Trends in Genetics

CellPress OPEN ACCESS

Opinion

Q-rich activation domains: flexible 'rulers' for transcription start site selection?

Andrea Bernardini[1,*] and Roberto Mantovani[1,*]

Recent findings broadened the function of RNA polymerase II (Pol II) proximal promoter motifs from quantitative regulators of transcription to important determinants of transcription start site (TSS) position. These motifs are recognized by transcription factors (TFs) that we propose to term 'ruler' TFs (rTFs), such as NRF1, NF-Y, YY1, ZNF143, BANP, and members of the SP, ETS, and CRE families, sharing as a common feature a glutamine-rich (Q-rich) effector domain also enriched in valine, isoleucine, and threonine (QVIT-rich). We propose that rTFs guide TSS location by constraining the position of the pre-initiation complex (PIC) during its promoter recognition phase through a specialized, and still enigmatic, class of activation domains.

Highlights

A limited set of proximal DNA motifs known to regulate transcription levels also contribute to determine the location of transcription start sites (TSSs).

These 'ruler motifs' are recognized by a set of 'ruler' transcription factors (rTFs) belonging to distinct families.

With the exception of YY1, rTFs candidates are equipped with Q-rich activation domains (ADs) with similar amino acid composition (QVIT-rich).

We propose that the TSS-positioning function of rTFs resides in their Q-rich ADs.

rTFs likely work through direct interactions with the pre-initiation complex (PIC), imposing a soft position-dependent constraint on the initial steps of PIC promoter recognition exploiting the dynamic properties of their Q-rich ADs.

Start site selection: not only core promoter motifs

How RNA Pol II is recruited at specific genomic locations and how specific DNA base-pairs are 'chosen' as start sites for transcription remain fundamental questions in biology. Pol II transcription can start at a single major site or from several, nonrandom positions in a broad 50–100-base pair (bp) window within the same promoter/enhancer region [1]. The pattern of **TSSs** (see Glossary) is specific to each transcribed region, suggesting that it depends on the combination of the underlying DNA sequence (a 'TSS code') and ubiquitous protein factors [2]. In the classical model, TSS selection is attributed to sequence motifs specifically located in the **core promoter** [3]. These elements, including the TATA-box, Initiator (Inr), and other motifs downstream of the TSS [**downstream promoter region (DPR)**, downstream promoter element (DPE), and motif ten element (MTE)], are contacted by components of the **PIC**. Structural studies on TATA and TATA-less core promoters recently visualized direct contacts between distinct promoter regions and **TFIID**, one of the six **general TFs (GTFs)**, in the early steps of mammalian PIC assembly [4,5]. These observations of *in vitro* reconstituted complexes rationalized earlier studies on core promoter motif recognition by GTFs.

Although shown to be sufficient to drive basal activity in transcription assays *in vitro*, the mere presence of core motifs is not sufficient to impart transcription competency in a native context. Indeed, the upstream proximal promoter region (approximately –200 to –40) is required to provide meaningful transcription levels [6,7]. This region is characterized by the presence of binding sites for sequence-specific TFs. According to textbook models, TFs regulate transcription at multiple levels (recruitment of cofactors and nucleosome remodeling complexes, enhancer–promoter communication, promotion of PIC assembly, and Pol II pause-release), ultimately impacting quantitatively on the average amount of RNA molecules produced [8,9].

Recent studies based on machine-learning (ML) models trained on large TSS-mapping data sets [10–12] and on functional mapping of TSSs [13] expand the function of a selected group of sequence-specific TFs beyond the above paradigm, providing a qualitative effect on TSS

[1]Dipartimento di Bioscienze, Università degli Studi di Milano, Via Celoria 26, 20133, Milano, Italy

*Correspondence:
andrea.bernardini@unimi.it (A. Bernardini) and
mantor@unimi.it (R. Mantovani).

positioning. These studies complement each other by using different subsets of training data and unbiased models, and by focusing on different aspects of transcription initiation, converging on a limited set of DNA motifs that regulate TSS position: SP, ETS, NF-Y (CCAAT-box), CRE, NRF1, YY1, ZNF143 (GFY), and BANP (CGCG motif or Clus1) (Figure 1A), in addition to the expected core promoter elements (TATA-box, Inr, and DPR) [3].

New sequence determinants of transcription start sites

By leveraging a vast set of TSS-mapping data (CAGE, RAMPAGE, PRO-cap, and GRO-cap) in different human cell lines, Dudnyk and colleagues trained an interpretable deep-learning model of transcription initiation at base-pair resolution (Puffin) and tested how initiation(s) depends on DNA sequence patterns in the promoter [10]. The model provides a position-dependent 'effect curve' that defines the effect a motif has on TSS positioning in the nearby region as a function of the distance from the motif itself. The effect curves can be thought as activity fingerprints of each motif, which underlie its mode of action at the molecular level. No single motif is found at all promoters, while most promoters harbor at least one ruler motif (two or three on average). These motifs are evolutionarily conserved, at least in mammals, implying that a limited set of TFs have evolved, possibly independently, a shared mechanism of action.

He and Danko used PRO-cap data from genetically distinct lymphoblastoid cell lines to train a model (CLIPNET) that incorporates information on cell type-specific TF activity and the effect of natural genetic variability on TSS choice and magnitude of the transcriptional output. In addition to those mentioned, IRF4 and CTCF motifs are present, with lower frequencies, while ZNF143 and BANP motifs are absent [11]. Cochran *et al.* developed ProCapNet to model both magnitude and shape of the initiation profile given the underlying DNA sequence [12]. They detected AP1, also detected by Duttke *et al.* [13], and CTCF motifs; among other observations, they reported

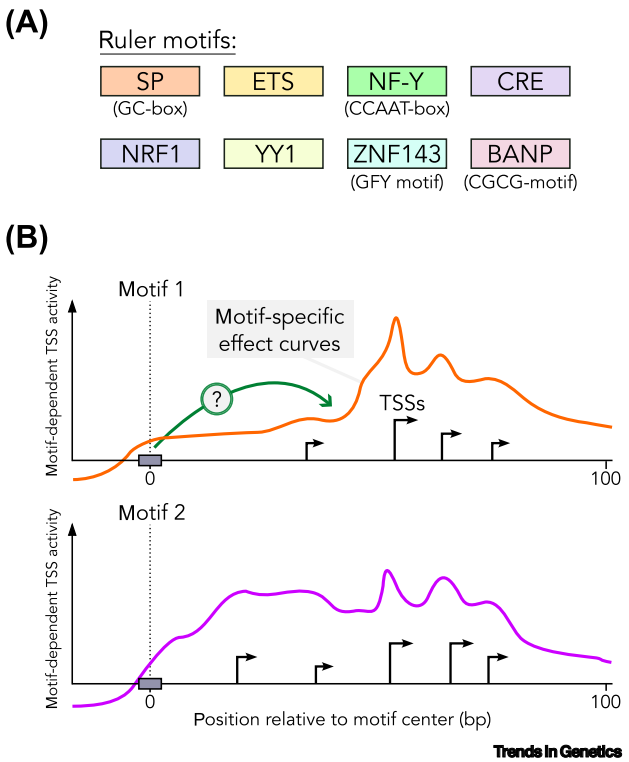

Figure 1. Ruler motifs and their effects on start site selection. (A) A set of eight transcription factor (TF)-binding motifs identified *de novo* as contributing to determine transcription start site (TSS) position ('ruler' motifs). Core promoter motifs are omitted, because their function in TSS selection is well established. Alternative motif names are indicated below the respective box. Motifs are ordered in descending frequency of occurrence at transcription start regions (TSRs) (top-left to bottom-right). (B) Ruler motifs influence TSS selection range according to a position-dependent activity curve specific to each motif, through an uncharacterized mechanism. The graphs show hypothetical effect profiles for two distinct ruler motifs. The motif-dependent TSS activity, defined as the relative contribution of the motif to the usage of each TSS (positive = activation; negative = repression), changes as a function of the distance from motif center in a non-uniform way, defining an 'activity fingerprint'. Refer to [10,11] for the actual motif-specific effect profiles.

Glossary

Activation domain (AD): protein region able to increase transcriptional activity of the target gene; also referred to as a *trans*-activation domain (tAD). ADs are often validated by fusing them with a heterologous DBD in reporter assays.
Core promoter: DNA region physically covered by the PIC, which includes the TSS (from –40 to +40 approximately). Core promoters can harbor specific signals to help direct PIC positioning, named core promoter motifs (e.g., TATA, Inr, and DPE).
Downstream promoter region (DPR): GC-rich DNA region spanning positions +17 to +35 relative to the TSS (+1). DPR is directly contacted by TFIID in the PIC. Other core promoter elements, such as downstream promoter element (DPE) or motif ten element (MTE) spatially overlap with DPR subregions.
General transcription factors (GTFs): set of six protein factors (TFIID, TFIIA, TFIIB, TFIIF, TFIIE, and TFIIH) that, together with Pol II and Mediator, constitute the PIC. GTFs are thought to assemble at every core promoter to assist and regulate Pol II engagement with DNA and initiate transcription.
Intrinsically disordered region (IDR): protein stretch refractory to adopting a stable fold; characterized by dynamic conformational ensembles rather than by a unique tertiary structure. IDRs can fold upon binding or establish 'fuzzy' dynamic interactions with their partners.
Pre-initiation complex (PIC): ~4.1-MDa complex that assembles at core promoters and mediates Pol II transcription initiation. PIC comprises Pol II itself, the six GTFs and Mediator (76 polypeptides), and integrates signals to regulate transcription initiation.
QVIT-rich domain: term proposed here to refer to a compositional class of ADs characterized by an enrichment in glutamine (Q), valine (V) and/or isoleucine (I), and threonine (T). Aromatic and charged residues instead tend to be rare. QVIT-rich domains are a subset of Q-rich domains, distinct from poly-Q regions. All Q-rich domains considered here are also QVIT-rich, therefore, the two terms are used interchangeably in the text.
Ruler motif: term proposed here to refer to a set of DNA motifs that contribute to define TSS position but are distinct from classical core promoter

initiator-like composite motifs in a subset of specific TF motifs that might be involved in defining local TSS position [12]. Importantly, both these studies analyzed all known **transcription start regions (TSRs)**, including enhancer RNAs (eRNAs), showing that the DNA syntax rules fundamentally apply also at enhancers, which differ mostly in terms of number and strength of the initiation-associated motifs, rather than relying on a different set of motifs [11,12]. Duttke *et al.* identified the same matrices with NRF1, NF-Y, SP, and ETS at the top of the frequencies [13]. The positions of these motifs with respect to the TSS are nonrandomly distributed and the shape of their frequency patterns is specific to each one, with characteristic frequency peaks and valleys at defined distances from the TSS, suggesting specific spatial constraints or distance optimality for function [10,13]. The distribution shape of a single motif mirrors its activity profile learned by ML models and further experimentally determined by motif swiping reporter assays [10,13]; that is, activity is position dependent and optimal at defined distances from the motif center (Figure 1B). Certain motifs, most notably NF-Y, show a striking 10.5-bp periodicity in their frequency distribution relative to TSSs that matches the DNA helical turn [10–13]. The identified matrices are generally located upstream from the TSS, associated with the promotion of bidirectional sense/antisense transcripts, with the exception of YY1, which is found within the downstream regions of unidirectional units. Collectively, the studies reported common motifs with descending frequencies in the order as reported above (Figure 1A), as well as the co-presence of more than one motif in essentially all regions examined.

In reality, this core set of elements is hardly a novelty for promoter structure analysis, given that they repeatedly emerged from *de novo* motif discovery in a ~200-bp window around the TSRs, coinciding with nucleosome-depleted regions [14–20]. These motifs can now be considered as regulators of transcription start location: for this reason, we propose the term '**ruler motif**' (Figure 1A). The obvious question is which are the **rTFs** that instruct TSS location in a position-dependent way and how are rTFs mechanistically carrying out the positioning of the PIC?

Linking motifs to their protein effectors

The human genome harbors ~1600 TF genes estimated to recognize more than 500 motif groups [21]. Single effectors are known for some ruler motifs (YY1, NF-Y, and NRF1) and their sequence specificity is understood at the atomic level [22–24]. The ZNF143/GFY and BANP matrices, despite some controversies over their recognition by THAP11/Ronin and ZBTB33/Kaiso, respectively, appear to bind ZNF143 [25] and BANP TFs [26,27]. All five candidate rTFs (YY1, NF-Y, NRF1, ZNF143, and BANP) are widely expressed and bind largely, if not predominantly, promoter sequences [25,26,28–30]. Importantly, the effects of NF-Y, YY1, NRF1, and ZNF143 depletion on TSS usage were validated in wet-lab experiments [10,13,25]. In fact, the impact of NF-Y on TSS selection, as well as on maintenance of the upstream border of targeted promoters free of nucleosomes, was previously reported by the Jothi laboratory [31]. In summary, these are realistic unique rTF candidates for the respective motifs.

Matters are more complex for the SP, ETS, and CRE motifs, since they can be recognized by multiple members of each TF family. Are all members involved? The first caveat could be that, due to the largely invariant nature of the TSS shape profiles among different tissues, the relevant rTFs are likely to be ubiquitously expressed with low variability across tissues. Inducible members, active upon a stimulus, or strictly tissue-specific ones, are unlikely to work as 'general' rTFs. Figure 2A shows the members of the SP, ETS, and CRE families ordered from lowest to highest variability in expression levels across tissues. If we focus on the top five ubiquitous candidates in each family, and exclude those annotated as repressors (KLF3, ERF, and ETV3) or strictly stress inducible (ATF7 and CREB3), we end up with four rTF candidates for the SP family (SP1, SP2, SP3, and KLF7), three for the ETS family (GABPA, ELF1, and ELF2), and three for the CRE family

motifs. The set of ruler motifs considered here is listed in the Abstract. The position of ruler motifs with respect to TSS shows a nonrandom distribution. Ruler motifs are targets of rTFs.

Ruler transcription factor (rTF): term proposed here to define a subclass of TFs exerting a TSS-positioning function. rTFs bind ruler motifs with specific positional biases from TSSs. Depletion of an rTF leads to a change in TSSs pattern at target TSRs.

TFIID: ~1.3-MDa multiprotein complex GTF, comprising TBP and 13 different TBP-associated factors (TAFs). According to the canonical model, TFIID is the first GTF to bind core promoter DNA, thus setting the location for the incoming PIC components.

Transcription start region (TSR): collective term referring to any DNA region showing transcription initiation activity, including promoters and transcribed enhancer regions.

Transcription start site (TSS): start position of a transcription event along a DNA molecule. It corresponds to the first nucleotide incorporated into the nascent RNA molecule.

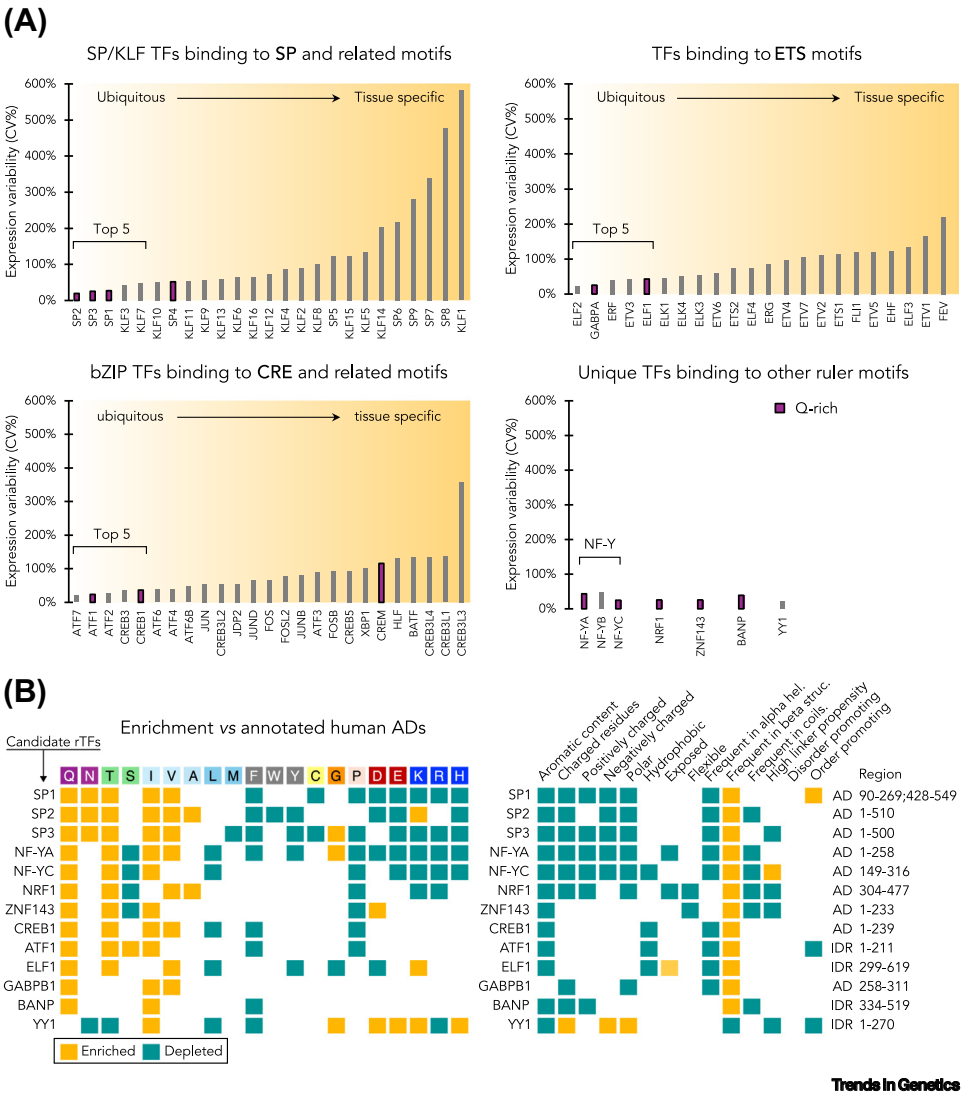

Figure 2. Ruler transcription factor candidates share a similar composition of their effector domains. (A) Variability in expression across human tissues for potential ruler transcription factors (TFs) grouped by motif family (SP, ETS, and CRE). TFs are ordered by increasing coefficient of variation (CV%) of their expression levels across tissues obtained from [21]. Q-rich TFs are indicated in purple. Motifs recognized by unique TFs are grouped in a single plot (bottom-right). (B) Enrichment/depletion patterns for single amino acids (left) and for resulting chemical/structural properties (right) for the activation domains or intrinsically disordered regions (IDRs) of the selected group of candidate ruler TFs. The statistically enriched/depleted features were calculated with Composition Profiler with default parameters [75], using the protein regions indicated on the right as queries, and all annotated human TFs activation domains as reference background (listed in [76]) with a statistical significance threshold set to P <0.05. GABPB1 is the subunit of the GABP complex harboring the activation domain; NF-Y subunits are listed separately.

(CREB1, ATF1, and ATF2). A second feature is related to genomic locations: one expects to find rTFs preferentially bound in proximity to active TSRs rather than at distal positions. A recent study directly compared the genomic distributions of several such TFs [32]: among ETS members, GABPA was almost exclusively, and ELF1 predominantly, located at TSS proximal regions, which extends a previous report [33]. Among 13 bZIPs tested, CREB1 showed an almost exclusive TSS proximal localization, while the related ATF2 was distributed more distally. In addition,

the DNA specificities of the GABPA/ELF and CREB subfamilies, as known from biochemical, structural, and *in vivo* approaches, match most closely the ETS and CRE logos found by the four studies [10–13]. As for the SP/KLF members, SP1, SP2, and SP3 bind mostly promoter sequences [34,35]. In summary, based on a common set of matrices, we propose here a short list of best candidate rTFs, but are there structural features common among them?

Candidate rTFs share similar amino acid composition of their activation domains

TFs minimally comprise a DNA-binding domain (DBD) and an effector domain, in charge of 'executing' a transcriptional effect, either transcription activation by an **activation domain (AD)**, or inhibition by a repression domain. According to the canonical classification based on the DBD, the candidate rTFs do not share similarities: the basis for their common function might then reside in their effector domains. With the notable exception of ligand-binding ADs of nuclear receptors, ADs are usually part of **intrinsically disordered regions (IDRs)** and have been broadly classified based on their amino acid composition in acidic, proline-rich, serine-rich, or glutamine-rich (Q-rich). SP1 is the founder member of TFs containing Q-rich ADs [36]; two other ruler motifs are recognized by TFs containing Q-rich ADs, namely NF-Y and NRF1 [37,38]. Is it conceivable that Q-rich ADs could be a common feature of rTFs?

From our list of candidate rTFs, we analyzed the amino acid composition of annotated ADs (or predicted IDRs for ATF1, ELF1, and BANP, the function of which as an AD is yet to be demonstrated). Figure S1 in the supplemental information online provides an example of a validated Q-rich AD and a Q-rich IDR from a candidate rTF. A set of shared features emerges (Figure 2B): (i) all regions are enriched in glutamine (Q) residues, with threonine (T) and aliphatic hydrophobic residues, specifically isoleucine and valine (I and V), also enriched; (ii) aromatic residues (F, Y, and W) are mostly depleted; (iii) acidic residues (D and E) are generally not enriched, with a subgroup of candidates (SP1, SP2, SP3, and NF-Y) clearly depleted of these residues; and (iv) all candidate regions are enriched in residues that promote potential β-structures. Given that these are IDRs, it is likely that these putative β-structures form only transiently. The features above distinguish these Q-rich domains from acidic ADs, the function of which relies on the exposure of bulky hydrophobic and aromatic residues interspersed with acidic amino acids [39–43], as also assessed in human cells [44,45]. Most importantly, despite the global enrichment in glutamine residues, rTFs do not have long poly-Q tracts found in other transcriptional regulators (such as TBP). Thus, the candidate rTFs selected here share commonalities: short patches of hydrophobic residues interspersed in a polar environment enriched in glutamine and threonine (**QVIT-rich domains**), but mostly depleted of aromatics and charge (Figure 2B). YY1 represents an apparent exception, since it is devoid of a Q-rich AD and its molecular target(s) and mechanism of action might be different, given that it activates transcription from a downstream position. A caveat is the absence of other TFs with functionally characterized Q-rich ADs in the list of ruler motifs: notably, OCT1/POU2F1, which has features of rTFs, such as the constitutive, ubiquitous presence, activation of 'basal' transcription [46], as well as prevalence of promoter recognition [47].

Specialized features of Q-rich ADs

What differentiates Q-rich ADs from other classes of ADs? A first feature is activity from proximal, as opposed to distal, binding sites. Since the earliest observations using plasmid-based experiments, it was shown that acidic activators supported transcription both from a proximal and a remote location [48]. By contrast, Q-rich ADs were unable to stimulate transcription from a distal (enhancer) position [49]. Nevertheless, proximally bound Q-rich TFs efficiently synergized with distal acidic activators. A tropism for promoter regions has been documented for several ruler motifs in *de novo* promoters selected from random sequences, as opposed to enhancers, across different human cell lines [50]. Although ruler motifs might strictly act *in cis*, they may well have a

function when embedded in an enhancer context, locally promoting chromatin accessibility and eRNA transcription. Overall, the short-range activity of Q-rich ADs fits well with the positional bias of the ruler motifs, entailing an evident constraint relative to the start site. Another notable feature is the relative 'strength' of activation potential, which is far lower than that of acidic activators, for example. In a recent survey, Udupa *et al.* reported on the rarity of Q-rich ADs, accounting for only 2.5% of the total based on high-throughput assays, with most being acidic ADs [51]. On the one hand, we argue that this rarity matches the apparent paucity of ruler matrices retrieved within the vast compendium of TF motifs; on the other hand, because of their intrinsic lower 'strength' in reporter assays, Q-rich ADs might be underestimated/missed in high-throughput screenings.

Q-rich TFs are likely to regulate the early steps of transcription initiation rather than of pause-release and elongation, as shown for SP1, and selected TFs, by combining nuclear run-on and RNase protection experiments [52]. More in-depth kinetic analyses showed that SP1 stabilized the binding of a reaction intermediate on DNA, rather than increasing the rate of its formation or promoter clearance [53,54]. A mixture of CRE-binding TFs (most likely CREB1 and ATF1) were shown to be involved at an early step of PIC assembly *in vitro* [55]. This effect was mediated by cooperative stabilization of TFIID binding to promoter DNA by ATF factors bound to the most proximal site, which led to the extension of the TFIID footprint to the region downstream of the TSS [56]. A hint for the recruitment of TFIID was also shown for NF-Y [57]. CREB1 represents a classic example of 'dual' activation: besides the Q-rich constitutive AD, it contains a well-studied kinase-inducible domain (KID), which mediates interactions with the co-activators CBP/P300. The two domains regulate different steps of transcription activation, with the Q-rich AD being involved in PIC formation, and KID acting in promoter clearance or re-initiation [58]. Thus, the presence of a Q-rich AD in rTFs does not exclude the coexistence of additional (sub)domains mediating different functions.

An additional point is evolutionary conservation, and the histone-like heterotrimeric NF-Y, present in all eukaryotes, could be a paradigm. While NF-YA and NF-YC Q-rich ADs are present in deuterostomes, and insects [37], the yeast trimer HAP2/3/5 needs to associate with a fourth subunit (HAP4) containing an acidic AD required for activation of targeted genes [59].

What are the targets of rTFs Q(VIT)-rich ADs?

Although ADs are known to interact with all kinds of transcriptional cofactors, we hypothesize that rTFs directly interact with a component of the PIC, which would consequently restrict TSS selection area (funnel model, Figure 3A). A first candidate would be TFIID, since it represents a node connecting activators, promoter recognition, and PIC assembly. TFIID size provides plenty of surfaces and domains for a diverse set of activators (co-activator function) [60]. Moreover, TFIID is equipped with modules able to physically contact and read core promoter DNA at TATA, Inr, and DPR locations, thus providing the structural means to fine-tune TSS selection once recruited/stabilized by the interactions with an rTF bound nearby [3–5,9,61] (Figure 3A). The relationships envisioned between ruler and core motifs related to their qualitative and quantitative aspects are outlined in Figure 3B.

Direct interactions with TFIID were shown for three rTF candidates reported here, namely SP1, CREB1, and NF-Y [34,62–67]. Intriguingly, the Q-rich ADs of both SP1 and CREB1 target the same moiety on the TAF4 subunit in TFIID [68], matching a short hydrophobic patch embedded in a long Q-rich IDR. Biophysical experiments suggested that these Q-rich ADs do not acquire a defined tertiary structure upon binding to TAF4 [69,70], but rather establish dynamic interactions, referred to as 'fuzzy' complexes [71]. It remains to be explored whether other candidate rTFs target TFIID. We speculate that seemingly different Q-rich ADs evolved independently in each

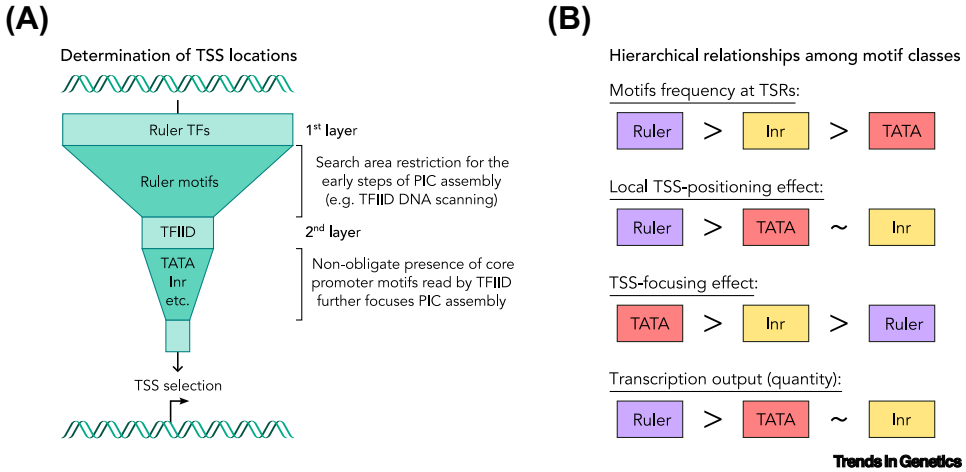

Figure 3. A hierarchical model of the relationships among ruler and core promoter motifs in transcription initiation. (A) A simple DNA-guided two-layer hypothetical 'funnel model' for transcription start site (TSS) selection mediated by ruler transcription factors (rTFs; first layer), which impose a soft constrain to the DNA region recognized by an early pre-initiation complex (PIC) assembly intermediate, possibly TFIID. In the presence of core promoter motifs (TATA, Inr, etc.), the initiation sites are further focused by direct recognition operated by TFIID (second layer). (B) Model of how ruler and core promoter motifs differ in their relative contribution to define distinct aspects of transcription. Ruler motifs dominate in defining the local region competent for transcription initiation and regulating the quantity of the transcription output. Strong core motifs occurrence within the area defined by ruler motifs show a dominant effect in focusing TSS selection.

rTF to recognize the same molecular targets given the intrinsic plasticity that characterizes these interactions.

A second target candidate would be Mediator, the archetype cofactor for a multitude of activators, especially enhancer-bound TFs. Mediator offers a diverse set of activator-binding domains and directly interacts with Pol II and other GTFs, potentially providing a mean to facilitate PIC formation. We are not aware of Q-rich TFs directly targeting Mediator, while there are plenty of examples of interactions with acidic and proline-rich activators.

TSS selection guided by rTFs

The hypothetical scenario that emerges from the above observations is the following. Promoters harboring at least one ruler motif recruit the correspondent rTF. Once bound, the rTF would stabilize an early step in PIC assembly, such as TFIID–DNA complex formation, through direct interactions. TFIID might either be directly recruited or be stabilized at defined positions once already bound to DNA. This might occur by a simple cooperative effect, with the rTF and TFIID reciprocally increasing their residence time on DNA through direct interactions. TFIID might initially contact DNA nonspecifically in a 'scanning' phase, until an isomerization step occurs that locks TFIID on DNA [72]. Possibly, this step corresponds to the transition of TFIID to its rearranged conformation, where it delivers TBP to the upstream core promoter followed by stable DNA bending [4]. This isomerization step might be facilitated by nearby TFs. Once TFIID is stably engaged on DNA, PIC formation would proceed and TSS location is established.

Direction of transcription relies on the relative order of promoter motifs rather than on motif orientation [73]: most ruler motifs induce a bidirectional activation effect downstream on the sense strand and upstream on the reverse strand [10,11,74], while seemingly repressing sense transcription starting upstream of the motif [13] (Figure 4A). Upstream repression might depend on a 'roadblock' effect of a TF positioned along the Pol II track, combined with the intrinsic asymmetry of the PIC structure

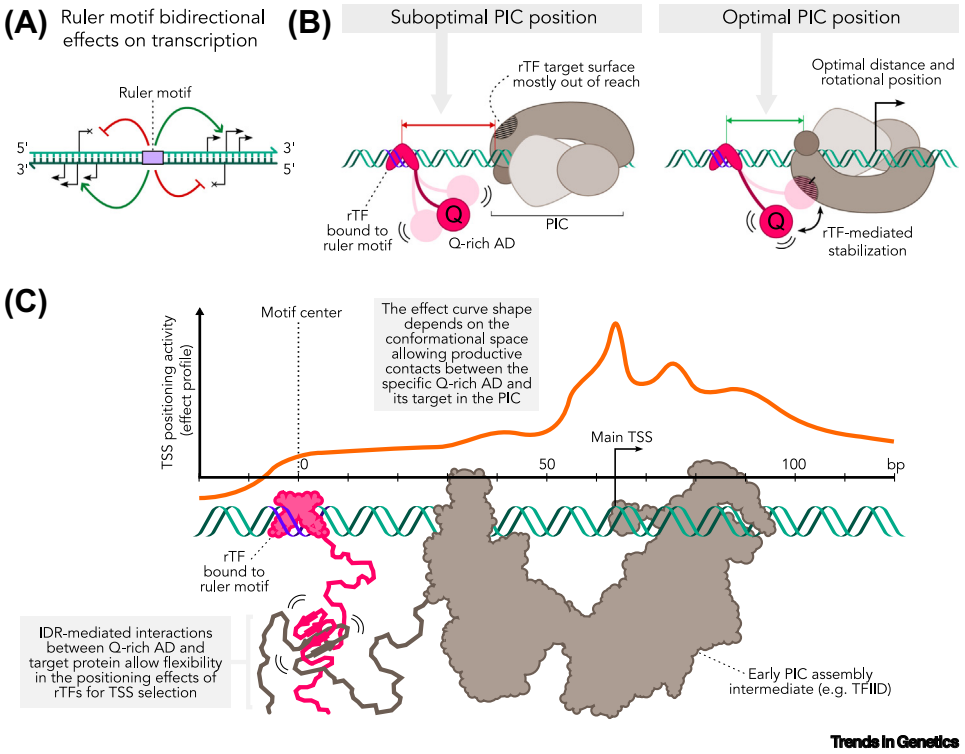

Figure 4. Mechanistic model of the function of ruler TFs on start site selection. (A) General effects of ruler motifs on the local transcription start site (TSS) activity. All ruler motifs (except YY1) have bidirectional activity on downstream transcription while repressing transcription from upstream sites. (B) Ruler transcription factors (rTFs) favor TSS activity in a position-dependent way. When the pre-initiation complex (PIC) assembles in an optimal position relative to the ruler motif (right), we posit that a cooperative interaction mediated by the Q-rich activation domain (AD) is established, increasing the residence time of PIC components on DNA and subsequent initiation at that position. The orientation of the two moieties relative to DNA helicity contributes to define the local preference for PIC formation. (C) Scheme of our interpretation of a ruler motif TSS selection effect profile (orange curve). Local optimal position of an early PIC assembly intermediate (likely TFIID) favors cooperative interactions with an upstream rTF (bottom, drawn approximately to scale). The interactions are mediated by the flexible Q-rich AD carried by the rTF. As represented here, the target of the Q-rich AD might also be a motif embedded in a second intrinsically disordered region (IDR) belonging to a PIC component. IDR-mediated interactions in this context would allow for a certain degree of flexibility in the distance and relative rotation of the two factors (rTF and the PIC) along the DNA, explaining the rather smooth appearance of the local peaks and valleys of the effect profile curve. Note that the schematic represents only one of the several potential locations of the PIC relative to the rTF that contribute to the global effect profile curve.

(i.e., the PIC surface targeted by an rTF is not reachable if the PIC assembles upstream of the ruler motif, leading to a stereo-positioning effect).

A relevant point is that the stabilization effect of the rTF on the PIC would be dependent on the relative position of the two interaction moieties, both in terms of distance and rotational phase along the DNA double helix (Figure 4B). The intrinsically disordered nature of the QVIT-rich ADs introduces flexibility to the system, thus softening the position-dependent constraints of rTF activity and leading to the observed 'smooth' activation profiles (Figure 4C). Instead, a 'construction block' interaction between two fixed, structured domains would generate a sharp on/off distance boundary, as occurs for the TSS positioning effect mediated by the TATA/TBP interaction within the PIC itself.

In conclusion, we speculate that ruler motifs would operate through cognate rTFs equipped with specialized QVIT-rich ADs that impose a soft local constraint in the initial steps of PIC promoter

recognition by stabilizing an early intermediate assembly at preferential distances from the ruler motif. The (nonobligate) presence of core promoter motifs would further focus TSS selection through direct recognition mediated by TFIID. The relative QVIT-rich AD flexibility would allow local adjustments of TFIID to dock on core promoter elements, if present, during the DNA-scanning phase.

Concluding remarks and future perspectives

We put forward the hypothesis that the limited set of rTFs contribute to organizing TSS location maps using a specialized class of Q-rich effector domains (QVIT-rich domains) that would directly establish interactions with components of the transcription machinery (TFIID as a first candidate). Several questions remain and demand dedicated studies in the near future (see Outstanding questions). To verify our model, it is necessary to test whether a QVIT-rich AD derived from an rTF would reproduce the same TSS profile when transplanted onto a different DBD. The competition for the same binding site among rTFs and other co-expressed TFs of the same family opens the possibility of a 'division of labor' model, whereby the same motif could be exploited to promote different steps in the transcription cycle depending on the identity of the effector domain brought by the bound TF. The molecular targets of each rTF would need to be identified and possibly structurally visualized. Finally, the study of the evolutionary trace of rTFs QVIT-rich ADs across metazoans, and beyond, might be holding informative clues about their function.

Acknowledgments

We thank László Tora and Nerina Gnesutta for critically reading our manuscript and for their helpful suggestions. This work is supported by MUR-PRIN #2022KWFA7C to R.M. and by the European Union - Next Generation EU, Mission 4, Component 2, CUP B93D21010860004.

Declaration of interests

The authors declare no competing interests.

Supplemental information

Supplemental information to this article can be found, in the online version, at https://doi.org/10.1016/j.tig.2024.11.008.

References

1. Carninci, P. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* 38, 626–635
2. Frith, M.C. *et al.* (2008) A code for transcription initiation in mammalian genomes. *Genome Res.* 18, 1–12
3. Vo, Ngoc L. *et al.* (2017) The punctilious RNA polymerase II core promoter. *Genes Dev.* 31, 1289–1301
4. Patel, A.B. *et al.* (2018) Structure of human TFIID and mechanism of TBP loading onto promoter DNA. *Science* 362, eaau8872
5. Chen, X. *et al.* (2021) Structural insights into preinitiation complex assembly on core promoters. *Science* 372, eaba8490
6. Jones, K.A. *et al.* (1985) Two distinct transcription factors bind to the HSV thymidine kinase promoter in vitro. *Cell* 42, 559–572
7. Miyamoto, N.G. *et al.* (1985) Specific interaction between a transcription factor and the upstream element of the adenovirus-2 major late promoter. *EMBO J.* 4, 3563–3570
8. Haberle, V. and Stark, A. (2018) Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.* 19, 621–637
9. Malik, S. and Roeder, R.G. (2023) Regulation of the RNA polymerase II pre-initiation complex by its associated coactivators. *Nat. Rev. Genet.* 24, 767–782
10. Dudnyk, K. *et al.* (2024) Sequence basis of transcription initiation in the human genome. *Science* 384, eadj0116
11. He, A.Y. and Danko, C.G. (2024) Dissection of core promoter syntax through single nucleotide resolution modeling of transcription initiation. *bioRxiv*, Published online September 17, 2024. https://doi.org/10.1101/2024.03.13.583868
12. Cochran, K. *et al.* (2024) Dissecting the cis-regulatory syntax of transcription initiation with deep learning. *bioRxiv*, Published online May 23, 2024. https://doi.org/10.1101/2024.05.28.596138
13. Duttke, S.H. *et al.* (2024) Position-dependent function of human sequence-specific transcription factors. *Nature* 631, 891–898
14. Santana, J.F. *et al.* (2022) Differential dependencies of human RNA polymerase II promoters on TBP, TAF1, TFIIB and XPB. *Nucleic Acids Res.* 50, 9127–9148
15. Suzuki, Y. *et al.* (2001) Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res.* 11, 677–684
16. FitzGerald, P.C. *et al.* (2004) Clustering of DNA sequences in human promoters. *Genome Res.* 14, 1562–1574
17. Mariño-Ramírez, L. *et al.* (2004) Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Res.* 32, 949–958
18. Xie, X. *et al.* (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 434, 338–345
19. Benner, C. *et al.* (2013) Decoding a signature-based model of transcription cofactor recruitment dictated by cardinal cis-regulatory elements in proximal promoter regions. *PLoS Genet.* 9, e1003906

Outstanding questions

Our current understanding of the sequence boundaries of the functional domains of rTFs remains incomplete. Can we establish high-throughput methods to experimentally test functional regions of rTFs in terms of TSS positioning and transcription activity?

Apart from isolated examples, we still lack a full list of the direct targets of rTFs. Can we establish a comprehensive list of QVIT-rich ADs targets by combining genetic, computational, and proteomic approaches?

The structural logic of QVIT-rich ADs remains elusive. What are the features and ensemble properties of QVIT-rich ADs in complex with their targets? Are they fundamentally distinct from those of acidic ADs?

ADs other than QVIT-rich ones have also been shown to interact with the basal transcription machinery. Which types of ADs, if any, impact TSS selection and what molecular mechanisms do they use to do so?

TFs that contain a Q-rich AD but do not recognize motifs in the current lists of ruler motifs do exist. Might these be additional rTFs that impact TSS positioning in pathway-specific, inducible, or tissue-specific units?

The QVIT-rich definition of rTFs ADs might hide a specific sequence syntax logic. Can we predict whether a TF has a TSS-positioning function given the sequence of its AD?

What is the role, if any, of QVIT-rich ADs in the specificity of rTFs genomic locations?

Although the TSS-positioning function of ruler motifs is conserved in mammals, dedicated efforts need to explore when this function evolved. What is the evolutionary history of rTFs ADs throughout eukaryotes?

20. Barbadilla-Martínez, L. *et al.* (2024) The regulatory grammar of human promoters uncovered by MPRA-trained deep learning. *bioRxiv*, Published online July 13, 2024. https://doi.org/10.1101/2024.07.09.602649
21. Lambert, S.A. *et al.* (2018) The human transcription factors. *Cell* 175, 598–599
22. Houbaviy, H.B. *et al.* (1996) Cocrystal structure of YY1 bound to the adeno-associated virus P5 initiator. *Proc. Natl. Acad. Sci. USA* 93, 13577–13582
23. Nardini, M. *et al.* (2013) Sequence-specific transcription factor NF-Y displays histone-like DNA binding and H2B-like ubiquitination. *Cell* 152, 132–143
24. Liu, K. *et al.* (2024) Molecular mechanism of specific DNA sequence recognition by NRF1. *Nucleic Acids Res.* 52, 953–966
25. Dong, J. *et al.* (2024) ZNF143 binds DNA and stimulates transcription initiation to activate and repress direct target genes. *bioRxiv*, Published online May 15, 2024. https://doi.org/10.1101/2024.05.13.594008
26. Grand, R.S. *et al.* (2021) BANP opens chromatin and activates CpG-island-regulated genes. *Nature* 596, 133–137
27. Liu, K. *et al.* (2023) Structural insights into DNA recognition by the BEN domain of the transcription factor BANP. *J. Biol. Chem.* 299, 104734
28. Ronzio, M. *et al.* (2024) Genomic binding of NF-Y in mouse and human cells. *Genomics* 116, 110895
29. Magnitov, M.D. *et al.* (2024) ZNF143 is a transcriptional regulator of nuclear-encoded mitochondrial genes that acts independently of looping and CTCF. *bioRxiv*, Published online March 12, 2024. https://doi.org/10.1101/2024.03.08.583864
30. Vella, P. *et al.* (2012) Yin Yang 1 extends the Myc-related transcription factors network in embryonic stem cells. *Nucleic Acids Res.* 40, 3403–3418
31. Oldfield, A.J. *et al.* (2019) NF-Y controls fidelity of transcription initiation at gene promoters through maintenance of the nucleosome-depleted region. *Nat. Commun.* 10, 3072
32. Nagy, G. *et al.* (2024) Lineage-determining transcription factor-driven promoters regulate cell type-specific macrophage gene expression. *Nucleic Acids Res.* 52, 4234–4256
33. Curina, A. *et al.* (2017) High constitutive activity of a broad panel of housekeeping and tissue-specific cis-regulatory elements depends on a subset of ETS proteins. *Genes Dev.* 31, 399–412
34. Suske, G. (1860) (2017) NF-Y and SP transcription factors - new insights in a long-standing liaison. *Biochim. Biophys. Acta Gene Regul. Mech.* 5, 590–597
35. Hasegawa, Y. and Struhl, K. (2021) Different SP1 binding dynamics at individual genomic loci in human cells. *Proc. Natl. Acad. Sci. USA* 118, e2113579118
36. Courey, A.J. and Tjian, R. (1988) Analysis of Sp1 in vivo reveals multiple transcriptional domains, including a novel glutamine-rich activation motif. *Cell* 55, 887–898
37. Bernardini, A. *et al.* (2022) Phylogeny of NF-YA trans-activation splicing isoforms in vertebrate evolution. *Genomics* 114, 110390
38. Gugneja, S. *et al.* (1996) Nuclear respiratory factors 1 and 2 utilize similar glutamine-containing clusters of hydrophobic residues to activate transcription. *Mol. Cell. Biol.* 16, 5708–5716
39. Gill, G. and Ptashne, M. (1987) Mutants of GAL4 protein altered in an activation function. *Cell* 51, 121–126
40. Erijman, A. *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
41. Kotha, S.R. and Staller, M.V. (2023) Clusters of acidic and hydrophobic residues can predict acidic transcriptional activation domains from protein sequence. *Genetics* 225, iyad131
42. Ravarani, C.N. *et al.* (2018) High-throughput discovery of functional disordered regions: investigation of transactivation domains. *Mol. Syst. Biol.* 14, e8190
43. Sanborn, A.L. *et al.* (2021) Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to Mediator. *eLife* 10, e68068
44. Alerasool, N. *et al.* (2022) Identification and functional characterization of transcriptional activators in human cells. *Mol. Cell* 82, 677–695.e7
45. DelRosso, N. *et al.* (2023) Large-scale mapping and mutagenesis of human transcriptional effector domains. *Nature* 616, 365–372
46. Pance, A. (2016) Oct-1, to go or not to go? That is the PolII question. *Biochim. Biophys. Acta* 1859, 820–824
47. Song, S. *et al.* (2021) OBF1 and Oct factors control the germinal center transcriptional program. *Blood* 137, 2920–2934
48. Sadowski, I. *et al.* (1988) GAL4-VP16 is an unusually potent transcriptional activator. *Nature* 335, 563–564
49. Seipel, K. *et al.* (1992) Different activation domains stimulate transcription from remote ('enhancer') and proximal ('promoter') positions. *EMBO J.* 11, 4961–4968
50. Sahu, B. *et al.* (2022) Sequence determinants of human gene regulatory elements. *Nat. Genet.* 54, 283–294
51. Udupa, A. *et al.* (2024) Commonly asked questions about transcriptional activation domains. *Curr. Opin. Struct. Biol.* 84, 102732
52. Blau, J. *et al.* (1996) Three functional classes of transcriptional activation domain. *Mol. Cell. Biol.* 16, 2044–2055
53. Narayan, S. and Wilson, S.H. (2000) Kinetic analysis of Sp1-mediated transcriptional activation of the human DNA polymerase β promoter. *Oncogene* 19, 4729–4735
54. Yean, D. and Gralla, J. (1996) Transcription activation by GC-boxes: evaluation of kinetic and equilibrium contributions. *Nucleic Acids Res.* 24, 2723–2729
55. Hai, T.W. *et al.* (1988) Analysis of the role of the transcription factor ATF in the assembly of a functional preinitiation complex. *Cell* 54, 1043–1051
56. Horikoshi, M. *et al.* (1988) Transcription factor ATF interacts with the TATA factor to facilitate establishment of a preinitiation complex. *Cell* 54, 1033–1042
57. Frontini, M. *et al.* (2002) NF-Y recruitment of TFIID, multiple interactions with histone fold TAF(II)s. *J. Biol. Chem.* 277, 5841–5848
58. Kim, J. *et al.* (2000) Distinct cAMP response element-binding protein (CREB) domains stimulate different steps in a concerted mechanism of transcription activation. *Proc. Natl. Acad. Sci. USA* 97, 11292–11296
59. Bolotin-Fukuhara, M. (2017) Thirty years of the HAP2/3/4/5 complex. *Biochem. Biophys. Acta BBA Gene Regul. Mech.* 1860, 543–559
60. Liu, W.-L. *et al.* (2009) Structures of three distinct activator–TFIID complexes. *Genes Dev.* 23, 1510–1521
61. Bernardini, A. *et al.* (2023) Transcription factor IID parks and drives preinitiation complexes at sharp or broad promoters. *Trends Biochem. Sci.* 48, 839–848
62. Ferreri, K. *et al.* (1994) The cAMP-regulated transcription factor CREB interacts with a component of the TFIID complex. *Proc. Natl. Acad. Sci. USA* 91, 1210–1213
63. Gill, G. *et al.* (1994) A glutamine-rich hydrophobic patch in transcription factor Sp1 contacts the dTAFII110 component of the Drosophila TFIID complex and mediates transcriptional activation. *Proc. Natl. Acad. Sci. USA* 91, 192–196
64. Xing, L. *et al.* (1995) cAMP response element-binding protein (CREB) interacts with transcription factors IIB and IID. *J. Biol. Chem.* 270, 17488–17493
65. Chiang, C.M. and Roeder, R.G. (1995) Cloning of an intrinsic human TFIID subunit that interacts with multiple transcriptional activators. *Science* 267, 531–536
66. Guermah, M. *et al.* (1998) Involvement of TFIID and USA components in transcriptional activation of the human immunodeficiency virus promoter by NF-kappaB and Sp1. *Mol. Cell. Biol.* 18, 3234–3244
67. Coustry, F. *et al.* (1998) The two activation domains of the CCAAT-binding factor CBF interact with the dTAFII110 component of the Drosophila TFIID complex. *Biochem. J.* 331, 291–297
68. Rojo-Niersbach, E. *et al.* (1999) Genetic dissection of hTAFII130 defines a hydrophobic surface required for interaction with glutamine-rich activators. *J. Biol. Chem.* 274, 33778–33784
69. Hibino, E. *et al.* (2017) Identification of heteromolecular binding sites in transcription factors Sp1 and TAF4 using high-resolution nuclear magnetic resonance spectroscopy. *Protein Sci.* 26, 2280–2290
70. Martinez-Yamout, M.A. *et al.* (2023) Glutamine-rich regions of the disordered CREB transactivation domain mediate dynamic intra-

and intermolecular interactions. *Proc. Natl. Acad. Sci. USA* 120, e2313835120
71. Tompa, P. and Fuxreiter, M. (2008) Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. *Trends Biochem. Sci.* 33, 2–8
72. Yakovchuk, P. *et al.* (2010) RNA polymerase II and TAFs undergo a slow isomerization after the polymerase is recruited to promoter-bound TFIID. *J. Mol. Biol.* 397, 57–68
73. O'Shea-Greenfield, A. and Smale, S.T. (1992) Roles of TATA and initiator elements in determining the start site location and direction of RNA polymerase II transcription. *J. Biol. Chem.* 267, 1391–1402
74. Core, L.J. *et al.* (2014) Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* 46, 1311–1320
75. Vacic, V. *et al.* (2007) Composition Profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinforma.* 8, 211
76. Soto, L.F. *et al.* (2021) Compendium of human transcription factor effector domains. *Mol. Cell* 82, 514–526