DNA mismatches reveal conformational penalties in protein–DNA recognition

https://doi.org/10.1038/s41586-020-2843-2

Received: 5 January 2019

Accepted: 17 September 2020

Published online: 21 October 2020

Check for updates

Ariel Afek[1,2], Honglue Shi[3], Atul Rangadurai[4], Harshit Sahay[1,5], Alon Senitzki[6], Suela Xhani[7], Mimi Fang[8,9], Raul Salinas[4], Zachery Mielko[1,10], Miles A. Pufall[8,9], Gregory M. K. Poon[7,11], Tali E. Haran[6], Maria A. Schumacher[4], Hashim M. Al-Hashimi[3,4] ✉ & Raluca Gordân[1,2,12,13] ✉

Transcription factors recognize specific genomic sequences to regulate complex gene-expression programs. Although it is well-established that transcription factors bind to specific DNA sequences using a combination of base readout and shape recognition, some fundamental aspects of protein–DNA binding remain poorly understood[1,2]. Many DNA-binding proteins induce changes in the structure of the DNA outside the intrinsic B-DNA envelope. However, how the energetic cost that is associated with distorting the DNA contributes to recognition has proven difficult to study, because the distorted DNA exists in low abundance in the unbound ensemble[3–9]. Here we use a high-throughput assay that we term SaMBA (saturation mismatch-binding assay) to investigate the role of DNA conformational penalties in transcription factor–DNA recognition. In SaMBA, mismatched base pairs are introduced to pre-induce structural distortions in the DNA that are much larger than those induced by changes in the Watson–Crick sequence. Notably, approximately 10% of mismatches increased transcription factor binding, and for each of the 22 transcription factors that were examined, at least one mismatch was found that increased the binding affinity. Mismatches also converted non-specific sites into high-affinity sites, and high-affinity sites into 'super sites' that exhibit stronger affinity than any known canonical binding site. Determination of high-resolution X-ray structures, combined with nuclear magnetic resonance measurements and structural analyses, showed that many of the DNA mismatches that increase binding induce distortions that are similar to those induced by protein binding—thus prepaying some of the energetic cost incurred from deforming the DNA. Our work indicates that conformational penalties are a major determinant of protein–DNA recognition, and reveals mechanisms by which mismatches can recruit transcription factors and thus modulate replication and repair activities in the cell[10,11].

A comprehensive survey of high-resolution structures of transcription factor (TF)-bound DNA revealed that more than 40% of the complexes contain base pairs with geometries that deviate substantially from the B-form envelope of naked DNA duplexes (Extended Data Fig. 1, Methods). The energy required to distort the DNA must come from favourable intermolecular interactions that take place upon complex formation[12,13]. This energetic cost could vary with sequence and contribute to protein–DNA binding affinity and selectivity[1,14,15]. Assessing conformational penalties experimentally is challenging because it requires accurate measurement of the abundance of these distorted DNA conformations in the unbound ensemble—conformations that are difficult to even detect using existing biophysical methods[3,4].

In a similar manner to the effects of TFs, mismatched base pairs can also induce distortions to the DNA ensemble that are much greater than those that occur in naked Watson–Crick sequences (Fig. 1a–c, Extended Data Fig. 2). For example, purine–purine mismatches such as G-G and G-A widen the base pair and can also flip the base into the *syn* conformation; pyrimidine–pyrimidine mismatches such as C-T and T-T constrict the base pair; wobble G-T and T-T mismatches change the shear; and A-A and C-C with only a single hydrogen bond can adopt a variety of conformations including partially melted states

[1]Center for Genomic and Computational Biology, Duke University School of Medicine, Durham, NC, USA. [2]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA. [3]Department of Chemistry, Duke University, Durham, NC, USA. [4]Department of Biochemistry, Duke University School of Medicine, Durham, NC, USA. [5]Program in Computational Biology and Bioinformatics, Duke University School of Medicine, Durham, NC, USA. [6]Department of Biology, Technion–Israel Institute of Technology, Haifa, Israel. [7]Department of Chemistry, Georgia State University, Atlanta, GA, USA. [8]Department of Biochemistry, Carver College of Medicine, University of Iowa, Iowa City, IA, USA. [9]Holden Comprehensive Cancer Center, University of Iowa, Iowa City, IA, USA. [10]Program in Genetics and Genomics, Duke University School of Medicine, Durham, NC, USA. [11]Center for Diagnostics and Therapeutics, Georgia State University, Atlanta, GA, USA. [12]Department of Computer Science, Duke University, Durham, NC, USA. [13]Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Durham, NC, USA. ✉e-mail: hashim.al.hashimi@duke.edu; raluca.gordan@duke.edu

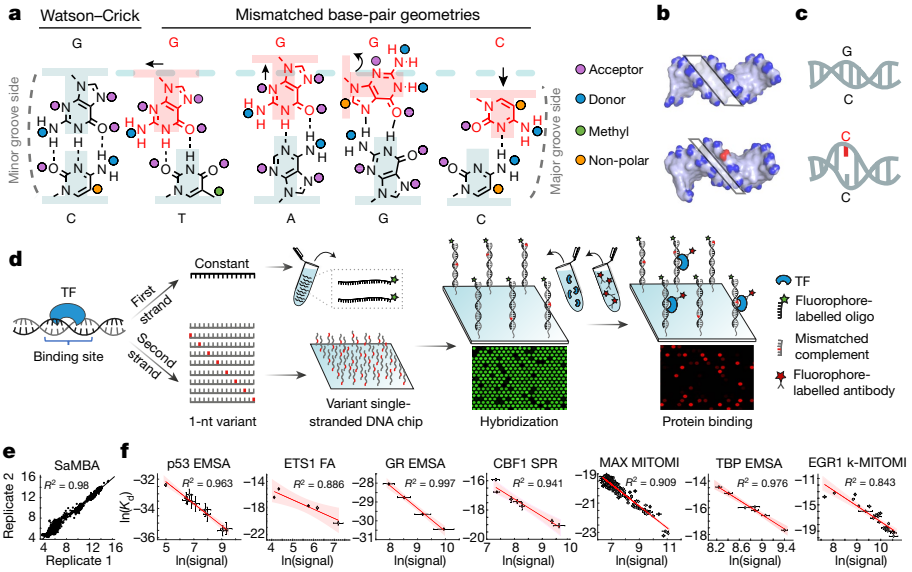


Fig. 1 | SaMBA measures the effects of mismatches on protein–DNA binding in high throughput. a–**c**, Mismatches change the local DNA geometry (**a**), affect global features such as the minor groove width (**b**) and destabilize the DNA (**c**). **d**, SaMBA is a chip-based assay for testing the binding of TFs to thousands of DNA mismatches and Watson–Crick sequences (Methods). DNA hybridization and protein–DNA binding are quantified using fluorophore-labelled oligos and antibodies, respectively. **e**, Reproducibility of SaMBA data, for technical replicates of ETS1 at 125 nM. Axes show the base 2 logarithm of the median fluorescent intensity signal corresponding to the bound ETS1 protein ($n = 12$ replicate spots for Watson–Crick sequences, and $n = 8$ for mismatched sequences). **f**, Protein binding levels measured by SaMBA (shown here for p53, ETS1, the glucocorticoid receptor (GR), CBF1, MAX, TBP and EGR1) correlate linearly with independent K_d measurements from a variety of experimental methods (FA, fluorescence anisotropy; MITOMI, mechanically induced trapping of molecular interactions; k-MITOMI, 'kinetic MITOMI'; SPR, surface plasmon resonance), allowing calibration of SaMBA data. Similarly to related array-based techniques[20], median values over replicate DNA spots are shown for SaMBA (error bars, median absolute deviation). Average values over replicates are shown for the orthogonal methods (error bars, s.d., when available). See Methods for the number of replicates ($n \geq 3$) for each experiment. Red shaded region, 95% confidence interval for Pearson's correlation.

(Extended Data Fig. 2a). Mismatches can also affect the geometry of the DNA minor and major grooves and base-step parameters, albeit to a smaller extent (Extended Data Fig. 2c). In addition, mismatches destabilize the DNA duplex by an amount (3.5–10 k_BT, in which k_B is the Boltzmann constant and T is temperature) (Extended Data Fig. 2b) comparable to the typical energetic cost of distorting the DNA upon protein binding (3–8 k_BT)[16].

SaMBA

To gain insights into the role of DNA conformational penalties in protein–DNA recognition, we developed a new high-throughput approach that we name SaMBA (saturation mismatch-binding assay), which leverages the DNA distortions induced by mismatches. We reasoned that different types of mismatches could redistribute the unbound DNA ensemble in various ways and lead, in some cases, to an increased abundance of distorted DNA states that are recognized by TFs. By pre-paying some of the energetic cost of deforming the DNA, mismatches could in turn increase the TF–DNA binding affinity, provided that the reduction in conformational penalty outweighs any effects caused by the potential loss of protein–DNA contacts. A conceptually similar strategy was used previously to assess conformational penalties in RNA–RNA association[9].

In SaMBA experiments, mismatches are generated by introducing every possible single-base variation in known DNA-binding sites of TFs in a high-throughput manner on a high-density DNA chip (Fig. 1d, Extended Data Fig. 3a–d, Methods). Mismatches are introduced by changing the sequence on one strand at a time (for example, **G**-C → **A**-C, **T**-C and **C**-C (the bases that are changed are shown in bold)). Protein-binding measurements are then conducted directly on the chip, with high reproducibility (Fig. 1e). The SaMBA signal intensities can be calibrated to equilibrium dissociation constants (K_d) using binding measurements from a variety of independent experimental methods (Fig. 1f), thus providing a route for determining binding energetics in a high-throughput manner (Methods, Extended Data Fig. 3e–h, Supplementary Table 3).

Beyond investigating the role of conformational penalties in TF–DNA recognition, SaMBA can be used more broadly to examine the effect of mismatches on protein–DNA binding landscapes and the proposed role of TF-bound mismatches in mutagenesis[10,17–19], including in cases in which mismatches enhance binding by creating or reinforcing favourable interactions that involve hydrogen bonding, electrostatics and stacking (as discussed below).

Mismatches enhance the binding of TFs to DNA

For 22 TFs from 15 distinct protein families, we used SaMBA to obtain saturation mismatch-binding profiles that show the quantitative changes in protein-binding signal induced by the introduction of every possible mismatch to known TF-binding sites and their flanking regions (Fig. 2a, Supplementary Table 1). Although two thirds of the mismatches introduced within TF-binding sites substantially weakened binding, around 10% increased binding. Notably, for each of the 22 TFs examined, at least 1 mismatch was found that increased the binding affinity when introduced within the binding site (Fig. 2a). In some cases, single mismatches created 'super sites' that exhibit a stronger binding affinity than the best canonical Watson–Crick binding sites (for example, in the case of p53) (Supplementary Table 1b). In other cases, mismatches introduced in non-specific DNA sites increased TF binding (Supplementary Table 1d) to levels similar to those observed for specific binding sites, thus effectively creating novel binding sites within non-specific DNA. For ETS1, the protein with the largest mismatch-driven effects outside of specific binding sites, we verified that mismatches could indeed increase TF binding beyond the distribution of non-specific binding

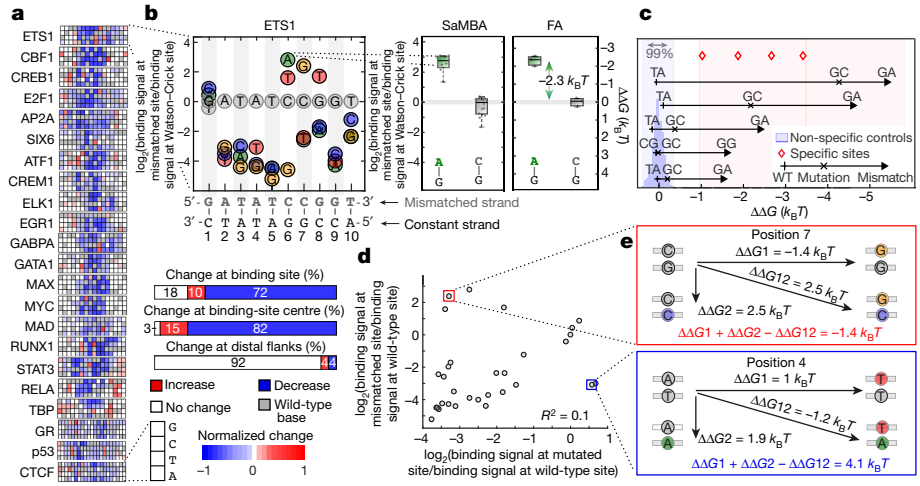


Fig. 2 | The effects of DNA mismatches on TF binding. a, SaMBA profiles for the 22 tested TFs. Heat maps show the effects of mismatches on TF binding, normalized so that −1 corresponds to the largest decrease (Methods). **b**, SaMBA profile for ETS1, with a representative mismatch-induced binding increase that was independently validated by fluorescence anisotropy. The *y* axis shows the base 2 logarithm of the ratio between the ETS1 binding signal at the mismatched site versus the Watson–Crick site, where binding signals are computed as median fluorescent signal intensities over replicate DNA spots. Coloured circles indicate significant changes (P value < 0.05, one-sided Mann–Whitney U test with Benjamini–Hochberg correction). Box plots show median signals over replicate DNA spots for SaMBA ($n = 8$ and $n = 12$ for the mismatched and Watson–Crick site, respectively) and replicate experiments for EMSA ($n = 3$). Boxes extend to the 25th and 75th percentiles; whiskers extend to the most-extreme data points. **c**, Five validated examples of mismatches in non-specific sequences that increase ETS1 binding to levels similar to specific sites (Methods). Each arrow corresponds to one mismatch in a particular non-specific sequence (Supplementary Table 2c). In some cases, Watson–Crick mutations also increase binding affinity, albeit to a smaller extent, indicating that the identity of the newly introduced base is important for enhanced binding affinity (Supplementary Table 2, Extended Data Fig. 5). **d**, Comparison of mismatch versus mutation effects for the ETS1 site in **b**, for mismatches on the upper strand. Values represent medians over replicate spots ($n = 8$). **e**, The energetic effects of base-pair mutations (diagonal) are different from the sum of the energetic effects of the two corresponding mismatches, demonstrating deviations from an additive model.

affinities (defined here as the 99th percentile of random sites) and towards high affinities characteristic of specific binding sites (defined here as sites with ETS1-bound nuclear magnetic resonance (NMR) structures or crystal structures) (Methods, Fig. 2c, Supplementary Table 2).

We verified representative examples of mismatch-induced enhancements in TF-binding sites using fluorescence anisotropy and electrophoretic mobility shift assays (EMSA), and found binding increases of 0.7–2.3 k_BT relative to consensus Watson–Crick binding sites (Fig. 2b, Extended Data Fig. 3e). Overall, the magnitude of mismatch-induced effects on TF binding was comparable to the magnitude of the effects of mutations in Watson–Crick binding sites (Extended Data Fig. 4a), although the directionality of these effects was sometimes opposite for mismatches versus their nearest mutations (for example, C-G → G-G increases binding, whereas C-G → G-C decreases binding) (Fig. 2d, e, Extended Data Fig. 4b). This shows that mismatches can provide an additional layer of information about TF–DNA interactions beyond what can be learned from analysing the effects of mutations in Watson–Crick DNA using traditional high-throughput methods[20–24].

Mismatches versus Watson–Crick mutations

The simplest explanation for the observed mismatch-induced increase in TF binding affinity is that the mutated base forms more favourable interactions with the TF, in a manner that is independent of the mismatch shape. In this simple additive model, each base in a base pair contributes independently to the TF binding energetics. Such a model predicts that the sum of the energetic changes (gains or losses) from the two individual single-base mutations is equal to the change in binding energy due to the double mutation (for example, $\Delta\Delta G_{\mathrm{CG>CT}} + \Delta\Delta G_{\mathrm{CG>AG}} = \Delta\Delta G_{\mathrm{CG>AT}}$ (mutated bases are shown in bold)). On the other hand, any mismatch-shape-dependent contribution to increased TF binding—including changes in the DNA ensemble that might help offset the energetic cost of DNA deformation—could lead to deviations from the additive model. We tested this simple model for the seven TFs for which calibration data were available in our study (Methods, Extended Data Fig. 4c, Supplementary Table 4). We found that additivity holds, within experimental error, in around 42% of cases in which mismatches significantly affect TF binding (for example, for ETS1 we found that $\Delta\Delta G_{\mathrm{AT>AG}} + \Delta\Delta G_{\mathrm{AT>CT}} \approx \Delta\Delta G_{\mathrm{AT>CG}}$ for position 7 in the binding site) (Extended Data Fig. 4c). For the remaining cases (around 58%), Watson–Crick mutations had a different energetic effect on TF binding compared to the sum of the two corresponding mismatches (Fig. 2e, Extended Data Fig. 4c, Supplementary Table 4)—indicating that the contributions of the mispaired bases are non-additive. Although non-additive models have been previously tested with regard to base pairs in Watson–Crick binding sites[25,26], our TF–mismatch binding data provide a unique opportunity to investigate dependencies between bases in a base pair.

Mismatches prepay distortion penalty

Deviations from the simple additive model can arise from various mechanisms. These include non-native interactions with the newly formed mismatch-dependent DNA shape (including the bases), and the reinforcement of native interactions, owing to mismatch-specific changes in the DNA ensemble that offset the conformational penalties associated with distorting the DNA upon TF binding. For the latter case, we would expect the mismatches to be located in regions that are distorted in the protein-bound DNA structure. Indeed, for the subset of 12 TFs for which structures were available at the RCSB Protein Data Bank (PDB), we found that the binding-site positions for which mismatches enhanced TF binding affinity were significantly more distorted than the rest of the binding-site positions, in terms of either the magnitude of the distortions ($P = 0.017$) or the number of distorted features ($P = 0.015$) (Methods, Supplementary Table 5).

If mismatches increase binding affinity in part by prepaying conformational penalties, we would also expect mismatches to bias local

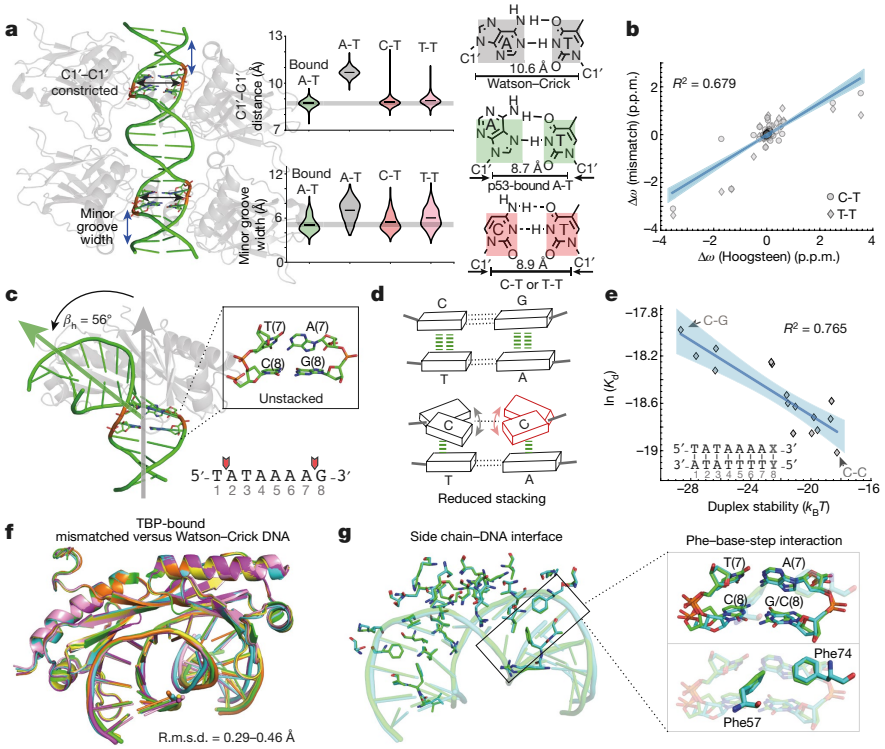


Fig. 3 | DNA mismatches that exhibit geometries similar to distorted base pairs in TF-bound DNA lead to increased binding affinity. a, Crystal structure of the p53–DNA complex shows a constricted Hoogsteen conformation at the positions marked in red. C-T and T-T mismatches, which increase p53-DNA binding affinity, mimic Hoogsteen base-pairing by constricting the C1′–C1′ distance and minor groove width. Violin plots show the distributions of the C1′–C1′ distance and minor groove width according to MD simulation data (Methods). **b**, NMR results confirm that T-T and C-T mismatches mimic Hoogsteen A-T geometry. Plot shows the chemical shift differences in the sugar C1′, C3′ and C4′ carbon atoms for T-T and C-T mismatches versus a locked Hoogsteen conformation (using N^1-methyladenosine[30]), relative to the Watson–Crick base-paired duplex (Methods). Blue shaded region, 95% confidence interval for Pearson's correlation. $\Delta\omega$ is the chemical shift difference between the mismatched (or Hoogsteen) duplex and the Watson–Crick duplex. **c**, Crystal structure of the TBP–DNA complex shows destabilization at an ApG base-pair step (positions 7–8) critical for TBP binding[8,31,32]. β_h, bending magnitude (Methods). **d**, The C-C mismatch destabilizes the DNA and has the lowest stacking propensity[33]. **e**, High correlation between TBP binding levels (medians over 9 replicate spots) and DNA duplex stability (Methods), computed over all base-pair variants at position 8 in the TBP site, suggests that prepaying the energetic cost for melting this base-pair modulates TBP binding affinity. Blue shaded region, 95% confidence interval for Pearson's correlation. **f**, Structural overlay of six TBP–DNA complexes shows that the complexes have nearly identical structures. Green, PDB 1QNE, Watson–Crick site 5′-TATAAAAG-3′; cyan, TBP-CC(2), 5′-TATAAAA**G**-3′ with CC at position 8; orange, TBP-AC, 5′-TATAAA**A**G-3′ with AC at position 7; yellow, PDB 6NJQ, Watson–Crick site 5′-TATAAACG-3′; purple, TBP-CC(1a) and pink, TBP-CC(1b), 5′-TATAAA**C**G-3′ with CC at position 7. Bold font shows the positions where mismatches were introduced. **g**, Overlay of the TBP–DNA interfaces (for 1QNE and TBP-CC(2)) demonstrates that interactions are highly similar between Watson–Crick and mismatched sites, including Phe interactions at the position of the mismatch (black rectangle).

or global aspects of the DNA structural ensemble to better mimic the structure of the DNA when bound to the TF. Because such ensembles are difficult to obtain, we used high-resolution crystal structure data (which were available for 12 TFs in our study) to compare the distortions in the TF-bound DNA with the distortions induced by mismatches (Methods). We observed some form of structural mimicry in 66% of cases (Supplementary Table 6). Returning to the example of ETS1, we found that the G-A mismatch at position 6—which increases binding by around 2.3 k_BT (Fig. 2b)—mimics the stretch, the C1′–C1′ distance and the minor groove width of ETS1-bound DNA (Extended Data Fig. 5b, Supplementary Table 6d). In addition, molecular dynamics (MD) simulations of the bound mismatched and Watson–Crick DNA for this and other mismatches that increase TF binding (Extended Data Fig. 5c, Supplementary Table 7) suggest that the formation of new protein–DNA contacts might also contribute to the enhanced binding affinity. Together, these data indicate that a single mismatch can affect the energetics of several types of interactions, including base readout, shape readout and conformational penalties.

To better isolate contributions from the energetic penalty, we focused on mismatches that enhanced the binding of p53 and TATA-binding protein (TBP). These mismatches were selected because they showed structural mimicry in base-pair features that deviate most strongly from the B-form envelope, and they occurred at positions that lack hydrogen bonds with the bases (Supplementary Table 5). In the case of p53, two positions in each p53 half-site have a preference for adopting non-canonical Hoogsteen conformations[27,28] (Fig. 3a). Hoogsteen base pairs represent an example of alternative, sparsely populated conformations in apo-DNA ensembles; they form with an abundance of less than 1%, at an estimated energetic cost of 3–7 k_BT[4,29,30]. The Hoogsteen pairing is achieved by flipping the purine base from an *anti* to a *syn* conformation, followed by a reduction of around 2 Å in the helical diameter and the C1′–C1′ distance. This reduction in DNA diameter at the p53-binding site allows p53 monomers to come into closer proximity, thus stabilizing the p53 tetramer[28]. Notably, our SaMBA data revealed that replacing the A-T at Hoogsteen sites with T-T or C-T mismatches, which also reduce the C1′–C1′ distance (Fig. 3a), enhanced the binding affinity of p53 by around 0.4–1.8 k_BT (Supplementary Tables 3, 4)—comparable to the magnitude of changes in p53 binding affinity caused by base-pair mutations (Extended Data Fig. 4a).

NMR analysis confirmed that the perturbations induced by A-T Hoogsteen base pairs[30] are similar to those induced by T-T and C-T mismatches (Fig. 3b, Extended Data Fig. 6c, d). The T-T and C-T

mismatches also induced narrowing of the minor groove width, thus resulting in an enhanced negative electrostatic potential[12], and the C-T mismatch led to over-twisting of the DNA helix, mimicking the p53-bound Watson–Crick structure (Fig. 3a, Extended Data Fig. 6e). These results indicate that T-T and C-T mismatches effectively mimic, in naked DNA, structural features of the Hoogsteen pairing favoured by p53, and thereby prepay some of the energetic penalty to form the preferred bound structure. As T-T and C-T mismatches do not increase the binding energetics to the same extent as the cost of forming Hoogsteen base pairs, it is possible that the mismatches do not mimic all aspects of the Hoogsteen conformation, and/or that the Hoogsteen conformation is not fully populated in the protein-bound state of the Watson–Crick DNA.

To test whether the reduction in DNA diameter is causing the increased binding of p53 to mismatched DNA, we measured the effects of all mismatches at the four Hoogsteen positions in the p53-binding site, using not only single-base variations (which are typical for SaMBA assays) but also double-base variations (Methods, Supplementary Table 4). As expected, pyrimidine–pyrimidine mismatches (C-T, T-C, T-T and C-C) enhanced p53 binding affinity, whereas all other mismatches at these positions either decreased binding or had non-significant effects (Extended Data Fig. 6f), consistent with our hypothesis. These findings are in line with a previous study in which modified bases were shown to induce Hoogsteen conformations and increase p53 binding affinity in a similar manner[28].

For TBP, previous studies have shown that partial intercalation of Phe residues at the first and last base steps of the **TA**TAAA**AG** binding site (base steps are shown in bold) leads to a loss of base stacking and the formation of a sharp kink as a key feature of the bound DNA[8,31,32] (Fig. 3c). Mismatches also destabilize the DNA duplex, with C-C having the least-favourable stacking interactions[33] (Fig. 3d). Notably, introducing mismatches at position 8 in the TBP-binding site, which is one of the highly unstacked positions, resulted in an increase in TBP binding affinity, with C-C having the largest effect (Extended Data Fig. 7a). This indicates that mismatches increase affinity by prepaying the energetic cost to partially melt the base pairs. If this were true, we would expect an inverse correlation between the increase in binding affinity and the stability of the mismatch. To test this prediction, we performed additional TBP-binding measurements for all mismatches and base-pair mutations at each position in the TBP-binding site using a modified SaMBA protocol (Methods, Supplementary Table 4). We compared these binding measurements to predicted destabilization energies (Methods) and observed a strong correlation ($R^2 = 0.765$) (Fig. 3e). Analysis of the other positions in the TBP-binding site revealed high correlations between destabilization energies and TBP binding ($R^2 > 0.4$) at three of the four unstacked positions (Extended Data Fig. 7b). No significant correlations were observed at other positions in the binding site, consistent with our hypothesis.

To further examine how mismatches affect protein–DNA binding, we solved four X-ray structures of TBP bound to DNA containing C-C and A-C mismatches at the unstacked positions 7 and 8, which increase the TBP binding affinity by 0.8–1.4 k_BT (resolution 2.0–2.5 Å) (Fig. 3f, Extended Data Fig. 7c, Supplementary Table 4). These structures are the first, to our knowledge, examples of structures of mismatch-containing DNA bound by a TF, and shed light on how mismatches might increase binding affinity. The heavy atoms of the structures superimpose with a root mean square deviation (r.m.s.d.) of 0.29–0.49 Å, which suggests that TBP interacts with mismatched and Watson–Crick DNA sites in a nearly identical manner, including in and around the mismatches (Fig. 3f, Extended Data Fig. 7c, Supplementary Discussion). Notably, the four TBP–DNA structures were obtained from distinct crystal forms (Extended Data Table 1), indicating that packing was not a factor in the similar DNA conformations. In all cases, no evidence was found for new contacts with the mismatches that would explain the large increases in TBP binding. This provides further evidence that mismatch-induced enhancements in protein binding can arise from prepaying energetic penalties that are invisible to detection based on X-ray structures.

Native and non-native interactions

In addition to prepaying conformational penalties and thus reinforcing native interactions (that is, hydrogen bonds and water-mediated, electrostatic and other interactions that would also form in Watson–Crick DNA), our MD simulation data also suggest that mismatches can enhance TF binding by promoting non-native interactions with the mismatched DNA, through changes in both the base identity and the DNA conformation at the mismatch and/or neighbouring sites. For example, in the case of the T-G mismatch at position 6 in the ETS1-binding site, for which no structural mimicry was identified in our analyses (Supplementary Table 6), MD simulations of protein-bound mismatched and Watson–Crick DNA revealed that the wobble conformation positions the mismatched T base to form non-native contacts with protein side chains (Extended Data Fig. 5e, Supplementary Table 7). Non-native interactions were also observed in MD simulations of non-specific sites that are rendered high-affinity ETS1-binding sites by specific mismatches (Extended Data Fig. 5i, j, Supplementary Table 7). In addition, a combination of non-native interactions and structural mimicry is observed in the case of A-G at position 6 in the ETS1-binding site (Extended Data Fig. 5b, e, h). Determining the structures of these complexes may help to reveal the nature of the non-native interactions, which could also include water-mediated hydrogen bonds and electrostatic interactions that might enhance the binding energetics (Extended Data Fig. 8).

Summary

Our study provides the largest analysis to date, to our knowledge, of the effects of DNA mismatches on protein binding, and reveals that DNA conformational penalties are an important determinant of protein–DNA binding affinity and selectivity. Our assay can be extended to include distortions in DNA shape that are induced by multiple mismatches, insertions and deletions, as well as damaged and epigenetically modified nucleotides, and can thus be used to thoroughly investigate these penalties in a high-throughput and unbiased manner. In addition, regardless of the precise mechanisms by which mismatches enhance TF binding, these high-affinity interactions could provide a biophysical mechanism for inhibiting the repair of specific mismatched sites, which would consequently contribute to the formation of genetic mutations in the cell[11] (Extended Data Fig. 9, Supplementary Discussion).

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2843-2.

1. Rohs, R. et al. Origins of specificity in protein–DNA recognition. *Annu. Rev. Biochem.* **79**, 233–269 (2010).
2. Siggers, T. & Gordân, R. Protein–DNA binding: complexities and multi-protein codes. *Nucleic Acids Res.* **42**, 2099–2111 (2014).
3. Guéron, M., Kochoyan, M. & Leroy, J.-L. A single mode of DNA base-pair opening drives imino proton exchange. *Nature* **328**, 89–92 (1987).
4. Nikolova, E. N. et al. Transient Hoogsteen base pairs in canonical duplex DNA. *Nature* **470**, 498–502 (2011).
5. Fischer, M., Coleman, R. G., Fraser, J. S. & Shoichet, B. K. Incorporation of protein flexibility and conformational energy penalties in docking screens to improve ligand discovery. *Nat. Chem.* **6**, 575–583 (2014).
6. Fraser, J. S. et al. Hidden alternative structures of proline isomerase essential for catalysis. *Nature* **462**, 669–673 (2009).
7. Lorch, Y., Davis, B. & Kornberg, R. D. Chromatin remodeling by DNA bending, not twisting. *Proc. Natl Acad. Sci. USA* **102**, 1329–1332 (2005).

8. Parvin, J. D., McCormick, R. J., Sharp, P. A. & Fisher, D. E. Pre-bending of a promoter sequence enhances affinity for the TATA-binding factor. *Nature* **373**, 724–727 (1995).
9. Denny, S. K. et al. High-throughput investigation of diverse junction elements in RNA tertiary folding. *Cell* **174**, 377–390 (2018).
10. Reijns, M. A. M. et al. Lagging-strand replication shapes the mutational landscape of the genome. *Nature* **518**, 502–506 (2015).
11. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264–267 (2016).
12. Rohs, R. et al. The role of DNA shape in protein–DNA recognition. *Nature* **461**, 1248–1253 (2009).
13. Zeiske, T. et al. Intrinsic DNA shape accounts for affinity differences between Hox-cofactor binding sites. *Cell Rep*. **24**, 2221–2230 (2018).
14. Azad, R. N. et al. Experimental maps of DNA structure at nucleotide resolution distinguish intrinsic from protein-induced DNA deformations. *Nucleic Acids Res*. **46**, 2636–2647 (2018).
15. Olson, W. K., Gorin, A. A., Lu, X.-J., Hock, L. M. & Zhurkin, V. B. DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proc. Natl Acad. Sci. USA* **95**, 11163–11168 (1998).
16. Battistini, F. et al. How B-DNA dynamics decipher sequence-selective protein recognition. *J. Mol. Biol*. **431**, 3845–3859 (2019).
17. Kunkel, T. A. & Erie, D. A. Eukaryotic mismatch repair in relation to DNA replication. *Annu. Rev. Genet*. **49**, 291–313 (2015).
18. Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715 (1993).
19. Pich, O. et al. Somatic and germline mutation periodicity follow the orientation of the DNA minor groove around nucleosomes. *Cell* **175**, 1074–1087 (2018).
20. Berger, M. F. et al. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol*. **24**, 1429–1435 (2006).
21. Shen, N. et al. Divergence in DNA specificity among paralogous transcription factors contributes to their differential in vivo binding. *Cell Syst*. **6**, 470–483 (2018).
22. Veprintsev, D. B. & Fersht, A. R. Algorithm for prediction of tumour suppressor p53 affinity for binding sites in DNA. *Nucleic Acids Res*. **36**, 1589–1598 (2008).
23. Jolma, A. et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res*. **20**, 861–873 (2010).
24. Warren, C. L. et al. Defining the sequence-recognition profile of DNA-binding molecules. *Proc. Natl Acad. Sci. USA* **103**, 867–872 (2006).
25. Benos, P. V., Bulyk, M. L. & Stormo, G. D. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res*. **30**, 4442–4451 (2002).
26. Chattopadhyay, A., Zandarashvili, L., Luu, R. H. & Iwahara, J. Thermodynamic additivity for impacts of base-pair substitutions on association of the Egr-1 zinc-finger protein with DNA. *Biochemistry* **55**, 6467–6474 (2016).
27. Kitayner, M. et al. Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. *Nat. Struct. Mol. Biol*. **17**, 423–429 (2010).
28. Golovenko, D. et al. New insights into the role of DNA shape on its recognition by p53 proteins. *Structure* **26**, 1237–1250 (2018).
29. Alvey, H. S., Gottardo, F. L., Nikolova, E. N. & Al-Hashimi, H. M. Widespread transient Hoogsteen base pairs in canonical duplex DNA with variable energetics. *Nat. Commun*. **5**, 4786 (2014).
30. Shi, H. et al. Atomic structures of excited state A-T Hoogsteen base pairs in duplex DNA by combining NMR relaxation dispersion, mutagenesis, and chemical shift calculations. *J. Biomol. NMR* **70**, 229–244 (2018).
31. Kim, J. L., Nikolov, D. B. & Burley, S. K. Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature* **365**, 520–527 (1993).
32. Mondal, M., Mukherjee, S. & Bhattacharyya, D. Contribution of phenylalanine side chain intercalation to the TATA-box binding protein-DNA interaction: molecular dynamics and dispersion-corrected density functional theory studies. *J. Mol. Model*. **20**, 2499 (2014).
33. Peyret, N., Seneviratne, P. A., Allawi, H. T. & SantaLucia, J., Jr. Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A, C.C, G.G, and T.T mismatches. *Biochemistry* **38**, 3468–3477 (1999).

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Structural survey of Watson–Crick and mismatched base pairs

We performed a comprehensive survey of DNA base-pair structures deposited in the RCSB PDB[34]. X-ray crystal structures (resolution < 3 Å) and NMR solution structures containing DNA were downloaded from the RCSB web server and organized into a searchable database[35]. Base-pair parameters (shear, stretch, stagger, buckle, propeller twist, opening and C1′–C1′ distance) of a given base pair, as well as base-step parameters (shift, slide, rise, tilt, roll and twist) were computed using X3DNA-DSSR[36] as described previously[37]. Base-pair parameters (except C1′–C1′ distance) and base-step parameters of bases with *syn* conformation (for example, in Hoogsteen base pairs and G-A and G-G mismatches) were not computed owing to incorrect reference frame.

The overall shape of the DNA was characterized by analysing the following shape parameters: minor groove width, major groove width, local helical bending, bending direction and local helical twisting. Minor and major groove widths were calculated using the P-P definition[38] by X3DNA-DSSR[36]. A well-established inter-helical Euler angle approach was used to quantify DNA local bending, including the bending magnitude (β_h, $0° \le \beta_h \le 180°$), the bending direction (γ_h, $-180° \le \gamma_h \le 180°$) and the helical twist (ζ_h, $-180° \le \zeta_h \le 180°$) of two helices across a given base-pair junction[35,37,39,40]. All calculations with poor alignment to the idealized helices (r.m.s.d. >2 Å for sugar and backbone atoms[39]) were omitted from analysis. Global parameters were analysed at the mismatch positions as well as ±1 base pair or base step.

A total of 903 A-T and 746 G-C standard Watson–Crick base pairs in naked DNA were identified (Supplementary Methods) and used to define the B-DNA envelope (Extended Data Fig. 1a, Supplementary Table 8). A total of 613 TF–DNA structures in the PDB[34] were used to identify Watson–Crick base pairs for which at least one base-pair parameter deviates from the free B-DNA envelope by three standard deviations or is completely outside the envelope. The statistics of these distorted Watson–Crick base pairs in TF-bound DNA are summarized in Extended Data Fig. 1 and Supplementary Table 8. To survey the DNA mismatch structure and geometry, all possible single mismatches (excluding modified bases) surrounded by at least two canonical Watson–Crick base pairs on both sides were identified and subjected to manual inspection (Supplementary Table 9). Of the 110 identified mismatches, 26 were in free DNA and not mediated by heavy metals (8 G-T, 7 G-A, 5 A-C, 3 T-T, 2 G-G and 1 C-T) (Supplementary Table 9, Extended Data Fig. 2a).

DNA melting analysis

Thermodynamic parameters for mismatch formation were computed using MELTING v.5.2.0 (ref. [41]) as an average over all possible sequence contexts surrounding each mismatch. Default options for nearest neighbour thermodynamic parameters and ion correction terms were used along with a sodium ion concentration of 150 mM. The energetic terms for helix initiation and symmetry were set to zero to mimic the placement of the mismatch within the context of a non-palindromic duplex.

Molecular dynamics simulations

All MD simulations were performed using the AMBER ff99 force field[42] with bsc0 corrections for DNA[43] and ff14SB corrections for proteins[44], and using standard periodic boundary conditions as implemented in the AMBER MD package[45]. To systematically analyse the ensemble behaviour of all mismatches, we performed MD simulations on unbound DNA for all possible Watson–Crick and mismatched base pairs embedded in constant flanking sequences: 5′-CTCTGCC**A**CGTGGGTCGT-3′ (the variable position is shown in bold). For G-A and G-G, we simulated two possible geometries: G(*anti*)-A(*anti*), G(*anti*)-A(*syn*) and G(*anti*)-G(*syn*), G(*syn*)-G(*anti*), in which one of the bases was manually rotated around the glycosidic bond by 180° to generate a *syn* conformation. Production runs of 500 ns were carried out and extended to achieve convergence of the r.m.s.d. of the DNA if necessary. Summary descriptions of the ensemble behaviour of different mismatches in the unbound DNA simulations are presented in Extended Data Fig. 2c. The dynamics of DNA mismatches in MD simulations are in good agreement with previous work[46].

For MD simulations of protein–DNA complexes, starting structures corresponding to the MYC/MAX, ETS1, p53, MAX/MAX, CTCF, EGR1, GR, ELK1 and RELA systems were obtained from PDB entries 1NKP, 2NNY, 3KZ8, 1AN2, 5KKQ, 1P47, 1R4R, 1DUX and 5U01, respectively (see Supplementary Methods for details). The TFs were chosen according to the availability of TF–DNA structures for DNA sequences similar to the ones tested by SaMBA. Production runs of 200 or 500 ns were carried out and extended to achieve convergence of the r.m.s.d. of the protein–DNA complex if necessary. For proteins bound to mismatched DNA sites, we chose not to simulate the mismatches A-A, A-C and C-C, given the lack of a stable base-pairing geometry for A-A[47] and the tendency of A-C and C-C to undergo protonation-dependent structural changes to form stable base-pairing geometries[48,49]. Protonation-dependent base-pairing conformational equilibria are susceptible to being highly influenced by protein binding[50], and are also difficult to model computationally[51]. We simulated one mismatch per protein, focusing on G-T and C-T mismatches, as well as T-T, G-G and G-A in specific cases, given their stable base-pairing geometries[52–56] and ability to be reliably modelled computationally[46]. The simulation results were used to analyse protein–DNA contacts and the buried surface area (Supplementary Table 7), as described in Supplementary Methods.

Protein expression and purification

For SaMBA experiments, full-length human proteins ETS1, ELK1, GABPA, RUNX1, E2F1, SIX6, AP2A, GATA1, MYC (c-MYC), MAX and MAD (MAD1), human EGR1 residues 335–423, human RELA residues 20–290, human GR residues 418–517, human STAT3 residues 128–715 and full-length *Saccharomyces cerevisiae* Cbf1 were expressed and purified as described previously[21,57–60]. Full-length human p53, TBP, CTCF, CREB1, CREM and ATF1 were obtained commercially (Supplementary Methods). For X-ray crystallography, the *Arabidopsis thaliana* TBP DNA-binding domain was produced as described previously[61]. For ETS1 fluorescence anisotropy binding assays, mouse ETS1 (residues 280–440) was produced as described previously[62]. For EMSA binding-affinity measurements, the human GR DNA-binding domain (residues 418–506) and human p53 (residues 94–360) were expressed and used as described previously[63–65].

SaMBA library design and measurements

SaMBA was performed as follows. Five custom DNA libraries (v.1–v.5), each containing around 15,000 single-stranded 60-base oligonucleotides, were designed computationally based on TF binding site sequences for 22 TFs (Supplementary Table 1a, e–j). The binding sites for each TF were selected on the basis of published data showing specific TF binding to these sites. Sites were selected to contain central 8-mers with protein-binding microarray (PBM) enrichment scores (E-score) of 0.35 or higher, which is indicative of specific protein binding[20,21]. For CTCF, p53 and RELA we selected strong binding sites on the basis of their DNA-binding motifs reported in the literature (CTCF[66], p53[22], RELA[67]). For GR, we used two identical half-sites of an idealized glucocorticoid response element with the preferred 3-base-pair spacer, as described and used previously[68].

Each DNA library was designed to contain multiple replicates (8–20) of both wild-type binding site sequences and all possible single-base variants of the same sites. For SaMBA library v.1, we used 14 replicate spots for each wild-type sequence and 8 replicate spots for each

mismatch. For SaMBA libraries v.2, v.3, v.4 and v.5, we used 20 replicate spots for each wild-type sequence and 10 replicate spots for each mismatch. The DNA libraries were commercially synthesized on DNA microarray chips (Agilent). Next, double-stranded DNA-binding sites were generated on the chip by hybridization with the wild-type reverse complement oligonucleotides in solution (variant complements were absent from the hybridization solution). For each wild-type sequence, the solution contained around 2.5 μM (large excess) unlabelled oligonucleotides purified by high-performance liquid chromatography (HPLC) and around 0.25 μM FAM/Cy3-labelled HPLC-purified oligonucleotides (Integrated DNA Technologies). For the variant sequences on the chip, the absence of perfect complements in solution ensured successful hybridization with the wild-type complements. The small fraction of fluorescently labelled oligonucleotides allowed us to assess the successful formation of mismatched duplexes on the chip (Extended Data Fig. 3a–d, Supplementary Table 10).

The reaction buffer mixture for the hybridization step was 100 μl 10× reaction buffer (260 mM Tris-HCl, pH 9.5, 65 mM $MgCl_2$) in a total volume of 1,000 μl, similarly to a previous study[20]. The chip was incubated with reaction mixture in a hybridization oven using a pre-warmed stainless-steel chamber and gasket cover slip. After a 5-h incubation (85 °C for 10 min, 75 °C for 10 min, 65 °C for 60 min, 60 °C for 120 min and 55 °C for 100 min), the hybridization chamber was disassembled in a glass staining dish in 500 ml phosphate buffered saline (PBS)/0.01% Triton X-100 at 37 °C. The chip was transferred to a second staining dish, washed for 10 min in PBS/0.01% Triton X-100 at 37 °C and washed once more for 3 min in PBS at room temperature, similarly to a previous study[20]. The fluorescent signal (Cy3/FAM) of hybridized oligonucleotides was measured using a GenePix 4400A microarray scanner to confirm that the hybridization was successful and reproducible, and that no detectable cross-hybridization occurred (Extended Data Fig. 3b, Supplementary Table 10).

Protein binding and antibody steps were performed similarly to PBM assays[20] (Supplementary Methods). The fluorescent signal of bound TF for each DNA spot was measured using a GenePix 4400A microarray scanner and the GenePix Pro 7.0 software. Multiple replicates of each sequence were used to quantitatively compare the binding signals between sequences and to statistically assess the significance of binding differences using a one-sided Wilcoxon-Mann–Whitney test, corrected for multiple hypotheses testing using the Benjamini–Hochberg procedure. SaMBA profiles (for example, Fig. 2b) representing the effect on TF binding for each possible mismatch along each parent sequence were produced by calculating the $\log_2$-transformed ratio between each mismatch and its corresponding wild-type parent sequence (Supplementary Table 1b). As the magnitudes of these ratios vary widely between proteins, for each parent site all ratios were also divided by the ratio of the largest decrease at the same site and multiplied by −1, so that the largest decrease for each parent sequence became −1 (Fig. 2a).

Validation and calibration of SaMBA data using measurements of TF binding affinity

DNA-binding affinity measurements for p53 were performed using EMSA, as described previously[64,69] (Supplementary Methods). The macroscopic dissociation binding constants for the dominant p53 tetrameric species were computed for ten different hairpin duplexes: four Watson–Crick and six containing mismatches (Supplementary Table 3). Six replicate measurements were performed for each duplex, and the average binding affinities were used in comparisons with SaMBA data (Fig. 1f, Extended Data Fig. 3e).

Binding affinity measurements for ETS1 (residues 280–440, termed ETS1(ΔN280)) were performed using steady-state fluorescence polarization, as described previously[70], using a Cy3-labelled DNA probe encoding the ETS1-binding sequence 5′-CGCACCGGATATCGCA-3′. In brief, 0.5 nM of DNA probe and 10 nM ETS1(ΔN280) were co-titrated with one of five unlabelled DNA duplexes: two Watson–Crick and three containing a mismatch (Supplementary Table 3). Triplicate measurements were performed for each duplex. The data confirmed both increased and decreased ETS1 binding owing to mismatches, as revealed by SaMBA (Fig. 1f, Extended Data Fig. 3e).

Binding affinity measurements for GR were performed using EMSA, as described previously[63] (Supplementary Methods). One Watson–Crick and three mismatched sites were tested (Supplementary Table 3). To avoid self-hybridization of the probes in EMSA, one of the two GR half-sites and the spacer between them were mutated compared to the SaMBA site. Positions known to be critical for GR binding were kept constant. Measurements were performed in triplicate, and the average binding affinities were used in comparisons with SaMBA data (Fig. 1f, Extended Data Fig. 3e).

The measurements described above were used both to validate TF binding increases and decreases due to mismatches, and to calibrate SaMBA data. To calibrate SaMBA data for additional TFs, we leveraged publicly available binding affinity data for Watson–Crick sequences by using a modified SaMBA protocol to test, for each TF of interest, multiple Watson–Crick sites with available affinity measurements (in addition to the wild-type and mismatched binding sites tested in a typical SaMBA assay). In our modified protocol, 60-mer DNA probes were designed to form hairpin duplexes with and without mismatches, and binding measurements were performed similarly to regular SaMBA assays (Supplementary Table 4). The following TF binding affinity datasets were used: surface plasmon resonance (SPR) data for CBF1[71], mechanically induced trapping of molecular interactions (MITOMI) data for CBF1 and MAX[72], fluorescence anisotropy (FA) data for p53[22], k-MITOMI data for EGR1[73], and EMSA data for TBP (from sites with consistent measurements in previous reports[74,75]).

Calibration of SaMBA data into free energy terms was performed as shown in Extended Data Fig. 3g, h, on the basis of the correlation between the EMSA, FA, SPR or MITOMI-derived affinities and the logarithm of the binding signal obtained in SaMBA (Supplementary Table 3). DNA libraries used for calibration also included all possible mismatches and mutations over a small number of DNA sites: two binding sites for ETS1, MAX and TBP, one binding site for CBF1, EGR1, p53 and GR, and two non-specific sites for ETS1 (Supplementary Table 4). The data for these 12 sites were used to directly compare the effects of mutations versus mismatches (Extended Data Fig. 4 and related text). When comparing the effect of base-pair mutations versus the sum of the effects of the corresponding one-base mismatch variants (Extended Data Fig. 4c and related text), the significance of the difference between these quantities was assessed using a two-sided *t*-test with Benjamini–Hochberg correction for multiple hypotheses testing; significant differences were called at a cut-off of 0.05 for the corrected *P* value. The effect of mutations on TF binding was also measured using the standard PBM protocol[20] (Supplementary Table 1). Consistent with the results obtained using the SaMBA libraries, the PBM libraries show that mutations have different effects on TF binding compared to mismatches (Fig. 2d, Supplementary Table 1). For all analyses presented here, proteins p53, ETS1 and GR were calibrated using new binding measurements for mismatched and Watson–Crick DNA sites, whereas CBF1, MAX, TBP and EGR1 were calibrated using data for Watson–Crick binding sites available in the literature (Supplementary Table 3).

ETS1 non-specific binding analysis

Owing to the high density of the DNA chips used in our experiments, each SaMBA DNA library can accommodate binding sites for several TFs (Supplementary Table 1). Thus, each TF was tested not only against its specific binding site(s), but also a small number of non-specific sites, which were specific to other TFs (Supplementary Table 1c, d). For all proteins examined, the introduction of mismatches increased binding even at non-specific DNA sites (Supplementary Table 1d) and, notably, in some cases the new binding levels were similar to those observed

for specific binding sites, thus effectively creating novel binding sites within non-specific DNA. To further test the significance and the magnitude of such increases, a new DNA sequence library was designed to measure the effects of mismatches that enhanced ETS1 binding at sites that were not originally designed for ETS1 (that is, sites that were specific to other TFs). This new library (Supplementary Table 2) contained positive and negative control groups of 'specific' and 'non-specific' sites, respectively, to enable accurate assessment of the relative binding strength of each of the sites of interest. The negative control group was composed of a set of 1,000 random DNA sequences. As specific sites can randomly appear among these sequences, we defined the non-specific binding affinity range by excluding the top 1% of the strongest bound sequences in this group. The positive control sequences were selected from crystal and NMR structures of ETS1–DNA complexes in which the ETS1 was shown to specifically bind the ETS-binding core GGA(A/T) (PDB codes: 2NNY, 2STT, 3MFK, 3RI4). Figure 2c shows five representative examples in which mismatches introduced in a non-specific site (that is, a site with binding affinity below the 99th percentile of random sites) increases the affinity to reach the specific range (that is, the range observed for sites with ETS1-bound crystal or NMR structures). The full dataset is available in Supplementary Table 2.

NMR experiments

We prepared A_6-DNA duplexes containing A-T, m^1A-T, T-T and C-T base pairs. The m^1A-containing single strand was purchased from Yale Keck Oligonucleotide Synthesis Facility with HPLC purification. All unmodified single strands were purchased from IDT with standard desalting purification. Concentrations were measured using a Nanodrop 3000, with the extinction coefficients for single and double strands obtained using the ADT bio oligo calculator. After resuspension in water, equimolar amounts of single strands were mixed together to form the duplexes. The duplexes were annealed by heating to 95 °C for 5 min and cooling at room temperature for around 1 h. They were then exchanged into NMR buffer (15 mM sodium phosphate, 25 mM sodium chloride, 0.1 mM EDTA, pH 6.9) using centrifugal concentrators. Duplex samples containing 10% D_2O after buffer exchange were lyophilized into 100% D_2O. Assignments of the sugar resonances were performed using a combination of two-dimensional (2D) ^{1}H-^{1}H NOESY, 2D ^{1}H-^{1}H TOCSY and 2D ^{1}H-^{13}C HSQC experiments. All measured chemical shift differences are available in Supplementary Table 11.

Structural analyses of mismatches that enhance TF binding

We used existing PDB structures of TF–DNA complexes to examine whether DNA mismatches can indeed mimic distorted conformations in native TF-bound DNA, which could explain the increased binding affinity of TFs to DNA mismatches. Structures of protein–DNA complexes are available in the PDB for 15 of the 22 TFs examined by SaMBA. For 3 of the 15 proteins (GATA1, MAD and STAT3), the base-pair position(s) at which mismatches increase TF binding are different in the crystal structure sequence compared to the sequences tested in SaMBA. We thus focused our structural analyses on the remaining 12 proteins (Supplementary Table 5). When multiple structures were available for the same TF, we chose the one with the DNA sequence most similar to the one tested in SaMBA. For the selected structures, we focused on the regions in common between the crystal structure and the SaMBA sequence, and at each position we computed the extent to which each structural feature deviates from the B-DNA envelope (Supplementary Table 5). For each position we also computed the largest deviation observed across all structural parameters, as well as the number of structural features with mean values more than one standard deviation above the mean observed for naked B-DNA (Supplementary Table 8a). We applied Mann–Whitney *U* tests on these summary statistics to ask whether the positions with mismatch-enhanced binding are more distorted than the other positions in TF-binding sites ($P = 0.017$ for the largest deviation; $P = 0.015$ for the fraction of distorted features).

Next, focusing on the regions that were identical between the crystal structure and the SaMBA sequence (underlined in Supplementary Table 6a), we identified 23 positions at which we found increased TF binding, owing to a total of 32 mismatches (for some positions we found several mismatches that lead to increased levels of TF binding). For these 23 positions, we comprehensively annotated all local and global distortions of DNA, defined as deviations in a structural parameter that are greater than one standard deviation above the mean of that parameter in free B-DNA structures (Supplementary Table 6b). Next, we examined the mismatch structures to determine whether the mismatches are inducing structural features that mimic bound geometries. Owing to the lack of available PDB structures of DNA mismatches embedded in Watson–Crick contexts, we systematically performed MD simulations of free DNA containing each mismatch. Similarly to our analyses of distortions in protein-bound DNA, we identified the local and global distortions caused by mismatches by comparing the distributions of structural parameters for mismatched DNA versus Watson–Crick DNA, according to the MD simulations (Supplementary Table 6c). By intersecting the lists of distortions identified in mismatched DNA versus TF-bound DNA, we identified all candidate features that are potentially mimicked by the mismatches that increased TF binding. We found such candidate features for 21 of the 32 mismatches (66%) (Supplementary Table 6d).

Crystallization and determination of the structure of TBP–mismatch DNA complexes

TBP–DNA complexes were prepared and used for vapour diffusion crystallization screens (Supplementary Methods), resulting in large, well-diffracting crystals suitable for data collection after optimization of initial hits. Data for all the crystals were collected at the Advanced Light Source (ALS) on beamlines 8.3.1 and 5.0.1. The data were processed with MOSFLM and scaled with SCALA[76,77]. The structures were solved by molecular replacement (with MolRep) using a previous structure of TBP (PDB 1QNE) with the water molecules removed, as a search model. After refinement in PHENIX[78], the structures were manually rebuilt in O[79]. MolProbity was used to guide the process of refitting and refinement[80]. See Extended Data Table 1 for the final data collection and refinement statistics for each structure.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The data that support the findings in this study are available as Supplementary Tables in Excel format. Coordinates and structure factor amplitudes for the TBP-AC, TBP-CC(1a), TBP-CC(1b) and TBP-CC(2) structures have been deposited in the PDB under the accession codes 6UEO, 6UEP, 6UER and 6UEQ, respectively. The raw SaMBA data have been deposited in the Gene Expression Omnibus (GEO) under accession number GSE156375. The PDB entries used in this study are available in Extended Data Figs. 1, 2, 5, 7 and Supplementary Tables 5–7, 9. High-resolution gel images for the EMSA data are available at https://figshare.com/projects/DNA_mismatches_reveal_conformational_penalties_in_protein-DNA_recognition/83663.

Code availability

The code used for the structural analyses presented in this study is available in GitHub at https://github.com/alhashimilab/TF_MM.

34. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
35. Zhou, H. et al. New insights into Hoogsteen base pairs in DNA duplexes from a structure-based survey. *Nucleic Acids Res.* **43**, 3420–3433 (2015).

36. Lu, X.-J., Bussemaker, H. J. & Olson, W. K. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.* **43**, e142 (2015).
37. Sathyamoorthy, B. et al. Insights into Watson–Crick/Hoogsteen breathing dynamics and damage repair from the solution structure and dynamic ensemble of DNA duplexes containing m^{1}A. *Nucleic Acids Res.* **45**, 5586–5601 (2017).
38. El Hassan, M. A. & Calladine, C. R. Two distinct modes of protein-induced bending in DNA. *J. Mol. Biol.* **282**, 331–343 (1998).
39. Bailor, M. H., Mustoe, A. M., Brooks, C. L., III & Al-Hashimi, H. M. 3D maps of RNA interhelical junctions. *Nat. Protocols* **6**, 1536–1545 (2011).
40. Bailor, M. H., Sun, X. & Al-Hashimi, H. M. Topology links RNA secondary structure with global conformation, dynamics, and adaptation. *Science* **327**, 202–206 (2010).
41. Le Novère, N. MELTING, computing the melting temperature of nucleic acid duplex. *Bioinformatics* **17**, 1226–1227 (2001).
42. Cheatham, T. E. III, Cieplak, P. & Kollman, P. A. A modified version of the Cornell *et al.* force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.* **16**, 845–862 (1999).
43. Pérez, A., Luque, F. J. & Orozco, M. Dynamics of B-DNA on the microsecond time scale. *J. Am. Chem. Soc.* **129**, 14739–14745 (2007).
44. Maier, J. A. et al. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
45. Salomon-Ferrer, R., Götz, A. W., Poole, D., Le Grand, S. & Walker, R. C. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. explicit solvent particle mesh Ewald. *J. Chem. Theory Comput.* **9**, 3878–3888 (2013).
46. Rossetti, G. et al. The structural impact of DNA mismatches. *Nucleic Acids Res.* **43**, 4309–4321 (2015).
47. Arnold, F. H., Wolk, S., Cruz, P. & Tinoco, I. Jr. Structure, dynamics, and thermodynamics of mismatched DNA oligonucleotide duplexes d(CCCAGGG)2 and d(CCCTGGG)2. *Biochemistry* **26**, 4068–4075 (1987).
48. Kouchakdjian, M., Li, B. F., Swann, P. F. & Patel, D. J. Pyrimidine.pyrimidine base-pair mismatches in DNA. A nuclear magnetic resonance study of T.T pairing at neutral pH and C.C pairing at acidic pH in dodecanucleotide duplexes. *J. Mol. Biol.* **202**, 139–155 (1988).
49. Boulard, Y. et al. The pH dependent configurations of the C.A mispair in DNA. *Nucleic Acids Res.* **20**, 1933–1941 (1992).
50. Peng, Y. & Alexov, E. Computational investigation of proton transfer, pKa shifts and pH-optimum of protein-DNA and protein-RNA complexes. *Proteins* **85**, 282–295 (2017).
51. Chen, W., Morrow, B. H., Shi, C. & Shen, J. K. Recent development and application of constant pH molecular dynamics. *Mol. Simul.* **40**, 830–838 (2014).
52. Rangadurai, A. et al. Why are Hoogsteen base pairs energetically disfavored in A-RNA compared to B-DNA? *Nucleic Acids Res.* **46**, 11099–11114 (2018).
53. Patel, D. J., Kozlowski, S. A., Ikuta, S. & Itakura, K. Deoxyguanosine-deoxyadenosine pairing in the d(C-G-A-G-A-A-T-T-C-G-C-G) duplex: conformation and dynamics at and adjacent to the dG X dA mismatch site. *Biochemistry* **23**, 3207–3217 (1984).
54. Webster, G. D. et al. Crystal structure and sequence-dependent conformation of the A.G mispaired oligonucleotide d(CGCAAGCTGGCG). *Proc. Natl Acad. Sci. USA* **87**, 6693–6697 (1990).
55. Allawi, H. T. & SantaLucia, J., Jr. NMR solution structure of a DNA dodecamer containing single G.T mismatches. *Nucleic Acids Res.* **26**, 4925–4934 (1998).
56. Boulard, Y., Cognet, J. A. & Fazakerley, G. V. Solution structure as a function of pH of two central mismatches, C. T and C. C, in the 29 to 39 K-*ras* gene sequence, by nuclear magnetic resonance and molecular dynamics. *J. Mol. Biol.* **268**, 331–347 (1997).
57. Gordân, R. et al. Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.* **3**, 1093–1104 (2013).
58. Frank, F., Okafor, C. D. & Ortlund, E. A. The first crystal structure of a DNA-free nuclear receptor DNA binding domain sheds light on DNA-driven allostery in the glucocorticoid receptor. *Sci. Rep.* **8**, 13497 (2018).
59. Takayama, Y., Sahu, D. & Iwahara, J. NMR studies of translocation of the Zif268 protein between its target DNA Sites. *Biochemistry* **49**, 7998–8005 (2010).
60. Belo, Y. et al. Unexpected implications of STAT3 acetylation revealed by genetic encoding of acetyl-lysine. *Biochim. Biophys. Acta* **1863**, 1343–1350 (2019).
61. Stelling, A. L. et al. Infrared spectroscopic observation of a G-C$^+$ Hoogsteen base pair in the DNA:TATA-box binding protein complex under solution conditions. *Angew. Chem. Int. Edn Engl.* **58**, 12010–12013 (2019).
62. Stephens, D. C. & Poon, G. M. Differential sensitivity to methylated DNA by ETS-family transcription factors is intrinsically encoded in their DNA-binding domains. *Nucleic Acids Res.* **44**, 8671–8681 (2016).
63. Zhang, L. et al. SelexGLM differentiates androgen and glucocorticoid receptor DNA-binding preference over an extended binding site. *Genome Res.* **28**, 111–121 (2018).
64. Vyas, P. et al. Diverse p53/DNA binding modes expand the repertoire of p53 response elements. *Proc. Natl Acad. Sci. USA* **114**, 10624–10629 (2017).
65. Weinberg, R. L., Veprintsev, D. B. & Fersht, A. R. Cooperative binding of tetrameric p53 to DNA. *J. Mol. Biol.* **341**, 1145–1159 (2004).
66. Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32**, D91–D94 (2004).
67. Siggers, T. et al. Principles of dimer-specific gene regulation revealed by a comprehensive characterization of NF-κB family DNA binding. *Nat. Immunol.* **13**, 95–102 (2012).
68. Luisi, B. F. et al. Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature* **352**, 497–505 (1991).
69. Beno, I., Rosenthal, K., Levitine, M., Shaulov, L. & Haran, T. E. Sequence-dependent cooperative binding of p53 to DNA targets and its relationship to the structural properties of the DNA targets. *Nucleic Acids Res.* **39**, 1919–1932 (2011).
70. Stephens, D. C. et al. Pharmacologic efficacy of PU.1 inhibition by heterocyclic dications: a mechanistic analysis. *Nucleic Acids Res.* **44**, 4005–4013 (2016).
71. Siggers, T., Duyzend, M. H., Reddy, J., Khan, S. & Bulyk, M. L. Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Mol. Syst. Biol.* **7**, 555 (2011).
72. Maerkl, S. J. & Quake, S. R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**, 233–237 (2007).
73. Geertz, M., Shore, D. & Maerkl, S. J. Massively parallel measurements of molecular interaction kinetics on a microfluidic platform. *Proc. Natl Acad. Sci. USA* **109**, 16540–16545 (2012).
74. Drachkova, I. et al. Effect of TATA box polymorphisms in human β-globin gene promoter associated with β-thalassemia on interaction with TATA-binding protein. *Russ. J. Genet. Appl. Res.* **1**, 183–188 (2011).
75. Drachkova, I. et al. The mechanism by which TATA-box polymorphisms associated with human hereditary diseases influence interactions with the TATA-binding protein. *Hum. Mutat.* **35**, 601–608 (2014).
76. Leslie, A. G. The integration of macromolecular diffraction data. *Acta Crystallogr. D* **62**, 48–57 (2006).
77. Potterton, E., Briggs, P., Turkenburg, M. & Dodson, E. A graphical user interface to the CCP4 program suite. *Acta Crystallogr. D* **59**, 1131–1137 (2003).
78. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
79. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
80. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
81. Yang, S., Salmon, L. & Al-Hashimi, H. M. Measuring similarity between dynamic ensembles of biomolecules. *Nat. Methods* **11**, 552–554 (2014).
82. Hombauer, H., Srivatsan, A., Putnam, C. D. & Kolodner, R. D. Mismatch repair, but not heteroduplex rejection, is temporally coupled to DNA replication. *Science* **334**, 1713–1716 (2011).
83. Krokan, H. E., Drabløs, F. & Slupphaug, G. Uracil in DNA—occurrence, consequences and repair. *Oncogene* **21**, 8935–8948 (2002).
84. Shen, J. C., Rideout, W. M., III & Jones, P. A. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res.* **22**, 972–976 (1994).

Acknowledgements We dedicate this paper to the memory of Dr Rosalind E. Franklin, on the occasion of her 100th birthday anniversary. Dr Franklin's legacy, including her crucial contribution to the discovery of the molecular structure of DNA, continues to inspire generations of diverse scientists around the world. We thank S. Adar for discussions that initiated this project; D. Herschlag for discussions and comments; E. Arbely, D. Golovenko, J. Iwahara, E. Ortlund and R. Young for providing recombinant purified protein; and L. McIntosh for providing expression plasmids. This work was supported by the National Institutes of Health (NIH) grants R01-GM135658 and R01-GM117106 (to R.G.) and R01-GM089846 (to H.M.A.-H.); a Duke University GCB Pilot Grant (to R.G. and H.M.A.-H.); and an Integrated DNA Technologies postdoctoral fellowship award (to A.A.). R.S. and M.A.S. were supported by NIH grant R35-GM130290 (to M.A.S.); A.S. and T.E.H. were supported by the Israel Science Foundation grant 1517/14 (to T.E.H.); S.X. and G.M.K.P. were supported by a National Science Foundation (NSF) grant MCB-2028902 (to G.M.K.P.); and M.F. and M.A.P. were supported by a NSF CAREER award MCB-1552862 (to M.A.P.) High-performance computing was partially supported by the Duke Center for Genomic and Computational Biology. We acknowledge the Advanced Light Source (ALS) at the Lawrence Berkeley National Laboratory for X-ray diffraction data collection on beamlines 8.3.1 and 5.0.1. Beamline 8.3.1 at the ALS is operated by the University of California Office of the President, Multicampus Research Programs and Initiatives grant MR-15-328599, the NIH (R01GM124149 and P30GM124169), Plexxikin and the Integrated Diffraction Analysis Technologies program of the US Department of Energy Office of Biological and Environmental Research. The Pilatus detector on beamline 5.0.1 was funded under NIH grant S10OD021832. The ALS-ENABLE beamlines are supported in part by the NIH National Institute of General Medical Sciences grant P30 GM124169. The ALS is a national user facility operated by Lawrence Berkeley National Laboratory on behalf of the US Department of Energy under contract number DE-AC02-05CH11231, Office of Basic Energy Sciences. The Berkeley Center for Structural Biology is supported in part by the Howard Hughes Medical Institute.

Author contributions A.A., H.M.A.-H. and R.G. designed and supervised the study. A.A. generated high-throughput protein–DNA binding data. A.A., H. Shi, A.R. and H. Sahay analysed the data. H. Shi and A.R. contributed NMR data. A.S., S.X., M.F., M.A.P., G.K.M.P. and T.E.H. contributed experimental data on protein–DNA binding affinities: p53 (A.S., T.E.H.), ETS1 (S.X., G.M.K.P.) and GR (M.F., M.A.P.). Z.M. contributed high-throughput protein–DNA binding data. R.S. and M.A.S. contributed X-ray crystallography data. A.A., H. Shi, A.R., H.M.A.-H. and R.G. wrote the manuscript, with input from all authors. All of the authors critically reviewed the manuscript and approved the final version.

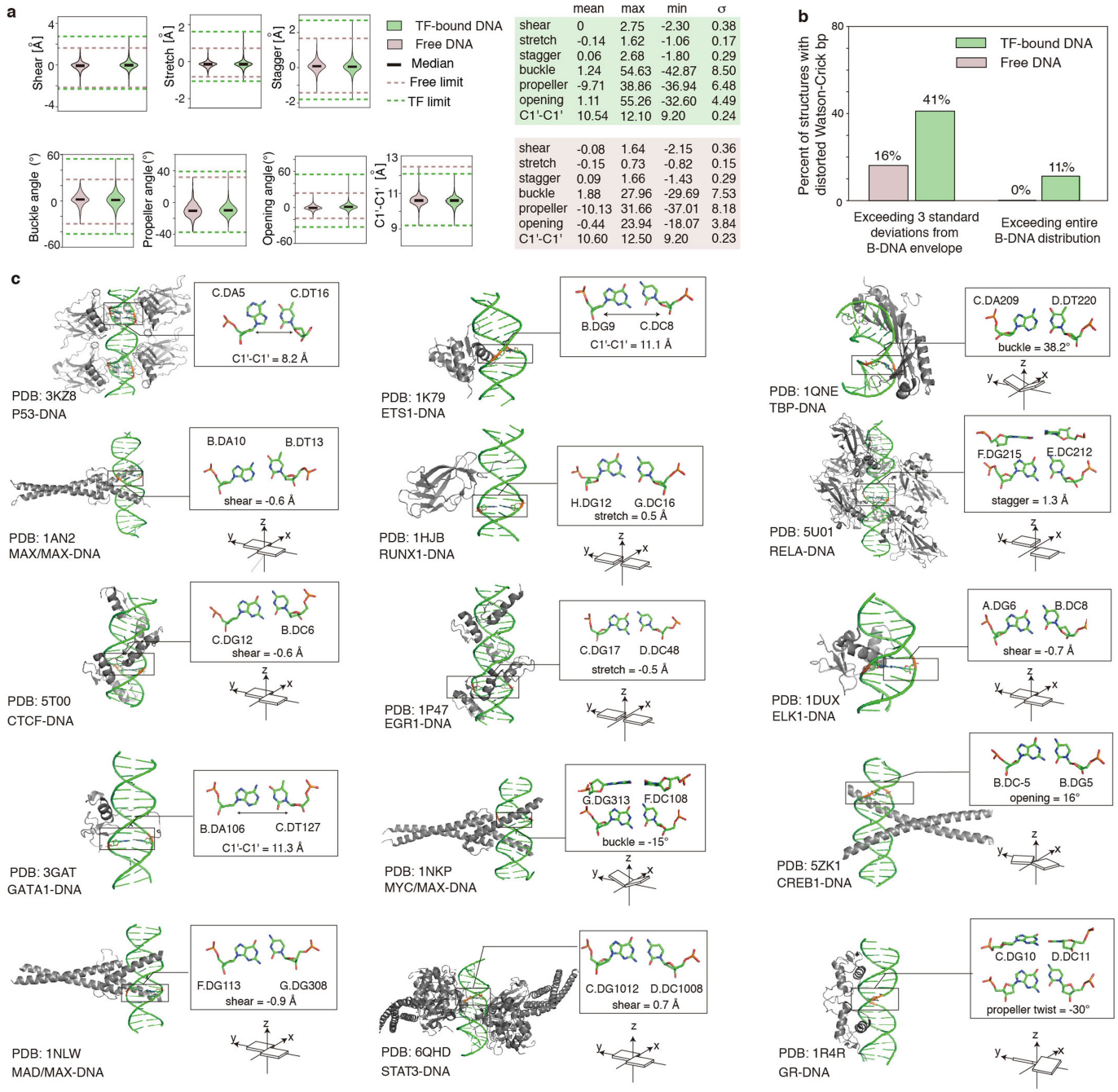
Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2843-2.

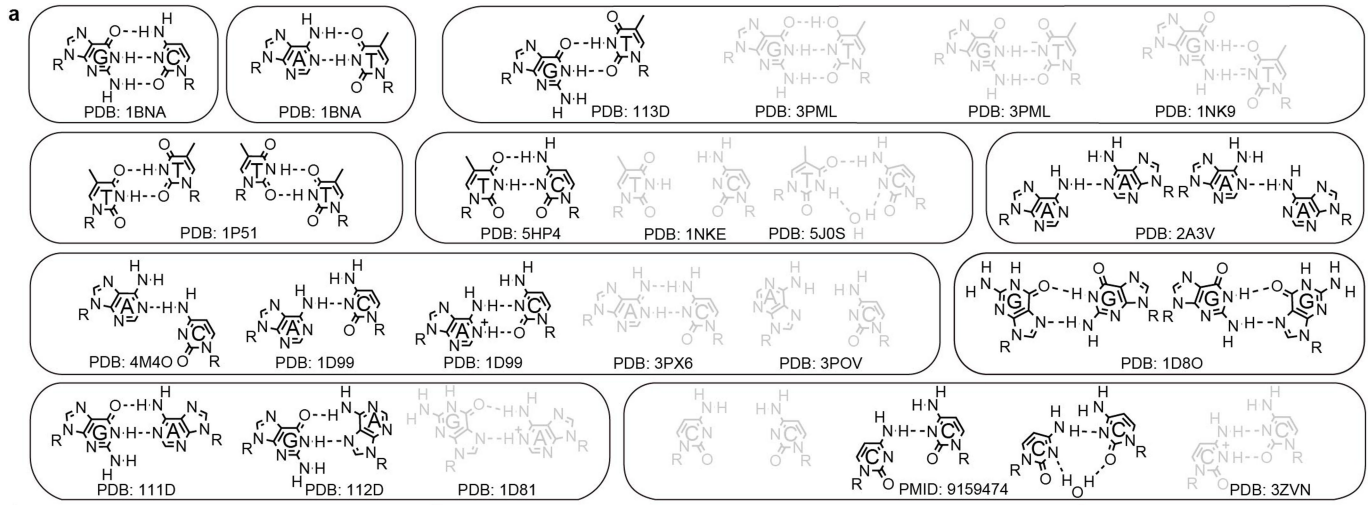
Correspondence and requests for materials should be addressed to H.M.A.-H. or R.G.

Peer review information *Nature* thanks James Fraser, Remo Rohs and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



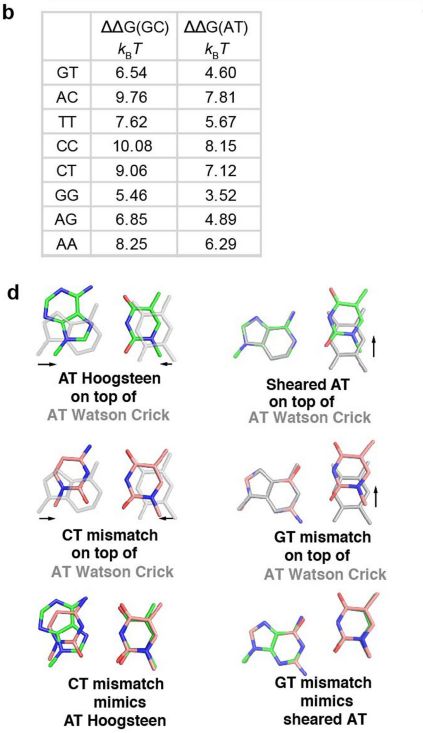
	mean	max	min	σ
shear	0	2.75	-2.30	0.38
stretch	-0.14	1.62	-1.06	0.17
stagger	0.06	2.68	-1.80	0.29
buckle	1.24	54.63	-42.87	8.50
propeller	-9.71	38.86	-36.94	6.48
opening	1.11	55.26	-32.60	4.49
C1'-C1'	10.54	12.10	9.20	0.24
shear	-0.08	1.64	-2.15	0.36
stretch	-0.15	0.73	-0.82	0.15
stagger	0.09	1.66	-1.43	0.29
buckle	1.88	27.96	-29.69	7.53
propeller	-10.13	31.66	-37.01	8.18
opening	-0.44	23.94	-18.07	3.84
C1'-C1'	10.60	12.50	9.20	0.23

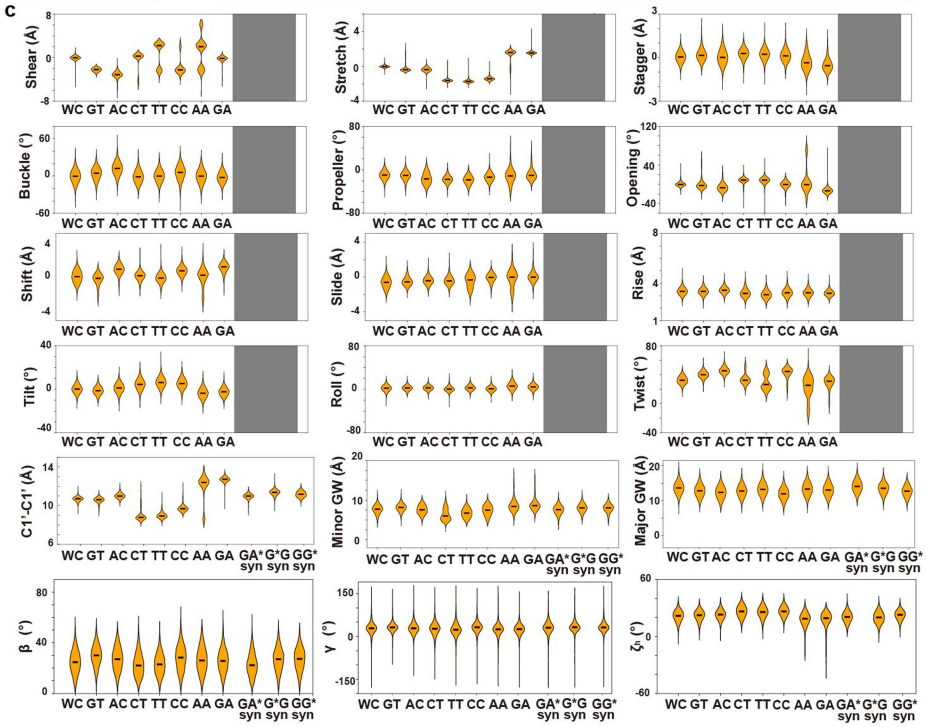
Extended Data Fig. 1 | Structural deformations in TF-bound and unbound DNA. a, Distributions of base-pair parameters in free and TF-bound DNA, from PDB[34] survey. Solid lines denote the median value of each parameter. Dashed lines denote the upper and lower bounds of the distribution for free (pink) and bound (green) DNA. 613 TF-bound structures and 409 free B-DNA structures, all with resolution < 3 Å, were used in the analysis (Methods). **b**, Percentage of structures with base pairs outside the B-DNA envelope. Among the 613 TF-bound structures, 41.1% (that is, 252) contain severe distortions of at least one base pair outside the free B-DNA envelope, with the envelope defined as at most 3 standard deviations above or below the mean. Only 16% (that is, 65) of the free B-DNA structures satisfy this criterion. (Using a less stringent definition of the B-DNA envelope, by considering two standard deviations above or below the mean, we found that 80.8% of the TF-bound structures contain at least one base pair outside the free B-DNA envelope, approximately twice the frequency observed in free DNA, which was 41.8%.) Considering the full range of base-pair parameter values as defining the free B-DNA envelope, we found that 11.3% (that is, 69) of the TF-bound structures contain at least one base pair with an extreme deformation that was never observed in any free DNA structure. **c**, Local deformations of base pairs observed in diverse TF-DNA complex structures. Left, 3D structures with the distorted base pairs highlighted in black boxes. Upper right, enlarged view of the base-pair structures with their base-pair parameters labelled. Lower right, schematic diagram of the corresponding base-pair parameters.



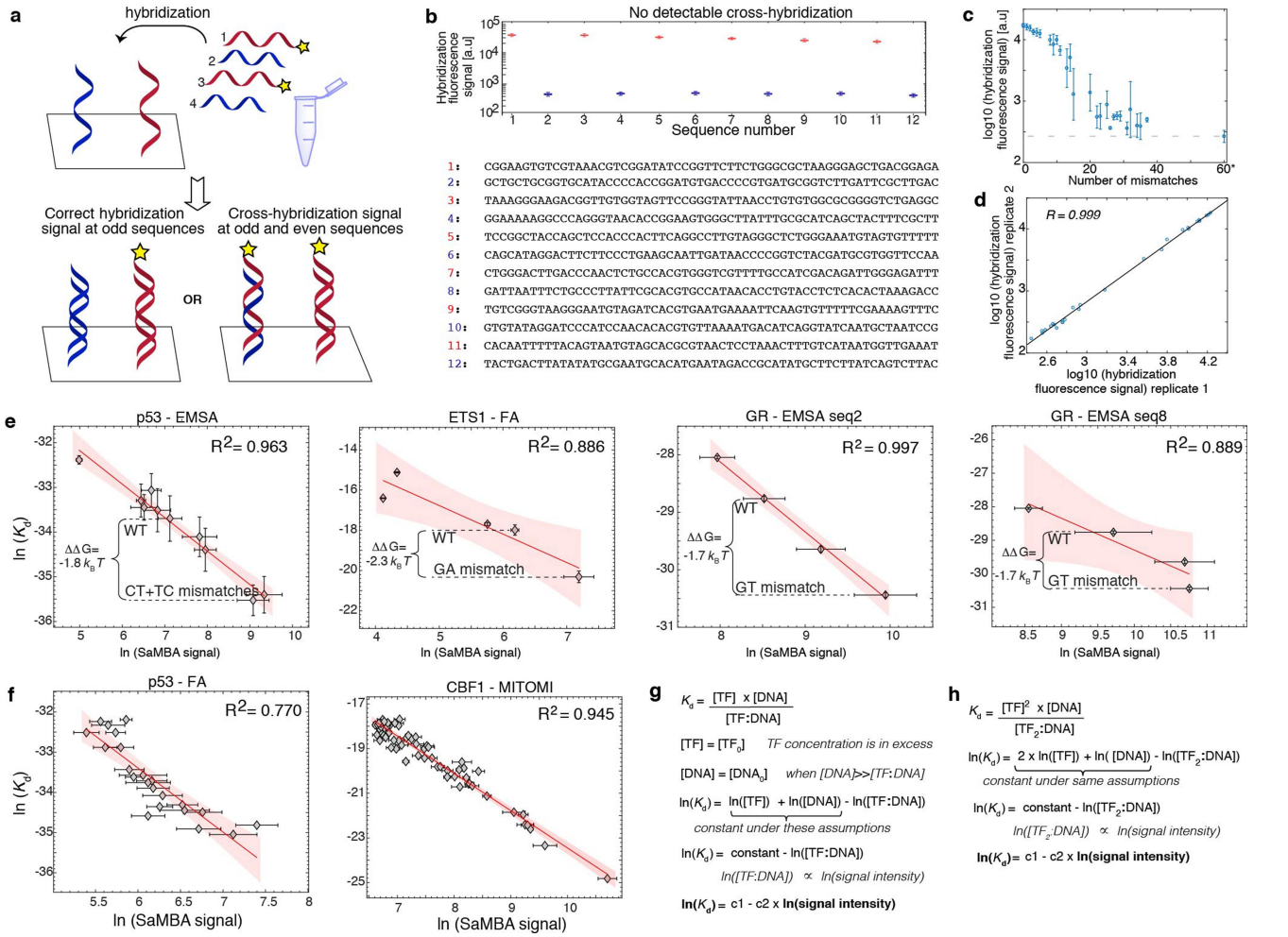
b

	ΔΔG(GC) k_BT	ΔΔG(AT) k_BT
GT	6.54	4.60
AC	9.76	7.81
TT	7.62	5.67
CC	10.08	8.15
CT	9.06	7.12
GG	5.46	3.52
AG	6.85	4.89
AA	8.25	6.29

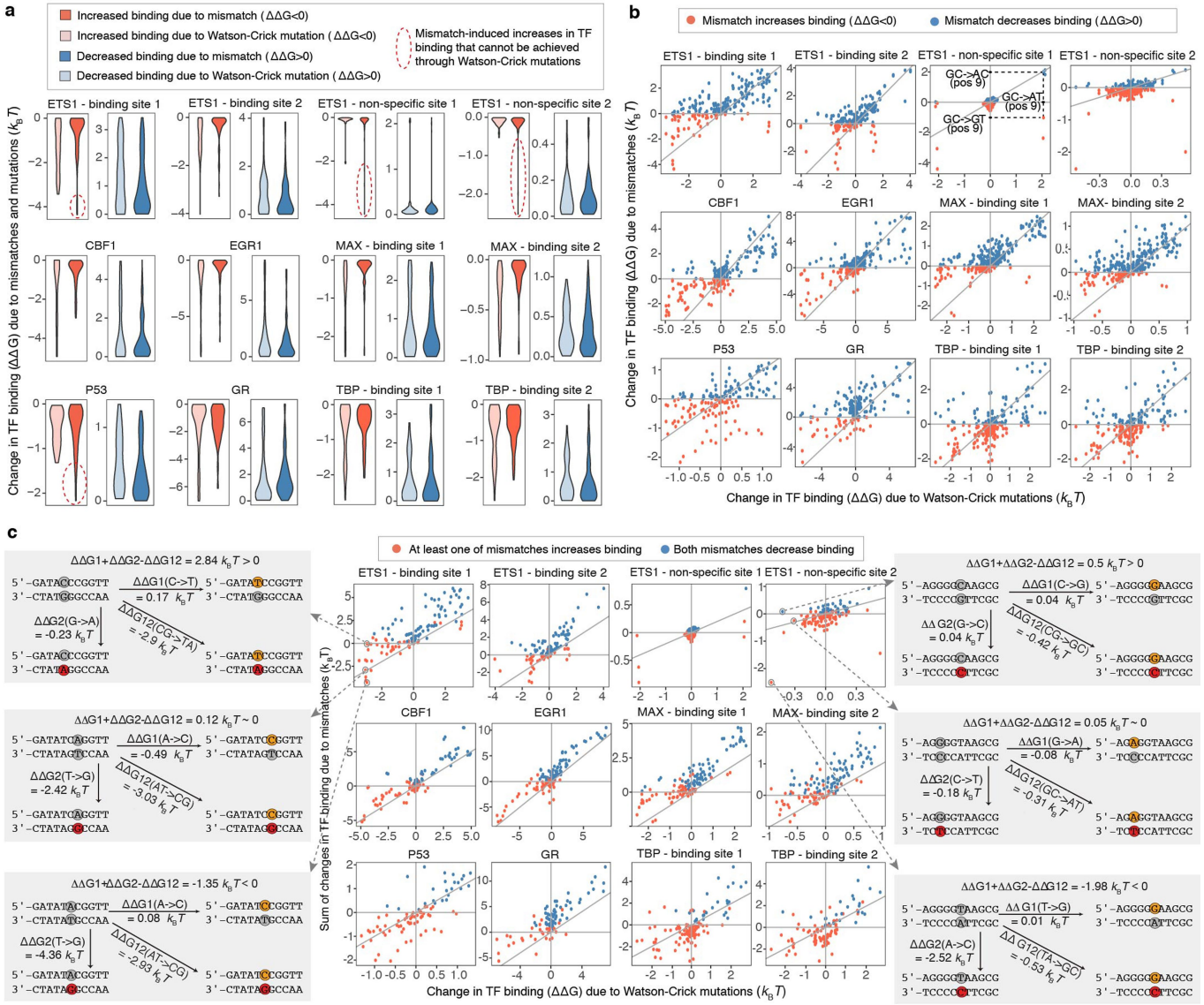




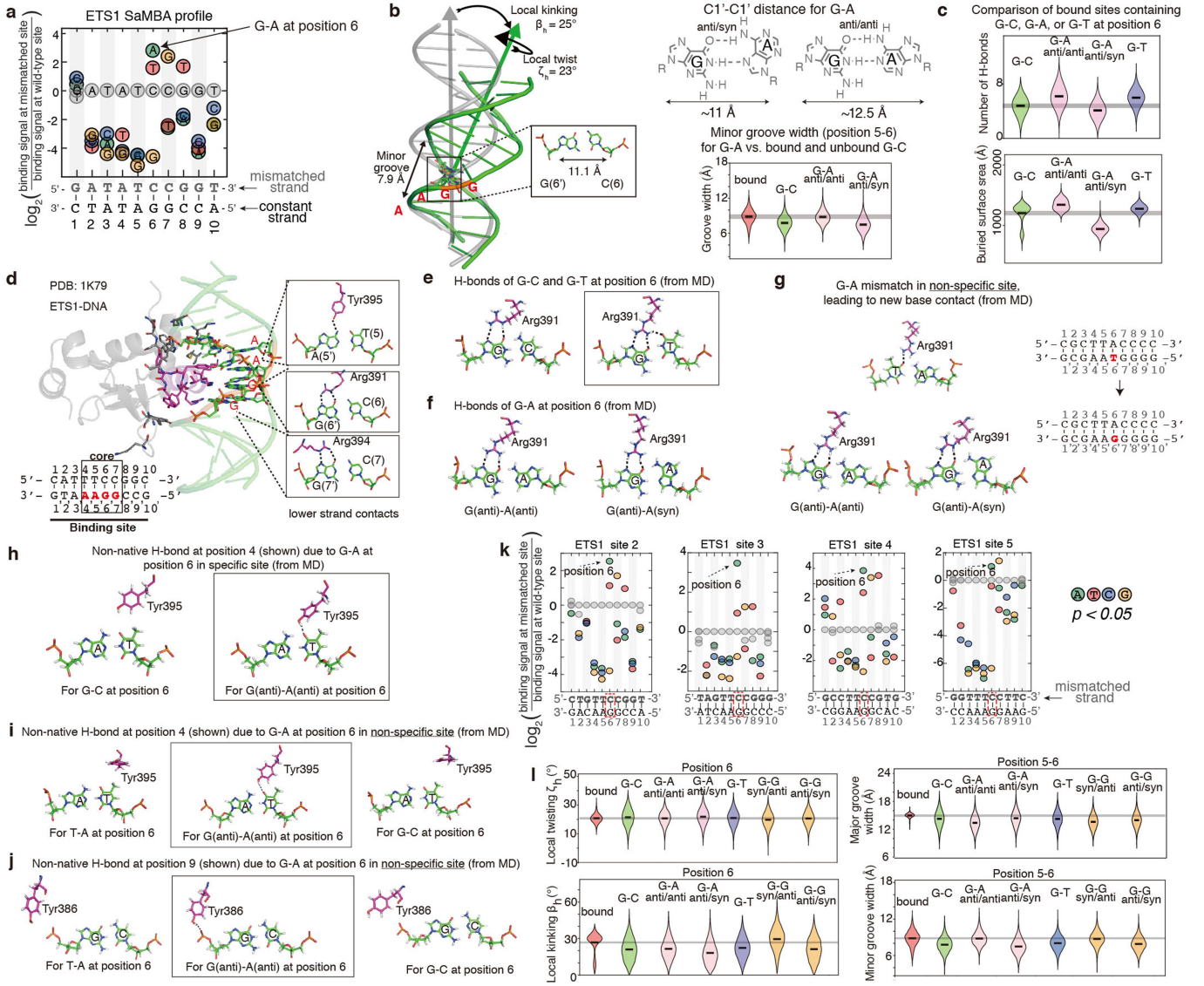
Extended Data Fig. 2 | Structural characteristics of DNA mismatches. **a**, Base-pairing geometry of Watson–Crick base pairs and mismatches, obtained from a survey of crystal structures in the PDB[34]. Mismatches with modified bases and those that were metal-mediated were excluded from analysis (Methods). Predominant base-pairing geometries under neutral pH conditions are shown in black. Minor geometries are shown in grey. **b**, Melting energies for DNA mismatches relative to G-C and A-T Watson–Crick base pairs. See Methods for details. **c**, Distributions of structural parameters in Watson–Crick and mismatched DNA, from MD simulations. Solid lines denote the median value of each parameter. Observations from the MD simulation results: (1) G-T retains wobble geometry during the MD simulation, with sheared conformation (|shear| around 2 Å) accompanied by a slight stretch. (2) T-T shows wobble geometry with sheared conformation (|shear| around 2 Å). Different from G-T, the T-T mismatch shows rapid dynamic equilibrium of both wobble geometries with either one of the Ts shifted to the minor groove direction. Despite this rapid dynamic equilibrium, the T-T base pair is still constricted with C1′–C1′ distance 8–9.5 Å. (3) Similar to T-T, the C-T mismatch is also constricted with two hydrogen bonds stably formed for most of the time. However, C-T mismatch can transiently adopt a high-energy conformation with only one hydrogen bond and is not constricted anymore (C1′–C1′ distance around 10 Å), potentially owing to the close contact between T-O2 and C-O2. The entire C-T MD trajectory is comprised of approximately 5% of these high-energy species. (4) C-C is partially constricted with C1′–C1′ distance around 9.8 Å owing to unstable hydrogen bonding. (5) All pyrimidine-pyrimidine mismatches are stacked in the helix without swing out of the helix in the MD trajectories. (6) G-G does not experience *anti-syn* equilibrium during the simulation. The C1′–C1′ distance of G-G (G(*syn*)-G(*anti*) or G(*anti*)-G(*syn*)) is around 11.2–11.5 Å, which is larger than the canonical G-C base pair. (7) G(*anti*)-A(*syn*) is not constricted (C1′–C1′ distance around 11 Å) and G(*anti*)-A(*anti*) reveals large C1′–C1′ distance around 12.8 Å. Base-pair and base-step parameters of bases with *syn* conformation (marked with *) were not computed, and are thus greyed out, owing to an ill-defined coordinate frame (Methods). The C1′–C1′ distance is shown, as it is not affected by the change of coordinate frame. **d**, Mismatches can mimic distorted base-pair geometries observed in protein-bound DNA. Overlays of distorted (coloured) and idealized Watson–Crick (grey) base pairs from 3DNA (top); mismatches (coloured) and idealized Watson–Crick (grey) base pairs (middle); and mismatched and distorted Watson–Crick base pairs (right). The mismatched conformations are of free DNA and were obtained from MD simulations (Methods). The C-T mismatch can mimic an A-T Hoogsteen base pair by constricting the C1′–C1′ distance (taken from PDB 3KZ8). The G-T mismatch can mimic a sheared A-T base pair by shifting the T to the major groove direction (taken from PDB 4MZR).



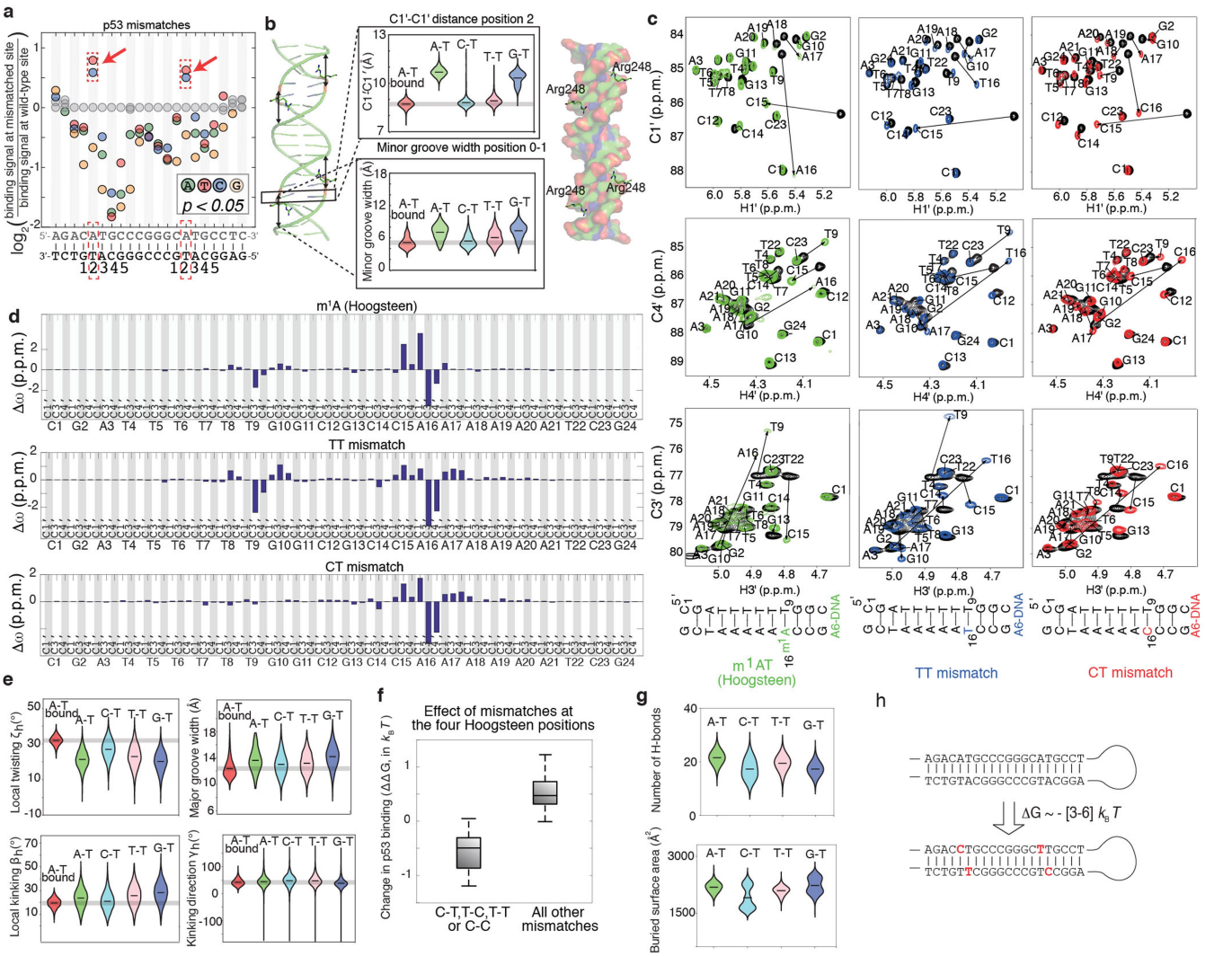
Extended Data Fig. 3 | Validation and calibration of SaMBA measurements. **a**, Schematic representation of our experimental workflow to detect cross-hybridization. To check whether certain oligonucleotides hybridize with non-target complementary oligonucleotides, we designed an experiment in which only certain oligonucleotides (red) were labelled. If significant cross-hybridization occurred, we would have detected fluorescent signal on the chip even for sequences without fluorescent complements in the hybridization solution (that is, for the sequences shown in blue). **b**, No significant cross-hybridization was detected. Bottom, list of 12 sequences used in the hybridization solution of one SaMBA experiment (red: fluorescently labelled oligonucleotides; blue: unlabelled). Top, fluorescent signal from the hybridization of these 12 sequences on the chip. For the sequences on the chip for which their complement is not labelled, the fluorescent signal is practically undetectable (blue), and it is several orders of magnitude lower than the sequences with a labelled complementary strand (red). Box plots show median signals over replicate DNA spots, with the bottom and top edges of each box indicating the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers. **c**, The effect of mismatches on hybridization. To estimate the efficiency of our hybridization protocol, we measured the hybridization signal of one specific sequence (sequence #3 for library v.1; see Methods, Supplementary Table 10), to different sequences containing multiple mismatches (0 to around 40), and a completely different sequence ('60*'). As expected, the hybridization was less efficient for sequences with large numbers of mismatches. However, for small numbers of mismatches the hybridization was highly efficient. Longer incubation time, higher oligonucleotide concentration, and normalization of the signal could enable the use of SaMBA for larger numbers of mismatches. Plot shows medians and standard deviations over all sequences containing the same number of mismatches, with 6 replicate spots per sequence. Mismatches were introduced randomly by generating N random base changes (N = 1–5, 10, 15, 25, 35, 45) to sequence #3, and repeating the procedure ten times for each N. This led to duplexes with 1 to 37 mismatches compared to the original sequence. **d**, Hybridization signal is highly reproducible. The correlation of hybridization signals between two replicate experiments was very high (R^2 = 0.99). Plot shows median values, computed over six replicate spots, based on data shown in **c**. **e**, Validation of mismatch effects by orthogonal methods. For p53, ETS1, and GR proteins, the log-transformed SaMBA binding intensities correlate with independent affinity measurements performed on mismatched and non-mismatched DNA sites (Methods). Similarly to PBM experiments, median values over all replicates were used for SaMBA (n = 10 replicate spots); error bars show the median absolute deviation. Average values over replicates were used for the orthogonal methods (n = 6 independent measurements for p53, and n = 3 independent measurements for ETS1 and GR), with error bars showing the standard deviation. Red shaded region, 95% confidence interval for Pearson's correlation. Binding free energy differences ($\Delta\Delta G$) are shown between native Watson–Crick binding sites and the highest increase in binding due to a mismatch. Two SaMBA sites were tested for GR (see Methods). **f**, Correlation between binding data obtained by SaMBA versus independent methods. For SaMBA data the plots show the median values over replicate spots (n = 10 replicate spots), with error bars showing the median absolute deviation. For independent data (Methods) the plots show the binding affinities as reported in the respective papers. Red shaded region, 95% confidence interval for Pearson's correlation. **g**, Standard equilibrium thermodynamics equations demonstrate that the logarithm of the K_d values of the TF–DNA complex is linearly proportional to the logarithm of the TF–DNA complex fluorescence signal, under certain conditions in which the TF concentration and the free DNA concentration are in excess compared to the concentration of the bound complex (and those remain constant during the reaction). **h**, Similar to **g**, for cases in which the DNA-bound species is a dimer.



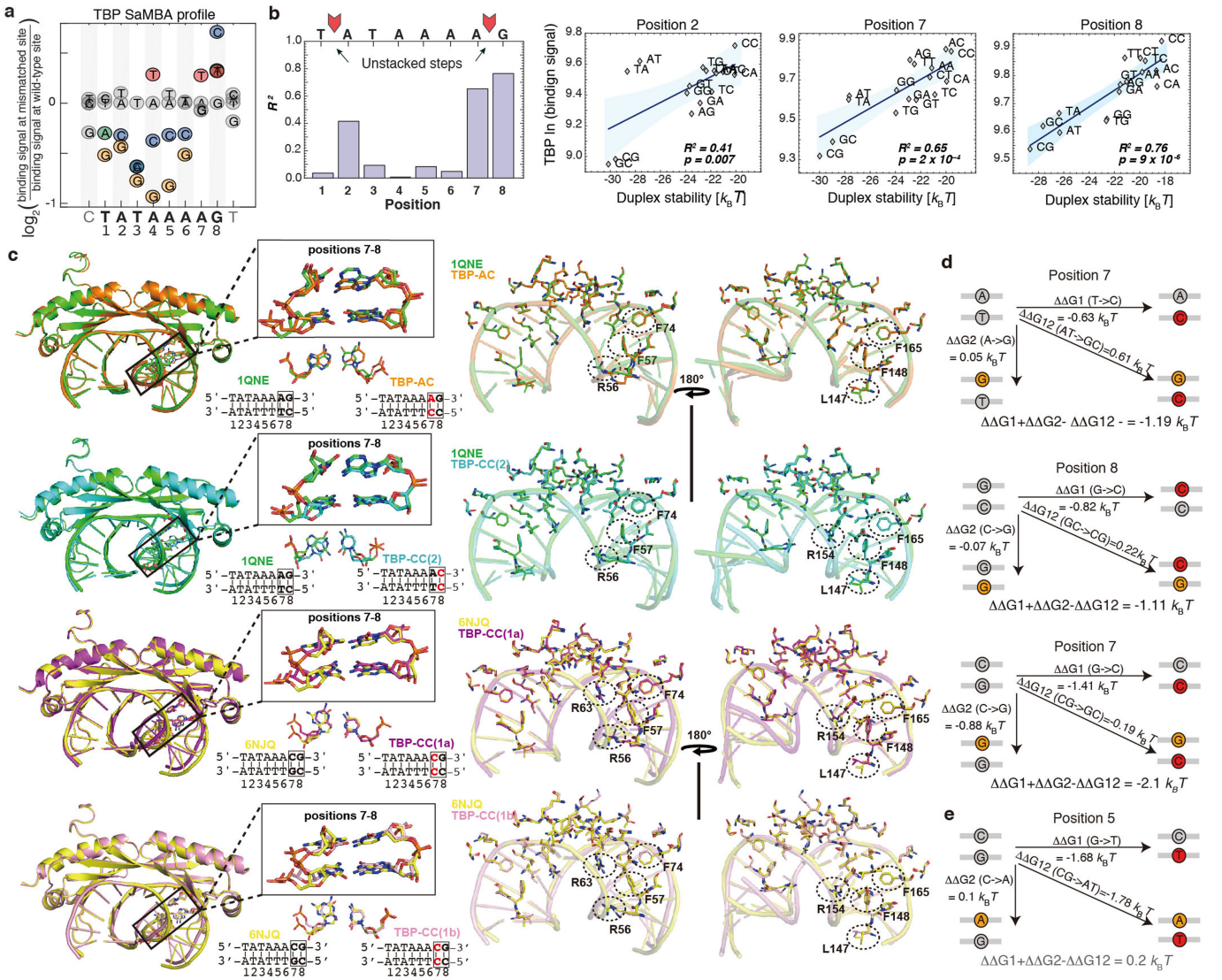
Extended Data Fig. 4 | Comparing the effects of mutations versus mismatches on TF binding. a, The magnitude of the energetic effects of mutations (light colours) and mismatches (dark colours) is similar. The effects were computed for all 7 proteins with available calibration data in our study, and for a total of 12 DNA sites (Methods). The effects of mismatches were calculated relative to the two closest Watson–Crick sequences (for example, for a G-T mismatch the closest Watson–Crick base pairs are G-C and A-T; the mismatch plots include both $\Delta\Delta G$(G-C > G-T) and $\Delta\Delta G$(A-T > G-T)). **b**, Mismatches and their corresponding mutations have different, even opposite effects on TF binding. Each mutation is compared to the two closest mismatches (for example, G-C > A-T is compared to both G-C > A-C and G-C > G-T). Top left quadrant, mutations increase binding, mismatches decrease binding. Top right quadrant, both mutations and mismatches decrease binding. Bottom left quadrant, both mutations and mismatches increase binding. Bottom right quadrant, mutations decrease binding, mismatches increase binding. The *x* axis and *y* axis show calibrated binding measurements computed from the median SaMBA signal intensities (over $n = 10$ replicate spots). **c**, Comparing the effect of mutations versus the cumulative effects of the two closest mismatches. Points close to the diagonal correspond to cases in which the effect of the mutation is approximately equal (within experimental noise) to the sum of the effects of the two mismatches. Points above the diagonal correspond to cases in which Watson–Crick mutations have either a more beneficial or a less detrimental effect on TF binding compared to the cumulative effect of the two mismatches. Points below the diagonal correspond to cases in which Watson–Crick mutations have either a less beneficial or a more detrimental effect on TF binding compared to the cumulative effect of the two mismatches. The *x* axis and *y* axis show calibrated binding measurements computed from the median SaMBA signal intensities (over $n = 10$ replicate spots). Please see Supplementary Table 4 for the raw binding data used to compute the measurements shown in this figure.



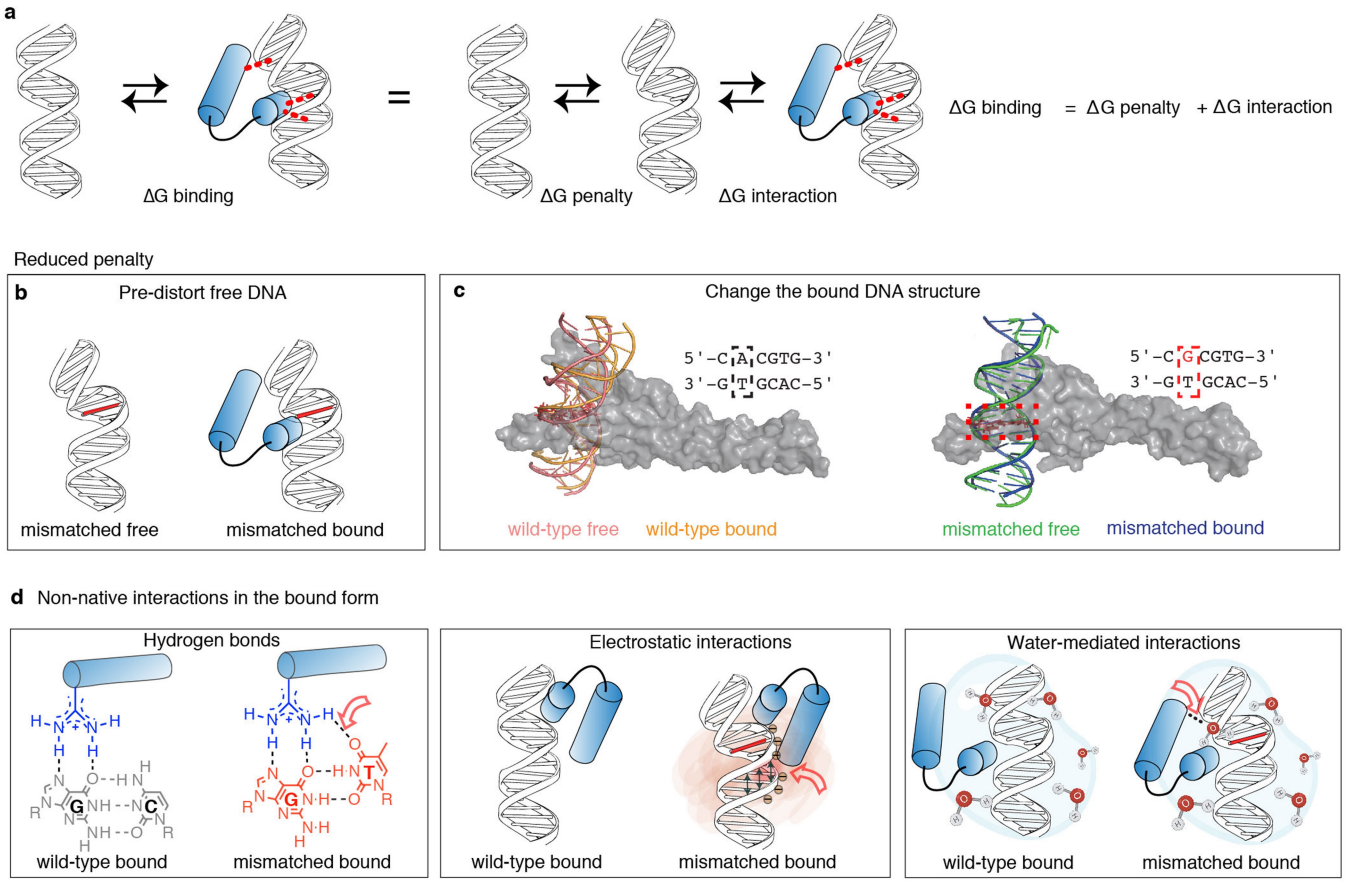
Extended Data Fig. 5 | The effects of mismatches on ETS1–DNA binding. **a**, SaMBA profile for an ETS1-binding site, highlighting the G-A mismatch at position 6, which shows the largest increase in binding affinity. **b**, Distortions. In the bound ETS1–DNA complex (PDB ID: 1K79), the positions at which the recognition helix is inserted into the DNA major groove are significantly distorted, with bending ($\beta_h = 23°$) towards the major groove, local unwinding ($\zeta_h = 23°$), and minor groove widening. Position 6, the middle position of the GGA core binding region, is highlighted to show the expanded C1′–C1′ distance. The G-A mismatch at this position mimics the C1′–C1′ distance of the bound DNA. Violin plots of the MD simulation data show that the G-A mismatch in *anti-anti* configuration also mimic the minor groove width of the bound G-C. **c**, Base readout. According to MD simulation results, G-A (*anti/anti*) and G-T mismatches increase the overall number of hydrogen bonds and the buried surface area at the ETS1-DNA interface, compared to the Watson–Crick G-C pair (Methods). **d**, ETS1–DNA interface in the GGAA core binding region. Contacting residues in the recognition helix are shown in magenta. Direct hydrogen bond contacts with the bases are highlighted; such contacts occur only at the GGA bases, on the 'lower' strand of the shown Watson–Crick DNA site. **e**, **f**, Representative snapshots of different hydrogen bond interactions between Arg391 and the base pair at position 6, from MD simulations. The G-T mismatch shows an additional hydrogen bond compared to G-C and G-A. **g**, In a non-specific site where G-A increases the affinity to reach the specific range, MD simulations show that the G-A mismatch forms hydrogen bonds similar to those formed in specific sites (shown in panel f). **h**, Non-native hydrogen bond at position 4, owing to the G-A mismatch at position 6 in the specific ETS1-binding site. **i**, **j**, Non-native hydrogen bond interactions created in a non-specific site (**g**) at positions neighbouring the positions of the mismatch, either with the base (**i**) or the backbone (**j**). **k**, SaMBA profiles for additional ETS1-binding sites. We measured the effect of mismatches in four ETS1-binding sites in addition to the one shown in **a**. Although the profiles for different sites are quantitatively different and dependent on the flanks, the trends for increased binding due to mismatches are similar. For all cases, the A-G mismatch at position 6 significantly increases ETS1 binding. **l**, Structural features at the mismatch position. Violin plots show the local twisting and kinking at position 6, and the minor and major groove width at position 5–6 of ETS1-bound DNA, as well as the naked DNA for different base pairs, according to MD.



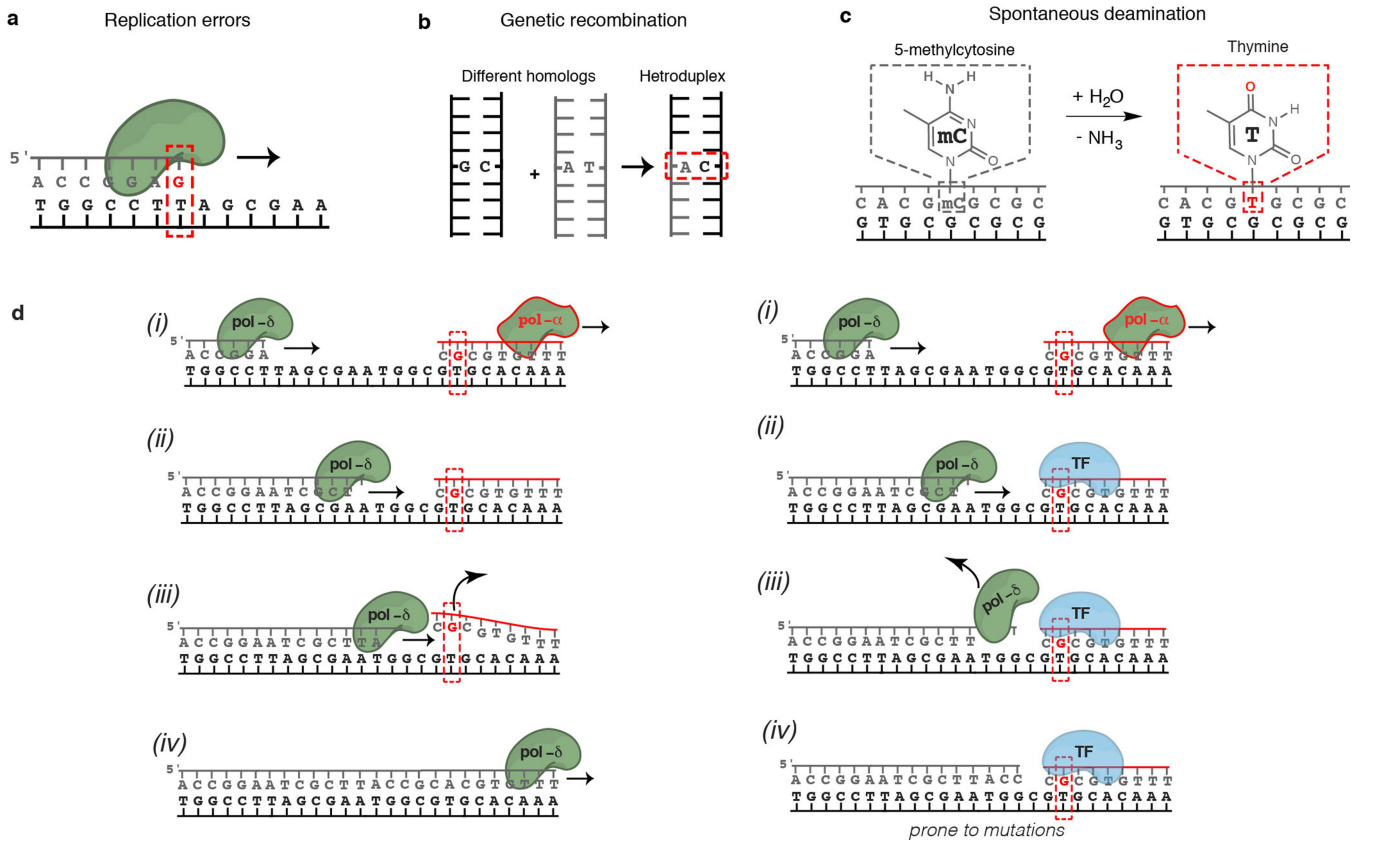
Extended Data Fig. 6 | The effects of mismatches on p53–DNA binding. **a**, Mismatch profile for p53 reveals that increased TF binding occurs only due to C-T and T-T mismatches (red rectangle) at the same positions at which the Hoogsteen conformation is observed in p53–DNA complexes (PDB 3KZ8). **b**, MD simulation-based violin plots of C1′–C1′ distance at position 2, as well as the minor grove width (at position 0–1), for p53-bound DNA and naked DNA (wild-type and mismatched) reveals that the minor groove for C-T and T-T mismatches is more similar to the bound form compared to the free A-T base pair. Plot also shows that the G-T mismatch, which reduces p53 binding, does not mimic these distortions seen in the bound DNA. Notably, a narrower minor grove at position 0–1 was previously suggested to be important for the interaction of the DNA with the Arg248 residue in p53[27]. **c**, **d**, NMR validation showing that T-T and C-T mimic the reduced C1′–C1′ distance observed in p53-bound DNA[27,28]. **c**, Chemical shift overlays of the 2D HSQC NMR spectra of the C1′–H1′, C4′–H4′ and C3′–H3′ regions for A6-DNA m^{1}A in which the m^{1}AT base pair is in the Hoogsteen conformation[30] (left, green), A6-DNA TT (middle, blue) and A6-DNA CT (right, red) with unmodified A6-DNA (black) at pH 6.9, 25 °C. **d**, Bar plots of the individual chemical shift differences (relative to unmodified A6-DNA) of the C1′, C3′ and C4′ carbon atoms of A6-DNA m^{1}A (top), A6-DNA TT (middle) and A6-DNA CT (bottom). Similarity between the Hoogsteen induced chemical shift differences and mismatch shifts (relative to the Watson–Crick wild-type) is observed for both T-T and C-T. **e**, Additional comparisons of global features (twisting angle, local kinking, and kinking direction at position 2 and major groove width at position 0–1) reveal additional mimicry between C-T mismatch and the Hoogsteen conformation local twisting angle. **f**, Pyrimidine–pyrimidine mismatches (C-T, T-C, T-T and C-C) in all four positions in which Hoogsteen conformation is observed (n = 16 mismatches total), increased p53 binding. However, all other mismatches at these positions (n = 32 mismatches total) decreased p53 binding, or had non-significant effects. $\Delta\Delta G$ represents the differences between the p53-DNA binding energy of each mismatch versus the wild-type sequence, and was estimated using the calibration with EMSA measurements (Methods). Box plots show median signals over all mismatches, with the bottom and top edges of each box indicating the 25th and 75th percentiles, respectively. The whiskers extend to the most-extreme data points that are not considered outliers. **g**, Number of p53-DNA hydrogen bonds and buried surface area at p53-DNA interface, obtained from MD simulations, failed to explain the observed increase in p53 binding, consistent with the prepaying mechanism being a key determinant for binding in this case. **h**, DNA hairpin with four mismatches (in the four positions for which the Hoogsteen conformation was previously observed), strongly binds p53: 3–6 k_BT stronger (depending on the data used for validation, Supplementary Tables 3, 4) compared to the highest-affinity p53-binding sites previously reported[22]. Notably, we expect the difference in binding affinity to other genomic p53 sites ($\Delta\Delta G$) to be even larger, as most p53-binding sites in the genome are of lower binding affinities[22].



Extended Data Fig. 7 | The effects of mismatches on TBP–DNA binding. **a**, Mismatch profile for TBP. **b**, Correlations between TBP-binding levels and DNA duplex stability were computed over all 16 base-pair variants at positions 1 to 8 in the TBP site. Bar plots (left) represent the squared Pearson correlation coefficient (R^2) at each position. For the only three positions with significant correlations (positions 2, 7, and 8) the scatter plot correlation is presented (right), with binding signals representing medians over 9 replicate spots. Blue shaded regions, 95% confidence interval for Pearson's correlation. The sequences of the Watson–Crick and mismatched base pairs are shown in each scatter plot (for example, for position 8, GC stands for the wild-type G-C base-pair in bold in the TBP site TATAAAA**G**, CC stands for C-C at this position, and so on). These high correlations are observed only in the unstacked base step positions. **c**, Left, structural overlays between TBP–DNA complexes with DNA mismatches (TBP-AC, orange; TBP-CC(2), cyan; TBP-CC(1a), purple; TBP-CC(1b), pink) and their corresponding Watson–Crick counterparts with single base substitutions (1QNE, green; 6NJQ, yellow). The base steps at position 7–8 are magnified and highlighted in black boxes. The structural overlay of the mismatch and the Watson–Crick base pairs are shown below each box, with their DNA sequences. Right, overlays of protein-DNA interfaces of TBP-DNA complexes, comparing mismatched and Watson–Crick sites. Four phenylalanine residues, as well as other amino acids that are discussed in the Supplementary Discussion are highlighted with dashed circles. **d**, Comparisons of the effects of Watson–Crick mutations versus the cumulative effects of the two closest mismatches, shown for the mismatches with new crystal structures. In all three cases the mismatches have significantly larger effects than the Watson–Crick mutations (see also Methods and Supplementary Table 4). $\Delta\Delta G$ values for TBP_site_1 in Supplementary Table 4 were used in these comparisons. **e**, Example of a Watson–Crick mutation that has a similar effect (within experimental error, Supplementary Table 4) to the sum of the two closest mismatches. $\Delta\Delta G$ values for TBP_site_1 in Supplementary Table 4 were used in these comparisons.



Extended Data Fig. 8 | Potential mechanisms for mismatch-enhanced TF binding. a, TF–DNA complex formation involves creation of intermolecular interactions, as well as DNA conformational changes. Thermodynamically, these processes can be separated into two independent events, and thus an increase in binding affinity could stem from additional interactions (decrease of $\Delta G_{interaction}$), and/or a reduction in the penalty to change the DNA conformation (decrease of $\Delta G_{penalty}$). **b**, A reduction in the energetic penalty to distort the DNA ($\Delta G_{penalty}$) could originate from DNA conformational changes owing to the mismatch, that is, before binding (for example, p53 and TBP, as described in the main text). **c**, A reduction in the energetic penalty for DNA distortion ($\Delta G_{penalty}$) could also originate from changes in the bound DNA. For example, MD simulations of the DNA conformations in free form and in the MYC–DNA complex (for the wild-type A-T and the mismatch G-T) suggest that the reduced penalty in this case is primarily due to changes in the mismatched bound form. The extent of overlap of the kinking direction (γ_h) obtained from the MD simulations was: $\Omega = 0.34$ (wild type) versus $\Omega = 0.15$ (G-T mismatch), and was analysed using a revised Jensen–Shannon divergence score (Ω)[81]. Representative structures of the DNA sites are shown for wild-type free (pink), wild-type bound (orange), G-T free (green) and G-T bound (blue). The MYC–MAX heterodimer is shown as a grey surface. **d**, Mismatches could lead to the formation of non-native interactions such as hydrogen bonds (left), electrostatic potential and shape sensing (centre), and water-mediated interactions (right). Red empty arrows point to the locations of the change. These changes could occur directly at the position of the mismatched base (for example, the G-T mismatch for ETS1), as well as at the positions of other bases and/or the backbone, owing to non-native structures (for example, the G-A mismatch for ETS1). Notably, mismatches not only alter the potential interacting chemical groups of the replaced base, but can also alter the relative orientation of the interacting bases (as observed for the T in the wobble geometry on the left).



Extended Data Fig. 9 | DNA mismatches in the cell. a, Mismatches can result from misincorporation of bases during DNA replication by DNA polymerases. The average rate at which replication errors are generated and escape proofreading is low in healthy cells (around 10^{-9}), but high in certain cancers and cells with Pol-ε or Pol-δ mutations. Even in healthy cells, the rates of generation of individual mismatches vary by more than a million fold[17] depending on the sequence context and the type of mismatch. **b**, Mismatches result from genetic recombination. A characteristic feature of homologous recombination is the exchange of DNA strands, which results in the formation of heteroduplex DNA. Mismatches can result from genetic recombination when the parental chromosomes contain non-identical sequences. In addition, mismatches can arise during DNA synthesis associated with recombination repair. The repair of these mismatches might be less efficient, as it was previously shown[82] that there is a strong temporal coupling between DNA replication and mismatch repair but a lack of temporal coupling for heteroduplex rejection[82]. **c**, Spontaneous deamination is common and estimated to occur 100−500 times per cell per day in humans[83]. G-T mismatches generated by deamination of 5-methylcytosine (5-meC) are not repaired by the DNA mismatch repair pathway and have considerably lower repair efficiency[83]. The high rate of 5-meC deamination, combined with their relatively slow repair in mammalian cells, contribute to making 5-meC a preferential target for point mutations (about 40-fold) compared to other nucleotides in the genome[84], and one of the major sources of the frequent C-to-T mutations observed in human cells[18]. **d**, Transcription factors bound to mismatched DNA could interfere with Pol-δ strand displacement activity. Left, DNA synthesized by non-proofreading mismatch-prone Pol-α is normally displaced by the proofreading non-error-prone Pol-δ. Right, it was previously shown[10] that increased mutation signals arise from regions synthesized by Pol-α that contain TF-binding sites. This study suggested that mismatched DNA synthesized by non-proofreading Pol-α is rapidly bound by TFs that act as barriers to Pol-δ displacement of Pol-α-synthesized DNA, resulting in locally increased mutation rates in subsequent rounds of replication.

Extended Data Table 1 | Data collection and refinement statistics for TBP–DNA mismatch structures

TBP-DNA structure	TBP-AC	TBP-CC(1a)	TBP-CC(1b)	TBP-CC(2)
Pdb code	6UEO	6UEP	6UER	6UEQ
Space group	P1	C2	$P2_12_12_1$	$P222_1$
Cell constants (Å)	a=42.4	a=113.6	a=88.9	a=45.4
	b=55.5	b=46.7	b=91.2	b=45.6
	c=146.3	c=146.3	c=97.6	c=155.2
Cell angles (°)	α=89.97	α=90.0	α=90.0	α=90.0
	β=90.0	β=95.5	β=90.0	β=90.0
	γ=90.14	γ=90.0	γ=90.0	γ=90.0
TBP-DNA complexes In ASU	4	2	2	1
Resolution (Å)	73.1-2.00	145.6-2.05	65.7-2.50	155.2-2.40
R_{sym}(%)[a]	5.8 (19.8)[b]	3.5 (35.3)	11.7 (65.4)	6.6 (46.1)
R_{pim} (%)	5.7 (19.6)	2.3 (26.7)	6.6 (41.6)	3.5 (24.5)
Overall I/σ(I)	7.1 (2.9)	18.4 (2.7)	6.6 (2.3)	10.3 (2.0)
#Unique Reflections	77170	83131	63767	12803
#Total Reflections	128765	143102	280131	82092
% Complete	90.7 (87.0)	93.0 (65.4)	99.7 (94.0)	96.4 (92.0)
CC(1/2)	0.998 (0.965)	0.999 (0.838)	0.989 (0.945)	0.998 (0.896)
Refinement Statistics				
Resolution (Å)	73.1-2.00	145.6-2.09	65.7-2.50	155.2-2.40
R_{work}/R_{free}(%)[c]	21.3/24.7	17.1/19.2	19.9/23.9	21.2/24.9
Rmsd				
Bond angles (°)	0.620	1.04	0.657	0.568
Bond lengths (Å)	0.004	0.010	0.004	0.003
Ramachandran analysis				
Favored (%/)	95.9	98.1	95.1	96.2
Disallowed(%)	0.0	0.0	0.0	0.0

[a] R_{sym} = ΣΣ|Ihkl − Ihkl(*j*)|/ΣIhkl, where Ihkl(*j*) is the observed intensity and Ihkl is the final average value of intensity.

[b] Values in parentheses are for the highest-resolution shell.

[c] $R_{work} = \Sigma||F_{obs}| - |F_{calc}||/\Sigma|F_{obs}|$ and $R_{free} = \Sigma||F_{obs}| - |F_{calc}||/\Sigma|F_{obs}|$; where all reflections belong to a test set of 5% randomly selected data.

nature research

Corresponding author(s): Raluca Gordan, Hashim Al-Hashimi

Last updated by author(s): Aug 17, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	SaMBA data were collected using GenePix Pro software (version 7.0). Crystallization data were processed with MOSFLM 7.3.0 and scaled with SCALA in Ccp4i version 7.0.078.
Data analysis	Custom Python 2.7.16 code (available at https://github.com/alhashimilab/TF_MM) was used to analyze NMR and crystal structures, using X3DNA-DSSR 1.6.5, Numpy 1.16.6, Matplotlib 2.2.5, and pandas 0.24.2. Thermodynamic parameters for mismatch formation, for Watson-Crick and mispairs, were computed using MELTING 5.2.0. Molecular dynamics simulations were performed using the AMBER 16.0 and AmberTools 17.0. Crystal structures were solved by molecular replacement (with MolRep in ccp4i version 7.0.078). After refinement in Phenix (version 1.17), the structures were manually rebuilt in O (version 8). MolProbity (version 4.5) was used to guide the process of refitting and refinement. Protein-DNA structures were illustrated using PyMOL 1.5.0.4. For the p53 EMSA assays, the bands were analyzed using Cliqs version 1.1. For the GR EMSA assays, the bands were analyzed using LiCor Image Studio version 5.2.5.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data that support the findings in this study are available as Supplementary Tables, in Excel format. Coordinates and structure factor amplitudes for the TBP-AC,

TBP-CC(1a), TBP-CC(1b) and TBP-CC(2) structures have been deposited in the RCSB Protein Data Bank (PDB) under the accession codes 6UEO, 6UEP, 6UER, and 6UEQ, respectively. The raw SaMBA data has been deposited in the Gene Expression Omnibus (GEO) under accession number GSE156375. The RCSB PDB entries used in this study are available in Extended Data Figures 1, 2, 5, and 7, and Supplementary Tables 5, 6, 7, and 9. High-resolution gel images for the EMSA data are available at https://figshare.com/projects/DNA_mismatches_reveal_conformational_penalties_in_protein-DNA_recognition/83663 .

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were chosen based on previous work (Shen et al, Cell Systems 2018), which used 6 replicate spots per unique DNA sequence. In previous work we found that using 6 replicate spots and computing the median signal over these replicates, was sufficient to obtain highly reproducible measurements of protein-DNA binding levels (with R^2>0.95 between independent experiments). On the SaMBA arrays we had sufficient space for 8-20 replicates spots per sequences; we use the maximum number of replicates possible on each array.
Data exclusions	No data were excluded
Replication	Scatterplot in Figure 1e shows highly reproducible data between the two independent SaMBA binding experiments performed for Ets1 (R^2=0.98). Scatterplot in Extended Data Figure 3d shows highly reproducible data between the two independent hybridization experiments performed (R^2=0.98). EMSA experiments for p53 and GR used six and three replicates, respectively. FA experiments for Ets1 used three replicates. All attempts at replication were successful.
Randomization	For each SaMBA DNA libray, replicate spots were randomly distributed across the DNA array surface. Randomization was not applicable to other experiments performed in this study.
Blinding	Blinding is not applicable to this study, as protein DNA samples are not required to be allocated into experimental groups in protein binding studies. No animals or human research participants were involved in this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Primary antibody: RelA rabbit polyclonal antibody (Origene Catalog #: TA890002). Anti-tag antibodies: Alexa647-conjugated GST antibody (Cell Signaling Technology, Catalog #3445), Alexa488-conjugated GST antibody (Invitrogen, Catalog #: A-11131), Penta·His Alexa647-conjugated antibody (Qiagen, Catalog #: 35370), Penta·His Alexa488-conjugated antibody (Qiagen, Catalog #: 35310). Secondary antibody: Goat anti-Rabbit IgG Alexa647 antibody (ThermoFisher, Catalog #: A21244).
Validation	The RelA primary antibody (mouse anti-human) was tested by Origene and guaranteed activity in applications: WB, IHC. Please see manufacturer's website for details: https://www.origene.com/catalog/antibodies/primary-antibodies/ta890002s/nf-kb-p65-rela-rabbit-polyclonal-antibody . For anti-tag antibodies, please see manufacturer's website for details on quality control and validation: https://www.cellsignal.com/products/antibody-conjugates/gst-26h1-mouse-mab-alexa-fluor-647-conjugate/3445 ; https://www.thermofisher.com/antibody/product/GST-Tag-Antibody-Polyclonal/A-11131 ; https://www.qiagen.com/us/products/discovery-and-translational-research/protein-purification/tagged-protein-expression-purification-detection/penta-his-alexa-fluor-647-conjugate/#productdetails ; https://www.qiagen.com/us/products/discovery-and-translational-research/protein-purification/tagged-protein-expression-purification-detection/penta-his-alexa-fluor-488-conjugate/#productdetails . For information on the

secondary antibody used in this study, please see the manufacturer's website: https://www.thermofisher.com/antibody/product/Goat-anti-Rabbit-IgG-H-L-Cross-Adsorbed-Secondary-Antibody-Polyclonal/A-21244 .