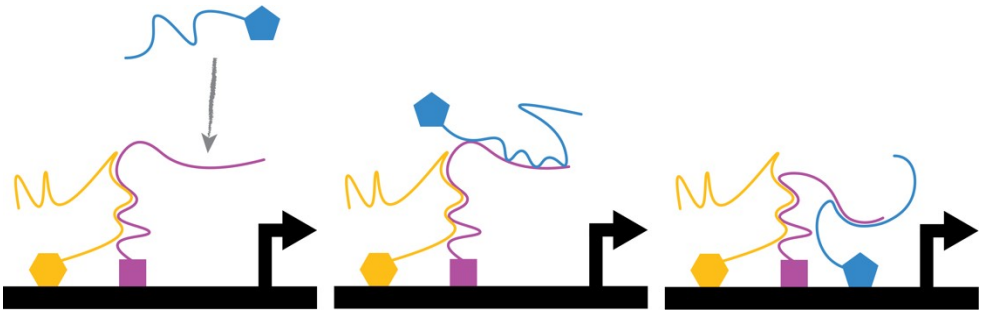
Transcription factors perform a two-step search of the nucleus

Max Valentín Staller
Center for Computational Biology, University of California, Berkeley, CA 94720
mstaller@berkeley.edu

Abstract:
Transcription factors regulate gene expression by binding to regulatory DNA and recruiting regulatory protein complexes. The DNA-binding and protein-binding functions of transcription factors are traditionally described as independent functions performed by modular protein domains. Here, I argue that genome binding can be a two-part process with both DNA-binding and protein-binding steps, enabling transcription factors to perform a two-step search of the nucleus to find their appropriate binding sites in a eukaryotic genome. I support this hypothesis with new and old results in the literature, discuss how this hypothesis parsimoniously resolves outstanding problems, and present testable predictions.

Key words:
Transcription factor, intrinsically disordered region, DNA binding domain, transcriptional condensate, liquid-liquid phase separation, transcription, gene regulation, transcriptional activation domain, transcription initiation



Main Text
Transcription factors have two jobs: binding DNA and regulating transcription. Site-specific transcription factors bind short DNA sequences, called motifs, with DNA binding domains. Eukaryotic transcription factors regulate transcription with effector domains that bind to regulatory complexes: repression domains bind corepressors and activation domains bind coactivators. Transcription factors have other functions, but most of their other domains (e.g., dimerization domains, degrons, and ligand-binding domains) modulate DNA binding or coregulator binding. In this review, I argue that the standard model is incomplete and that some transcription

factors search the nucleus in a two-step process. These transcription factors use protein-protein interactions to perform a *global search* of the nucleus to find a ‘protein cloud’ and then use DNA binding domains to perform a *local search* of the DNA within that protein cloud. This expanded model is motivated by examples where deleting the DNA binding domain does not prevent transcription factors from localizing to the correct promoters (Brodsky *et al.* 2020; Gera *et al.* 2022), which I discuss in detail below. The global search with protein-protein interactions localizes the transcription factor to the right region of the nucleus, and then the DNA binding domain scans the DNA in that region and dwells on the cognate motif. Critically, the protein-protein interactions that perform the global search for the protein cloud require protein sequences outside the DNA binding domain.

I have chosen the term ‘protein cloud’ to emphasize that this idea is still cloudy. I am picturing a non-stoichiometric cluster of transcription factors engaged in both homotypic interactions between multiple copies of the same transcription factor and heterotypic interactions between different transcription factors. This cluster may or may not include coactivator proteins, which could, in principle, bridge multiple TF molecules (Tuttle *et al.* 2018; Sanborn *et al.* 2021). I am not invoking a large, energetically stable liquid-liquid-phase-separated droplet, but something more dynamic, in line with the original definition of a condensate or with a transcription factor hub (Shin and Brangwynne 2017; Chong *et al.* 2018). I am picturing dozens of molecules, not hundreds. In plants, the AUXIN RESPONSE FACTOR (ARF)7, ARF19 and EARLY FLOWERING3 (ELF3) transcription factors each become inactive when they enter a condensate (Powers *et al.* 2019; Jung *et al.* 2020). In human cell culture, much of the attention on transcriptional condensates has focused on transcriptional activation. Although I assume a rather explicit mechanism for transcriptional activation (see below), this hypothesis is not about activation. Instead, it addresses the problem of selecting active regions of the genome. It is related to the problem of identifying where transcriptional condensates or hubs form, which is the same as the old problem of why a region of the genome is an active enhancer in one cell type and inert in another.

In transcription factor biology, we know a lot more about DNA binding domains than we know about the rest of the protein. DNA binding domains are structured, conserved, and predictable based on protein sequences (Latchman 2008; El-Gebali *et al.* 2019). DNA binding domains are the basis for transcription factor family organization schemes (Lambert *et al.* 2018). There are many methods for measuring protein-DNA interactions *in vitro* and *in vivo* (Stormo 2013). Outside of the DNA binding domain, transcription factors are primarily composed of long intrinsically disordered regions (IDRs) that do not fold into a single 3D structure and instead exhibit multiple conformations (Liu *et al.* 2006; van der Lee *et al.* 2014). The sequence of an IDR controls whether these ensembles are expanded, collapsed, or form

hairpins (Das and Pappu 2013). The nomenclature in the literature is confusing: some IDRs have been called Low Complexity Domains because they contain only a few types of amino acids (Chong *et al.* 2018; Cascarina *et al.* 2020). The terms activation domain, transactivation domain, or activator domain have been used to refer to everything outside of the DNA binding domain or to minimized, highly active regions (Latchman 2008; Staller *et al.* 2018; Tycko *et al.* 2020). Here, I use the term activation domain to refer to short, highly active regions that directly contact coactivators, and I use the term IDR to refer to extended regions outside of DNA binding domains and other folded domains. I use the term IDR to refer to regions described as the ‘non-DBD’ by Brodsky *et al.* 2020.

Classically, it was argued that DNA binding domains and activation domains were independent, modular components, but this idea is approaching the end of its usefulness. In the few cases that have been carefully examined, activation domains can modulate DNA affinity, increase specificity for cognate motifs, or increase affinity for random DNA (Liu *et al.* 2008; Krois *et al.* 2018; Baughman *et al.* 2022). For the remainder of this piece, I assume that true modularity is rare. All activation domains are disordered in solution, and many fold upon binding to partners (Dyson and Wright 2016). The one known exception is IRF3, which is natively folded (Qin *et al.* 2003). There are a handful of well studied repression domains, notably the KRAB and POZ/BTB domains, but aside from these two types, there are no good predictors of repression domains (Bintu *et al.* 2016; Soto *et al.* 2021). There is a rich body of work examining activation domain coactivator interactions with NMR; for example, p53, RelA, the ETV family, Hif1a, and CITED2 (Dyson and Wright 2016; Raj and Attardi 2017; Currie *et al.* 2017; Berlow *et al.* 2022) in human and Gcn4 and Gal4 in yeast (Brzovic *et al.* 2011; Hahn and Young 2011; Tuttle *et al.* 2021). There has been some progress predicting acidic activation domains from protein sequence in yeast and human proteomes (Ravarani *et al.* 2018; Erijman *et al.* 2020; Sanborn *et al.* 2021; Staller *et al.* 2022), but it has been difficult to distill the features of other classes, such as proline-rich or glutamine-rich activation domains (Latchman 2008). In recent work, I argued the critical sequence feature of acidic activation domains is the balance between acidic residues and aromatic and leucine residues (Staller *et al.* 2022).

This two-step nuclear search hypothesis is motivated by a result from Naama Barkai and colleagues showing how IDRs of Msn2 and Yap1 are necessary and sufficient for targeting a transcription factor to the correct promoter in yeast (Brodsky *et al.* 2020, Gera *et al.* 2022). This hypothesis is further influenced by single-molecule imaging of transcription factor dynamics in living nuclei, where the IDRs of Hif1a and Hif2a are necessary and sufficient to control the fraction of molecules bound to chromatin and the diffusion rates of mobile molecules (Chen *et al.* 2021). However, this hypothesis can also explain several puzzling results from genomics over the last two

decades and reemphasizes outstanding questions. In the following sections, I develop this hypothesis, contrast it with several models in the literature, and discuss testable predictions.

Assumptions

Implicit in the two-step nuclear search hypothesis are several assumptions about how transcription factors work together to activate transcription. First, I assume a thermodynamic framework, where protein-protein interactions and transcription factor-DNA interactions occur quickly enough to come to equilibrium. Protein clouds can nucleate anywhere, but they preferentially accumulate at genomic sites with many transcription factor binding sites. Traditionally, the thermodynamic framework assumed constant microscopic on-rates and slower off-rates at cognate sites, but there is accumulating evidence that DNA sequence modulates transcription factor-DNA on rates (Marklund *et al.* 2022). Second, I assume a key feature of transcriptional regulation is enhancer occupancy, or the total fraction of time an enhancer is bound by transcription factors (and not the residence times of individual molecules, which are generally less than 15 seconds) (Sherman and Cohen 2012; Stormo 2013; Chen *et al.* 2014, 2021; Hansen *et al.* 2018). Genome specificity is achieved thermodynamically by equilibrium binding of transcription factors. Third, I assume that all transcriptional regulation is combinatorial: namely, that multiple transcription factors must simultaneously achieve high occupancy to activate transcription. It is not yet clear whether each transcription factor brings in a different coactivator or if multiple transcription factor molecules together recruit one coactivator (e.g. a p53 tetramer binding four domains of p300 (Ferreon *et al.* 2009)). Fourth, I assume that an enhancer acts as a scaffold to bring together the multiple biochemical activities necessary to progress through the steps of the transcription cycle (e.g. opening chromatin, assembling the basal transcriptional machinery, forming the polymerase initiation complex, initiating polymerase, and releasing paused polymerase) (Fuda *et al.* 2009). While it is clear that there is more than one step in transcription, it is not clear how many of these steps are near rate-limiting at a given gene. For a thorough and highly accessible discussion of kinetic control of transcription see Scholes *et al.* 2017. Fifth, I assume that multivalent binding ‘cycles’ that bridge multiple molecules are a critical feature: transcription factors simultaneously bind DNA and other proteins and simultaneous release of all contacts is rare, slowing transcription factor escape from a protein cloud (Deeds *et al.* 2012, Sanborn *et al.* 2021). Sixth, I will assume that histone modifications are the time integral of recent transcription factor binding activity, serving as a short-term memory for occupancy (Long *et al.* 2016).

A new phenomenon requires a new model

The crucial new data motivating the two-step nuclear search hypothesis is the recent work from Naama Barkai and colleagues (Brodsky *et al.* 2020) showing that long IDRs are necessary and sufficient to target Msn2 and Yap1 to the correct promoters in yeast. Critically, the DNA binding domain is dispensable for targeting to the correct promoter: transcription factors lacking the DNA binding domain lost the sharp peak in binding signal over the DNA motif, but they retained substantial binding throughout the

promoter. The integral of the binding signal over the full promoter was largely unchanged between full length Msn2 and the DNA binding domain deletion. In contrast, the Msn2 DNA binding domain alone bound some, but not all, of the same promoters and bound to new promoters. For promoters that retained binding of the DNA binding domain only, the integral of the binding signal was reduced and the remaining binding shifted to motifs in the nucleosome free region (the ~100 bp upstream of a transcription start site). For Msn2, the binding signal over the promoter decreased as the IDR was shortened. Notably, the annotated activation domains were dispensable for proper promoter targeting. One important coactivator subunit, Med15, was also dispensable for proper promoter targeting. In reciprocal chimeras that exchanged the IDRs and DNA binding domains of Msn2 and Nrg2, the IDR dominated promoter selection. This result upends the classical picture of a modular transcription factor where the DNA binding domain is solely responsible for localization to the correct genomic locations.

The two-step nuclear search hypothesis can explain this result: the IDR localizes the transcription factor to the protein cloud at the correct target promoters and the DNA binding domain scans this promoter and binds to its cognate motif. AD-coactivator interactions may contribute to localizing a transcription factor to the right protein cloud, but they are neither necessary nor sufficient (Brodsky *et al.* 2020). Targeting the transcription factor to the protein cloud requires additional protein-protein interactions. I anticipate these interactions will include both homotypic interactions between multiple copies of the same transcription factor and heterotypic interactions between different transcription factors. There is direct evidence for homotypic clusters of Sp1, Mig1, and Msn2 (Su *et al.* 1991; Wollman *et al.* 2017; Chong *et al.* 2018). This IDR-mediated nuclear search is primarily used to find existing protein clouds at specific genomic locations, not nucleate new ones. I discuss below how these protein clouds nucleate at specific genomic regions.

Importantly, Brodsky *et al*. could not detect this phenomenon with traditional ChIP-seq and required a more sensitive method, ChEC-seq (Brodsky *et al.* 2020). Independent work using Calling Cards, an orthogonal method, found that for two paralogous yeast transcription factors, regions outside the DNA binding domain control targeting to the correct promoters (Shively *et al.* 2019). Gera *et al*. examined 30 pairs of transcription factor paralogs and showed that for 18 pairs, genomic localization is determined primarily by regions outside the DNA binding domain (Gera *et al.* 2022). The remaining 12 behaved like traditional transcription factors, with the DNA binding domain determining promoter selection.

It is likely that Chen *et al*. are observing the same phenomenon as Brodsky *et al*. and Gera *et al.* at the single molecule level (Chen *et al.* 2021). By comparing chimeras of two paralogous transcription factors, they have shown that the fraction of molecules immobilized on the chromatin and the

diffusion rate of mobile proteins are determined primarily by the IDR and not the DNA binding domain. The different diffusion rates of the mobile fractions can be explained by the IDRs orchestrating distinct constellations of protein-protein interactions; namely, distinct clusters that wander the nucleus at different rates. The changes in the fraction of molecules bound to chromatin is hard to rationalize without something akin to the two-step nuclear search hypothesis. The two-step nuclear search explains both of these single molecule phenomena.

A two-step search solves old problems

Invoking a two-step nuclear search solves three old problems: 1) Why do only a minority of residues in transcription factors have known functions? 2) Why are only a tiny fraction of transcription factor motifs in a metazoan genome bound *in vivo*? 3) Why do many genome regions detected by ChIP-seq assays not contain motifs for the precipitated transcription factor?

First, the known functional domains in most transcription factors cover only a minority of residues (Lambert *et al.* 2018; Soto *et al.* 2021). Most eukaryotic transcription factors have a short, structured, and conserved DNA binding domain, while the majority of the protein is intrinsically disordered and poorly conserved. Even in well-characterized transcription factors, the known activation domains, repression domains, ligand binding domains, dimerization domains, and other Pfam domains cover only the minority of residues (Soto *et al.* 2021). What is the rest of the protein doing? Some of these residues are flexible linkers between activation domains and are necessary for multivalent, fuzzy binding to coactivators (Harmon *et al.* 2017; Tuttle *et al.* 2018). However, we should be skeptical of the idea that the majority of residues in a transcription factor are linkers. We must also grant that most effector domains are not yet annotated, but known examples are short, with a median length of 91 residues (Soto *et al.* 2021). Under the two-step nuclear search hypothesis, some of these long IDRs bind other IDRs to localize transcription factors to a protein cloud at target promoters. Metazoan transcription factors have expanded IDRs (Liu *et al.* 2006; Jana *et al.* 2021), which may result from neutral drift (Lynch *et al.* 2016) but may enable the expansion of protein-protein interactions that accompanied multicellularity (Dunker *et al.* 2015). There is evidence that long IDRs can mediate homotypic and heterotypic interactions that cause clustering in the nucleus (Chong *et al.* 2018; Boija *et al.* 2018). Under the two-step nuclear search hypothesis, the unannotated regions of IDRs perform the global search.

Second, how do transcription factors avoid getting lost in the genome? Only a tiny fraction of predicted transcription factor binding sites in a metazoan genome are bound by a transcription factor: there are millions of predicted motifs, thousands of which are bound in ChIP-seq assays and a subset of which are active in reporter gene assays. What distinguishes the bound sites

from the unbound sites? This problem has enthralled genomicists for over 20 years (Harbison *et al.* 2004; Harrison *et al.* 2011; White *et al.* 2013). For a thorough review of the specificity problem see Brodsky *et al.* 2021. This problem has been formalized with information theory: metazoan genomes are large and transcription factor motifs are short, so there is not enough information in a single motif occurrence to uniquely define genomic addresses (Wunderlich and Mirny 2009). In the human genome, a cluster of 10-15 sites are necessary to uniquely encode a 500-1000 bp genomic location. In the two-step nuclear search hypothesis, the IDR performs the global search, contributing additional information to find the right loci. Once the transcription factor is in the protein cloud, the DNA binding domain is only responsible for the local search of a much smaller amount of DNA. The local search then becomes efficient, leading to high occupancy and sharp peaks over cognate motifs in ChEC-seq (Brodsky *et al.* 2020). The two-step search similarly explains how large clusters of Ultrabithorax (Ubx) protein can accumulate at low-affinity transcription factors binding sites that control development of bristles in fly (Crocker *et al.* 2015). A protein cloud with dozens of members, each with an expanded IDR, also offers a larger search target than a single DNA binding site.

Third, genome-wide ChIP-seq data contain a second paradox: many peaks do not contain a DNA motif for the precipitated transcription factor. By some estimates 30-70% of called ChIP-seq peaks do not contain a motif for the precipitated transcription factor (Harrison *et al.* 2011; Spitz and Furlong 2012; reviewed in Jana *et al. 2021*). There are at least three classes of peaks without motifs: 1) “Hyperchipable” regions caused by DNA/RNA hybrids, high expression, and other fixation artifacts (Teytelman *et al.* 2013). 2) Highly occupied target (HOT) regions of highly open chromatin that are bound by practically every transcription factor and are sometimes computationally removed as an artifact (Kvon *et al.* 2012). 3) True enhancers bound by partner transcription factors. The third class motivated the *transcription factor collective model*: active enhancers are bound by a group of cell-type specific transcription factors that together activate expression (Spitz and Furlong 2012). Any given enhancer has binding sites for most but not all transcription factors in this group. Under the two-step nuclear search hypothesis, a transcription factor will spend significant time in all compatible protein clouds, not just those with cognate binding sites, and these clouds will provide ChIP-seq signal. Some will consider this two-step nuclear search hypothesis to be a restatement of the transcription factor collective model, but I argue below this hypothesis makes several more precise predictions.

Additional support from the literature

Further support of the two-step nuclear search hypothesis comes from ChIP-exo and single-particle tracking experiments on transcription factor mutants that remove the IDR or mutate the DNA binding domain (Chen *et al.* 2014). Compared to the full-length protein, the Sox2 DNA binding domain

alone spent less time in 3D diffusion, had double the number of ChIP-exo peaks, and its mean dwell time on chromatin was shorter. This result was interpreted as more binding to ‘pseudotargets’ with lower quality motifs (more ChIP-exo peaks and shorter binding times to these lower quality motifs). The reciprocal perturbation, a mutation disrupting the Sox2 DNA binding domain, still bound ~26% of original genomic loci, showing that the IDR is sufficient for genomic localization, similar to Msn2 in yeast (Brodsky *et al.* 2020). Compared to the full protein, the DNA binding domain-inactivating mutant spent more time in 3D diffusion, had a lower fraction of immobilized molecules, and these immobile molecules had longer dwell times. These results imply that the IDR is reducing binding to incorrect genomic loci, either by increasing time spent in protein clouds at the correct loci or by other means (like directly competing with the DNA binding domain (Krois *et al.* 2018)). The results are not intuitive but can be interpreted as follows: the DNA binding domain contributes both short-lived binding at random DNA and medium-lived binding at motifs, while the IDR contributes long-lived binding to protein clouds. The WT protein is a convolution of these three binding modes. Reciprocally, WT 3D search can be interrupted by DNA binding to a true motif, nonspecific DNA binding to random open DNA, or IDR binding to a protein cloud. Under the two-step nuclear search hypothesis, the interpretation of these data is that the protein-protein interactions that retain transcription factors in protein clouds have slower off-rates (longer dwell times) than DNA binding interactions at low quality motifs. Also consistent with the two-step nuclear search, single-particle tracking of the glucocorticoid receptor observed low mobility (confined) and chromatin-bound states (Garcia *et al.* 2021). Deleting the IDR caused a loss of the confinement state and the majority of ChIP-seq peaks.

Relationship to other models

The two-step nuclear search hypothesis is a reimagining of the *Transcription Factor Funnel Model* where the funnel is protein-protein interactions instead of DNA (Castellanos *et al.* 2020). In the DNA funnel model, partial transcription factor binding sites near a ‘real’ transcription factor binding site can slow down a DNA binding domain during 1D scanning of DNA, effectively concentrating the transcription factor near the real binding sites (Wunderlich and Mirny 2008). The transcription factor funnel model has always been hard to rationalize with eukaryotic chromatin and its short regions of naked DNA between histones. The observed partial sites can just as easily be the product of binding site turnover (Ludwig *et al.* 2000; Hare *et al.* 2008). By contrast, the two-step nuclear search hypothesis uses protein-protein interactions rather than DNA binding domain-DNA interactions to concentrate protein at active enhancers.

The two-step nuclear search hypothesis is compatible with the original formulation of the Pioneer Factor Hypothesis. Pioneer factors are transcription factors with specialized DNA binding domains and specialized

activation domains that bind closed chromatin and open it up for other transcription factors, defining the active enhancer landscape and specifying cell types (Zaret 2020). This function is analogous to nucleating and localizing the protein clouds. The two-step nuclear search hypothesis is more useful for explaining global gene regulation if only some transcription factors follow it: some transcription factors define the locations of the protein clouds with DNA binding domains and others are followers with IDRs. For example, on long time scales developmental master regulator transcription factors would localize the protein clouds at cell-type-specific enhancers, then fast acting, signaling effector transcription factors could simply join these clouds (e.g. Glucocorticoid receptor (Barolo and Posakony 2002; Vockley *et al.* 2016)). The IDR-dominated Msn2, Hif1α and Hif2α are stress response transcription factors (Brodsky *et al.* 2020, Chen *et al.* 2021).

However, the two step-nuclear search hypothesis is equally compatible with the Collaborative Competition Model, where transcription factors work together to evict nucleosomes and open chromatin (Polach and Widom 1996; Mirny 2010). Once formed, a protein cloud has many DNA binding domains that together outcompete nucleosomes. In the Collaborative Competition Model, DNA binding domains have quantitatively different affinities for DNA rather than specialized subclasses.

It bears noting that Brodsky and colleagues offer two other explanations for their observed phenomena (Brodsky *et al.* 2020, Jana *et al.* 2021, Brodsky *et al.* 2021, Gera *et al.* 2022.) They propose that the IDR-mediated nuclear localization could be driven by condensates. More intriguingly, they propose the IDR can directly bind to specific DNA sequences in a highly distributed manner. In vitro experiments may be necessary to distinguish these two models or the two-step nuclear search.

Do transcription factors hunt the genome for binding sites in packs or as lone wolves?

Most cartoons of transcription factor function depict a single protein molecule diffusing through the nucleoplasm searching for its cognate binding site. The implicit assumption is that transcription factors are lone wolves that search for their binding sites by themselves (Figure 1A).

A corollary to the two-step nuclear search hypothesis is that clusters of transcription factors could search the nucleoplasm together as a single unit, collaboratively hunting for binding sites, like a wolf pack (Figure 1B). This cluster of transcription factors, or nascent protein cloud, would have many DNA binding domains that together contain enough motif information to uniquely specify regions of the genome. A heterotypic cluster of transcription factors matches the clusters of heterotypic binding sites in an enhancer. These transcription factor wolf packs would have variable sizes, which can explain why some transcription factors have a broad range of apparent

diffusion constants in single-particle tracking experiments (Heckert *et al.* 2021; Chen *et al.* 2021). For some transcription factors, like Mig1 and Msn2, the functional unit is likely a small cluster (Wollman *et al.* 2017). Notably, a wolf pack would complicate some models of cooperative activation of transcription (Estrada *et al.* 2016; Angela H. DePace, personal communication).

It is not clear if a wolf pack would speed up or slow down nuclear search kinetics. More DNA binding domains would increase the number of non-specific DNA binding events, which could slow the search. More DNA binding domains would also slow the off-rate at real target sites, ensuring that more collisions with real targets are productive. Under the assumption of a thermodynamic framework here, the wolf pack aids in the selection of correct genomic locations. The transcription factors that establish the protein cloud could search the nucleus as a wolf pack and signal response effector transcription factors would join the clouds by performing the two-step nuclear search.

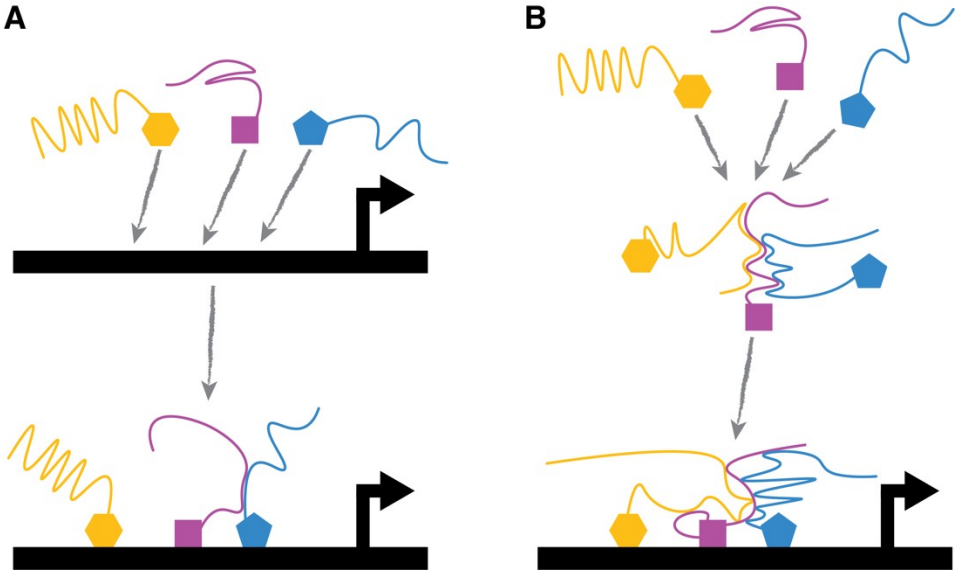


Figure 1: Transcription factors could hunt the genome for binding sites in wolf packs.
A) In the traditional model, transcription factors arrive at a promoter independently, hunting for binding sites like lone wolves. They often bind cooperatively on arrival. B) Some transcription factors can form clusters in the nucleoplasm and search for promoters as a group, hunting the genome like a wolf pack.

Am I kicking the can down the road?

The biggest weakness with the two-step nuclear search hypothesis is the lingering question of specificity. How does the protein cloud form at or localize to the right parts of the genome? This weakness is a restatement of other important problems: what distinguishes active enhancers in a cell? or

what nucleates transcriptional condensates? One answer comes from the thermodynamic framework, where all euchromatin is sampled with approximately the same on-rate, and slower off-rates at clusters of binding sites nucleate the protein clouds. Protein clouds emerge at clusters of binding sites by equilibrium binding of transcription factors to DNA. Protein-protein interactions between the transcription factors stabilize the clouds in a feed-forward manner. In the wolf pack framework, master regulator transcription factors bind each other in the nucleoplasm and search the genome as a unit. Once they find a cognate transcription factor binding site cluster, they would have an extended dwell time. Individual molecules would still have short residence times, but the protein cloud would have a longer dwell time, resulting in higher DNA occupancy (Sanborn *et al. 2021)*.

A parallel problem with the two-step nuclear search hypothesis is the issue of protein cloud diversity. Do individual transcription factors join multiple types of protein clouds? Are all the clouds similar? It is safe to assume that many different clouds will eventually activate transcription by recruiting coactivators like p300/CBP, Mediator, SAGA, and TFIID (Latchman 2008). Do these transcription factor-coactivator interactions occur before or after a cloud settles on a genomic locus? It follows that transcription factor-coactivator interactions are poor candidates for protein-protein interactions to nucleate protein clouds because coactivators must be able to activate many (sometimes all) genes and must be able to enter potentially all protein clouds. For example, in yeast, it has been argued that Mediator is necessary for transcription of virtually all genes (Petrenko *et al.* 2017), but degrading Mediator with degrons changes the expression of only 6% of genes (Warfield *et al.* 2021). Degrading Mediator in human cells has similarly modest effects (El Khattabi *et al.* 2019). If instead, the dominant force creating protein clouds is transcription factor-transcription factor interactions (homotypic or heterotypic), then it is easy to create diverse protein clouds.

Combining DNA binding domain-driven and IDR-driven nuclear search–allowing for a diversity of transcription factors

So far, I have drawn a strong contrast between traditional DNA binding domain-driven nuclear search and a two-step, IDR-driven nuclear search, but biology rarely works in absolutes. We can imagine a continuum between a DNA binding domain-only mode and an IDR-only mode of genomic site selection. This continuum is anchored by Max, which contains only a DNA binding domain, and the Notch Intracellular Domain, which has no DNA binding domain (Grandori *et al.* 2000, Hori *et al.* 2013). The Notch signaling protein is cleaved in response to extracellular signals, allowing the Notch Intracellular Domain to enter the nucleus and bind to CSL (also known as Suppressor of Hairless in flies or Lag1 in worms), displacing corepressors and recruiting mastermind and other coactivators (Hori *et al.* 2013). Notch lacks a DNA binding domain and performs the global search using its IDR. Other transcription factors would lie on this continuum between Max/Max dimers

and the Notch intracellular domain. For each transcription factor, genomic site selection would be the combination of the DNA binding domain contribution and the IDR contribution. This combination may or may not be a simple sum. The transcription factors that are DNA binding domain-dependent would establish the protein clouds while the transcription factors that are IDR-dependent would go to the existing clouds. The two-step nuclear search is more useful for gene regulation if some transcription factors set up the protein clouds and others follow.

Moreover, it is formally possible that the same transcription factor might find different binding sites in the genome with different mixtures of the two parts of the two-step search: that some genomic sites will be selected by the DNA binding domain and other genomic sites will be selected by the IDR. New work from the Barkai group found that 12 pairs of transcription factor paralogs had largely overlapping genomic localization. For 12 other pairs, the IDR dominated promoter localization; for the remaining 6 pairs, both the IDR and the DNA binding domain contributed (Gera *et al.* 2022). This blend of genomic site selection parallels the recent argument that transcription factors can have pioneering activity at specific genomic sites (Hansen *et al.* 2022). A transcription factor might help establish a protein cloud at the genomic locations with high quality motifs for its DNA binding domain and be a follower with its IDR at other genomic locations.

Testing the two-step nuclear search model

A model is most useful when it can make testable predictions. The two-step nuclear search hypothesis predicts that more transcription factors will behave like the Brodsky *et al* data: transcription factors without DNA binding domains (or with mutant DNA binding domains) will continue to localize to the correct enhancers and promoters, but lose the focal peaks above motifs. Truncating transcription factor IDRs will gradually shift genome binding from endogenous targets towards DNA binding domain-only targets, which will be more enriched for motifs and general open chromatin.

There are three more predictions. First, there will be regions of the protein that are responsible for genomic localization outside of the activation domains and DBDs. They will be necessary and sufficient for the global search. Brodsky *et al*. have demonstrated this prediction genome wide and Chen *et al*. have demonstrated it for single molecules. All that remains is to find more examples and exceptions. Second, the reciprocal prediction is that if we cut out the internal ‘inert’ regions of transcription factors, there will be genome localization defects, i.e. a minimal transcription factor with all the known minimal activation domains and DNA binding domain will not bind to and activate endogenous targets. Third, chimeras that swap DNA binding domains and IDRs between pairs of transcription factors could reveal more cases where the IDR or the DNA binding domain dominates genome binding. These experiments are now feasible.

Conclusion

I have proposed that transcription factors search the nucleus for binding sites with a combination of a global search with protein-protein interactions mediated by the IDR and a local search with protein-DNA interactions mediated by the DNA binding domain. This two-step nuclear search hypothesis can explain several long-standing irregularities in the literature. It follows that these protein-protein interactions may initiate off of the DNA, yielding small wolf packs of transcription factors that together hunt the nucleus for binding sites. So far, I have discussed this idea only in the context of active euchromatin. If the tight meshwork of heterochromatin (Ou *et al.* 2017) precludes transcription factor wolf packs from entering, this would further ensure a tight off state, reduce the genomic search space, and speed up nuclear searches.

Data availability

No data was generated for this article.

Acknowledgements

The author would like thank Alex Holehouse, Zeba Wunderlich, Vincent Fan, Yu Chen, Ben Vincent, Thomas Graham, Angela DePace, Ryan Friedman, Xavier Darzacq, Clarice Kit Hong, Robert Tjian, Michael Gabriel Hayes, Jordan Stefani, Abrar Abidi, Michael White, Nicholas Morffy, and members of the Tjian-Darzacq labs for helpful discussions and comments on the manuscript. The author would also like to thank Jasper Rine for editorial feedback.

Funding

This work was supported by the Burroughs Wellcome Fund Postdoctoral Enrichment Program grant 1017384 and NSF grant 2112057.

Conflict of Interest

The author reports no conflicts of interest.

References

Barolo S., and J. W. Posakony, 2002 Three habits of highly effective signaling pathways: principles of transcriptional control by developmental cell signaling. Genes Dev. 16: 1167–1181.

Baughman H. E. R., D. Narang, W. Chen, A. C. Villagrán Suárez, J. Lee, *et al.*, 2022 An intrinsically disordered transcription activation domain alters the DNA binding affinity and specificity of NFκB p50/RelA. bioRxiv 2022.04.11.487922. [accessed 2022 July 10].

Berlow R. B., H. Jane Dyson, and P. E. Wright, 2022 Multivalency enables unidirectional switch-like competition between intrinsically disordered proteins. Proceedings of the National Academy of Sciences 119.

Bintu L., J. Yong, Y. E. Antebi, K. McCue, Y. Kazuki, *et al.*, 2016 Dynamics of epigenetic regulation at the single-cell level. Science 351: 720–724.

Boija A., I. A. Klein, B. R. Sabari, A. Dall'Agnese, E. L. Coffey, *et al.*, 2018 Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. Cell 175: 1842–1855.e16.

Brodsky S., T. Jana, K. Mittelman, M. Chapal, D. K. Kumar, *et al.*, 2020 Intrinsically Disordered Regions Direct Transcription Factor In Vivo Binding Specificity. Mol. Cell 79: 459–471.e4.

Brodsky S., T. Jana, and N. Barkai, 2021 Order through disorder: The role of intrinsically disordered

regions in transcription factor binding specificity. Curr. Opin. Struct. Biol. 71: 110–115.
Brzovic P. S., C. C. Heikaus, L. Kisselev, R. Vernon, E. Herbig, *et al.*, 2011 The acidic transcription activator Gcn4 binds the mediator subunit Gal11/Med15 using a simple protein interface forming a fuzzy complex. Mol. Cell 44: 942–953.
Cascarina S. M., M. R. Elder, and E. D. Ross, 2020 Atypical structural tendencies among low-complexity domains in the Protein Data Bank proteome. PLoS Comput. Biol. 16: e1007487.
Castellanos M., N. Mothi, and V. Muñoz, 2020 Eukaryotic transcription factors can track and control their target genes using DNA antennas. Nat. Commun. 11: 540.
Chen J., Z. Zhang, L. Li, B.-C. Chen, A. Revyakin, *et al.*, 2014 Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. Cell 156: 1274–1285.
Chen Y., C. Cattoglio, G. Dailey, Q. Zhu, R. Tjian, *et al.*, 2021 Mechanisms Governing Target Search and Binding Dynamics of Hypoxia-Inducible Factors. bioRxiv 2021.10.27.466110. [accessed 2022 July 10].
Chong S., C. Dugast-Darzacq, Z. Liu, P. Dong, G. M. Dailey, *et al.*, 2018 Imaging dynamic and selective low-complexity domain interactions that control gene transcription. Science 361. https://doi.org/10.1126/science.aar2555
Crocker J., N. Abe, L. Rinaldi, A. P. McGregor, N. Frankel, *et al.*, 2015 Low affinity binding site clusters confer hox specificity and regulatory robustness. Cell 160: 191–203.
Currie S. L., J. J. Doane, K. S. Evans, N. Bhachech, B. J. Madison, *et al.*, 2017 ETV4 and AP1 Transcription Factors Form Multivalent Interactions with three Sites on the MED25 Activator-Interacting Domain. J. Mol. Biol. 429: 2975–2995.
Das R. K., and R. V. Pappu, 2013 Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proc. Natl. Acad. Sci. U. S. A. 110: 13392–13397.
Deeds E. J., J. A. Bachman, and W. Fontana, 2012 Optimizing ring assembly reveals the strength of weak interactions. Proc. Natl. Acad. Sci. U. S. A. 109: 2348–2353.
Dunker A. K., S. E. Bondos, F. Huang, and C. J. Oldfield, 2015 Intrinsically disordered proteins and multicellular organisms. Semin. Cell Dev. Biol. 37: 44–55.
Dyson H. J., and P. E. Wright, 2016 Role of Intrinsic Protein Disorder in the Function and Interactions of the Transcriptional Coactivators CREB-binding Protein (CBP) and p300. J. Biol. Chem. 291: 6714–6722.
El-Gebali S., J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, *et al.*, 2019 The Pfam protein families database in 2019. Nucleic Acids Res. 47: D427–D432.
El Khattabi L., H. Zhao, J. Kalchschmidt, N. Young, S. Jung, *et al.*, 2019 A Pliable Mediator Acts as a Functional Rather Than an Architectural Bridge between Promoters and Enhancers. Cell 178: 1145–1158.e20.
Erijman A., L. Kozlowski, S. Sohrabi-Jahromi, J. Fishburn, L. Warfield, *et al.*, 2020 A High-Throughput Screen for Transcription Activation Domains Reveals Their Sequence Features and Permits Prediction by Deep Learning. Mol. Cell 78: 890–902.e6.
Estrada J., F. Wong, A. DePace, and J. Gunawardena, 2016 Information Integration and Energy Expenditure in Gene Regulation. Cell 166: 234–244.
Ferreon J. C., C. W. Lee, M. Arai, M. A. Martinez-Yamout, H. J. Dyson, *et al.*, 2009 Cooperative regulation of p53 by modulation of ternary complex formation with CBP/p300 and HDM2. Proc. Natl. Acad. Sci. U. S. A. 106: 6591–6596.
Fuda N. J., M. B. Ardehali, and J. T. Lis, 2009 Defining mechanisms that regulate RNA polymerase II transcription in vivo. Nature 461: 186–192.
Garcia D. A., T. A. Johnson, D. M. Presman, G. Fettweis, K. Wagh, *et al.*, 2021 An intrinsically disordered region-mediated confinement state contributes to the dynamics and function of transcription factors. Mol. Cell 81: 1484–1498.e6.
Gera T., F. Jonas, R. More, and N. Barkai, 2022 Evolution of binding preferences among whole-genome duplicated transcription factors. eLife 11.
Grandori C., S. M. Cowley, L. P. James, and R. N. Eisenman, 2000 The Myc/Max/Mad network and the transcriptional control of cell behavior. Annu. Rev. Cell Dev. Biol. 16: 653–699.
Hahn S., and E. T. Young, 2011 Transcriptional regulation in Saccharomyces cerevisiae: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. Genetics 189: 705–736.
Hansen A. S., M. Woringer, J. B. Grimm, L. D. Lavis, R. Tjian, *et al.*, 2018 Robust model-based analysis of single-particle tracking experiments with Spot-On. Elife 7. https://doi.org/10.7554/eLife.33125
Hansen J. L., K. J. Loell, and B. A. Cohen, 2022 The pioneer factor hypothesis is not necessary to explain ectopic liver gene activation. Elife 11. https://doi.org/10.7554/eLife.73358
Harbison C. T., D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, *et al.*, 2004 Transcriptional

regulatory code of a eukaryotic genome. Nature 431: 99–104.
Hare E. E., B. K. Peterson, V. N. Iyer, R. Meier, and M. B. Eisen, 2008 Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLoS Genet. 4: e1000106.
Harmon T. S., A. S. Holehouse, M. K. Rosen, and R. V. Pappu, 2017 Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. Elife 6. https://doi.org/10.7554/eLife.30294
Harrison M. M., X.-Y. Li, T. Kaplan, M. R. Botchan, and M. B. Eisen, 2011 Zelda binding in the early Drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. PLoS Genet. 7: e1002266.
Heckert A., L. Dahal, R. Tjian, and X. Darzacq, 2021 Recovering mixtures of fast diffusing states from short single particle trajectories. bioRxiv 2021.05.03.442482. [accessed 2022 July 10].
Hori K., A. Sen, and S. Artavanis-Tsakonas, 2013 Notch signaling at a glance. J. Cell Sci. 126: 2135–2140.
Jana T., S. Brodsky, and N. Barkai, 2021 Speed–Specificity Trade-Offs in the Transcription Factors Search for Their Genomic Binding Sites. Trends Genet. 37: 421–432.
Jung J.-H., A. D. Barbosa, S. Hutin, J. R. Kumita, M. Gao, *et al.*, 2020 A prion-like domain in ELF3 functions as a thermosensor in Arabidopsis. Nature 585: 256–260.
Krois A. S., H. J. Dyson, and P. E. Wright, 2018 Long-range regulation of p53 DNA binding by its intrinsically disordered N-terminal transactivation domain. Proc. Natl. Acad. Sci. U. S. A. 115: E11302–E11310.
Kvon E. Z., G. Stampfel, J. O. Yáñez-Cuna, B. J. Dickson, and A. Stark, 2012 HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature. Genes Dev. 26: 908–913.
Lambert S. A., A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, *et al.*, 2018 The Human Transcription Factors. Cell 175: 598–599.
Latchman D. S., 2008 *Eukaryotic Transcription Factors*. Elsevier Science.
Lee R. van der, M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, *et al.*, 2014 Classification of intrinsically disordered regions and proteins. Chem. Rev. 114: 6589–6631.
Liu J., N. B. Perumal, C. J. Oldfield, E. W. Su, V. N. Uversky, *et al.*, 2006 Intrinsic disorder in transcription factors. Biochemistry 45: 6873–6888.
Liu Y., K. S. Matthews, and S. E. Bondos, 2008 Multiple intrinsically disordered sequences alter DNA binding by the homeodomain of the Drosophila hox protein ultrabithorax. J. Biol. Chem. 283: 20874–20887.
Long H. K., S. L. Prescott, and J. Wysocka, 2016 Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell 167: 1170–1187.
Ludwig M. Z., C. Bergman, N. H. Patel, and M. Kreitman, 2000 Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403: 564–567.
Lynch M., M. S. Ackerman, J.-F. Gout, H. Long, W. Sung, *et al.*, 2016 Genetic drift, selection and the evolution of the mutation rate. Nat. Rev. Genet. 17: 704–714.
Marklund E., G. Mao, J. Yuan, S. Zikrin, E. Abdurakhmanov, *et al.*, 2022 Sequence specificity in DNA binding is mainly governed by association. Science 375: 442–445.
Mirny L. A., 2010 Nucleosome-mediated cooperativity between transcription factors. Proc. Natl. Acad. Sci. U. S. A. 107: 22534–22539.
Ou H. D., S. Phan, T. J. Deerinck, A. Thor, M. H. Ellisman, *et al.*, 2017 ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. Science 357. https://doi.org/10.1126/science.aag0025
Petrenko N., Y. Jin, K. H. Wong, and K. Struhl, 2017 Evidence that Mediator is essential for Pol II transcription, but is not a required component of the preinitiation complex in vivo. eLife 6.
Polach K. J., and J. Widom, 1996 A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. J. Mol. Biol. 258: 800–812.
Powers S. K., A. S. Holehouse, D. A. Korasick, K. H. Schreiber, N. M. Clark, *et al.*, 2019 Nucleo-cytoplasmic Partitioning of ARF Proteins Controls Auxin Responses in Arabidopsis thaliana. Mol. Cell 76: 177–190.e5.
Qin B. Y., C. Liu, S. S. Lam, H. Srinath, R. Delston, *et al.*, 2003 Crystal structure of IRF-3 reveals mechanism of autoinhibition and virus-induced phosphoactivation. Nat. Struct. Biol. 10: 913–921.
Raj N., and L. D. Attardi, 2017 The Transactivation Domains of the p53 Protein. Cold Spring Harb. Perspect. Med. 7. https://doi.org/10.1101/cshperspect.a026047
Ravarani C. N., T. Y. Erkina, G. De Baets, D. C. Dudman, A. M. Erkine, *et al.*, 2018 High-throughput discovery of functional disordered regions: investigation of transactivation domains. Mol. Syst. Biol. 14: e8190.

Sanborn A. L., B. T. Yeh, J. T. Feigerle, C. V. Hao, R. J. Townshend, *et al.*, 2021 Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to Mediator. Elife 10. https://doi.org/10.7554/eLife.68068
Scholes C., A. H. DePace, and Á. Sánchez, 2017 Combinatorial Gene Regulation through Kinetic Control of the Transcription Cycle. Cell Syst 4: 97–108.e9.
Sherman M. S., and B. A. Cohen, 2012 Thermodynamic state ensemble models of cis-regulation. PLoS Comput. Biol. 8: e1002407.
Shin Y., and C. P. Brangwynne, 2017 Liquid phase condensation in cell physiology and disease. Science 357. https://doi.org/10.1126/science.aaf4382
Shively C. A., J. Liu, X. Chen, K. Loell, and R. D. Mitra, 2019 Homotypic cooperativity and collective binding are determinants of bHLH specificity and function. Proc. Natl. Acad. Sci. U. S. A. 116: 16143–16152.
Soto L. F., Z. Li, C. S. Santoso, A. Berenson, I. Ho, *et al.*, 2021 Compendium of human transcription factor effector domains. Molecular Cell.
Spitz F., and E. E. M. Furlong, 2012 Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. 13: 613–626.
Staller M. V., A. S. Holehouse, D. Swain-Lenz, R. K. Das, R. V. Pappu, *et al.*, 2018 A High-Throughput Mutational Scan of an Intrinsically Disordered Acidic Transcriptional Activation Domain. Cell Syst 6: 444–455.e6.
Staller M. V., E. Ramirez, S. R. Kotha, A. S. Holehouse, R. V. Pappu, *et al.*, 2022 Directed mutational scanning reveals a balance between acidic and hydrophobic residues in strong human activation domains. Cell Syst. https://doi.org/10.1016/j.cels.2022.01.002
Stormo G. D., 2013 *Introduction to protein-DNA interactions: structure, thermodynamics, and bioinformatics*. Cold Spring Harbor Laboratory Press.
Su W., S. Jackson, R. Tjian, and H. Echols, 1991 DNA looping between sites for transcriptional activation: self-association of DNA-bound Sp1. Genes Dev. 5: 820–826.
Teytelman L., D. M. Thurtle, J. Rine, and A. van Oudenaarden, 2013 Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. Proc. Natl. Acad. Sci. U. S. A. 110: 18602–18607.
Tuttle L. M., D. Pacheco, L. Warfield, J. Luo, J. Ranish, *et al.*, 2018 Gcn4-Mediator Specificity Is Mediated by a Large and Dynamic Fuzzy Protein-Protein Complex. Cell Rep. 22: 3251–3264.
Tuttle L. M., D. Pacheco, L. Warfield, D. B. Wilburn, S. Hahn, *et al.*, 2021 Mediator subunit Med15 dictates the conserved “fuzzy” binding mechanism of yeast transcription activators Gal4 and Gcn4. Nat. Commun. 12: 1–11.
Tycko J., N. DelRosso, G. T. Hess, Aradhana, A. Banerjee, *et al.*, 2020 High-Throughput Discovery and Characterization of Human Transcriptional Effectors. Cell 183: 2020–2035.e16.
Vockley C. M., A. M. D’Ippolito, I. C. McDowell, W. H. Majoros, A. Safi, *et al.*, 2016 Direct GR Binding Sites Potentiate Clusters of TF Binding across the Human Genome. Cell 166: 1269–1281.e19.
Warfield L., R. Donczew, L. Mahendrawada, and S. Hahn, 2021 Mediator is broadly recruited to gene promoters via a Tail-independent mechanism. bioRxiv 2021.12.21.473728. [accessed 2022 July 10].
White M. A., C. A. Myers, J. C. Corbo, and B. A. Cohen, 2013 Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. Proc. Natl. Acad. Sci. U. S. A. 110: 11952–11957.
Wollman A. J., S. Shashkova, E. G. Hedlund, R. Friemann, S. Hohmann, *et al.*, 2017 Transcription factor clusters regulate genes in eukaryotic cells. Elife 6. https://doi.org/10.7554/eLife.27451
Wunderlich Z., and L. A. Mirny, 2008 Spatial effects on the speed and reliability of protein–DNA search. Nucleic Acids Res. 36: 3570–3578.
Wunderlich Z., and L. A. Mirny, 2009 Different gene regulation strategies revealed by analysis of binding motifs. Trends Genet. 25: 434–440.
Zaret K. S., 2020 Pioneer Transcription Factors Initiating Gene Network Changes. Annu. Rev. Genet. 54: 367–385.