# Non-consensus flanking sequence of hundreds of base pairs around in vivo binding sites: statistical beacons for transcription factor scanning

**Kateřina Faltejsková**[1,2,✉], **Josef Šulc**[1,3], **and Jiří Vondrášek**[1,3,✉]

[1]Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Prague 6 166 10, Czech Republic
[2]Computer Science Institute of Charles University, Faculty of Mathematics and Physics, Charles University, Malostranské náměstí 25, Prague 118 00, Czech Republic
[3]Faculty of Science, Charles University, Albertov 2038, Prague 128 00, Czech Republic

**It was long suspected that for specific DNA binding by a transcription factor, the flanks of the binding motifs can play an important role. By a thorough analysis of the DNA sequence in the broad context (± 5000 bp) of *in vivo* binding sites (as identified in a ChIP-seq or a Cut&Tag experiment), we show that the average GC content is in most cases statistically significantly increased around the binding site in a patch spanning 1000–1500 bp. This increase was observed consistently in experiment targeting the same TF in different cell lines. The surrounding of binding sites of certain TFs like MYC display a directional alteration of dinucleotide frequencies. We attempt to explain these preferences by alteration in DNA shape features as well as by potential cooperation with other TFs. We observed differences in sequence affinity to various potential cooperating TFs between cell lines. Altogether, we propose that the observed feature distortion is indicative of a coarse scanning mechanism that helps TFs find the target binding site.**

**genomics | gene regulation | transcription factors | transcription factor binding | ChIP-seq**

**Correspondence:** ***katerina.faltejskova@uochb.cas.cz***

## Introduction

Transcription factors (TFs) facilitate the transmission of information from DNA to downstream molecular processes. These molecules specifically bind motifs of mostly 6-12 bp (1, 2). Such a length (averaging ∼10 bp) balances binding specificity and the ability of TFs to regulate multiple genomic loci efficiently, although longer motifs have been observed (3).

It has been known for some time that only a fraction of motifs capable of being bound are truly bound in vivo at a given point in time (1). There appear to be several reasons for this selectivity. Firstly, many TFs do not function as monomers. The binding motifs of monomeric TFs and TFs in dimers can differ significantly (4, 5). These composite motifs are important for acquiring specificity during cell differentiation (5). Additionally, the affinity and biological role of TFs can change depending on cooperating molecules, highlighting a flexible functionality (6, 7). For instance, yeast contains more TFs with a wide activity range than those with uniquely defined functions (7).

When describing TF binding, increasing attention is being paid to the structure of the binding site. Describing the shape of the binding site can be more informative for binding predictions than the motif sequence alone (8–10). The structure and flexibility of DNA flanking the motif also significantly influence the overall affinity of a TF for its binding site (11–13). Recent studies emphasized flexibility of DNA flanking regions as crucial determinants of TF binding specificity (11).

By gradually shifting from a detailed to a broad perspective, we can conceptualize the recognition process as comprising several overlapping levels. At the lowest level, base readout involves direct recognition through chemical contacts, including hydrogen bonds, electrostatic interactions, water-mediated bonds, and hydrophobic contacts, all contributing directly to binding specificity and affinity. Namely, the preference for a given nucleotide at a specific position is mainly determined by interactions between the amino acid side chains of the TF and the accessible edges of the base pairs that are contacted.

Building on these chemical contacts, TFs also often read the genetic code through contacts in the major and minor grooves of DNA. In the major groove, side chains of amino acids such as arginine and asparagine form specific hydrogen bonds with the functional groups of base pairs, allowing discrimination between G•C and A•T pairs (14). For instance, zinc-finger protein Zif268 recognizes DNA via residues positioned at -1, +3, and +6 in its three-finger array, creating a combinatorial code by interacting specifically within successive three-base-pair subsites (14).

Minor groove recognition, conversely, often relies on DNA shape variations such as groove width and electrostatic potential rather than direct chemical complementarity. The TATA-binding protein (TBP) exemplifies this strategy by wedging phenylalanine residues into widened minor grooves to stabilize the complex (15).

This readout mechanism is often shown as an example of shape readout. Shape readout refers to transcription factor (TF) recognition based on structural alterations of the DNA helix (e.g., roll, twist, tilt, groove geometry) that vary with sequence context (8). For example, Hox proteins promote

bending through such perturbations, generating unique conformations that favor binding (11).

Incorporating shape data into previously sequence-only predictors significantly improves TF-binding site predictions (8, 16–18). Indeed, it has been shown on numerous occasions that integrating DNA-shape information into previously sequence-only based predictors gives consistent improvement in binding site prediction in a significant number of cases (8, 16–18).

For example, in the E2F4–DP2 complex, four arginines create eight hydrogen bonds specifically to G in the core GCGCG motif, selecting both the chemical groups and the rotational propeller twist of the G/C steps (16). It has also been shown that some TFs recognize structural features, such as DNA bending and unwinding (19–21). This shape readout strategy reads static and dynamic properties of DNA structure, as well as electrostatic changes of the DNA. It is considered very likely that most proteins use the interplay of base and shape readout to recognize their cognate binding sites (22, 23). Experimental techniques, including hydroxyl radical cleavage and atomic force microscopy, also confirm that sequence-dependent shape variations significantly influence protein binding (8, 17, 18, 24, 25).

Several studies have also indicated that TF preferences for specific DNA shapes extend beyond core motifs, with sequences immediately flanking TF-binding sites contributing substantially to binding specificity (12, 26, 27). Clearly, incorporating the sequence and structural context surrounding TF binding sites improves computational predictions and better aligns with observed biological outcomes. Additionally, it improves the fit between our models and *in vivo* reality (28).

Apart from the precise interplay of base and shape readout, less local influences help guide TF diffusion kinetics, optimizing search efficiency for target DNA sequences. These indirect interactions optimize the search time needed to find the right target DNA element.

Considering random naive 3D diffusion within the crowded nuclear environment, the likelihood of locating a 1 bp target (approximately 0.34 nm) is extremely low. Experimentally measured association rates are $\sim$100-fold faster than the maximum allowed by simple 3D diffusion (29).

To accelerate target-site localization, TFs utilize facilitated diffusion, combining 1D scanning (sliding, hopping, intersegmental transfer) with short-range 3D diffusion, significantly improving search efficiency (29–31). This is clearly seen in some larger organizational elements in the genome. Noticeably, repeating oligomeric poly(dA:dT) and poly(dC:dG) tracts seem to attract or repulse TFs to/from (non)specific binding sites, as has also been observed *in vitro* (32–34).

Positive design elements—extended poly(dA:dT) tracts—create low-energy 'highways' that statistically enhance nonspecific TF–DNA affinity and funnel TFs toward nearby cognate sites (34). Homo-oligonucleotide sequence correlations, such as these tracts, broaden the TF–DNA binding energy distribution, thus lowering overall nonspecific binding free energy. In contrast, DNA with alternating correlations narrows the energy spectrum and raises the nonspecific TF–DNA binding free energy (34). These factors allow for both positive and negative non-consensus design, greatly expanding the targeting system of the DNA transcription (33). Collectively, these elements shape DNA's geometry and electrostatic landscape, forming regions that attract or repel TFs over longer distances. This finely tuned structure optimizes the speed and specificity of TF binding.

Transcription factor (TF) targeting is significantly influenced by genome packaging. Nucleosomes, consisting of $\approx$ 147 bp of DNA wrapped around histone octamers, obscure binding sites through steric hindrance (35). Pioneer TFs uniquely access nucleosomal DNA, recognizing regions within tightly packed chromatin and initiating chromatin remodeling to permit subsequent transcriptional activation, unlike conventional TFs (34). Within fully wrapped nucleosomes, DNA sites remain largely hidden by histones. However, transient "nucleosome breathing" near DNA entry and exit points briefly exposes these hidden sites, enabling pioneer TFs to access closed chromatin (35). Subsequently, ATP-dependent chromatin remodelers (e.g., SWI/SNF, ISWI) extend these accessible regions by sliding or evicting nucleosomes, thus increasing DNA accessibility and facilitating transcriptional activation (36).

Chromatin compaction at larger scales into fibers and loops further increases molecular crowding. Live-cell single-particle tracking studies show TF localization predominantly in less dense chromatin regions, with slower diffusion upon encountering denser regions (37, 38). At megabase scales, topologically associating domains (TADs) and CTCF/cohesin-mediated loops funnel TFs into regulatory neighborhoods, thereby reducing the search space (37).

Combined with base-specific and shape-specific DNA recognition, as well as facilitated diffusion, these chromatin characteristics sculpt a multiscale free-energy landscape, directing TF movement from individual base pairs to higher-order genome structures.

TFs frequently contain intrinsically disordered regions (IDRs) facilitating multivalent interactions with DNA, histones, and additional regulatory proteins, significantly elevating local TF concentrations, enabling cooperative assembly, and enhancing regulatory efficiency (38). TF binding reshapes DNA geometry or exposes new protein-protein interfaces, facilitating cooperative binding with nearby TFs—a DNA sequence-dependent phenomenon termed TF cooperativity (39, 40). Concurrently, coactivator complexes (Mediator, p300) bridge multiple TFs and recruit chromatin remodelers, stabilizing large enhanceosome assemblies. This promotes sustained chromatin openness and non-linear transcriptional regulation (41).

Moreover, many TFs and coactivators undergo liquid-liquid phase separation at super-enhancers, forming dynamic condensates that concentrate DNA and regulatory proteins. Such condensates significantly reduce the dimensionality of TF searches (42).

**Alternative DNA Structures.** Regulatory DNA regions frequently adopt alternative (non-B) structures, providing dis-

tinct binding platforms for specialized TFs. G-quadruplexes (G4s) form in guanine-rich promoter and enhancer regions, attracting specialized G4-binding proteins that influence local chromatin architecture and transcription. Additionally, transcription-generated R-loops (RNA:DNA hybrids) create single-stranded DNA segments and alter local DNA supercoiling, modulating TF binding and diffusion properties (43).
In this work, we attempt to thoroughly describe the sequence features of the *in vivo* binding sites for 10 various TFs and several different cell lines. Building on the work of Dror et al., who proposed that TF binding sites are surrounded by a patch with altered GC content (increased around binding sites of transcription factors from *e. g.* SP1 and ELF1, decreased around binding sites of homeodomain transcription factors), we describe the sequence preferences and derivable features within the vast surrounding region of the binding sites. We discover meaningful distortion of GC content and dinucleotide frequencies spanning thousands of base pairs, as well as distortions of some DNA shape features and affinities to known transcription factors. We propose that these relatively far-reaching distortions are indicative of the 1D scanning, a coarse search mechanism that guides the TF towards a general region. Once a TF arrives to the vicinity of the target binding site, short-range 3D diffusion assumes a leading role in binding the target DNA site.

## Methods

**Data curation.** For our analysis, we curated a set of human TF ChIP-seq experiments from the ENCODE portal (44). We chose experiments targeting one of the following transcription factors: CTCF, FOXK2, IRF1, MEF2A, MYC, NANOG, NFRKB, RUNX1, SPI1, TP53. We limited our search to experiments without red flags in the audit category. From this set of experiments, we selected only the experiments associated with a biosample that appears at least three times (*i. e.*, H1, K562, HepG2, A549, MCF-7, GM12878, WTC11 and GM23338). For each available combination of transcription factor and biosample, we included a single experiment. In most cases, no more data was available.
As a representation of each experiment, we took the "conservative IDR thresholded peaks" file. If no such file was available, we represented the experiment with the "IDR thresholded peaks" file. There was only one such case (ENCFF692RPA, CTCF, H1 biosample). The files were downloaded on 17 January 2025. The final identifiers can be found in Table 1.
To validate our results using a different experimental method, we included a Cut&Tag experiment GSM3560258 (45). Using this data source, we replicated the peak calling process using MACS2 (46) as per the original work (45).

**Clustering of peaks based on GC content.** For each peak sequence in the ChIP-seq experiment results, we assigned the GC content as the percentage of guanine or cytosine bases. No peak size normalization was used.
Then, for each pair of experiments, we uniformly randomly sampled 1000 peak sequences from each experiment and we tried to separate them using agglomerative clustering (scikit-learn implementation (47)) and a trivial GC content difference distance function (see Results). For clustering, we try out single, complete and average linkage function and we report the result with the best accuracy. For each experiment pair, we repeat this process ten times.

**Feature extraction.** From each ChIP-seq experiment, 10 000 peaks were uniformly randomly sampled (if there were fewer peaks, all were included). For each of the acquired peaks, its center was identified. Each of the peaks was then prolonged 5 000 bp both upstream and downstream of the peak center, resulting in a 10 000 bp–long patch of DNA sequence for each sampled peak. We denoted this patch a "prolonged peak". We assume the binding site to be in the middle of a ChIP-seq peak (48). In the case of Cut&Tag experiments, we have used the summits as peak centers to prolong.
From each prolonged peak, we extracted 50 bp windows that span the entire length of the prolonged peak, with each window overlapping the previous one by 25 bp. For some analysis, we only use the non-overlapping windows.
To each of the windows, we assign the following values:

- **GC content**: the number of guanine (G) and cytosine (C) bases in the window sequence.
- **dinucleotide frequencies**: the number of a particular dinucleotide (*i. e.* CG or AC) in the window sequence.
- **TF affinity z-score**: for each window, we calculate maximum affinity to human transcription factors from the HOCOMOCO v12 non-redundant set (49) (both on the forward and reverse strand). This affinity score is then z-score normalized. In this paper, we work with two normalization datasets. To show that affinity to various TFs changes around the binding site, the affinity score is normalized with regards to affinity to sequences 5000 bases upstream of annotated transcription starts of RefSeq genes (hg19 genome assembly, downloaded from UCSC Genome Browser database (50), hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/).
- **presence of non-B DNA features**: indication whether the window has at least 5 bp overlap with a G-quadruplex or a Z-DNA motif. Information on these features was downloaded from non-B DB (51) (hg38 genome assembly, nonb-abcc.ncifcrf.gov/apps/ftp/).
- **intra- and inter-base DNA shape features**: for each window, we predicted DNA shape features from the sequence using deepDNAshape (52). The DNA shape feature include buckle, helix twist, minor groove width, opening, rise, roll, shear, shift, slide, stagger, stretch and tilt (53). For prediction, we set the number of layers used by deepDNAshape to 7. As the prediction is available for every base-pair (or every space between two of them), we took every tenth value in

the window. The prediction on the window edges were excluded so that every prediction has the context sequence available.

**Assessing the presence of conserved sequence motifs by sequence alignment.** As a control experiment, we constructed two multiple sequence alignments (MSA) for each BED file in our input dataset. For the first one, we randomly sampled 1000 peaks and took 150 bp upstream and downstream from each peak center (resulting in 300 bp sequences). The second one was constructed by sampling 100 prolonged peaks (see above). For alignment, we used the Clustal Omega command-line tool in default setting for DNA sequences (54). We will refer to these two alignments as the "MSA on peaks" and "MSA on prolonged peaks".

**Overlap of the prolonged peak and known regulatory features.** To link our finding to known regulatory features, we have also assessed whether the windows has at least 5 bp overlap with any known ENSEMBL regulatory features (enhancer, promoter, open chromatin region). The information on regulatory features was dowloaded from ENSEMBL (55), release 113.

**Statistical analysis.** To identify excessive patches in the prolonged peaks, we defined an "edge" of the prolonged peak as the first and last 500 bp of the prolonged peak. Then, we adopted a one-dimensional cluster-based permutation approach to identify contiguous runs of significant sequence-feature deviation around each peak. We then subtracted this edge mean from each peak's profile to yield a peak$\times$offset matrix of deviations. sparse adjacency matrix was constructed linking each offset to its immediate neighbors, and MNE's permutation_cluster_1samp_test was applied (1,000 permutations; cluster-forming threshold $|t| > 2.0$; two-sided) (56). Clusters whose summed t-mass exceeded the 95th percentile of the permutation null distribution ($p < 0.05$) were called significant, and their start/end offsets were recorded.
As stated above, for each feature and for each ChIP-seq or Cut&Tag experiment, we first summarized the 10 kb prolonged peak as a series of equally spaced windows (50 bp with 25 bp overlap).
Then, we attempted to identify excessive patches in the individual prolonged peaks. For every feature and every window, we calculated the respective z-score using the "edges" defined above as background distribution. We declared a window/feature pair excessive if it had absolute z-score of more than a predefined threshold, gaining a zero-one vector for every feature. On this vector, we then applied rolling mean on every five windows. Based on this vector, we declared a window/feature pair excessive after smoothing if its value is bigger than 0.5, *i. e.* if there is at least 3 out of 5 window/feature pairs around the given window that have z-score higher than 3. We consider a window $w$ excessive if there is any feature $f$ for which window $w$/feature $f$ pair is excessive after smoothing. This evaluation was used for visualization only; we did not derive further inference beside simple demonstration.

**Assessing biological relevance.** In recent years, the representation of binding site sequences has gained significant nuance. To evaluate the long-range effects of DNA sequence context, we chose not to incorporate predefined motif models into our analysis. Instead, we relied on experimental evidence: assuming that ChIP-seq or Cut&Tag peaks provide a credible description of the transcription factor's binding landscape within the cell (48, 57).
To ensure our long-range sequence-bias signals are not driven by low-confidence peaks, we ranked all ChIP-seq binding events by their MACS2 FDR (q-value) and then re-ran the analysis above on increasingly stringent subsets: a dataset consisting of (1) the top 75%, (2) the top 50% and (3) the top 25% of peaks.

**Use of artificial intelligence in preparing this manuscript.** We acknowledge the use of OpenAI's ChatGPT in preparing this manuscript. It provided support in drafting simple code snippets, smoothing English phrasing, and offering basic editing suggestions. All scientific ideas, analyses and conclusions, however, are solely those of the authors.

## Results

As can be seen on the Figure 1.1, there is considerable variability in the nucleotide frequency of a peak captured in different ChIP-seq experiments depending on the target TF as well as the biosample used in the experiment. In particular, for NFRKB and IRF1 the difference between median peak GC content is over 10 % depending on which cell line was used in the experiment.
These differences in GC content are, in some cases, sufficient to distinguish between loci bound in two different experiments. We defined a trivial 'distance' function as

$$\mathrm{dst}(S_0, S_1) := |\mathrm{GC}(S_0) - \mathrm{GC}(S_1)|,$$

where $\mathrm{GC}(S)$ denotes the fraction of G or C bases in sequence $S$. Using this function, we were able to cluster the sequences using agglomerative clustering with up to 70–80 % accuracy (see Figure 1.2a and b).
Beyond the GC content fraction, traditional bioinformatic methods reveal no similarity or consensus. In both MSA on peaks and MSA on prolonged peaks, we found little to no shared subsequences. In most alignments, a peak sequence or a prolonged peak does not even intersect more than two thirds of other peak sequences. We were not able to retrieve distinct consensus motifs, not even local ones.

**GC content and dinucleotide frequencies show a differently distributed patch around the binding site.** First, we attempted to assess the size of the altered span. To do that, we used sequences spanning 10000 bp centered in the peak center (prolonged peaks). Using the cluster-based permutation test, we found that in general, GC content is increased around binding site (when compared to the edges of the prolonged peaks), spanning altogether 2000–4000 bp (see Tables 2 and 3). The only exception we were able to record

were experiments targeting NANOG in GM233338 (where GC content around the binding site creates a "W"-shaped curve, being strongly decreased 500 bp to 200 bp (both upstream and downstream) and "bouncing back" ±100 bp from the peak center) and experiments targeting NFRKB in MCF-7 (with a decrease of GC content ±300 bp from the peak center). In shorter span, the general rule of GC content enrichment around the binding site was observed (27). From our observations, it appears that 1000–2000 bp (depending on the target TF) upstream of the TF binding site, the frequency of GC base pairs starts increasing. Downstream of the binding site, GC content appears to drop symmetrically (see Figure 2, right column).

It is worth noting that there are only minor differences between the patches found in experiments on different cell lines, except for NFRKB (see top left plot in Figure 2). For this TF, there are immediately apparent differences: positions of increases and decreases directly contradict each other on a 1000 bp–long span. In experiments targeting NANOG, FOXK2, and IRF1 (for dinucleotide frequencies), minor differences between cell lines can be found as well. The contrast is less obvious. In some cases, one cell line displays a "normal" level (without excess) and the other shows an elevated presence. In others, we see a short patch (≈200 bp) with decrease in frequency in some and a longer patch (1000 bp and more) increase in frequency in others. The GC content frequencies can be seen in Supplementary Section 1.

While GC content around the binding site appeared symmetrical in the studied, it was not the case for dinucleotide frequency. The differences were most prominent in experiments targeting MYC, FOXK2, p53, NFRKB and IRF1. In particular, MYC binding sites are linked with an increase in AA, CA and AC dinucleotide frequencies, ending at -3000 – -1500 bp, followed by a decrease starting about -1000 bp (AA, CA) or about -300 bp (AC). This is mirrored downstream: there is a decrease in TT, TG and GT frequency starting far upstream of the binding site and ending 1000 bp (for TT and TG, 300 bp for GT) downstream of the binding site. Frequency of the GT dinucleotide starts rising 1000 bp from the binding site, frequency of TT and TG as well, 2000 bp from the binding site. The patches are depicted in Figure 2.

Complete observed dinucleotide frequencies for each of the probed TF can be found in the Supplementary Section 1. This trend (patches with increased AA, CA and AC frequency before the binding site and decreased downstreamed of it, combined with upstream decrease and downstream increase of their reverse complement TT, TG and GT) was observed for FOXK2, p53, IRF1 and NFRKB in certain cell lines (not MCF7). MYC and p53 are involved in the cell cycle regulation (58, 59), while proteins from IRF family including the IRF1 has a role during the immune cell differentiation (60). We find it interesting to note that these TFs with very important roles for the cell fate share similar pattern in dinucleotide frequencies around the binding site. The AA/CA/AC frequencies mirrored by TT/TG/GT frequencies downstream appear to create a funnel around the binding site (33).

**DNA shape features around the peak center.** DNA shape has been shown to affect the TF binding in several studies (2, 8, 10). We attempted to assess its role in the context of the prolonged peaks. To that end, we have predicted inter–base pair and intra–base pair DNA shape features using the deepDNAshape (52) predictor, proccuring predictions for all features this tool offers us. We are sampling every tenth base pair.

As with the (di)nucleotide frequencies, the average of several shape feature value clearly marks the assumed binding site (see Figure 3). In Table 4, we report the occurence and size for helical and propeller twist, minor groove width, opening, roll, stagger and stretch. Other DNA shape features did not exhibit changing signals.

The direction of change (increase or decrease in value) often agree between different TFs and cell lines (exception being NFRKB targetting MCF-7). Generally, a binding site is accompanied with a patch that displays decrease in propeller and hellical twist, increase in opening, roll or stagger as well as a wider minor groove and decreased values of stretch. Such a site would be linked with increased flexibility (consistent with the results of Yella et al.) and possible unwinding and bending (hellical twist, opening). The values of other DNA shape features might suggest local assymetries and altered base stacking, however, the values presented here come are based on predictions, which might introduce a degree of uncertainty to our findings. The complete overview of DNA shape feature increases and decreases are included in Supplementary Section 2.

As detailed in Table 4, there are differences in the size of the affected patch. However, for most of the TF–feature combination listed above, the altered patch can be several hundred base pairs long (even more than a thousand). Nevertheless, the magnitude of feature change is small (see Figure 3).

Speaking of the DNA structural properties related to transcription, several works have shown some involvement of Z-DNA conformation and gene regulation (61–63). Similarly, formation of G-quadruplexes was observed to have a role in transcription regulation (64, 65). On top of that, both of these structural features could explain some of the observed increases and decreases in (di)nucleotide frequencies described above. Z-DNA are more likely to form in sequences rich in CG dinucleotide (66) (CG dinucleotide were often increased around the binding site, see Supplementary Section 1). G-quadruplexes are, as the name suggests, linked with a specific guanine-rich motif (65).

Therefore, we incorporate information on the presence of both Z-DNA and G-quadruplexes in the prolonged peaks to complement the available predictions on DNA shape. The information is taken from the non-B DB database (51).

Some transcription factor binding sites and their vicinity display increased presence of Z-DNA motifs but not all of them (see Figure 3, bottom). Restricting ourselves to the top 25 % of peaks in each experiment, we observed this phenomenon for FOXK2, MYC and p53 (Mann-Whitney U test, significance level of 5 %, one-sided alternative hypothesis: are the fractions of windows with a Z-DNA motif higher at the bind-

ing site (± 100 bp) than on the edge of the prolonged peak?). However, even for these TFs, the change is too small to fully explain the nucleotide frequencies reported above: on average, about ≈ 1% of windows on the edge of the prolonged peak contain Z-DNA motif. At the binding site, it is 2 % (p53) or 6 % (MYC, FOXK2).

For G-quadruplex motifs, we found that CTCF, FOXK2, MEF2A, MYC and p53 are more likely to have a G-quadruplex at the binding site than at the edge of the binding site (Mann-Whitney U test, significance level of 5 %, one-sided alternative hypothesis: are the fractions of windows with a G-quadruplex motif higher at the binding site (± 100 bp; see Figure 3) than on the edge of the prolonged peak in each of the cell lines?). Again, the magnitude of the increase is fairly small: from ≈1.5 % to 3–6 % on average.

**Transcription factor affinity around the peak center.** To gain insight into possible cooperation of other TFs with the target TF in the given cell line, we scored the prolonged peaks with affinity models of more than 400 TFs from the HOCOMOCO database (49). For each window and TF combination, we work with the maximum affinity (or its z-score) per window. The reported results were acquired by processing the top 25 % peaks in each experiment.

In the top plots of Figure 4, we report the number of TFs from the non-redundant HOCOMOCO set for which we observed increased or decreased affinity (when compared to the general gene-upstream background). For most of the probed experiments, about 100–150 TFs showed increased affinity to the vicinity of the binding site, often matching the size of the area with increased GC content (± 1500 bp). MEF2A and Nanog constitute exceptions; the affected area in these cases spans 'only' ≈ ± 300 bp and affected ≈ 60 TFs from the non-redundant dataset. When we compare the average affinity z-score in the affected patch and the average affinity z-score in the unaffected patches, we observed that the differences rarely exceed 0.25 z-score points. As can be seen on the top right plot of the Figure 4, tens of TFs (20–80 for increased affinity, 40–120 for decreased affinity, not counting MEF2A and Nanog) differ for more than 0.2 z-score points.

It is interesting to note that between cell lines, the number of affected TFs (either way) does not differ (for CTCF and MYC, see Figure 4, middle plot). The composition of the affected TFs, on the other hand, exhibits differences, mainly when a higher difference threshold is applied (see Table 5 and Figure 4). In particular, for MYC, the intersection over union score between transcription factors with increased affinity to the prolonged peak for at least 0.2 z-score points is 0.52 — indicating that approximately half of the unique elements across all the sets are shared. The case is similar for the TFs with decreased affinity for at least 0.2 z-score points. When calculating the same score on cell line pairs (see Figure 4, bottom plots), the lowest achieved IoU scores were 0.85 for TFs with decreased affinity and 0.78 for TFs with increased affinity, indicating (together with the number of affected TFs) that differences between the affected TF sets may contribute to cell line specificity.

## Discussion

Our study provides compelling evidence that TF binding sites exhibit not only local sequence motifs but also broader contextual signals. Figuratively speaking, the binding site resembles a distortion point, influencing the distribution of its flanking DNA over distances spanning hundreds of base pairs.

We report finding a large non-consensus sequence signal in almost every experiment in our dataset. In case of GC content, similar finding have been reported previously (27), indicating that this phenomenon is widespread and consistent. We discovered increased GC content around the binding sites of TFs with various functions, ranging from the pioneer transcription factor CTCF (67) to MYC or p53 with complex role in cell cycle regulation (58, 59).

Beyond GC content, we report a directional pattern in dinucleotide frequencies. This distortion is most notable in the cases of MYC, p53, FOXK2, NFRKB and IRF1, where the mirrorring frequency patterns (AA/CA/AC upstream, TT/TG/GT downstream) suggest the presence of an electrostatic funnel (34) guiding the TF towards the binding site.

We also observed that DNA shape features (such as propeller twist, minor groove width, and roll) show consistent alterations in the vicinity of the binding sites. Although these changes are modest in magnitude, they span several hundred to over a thousand base pairs, indicating that the structural flexibility of the DNA around the binding site is an inherent feature of these genomic regions. The flexibility of the flanking sequence has been shown to affect TF affinity to its binding site in several studies (11–13). In our dataset, the largest shape distortions were observed near MYC binding sites. However, these shape features are derived from sequence only. On the level of hundreds of base-pairs, several other factors as nucleosome packaging comes into play when it comes to the *in vivo* shape of the DNA helix.

Additionally, our analysis of TF affinities within these regions reveals that tens of TFs experience altered affinity in the vicinity of a binding site—either enhanced or diminished. On one hand, this could be partly attributed to the altered nucleotide composition (although when the GC-aware normalization set was used, the same results were achieved). On the other hand, there were differences in the set composition of TFs with highly increased/highly decreased affinity between different cell lines for MYC, while the (di)nucleotide composition was consistent. These observations suggest that the broader genomic context and potential cooperative interactions shape TF binding landscapes in a cell-type-specific manner.

Collectively, our findings support a model in which TF binding site recognition is a multi-layered process. This process involves not only direct sequence and shape readout at the motif level but also a broader, kilobase-scale environment that modulates the search and recognition dynamics of TFs. We believe the observed sequence distortions affect the binding site kinetics, as the search for the binding sites is influenced by subtle changes of the DNA shape and electrostatic potential (29, 31, 35, 37). Likely, these direct sequence-

driven effects combine with DNA secondary structures (*e.g.*, Z-DNA and G-quadruplexes) and influences of chromatin architecture and nucleosome positioning into a nuanced binding landscape.

## Conclusion

In summary, our genome-wide analysis reveals that TF binding sites are consistently flanked by an extended symmetrical region (1000–1500 bp) of elevated GC content, along with directional dinucleotide biases in certain TFs (such as MYC). We believe that these long non-consensus flanking sequences act as "beacons" to guide TFs toward their target binding sites. The presence of DNA shape alterations in these regions further supports a model where increased local flexibility and malleability of the DNA helix contribute to TF targeting and specificity.

Moreover, our observation of TF affinity profiles – in particular, the recorded differences between cell lines despite similar underlying sequence features (MYC) – suggests an interplay of several TFs, where the preferred and undesired binding partners are selected on the sequence level.

Altogether, our findings underscore the importance of considering not only the core motifs but also the broader sequence context when studying specific recognition of the DNA by a TF. We believe that our findings may contribute to a more nuanced understanding of gene regulation in different cellular contexts and improvement in TF–binding predictions.

## Data availability

Source codes, complete list of used ChIP-seq experiments and final description of excessive patches for each experiment is stored in Zenodo archive under DOI 10.5281/zenodo.15472407.

## Acknowledgements

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic This work was supported by the ELIXIR CZ Research Infrastructure under Grant (ID LM2023055, MEYS CR); and Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences under Grant (RVO: 61388963).

## Bibliography

1. Seungsoo Kim and Joanna Wysocka. Deciphering the multi-scale, quantitative cis-regulatory code. *Molecular cell*, 83(3):373–392, 2023.
2. Matthew Slattery, Tianyin Zhou, Lin Yang, Ana Carolina Dantas Machado, Raluca Gordân, and Remo Rohs. Absence of a simple code: how transcription factors read the genome. *Trends in biochemical sciences*, 39(9):381–399, 2014.
3. Alexander J Stewart, Sridhar Hannenhalli, and Joshua B Plotkin. Why transcription factor binding sites are ten nucleotides long. *Genetics*, 192(3):973–985, 2012.
4. Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, et al. Dna-binding specificities of human transcription factors. *Cell*, 152(1):327–339, 2013.
5. Zhiyuan Xie, Ilya Sokolov, Maria Osmala, Xue Yue, Grace Bower, J Patrick Pett, Yinan Chen, Kai Wang, Ayse Derya Cavga, Alexander Popov, et al. Dna-guided transcription factor interactions extend human gene regulatory code. *Nature*, pages 1–10, 2025.
6. Rosa Martinez-Corral, Dhana Friedrich, Robert Frömel, Lars Velten, Jeremy Gunawardena, and Angela H DePace. Emergence of activation or repression in transcriptional control under a fixed molecular context. *bioRxiv*, 2024.
7. Lakshmi Mahendrawada, Linda Warfield, Rafal Donczew, and Steven Hahn. Low overlap of transcription factor dna binding and regulatory targets. *Nature*, pages 1–9, 2025.
8. Anthony Mathelier, Beibei Xin, Tsu-Pei Chiu, Lin Yang, Remo Rohs, and Wyeth W Wasserman. Dna shape features improve transcription factor binding site predictions in vivo. *Cell systems*, 3(3):278–286, 2016.
9. Md Abul Hassan Samee, Benoit G Bruneau, and Katherine S Pollard. A de novo shape motif discovery algorithm reveals preferences of transcription factors for dna shape beyond sequence motifs. *Cell systems*, 8(1):27–42, 2019.
10. Nanjun Chen, Jixiang Yu, Zhe Liu, Lingkuan Meng, Xiangtao Li, and Ka-Chun Wong. Discovering dna shape motifs with multiple dna shape features: generalization, methods, and validation. *Nucleic Acids Research*, 52(8):4137–4150, 2024.
11. Debostuti Ghoshdastidar and Manju Bansal. Flexibility of flanking dna is a key determinant of transcription factor affinity for the core motif. *Biophysical Journal*, 121(20):3987–4000, 2022.
12. Venkata Rajesh Yella, Devesh Bhimsaria, Debostuti Ghoshdastidar, José A Rodríguez-Martínez, Aseem Z Ansari, and Manju Bansal. Flexibility and structure of flanking dna impact transcription factor affinity for its core motif. *Nucleic acids research*, 46(22):11883–11897, 2018.
13. Tsu-Pei Chiu, Jinsen Li, Yibei Jiang, and Remo Rohs. It is in the flanks: Conformational flexibility of transcription factor binding sites. *Biophysical Journal*, 121(20):3765–3767, 2022.
14. N. P. Pavletich and C. O. Pabo. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science (New York, N.Y.)*, 252(5007):809–817, May 1991. ISSN 0036-8075. doi: 10.1126/science.2028256.
15. Joseph L. Kim and Stephen K. Burley. 1.9 Å resolution refined structure of TBP recognizing the minor groove of TATAAAAG. *Nature Structural Biology*, 1(9):638–653, September 1994. ISSN 1545-9985. doi: 10.1038/nsb0994-638. Publisher: Nature Publishing Group.
16. Remo Rohs, Xiangshu Jin, Sean M West, Rohit Joshi, Barry Honig, and Richard S Mann. Origins of specificity in protein-DNA recognition. *Annual review of biochemistry*, 79:233–269, January 2010. ISSN 1545-4509. doi: 10.1146/annurev-biochem-060408-091030.
17. Tianyin Zhou, Ning Shen, Lin Yang, Namiko Abe, John Horton, Richard S. Mann, Harmen J. Bussemaker, Raluca Gordân, and Remo Rohs. Quantitative modeling of transcription factor binding specificities using DNA shape. *Proceedings of the National Academy of Sciences*, 112(15):4654–4659, April 2015. doi: 10.1073/pnas.1422023112. Publisher: Proceedings of the National Academy of Sciences.
18. Siguo Wang, Qinhu Zhang, Zhen Shen, Ying He, Zhen-Heng Chen, Jianqiang Li, and De-Shuang Huang. Predicting transcription factor binding sites using DNA shape features based on shared hybrid deep learning architecture. *Molecular Therapy - Nucleic Acids*, 24:154–163, June 2021. ISSN 2162-2531. doi: 10.1016/j.omtn.2021.02.014.
19. Stefano Stella, Duilio Cascio, and Reid C. Johnson. The shape of the DNA minor groove directs binding by the DNA-bending protein Fis. *Genes & Development*, 24(8):814–826, April 2010. ISSN 1549-5477. doi: 10.1101/gad.1900610.
20. Stephen P. Hancock, Tahereh Ghane, Duilio Cascio, Remo Rohs, Rosa Di Felice, and Reid C. Johnson. Control of DNA minor groove width and Fis protein binding by the purine 2-amino group. *Nucleic Acids Research*, 41(13):6750–6760, July 2013. ISSN 1362-4962. doi: 10.1093/nar/gkt357.
21. Yongheng Chen, Xiaojun Zhang, Ana Carolina Dantas Machado, Yuan Ding, Zhuchu Chen, Peter Z. Qin, Remo Rohs, and Lin Chen. Structure of p53 binding to the BAX response element reveals DNA unwinding and compression to accommodate base-pair insertion. *Nucleic Acids Research*, 41(17):8368–8376, September 2013. ISSN 1362-4962. doi: 10.1093/nar/gkt584.
22. Yongheng Chen, Darren L. Bates, Raja Dey, Po-Han Chen, Ana Carolina Dantas Machado, Ite A. Laird-Offringa, Remo Rohs, and Lin Chen. DNA binding by GATA transcription factor suggests mechanisms of DNA looping and long-range gene regulation. *Cell Reports*, 2(5):1197–1206, November 2012. ISSN 2211-1247. doi: 10.1016/j.celrep.2012.10.012.
23. Xiaojun Zhang, Ana Carolina Dantas Machado, Yuan Ding, Yongheng Chen, Yan Lu, Yankun Duan, Kenneth W. Tham, Lin Chen, Remo Rohs, and Peter Z. Qin. Conformations of p53 response elements in solution deduced using site-directed spin labeling and Monte Carlo sampling. *Nucleic Acids Research*, 42(4):2789–2797, February 2014. ISSN 1362-4962. doi: 10.1093/nar/gkt1219.
24. Pengju Ding, Yifei Wang, Xinyu Zhang, Xin Gao, Guozhu Liu, and Bin Yu. DeepSTF: predicting transcription factor binding sites by interpretable deep neural networks combining sequence and shape. *Briefings in Bioinformatics*, 24(4):bbad231, July 2023. ISSN 1477-4054. doi: 10.1093/bib/bbad231.
25. Jichen Yang and Stephen A. Ramsey. A DNA shape-based regulatory score improves position-weight matrix-based recognition of transcription factor binding sites. *Bioinformatics (Oxford, England)*, 31(21):3445–3450, November 2015. ISSN 1367-4811. doi: 10.1093/bioinformatics/btv391.
26. Raluca Gordân, Ning Shen, Iris Dror, Tianyin Zhou, John Horton, Remo Rohs, and Martha L. Bulyk. Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Reports*, 3(4):1093–1104, April 2013. ISSN 2211-1247. doi: 10.1016/j.celrep.2013.03.014.
27. Iris Dror, Tamar Golan, Carmit Levy, Remo Rohs, and Yael Mandel-Gutfreund. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome research*, 25(9):1268–1280, 2015.
28. Lin Yang, Yaron Orenstein, Arttu Jolma, Yimeng Yin, Jussi Taipale, Ron Shamir, and Remo Rohs. Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Molecular Systems Biology*, 13(2):910, February 2017. ISSN 1744-4292. doi: 10.15252/msb.20167238.
29. Michael Slutsky and Leonid A. Mirny. Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential. *Biophysical Journal*, 87(6):4021–4035, December 2004. ISSN 0006-3495. doi: 10.1529/biophysj.104.050765.
30. Stephen E. Halford and John F. Marko. How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Research*, 32(10):3040–3052, May 2004. ISSN 0305-1048. doi: 10.1093/nar/gkh624.
31. R. B. Winter, O. G. Berg, and P. H. von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The Escherichia coli lac repressor–operator interaction:

kinetic measurements and conclusions. *Biochemistry*, 20(24):6961–6977, November 1981. ISSN 0006-2960. doi: 10.1021/bi00527a030.

32. Ariel Afek and David B. Lukatsky. Genome-wide organization of eukaryotic preinitiation complex is influenced by nonconsensus protein-DNA binding. *Biophysical Journal*, 104(5): 1107–1115, March 2013. ISSN 1542-0086. doi: 10.1016/j.bpj.2013.01.038.
33. Ariel Afek and David B. Lukatsky. Positive and negative design for nonconsensus protein-DNA binding affinity in the vicinity of functional binding sites. *Biophysical Journal*, 105(7): 1653–1660, October 2013. ISSN 1542-0086. doi: 10.1016/j.bpj.2013.08.033.
34. Itamar Sela and David B. Lukatsky. DNA sequence correlations shape nonspecific transcription factor-DNA binding affinity. *Biophysical Journal*, 101(1):160–166, July 2011. ISSN 1542-0086. doi: 10.1016/j.bpj.2011.04.037.
35. Itai Leven and Yaakov Levy. Quantifying the two-state facilitated diffusion model of protein–DNA interactions. *Nucleic Acids Research*, 47(11):5530–5538, June 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz308.
36. Carmen Al Masri, Biao Wan, and Jin Yu. Nonspecific vs. specific DNA binding free energetics of a transcription factor domain protein. *Biophysical Journal*, 122(22):4476–4487, November 2023. ISSN 0006-3495, 1542-0086. doi: 10.1016/j.bpj.2023.10.025. Publisher: Elsevier.
37. Maximilian Bauer and Ralf Metzler. Generalized Facilitated Diffusion Model for DNA-Binding Proteins with Search and Recognition States. *Biophysical Journal*, 102(10):2321–2330, May 2012. ISSN 0006-3495, 1542-0086. doi: 10.1016/j.bpj.2012.04.008. Publisher: Elsevier.
38. Jérôme Cartailler and Jürgen Reingruber. Facilitated diffusion framework for transcription factor search with conformational changes. *Physical Biology*, 12(4):046012, July 2015. ISSN 1478-3975. doi: 10.1088/1478-3975/12/4/046012. arXiv:1507.02452 [q-bio].
39. Hiroaki Matsuda, Gregory Garbès Putzel, Vadim Backman, and Igal Szleifer. Macromolecular Crowding as a Regulator of Gene Transcription. *Biophysical Journal*, 106(8):1801–1810, April 2014. ISSN 0006-3495, 1542-0086. doi: 10.1016/j.bpj.2014.02.019. Publisher: Elsevier.
40. Samantha Kent, Kyle Brown, Chou-hsun Yang, Njood Alsaihati, Christina Tian, Haobin Wang, and Xiaojun Ren. Phase-Separated Transcriptional Condensates Accelerate Target-Search Process Revealed by Live-Cell Single-Molecule Imaging. *Cell Reports*, 33(2): 108248, October 2020. ISSN 2211-1247. doi: 10.1016/j.celrep.2020.108248.
41. Christian P. Bacher, Michaela Reichenzeller, Chaitanya Athale, Harald Herrmann, and Roland Eils. 4-D single particle tracking of synthetic and proteinaceous microspheres reveals preferential movement of nuclear particles along chromatin – poor tracks. *BMC Cell Biology*, 5(1):45, November 2004. ISSN 1471-2121. doi: 10.1186/1471-2121-5-45.
42. Kaustubh Wagh, Diana A. Stavreva, and Gordon L. Hager. Transcription dynamics and genome organization in the mammalian nucleus: Recent advances. *Molecular Cell*, 85(2): 208–224, January 2025. ISSN 1097-2765. doi: 10.1016/j.molcel.2024.09.022. Publisher: Elsevier.
43. Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, 326(5950):289–293, October 2009. ISSN 1095-9203. doi: 10.1126/science.1181369.
44. Yunhai Luo, Benjamin C Hitz, Idan Gabdank, Jason A Hilton, Meenakshi S Kagda, Bonita Lam, Zachary Myers, Paul Sud, Jennifer Jou, Khine Lin, et al. New developments on the encyclopedia of dna elements (encode) data portal. *Nucleic acids research*, 48(D1):D882–D889, 2020.
45. Hatice S Kaya-Okur, Steven J Wu, Christine A Codomo, Erica S Pledger, Terri D Bryson, Jorja G Henikoff, Kami Ahmad, and Steven Henikoff. Cut&tag for efficient epigenomic profiling of small samples and single cells. *Nature communications*, 10(1):1930, 2019.
46. Jianxing Feng, Tao Liu, Bo Qin, Yong Zhang, and Xiaole Shirley Liu. Identifying chip-seq enrichment using macs. *Nature protocols*, 7(9):1728–1740, 2012.
47. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
48. Marius Gheorghe, Geir Kjetil Sandve, Aziz Khan, Jeanne Chèneby, Benoit Ballester, and Anthony Mathelier. A map of direct tf–dna interactions in the human genome. *Nucleic Acids Research*, 47(4):e21–e21, 12 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1210.
49. Ivan V Kulakovskiy, Ilya E Vorontsov, Ivan S Yevshin, Ruslan N Sharipov, Alla D Fedorova, Eugene I Rumynskiy, Yulia A Medvedeva, Arturo Magana-Mora, Vladimir B Bajic, Dmitry A Papatsenko, Fedor A Kolpakov, and Vsevolod J Makeev. Hocomoco: towards a complete collection of transcription factor binding models for human and mouse via large-scale chip-seq analysis. *Nucleic Acids Research*, 46(D1):D252–D259, 11 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx1106.
50. Gerardo Perez, Galt P Barber, Anna Benet-Pages, Jonathan Casper, Hiram Clawson, Mark Diekhans, Clay Fischer, Jairo Navarro Gonzalez, Angie S Hinrichs, Christopher M Lee, et al. The ucsc genome browser database: 2025 update. *Nucleic Acids Research*, 53(D1): D1243–D1249, 2025.
51. Regina Z Cer, Duncan E Donohue, Uma S Mudunuri, Nuri A Temiz, Michael A Loss, Nathan J Starner, Goran N Halusa, Natalia Volfovsky, Ming Yi, Brian T Luke, et al. Non-b db v2. 0: a database of predicted non-b dna-forming motifs and its associated tools. *Nucleic acids research*, 41(D1):D94–D100, 2012.
52. Jinsen Li, Tsu-Pei Chiu, and Remo Rohs. Predicting dna structure using a deep learning method. *Nature communications*, 15(1):1243, 2024.
53. Jinsen Li, Jared M Sagendorf, Tsu-Pei Chiu, Marco Pasi, Alberto Perez, and Remo Rohs. Expanding the repertoire of dna shape features for genome-scale studies of transcription factor binding. *Nucleic acids research*, 45(22):12877–12887, 2017.
54. Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular systems biology*, 7(1):539, 2011.
55. Daniel R Zerbino, Steven P Wilder, Nathan Johnson, Thomas Juettemann, and Paul R Flicek. The ensembl regulatory build. *Genome biology*, 16:1–8, 2015.
56. Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7 (267):1–13, 2013. doi: 10.3389/fnins.2013.00267.
57. Rebecca Worsley Hunt, Anthony Mathelier, Luis Del Peso, and Wyeth W Wasserman. Improving analysis of transcription factor binding sites within chip-seq data based on topological motif enrichment. *BMC genomics*, 15:1–19, 2014.
58. Gabriel Bretones, M Dolores Delgado, and Javier León. Myc and cell cycle control. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1849(5):506–516, 2015.
59. Luciana E Giono and James J Manfredi. The p53 tumor suppressor participates in multiple cell cycle checkpoints. *Journal of cellular physiology*, 209(1):13–20, 2006.
60. Sophia Giang and Antonio La Cava. Irf1 and batf: key drivers of type 1 regulatory t-cell differentiation. *Cellular & molecular immunology*, 14(8):652–654, 2017.
61. Adriana CernÃ, Ana Cuadrado, N Jouve, SM DÃaz de la Espina, C De la Torre, et al. Z-dna, a new in situ marker for transcription. *European Journal of Histochemistry*, 48(1): 49–56, 2004.
62. So-I Shin, Seokjin Ham, Jihwan Park, Seong Hye Seo, Chae Hyun Lim, Hyeongrin Jeon, Jounghyun Huh, and Tae-Young Roh. Z-dna-forming sites identified by chip-seq are associated with actively transcribed regions in the human genome. *DNA Research*, 23(5): 477–486, 2016.
63. Nazar Beknazarov, Dmitry Konovalov, Alan Herbert, and Maria Poptsova. Z-dna formation in promoters conserved between human and mouse are associated with increased transcription reinitiation rates. *Scientific Reports*, 14(1):17786, 2024.
64. Jenna Robinson, Federica Raguseo, Sabrina Pia Nuccio, Denise Liano, and Marco Di Antonio. Dna g-quadruplex structures: more than simple roadblocks to transcription? *Nucleic Acids Research*, 49(15):8419–8431, 07 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab609.
65. Julian L. Huppert and Shankar Balasubramanian. G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Research*, 35(6):2105–2105, 02 2007. ISSN 0305-1048. doi: 10.1093/nar/gkm109.
66. Daniel Svozil, Jan Kalina, Marek Omelka, and Bohdan Schneider. Dna conformations and their sequence preferences. *Nucleic acids research*, 36(11):3690–3706, 2008.
67. Amandine Barral and Kenneth S Zaret. Pioneer factors: roles and their regulation in development. *Trends in Genetics*, 40(2):134–148, 2024.

## Figures and tables

| TF | Biosample | ENCODE accessions |
|---|---|---|
| CTCF | A549 | ENCFF966KGI |
| CTCF | GM12878 | ENCFF536RGD, ENCFF017XLW |
| CTCF | H1 | ENCFF692RPA |
| CTCF | HepG2 | ENCFF252PLM, ENCFF123DAC |
| CTCF | K562 | ENCFF769AUF |
| CTCF | MCF-7 | ENCFF962IZR, ENCFF820BVW |
| CTCF | WTC11 | ENCFF503DSN |
| FOXK2 | GM12878 | ENCFF855EOH, ENCFF578XIP |
| FOXK2 | HepG2 | ENCFF008HBF, ENCFF997KQI |
| FOXK2 | K562 | ENCFF022HGL, ENCFF154SRU |
| FOXK2 | MCF-7 | ENCFF583QAD, ENCFF356XXS |
| IRF1 | HepG2 | ENCFF174QAR |
| IRF1 | K562 | ENCFF019LOH, ENCFF251ZSJ |
| IRF1 | WTC11 | ENCFF910ZUK |
| MEF2A | GM12878 | ENCFF131JIX, ENCFF775DWU |
| MEF2A | HepG2 | ENCFF987BRO |
| MEF2A | K562 | ENCFF288TSN, ENCFF190IHG |
| MYC | A549 | ENCFF674RQX, ENCFF356PWE |
| MYC | GM12878 | ENCFF765CKK |
| MYC | H1 | ENCFF372RQZ, ENCFF700CXD |
| MYC | HepG2 | ENCFF800JFG |
| MYC | K562 | ENCFF043WTJ |
| MYC | MCF-7 | ENCFF149BIS |
| NANOG | GM23338 | ENCFF897LBK, ENCFF463AFM |
| NANOG | H1 | ENCFF655XZF, ENCFF722JFZ |
| NFRKB | HepG2 | ENCFF252JYE, ENCFF748JYG |
| NFRKB | K562 | ENCFF594KPG, ENCFF048HYG |
| NFRKB | MCF-7 | ENCFF360OXK, ENCFF768VQG |
| RUNX1 | K562 | ENCFF264QLP, ENCFF015ADG |
| SPI1 | GM12878 | ENCFF881ARS, ENCFF198SPL |
| SPI1 | K562 | ENCFF185DGL, ENCFF456BUE |
| TP53 | A549 | ENCFF488WQN |
| TP53 | HepG2 | ENCFF701DTE |
| TP53 | WTC11 | ENCFF280RDQ |

**Table 1.** ENCODE identifiers of the used TF ChIP-seq experiments.

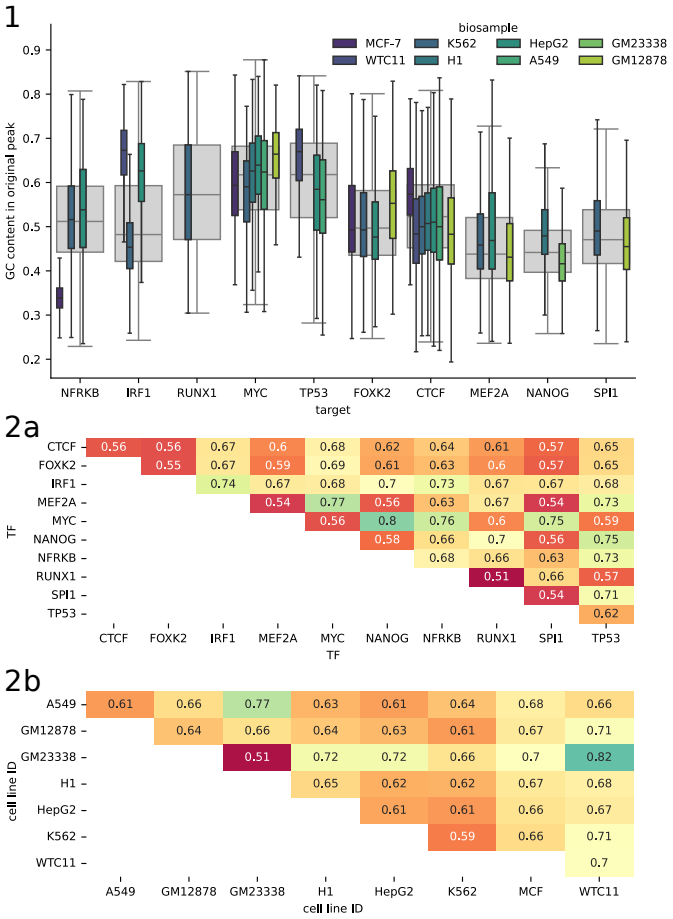



**Fig. 1. (1)** distribution of GC content in peaks identified by ChIP-seq experiment, grouped by cell line and target TF. The grey boxplot describes the distribution of all the peaks captured in a TF ChIP-seq experiment targeting the corresponding TF; **(2a)** average accuracy of agglomerative clustering based on trivial GC content difference, grouped by transcription factor; **(2b)** average accuracy of agglomerative clustering based on trivial GC content difference, grouped by cell line

| feature | CTCF | SPI1 | MEF2A | FOXK2 | NFRKB | TP53 | IRF1 | MYC | NANOG | RUNX1 |
|---|---|---|---|---|---|---|---|---|---|---|
| GC content | | | | | 637.5 (17.7) | | | | 787.5 (88.4) | |
| AA freq. | 2297.7 (601.5) | 1962.5 (727.6) | 670.0 (622.6) | 2565.6 (572.7) | 3218.8 (1037.7) | 3783.3 (1634.1) | 2531.2 (552.4) | 4506.2 (1120.3) | 300.0 (0.0) | 1800.0 (247.5) |
| AC freq. | 330.6 (239.4) | 537.5 (571.7) | 275.0 (325.0) | 1781.2 (1056.6) | 3150.0 (1726.7) | 3225.0 (1677.2) | 1968.8 (1698.0) | 3893.8 (1379.0) | 368.8 (100.8) | 375.0 (35.4) |
| AG freq. | 275.0 | | 100.0 | 335.7 (190.9) | 450.0 (234.0) | 783.3 (431.1) | 950.0 | 1053.1 (422.5) | 162.5 (17.7) | 75.0 (0.0) |
| AT freq. | 3413.6 (928.3) | 3143.8 (682.0) | 2170.0 (697.9) | 2940.6 (490.1) | 3037.5 (609.1) | 3608.3 (267.3) | 3425.0 (388.4) | 3562.5 (512.7) | 1031.2 (785.6) | 2500.0 (70.7) |
| CA freq. | 863.6 (449.4) | 1200.0 (1227.6) | 940.0 (675.8) | 3346.9 (1229.4) | 3241.7 (2426.3) | 4558.3 (1112.5) | 3625.0 (1558.0) | 5015.6 (1118.7) | 381.2 (154.6) | 1400.0 (141.4) |
| CC freq. | | | 150.0 | | 500.0 (0.0) | | | | 200.0 (0.0) | |
| CG freq. | | | | 225.0 | 575.0 (70.7) | | | | 262.5 (88.4) | |
| CT freq. | 75.0 | | 75.0 | 416.7 (315.3) | 308.3 (109.2) | 750.0 (565.1) | 575.0 (495.0) | 1078.1 (509.5) | | 100.0 (0.0) |
| GA freq. | 50.0 (16.7) | | 100.0 | 200.0 | 158.3 (80.4) | 350.0 | 300.0 | 181.2 (65.1) | | 75.0 (0.0) |
| GC freq. | | | | 125.0 | 587.5 (123.7) | 200.0 | | | 275.0 (106.1) | |
| GG freq. | | | 75.0 | | 462.5 (53.0) | | | 650.0 | 337.5 (53.0) | |
| GT freq. | 361.1 (341.7) | 583.3 (335.7) | 475.0 | 1696.4 (1163.4) | 3656.2 (508.4) | 3283.3 (963.8) | 1883.3 (1915.8) | 4143.8 (1144.8) | 356.2 (31.5) | 462.5 (53.0) |
| TA freq. | 2838.6 (813.2) | 2650.0 (523.2) | 1360.0 (800.9) | 2578.1 (316.9) | 2681.2 (577.8) | 3125.0 (525.0) | 2481.2 (827.2) | 2975.0 (404.9) | 618.8 (269.5) | 2100.0 (70.7) |
| TC freq. | 55.6 (16.7) | | 87.5 (17.7) | 200.0 | 231.2 (80.0) | 237.5 (123.7) | 275.0 | 150.0 (25.0) | | |
| TG freq. | 1170.5 (679.4) | 1306.2 (911.1) | 893.8 (692.3) | 3600.0 (1417.0) | 3915.0 (2373.9) | 4825.0 (883.5) | 2781.2 (1858.6) | 5321.9 (792.6) | 312.5 (136.2) | 1837.5 (477.3) |
| TT freq. | 2377.3 (587.2) | 2118.8 (1262.3) | 590.0 (479.8) | 2459.4 (758.2) | 3362.5 (955.8) | 3791.7 (748.5) | 2343.8 (1603.3) | 4093.8 (1176.5) | 350.0 (35.4) | 2212.5 (53.0) |

**Table 2.** Total size of area with significant frequency **decrease**. The values shown are the average over all experiments targetting a particular TF (standard deviation in parenthesis). The results shown were extimated from peaks above the FDR percentile threshold of 50 %.

| feature | CTCF | SPI1 | MEF2A | FOXK2 | NFRKB | TP53 | IRF1 | MYC | NANOG | RUNX1 |
|---|---|---|---|---|---|---|---|---|---|---|
| GC content | 4018.2 (968.8) | 3631.2 (920.2) | 2265.0 (670.0) | 3071.9 (279.5) | 3168.8 (747.0) | 3675.0 (606.2) | 3581.2 (260.9) | 3596.9 (360.7) | 1193.8 (1090.9) | 2750.0 (282.8) |
| AA freq. | 200.0 | 112.5 (53.0) | | 700.0 (807.5) | 787.5 (396.5) | 966.7 (469.3) | 687.5 (229.8) | 1789.3 (1189.4) | 625.0 (353.6) | |
| AC freq. | 712.5 (830.9) | 175.0 (106.1) | 183.3 (52.0) | 1303.1 (1250.3) | 2275.0 (732.3) | 2650.0 (303.1) | 1137.5 (1431.9) | 3271.9 (1119.7) | | 512.5 (17.7) |
| AG freq. | 334.1 (274.2) | 375.0 (144.3) | 166.7 (94.6) | 200.0 (212.1) | 162.5 (17.7) | | 612.5 (88.4) | | 318.8 (37.5) | |
| AT freq. | | | | | 562.5 (53.0) | | | | | |
| CA freq. | 68.2 (44.8) | 150.0 (35.4) | | 862.5 (970.9) | 1418.8 (423.0) | 1541.7 (450.2) | 812.5 (866.2) | 2146.9 (1058.1) | 131.2 (62.5) | 400.0 (35.4) |
| CC freq. | 3031.8 (609.1) | 2287.5 (501.0) | 1080.0 (894.7) | 2440.6 (428.2) | 2856.2 (878.5) | 3433.3 (267.3) | 3143.8 (559.2) | 3215.6 (374.4) | 506.2 (445.1) | 2087.5 (194.5) |
| CG freq. | 3768.2 (623.4) | 4162.5 (1191.2) | 2985.0 (887.2) | 4200.0 (858.5) | 4743.8 (205.5) | 4300.0 (525.0) | 4112.5 (163.9) | 4625.0 (279.3) | 1868.8 (1489.6) | 3450.0 (70.7) |
| CT freq. | 382.5 (273.1) | 450.0 (120.8) | 75.0 (0.0) | 87.5 (17.7) | 162.5 (17.7) | | 562.5 (123.7) | | 268.8 (65.7) | |
| GA freq. | 93.8 (23.9) | 337.5 (77.7) | 50.0 | 350.0 (308.7) | 643.8 (144.9) | | 341.7 (305.5) | 150.0 (106.1) | 175.0 (35.4) | 150.0 |
| GC freq. | 3152.3 (662.1) | 2693.8 (650.4) | 1555.0 (842.1) | 2615.6 (421.3) | 2737.5 (531.7) | 3058.3 (470.6) | 2918.8 (258.5) | 3131.2 (195.8) | 837.5 (679.6) | 2275.0 (0.0) |
| GG freq. | 2852.3 (343.8) | 2781.2 (657.4) | 1380.0 (703.9) | 2718.8 (478.2) | 2693.8 (457.1) | 3250.0 (591.1) | 2825.0 (370.8) | 3006.2 (318.1) | 631.2 (558.0) | 2187.5 (17.7) |
| GT freq. | 100.0 | 137.5 (17.7) | 216.7 (137.7) | 1209.4 (964.4) | 2637.5 (709.6) | 2558.3 (1572.1) | 1087.5 (1203.6) | 2815.6 (1316.2) | | 400.0 (141.4) |
| TA freq. | | | | | 575.0 (70.7) | | | | | |
| TC freq. | 210.0 (133.0) | 431.2 (173.7) | 50.0 | 433.3 (226.2) | 962.5 (366.6) | 275.0 (70.7) | 366.7 (144.3) | 362.5 (222.2) | 75.0 (0.0) | |
| TG freq. | 52.8 (19.5) | 125.0 | | 357.1 (292.5) | 1050.0 (523.2) | 1425.0 (1073.3) | 700.0 (388.9) | 1768.8 (1096.1) | 100.0 (0.0) | |
| TT freq. | | 100.0 | | 457.1 (268.4) | 666.7 (428.6) | 1125.0 (35.4) | 408.3 (407.2) | 1807.1 (1115.2) | 575.0 (0.0) | |

**Table 3.** Total size of area with significant frequency **increase**. The values shown are the average over all experiments targetting a particular TF (standard deviation in parenthesis). The results shown were extimated from peaks above the FDR percentile threshold of 50 %.

| tf | CTCF | FOXK2 | IRF1 | MEF2A | MYC | NANOG | NFRKB | RUNX1 | SPI1 | TP53 |
|---|---|---|---|---|---|---|---|---|---|---|
| HelT, increase | 0/10, - | 0/8, - | 0/4, - | 0/5, - | 0/8, - | 0/4, - | 0/6, - | 0/2, - | 1/4, 20.0 | 0/3, - |
| MGW, increase | 10/10, 110 bp | 8/8, 440 bp | 2/4, 825 bp | 3/5, 80 bp | 8/8, 945 bp | 4/4, 75 bp | 4/6, 820 bp | 2/2, 450 bp | 3/4, 20 bp | 3/3, 950 bp |
| Opening, increase | 10/10, 1350 bp | 8/8, 1870 bp | 4/4, 1365 bp | 5/5, 190 bp | 8/8, 2275 bp | 4/4, 190 bp | 4/6, 2105 bp | 2/2, 1535 bp | 4/4, 1130 bp | 3/3, 2230 bp |
| ProT, increase | 0/10, - | 0/8, - | 1/4, 20 bp | 1/5, 10 bp | 0/8, - | 1/4, 30 bp | 2/6, 25 bp | 0/2, - | 0/4, - | 0/3, - |
| Roll, increase | 10/10, 40 bp | 7/8, 120 bp | 3/4, 80 bp | 0/5, - | 8/8, 215 bp | 1/4, 10 bp | 3/6, 120 bp | 2/2, 50 bp | 0/4, - | 2/3, 300 bp |
| Stagger, increase | 10/10, 525 bp | 8/8, 1150 bp | 4/4, 760 bp | 5/5, 110 bp | 8/8, 1675 bp | 2/4, 215 bp | 4/6, 1415 bp | 2/2, 650 bp | 4/4, 485 bp | 3/3, 1410 bp |
| Stretch, increase | 0/10, - | 1/8, 20 bp | 0/4, - | 0/5, - | 0/8, - | 1/4, 20 bp | 1/6, 30 bp | 0/2, - | 0/4, - | 0/3, - |
| HelT, decrease | 10/10, 210 bp | 8/8, 665 bp | 4/4, 500 bp | 2/5, 25 bp | 8/8, 1305 bp | 3/4, 190 bp | 4/6, 1020 bp | 2/2, 580 bp | 3/4, 330 bp | 3/3, 1060 bp |
| MGW, decrease | 0/10, - | 0/8, - | 1/4, 30 bp | 0/5, - | 0/8, - | 2/4, 75 bp | 0/6, - | 0/2, - | 3/4, 10 bp | 0/3, - |
| Opening, decrease | 0/10, - | 0/8, - | 0/4, - | 0/5, - | 0/8, - | 1/4, 20 bp | 2/6, 60 bp | 0/2, - | 0/4, - | 0/3, - |
| ProT, decrease | 10/10, 895 bp | 8/8, 1480 bp | 4/4, 1030 bp | 5/5, 30 bp | 8/8, 1980 bp | 4/4, 170 bp | 4/6, 1785 bp | 2/2, 1130 bp | 4/4, 890 bp | 3/3, 1890 bp |
| Roll, decrease | 0/10, - | 2/8, 25 bp | 1/4, 20 bp | 0/5, - | 0/8, - | 2/4, 65 bp | 1/6, 10 bp | 0/2, - | 3/4, 10 bp | 0/3, - |
| Stagger, decrease | 0/10, - | 0/8, - | 0/4, - | 0/5, - | 0/8, - | 0/4, - | 2/6, 40 bp | 0/2, - | 0/4, - | 0/3, - |
| Stretch, decrease | 10/10, 665 bp | 8/8, 1410 bp | 4/4, 1035 bp | 5/5, 80 bp | 8/8, 1995 bp | 4/4, 200 bp | 4/6, 1780 bp | 2/2, 970 bp | 4/4, 790 bp | 3/3, 1900 bp |

**Table 4.** Presence and size of patches with altered DNA shape feature values. The indicated value shows a number of experiments (peak files) that exhibit such behaviour out of all experiments in out dataset, as well as the median size of the affected patch.

| TF | difference threshold | IoU: increased affinity | IoU: decreased affinity |
|---|---|---|---|
| CTCF | 0.05 | 0.82 | 0.73 |
| FOXK2 | 0.05 | 0.69 | 0.78 |
| IRF1 | 0.05 | 0.65 | 0.57 |
| MEF2A | 0.05 | 0.05 | 0.07 |
| MYC | 0.05 | 0.85 | 0.88 |
| NANOG | 0.05 | 0.45 | 0.12 |
| NFRKB | 0.05 | 0.00 | 0.00 |
| RUNX1 | 0.05 | 0.94 | 0.95 |
| SPI1 | 0.05 | 0.81 | 0.59 |
| TP53 | 0.05 | 0.90 | 0.84 |
| CTCF | 0.20 | 0.00 | 0.00 |
| FOXK2 | 0.20 | 0.01 | 0.03 |
| IRF1 | 0.20 | 0.00 | 0.00 |
| MEF2A | 0.20 | 0.00 | 0.00 |
| MYC | 0.20 | 0.52 | 0.61 |
| NANOG | 0.20 | 0.00 | 0.00 |
| NFRKB | 0.20 | 0.00 | 0.00 |
| RUNX1 | 0.20 | 0.65 | 0.69 |
| SPI1 | 0.20 | 0.25 | - |
| TP53 | 0.20 | 0.18 | 0.26 |

**Table 5.** Similarity assessment of the sets of TFs with increased/decreased affinity for the prolonged peak between experiments targeting the same TF. For each experiment, a set of TFs exhibiting affected affinity (respectively) at the binding site $\pm$ 1500 bp was collected. The intersection over union score reported in the table is calculated over all available experiments targetting a given TF in our datasets. Results for RUNX1 (in particular) should be taken in moderation: only one cell line is available for this TF.

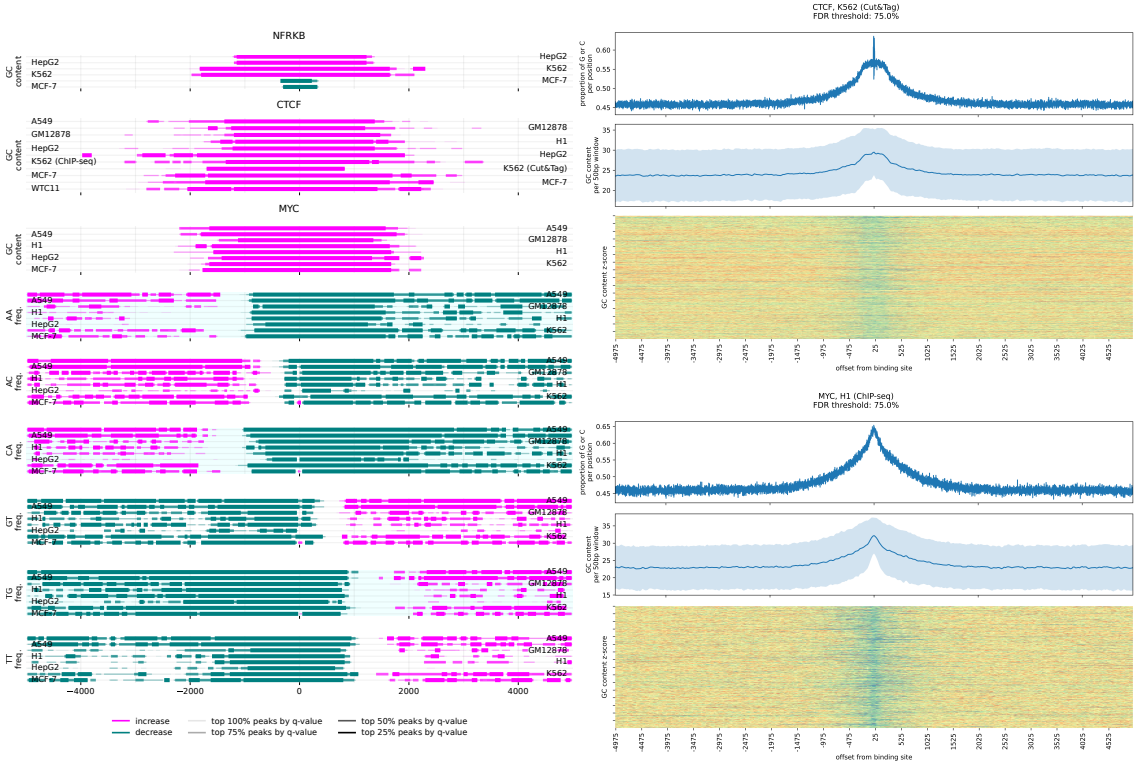



**Fig. 2. Left:** Spans of selected (di)nucleotide frequencies for NFRKB, CTCF and MYC for each cell line available. Decrease is depicted in teal, increase in magenta. Width of the line denotes the source dataset as per the peak FDR score (the thickest corresponds to the most stringently pruned dataset); **Right:** Overview of GC content on the most relevant 25 % of peaks in the Cut&Tag experiment targeting CTCF in K562 (above) and in the ChIP-seq experiment targeting MYC in H1 (ENCFF372RQZ, below). In each of the overviw plots, the top plot depicts G or C base fraction per site. The middle plot shows average GC content per 50 bp window $\pm$ its standard deviation. The bottom plot shows GC content z-score (edges were used as the background distribution) per window in individual prolonged peaks.

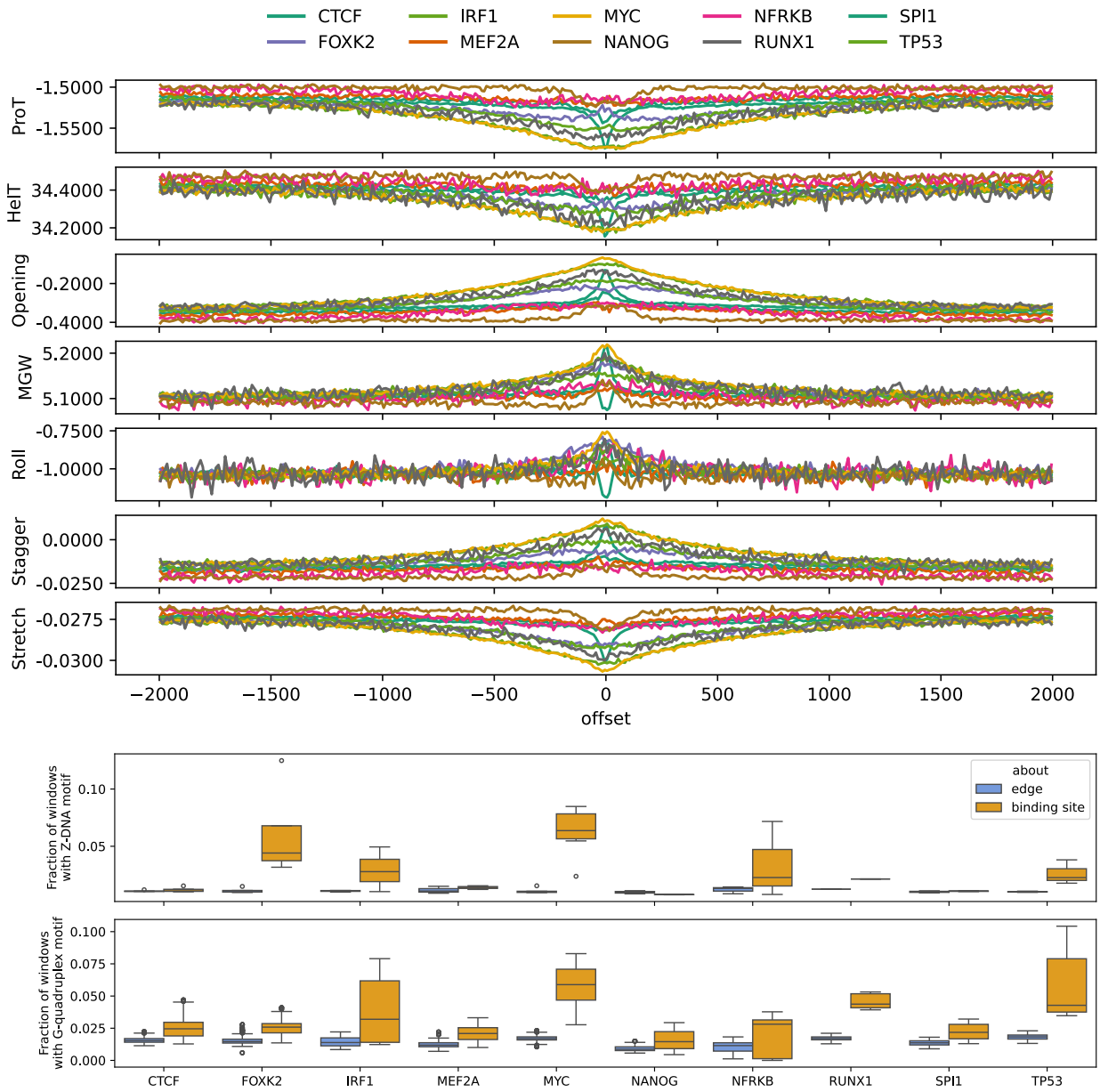



**Fig. 3. Top:** Average values of DNA shape features between -2000 bp and 2000 bp w. r. to the assumed binding site, grouped by TF; **bottom:** comparison of Z-DNA and G-quadruplex presence at the binding site (offset w. r. to the binding site between -100 and 100) and at the edge of the prolonged peak (3000 bp and further from the binding site). For each experiment and group, the largest observed fraction of the respective DNA feature is taken.

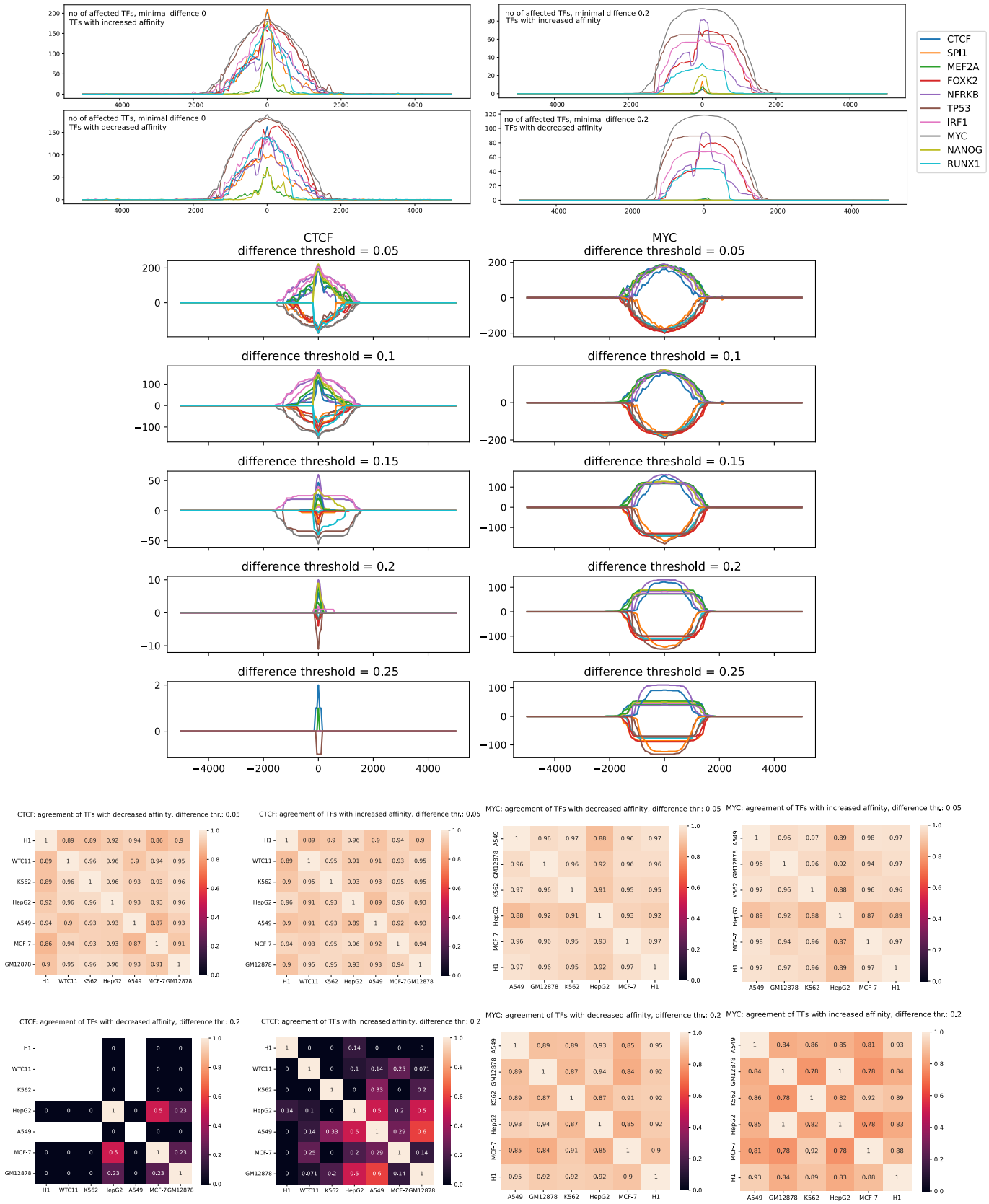
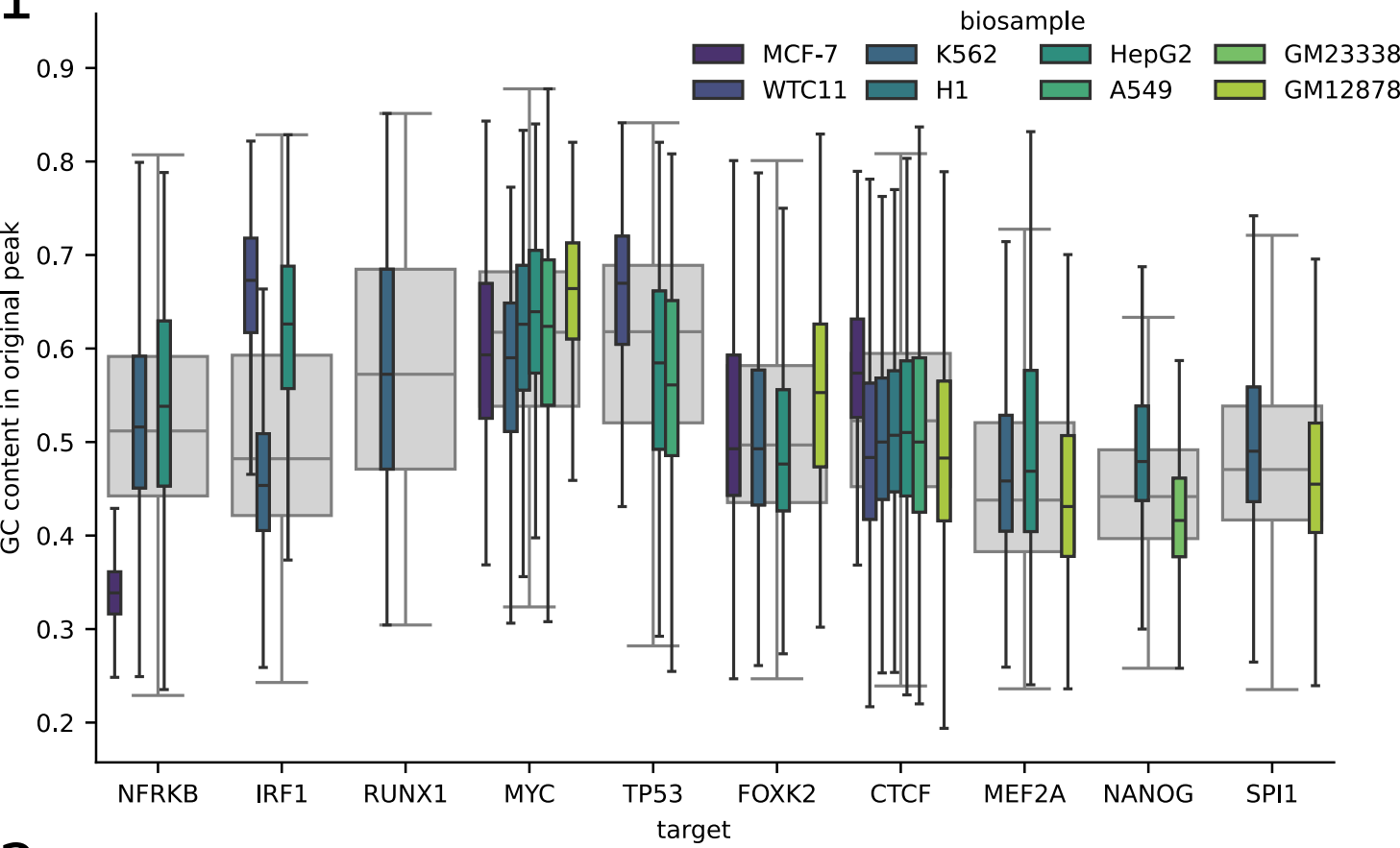



**Fig. 4. Top:** average number of TFs from the non-redundant HOCOMOCO set with increased/decreased affinity to the prolonged peak (with general gene–upstream sequence as a background). The two top left plots denote the number of affected TFs without any difference thresholding, the two top right contains only TFs with at least 0.2 difference between average z-score in affected patch and in unaffected sequence. **Middle:** number of affected TFs per cell line and difference threshold for CTCF (left column) and MYC (right column). TFs with increased affinity are counted in positive numbers, those with decreased affinity are counted in negative numbers. **Bottom:** Highest observed intersection over union on a position $\pm$ 1500bp on the set of TFs with increased (or decreased) affinity between cell line pairs for experiments targetting CTCF and MYC.

1

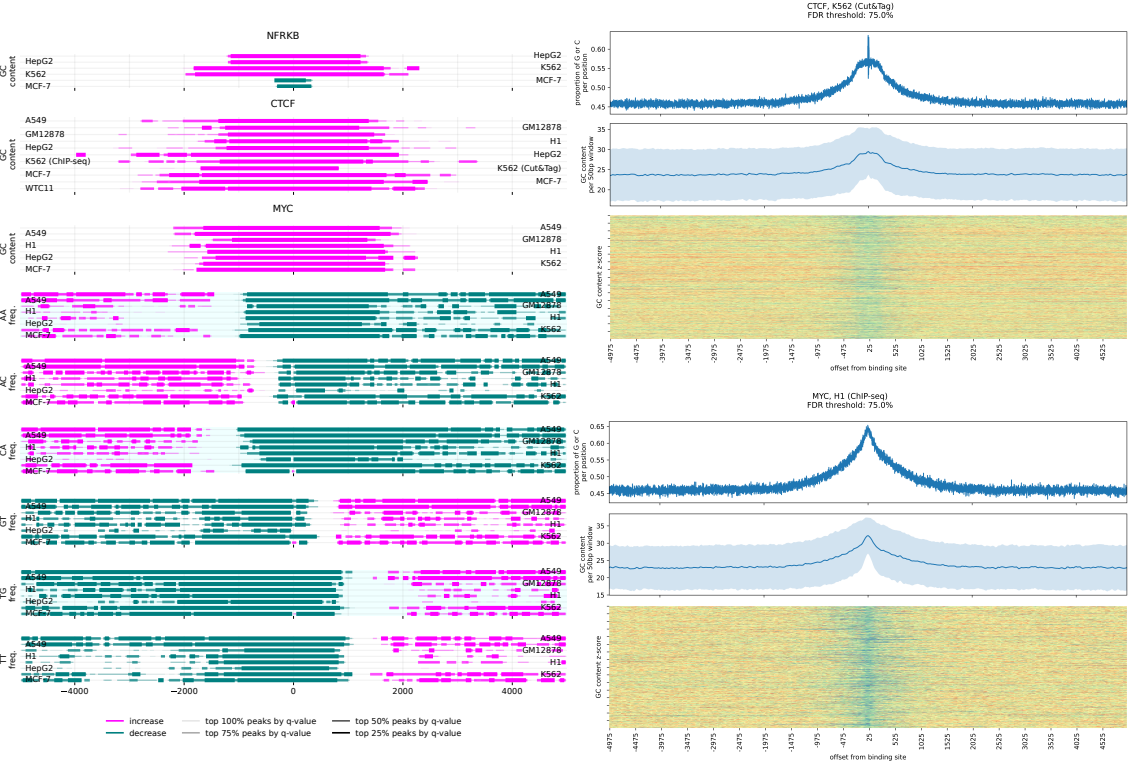


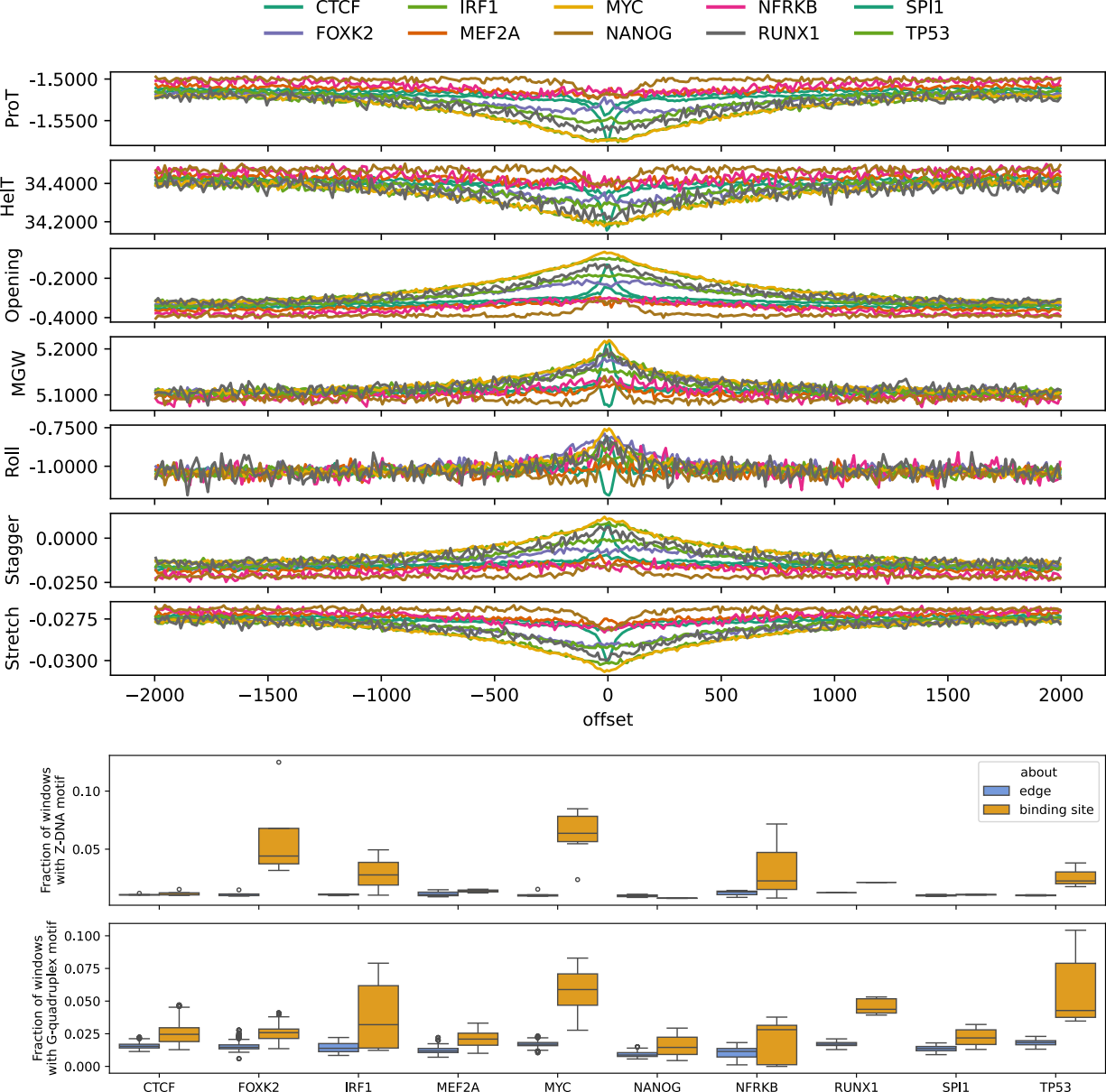

2a

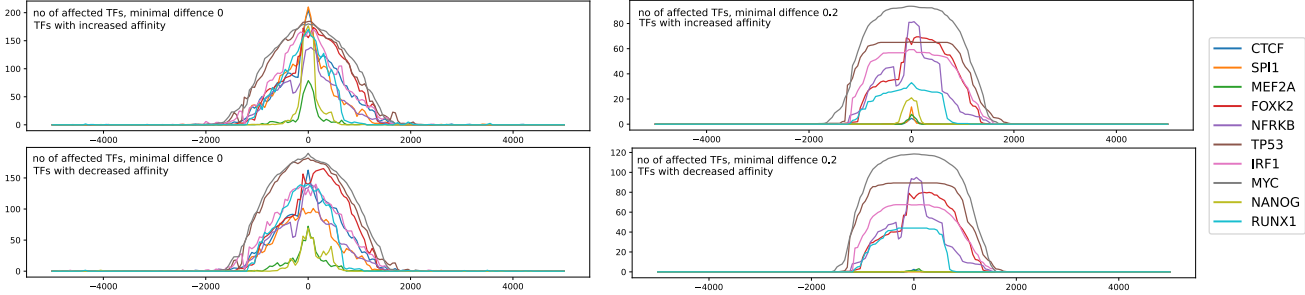
| TF | CTCF | FOXK2 | IRF1 | MEF2A | MYC | NANOG | NFRKB | RUNX1 | SPI1 | TP53 |
|---|---|---|---|---|---|---|---|---|---|---|
| CTCF | 0.56 | 0.56 | 0.67 | 0.6 | 0.68 | 0.62 | 0.64 | 0.61 | 0.57 | 0.65 |
| FOXK2 | | 0.55 | 0.67 | 0.59 | 0.69 | 0.61 | 0.63 | 0.6 | 0.57 | 0.65 |
| IRF1 | | | 0.74 | 0.67 | 0.68 | 0.7 | 0.73 | 0.67 | 0.67 | 0.68 |
| MEF2A | | | | 0.54 | 0.77 | 0.56 | 0.63 | 0.67 | 0.54 | 0.73 |
| MYC | | | | | 0.56 | 0.8 | 0.76 | 0.6 | 0.75 | 0.59 |
| NANOG | | | | | | 0.58 | 0.66 | 0.7 | 0.56 | 0.75 |
| NFRKB | | | | | | | 0.68 | 0.66 | 0.63 | 0.73 |
| RUNX1 | | | | | | | | 0.51 | 0.66 | 0.57 |
| SPI1 | | | | | | | | | 0.54 | 0.71 |
| TP53 | | | | | | | | | | 0.62 |

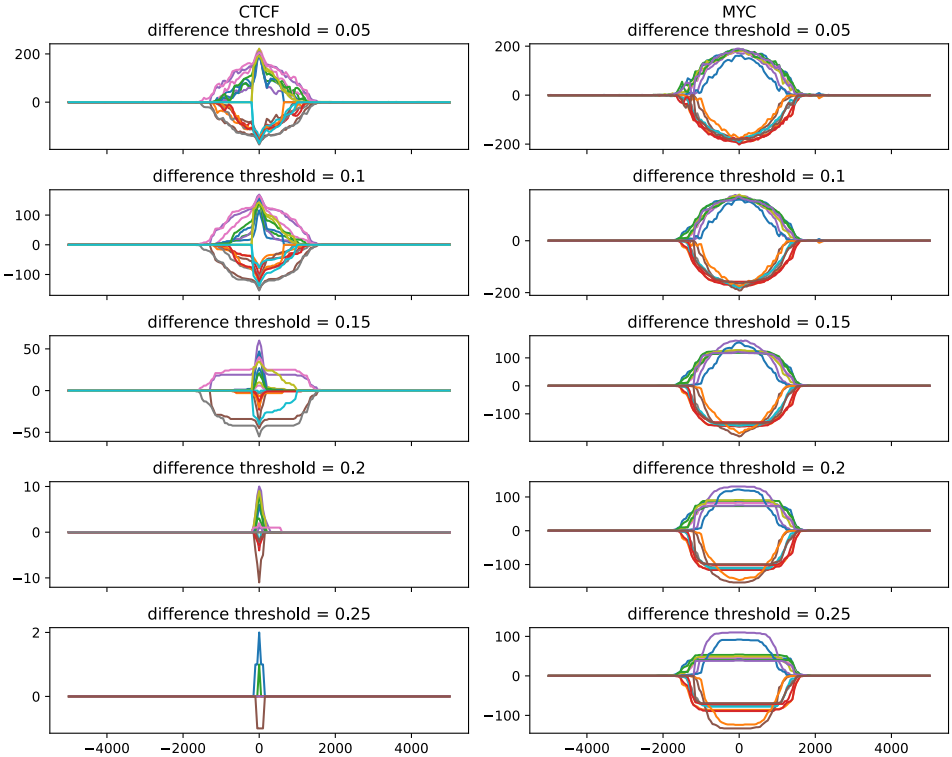
2b

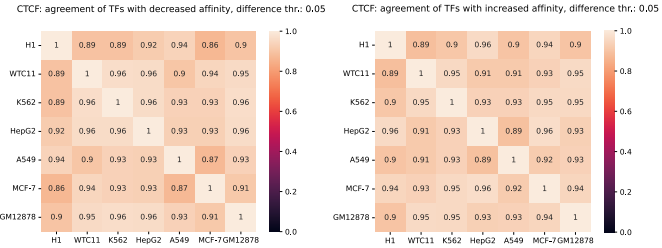
| cell line ID | A549 | GM12878 | GM23338 | H1 | HepG2 | K562 | MCF | WTC11 |
|---|---|---|---|---|---|---|---|---|
| A549 | 0.61 | 0.66 | 0.77 | 0.63 | 0.61 | 0.64 | 0.68 | 0.66 |
| GM12878 | | 0.64 | 0.66 | 0.64 | 0.63 | 0.61 | 0.67 | 0.71 |
| GM23338 | | | 0.51 | 0.72 | 0.72 | 0.66 | 0.7 | 0.82 |
| H1 | | | | 0.65 | 0.62 | 0.62 | 0.67 | 0.68 |
| HepG2 | | | | | 0.61 | 0.61 | 0.66 | 0.67 |
| K562 | | | | | | 0.59 | 0.66 | 0.71 |
| WTC11 | | | | | | | | 0.7 |

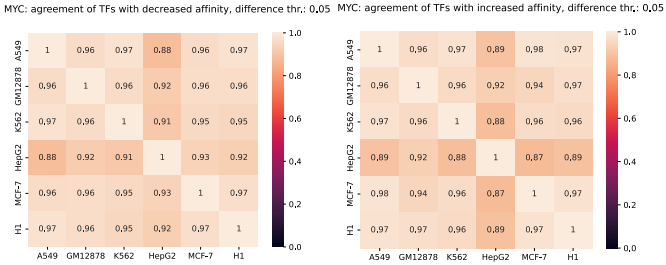

NFRKB
CTCF
MYC
GC content
AA freq
AC freq
CA freq
GT freq
TG freq
TT freq
HepG2
K562
MCF-7
A549
GM12878
H1
K562 (ChIP-seq)
WTC11
K562 (Cut&Tag)
−4000
−2000
0
2000
4000
CTCF, K562 (Cut&Tag)
FDR threshold: 75.0%
MYC, H1 (ChIP-seq)
proportion of G or C
GC content per 50bp window
GC content z-score
offset from binding site
increase
decrease
top 100% peaks by q-value
top 75% peaks by q-value
top 50% peaks by q-value
top 25% peaks by q-value

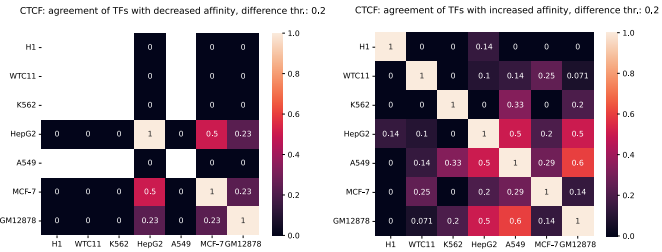

CTCF
FOXK2
IRF1
MEF2A
MYC
NANOG
NFRKB
RUNX1
SPI1
TP53
ProT
HelT
Opening
MGW
Roll
Stagger
Stretch
offset
about
edge
binding site
Fraction of windows with Z-DNA motif
Fraction of windows with G-quadruplex motif

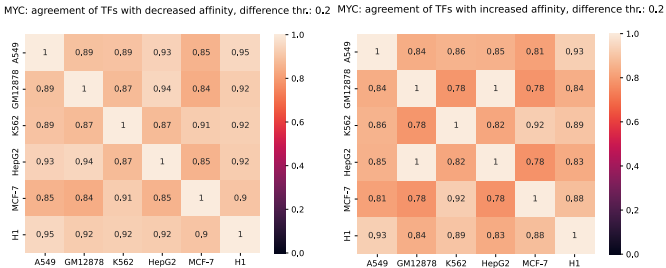

no of affected TFs, minimal diffence 0
TFs with increased affinity
no of affected TFs, minimal diffence 0.2
TFs with increased affinity
no of affected TFs, minimal diffence 0
TFs with decreased affinity
no of affected TFs, minimal diffence 0.2
TFs with decreased affinity
CTCF
SPI1
MEF2A
FOXK2
NFRKB
TP53
IRF1
MYC
NANOG
RUNX1


CTCF
MYC
difference threshold = 0.05
difference threshold = 0.1
difference threshold = 0.15
difference threshold = 0.2
difference threshold = 0.25


CTCF: agreement of TFs with decreased affinity, difference thr.: 0.05
CTCF: agreement of TFs with increased affinity, difference thr.: 0.05


MYC: agreement of TFs with decreased affinity, difference thr.: 0.05
MYC: agreement of TFs with increased affinity, difference thr.: 0.05


CTCF: agreement of TFs with decreased affinity, difference thr.: 0.2
CTCF: agreement of TFs with increased affinity, difference thr.: 0.2


MYC: agreement of TFs with decreased affinity, difference thr.: 0.2
MYC: agreement of TFs with increased affinity, difference thr.: 0.2

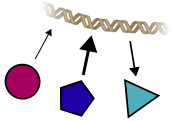
**Dataset:**
publicly available ChIP-seq and Cut&Tag experiments

| **transcription factors:** | | **cell lines:** |
|---|---|---|
| CTCF | × | A549 |
| SPI1 | | GM12878 |
| MEF2A | | GM23338 |
| FOXK2 | | H1 |
| p53 | | HepG2 |
| IRF1 | | K562 |
| MYC | | WTC11 |
| NANOG | | |
| RUNX1 | | |

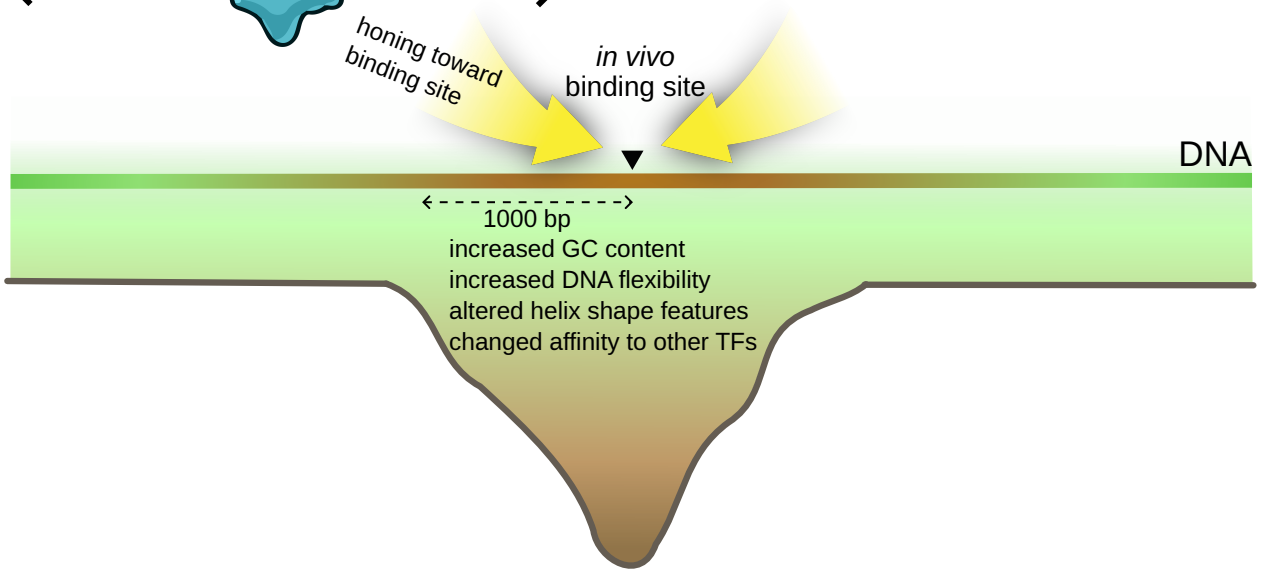
**Analysis of sequence features in large span around the binding site:**

DNA (di)nucleotide preferences

DNA shape features



affinity to various other known TFs

**Proposed aspect of binding site search *in vivo***

TF is coarsely scanning for the binding site