



universität
wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation / Title of the Doctoral Thesis

„Enhancer-responsiveness and -specificity of
Core Promoters in Gene Transcription“

verfasst von / submitted by

Muhammad Mamduh Ahmad Zabidi

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien, 2017 / Vienna 2017

Studienkennzahl lt. Studienblatt /
degree programme code as it appears on the student
record sheet:

A 794 685 490

Dissertationsgebiet lt. Studienblatt /
field of study as it appears on the student record sheet:

Molekulare Biologie

Betreut von / Supervisor:

Dipl.-Biochem. Dr. Alexander Stark

Acknowledgements

First I thank my whole family especially Mum for raising me in this World and teaching me all the good ethics.

I thank Alex who made a risky hire on an unproven computational person, for close supervision and all aspects of this thesis.

I thank Cosmas for leading the wetlab team to produce the data and for picking up my slack. For Michi and Martina, I thank them for their work in the wetlab, and for their motherly nature that they carry to the lab. I also thank Kathi and Olga for their contributions in the wetlab. I thank Tomáš, Daniel Gerlach and Omar for lots of help when I first started at the computer. Also thanks for Tomáš again for many extra help. Thanks to the rest of the Stark group: Dasha, Lukasz, Fanny, Daniel Spies, Anaïs, Felix, Evgeny, Gerald, Sebastian, Rui, Christoph Neumayr, Christoph Stelzer, Vanja, Ashley, Leo, Antonio, Lukas, Evgeniia, Filip, Mona, Mayela and Ivan.

I thank Petar, Lux and Hannes for the high-performance computing cluster. Thanks also to the Vienna Biocenter Core Facilities, especially the next generation sequencing, and the members of the DoktoratsKolleg RNA Biology. Special thanks for Karlo for providing me with a quiet space in the library where I can really delve into the code and engineer the solutions for the biological problems in this thesis.

I thank the members of my thesis committee, Julius, Luisa and Florian. I also thank Inês and Chris for an excellent PhD program in the best city in the World.

Thanks also to other people from the campus: Dani, Ian, Gordana, David, Kota, June, Olie, Zahra, and many more that I probably have forgotten here.

I thank my German teachers from Internationales Kulturinstitut Wien, especially Magdalena who taught me for the first 2 years, as well as Anna, Birgit, and Andreas. The

German knowledge that I have accumulated certainly helps in Vienna. I also thank the small Malaysian community here in Vienna.

I especially thank Venu and Harini. Nobody has ever been so kind as to take me in while I recuperate after my ACL surgery. I don't know how to repay you.

And for many more people who I forgot to mention here.

Core promoters are made up of core promoter elements that recruit *trans* factors. You are made up of some traits, with which you would attract people and events, whether you want it or not, knowingly or not. Some life patterns will happen to you and teach you, again and again and again. Until you learn.

Core promoters are directional. Life is directional. You waste it, you ain't getting it back. You can only move forward. So stay healthy. And don't wander.

Core promoters are short. Life is short.

When the organism is dead, core promoters don't matter no more.

When you're dead, nothing matters anymore.

I hope other human beings and humanity in general benefit from this thesis.

Sincerely yours.

Table of Contents

Table of Contents	5
Summary	7
Zusammenfassung	9
Introduction	11
The Genome and Transcription Regulation.....	11
Core Promoters as Genetic Elements	13
Gene Regulation via Core Promoters: Variegated Forms of Core Promoter Complexes	16
Gene Regulation via Core Promoters: Early Steps of Transcription	18
Gene Regulation via Core Promoters: RNA Polymerase II Pausing.....	21
Enhancer-specificity of Core Promoters	22
Enhancer-responsiveness of Core Promoters	25
Aims of the Thesis	29
Results and Discussion	31
Paper #1: Enhancer–core-promoter Specificity Separates Developmental and Housekeeping Gene Regulation	31
Paper #2: Regulatory Enhancer–Core-Promoter Communication via Transcription Factors and Cofactors	49
Paper #3: Genome-wide Assessment of Sequence-intrinsic Enhancer Responsiveness at Single-base-pair Resolution	65
Conclusions and Perspectives	91
Enhancer-specificity and Biochemical Compatibility	91
Enhancer-specificity in Other Transcription Programs.....	93
Decoding Enhancer-responsiveness.....	93
Deciphering the Initiation Code.....	94
Enhancer-responsiveness and Transcription Regulation	95
Beyond Enhancer-responsiveness and –specificity	96
References	99

Summary

Animal development is attributed to differential gene expression that is tightly regulated. Genes are transcribed from core promoters, sequences of around 100 base pairs (bp) surrounding the transcription start sites (TSSs) at which the RNA Polymerase II (Pol II) complex assembles. The cell-type specific transcriptional activities of core promoters are dependent on a second type of genomic regulatory element termed enhancers. During my PhD, I am interested in the specificity and responsiveness of core promoters towards enhancers.

An open question in the study of transcription is whether core promoters display intrinsic preferences towards enhancers. To address this hypothesis, we tested genome-wide enhancer candidates for their ability to activate core promoters that represent the housekeeping or developmental transcription programs, respectively. In *Drosophila melanogaster* cell lines, the two core promoter types exhibit differential preferences towards thousands of enhancers. Housekeeping core promoters are activated by enhancers that are active across cell types, while developmental core promoters are activated by enhancers that are highly cell-type specific. These two enhancer classes also differ in their genomic location, the protein factors that they bind, and the function of the neighbouring genes.

Different core promoters do not always support transcription at the same level, differences that have been mainly attributed to the wide range of enhancer strengths. Little is known about the intrinsic responsiveness of core promoters towards enhancers. We used single defined enhancers to test the enhancer-responsiveness of genome-wide core promoter candidates from *Drosophila melanogaster*. Core promoters vary widely in their enhancer-responsiveness, with differences up to three orders of magnitude. The differences correlate with sequence signatures and are associated with genes of different function.

In summary, the results obtained during my PhD thesis project show that the preference of core promoters towards enhancers represent another mechanism of

enhancer–core-promoter communication, and the thousands of core promoters in the genome vary substantially in their enhancer-responsiveness. These findings demonstrate that core promoters are actively involved in the precisely regulated gene expression that drives animal development.

Zusammenfassung

Differenzielle Genexpression ist entscheidend für die Entwicklung mehrzelliger Lebewesen wie Menschen und Tiere und muss daher streng kontrolliert werden. Gene werden beginnend von einer etwa 100 Basenpaare langen DNA Sequenz transkribiert. Diese Sequenz wird als Kernpromotor oder Core Promotor bezeichnet und umgibt den Transkriptionsstartpunkt (TSS), außerdem enthält sie die mindest-notwendigen DNA Sequenzelemente, welche für die Assemblierung des RNA-Polymerase II Komplexes nötig sind. Die zelltypspezifische Aktivierung der Transkription an den Core Promotoren ist jedoch von einer zweiten Klasse regulatorischer DNA Elemente im Genom, genannt transkriptionelle Enhancer, abhängig. Während meiner Doktorarbeit konzentrierte ich mich auf die Spezifität und das Ansprechverhalten von Core Promotoren in Bezug auf transkriptionelle Enhancer.

Eine der offenen Fragen in der Transkriptionsforschung ist, ob Core Promotoren eine intrinsische Präferenz für bestimmte Enhancer besitzen. Um diese Frage zu beantworten, untersuchten wir Enhancerkandidaten auf deren Fähigkeit Core Promotoren zu aktivieren. Dabei testeten wir sowohl Core Promotoren, die sich an nicht-regulierten, konstitutiv exprimierten Genen, sogenannten Haushaltsgenen, befinden, als auch solche, die an entwicklungsspezifisch regulierten Genen liegen. Die zwei Core Promotor Typen weisen in *Drosophila melanogaster* unterschiedliche Präferenzen zu tausenden Enhancern auf. Enhancer, die Core Promotoren vom Haushaltstyp aktivieren, sind über mehrere Zelltypen hinweg aktiv, wohingegen Enhancer, welche entwicklungsspezifische Core Promotoren aktivieren, zelltypspezifische Aktivität aufweisen. Enhancer dieser zwei Klassen unterscheiden sich auch in ihrer Position innerhalb des Genoms, den Proteinen, die sie binden, und der Funktion ihrer Nachbargene.

Core Promotoren weisen unterschiedliche Transkriptionslevel auf, Unterschiede, die bislang der weiten Bandbreite an Stärke der Enhancer zugeschrieben wurden. Über das intrinsische Ansprechverhalten von Core Promotoren auf bestimmte Enhancer ist bisher wenig bekannt. Wir verwendeten einzelne, definierte Enhancer, um dieses Ansprechverhalten von sämtlichen Core Promotor Kandidaten im *Drosophila*

melanogaster Genom zu testen. Core Promotoren variieren stark im Enhancer-Ansprechverhalten, wobei wir Unterschiede von bis zu drei Größenordnungen feststellten. Die Unterschiede korrelieren mit bestimmten Sequenzeigenschaften und sind mit Genen verschiedener Funktionen assoziiert.

Zusammenfassend zeigen die Ergebnisse meiner Doktorarbeit, dass Core Promotoren Präferenzen zu unterschiedlichen Enhancern aufweisen, was einen weiteren Mechanismus der Kommunikation zwischen Enhancern und Core Promoter darstellt, und dass sich die tausenden Core Promotoren im Genom substantiell im Enhancer-Ansprechverhalten unterscheiden. Diese Erkenntnisse zeigen, dass Core Promotoren aktiv an der präzisen Regulation der Genexpression beteiligt sind, welche die Entwicklung höherer Organismen erst ermöglicht.

Introduction

The Genome and Transcription Regulation

Single-cell animal embryos develop into mature multicellular organisms that consist of diverse cell types and tissues. Amazingly, the individual cells are genetically identical throughout development. This observation indicates that cellular machineries interpret only the necessary information in the genome, without modifying or deleting any nonessential parts. Consistently, the genome from somatic adult cells can still drive the development of embryos into adult animals^{1,2}. Today, overwhelming evidence suggests that differential gene expression, the difference in how cells interpret their genomes, is the driver of animal development^{3,4}.

The two principal regulatory DNA elements in gene expression are core promoters and enhancers (Figure 1A). Core promoters are short sequences flanking the first genomic positions where gene transcription starts. Another genetic element termed enhancers carry the cell-type specific activities of core promoters. Typically several hundred base pairs in length, enhancers can function irrespective of their orientation and distance⁵. In parallel, they can be located at arbitrary genomic positions relative to their target core promoters^{6,7}.

Enhancers contain binding motifs that recruit sequence-specific transcription factors (TFs) and cofactors (COFs) to activate target core promoters⁶. These TFs typically come in various combinations⁸ to direct development (for instance, refs. 9,10). Some TF combinations transform terminally differentiated cells into cells that closely resemble their undifferentiated ancestors, which can (re)differentiate into other cell types^{11,12}. Some other TFs instruct direct transformation of one cell type into another, bypassing the undifferentiated intermediates¹³⁻¹⁸. These lines of evidence indicate that the introduction of TFs activate groups of highly instructive enhancers, which are influential on cell fate¹⁹.

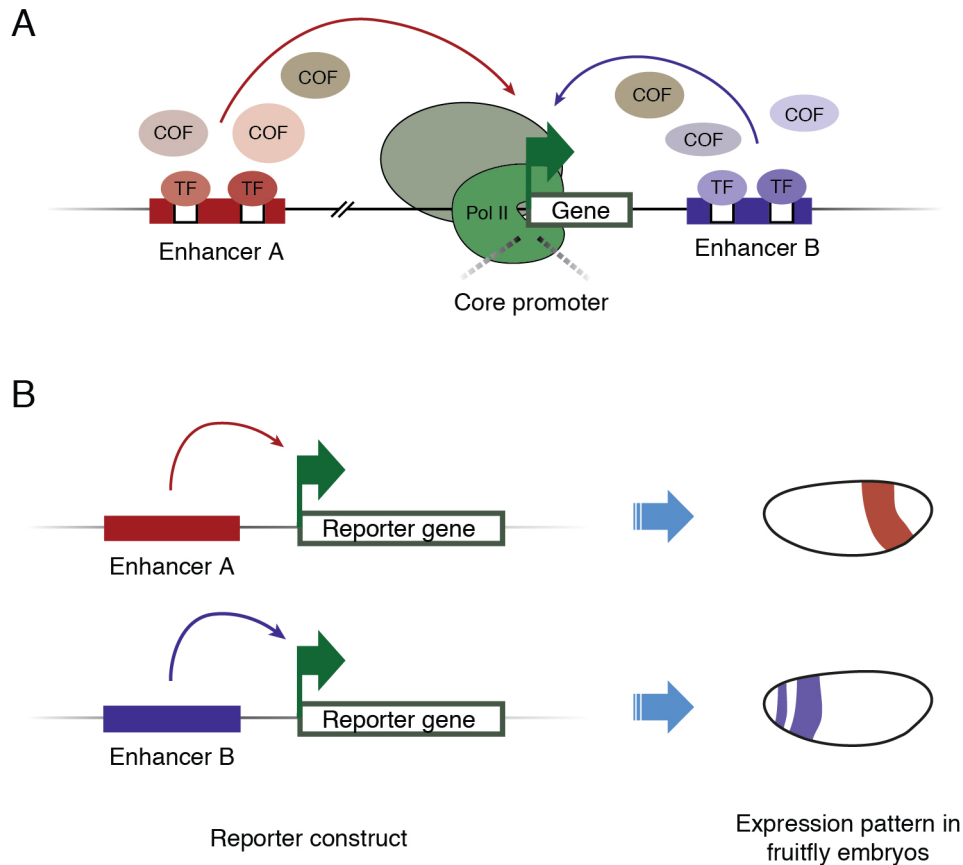


Figure 1 | Core promoters and enhancers, two main DNA elements in gene expression. A) Core promoters (green arrow) are short sequences that encompass transcription start sites (TSSs). Enhancers carry the cell-type specific activities of core promoters, and can lie at arbitrary locations. Enhancers recruit transcription factors (TFs) via TF binding motifs (white-filled boxes). These TFs in turn recruit cofactors (COFs) to transmit the signaling information to target core promoters. B) Enhancer activities are in the sequence, which can be studied via reporter assays. For example, enhancers A and B encode different expression patterns in *Drosophila* embryos.

Outside of their native contexts, enhancers frequently recapitulate the expression pattern of their endogenous target genes. This trait allows various aspects of enhancers to be studied^{6,7} (Figure 1B). For example, enhancers act in an additive manner, and some enhancers can specify the same expression patterns, thus allowing for robustness^{20,21}. Furthermore, enhancers do not always contain TF motifs with the best affinity, as suboptimal sequence signatures is opted for the correct activity²². Finally, the number of active enhancers scale directly with developmental stages²³.

Differential gene expression is typically ascribed to enhancers. Thousands of these *cis*-regulatory elements interpret signaling cues, ultimately converging at core promoters which convert the information into transcriptional output.

Core Promoters as Genetic Elements

Core promoters²⁴⁻²⁸ are minimal DNA sequences, typically around 80 to 100bp in length, that are sufficient for accurate initiation of transcription²⁹⁻³². In the genome, core promoters flank the TSSs and coincide with the 5' transcript ends. At protein-coding genes, core promoters nucleate the preinitiation complex (PIC)^{33,34} that consists of the general or basal transcription factors (GTFs) TFIIA, TFIIB, TFIID³⁵, TFIIE³⁶ TFIIF and TFIIH³⁷, the catalytic Pol II³⁸ as well as other auxiliary factors³⁹⁻⁴². By default, core promoters support only minimal levels of transcription, and their induced activity relies on enhancers⁴³⁻⁴⁵.

Core promoters can contain short nucleotide stretches with distinctive patterns⁴⁶⁻⁴⁸, termed core promoter motifs or elements (reviewed in refs. 28,49,50). Some core promoter elements are enriched at core promoters where the TSSs are concentrated at a few closely-spaced nucleotides. Such a “focused initiation” pattern is typically found at core promoters of developmentally-regulated genes with strong cell-type specific expression. Here, the position of the major TSS is denoted as the +1 position, which typically coincides with the adenosine (A) residue of the core promoter element initiator (Inr)⁵¹⁻⁵³ (Figure 2). Other core promoter elements such as the TATA box⁵⁴ and downstream promoter element (DPE)⁵⁵ are enriched at positions around -30 and +30, respectively. Other positionally-specific core promoter elements⁵⁶⁻⁶¹ include the motif ten element (MTE)^{46,62}, TFIIB response elements (BREs)^{63,64} and downstream core element (DCE)⁶⁵.

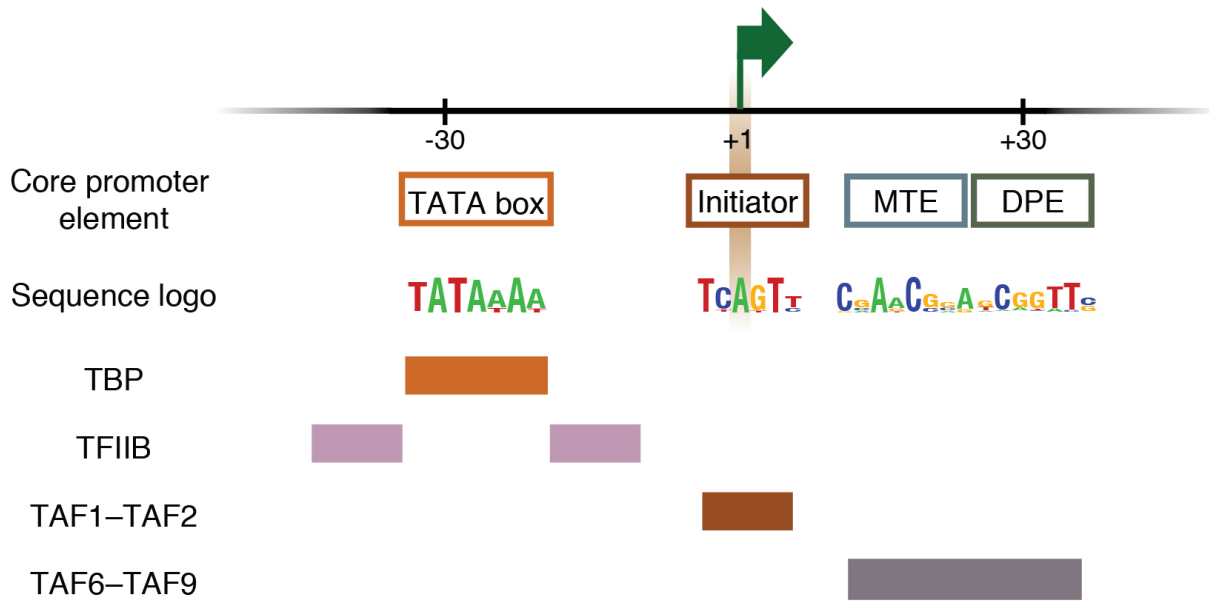


Figure 2 | Core promoter elements of focused core promoters, and regions of contact of the preinitiation complex (PIC). The positions of core promoter elements and their nucleotide frequencies^{46,66}, represented by sequence logos⁶⁷, are shown. Typically, the adenosine (A) residue of the Initiator coincides with the +1 position^{51,52}, the most prominent TSS (brown shading). Other core promoter elements such as the TATA box, motif ten element (MTE) and downstream promoter element (DPE) also show positional preferences. Also shown are the positions of contacts by preinitiation complex (PIC) components⁶⁸⁻⁷¹.

In contrast, the core promoter elements DRE⁷², as well as Ohler Motifs⁴⁶ 1, 5, 6, and 7 are typically not constrained within any specific core promoter subregion⁷³, and are enriched in core promoters of ubiquitously expressed housekeeping genes. Initiation at these core promoters tend to be spread over 40 to 100 nucleotides with no dominant TSS, a pattern also termed “dispersed initiation”. An exception is at translation-related genes where focused initiation occurs instead, with the TCT element is at around the +1 position^{74,75}. In vertebrates, dispersed core promoters could also be within CpG islands⁷⁶⁻⁷⁹, regions with elevated CG (also termed CpG) dinucleotide contents⁸⁰⁻⁸⁴.

Focused and dispersed initiation patterns are also related to distinct nucleosome configurations⁸⁵. In addition, many core promoters exhibit a mixture of both patterns (for instance, refs. 61,86,87). Some nucleotide variants that affect the affinity of core promoter elements also affect initiation pattern, resulting in expression variability among cells⁸⁸.

The occurrence of some core promoter elements and the dichotomy in initiation pattern is deeply conserved, yet species-specific innovations also exist. For example, focused plant core promoters are also enriched in the TATA box and Inr, in addition to

the plant-specific core promoter element Y-patch. Interestingly, plants lack CpG core promoters, and dispersed initiation in plants correlates with GA repeats with no preference for housekeeping genes^{89,90}. Additionally, the TATA box in *Saccharomyces cerevisiae* is located at positions around -40 to -120, instead of around -30 as in most species⁹¹.

Core promoter elements other than the TCT element are also associated with specific biological functions⁹²⁻⁹⁶. The TATA box is enriched in genes expressed before zygotic transcription⁹⁷, in terminally differentiated tissues^{94,95,98} and in strongly- and rhythmically-activated circadian genes⁹⁹. Interestingly, the TATA box affinity inversely correlates with gene length¹⁰⁰. The DPE is enriched at genes expressed during early embryogenesis and related to signaling⁹⁴. Some patterns of associations emerge between core promoter elements. For example, the TATA box and DPE occupy non-overlapping core promoter subregions and tend to co-occur with Inr. However, the TATA box and DPE rarely co-occur in the same core promoter^{46,95,101}. Additionally, some core promoter elements have distinct variants, such as Inr¹⁰² and the DPE⁹⁵. It seems that no single core promoter element occurs in all core promoters, and possibly more core promoter elements are yet to be discovered^{26,28}. Interestingly, most core promoter elements are directional (except, for example the TCT element or the palindromic DRE), consistent with the nonsymmetric nature of the PIC and the directionality of Pol II transcription activity⁴⁰.

As core promoters mark the beginning of genes, their recognition by the PIC is vital. The canonical view on core promoter recognition primarily centers around TFIID, an ~1 megadalton PIC component that consists of TATA-binding protein (TBP) and 14 TBP-associated factors (TAFs)³⁵. Specific TFIID subunits bind to core promoter elements to correctly position the PIC⁷⁰. For example, TBP straddles the DNA¹⁰³ and recognizes the TATA box¹⁰⁴, while TAF1 and TAF2 bind Inr¹⁰⁵ (Figure 2). However, TFIID is not the only PIC component that binds core promoters^{40,57}. For instance, TFIIB recognizes the BREs⁶³. Additionally, cooperativity between the PIC components is required for its correct positioning at core promoters. For example, TBP alone in solution has little preference to bind the TATA box in the correct orientation, and its orientational specificity is conferred by other GTFs¹⁰⁶⁻¹⁰⁹.

The discovery of the DCE underlines the importance of core promoter elements for core promoter function. It was first identified as a series of mutations in β -thalassemia patients¹¹⁰⁻¹¹⁴ that reduce TAF1 binding¹¹⁵. The mutations have no effect on transcript stability despite their location downstream of the +1 position⁶⁵. Thus, core promoter elements are functional subregions of core promoters that interact with *trans* factors.

Core promoter sequences were initially thought to be invariant as core promoters that were first isolated all contained the TATA box and Inr¹¹⁶. As more core promoters were identified, it was recognized that their sequence, and almost in parallel, their associated factors are also diverse. Indeed, emerging evidence supports that core promoters are directly involved in gene regulation.

Gene Regulation via Core Promoters: Variegated Forms of Core Promoter Complexes

The view on core promoter recognition by the PIC has initially revolved around the TATA box and Inr-containing sequences. However, many core promoters do not contain either of these, nor any other recognizable core promoter element. How are core promoters in the genome recognized, activated, and regulated?

A relatively straightforward way to differentially regulate transcription via core promoters is to differentially express the canonical PIC components. For example, the pluripotency state of mouse ES cells is achieved partly via maintaining the TFIID components at high levels. Disturbance in the level of these components lowers pluripotency¹¹⁷. Interestingly, human ES cells express varying levels of TFIID subunits, and these subunits make up at least two alternative PIC complexes associated with different gene subsets¹¹⁸. Relatively “simple” alternative TFIID complexes, such as one that lacks TBP, also exist that drive transcription independent of TATA box recognition¹¹⁹.

Interestingly, specific regulation at core promoters is achieved by more than controlling the levels or configurations of the canonical TFIID. Several of the core promoter factors have multiple paralogs, including TFIIA, TAFs and TBP. These paralogs

significantly increase the ability of core promoter complexes to differentially regulate core promoters (reviewed in refs. 120-122). First, some TFIIA paralogs are highly cell-type specific and potentially drive differential transcription activity¹²³⁻¹²⁵. Second, in an otherwise canonical TFIID complex, a single TAF paralog substitution could mediate differential core promoter activation¹²⁶. In other examples, single TAF paralogs are also important in several differentiation processes¹²⁷⁻¹²⁹. Additionally, core promoter recognition does not always involve single component substitution. For instance, a group of TAF paralogs are specifically co-expressed in testis to drive the differentiation of male germ cells¹³⁰. Indeed, many TAF paralogs in different tissues and species are critical for the expression of specific gene subsets¹³¹.

Third, TBP paralogs termed TBP-related factors (TRFs) could substitute TBP to drive cell-type specific gene programs¹³²⁻¹³⁴. TRF1 associates with multiple TAF subunits that are typically not members of the canonical TFIID¹³⁵, and can exert its activity via an upstream TC-rich motif located at the expected TATA box position¹³⁶. Additionally, TRF1 is also important in Pol III transcription^{137,138} (Pol III transcription is reviewed elsewhere¹³⁹⁻¹⁴²). Interestingly, TRF2 drives the expression from DPE-containing core promoters¹⁴³, considered as an opposite class from TATA box-containing core promoters, the expected target of TBP^{46,94,144,145}. TRF2 and a cell-type specific version of TFIIA also selectively disrupt TBP-dependent TATA box activation^{146,147}. Additionally, TRF2 is critical for core promoter activation of many housekeeping genes¹⁴⁸⁻¹⁵⁰.

The requirement of differential core promoter recognition complexes is exemplified at the *Drosophila* genes *tudor* (*tud*) and *proliferating cell nuclear antigen* (*PCNA*). The expression of both of these genes can be separately driven by closely-spaced alternative core promoters, which activation required the canonical TFIID complex at one, and either TRF1 or TRF2-containing complexes at the other¹³⁶. In addition, at the histone gene locus, the core promoters of the core histones and histone H1 are recognized by distinct but non-prototypical TFIID complexes which separately contain either TBP or TRF2¹³⁷. These complexes also display different binding kinetics, consistent with their distinctive expression timing patterns¹⁵¹.

Finally, TRF3 is a vertebrate-specific TBP paralog, and the most similar to TBP^{152,153}. TRF3 has a critical role in hematopoiesis¹⁵⁴ where it associates with TAF3 (ref. 155). TRF3–TAF3 association is also implicated in the transformation of muscle precursor cells (myoblasts), to terminally differentiated myotubes of skeletal muscles. This process is accompanied by a decrease in the prototypical TFIID complex level¹⁵⁶, yet the reduced TFIID level is still instrumental for muscle gene expression¹⁵⁷. These observations might stem from the close biochemical similarity between TBP and TRF3. TRF3 can also bind the TATA box, and this binding is similarly enhanced by TFIIA. Indeed, TRF3 can partially substitute TBP function^{153,158}.

The existence of many core promoter factors provides the cell with elaborate *trans* machinery components to meet the demands of specific spatiotemporal gene regulation. Consistently, the increase in complex body plan also parallels the expansion of core promoter factors during evolution^{159,160}. Furthermore, core promoters are also the site of distinct steps at which the transcriptional machinery undergoes major transitions, and some of these steps are modulated for gene regulation.

Gene Regulation via Core Promoters: Early Steps of Transcription

Transcription begins with several sequential events that take place within or very proximal to core promoters (Figure 3). First, the PIC assembles at core promoters to form a “closed complex”^{161,162}. Subsequently, a short stretch of the DNA is melted in an “open complex”^{163,164}. Pol II then transcribes the leading nucleotides of the mRNA while moving downstream in an “initial transcribing complex”¹⁶⁵⁻¹⁶⁹ and separates from other PIC components¹⁷⁰⁻¹⁷². Shortly after, Pol II may “pause” several tens of base pairs downstream^{173,174}, before it continues transcribing¹⁷⁵. Some of these steps are rate-limiting and responsive to signaling cues, thus allowing for regulation^{24,176-178}. Additionally, the sequence is also critical at some of these steps.

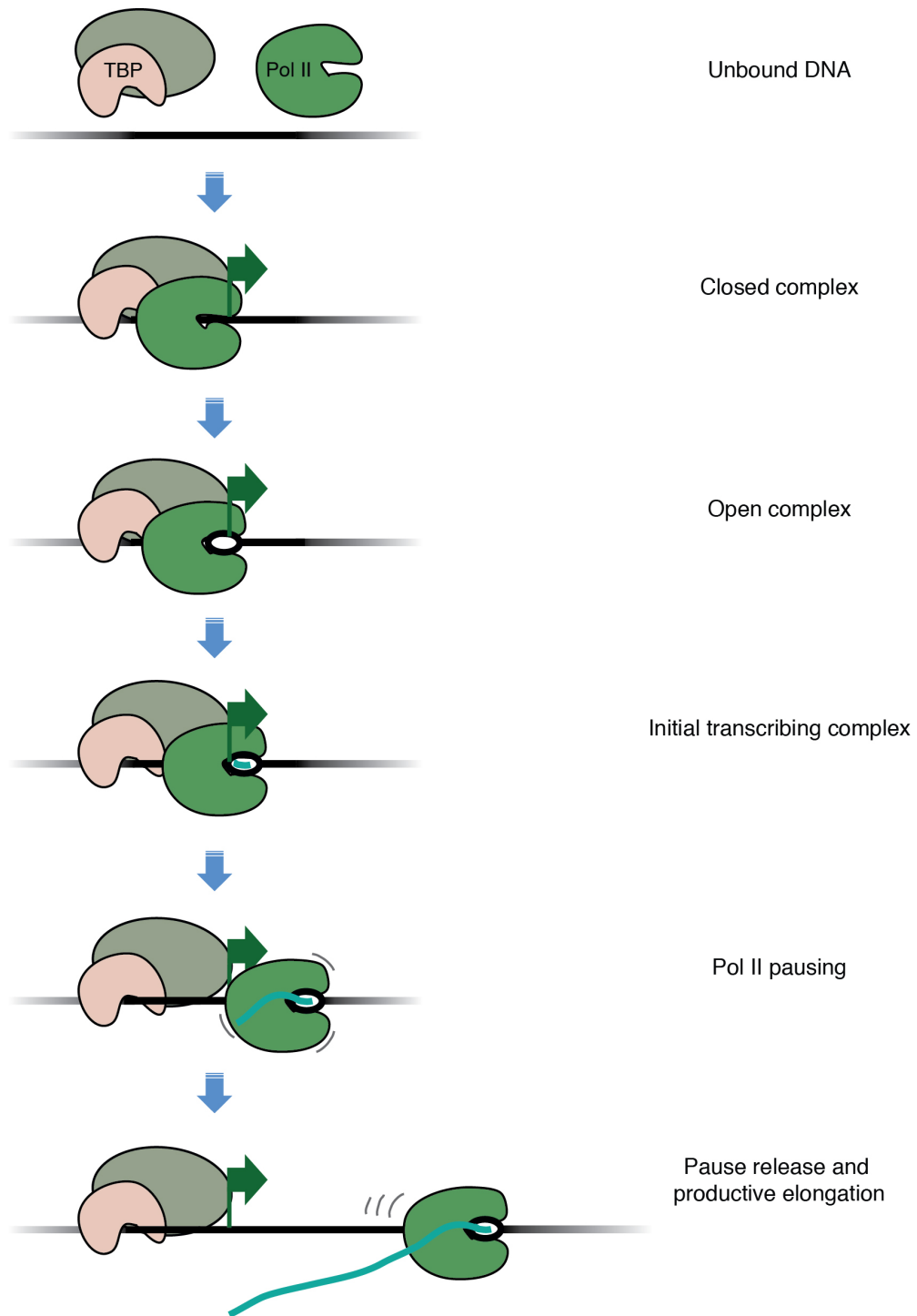


Figure 3 | Initial steps of transcription. Initially, RNA Polymerase II (Pol II), saddle-shaped TATA-binding protein (TBP) and other PIC components form a closed complex at core promoters. The closed complex then progresses into an open complex, where the DNA is locally melted and Pol II is positioned at the TSS (green arrow). The open complex transitions into an initial transcribing complex where Pol II synthesizes the leading nucleotides of mRNA. Pol II may subsequently pause after several tens of nucleotides, before continuing into productive transcript elongation. Core promoters can also wrap around nucleosomes before the PIC forms¹⁷⁹⁻¹⁸¹ (not shown). Reviewed in detail in refs. 40,68,177,182.

As core promoter elements bind TFIID, changes at core promoter elements such as the TATA box affect PIC binding efficiency^{183,184}. However, the reduced level of transcription observed upon mutating the TATA box is much lower than can be accounted for by the reduced PIC levels, suggesting that the TATA box may play also additional roles^{184,185}. One of such roles is exemplified in the differential activation of the p53 pathway transcriptional targets. In sensing DNA damage, the p53 pathway activates early and late transcriptional responses^{186,187}. Some early p53 core promoter targets contain the TATA box that become less responsive when p53 level increases¹⁸⁸. At the core promoter of the early response gene *p21*, the TATA box specifically aids in rapid PIC assembly. However, the *p21* core promoter only allows limited rounds of transcription. In contrast, the late response *FAS receptor (FasR)* core promoter supports inefficient PIC assembly, yet permits multiple rounds of transcription that leads to a sustained cellular response. The difference in PIC assembly kinetics is consistent with cellular requirements upon DNA damage, where the cell first activates the cell cycle arrest *p21*, followed by the proapoptotic *FasR*¹⁸⁹. Interestingly, p53 also activates the expression of an alternative *p21* transcript. This transcript is expressed from a TATA-less core promoter and is dependent on TRF2, a TBP paralog¹⁹⁰.

Other examples also exist where single signaling cues lead to differential activation of target core promoters via distinct regulatory strategies at the steps of transcription. For example, the *Drosophila* early embryogenesis regulator Zelda activates genes that may be regulated either via Pol II recruitment to the core promoter or via releasing paused Pol II¹⁹¹. Additionally, target core promoters of nuclear factor kappa B (NF-κB) signaling favor the recruitment of either positive or negative regulators of pause release¹⁹². Some enhancers have been identified that specifically regulate the early transcriptional steps, and interestingly, an enhancer could also modulate the different steps through the action of different TFs¹⁹³.

The differential regulation at core promoters can also involve thousands of genes simultaneously. In resting lymphocytes, the PIC docks at core promoters of future genes and is maintained in a closed complex. Upon activation of these immune cells, the expression of the GTF TFIIH leads to simultaneous DNA melting at these core promoters and rapid increase in mRNA production¹⁹⁴. Remarkably, the switch of transcription

programs can also be mediated at overlapping core promoters. In zebrafish embryos, maternal transcription is initiated from core promoters containing an upstream AT-rich motif. Upon progression into the zygotic stage, transcription is supported by core promoters that are mediated by nucleosome position. These zygotic core promoters sometimes overlap with their maternally-active counterparts¹⁹⁵.

Gene Regulation via Core Promoters: RNA Polymerase II Pausing

At thousands of metazoan core promoters, Pol II accumulates proximally downstream of the TSSs at positions +20 to +60 (ref. 174) while being stably engaged with the nascent transcript^{173,196-202}. The levels of these proximally-accumulated Pol II, commonly referred as being “paused”, vary and correlate with core promoter sequences (reviewed in refs. 182,203,204).

In *Drosophila*, paused Pol II is commonly found at core promoters that carry the core promoter elements DPE, MTE and Pause Button (PB), typical of developmental control genes^{197,205}. These sequence motifs have noticeable levels of G and C (GC) nucleotides, and are located downstream of the TSSs^{85,97,198,205}. Pausing in human and mouse also correlates with high GC content^{206,207}. Pausing tends to be low in TATA box-containing core promoters^{97,208}, and the deletion of the TATA box results in modest pause decrease²⁰⁹. While pausing is more prevalent and can be high in focused core promoters^{85,197}, pausing in dispersed core promoters is typically low to moderate^{85,202,210} and occurs further downstream²¹¹. Pausing at core promoters with distinct initiation patterns also correlates with different sequence features and is established by different factors^{198,210-214}.

The biological significance of pausing is potentially manifold. However, some recurrent themes emerge. Pausing was initially thought to be prepare genes for future activation^{203,204} or to facilitate rapid and synchronous responses to developmental signals²¹⁵. Additionally, pausing can reset core promoters to their ground state^{216,217}. As pausing is relatively stable over time²¹⁸⁻²²⁰, it may allow integration of regulatory signals²²⁰. In some cases, the degree of pause release also correlates with gene activity. Upon heat

shock, mammalian cells modulate gene activation and repression via respectively increasing or decreasing the release of the paused Pol II²²¹. Some enhancers indeed promote the release of pausing^{209,222}, and interestingly, pausing can also block enhancer activity²²³.

Enhancer-specificity of Core Promoters

Frequently, synthetic transcriptional units with heterologous enhancer–core-promoter pairing recapitulate the endogenous expression of the genes to which the enhancers are ascribed (for example, refs. 224,225). Through this strategy, thousands of enhancers in the genome have been identified (for instance, refs. 23,226-231). However, core promoters are also similarly numerous and highly diverse in sequence^{48,61,86,87,95,196,200,211,232-237} (reviewed in refs. 28,49,238,239). Even for the same gene, core promoters that drive the expression of alternative transcript isoforms may be highly dissimilar, with distinct GTF dependencies²⁴⁰.

These observations suggests that an inherent specificity between these elements exists²⁴¹⁻²⁴⁵, a notion that has been repeatedly postulated (for instance, reviewed in refs. 24,239,246,247). In line with this notion, some enhancer–core-promoter combinations result in unequal transcriptional output^{225,248-251}, while some other enhancer–core-promoter combinations do not result in substantial differences in transcriptional activity²⁵².

Differences in the level of transcriptional activity that are observed from different enhancer–core-promoter combinations indicate that both the protein factors and core promoters determine the transcriptional levels²⁵³⁻²⁶⁰. For instance, the sequence-specific TFs Caudal and Dorsal, which specify patterning of *Drosophila* embryos, preferentially activate DPE- over TATA box-containing core promoters²⁶¹⁻²⁶³. Other studies indicated that PIC components also play a role in the differential activation of core promoters²⁶⁴⁻²⁶⁷. In addition, many other TFs and COFs also show differential activity in various enhancer–core-promoter contexts²⁶⁸. The switch in transcription programs during maternal-to-zygotic transitions is also accompanied by differential activation of distinct core promoter

architectures^{97,195}, raising the possibility that distinct enhancer classes are utilized during the different stages.

Butler & Kadonaga investigated the existence of genomic enhancers that differentially activate TATA box or DPE-containing core promoters²⁴³. To this end, the authors performed enhancer trapping screens by randomly inserting transposon cassettes into the genome²⁶⁹⁻²⁷¹. These cassettes carry two markers that are driven by either TATA box or DPE-containing core promoters that can be differentially recombined (Figure 4A). Recombination of the parental lines produce sister lines that are genetically equivalent, except at the remaining marker²⁷². On differential recombination, 14 lines showed no differential marker activity. Interestingly, 3 lines showed preferential DPE-driven, while 1 line exhibited preferential TATA box-driven marker activity (Figure 4B). While the exact enhancers could not be identified, the results suggest for the existence of genomic enhancers that preferentially activate TATA box or DPE-containing core promoters.

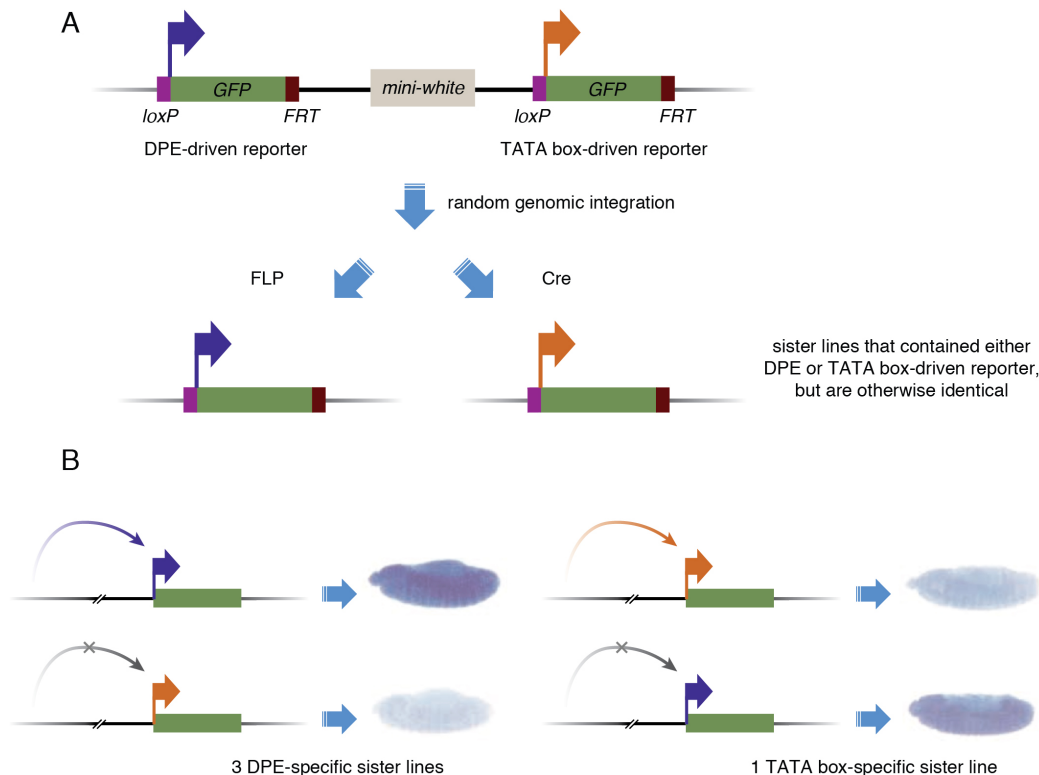


Figure 4 | Enhancer-trapping experiment supports the existence of core promoter-specific enhancers. A) Experimental layout of ref. 243. An enhancer-trapping vector consists of two *green fluorescent protein* (*GFP*) markers driven by either DPE- or TATA box-containing core promoters. Each of these reporters are flanked by *loxP* and *flippase recognition target* (*FRT*) sites, enabling them to be excised out *in vivo* by either Cre or Flippase (FLP) recombinase to produce sister lines. B) From this study, 3 DPE- and 1 TATA-box specific lines were identified. *In situ* embryo images are from the original publication²⁴³.

In another study, Merli et al. asked the question from a different angle. Given known enhancers, can they differently activate core promoters in the genome? At the locus of the *Drosophila* gene *decapentaplegic* (*dpp*), *dpp* enhancers correctly activate their cognate core promoter, located 35 kbp upstream. In contrast, the enhancers do not activate the more proximal *SLY-1* homologous (*Slh*) and *out at first* (*oaf*) core promoters (Figure 5). The authors substituted the endogenous *oaf* with *hsp70* core promoter, resulting in the activation of *oaf* in a manner that resembles the endogenous *dpp* expression²⁴⁴. This result showed that exchanging a core promoter is sufficient for gene activation by enhancers.

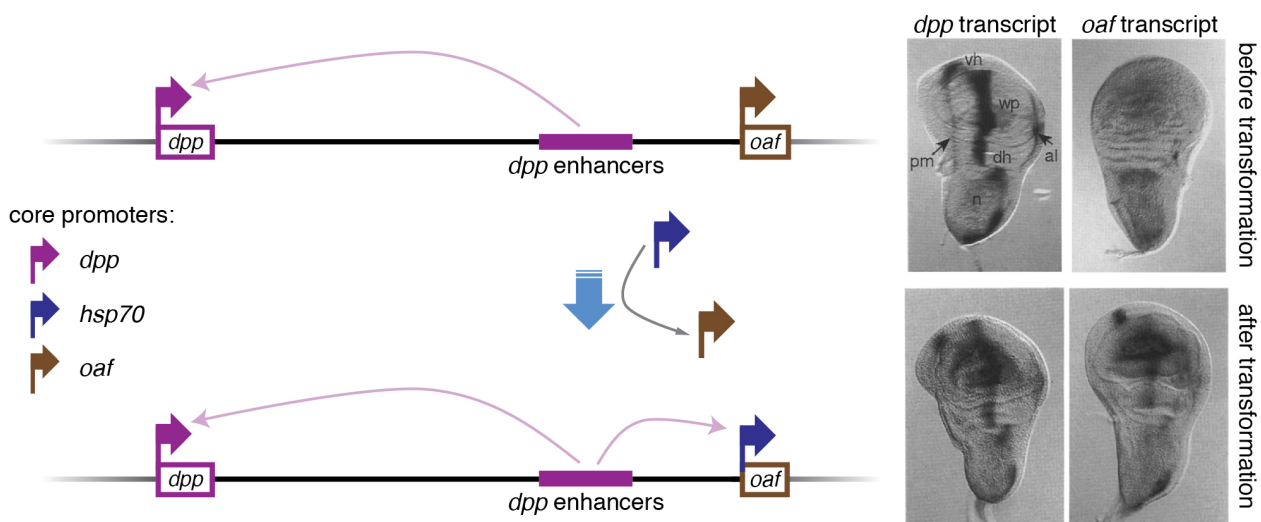


Figure 5 | Enhancer–core-promoter specificity at the locus of *decapentaplegic* (*dpp*) and *out at first* (*oaf*).

The *dpp* enhancers activates their cognate *dpp* core promoter, located 35 kilobase pair (kbp) away. Interestingly, the enhancers do not activate the proximal *oaf* core promoter (top). Substitution of the *oaf* core promoter with *hsp70* core promoter, previously known to be responsive to the *dpp* enhancers²⁷³, results in the activation of *oaf* (bottom). RNA *in situ* images are from the original publication²⁴⁴. For simplicity, several features were omitted: 1) *SLY-1* homologous (*Slh*) core promoter, which is divergent separated by 300 bp from the *oaf* core promoter. Like *oaf*, *Slh* also shows different expression than *dpp*. 2) Additionally, the depicted *dpp* enhancers comprise of at least 7 known enhancers.

Together, these studies suggest that core promoters have certain specificities towards enhancers. The enhancer-specificities of core promoters result in differences in transcription activity, potentially mediated via intermediate *trans* factors²⁷⁴.

Enhancer-responsiveness of Core Promoters

In their default state, core promoters are associated with minimal set of factors and exhibit low level of *basal activity* or *strength*⁴⁵. Upon activation by enhancers that recruit additional factors, core promoters exhibit induced activity or activated transcription²⁷⁵⁻²⁷⁸. *Enhancer-responsiveness* is how potently core promoters can be activated by enhancers^{27,52,279} (Figure 6).

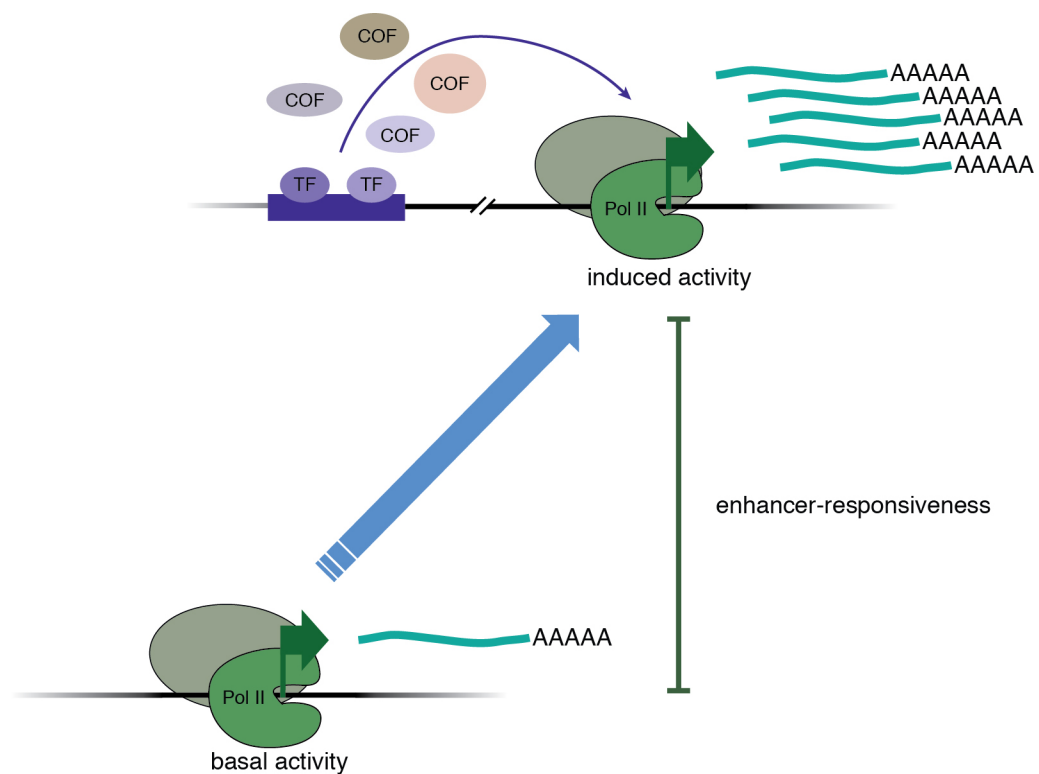


Figure 6 | Enhancer-responsiveness of core promoters. Basal activity or strength of core promoters (bottom) is characteristic of their default state. Here, core promoters are associated with Pol II and minimal set of factors to support basal transcription. Induced activity or activated transcription of core promoters (top) results from their activation by enhancers, mediated through TFs and COFs. Enhancer-responsiveness is the potency of core promoters to be activated. Core promoters can also be under the influence of negative-acting determinants, such as inaccessibility due to nucleosome configuration^{179,181,280-283}, which could also have positive effect²⁸⁴ (not shown). Adapted from refs. 43-45.

While gene expression has been credited to the cell-type specific activities of enhancers^{6,7}, the contribution of intrinsic activities of core promoters (their basal activity and enhancer-responsiveness) to gene expression has remained largely underexplored. Methods that quantify endogenous transcripts have been developed, such as cap analysis

of gene expression (CAGE)²⁸⁵ that measures steady state levels of capped transcripts, and global run-on sequencing (GRO-seq)^{196,211} that quantifies nascent transcripts. However, quantification of endogenous transcripts from these methods reflect the combination of intrinsic core promoter properties, enhancer activities, and other features such as DNA accessibility. Thus, basal activities and more importantly, enhancer-responsiveness of core promoters in the genome cannot easily be deconvolved.

Small scale studies indicate that core promoters possess highly different basal activities and enhancer-responsiveness. These traits may correlate with the properties of core promoter elements. Specifically, the nucleotide compositions at subregions of core promoter elements²⁸⁶⁻²⁹⁰ and the spacing of core promoter elements are crucial^{65,291,292}. Additionally, the combinations of core promoter elements have cooperative effect on the core promoter activities^{52,293}. Indeed, the optimization of the traits of core promoter elements has the guided design of highly responsive synthetic core promoters^{224,287,294,295}. Importantly, the enhancer-responsiveness of several core promoters seemed to both vary quite substantially and be divorced from cell type²⁹⁶. The maximum level of observed *in vivo* gene expression also correlates with the traits of core promoter elements. As these traits influence enhancer responsiveness, this strongly suggests that enhancer-responsiveness is involved in gene expression^{99,297}. Interestingly, the full extent of enhancer-responsiveness, and how it is employed in the genome, has not been investigated.

Furthermore, endogenous transcription initiation is also observed at positions distal from gene start positions^{196,298}, the expected locations of core promoters. When tested in reporter assays, such regions including the surrounding ~1 kb that also have genomic PIC binding can indeed support transcription²⁹⁹. Additionally, spurious intragenic initiation can happen, but is actively suppressed by the DNA methylation machinery of the gene body³⁰⁰. Alternatively, the intragenic initiation events that are observed may be a consequence of recapping of the derivatives from longer transcripts⁸⁷. Transcription initiation is also observed in open chromatin regions²⁸¹ such as upstream of gene starts, and at enhancers that produce enhancer RNAs (eRNAs)^{61,301-308}. The role of pervasive transcription at enhancers is still unclear, but there are cases where eRNA depletion was reported to impair transcription of associated genes³⁰⁹. Thus, determining enhancer-

responsiveness of core promoter candidates in the genome will help to elucidate whether distal endogenous initiation events are meaningful.

In conclusions, core promoters are increasingly recognized as active players in gene regulation. They are highly diverse in their primary sequences and their associated factors, can be regulated at the early steps of transcription, and are substrates for gene regulation^{24,239,310}. Core promoters require activation from enhancers, and together, they represent two main regulatory elements in the genome.

However, critical questions remain on the relationship between these two major genetic elements. First, can certain core promoters be differentially activated by certain enhancers? Second, how potently can core promoters be activated by enhancers?

These are the questions that motivate this thesis.

Aims of the Thesis

Core promoters are the sites of transcription initiation and are increasingly recognized as active contributors in gene regulation. However, much about core promoters is still unknown. The goals of this thesis are to determine if certain core promoters are preferentially activated by certain enhancers (enhancer–core-promoter specificity), and how potently core promoters can be activated in response to enhancers (enhancer-responsiveness).

Specifically, the aims of this thesis are:

1. Enhancer–core-promoter specificity:
 - To test the enhancer activity of genome-wide candidate fragments towards two core promoter classes
 - To understand the sequence and TF binding signatures at enhancers that mediate the specificity
2. Enhancer-responsiveness:
 - To determine the enhancer-responsiveness of genome-wide core promoter candidates towards single enhancers
 - To understand the sequence signatures that are associated with enhancer-responsiveness
 - To understand how core promoters with different enhancer-responsiveness are employed in the genome

This thesis comprises three papers. Paper #1 (ref. 311) tests the long-standing hypothesis of enhancer–core-promoter specificity, and shows that the specificity exists genome-wide. The specificity is sequence-encoded and employed to separate the housekeeping and the developmental transcription programs. Paper #2 (ref. 312) discusses the findings from Paper #1 in view of our current understanding of gene regulation. Paper #3 (ref. 313) determines enhancer-responsiveness of core promoters in the genome. Core promoters have a wide range of enhancer-responsiveness that is also invariant to enhancers and cell types. Core promoters with different enhancer-responsiveness are also associated with different biological functions.

Results and Discussion

Paper #1: Enhancer–core-promoter Specificity Separates Developmental and Housekeeping Gene Regulation

Muhammad A. Zabidi*, Cosmas D. Arnold*, Katharina Schernhuber, Michaela Pagani, Martina Rath, Olga Frank & Alexander Stark, *Nature*, 2015. 518(7540) pp. 556-559.

*These authors contributed equally.

In an animal genome, there are thousands of genes, associated with similarly numerous enhancers and core promoters. An unanswered question in gene regulation is whether core promoters have intrinsic preference towards enhancers. Isolated examples supporting this notion exist. However, the prevalence and the underlying properties of enhancer–core-promoter specificity have hitherto remained unanswered.

Core promoters of housekeeping and developmentally-related genes have distinct core promoter architectures. We therefore tested the activity of genome-wide enhancer candidate fragments from *Drosophila melanogaster* for their activities towards representatives of either of the two core promoter classes. We employed a genome-wide enhancer activity assay termed self-transcribing active regulatory region sequencing (STARR-seq). This assay enables us to test the candidate fragments for their ability to activate transcription from a given core promoter.

We show that thousands of enhancers in the genome exhibit high specificity towards housekeeping versus developmental-type core promoters. Housekeeping enhancers are active across cell types and associated with genes that perform housekeeping functions. Conversely, developmental enhancers are highly cell-type specific and are associated with tightly-regulated developmentally-related genes. These enhancer classes also carry distinct sequence and TF binding signatures. Housekeeping enhancers are enriched in DNA replicating element (DRE) and bound by the DRE factor (DREF), while developmental enhancers are enriched in GAGA motif, recognized by Trithorax-like (Trl) factor.

Our results show that the specificity between core promoters and enhancers is intrinsic in the sequence. This mechanism operates genome-wide, elegantly employed to separate the regulation of housekeeping and developmental gene transcription programs.

Author contributions

M.A.Z., C.D.A. and A.S. conceived the project. C.D.A., K.S., M.P., M.R. and O.F. performed the experiments and **M.A.Z.** the computational analyses. **M.A.Z.**, C.D.A. and A.S. wrote the manuscript.

Enhancer—core-promoter specificity separates developmental and housekeeping gene regulation

Muhammad A. Zabidi^{1*}, Cosmas D. Arnold^{1*}, Katharina Schernhuber¹, Michaela Pagani¹, Martina Rath¹, Olga Frank¹ & Alexander Stark¹

Gene transcription in animals involves the assembly of RNA polymerase II at core promoters and its cell-type-specific activation by enhancers that can be located more distally¹. However, how ubiquitous expression of housekeeping genes is achieved has been less clear. In particular, it is unknown whether ubiquitously active enhancers exist and how developmental and housekeeping gene regulation is separated. An attractive hypothesis is that different core promoters might exhibit an intrinsic specificity to certain enhancers^{2–6}. This is conceivable, as various core promoter sequence elements are differentially distributed between genes of different functions⁷, including elements that are predominantly found at either developmentally regulated or at housekeeping genes^{8–10}. Here we show that thousands of enhancers in *Drosophila melanogaster* S2 and ovarian somatic cells (OSCs) exhibit a marked specificity to one of two core promoters—one derived from a ubiquitously expressed ribosomal protein gene and another from a developmentally regulated transcription factor—and confirm the existence of these two classes for five additional core promoters from genes with diverse functions. Housekeeping enhancers are active across the two cell types, while developmental enhancers exhibit strong cell-type specificity. Both enhancer classes differ in their genomic distribution, the functions of neighbouring genes, and the core promoter elements of these neighbouring genes. In addition, we identify two transcription factors—Dref and Trl—that bind and activate housekeeping versus developmental enhancers, respectively. Our results provide evidence for a sequence-encoded enhancer—core-promoter specificity that separates developmental and housekeeping gene regulatory programs for thousands of enhancers and their target genes across the entire genome.

We chose the core promoter of *Ribosomal protein gene 12* (*RpS12*) and a synthetic core promoter derived from the *even skipped* transcription factor¹¹ as representative ‘housekeeping’ and ‘developmental’ core promoters, respectively (hereafter termed hkCP and dCP; Fig. 1a and Extended Data Figs 1, 2) and tested the ability of all candidate enhancers genome wide to activate transcription from these core promoters using self-transcribing active regulatory region sequencing (STARR-seq)¹² in *D. melanogaster* S2 cells. This set-up allows the testing of all candidates in a defined sequence environment, which differs only in the core promoter sequences but is otherwise constant^{12,13}.

Two hkCP STARR-seq replicates were highly similar (genome-wide Pearson correlation coefficient (PCC) 0.98; Extended Data Fig. 1c) and yielded 5,956 enhancers, compared with 5,408 enhancers obtained when we reanalysed dCP STARR-seq data¹² (Supplementary Table 1). Interestingly, the hkCP and dCP enhancers were largely non-overlapping (Fig. 1b, c) and the genome-wide enhancer activity profiles differed (PCC 0.38), as did the individual enhancer strengths: of the 11,364 enhancers, 8,144 (72%) activated one core promoter at least twofold more strongly than the other, a difference rarely seen in the replicate experiments for each of the core promoters (Fig. 1d). Indeed, 21 out of 24 hkCP-specific enhancers activated luciferase expression (>1.5-fold and t -test $P < 0.05$) from the hkCP versus 1 out of 24 from the dCP (Fig. 1e and Extended Data Fig. 3). Consistently, 10 out of 12 dCP-specific enhancers were positive with the dCP but only 2 out of 12 with the hkCP, a highly significant difference ($P = 5.1 \times 10^{-6}$, Fischer’s exact test) that confirms the enhancer—core-promoter specificity observed for thousands of enhancers across the entire genome.

Enhancers that were specific to either the hkCP or the dCP showed markedly different genomic distributions (Fig. 2a and Extended Data

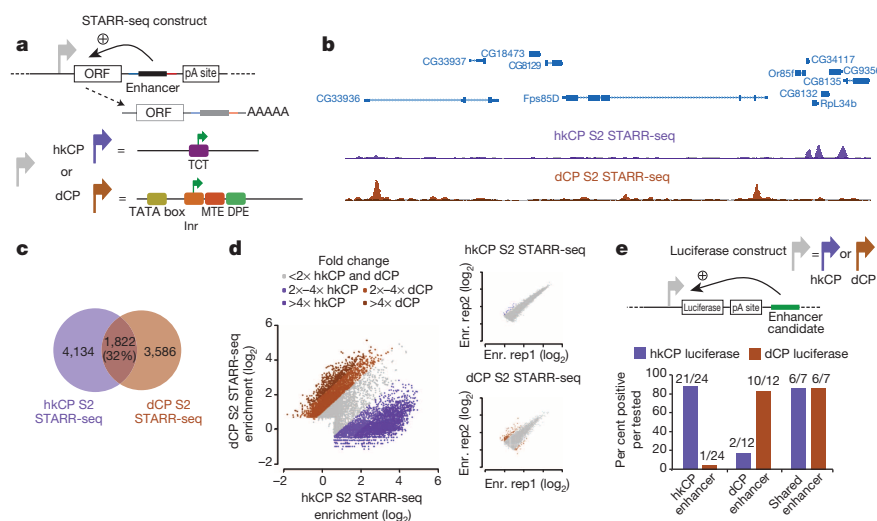


Figure 1 | Distinct sets of enhancers activate transcription from the hkCP and dCP in S2 cells. **a**, STARR-seq set-up using the hkCP housekeeping (*RpS12*; purple) and dCP developmental core promoters (*Drosophila* synthetic core promoter (DSCP)¹¹; brown) **b**, Genome browser screenshot depicting STARR-seq tracks for both core promoters. **c**, Overlap of hkCP and dCP enhancers. **d**, hkCP versus dCP STARR-seq enrichments at enhancers (insets show enrichment for replicates (Enr. rep) 1 versus 2 for hkCP and dCP; dCP data reanalysed from ref. 12). **e**, hkCP, dCP or shared enhancers that activate luciferase (>1.5-fold and $P < 0.05$ (one-sided t -test); $n = 3$; Extended Data Figs 3 and 5) from hkCP (purple) or dCP (brown; numbers show positive/tested).

¹Research Institute of Molecular Pathology IMP, Vienna Biocenter VBC, Dr Bohr-Gasse 7, 1030 Vienna, Austria.

*These authors contributed equally to this work.

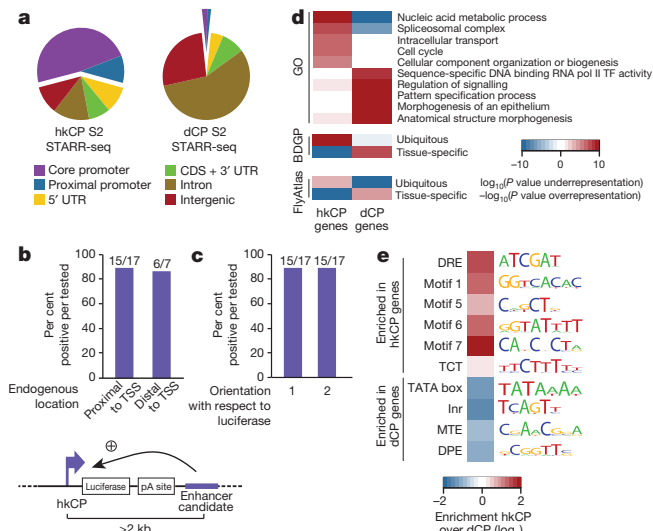


Figure 2 | hkCP and dCP enhancers differ in genomic distribution and flanking genes. **a**, Genomic distribution of hkCP and dCP enhancers. CDS, coding sequence; UTR, untranslated region. **b**, c, hkCP enhancers function distally in luciferase assays independent of their genomic positions (**b**) and orientation towards the luciferase TSS (**c**; orientation 1 from **b**; Extended Data Figs 3 and 5). **d**, e, GO (5 of the top 100 terms shown per column; Supplementary Table 11) and gene expression (terms curated from the Berkeley *Drosophila* Genome Project (BDGP) and FlyAtlas analyses (**d**) and enrichment of core promoter elements at TSSs (**e**) for genes next to hkCP and dCP enhancers. TF, transcription factor.

Fig. 4): whereas the majority (58.4%) of hkCP-specific enhancers overlapped with a transcription start site (TSS) or were proximal to a TSS (≤ 200 bp upstream; Fig. 2a), dCP-specific enhancers located predominantly to introns (56.5%) and intergenic regions (26.9%; Fig. 2a)¹². Importantly, despite the TSS-proximal location of most hkCP-specific enhancers, they activated transcription from a distal core promoter in STARR-seq (Fig. 1a and Extended Data Figs 1a, 2). Luciferase assays confirmed that they function from a distal position (>2 kb from the TSS) downstream of the luciferase gene and independently of their orientation towards the luciferase TSS (Fig. 2b, c and Extended Data Figs 3, 5). These results show that TSS-proximal sequences can act as bona fide enhancers¹⁴ and that developmental and housekeeping genes are both regulated through core promoters and enhancers, yet with a substantially different fraction of TSS-proximal enhancers (3.4% versus 58.4%).

hkCP and dCP enhancers were also located next to functionally distinct classes of genes according to gene ontology (GO) analyses: genes next to hkCP enhancers were enriched in diverse housekeeping functions

including metabolism, RNA processing and the cell cycle, whereas genes next to dCP enhancers were enriched for terms associated with developmental regulation and cell-type-specific functions (Fig. 2d, Extended Data Fig. 6a and Supplementary Tables 2–4). Consistently, hkCP enhancers were preferentially near ubiquitously expressed genes and dCP enhancers were near genes with tissue-specific expression (Fig. 2d and Supplementary Table 5).

The core promoters of the putative endogenous target genes of hkCP and dCP enhancers were also differentially enriched in known core promoter elements¹⁵ (Fig. 2e and Extended Data Fig. 6b): TSSs next to hkCP enhancers were enriched in Ohler motifs¹⁶ 1, 5, 6 and 7, consistent with the ubiquitous expression and housekeeping functions of these genes. In contrast, TSSs next to dCP enhancers were enriched in TATA box, initiator (Inr), motif ten element (MTE) and downstream promoter element (DPE) motifs, which are associated with cell-type-specific gene expression^{7,15}.

We next investigated whether the specificity that hkCP and dCP show to the two enhancer classes applies more generally. We selected three additional core promoters from housekeeping genes with different functions: from the *eukaryotic translation elongation factor 1 δ* (*eEF1 δ*), the putative splicing factor *x16*, and the cohesin loader *Nipped-B* (*NipB*). Importantly, all three contained combinations of core promoter elements that differed from that of hkCP, namely TCT⁸ and DNA-replication-related element (DRE) motifs (*eEF1 δ*), and Ohler motifs 1 and 6 (*x16* and *NipB*; Fig. 3a). In addition, we selected a DPE-containing core promoter of the transcription factor *pannier* (*pnr*) and the TATA-box core promoter of *Heat shock protein 70* (*Hsp70*), which can be activated by tissue-specific enhancers (for example, see ref. 17), thus covering the two most prominent core promoter types of regulated genes^{9,16,18}.

We performed STARR-seq for the five additional core promoters and grouped the genome-wide enhancer activity profiles of all seven core promoters by hierarchical clustering. This revealed two distinct clusters corresponding to the four housekeeping and the three developmental core promoters, respectively (Fig. 3b, Extended Data Fig. 7 and Supplementary Tables 6, 7), and the core promoters of both clusters indeed responded markedly differentially to individual genomic enhancers (Fig. 3c).

These results obtained for core promoters with diverse motif content and from genes with various functions suggest that the distinct enhancer preferences observed between hkCP and dCP apply more generally and that two broad classes of housekeeping and developmental (or regulated) core promoters exist. Differences within each class might correspond to differences in relative enhancer preferences of the core promoters^{2–6}, while similarities between both classes could reflect enhancers that are shared (Fig. 1c–e) or core promoters that can be activated to different extents by enhancers from both classes (for example,

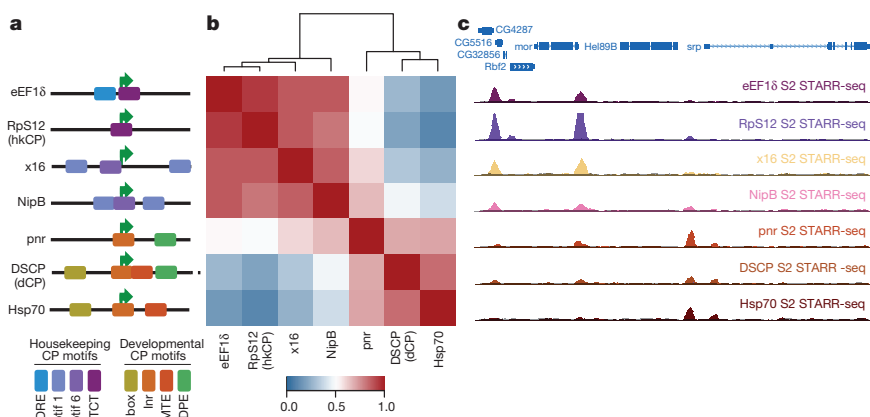


Figure 3 | Housekeeping and developmental core promoters differ characteristically in their enhancer preferences. **a**, Different housekeeping (top 4) and developmental-like (bottom 3) core promoters and their motif content (schematic). **b**, Bi-clustered heat map depicting pairwise similarities of STARR-seq signals (PCCs at peak summits). PCCs and dendrogram (top) show the separation between housekeeping and regulated core promoters. **c**, Genome browser screenshot depicting STARR-seq tracks for all seven core promoters.

RESEARCH LETTER

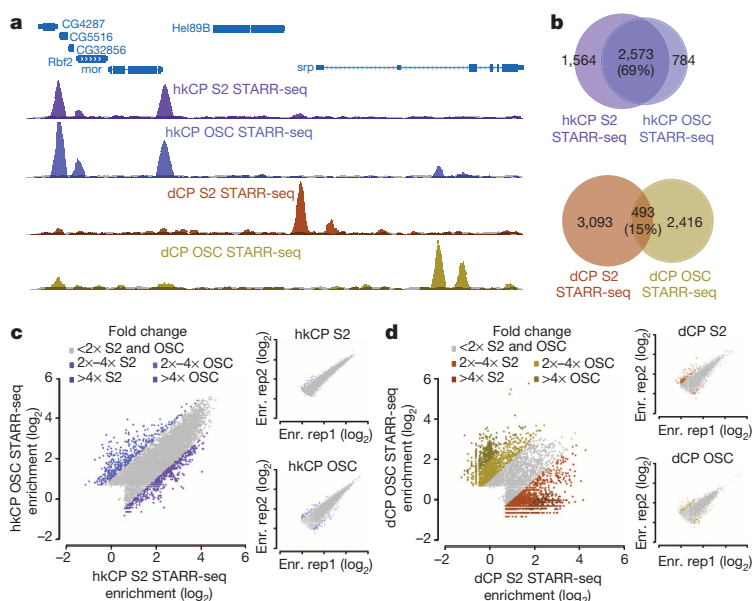


Figure 4 | hkCP enhancers are shared across cell types. **a**, Genome browser screenshot showing tracks for hkCP (top) and dCP STARR-seq (bottom) in S2 cells and OSCs. **b**, Overlap of hkCP (top) and dCP (bottom) enhancers between S2 cells and OSCs. **c**, **d**, hkCP (**c**) and dCP (**d**) STARR-seq enrichments in S2 cells versus OSCs at hkCP- or dCP-specific enhancers (insets show enrichments for replicates (Enr. rep) 1 versus 2; dCP data reanalysed from ref. 12).

NipB; Fig. 3b, c). The latter might be important if broadly expressed housekeeping genes need to be further activated in specific tissues.

To test whether hkCP enhancers function in different cell types, we performed STARR-seq using hkCP in OSCs, which differ strongly from S2 cells in gene expression and dCP enhancer activities¹². Two hkCP STARR-seq replicates in OSCs were highly similar (PCC 0.97) and yielded 6,217 enhancers (Supplementary Table 1), compared with 5,774 enhancers obtained for dCP data from OSCs¹². The OSC data confirmed the differences between hkCP and dCP enhancers observed in S2 cells (Extended Data Figs 8, 9 and Supplementary Tables 8–10). Strikingly,

the hkCP-specific enhancers in OSCs and S2 cells (3,357 and 4,137, respectively) were almost indistinguishable, whereas dCP-specific enhancers (2,909 in OSCs and 3,586 in S2 cells) differed strongly between the two cell types¹² and from the hkCP enhancers (Fig. 4a). The observation that hkCP enhancers showed similar activities in both cell types while dCP enhancers were cell-type specific was true genome wide when comparing genomic locations (69% versus 15% overlap) or enhancer strengths as measured by STARR-seq (PCC at peak summits 0.83 versus 0.05; Fig. 4b–d and Extended Data Fig. 9c). Together, these results show that hkCP enhancers are shared between two different cell types,

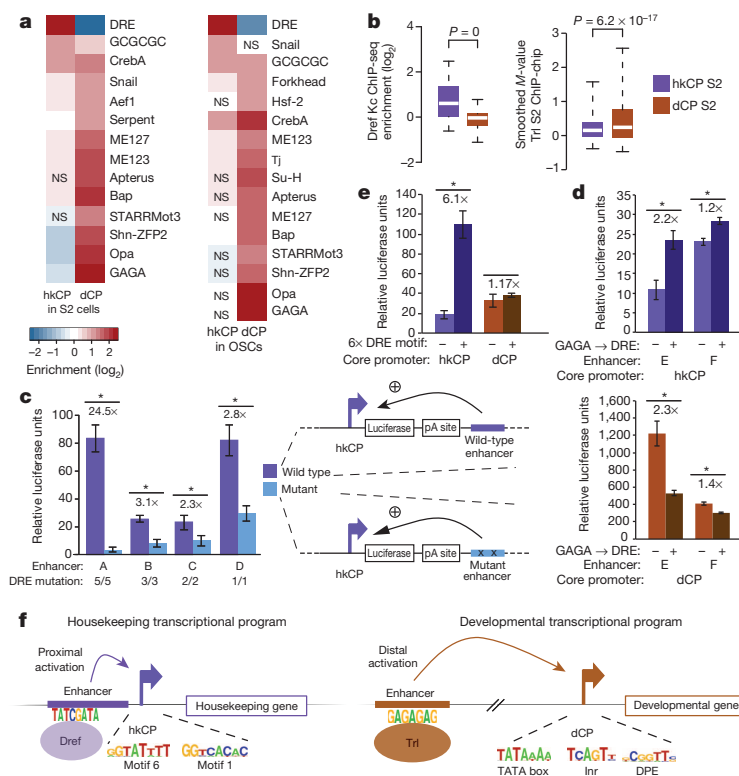


Figure 5 | hkCP and dCP enhancers depend on Dref and Trl, respectively. **a**, **b**, Motif enrichment (**a**) and ChIP signals for Dref and Trl (**b**) in hkCP and dCP enhancers. False discovery rate (FDR)-corrected hypergeometric $P > 0.01$; boxes: median and interquartile range; whiskers: 5th and 95th percentiles; two-sided Wilcoxon-rank-sum P values. NS, not significant. **c**, Luciferase assays for four wild-type and DRE-motif-mutant hkCP enhancers (numbers show mutated motifs). Error bars show standard deviation (s.d.) ($n = 3$, biological replicates). $*P < 0.005$ (one-sided t -test). **d**, Luciferase assays for two dCP enhancers (–) and their GAGA → DRE-mutant variants (+) with hkCP (top) and dCP (bottom; details as in c). **e**, Luciferase assays for an array of DRE motifs with hkCP and dCP (details as in c). **f**, Model: housekeeping genes contain Ohler motifs 1, 5, 6, 7 and/or the TCT motif and are activated by TSS-proximal hkCP enhancers via Dref. Regulated genes contain TATA box, Inr, MTE and/or DPE and are activated by distal dCP enhancers via Trl.

whereas dCP enhancers are cell-type specific¹², presumably representing ubiquitous housekeeping versus developmental and cell-type-specific gene expression programs.

To assess whether the marked core promoter specificities of the hkCP and dCP enhancers are encoded in their sequences, we analysed the *cis*-regulatory motif content of both classes of enhancers¹⁹. This revealed a strong enrichment of the DRE motif in hkCP enhancers (Fig. 5a and Supplementary Tables 11, 12), whereas dCP enhancers were strongly enriched in the GAGA motif of *Trithorax-like* (*Trl*) and other motifs previously described to be important for dCP enhancers²⁰. Published genome-wide chromatin immunoprecipitation (ChIP) data^{21,22} confirmed that DRE-binding factor (Dref) bound significantly more strongly to hkCP enhancers than to dCP enhancers (Wilcoxon $P = 0$; Fig. 5b), while the opposite was true for *Trl* (Wilcoxon $P = 6.2 \times 10^{-17}$). Considering only distal enhancers (> 500 bp from the closest TSS) yielded the same results (Extended Data Fig. 10a, b and Supplementary Tables 13, 14), suggesting that the differential occupancy is a property of both classes of enhancers rather than a consequence of the different extents to which they overlap with TSSs. Disrupting the DRE motifs in four different hkCP enhancers substantially reduced the activities of the enhancers as measured by luciferase assays in S2 cells (between 2.3- and 24.5-fold reduction; Fig. 5c), while dCP enhancers depend on GAGA motifs²⁰. Adding DRE motifs to 11 different dCP enhancers significantly increased luciferase expression from the hkCP for 9 of them (82%; Extended Data Fig. 10c), and changing the GAGA motifs of two dCP enhancers to DRE motifs significantly increased the activities of both enhancers towards the hkCP but decreased their activities towards the dCP (Fig. 5d). Furthermore, an array of six DRE motifs was sufficient to activate luciferase expression from the hkCP but not the dCP (Fig. 5e). Together, these results show that hkCP and dCP enhancers depend on DRE and GAGA motifs, respectively, and demonstrate that DRE motifs are required and sufficient for hkCP enhancer function.

Our results show that developmental and housekeeping gene regulation is separated genome wide by sequence-encoded specificities of thousands of enhancers to one of two types of core promoter, supporting the longstanding 'enhancer-core-promoter specificity' hypothesis^{2-6,23}. Our findings indicate that these specificities are probably mediated by defined biochemical compatibilities²⁴ between different *trans*-acting factors such as Dref versus *Trl* (at enhancers) and the different paralogues that exist for several components of the general transcription apparatus (at core promoters), presumably including the TATA-box-binding protein-related factor 2 (Trf2) at housekeeping core promoters^{25,26}. As such paralogues can have tissue-specific expression and stage-specific or promoter-selective functions^{27,28} (reviewed in refs 29, 30), sequence-encoded enhancer-core-promoter specificities could be used more widely to define and separate different transcriptional programs (Fig. 5f).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 22 May; accepted 20 October 2014.

Published online 15 December 2014; corrected online 25 February 2015 (see full-text HTML version for details).

- Levine, M., Cattoglio, C. & Tjian, R. Looping back to leap forward: transcription enters a new era. *Cell* **157**, 13–25 (2014).
- Li, X. & Noll, M. Compatibility between enhancers and promoters determines the transcriptional specificity of gooseberry and gooseberry neuro in the *Drosophila* embryo. *EMBO J.* **13**, 400–406 (1994).
- Ohtsuki, S., Levine, M. & Cai, H. N. Different core promoters possess distinct regulatory activities in the *Drosophila* embryo. *Genes Dev.* **12**, 547–556 (1998).
- Sharpe, J., Nonchev, S., Gould, A., Whiting, J. & Krumlauf, R. Selectivity, sharing and competitive interactions in the regulation of *Hoxb* genes. *EMBO J.* **17**, 1788–1798 (1998).
- Merli, C., Bergstrom, D. E., Cygan, J. A. & Blackman, R. K. Promoter specificity mediates the independent regulation of neighboring genes. *Genes Dev.* **10**, 1260–1270 (1996).
- Butler, J. E. & Kadonaga, J. T. Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes Dev.* **15**, 2515–2519 (2001).
- Kadonaga, J. T. Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip. Rev. Dev. Biol.* **1**, 40–51 (2012).
- Parry, T. J. et al. The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes Dev.* **24**, 2013–2018 (2010).
- Engström, P. G., Ho Sui, S. J., Drivenes, O., Becker, T. S. & Lenhard, B. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res.* **17**, 1898–1908 (2007).
- FitzGerald, P. C., Sturgill, D., Shyakhtenko, A., Oliver, B. & Vinson, C. Comparative genomics of *Drosophila* and human core promoters. *Genome Biol.* **7**, R53 (2006).
- Pfeiffer, B. D. et al. Tools for neuroanatomy and neurogenetics in *Drosophila*. *Proc. Natl Acad. Sci. USA* **105**, 9715–9720 (2008).
- Arnold, C. D. et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
- Shlyueva, D. et al. Hormone-responsive enhancer-activity maps reveal predictive motifs, indirect repression, and targeting of closed chromatin. *Mol. Cell* **54**, 180–192 (2014).
- Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β -globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).
- Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Rev. Genet.* **13**, 233–245 (2012).
- Ohler, U., Liao, G.-C., Niemann, H. & Rubin, G. M. Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol.* **3**, research0087.1–0087.12 (2002).
- Smith, D., Wohlgemuth, J., Calvi, B. R., Franklin, I. & Gelbart, W. M. *hobo* enhancer trapping mutagenesis in *Drosophila* reveals an insertion specificity different from *P* elements. *Genetics* **135**, 1063–1076 (1993).
- Kutach, A. K. & Kadonaga, J. T. The downstream promoter element DPE appears to be as widely used as the TATA box in *Drosophila* core promoters. *Mol. Cell Biol.* **20**, 4754–4764 (2000).
- Yáñez-Cuna, J. O., Dinh, H. Q., Kwon, E. Z., Shlyueva, D. & Stark, A. Uncovering *cis*-regulatory sequence requirements for context-specific transcription factor binding. *Genome Res.* **22**, 2018–2030 (2012).
- Yáñez-Cuna, J. O. et al. Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. *Genome Res.* **24**, 1147–1156 (2014).
- Gurudatta, B. V., Yang, J., Van Bortle, K., Donlin-Asp, P. G. & Corces, V. G. Dynamic changes in the genomic localization of DNA replication-related element binding factor during the cell cycle. *Cell Cycle* **12**, 1605–1615 (2013).
- modENCODE Consortium Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**, 1787–1797 (2010).
- Ohler, U. & Wassarman, D. A. Promoting developmental transcription. *Development* **137**, 15–26 (2010).
- van Arensbergen, J., van Steensel, B. & Bussemaker, H. J. In search of the determinants of enhancer-promoter interaction specificity. *Trends Cell Biol.* <http://dx.doi.org/10.1016/j.tcb.2014.07.004> (2014).
- Wang, Y.-L. et al. TRF2, but not TBP, mediates the transcription of ribosomal protein genes. *Genes Dev.* **28**, 1550–1555 (2014).
- Isogai, Y., Keles, S., Prestel, M., Hochheimer, A. & Tjian, R. Transcription of histone gene cluster by differential core-promoter factors. *Genes Dev.* **21**, 2936–2949 (2007).
- Hochheimer, A., Zhou, S., Zheng, S., Holmes, M. C. & Tjian, R. TRF2 associates with DREF and directs promoter-selective gene expression in *Drosophila*. *Nature* **420**, 439–445 (2002).
- Deato, M. D. E. & Tjian, R. Switching of the core transcription machinery during myogenesis. *Genes Dev.* **21**, 2137–2149 (2007).
- D'Alessio, J. A., Wright, K. J. & Tjian, R. Shifting players and paradigms in cell-specific transcription. *Mol. Cell* **36**, 924–931 (2009).
- Müller, F., Zaucker, A. & Tora, L. Developmental regulation of transcription initiation: more than just changing the actors. *Curr. Opin. Genet. Dev.* **20**, 533–540 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank L. Cochella and O. Bell for comments on the manuscript. Deep sequencing was performed at the CSF Next-Generation Sequencing Unit (<http://csf.ac.at>). M.A.Z. was supported by the Austrian Science Fund (FWF, F4303-B09) and C.D.A., K.S., M.R. and O.F. by a European Research Council Starting Grant (no. 242922) awarded to A.S. Basic research at the Research Institute of Molecular Pathology is supported by Boehringer Ingelheim GmbH.

Author Contributions M.A.Z., C.D.A. and A.S. conceived the project. C.D.A., K.S., M.P., M.R. and O.F. performed the experiments and M.A.Z. the computational analyses. M.A.Z., C.D.A. and A.S. wrote the manuscript.

Author Information All deep sequencing data are available at <http://www.starklab.org> and have been deposited in the Gene Expression Omnibus database under accession numbers GSE40739 and GSE57876. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.S. (stark@starklab.org).

RESEARCH LETTER

METHODS

hkCP STARR-seq vector. We derived the hkCP STARR-seq vector from the original STARR-seq vector¹² by replacing the DSCP sequence with the sequence of the *RpS12* core promoter (−50 to +50 bp relative to the TSS; TTGTACCAATAGCT AAAAAGTCAATCTCCAGCGCCATGCGGATTTTGTCTCTTCTTTCCG GTTGTCAAAAGGTACAGATGCTTGGATTTTATTTCTC). The STARR-seq vectors are available subject to a material transfer agreement (MTA). For both STARR-seq vectors, we confirmed that transcription initiates from within the respective core promoters' Inr (DSCP) and TCT (*RpS12*) motifs by 5' rapid amplification of cDNA ends (RACE; Extended Data Fig. 2). All other STARR-seq vectors were derived from the hkCP STARR-seq vector by replacing the 100 bp sequence encompassing the *RpS12* core promoter by the sequences indicated in Supplementary Table 15 using the BglII and SbfI restriction sites.

hkCP and dCP luciferase vectors. For the dCP luciferase vector, the SV40 promoter of the pGL3-Promoter Vector (Promega) was replaced by the DSCP¹¹ and a Gateway cassette was inserted downstream of the luciferase gene and the SV40 polyA-signal into the AfeI restriction site, to allow Gateway LR cloning of candidate sequences¹². For the hkCP luciferase vector, the SV40 promoter and the sequence until the translation start codon of the luciferase gene was replaced by the sequence encompassing the TSS of *RpS12* from −50 bp until its translation start codon: TTGTACCAATAGCTAAAACTCACATCTCCAGCGCCATGCCGA TTTTGTCTCTTTCTTTCCGGTTGTCAAAAGGTACAGATGCTTGGATTT TATTTCTCCGAAATGAAGAGGTTTCTTATCGAAAATGTAATAAATATG AACAAATTAATCTATCTTTCCAGTGCAGTGCATCCTTAACCGCAGAAAC. Constructs are available subject to an MTA.

Intrinsic activity of core promoters. All core promoters used in this study were cloned into the dCP luciferase vector (without the Gateway cassette), replacing the DSCP between the BglII and SbfI restriction site with the respective core promoter. For each core promoter, the intrinsic (or basal) activity was measured as firefly luciferase activity and is presented as relative luciferase units, normalized to *Renilla* luciferase signals.

Genome-wide STARR-seq screens. STARR-seq enhancer screens using the core promoters of *RpS12* (hkCP), *NipB*, *x16*, and *eFl1δ* (Supplementary Table 15) were performed in two biological replicates (independent transfections) as described previously¹² with the following exceptions. First, 1.6×10^9 S2 cells and OSCs³¹ were transfected per biological replicate. Second, first-strand cDNA synthesis was performed in 30–60 reactions with the STARR-seq RT primer (CTCATCAATGTAT CTATCATGTCTG) as reverse transcription primer. Last, next-generation sequencing (NGS) was performed on an Illumina HiSeq 2000 machine using multiplexing according to the manufacturer's instructions. STARR-seq data using the DSCP (dCP STARR-seq) and *Hsp70* core promoters are from ref. 12, but were reanalysed using the same pipeline as for hkCP STARR-seq.

Focused STARR-seq BAC screens. The DSCP is a 137-nucleotide-long synthetic core promoter derived from the core promoter of *even-skipped* (*eve*)¹¹. To assess the functional similarity of the DSCP, its 137-nucleotide-long wild-type counterpart from the *eve* locus, and a version defined identically to all other core promoter used here (−50 to +50 nucleotides around the TSS), we performed STARR-seq screens with libraries derived from 29 different BACs containing a total of ~5 Mb of *D. melanogaster* genomic DNA (Supplementary Table 16). For comparison, we also screened all other core promoters with this library. For library cloning, all BACs were grown in individual bacterial cultures and were then mixed equally according to measurements of their optical density at 600 nm ($OD_{600\text{ nm}}$) before BAC DNA isolation to achieve an equal distribution of all BACs. BAC DNA extraction, sonication and adaptor ligation was performed as described¹² and the same adaptor-ligated and PCR-amplified BAC DNA was used to clone all focused STARR-seq libraries. Per STARR-seq vector, four In-Fusion reactions were performed, which allowed five transformation reactions as described¹². Each library was grown in 4 l liquid culture (LB medium) to an $OD_{600\text{ nm}}$ of 2.0–2.5. Each BAC library was screened as described earlier for the genome-wide screens; however, only 1×10^8 S2 cells were used, accounting for the less complex library. Similarly, the number of reactions for all subsequent steps of the STARR-seq protocol was reduced fourfold.

Luciferase reporter assays. Luciferase assays were performed as described previously¹² with the exception that the candidate enhancers were cloned downstream of the luciferase gene and the polyA signal, more than 2 kb away from the respective core promoter (*RpS12* or DSCP). Candidate enhancers were selected manually based on different criteria to allow the systematic assessment of several aspects of this study, including enhancers that were (1) specific to one of the two different core promoters (24 hkCP and 12 dCP enhancers) or found in both screens (7 shared enhancers); (2) located proximally (17) or distally (7) to the hkCP; and (3) of different strengths according to STARR-seq (ranks 18 to 1,044). We cloned all candidates as described¹² (for their genomic coordinates and primer sequences see Supplementary Table 17), picking initially one orientation towards the luciferase TSS randomly. However, to test the influence of TSSs contained in the candidate

sequences, we cloned and tested all TSS proximal candidates (hkCP_01 to hkCP_17) in both orientations using both core promoters. Candidate enhancers with DRE mutations were cloned from synthesized DNA fragments (GeneArt Strings; Supplementary Table 18). Candidates with DRE motifs that replace GAGA motifs were cloned similarly using synthesized DNA fragments (gBlocks) obtained from Integrated DNA Technologies (Supplementary Table 19). We also added an array of 6× DRE motifs into the AfeI restriction site of the dCP and hkCP luciferase vectors and cloned dCP_01 to dCP_11 into the middle of the DRE motif array (using AfeI) of the hkCP luciferase vector, such that these sequences were each flanked by three DRE motifs (Supplementary Table 19).

Luciferase assay data analysis. For all luciferase assays, we calculated standard deviations and one-sided Student's *t*-tests from three biological replicates (independent transfections). Core promoters have intrinsic (basal) activities that can differ between different core promoters. Therefore, when comparing enhancer activities for different core promoters, normalization to the core promoters' intrinsic activities is required, which we assessed with three different negative control fragments (nine biological replicates in total). For all measurements, we normalized firefly luciferase values first to *Renilla* luciferase values (controlling for transfection efficiency) and then to the normalized luciferase values of the three negative control sequences. Candidates with a significant ($P < 0.05$) enrichment greater than 1.5 fold over negative were considered positive.

5' RACE of STARR-seq transcripts. To determine the exact TSSs of hkCP and dCP within the STARR-seq vectors we performed 5' RACE of STARR-seq transcripts using one enhancer for each (an intergenic enhancer of *TpnC41C* for hkCP and an intronic enhancer of *zfh1* (shared_01 from ref. 12) for dCP) which we cloned with EcoRV at the position of the selection cassette used during library cloning (Supplementary Table 20). We transfected 3.2×10^7 cells with each of the constructs and isolated total RNA using the RNeasy mini prep kit (Qiagen; two columns per construct) followed by polyA+ RNA isolation using oligo-dT Dynabeads (Life Technologies) according to the manufacturer's instructions. We then performed 5' RACE for both samples using the FirstChoice RLM-RACE Kit (Ambion; catalogue no. AM1700) according to the manufacturer's instructions. To reflect RNA processing of the STARR-seq pipeline, reverse transcription was, however, performed using SuperscriptIII (Invitrogen) according to the manufacturer's instructions and using the reverse transcription primer GFP-RT (Supplementary Table 20) as a gene-specific primer (using RNA amounts according to the FirstChoice manual). The first PCR was performed with the manufacturer-provided 5' RACE Outer Primer and the transcript-specific primer RACE-01-rv, using 2× KAPA HiFi Hot Start Ready Mix (98 °C for 45 s; followed by 35 cycles of 98 °C for 15 s, 69 °C for 30 s, 72 °C for 30 s) with 1 µl of cDNA as template. The nested PCR was performed similarly (primer: 5' RACE Inner Primer and RACE-02-rv; 98 °C for 45 s; followed by 30 cycles of 98 °C for 15 s, 67 °C for 30 s, 72 °C for 10 s). The PCR products were visualized on a 1% agarose gel. The PCR products for both samples were Sanger sequenced using the primer GFP-seq-rv (for all primer sequences see Supplementary Table 20).

STARR-seq NGS data processing. Paired-end STARR-seq and input read processing was performed as described³². The NGS data for dCP (DSCP) and *Hsp70* were obtained from ref. 12 and reanalysed. In the same cell line, a hkCP peak is considered to be 'specific' if the 501 bp window centred at the peak summit does not overlap with any such window for dCP peaks, and vice versa (note that this is only applied within each cell type, such that comparisons across cell types are not influenced). For screens with the BAC-derived libraries, we considered only fragments that originated from the BACs used and determined the relative abundance of each BAC from the NGS data of the respective inputs only. On the basis of this, we then adjusted both inputs and STARR-seq NGS data such that all BACs were equally represented and analysed the data as described earlier.

Venn diagrams and peak intersection. We used the same intersection method as described earlier, and plotted the Venn diagrams with areas proportional to the number of peaks.

Scatter plots. We calculated the STARR-seq enrichment over input at the summit positions of both data sets that were to be compared, using a pseudo count of 1, and computed the $\log_2$ of corrected ratio as described¹². This plots one data point for each enhancer—even for closely spaced ones—exactly at the enhancer's summit position. For visualizing replicates, we called peaks on the merged data sets and plotted the values from both replicates at these peaks' summits.

Enhancer-to-gene assignment. We performed three different strategies of enhancer-to-gene assignments: (1) 'closest TSS', whereby an enhancer is assigned to the closest TSS of an annotated transcript; (2) '1 kb TSS', whereby an enhancer is assigned to all TSSs that are within 1 kb; and (3) 'gene loci', whereby an enhancer is assigned to a gene provided that it falls within 5 kb upstream from the TSS, within the gene body itself, or 2 kb downstream of the gene (multiple assigned genes are possible). In all cases we used annotation from *D. melanogaster* FlyBase release 5.50.

Genomic distribution. We assigned a unique annotation for each nucleotide in the genome by using the following priority order: coding sequence (CDS), core promoter (± 50 bp around TSS), 5' UTR, 3' UTR, first intron, intron, proximal promoter (200 bp upstream of a TSS), intergenic region. We then assigned each peak to one of these categories by the annotation of the peak's summit.

GO analysis. We assessed whether genes assigned to hKCP or dCP enhancers were enriched for particular GO categories³³ by calculating hypergeometric *P* values for all categories, which we corrected for multiple comparisons (FDR-type correction in R). We then sorted all categories according to *P* values of overrepresentation, selected the top 100 of either hKCP or dCP, and removed redundant categories manually. For each category, we calculated $\log_{10}(P\text{-value underrepresentation}) - \log_{10}(P\text{-value overrepresentation})$, and sorted the terms in a descending order of difference between hKCP and dCP values. The colour intensity of the heat maps represents $\log_{10}(P\text{-value underrepresentation}) - \log_{10}(P\text{-value overrepresentation})$.

Gene expression analysis. We analysed enrichment in ubiquitous versus tissue-specific gene expression sets as described for the GO analysis above. To define the gene sets based on an *in situ* hybridization data set of fly embryos (BDGP³⁴), we first removed maternal (stages 1 to 3) annotations, as well as genes with the annotation 'no staining' in all stages. We required each gene to have annotations for at least three stage groupings. We called a gene 'tissue specific' if at most one of these annotations contains the word 'ubiquitous', and called it 'ubiquitous' if at least 60% of them contain word 'ubiquitous'. We also defined gene sets based on microarray data sets from dissected fly tissues (FlyAtlas³⁵). We defined genes as 'ubiquitous' if their expression does not change more than twofold compared with the whole fly for at least 15 out of 23 tissues. For this, we used the ratios and 'change_direction' calls from FlyAtlas directly and did not consider cell lines and carcasses. We similarly defined genes to be 'tissue specific' if they change more than twofold in at least three tissues. We do not consider genes with multiple conflicting entries as they can result from the use of multiple probes and removed genes that overlapped between the 'ubiquitous' and 'tissue-specific' gene sets from both sets.

Transcription factor motif and core promoter element enrichment analysis. We used previously employed position weight matrices (PWMs) for different transcription factors¹³ with a cut-off of $4^{-6} = 2.4 \times 10^{-4}$. We selected random control regions by controlling for genomic and chromosome distribution, and required that they did not overlap with any peak. We scored each motif for its enrichment in 401 bp windows centred on the peak summits by multiple testing (FDR) corrected hypergeometric *P* values. We considered only motifs that showed $\log_2(\text{confidence ratio of motif counts in peak windows/motif counts in random control regions}) > 1$ and *P* value < 0.01 in hKCP or dCP enhancers (or both) and reduced motif redundancy by removing highly similar motifs as in ref. 13 and references therein. We sorted the motifs in a descending order by difference in $\log_2(\text{hKCP enrichment}) - \log_2(\text{dCP enrichment})$. When assessing whether the observed motif distribution persisted for distal enhancers (Extended Data Fig. 10a), we kept the motifs and their order as in Fig. 5a and only re-evaluated their enrichment in distal enhancers. The colour intensity of the heat maps represents $\log_2(\text{confidence ratio of motif counts in peak windows/motif counts in random control regions})$. We used previously published PWMs or created PWMs from published nucleotide counts for TATA box, Inr, MTE, DPE and Ohler motifs¹⁶ 1, 5, 6, 7 and the TCT motif⁸ restricted to 8 bp. We scanned for motif occurrences using MAST from the MEME suite³⁶ (version 4.9.0) and parameters that ensured specificity and sensitivity for each motif (Supplementary Table 21). For enhancer-to-gene assignment methods 1 and 2 described earlier, we determined the presence of each core promoter element in the core promoter region of all genes uniquely assigned to either hKCP or dCP enhancers, respectively. For assignment method 3, we took the core promoter elements of the TSSs of the longest messenger RNA isoform. We assessed the differential distribution of each core promoter element between the core promoters assigned to hKCP or dCP enhancers by confidence ratios and hypergeometric *P* values.

Transcription factor motif and core promoter element *de novo* discovery. We used MEME³⁶ (version 4.9.0) to discover *de novo* motifs with lengths between 5 and 8 nucleotides in the enhancer regions we identified using STARR-seq and in

the core promoter regions around the nearest annotated transcription TSS. We provide all discovered motifs in Supplementary Table 22.

Core promoter similarity heat map. For all pairs of core promoters, we computed pair-wise PCCs between the respective STARR-seq fragment coverages at the summits of all peaks called in either of the two screens genome wide. We performed hierarchical clustering (complete linkage) in R, directly using the computed PCC values as similarities.

STARR-seq enrichment heat map. We computed the $\log_2$ of the corrected STARR-seq enrichment over input as described earlier, but for each nucleotide in a 20 kb window around all reference peak summit positions, and down-sampled the data points 50-fold by calculating one average data point per 50 nucleotides.

STARR-seq enrichment meta-profiles around TSSs. We calculated corrected STARR-seq enrichments ($\log_2$) as for the heat maps, but for 20 kb windows around TSSs, selected according to their core promoter motif content (see Extended Data Figs 4 and 8), corrected for the orientation of the TSSs within the genomic sequence. We then calculated the average for each position along the x-axis.

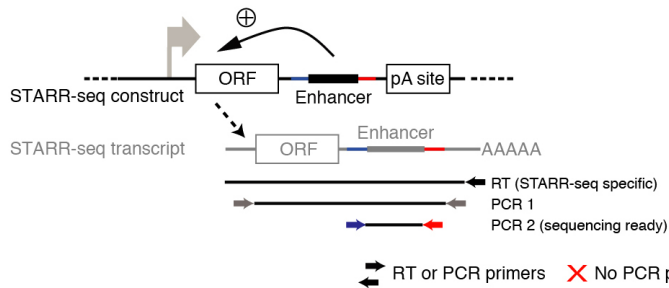
Boxplot. We obtained Dref ChIP-seq and input data (from Kc167 cells) from ref. 21 (Gene Expression Omnibus accession numbers GSM977024 and GSM762849) and mapped the 36-nucleotide reads using bowtie³⁷ (version 0.12.9) with the following parameters: -p 4 -q -v 3 -m 1 --best --strata --quiet. We extended the reads to 150 bp, calculated the coverage for ChIP-seq and input at the STARR-seq peak summit, normalized the value to the number of input fragments, added a pseudo count of 1, and computed the confidence ratio of ChIP-seq over input. For the Trl ChIP-chip data obtained from ref. 22, we used the signal of the chip-array probe at the peak summit if available or inferred the signal by linear extrapolation from the two nearest flanking probes (one on each side) provided that they were both within 10 nucleotides of the peak summit. We calculated statistical significance via Wilcoxon's paired rank tests.

Coordinate intersections. We performed genomic coordinate intersections using the BEDTools suite³⁸ (version 2.17.0).

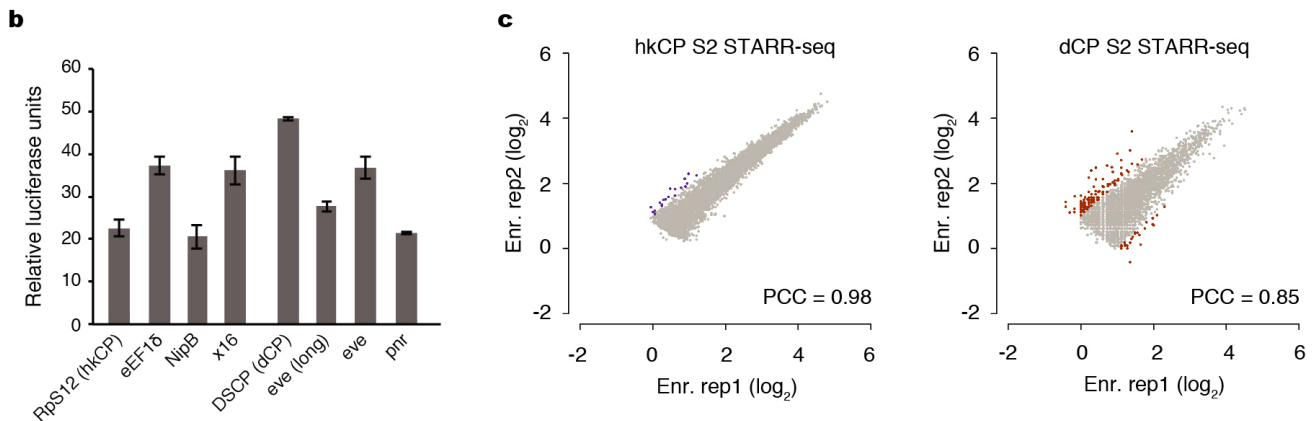
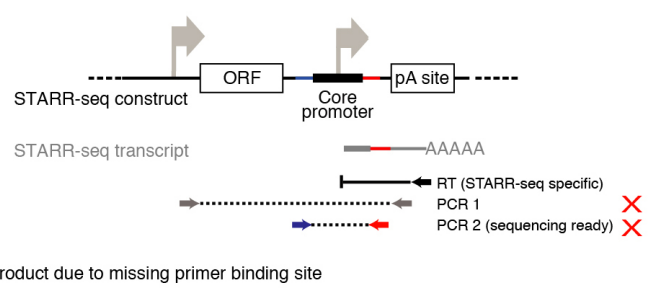
Statistics. We performed all statistical calculations and created graphical displays with R³⁹.

31. Saito, K. *et al.* A regulatory circuit for *piwi* by the large Maf gene *traffic jam* in *Drosophila*. *Nature* **461**, 1296–1299 (2009).
32. Arnold, C. D. *et al.* Quantitative genome-wide enhancer activity maps for five *Drosophila* species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nature Genet.* **46**, 685–692 (2014).
33. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29 (2000).
34. Tomancak, P. *et al.* Global analysis of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.* **8**, R145 (2007).
35. Chintapalli, V. R., Wang, J. & Dow, J. A. T. Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. *Nature Genet.* **39**, 715–720 (2007).
36. Bailey, T. L. & Gribskov, M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**, 48–54 (1998).
37. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
38. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
39. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2010).
40. Zeitlinger, J. & Stark, A. Developmental gene regulation in the era of genomics. *Dev. Biol.* **339**, 230–239 (2010).
41. Kvon, E. Z. *et al.* Genome-scale functional characterization of *Drosophila* developmental enhancers *in vivo*. *Nature* **512**, 91–95 (2014).
42. Soler, E. *et al.* The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes Dev.* **24**, 277–289 (2010).
43. Chen, K. *et al.* A global change in RNA polymerase II pausing during the *Drosophila* midblastula transition. *eLife* **2**, e00861 (2013).
44. Lagha, M. *et al.* Paused Pol II coordinates tissue morphogenesis in the *Drosophila* embryo. *Cell* **153**, 976–987 (2013).
45. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950–953 (2013).

RESEARCH LETTER

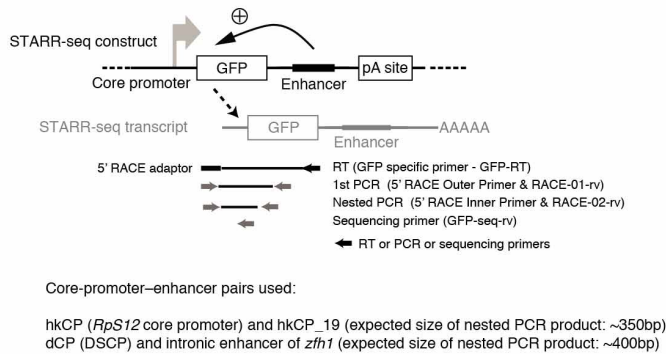
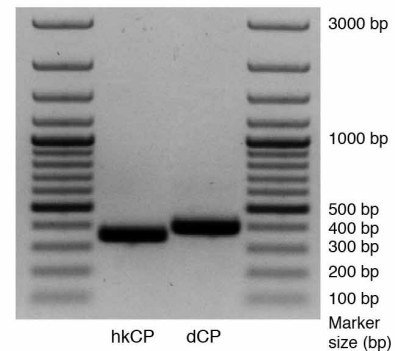
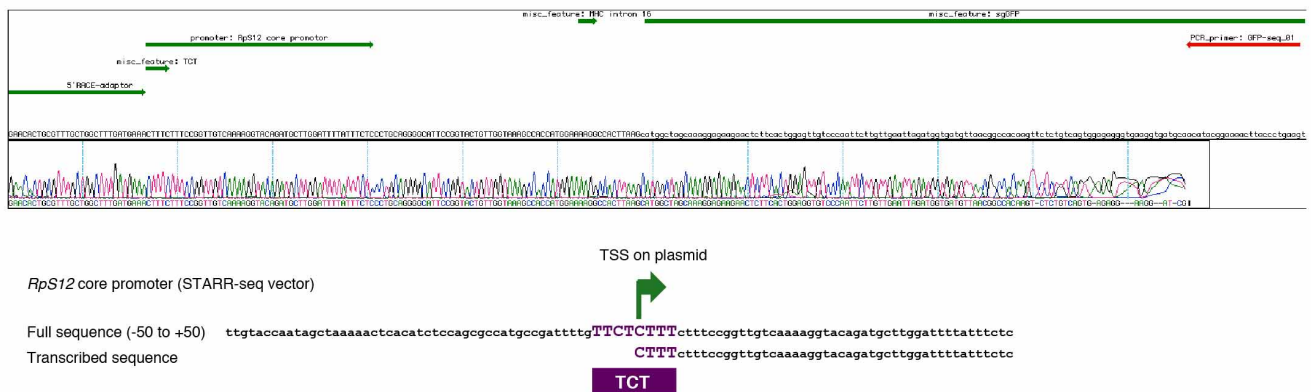
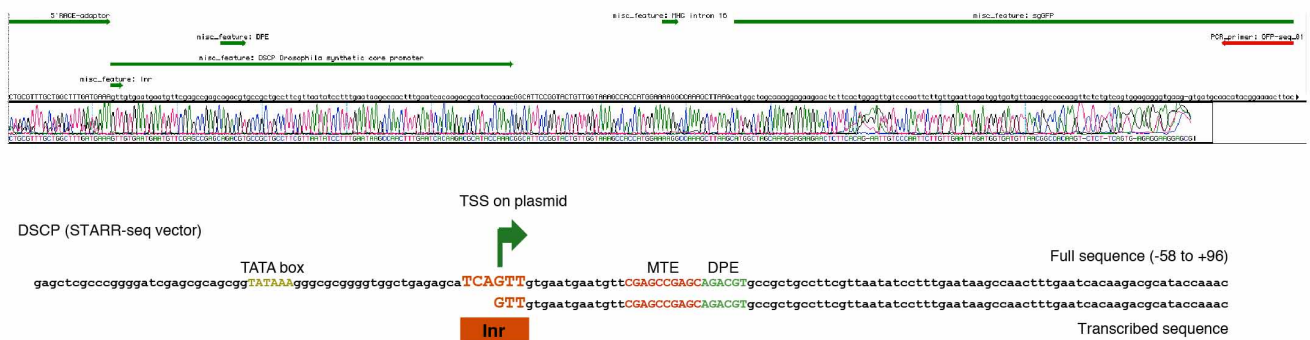
a Candidate fragment is an enhancer → detected

Candidate fragment is a core promoter → not detected



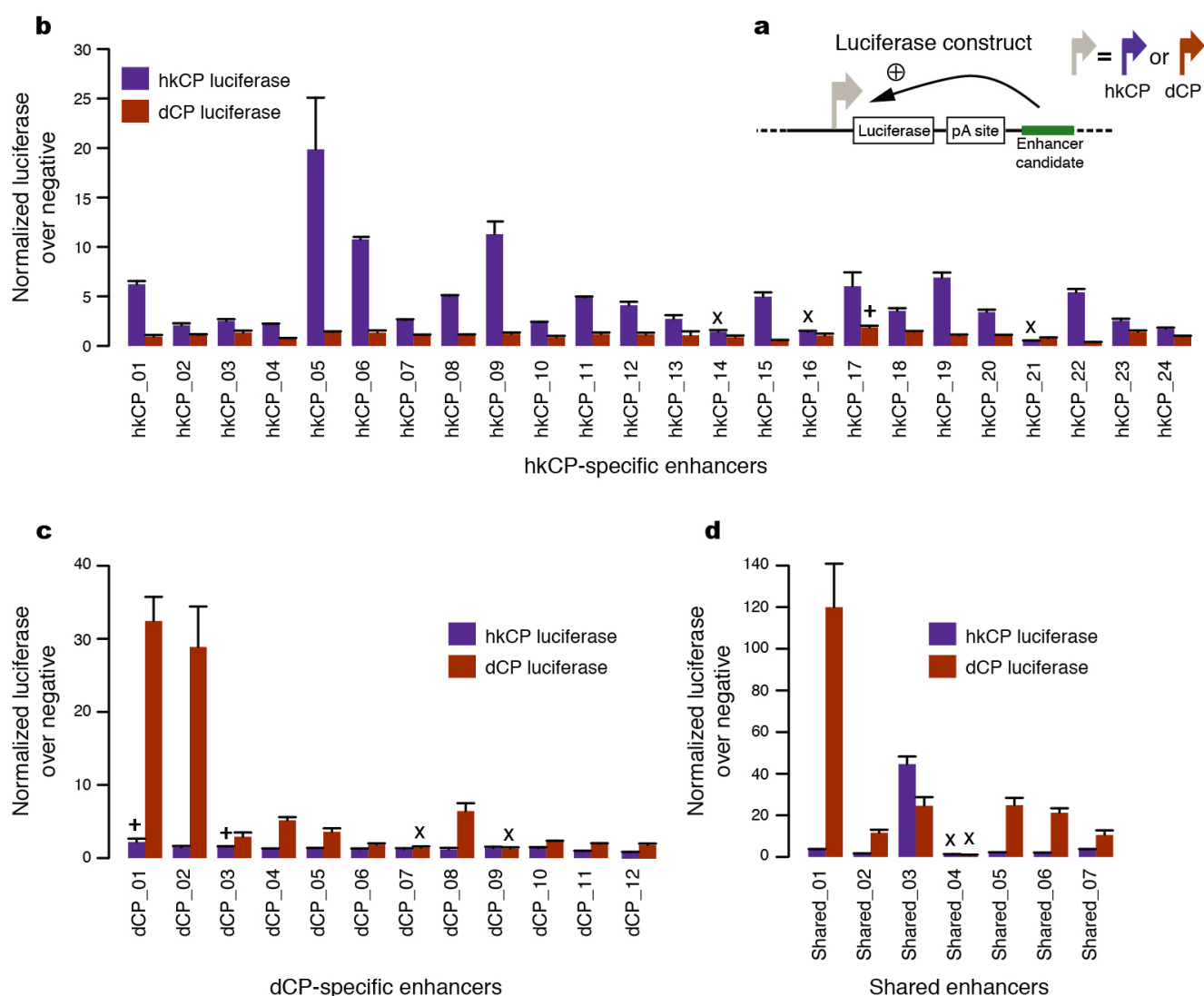
Extended Data Figure 1 | Set-up of STARR-seq with different core promoters. **a**, STARR-seq detects enhancers but no promoters (reproduced with permission from ref. 12). Left, STARR-seq couples the enhancer activities of candidate fragments to the sequences of the candidates in *cis* by placing the candidates to a position within the reporter transcript. Enhancer activities can therefore be assessed by the presence of candidates among cellular messenger RNAs, which allows the parallel assessment of millions of candidates, enabling genome-wide screens. Sequences that activate transcription from the intended core promoter of the STARR-seq vector lead to a full-length reporter transcript and can be detected by STARR-seq. Shown are the reverse transcription (RT) and nested polymerase chain reaction (PCR) steps of the STARR-seq reporter RNA processing protocol that ensure this. Right, in contrast, STARR-seq does not detect truncated transcripts that result if a candidate fragment functions as a promoter to initiate transcription. Thus, core-promoter-containing (that is, TSS-overlapping) sequences that are detected by STARR-seq exhibit enhancer activity as they can activate transcription from a remote position, in addition to their ability to serve as core

promoters endogenously¹². **b**, Luciferase signals (firefly/*Renilla*) assessing the intrinsic (or basal) activity of the core promoters used in this study. The luciferase reporter constructs do not contain any enhancer and differ only in the respective core promoter sequences. The basal activities differ as expected, but do not differ consistently between housekeeping (*RpS12*, *eEF1δ*, *NipB*, *x16*) and developmental (*DSCP*, *eve* (long), *eve* and *pnr*) core promoters, nor between core promoters for which the STARR-seq screens appear most similar (for example, *RpS12* and *eEF1δ*; see Fig. 3). Note that all luciferase assays and STARR-seq screens are corrected for differences in intrinsic activity. **c**, Reproducibility of hkCP and dCP STARR-seq in *D. melanogaster* S2 cells. The reproducibility of hkCP and dCP STARR-seq as assessed by the STARR-seq enrichments (replicate 1 versus 2) at the summits of enhancer peaks called in the merged experiments (hkCP: 5,956; dCP: 5,408). Scatter plots are enlarged versions of the insets in Fig. 1d. “Enr. rep X”, STARR-seq enrichment in replicate X. Note that the raw data for dCP have been re-analysed from ref. 12.

a Schematic overview of 5' RACE on the STARR-seq vectors**b** Agarose gel electrophoresis of 5' RACE nested PCRs (PCR2)**c** Sequence alignment and chromatogram of Sanger sequencing of the PCR product (nested PCR) of 5' RACE of the hkCP STARR-seq vector**d** Sequence alignment and chromatogram of Sanger sequencing of the PCR product (nested PCR) of 5' RACE of the dCP STARR-seq vector

Extended Data Figure 2 | Transcription initiates within the core promoter of the STARR-seq construct. **a–d**, 5' Rapid amplification of cDNA ends (5' RACE) demonstrates that transcription initiates at the TCT and Inr motifs within the hkCP and dCP, respectively. **a**, Set-up of the 5' RACE experiment, including the STARR-seq plasmid, used here with two defined enhancers, the STARR-seq transcript and the location of all primers used to specifically amplify 5'-capped STARR-seq transcripts. **b**, 5' RACE nested PCR products separated on a 1% agarose gel. **c**, Screenshot of Sanger sequencing results

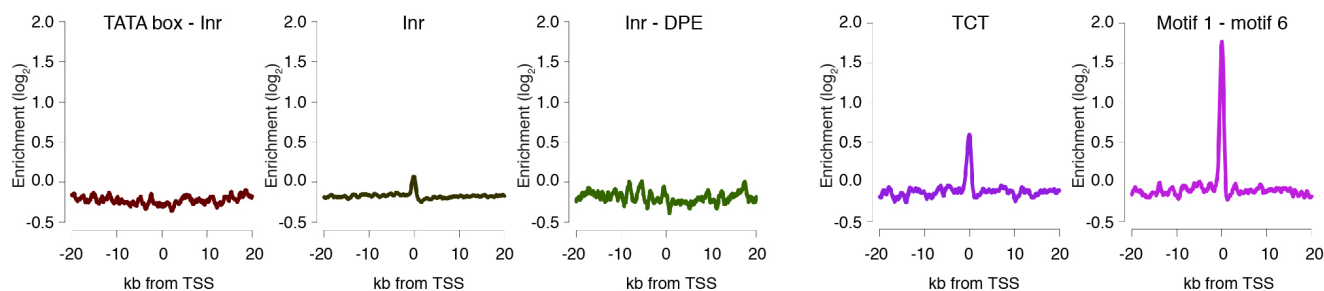
(chromatogram and called bases) compared with the template sequence. Annotations are shown in green, in the following order: 5' RACE adaptor, hkCP with TCT motif (only the part downstream of the TSS is annotated, as the 5' part is not present in the sequenced complementary DNA), spliced intron, green fluorescent protein (GFP); the sequencing primer is shown in red (top). Also shown is a version that displays the template and Sanger sequencing results for the core promoter region only (zoom in). **d**, Same as in **c** but for the dCP for which transcription initiates within the Inr motif.



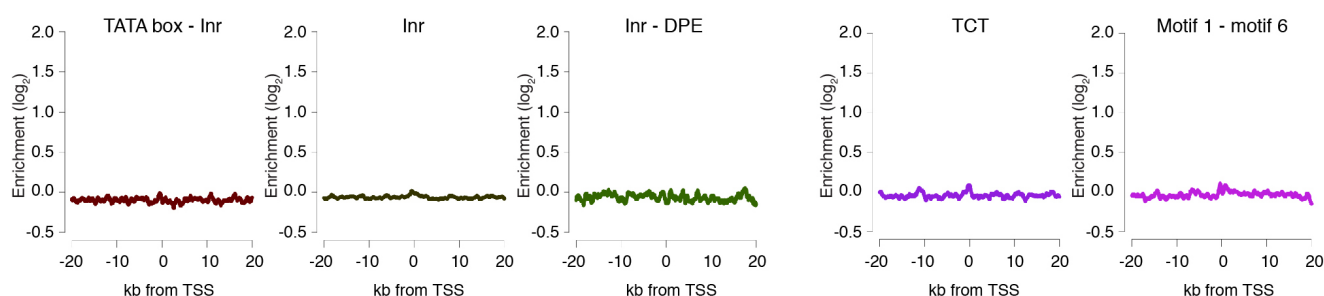
Extended Data Figure 3 | Specificity of hkCP and dCP enhancers to the hkCP and dCP assessed by luciferase assays. **a**, Luciferase reporter set-up with the hkCP or dCP (see also Fig. 1e). **b**, Luciferase signals of 24 hkCP-specific enhancers tested in a hkCP- (purple bars) as well as in a dCP-containing (brown bars) luciferase reporter. Twenty-one out of 24 hkCP enhancers showed luciferase activity (>1.5 fold over negative, $P < 0.05$ via one-sided unpaired Student's t -test, $n = 3$) with the hkCP, while only 1 out of 24 showed activity with the dCP (error bars are s.d. of three biological replicates, 'x' indicates candidates that are not active with the correct core promoter, and '+'

indicates candidates for which the activity with the wrong core promoter is above the threshold (note that the activity with the correct core promoter is still higher in all three cases). **c**, As in **b** but testing dCP-specific enhancers. Ten out of 12 are positive with the dCP whereas only 2 out of 12 are positive with the hkCP. **d**, As in **b** and **c** but testing shared enhancers that were found by STARR-seq with hkCP and dCP; 6 out of 7 are active with both core promoters. See Supplementary Table 17 for the genomic coordinates of the enhancers and the primers used to amplify them.

hkCP S2 STARR-seq signal around core promoter types

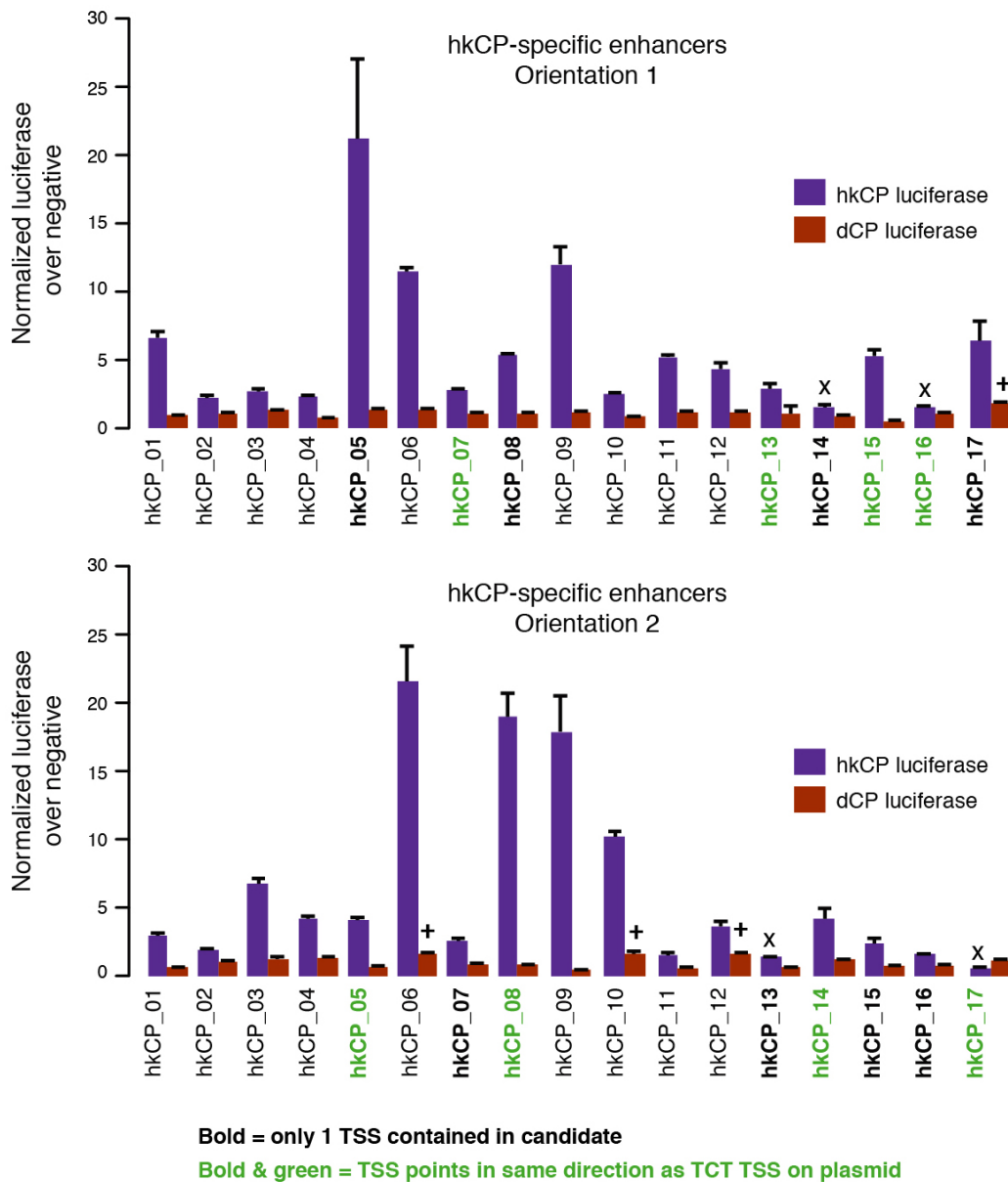


dCP S2 STARR-seq signal around core promoter types



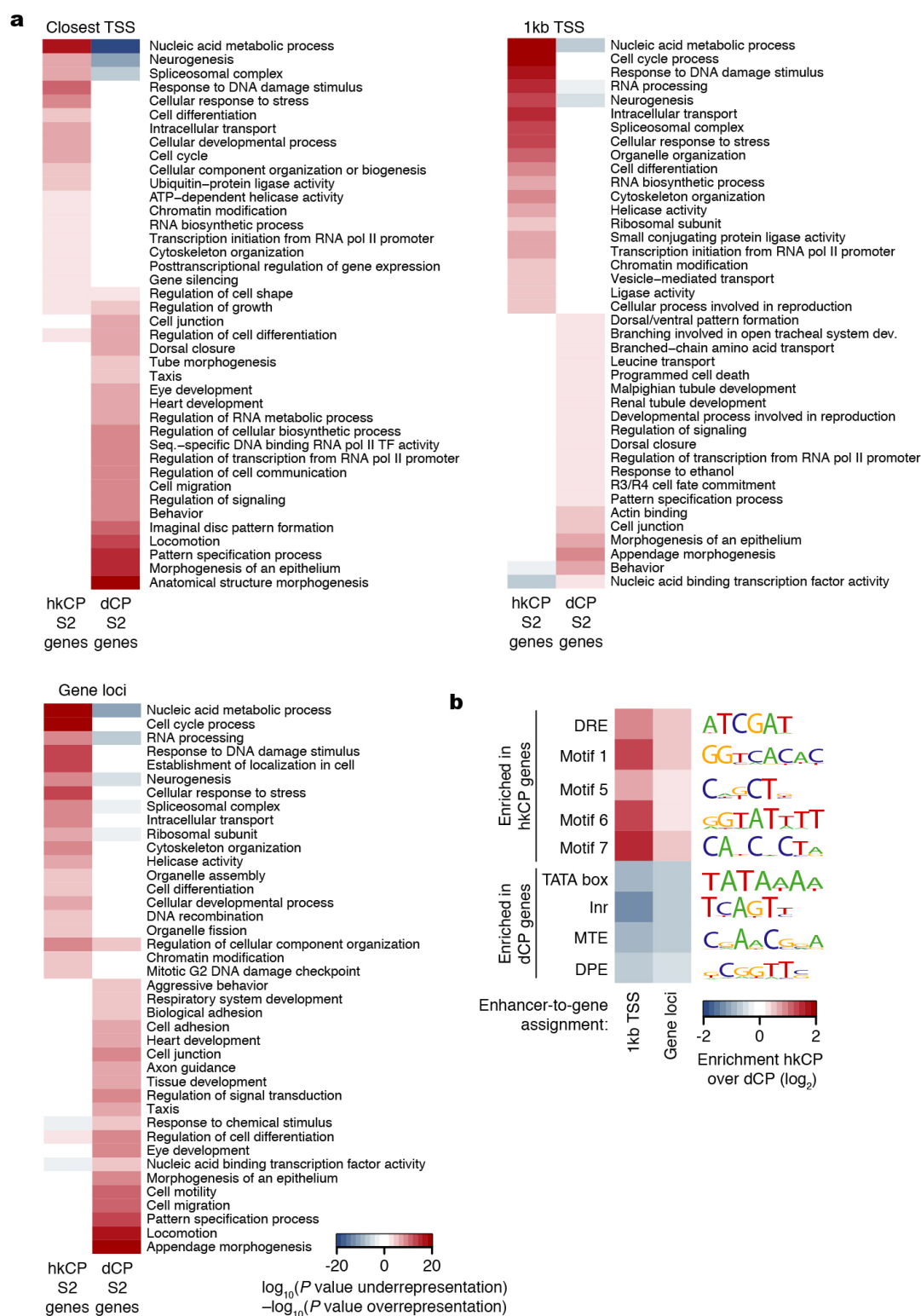
Extended Data Figure 4 | hkCP and dCP STARR-seq signal in S2 cells around different core promoter types. Average hkCP (top) and dCP (bottom) S2 STARR-seq enrichment in 40 kb intervals around TSSs that contain different combinations of known core promoter motifs. Shown are (left to right) TATA box–Inr (179 TSSs), Inr (that do not contain either TATA box or DPE; 1,901), Inr–DPE (100), TCT (303) and motif 1–motif 6 (266). According to

their motif contents, the first three are developmental-type core promoters and the last two are housekeeping-type core promoters. Indeed, only the housekeeping-type core promoters show a strong enrichment of hkCP S2 STARR-seq signals at the TSS, which is not seen for the dCP STARR-seq signal (owing to enhancer–core-promoter specificity) nor for the developmental-type core promoters (owing to the dCP enhancers location at more distal sites).



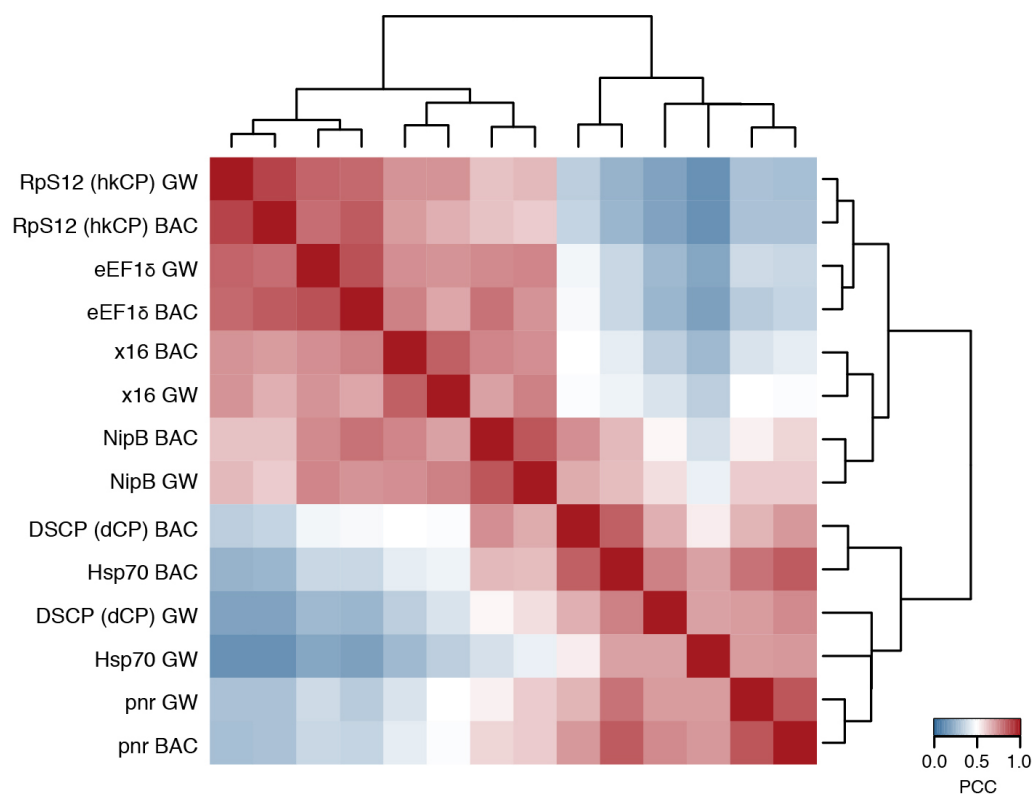
Extended Data Figure 5 | TSS-overlapping hkCP enhancers function independent of their orientation. Luciferase signals for all 17 TSS-overlapping hkCP enhancers (that is, containing one TSS or two divergent TSSs; see Supplementary Table 17) from Extended Data Fig. 3 cloned in the second orientation with respect to the TSS of the luciferase gene (bottom bar plot; the top bar plot corresponds to the initial orientation as in Extended Data Fig. 3 and is shown for comparison). In both orientations, 15 out of 17 enhancers showed activity towards the hkCP (details as in Extended Data Fig. 3). These results together with the findings in Extended Data Fig. 3 challenge the widespread notion that TSS-proximal sequences are promoters and even the concept of promoters more generally: sequences that

autonomously activate gene expression—and are therefore often termed promoters—might in fact be the combination of a core promoter and a proximal enhancer. The TSS-proximal location of many housekeeping enhancers might be evolutionarily more ancient, consistent with regulatory mechanisms in simple eukaryotes such as yeast. In contrast, enhancers of genes with more complex regulation are typically located more distally, potentially simply because the several different cell-type-specific enhancers of these genes would not all fit to positions near TSSs. Consistently, such genes frequently have larger intergenic and intragenic regions⁴⁰ known to accommodate enhancers with diverse activity patterns⁴¹.



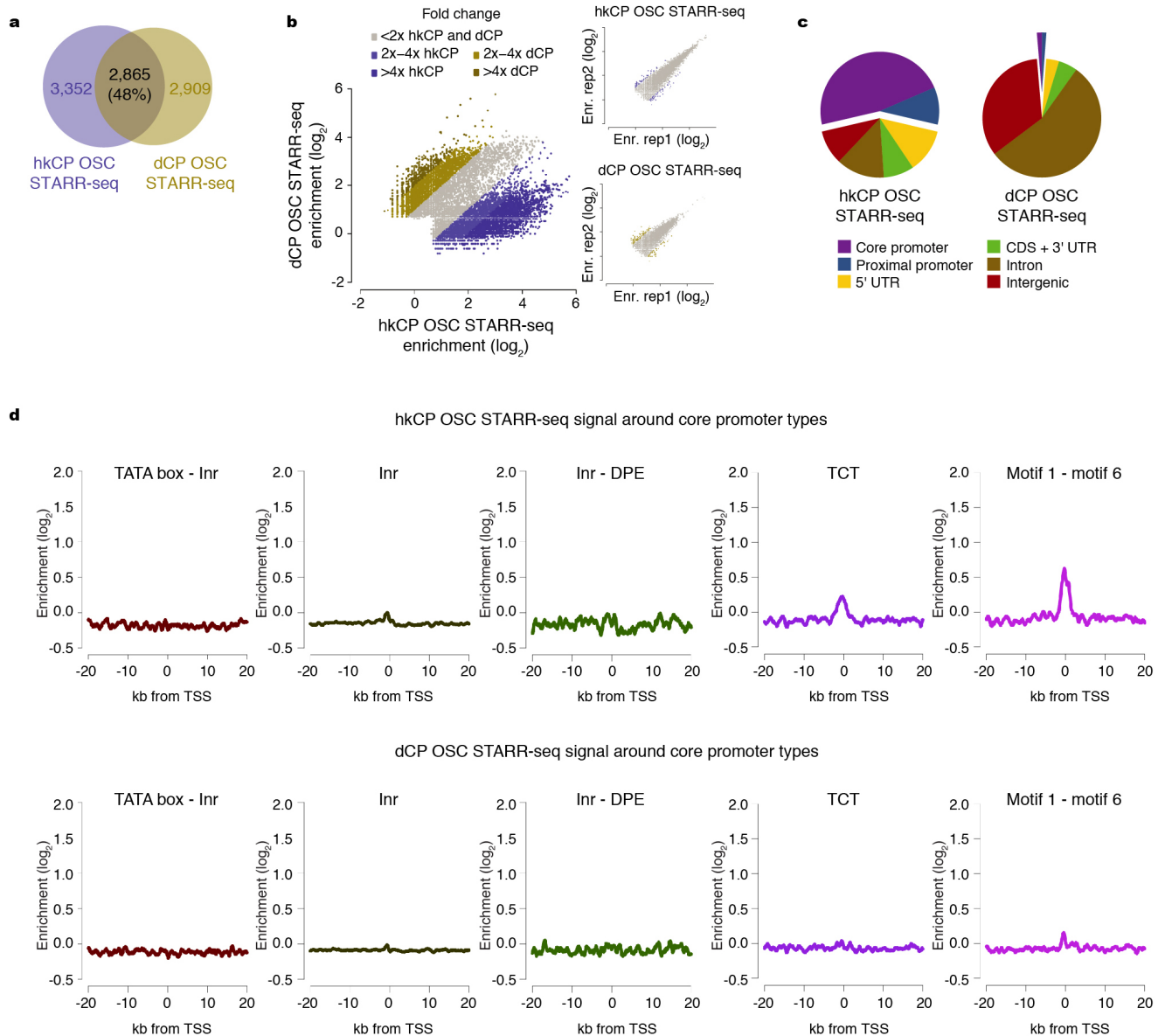
Extended Data Figure 6 | hkCP and dCP enhancers in S2 cells are associated with genes of different functions and core promoter elements. a, GO analysis of genes next to hkCP- and dCP-specific enhancers in S2 cells using different enhancer-to-gene assignment strategies (top left, 'closest TSS' as in Fig. 2; top right, '1 kb TSS'; bottom left, 'gene loci'; see Methods for details). Shown are 20 non-redundant GO categories selected from the 100 most significantly enriched categories associated with each enhancer class (see Supplementary

Tables 2–4 for all categories). **b,** Enrichment of core promoter elements at genes next to hkCP- and dCP-specific enhancers in S2 cells. Similar analysis as in Fig. 2e, but using different enhancer-to-gene assignment strategies (see Methods for details). Consistent with Fig. 2e, core promoters of genes assigned to hkCP-specific enhancers are enriched in motifs 1, 5, 6, 7 and DRE, while core promoters of genes assigned to dCP-specific enhancers are enriched for TATA box, Inr, MTE and DPE motifs, irrespective of the assignment strategy.



Extended Data Figure 7 | Housekeeping and developmental core promoters differ characteristically in their global enhancer preferences. As in Fig. 3b but including biological replicates with independently cloned focused bacterial artificial chromosome (BAC) libraries covering around 5 Mb of genomic sequence (BAC) and assessing the PCC at each position along these regions. GW, genome-wide screens as in Fig. 3b. The similarity observed for the TATA

box- and DPE-containing core promoters (*Hsp70*, *pnr* and DSCP (dCP)) suggest that differences related to these core promoter elements might be more subtle or related to alternative mechanisms, including the potential preferences of more proximal or distal enhancers⁴² or RNA polymerase II pausing and the dynamics versus stochasticity of initiation and elongation^{43,44,45}.

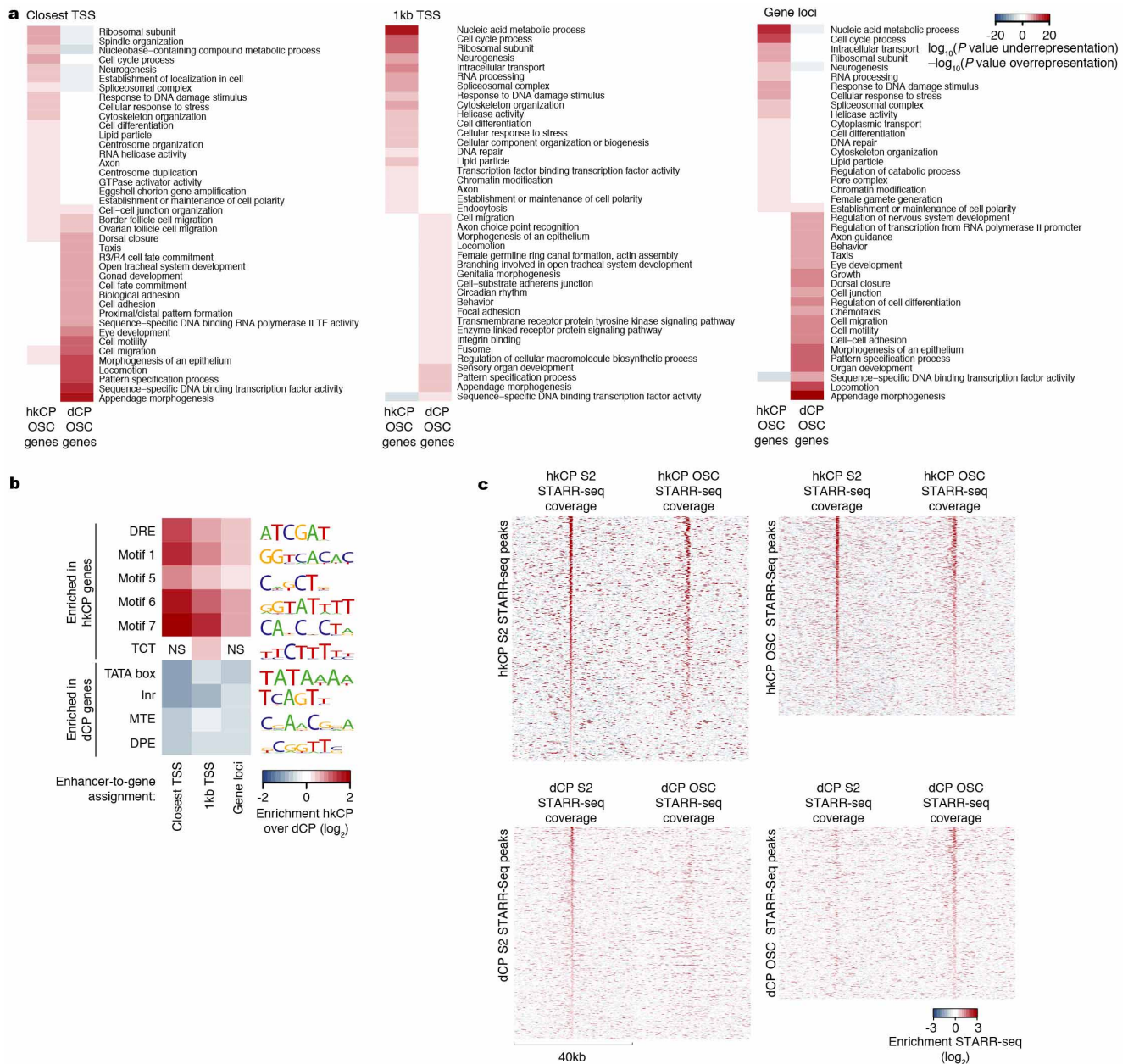


Extended Data Figure 8 | hkCP and dCP enhancers differ in OSCs.

a, b, Different enhancers activate transcription from hkCP and dCP in OSCs. As Fig. 1c, d but for OSCs rather than S2 cells (data in bottom inset of **b** are re-analysed from ref. 12). **c**, Genomic distribution of hkCP and dCP

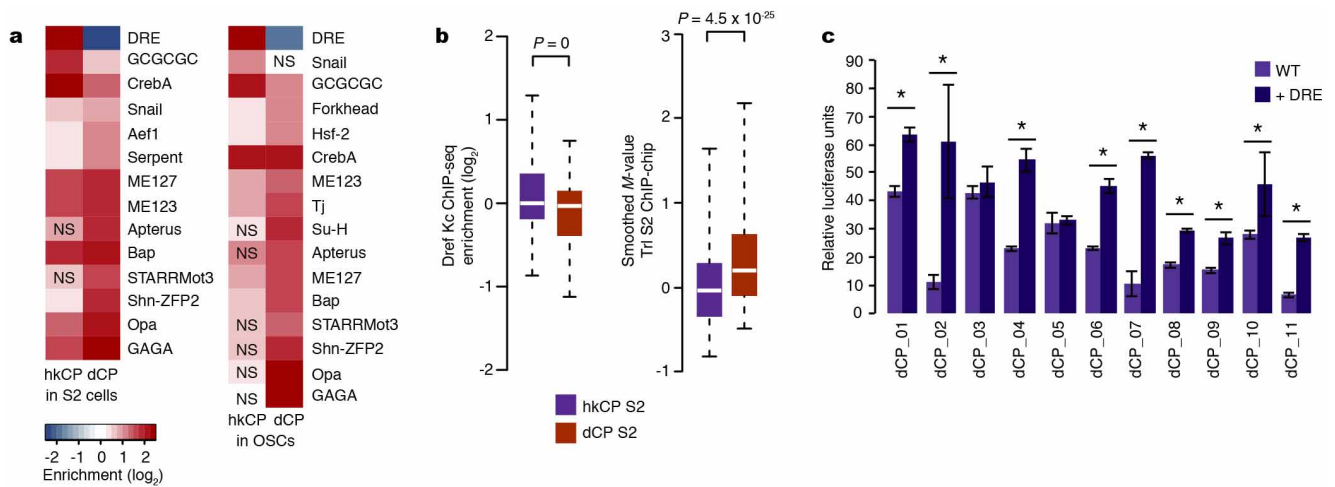
enhancers in OSCs. As Fig. 2a but for OSCs rather than S2 cells. **d**, hkCP and dCP STARR-seq signal in OSCs around different core promoter types. As Extended Data Fig. 4 but for OSCs rather than S2 cells.

RESEARCH LETTER



Extended Data Figure 9 | Differences between hkCP and dCP enhancers in OSCs. **a**, GO analysis of genes next to hkCP- and dCP-specific enhancers in OSCs. As Extended Data Fig. 6a but for OSCs rather than S2 cells (see Supplementary Tables 8–10 for all categories). **b**, Enrichment of core promoter elements at genes next to hkCP- and dCP-specific enhancers in OSCs. As Fig. 2e

and Extended Data Fig. 6b but for OSCs rather than S2 cells. NS, not significant (hypergeometric $P > 0.05$). **c**, Heat maps of hkCP (top) and dCP (bottom) STARR-seq enrichments in S2 cells and OSCs. Heat maps on the left and right are centred on the summits of core-promoter-type-specific enhancers in S2 and OSCs, respectively.



Extended Data Figure 10 | The activities of hkCP and dCP enhancers are dependent on DRE and GAGA motifs, respectively. **a**, Differential motif enrichment in distally located hkCP- and dCP-specific enhancers (as in Fig. 5a but assessing enrichments of the same motif PWMs exclusively at distal enhancers >500 bp away from the closest TSSs). Key motifs including DRE and GAGA are also differentially enriched in distal hkCP- and dCP-specific enhancers. NS, not significant (FDR-corrected hypergeometric $P > 0.01$). S2 cells: hkCP $n = 790$, dCP $n = 3,013$; OSCs: hkCP $n = 556$, dCP $n = 2,555$. **b**, Distal hkCP- and dCP-specific enhancers are differentially bound by Dref and Trl, respectively. ChIP enrichments of Dref (left) and Trl (right) at S2 hkCP- and dCP-specific enhancers that are distal (>500 bp) from the closest TSSs. Equivalent to Fig. 5b, but considering exclusively TSS-distal enhancers to

exclude potentially confounding effects for TSS-proximal enhancers for which it is not possible to discern whether binding occurs due to the enhancer sequence or core promoter function. The differential binding between Dref and Trl to hkCP- and dCP-specific enhancers, respectively, is also found in Kc167 cells, in which the Dref ChIP-seq experiment had been performed (data not shown). **c**, Addition of DRE motifs to dCP enhancers increases their activity towards hkCP. Relative luciferase activity values (firefly/*Renilla*) for 11 dCP enhancers without DRE motifs (wild type (WT), light purple) and with 3 DRE motifs flanking the enhancers on each side (+DRE, dark purple). * $P < 0.05$, one-sided unpaired Student's *t*-test; error bars denote the s.d. of three biological replicates.

Paper #2: Regulatory Enhancer–Core-Promoter Communication via Transcription Factors and Cofactors

Muhammad A. Zabidi & Alexander Stark, *Trends in Genetics*, 2016. 32(12) pp. 801-814.

The two principal *cis*-regulatory DNA elements in an animal genome are enhancers and core promoters. Enhancers can lie at arbitrary distances from their cognate core promoters. It is not clear how enhancers target the correct core promoters in the genome. Motivated by this intriguing picture, a principal question in gene regulation is how enhancers activate their core promoters.

In this review, we summarize the current knowledge on enhancer–core-promoter communication. First, genomic regions are physically organized into topological domains that restrict enhancer functions. Alternatively, tethering elements and DNA accessibility may also specify enhancer–core-promoter communication.

The specificity between enhancers and core promoters has been observed in several isolated cases and in our genome-wide study. The specificities are conferred by the sequence that recruits distinct factors. Other factors also exhibit differential occupancy and activity towards core promoters, supporting that the enhancer–core-promoter specificity is mediated via biochemical compatibility of the factors, many of which also have enzymatic activities. These factors assemble at high concentration in regulatory microenvironments, where post-translational modifications (PTMs) may act as conduits to transfer the signaling information.

We end the review with perspectives for further development in the field. Other sequence-mediated specificities may also be employed in other transcription programs. We also propose the reannotation of regulatory genomic regions based on their functions, instead of the proximity of the regions.

Author contributions

M.A.Z. and A.S. designed and created the figures and wrote the manuscript.

Review

Regulatory Enhancer–Core-Promoter Communication via Transcription Factors and Cofactors

Muhammad A. Zabidi¹ and Alexander Stark^{1,*}

Gene expression is regulated by genomic enhancers that recruit transcription factors and cofactors to activate transcription from target core promoters. Over the past years, thousands of enhancers and core promoters in animal genomes have been annotated, and we have learned much about the domain structure in which regulatory genomes are organized in animals. Enhancer–core-promoter targeting occurs at several levels, including regulatory domains, DNA accessibility, and sequence-encoded core-promoter specificities that are likely mediated by different regulatory proteins. We review here current knowledge about enhancer–core-promoter targeting, regulatory communication between enhancers and core promoters, and the protein factors involved. We conclude with an outlook on open questions that we find particularly interesting and that will likely lead to additional insights in the upcoming years.

Trends

The specificity of enhancers for core promoters is encoded in the sequence.

Motifs within enhancer and core-promoter sequences recruit *trans*-acting factors that mediate regulatory enhancer–core-promoter communication and specificity.

The communication is mediated by a high local concentration of cofactors that interact dynamically with, and possibly post-transcriptionally modify, each other and the RNA polymerase.

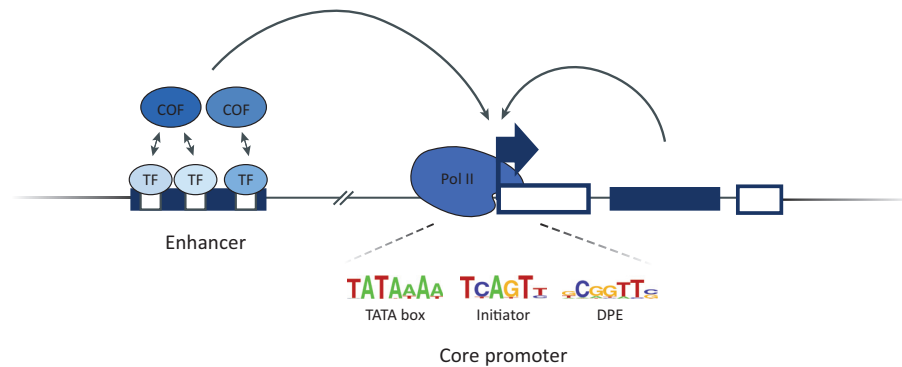
Gene expression and its spatiotemporal regulation are central to the development and adult physiology of all multicellular organisms. It enables the formation of distinct cell types with specialized morphologies and functions by allowing different genes to be activated in a cell type-specific manner. Transcription constitutes the first and one of the most intensely regulated steps of gene expression. Indeed, the RNA content of different cell types differs greatly and correlates well with protein abundance for many genes [1]. Furthermore, misregulation at the level of transcription underlies several developmental disorders and diseases such as cancer [2,3]. Transcription initiates within **core promoters** (see [Glossary](#)), short sequences of ~100 bp surrounding the transcription start-sites (TSSs) at the 5' ends of genes. Core promoters recruit RNA polymerase II (Pol II), assemble the **pre-initiation complex** (PIC), and dictate the accurate position of initiation and the direction of transcription [4]. Typically, however, core promoters on their own cannot support efficient transcription and exhibit only low basal activities. Instead, their cell type-specific activities are typically determined by **enhancers**, the second key type of transcriptional regulatory elements [5,6].

Enhancers are genomic DNA elements of up to several hundred bp in length which contain short **transcription factor (TF) recognition sequences or binding sites**. Through these sites, combinations of TFs are recruited to enhancers and in turn recruit cofactors with a variety of biochemical functions ([Figure 1](#), Key Figure). Through the combined activating or repressive cues of the different TFs and **cofactors**, enhancers exert their overall regulatory function to control transcription from target core promoters irrespective of their orientation and distance [7]. Because enhancers can act over short and long distances, in other words their positions with

¹Research Institute of Molecular Pathology (IMP), Vienna Biocenter (VBC), Dr. Bohr-Gasse 7, 1030 Vienna, Austria

*Correspondence: stark@starklab.org (A. Stark).

Key Figure

Enhancers and Core Promoters, Two Major Classes of *Cis*-Regulatory Elements

Trends in Genetics

Figure 1. An enhancer contains binding sites for sequence-specific transcription factors (TFs). These in turn recruit cofactors (COFs) that typically mediate the regulatory communication between the core promoter and the enhancer, in other words relay the regulatory cues from enhancers to target core promoters. Core promoters encompass short sequences of ~100 bp surrounding the transcription start site (TSS), where RNA polymerase II (Pol II) assembles and initiates transcription. Core promoters typically contain characteristic core-promoter elements or motifs, for example the TATA box, Initiator, or downstream promoter element (DPE).

respect to their target core promoters can be arbitrary, they do not always regulate the nearest gene.

An important question in biology is how activating regulatory cues are communicated from enhancers to their correct target core promoters. We discuss different modes of specifying enhancer–core-promoter communication, starting with the organization of the regulatory genome into chromatin domains, followed by the specification of enhancer–core-promoter contacts by DNA accessibility and enhancer–core-promoter tethering, and finally how enhancer and core-promoter sequences recruit different regulatory proteins that mediate the regulatory communication and specificity.

Core Promoter Targeting within Topologically Associating Domains (TADs)

Cis-regulatory genomic regions in animals are organized locally into domains that can be several kb or Mb in length (Figure 2A). These domains are typically delineated by insulator proteins such as CTCF (reviewed in [8–10]) or by broadly expressed housekeeping genes [11–13]. Within these **topologically associating domains** (TADs) [12,14,15], chromatin contacts are more frequent than elsewhere, as measured by chromosome conformation capture (3C) and variant techniques [16–22], or by fluorescence *in situ* hybridization (FISH) [23].

Ample evidence supports a role of TADs in restricting or directing enhancer function during transcriptional regulation: for example, TAD boundaries and insulator binding are depleted between enhancers and their target core promoters [24], and TAD boundaries curb the spreading of chromatin marks associated with transcriptional activity [12,25–27]. Moreover, while enhancers function generally independently of their orientation, enhancers with proximal CTCF binding sites can show directional activities *in vivo* that can be inverted by inverting the

Glossary

Cofactors: regulatory protein factors that are typically unable to bind to DNA themselves and are recruited to enhancers by transcription factors (TFs). Cofactors can have enzymatic functions, for example catalyze post-translational modifications (PTMs) of proteins, and mediate the regulatory function of the enhancers.

Core promoter: a short sequence around the transcription start-site (TSS) that can direct the recruitment of RNA polymerase II (Pol II) and transcription initiation. Core promoters typically have low basal activities in the absence of enhancers and are also termed minimal promoters.

Enhancer: a sequence that boosts transcription from a target core promoter in a cell type-specific manner and independently of the relative orientation and distance of the enhancer. Known enhancers are several tens to several hundred bp in length.

Insulator: a DNA element that can block regulatory enhancer–core-promoter communication, typically by the recruitment of insulator proteins such as CTCF. Insulators can thus demarcate the range of enhancer activity and define topologically associating domains (TAD) borders.

Pre-initiation complex (PIC): an assembly of proteins, including Pol II and general transcription factors, that first nucleates at the core promoter before transcription initiation, and renders Pol II transcription-competent.

Promoter: sequence up to several kb upstream of TSSs that can autonomously drive transcription. This functionality and the fact that some promoters can activate transcription from a distal core promoter in reporter assays is in line with promoters consisting of a core promoter and a proximal or overlapping enhancer.

Promoter-proximal tethering element (PTE): a sequence proximal to core promoters that enables or facilitates its interaction with enhancers.

Promoter-targeting sequence: a sequence proximal to enhancers that enables or facilitates their interactions with core promoters.

Self-transcribing active regulatory region sequencing (STARR-seq): enhancer-activity assay in which

DNA fragment containing the enhancer and CTCF binding site [28]. Within TADs, individual enhancers are able to activate reporter genes irrespective of the positions into which the reporters were integrated, suggesting that enhancer–core-promoter communication within TADs is not restricted to specific positions [29,30].

Furthermore, disruptions of TAD boundaries lead to gene deregulation, manifesting in conditions such as polydactylies in human patients and in mouse models [31]. In addition, chromosomal rearrangements [32] or reduced CTCF binding because of hypermethylation of CTCF binding sites [33] can impair boundary function and have been implicated in cancer. Such alterations create new enhancer–core-promoter interactions, leading to gene misexpression and increased cancer cell oncogenicity (reviewed in [34]), and, indeed, mutations of CTCF sites are enriched in cancer-associated SNPs [35]. Together, these observations suggest that TADs act as gene regulatory units.

Core-Promoter Selection via DNA Accessibility and Tethering

While genes within TADs are indeed often coordinately regulated [36], many are not coexpressed, and enhancers and their target core promoters are not necessarily adjacent and colinear in the genomic sequence. For example, during *Drosophila* embryogenesis, the neighboring genes *Sex combs reduced* (*Scr*) and *fushi tarazu* (*ftz*) are expressed in different patterns and their enhancers are not colinear (Figure 2B): the enhancer of *Scr* is located 3' of the *ftz* gene and the enhancer of *ftz* lies between the two genes. The selectivity of the *Scr* enhancer for the *Scr* promoter can be recapitulated in reporter assays and depends on a **promoter-proximal tethering element** (PTE) [37] that might mediate specific enhancer–core-promoter spatial proximities or contacts. Interestingly, another PTE identified at the *Bithorax* complex locus even enables enhancer communication across **insulators** [38,39]. Equivalent sequences have also been described within or next to enhancers, and these mediate the contacts of enhancers to their target promoter, thus termed **promoter-targeting sequences** [40]. Such enhancer–core-promoter tethering would be compatible with the observation of looping and stable enhancer–polymerase contacts as observed during development [41]. In experimental setups with forced enhancer–core-promoter proximity, tethering approaches have also been shown to activate transcription in defined systems [42–44].

Alternatively, the promoter-proximal sequences might, in a cell type-specific manner, regulate the availability of a core promoter by modulating its DNA accessibility within chromatin (Figure 2C). Transcriptional inactivity due to promoter inaccessibility is also found more generally in *Drosophila* when enhancers skip neighboring genes to activate more distal core promoters [24], and might also explain the different expression patterns of the divergently transcribed homeobox genes *gooseberry* (*gsb*) and *gooseberry-neuro* (*gsb-n*) [45]: while divergently transcribed homologous genes are often coexpressed, *gsb* and *gsb-n* seem to be differentially active at different embryonic stages, potentially reflecting differential accessibilities of the core-promoter sequences. The regulation of DNA accessibility is also important in other species to regulate core-promoter activity and gene expression (e.g., in worm [46,47]) or the transcriptional activity of entire gene loci [48–50], and is likely involved in controlling enhancer–core-promoter targeting more generally.

Sequence-Encoded Enhancer–Core-Promoter Specificity

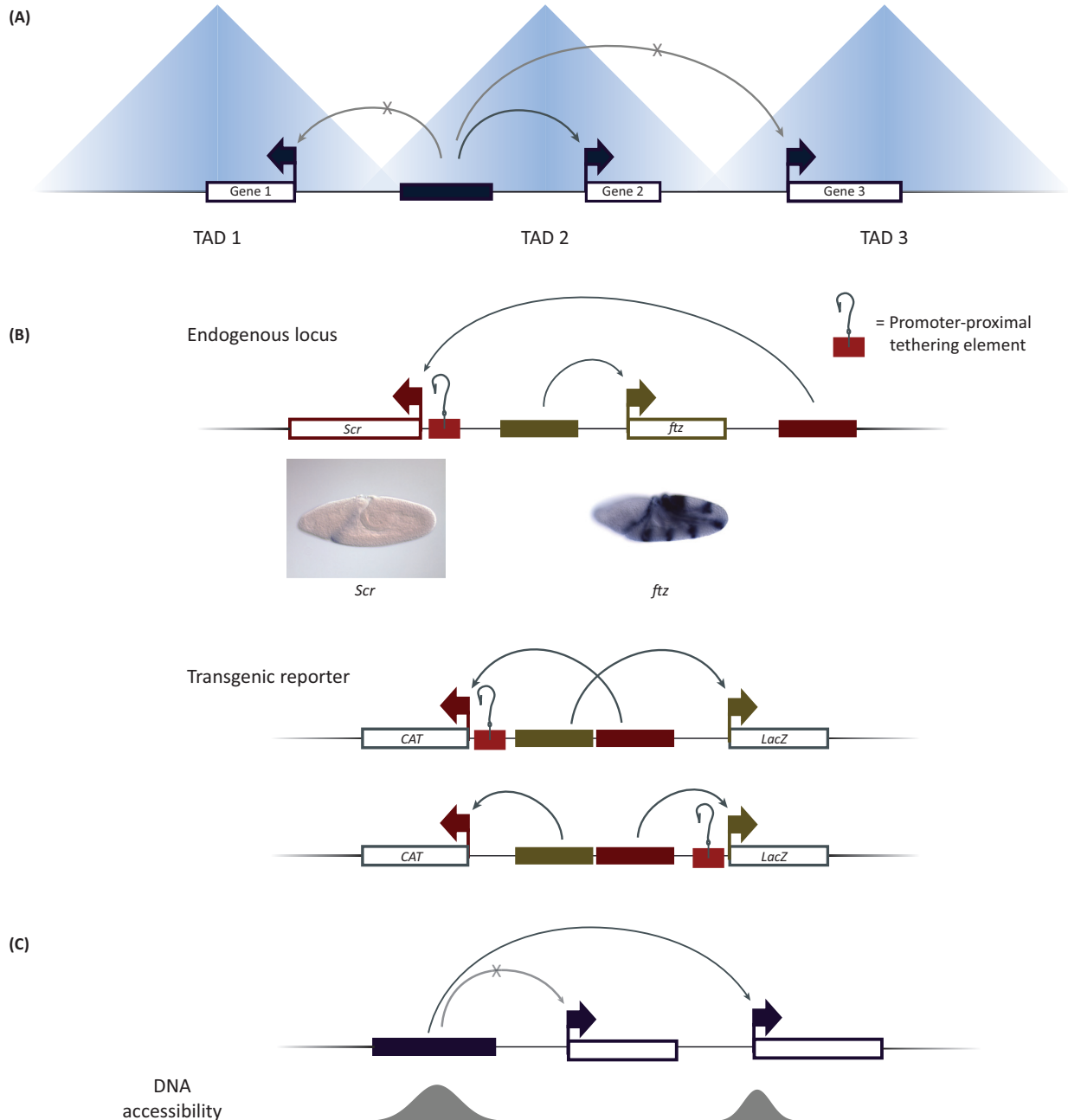
DNA accessibility, however, does not always explain why neighboring genes are differentially expressed. In *Drosophila*, neighboring genes *out at first* (*oaf*) and *decapentaplegic* (*dpp*) are expressed in different patterns at the same stage, despite the *oaf* core promoter being more proximal to the *dpp* enhancer. Furthermore, if the *oaf* core promoter is replaced by the *hsp70* core promoter, *oaf* is activated in the *dpp* expression pattern, arguing that the respective core-promoter sequences are important determinants of enhancer targeting [51]. In addition, during

candidate DNA fragments are positioned in the 3'-untranslated region of a reporter gene such that active enhancers transcribe themselves and the activity of the sequence can be assessed by measuring the abundance of their transcripts among cellular RNAs [61]. The coupling of candidate sequences to enhancer activity in *cis*, such that each enhancer serves as its own barcode, allows the parallel assessment of millions of DNA fragments from arbitrary sources and enables genome-wide functional enhancer screens.

TF recognition sequences or binding sites (vs motifs): short DNA sequences that bind to TFs. Consensus sequences that summarize the binding preferences of a TF are known as TF motifs.

Transcription factors (TFs): proteins that bind to DNA sequences in enhancers through their DNA-binding domains and that activate or repress transcription, usually via the recruitment of cofactors.

Topologically associating domains (TADs): large genomic regions, typically several kb or Mb in length, within which frequent chromatin contacts occur, as measured by chromosome conformation capture (3C) and variant techniques. As discussed in the main text, the importance of TADs for transcriptional regulation is increasingly being recognized.



Trends in Genetics

Figure 2. Topologically Associating Domains (TADs), Promoter-Proximal Tethering Elements (PTEs), and DNA Accessibility. (A) Enhancer function is typically restricted to activate core promoters within the same TAD [12,31], the boundaries of which are enriched in insulator protein binding. (B) PTEs are sequences proximal to core promoters that promote preferential interaction between enhancers and core promoters [37–39]. For example, the PTE of the *Sex combs reduced* (*Scr*) core promoter enables its activation by the distally located *Scr* enhancer, skipping the intervening gene *fushi tarazu* (*ftz*). In a transgenic reporter, relocation of the PTE to a proximal position of the *ftz* core promoter results in activation of the latter by the *Scr* enhancer. *In situ* staining of embryo images are from the Berkeley *Drosophila* Genome Project [161]. (C) Inaccessible DNA can prevent the activation of a core promoter by a nearby enhancer. For example, *Drosophila* enhancers skip the proximal and inaccessible core promoters to activate more-distal and accessible core promoters [24]. Abbreviations: CAT, chloramphenicol acetyltransferase; LacZ, gene of β -galactosidase.

maternal and early zygotic transcription many genes initiate at different TSS that can be very closely spaced but which are located within AT- versus CG-rich core promoters, suggesting that core-promoter sequences are involved [52]. This is further supported by the observation that reporter genes under the control of different enhancer–core-promoter combinations can also exhibit distinct expression patterns [53–55], even when integrated at identical genomic positions [56]. In addition, the regulation of several *Hox* genes including *ftz* by the *caudal* (*cad*) TF seems to depend on specific sequence elements within the core promoters, particularly the downstream promoter element (DPE) [57].

Apart from the DPE, other defined core-promoter motifs or elements also exist and are differentially distributed in core promoters of genes with different functions. For example, core promoters of developmentally regulated genes tend to contain TATA box, Initiator (Inr), and DPE motifs, whereas core promoters of housekeeping genes contain motifs such as DNA replicating elements (DREs), and Ohler motifs 1, 6, 7, and 8 (reviewed in [58,59]). This sharp dichotomy strongly indicates the involvement of core-promoter sequences in enhancer specificity.

Given the differential core-promoter element distribution between the developmental and housekeeping gene regulatory programs, the activity of millions of *Drosophila* enhancer candidates towards several housekeeping and developmental core promoters [60] were recently tested using **self-transcribing active regulatory region sequencing** (STARR-seq) [61]. This revealed thousands of enhancers with strong preference for either one of the two core-promoter classes (Figure 3A). In the defined reporter set-up the core promoters were the only variable: proximity, DNA accessibility, insulators, or other DNA elements cannot explain the marked enhancer–core-promoter specificity, suggesting that core-promoter sequences fall into different functional classes that are activated by distinct types of enhancers.

Differential Occupancy of *Trans* Factors at Housekeeping versus Developmental Enhancers

The identification of thousands of enhancers with preference for either one of two core-promoter classes enabled comparisons of the DNA sequences to identify sequence features that underlie this specificity. As expected, developmental enhancers were enriched for motifs of cell type-specific TFs such as Serpent (Srp), Traffic jam (Tj), and Chorion factor 2 (Cf2), and also for dinucleotide repeats, particularly GA repeats that are bound by Trithorax-like (Trl) [62]. By contrast, housekeeping enhancers were enriched for DREs, recognized by the DNA replication-related element factor (DREF). DREs were necessary and sufficient for housekeeping enhancer activity and allowed reprogramming of developmental into housekeeping enhancers. Consistently, Trl and DREF were also differentially bound to both types of enhancers. In addition, depletion or specific inhibition of Trl or DREF results in different gene expression changes [63–67], providing further evidence that the two factors play different roles in gene regulation. The distinctive distribution of TF motifs and the differential binding of the corresponding TFs suggest that the core-promoter specificity is encoded in enhancer sequences and is mediated by *trans*-activating factors [60] (Figure 3B).

Differential Activity of *Trans* Factors at Housekeeping versus Developmental Core Promoters

While sequence motifs and TF occupancy are differentially distributed between housekeeping versus developmental enhancers, these observations do not shed light on how the TFs exert their functions and whether they themselves have intrinsic core-promoter preferences as well. The regulatory activity of different TFs or cofactors on core promoters can be assessed by directly tethering the respective factors via heterologous DNA-binding domains (DBDs) to positions upstream of reporter gene core promoters in activator bypass experiments [42–44,68]. Following this logic, 812 *Drosophila* TFs and cofactors were recruited to a housekeeping (hkCP) and a

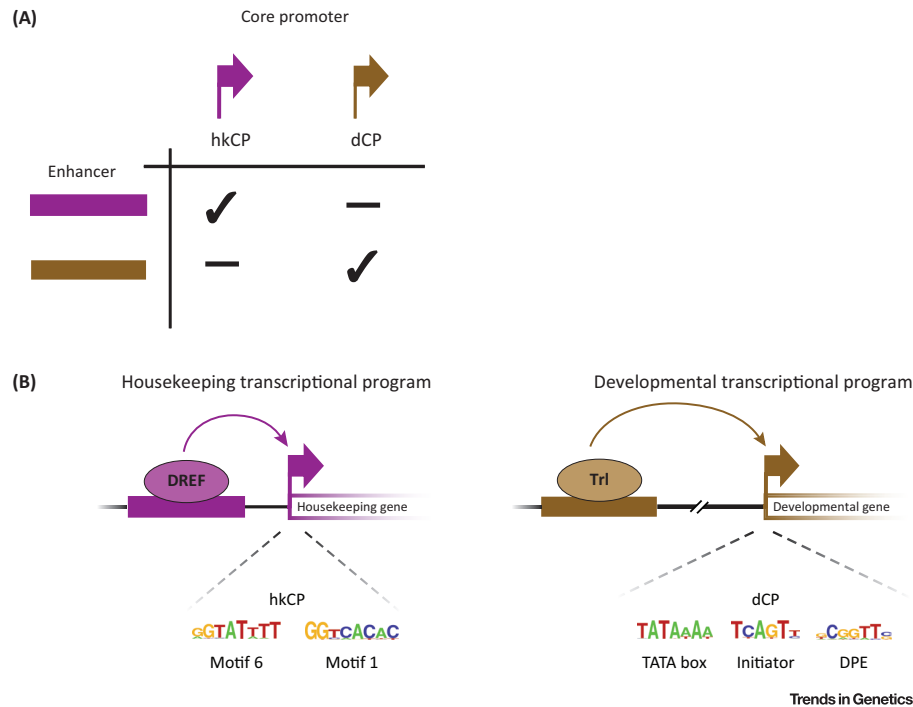
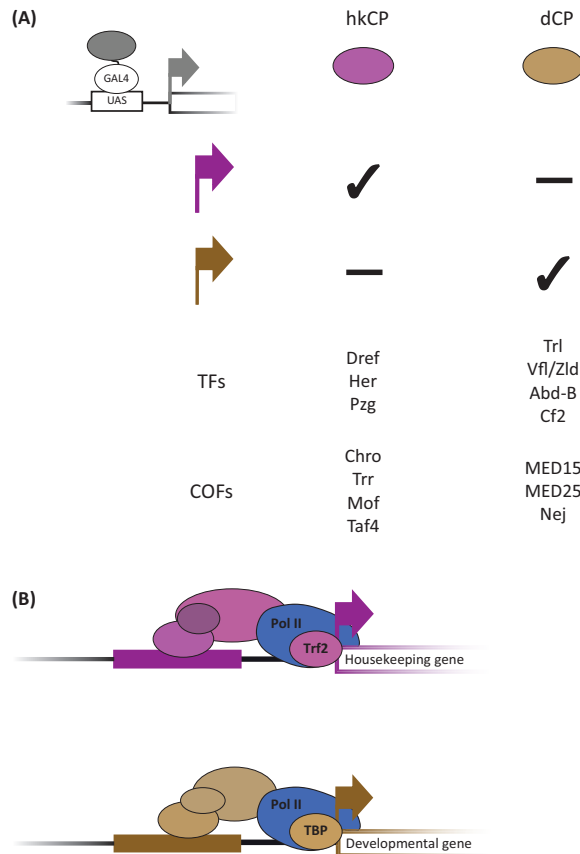


Figure 3. Sequence-Mediated Enhancer–Core-Promoter Specificity. (A) Housekeeping core promoters (hkCPs; purple) show preferences for housekeeping over developmental enhancers. The reverse is true for developmental core promoters (dCPs; ochre) [60]. (B) Housekeeping enhancers bind to DREF and activate transcription from hkCPs, which typically contain Ohler motif 1 and 6 core-promoter elements (indicated by motif logos). Developmental enhancers meanwhile bind Trl and activate transcription from dCPs that contain different core-promoter elements, such as the TATA box, Initiator, and DPE. Compared to housekeeping enhancers which are TSS-proximal or even overlapping, developmental enhancers are found at various positions, including within introns or very distal.

developmental (dCP) core promoter [69]. Consistent with the differential TF binding to housekeeping and developmental enhancers [60], recruitment of DREF and Trl recapitulated the differential activation of hkCP (by DREF) and dCP (by Trl; Figure 4A).

Interestingly, many additional TFs show differential activities towards the two core promoters (Figure 4A). For example, Putzig (Pzg) preferentially activates hkCPs and is indeed important for housekeeping gene expression [70,71]. On the other hand, TFs that preferentially activate dCPs represent factors important during fly development, including the *Hox* TF Abdominal-B (Abd-B), the early zygotic activator Zelda (Zld), and the developmental TFs Cf2 and Pointed (Pnt).

Transcriptional cofactors similarly exhibit core-promoter preferences: dCPs are strongly activated by the Mediator subunits MED15 and MED25 as well as the *Drosophila* CBP/p300 ortholog Nejire (Nej), while hkCPs are strongly activated by Chromator, Males absent on the first (Mof), TBP-associated factor 4 (Taf4), and Trithorax-related (Trr). The latter type of factors indeed play roles in cell maintenance: for example, Mof is the acetyltransferase component of the Male-specific lethal (MSL) [72–75] and Non-specific lethal (NSL) [76,77] complexes that control transcriptional dosage compensation for male X chromosome and housekeeping genes, respectively, while Chromator is important in maintaining spindle dynamics during mitosis [78] and regulates chromatin structure [79]. Unsurprisingly, there is much evidence that underscores the importance of cofactor recruitment via sequence-specific TFs during transcriptional regulation in different physiological processes. During dauer formation in worms, for example,



Trends in Genetics

Figure 4. *Trans* Factors Mediate Sequence-Directed Enhancer–Core–Promoter Specificity. (A) Activator-bypass experiments show that different transcription factors (TFs) and cofactors (COFs) can differentially activate housekeeping core promoters (hkCPs) over developmental core promoters (dCPs), or vice versa. Activation is indicated by a check-mark, and examples of TFs and cofactors that function accordingly are shown below. (B) hkCPs and dCPs recruit TRF2- and TBP-containing complexes, respectively. The members of these complexes also potentially differ. Abbreviation: Pol II, RNA polymerase II.

the DAF-16 TF recruits the SWI/SNF complex to activate longevity and stress-resistance target genes [80], while the sterol regulatory element binding protein (SREBP) TF recruits p300 and MED15 during lipid homeostasis [81]. Other examples also exist in different systems [82–84]. Interestingly, cofactors can also provide feedback to the DNA-binding activity of TFs [85]. Finally, mutation or overexpression of cofactors deregulates communication between sequence-specific TF binding to target genes, resulting in human pathologies, for example intellectual disability [86] and colorectal cancers [87]. Thus, the ability of enhancers to trigger transcription from their target core promoters is highly dependent on channeling of the regulatory information at enhancers to core promoters via cofactors. Collectively, the existence of factors that differentially activate the two types of core promoters suggest that these factors might be the *trans*-determinants of enhancer–core-promoter specificity.

Differential Protein Occupancy at Housekeeping versus Developmental Core Promoters

The distinct transcriptional outcomes when tethering the same cofactor to different core promoters in activator-bypass experiments suggests that the protein factors at the two

core-promoter types are also different (Figure 4B). Consistently, different core promoters that are activated by the same TF are differentially affected by depletion of different cofactors, suggesting that they rely on different *trans* factors [88].

Biochemical studies of proteins bound to core promoters typically used sequences with the canonical TATA box and Inr motifs. Such studies and the recent elucidation of protein-complex structures have revealed that, in an archetypical PIC, TFIID serves as the main recognition factor for the core promoter [89–91]. TFIID consists of TBP together with 12–14 TAF subunits: TBP specifically recognizes the TATA box [92–94], while some TAF subunits recognize other core-promoter elements, for example TAF1 and TAF2 bind to Inr [95], and TAF6 and TAF9 bind to DPE [96] (reviewed in [97]). In addition, TAFs can also relay the communication from the enhancers via direct contact with sequence-specific TFs [98] or by being the targets of cofactors (reviewed in [99]).

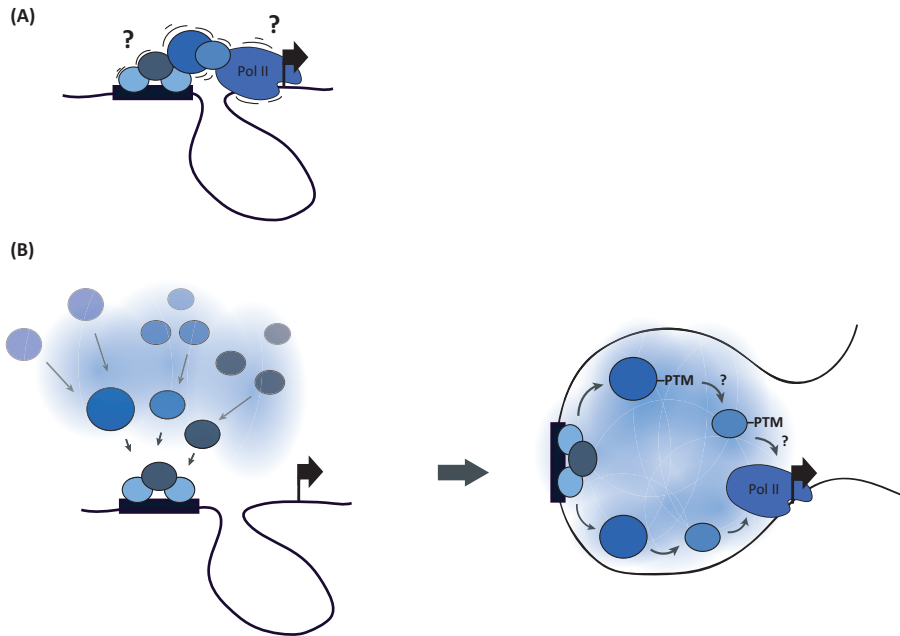
However, the PIC is far from uniform. Distinct TFIID complexes exist [100–102] that can bind to the TATA box and Inr. Further, some of the components seem to be dispensable, including for example TBP that appears not to be required for transcription, not even at TATA- and Inr-containing core promoters [103–105]. Furthermore, ‘canonical’ TAFs are dispensable in accurately positioning the PIC at a core promoter that contains a hitherto unknown core-promoter element [106], suggesting different factor requirements for different core promoter sequences.

Evidence suggests that housekeeping core promoters, which typically do not contain TATA box or Inr elements, assemble different complexes: the binding proteins for two of the housekeeping core-promoter motifs are known: Motif 1 binding protein (M1BP) recognizes Ohler motif 1 [107], while DREF recognizes DREs [108]. DREF is a part of a large complex that includes other factors including Pzgf, an hkCP-specific cofactor (see above), TATA box binding protein-related factor 2 (TRF2) [109], and components of the nucleosome-remodeling factor complex (NURF) that catalyzes nucleosome sliding downstream of active housekeeping core promoters [110].

TRF2 has been shown to bind at non-overlapping positions from TBP in the histone gene cluster [111] to control the expression of ribosomal genes [109], and it might also function at DPE-containing core promoters [112]. Further, while it was found that TAF4 highly activates the hkCP, it also exhibits preferences for DPE- over TATA-containing core promoters [113]. Some instances have been found where enhancers [53,56] as well as factors [114–116] show TATA-over DPE-specific activation, suggesting that DPE- and TATA-containing core promoters might represent different subclasses of developmental core promoters [58]. Indeed, the TATA-box and DPE motifs rarely occur together in the same core promoters [117,118].

Transcriptional Activation Mediated via Activating Microenvironments

While it is well established that enhancers and core promoters are either proximal along the linear DNA or spatially close in 3D [41,119,120], the exact details on how regulatory cues are communicated between enhancers and core promoters are not entirely clear. Increasing evidence suggests that static and rigid protein–protein interactions between these factors are not involved or at least not required (Figure 5A). For example, the activity of intronic enhancers is not disrupted when Pol II crosses them during the transcription of the host gene [121,122]. Further, the finding that a single enhancer can simultaneously activate transcription from two core promoters that are 15 kb apart in a reporter setup in transgenic *Drosophila* embryos [123] also speaks against static protein–protein contacts. On the other hand, the transcription dynamics of mammalian β - and γ -globins on a single allele are more consistent with rapid switching of enhancer–core-promoter contacts, suggesting that such dynamics might depend on the experimental system [42]. These observations, as well as others, are



Trends in Genetics

Figure 5. Enhancers Activate Core Promoters in a Microenvironment. (A) Static model of transcription regulation in which defined protein complexes formed by static protein–protein interactions at enhancers exert their function on core promoters. This model is incompatible with some observations such as the simultaneous activation of two core promoters by a single enhancer [123]. (B) Transcription regulation might alternatively occur via an activating microenvironment in which enhancers and core promoters recruit *trans* factors to create a high concentration of regulatory proteins that dynamically interact with each other, and enable regulatory communication through post-transcriptional modifications (PTMs, top) or via dynamic protein–protein interactions and recruitment (bottom).

incompatible with a scanning model in which Pol II is recruited to enhancers and then ‘scans’ to target core promoters [124,125], or with a model in which Pol II is ‘handed over’ from enhancers to core promoters (recently reviewed in [126,127]).

Such observations are compatible with an activating microenvironment around the enhancer in which enhancer-bound TFs recruit cofactors, thereby increasing their local concentration and ability to activate nearby core promoters (Figure 5B; [128,129] for more detailed review). Because different activating cofactors possess enzymatic activities, for example to post-transcriptionally modify other proteins, these activities and post-translational modifications (PTMs) might be involved in transcriptional regulation. Indeed, the activities of many TFs, general TFs of the PIC, Pol II, and some cofactors are modulated by PTMs. The communication from enhancer to core promoter thus might conceivably be transmitted via PTMs of different factors, presenting an attractive alternative mechanism compared to static interactions of the factors. For example, Pol II is acetylated by P300/CBP at specific growth factor-responsive genes, and this PTM is required for Pol II activity at these genes [130]. Similarly, P300/CBP acetylates hematopoietic TFs, leading to the recruitment of BRD4 and promoting the expression of leukemia maintenance genes [131,132]. PTMs of TFs can also modulate TF binding site preferences and thus target genes, as for example has been shown for Mef2 in *Drosophila* [133]. PTMs are also important in regulating specific steps of transcription, for example the release of paused Pol II, and it is conceivable that some enhancers specifically regulate this step [134]. These examples highlight the role of PTMs in relaying regulatory information from enhancers to target gene core promoters.

Concluding Remarks and Future Directions

The past years have witnessed enormous progress in our understanding of transcriptional regulation, the organization of animal regulatory genomes, and enhancer–core-promoter communication: insulators restrict enhancer activities and delineate the genome into regulatory domains, long-range enhancer–promoter contacts enable distal regulation, and the cell type-specific availability of core promoters can be regulated by DNA accessibility within chromatin. Furthermore, it is increasingly clear that enhancer–core-promoter specificities can be determined by the sequences of both elements: differential motif distribution allows distinct sets of factors to be recruited at enhancers and core promoters such that biochemical compatibilities between the factors determine core-promoter targeting and effective regulatory communication.

Sequence-Mediated Core–Promoter–Enhancer Specificity in Other Transcription Programs

As known core-promoter elements correlate with biological functions [58,118,135], and diverse TAFs and their paralogs exist that relate to specific transcription programs (reviewed in [136,137]), it is possible that more transcription programs are separated at the level of enhancer–core-promoter specificity. For example, during *Drosophila* spermatogenesis, an alternative TFIID complex that consists of testis-specific TAFs (tTAFs) is required for the transcription of spermatid differentiation genes, but is dispensable for meiotic cell-cycle genes [138,139]. The regulatory information comes from sequence-specific DNA binding of the tMAC complex, which relays the information through MED22 to achieve gene selectivity of tTAFs [140]. It will be exciting to learn about additional transcriptional programs, for example in germ cells.

Sequence-Encoded Enhancer and Core-Promoter Function

Genomic regulatory regions have often been annotated according to their location with respect to genes, including ‘promoter’ regions as regions upstream of gene transcription sites and ‘enhancers’ as TSS-distal. Our ability to measure transcription (initiation) with increasing sensitivity has revealed that both enhancer and promoter regions can initiate transcription and frequently contain (degenerate) core-promoter sequence elements [141–143]. Transcription initiation within enhancers has for example been used to predict enhancers [141,144–148]. Furthermore, functional assays have demonstrated that *Drosophila* promoters can partly act as enhancers to activate transcription from a distal target core promoter [60]. Consistently, human TSS-proximal sequences that can recruit cofactors display transcription-initiating (‘promoter’) activity when tested without an enhancer, as well as enhancer activity [149]. Typically, however, core-promoter activities have been measured in the presence of activating enhancers [55,150,151]. These observations suggest that future definitions might better be based on functional assays that specifically assess enhancer and core-promoter function separately. Because enhancers are typically assessed by their ability to activate transcription from a given core promoter compared to negative controls (reviewed in [6,152]), we propose to assess core-promoter functionality by the ability to convert the activity of a given enhancer into transcription initiation events, in other words by the enhancer-responsiveness of a core promoter measured as its induced versus basal activity. It will be interesting to learn which sequences within a large animal genome possess high enhancer-responsiveness and what the determining sequence properties will be.

Rethinking Enhancer–Core-Promoter Communication: Activating Microenvironments and Biochemical Compatibilities

We are excited about the prospects of learning how enhancer-bound TFs and the cofactors they recruit mediate regulatory communication with core promoters. It is clear that the proteins involved do not form rigid or stable complexes; instead a dynamic and flexible microenvironment is generated that can activate more than one core promoter simultaneously [123] and the function of which is resilient to structural perturbation by transcription through the enhancer [121,122]. Which combinations of TFs and cofactors are able to activate transcription, how

Outstanding Questions

What are the defining sequence features of a core promoter that allow high enhancer-responsiveness?

How many functionally distinct types of core promoters exist? Are additional types of enhancer–core-promoter specificities employed for other transcription programs (e.g., in germ cells)?

How are enhancer–core-promoter specificities implemented molecularly?

How is the information from the enhancer communicated to the core promoter? Do different enhancers use different modes of communication?

Which combinations of TFs activate transcription? How is combinatorial control achieved molecularly?

How many different types of TFs exist, and how many are obligately combinatorial?

Which TFs recruit which of the cofactors, and what are the respective protein domains or interaction surfaces?

Which PTMs are involved, and what are the key PTM target proteins? Are there consensus signals?

biochemical compatibility is implemented, and which PTMs on TFs, cofactors, histones, or PIC components are involved are exciting open questions for future studies (see Outstanding Questions). Perhaps in the future such knowledge could be used for therapeutic purposes, for example to very precisely activate or repress gene transcription using TALE or Cas9-derived regulators [153,154], entirely synthetic proteins [155,156], or small-molecule mimetics [157] or inhibitors [3,158–160].

Acknowledgments

We thank current and former members of group of A.S. (IMP) for discussion and feedback. The group of A.S. is supported by the European Research Council (ERC) under the European Commission Horizon 2020 research and innovation programme (grant agreement 647320) and by the Austrian Science Fund (FWF, F4303-B09). Basic research at the IMP is supported by Boehringer Ingelheim GmbH and the Austrian Research Promotion Agency (FFG).

References

- Schwanhäusser, B. *et al.* (2011) Global quantification of mammalian gene expression control. *Nature* 473, 337–342
- Arrowsmith, C.H. *et al.* (2012) Epigenetic protein families: a new frontier for drug discovery. *Nat. Rev. Drug Discov.* 11, 384–400
- Dawson, M.A. and Kouzarides, T. (2012) Cancer epigenetics: from mechanism to therapy. *Cell* 150, 12–27
- Roeder, R.G. (1996) The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.* 21, 327–335
- Spitz, F. and Furlong, E.E.M. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* 13, 613–626
- Shlyueva, D. *et al.* (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* 15, 272–286
- Banerji, J. *et al.* (1981) Expression of a β -globin gene is enhanced by remote SV40 DNA sequences. *Cell* 27, 299–308
- Gibcus, J.H. and Dekker, J. (2013) The hierarchy of the 3D Genome. *Mol. Cell* 49, 773–782
- Dixon, J.R. *et al.* (2016) Chromatin domains: the unit of chromosome organization. *Mol. Cell* 62, 668–680
- Vietri Rudan, M. and Hadjur, S. (2015) Genetic tailors: CTCF and cohesin shape the genome during evolution. *Trends Genet.* 31, 651–660
- Ulianov, S.V. *et al.* (2016) Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Res.* 26, 70–84
- Dixon, J.R. *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376–380
- Hou, C. *et al.* (2012) Gene density, transcription, and insulators contribute to the partition of the *Drosophila* genome into physical domains. *Mol. Cell* 48, 471–484
- Sexton, T. *et al.* (2012) Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* 148, 458–472
- Nora, E.P. *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485, 381–385
- Dekker, J. (2002) Capturing chromosome conformation. *Science* 295, 1306–1311
- Simonis, M. *et al.* (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* 38, 1348–1354
- Zhao, Z. *et al.* (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* 38, 1341–1347
- Dostie, J. *et al.* (2006) Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* 16, 1299–1309
- Lieberman-Aiden, E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293
- Fullwood, M.J. *et al.* (2009) An oestrogen-receptor- α -bound human chromatin interactome. *Nature* 462, 58–64
- Dekker, J. *et al.* (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* 14, 390–403
- Mahy, N.L. *et al.* (2002) Spatial organization of active and inactive genes and noncoding DNA within chromosome territories. *J. Cell Biol.* 157, 579–589
- Kvon, E.Z. *et al.* (2014) Genome-scale functional characterization of *Drosophila* developmental enhancers in vivo. *Nature* 512, 91–95
- Montavon, T. *et al.* (2011) A regulatory archipelago controls Hox genes transcription in digits. *Cell* 147, 1132–1145
- Tsujimura, T. *et al.* (2015) A discrete transition zone organizes the topological and regulatory autonomy of the adjacent *Tfap2c* and *Bmp7* genes. *PLoS Genet.* 11, e1004897
- Narendra, V. *et al.* (2015) CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science* 347, 1017–1021
- Guo, Y. *et al.* (2015) CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell* 162, 900–910
- Marinić, M. *et al.* (2013) An integrated holo-enhancer unit defines tissue and gene specificity of the *Fgf8* regulatory landscape. *Dev. Cell* 24, 530–542
- Symmons, O. *et al.* (2014) Functional and topological characteristics of mammalian regulatory domains. *Genome Res.* 24, 390–400
- Lupiáñez, D.G. *et al.* (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* 161, 1012–1025
- Gröschel, S. *et al.* (2014) A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell* 157, 369–381
- Flavahan, W.A. *et al.* (2016) Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* 529, 110–114
- Valton, A.-L. and Dekker, J. (2016) TAD disruption as oncogenic driver. *Curr. Opin. Genet. Dev.* 36, 34–40
- Hnisz, D. *et al.* (2016) Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* 351, 1454–1458
- Lin, Y.C. *et al.* (2012) Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. *Nat. Immunol.* 13, 1196–1204
- Calhoun, V.C. *et al.* (2002) Promoter-proximal tethering elements regulate enhancer–promoter specificity in the *Drosophila* Antennapedia complex. *Proc. Natl. Acad. Sci. U.S.A.* 99, 9243–9247
- Akbari, O.S. *et al.* (2008) A novel promoter-tethering element regulates enhancer-driven gene expression at the bithorax complex in the *Drosophila* embryo. *Development* 135, 123–131
- Ho, M.C.W. *et al.* (2011) Disruption of the abdominal-B promoter tethering element results in a loss of long-range enhancer-directed Hox gene expression in *Drosophila*. *PLoS ONE* 6, e16283

40. Zhou, J. and Levine, M. (1999) A novel cis-regulatory element, the PTS, mediates an anti-insulator activity in the *Drosophila* embryo. *Cell* 99, 567–575
41. Ghavi-Helm, Y. *et al.* (2014) Enhancer loops appear stable during development and are associated with paused polymerase. *Nature* 512, 96–100
42. Bartman, C.R. *et al.* (2016) Enhancer regulation of transcriptional bursting parameters revealed by forced chromatin looping. *Mol. Cell* 62, 237–247
43. Deng, W. *et al.* (2012) Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* 149, 1233–1244
44. Deng, W. *et al.* (2014) Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell* 158, 849–860
45. Li, X. and Noll, M. (1994) Compatibility between enhancers and promoters determines the transcriptional specificity of gooseberry and gooseberry neuro in the *Drosophila* embryo. *EMBO J.* 13, 400–406
46. Fakhouri, T.H.I. *et al.* (2010) Dynamic chromatin organization during foregut development mediated by the organ selector gene PHA-4/FoxA. *PLoS Genet.* 6, e1001060
47. Cochella, L. and Hobert, O. (2012) Embryonic priming of a miRNA locus predetermines postmitotic neuronal left/right asymmetry in *C. elegans*. *Cell* 151, 1229–1242
48. Tuan, D. *et al.* (1985) The 'beta-like-globin' gene domain in human erythroid cells. *Proc. Natl. Acad. Sci. U.S.A.* 82, 6384–6388
49. Forrester, W.C. *et al.* (1986) A developmentally stable chromatin structure in the human beta-globin gene cluster. *Proc. Natl. Acad. Sci. U.S.A.* 83, 1359–1363
50. Groudine, M. *et al.* (1983) Human fetal to adult hemoglobin switching: changes in chromatin structure of the beta-globin gene locus. *Proc. Natl. Acad. Sci. U.S.A.* 80, 7551–7555
51. Meril, C. *et al.* (1996) Promoter specificity mediates the independent regulation of neighboring genes. *Genes Dev.* 10, 1260–1270
52. Haberle, V. *et al.* (2014) Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature* 507, 381–385
53. Ohtsuki, S. *et al.* (1998) Different core promoters possess distinct regulatory activities in the *Drosophila* embryo. *Genes Dev.* 12, 547–556
54. Sharpe, J. *et al.* (1998) Selectivity, sharing and competitive interactions in the regulation of Hoxb genes. *EMBO J.* 17, 1788–1798
55. Gehrig, J. *et al.* (2009) Automated high-throughput mapping of promoter-enhancer interactions in zebrafish embryos. *Nat. Methods* 6, 911–916
56. Butler, J.E. and Kadonaga, J.T. (2001) Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes Dev.* 15, 2515–2519
57. Juven-Gershon, T. *et al.* (2008) Caudal, a key developmental regulator, is a DPE-specific transcriptional factor. *Genes Dev.* 22, 2823–2830
58. Kadonaga, J.T. (2012) Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip. Rev. Dev. Biol.* 1, 40–51
59. Sandelin, A. *et al.* (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat. Rev. Genet.* 8, 424–436
60. Zabidi, M.A. *et al.* (2015) Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* 518, 556–559
61. Arnold, C.D. *et al.* (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339, 1074–1077
62. Yáñez-Cuna, J.O. *et al.* (2014) Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. *Genome Res.* 24, 1147–1156
63. Hyun, J. *et al.* (2005) DREF is required for efficient growth and cell cycle progression in *Drosophila* imaginal discs. *Mol. Cell. Biol.* 25, 5590–5598
64. Yoshida, H. *et al.* (2004) DREF is required for EGFR signalling during *Drosophila* wing vein development. *Genes Cells* 9, 935–944
65. Farkas, G. *et al.* (1994) The Trithorax-like gene encodes the *Drosophila* GAGA factor. *Nature* 371, 806–808
66. Killip, L.E. and Grewal, S.S. (2012) DREF is required for cell and organismal growth in *Drosophila* and functions downstream of the nutrition/TOR pathway. *Dev. Biol.* 371, 191–202
67. Fuda, N.J. *et al.* (2015) GAGA factor maintains nucleosome-free regions and has a role in RNA polymerase II recruitment to promoters. *PLoS Genet.* 11, e1005108
68. Ptashne, M. and Gann, A. (1997) Transcriptional activation by recruitment. *Nature* 386, 569–577
69. Stampfel, G. *et al.* (2015) Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature* 528, 147–151
70. Hochheimer, A. *et al.* (2002) TRF2 associates with DREF and directs promoter-selective gene expression in *Drosophila*. *Nature* 420, 439–445
71. Kugler, S.J. and Nagel, A.C. (2007) Putzig is required for cell proliferation and regulates Notch activity in *Drosophila*. *Mol. Biol. Cell* 18, 3733–3740
72. Akhtar, A. and Becker, P.B. (2000) Activation of transcription through histone H4 acetylation by MOF, an acetyltransferase essential for dosage compensation in *Drosophila*. *Mol. Cell* 5, 367–375
73. Kind, J. *et al.* (2008) Genome-wide analysis reveals MOF as a key regulator of dosage compensation and gene expression in *Drosophila*. *Cell* 133, 813–828
74. Hilfiker, A. *et al.* (1997) Mof, a putative acetyl transferase gene related to the Tip60 and MOZ human genes and to the SAS genes of yeast, is required for dosage compensation in *Drosophila*. *EMBO J.* 16, 2054–2060
75. Smith, E.R. *et al.* (2000) The *Drosophila* MSL complex acetylates histone H4 at lysine 16, a chromatin modification linked to dosage compensation. *Mol. Cell. Biol.* 20, 312–318
76. Raja, S.J. *et al.* (2010) The nonspecific lethal complex is a transcriptional regulator in *Drosophila*. *Mol. Cell* 38, 827–841
77. Lam, K.C. *et al.* (2012) The NSL complex regulates housekeeping genes in *Drosophila*. *PLoS Genet.* 8, e1002736
78. Ding, Y. *et al.* (2009) Chromator is required for proper microtubule spindle formation and mitosis in *Drosophila*. *Dev. Biol.* 334, 253–263
79. Rath, U. *et al.* (2006) The chromodomain protein, Chromator, interacts with JIL-1 kinase and regulates the structure of *Drosophila* polytene chromosomes. *J. Cell. Sci.* 119, 2332–2341
80. Riedel, C.G. *et al.* (2013) DAF-16 employs the chromatin remodeller SWI/SNF to promote stress resistance and longevity. *Nat. Cell Biol.* 15, 491–501
81. Yang, F. *et al.* (2006) An ARC/Mediator subunit required for SREBP control of cholesterol and lipid homeostasis. *Nature* 442, 700–704
82. Ge, K. *et al.* (2002) Transcription coactivator TRAP220 is required for PPAR gamma 2-stimulated adipogenesis. *Nature* 417, 563–567
83. Yin, J.-W. *et al.* (2012) Mediator MED23 plays opposing roles in directing smooth muscle cell and adipocyte differentiation. *Genes Dev.* 26, 2192–2205
84. Boube, M. *et al.* (2014) *Drosophila melanogaster* Hox transcription factors access the RNA polymerase II machinery through direct homeodomain binding to a conserved motif of mediator subunit Med19. *PLoS Genet.* 10, e1004303
85. Alpern, D. *et al.* (2014) TAF4, a subunit of transcription factor IID, directs promoter occupancy of nuclear receptor HNF4A during post-natal hepatocyte differentiation. *Elife* 3, e03613
86. Hashimoto, S. *et al.* (2011) MED23 mutation links intellectual disability to dysregulation of immediate early gene expression. *Science* 333, 1161–1163
87. Morris, E.J. *et al.* (2008) E2F1 represses beta-catenin transcription and is antagonized by both pRB and CDK8. *Nature* 455, 552–556
88. Marr, M.T. *et al.* (2006) Coactivator cross-talk specifies transcriptional output. *Genes Dev.* 20, 1458–1469

89. Louder, R.K. *et al.* (2016) Structure of promoter-bound TFIIID and model of human pre-initiation complex assembly. *Nature* 531, 604–609
90. Cianfrocco, M.A. *et al.* (2013) Human TFIIID binds to core promoter DNA in a reorganized structural state. *Cell* 152, 120–131
91. Sainsbury, S. *et al.* (2015) Structural basis of transcription initiation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* 16, 129–143
92. Hoffman, A. *et al.* (1990) Highly conserved core domain and unique N terminus with presumptive regulatory motifs in a human TATA factor (TFIID). *Nature* 346, 387–390
93. Kao, C.C. *et al.* (1990) Cloning of a transcriptionally active human TATA binding factor. *Science* 248, 1646–1650
94. Peterson, M.G. *et al.* (1990) Functional domains and upstream activation properties of cloned human TATA binding protein. *Science* 248, 1625–1630
95. Chalkley, G.E. and Verrijzer, C.P. (1999) DNA binding site selection by RNA polymerase II TAFs: a TAF(II)250–TAF(II)150 complex recognizes the initiator. *EMBO J.* 18, 4835–4845
96. Burke, T.W. and Kadonaga, J.T. (1996) *Drosophila* TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev.* 10, 711–724
97. Smale, S.T. and Kadonaga, J.T. (2003) The RNA polymerase II core promoter. *Annu. Rev. Biochem.* 72, 449–479
98. Liu, W.-L. *et al.* (2009) Structures of three distinct activator–TFIID complexes. *Genes Dev.* 23, 1510–1521
99. Näär, A.M. *et al.* (2001) Transcriptional coactivator complexes. *Annu. Rev. Biochem.* 70, 475–501
100. Brou, C. *et al.* (1993) Distinct TFIID complexes mediate the effect of different transcriptional activators. *EMBO J.* 12, 489–499
101. Jacq, X. *et al.* (1994) Human TAFII30 is present in a distinct TFIID complex and is required for transcriptional activation by the estrogen receptor. *Cell* 79, 107–117
102. Bertolotti, A. *et al.* (1996) hTAF_{II}68, a novel RNA/ssDNA-binding protein with homology to the pro-oncoproteins TLS/FUS and EWS is associated with both TFIID and RNA polymerase II. *EMBO J.* 15, 5022–5031
103. Wieczorek, E. *et al.* (1998) Function of TAF_{II}-containing complex without TBP in transcription by RNA polymerase II. *Nature* 393, 187–191
104. Usheva, A. and Shenk, T. (1994) TATA-binding protein-independent initiation: YY1, TFIIB, and RNA polymerase II direct basal transcription on supercoiled template DNA. *Cell* 76, 1115–1121
105. Hansen, S.K. *et al.* (1997) Transcription properties of a cell type-specific TATA-binding protein. *TRF. Cell* 91, 71–83
106. Anish, R. *et al.* (2009) Characterization of transcription from TATA-less promoters: identification of a new core promoter element XCP2 and analysis of factor requirements. *PLoS ONE* 4, e5103
107. Li, J. and Gilmour, D.S. (2013) Distinct mechanisms of transcriptional pausing orchestrated by GAGA factor and M1BP, a novel transcription factor. *EMBO J.* 32, 1829–1841
108. Matsukage, A. *et al.* (2008) The DRE/DREF transcriptional regulatory system: a master key for cell proliferation. *Biochim. Biophys. Acta* 1779, 81–89
109. Wang, Y.-L. *et al.* (2014) TRF2, but not TBP, mediates the transcription of ribosomal protein genes. *Genes Dev.* 28, 1550–1555
110. Kwon, S.Y. *et al.* (2016) Genome-wide mapping targets of the metazoan chromatin remodeling factor NURF reveals nucleosome remodeling at enhancers, core promoters and gene insulators. *PLoS Genet.* 12, e1005969
111. Isogai, Y. *et al.* (2007) Transcription of histone gene cluster by differential core-promoter factors. *Genes Dev.* 21, 2936–2949
112. Kedmi, A. *et al.* (2014) *Drosophila* TRF2 is a preferential core promoter regulator. *Genes Dev.* 28, 2163–2174
113. Wright, K.J. *et al.* (2006) TAF4 nucleates a core subcomplex of TFIID and mediates activated transcription from a TATA-less promoter. *Proc. Natl. Acad. Sci. U.S.A.* 103, 12347–12352
114. Willy, P.J. (2000) A basal transcription factor that activates or represses transcription. *Science* 290, 982–984
115. Hsu, J.-Y.J. *et al.* (2008) TBP, Mot1, and NC2 establish a regulatory circuit that controls DPE-dependent versus TATA-dependent transcription. *Genes Dev.* 22, 2353–2358
116. Lewis, B.A. *et al.* (2005) Functional characterization of core promoter elements: DPE-specific transcription requires the protein kinase CK2 and the PC4 coactivator. *Mol. Cell* 18, 471–481
117. Ohler, U. *et al.* (2002) Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol.* 3, RESEARCH0087
118. Ohler, U. (2006) Identification of core promoter modules in *Drosophila* and their application in accurate transcription start site prediction. *Nucleic Acids Res.* 34, 5943–5950
119. Tolhuis, B. *et al.* (2002) Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol. Cell* 10, 1453–1465
120. Carter, D. *et al.* (2002) Long-range chromatin regulatory interactions in vivo. *Nat. Genet.* 32, 623–626
121. Mitchell, J.A. and Fraser, P. (2008) Transcription factories are nuclear subcompartments that remain in the absence of transcription. *Genes Dev.* 22, 20–25
122. Palstra, R.-J. *et al.* (2008) Maintenance of long-range DNA interactions after inhibition of ongoing RNA polymerase II transcription. *PLoS ONE* 3, e1661
123. Fukaya, T. *et al.* (2016) Enhancer control of transcriptional bursting. *Cell* 166, 358–368
124. Heuchel, R. *et al.* (1989) Two closely spaced promoters are equally activated by a remote enhancer: evidence against a scanning model for enhancer action. *Nucleic Acids Res.* 17, 8931–8947
125. Müller, H.P. *et al.* (1990) A transcriptional terminator between enhancer and promoter does not affect remote transcriptional control. *Somat. Cell Mol. Genet.* 16, 351–360
126. Beagrie, R.A. and Pombo, A. (2016) Gene activation by metazoan enhancers: diverse mechanisms stimulate distinct steps of transcription. *Bioessays* 38, 881–893
127. Vernimmen, D. and Bickmore, W.A. (2015) The hierarchy of transcriptional activation: from enhancer to promoter. *Trends Genet.* 31, 696–708
128. Kulaeva, O.I. *et al.* (2012) Distant activation of transcription: mechanisms of enhancer action. *Mol. Cell. Biol.* 32, 4892–4897
129. Lemon, B. and Tjian, R. (2000) Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev.* 14, 2551–2569
130. Schröder, S. *et al.* (2013) Acetylation of RNA polymerase II regulates growth-factor-induced gene transcription in mammalian cells. *Mol. Cell* 52, 314–324
131. Roe, J.-S. *et al.* (2015) BET bromodomain inhibition suppresses the function of hematopoietic transcription factors in acute myeloid leukemia. *Mol. Cell* 58, 1028–1039
132. Bhagwat, A.S. *et al.* (2016) BET bromodomain inhibition releases the mediator complex from select cis-regulatory elements. *Cell Rep.* 15, 519–530
133. Clark, R.I. *et al.* (2013) MEF2 is an in vivo immune-metabolic switch. *Cell* 155, 435–447
134. Adelman, K. and Lis, J.T. (2012) Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.* 13, 720–731
135. Katzenberger, R.J.R. *et al.* (2011) The *Drosophila* translational control element (TCE) is required for high-level transcription of many genes that are specifically expressed in testes. *PLoS ONE* 7, e45009
136. D'Alessio, J.A. *et al.* (2009) Shifting players and paradigms in cell-specific transcription. *Mol. Cell* 36, 924–931
137. Müller, F. *et al.* (2010) Developmental regulation of transcription initiation: more than just changing the actors. *Curr. Opin. Genet. Dev.* 20, 533–540
138. Hiller, M. *et al.* (2004) Testis-specific TAF homologs collaborate to control a tissue-specific transcription program. *Development* 131, 5297–5308
139. Chen, X. *et al.* (2005) Tissue-specific TAFs counteract Polycomb to turn on terminal differentiation. *Science* 310, 869–872

140. Lu, C. and Fuller, M.T. (2015) Recruitment of Mediator complex by cell type and stage-specific factors required for tissue-specific TAF dependent gene activation in an adult stem cell lineage. *PLoS Genet.* 11, e1005701
141. 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
142. Andersson, R. (2015) Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *Bioessays* 37, 314–323
143. Kim, T.-K. and Shiekhattar, R. (2015) Architectural and functional commonalities between enhancers and promoters. *Cell* 162, 948–959
144. Hah, N. *et al.* (2013) Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res.* 23, 1210–1223
145. Li, W. *et al.* (2013) Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* 498, 516–520
146. Scruggs, B.S. *et al.* (2015) Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. *Mol. Cell* 58, 1101–1112
147. Andersson, R. *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455–461
148. Schwalb, B. *et al.* (2016) TT-seq maps the human transient transcriptome. *Science* 352, 1225–1228
149. Nguyen, T.A. *et al.* (2016) High-throughput functional comparison of promoter and enhancer activities. *Genome Res.* 26, 1023–1033
150. Ede, C. *et al.* (2016) Quantitative analyses of core promoters enable precise engineering of regulated gene expression in mammalian cells. *ACS Synthetic Biol.* 5, 395–404
151. Lubliner, S. *et al.* (2013) Sequence features of yeast and human core promoters that are predictive of maximal promoter activity. *Nucleic Acids Res.* 41, 5569–5581
152. White, M.A. (2015) Understanding how cis-regulatory function is encoded in DNA sequence using massively parallel reporter assays and designed sequences. *Genomics* 106, 165–170
153. Doudna, J.A. and Charpentier, E. (2014) The new frontier of genome engineering with CRISPR–Cas9. *Science* 346, 1258096
154. Dominguez, A.A. *et al.* (2015) Beyond editing: repurposing CRISPR–Cas9 for precision genome regulation and interrogation. *Nat. Rev. Mol. Cell Biol.* 17, 5–15
155. Mapp, A.K. *et al.* (2000) Activation of gene expression by small molecule transcription factors. *Proc. Natl. Acad. Sci. U.S.A.* 97, 3930–3935
156. Xiao, X. *et al.* (2007) Design and synthesis of a cell-permeable synthetic transcription factor mimic. *J. Comb. Chem.* 9, 592–600
157. Nyanguile, O. *et al.* (1997) A nonnatural transcriptional coactivator. *Proc. Natl. Acad. Sci. U.S.A.* 94, 13402–13406
158. Dawson, M.A. *et al.* (2011) Inhibition of BET recruitment to chromatin as an effective treatment for MLL-fusion leukaemia. *Nature* 478, 529–533
159. Zuber, J. *et al.* (2011) RNAi screen identifies Brd4 as a therapeutic target in acute myeloid leukaemia. *Nature* 478, 524–528
160. Knutson, S.K. *et al.* (2012) A selective inhibitor of EZH2 blocks H3K27 methylation and kills mutant lymphoma cells. *Nat. Chem. Biol.* 8, 890–896
161. Tomancak, P. *et al.* (2002) Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.* 3, RESEARCH0088

Paper #3: Genome-wide Assessment of Sequence-intrinsic Enhancer Responsiveness at Single-base-pair Resolution

Cosmas D. Arnold*, **Muhammad A. Zabidi***, Michaela Pagani, Martina Rath, Katharina Schernhuber, Tomáš Kazmar & Alexander Stark, *Nature Biotechnology*, 2017. 35(2) pp. 136-144.

*These authors contributed equally.

Core promoters convert signaling information from enhancers into transcription initiation. Accordingly, a principal property of core promoters is *enhancer-responsiveness* that describes their inducibility upon activation by enhancers. However, the extent of the enhancer-responsiveness of core promoters, and how enhancer-responsiveness is important in gene expression, have remained unknown. Transcription initiation has also been observed at genomic positions that are distal from gene starts, the expected locations of core promoters. Whether these distal positions represent bona fide core promoters have remained unexplained.

Available genome-wide methods do not enable the quantification of enhancer-responsiveness. These methods quantify endogenous transcripts, the sum of intrinsic traits of core promoters, enhancer activities and other determinants. To measure enhancer-responsiveness of core promoters in the genome, we developed self-transcribing active core promoter sequencing (STAP-seq). This reporter assay enables the quantification of the potential of genome-wide core promoter candidate fragments to act as core promoters towards single and defined enhancers. Importantly, this setup allows direct comparison of the intrinsic activity of the core promoter candidates.

Core promoters in the genome exhibit a wide range of enhancer-responsiveness. Enhancer-responsiveness correlate with gene functions: highly responsive core promoters are associated with TFs, and lowly responsive core promoters are associated with cell-type specific enzymes. Core promoter sequences are also predictive of their enhancer-responsiveness. Finally, enhancer-responsiveness is significantly lower at distal positions of endogenous initiation. This observation supports that such positions are the sites of transcription initiation as they associate with enhancers during endogenous transcription.

Author contributions

C.D.A., **M.A.Z.**, and A.S. conceived the project. C.D.A., M.P., and M.R. performed the experiments with the help of K.S., and **M.A.Z.** the computational analyses. T.K. performed the k-mer based predictions. C.D.A., **M.A.Z.**, and A.S. wrote the manuscript. A.S. supervised the project.

ARTICLES

Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution

Cosmas D Arnold^{1,2}, Muhammad A Zabidi^{1,2}, Michaela Pagani¹, Martina Rath¹, Katharina Schernhuber¹, Tomáš Kazmar¹ & Alexander Stark¹

Gene expression is controlled by enhancers that activate transcription from the core promoters of their target genes. Although a key function of core promoters is to convert enhancer activities into gene transcription, whether and how strongly they activate transcription in response to enhancers has not been systematically assessed on a genome-wide level. Here we describe self-transcribing active core promoter sequencing (STAP-seq), a method to determine the responsiveness of genomic sequences to enhancers, and apply it to the *Drosophila melanogaster* genome. We cloned candidate fragments at the position of the core promoter (also called minimal promoter) in reporter plasmids with or without a strong enhancer, transfected the resulting library into cells, and quantified the transcripts that initiated from each candidate for each setup by deep sequencing. In the presence of a single strong enhancer, the enhancer responsiveness of different sequences differs by several orders of magnitude, and different levels of responsiveness are associated with genes of different functions. We also identify sequence features that predict enhancer responsiveness and discuss how different core promoters are employed for the regulation of gene expression.

Animal development is coordinated by differential gene expression that is commonly attributed to the dynamic and cell-type-specific activities of transcriptional enhancer sequences^{1,2}. Enhancers are genomic regulatory elements that recruit transcription factors and cofactors to activate transcription from their target core promoters, short sequences at the 5' ends of genes at which RNA polymerase II (Pol II) assembles and gene transcription initiates^{3,4}. However, sensitive methods to assess endogenous transcription have revealed many positions outside gene starts that initiate transcription, blurring the distinction between promoters and other genomic regions^{5,6}. We reasoned that one of the key functions of bona fide core promoters that is essential for differential gene expression is their ability to strongly respond to enhancers, that is, to efficiently convert the enhancers' activating cues into productive gene transcription. However, despite the central importance of this enhancer responsiveness, it has remained unclear how many sequences in large animal genomes respond to a given enhancer, which sequences respond most strongly, and how wide the range of response strength is. Knowing the range of enhancer responsiveness and the sequence features of strongly versus weakly responding candidates is critical for determining how enhancer responsiveness is sequence-encoded and differentially employed for the regulation of gene expression. It is also important for our understanding of transcription and might explain the cause of endogenous transcription initiation within different genomic regions.

However, the quantitative assessment of enhancer responsiveness in a standardized manner is challenging and cannot be derived from measures of endogenous transcription. Endogenous initiation rates

result from activation by different—possibly many^{7,8}—enhancers of different strengths, and the combined contributions of core promoter and enhancer functionality cannot be deconvoluted. Therefore, the responsiveness of candidate sequences to an enhancer needs to be assessed in standardized reporter assays, under the influence of defined enhancers that are kept constant, a technique equivalent to the widely used enhancer activity assays that operate with constant core promoters². Such assays have been performed for individual core promoters in human and *Drosophila* cells^{9–11}, and systematic medium-scale tests of yeast core promoters¹² and of mammalian core promoters *in vitro*¹³ exist. However, a method to systematically assess enhancer responsiveness for millions of candidate fragments across large genomes is lacking.

Here we present STAP-seq, a method to assess the sequence-intrinsic enhancer responsiveness of millions of candidate sequences at single-base-pair resolution and perform a genome-wide analysis of their enhancer responsiveness in *Drosophila melanogaster* cells. We find that in the presence of a single strong enhancer, thousands of sequences exhibit enhancer-responsive transcriptional activation with strengths that vary over three orders of magnitude. The strength of the response can be predicted from the candidates' primary sequences, and the host genes of candidates that respond strongly or weakly to enhancers differ characteristically in their function. Positions within enhancers that initiate transcription endogenously exhibit low enhancer responsiveness compared to positions within bona fide core promoters. Overall, our systematic analysis identifies sequences that efficiently convert enhancer activity into transcription initiation events and shows how

¹Research Institute of Molecular Pathology (IMP), Vienna Biocenter (VBC), Vienna, Austria. ²These authors contributed equally to this work. Correspondence should be addressed to A.S. (stark@starklab.org).

Received 12 April; accepted 8 November; published online 26 December 2016; doi:10.1038/nbt.3739

sequences with different enhancer responsiveness are employed for the regulation of gene expression.

RESULTS

Determining sequence-intrinsic enhancer responsiveness

To identify genomic sequences that can initiate transcription in response to enhancers and to directly assess the strength of the response (that is, the candidate fragments' inducibility or enhancer responsiveness), we developed STAP-seq (Fig. 1a). We separately determine the induced and basal activities for candidate DNA fragments at single-base-pair resolution in standardized reporter setups that do (STAP-seq^{enh}) or do not (STAP-seq^{ctrl}) contain a single defined enhancer. For STAP-seq^{enh}, we randomly sheared *D. melanogaster* genomic DNA into short fragments (median length, 192 bp; Supplementary Fig. 1a) and cloned them in bulk into a reporter plasmid at the position of the core promoter (also called the minimal promoter), between a defined strong developmental enhancer (from the transcription factor *Zn finger homeodomain 1* (*zfh1*)^{10,14}) and a protein-coding open reading frame (ORF; Fig. 1a). Core promoters are typically ~100 bp long, and the candidate fragments thus also included flanking sequences up- and/or downstream of potential core promoters^{15,16}. This, however, should not influence our ability to assess enhancer responsiveness as negligible differences were observed between sequences within this length range¹⁰ (see also below for a demonstration that most candidates have low basal activities). If a fragment initiates transcription in response to the enhancer, it will produce reporter transcripts, and the number of reporter transcripts generated will directly reflect the fragment's induced activity. Because all reporter transcripts will be identical except for short 5' sequence tags that originate from the respective candidate, this allows the quantification of the candidate's activity in the respective reporter setup (Fig. 1a). For STAP-seq^{ctrl}, we repeated the above with an enhancer-less reporter setup to assess the candidate fragments' basal activities and subsequently their enhancer responsiveness (see below).

We transfected *D. melanogaster* S2 cells with each reporter library, isolated polyadenylated RNA, and followed a modified CapSeq protocol^{17,18} to selectively capture the reporter mRNAs' 5' sequence tags, which enable the precise mapping of the transcription start sites (TSSs; i.e., the +1 nucleotides) throughout the genome. Briefly, all non-5' capped RNA species were rendered ligation-incompetent by dephosphorylating their 5' ends with calf intestinal phosphatase. Subsequently, the 5' caps were removed with tobacco acid pyrophosphatase, and RNA oligonucleotides, each containing an 8-nucleotide (nt) random barcode as a unique molecular identifier, were ligated to the resulting 5' phosphate RNA molecules. The 5' sequence tags of the reporter transcripts were then selectively reverse transcribed, amplified, and paired-end sequenced. The paired-end reads were aligned to the *D. melanogaster* genome and the initiation events at each genomic position were quantified in a strand-specific manner by the number of unique sequence tags, as identified by the unique molecular identifiers. We performed two technical replicates for both STAP-seq^{enh} and STAP-seq^{ctrl}, which in each case were highly similar (Pearson correlation coefficient (PCC) = 0.99 and 0.93, respectively; see below for replicates with independent libraries and transfections and an analysis of the variance of individual data points), and combined them for further analyses. We will now first discuss the induced activities obtained from STAP-seq with the *zfh1* enhancer (STAP-seq^{zfh1}) and compare them with measures of endogenous initiation (basal activities cannot be assessed in endogenous contexts) and with STAP-seq screens using different enhancers and another cell type, before we discuss enhancer responsiveness as the normalized ratio STAP-seq^{enh} versus STAP-seq^{ctrl}.

STAP-seq identifies endogenous transcription start sites

STAP-seq^{zfh1} revealed a highly distinctive genomic profile of candidate transcription initiation events with specific signals that overlapped TSSs annotated by FlyBase (aTSSs; Fig. 1b). Indeed, even though candidates from across the entire genome were assayed, over half (55%) of all 28,509 genomic positions with ≥5 tags mapped to within 50 bp of an aTSS, and the degree of alignment with aTSSs improved with higher tag counts: 66% of all positions with ≥10 tags and 71% of all positions with ≥20 tags were within 50 bp, and 30% (≥5 tags), 39% (≥10 tags), or 45% (≥20 tags) were within 5 bp, respectively (Supplementary Fig. 2a). The latter corresponds to more than 140-fold enrichment over the genome (only 0.32% of the genome is within 5 bp of an aTSS; Supplementary Fig. 2a,b).

Furthermore, plotting the cumulative STAP-seq^{zfh1} tag count around all aTSSs revealed a strong enrichment, with the highest value precisely at the +1 position (Fig. 1c and Supplementary Fig. 2c). Data sets that measure endogenous transcription initiation^{18,19} and the analysis of the Initiator (Inr) motif, known to coincide with TSSs²⁰, also supported STAP-seq-defined TSSs (experimentally defined TSSs or eTSSs; see Online Methods), even if they did not map to aTSSs (Fig. 1d,e and Supplementary Fig. 2d–h).

Overall, these results suggest that STAP-seq identifies positions that initiate transcription endogenously and that aTSSs at annotated gene starts are distinguished among the many genome-wide candidate fragments by their ability to efficiently convert enhancer activities into transcription initiation events. Because we tested short fragments in a defined reporter setup outside their endogenous sequence and chromatin contexts, these results confirm that the ability to convert enhancer activities into transcription initiation events and the precise position of transcription initiation are encoded in the DNA sequence.

Induced activities are consistent for three developmental enhancers

We next asked whether induced activities are influenced by the choice of enhancer. We therefore repeated STAP-seq with the *zfh1* enhancer and with additional enhancers for a focused candidate library of reduced complexity, derived from 34 bacterial artificial chromosomes (BACs) that cover about 5% of the *D. melanogaster* genome and contain ~1,100 aTSSs. For these screens, we chose an intronic enhancer within *sugarless* (*sgl*) and an intergenic enhancer near *hamlet* (*ham*), both developmental enhancers weaker than the *zfh1* enhancer¹⁰. In addition, we chose two housekeeping enhancers close to *nucampholin* (*ncm*) and to *short spindle 3* (*ssp3*), respectively. As they were expected to specifically activate housekeeping- but not developmental-type core promoters¹⁰, they served as an outgroup in our subsequent analysis.

For the screens using developmental enhancers, strong signals were observed at aTSSs of developmentally regulated genes, while the housekeeping enhancers produced strong signals at aTSSs of housekeeping genes (Fig. 2a). This confirmed the expected core promoter specificity¹⁰ of the three developmental and the two housekeeping enhancers, and suggests that within each of these two broad transcriptional programs¹⁰—but not across programs—the tested fragments respond similarly to different enhancers.

Indeed, the focused screens with all three developmental enhancers were highly similar (all PCCs ≥ 0.83; Fig. 2b and Supplementary Fig. 3a) and also agreed well with the genome-wide screen (PCC = 0.86 between focused and genome-wide STAP-seq^{zfh1}; Supplementary Fig. 3b; for an analysis of the variance of individual data points, see Supplementary Fig. 3c). The most highly induced sequences in STAP-seq^{zfh1} were also the most highly induced ones when using the weaker

ARTICLES

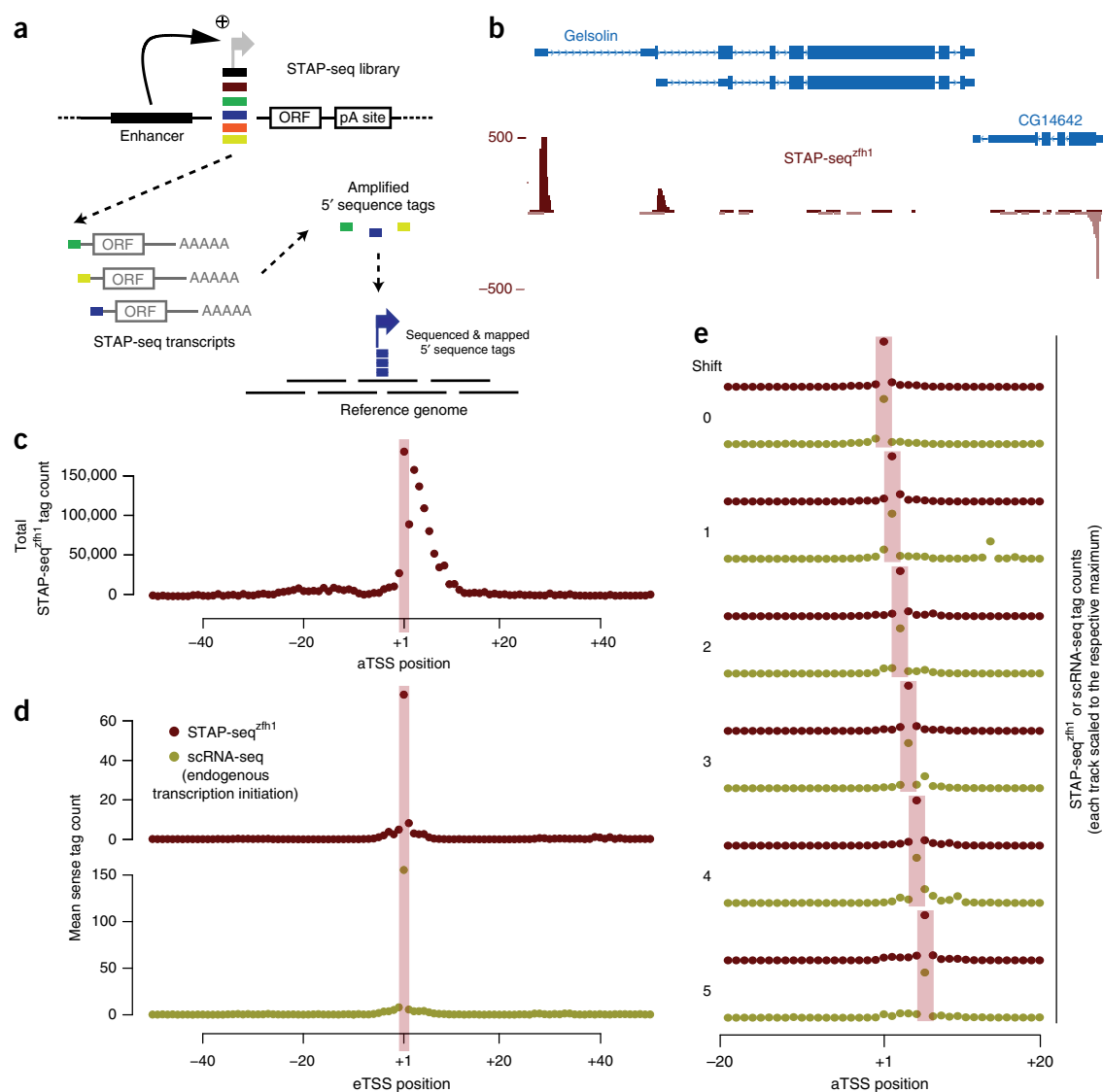


Figure 1 STAP-seq identifies position and orientation of transcription initiation within arbitrary candidate fragments. **(a)** Experimental setup of STAP-seq. Short candidate DNA fragments are cloned into a reporter construct that provides an enhancer and a reporter gene (short open reading frame (ORF)). Active candidates initiate reporter transcripts that start with sequence tags depicting the exact TSS. These tags are then sequenced and mapped to the reference genome. **(b)** UCSC Genome browser screenshot depicting STAP-seq using the *zfh1* enhancer. Tag coverage is shown in a strand-specific manner. **(c)** Cumulative STAP-seq^{zfh1} tag counts around FlyBase-annotated TSSs (aTSS). **(d)** Metagene profile of STAP-seq^{zfh1} tag counts and short-nuclear-capped-RNA-seq (scRNA-seq)¹⁸ signals at experimentally determined STAP-seq TSSs (eTSSs). **(e)** Agreement of STAP-seq^{zfh1} and scRNA-seq¹⁸ for eTSSs that are shifted with respect to aTSSs by 1–5 nt.

developmental *sgl* and *ham* enhancers (Fig. 2b,c and Supplementary Fig. 3a), and their rankings agreed well (Spearman's correlation coefficient (SCC) = 0.75 for *sgl* versus *zfh1* and for *ham* versus *zfh1*). The similarity between the induced activities with each of the different developmental enhancers became particularly apparent in the comparison to the outgroup. Whereas the screens performed with the two housekeeping enhancers were highly similar (PCC = 0.88; Fig. 2b), they differed characteristically from the screens with the developmental enhancers (*sgl* versus *ncm* and *ssp3* had PCC = 0.18 and 0.16, respectively; *ham* versus *ncm* and *ssp3* had PCC = 0.16 and 0.14, respectively; Fig. 2b,c). Indeed, when we grouped all five screens by hierarchical clustering, the three developmental screens clustered

tightly as did the two housekeeping screens, forming an outgroup as expected (Fig. 2c).

These results show that different sequences responded to developmental versus housekeeping enhancers, recapitulating the previously reported enhancer–core-promoter specificity¹⁰. They demonstrate that STAP-seq is a tool that can probe enhancer responsiveness of millions of candidate fragments and identify strongly responding sequences for different types of enhancers. Notably, the responses were consistent across the three different developmental enhancers and across the two different housekeeping enhancers we tested, suggesting that enhancer responsiveness within a given transcriptional program is independent of the

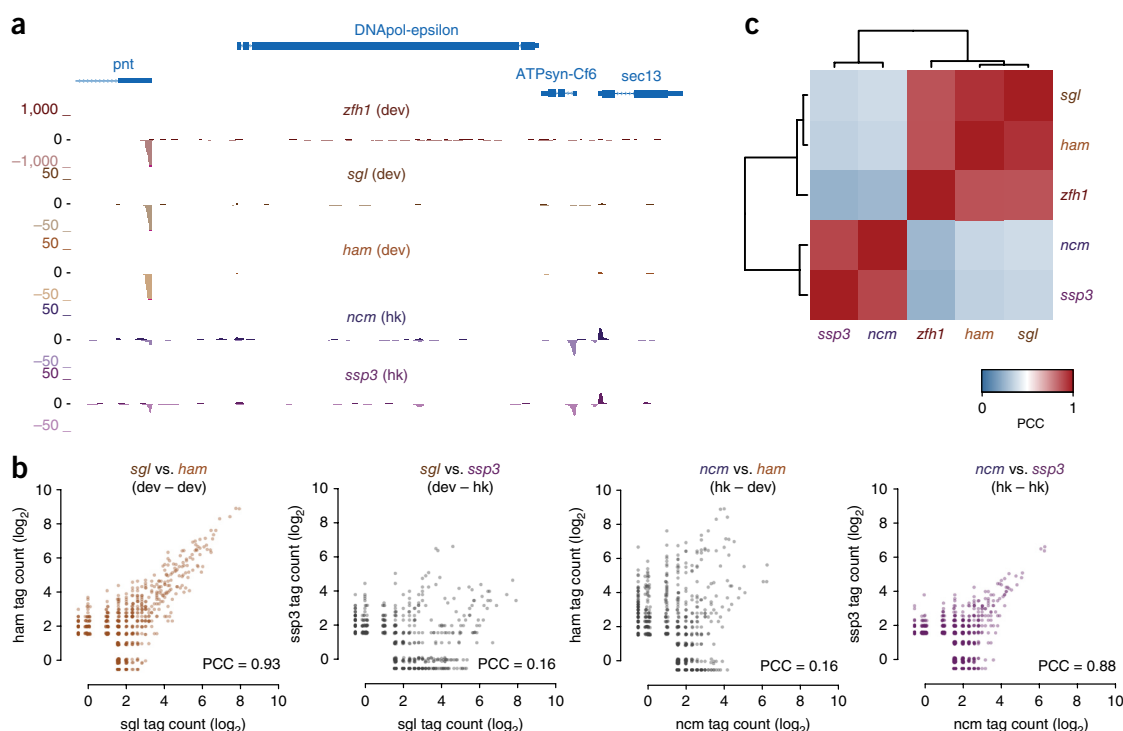


Figure 2 Induced activities are consistent across developmental enhancers. (a) UCSC genome browser screenshot showing STAP-seq signals of the focused screens using the indicated developmental (dev; *zfh1*, *sgl*, and *ham*) and housekeeping (hk; *ncm* and *ssp3*) enhancers. The depicted locus covers developmental and housekeeping genes. *Pointed* (*pnt*) codes for a transcription factor, whereas *ATP synthase, coupling factor 6* (*ATPsyn-Cf6*), and *Secretory 13* (*sec13*) code for components of ATP synthase and nuclear pore complex, respectively. (b) Scatterplots depicting the similarity of STAP-seq screens with two developmental (left) and two housekeeping (right) enhancers, respectively, and the dissimilarity between developmental- and housekeeping-enhancer screens (middle). (c) Bi-clustered heatmap depicting pairwise similarities. Pearson correlation coefficients (PCCs) of STAP-seq tag counts for three developmental and two housekeeping enhancers.

particular enhancers used and thus constitutes a functional sequence feature of general importance.

Activities are consistent across two different cell types

To investigate whether the different induced activities observed for different sequences are consistent across different cell types or vary with cell-type-specific gene expression, we repeated STAP-seq with the focused library in *D. melanogaster* ovarian somatic cells (OSCs)²¹. OSCs differ from S2 cells in gene expression and enhancer activities^{10,14}, which required exchanging the S2-specific *zfh1* enhancer with an OSC-specific developmental enhancer from *traffic jam* or *tj*¹⁴. Notably, STAP-seq^{zfh1} in S2 cells and STAP-seq^{tj} in OSCs were highly similar (PCC = 0.85; Fig. 3a,b and Supplementary Fig. 4), reminiscent of the screens with three different developmental enhancers in S2 cells (Fig. 2b,c and Supplementary Fig. 3a), and much above the similarity of endogenous transcription between the two cell types as assessed by Global Run-On sequencing (GRO-seq)^{22,23} (PCC = 0.31 at aTSSs; Fig. 3c). Even sequences that are endogenously active exclusively in S2 cells or OSCs, respectively, behave similarly across the two cell types in STAP-seq (Kolmogorov–Smirnov test *P* value > 0.1; Fig. 3a and Supplementary Fig. 4).

Together, these results suggest that the enhancer responsiveness of a DNA sequence is an intrinsic property that is independent of cell-type-specific gene expression, confirming previous observations during transgene expression with widely used minimal promoters (e.g., the *Drosophila* synthetic core promoter in different tissues in transgenic flies^{8,24}). These results also demonstrate the complementarity

of STAP-seq and methods that assess endogenous transcription initiation, such as GRO-seq, that detects positions of transcriptionally engaged Pol II, and short nuclear-capped RNA-seq (scRNA-seq) that measures short nascent transcripts *in vivo* (Fig. 3d–f and Supplementary Fig. 5); whereas endogenous transcription initiation reflects cell-type-specific gene expression, STAP-seq measures the sequence-intrinsic ability of DNA fragments to initiate transcription in response to an enhancer, that is, it measures the fragments' enhancer responsiveness. Notably, enhancer responsiveness appears to be consistent across different enhancers and cell types.

A wide and continuous range of enhancer responsiveness

A notable aspect of the STAP-seq data is the very wide range of induced activities for the different candidate fragments. Whereas the vast majority of the tested genomic positions did not initiate transcription in STAP-seq^{zfh1} (0 tags), 1,864 eTSSs had ≥100 tags at their +1 positions, 136 had ≥1,000, and the strongest eTSS had 14,249—even though all fragments were tested using the same enhancer. The consistency of these differences between different enhancers and across different cell types suggests that the ability to efficiently convert enhancer-activity into transcription initiation events is a sequence-intrinsic property and an important contributor to transcription regulation.

We therefore define the inducibility or enhancer responsiveness of a candidate sequence as the ratio of its induced versus basal activity, measured by STAP-seq^{enh} and STAP-seq^{ctrl}, respectively. To assess enhancer responsiveness genome-wide, we repeated STAP-seq

ARTICLES

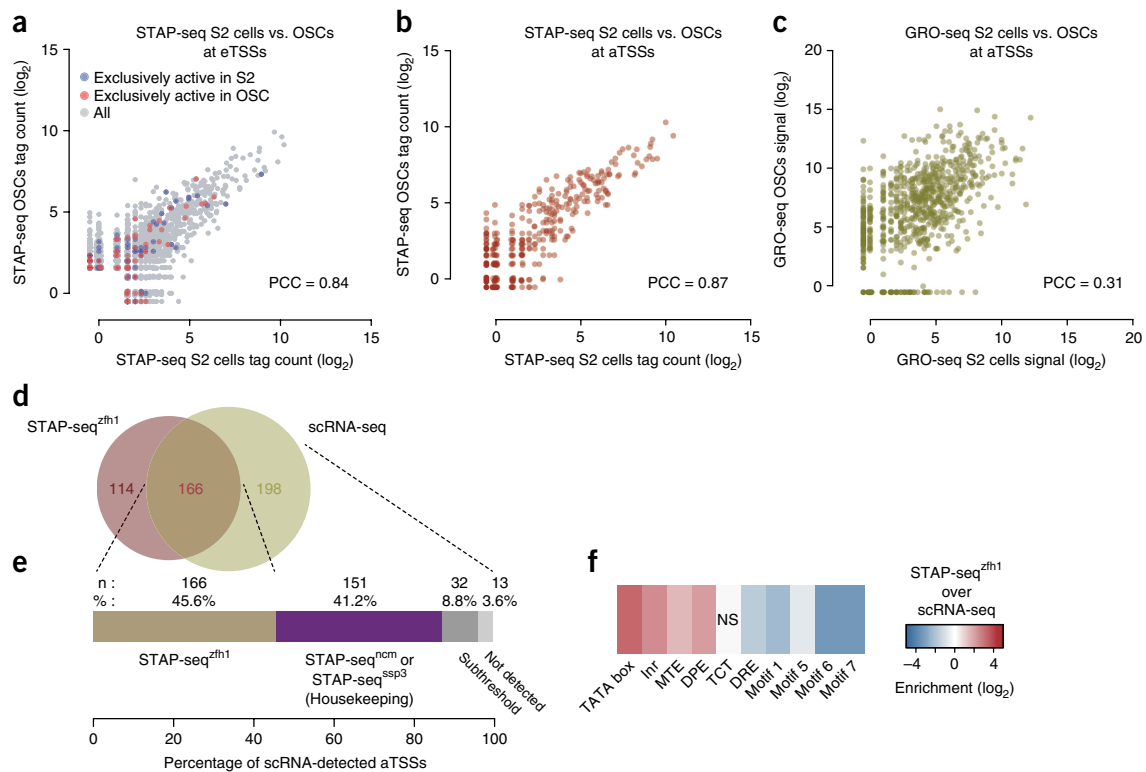


Figure 3 Induced activities are consistent across cell types. (a) Scatterplot depicting STAP-seq tag counts for STAP-seq^{zhf1} in S2 cells (x-axis) versus STAP-seq^{ti} in OSCs (y-axis) and their similarity (PCC). TSSs that endogenously—as measured by GRO-seq^{22,23}—are exclusively active in S2 cells or OSCs are labeled blue or red, respectively (see also **Supplementary Fig. 4**). (b,c) Scatterplots depicting comparisons between STAP-seq in b, and GRO-seq^{22,23} in c, in S2 cells and OSCs at aTSSs. (d) Venn diagram depicting the overlap of aTSSs detected by STAP-seq^{zhf1} and scRNA-seq¹⁸ in S2 cells for genomic regions covered in the focused STAP-seq screens. (e) Breakdown of aTSSs detected by scRNA-seq¹⁸: 45.6% are also detected by STAP-seq^{zhf1} and essentially all other aTSSs are detected by the focused STAP-seq screens with housekeeping enhancers. Only 13 aTSSs (3.6%) are not found by developmental and housekeeping STAP-seq screens combined. (f) Core promoter motif-enrichment analyses of aTSSs uniquely detected by either STAP-seq^{zhf1} or scRNA-seq¹⁸. NS, not significant.

without an enhancer (STAP-seq^{ctrl}; **Supplementary Fig. 1b**), again obtaining two highly similar replicates (PCC = 0.93), and divided the induced by the basal activity for each eTSS, normalizing to spike-in controls present in both samples (Online Methods). This revealed a very wide range of developmental enhancer responsiveness with an up to 1,000-fold difference between the highest and lowest inducibility (**Supplementary Fig. 6a–c**; housekeeping enhancer responsiveness had a much reduced dynamic range, see **Supplementary Fig. 6d**). We also found a similarly wide range of responsiveness for the known aTSSs (**Fig. 4a**), particularly when we corrected their positions based on short nuclear-capped RNA-seq (scRNA-seq) data^{18,25} or restricted the analysis to corrected aTSSs containing exclusively TATA box, Inr, DPE, or MTE motifs (**Supplementary Fig. 6b** and Online Methods). By contrast, analyzing the same number of randomly selected positions (**Fig. 4b**) or antisense initiation at the +1 positions of eTSSs (**Fig. 4c**) revealed only very weak enhancer responsiveness.

To validate the different levels of enhancer responsiveness determined by STAP-seq, we tested 30 sequences (from ranks 2 to 4,675; 19 known aTSSs and 11 eTSSs that to our knowledge had not been described previously, including 2 candidates that overlap exons) and 16 negative controls (12 aTSSs and 4 candidates without TSS annotation) individually in luciferase assays. The enhancer responsiveness determined by STAP-seq and luciferase induction showed a high linear agreement (PCC = 0.96; **Fig. 4d**), much higher than the PCCs observed between the luciferase results and methods that measure endogenous

transcription initiation (luciferase versus scRNA-seq¹⁸, CAGE²⁶, and GRO-seq²³ show PCCs of 0.35, 0.08, and 0.4, respectively). Together, these results validate the wide range of enhancer responsiveness and establish STAP-seq as a quantitative genome-wide assay to functionally quantify this measure, which does not necessarily correlate with endogenous transcription rates in any particular cell type.

Enhancer responsiveness correlates with gene function

The results above reveal that sequences in the genome vary widely in their ability to convert enhancer activities into transcription initiation events. Strong additive activation by multiple enhancers might require highly inducible TSSs, because weaker ones might otherwise limit transcription rates. Indeed, sequences proximal to aTSSs of genes with five or more developmental enhancers^{10,14} to the developmental *zhf1* enhancer were significantly more inducible than those with only one or two developmental enhancers (**Fig. 4e**; P value = 2.66×10^{-6}). Furthermore, sequences with different levels of enhancer responsiveness to the developmental *zhf1* enhancer were also located near genes of different biological functions (**Supplementary Fig. 7a**). The overall most responsive sequences tended to be near genes involved in development, regulation of gene expression, and response to stimuli, whereas the overall weakest ones were next to housekeeping genes, as expected, given the incompatibility of the developmental *zhf1* enhancer, used in STAP-seq^{zhf1}, with the core promoters of housekeeping genes¹⁰. When we restricted the analysis to eTSSs that contain only TATA box, Inr,

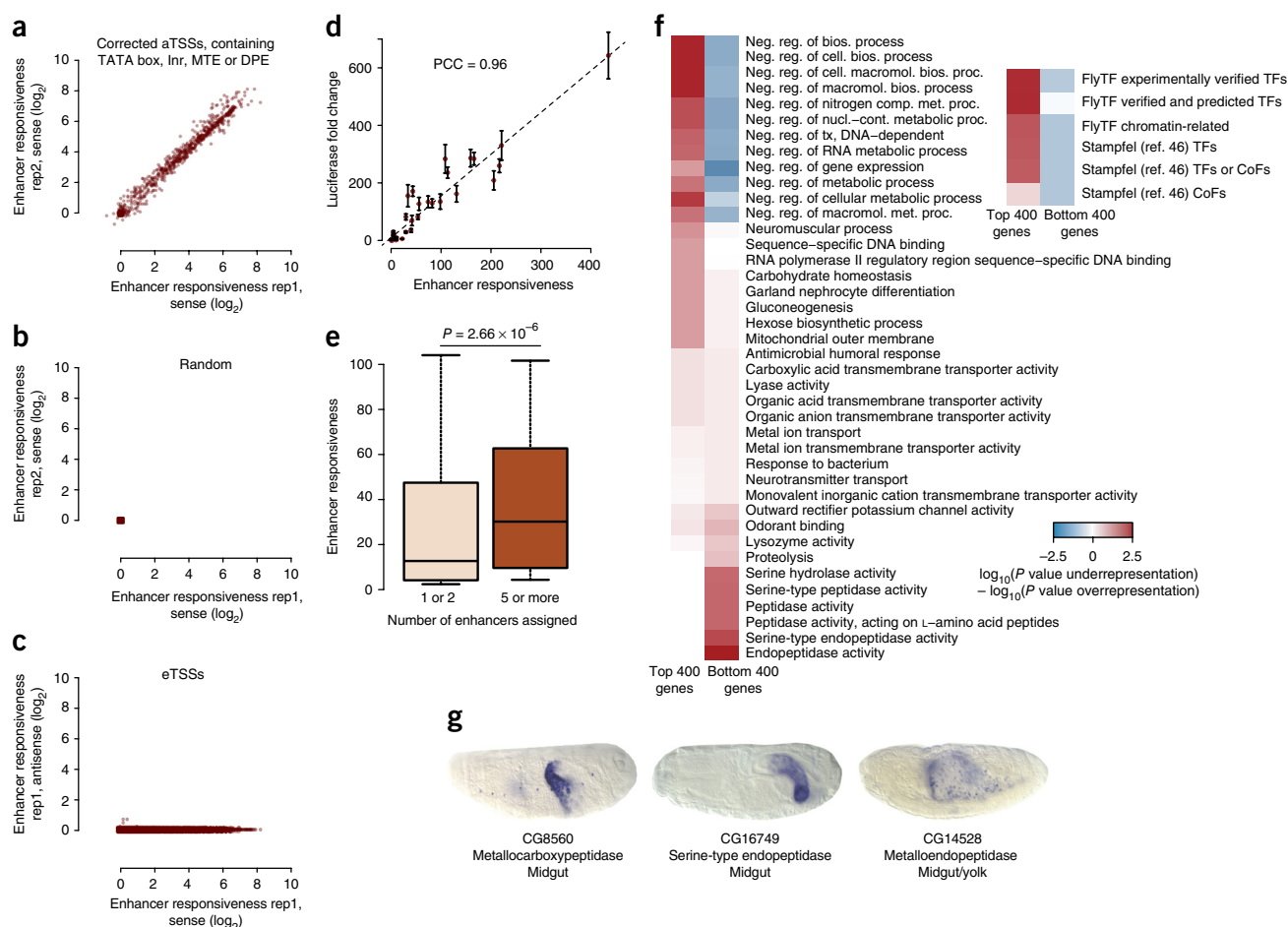


Figure 4 Wide range of enhancer responsiveness and associated biological functions. (a–c) Scatterplots showing the range of enhancer responsiveness at corrected aTSSs that contain exclusively TATA box, Inr, MTE, or DPE in **a**, random positions in **b**, and eTSSs in **c**, depicting replicate 1 versus 2 in **a**, and **b**, and sense versus antisense signals of replicate 1 in **c**. (d) Enhancer responsiveness according to STAP-seq versus luciferase induction by the *zfh1* enhancer. Error bars, s.d.; $n = 3$. (e) Boxplot showing enhancer responsiveness for aTSSs of genes that are surrounded by 1 or 2 versus 5 or more enhancers ($n = 1,325$ and 139, respectively; Wilcoxon P value). Center line: median; limits: interquartile range; whiskers: 10th and 90th percentiles. (f) Heatmaps depicting enrichments for the most differentially enriched Gene Ontology (GO) categories and for defined sets of transcription factors among the 400 genes associated with the strongest or weakest eTSSs that contain exclusively TATA box, Inr, MTE, or DPE. (g) Berkeley *Drosophila* Genome Project (BDGP)³⁹ *in situ* embryo images for genes representing the GO categories most strongly enriched near weak eTSSs.

MTE, or DPE motifs (i.e., those that preferentially function with developmental enhancers¹⁰), we found that the most responsive eTSSs were enriched near genes coding for transcription factors, whereas weak ones were predominantly near genes for cell-type-specific enzymes (Fig. 4f). For example, CG8560, CG16749, and CG14528 are all annotated as peptidases and are expressed in midgut and/or yolk (Fig. 4g).

This suggests that highly responsive non-housekeeping core promoters might regulate genes that require rapid induction (e.g., transcription factors), whereas weakly responsive ones could be employed at genes with potentially lower transcription kinetics (e.g., enzymes; Fig. 4f,g). Together, these results suggest that core promoters with different levels of enhancer responsiveness are employed for the transcription of genes with different functions and different regulatory characteristics.

The DNA sequence predicts enhancer responsiveness

We next investigated whether sequences that respond very differently to developmental enhancers have recognizable features that can predict enhancer responsiveness. We binned eTSSs according to their

responsiveness and visualized the nucleotide preferences, using the +1 positions of the eTSSs as anchor points. The resemblance with the established Inr motif²⁰ correlated with responsiveness and was higher for strongly responsive sequences than for weaker ones (Fig. 5a). In addition, the most responsive sequences had a preference for guanine (G) around the +30 position and a consensus sequence that resembled the downstream promoter element (DPE) motif²⁷. They also showed increasing information content around positions +15 to +20, where no known core promoter element resides, especially a prominent TC dinucleotide at position +17, just upstream of the motif-ten-element (MTE²⁸; Fig. 5a).

This observation prompted us to speculate whether the presence of core promoter motifs in the sequences (i.e., match quality/score or affinity) might determine the sequences' enhancer responsiveness. Indeed, eTSSs with higher responsiveness showed greater similarity to the canonical Inr, TATA box, and DPE motifs²⁰ (Fig. 5b), and the combined similarity scores for these motifs correlated with the responsiveness of eTSSs (Fig. 5c). This is also reflected by an enrichment of each of these motifs in eTSS sequences compared to

ARTICLES

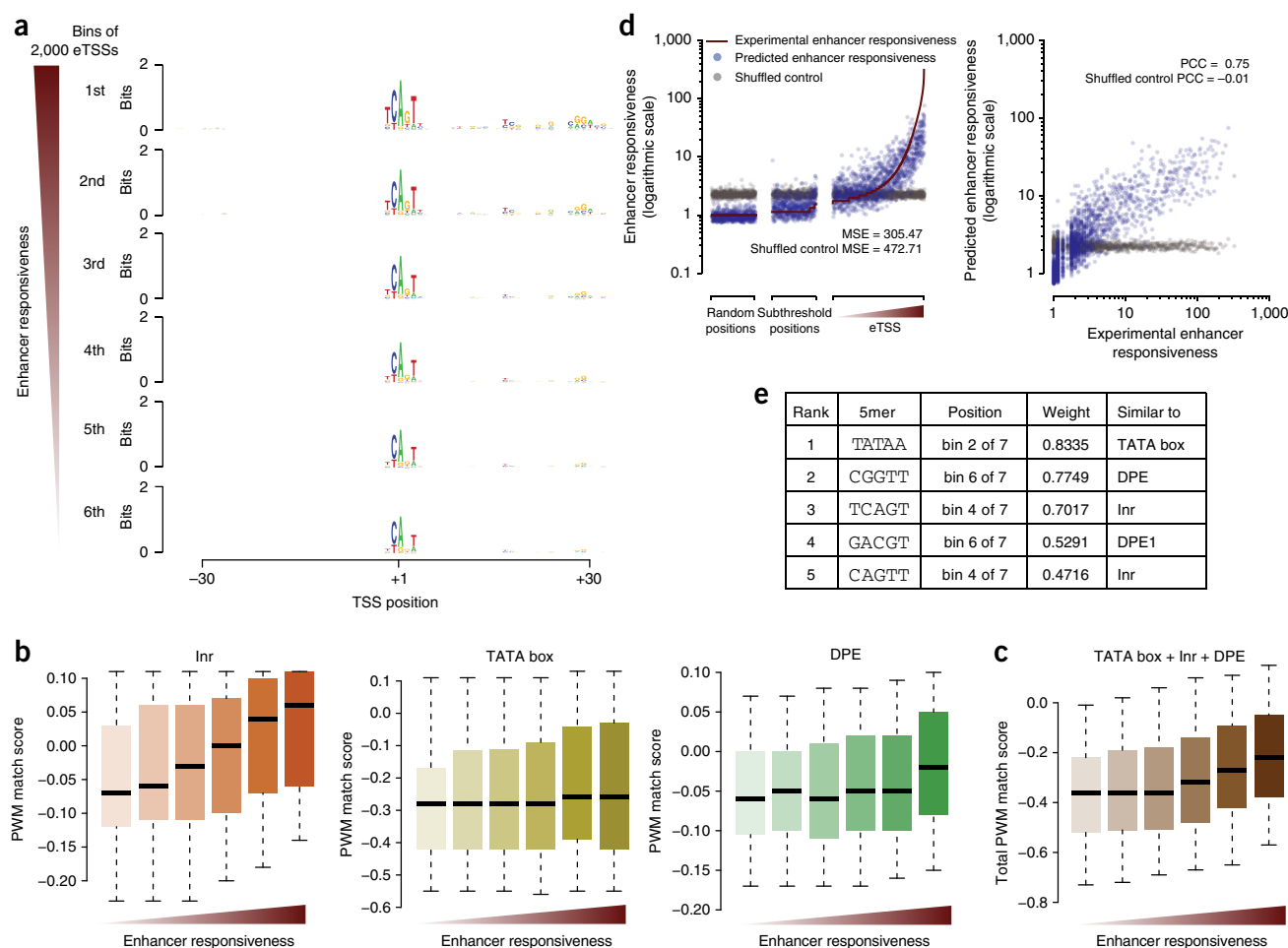


Figure 5 Candidate sequences are predictive of responsiveness to developmental enhancers. **(a)** Sequence logos summarizing position-specific nucleotide frequencies for eTSSs (bins of 2,000 sequences) ranked by decreasing enhancer responsiveness. **(b)** Position weight matrix (PWM) match scores for TATA box, Inr, and DPE motifs at eTSSs ranked by enhancer responsiveness **(b)**, and aggregate quality scores of all three motifs from **b** **(c)**. Center line: median; limits: interquartile range; whiskers: 5th and 95th percentiles. **(d)** Scatterplots of experimentally determined and predicted enhancer responsiveness for eTSSs, subthreshold positions, and random positions. Also included is predicted enhancer responsiveness after randomizing the assignment between the sequences and responsiveness (gray). MSE: mean square error. **(e)** Five most predictive 5mers, their positions (bin out of 7 bins along the sequence), and weights, as well as the most similar known core promoter motifs.

random sequences, which increases toward more responsive eTSSs (Supplementary Fig. 7b), even though, for example, TATA-box- and DPE-containing core promoters are typically found at different genes with distinct expression properties^{29,30}. Indeed, although both TATA box and DPE are increasingly enriched, they less frequently occur in the same eTSS, consistent with previous reports²⁰ and with their non-overlapping spatiotemporal expression patterns and biological functions^{29,30} (Supplementary Fig. 7c). For more responsive eTSSs, the Inr, TATA box, and DPE motifs also aligned increasingly well to their consensus positions at +1, -27, and +30, respectively⁴, and the distance to these consensus positions increased for less responsive eTSSs (Supplementary Fig. 7d,e; the absence of the TATA box in the sequence logo in Fig. 5a, despite its enrichment, stems from a reduced positional constraint³¹, see Supplementary Fig. 7e).

Moreover, the positional occurrence of specific 5mers relative to the eTSSs is predictive of the sequences' enhancer responsiveness by a linear model³² using fivefold cross-validation, leading to a PCC of 0.75 between the predicted and experimentally determined enhancer

responsiveness (Fig. 5d,e). The 5mers with the highest weights resemble known core promoter motifs and are specifically enriched at the canonical positions of these motifs (Fig. 5e). Together, these results suggest that enhancer responsiveness is determined by core promoter motif affinity (i.e., match quality) and positioning, providing a potential explanation for the positional preferences of these motifs.

DISCUSSION

The ability to efficiently convert enhancer activities into transcription initiation events is of central importance for differential gene expression. Here, we develop a functional reporter assay, STAP-seq, to quantitatively assess enhancer responsiveness systematically for millions of candidate fragments across entire genomes. Thousands of short fragments of genomic DNA are able to specifically initiate transcription to very different levels when activated by a single strong enhancer. For annotated gene TSSs, the different levels of enhancer responsiveness correlate with the function and the number of enhancers of the respective host gene, suggesting that strongly responsive

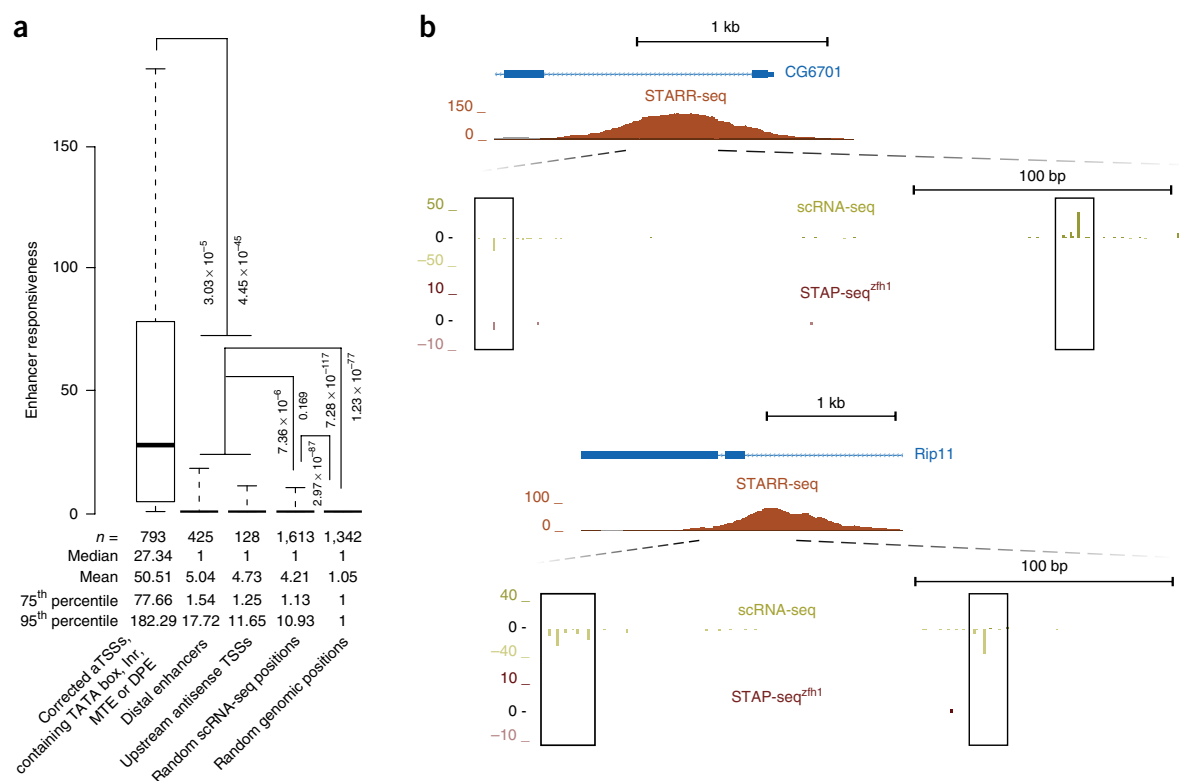


Figure 6 Positions of endogenous transcription initiation in developmental enhancers and upstream of aTSSs have weak sequence-intrinsic enhancer responsiveness. **(a)** Boxplot depicting enhancer responsiveness of positions that initiate transcription in S2 cells (≥ 5 scRNA-seq¹⁸ tags; left-most four boxes) or are randomly selected from the *D. melanogaster* genome (rightmost box, 'Random genomic positions'). 'Corrected aTSSs, containing TATA box, Inr, MTE or DPE', are position-corrected according to scRNA-seq¹⁸ as in **Figure 4a** and **Supplementary Figure 6b**. For 'Distal enhancers', we used STARR-seq enhancers¹⁴ that are more than 500 bp away from the nearest aTSS and for each enhancer considered the position with the highest scRNA-seq signal within ± 250 bp around the STARR-seq peak summit on either strand (disregarding enhancers for which this signal was below 5 tags). For 'Upstream antisense TSSs', we considered the position with the highest scRNA-seq signal upstream and antisense of aTSSs until the 3' end or—for divergent gene pairs—until 500 bp upstream of the 5' end (aTSS) of the next gene. 'Random scRNA-seq positions' are aTSS- and enhancer-distal and not closely spaced with respect to each other. Also shown are *P* values via one-sided Wilcoxon's rank-sum test between the categories. Center line: median; limits: interquartile range; whiskers: 5th and 95th percentiles. **(b)** UCSC Genome browser screenshots exemplifying representative loci of endogenous transcription initiation within enhancers as measured by scRNA-seq¹⁸ that have only weak STAP-seq signals.

core promoters might be required to reach high transcription rates, whereas those weakly responsive might serve to limit transcription and thus enhancer additivity.

Our observations of both a continuum of responsiveness and its degree of variation across three orders of magnitude suggest that enhancer responsiveness is an important measure to characterize and classify transcriptional regulatory elements. The continuum of activity suggests that many sequences can initiate transcription at very low levels (at or below the thresholds used here) when brought into the vicinity of strong enhancers. This could explain recent observations that enhancers and positions upstream of active aTSSs can be sites of transcriptional initiation^{15,16,33–38}. When we measured the sequence-intrinsic enhancer responsiveness of these positions with STAP-seq, we found it slightly higher than at control positions (random positions with endogenous initiation and arbitrarily chosen genomic positions), but substantially weaker than at positions within bona fide core promoters (**Fig. 6**). This finding suggests that in the vicinity of strong enhancers, accessible DNA might unavoidably initiate transcription, preferentially at sites of (degenerate) core promoter motifs, even if the respective DNA sequence is responsive to the enhancer only at the level of random sequences.

Our results further suggest that autonomously active promoters might consist of an enhancer-responsive core promoter and a TSS-proximal or TSS-overlapping enhancer. Whereas STAP-seq^{ctrl} had generally only a few tags, consistent with low basal activities of core promoters (**Supplementary Fig. 8**), the genomic positions of candidates with the highest activity in STAP-seq^{ctrl} frequently overlapped those of enhancers, predominantly housekeeping enhancers, suggesting that such autonomous promoter activities stem from proximal enhancers (**Supplementary Fig. 9**). This provides a simple explanation for the previously observed similarity between promoters and enhancers^{5,6} in that both have enhancer functionality yet differ in the presence of strongly responsive core promoters that support productive transcription.

Our results could explain the source of transcription initiation within enhancers and suggest how core promoters and proximal enhancers can form autonomously functioning promoters. Even though high enhancer activity and enhancer responsiveness can co-occur within a given DNA fragment, the two functions are generally uncoupled, re-emphasizing the difference and importance of the two key types of transcription regulatory elements and the functionalities they encode. STAP-seq will prove useful to assess enhancer responsiveness

ARTICLES

genome-wide and select the most responsive sequences, which should allow the highly efficient expression of transgenes, potentially beyond what is currently possible. STAP-seq will also be useful for studying the mechanisms of transcriptional initiation and its regulation, questions of fundamental importance especially today, when the key role of transcriptional regulation during development, evolution, and disease is becoming exceedingly clear.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the [online version of the paper](#).

Accession codes. GEO: [GSE78886](#). Vectors are available from Addgene (http://www.addgene.org/Alexander_Stark).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank L. Cochella and members of the Stark group for comments on the manuscript and Life Science Editors (<http://lifescienceeditors.com>) for editorial support. We are grateful to P. Heine and E. Jans (MaxCyte) for help setting up efficient plasmid transfection. Deep sequencing was performed at the Vienna BioCenter Core Facilities GmbH (VBCF) Next-Generation Sequencing Unit (<http://vbcbf.ac.at>). The Stark group is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 647320) and by the Austrian Science Fund (FWF, F4303-B09). Basic research at the IMP is supported by Boehringer Ingelheim GmbH and the Austrian Research Promotion Agency (FFG).

AUTHOR CONTRIBUTIONS

C.D.A., M.A.Z., and A.S. conceived the project. C.D.A., M.P., and M.R. performed the experiments with the help of K.S., and M.A.Z. the computational analyses. T.K. performed the k-mer based predictions. C.D.A., M.A.Z., and A.S. wrote the manuscript. A.S. supervised the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

- Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β -globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).
- Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–286 (2014).
- Roeder, R.G. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.* **21**, 327–335 (1996).
- Kadonaga, J.T. Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip. Rev. Dev. Biol.* **1**, 40–51 (2012).
- Core, L.J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–1320 (2014).
- Kim, T.-K. & Shiekhattar, R. Architectural and functional commonalities between enhancers and promoters. *Cell* **162**, 948–959 (2015).
- Spitz, F. & Furlong, E.E.M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).
- Kvon, E.Z. *et al.* Genome-scale functional characterization of *Drosophila* developmental enhancers *in vivo*. *Nature* **512**, 91–95 (2014).
- Juven-Gershon, T., Cheng, S. & Kadonaga, J.T. Rational design of a super core promoter that enhances gene expression. *Nat. Methods* **3**, 917–922 (2006).
- Zabidi, M.A. *et al.* Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556–559 (2015).
- Ede, C., Chen, X., Lin, M.-Y. & Chen, Y.Y. Quantitative analyses of core promoters enable precise engineering of regulated gene expression in mammalian cells. *ACS Synth. Biol.* **5**, 395–404 (2016).
- Lubliner, S. *et al.* Core promoter sequence in yeast is a major determinant of expression level. *Genome Res.* **25**, 1008–1017 (2015).
- Patwardhan, R.P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).
- Arnold, C.D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
- Duttke, S.H.C. *et al.* Perspectives on unidirectional versus divergent transcription. *Mol. Cell* **60**, 348–349 (2015).
- Andersson, R. *et al.* Human gene promoters are intrinsically bidirectional. *Mol. Cell* **60**, 346–347 (2015).
- Gu, W. *et al.* CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as *C. elegans* piRNA precursors. *Cell* **151**, 1488–1500 (2012).
- Nechaev, S. *et al.* Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. *Science* **327**, 335–338 (2010).
- Ni, T. *et al.* A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nat. Methods* **7**, 521–527 (2010).
- Ohler, U., Liao, G.-C., Niemann, H. & Rubin, G.M. Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol.* **3**, RESEARCH0087 (2002).
- Saito, K. *et al.* A regulatory circuit for piwi by the large Maf gene traffic jam in *Drosophila*. *Nature* **461**, 1296–1299 (2009).
- Sienski, G., Dönertas, D. & Brennecke, J. Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. *Cell* **151**, 964–980 (2012).
- Core, L.J. *et al.* Defining the status of RNA polymerase at promoters. *Cell Rep.* **2**, 1025–1035 (2012).
- Pfeiffer, B.D. *et al.* Tools for neuroanatomy and neurogenetics in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **105**, 9715–9720 (2008).
- Adelman, K. & Lis, J.T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.* **13**, 720–731 (2012).
- modENCODE Consortium. *et al.* Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**, 1787–1797 (2010).
- Burke, T.W. & Kadonaga, J.T. *Drosophila* TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev.* **10**, 711–724 (1996).
- Lim, C.Y. *et al.* The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev.* **18**, 1606–1617 (2004).
- Zeitlinger, J. *et al.* RNA polymerase stalling at developmental control genes in the *Drosophila melanogaster* embryo. *Nat. Genet.* **39**, 1512–1516 (2007).
- Engström, P.G., Ho Sui, S.J., Drivenes, O., Becker, T.S. & Lenhard, B. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res.* **17**, 1898–1908 (2007).
- Ponjavic, J. *et al.* Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. *Genome Biol.* **7**, R78 (2006).
- Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).
- Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
- De Santa, F. *et al.* A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol.* **8**, e1000384 (2010).
- Lam, M.T.Y., Li, W., Rosenfeld, M.G. & Glass, C.K. Enhancer RNAs and regulated transcriptional programs. *Trends Biochem. Sci.* **39**, 170–182 (2014).
- Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
- Scruggs, B.S. *et al.* Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. *Mol. Cell* **58**, 1101–1112 (2015).
- Hah, N. *et al.* A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell* **145**, 622–634 (2011).
- Tomancak, P. *et al.* Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.* **3**, RESEARCH0088 (2002).

ONLINE METHODS

STAP-seq screening vector. For STAP-seq in *Drosophila* cells we constructed a screening vector based on the pGL3-Promoter backbone (Promega; cat. no. E1751) by replacing the sequence between BglII and FseI with the following sequence, containing a ccdB suicide gene flanked by homology arms (used for cloning the candidates during library generation), an intron (mhc16), an ORF (truncated sgGFP, Qbiogene, Inc.), followed by the pGL3's SV40 late polyA-signal. The full sequence is available at <http://www.addgene.org>. The enhancers were cloned between the KpnI and BglII sites (for coordinates and sequences of the enhancers please see **Supplementary Table 1**). The control screens were performed with the STAP-seq vector not harboring any enhancer.

STAP-seq library generation. Genomic DNA (genome-wide libraries) or BAC DNA (focused libraries; **Supplementary Table 2**) was isolated as described previously^{10,14}. The DNA was sheared by sonication (Covaris S220) and DNA fragments (100– to 250-bp length) were size-selected using a 1% agarose gel. Illumina NEBnext Multiplexing Adaptors (New England BioLabs (NEB); cat. no. E7335 or E7500) were ligated to 1 µg of size-selected DNA fragments using NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB; cat. no. E7645L) following the manufacturer's instructions, except the final PCR amplification step. Ten PCR reactions (98 °C for 45 seconds (s); followed by 10 cycles of 98 °C for 15 s, 65 °C for 30 s, 72 °C for 10 s) with 1 µl adaptor ligated DNA as template were performed, using KAPA HiFi Hot Start Ready Mix (KAPA Biosystems; cat. no. KK2602) and primers (fw: TAGAGCATGCACCGACA CTCTTTCCCTACACGACGCTCTTCCGATCT and rev: GGCCGAATTCG TCGAGTGAAGTTCAGACGTGTGCTCTTCCGATCT), which add a specific 15-nt extension to both adapters for directional cloning using recombination (Clontech In-Fusion HD; cat. no. 639650). Each five PCR reactions were pooled, purified, and size selected with Agencourt AMPureXP DNA beads (ratio beads/PCR 1.25; cat. no. A63881), followed by column purification (QIAquick PCR purification kit; cat. no. 28106). Cloning of the fragments into the vector was performed as described previously¹⁴.

STAP-seq spike-in controls. In order to control for transfection efficiency and to normalize all STAP-seq screens we used spike-in controls. We generated four STAP-seq spike-in control plasmids that are driven by the *zfh1* enhancer and harbor a single sequence each, which were derived from the *Drosophila pseudoobscura* orthologs of *even skipped* (*eve*), *CG32369*, and two alternative TSSs from *u-shaped* (*ush*). Reads derived from those sequences map uniquely to the *Drosophila pseudoobscura* (dp3 assembly), but not to the *D. melanogaster* genome. We cloned the sequences into the STAP-seq vector using the same strategy as we used for library generation (see above; **Supplementary Table 3**). A mix of the four spike-in control plasmids was added to the genome-wide STAP-seq libraries before transfection at a final dilution of 1:1,000,000. For the focused libraries (BAC) we only used the *eve* spike-in control plasmid at a dilution of 1:100,000.

Cell culture and transfection. S2 cells were obtained from Life Technologies and cultured as described previously¹⁴. Transfection of the STAP-seq libraries was performed with 1×10^8 (focused libraries) or 1.2×10^9 cells (genome-wide libraries) at 70–80% confluence using the MaxCyte STX Scalable Transfection System. Cells were transfected at a density of 1×10^9 cells per milliliter in MaxCyte HyClone buffer using OC-100 or OC-400 processing assemblies and 50 µg library per milliliter of cells. S2 were pulsed with the pre-set program Optimization 1. Cells were transferred to a cell culture flask and mixed with 10% DNaseI (2,000 U/ml) and incubated for 30 min at 27 °C, before resuspension in full medium. Cells were incubated after electroporation for 24 h before RNA isolation. OSCs²¹ originally isolated by the M. Siomi laboratory (Keio University School of Medicine) were obtained from the laboratory of J. Brennecke (Institute of Molecular Biotechnology (IMBA)) and cultured as described previously¹⁴. Transfection of the focused library was performed using all cells from a 70–80% confluent square dish (24.5 cm × 24.5 cm) in an OC-400 processing assembly in 400 µl MaxCyte HyClone buffer mixed 1:1 with the OSC culture medium without supplements and 20 µg of library (pre-set program Optimization 5). Cells were transferred to a cell culture flask and mixed with 10% DNaseI (2,000 U/ml) and incubated for 30 min at 27 °C, before resuspension in full medium. The cells were plated on a square dish

(24.5 cm × 24.5 cm) for 24 h after electroporation. For the focused screens, we performed three STAP-seq^{zfh1} (PCC ≥ 0.97 with each other), and two each for STAP-seq^{egl} (PCC = 0.95), STAP-seq^{ham} (0.98), STAP-seq^{ssp3} (0.87), STAP-seq^{ncm} (0.78), and STAP-seq^{ij} (0.98). All cell lines used are checked for mycoplasma contamination on a regular basis.

STAP-seq RNA processing. 24 h after electroporation total RNA was isolated followed by polyA⁺ RNA purification and DNaseI treatment, as described previously¹⁴. 10–20 µg (focused) or 200 µg (genome-wide) of DNaseI-treated RNA was incubated with calf intestinal alkaline phosphatase (CIP; NEB cat. no. M0290L). Per 1 µg RNA, 0.5 µl CIP was used. The reactions were cleaned up using Qiagen RNeasy MinElute reaction clean-up kit (cat. no. 74204) according to the manufacturer's protocol, adding beta-Mercaptoethanol to the RLT buffer. Subsequently all RNA was processed during all further reactions. The CIP-treated RNA was then incubated with 0.05 µl Tobacco Alkaline Phosphatase (TAP; Epicentre, discontinued, now available as Cap-Clip Acid Pyrophosphatase (cat. no. C-CC15011H) from CELLSERIP) per 1 µg RNA to remove the 5' cap of all 5'-capped RNA species. The reactions were cleaned up using Agencourt RNAClean XP (BeckmanCoulter, cat. no. A63987) at a ratio of 1.8 of beads to RNA. To the 5' ends of the TAP-treated RNA 10 µM RNA oligo-nucleotide (GUUCAGAGUUCUACAGUCCGACGAUCNNNNNNNN) was ligated per 1 µg RNA at 16 °C for 16 h using 0.2 µl T4 RNA Ligase 1 (ssRNA Ligase, NEB, cat. no. M0204L). The eight random nucleotides at the 3' end of the 5' RNA linker are used as a Unique-Molecular-Identifier (UMI) to count reporter mRNAs (see below), but also minimizes sequence preferences during the T4 RNA Ligase 1 reaction⁴⁰. The reactions were cleaned up using Agencourt RNAClean XP (BeckmanCoulter, cat. no. A63987) at a ratio of 1.0 of beads to RNA. First strand cDNA synthesis was performed with 1 µl of Invitrogen's SuperscriptIII (50 °C for 60 min, 70 °C for 15 min; cat. no. 18080085) using a reporter-RNA-specific primer (CAAACATCAATGTATCTTATCATG) for 2.5–5 µg of polyA⁺ RNA in 20 µl total volume. Five reactions were pooled and 1 µl of 10 mg/ml RNaseA was added (37 °C for 1 h) followed by bead purification (Agencourt AMPureXP DNA beads (ratio beads/RT reaction 1.8). We amplified the total amount of reporter cDNA obtained from reverse transcription (above) for Illumina sequencing. For the focused libraries we performed two PCR reactions using the KAPA real-time library amplification kit (KAPA Biosystems, cat. no. KK2702) according to the manufacturer's protocol. The genome-wide screens were amplified using KAPA HiFi Hot Start Ready Mix (KAPA Biosystems; cat. no. KK2602) in 32 PCRs. As forward primer we used AATGATACGGCGACCACCGAGATCTACACGTTTCAGAGTTCTACAGT CCGA and as reverse primer NEBnext Multiplex Oligos for Illumina (NEB; cat. no. E7335 or E7500). PCR products were purified with Agencourt AMPureXP DNA beads (ratio beads/PCR 1.25).

Illumina sequencing. All samples were sequenced by the VBCF's NGS unit on an Illumina HiSeq2500 platform, following manufacturer's protocol. All deep sequencing data are available at <http://www.starklab.org/> and are deposited in GEO.

Luciferase reporter assays. We replaced the SV40 promoter of the pGL3-promoter plasmid (Promega) by the candidate sequences (see **Supplementary Table 4** for coordinates and primers) between the BglII and SbfI restriction sites. As in STAP-seq the *zfh1* enhancer (inserted in the KpnI restriction site) was used to drive transcription from the candidates. To determine the basal activities of the candidates they were also cloned into the luciferase vector not harboring any enhancer. Individual constructs were tested by co-transfecting 100,000 cells with 95 ng of the respective pGL3 firefly construct and 5 ng of a *Renilla* control plasmid (driven by the *ubiquitin-63E* promoter) that is based on the pRL plasmid (Promega) using FuGENE HD Transfection Reagent (Promega; cat. no. E2312). Using the Promega Dual Luciferase Assay kit (cat. no. E1960), we measured luciferase activity at a Bio-Tek Synergy H1 fluorescence plate reader.

Candidate selection for Luciferase validation. The candidate TSSs were selected to not have basal activity (STAP-seq^{ctrl} ≤ 3 tags) but are active in STAP-seq^{zfh1} (≥ 5 tags). According to this rule we selected 30 positive regions across the entire range of strength (from rank 2 to rank 4,675; 19 aTSS and 11 eTSS, including 2 candidates that overlap exons). We in addition selected 12

aTSS for which we did not observe any STAP-seq signal and 4 genomic regions without any TSS annotation as negative controls.

STAP-seq NGS data processing. Paired-end STAP-seq reads were trimmed to 44 bp, with the first 8 bp as the unique molecular barcode identifier (UMI). The reads were mapped using the remaining 36 bp to dm3 and dp3 (for spike-in controls) genome assemblies using Bowtie⁴¹ version 0.12.9 as in ref. 10. The dp3 spike-in controls were selected from their dm3 orthologs that allows unambiguously unique mapping to dp3 with the same mapping parameter. For paired-end reads that are mapped to the same positions, we collapsed those that have identical UMIs as well as those for which the UMIs differed by 1 bp to ensure the counting of unique reporter mRNAs (we removed all mapped reads that had N's in their UMIs). Tag counts at each position represent the sum of the 5'-most position of collapsed fragments. For focused STAP-seq screens in S2, we subsampled all reads at 700,000 reads before mapping and removed fragments outside of the focused regions (see **Supplementary Table 2** for full list of BACs) from analysis (**Supplementary Table 5**). For analysis of STAP-seq^{zh1} versus STAP-seq^{ij}, we subsampled 300,000 of the mapped BAC fragments.

Genomic distribution. We assigned a unique annotation for each nucleotide in the genome via the following priority order: ± 5 , ± 10 , ± 50 bp around aTSS, CDS, 5'-untranslated region (UTR), 3'-UTR, intron, intergenic region. We then assigned each eTSSs to one of these categories by the annotation of the +1 position of the eTSSs.

eTSSs calling. We selected candidate positions as those with at least 5 tag counts in the STAP-seq^{zh1} experiment. At 5 tag counts >98% of eTSSs could be recovered in the replicates. For normalization, we used the tag counts at +1 position of the dp3 *eve* (focused STAP-seq^{zh1}) or together with the tag counts from the two most prominent initiation positions from the *ush-1* spike-in TSSs (genome-wide STAP-seq^{zh1}; **Supplementary Table 3**). We used these tag counts as the probability P in determining the binomial distribution, and determined the enhancer responsiveness of each candidate position as the corrected ratio of STAP-seq^{zh1}/STAP-seq^{ctrl} tag using a pseudo count of 1. We considered only positions that were more than 1.5-fold enriched in STAP-seq^{zh1} over STAP-seq^{ctrl} with P value ≤ 0.05 , before merging those that were within ± 10 bp around each other. We determined within each of these regions the position with the highest enhancer responsiveness, with which we ranked the eTSSs, as the +1 positions.

Metagene plots. We obtained raw reads from scRNA-seq obtained from ref. 18 (GSM463298), and mapped the 38 bp reads as 36-nt reads using Bowtie⁴¹ with the following parameters: -p 4 -q -v 3 -m 1-best-strata -quiet. PEAT data were obtained from ref. 19 (https://ohlerlab.mdc-berlin.de/research/Download_The_Data_97/). Raw tag counts were directly derived from the strand-specific log₂-transformed bigwig files. To test the accuracy of identification of the highest position as the +1 position in STAP-seq (**Supplementary Fig. 2c**), we bootstrapped the STAP-seq^{zh1} tag counts, normalized to the spike-ins, at each position 100 times, calculated the mean, and plotted the 5th, 50th, and 95th percentiles.

Sequence logo. We downloaded genomic sequence using fastacmd version 2.2.10, and used WebLogo version 2.8 (ref. 42).

Scatterplots. For tag count scatterplots, we considered only positions that have at least 3 tag counts. For enhancer responsiveness scatterplots, we calculated the corrected ratio of STAP-seq^{zh1}/STAP-seq^{ctrl} tag using a pseudo count of 1, and computed the log₂ values. We added a pseudo count of 0.7071 to any zero values, and a dithering factor of 0.1 for nonzero coordinates that are plotted more than once. For comparison between STAP-seq and GRO-seq, we took the sum of STAP-seq tag counts within ± 5 bp of aTSSs, or the sum of GRO-seq fragment 5' positions in a window spanning 101 bp downstream. To depict variation of STAP-seq, we plotted the sample s.d. from three focused STAP-seq^{zh1} replicates at eTSS positions, called on the combined three replicates, on a scatterplot on two of the replicates with no dithering. For scatterplot of enhancer responsiveness in focused screens, we summed up the tag counts

of positions ± 50 bp around aTSSs for the induced and basal activities, and calculated the enhancer responsiveness as above.

Luciferase assay analysis. We first normalized firefly over *Renilla* luciferase values for each of the three independent transfections per construct individually and then calculated the mean and s.d. for these normalized values. Finally, we used these means and s.d. to calculate the fold change of the *zfh1* enhancer-driven luciferase signals over the enhancer-less control.

STAP-seq correlation heatmap. For all pairs of enhancers, we computed pairwise Pearson correlation coefficients (PCCs) between the respective STAP-seq tag counts at positions that are covered by at least 3 tag counts in either enhancer. We performed hierarchical clustering (complete linkage) in R, directly using the correlation values as similarities as previously¹⁰.

Cell-type specificity analysis. We obtained GRO-seq raw reads performed in S2 (ref. 23) (GSM577244) and OSCs²² (GSM1027403) and mapped them as 36-nt reads using Bowtie⁴¹ with the following parameters: -p 4 -q -v 3 -m 1-best-strata-quiet. We considered all positions that were ≥ 3 tag counts in either STAP-seq^{zh1} S2 or STAP-seq^{ij} OSC and determined the sum of GRO-seq fragment 5' positions in a window spanning 101 bp downstream. We classified a position to be exclusively active in a cell type if it had GRO-seq start fragments at least 15 in a cell type but not in the other. As this gives more OSC active TSSs, we considered only OSC TSSs with the highest GRO-seq signal to be the same number as in S2 cells. For the cumulative distribution plot, we randomized "all" positions and took the same number as S2 active positions. We performed two-sided Kolmogorov-Smirnov between S2 and OSC active positions.

Global comparison of STAP-seq^{zh1} and other endogenous methods. We used our previously published RNA-seq¹⁴. We obtained raw reads from two CAGE replicates in S2 (SRX142946 and SRX144189, modENCODE submission 5331)²⁶ and mapped the 27-nt reads using Bowtie⁴¹ with the following parameters: -p 4 -q -v 2 -m 1-best-strata -quiet. As CAGE reads often carries nontemplate G nucleotide at the 5' end⁴³, we shifted the start position of the fragment 1 bp downstream when the first position mapped to a G in the genome, and combined the two replicates. We computed Pearson correlation coefficients (PCCs) only for aTSSs that in either experiment had the following cutoffs: around ± 5 bp aTSSs with sum of signal ≥ 3 for STAP-seq^{zh1}, ≥ 5 for scRNA and CAGE, or ≥ 15 for 101 bp downstream of aTSSs for GRO-seq. For comparisons with RNA-seq, we took the aTSS with the highest signal from each gene, and considered only genes with ≥ 3 RPKM. We used our core promoter motif counts and calculated their enrichment as previously¹⁰. To specifically compare STAP-seq^{zh1} and scRNA-seq¹⁸ (**Fig. 3d-f** and **Supplementary Fig. 5**), an aTSSs was considered to be detected if within ± 5 bp of either side it has ≥ 3 STAP-seq^{zh1} tag counts, or ≥ 5 by scRNA-seq tag counts in total. We calculated the binomial distribution of DHS-seq¹⁴ within ± 250 -bp window and called it to be closed for a P value > 0.05 . We obtained bigwig RAMPAGE data⁴⁴ (GSE36212), and determined the mean RAMPAGE signal at shifted controls 200 bp upstream of aTSSs to be 10, and therefore used this value as cutoff. For aTSS detection in the focused screens, an aTSSs was considered to be detected in a housekeeping screen if it had ≥ 3 tag counts in either *ncm* or *ssp3* screens within ± 50 bp to account for broad initiation of housekeeping genes, or if it had 1 or 2 tag counts to be considered to be detected at subthreshold level.

Density plot of enhancer responsiveness. Kernel density was calculated using density function in R from the log₂ values of the enhancer responsiveness. We first calculated density parameters using aTSSs, for which we obtained bandwidths of 0.0903 and 0.09597, respectively, for replicates 1 and 2. We plotted the density using polygon function, and added pseudo-positions at the ends of the estimated kernel density: x coordinates corresponded to the estimated end positions, and y coordinates were from the minimum y coordinates from the estimate.

Number of enhancer per gene analysis. We assigned a gene to an eTSS if the eTSS lies within ± 20 bp from the 5' end of a transcript from that gene.

We removed eTSSs and genes that are assigned more than once. We used annotation from *D. melanogaster* FlyBase release 5.50. As a gene could have multiple aTSS, we only considered aTSS with the highest enhancer responsiveness. To count the number of enhancers per gene, we used our previous assignment¹⁰. Briefly, developmental enhancers were assigned to a gene provided that they fall anywhere within 5 kb upstream from the TSS to 2 kb downstream from the gene end of the longest isoform.

Correction of aTSS +1 positions. For “corrected aTSSs”, we realigned the +1 positions to be the highest scRNA position that is at least 5 within a window of up to 20 bp on either side of the aTSS. For aTSSs that contain TATA box, Inr, MTE or DPE, we considered only “corrected aTSSs” that contain either of these motifs but not TCT, DRE, Ohler motifs 1, 5, 6, and 7 as previously¹⁰.

Gene Ontology (GO) and TF enrichment analysis. We used aTSS-to-eTSS assignments as above. We ranked the genes based on their enhancer responsiveness and divided them into two categories: the top and bottom 1,000 genes. We assessed whether genes assigned to an eTSS were enriched for any GO terms⁴⁵ by calculating hypergeometric *P* values and enrichment for all terms. For all terms that were enriched more than twofold among the top and bottom gene sets and were present at least 4 times in all assigned eTSSs, we sorted for their enrichment and counts. For each category, we calculated $\log_{10}(P \text{ value under-representation}) - \log_{10}(P \text{ value over-representation})$, and sorted the terms in a descending order of difference between values from the classes. The color intensity of the heatmaps represents $\log_{10}(P \text{ value under-representation}) - \log_{10}(P \text{ value over-representation})$. To investigate the relationship between eTSSs that contain TATA box, Inr, MTE, or DPE, and their biological function, we considered only eTSSs that have any of these motifs but not TCT, DRE, Ohler motifs 1, 5, 6, and 7, terms that were enriched more than 1.5-fold among the top and bottom gene sets and were present at least twice in all assigned eTSSs. For TF enrichment analysis, we used our curated TF and cofactor lists⁴⁶ as well as sets of factors annotated by the *Drosophila* Transcription Factor Database (<http://www.flytf.org/>)⁴⁷ (“experimentally verified site-specific TFs”, “equivalent to release v1—trusted TFs”, and “proteins involved in chromatin-related processes”).

Analysis of enhancer responsiveness and length of fragments. For +1 of eTSS positions, as well as STAP-seq^{zh1} positions that are $\geq 1-5$ tag counts, we intersected with the sequenced input STAP-seq^{zh1} fragments. We considered only fragments that intersect on the same strand and cover ± 30 bp around the positions, and determined the longest ones. We also divided the fragments into 4 groups around the median: 80 to 140 bp, 141 to 190 bp, 191 to 240 bp, and 241 to 300 bp, and determined the eTSSs that obey the same intersection rule as above.

Core promoter element enrichment and position heatmaps. We scanned for motif occurrences using MAST from the MEME suite⁴⁸ (version 4.9.0) and used parameters that ensured specificity and sensitivity (for enrichment heatmap) or sensitivity only (for position heatmaps) for each motif as previously¹⁰. For enrichment heatmap, we calculated enrichment and hypergeometric distribution, and considered an enrichment to be significant for *P* value ≤ 0.05 .

Core promoter element quality and position boxplots. We scanned for motif occurrences as above to obtain motif match scores, and added the individual match scores to derive the aggregate match scores. We determined the median position for the core promoter element from these positions, and computed the deviation of each sequence for the position boxplot.

k-mer-based prediction of enhancer responsiveness. We considered all 13,218 eTSSs, and also included two categories of control positions with no STAP-seq^{ctrl} signal: 6,405 each for subthreshold positions with STAP-seq^{zh1} less than 5, and random positions with no STAP-seq^{zh1} signal (from random positions for motif enrichment analysis). This selection covers a large span of responsiveness and made the sequence-based discrimination tractable. We counted the occurrences of 5mers in seven equally spaced sectors around the +1 position: -50 to -37, -36 to -23, -22 to -9, -8 to +6, +7 to +20, +21 to +34, +35 to +48. To simplify the learning, we de-noised the data by binning every

ten consecutive sequences: the k-mer counts were summed in each bin and a median of responsiveness was considered as the responsiveness of the bin, resulting in 10× smaller data set and larger k-mer counts. We $\log_{10}$ -transformed the responsiveness to decrease their dynamic range and exponential growth, and used an L_1 -regularized linear model (ref. 32, implemented in the scikit-learn Python package⁴⁹). We kept the regularization coefficient α fixed at 10^{-3} , and estimated the mean squared error and correlation coefficient of the predicted responsiveness using a fivefold cross-validation.

Enhancer responsiveness analysis at positions of endogenous distal transcription. For distal enhancers, we defined the position within ± 250 bp of distal enhancer summits as previously defined¹⁰ with the highest scRNA signal ≥ 5 . For antisense upstream TSSs, we looked only upstream of positions of active aTSSs (total scRNA signal within ± 5 bp either side of the +1 position to be ≥ 5) of the longest 5' isoform of each gene, and considered the highest positions that were not in the gene body of another gene, or 500 bp upstream of another aTSSs, and had scRNA signal ≥ 5 . For random scRNA positions, we first selected positions with scRNA signal ≥ 5 that did not overlap 500 bp of all aTSSs, developmental enhancers, as well as antisense upstream TSSs on both strands, merged those that were within 10 bp of each other, and considered only the highest positions. For random genomic positions, we considered 2,000 positions and controlled for their chromosome distributions. For all of the positions defined as above, we removed those that have gaps in the input coverage in the same strand, and calculated the sum of STAP-seq^{zh1} and STAP-seq^{ctrl} within ± 5 bp and calculated enhancer responsiveness as above (see eTSSs calling). We calculated the *P* value via one-sided Wilcoxon's rank-sum test.

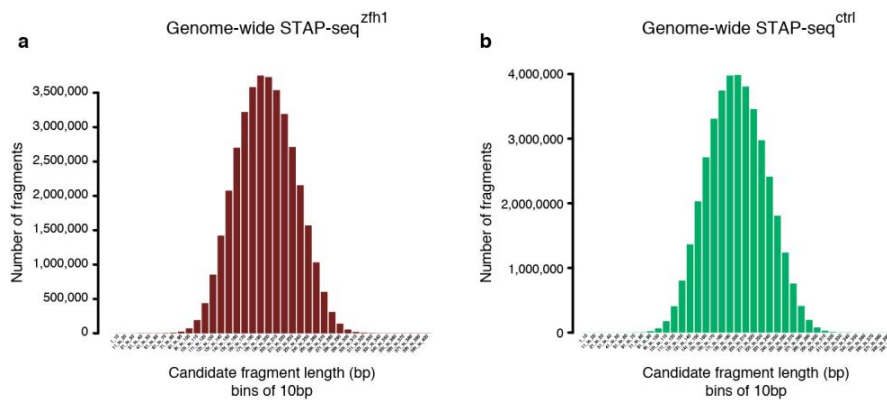
Generation of random TSSs (motif enrichment). We aimed the random positions to be the same number of all aTSSs, considered positions that do not overlap 1 kb surrounding aTSS and eTSSs. We further merged positions that are within ± 50 bp of each other, recentered the position, and removed positions with undefined nucleotides (Ns) within ± 50 bp.

Generation of random TSSs (scatter and density plots). We generated random positions that do not overlap ± 50 bp of eTSS and aTSSs, aimed to be the same number of all aTSSs.

Coordinate intersections. We performed genomic coordinate intersections using the BEDTools suite⁵⁰ version 2.17.0.

Statistics. We performed all statistical calculations and created graphical displays with R⁵¹.

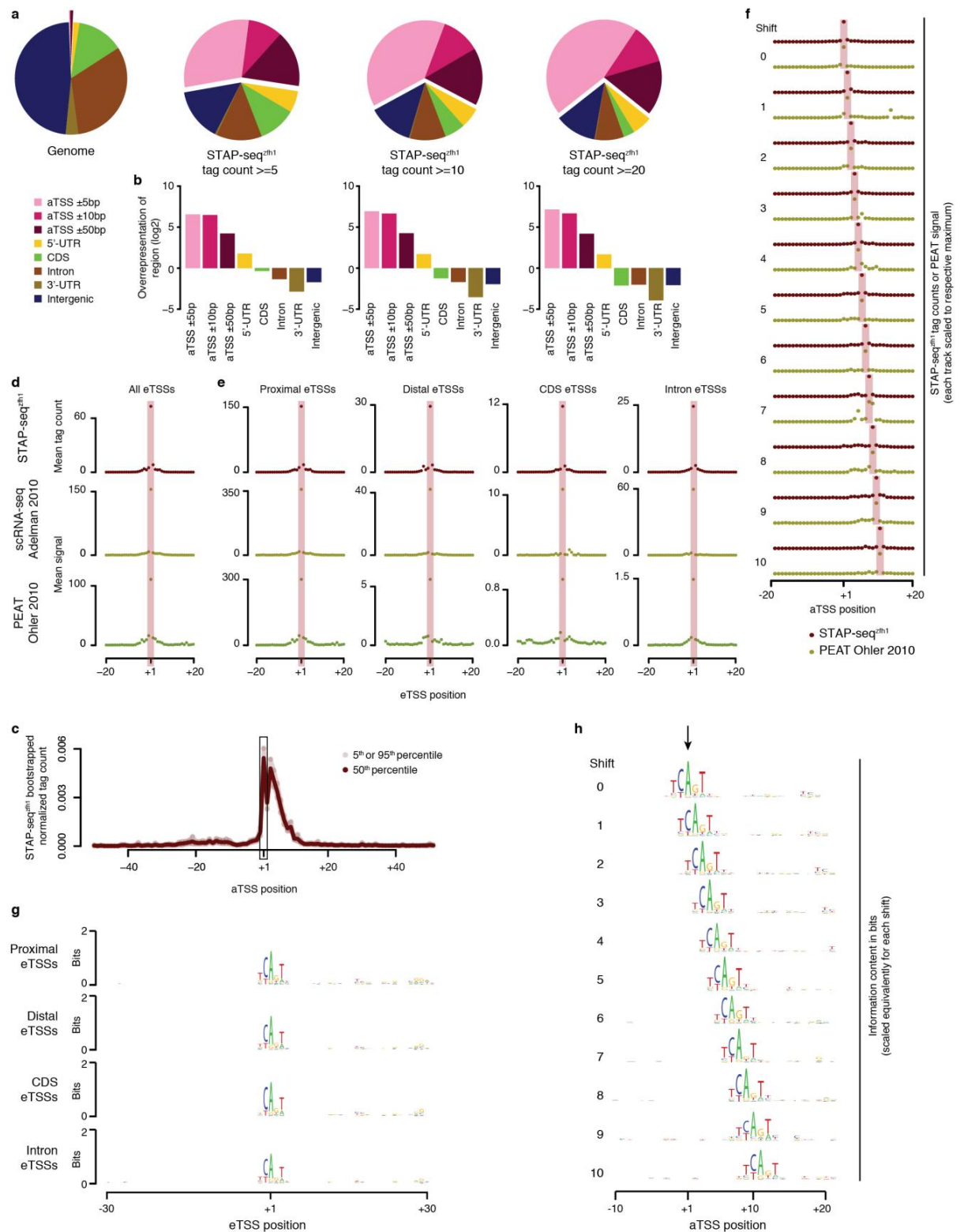
40. Jayaprakash, A.D., Jabado, O., Brown, B.D. & Sachidanandam, R. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Res.* **39**, e141 (2011).
41. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
42. Crooks, G.E., Hon, G., Chandonia, J.-M. & Brenner, S.E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
43. Kodzius, R. *et al.* CAGE: cap analysis of gene expression. *Nat. Methods* **3**, 211–222 (2006).
44. Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T.R. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res.* **23**, 169–180 (2013).
45. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
46. Stampfel, G. *et al.* Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature* **528**, 147–151 (2015).
47. Adryan, B. & Teichmann, S.A. FlyTF: a systematic review of site-specific transcription factors in the fruit fly *Drosophila melanogaster*. *Bioinformatics* **22**, 1532–1533 (2006).
48. Bailey, T.L. & Gribskov, M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**, 48–54 (1998).
49. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
50. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
51. R Development Core Team. *R: a Language and Environment for Statistical Computing* (Vienna, Austria, 2012).



Supplementary Figure 1

Candidate fragment length distribution.

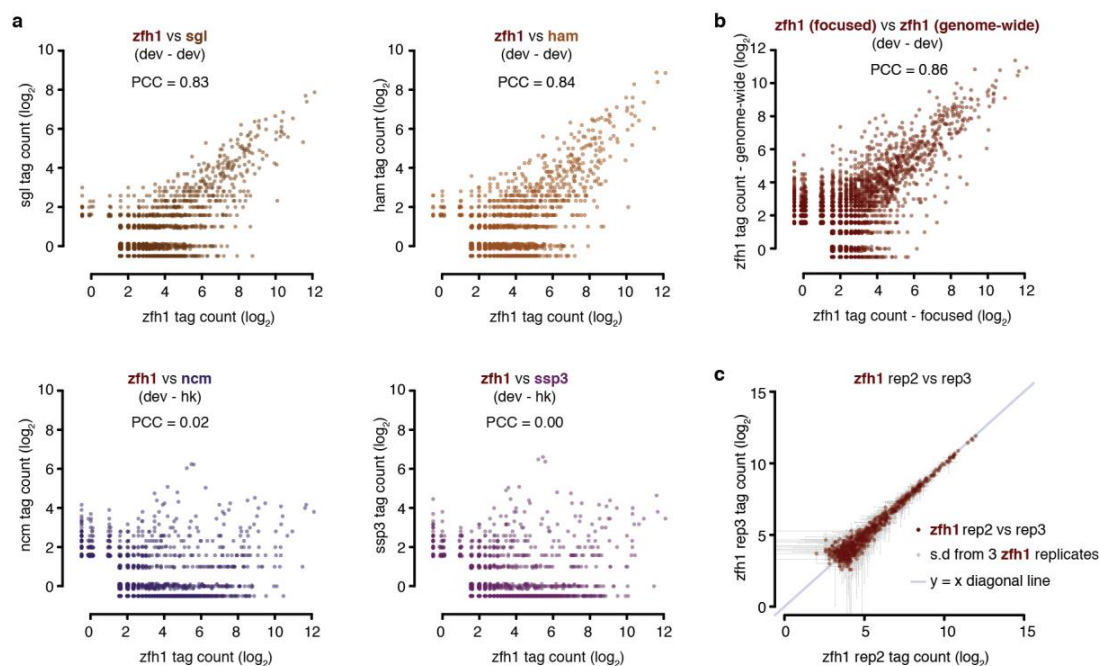
(a) Candidate fragment length distributions of STAP-seq^{zfh1} and (b) STAP-seq^{ctrl} input fragments in bins of 10bp (40 bins in total).



Supplementary Figure 2

Genomic distribution of eTSSs and agreement of eTSS positions with endogenous TSSs and established sequence motifs.

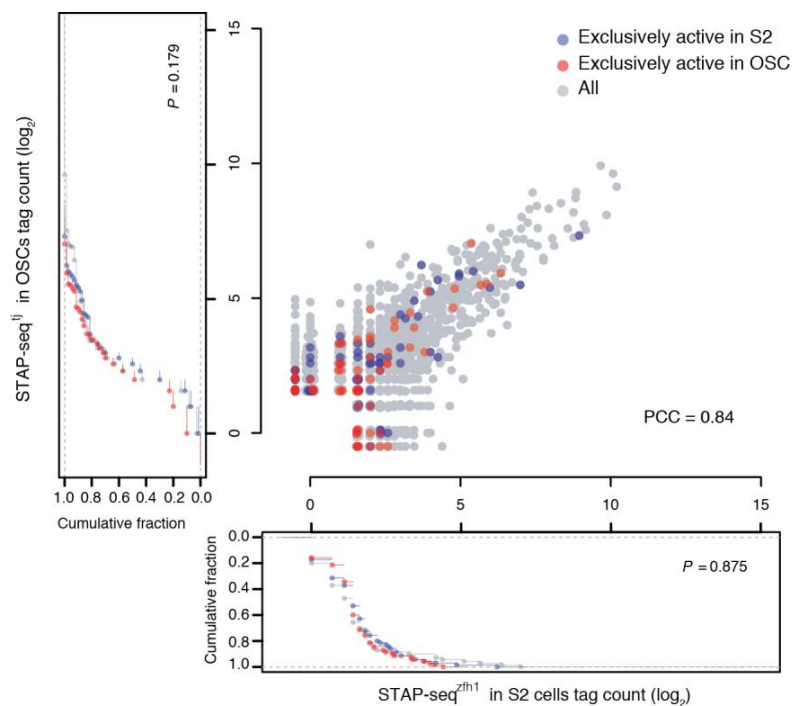
(a) Pie charts of the genomic distribution of all genomic sequences that initiate transcription with tag counts ≥ 5 , ≥ 10 , ≥ 20 , respectively, in response to the *zfh1* enhancer compared to the *D. melanogaster* genome (leftmost pie chart; the sector indicates aTSSs ± 50 bp that make up 0.32% of the genome). (b) Bar plots visualizing the enrichment of the regions from a, over the genome. (c) Metagene profile of average normalized STAP-seq^{zfh1} tag counts around aTSSs, including the 5th and 95th percentiles determined by bootstrapping. (d) Metagene profiles of STAP-seq^{zfh1}, scRNA-seq¹⁸ and PEAT¹⁹ at all eTSSs. (e) As in d, but specifically for proximal, distal, within coding DNA sequence (CDS) and intronic eTSSs. (f) Agreement of STAP-seq eTSSs and embryo-derived TSSs by PEAT¹⁹ that are shifted from aTSSs between 1 to 10 nucleotides (each row is scaled to the respective maximum; see **Figure 1e** for the equivalent comparison with scRNA-seq from S2 cells¹⁸). (g) Sequence-logos depicting position-specific nucleotide frequencies for eTSSs that are aTSS-proximal or -distal, within CDS, or intronic. (h) Sequence logos of eTSSs that coincide with aTSS (shift 0, top row, the +1 position of aTSSs is indicated by the arrow) or are proximally-misaligned by 1 to 10 base pairs (bps) from aTSSs (logos are moved closer together but not scaled differently relatively to each other).



Supplementary Figure 3

Reproducibility of STAP-seq^{zf1} and comparisons between STAP-seq screens with different enhancers.

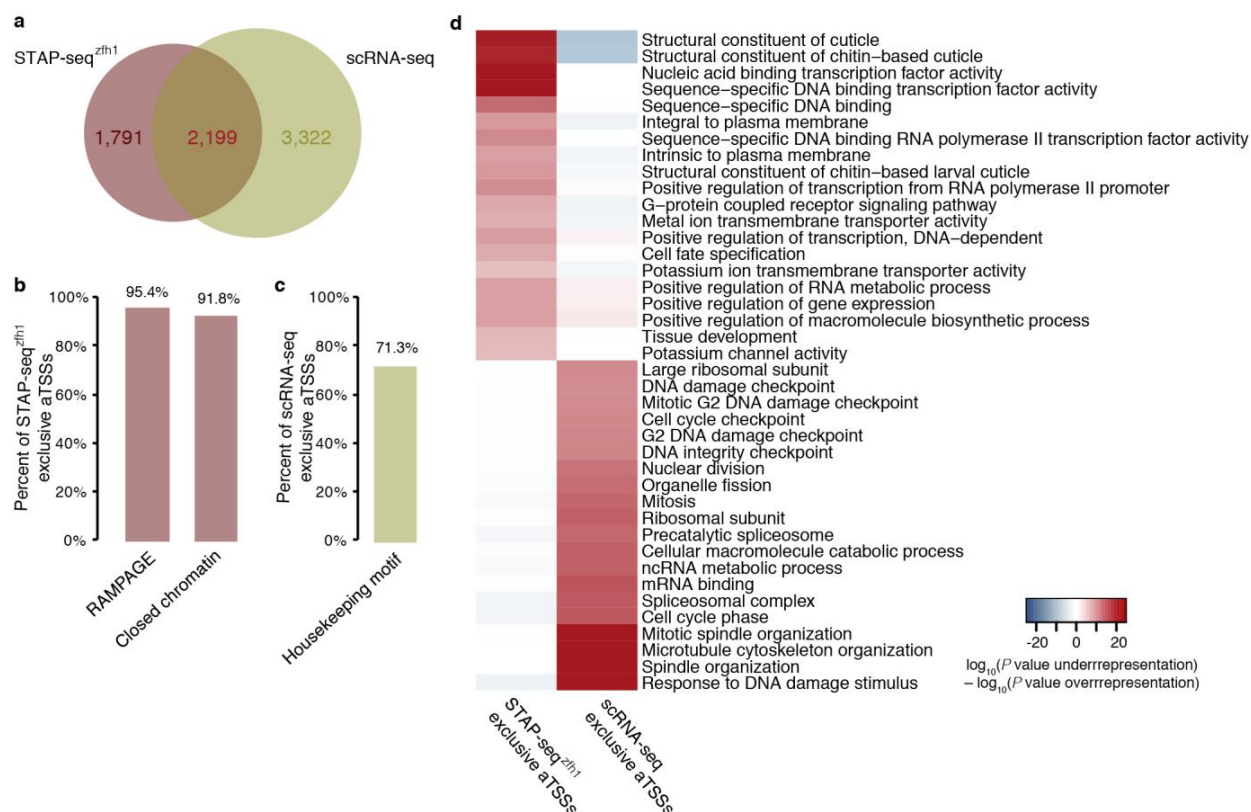
(a) Scatterplots comparing focused STAP-seq^{zf1} screens with different screens with diverse developmental (dev; top) and housekeeping enhancers (hk; bottom; see **Figure 2** for further comparisons; PCC: Pearson correlation coefficient). (b) Scatterplot depicting the similarity between focused and genome-wide STAP-seq^{zf1} screens. (c) Scatterplot for two independent biological replicates of focused STAP-seq^{zf1} screens, including standard deviations (s.d.) calculated across three independent biological replicates (error bars).



Supplementary Figure 4

Induced activities are consistent across cell types.

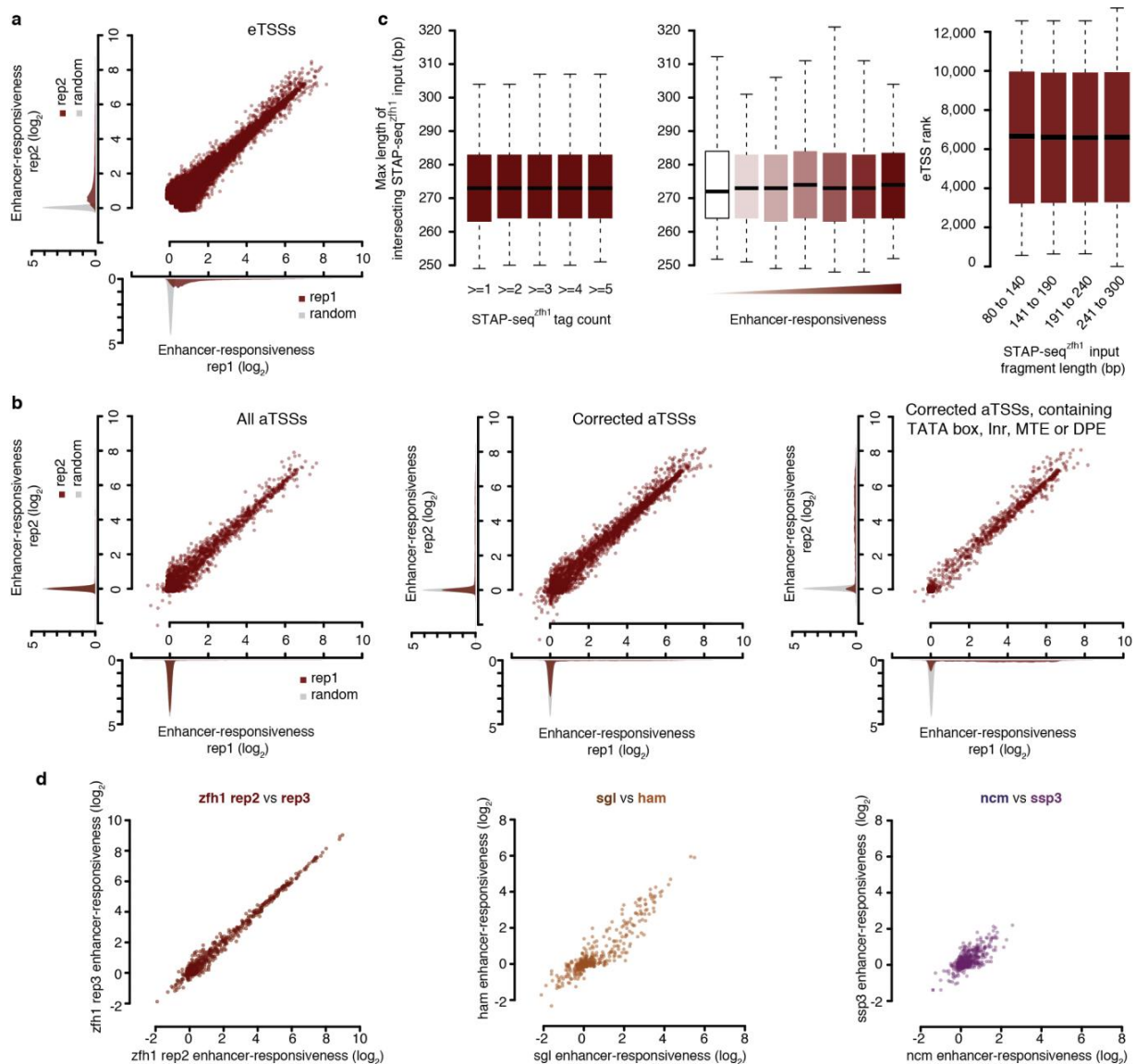
Scatterplot depicting STAP-seq tag counts for STAP-seq^{zh1} in S2 cells (x-axis) versus STAP-seq^{ij} in OSCs (y-axis) and their similarity (expressed as PCC: Pearson correlation coefficient). TSSs that are endogenously – as measured by GRO-seq^{22,23} – and exclusively active in S2 cells or OSCs are labeled blue or red, respectively. Aligned to each axis are the respective cumulative distributions used to assess the difference between TSSs that are endogenously exclusively active in either S2 cells or OSCs by Kolmogorov–Smirnov tests (P values indicated). The scatter plot corresponds to **Figure 3a**.



Supplementary Figure 5

STAP-seq is complementary to methods that assess endogenous transcription initiation.

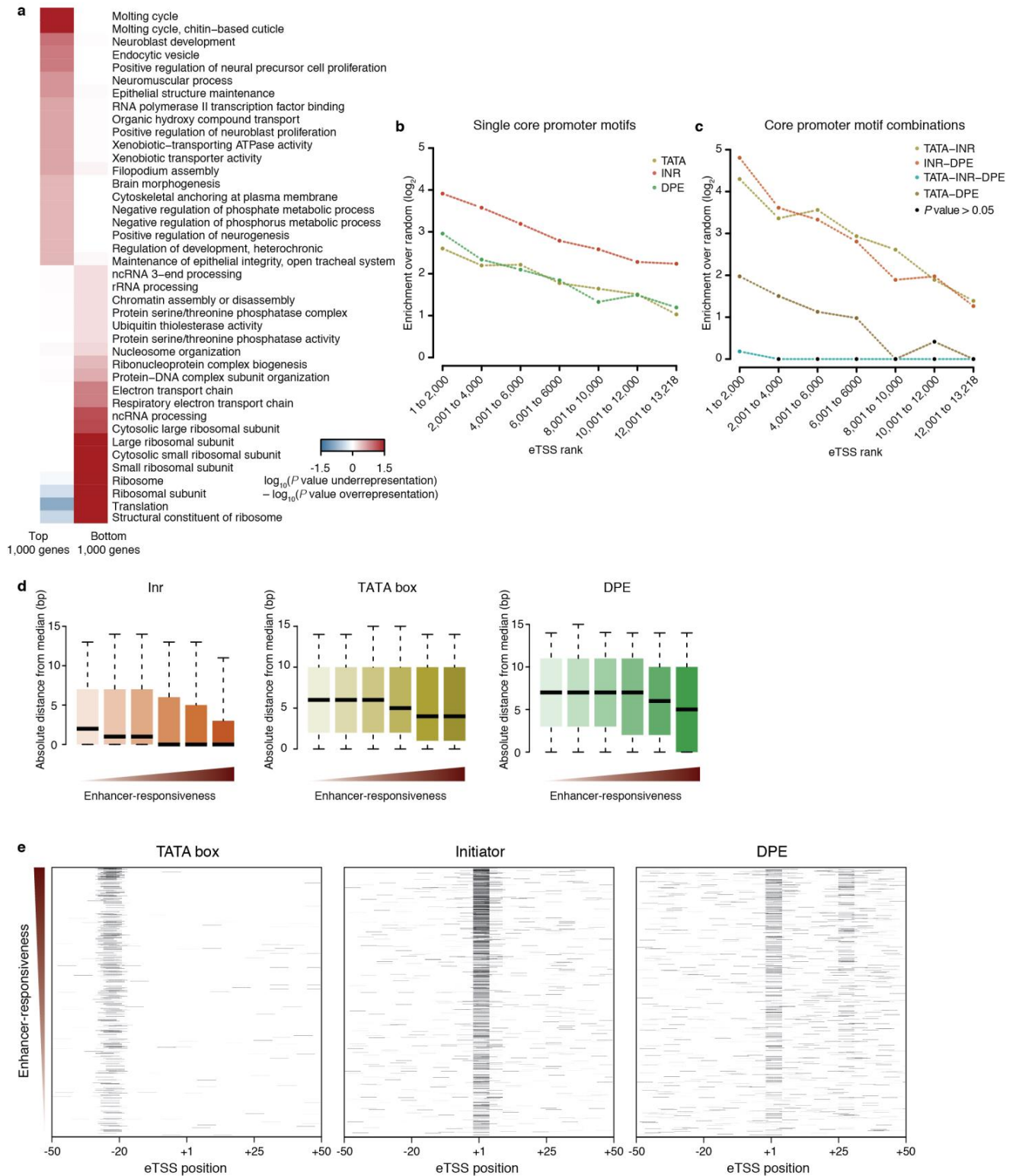
(a) Venn diagram depicting the overlap of aTSSs detected by STAP-seq^{zfh1} and scRNA-seq in S2 cells genome-wide (see **Figures 3d** and **e** for an equivalent analysis of the focused STAP-seq screens). (b) Proportion of aTSSs uniquely detected by STAP-seq^{zfh1} that are also detected during different developmental stages by RAMPAGE⁴⁴ (left bar) or are in closed chromatin in S2 cells (right bar). (c) aTSSs uniquely detected by scRNA-seq that contain housekeeping core promoter motifs. (d) Gene Ontology analysis of aTSSs uniquely detected by either STAP-seq^{zfh1} or scRNA-seq. The results from **b-d** suggest that aTSSs uniquely detected by scRNA-seq are housekeeping-type core promoters (see also **Figures 3e** and **f**), while aTSSs uniquely detected by STAP-seq^{zfh1} are endogenously not active in S2 cells (b, right bar) but in other cell types (b, left bar).



Supplementary Figure 6

A wide range of enhancer-responsiveness.

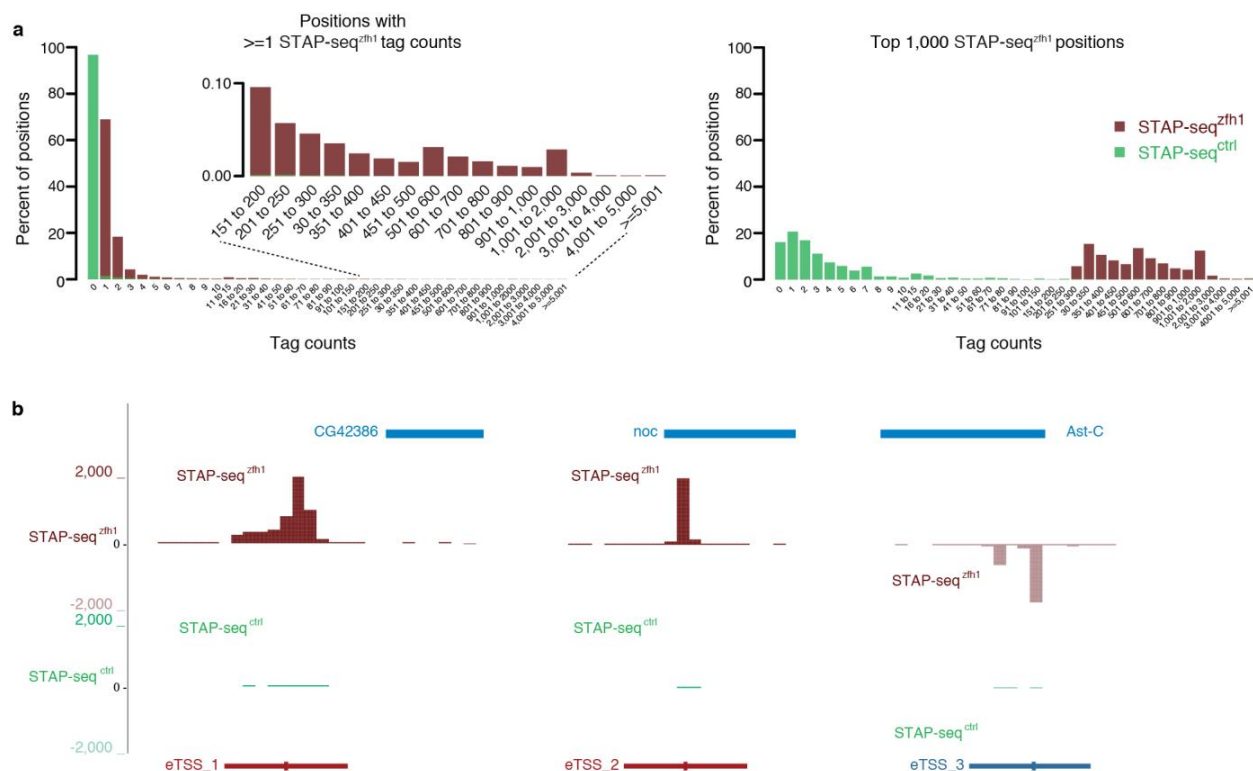
(a) Scatterplots depicting the range of enhancer-responsiveness as determined by STAP-seq^{zh1} over STAP-seq^{ctrl} at eTSSs for replicate (rep) 1 versus 2. The distributions of enhancer-responsiveness at the respective TSSs (brown) and at random positions (grey) are shown by density plots along the axes. **(b)** As in **a**, but, at all, position-corrected, and TATA box, Initiator (Inr), MTE or DPE-containing and position-corrected aTSSs. **(c)** TSS strengths are independent of candidate lengths. Boxplots depicting maximal length of candidate STAP-seq^{zh1} fragments intersecting positions with the corresponding STAP-seq^{zh1} tag counts (left) or eTSSs of different enhancer-responsiveness (middle), and ranks of eTSSs which +1 positions intersect candidate STAP-seq^{zh1} fragments of different lengths (right). Center line: median; limits: interquartile range; whiskers: 5th and 95th percentiles. **(d)** Housekeeping enhancers activate transcription from candidate sequences at a reduced dynamic range compared to developmental enhancers. Scatterplots depicting the range of enhancer-responsiveness of the indicated focused STAP-seq screens at aTSSs in genomic regions covered by the focused candidate libraries for developmental (left and middle) and housekeeping (right) enhancers. To account for the broad nature of initiation at housekeeping core promoters, enhancer-responsiveness in **d**, is calculated in a window of ± 50 bp around aTSSs.



Supplementary Figure 7

Biological significance and sequence properties of eTSSs with different enhancer-responsiveness.

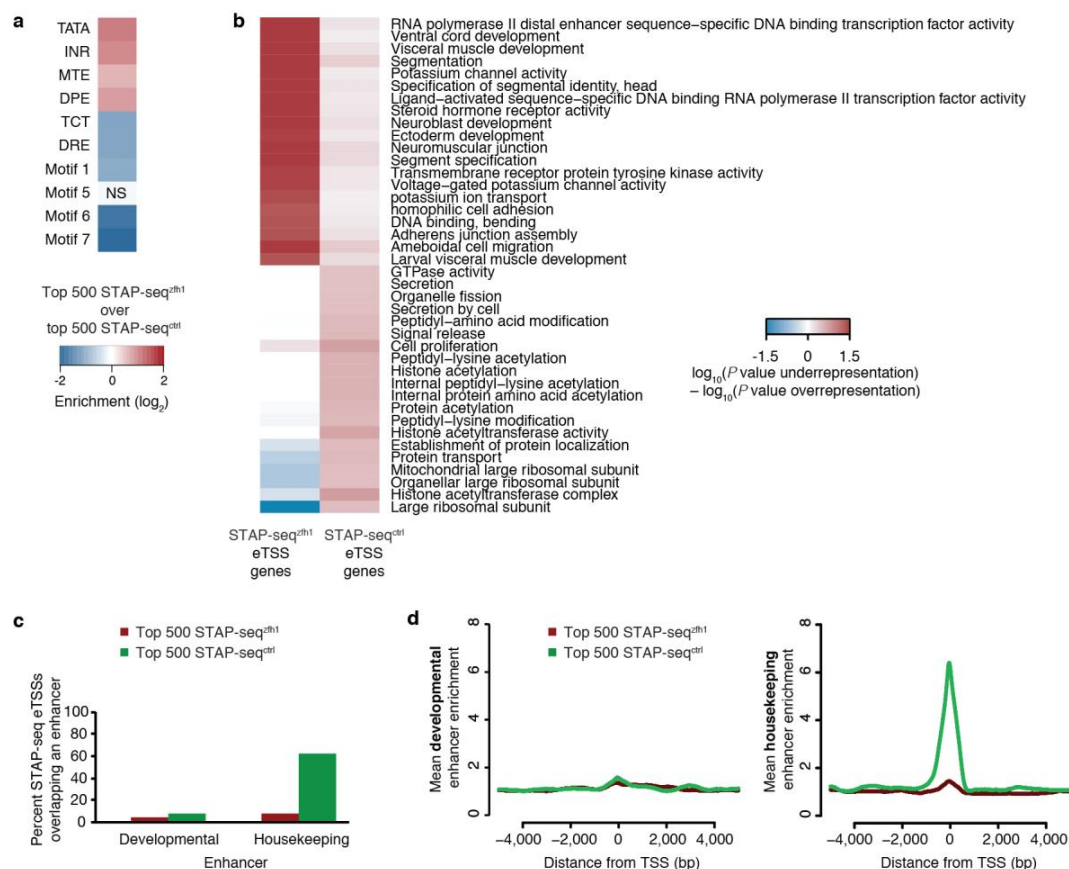
(a) Gene Ontology analysis of the top and bottom 1,000 genes associated with the strongest and weakest eTSSs, respectively (see **Figure 4f** for an equivalent analysis restricted to eTSSs containing exclusively TATA box, Inr, MTE or DPE). (b-c) Enrichment of single (b) or combinations (c) of core promoter motifs in eTSSs of different ranks compared to random genomic sequences. As has been observed before²⁰, the combination of TATA box and DPE motif is less strongly enriched compared to the TATA box and Inr or Inr and DPE motif pairs and the combination of all three motifs rarely occurs in the same eTSS. (d) Boxplots depicting motif deviation (distance in base pairs, bp) from their consensus positions (determined as the average position across all eTSSs) versus enhancer-responsiveness of eTSSs. Center line: median; limits: interquartile range; whiskers: 5th and 95th percentiles (e) Heatmaps depicting positional occurrences of core promoter motifs around eTSSs. Note that TATA boxes are somewhat less positionally constraint, as previously observed for mammalian core promoters³¹.



Supplementary Figure 8

Induced and basal activities measured by STAP-seq^{zfh1} and STAP-seq^{ctrl}, respectively, reveal low basal activities for top eTSSs.

(a) Histograms depicting distributions of normalized STAP-seq^{zfh1} and STAP-seq^{ctrl} tag counts at positions that are covered by at least 1 tag (left) or the top 1,000 STAP-seq^{zfh1} positions (right). The inset on the left shows that STAP-seq^{zfh1} but not STAP-seq^{ctrl} reaches high normalized tag counts. (b) Representative screenshots of the top three eTSSs, depicting high STAP-seq^{zfh1} and low STAP-seq^{ctrl} tag counts.



Supplementary Figure 9

Sequences with the highest basal activities are at housekeeping genes and overlap housekeeping-type enhancers.

(a) Core promoter motif enrichment analysis of candidate sequences with the highest basal (STAP-seq^{ctrl}) versus the highest induced (STAP-seq^{th1}) activities (top 500 each). NS: not significant. (b) Gene Ontology (GO) analysis for genes associated with eTSSs that show the highest basal (STAP-seq^{ctrl}, right) versus the highest induced (STAP-seq^{th1}, left) activities. (c) Fraction of candidates as in a, that show STARR-seq enrichment values of at least 3-fold. Developmental and housekeeping STARR-seq data are from ref. 10. (d) As in c, but plotting average STARR-seq signals around the most prominent, or the center of dispersed TSSs.

Conclusions and Perspectives

Enhancer-specificity and Biochemical Compatibility

Previously, the specificity of core promoters towards enhancers have been observed in a few cases. However, how the specificity is mediated could not be determined. Here I have shown that the enhancer-specificity of core promoters correlates globally with enhancer sequence signatures, through which TFs are differentially recruited. Core promoters, including housekeeping and developmental-type core promoters, have been known to differentially recruit and be activated by *trans* factors^{121,122}. Together, this suggests that enhancer–core-promoter specificity is directly mediated via biochemical compatibility between the factors recruited by core promoters and enhancers.

Separately, our group has determined that many TFs and COFs have differential activities towards housekeeping and developmental core promoters²⁶⁸. These results, together with the known biological functions of the factors, suggest that these factors mediate the enhancer–core-promoter communication of the two transcription programs. The biological functions of these factors (apart from those already mentioned in Paper #2) are also consistent with their activities towards the different core promoters. For example, the TF *relative-of-woc* (*row*) preferentially activates housekeeping core promoters, and is involved in the correct recognition of heterochromatin³¹⁴. Conversely, Dorsal-related immunity factor (*Dif*) and Pointed (*Pnt*) preferentially activate developmental core promoters. These TFs are respectively crucial for antimicrobial response³¹⁵, and neural and eye development^{316,317} in living *Drosophila*.

The next step is to determine, via proteomics-based approach, whether these factors exist and are important in *in vivo* complexes that mediate the enhancer–core-promoter activity. Additionally, depletion experiments might be able to validate the requirement of these factors for the activity of the complexes. To this end, the enhancers identified in Paper #1, and the core promoters in Paper #3, provide a ready set of *cis*-regulatory elements with which the protein complexes could be isolated.

The next question is how the message from enhancers is relayed to core promoters. At least two scenarios can be envisioned. First, the message is transmitted via the conformational change of the complexes. Alternatively, as many of the COFs carry enzymatic activities, the signal can be relayed by PTMs. The Mediator complex³¹⁸ has been extensively studied and shown to act as a “bridge” that communicate the regulatory signal between enhancers and core promoters for most Pol II genes³¹⁹⁻³²¹. Both the membership of the Mediator subunits and their PTMs can affect the biological functions of the Mediator³²². Interestingly, some Mediator subunits show strongly differential activity towards the developmental versus housekeeping core promoters²⁶⁸. For example, Mediator complex subunits 15 and 25 (MED15 and MED25) strongly activate developmental core promoters compared to housekeeping core promoters, implying that they function differentially in these transcription programs. Further investigations on these factors, as well as other factors that show differential activity towards the two core promoters, will determine how they function within their preferred programs, as well as how potential communication leakage across the programs is prevented.

Shortly after Paper #2 was published, another group has validated our findings and extended the characterization of the two enhancer classes. Specifically, they investigated enhancer–core-promoter communication of the two transcription programs from the perspective of genome organization. To this end, they determined the binding signatures of architectural proteins at the identified enhancers³²³. Strikingly, the two enhancer classes exhibit differential occupancy of architectural proteins. Furthermore, housekeeping enhancers associate with multiple core promoters at topologically associated domain (TAD) borders, while developmental enhancers are more likely to interact with single core promoters within TADs. Interestingly, they also observed differential enrichment of histone PTMs surrounding the two enhancer classes, suggesting for their differential association with histone-modifying enzymes. It will be interesting to see whether these enzymes are also involved in PTM-mediated signaling within these transcription programs. Additionally, the results from their study suggests that differential assembly of architectural proteins could be either an alternative, or a non-overlapping mechanism, from biochemical compatibility in mediating enhancer–core-promoter specificity.

Enhancer-specificity in Other Transcription Programs

The housekeeping and developmental transcription programs are active in several somatic *Drosophila* cell lines that we investigated. Other transcription programs, also separated via enhancer–core-promoter specificity, might also exist in these cell lines. In addition, many transcription programs that are dedicated for specific cell types exist. Especially interesting are transcription programs in cell types with very different properties, such as stem or germ cells. Some of the critical core promoter factors for these transcription programs have been identified. For example, gonad-specific TAFs have been identified and shown to drive the pluripotency transcription programs^{128,130,324-326}. It is plausible that enhancer–core-promoter specificity is also employed to separately drive gonad-specific transcription programs from other co-existing transcription programs.

Decoding Enhancer-responsiveness

Core promoter activities have previously been determined mainly via *in vitro* transcription and reporter assays. Most of these studies were limited in scale, and the range of enhancer-responsiveness of core promoters in the genome could not be determined. Here, I have shown that the wide range of enhancer-responsiveness is related to the affinity and position of core promoter elements. This observation implies that the difference in sequence signatures translates into enhancer-responsiveness via the different efficiency, stability or kinetics of the *trans* factors at core promoters. Consistently, the affinity of the TATA box correlates with the conformational change and stability of the core promoter-TBP containing complex^{327,328}. Thus, given the constant enhancer input as in STAP-seq, each core promoter has differential behavior of the *trans* factors, translating the difference in the core promoter sequence into different levels of transcription initiation.

Enhancer-responsiveness of housekeeping core promoters has reduced dynamic range. For translation-related core promoters that contain the positionally-constrained TCT element and support focused initiation, enhancer-responsiveness might be similarly related to the affinities and the positioning of the TCT element. However, other

housekeeping core promoters support dispersed initiation, and carry non positionally-constrained core promoter elements. Thus, enhancer-responsiveness at these core promoters might rely on the affinities of the core promoter elements, or other determinants such as the overall sequence composition. Furthermore, it will be interesting to determine whether enhancer-responsiveness has any correlation with the “disperseness” of initiation at housekeeping core promoters.

Deciphering the Initiation Code

At core promoters identified by STAP-seq, Inr stably aligns around the +1 position. However, there are still many genomic instances of Inr with no STAP-seq activity. Within these fragments, there could be short motifs that negatively affect the PIC^{279,329,330}, or that affect the stability of the chimeric STAP-seq transcript, for example the poly (A) signal (PAS) motif^{331,332}. A deeper analysis of the Inr-containing fragments with low enhancer-responsiveness will elucidate how the transcriptional machinery surveys the genome, correctly recognizes and activates core promoters while ensuring fidelity in transcription.

Recently, it was shown that CpG methylation represses spurious transcription initiation from intragenic Inr-like sequences in mammals³⁰⁰. Interestingly, STAP-seq also identifies core promoters that are distal from annotated gene starts. As *Drosophila* lacks genome methylation, it will also be interesting to determine how these sequences are silenced in the genome. Finally, STAP-seq also allows the identification of core promoters that do not belong to the housekeeping or developmental classes. This can be done by identifying core promoters that are highly active in other tissues or developmental stage, but not active in STAP-seq using housekeeping and developmental enhancers in somatic cell lines. Such analysis will shed light into the nature of these core promoters, as well as the transcription programs to which they belong.

Enhancer-responsiveness and Transcription Regulation

Enhancer-responsiveness is also invariant to cell type, consistent with previous studies of single core promoters. Furthermore, this is in line with the notion that core promoters are the sites of the nucleation of the PIC. Across cell types, the PIC is (relatively) uniform compared to the complexes that form at enhancers, directly suggesting that the cell-type specific activity of genes lies mainly at enhancers. As non-canonical PIC components, particularly cell-type specific TAFs also exist, it will be interesting to see what is the nature of enhancer-responsive on recognition by complexes that contain such cell-type specific TAFs. Additionally, there are also individual core promoters with highly specific cell-type activities, implying that some sequence-specific TF motifs could be present in core promoters³³³⁻³³⁶.

The wide range of enhancer-responsiveness of core promoters is related to gene functions and the number of enhancers. Genes with highly responsive core promoters are associated with more enhancers, compared to genes with core promoters that have reduced responsiveness. Thus, enhancer-responsiveness of core promoters imposes a ceiling level for a gene to be expressed. Consistently, core promoter sequences is related to their maximally-observed *in vivo* activity²⁸⁶.

Consider a scenario where core promoters have highly similar enhancer-responsiveness. For these hypothetical core promoters, they have a constant degree of activation irrespective of enhancer input. The transcriptional output is dependent instead on their intrinsic basal activity and DNA accessibility. An example exists where it is beneficial to regulate a core promoter via its basal activity. Two well-studied core promoters from adenovirus are the E4 and the major late (ML) core promoters. Compared to the ML core promoter, the E4 core promoter is highly responsive towards activators. The ML core promoter has lower responsiveness but strong basal activity, and is proposed to function by being responsive towards the viral genome copy number³³⁰. For multicellular animals with stable ploidy and gene copy number however, it is the wide range of enhancer-responsiveness of the core promoters that is critical for differential gene regulation.

Beyond Enhancer-responsiveness and –specificity

The studies in Papers #1 and #3 were performed in *Drosophila*. Many aspects of transcription are conserved between *Drosophila* and mammals, including several core promoter elements, their biological association (such as the TCT element with translation-associated genes), and the dichotomy of the initiation patterns. It is likely that enhancer-specificity and –responsiveness exists and is also sequence-encoded in mammals and other species.

How do enhancer-specificity and -responsiveness of core promoters relate? Enhancer-specificity is an extreme trait of enhancer-responsiveness. Core promoters that are highly responsive towards one transcription program seem to be lowly responsive, or could even be repressive, towards the other. By having mutually exclusive activity towards the different enhancer classes, erroneous activation of core promoters by the other enhancer class can be prevented. This notion is consistent with previous observations that core promoter elements that are related to either the housekeeping or the developmental programs seldom co-occur^{46,101}. A thorough examination of core promoter activities from Paper #3 will elucidate whether this notion is true.

Knowledge on enhancer-specificity and –responsiveness should aid in the design of transcriptional reporter and transgene expression constructs (for example, discussed in ref. 225). For most transgenic experiments, it is important that the constructs exhibit strong cell-type specific activity. This can be done by pairing a cell-type specific enhancer with a developmental core promoter. The correct level of reporter expression can be tuned via selecting the core promoter: a core promoter with high enhancer-responsiveness and low basal activity will result in high expression and low ectopic expression. Conversely, a housekeeping enhancer–core-promoter pairing will ensure construct activity with minimal stage or cell-type specificity.

In conclusion, I have shown that core promoters have sequence-intrinsic enhancer-specificity and a wide range of enhancer-responsiveness. These traits determine the transcription output of core promoters on activation by enhancers. Importantly, these traits are central in explaining differential gene expression, how cell translate the

information from the genome and drive the development of multicellular animals. Additionally, the identification of housekeeping and developmental enhancers, and the quantification of enhancer-responsiveness of core promoters, should be a valuable resource for future studies.

References

1. Wilmut, I. *et al.* Somatic cell nuclear transfer. *Nature* **419**, 583–587 (2002).
2. Wilmut, I., Schnieke, A. E., McWhir, J., Kind, A. J. & Campbell, K. H. S. Viable offspring derived from fetal and adult mammalian cells. *Nature* **385**, 810–813 (1997).
3. Carroll, S. B. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell* **134**, 25–36 (2008).
4. Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424**, 147–151 (2003).
5. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β -globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).
6. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **15**, 272–286 (2014).
7. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613–626 (2012).
8. Reiter, F., Wienerroither, S. & Stark, A. Combinatorial function of transcription factors and cofactors. *Curr. Opin. Genet. Dev.* **43**, 73–81 (2017).
9. Spence, J. R. *et al.* Directed differentiation of human pluripotent stem cells into intestinal tissue in vitro. *Nature* **470**, 105–109 (2010).
10. D'Amour, K. A. *et al.* Production of pancreatic hormone-expressing endocrine cells from human embryonic stem cells. *Nat. Biotechnol.* **24**, 1392–1401 (2006).
11. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**, 663–676 (2006).
12. Takahashi, H., Lassmann, T., Murata, M. & Carninci, P. 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat Protoc* **7**, 542–561 (2012).
13. Ieda, M. *et al.* Direct Reprogramming of Fibroblasts into Functional Cardiomyocytes by Defined Factors. *Cell* **142**, 375–386 (2010).
14. Zhou, Q., Brown, J., Kanarek, A., Rajagopal, J. & Melton, D. A. In vivo reprogramming of adult pancreatic exocrine cells to beta-cells. *Nature* **455**, 627–632 (2008).

15. Xie, H., Ye, M., Feng, R. & Graf, T. Stepwise reprogramming of B cells into macrophages. *Cell* **117**, 663–676 (2004).
16. Vierbuchen, T. *et al.* Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* **463**, 1035–1041 (2010).
17. Lo, H.-Y. G. *et al.* A single transcription factor is sufficient to induce and maintain secretory cell architecture. *Genes Dev.* **31**, 154–171 (2017).
18. Choi, J. *et al.* MyoD converts primary dermal fibroblasts, chondroblasts, smooth muscle, and retinal pigmented epithelial cells into striated mononucleated myoblasts and multinucleated myotubes. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 7988–7992 (1990).
19. Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.* **16**, 144–154 (2015).
20. Cannavò, E. *et al.* Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. *Current Biology* **26**, 38–51 (2016).
21. Xiong, N., Kang, C. & Raulet, D. H. Redundant and Unique Roles of Two Enhancer Elements in the TCR γ Locus in Gene Regulation and $\gamma\delta$ T Cell Development. *Immunity* **16**, 453–463 (2002).
22. Farley, E. K. *et al.* Suboptimization of developmental enhancers. *Science* **350**, 325–328 (2015).
23. Kvon, E. Z. *et al.* Genome-scale functional characterization of *Drosophila* developmental enhancers in vivo. *Nature* **512**, 91–95 (2014).
24. Smale, S. T. Core promoters: active contributors to combinatorial gene regulation. *Genes Dev.* **15**, 2503–2508 (2001).
25. Smale, S. T. & Kadonaga, J. T. The RNA Polymerase II Core Promoter. *Annu. Rev. Biochem.* **72**, 449–479 (2003).
26. Butler, J. E. F. & Kadonaga, J. T. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.* **16**, 2583–2592 (2002).
27. Gross, P. & Oelgeschläger, T. Core promoter-selective RNA polymerase II transcription. *Biochem. Soc. Symp.* 225–236 (2006).
28. Kadonaga, J. T. Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip Rev Dev Biol* **1**, 40–51 (2012).
29. Luse, D. S. & Roeder, R. G. Accurate transcription initiation on a purified mouse

- beta-globin DNA fragment in a cell-free system. *Cell* **20**, 691–699 (1980).
30. Weil, P. A., Luse, D. S., Segall, J. & Roeder, R. G. Selective and accurate initiation of transcription at the Ad2 major late promotor in a soluble system dependent on purified RNA polymerase II and DNA. *Cell* **18**, 469–484 (1979).
 31. Kamakaka, R. T., Tyree, C. M. & Kadonaga, J. T. Accurate and efficient RNA polymerase II transcription with a soluble nuclear fraction derived from *Drosophila* embryos. *Proc. Natl. Acad. Sci. U.S.A.* **88**, 1024–1028 (1991).
 32. Dignam, J. D., Lebovitz, R. M. & Roeder, R. G. Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Research* **11**, 1475–1489 (1983).
 33. Grünberg, S. & Hahn, S. Structural insights into transcription initiation by RNA polymerase II. *Trends in Biochemical Sciences* **38**, 603–611 (2013).
 34. Vannini, A. & Cramer, P. Conservation between the RNA polymerase I, II, and III transcription initiation machineries. *Mol. Cell* **45**, 439–446 (2012).
 35. Matsui, T., Segall, J., Weil, P. A. & Roeder, R. G. Multiple factors required for accurate initiation of transcription by purified RNA polymerase II. *J. Biol. Chem.* **255**, 11992–11996 (1980).
 36. Reinberg, D. & Roeder, R. G. Factors involved in specific transcription by mammalian RNA polymerase II. Purification and functional analysis of initiation factors IIB and IIE. *J. Biol. Chem.* **262**, 3310–3321 (1987).
 37. Flores, O., Lu, H. & Reinberg, D. Factors involved in specific transcription by mammalian RNA polymerase II. Identification and characterization of factor IIH. *J. Biol. Chem.* **267**, 2786–2793 (1992).
 38. Roeder, R. G. & Rutter, W. J. Multiple ribonucleic acid polymerases and ribonucleic acid synthesis during sea urchin development. *Biochemistry* **9**, 2543–2553 (1970).
 39. Orphanides, G., Lagrange, T. & Reinberg, D. The general transcription factors of RNA polymerase II. *Genes Dev.* **10**, 2657–2683 (1996).
 40. Hahn, S. Structure and mechanism of the RNA polymerase II transcription machinery. *Nat. Struct. Mol. Biol.* **11**, 394–403 (2004).
 41. Conaway, R. C. & Conaway, J. W. General Initiation Factors for RNA Polymerase II. *Annu. Rev. Biochem.* **62**, 161–190 (1993).
 42. Thomas, M. C. & Chiang, C.-M. The general transcription machinery and general

- cofactors. *Crit. Rev. Biochem. Mol. Biol.* **41**, 105–178 (2006).
43. Gross, P. & Oelgeschläger, T. Core promoter-selective RNA polymerase II transcription. *Biochem. Soc. Symp.* 225–236 (2006).
 44. Burley, S. K. & Roeder, R. G. TATA box mimicry by TFIID: autoinhibition of pol II transcription. *Cell* **94**, 551–553 (1998).
 45. Roeder, R. G. Role of General and Gene-specific Cofactors in the Regulation of Eukaryotic Transcription. *Cold Spring Harbor Symposia on Quantitative Biology* **63**, 201–218 (1998).
 46. Ohler, U., Liao, G.-C., Niemann, H. & Rubin, G. M. Computational analysis of core promoters in the Drosophila genome. *Genome Biol.* **3**, RESEARCH0087 (2002).
 47. Xi, H. *et al.* Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. *Genome Research* **17**, 798–806 (2007).
 48. FitzGerald, P. C., Shlyakhtenko, A., Mir, A. A. & Vinson, C. Clustering of DNA sequences in human promoters. *Genome Research* **14**, 1562–1574 (2004).
 49. Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet* **13**, 233–245 (2012).
 50. Haberle, V. & Lenhard, B. Promoter architectures and developmental gene regulation. *Seminars in Cell & Developmental Biology* **57**, 11–23 (2016).
 51. Corden, J. *et al.* Promoter sequences of eukaryotic protein-coding genes. *Science* **209**, 1406–1414 (1980).
 52. Smale, S. T. & Baltimore, D. The ‘initiator’ as a transcription control element. *Cell* **57**, 103–113 (1989).
 53. Vo Ngoc, L., Cassidy, C. J., Huang, C. Y., Duttke, S. H. C. & Kadonaga, J. T. The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters. *Genes Dev.* **31**, 6–11 (2017).
 54. Goldberg, M. L. Sequence analysis of Drosophila histone genes. (1979).
 55. Burke, T. W. & Kadonaga, J. T. Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev.* **10**, 711–724 (1996).
 56. Hetzel, J., Duttke, S. H., Benner, C. & Chory, J. Nascent RNA sequencing reveals distinct features in plant transcription. *Proceedings of the National Academy of*

- Sciences* **113**, 12316–12321 (2016).
57. Marbach-Bar, N. *et al.* DTIE, a novel core promoter element that directs start site selection in TATA-less genes. *Nucleic Acids Research* **44**, 1080–1094 (2016).
 58. Nepal, C. *et al.* Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. *Genome Research* **23**, 1938–1950 (2013).
 59. Tokusumi, Y., Ma, Y., Song, X., Jacobson, R. H. & Takada, S. The new core promoter element XCPE1 (X Core Promoter Element 1) directs activator-, mediator-, and TATA-binding protein-dependent but TFIID-independent RNA polymerase II transcription from TATA-less promoters. *Mol. Cell. Biol.* **27**, 1844–1858 (2007).
 60. Anish, R., Hossain, M. B., Jacobson, R. H. & Takada, S. Characterization of transcription from TATA-less promoters: identification of a new core promoter element XCPE2 and analysis of factor requirements. *PLoS One* **4**, e5103 (2009).
 61. Li, H. *et al.* Genome-wide analysis of core promoter structures in *Schizosaccharomyces pombe* with DeepCAGE. *RNA Biology* **12**, 525–537 (2015).
 62. Lim, C. Y. *et al.* The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev.* **18**, 1606–1617 (2004).
 63. Lagrange, T., Kapanidis, A. N., Tang, H., Reinberg, D. & Ebright, R. H. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev.* **12**, 34–44 (1998).
 64. Deng, W. & Roberts, S. G. E. A core promoter element downstream of the TATA box that is recognized by TFIIB. *Genes Dev.* **19**, 2418–2423 (2005).
 65. Lewis, B. A., Kim, T. K. & Orkin, S. H. A downstream element in the human beta-globin promoter: evidence of extended sequence-specific transcription factor IID contacts. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 7172–7177 (2000).
 66. Stormo, G. D., Schneider, T. D. & Gold, L. Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucleic Acids Research* **14**, 6661–6679 (1986).
 67. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Research* **14**, 1188–1190 (2004).
 68. Sainsbury, S., Bernecky, C. & Cramer, P. Structural basis of transcription initiation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **16**, 129–143 (2015).

69. Cianfrocco, M. A. & Nogales, E. Regulatory interplay between TFIID's conformational transitions and its modular interaction with core promoter DNA. *Transcription* **4**, (2013).
70. Louder, R. K. *et al.* Structure of promoter-bound TFIID and model of human pre-initiation complex assembly. *Nature* **531**, 604–609 (2016).
71. Nogales, E., Fang, J. & Louder, R. K. Structural dynamics and DNA interaction of human TFIID. *Transcription* **8**, 55–60 (2017).
72. Hirose, F., Yamaguchi, M., Handa, H., Inomata, Y. & Matsukage, A. Novel 8-base pair sequence (*Drosophila* DNA replication-related element) and specific binding factor involved in the expression of *Drosophila* genes for DNA polymerase alpha and proliferating cell nuclear antigen. *J. Biol. Chem.* **268**, 2092–2099 (1993).
73. Ince, T. A. & Scotto, K. W. A conserved downstream element defines a new class of RNA polymerase II promoters. *J. Biol. Chem.* **270**, 30249–30252 (1995).
74. Perry, R. P. The architecture of mammalian ribosomal protein promoters. *BMC Evol. Biol.* **5**, 15 (2005).
75. Parry, T. J. *et al.* The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes Dev.* **24**, 2013–2018 (2010).
76. Laflamme, K. *et al.* Functional analysis of a novel cis-acting regulatory region within the human ankyrin gene (ANK-1) promoter. *Mol. Cell. Biol.* **30**, 3493–3502 (2010).
77. Gallagher, P. G. A dinucleotide deletion in the ankyrin promoter alters gene expression, transcription initiation and TFIID complex formation in hereditary spherocytosis. *Hum. Mol. Genet.* **14**, 2501–2509 (2005).
78. Schug, J. *et al.* Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.* **6**, R33 (2005).
79. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 1412–1417 (2006).
80. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25**, 1010–1022 (2011).
81. Bird, A. P. CpG-rich islands and the function of DNA methylation. *Nature* **321**, 209–213 (1986).

82. Tykocinski, M. L. & Max, E. E. CG dinucleotide clusters in MHC genes and in 5' demethylated genes. *Nucleic Acids Research* **12**, 4385–4396 (1984).
83. Cooper, D. N., Taggart, M. H. & Bird, A. P. Unmethlated domains in vertebrate DNA. *Nucleic Acids Research* **11**, 647–658 (1983).
84. Bird, A., Taggart, M., Frommer, M., Miller, O. J. & Macleod, D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40**, 91–99 (1985).
85. Rach, E. A. *et al.* Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLoS Genetics* **7**, e1001274 (2011).
86. Hoskins, R. A. *et al.* Genome-wide analysis of promoter architecture in *Drosophila melanogaster*. *Genome Research* **21**, 182–192 (2011).
87. Ni, T. *et al.* A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nat. Methods* **7**, 521–527 (2010).
88. Schor, I. E. *et al.* Promoter shape varies across populations and affects promoter evolution and expression noise. *Nat. Genet.* **49**, 550–558 (2017).
89. Yamamoto, Y. Y. *et al.* Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC Genomics* **8**, 67 (2007).
90. Yamamoto, Y. Y. *et al.* Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis. *Nucleic Acids Research* **35**, 6219–6226 (2007).
91. Hampsey, M. Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiol. Mol. Biol. Rev.* **62**, 465–503 (1998).
92. Mitra, S. & Narlikar, L. No Promoter Left Behind (NPLB): learnde novopromoter architectures from genome-wide transcription start sites. *Bioinformatics* **32**, 779–781 (2015).
93. Narlikar, L. Multiple novel promoter-architectures revealed by decoding the hidden heterogeneity within the genome. *Nucleic Acids Research* **42**, 12388–12403 (2014).
94. Engström, P. G., Ho Sui, S. J., Drivenes, O., Becker, T. S. & Lenhard, B. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Research* **17**, 1898–1908 (2007).
95. FitzGerald, P. C., Sturgill, D., Shyakhtenko, A., Oliver, B. & Vinson, C. Comparative genomics of *Drosophila* and human core promoters. *Genome Biol.*

- 7, R53 (2006).
96. Katzenberger, R. J. R., Rach, E. A. E., Anderson, A. K. A., Ohler, U. U. & Wassarman, D. A. D. The *Drosophila* Translational Control Element (TCE) Is Required for High-Level Transcription of Many Genes That Are Specifically Expressed in Testes. *PLoS One* **7**, e45009–e45009 (2011).
 97. Chen, K. *et al.* A global change in RNA polymerase II pausing during the *Drosophila* midblastula transition. *Elife* **2**, (2013).
 98. Sloutskin, A. *et al.* Element: a computational tool for detecting core promoter elements. *Transcription* **6**, 41–50 (2015).
 99. Westermarck, P. O. Linking Core Promoter Classes to Circadian Transcription. *PLoS Genetics* **12**, e1006231 (2016).
 100. Moshonov, S., Elfakess, R., Golan-Mashiach, M., Sinvani, H. & Dikstein, R. Links between core promoter and basic gene features influence gene expression. *BMC Genomics* **9**, 92 (2008).
 101. Ohler, U. Identification of core promoter modules in *Drosophila* and their application in accurate transcription start site prediction. *Nucleic Acids Research* **34**, 5943–5950 (2006).
 102. Yarden, G., Elfakess, R., Gazit, K. & Dikstein, R. Characterization of sINR, a strict version of the Initiator core promoter element. *Nucleic Acids Research* **37**, 4234–4246 (2009).
 103. Nikolov, D. B. *et al.* Crystal structure of TFIID TATA-box binding protein. *Nature* **360**, 40–46 (1992).
 104. Horikoshi, M. *et al.* Cloning and structure of a yeast gene encoding a general transcription initiation factor TFIID that binds to the TATA box. *Nature* **341**, 299–303 (1989).
 105. Chalkley, G. E. & Verrijzer, C. P. DNA binding site selection by RNA polymerase II TAFs: a TAF(II)250-TAF(II)150 complex recognizes the initiator. *EMBO J.* **18**, 4835–4845 (1999).
 106. Librizzi, M. D., Moir, R. D., Brenowitz, M. & Willis, I. M. Expression and purification of the RNA polymerase III transcription specificity factor IIIB70 from *Saccharomyces cerevisiae* and its cooperative binding with TATA-binding protein. *J. Biol. Chem.* **271**, 32695–32701 (1996).
 107. Librizzi, M. D., Brenowitz, M. & Willis, I. M. The TATA element and its context

- affect the cooperative interaction of TATA-binding protein with the TFIIB-related factor, TFIIB70. *J. Biol. Chem.* **273**, 4563–4568 (1998).
108. Kays, A. R. & Schepartz, A. Virtually unidirectional binding of TBP to the AdMLP TATA box within the quaternary complex with TFIIA and TFIIB. *Chem. Biol.* **7**, 601–610 (2000).
 109. Cox, J. M. *et al.* Bidirectional binding of the TATA box binding protein to the TATA box. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 13475–13480 (1997).
 110. Wong, C. *et al.* Characterization of β -thalassaemia mutations using direct genomic sequencing of amplified single copy DNA. *Nature* **330**, 384–386 (1987).
 111. Ho, P. J. *et al.* Moderate reduction of beta-globin gene transcript by a novel mutation in the 5' untranslated region: a study of its interaction with other genotypes in two families. *Blood* **87**, 1170–1178 (1996).
 112. Gonzalez-Redondo, J. M. *et al.* A C→T substitution at nt-101 in a conserved DNA sequence of the promotor region of the beta-globin gene is associated with 'silent' beta-thalassemia. *Blood* **73**, 1705–1711 (1989).
 113. Cai, S. P. *et al.* Two novel beta-thalassemia mutations in the 5' and 3' noncoding regions of the beta-globin gene. *Blood* **79**, 1342–1346 (1992).
 114. Athanassiadou, A., Papachatzopoulou, A., Zoumbos, N., Maniatis, G. M. & Gibbs, R. A novel beta-thalassaemia mutation in the 5' untranslated region of the beta-globin gene. *Br. J. Haematol.* **88**, 307–310 (1994).
 115. Lee, D.-H. *et al.* Functional characterization of core promoter elements: the downstream core element is recognized by TAF1. *Mol. Cell. Biol.* **25**, 9674–9686 (2005).
 116. Breathnach, R. & Chambon, P. Organization and Expression of Eucaryotic Split Genes Coding for Proteins. *Annu. Rev. Biochem.* **50**, 349–383 (1981).
 117. Pijnappel, W. W. M. P. *et al.* A central role for TFIID in the pluripotent transcription circuitry. *Nature* **495**, 516–519 (2013).
 118. Maston, G. A. *et al.* Non-canonical TAF complexes regulate active promoters in human embryonic stem cells. *Elife* **1**, e00068 (2012).
 119. Wieczorek, E., Brand, M., Jacq, X. & Tora, L. Function of TAF(II)-containing complex without TBP in transcription by RNA polymerase II. *Nature* **393**, 187–191 (1998).
 120. Goodrich, J. A. & Tjian, R. Unexpected roles for core promoter recognition factors

- in cell-type-specific transcription and gene regulation. *Nat Rev Genet* **11**, 549–558 (2010).
121. Müller, F. & Tora, L. The multicoloured world of promoter recognition complexes. *EMBO J.* **23**, 2–8 (2003).
 122. Hochheimer, A. & Tjian, R. Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression. *Genes Dev.* **17**, 1309–1320 (2003).
 123. Zeidler, M. P., Yokomori, K., Tjian, R. & Mlodzik, M. Drosophila TFIIA-S is up-regulated and required during Ras-mediated photoreceptor determination. *Genes Dev.* **10**, 50–59 (1996).
 124. Ozer, J., Moore, P. A. & Lieberman, P. M. A testis-specific transcription factor IIA (TFIIAtau) stimulates TATA-binding protein-DNA binding and transcription activation. *J. Biol. Chem.* **275**, 122–128 (2000).
 125. Upadhyaya, A. B., Lee, S. H. & DeJong, J. Identification of a general transcription factor TFIIAalpha/beta homolog selectively expressed in testis. *J. Biol. Chem.* **274**, 18040–18048 (1999).
 126. Liu, W.-L. et al. Structural changes in TAF4b-TFIID correlate with promoter selectivity. *Mol. Cell* **29**, 81–91 (2008).
 127. Zhou, H. et al. Dual functions of TAF7L in adipocyte differentiation. *Elife* **2**, e00170 (2013).
 128. Zhou, H. et al. Taf7l cooperates with Trf2 to regulate spermiogenesis. *Proceedings of the National Academy of Sciences* **110**, 16886–16891 (2013).
 129. Zhou, H., Wan, B., Grubisic, I., Kaplan, T. & Tjian, R. TAF7L modulates brown adipose tissue formation. *Elife* **3**, (2014).
 130. Hiller, M. A., Lin, T. Y., Wood, C. & Fuller, M. T. Developmental regulation of transcription by a tissue-specific TAF homolog. *Genes Dev.* **15**, 1021–1030 (2001).
 131. Albright, S. R. & Tjian, R. TAFs revisited: more data reveal new twists and confirm old ideas. *Gene* **242**, 1–13 (2000).
 132. Reina, J. H. & Hernandez, N. On a roll for new TRF targets. *Genes Dev.* **21**, 2855–2860 (2007).
 133. Akhtar, W. & Veenstra, G. J. C. TBP-related factors: a paradigm of diversity in transcription initiation. *Cell Biosci* **1**, 23 (2011).

134. Zehavi, Y., Kedmi, A., Ideses, D. & Juven-Gershon, T. TRF2: TRansForming the view of general transcription factors. *Transcription* **6**, 1–6 (2015).
135. Hansen, S. K., Takada, S., Jacobson, R. H., Lis, J. T. & Tjian, R. Transcription Properties of a Cell Type-Specific TATA-Binding Protein, TRF. *Cell* **91**, 71–83 (1997).
136. Holmes, M. C. & Tjian, R. Promoter-selective properties of the TBP-related factor TRF1. *Science* **288**, 867–870 (2000).
137. Isogai, Y., Takada, S., Tjian, R. & Keles, S. Novel TRF1/BRF target genes revealed by genome-wide analysis of Drosophila Pol III transcription. *EMBO J.* **26**, 79–89 (2007).
138. Takada, S., Lis, J. T., Zhou, S. & Tjian, R. A TRF1:BRF complex directs Drosophila RNA polymerase III transcription. *Cell* **101**, 459–469 (2000).
139. Teichmann, M., Dieci, G., Pascali, C. & Boldina, G. General transcription factors and subunits of RNA polymerase III: Paralogs for promoter- and cell type-specific transcription in multicellular eukaryotes. *Transcription* **1**, 130–135 (2010).
140. Schramm, L. Recruitment of RNA polymerase III to its target promoters. *Genes Dev.* **16**, 2593–2620 (2002).
141. Turowski, T. W. & Tollervey, D. Transcription by RNA polymerase III: insights into mechanism and regulation. *Biochem. Soc. Trans.* **44**, 1367–1375 (2016).
142. Geiduschek, E. P. & Kassavetis, G. A. The RNA polymerase III transcription apparatus. *J. Mol. Biol.* **310**, 1–26 (2001).
143. Kedmi, A. *et al.* Drosophila TRF2 is a preferential core promoter regulator. *Genes Dev.* **28**, 2163–2174 (2014).
144. Kutach, A. K. & Kadonaga, J. T. The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. *Mol. Cell. Biol.* **20**, 4754–4764 (2000).
145. Pennington, K. L., Marr, S. K., Chirn, G.-W. & Marr, M. T. Holo-TFIID controls the magnitude of a transcription burst and fine-tuning of transcription. *Proceedings of the National Academy of Sciences* **110**, 7678–7683 (2013).
146. Martianov, I., Velt, A., Davidson, G., Choukrallah, M.-A. & Davidson, I. TRF2 is recruited to the pre-initiation complex as a testis-specific subunit of TFIIA/ALF to promote haploid cell gene expression. *Sci Rep* **6**, 32069 (2016).
147. Suzuki, H., Isogai, M., Maeda, R., Ura, K. & Tamura, T.-A. TBP-like protein (TLP)

- interferes with Taspase1-mediated processing of TFIIA and represses TATA box gene expression. *Nucleic Acids Research* **43**, 6285–6298 (2015).
148. Wang, Y.-L. *et al.* TRF2, but not TBP, mediates the transcription of ribosomal protein genes. *Genes Dev.* **28**, 1550–1555 (2014).
 149. Isogai, Y., Keles, S., Prestel, M., Hochheimer, A. & Tjian, R. Transcription of histone gene cluster by differential core-promoter factors. *Genes Dev.* **21**, 2936–2949 (2007).
 150. Fan, W. *et al.* Drosophila TRF2 and TAF9 regulate lipid droplet size and phospholipid fatty acid composition. *PLoS Genetics* **13**, e1006664 (2017).
 151. Guglielmi, B., La Rochelle, N. & Tjian, R. Gene-specific transcriptional mechanisms at the histone gene cluster revealed by single-cell imaging. *Mol. Cell* **51**, 480–492 (2013).
 152. Persengiev, S. P. *et al.* TRF3, a TATA-box-binding protein-related factor, is vertebrate-specific and widely expressed. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 14887–14891 (2003).
 153. Bártfai, R. *et al.* TBP2, a vertebrate-specific member of the TBP family, is required in embryonic development of zebrafish. *Curr. Biol.* **14**, 593–598 (2004).
 154. Hart, D. O., Raha, T., Lawson, N. D. & Green, M. R. Initiation of zebrafish haematopoiesis by the TATA-box-binding protein-related factor Trf3. *Nature* **450**, 1082–1085 (2007).
 155. Hart, D. O., Santra, M. K., Raha, T. & Green, M. R. Selective interaction between Trf3 and Taf3 required for early development and hematopoiesis. *Dev. Dyn.* **238**, 2540–2549 (2009).
 156. Deato, M. D. E. & Tjian, R. Switching of the core transcription machinery during myogenesis. *Genes Dev.* **21**, 2137–2149 (2007).
 157. Malecová, B. *et al.* TBP/TFIID-dependent activation of MyoD target genes in skeletal muscle cells. *Elife* **5**, (2016).
 158. Jallow, Z., Jacobi, U. G., Weeks, D. L., Dawid, I. B. & Veenstra, G. J. C. Specialized and redundant roles of TBP and a vertebrate-specific TBP paralog in embryonic gene regulation in *Xenopus*. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 13525–13530 (2004).
 159. Duttke, S. H. C. Evolution and diversification of the basal transcription machinery. *Trends in Biochemical Sciences* **40**, 127–129 (2015).

160. Duttke, S. H. C., Doolittle, R. F., Wang, Y.-L. & Kadonaga, J. T. TRF2 and the evolution of the bilateria. *Genes Dev.* **28**, 2071–2076 (2014).
161. Murakami, K. *et al.* Formation and fate of a complete 31-protein RNA polymerase II transcription preinitiation complex. *Journal of Biological Chemistry* **288**, 6325–6332 (2013).
162. Buratowski, S., Hahn, S., Guarente, L. & Sharp, P. A. Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell* **56**, 549–561 (1989).
163. Kostrewa, D. *et al.* RNA polymerase II-TFIIB structure and mechanism of transcription initiation. *Nature* **462**, 323–330 (2009).
164. Plaschka, C. *et al.* Transcription initiation complex structures elucidate DNA opening. *Nature* **533**, 353–358 (2016).
165. Holstege, F. C., Fiedler, U. & Timmers, H. T. Three transitions in the RNA polymerase II transcription complex during initiation. *EMBO J.* **16**, 7468–7480 (1997).
166. Cramer, P., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* **292**, 1863–1876 (2001).
167. Kugel, J. F. & Goodrich, J. A. Translocation after synthesis of a four-nucleotide RNA commits RNA polymerase II to promoter escape. *Mol. Cell. Biol.* **22**, 762–773 (2002).
168. Luse, D. S. & Jacob, G. A. Abortive initiation by RNA polymerase II in vitro at the adenovirus 2 major late promoter. *J. Biol. Chem.* **262**, 14990–14997 (1987).
169. Zuo, Y. & Steitz, T. A. A structure-based kinetic model of transcription. *Transcription* **8**, 1–8 (2016).
170. Zawel, L., Kumar, K. P. & Reinberg, D. Recycling of the general transcription factors during RNA polymerase II transcription. *Genes Dev.* **9**, 1479–1490 (1995).
171. Hawley, D. K. & Roeder, R. G. Functional steps in transcription initiation and reinitiation from the major late promoter in a HeLa nuclear extract. *J. Biol. Chem.* **262**, 3452–3461 (1987).
172. Yudkovsky, N., Ranish, J. A. & Hahn, S. A transcription reinitiation intermediate that is stabilized by activator. *Nature* **408**, 225–229 (2000).
173. Rougvie, A. E. & Lis, J. T. The RNA polymerase II molecule at the 5' end of the uninduced hsp70 gene of *D. melanogaster* is transcriptionally engaged. *Cell* **54**, 795–804 (1988).

174. Gilmour, D. S. & Lis, J. T. RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in *Drosophila melanogaster* cells. *Mol. Cell. Biol.* **6**, 3984–3989 (1986).
175. Jonkers, I. & Lis, J. T. Getting up to speed with transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **16**, 167–177 (2015).
176. Danino, Y. M., Even, D., Ideses, D. & Juven-Gershon, T. The core promoter: At the heart of gene expression. *Biochim. Biophys. Acta* **1849**, 1116–1131 (2015).
177. Fuda, N. J., Ardehali, M. B. & Lis, J. T. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* **461**, 186–192 (2009).
178. Michel, M. & Cramer, P. Transitions for regulating early transcription. *Cell* **153**, 943–944 (2013).
179. Cosma, M. P. Ordered Recruitment. *Mol. Cell* **10**, 227–236 (2002).
180. Narlikar, G. J., Fan, H.-Y. & Kingston, R. E. Cooperation between complexes that regulate chromatin structure and transcription. *Cell* **108**, 475–487 (2002).
181. Imbalzano, A. N., Kwon, H., Green, M. R. & Kingston, R. E. Facilitated binding of TATA-binding protein to nucleosomal DNA. *Nature* **370**, 481–485 (1994).
182. Nechaev, S. & Adelman, K. Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation. *Biochim. Biophys. Acta* **1809**, 34–45 (2011).
183. Yean, D. & Gralla, J. Transcription reinitiation rate: a special role for the TATA box. *Mol. Cell. Biol.* **17**, 3809–3816 (1997).
184. Jacob, G. A., Kitzmiller, J. A. & Luse, D. S. RNA polymerase II promoter strength in vitro may be reduced by defects at initiation or promoter clearance. *J. Biol. Chem.* **269**, 3655–3663 (1994).
185. Ranish, J. A., Yudkovsky, N. & Hahn, S. Intermediates in formation and activity of the RNA polymerase II preinitiation complex: holoenzyme recruitment and a postrecruitment role for the TATA box and TFIIB. *Genes Dev.* **13**, 49–63 (1999).
186. Ljungman, M. & Lane, D. P. Opinion: Transcription — guarding the genome by sensing DNA damage. *Nature Reviews Cancer* **4**, 727–737 (2004).
187. Vousden, K. H. & Lane, D. P. p53 in health and disease. *Nat. Rev. Mol. Cell Biol.* **8**, 275–283 (2007).
188. Mack, D. H., Vartikar, J., Pipas, J. M. & Laimins, L. A. Specific repression of TATA-mediated but not initiator-mediated transcription by wild-type p53. *Nature*

- 363**, 281–283 (1993).
189. Morachis, J. M., Murawsky, C. M. & Emerson, B. M. Regulation of the p53 transcriptional response by structurally diverse core promoters. *Genes Dev.* **24**, 135–147 (2010).
 190. Suzuki, H., Ito, R., Ikeda, K. & Tamura, T.-A. TATA-binding protein (TBP)-like protein is required for p53-dependent transcriptional activation of upstream promoter of p21Waf1/Cip1 gene. *Journal of Biological Chemistry* **287**, 19792–19803 (2012).
 191. Saunders, A., Core, L. J., Sutcliffe, C., Lis, J. T. & Ashe, H. L. Extensive polymerase pausing during *Drosophila* axis patterning enables high-level and pliable transcription. *Genes Dev.* **27**, 1146–1158 (2013).
 192. Amir-Zilberstein, L. *et al.* Differential regulation of NF-kappaB by elongation factors is determined by core promoter type. *Mol. Cell. Biol.* **27**, 5246–5259 (2007).
 193. Hang, S. & Gergen, J. P. Different modes of enhancer-specific regulation by Runt and Even-skipped during *Drosophila* segmentation. *Mol. Biol. Cell* **28**, 681–691 (2017).
 194. Kouzine, F. *et al.* Global Regulation of Promoter Melting in Naive Lymphocytes. *Cell* **153**, 988–999 (2013).
 195. Haberle, V. *et al.* Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature* **507**, 381–385 (2014).
 196. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science* **322**, 1845–1848 (2008).
 197. Gilchrist, D. A. *et al.* Pausing of RNA Polymerase II Disrupts DNA-Specified Nucleosome Organization to Enable Precise Gene Regulation. *Cell* **143**, 540–551 (2010).
 198. Lee, C. *et al.* NELF and GAGA factor are linked to promoter-proximal pausing at many genes in *Drosophila*. *Mol. Cell. Biol.* **28**, 3290–3300 (2008).
 199. Min, I. M. *et al.* Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev.* **25**, 742–754 (2011).
 200. Nechaev, S. *et al.* Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. *Science* **327**, 335–338 (2010).

201. Rahl, P. B. *et al.* c-Myc Regulates Transcriptional Pause Release. *Cell* **141**, 432–445 (2010).
202. Zeitlinger, J. *et al.* RNA polymerase stalling at developmental control genes in the *Drosophila melanogaster* embryo. *Nat. Genet.* **39**, 1512–1516 (2007).
203. Gaertner, B. & Zeitlinger, J. RNA polymerase II pausing during development. *Development* **141**, 1179–1183 (2014).
204. Adelman, K. & Lis, J. T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet* **13**, 720–731 (2012).
205. Hendrix, D. A., Hong, J.-W., Zeitlinger, J., Rokhsar, D. S. & Levine, M. S. Promoter elements associated with RNA Pol II stalling in the *Drosophila* embryo. *Proceedings of the National Academy of Sciences* **105**, 7762–7767 (2008).
206. Fenouil, R. *et al.* CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome Research* **22**, 2399–2408 (2012).
207. Day, D. S. *et al.* Comprehensive analysis of promoter-proximal RNA polymerase II pausing across mammalian cell types. *Genome Biol.* **17**, 120 (2016).
208. Gaertner, B. *et al.* Poised RNA polymerase II changes over developmental time and prepares genes for future expression. *Cell Rep* **2**, 1670–1683 (2012).
209. Krumm, A., Hickey, L. B. & Groudine, M. Promoter-proximal pausing of RNA polymerase II defines a general rate-limiting step after transcription initiation. *Genes Dev.* **9**, 559–572 (1995).
210. Li, J. & Gilmour, D. S. Distinct mechanisms of transcriptional pausing orchestrated by GAGA factor and M1BP, a novel transcription factor. *EMBO J.* **32**, 1829–1841 (2013).
211. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950–953 (2013).
212. Lee, H., Kraus, K. W., Wolfner, M. F. & Lis, J. T. DNA sequence requirements for generating paused polymerase at the start of hsp70. *Genes Dev.* **6**, 284–295 (1992).
213. Glaser, R. L., Thomas, G. H., Siegfried, E., Elgin, S. C. & Lis, J. T. Optimal heat-induced expression of the *Drosophila* hsp26 gene requires a promoter sequence containing (CT)_n.(GA)_n repeats. *J. Mol. Biol.* **211**, 751–761 (1990).
214. Duarte, F. M. *et al.* Transcription factors GAF and HSF act at distinct regulatory

- steps to modulate stress-induced gene activation. *Genes Dev.* **30**, 1731–1746 (2016).
215. Boettiger, A. N. & Levine, M. Rapid transcription fosters coordinate snail expression in the *Drosophila* embryo. *Cell Rep* **3**, 8–15 (2013).
 216. Ghosh, S. K. B., Missra, A. & Gilmour, D. S. Negative elongation factor accelerates the rate at which heat shock genes are shut off by facilitating dissociation of heat shock factor. *Mol. Cell. Biol.* **31**, 4232–4243 (2011).
 217. Gilchrist, D. A. *et al.* Regulating the regulators: the pervasive effects of Pol II pausing on stimulus-responsive gene networks. *Genes Dev.* **26**, 933–944 (2012).
 218. Buckley, M. S., Kwak, H., Zipfel, W. R. & Lis, J. T. Kinetics of promoter Pol II on Hsp70 reveal stable pausing and key insights into its regulation. *Genes Dev.* **28**, 14–19 (2014).
 219. Jonkers, I., Kwak, H. & Lis, J. T. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife* **3**, e02407 (2014).
 220. Henriques, T. *et al.* Stable Pausing by RNA Polymerase II Provides an Opportunity to Target and Integrate Regulatory Signals. *Mol. Cell* **52**, 517–528 (2013).
 221. Mahat, D. B., Salamanca, H. H., Duarte, F. M., Danko, C. G. & Lis, J. T. Mammalian Heat Shock Response and Mechanisms Underlying Its Genome-wide Transcriptional Regulation. *Mol. Cell* **62**, 63–78 (2016).
 222. Liu, W. *et al.* Brd4 and JMJD6-Associated Anti-Pause Enhancers in Regulation of Transcriptional Pause Release. *Cell* **155**, 1581–1595 (2013).
 223. Chopra, V. S., Cande, J., Hong, J.-W. & Levine, M. Stalled Hox promoters as chromosomal boundaries. *Genes Dev.* **23**, 1505–1509 (2009).
 224. Pfeiffer, B. D. *et al.* Tools for neuroanatomy and neurogenetics in *Drosophila*. *Proceedings of the National Academy of Sciences* **105**, 9715–9720 (2008).
 225. Gehrig, J. *et al.* Automated high-throughput mapping of promoter-enhancer interactions in zebrafish embryos. *Nat. Methods* **6**, 911–916 (2009).
 226. Dickel, D. E. *et al.* Function-based identification of mammalian enhancers using site-specific integration. *Nat. Methods* **11**, 566–571 (2014).
 227. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**, 265–270 (2012).
 228. Arnold, C. D. *et al.* Quantitative genome-wide enhancer activity maps for five

- Drosophila* species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nat. Genet.* **46**, 685–692 (2014).
229. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
 230. Shlyueva, D. *et al.* Hormone-responsive enhancer-activity maps reveal predictive motifs, indirect repression, and targeting of closed chromatin. *Mol. Cell* **54**, 180–192 (2014).
 231. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).
 232. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**, 626–635 (2006).
 233. Consortium, T. F., PMI, T. R. & DGT, C. A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
 234. Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T. R. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Research* **23**, 169–180 (2013).
 235. Graveley, B. R. *et al.* The developmental transcriptome of *Drosophila melanogaster*. *Nature* **471**, 473–479 (2011).
 236. Frith, M. C. *et al.* A code for transcription initiation in mammalian genomes. *Genome Research* **18**, 1–12 (2008).
 237. Mwangi, S., Attardo, G., Suzuki, Y., Aksoy, S. & Christoffels, A. TSS seq based core promoter architecture in blood feeding Tsetse fly (*Glossina morsitans morsitans*) vector of Trypanosomiasis. *BMC Genomics* **16**, 722 (2015).
 238. Sandelin, A. *et al.* Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet* **8**, 424–436 (2007).
 239. Ohler, U. & Wassarman, D. A. Promoting developmental transcription. *Development* **137**, 15–26 (2010).
 240. Hansen, S. K. & Tjian, R. TAFs and TFIIA mediate differential utilization of the tandem *Adh* promoters. *Cell* **82**, 565–575 (1995).
 241. Sharpe, J., Nonchev, S., Gould, A., Whiting, J. & Krumlauf, R. Selectivity, sharing and competitive interactions in the regulation of *Hoxb* genes. *EMBO J.* **17**, 1788–1798 (1998).

242. Ohtsuki, S., Levine, M. & Cai, H. N. Different core promoters possess distinct regulatory activities in the *Drosophila* embryo. *Genes Dev.* **12**, 547–556 (1998).
243. Butler, J. E. & Kadonaga, J. T. Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes Dev.* **15**, 2515–2519 (2001).
244. Merli, C., Bergstrom, D. E., Cygan, J. A. & Blackman, R. K. Promoter specificity mediates the independent regulation of neighboring genes. *Genes Dev.* **10**, 1260–1270 (1996).
245. Kikuta, H. *et al.* Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Research* **17**, 545–555 (2007).
246. van Arensbergen, J., van Steensel, B. & Bussemaker, H. J. In search of the determinants of enhancer-promoter interaction specificity. *Trends Cell Biol.* **24**, 695–702 (2014).
247. Atkinson, T. J. & Halfon, M. S. Regulation of gene expression in the genomic context. *Comput Struct Biotechnol J* **9**, e201401001 (2014).
248. Choi, O. R. & Engel, J. D. Developmental regulation of beta-globin gene switching. *Cell* **55**, 17–26 (1988).
249. Li, X. & Noll, M. Compatibility between enhancers and promoters determines the transcriptional specificity of gooseberry and gooseberry neuro in the *Drosophila* embryo. *EMBO J.* **13**, 400–406 (1994).
250. Corbin, V. & Maniatis, T. The role of specific enhancer-promoter interactions in the *Drosophila* Adh promoter switch. *Genes Dev.* **3**, 2191–2120 (1989).
251. Devlin, B. H., Wefald, F. C., Kraus, W. E., Bernard, T. S. & Williams, R. S. Identification of a muscle-specific enhancer within the 5'-flanking region of the human myoglobin gene. *J. Biol. Chem.* **264**, 13896–13901 (1989).
252. Kermekchiev, M., Pettersson, M., Matthias, P. & Schaffner, W. Every enhancer works with every promoter for all the combinations tested: could new regulatory pathways evolve by enhancer shuffling? *Gene Expr.* **1**, 71–81 (1991).
253. Wefald, F. C., Devlin, B. H. & Williams, R. S. Functional heterogeneity of mammalian TATA-box sequences revealed by interaction with a cell-specific enhancer. *Nature* **344**, 260–262 (1990).
254. Yang-Tsung, C. & Keller, E. B. The TATA-dependent and TATA-independent promoters of the *Drosophila melanogaster* actin 5C-encoding gene. *Gene* **106**,

- 237–241 (1991).
255. Corbin, V. & Maniatis, T. The role of specific enhancer-promoter interactions in the *Drosophila* Adh promoter switch. *Genes Dev.* **3**, 2191–2120 (1989).
 256. Emami, K. H., Navarre, W. W. & Smale, S. T. Core promoter specificities of the Sp1 and VP16 transcriptional activation domains. *Mol. Cell. Biol.* **15**, 5906–5916 (1995).
 257. Conkright, M. D. *et al.* Genome-wide analysis of CREB target genes reveals a core promoter requirement for cAMP responsiveness. *Mol. Cell* **11**, 1101–1108 (2003).
 258. Ernst, P. *et al.* A potential role for Elf-1 in terminal transferase gene regulation. *Mol. Cell. Biol.* **16**, 6121–6131 (1996).
 259. Martinez, E. Core promoter-selective coregulators of transcription by RNA polymerase II. *Transcription* **3**, (2012).
 260. Simon, M. C., Fisch, T. M., Benecke, B. J., Nevins, J. R. & Heintz, N. Definition of multiple, functionally distinct TATA elements, one of which is a target in the hsp70 promoter for E1A regulation. *Cell* **52**, 723–729 (1988).
 261. Juven-Gershon, T., Hsu, J.-Y. & Kadonaga, J. T. Caudal, a key developmental regulator, is a DPE-specific transcriptional factor. *Genes Dev.* **22**, 2823–2830 (2008).
 262. Zehavi, Y., Kuznetsov, O., Ovadia-Shochat, A. & Juven-Gershon, T. Core promoter functions in the regulation of gene expression of *Drosophila* dorsal target genes. *Journal of Biological Chemistry* **289**, 11993–12004 (2014).
 263. Shir-Shapira, H. *et al.* Structure-Function Analysis of the *Drosophila melanogaster* Caudal Transcription Factor Provides Insights into Core Promoter-preferential Activation. *Journal of Biological Chemistry* **290**, 17293–17305 (2015).
 264. Hsu, J.-Y. J. *et al.* TBP, Mot1, and NC2 establish a regulatory circuit that controls DPE-dependent versus TATA-dependent transcription. *Genes Dev.* **22**, 2353–2358 (2008).
 265. Willy, P. J. A Basal Transcription Factor That Activates or Represses Transcription. *Science* **290**, 982–984 (2000).
 266. Lewis, B. A., Sims, R. J., III, Lane, W. S. & Reinberg, D. Functional Characterization of Core Promoter Elements: DPE-Specific Transcription Requires the Protein Kinase CK2 and the PC4 Coactivator. *Mol. Cell* **18**, 471–481 (2005).
 267. Xu, M. *et al.* Core promoter-selective function of HMGA1 and Mediator in

- Initiator-dependent transcription. *Genes Dev.* **25**, 2513–2524 (2011).
268. Stampfel, G. *et al.* Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature* **528**, 147–151 (2015).
 269. O'Kane, C. J. & Gehring, W. J. Detection in situ of genomic regulatory elements in *Drosophila*. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 9123–9127 (1987).
 270. Bellen, H. J. *et al.* P-element-mediated enhancer detection: a versatile method to study development in *Drosophila*. *Genes Dev.* **3**, 1288–1300 (1989).
 271. Bier, E. *et al.* Searching for pattern and mutation in the *Drosophila* genome with a P-lacZ vector. *Genes Dev.* **3**, 1273–1287 (1989).
 272. Siegal, M. L. & Hartl, D. L. Transgene Coplacement and high efficiency site-specific recombination with the Cre/loxP system in *Drosophila*. *Genetics* **144**, 715–726 (1996).
 273. Blackman, R. K., Sanicola, M., Raftery, L. A., Gillevet, T. & Gelbart, W. M. An extensive 3' cis-regulatory region directs the imaginal disk expression of decapentaplegic, a member of the TGF-beta family in *Drosophila*. *Development* **111**, 657–666 (1991).
 274. Tjian, R. & Maniatis, T. Transcriptional activation: a complex puzzle with few easy pieces. *Cell* **77**, 5–8 (1994).
 275. Chen, J. L., Attardi, L. D., Verrijzer, C. P., Yokomori, K. & Tjian, R. Assembly of recombinant TFIID reveals differential coactivator requirements for distinct transcriptional activators. *Cell* **79**, 93–105 (1994).
 276. Thut, C. J., Chen, J. L., Klemm, R. & Tjian, R. p53 transcriptional activation mediated by coactivators TAFII40 and TAFII60. *Science* **267**, 100–104 (1995).
 277. Takiya, S., Hui, C. C. & Suzuki, Y. A contribution of the core-promoter and its surrounding regions to the preferential transcription of the fibroin gene in posterior silk gland extracts. *EMBO J.* **9**, 489–496 (1990).
 278. Gorski, K., Carneiro, M. & Schibler, U. Tissue-specific in vitro transcription from the mouse albumin promoter. *Cell* **47**, 767–776 (1986).
 279. Chen, Z. & Manley, J. L. Core promoter elements and TAFs contribute to the diversity of transcriptional activation in vertebrates. *Mol. Cell. Biol.* **23**, 7350–7362 (2003).
 280. Lorch, Y., LaPointe, J. W. & Kornberg, R. D. Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. *Cell*

- 49**, 203–210 (1987).
281. Han, M. & Grunstein, M. Nucleosome loss activates yeast downstream promoters in vivo. *Cell* **55**, 1137–1145 (1988).
 282. Prioleau, M. N., Huet, J., Sentenac, A. & Méchali, M. Competition between chromatin and transcription complex assembly regulates gene expression during early development. *Cell* **77**, 439–449 (1994).
 283. Workman, J. L. & Roeder, R. G. Binding of transcription factor TFIID to the major late promoter during in vitro nucleosome assembly potentiates subsequent initiation by RNA polymerase II. *Cell* **51**, 613–622 (1987).
 284. Nagai, S., Davis, R. E., Mattei, P.-J., Eagen, K. P. & Kornberg, R. D. Chromatin potentiates transcription. *Proceedings of the National Academy of Sciences* **114**, 1536–1541 (2017).
 285. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 15776–15781 (2003).
 286. Lubliner, S. *et al.* Core promoter sequence in yeast is a major determinant of expression level. *Genome Research* **25**, 1008–1017 (2015).
 287. Juven-Gershon, T., Cheng, S. & Kadonaga, J. T. Rational design of a super core promoter that enhances gene expression. *Nat. Methods* **3**, 917–922 (2006).
 288. Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).
 289. Kraus, R. J. *et al.* Experimentally determined weight matrix definitions of the initiator and TBP binding site elements of promoters. *Nucleic Acids Research* **24**, 1531–1539 (1996).
 290. Revyakin, A. *et al.* Transcription initiation by human RNA polymerase II visualized at single-molecule resolution. *Genes Dev.* **26**, 1691–1702 (2012).
 291. Burke, T. W. & Kadonaga, J. T. The downstream core promoter element, DPE, is conserved from *Drosophila* to humans and is recognized by TAFII60 of *Drosophila*. *Genes Dev.* **11**, 3020–3031 (1997).
 292. O'Shea-Greenfield, A. & Smale, S. T. 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 (1992).
 293. Colgan, J. & Manley, J. L. Cooperation between core promoter elements

- influences transcriptional activity in vivo. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 1955–1959 (1995).
294. Portela, R. M. C. *et al.* Synthetic Core Promoters as Universal Parts for Fine-Tuning Expression in Different Yeast Species. *ACS Synthetic Biology* **6**, 471–484 (2017).
 295. Even, D. Y. *et al.* Engineered Promoters for Potent Transient Overexpression. *PLoS One* **11**, e0148918 (2016).
 296. Ede, C., Chen, X., Lin, M.-Y. & Chen, Y. Y. Quantitative Analyses of Core Promoters Enable Precise Engineering of Regulated Gene Expression in Mammalian Cells. *ACS Synthetic Biology* **5**, 395–404 (2016).
 297. Lubliner, S., Keren, L. & Segal, E. Sequence features of yeast and human core promoters that are predictive of maximal promoter activity. *Nucleic Acids Research* **41**, 5569–5581 (2013).
 298. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
 299. Denissov, S. *et al.* Identification of novel functional TBP-binding sites and general factor repertoires. *EMBO J.* **26**, 944–954 (2007).
 300. Neri, F. *et al.* Intragenic DNA methylation prevents spurious transcription initiation. *Nature* **543**, 72–77 (2017).
 301. Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
 302. De Santa, F. *et al.* A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol.* **8**, e1000384 (2010).
 303. Scruggs, B. S. *et al.* Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Mol. Cell* **58**, 1101–1112 (2015).
 304. Hah, N. *et al.* A Rapid, Extensive, and Transient Transcriptional Response to Estrogen Signaling in Breast Cancer Cells. *Cell* **145**, 622–634 (2011).
 305. Michel, M. *et al.* TT-seq captures enhancer landscapes immediately after T-cell stimulation. *Mol. Syst. Biol.* **13**, 920 (2017).
 306. Schwalb, B. *et al.* TT-seq maps the human transient transcriptome. *Science* **352**, 1225–1228 (2016).
 307. Wu, H. *et al.* Tissue-specific RNA expression marks distant-acting developmental enhancers. *PLoS Genetics* **10**, e1004610 (2014).

308. Seila, A. C. *et al.* Divergent Transcription from Active Promoters. *Science* **322**, 1849–1851 (2008).
309. Bose, D. A. *et al.* RNA Binding to CBP Stimulates Histone Acetylation and Transcription. *Cell* **168**, 135–149.e22 (2017).
310. Juven-Gershon, T. & Kadonaga, J. T. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Developmental Biology* **339**, 225–229 (2010).
311. Zabidi, M. A. *et al.* Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556–559 (2015).
312. Zabidi, M. A. & Stark, A. Regulatory Enhancer–Core-Promoter Communication via Transcription Factors and Cofactors. *Trends Genet.* **32**, 801–814 (2016).
313. Arnold, C. D. *et al.* Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. *Nat. Biotechnol.* **35**, 136–144 (2016).
314. Font-Burgada, J., Rossell, D., Auer, H. & Azorín, F. Drosophila HP1c isoform interacts with the zinc-finger proteins WOC and Relative-of-WOC to regulate gene expression. *Genes Dev.* **22**, 3007–3023 (2008).
315. Matova, N. & Anderson, K. V. Rel/NF-kappaB double mutants reveal that cellular immunity is central to Drosophila host defense. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 16424–16429 (2006).
316. Baonza, A., Murawsky, C. M., Travers, A. A. & Freeman, M. Pointed and Tramtrack69 establish an EGFR-dependent transcriptional switch to regulate mitosis. *Nat. Cell Biol.* **4**, 976–980 (2002).
317. Shwartz, A., Yogev, S., Schejter, E. D. & Shilo, B.-Z. Sequential activation of ETS proteins provides a sustained transcriptional response to EGFR signaling. *Development* **140**, 2746–2754 (2013).
318. Kim, Y. J., Bjorklund, S., Li, Y., Sayre, M. H. & Kornberg, R. D. A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II. *Cell* **77**, 599–608 (1994).
319. Taatjes, D. J. The human Mediator complex: a versatile, genome-wide regulator of transcription. *Trends in Biochemical Sciences* **35**, 315–322 (2010).
320. Allen, B. L. & Taatjes, D. J. The Mediator complex: a central integrator of transcription. *Nat. Rev. Mol. Cell Biol.* **16**, 155–166 (2015).

321. Conaway, R. C. & Conaway, J. W. Origins and activity of the Mediator complex. *Seminars in Cell & Developmental Biology* **22**, 729–734 (2011).
322. Poss, Z. C., Ebmeier, C. C. & Taatjes, D. J. The Mediator complex and transcription regulation. *Crit. Rev. Biochem. Mol. Biol.* **48**, 575–608 (2013).
323. Cubeñas-Potts, C. *et al.* Different enhancer classes in *Drosophila* bind distinct architectural proteins and mediate unique chromatin interactions and 3D architecture. *Nucleic Acids Research* (2016). doi:10.1093/nar/gkw1114
324. Falender, A. E. *et al.* Maintenance of spermatogenesis requires TAF4b, a gonad-specific subunit of TFIID. *Genes Dev.* **19**, 794–803 (2005).
325. Voronina, E. *et al.* Ovarian granulosa cell survival and proliferation requires the gonad-selective TFIID subunit TAF4b. *Developmental Biology* **303**, 715–726 (2007).
326. Geles, K. G. *et al.* Cell-type-selective induction of c-jun by TAF4b directs ovarian-specific transcription networks. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 2594–2599 (2006).
327. Starr, B. D., Hoopes, B. C. & Hawley, D. K. DNA Bending is an Important Component of Site-specific Recognition by the TATA Binding Protein. *J. Mol. Biol.* **250**, 434–446 (1995).
328. Arkova, O. V., Kuznetsov, N. A., Fedorova, O. S., Kolchanov, N. A. & Savinkova, L. K. Real-Time Interaction between TBP and the TATA Box of the Human Triosephosphate Isomerase Gene Promoter in the Norm and Pathology. *Acta Naturae* **6**, 36–40 (2014).
329. Evans, R. Activator-mediated disruption of sequence-specific DNA contacts by the general transcription factor TFIIB. *Genes Dev.* **15**, 2945–2949 (2001).
330. Wolner, B. S. & Gralla, J. D. Roles for non-TATA core promoter sequences in transcription and factor binding. *Mol. Cell. Biol.* **20**, 3608–3615 (2000).
331. Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B. & Sharp, P. A. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**, 360–363 (2013).
332. Proudfoot, N. J. & Brownlee, G. G. 3' Non-coding region sequences in eukaryotic messenger RNA. *Nature* **263**, 211–214 (1976).
333. Tamura, T., Sumita, K., Hirose, S. & Mikoshiba, K. Core promoter of the mouse myelin basic protein gene governs brain-specific transcription in vitro. *EMBO J.* **9**,

- 3101–3108 (1990).
334. Dierich, A., Gaub, M. P., LePennec, J. P., Astinotti, D. & Chambon, P. Cell-specificity of the chicken ovalbumin and conalbumin promoters. *EMBO J.* **6**, 2305–2312 (1987).
335. Theill, L. E., Wiborg, O. & Vuust, J. Cell-specific expression of the human gastrin gene: evidence for a control element located downstream of the TATA box. *Mol. Cell. Biol.* **7**, 4329–4336 (1987).
336. Tamura, T. *et al.* Tissue-specific in vitro transcription from the mouse myelin basic protein promoter. *Mol. Cell. Biol.* **9**, 3122–3126 (1989).