Eukaryotic transcription factors can track and control their target genes using DNA antennas

Milagros Castellanos1,2, Nivin Mothi3 & Victor Muñoz1,2,3*

Eukaryotic transcription factors (TF) function by binding to short 6-10 bp DNA recognition sites located near their target genes, which are scattered through vast genomes. Such process surmounts enormous specificity, efficiency and celerity challenges using a molecular mechanism that remains poorly understood. Combining biophysical experiments, theory and bioinformatics, we dissect the interplay between the DNA-binding domain of Engrailed, a Drosophila TF, and the regulatory regions of its target genes. We find that Engrailed binding affinity is strongly amplified by the DNA regions flanking the recognition site, which contain long tracts of degenerate recognition-site repeats. Such DNA organization operates as an antenna that attracts TF molecules in a promiscuous exchange among myriads of intermediate affinity binding sites. The antenna ensures a local TF supply, enables gene tracking and fine control of the target site’s basal occupancy. This mechanism illuminates puzzling gene expression data and suggests novel engineering strategies to control gene expression.

1 Instituto Madrileño de Estudios Avanzados en Nanociencia (IMDEA Nanociencia), Faraday 9, Campus de Cantoblanco, Madrid 28049, Spain. 2 Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas (CSIC), Darwin 3, Campus de Cantoblanco, Madrid 28049, Spain. 3 Department of Bioengineering, School of Engineering, University of California, 95343 Merced, CA, USA. *email: vmunoz3@ucmerced.edu
transcription factors (TF) control gene expression by binding to their target DNA site to recruit, or block, the transcription machinery onto the promoter region of the gene of interest. Their function relies on the ability to find their target site quickly and selectively. In living cells TFs are present in nM concentrations and bind the target site with comparable affinity, but they also bind any DNA sequence (nonspecific binding), resulting in millions of low affinity (i.e., >10⁻⁸ M) competing sites. Nonspecific binding facilitates the search for the target site by allowing the TF to slide along DNA via a relatively slow, but more efficient, one dimensional diffusive motion (D < 10⁻⁸ cm² s⁻¹) that involves rotation about the DNA axis and covers distances between 300 and 10,000 bp. Another mechanism of facilitated diffusion occurs when the TF is transferred between DNA regions in transient spatial proximity. These various nonspecific binding modes act jointly to speed up the TF recognition of its target site. For instance, in vivo imaging experiments in bacteria indicate that the combination of these molecular elements suffices to explain the homing, selectivity and occupancy of prokaryotic TFs.

Eukaryotic gene expression is much more complex and operates in multiple layers, including dynamic control over the chromatin structure, and epigenetic factors. But even at the molecular level, achieving efficient transcription control is much more challenging than in prokaryotes. Eukaryotic genomes are orders of magnitude larger and their TFs feature much shorter DNA recognition sites, leading to hundreds of random occurrences for any consensus sequence, and thus to inevitably impaired specificity and site occupancy. Moreover, each eukaryotic TF controls tens to hundreds of genes scattered throughout the genome, and expressing each gene needs various TFs simultaneously binding to their sites to form the transcription complex, an extremely rare event in probabilistic terms. As result, the in vivo site occupancy patterns of eukaryotic TFs are more complex than predicted by their in vitro site-specific binding profiles and do not strongly correlate with the actual levels of gene expression. Intriguingly, single-molecule fluorescence tracking in mammalian cells has shown that the TF Sox2 finds one of its target sites in fewer than 100 binding attempts, suggesting that it only samples a miniscule fraction of the transcriptionally accessible genome.

An interesting feature highlighted by genome analysis is an accumulation of potential TF binding sites in regions flanking eukaryotic genes. Such clusters of degenerate recognition sites are assumed to be key for transcription control, and thus are generally classified as gene regulatory regions (RR). The potential roles that repetitive sequence patterns flanking the cognate site may play on how eukaryotic TFs find their target site have been recently subject to intense scrutiny. For instance, when surrounding the target site, certain symmetric sequence repeats can affect the binding preference by simple statistical (or entropic) factors. However, the molecular aspects of the interaction between TF and these flanking DNA regions have not yet been established, nor is there a mechanism available that integrates binding to these regions with canonical specific and nonspecific DNA binding to enable efficient eukaryotic transcription. Here we address this fundamental question investigating the interactions between the DNA-binding domain of a eukaryotic TF and the regulatory regions of genes under its transcriptional control. We utilize biophysical methods to dissect the binding process, statistical mechanical modeling to integrate and rationalize the results, and bioinformatics analysis to further explore the functional implications.

**Results**

A model of eukaryotic transcription factor binding to target gene. We focus on Engrailed, a TF from *Drosophila melanogaster* involved in defining embryonic parasegmental subdivision and maintaining parasegmental borders in adult appendages. Engrailed controls the expression of over 200 different genes in *Drosophila*. Its DNA-binding domain (EngHD) folds into a three-helix bundle that recognizes the palindromic sequence TAATTA as its consensus site. EngHD binds to DNA inserting its C-terminal α-helix into the DNA major groove to engage in specific interactions with consensus bases (Fig. 1a, b). Binding is reinforced by electrostatic interactions with the phosphate backbone (Fig. 1b). As example of gene regulatory region (RR), we selected a fragment from the β-tubulin gene, which is repressed by Engrailed. This fragment (bases 2769–2918 from the transcription start site) falls within the first intron, and contains one specific binding site that slightly diverges from the consensus (TAATTTG), but retains relatively high affinity. We employed fluorescence correlation spectroscopy (FCS) as biophysical method to characterize the interactions between EngHD and the β-tubulin RR in quasi-single particle conditions. In FCS, a miniscule confocal volume (~1 fL) is illuminated so that the diffusive paths of individual fluorescent molecules are detected from correlated fluctuations in fluorescence emission. Free and bound molecules are identified based on their different diffusive properties. In our case, we use a fluorescence-labeled version of EngHD at ~1 nM in the presence of varying concentrations of unlabeled dsDNA, and determine the fraction of EngHD bound to DNA from the retarded diffusion of the complex (Fig. 1c) (see Methods). FCS is optimally suited for our purposes because it is sensitive to binding in sub-nM to μM range and measures DNA association directly, whether such association comes from one (specific) or many (nonspecific) binding sites.

Specific versus nonspecific DNA-binding contributions. We analyzed the contributions to EngHD’s binding affinity using a series of DNA molecules based on the 75-bp central segment of the β-tubulin intron in which we modified the specific binding site (Supplementary Table 1). FCS experiments on the consensus variant (TAATTA) rendered a dissociation constant ($K_D$) of $4 \cdot 10^{-9}$ M (Fig. 2a, Supplementary Table 2) that is consistent with previous measurements on a non-natural DNA sequence using gel-shift assays. Experiments on the remaining DNAs showed affinity decreases proportional to the divergence from the consensus sequence (Fig. 2a, Supplementary Table 2). TAATTT exhibited a 4-fold decrease in affinity. Permutation of the two central bases (TATATA) and replacement of the last base by G rendered 10-fold decreases. The simultaneous change of the two end bases to G, or a highly divergent binding site (CGTGTTT) resulted in 19- and 22-fold affinity drops, respectively. These experiments reveal evident changes in affinity. However, the affinity decrease is small relative to how much the binding site diverges from the consensus, most notably for CGTGTTT in which only one consensus base is retained. To investigate this issue, we compared these results with the position weight matrix (PWM) for Engrailed obtained from bacterial one-hybrid assays. The PWM recapitulates the consensus binding site (Fig. 2b) and predicts a decreasing binding probability as the target site diverges from consensus. However, the PWM predicts changes many orders of magnitude larger than what we find.
experimental (Fig. 2b). The PWM also predicts huge differences in binding to target sites that exhibit very similar affinity in the context of the 75 bp β3-tubulin intron DNA. The implication is that EngHD DNA binding is more complex than dictated by specific interactions with the consensus motif. The most likely explanation is that EngHD binds promiscuously to the flanking DNA sequence, thereby buffering the degradation or even elimination of the consensus site.

An obvious factor driving promiscuity are electrostatic interactions, which contribute to the stabilization of the specific binding site[42], but also promote nonspecific binding to any other site along a given DNA molecule[3]. The 3D structure of the EngHD-DNA complex[40,43] highlights attractive electrostatic interactions formed between positively charged side-chains in EngHD and the DNA phosphate backbone (Fig. 1b), consistently with reports on other DNA-binding domains[42,44,45]. To establish the role of these interactions, we investigated how ionic strength affects EngHD-DNA complex[40,43] highlights attractive electrostatic interactions formed between positively charged side-chains in EngHD and the DNA phosphate backbone (Fig. 1b), consistently with reports on other DNA-binding domains[42,44,45]. To establish the role of these interactions, we investigated how ionic strength affects EngHD-DNA complex[40,43] highlights attractive electrostatic interactions formed between positively charged side-chains in EngHD and the DNA phosphate backbone (Fig. 1b), consistently with reports on other DNA-binding domains[42,44,45]. To establish the role of these interactions, we investigated how ionic strength affects EngHD-DNA complex[40,43] to the experimental data and the resulting parameters are described in Methods. We find that the PWM model reproduces all of the data at a semiquantitative level using only one fitting parameter, whereas the structure-based model fits the data over 50-fold better using four parameters (see Supplementary Fig. 1). A Fisher test indicates that the probability that the statistically simpler model (PWM) is equivalent to the fundamentally complex (structure-based) model is below 10−9. We thus focused on the structure-based model (Fig. 3b) for all subsequent analyses. The fits to, and predictions from, the structure-based model are shown as red circles and/or red curves throughout the article (e.g. Fig. 2a, c).

The ability of the structure-based model to reproduce the non-trivial changes in affinity that we observe suggests that it captures the fundamental energetics of EngHD binding to DNA. Such binding energetics confirm the existence of a third, non-canonical, DNA-binding mode in which EngHD binds promiscuously to degenerate consensus repeats. From here onwards we denote these three DNA-binding modes as: (1) specific binding to a consensus site (SB); (2) degenerate consensus binding (DCB), which refers to binding to any other site with a partial consensus sequence; (3) nonspecific binding (NSB), defined as sequence-independent, electrostatically driven binding[9]. A similar, non-specific binding of eukaryotic TFs to clusters of degenerate consensus repeats around a cognate site has been proposed to increase the site's occupancy in vivo[30,31] and to accelerate the search for the target site[33]. Here we determine its actual contribution to binding and dissect its molecular mechanism. The key questions that emerge are: how does the interplay of these three binding modes define the overall binding behavior of EngHD? And what are its functional implications?

A simple theoretical model of EngHD-DNA-binding energetics. To quantitatively rationalize the complex DNA-binding properties of EngHD we built a statistical mechanical model that considers binding to any 6-bp site available in any given DNA molecule (Fig. 3). Particularly, we implemented two versions of the model energetics: one empirical version based on the Engrailed PWM matrix, and another version inspired by the 3D structure of the EngHD-DNA complex[40] (Fig. 3a). Both models, the fitting to the experimental data and the resulting parameters are described in Methods. We find that the PWM model reproduces all of the data at a semiquantitative level using only one fitting parameter, whereas the structure-based model fits the data over 50-fold better using four parameters (see Supplementary Fig. 1). A Fisher test indicates that the probability that the statistically simpler model (PWM) is equivalent to the fundamentally complex (structure-based) model is below 10−9. We thus focused on the structure-based model (Fig. 3b) for all subsequent analyses. The fits to, and predictions from, the structure-based model are shown as red circles and/or red curves throughout the article (e.g. Fig. 2a, c).

The ability of the structure-based model to reproduce the non-trivial changes in affinity that we observe suggests that it captures the fundamental energetics of EngHD binding to DNA. Such binding energetics confirm the existence of a third, non-canonical, DNA-binding mode in which EngHD binds promiscuously to degenerate consensus repeats. From here onwards we denote these three DNA-binding modes as: (1) specific binding to a consensus site (SB); (2) degenerate consensus binding (DCB), which refers to binding to any other site with a partial consensus sequence; (3) nonspecific binding (NSB), defined as sequence-independent, electrostatically driven binding[9]. A similar, non-specific binding of eukaryotic TFs to clusters of degenerate consensus repeats around a cognate site has been proposed to increase the site's occupancy in vivo[30,31] and to accelerate the search for the target site[33]. Here we determine its actual contribution to binding and dissect its molecular mechanism. The key questions that emerge are: how does the interplay of these three binding modes define the overall binding behavior of EngHD? And what are its functional implications?

The binding free energy landscape flanking the target site. Our theoretical analysis points to DCB as modulator of EngHD’s affinity for the β3-tubulin first intron. The DNA sequence flanking the target site is indeed rich in A/T clusters[39] that feature many potential DCB sites (see m2 in Fig. 4a). The calculated
Fig. 2 Mapping the energetics of EngHD binding to DNA. a (top) Experimental FCS autocorrelation decays of EngHD in the presence of 50 nM of each of the six 75-bp DNA molecules based on the β3 tubulin gene with variations in the SB site. The average diffusion time relative to diffusion of the saturated complex (black curve) reflects the fraction of bound molecules. (bottom) Dissociation constants for the six 75-bp DNAs determined experimentally by FCS from three independent experiments and calculated by the statistical mechanical model (red circles). b Experimental changes in binding affinity compared with the changes expected by pure consensus binding. The upper panel shows the consensus binding logo for engrailed obtained from bacterial one-hybrid high throughput assays. The lower panel shows the correlation between the experimental changes in binding affinity, $K_D$(variant) · $K_D$(specific)$^{-1}$, in the abscissa; and the inverted relative probability of binding calculated from the position weight matrix, $p_{PWM}$(specific) · $p_{PWM}$(variant)$^{-1}$, in the ordinate. c Ionic strength dependence of EngHD binding to the 75-bp DNA molecule bearing the natural TAATTG high affinity site (dark blue in a). Experimental data are shown as black open circles and the statistical mechanical model calculation is shown as a red curve. Bars delimit the 95% confidence interval, see Supplementary Table 2. Source data are provided as a Source Data file.

Fig. 3 Statistical mechanical model to describe the DNA-binding modes of EngHD. a Scheme of the various interactions existing between EngHD and DNA as observed in the 3D structure of the EngHD-DNA complex (based on ref. 40, PDB: 1HDD). b (top) Representation of how the statistical mechanical model calculates all possible binding events using a 6-bp sliding window that runs 5′–3′ through both strands. (bottom) detailed binding energetics for the two exemplary binding sites selected from the top (shown on top as gray boxes). The upper example includes all the specific binding interactions of a full consensus site, and the lower is an example of a degenerate consensus site. The model and interactions are described in the text, and the parameters obtained after global optimization against all experimental data (Figs. 2, 4 and 5) together with their statistical significance (one standard deviation) are given in the box. Black symbols correspond to consensus core tetrad interactions ($\Delta G_{\text{consensus core}}$), gray asterisks correspond to degenerate consensus interactions with A or T ($\Delta G_{\text{degenerate AT}}$), orange circles correspond to electrostatic (nonspecific) interactions, which extend over 8-bp ($\Delta G_{\text{elec}}$) and cp is the cooperative interaction that takes place when the site includes the full consensus sequence ($2\Delta G_{\text{cp}}$; see model description).
binding profile of the whole DNA molecule shows a rugged binding free energy landscape with many minima. The local minima concentrate around the target site, producing an overall funnel for EngHD binding (m1 in Fig. 4c), a property reminiscent of the energy landscapes associated to protein folding, binding and function. Interestingly, binding to many mid-affinity sites around the target site could be a strategy to enhance overall affinity without involving strong specific interactions, that is, maintaining relatively fast dissociation rates. Such behavior is consistent with theoretical predictions. A rugged funneled binding landscape also introduces resilience to mutations on the target site, exactly as we observe experimentally (Fig. 2a). This binding landscape also introduces resilience to mutations on the absence of a competing SB site see their occupancy raised, thereby buffering the overall drop in affinity.

To further investigate this phenomenon, we designed a 75 bp DNA that carries the original β3-tubulin high affinity site (TAATTG), but minimizes A–T content everywhere else, and thus eliminates DCB sites (m3 in Fig. 4a). m3 shows a 21-fold decrease in binding affinity relative to m2 (KD of ~0.7 · 10^-6 M) (Fig. 4b). Such drop is striking because the two DNAs have the same target site. We hence confirm that DCB dominates the overall binding to the β3-tubulin intron region. The statistical mechanical model underestimates the affinity drop, presumably because this simple model does not account for the formation of secondary structure that takes place in this G/C-rich DNA (Supplementary Fig. 2) and which is likely to further impair EngHD binding. In fact, our experimental result is close to the model prediction for the same flanking sequence bearing the low affinity CGTGTT on the target site (open circle in Fig. 4b). An alternative explanation for this result could be the potential accumulation of symmetric base repeats in the flanking region, a factor proposed, and found in the TF MAX, to entropically stimulate binding in the absence of base-specific interactions. The β3-tubulin intron fragment does indeed contain many base repeats (m2 in Fig. 4a). We thus tested this possibility using a DNA that maintains the β3-tubulin original base composition, but it eliminates base repeats (m5 in Fig. 4a). We also used the low affinity CGTGTT as target site, aiming to minimize SB contributions and thus increase the experimental sensitivity to differences between DCB and NSB. The statistical mechanical model does not include nonspecific effects from symmetric base repeats, and, accordingly, it calculates minimal binding differences between m5 and the β3-tubulin sequence (m4). FCS experiments also show minimal differences (Fig. 4b), confirming that the flanking DNA effects we see in EngHD arise from promiscuous DCB instead of from nonspecific base repeats. Interestingly, the flanking effects appear to extend over relatively long distances, as suggested by the two-fold higher affinity of the 150-bp β3-tubulin fragment (m6) relative to the 75-bp version (m2; Fig. 4b). Experiments on other 150-bp DNA molecules further confirm that, in absence of a SB site, the affinity increases proportionally to the availability of DCB sites. For instance, shuffling the entire 150-bp sequence (m7) or the external region (m8), does not change the affinity in either experiments or model calculations (red circles for m6, m7 and m8 in Fig. 4b).

Fig. 4 The region around the specific site produces a rugged free energy landscape. a DNA molecules used to explore the effects of the sequence of the region flanking the specific site. b Dissociation constants for the DNA molecules shown in panel a obtained by FCS from three independent experiments, and calculated by the statistical mechanical model (red circles). c Examples of binding free energy landscapes and EngHD occupancy profiles obtained by the model. BS index indicates the position of the first base of each potential binding site in the coding or complementary strands. The free energy scale is given on the left, and the probability of each site being occupied on the right y-axis. The occupancy profiles have been calculated at the same DNA concentration of 10 nM: just above the KD for m1 and below the KD for m2. Bars delimit the 95% confidence interval, see Supplementary Table 2. Source data are provided as a Source Data file.

Contribution from promiscuous binding to degenerate consensus sites. A key question is whether promiscuous DCB is just localized near the SB or propagates over the full RR of EngHD target genes. This consideration is important given that eukaryotic cis-acting RR extend over thousands of bp, and are often located far (>50 Kb) from the transcription starting site. The long eukaryotic RRs could potentially exploit DCB to massively amplify the binding affinity of relevant transcription factors. To investigate this hypothesis, we designed a series of DNA
molecules based on the β3-tubulin intron but with varying size (38, 75, 150, 300, and 600 bp). In FCS experiments these DNA molecules diffuse with coefficients that decrease proportionally to their size (Fig. 5a) as expected from the known length dependence of DNA’s translational diffusion coefficient(50). Binding experiments showed a remarkably strong amplification of EngHD binding as its natural DNA partner grows in length: from $K_D \sim 5.6 \times 10^{-8}$ M for the 38-bp DNA to $\sim 2.2 \times 10^{-9}$ M for the 600-bp molecule, or a $\sim 25$-fold increase for a 15-fold longer DNA that does not incorporate extra SB sites (Fig. 5b). Exploring by FCS the flanking sequence effects over shorter or longer DNA scales is difficult due to technical limitations: 38-bp is the shortest DNA that results in a complex with diffusion coefficient clearly distinguishable from that of free EngHD, and the affinity of the 600-bp is close to the detection limit. However, the statistical mechanical model, which recapitulates these experimental trends (red circles in Fig. 5b), predicts a 300-fold affinity increase for the entire β3-tubulin gene(51), relative to its unique high affinity SB site (Fig. 5b). This strong amplification implies that EngHD binding to the β3-tubulin gene is in fact dominated by promiscuous binding to DCB sites, which win over SB by virtue of their vast numbers, even though each site has relatively low affinity (i.e., $\sim 10^{-7}$ M).

To determine whether binding amplification scales up to full genes, we synthesized a 7.2 kbp DNA encompassing the β3-tubulin gene sequence (including introns, but without 5′ and 3′ UTR). FCS experiments of EngHD in the presence of pM concentrations of this DNA molecule confirm very strong binding, which is noticeable even at $\sim 150$ pM (Fig. 5c), indicating a $K_D$ in the sub-nM range (Fig. 5b). These experiments demonstrate that the binding amplification induced by DCB extends over the DNA scales of full genes. As result, the affinity for the entire β3-tubulin RR is orders of magnitude higher than binding to just its original, high affinity target site. In other words, DCB transforms the β3-tubulin RR onto a potent attractor for EngHD molecules. It follows that such binding pattern ensures local availability of the transcription factor as well as low occupancy of the specific site, and thus may operate as a transcription antenna.

**The binding profile of a transcription antenna.** The fingerprint of transcription antennas would be the accumulation of DCB on the regulatory (noncoding) regions of the gene as opposed to the coding regions (exons). We can investigate this question bioinformatically by calculating the EngHD binding profiles for other gene sequences. Before embarking on large-scale DNA sequence profiling, however, we tested the biological significance of the predictions of our model by calculating the binding affinity of the 2226 DNA fragments (each between 100 and 500 bp long) that have been identified in ChIP-Seq experiments as DNA regions that bind Engrailed in vivo. The model predicts high binding affinity for all ChIP-Seq fragments, with over 90% of the fragments’ predicted $K_D$ values between 1 and 10 nM (Fig. 6a). These affinities are comparable to what we have measured in vitro for the 150 bp segment from the first intron of β3-tubulin carrying the consensus site (Fig. 4b). We can thus conclude that the binding predictions of the statistical mechanical model are biologically significant. In this regard, we note that the binding profile of the full β3-tubulin gene (containing the 5′ and 3′ UTR regions) has a distinct pattern of dense local clusters of DCB sites found in the noncoding regions (magenta in Fig. 6b) together with an absence of them in exons (orange in Fig. 6b), which is what we expect for a transcription antenna. Hence, the β3-tubulin gene sequence maximizes EngHD binding along the RR so that molecules of EngHD become localized around the target gene via promiscuous DCB binding, whereas the exons remain unoccupied. In Drosophila cells, EngHD is present at 1–10 nM concentrations(52), which suggests that the β3-tubulin RR will host several EngHD molecules at all times. Here it is important to note that the overall affinity is high, but each DCB binding event has moderate affinity ($K_D \sim 10^{-7}$ M), and therefore a relatively high dissociation rate that could permit the fast interconversion between sites.

**A kinetic model of transcription antennas.** The binding profile of the β3-tubulin RR immediately suggests a role as efficient tracking device that limits the TF search for the SB to within the antenna limits, rather than over the entire genome (or cell).
Such a transcription antenna could be also used to control the SB occupancy at the levels required for biological function. To quantitatively explore these ideas, we built a very simple kinetic model of a transcription antenna. The model is composed of four species connected kinetically as shown in Fig. 6c, and it assumes that the region of interest is transcriptionally active. For any such accessible gene, the relevant TF is in either of the four states: bound to DCB sites in the antenna \((D_{\text{deg}})\), bound to the specific site \((S\beta)\), in the small cellular volume surrounding the antenna \((A\beta)\), or in the cellular milieu \((C\beta)\). Unbound TF molecules enter and escape the antenna space by diffusion (modeled with rate constants \(k_1\) and \(k_{-1}\)). Once inside the antenna, free TF molecules \((A\beta)\) bind to any DCB site \((D_{\text{deg}})\) via \(k_2\), and to \(SB\) via \(k_3\). The TF can be released back to \(A\beta\) from the bound states via \(k_{-2}\) and \(k_{-3}\), respectively. Finally, the TF can reach \(SB\) from sites within \(D_{\text{deg}}\) or vice versa via one dimensional diffusion (sliding and/or intersegment transfer), represented by \(k_{4}\) and \(k_{-4}\). The details of the model, rate matrix, and the interpretation of its eigenvalues are provided in Methods. To explore its functioning, we implemented it with parameters (also given in Methods) inspired by our results on the \(\beta\) tubulin gene and general properties of Engrailed and \textit{Drosophila} cells, aiming simply to represent a plausible scenario. These calculations confirm the anticipated mode of operation, which is shown graphically in Fig. 6d. As illustrated in the figure, the antenna accumulates the largest population of TF molecules in very slow exchange with the cell milieu (TF either free or bound somewhere else in DNA) because the diffusive exchange depends on the infinitesimally small probability of finding an unbound TF within the antenna \((A\beta)\), square bracketed species in Fig. 6d). Therefore, the antenna effectively traps TF molecules in a variety of binding events, thereby locally buffering any changes in accessibility of distant DNA due to chromatin dynamics\(^{2,3}\), or in TF concentration at the cellular level. In contrast, TF molecules within the antenna (including the target site, \(SB\)) remain in fast, millisecond exchange \((k_{\text{ex}}\) in Fig. 6d), either by dissociation—facilitated by the relatively high dissociation rates of DCB sites—and quick rebinding (1 in Fig. 6d), or through sliding (2 in Fig. 6d). These general effects are
also consistent with recent theoretical work 33, and demonstrate how a transcription antenna can guarantee a local supply of TF, quick turnover over the SB site, and fine control of SB occupancy via binding competition with DCB sites.

### Genes regulated by Engrailed contain transcription antennas.

A follow-up question is whether transcription antennas are specific to the β3 tubulin gene or broadly used by Engrailed. At least 203 Drosophila genes are controlled by Engrailed 36. The Drosophila genome is in general A/T rich (about 57%) 54, but it has a spike in A/T content in the gene promoter regions and a downward trend after the transcription site (Fig. 7a) 55. This pattern is consistent with transcription antennas. For further bioinformatic analysis, we focused on the 39 genes that are best characterized as being under Engrailed control (Supplementary Table 3). The profile of transcription antennas in these genes is apparent even via simple sequence analysis. For instance, the consensus sextuplet (TAATTA), the two possible quintets and the consensus sextuplet (TAATTA), the two possible quintets and quartet consensus sites in all of the D. melanogaster genes known to be regulated by Engrailed. RR, regulatory regions; CDS, coding sequence. The pro

**Discussion**

Eukaryotic TFs track their target genes, control site occupancy, and coordinate binding with partners to form the transcription complex. These processes must involve modes of interaction with DNA that go beyond nonspecific binding and facilitated 1D and/or 2D diffusion 56,57. By focusing on the DNA regions flanking the target site of real genes, we have discovered, and characterized, a molecular mechanism that enables such functions. The mechanism exploits the natural tendency of biomolecules to exhibit energetic frustration 46,47, in this case manifested by binding promiscuity. Particularly, we find that the affinity of the Drosophila TF Engrailed to the RRs of its target genes is strongly amplified by long tracts of degenerate consensus repeats that are present in such regions. The combination of a promiscuous TF and a DNA region rich in DCB repeats operates as a transcription antenna. Once the DNA region becomes accessible by chromatin dynamics 52, and thus transcriptionally active, the antenna attracts TF molecules that remain loosely associated to the gene of interest through a highly dynamic exchange among the hundreds of mid-affinity binding sites (<<SB but >> NSB) within the antenna. In this light, we confirm that the short recognition sequences and promiscuous specific binding of eukaryotic TFs are a functional strategy to ensure their colocalization with the relevant genes, as it has been postulated by other authors 29,31. For instance, there are ~30,000 copies of Engrailed per cell 17 and about 200 genes estimated to be under its control 36. Taking the β3 tubulin antenna as example (Fig. 6b), it follows that each of these 200 genes will contain on average ~150 EngHD molecules trapped in its antenna, whereas fewer than 15 molecules will be found anywhere else in the cell (bound non-specifically or free). The pool of TF molecules inside an antenna will be in exchange between sites that are relatively weak binders, so their faster dissociation rates facilitate turnover over the SB site, and thus enable a nimble gene expression response (Fig. 6d).

The antenna mechanism sheds new light onto some puzzling observations of eukaryotic gene expression. The mechanism predicts two “specific” binding modes: a frequent, still physically localized, but weaker binding event to antenna DCB sites, and a
rare, high affinity binding to target sites (SB). These properties are in striking accord with single-molecule TF tracking experiments in mammalian cells, which have reported that only ~1% of detectable binding events (with lifetimes >0.5 s) were to high affinity sites, whereas the remainder involved moderately weak binding events that seem too static to represent TF molecules sliding at 10^5 bp/s^{-1} over DNA. Antennas also enable control of the target site’s occupancy by competing locally for binding. A relatively distant (e.g., few kbp away) antenna can keep the basal SB occupancy of an activator at a suitable minimum and TF supply still relatively local. In contrast, an antenna surrounding the target recognition site can amplify a repressor’s effect. Binding events concentrated on long antennas provides a simple explanation of why crosslinking data on eukaryotes produces many more events than expected from the number of genes under control of the given TF, and relatively weak correlations between site occupancy and gene expression levels.

Furthermore, the use of transcription antennas could greatly facilitate the synchronous recruitment of various TFs to assemble into the transcription machinery. Summarizing, DNA antennas provide an elegant mechanism that sheds new light on how eukaryotic TFs operate at the molecular level and explains several paradoxes of existing eukaryotic gene expression data. These molecular devices provide an additional layer of eukaryotic transcriptional control in which the size and sequence profile of the antenna can be engineered, whether by evolution or by scientists, to modulate site occupancy, response swiftness and levels of gene expression.

### Methods

#### Statistical mechanical model for EngHD binding to DNA

The binding of EngHD to a DNA molecule of N base pairs contains a total of 2 · (N − 5) potential binding sites (binding to all possible 6 bp sites in either strand). Defining the unbound state as reference (statistical weight, \( w_{\text{free}} = 1 \)), the partition function for EngHD binding to DNA is thus:

\[
Q([\text{DNA}]) = \sum_{\alpha = 0}^{N-5} w_\alpha \sum_{\beta = 0}^{N-5} w_\beta + 1
\]

(1)

where \( \alpha \) and \( \beta \) are dummy indexes that indicate the position in the DNA sequence of the first base of each 6-bp binding site on the coding and complementary strands, respectively, both running from the 5′-3′ end. The statistical weight for EngHD bound to DNA site \( x \) (whether on the coding or the complementary strand) is defined by the following equation:

\[
w_x([\text{DNA}]) = w_{\text{free}} [\text{DNA}] \exp (-\Delta G^{\text{binding}} / RT)
\]

(2)

where \( w_x \) is a proportionality constant that represents the ratio between the diffusion controlled rate constant for association (\( k^{\text{on}} \) in M^{-1}s^{-1}) and the dissociation rate constant in the absence of interactions between EngHD and DNA (\( k^{\text{off}} \) in s^{-1}). \( \Delta G^{\text{binding}} \) is the binding free energy between EngHD and site \( x \). Here we set \( w_0 \) to 5 · 10^{-4} M^{-1}, consistently with EngHD’s diffusion coefficient (determined by FCS as ~122 μm^2/s, see results). However, its exact value is of no practical consequence since it just scales \( \Delta G^{\text{binding}} \). To define \( \Delta G^{\text{binding}} \) as a function of the site’s sequence we developed two complementary interaction models inspired by the PWM and the 3D structure of the EngHD-DNA complex [2, 28]. Both models follow the general formula:

\[
\Delta G^{\text{binding}} = (5 + \delta) \Delta G_{\text{diss}} + \sum_{k=5}^{\infty} \Delta G_k \sum_{\beta = 0}^{N-5} \left[ \delta \beta_\beta + (\delta_\beta + \delta_{\beta \beta})/2 \right] \Delta G_{\beta \beta}
\]

(3)

where \( \delta \) is a Kronecker delta that takes a value of 1 for any binding site in a central position of the DNA (\( x \geq 10 \) and \( x \leq N-10 \)) or 0 for any binding site located less than 1 turn from either DNA end (\( x \leq 10 \) or \( x \geq N-10 \)) to account for end effects on the statistical interaction. \( \Delta G_{\text{diss}} \) is the free energy of the EngHD-DNA specific interactions with each base on site \( x \). In the PWM-based model, \( \Delta G_k \) is directly defined according to the 6 x 4 position weight matrix of the EngHD-DNA core consensus (Fig. 2b) as \( \Delta G_k = -RT \ln (p_k^\text{PWM} / 0.25) \), where \( p_k^\text{PWM} \) is the probability of base \( k \) to be found in position \( k \) of the site, and \( \Delta G_0 = 0 \). This is a model with 24 predetermined and one free fitting parameter (\( \Delta G_{\text{diss}} \)). In the structure-based model, \( \Delta G_k \) is defined by two types of specific interactions: one for each consensus base in the core tetrad (\( A_3, A_5, T_4, T_5 \)) and one interaction for any other T or A present in the site (\( A/T_1, T_2, T_5, A_3, A_5, \) and \( A/T_5 \) ) (asterisk in Fig. 3b).

#### Kinetic model of a gene transcription antenna

The model defines four different states of the TF in reference to its binding status to the DNA: (1) \( C_0 \) corresponds to the pool of TF on the cell milieu; (2) \( A_1 \) corresponds to the TF unbound and diffusing within the small cellular volume occupied by the DNA antenna; (3) \( D_{\text{up}} \) corresponds to the TF bound to the cluster of degenerate consensus repeats that conform the antenna; (4) \( B \) corresponds to the TF associated to the specific binding site. Obviously, in a real cell the TF will also have the opportunity to bind to other regions of the genome. However, from the point of view of the control of one gene, binding to other regions in the genome (other gene RRs controlled by the same TF and nonspecific binding to any DNA sequence) will simply decrease the overall availability of TF to be bound to the antenna, and thus it can be implicitly assumed as part of \( C_0 \). The four states are kinetically connected according to the scheme depicted in Fig. 6c. In this scheme, the unbound protein can get in and out of the antenna space by simple diffusion. This transport is modeled kinetically through the rate constants \( k_1 \) (onto the antenna) and \( k_4 \) (off the antenna) space by simple diffusion. This transport is modeled kinetically through the rate constants \( k_1 \) (onto the antenna) and \( k_4 \) (off the antenna). For a given set of rate constants, this rate matrix can be solved as an eigenvalue problem using standard matrix calculus. The three nonzero eigenvalues provide the kinetic phases of the model and the eigenvectors provide the kinetic amplitudes (or the equilibrium populations for the zero eigenvalue) for a given set of initial conditions (i.e., initial populations). To illustrate the functioning of such an antenna system with conditions that are functionally relevant, we parameterized the rate matrix as: (a) assuming a concentration of TF in the cell of 1 nM and a diffusion coefficient for a protein inside a cell of \( D = 3 \cdot 10^{-8} \) cm^2/s. The rate constant to get into the antenna from the cell milieu (\( k_1 \)) was set to \( 10^6 \) M^{-1}s^{-1}. For a given set of rate constants, this rate matrix can be solved as an eigenvalue problem using standard matrix calculus.
The rate constant to escape from the antenna \((k_{\text{E}})\) is defined by the ratio between the cell and antenna volumes \((1000 \mu\text{m}^2/0.033 \mu\text{m}^2)\) to keep detailed balance, thus resulting in a mean of 1. The antenna participation factors \(\beta_i\) components of high purity consensus sites, each one of them with \(K_{\text{c}} = 10^{-7} \text{ M}\) and a single specific binding site \((SB)\) with \(K_{\text{SB}} = 10^{-8} \text{ M}\). The dissociation rate constant from \(SB\) to \(A\) was set to \(k_{-1} = 10^{5} \text{ M}^{-1} \text{s}^{-1}\) and the global dissociation constant from \(D_{\text{p}}\) to \(A\) was set to \(k_{-2} = 10^{2} \text{ M}^{-1} \text{s}^{-1}\). The rate constants for binding from \(A_i\) to \(SB\) and \(D_{\text{p}}\) are directly set by detailed balance to \(k_{1} = 10^{-4} \text{ s}^{-1}\) and \(k_{2} = 8 \times 10^{-5} \text{ s}^{-1}\), respectively. Finally, \(SB\) and \(D_{\text{p}}\) can interconvert by 1D sliding. Using a sliding motion of approximately 200 bp·ms⁻¹ and a mean 200 bp separation of 2.5 Kbp between \(SB\) and any degenerate consensus site within \(D_{\text{p}}\) we estimated the mean rate constant to slide onto \(SB\) as \(k_{3} = 8 \times 10^{-5} \text{ s}^{-1}\) and thus, by detailed balance the rate to slide off \(SB\) is \(k_{4} = 0.89 \times 10^{-5} \text{ s}^{-1}\). Using these numbers, the three nonzero eigenvalues of the rate matrix \((E)\) are: \(E_1 = 2 \times 10^{-3}\), \(E_2 = 1.0 \times 10^{-3}\) and \(E_3 = 0.0099\). The equilibrium populations are \(c \approx 0.0099, A_3 = 3.3 \times 10^{-5}, D_0 = 0.018\), and \(SB = 0.0099\) (Fig. 6d). Analysis of the amplitudes (eigenvectors) indicates that the slowest nonzero eigenvalue \((\lambda_1)\) reflects the very slow re-equilibration between the cell milieu and the antenna. The fastest eigenvalue \((\lambda_3)\) reflects the extremely fast equilibration between the free and bound TF molecules within the antenna (the population of \(A_1\) is very small, and thus its equilibrium is nearly instantaneous). Finally, the intermediate eigenvalue \((\lambda_2)\) includes all the kinetic flux between \(SB\) and \(D_{\text{p}}\), and thus it indicates how much time it takes a TF molecule located within the antenna to find the specific binding site (slightly over 10 ms). The response time is thus very fast, even though the specific binding site occupancy is maintained at a minimum (about 1%). These results are incorporated onto the graphical representation of the β3 tubulin gene antenna of Fig. 6d.

Sequence analysis of engrafted transcription antennas. Engrailed is thought to regulate the expression of at least 203 genes whether by itself or with the participation of other transcription factors36. For the analysis of transcription sequences, the accession numbers for 203 genes that encode for TF molecules within the antenna (the population of \(A_1\) is very small, and thus its equilibrium is nearly instantaneous). Finally, the intermediate eigenvalue \((\lambda_2)\) includes all the kinetic flux between \(SB\) and \(D_{\text{p}}\), and thus it indicates how much time it takes a TF molecule located within the antenna to find the specific binding site (slightly over 10 ms). The response time is thus very fast, even though the specific binding site occupancy is maintained at a minimum (about 1%). These results are incorporated onto the graphical representation of the β3 tubulin gene antenna of Fig. 6d.

Fluorescent labeling for FCS experiments. EngrHd was labeled with Alexa 488 chroomophore at the C-terminus (manufactured by BioSciences, Inc. - Molecular Probes). Briefly, 3 mg of EngrHd was mixed by drop with 1 mg of dye to a final volume of 3 ml in PBS buffer in presence of 2 mM TCEP-HCl. The sample was incubated in total darkness for 2 h at room temperature or overnight at 4 °C, after which the reaction was stopped adding 2-Mercaptoethanol (SHAG). The protein was purified by anion exchange chromatography using a SP Sepharose Fast Flow column (GE Healthcare). The equilibrium buffer contained 25 mM Tris-HCl pH 7.5, 0.1 mM EDTA, 0.1 M NaCl and 2 mM TCEP-HCl. Elution was performed in the same buffer but increasing salt concentration in a gradient of NaCl concentration up to 0.5 M. Fractions containing EngrHd were analyzed by PAGE-SDS and dialyzed against PBS. The resulting EngrHd samples were subsequently concentrated using Amicon Ultra-15 3.000 MWL, Da (Millipore) and quantified by A280nm and PAGE-SDS.

Protein expression and purification. The coding sequence of Enghd (PDR: 3HDD) containing a Cys residue at the C-terminal was cloned from TopGene Technologies cloned into the pBRT 4 vector36. Sequence identity was confirmed by DNA sequencing. Recombinant protein expression was performed in transformed BL21 (DE3) cells (Novagen) grown at 37 °C/220 rpm and inducing with 1 mM IPTG at an OD600nm~1 for 4 h. Cells were pelleted, lysed by multiple freeze-thaw cycles in lysis buffer (100 mM Tris-HCl, pH 8, 200 mM KCl, 1 mM EDTA, 2 mM CaCl2, 2 mM MgCl2, 50% glycerol, 2 mM TCEP-HCL, 1 mM PMFS), and centrifuged at 30,000 rpm for 1 h. The supernatant was purified by cation exchange chromatography using a SP Sepharose Fast Flow column (GE Healthcare). The equilibrium buffer contained 25 mM Tris-HCl pH 7.5, 0.1 mM EDTA, 0.1 M NaCl and 2 mM TCEP-HCl. Elution was performed in the same buffer but increasing salt concentration in a gradient of NaCl concentration up to 0.5 M. Fractions containing Enghd were analyzed by PAGE-SDS and dialyzed against PBS. The resulting Enghd samples were subsequently concentrated using Amicon Ultra-15 3.000 MWL, Da (Millipore) and quantified by A280nm and PAGE-SDS.

fluorescent labeling for FCS experiments. Enghd was labeled with Alexa 488 chroomophore at the C-terminus (manufactured by BioSciences, Inc. - Molecular Probes). Briefly, 3 mg of Enghd was mixed by drop with 1 mg of dye to a final volume of 3 ml in PBS buffer in presence of 2 mM TCEP-HCl. The sample was incubated in total darkness for 2 h at room temperature or overnight at 4 °C, after which the reaction was stopped adding 2-Mercaptoethanol (SHAG). The protein was purified by anion exchange chromatography using a SP Sepharose Fast Flow column (GE Healthcare). The equilibrium buffer contained 25 mM Tris-HCl pH 7.5, 0.1 mM EDTA, 0.1 M NaCl and 2 mM TCEP-HCl. Elution was performed in the same buffer but increasing salt concentration in a gradient of NaCl concentration up to 0.5 M. Fractions containing Enghd were analyzed by PAGE-SDS and dialyzed against PBS. The resulting Enghd samples were subsequently concentrated using Amicon Ultra-15 3.000 MWL, Da (Millipore) and quantified by A280nm and PAGE-SDS.

Design of the DNA molecules. All the double stranded DNA molecules used in our experiments are based on the sequence of the β3 tubulin gene (Genbank ID: 37888), which is a natural gene from Drosophila melanogaster known to be regulated by Engrailed. Specifically, we choose a 150 bp region (nt 2525–2674, counting from exon 1) that contains only one high affinity (but not consensus) binding site (TAATTG) with a predicted affinity in the nM range that is ideal for FCS experiments. The designed DNA sequences contain the high affinity TAATTG site in their central position and include segments from the β3 tubulin gene sequence that expand in both directions to complete a total of 38, 75, or 150 bp for each probe. Sample purity was confirmed by MALDI-TOF mass spectrometry. The degree of labeling (−99%) was obtained as the ratio between fluorophore and protein determined by absorbance, comparing the A280nm for the fluorophore and the Abs280nm for the protein and applying a correction factor of 0.11 for the fluorophore’s contribution to A280nm (Molecular Probes).

Fluorescent labeling for FCS experiments. Enghd was labeled with Alexa 488 chroomophore at the C-terminus (manufactured by BioSciences, Inc. - Molecular Probes). Briefly, 3 mg of Enghd was mixed by drop with 1 mg of dye to a final volume of 3 ml in PBS buffer in presence of 2 mM TCEP-HCl. The sample was incubated in total darkness for 2 h at room temperature or overnight at 4 °C, after which the reaction was stopped adding 2-Mercaptoethanol (SHAG). The protein was purified by anion exchange chromatography using a SP Sepharose Fast Flow column (GE Healthcare). The equilibrium buffer contained 25 mM Tris-HCl pH 7.5, 0.1 mM EDTA, 0.1 M NaCl and 2 mM TCEP-HCl. Elution was performed in the same buffer but increasing salt concentration in a gradient of NaCl concentration up to 0.5 M. Fractions containing Enghd were analyzed by PAGE-SDS and dialyzed against PBS. The resulting Enghd samples were subsequently concentrated using Amicon Ultra-15 3.000 MWL, Da (Millipore) and quantified by A280nm and PAGE-SDS.

Sequence analysis of genes not under engrafted control. Identifying genes that are demonstrated not to be under Engrailed control is far from trivial because, as key transcription factor in development, Engrailed controls ubiquitous genes involved in many fundamental morphogenesis, communication and signaling processes, including genes that encode for signal proteins, receptors, protein kinases, transcription factors, transcription co-factors, transcription co-repressors, etc. We thus looked into the Interactive Fly database https://www.sdbonline.org/sites/ flyaimain/3a-dtest.htm, which contains a large number of D. melanogaster genes with detailed functional annotations, and eliminated all genes that are involved in any of those processes mentioned above and/or that are known to be controlled by any of the factors listed above (e.g., most homeoboxes). After applying this filter, we identified a group of 24 genes unlikely to be controlled by Engrailed based on their known function, but still with AT contents close to the 57% of the overall Drosophila genome.
DNA purification and hybridization. Single stranded DNA molecules corresponding to both strands of the 38-150 bp DNA molecules were ordered from Sigma. The ssDNA molecules (8–s) were re-suspended in sterile MilliQ water and their concentration determined by Abs260nm. Hybridization of the pairs of complementary strands were carried out by performing a temperature ramp from 95 °C to 4 °C in 45 min on a thermocycler. dsDNAs were stored at −20 °C in aliquots. The longer DNA molecules (300 bp and 600 bp) were synthesized by TOP Gene Technologies and delivered cloned into the pBABE vector. Plasmids were transformed into DH5α cells (Invitrogen) and purified by Maxiprep (Qiagen). The 300 and 600 bp dsDNA fragments were obtained by enzymatic digestion (Xhol/HindIII for 300 bp DNA or NcoI/HindIII for 600 bp DNA, NEB) and isolated from a 2% agarose gel using the Wizard SV Gel and PCR Clean-Up System (Promega). The 7219 bp DNA was obtained by enzymatic digestion (HindIII-KpnI) of the full β-tubulin gene cloned into the pUC57 vector (Bio Basic) and isolated from a 2% agarose gel using the Wizard SV Gel and PCR Clean-Up System (Promega). dsDNA bands were eluted in sterile MilliQ water and quantified by Abs260nm. These DNA molecules were re-hybridized and stored following the same protocol used for the shorter DNAs. Controls were set to determine the translational diffusion coefficient of the DNA molecules by FCS, we ordered 5′-fluorescently labeled (with Alexa 488) ssDNA of 33 and 75 bases from IBA-lifesciences and ATD bio, respectively. These ssDNA molecules were, hybridized with their complementary unlabeled strand and stored using the same protocol described above.

Glass coverslips preparation. Glass coverslips 24 × 24 mm #1 (Menzel-Glaser) were immersed in a solution containing 1vol. acetone/1vol. methanol/2vol. water and incubated for 5 min, followed by sonication for 30 min. in a bath with a 1 M KOH solution, a thorough rinse with water and immersion in acetone for 5 min. Cleaned coverslips were incubated for 5 min. in Vectabond reagent (Vector labs) using a ratio of 6:0.2 ml reagent for each ml of acetone. This product chemically modifies the glass to form a highly adherent surface. After Vectabond treatment, the coverslips were rinsed with water and stored in water (for no more than 2 months) until pegylation. As first step in the pegylation procedure, the coverslips were dried and pressed glued to a press-to-seal silicone isolator (Molecular Probes) while being dried. For the longest, non-rigid, DNAS (300 and 600-bp) we obtained the concentration determined by Abs260nm. Hybridization of the pairs of complementary strands were carried out by globally determining the overall diffusion coefficient of the DNA molecules by FCS, we ordered 5′-fluorescently labeled (with Alexa 488) ssDNA of 33 and 75 bases from IBA-lifesciences and ATD bio, respectively. These ssDNA molecules were, hybridized with their complementary unlabeled strand and stored using the same protocol described above.

FCS sample preparation. FCS experiments were performed at 296–298 K with 30 μl of solution (10 mM Tris-HCl pH 7.5, 0.1 mM EDTA, 50 mM NaCl) prepared with a fixed EngHD concentration of 2 or 5 nM. The dependence on the ionic strength was studied using the same buffer but increasing the NaCl concentration (150 mM or 350 mM). A series of experiments at various ssDNA concentrations was performed for each DNA molecule and condition (salt concentration) to ensure coverage of the entire binding isotherm (from the pM to the mM range). To avoid sample evaporation during the measurement, the well was covered with parafilm.

FCS instrument configuration. All measurements were carried out on a Microtime 200 confocal microscope (PicoQuant). A small fraction (~70 kV) of the 485 nm light emitted from a pulsed diode laser (Model LDH-D-C; 40 MHz pulse repetition rate and 50 ps pulse width) was reflected by a 510 nm long-pass dichroic mirror (Chroma) onto the back of the microscope objective (UPlanSApo 60 × /1.2 numerical aperture Olympus water immersion) and focused to a limited spot in the sample. A solution of 0.2 mg/μL of NHS-PEG in 100 mM sodium borate at pH 8.5 was added to each well and incubated for 3 h. This solution was removed after the 3 h incubation and the wells were rinsed with abundant water. The entire process was performed at room temperature, the used reagents were spectroscopic grade and the water was passed through a 0.22 μm filter (Millipore). The pegylated coverslips were stored in a closed box at 277 K with a drop of water removed after the 3 h incubation and the wells were rinsed with abundant water.

Confocal volume calibration. The confocal volume was estimated to be 1.6 femtoliters for all the experiments described in this work.

Theoretical calculations of diffusion coefficients. Estimates of the diffusion coefficient for EngHD and for the EngHD-DNA complexes (assuming diffusion of the complex can be approximated by the diffusion of the dsDNA molecule alone given the ssDNA content expected for the four proteins of rigid ~7 kDa) were obtained from theoretical calculations with the software Hydroprop (for EngHD and all DNAs up to 150-bp, which is the limit for the rigid rod approximation given the DNA’s persistence length) and from experimental determination from the autocorrelation decay in FCS (for EngHD and dDNAs of 33-bp and 75-bp labeled with Atto488). Agreement between estimates from both methods was excellent.

Determination of the Kd from FCS experiments. Determination of the Kr for a given EngHD-DNA complex was carried out by globally fitting the autocorrelation function decays for a series of FCS experiments at varying concentration of DNA in which the decay for each experiment at a given DNA concentration was analyzed using equation:

\[ G(t) = G_0(t) \times G_1(t) \times G_2(t) \]

In this equation, the contribution to the fluorescence autocorrelation function corresponding to the diffusion of labeled EngHD molecules (free and in complex with DNA) in and out of the confocal volume is represented by

\[ G_0(t) = \left( \frac{1}{\sigma^2} \right) \left( 1 + \left( \frac{t}{\tau_{\text{prot}}} \right)^2 \right)^{-2} \]

where \( N^{\text{free}} = (N_{\text{prot}} + N_{\text{comp}}) \) is the average number of molecules (combination of free and complexed EngHD), \( \tau_{\text{prot}} = \omega_{\text{prot}}/D_{\text{prot}} \) is the mean diffusion time for fluorescently labeled EngHD in free form, \( \tau_{\text{comp}} = \omega_{\text{comp}}/D_{\text{comp}} \) is the mean diffusion time for the labeled DNA alone (i.e., assuming the same diffusion coefficient for DNA alone and protein-DNA complex, see above), \( \omega \) is the dye brightness and

\[ P_{\text{prot}} = \frac{1}{1 + K_r D_{\text{prot}}} \]

where \( K_r = \frac{K_{\text{prot}}}{K_{\text{DNA}}} \) is the fraction of EngHD present in free form at each concentration of DNA. The confocal volume dimensions \( \omega_{\text{prot}} \) and \( \omega_{\text{comp}} \) were determined using Atto488 as a reference of known diffusion coefficient and brightness (as described above).

The component of the fluorescence autocorrelation function due to triplet build up is obtained as

\[ G_1(t) = \left( 1 + A_2 e^{-t/\tau_{\text{2}}} \right) \]

where \( A_2 = \frac{1}{\tau_{\text{2}}} \) is the fraction of A488 molecules in the triplet state. And the contributions to the fluorescence autocorrelation function arising from the presence of fluorophore molecules not associated to EngHD are obtained as

\[ G_2(t) = \left( 1 + A_3 e^{-t/\tau_{\text{3}}} \right) \]

where \( A_3 = \frac{1}{\tau_{\text{3}}} \) is the fraction of A488 molecules that are detached from EngHD.

Fluorescence autocorrelation decays were delimited in time to eliminate after-pulsing effects. The complete series of FCS decays for EngHD in association with a given DNA molecule (varying the DNA concentration) were fitted globally to Eqs. (8)–(12) in which the diffusion coefficient for the free protein and for the complex, the apparent \( K_r \) for binding, fraction of A488 molecules detached from EngHD, the dye brightness on the free protein, and dye brightness on the complex, are all global parameters. The diffusion coefficient for free EngHD and EngHD-DNA complex were bound to the values that we determined in independent FCS experiments and theoretical considerations using Hydroprop. For EngHD we obtained a value of 122 μm² s⁻¹, in good agreement with the value of 131 μm² s⁻¹ estimated by Hydropro. For the longest, non-rigid, DNA (300 and 600-bp) we obtained the bounded values for the diffusion coefficient from values previously reported by other authors and from independent FCS measurements performed in conditions of DNA saturation (100% complex). The non-linear fitting of the fluorescence correlation function to Eqs. (8)–(12) performed using a custom-built MATLAB routine implemented with the lsqcurvefit function for least-squares optimization. Equations (8)–(10) assume a simple two-state model for the binding of EngHD to DNA (each EngHD molecule is either free or bound), which is a reasonable assumption for our EngHD-DNA complex. However, when EngHD-DNA is labeled, the much larger DNA molecule determines the overall diffusion coefficient of the complex, and the experiments are performed at DNA concentrations equal or larger than the protein concentration (to ensure that the probability of two EngHD molecules binding to the same DNA molecule is small). The fitting procedure was performed for each series of FCS experiments at different concentrations of a given DNA molecule (FCS titrations). FCS titrations were repeated three independent times, and the fit was performed for each experiment independently. The fits were carried out expressing the \( K_r \) in base 10 logarithms to ensure linear sampling and fit convergence. The global fit to each individual FCS titration rendered an estimate of the \( K_r \) and uncertainty (at 95% confidence, or two standard deviations). The reported values are weighted averages of the \( K_r \) and uncertainty (at 95% confidence) from the multiple FCS titrations performed for each DNA molecule. The weighted mean \( K_r \), uncertainty, \( K_r \) estimated for each individual titration experiment and their statistical weight are given in Supplementary Table 2. The statistical weight for each FCS titration of a given molecule was defined as the inverse of the squared uncertainty obtained from the global fit to all the FCS curves at different DNA concentrations. The statistical
weights for each experiment were divided by the sum of the inverse of squared uncertainties for the three experiments to ensure proper normalization.

Comparison with in vivo ChIP-seq profile. ChIP-seq experiments on transgenic flies expressing en-GFP fusion proteins, and IP using an anti-GFP antibody were retrieved from the ENCODE Project database (experiment ENCSR952TDU) and profiled for EngHD binding in vitro using the structure-based statistical mechanical model described in this work. Optimal idr thresholded peaks (file ENCF6809M1, dm6 D. melanogaster last whole genome release in 2014) resulted in a collection of 2226 DNA fragments (<500 bp) that were translated to FASTA. The sequences of the 2226 fragments were profiled for in vitro binding to EngHD with the statistical mechanical model. The overall dissociation constant (i.e., for the entire fragment) was calculated for each fragment.

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

Data availability. All data and the gene expression plasmid containing the EngHD are available upon request. The sequence of all the DNA molecules used in this study are given as supplementary information (Supplementary Table 1). DNA molecules from 38 to 150 bp were produced by chemical synthesis and purchased directly from Sigma–Aldrich (Missouri). DNA molecules of 300 bp and 600 bp were synthesized and cloned into the pBAT 4 vector by Top Gene Technologies (Canada). The 7219 bp full β3 tubulin was synthesized and cloned into the pUC57 vector by Basic (Canada). The raw data resulting in Figs. 1c, 2a–c, 4b, 5a–c, 6a, b, 7a, b and Supplementary Figs. 1–3 are provided as a Source Data File.

Code availability. Analysis of the individual FCS curves, the global fitting of FCS curves obtained at different DNA concentrations for each DNA molecule to determine their experimental $K_D$, the two versions of the statistical mechanical model for EngHD binding to DNA, routines for the global fitting of the experimental $K_D$ values for all DNA molecules with the statistical mechanical models, and the kinetic model of transcription antenna have been performed using custom code for Matlab. All the Matlab scripts are available upon request to the corresponding author.

Received: 18 June 2018; Accepted: 12 December 2019; Published online: 28 January 2020

References

1. Bintu, L. et al. Transcriptional regulation by the numbers: models. Curr. Opin. Genet. Dev. 15, 156–165 (2005).
2. Milo, R., Jorgensen, P., Moran, U., Weber, G. & Springer, M. BioNumbers—the database of key numbers in molecular and cell biology. Nucleic Acids Res. 38, D106 (2010).
3. Cairns, B. R. The logic of chromatin architecture and remodelling at cell division. Cell 124, 156–166 (2006).
4. Wang, Y. M., Austin, R. H. & Cox, E. C. Single molecule measurements of protein translocation on nucleic acids. 1. Models and theory. J. Biol. Chem. 276, 41240–41246 (2001).
5. Blainey, P. C. et al. Nonspecific protein binding to repetitive DNA sequence elements signiﬁcantly affects eukaryotic genomes. PLoS Comput. Biol. 11, e1004429 (2015).
6. Malin, J. et al. Crowdsourcing: Spatial clustering of low-afﬁnity binding sites ampliﬁes in vivo transcription factor occupancy. BioRxiv (2015). https://www.biorxiv.org/content/10.1101/024398v1.
7. Shvets, A. A. & Kolomeisky, A. B. Sequence heterogeneity accelerates protein search for targets on DNA. J. Chem. Phys. 143, 12B639–631B639 (2015).
8. Lange, M., Kochugaeva, M. & Kolomeisky, A. B. Dynamics of the protein search for targets on DNA in the presence of traps. J. Phys. Chem. B 119, 12410–12416 (2015).
9. Ingham, P., Martinez-Arias, A., Lawrence, P. A. & Howard, K. Expression of engrailed in the parasegment of Drosophila. Nature 317, 634–636 (1985).
10. Morata, G. & Lawrence, P. A. Control of compartment development by the engrailed gene in Drosophila. Nature 253, 614–617 (1975).
11. Solano, P. J. et al. Genome-wide identification of in vivo Drosophila Engrailed-binding DNA fragments and related target genes. Development 130, 1243–1254 (2003).
12. Ades, S. E. & Sauer, R. T. Differential DNA-binding speciﬁcity of the engrailed homeodomain: the role of residue 30. Biochemistry 33, 9187–9194 (1994).
13. Serrano, N., Brock, H. W. & Maschat, F. beta3-tubulin is directly repressed by the engrailed protein in Drosophila. Development 124, 2527–2536 (1997).
14. Hinz, U., Wolk, A. & Renkawitz-Pohl, R. Ultrabithorax is a regulator of beta 3 tubulin expression in the Drosophila visceral mesoderm. Development 116, 543–554 (1992).
15. Kissingcr, C. R., Liu, B. S., Martin-Blanco, E., Kornberg, T. B. & Pabo, C. O. Crystal structure of an engrailed homeodomain-DNA complex at 2.8 A resolution: a framework for understanding homeodomain-DNA interactions. Cell 63, 579–590 (1990).
16. Shokri, L. et al. A comprehensive drosophila melangaster transcription factor interactome. Cell Rep. 27, 935–970.e957 (2019).
17. Dragan, A. I. et al. Forces driving the binding homeodomains to DNA. Biochemistry 45, 141–151 (2006).
18. Fraenkel, E., Roold, M. A., Chambers, K. A. & Pabo, C. O. Engrailed homeodomain-DNA complex at 2.2 A resolution: a detailed view of the interface and comparison with other engrailed structures. J. Mol. Biol. 284, 351–361 (1998).
19. Loregian, A., Sinigaglia, E., Mercorelli, B., Palu, G. & Coen, D. M. Binding parameters and thermodynamics of the interaction of the human cytomegalovirus
DNA polymerase accessory protein, UL44, with DNA: implications for the processivity mechanism. Nucleic Acids Res. 35, 4779–4791 (2007).
45. Mou, Y., Yu, J. Y., Wannier, T. M., Guo, C. L. & Mayo, S. L. Computational design of co-assembling protein-DNA nanowires. Nature 525, 230–233 (2015).
46. Onuchic, J. N. & Wolynes, P. G. Theory of protein folding. Curr. Opin. Struct. Biol. 14, 70–75 (2004).
47. Ferreiro D. U., Komives E. A., Wolynes P. G. Frustration in biomolecules. Quart. Rev. Biophys. 47, 285–363 (2014).
48. Sosinsky, A., Honig, B., Mann, R. S. & Califano, A. Discovering transcriptional regulatory regions in Drosophila by a nonalignment method for phylogenetic footprinting. Proc. Natl Acad. Sci. USA 104, 6305–6310 (2007).
49. Yao, P. et al. Coexpression networks identify brain region-specific enhancer RNAs in the human brain. Nat. Neurosci. 18, 1168–1174 (2015).
50. Lukaess, G. L. et al. Site-dependent DNA mobility in cytoplasm and nucleus. J. Biol. Chem. 275, 1625–1629 (2000).
51. Gramates, L. S. et al. FlyBase at 25: looking to the future. Nucleic Acids Res. 45, D663–D671 (2017).
52. Hihara, S. et al. Local nucleosome dynamics facilitate chromatin accessibility across regulatory regions in Drosophila by a nonalignment method for phylogenetic footprinting. Proc. Natl Acad. Sci. USA 108, 892–898 (2011).
53. Chu, X., Muhoz, V. Roles of conformational disorder and downhill folding in modulating protein–DNA recognition. PCCP, 19 28527–28539 (2017).
54. MacArthur, S. et al. Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. Genome Biol. 10, R80 (2009).
55. Makunin, I. V., Kolesnikova, T. D. & Andreyenkova, N. G. Underreplicated regions in Drosophila melanogaster: are enriched with fast-evolving genes and highly conserved noncoding sequences. Genome Biol. Evol. 6, 2050–2060 (2014).
56. Zhou, H. X. Rapid search for specific sites on DNA through conformational switch of nonspecifically bound proteins. Proc. Natl Acad. Sci. USA 108, 8651–8656 (2011).
57. Perez, N. M. expressed, purified and labeled the protein EngHD, and performed and analyzed all the fluorescence correlation spectroscopy experimental data of EngHD binding to DNA but for the full-length β tubulin DNA. N.M. expressed and labeled the protein EngHD for binding studies to the 7219 kbp-long β tubulin DNA, prepared samples, and performed and co-analyzed the fluorescence correlation spectroscopy experimental data of EngHD binding to the β tubulin DNA. V.M. designed and coordinated the project, supervised, co-analyzed and interpreted the F.C.S. binding data, developed the statistical mechanical model for EngHD binding to DNA and the kinetic model of transcription antennas, and carried out the data fitting and overall statistical analysis. All authors participated in writing the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41467-019-14217-8.

Correspondence and requests for materials should be addressed to V.M.

Peer review information Nature Communications thanks Anatoly Kolomeisky, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020