Tcf7l2 with Oct4 and Klf4 has been observed previously in the *in vitro* formation of a Lef1 (part of the Tcf family), Ets1, and CREB complex, which found that Lef1 promoted interactions between the flanking motifs in a way dependent on the phase of CREB in the DNA helix (Giese et al., 1995). Our work extends this to show that DNA helix-influenced cofactor binding can occur across several helical turns and that such subtle effects of spatial positioning between transcription factor-binding motifs play important roles in determining binding in a genomic context.

While the DNA-binding domains of transcription factors are generally small and well folded, the remaining domains responsible for interactions with other proteins are often disordered or connected by flexible segments (Liu et al., 2006). This flexibility should allow interacting domains of two adjacently bound transcription factors to interact across a range of spacings. This predicted flexibility in interaction distance is consistent with results from small-scale enhancer reporter assays (Erceg et al., 2014), which found that varying the relative spacing between the pMAD and Tin motifs can affect expression of a reporter gene integrated in a developing fruit fly (*Drosophila*) embryo. In one tissue, shifting from a 2- to 8-bp gap is enough to abrogate reporter expression, whereas in a different tissue it only halves expression. Importantly, in both cases this change is gradual: gaps of 4 and 6 bp exhibit intermediate function. Similarly, Farley et al. (2016) tested variants of an enhancer that is active during sea squirt development (*Ciona intestinalis*) in an oligonucleotide reporter assay and found that small shifts in the distance between Ets- and Zicl-binding sites within can change the strength of the enhancer. Our work extends upon these studies by showing that these spatially flexible cofactor effects span over a larger spatial region at least up to 50 bp and can influence binding of transcription factors.

Our finding that interactions between transcription factors can occur over a range of distances but also change in strength in oscillating fashion by shifting spacing by 3 bp has implications for current approaches that model and predict transcription factor binding. Typically, computational models of motif interactions are designed to allow for a constant effect irrespective of their relative positions, or to occur at an invariant spacing. For example, position weight matrices, linear regression, and support vector machines (i.e., Ghandi et al., 2014), all utilize sequence representations based on short subsequences or fixed gap lengths, that do not readily generalize across different gap lengths. Pooling across several different positions, for example, as is utilized in convolutional neural networks for protein-DNA binding (Alipanahi et al., 2015; Kelley et al., 2016), similarly blurs over the smooth and oscillating spatial effects observed here. There are certainly cases in which TFs may require fixed spacing as with Sox2-Oct4 binding to a directly adjacent dual motif (Chen et al., 2008; Jolma et al., 2015). It is likely, however, that cases like the Tcf7l2-Oct4 interaction we identify with oscillating and variable spacing are at least as typical. The covariance functions used here could be used directly as the kernel in a support vector machine to model such effects, while convolutional neural networks could utilize pooling layers across every N^{th} (e.g., every 10 or 11 bp) position rather than locally. Interestingly, a similar periodic effect has been inferred for Nanog relative to the Sox2 motif with a convolutional neural network from genomic ChIP-nexus data in mESCs (Avsec et al., 2019).

From a data standpoint, evaluating the effect of transcription factor interactions across spacings requires many more data points than treating the interactions as spacing-agnostic or spacing-invariant. It is highly unlikely that we could have identified the Tcf7l2-Oct4 interaction patterns we found using genome-wide transcription factor binding data. The genome does not have enough examples of these two motifs at varying spacings, and each genomic site has many other adjacent binding sites that would complicate modeling. Thus, our approach of integrating a designed set of sequences into fixed genomic locations enables fine-grained dissection of transcription factor logic in a way that is not possible from observational genomic data types such as ChIP-seq, DNase-seq, or ATAC-seq. Even with our approach using thousands of designed phrases, we are limited in power for building up an accurate model of Tcf7l2 binding. In several cases the residual methylation signal often appears to line up with the periodic components over a few positions, indicating that a more complex cooperative interaction exists within the data than is captured by a linear combination of spatial effects on Tcf7l2 binding.

Similarly, we lack enough observations to account for changing strength of Oct4 and Klf4 motifs (unlike for Tcf7l2 where we can accurately capture the strength of the motif through changes in flanking nucleotides). Particularly, the effect of Oct4 on Tcf7l2 binding in different backbone phrases shows a range of magnitudes, which prevents combining observations from several backbone phrases into accurate estimates of shape profile of Tcf7l2 cofactor interactions. This paucity of backbone phrase variation may explain why the Sox2 motif significantly modulates Tcf7l2 binding in the larger first screen but not in the second screen with limited examples. Klf4 motifs also showed variable strength in influencing Tcf7l2 binding across the second library. Additionally, some Sp1-like motifs—similar to the Klf4 motif—appeared like they might influence binding, but these effects were neither strong enough to produce a reliable ‘‘footprint’’ for specific instances nor consistent enough to detect a consensus motif. A more complete and predictive understanding of Tcf7l2 binding logic will require tiling or varying the strength of Oct4 or Klf4 motifs across Tcf motif containing backbone phrases as well as directly measuring binding of both Oct4/Klf4 and Tcf7l2 in the same set of backbone phrases.

Our experimental design is also confounded by using ectopic expression of Tcf7l2 that is fused to an enzyme, so the altered levels of Tcf7l2 expression or altered function of the fusion protein may not perfectly mimic native Tcf7l2 binding. Extrapolation to native Tcf regulatory circuitry would also need to account for the differential expression and large splicing heterogeneity of other Tcf/Lef family numbers (Weise et al., 2010), which would be expected to both accommodate for changes in function and compete for binding to the same motifs. Compared with alternatives, such as the lossy ChIP assay, the gain in resolution offered by DamID makes this trade-off worthwhile when looking at individual loci, which occur at most twice in each cell.

Overall, this study demonstrates the power of massively parallel integration of DNA-sequence variants into a controlled locus to address aspects of transcription factor-binding logic that are difficult to address using observational genomic approaches such as ChIP-seq or *in vitro* approaches such as protein binding matrix arrays. In the future, this approach could be expanded to address co-binding logic by profiling binding of multiple transcription factors to the same collection of sequences, dynamic transcription factor binding by profiling binding in different cell types, or combined with gene expression readouts to link transcription factor-binding patterns to gene regulatory activity.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead Contact
  - ○ Materials Availability
  - ○ Data and Code Availability

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2020.08.004.

ACKNOWLEDGMENTS

This work is supported in part by a Human Frontier Science Program Young Investigator Grant (RGC0084/2014), the National Health and Medical Research Council of Australia (1105271), the National Heart Foundation of Australia (100848), the National Institutes of Health (1K01DK101684, 1RO1HG008363, 1R01HG008754, NWO, and the American Cancer Society.

AUTHOR CONTRIBUTIONS

T.S., J.W.K.H., and R.S. designed the experiments. T.S., L.C., and R.S. collected the data. T.S. analyzed the data and wrote the code. T.S. and R.S. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 23, 2020
Revised: June 3, 2020
Accepted: August 4, 2020
Published: September 23, 2020

REFERENCES

Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of dna- and rna-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838.

Arbab, M., Srinivasan, S., Hashimoto, T., Geijsen, N., and Sherwood, R.I. (2015). Cloning-free Crispr. Stem Cell Rep *5*, 908–917.

Arce, L., Yokoyama, N.N., and Waterman, M.L. (2006). Diversity of LEF/TCF action in development and disease. Oncogene *25*, 7492–7504.

Aronesty, E. (2011). Command-line tools for processing biological sequencing data *7*, 1–8.

Atchison, J., and Shen, S.M. (1980). Logistic-normal distributions: some properties and uses. Biometrika *67*, 261–272.

Avsec, Ž., Weilert, M., Shrikumar, A., Alexandari, A., Krueger, S., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., and Zeitlinger, J. (2019). Deep learning at base-resolution reveals motif syntax of the cis-regulatory code. bioRxiv https://www.biorxiv.org/content/10.1101/737981v1.

Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X., et al. (2009). Diversity and complexity in dna recognition by transcription factors. Science *324*, 1720–1723.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J., et al. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell *133*, 1106–1117.

Cole, M.F., Johnstone, S.E., Newman, J.J., Kagey, M.H., and Young, R.A. (2008). Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells. Genes Dev *22*, 746–755.

Cook, D.R., and Weisberg, S. (1982). Residuals and Influence in Regression (Taylor & Francis).

ENCODE-DREAM Consortium (2017). ENCODE-DREAM in vivo transcription factor binding site prediction challenge. https://www.synapse.org/#!Synapse:syn6131484/wiki/402026.

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

Erceg, J., Saunders, T.E., Girardot, C., Devos, D.P., Hufnagel, L., and Furlong, E.E.M. (2014). Subtle changes in motif positioning cause tissue-specific effects on robustness of an enhancer's activity. PLoS Genet *10*, e1004060.

Farley, E.K., Olson, K.M., Zhang, W., Rokhsar, D.S., and Levine, M.S. (2016). Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. Proc. Natl. Acad. Sci. USA *113*, 6508–6513.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. J. Stat. Software *33*, 1–22.

Frietze, S., Wang, R., Yao, L., Tak, Y.G., Ye, Z., Gaddis, M., Witt, H., Farnham, P.J., and Jin, V.X. (2012). Cell type-specific binding patterns reveal that Tcf7l2 can be tethered to the genome by association With Gata3. Genome Biol *13*, R52.

Ghandi, M., Lee, Dongwon, Mohammad-Noori, M., and Beer, M.A. (2014). Enhanced regulatory sequence prediction using gapped K-mer features. PLoS Comput. Biol. *10*, e1003711.

Giese, K., Kingsley, C., Kirshner, J.R., and Grosschedl, R. (1995). Assembly and function of a Tcr alpha enhancer complex is dependent on Lef-1-induced dna bending and multiple protein-protein interactions. Genes Dev *9*, 995–1008.

Grant, S.F.A., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., Helgadottir, A., et al. (2006). Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat. Genet. *38*, 320–323.

Hashimoto, T., Sherwood, R.I., Kang, D.D., Rajagopal, N., Barkal, A.A., Zeng, H., Emons, B.J.M., Srinivasan, S., Jaakkola, T., and Gifford, D.K. (2016). A synergistic DNA logic predicts genome-wide chromatin accessibility. Genome Res *26*, 1430–1440.

Iacovino, M., Bosnakovski, D., Fey, H., Rux, D., Bajwa, G., Mahen, E., Mitanoska, A., Xu, Z., and Kyba, M. (2011). Inducible cassette exchange: a rapid and efficient system enabling conditional gene expression in embryonic stem and primary cells. Stem Cells *29*, 1580–1588.

Jolma, A., Yin, Y., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature *527*, 384–388.

Kelley, D.R., Snoek, J., and Rinn, J.L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res *26*, 990–999.

Kladde, M.P., and Simpson, R.T. (1992). Positioned nucleosomes inhibit Dam methylation in vivo. Proc. Natl. Acad. Sci. USA *91*, 1361–1365.

Klug, A., and Lutter, L.C. (1981). The helical periodicity of DNA on the nucleosome. Nucleic Acids Res *9*, 4267–4283.

Korinek, V., Barker, N., Moerer, P., van Donselaar, E., Huls, G., Peters, P.J., and Clevers, H. (1998). Depletion of epithelial stem-cell compartments in the small intestine of mice lacking Tcf-4. Nat. Genet. *19*, 379–383.

Korinek, V., Barker, N., Morin, P.J., van Wichen, D., de Weger, R., Kinzler, K.W., Vogelstein, B., and Clevers, H. (1997). Constitutive transcriptional activation by a beta -catenin-Tcf complex in APC−/− colon carcinoma. Science *275*, 1784–1787.

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

Liu, J., Perumal, N.B., Oldfield, C.J., Su, E.W., Uversky, V.N., and Dunker, A.K. (2006). Intrinsic disorder in transcription factors. Biochemistry *45*, 6873–6888.

Love, J.J., Li, X., Case, D.A., Giese, K., Grosschedl, R., and Wright, P.E. (1995). Structural basis for DNA bending by the architectural transcription factor LEF-1. Nature *376*, 791–795.

Nelson, W.J., and Nusse, R. (2004). Convergence of Wnt, beta-catenin, and cadherin pathways. Science *303*, 1483–1487.

Norton, L., Chen, X., Fourcaudot, M., Acharya, N.K., DeFronzo, R.A., and Heikkinen, Sami. (2014). The mechanisms of genome-wide target gene regulation by Tcf7l2 in liver cells. Nucleic Acids Res *42*, 13646–13661.

Pereira, L., Yi, F., and Merrill, B.J. (2006). Repression of Nanog gene transcription by Tcf3 limits embryonic stem cell self-renewal. Mol. Cell. Biol. *26*, 7479–7491.

Rajagopal, N., Srinivasan, S., Kooshesh, K., Guo, Y., Edwards, M.D., Banerjee, B., Syed, T., Emons, B.J.M., Gifford, D.K., and Sherwood, R.I. (2016). High-throughput mapping of regulatory DNA. Nat. Biotechnol. *34*, 167–174.

Rasmussen, C.E., and Williams, C.K.I. (2006). Gaussian Processes for Machine Learning (The MIT Press).

Rhodes, D., and Klug, A. (1980). Helical periodicity of DNA determined by enzyme digestion. Nature *286*, 573–578.

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

Sun, D., Xi, Y., Rodriguez, B., Park, H.J., Tong, P., Meong, M., Goodell, M.A., and Li, W. (2014). Moabs: model based analysis of bisulfite sequencing data. Genome Biol *15*, R38.

Szczesnik, T., Ho, J.W.K., and Sherwood, R. (2019). Dam mutants provide improved sensitivity and spatial resolution for profiling transcription factor binding. Epigenet. Chromatin *12*, 36.

Teytelman, L., Thurtle, D.M., Rine, J., and van Oudenaarden, A. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. Proc. Natl. Acad. Sci. USA *110*, 18602–18607.

Vogel, M.J., Peric-Hupkes, D., and van Steensel, B. (2007). Detection of in vivo protein-DNA interactions using DamID in mammalian cells. Nat. Protoc. *2*, 1467–1478.

Wang, J.C. (1979). Helical repeat of dna in solution. Proc. Natl. Acad. Sci. USA *76*, 200–203.

Weise, A., Bruser, K., Elfert, S., Wallmen, B., Wittel, Y., Wöhrle, S., and Hecht, A. (2010). Alternative splicing of Tcf7l2 transcripts generates protein variants with differential promoter-binding and transcriptional activation properties at Wnt/β-catenin targets. Nucleic Acids Res *38*, 1964–1981.

Wetering, M., and Clevers, J.C. (1992). Sequence-specific interaction of the HMG-box factor TCF-1 occurs within the minor groove of a Watson-Crick double helix. EMBO J *11*, 3039–3044.

Wetering, M., Oosterwegel, M., Dooijes, D., and Clevers, J.C. (1991). Identification and cloning of TCF-1, a T cell-specific transcription factor containing a sequence-specific HMG box. EMBO J *10*, 123–132.

Wilson, A.G., and Adams, R.P. (2013). Gaussian process kernels for pattern discovery and extrapolation. arXiv https://arxiv.org/abs/1302.4245.

Wong, E.S., Schmitt, B.M., Kazachenka, A., Thybert, D., Redmond, A., Connor, F., Rayner, T.F., Feig, C., Ferguson-Smith, A.C., Marioni, J.C., et al. (2017). Interplay of cis and trans mechanisms driving transcription factor binding and gene expression evolution. Nat. Commun. *8*, 1092.

Zambrowicz, B.P., Imamoto, A., Fiering, S., Herzenberg, L.A., Kerr, W.G., and Soriano, P. (1997). Disruption of overlapping transcripts in the ROSA beta geo 26 gene trap strain leads to widespread expression of beta-galactosidase in mouse embryos and hematopoietic cells. Proc. Natl. Acad. Sci. USA *94*, 3789–3794.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). Pear: a fast and accurate Illumina paired-end read merger. Bioinformatics *30*, 614–620.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Tcf7l2 DamID on 12,000 oligo library	This paper	data/1st-library-raw-values.txt at https://gitlab.com/tszczesnik/tcf-grammar-analysis
Tcf7l2 DamID on 2,000 oligo library	This paper	data/2nd-library-raw-values.txt at https://gitlab.com/tszczesnik/tcf-grammar-analysis
Experimental Models: Cell Lines		
129P2/OlaHsd mouse embryonic stem cells	(Iacovino et al. 2011)	N/A
Oligonucleotides		
12,000 170-bp oligo pool	CustomArray	data/1st-library-sequences.fa at https://gitlab.com/tszczesnik/tcf-grammar-analysis
2,000 150-bp oligo pool	Twist Biosciences	data/2nd-library-sequences.fa at https://gitlab.com/tszczesnik/tcf-grammar-analysis
Recombinant DNA		
plasmid: DamN126A and Tcf7l2-DamN126A	(Szczesnik et al.,2019)	N/A
Software and Algorithms		
fastq-multx	(Aronesty 2011)	N/A
PEAR	(Zhang et al. 2014)	N/A
Bwa	(Li and Durbin 2009)	N/A
Glmnet	(Friedman et al., 2010)	N/A
negative-binomial normalisation	This paper	statistical-analysis/negative-binomial-dropout.stan at https://gitlab.com/tszczesnik/tcf-grammar-analysis
Gaussian process regression	This paper	gp-damid/ at https://gitlab.com/tszczesnik/tcf-grammar-analysis

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources should be directed to the Lead Contact, Dr Rich Sherwood (rsherwood@rics.bwh.harvard.edu).

Materials Availability

This study did not generate new materials.

Data and Code Availability

The data and code are available online at https://gitlab.com/tszczesnik/tcf-grammar-analysis.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All experiments were done in 129P2/OlaHsd mouse embryonic stem cells (mESC), which were cultured according to previously published protocols (ENCODE Project Consortium, 2012). mESCs were maintained on gelatin-coated plates feeder-free in mESC media composed of Knockout DMEM (Life Technologies) supplemented with 15% defined fetal bovine serum (FBS) (HyClone), 0.1mM nonessential amino acids (NEAA) (Life Technologies), Glutamax (GM) (Life Technologies), 0.55mM 2 β-mercaptoethanol (Sigma), 1X E SGRO LIF (Millipore), 5 nM GSK-3 inhibitor XV and 500 nM UO126. Cells were regularly tested for mycoplasma. The non-homologous end joining pathway was disabled by knocking out two necessary genes (Prkdc and Lig4), along with constitutive activation of Rbbp8, which together increase the rate of homologous recombination (Arbab et al., 2015).

METHOD DETAILS

Phrase Library Design

The first phrase library (12,000 170bp) was ordered from CustomArray. Two 20bp primer sites were located at each end. One end also included a short (11bp) barcode followed by another primer site (20bp) for separately amplifying it. 11bp is necessay to have enough unique barcodes for all 12,000 phrases such that they are separated by an edit (levenshtein) distance of at least 3, so that any single error to be automatically corrected. Due to the presence of truncated library phrases, however, this barcode couldn't uniquely identify phrases and wasn't used. The remaining 99-bp was used to test out possible Tcf7l2 binding sites.

A portion of phrases were generated from ChIP-seq peaks bound in either Tcf7l1 ChIP-seq in mESCs or Tcf7l2 ChIP-seq in intestinal endoderm cells, or in both (1200 phrases each, split into two groups of 600 with or without a clear Tcf motif), along with 2400 phrases with a clear Tcf motif, occuring near marks of H3K27ac, but distal to (>10kb) to any Tcf ChIP-seq binding site. Due to the similarity between Tcf7l1 and Tcf7l2, Tcf7l1 ChIP-seq is a reliable measure of Tcf7l2 binding; genome-wide DamID for Tcf7l2 in mESCs correlates stronger with the Tcf7l1 mESC ChIP-seq than with Tcf7l2 intestinal endoderm ChIP-seq (Szczesnik *et al.*, 2019).

Phrases made up of combinations of various binding motifs were generated as follows:

1. Sample 25 (out of 35 possible) combinations of the non-Tcf7l2 and non-pioneer motifs (Hnf4a, Gata3, Foxa1/o1, c-Myc, Oct4, Sox3, Smad3).
2. Each of these generates a combination of with and without a Tcf motif.
3. Each of these generates a combination of no pioneer, Ets, Klf, or both motifs.
4. Sample 3 permutations out of each of these.
5. Sample 3 different gap lengths for each phrase.
6. Sample 3 randomly generated sequences for the gap

The second oligo pool (2,000 149bp) was ordered from Twist Biosciences. Two 25bp primer sites are located at each end, each containing a GATC, flanking a 99-bp variable region. Backbone phrases were chosen from the first library based on the probability that the posterior distribution of each phrase's methylation rate was either lower than, contained, or higher than the average methylation rate of the library ($p < 0.05$). A set of ~30 phrases per backbone phrase were generated by tiling the Tcf motif (ATCAAAG) every 3bp from the starting position. Motifs were scrambled to a specific sequence as follows:

- Tcf: ATCAAAG -> AACGTCG
- Oct: ATGCAAAT -> ATCGGCAT
- Klf: GCCACACCCA -> GCGAGACGCA
- Sox: CWTTGT -> CGTACT
- Gata: AGATA -> ACCCA

Phrase Library Integration

Oligo pools were amplified with primers at both ends (40ul NEBnext, 0.2-ul library, Ta=65, 30 cycles) and the 170bp band purified on a 4% agarose gel (Qiagen gel purification). Phrases were extended with homology arms (Table S1) (1000ul NEBnext, Ta = 65, 30 cycles) and purified (Qiagen minelute) to prepare for electroporation. 20 ug CBh Cas9-BlastR plasmid, 20 ug U6-gRNA-HygroR plasmid, and purified phrases were vacuum centrifuged to a final volume of <20 ul, added to 120 ml EmbryoMax Electroporation Buffer (ES-003-D, Millipore), and mixed with mESCs pelleted from a 15cm plate (~2e7 cells). This was transferred into a 0.4-cm electroporation cuvette and electroporated using a BioRad electroporator (230 V, 0.500 mF, and maximum resistance). Cells were passaged three times following integration.

DamID

Constructs were made by fusing Dam or Dam-N126A to the N-terminus of Tcf7l2 with a short flexible linker (Szczesnik et al., 2019). Dam-Tcf7l2 fusion and Dam only constructs are expressed from a single-integration Dox-inducible transgene expression cassette (Iacovino et al., 2011). This puts the Dam constructs under control of a tet-responsive promoter, along with integrating a neomycin resistance gene that is selected for by culturing the cells in G418 (300μg/mL) for one week. mESCs were cultured in 15cm plates and split at low ratios to ensure a high library diversity was maintained. Following expression of Dam fusion protein (8 hours for wild-type, 24 hours for N126A), genomic DNA is extracted and split into two pools that are digested by either DpnI, which cuts all methylated phrases, or DpnII, which cuts all unmethylated phrases (16ug DNA in 100ul and 100U enzyme, for 16 hours at 37C). Integration of the phrase library and presence of methylation was measured by qPCR on the DpnI and DpnII digest for the integrated site and control genomic Tcf7l2 bound locations. Completion of the DpnII digests was tested by undetectable DNA in controls cells, and of DpnI by heavy methylation with long wild-type unfused Dam expression (>90% methylated). Since the same conditions remove all traces of methylated bacterial plasmids, we infer that the DpnI digest is also close to 100%.

Next Generation Sequencing

Phrases were PCR amplified following DpnI / DpnII digestion with primers outside the homology arms and spanning the GATC site (16 cycles). This makes integrated phrases at least 100 times more numerous than unintegrated phrases (measured by qPCR), so that they dominate the signal. Two further short PCRs extend each phrase with adapters for illumina sequencing and a unique barcode for each sample. The cycle number is determined for this PCR based on qPCR to obtain sufficient amplification for Tapestation-based sample pooling and NGS. The resulting phrases are directly sequenced on a next-seq with midoutput 300bp kit (150bp read one, 150bp read two).

QUANTIFICATION AND STATISTICAL ANALYSIS

Alignment

Reads from different samples were demultiplxed with fastq-multx (Aronesty, 2011) based on short barcodes incorporated at the start and end of each sequence. Prior to alignment, overlapping paired end reads were assembled into a single read using PEAR (default parameters) (Zhang et al., 2014), in order to reduce the false positive rate stemming from truncated phrases. The assembled reads are aligned to the ordered library of sequences using BWA (mem algorithm with default parameters) (Li and Durbin, 2009). Counts are generated for each sequence by summing up all exact matches to it.

Deduplication

A negative-binomial dropout model is used to estimate the degree of PCR amplification and read depth present in the sequencing data, and hence to infer the number of unique genomic phrases in the original pool of cells. In principle the genomic counts could be estimated from the numbers of cells and integration efficiency, but in practise the uncertainties in each measurement are high and do not account for the variable abundance of each phrase.

The initial distribution of counts before amplification (x) for each digest is modelled as a negative binomial with shape and rate parameters α and β. This is a natural choice for positive count data, and can fit the positive tail in the data and the underlying unimodal distribution (i.e. apart from the dropouts and staggering stemming from the amplification of low counts).

Observed read counts are modelled as coming from a Poisson distribution stemming from a linear amplification (rate γ) of these latent counts (Equation 1). A constant amplification rate across all phrases is assumed since they are of the same length and similar GC content. This leaves any phrases missing in the original distribution as 0, while 1 shifts to a peak centered on γ, 2 to 2γ, and so on, creating staggered peaks that eventually run into one another. β is constrained to be identical between the DpnI and DpnII digests, which prevents the model from assuming different methylation rates between digests, and is consistent with the total amount of methylation across the phrase library following a beta distribution.

$$\begin{aligned} p(y|\alpha,\beta,\gamma) = {} & nbinom(y|\alpha,\beta)nbinom(0|\alpha,\beta) \\ & + \sum_{x=1}^{\infty} pois(y|\gamma x)nbinom(x|\alpha,\beta) \end{aligned} \qquad \text{(Equation 1)}$$

In the first phrase library, we found that modelling the dropout reads as following the original unamplified background distribution works well, in comparison to low rate Poisson, since it captures the shape of the low level contamination from unintegrated background phrases better. In the second phrase library, due to the higher overall sequencing depth, we observed a clear background contaminating population that stems from unintegrated phrases persisting in the cells following electroporation and being amplified by the 2nd and 3rd library preparation PCR cycles. In this case we use a mixture model (R package mixtools) to fit two 2D log-normal distributions (equal variance, starting position means (DpnI count, DpnII count): (1000, 10) and (1000, 1000)), and zero-out the DpnII counts of the lower population prior to estimating the amplification rate (see Figure S6).

To avoid the expensive sum over all possible discrete counts (Equation 1), and since the important information comes from 0 and 1 counts, we approximate higher counts with a continuous distribution. We rewrite our original negative binomial as a Poisson-gamma mixture with an explicit latent count rate $\hat{x}$, shift the amplification rate to it and simplify down (Equation 2).

$$\begin{aligned} p(y) & = \int_0^{\infty} pois(y|\gamma\hat{x})gamma(\hat{x}|\alpha,\beta)d\hat{x} \\ & = nbinom\left(y\middle|\alpha,\frac{\beta}{\gamma}\right) \end{aligned} \qquad \text{(Equation 2)}$$

To avoid double counting the latent genomic counts between 0 and 1 we subtract away the portion of continous counts that would have given rise to 0 and 1 observed counts; $p(\hat{x}|x=0,1)$ (Equation 3).

$$\begin{aligned} p(y|x) & = p(y|\hat{x})p(\hat{x}|x) \\ & = \int_0^{\infty} pois\left(y\middle|\gamma\hat{x}\right)gamma\left(\hat{x}\middle|\alpha+x,\beta+1\right)d\hat{x} \\ & = nbinom\left(y\middle|\alpha+x,\frac{\beta+1}{\gamma}\right) \end{aligned} \qquad \text{(Equation 3)}$$

Equation 4 shows the final likelihood which combines the discrete distribution 0 to k low counts with continous higher counts, with the structure:

1. Likelihood of unamplified distribution coming from zero counts.
2. Counts from amplified discrete distribution.
3. Counts from amplified continuous distribution.
4. Subtract component of the continuous distribution that is already modelled discretely.

$$\begin{aligned} p(y|\alpha,\beta,\gamma) &= nbinom(y|\alpha,\beta)nbinom(0|\alpha,\beta) \\ &+ \sum_{x=1}^{k} pois\left(y\middle|\gamma x\right)nbinom\left(x\middle|\alpha,\beta\right) \\ &+ nbinom\left(y\middle|\alpha,\frac{\beta}{\gamma}\right) \\ &- \sum_{x=0}^{k} nbinom\left(y\middle|\alpha+x,\frac{\beta+1}{\gamma}\right)nbinom\left(x\middle|\alpha,\beta\right) \end{aligned} \tag{Equation 4}$$

α, β, and γ are estimated using Hamiltonian Monte Carlo sampling implemented in the statistical programming language Stan. Following normalisation the distribution appears to follow the expected independent and identically distributed binomial distribution between replicates, and so further inference the per phrase DpnI and DpnII counts are summed up across replicates.

Beta-Binomial Model

A beta binomial distribution was fit to the normalised DpnI / DpnII count data using the dbetabinom (VGAM package) and mle2 functions (bbmle package) in R. Starting parameters were $\alpha = \beta = 2$ (mean of 0.5 with large spread). Conjugacy between the binomial distribution (for read counts) and beta distribution (for methylation rates) results in a straightforward calculation for the posterior beta distribution of methylation rates for each phrases (Equation 5). Beta-binomial models are used to the same effect in reducing the fase positive rate of detecting methylated cytosines with few supporting reads in genome-wide bisulfite sequencing (Sun et al., 2014).

$$p(r_i|a_i,b_i) = beta(a_i+\widehat{\alpha}, b_i+\widehat{\beta}) \tag{Equation 5}$$

Generalised Linear Model

Effects of motifs or genomic context were calculated using a generalised linear model (R, glmnet package (Friedman et al., 2010)) with binomial output based on the methylated (DpnII) and unmethylated (DpnI) counts (i.e. logistic regression with multiple measures per point). Equation 6 shows the model log-likelihood: β are the linear weights, x_i the features for the ith phrase, y_i is an indicator for a methylated count. See (Friedman et al., 2010) for details on estimating β to maximise the log-likelihood. Lasso regularisation is used, with the penalty strength (λ) set by minimising the mean-squared error upon 10-fold cross-validation. This shrinks non-significant features to 0, and puts significant differences to be detected past ~0.02 effect size. By using the normalised DpnII and DpnI counts the variance in the estimate of each phrase's methylation is retained, preventing the logistic regression from overfitting past it.

$$logp(y|x,\beta) = \frac{1}{N}\sum_{i=1}^{N}\left[y_i\left(\beta_0 + x_i^T\beta\right) - log\left(1 + exp\left(\beta_0 + x_i^T\beta\right)\right)\right] - \lambda\left\|\beta\right\| \tag{Equation 6}$$

Gaussian Process Regression

Gaussian process regression is used to model the fraction methylated across the phrase library (for an overview of their use in machine learning see (Rasmussen and Williams, 2006)). For each (ith) phrase a logit-normal approximation that exactly matches the first two moments (μ_i, σ_i) to the posterior beta function (α_i, β_i) was used (Equation 7), where ψ and ψ_1 are the digamma and trigamma functions. This approximation does not hold at the boundary values of 0 or 1, however since such methylation fractions only occur at low read counts, the beta-binomial model shifts posterior mean for each phrase away from the boundary and towards the mean library methylation rate. The Kullback-Liebler divergence ranges across 2×10^{-6} to 5×10^{-2} for read counts on the order of up to 100 that have values shifted away from the boundaries (Atchison and Shen, 1980).

$$\begin{aligned} \mu_i &= \psi(\alpha_i) - \psi(\beta_i) \\ \sigma_i &= \psi_1(\alpha_i) + \psi_1(\beta_i) \end{aligned} \tag{Equation 7}$$

The Gaussian process fits the posterior mean (μ_i) with a linear mixture of several functions (weights w_a^2 for the *a*th function) based on the phrase (x_i). The posterior variance (σ_i) is included with a fixed weight ($w_0 = 1$), which prevents the Gaussian process from fitting beyond the sampling resolution of the data.

The Gaussian process models several effects which are listed in Table S2, these comprise of the individual phrase variance (posterior variance and residuals), constant linear effects (constant across all phrases, or constant across phrases from the same backbone phrase, and flanking nucleotide effects), and several non-linear position based functions designed to estimate how the position of Tcf motif affects the binding of Tcf7l2 to it. These non-linear position based functions are either shared across all phrases, phrases sharing the same Tcf motif orientation, or only those phrases generated from the same backbone phrase ($b = \{0,1,2\}$ is used to refer to these cases respectively). The class of position based functions is determined by a covariance function based on the relative position of the Tcf motif between two different phrases (p_i and p_j). Two such position based covariance functions are used:

- Smoothly varying functions (k_{rb}) modelled with a radial basis covariance function, parametrised by a length scale λ_b (Equation 8). For examples of functions specified by this covariance function see Figure S10A.

$$k_{rb}\left(p_i, p_j\right) = e^{\frac{-\left(p_i - p_j\right)^2}{2\lambda_b^2}} \qquad \text{(Equation 8)}$$

- Periodic functions (k_{pr}) modelled with a periodic covariance function parametrised by inverse periodicity (τ_b) and length scale (ρ_b) (Equation 9) (Wilson and Adams, 2013). For examples of functions specified by this covariance function see Figure S10B.

$$k_{pr}\left(p_i, p_j\right) = e^{-2\left(\pi\left(p_i - p_j\right)\rho_b\right)^2} cos\left(2\pi\left(p_i - p_j\right)\tau_b\right) \qquad \text{(Equation 9)}$$

Since covariance functions are closed under addition, a linear mixture of these positional effects along with the per phrase and other linear effects produces a valid (positive definite) covariance function. This is shown in Equation 10 using the Kronecker delta notation. *i'* and *j'* refer to the orientation of the Tcf motif, and *i*'' and *j*'' refer to the original backbone phrase (i.e. $\delta_{i'j'} = 1$ only if the phrases share the same Tcf motif orientation, and $\delta_{i''j''} = 1$ only if the phrases share the same original backbone phrase). f_{ic} is the position of the 5' flanking nucleotide of phrase i at position c {1, 2}; t_{ic} for the 3' flanking nucleotide.

$$\begin{aligned} k_{total}(x_i, x_j) &= \delta_{ij}\sigma_i\sigma_j + \delta_{ij}w_1^2 + w_2^2 + \delta_{i''j''}w_3^2 \\ &+ w_4^2 \sum_{c=1}^{2}\left(\delta_{f_{ic}f_{jc}} + \delta_{t_{ic}t_{jc}}\right) \\ &+ w_5^2 k_{rb}\left(p_i, p_j \middle| \lambda_0\right) + w_6^2 k_{pr}\left(p_i, p_j \middle| \tau_0, \rho_0\right) \\ &+ \delta_{i'j'}\left(w_7^2 k_{rb}\left(p_i, p_j \middle| \lambda_1\right) + w_8^2 k_{pr}\left(p_i, p_j \middle| \tau_1, \rho_1\right)\right) \\ &+ \delta_{i''j''}\left(w_9^2 k_{rb}\left(p_i, p_j \middle| \lambda_2\right) + w_{10}^2 k_{pr}\left(p_i, p_j \middle| \tau_2, \rho_2\right)\right) \end{aligned} \qquad \text{(Equation 10)}$$

The fit of the model is evaluated by the likelihood fit to the data (Equation 11). The covariance matrix *K* is constructed by evaluating the covariance function (k_totak) for all pairs of phrases.

$$\log p(\mu|x) = -\frac{1}{2}\left[\mu^T K^{-1}\mu + log|K| + n\log 2\pi\right] \qquad \text{(Equation 11)}$$

Different functions comprising the Gaussian process fit are extracted using Equation 12, where $\widehat{K}$ is constructed only from the portion of the covariance function to be extracted. Cross validated fits are similarly calculated as in Equation 12 based on a formula for linear smoothers, with $\widehat{K}$ excluding the posterior variance and residual components (Cook and Weisberg, 1982).

$$\widehat{\mu} = K\widehat{K}^{-1}\mu$$

$$\widehat{\mu}_{-i} = \mu_i - \left[I - K\widehat{K}^{-1}\right]_{ii}^{-1}\left(\mu_i - \widehat{\mu}_i\right) \qquad \text{(Equation 12)}$$

The weights (w_a) and the hyperparameters ($\lambda_b, \tau_b, \rho_b$) were found by gradient based optimisation of the likelihood using L-BFGS. The optimal periodicity ($\frac{1}{\tau_b}$) was found by grid search over all integer periodicities (up to 30bp) prior to gradient based optimisation. All algorithms in this section in Haskell through bindings to linear algebra (Blas and Lapack) and optimisation (libLBFGS) libraries. Cholesky decomposition is used for inverting the covariance matrix.