Large-scale analysis of *Drosophila* core promoter function using synthetic promoters

Zhan Qi

München 2019

Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

Large-scale analysis of *Drosophila* core promoter function using synthetic promoters

Zhan Qi
aus
Nanjing, Jiangsu, China

2019

Erklärung:

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Herrn Prof. Dr. Veit Hornung betreut.

Eidesstattliche Versicherung:

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, den 10.05.2019

Zhan Qi

Dissertation eingereicht am 02.04.2019

1. Gutachter: Prof. Dr. Veit Hornung

2. Gutachter: Prof. Dr. Karl-Peter Hopfner

Mündliche Prüfung am 30.04.2019

TABLE OF CONTENTS

ACKNOWLEDGMENTS

First of all, I would like to thank my supervisor Prof. Dr. Ulrike Gaul for giving me the opportunity to work on this exciting topic in her lab. I am very grateful for her support and trust throughout my Ph.D. work. I also would like to express my gratitude to Prof. Dr. Veit Hornung for stepping in during the last years and supporting me as my thesis advisor. I am grateful to other members of my TAC committee: Prof. Dr. Andreas Ladurner and Dr. Stefan Canzar. Their insightful suggestions and comments were constructive for refining the project. I also would like to thank Prof. Dr. Karl-Peter Hopfner, Prof. Dr. Julian Stingele and PD Dr. Dietmar Martin for their time and support as my examination board members.

I am indebted to Dr. Christophe Jung and Ulrich Unnerstall for their guidance through the project. They provided enlightening instructions and helpful suggestions for my project. I also appreciate the help from Dr. Jung on the detailed corrections for this thesis. Moreover, I thank Dr. Johannes Söding, Dr. Julia Philippou-Massier, Dr. Katja Frühauf and Dr. Miroslav Nikolov for their valuable early effort in this project.

During these years, I have learned a lot and got extensive help from Peter Bandilla for operating the liquid handling automation, from Claudia Ludwig for her experimental expertise, and from Dr. Mark Heron and Dr. Anja Sophie Kiesel for their execution of the XXmotif on core promoter sequence design and the initial data analysis. A huge thank goes to all of them. I also would like to express my big thanks to all the previous and current members of the Gaul lab, especially Dr. Sara Batelli, Dr. Alessio Renna, Dr. Andrea Storti, Dr. Marta Bozek, Dr. Roberto Cortini, Monika Hanf, Christine Rottenberger, Stefano Ceolin, Max Schnepf, Liliana Pedro, Shao-Yen Kao. I really miss our joyful lunch breaks and weekly cake schedules.

In addition, I am thankful for the Graduate School of Quantitative Bioscience Munich for their support of scientific study and creating the environment to connect and share thoughts with other great Ph.D. students.

Last but not least, I would like to thank my family, especially my parents and my grandparents. I could not achieve this without their endless support and encouragement. A special thank goes to my dear husband, Mingyan Wang. He can always find a way to comfort me and make me happy. We are the best partners in a lot of aspects. Thank you for always being by my side, and I'm sure that we will build a sweet home full of joy and success.

SUMMARY

The core promoter comprises the transcription start site (TSS) and approximately 150 bp of the flanking sequence. The accurate transcription initiation and basal expression level of a gene are primarily determined by differential recruitment of the transcription machinery, which is also known as the pre-initiation complex (PIC) consisting of the RNA polymerase II (Pol II) together with the general transcription factors (GTFs), to its core promoter region.

Genome-wide studies have revealed various properties of native core promoters, including a crucial sequence feature called sequence motifs which enrich as over-represented DNA sequences that mark the potential binding sites of GTFs. Correlating these motifs to gene sets with distinct expression features allowed the fine-grained classification of core promoter into regulated/stalled, highly regulated, housekeeping and ribosomal architectures, which also showed different chromatin properties. Genetic variations naturally occurred at the motif sites were recently proved to alter both promoter strength and TSS distribution significantly. Despite the enormous importance of the core promoter and its sequence features, how they encode or compute the intrinsic expression levels remains poorly understood. The systematic mutational approach which perturbs native sequences and measures consequent effects is a powerful tool to ascertain the functional influence of specific features on promoter activity. In this thesis, I report on a large-scale luciferase-assay-based method developed to quantitatively measure promoter activity with high reproducibility and sensitivity in the well-studied experimental model *Drosophila melanogaster* (*D. melanogaster*). We applied this technique to measure both basal and induced expressions of systematically designed promoters in *D. melanogaster*, decoding the sequence determinants of their activity.

The synthetic promoter constructs consist of: (1) a motif-rich core promoter region of 130 bp around TSS from our designed library with thousands of native and perturbative sequences representing different core promoter architectures; (2) a stimulus-response element for binding of the ecdysone receptors to recruit the steroid hormone ecdysone for transcriptional activation; and (3) the genomic -1 and +1 nucleosome positioning sequences to mimic the endogenous nucleosomal context. A high-throughput experimental pipeline using automated robot systems was implemented and optimized for reporter plasmids isolation, *Drosophila* S2 cell transient transfection, ecdysone treatment and dual luciferase assay, which enabled highly reproducible measurements of promoter activity. The entire measurements of all tested synthetic promoters covered a wide range over more than four orders of magnitude in activity level.

By extensively testing mutagenized core promoter sequences, we corroborated the functional specificity of sequence motifs and that their adequate strength (PWM score) and precise positioning are essential features of core promoter activity. Additionally, our highly sensitive measurements of single base pair mutations could be used to produce the expression-based position probability matrices (PPMs) and activity logos for core promoter motifs. The context sequences surrounding the motifs also played a role but usually less prominent in defining the activity.

Moreover, combinatorial motif mutations that altered both the strength and the positioning of all motifs often resulted in strong effects, which were then compared with the effects of individual motif mutations. Remarkably, we found a linear combination of these individual motif features could largely (~ 77%) account for the combinatorial effects on core promoter activity. When applying a similar analysis to the combination of sequence feature blocks containing motifs together with their flanking and context sequences, 66% of the variance in expression levels could be linearly explained.

Finally, we showed that the surrounding sequences of the core promoter region also influenced promoter activity, especially for the ecdysone response element (EcRE). The ecdysone responsiveness correlated with the core promoter architecture, that is, ecdysone could induce both developmental and constitutive core promoters but the induction was stronger with the developmental ones. We also found a negative correlation between the ecdysone inducibility and the basal expression level; this correlation was more significant for constitutive promoters. Finally, by testing the nucleosomal context sequences, we found that the TSS downstream nucleosome positioning sequence had a stronger influence on constitutive core promoter activity.

Overall, this large-scale quantitative core promoter activity analysis enabled the first comprehensive dissection of *Drosophila* core promoter features and shed light on their roles for better predictability of gene expression.

I INTRODUCTION

1. Transcriptional regulation and core promoters

Appropriate gene expression with the correct timing in the precise spatial range is crucial for the development, evolution and diversity of all organisms. The control of gene expression occurs primarily at the process of transcription (Levine & Tjian, 2003), in which the genetic information is conveyed from DNA to RNA. Binding of regulatory proteins known as transcription factors (TFs) to non-coding *cis*-regulatory elements (CREs) including promoters and enhancers fundamentally regulates the transcription of genes. In eukaryotic cells, the accessibility of CREs is determined by the local chromatin configuration, as most of genomic DNA is wrapped around histone octamers to form nucleosomes (Kornberg & Lorch, 1999). Active CREs often locate at the nucleosome-depleted region (NDR). They support the assembly of transcription machinery to initiate transcription and mediate binding of other TFs together with cofactors to further activate or suppress transcription (Hampsey, 1998; Roeder, 1996; Spitz & Furlong, 2012; Zabidi & Stark, 2016). The disruption of these CREs often associates with common diseases (Kundaje et al., 2015; Maurano et al., 2012). The RNA polymerase II (Pol II) core promoter is the minimal DNA sequence that is recognized by the basal transcription machinery to drive accurate transcription initiation of protein-coding genes (Juven-Gershon, Hsu, Theisen, & Kadonaga, 2008; Smale & Kadonaga, 2003; Thomas & Chiang, 2006). It makes an essential contribution for setting the gene expression level.

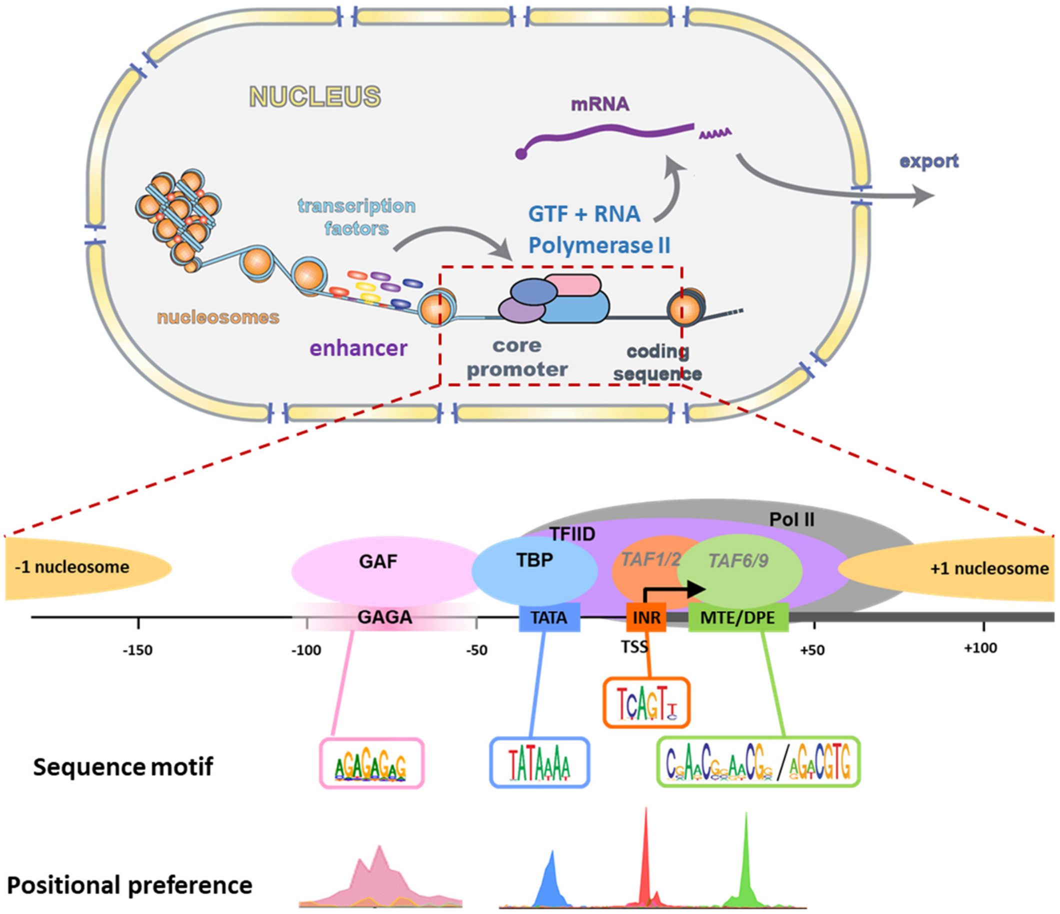


Figure 1. Transcriptional regulation at core promoter. Binding of transcription factors (TFs) to non-coding *cis*-regulatory elements (CREs) located at the nucleosome-depleted region (NDR) including promoters and enhancers fundamentally regulates the transcript synthesis. The core promoter comprises the transcription start site (TSS) and its flanking sequence around 150 bp. It is the binding site of the RNA polymerase II (Pol II) and the general transcription factors (GTFs) for accurate transcription initiation. The core promoter region contains various over-represented sequence motifs with distinct positional preferences, such as well-positioned INR and broadly-distributed GAGA.

1.1 Transcription initiation at core promoters

The core promoter comprises the transcription start site (TSS) and approximately 150 bp of the flanking sequence. It functions as the recognition and landing site for Pol II along with the general transcription factors (GTFs) to start transcription (Figure 1). The major GTFs required to form this pre-initiation complex (PIC) include TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH, among which TFIID binds primarily to core promoter elements and helps to nucleate the PIC (Orphanides, Lagrange, & Reinberg, 1996; Sainsbury, Bernecky, & Cramer, 2015; Thomas & Chiang, 2006). In general, subunits of TFIID including TATA-Box-binding protein (TBP) and TBP-associated factors (TAFs) firstly recognize and bind to specific sequences in core promoter region like TATA-Box, initiator (INR) and downstream promoter element (DPE). The TFIID-core-promoter complex is then recognized and stabilized by TFIIA and TFIIB, followed by the recruitment of Pol II-TFIIF complex. Finally, the binding of TFIIE and TFIIH complete the PIC assembly. The core promoter sequence melts and the transcription bubble is formed, allowing Pol II to initiate transcription and nascent transcript synthesis (Louder et al., 2016; Plaschka et al., 2016). In addition to this canonical view, diverse core promoter architectures as well as PIC compositions significantly contribute to cell-type-specific and gene-specific transcriptional regulation (Baptista et al., 2017; Goodrich & Tjian, 2010; Hansen, Takada, Jacobson, Lis, & Tjian, 1997; Hochheimer, Zhou, Zheng, Holmes, & Tjian, 2002a; Parry et al., 2010; Rabenstein, Zhou, Lis, & Tjian, 1999). For example, a TFIID-independent transcription driven by TBP-related factor 2 (TRF2) via TCT motif at the TSS was found in most of *Drosophila* ribosomal protein genes (Y.-L. Wang et al., 2014).

Pol II stalling usually occurs between transcription initiation and productive elongation after transcribing around 30-50 nucleotides of nascent RNA, which is also a step that limits transcription rates (Lis, 1998; Rougvie & Lis, 1988). Pol II pauses at the promoter-proximal region of many stimuli-responsive and developmental genes with various duration time (Gressel et al., 2017; Krebs et al., 2017; Muse et al., 2007; Shao & Zeitlinger, 2017; Zeitlinger et al., 2007), indicating its role at another layer of transcriptional regulation. The elongation factors including DRB sensitivity-inducing factor (DSIF) and negative elongation factor (NELF) are involved in triggering Pol II stalling (Missra & Gilmour, 2010; Qiu & Gilmour, 2017). In addition, specific sequence features and nucleosome organizations downstream of the TSS are sometimes required for paused Pol II as well (Hendrix, Hong, Zeitlinger, Rokhsar, & Levine, 2008; Weber, Ramachandran, & Henikoff, 2014).

1.2 TSS distribution and promoter shape

Various genome-wide methods, including cap analysis of gene expression (CAGE) (Shiraki et al., 2003; Takahashi, Lassmann, Murata, & Carninci, 2012), 5’ serial analysis of gene expression (5’ SAGE) (Hashimoto et al., 2004; Wei et al., 2004; Zhang & Dietrich, 2005) and other similar approaches (Gu et al., 2012; Ni et al., 2010; Valen et al., 2009), characterize endogenous TSS distributions based on analysis of the 5’ ends of transcripts and identify transcription initiation patterns together with corresponding core promoter regions in different organisms (Ahsan et al., 2009; Carninci et al., 2006; Chen et al., 2013; Forrest et al., 2014; Haberle et al., 2014; Hoskins et al., 2011; Ni et al., 2010).

Based on the distribution of TSSs, there are generally two patterns of transcription initiation: focused and dispersed. Focused initiation pattern has a single TSS or a narrow region of TSSs within ~ 5 bp. In contrast, multiple weak TSSs distribute over a wide region of around 50-100 bp in dispersed initiation pattern (Carninci et al., 2006; Juven-Gershon et al., 2008; Kadonaga, 2012). The corresponding core promoter types are termed as narrow peak (NP) for the focused transcription initiations and broad peak (BP) for the dispersed ones. Distinct promoter shapes are associated with gene-specific and sequence-specific features. Developmentally regulated and tissue-specific genes mostly contain NP promoters with strictly positioned core promoter motifs like TATA-Box, INR, motif ten element (MTE) and DPE (Rach, Yuan, Majoros, Tomancak, & Ohler, 2009), whereas BP promoters is mainly associated with housekeeping genes and tend to have weakly positioned motifs including CpG islands in mammals and DNA replication-related element (DRE), Ohler1, Ohler6 in *Drosophila melanogaster* (*D. melanogaster*) (Carninci et al., 2006; Hoskins et al., 2011; Rach et al., 2009). In addition, NP promoters were found to have a higher GC content than BP promoters (Rach et al., 2009). The two types of promoters also differ in nucleosome organization and Pol II stalling. NP promoters are usually characterized with imprecisely located nucleosomes which facilitate paused Pol II (Kwak, Fuda, Core, & Lis, 2013; Nechaev et al., 2010; Rach et al., 2011). Moreover, promoter shape is a widely conserved feature between species. Compared to evolutionary-constrained NP promoters, BP promoters are able to maintain their shape feature as well as promoter activity when a genetic variation affects one TSS owing to the buffer functions of other TSSs within the same distribution (Schor et al., 2017).

1.3 Core promoter motifs in *D. melanogaster*

Various short sequences with distinct functions constitute the core promoter. They are known as sequence motifs, which were mainly discovered by computational identification of over-represented sequences in the core promoter regions (Figure 1). Most of them serve as the binding sites for GTFs and other TFs that mediate PIC assembly and subsequent transcriptional processes. Natural genetic variants occurred at the motif sites were recently proved to alter both promoter strength and TSS distribution significantly (Schor et al., 2017). Nevertheless, the motif composition in promoter regions of genes with different functions in different species is usually diverse and non-universal. Besides, there are still many core promoters containing no known motifs. In this section, the main core promoter motifs found in *D. melanogaster* are discussed as follows.

The INR motif is one of the most frequently used motifs and it usually marks the TSS of focused initiation at "A" of its 3rd position (FitzGerald, Sturgill, Shyakhtenko, Oliver, & Vinson, 2006; Ohler, Liao, Niemann, & Rubin, 2002). The INR consensus in flies (TCAGTY) is more restrictive compared to that in human core promoters. It is mainly recognized and bound by TFIID subunits TAF1 and TAF2. The first discovered eukaryotic core promoter motif is TATA-Box (Lifton, Goldberg, Karp, & Hogness, 1978), which is a highly conserved A/T rich sequence located at around 25-30 bp upstream of the TSS. It also enriches in NP promoters but is less abundant than INR. TBP binds to TATA-Box and helps the Pol II recruitment. For core promoters that contain INR but are lack of TATA-Box, the DPE motif often occurs, which is positioned strictly ~ 30 bp downstream of the TSS (Burke & Kadonaga, 1997). It serves as the binding sites for another two TAF subunits of TFIID: TAF6 and TAF9. Similar to the required spacing between INR and TATA-Box for synergistic binding of TFIID (Emami, Jain, & Smale, 1997), the precise spacing between INR and DPE also coordinate the binding of TFIID to initiate transcription properly. Another downstream core promoter motif with a strong positional preference (+18 to +29 relative to TSS, sometimes overlapping with DPE) is MTE, which usually needs an INR and shows synergism with both TATA-Box and DPE (Lim et al., 2004). Which factor binds to it is still not clear, although the structure analysis of promoter-bound human TFIID revealed potential interactions of TAF1 and TAF2 with this region (Louder et al., 2016). The synergism of these four well-characterized and well-positioned motifs (TATA-Box, INR, MTE and DPE) was utilized for the construction of a super synthetic core promoter with high transcriptional activity (Juven-Gershon, Cheng, & Kadonaga, 2006).

In addition to the TFIID that binds to the core promoter motifs, another GTF TFIIB also interacts with the core promoter. The DNA sequences bound by TFIIB are next to the TATA-Box, known as TFIIB recognition element BRE^u and BRE^d located upstream and downstream of the TATA-Box, respectively (Deng & Roberts, 2005; Lagrange, Kapanidis, Tang, Reinberg, &

Ebright, 1998). Disruptions of either motif in different promoter contexts showed positive or negative effects on basal transcription level.

The core promoter motifs recognized by other TFs for proper transcriptional regulation include DRE and Ohler1. Both of them show rare positional preferences. DRE is an 8 bp palindromic sequence motif and is essential for cell-cycle and cell-proliferation genes regulation together with its binding TF called DREF (Hirose, Yamaguchi, Handa, Inomata, & Matsukage, 1993). It is thought to have the similar role of a TATA-Box since it can recruit DREF-associated TRF2 to specifically initiate transcription of constitutive genes (Hochheimer, Zhou, Zheng, Holmes, & Tjian, 2002b; Kopytova et al., 2006). Ohler1, also known as motif 1, was found initially by computational analysis of many core promoters as an over-represented sequence (Ohler et al., 2002). It enriches in BP promoters and is bound by a zinc-finger protein - motif 1 binding protein, M1BP (Li & Gilmour, 2013; Ohler, 2006; Rach et al., 2009). Ohler1 is also important for Pol II stalling that differs in the mechanism of GAGA-enriched paused genes. GAGA element and its binding protein GAF are mostly involved in paused Pol II at upstream of the +1 nucleosome in focused initiation pattern while M1BP-bound Ohler1 drives less-efficient Pol II stalling in BP promoters that is probably affected by the +1 nucleosome obstacle (Fuda & Lis, 2013; Li & Gilmour, 2013). Notably, another core promoter motif pause button (PB) also shows enrichment in stalled promoters containing the GAGA motif (Hendrix et al., 2008). E-Box that is bound by basic helix-loop-helix leucine zipper (bHLH-zip) transcription factors and sometimes located around the TSS is considered as a core promoter motif as well (FitzGerald et al., 2006).

There are several core promoter motifs still lacking the knowledge of their binding factors. One of the examples is the TCT motif that exists in almost all ribosomal protein gene promoters in *Drosophila* (Parry et al., 2010). It is also known as the polypyrimidine initiator with a consensus of YYCTTTYY. Transcription usually starts at “C” at the 3rd position and is mediated by TRF2 instead of TBP which is commonly used in TATA-Box dependent process (Y.-L. Wang et al., 2014). Besides, TCT motif can be converted into an active INR through a single T-to-A substitution. Ohler6 and Ohler7 are the other two core promoter motifs with unknown binding proteins (Ohler et al., 2002). They are computationally defined motifs with weak location bias. Furthermore, Ohler6 was found to co-occur with Ohler1 with a preferred spacing and the combination of them was postulated to function as an alternative of TATA-Box + INR pair (Ohler, 2006). Similarly, Ohler7 also tend to associate with DRE in BP promoters (Rach et al., 2009).

1.4 Core promoter types and enhancer–core-promoter specificity

Based on the motif co-occurrence in *D. melanogaster*, five core promoter modules were initially defined: TATA-Box + INR pair, INR + DPE pair, Ohler1 + Ohler6 pair, DRE only and INR only (Ohler, 2006). By associating with gene functions, these five modules can be further classified into three major types including tissue-specific core promoters with focused initiation patterns and an enrichment of TATA-Box and INR motifs; core promoters of ubiquitously expressed housekeeping genes with dispersed initiation patterns and broadly positioned motifs like Ohler1, Ohler6 and DRE; core promoters containing only INR or INR + DPE pair and associated with developmentally regulated genes (Engström, Sui, Drivenes, Becker, & Lenhard, 2007; Lenhard, Sandelin, & Carninci, 2012). The TCT motif occurrence (rarely existed, e.g., only in around 1% of NP promoters in humans; Vo Ngoc, Cassidy, Huang, Duttke, & Kadonaga, 2017) determines one extra minor type of core promoter with a focused initiation pattern.

Various types of core promoters not only differ in gene expression features but also influence transcriptional regulation via differential responses to enhancers. Certain kinds of enhancers activate transcription specifically from either TATA-Box-dependent core promoters or DPE-dependent core promoters (Butler & Kadonaga, 2001). Genome-wide assessment of enhancer activity upon housekeeping core promoters and developmental core promoters in *Drosophila* also suggests enhancer preferences which distinguish the two modes of transcription programs (Zabidi et al., 2015). Complementarily, by testing the responsiveness of a large number of core promoters to developmental or housekeeping enhancers, sequence-encoded specificity is again confirmed (Arnold et al., 2017).

1.5 Chromatin features of core promoters

The basal transcription machinery competes with nucleosomes, although characterized with low occupancy, to have access to the core promoter region for proper initiation of transcription. Core promoters with different initiation patterns and distinct motif content vary in nucleosome organization (Figure 2). BP promoters of housekeeping genes show the canonical nucleosome pattern where TSSs are largely depleted from nucleosomes (known as NDRs). The NDR is flanked by a strongly positioned -1 nucleosome upstream (usually sensitive to MNase digestion) and a well-positioned +1 nucleosome downstream of the TSS. Moreover, the +1 nucleosome is followed by a regular phasing of downstream nucleosomes (Figure 2A). Core promoters with TCT motifs have similar nucleosome patterns although they are more often characterized as focused initiation. In contrast, NP promoters containing TATA-Box, INR, MTE or DPE are associated with a disordered nucleosome organization (Figure 2B; Mavrich et al., 2008; Rach et al., 2011). The two distinct nucleosome patterns also differ in the dinucleotide frequencies around TSSs, albeit nucleosomes themselves generally show low AT content (Figure 2). The +1 nucleosome in canonical patterns appears with a sharp decrease of AA/TT dinucleotide frequencies. Furthermore, a linear model based on DNA sequence features found that the GC content plays a pivotal role in nucleosome occupancy in vitro (Tillo & Hughes, 2009), which can be explained by the correlated structural properties of DNA and reduced frequencies of poly(dA:dT) tracts.

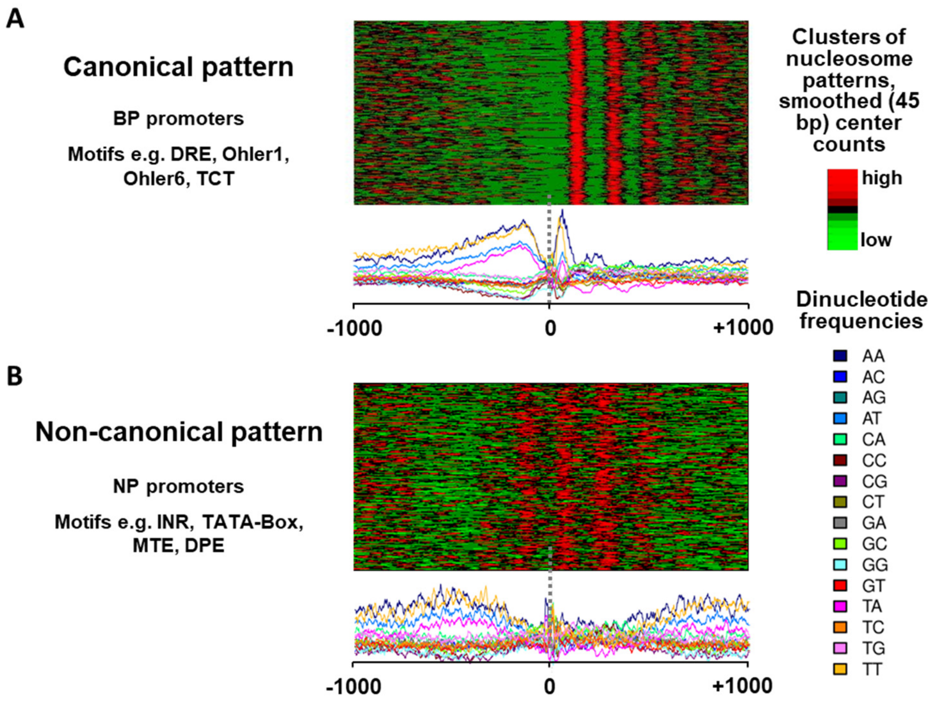


Figure 2. Nucleosome organization and dinucleotide frequencies around TSS (± 1 kb window) in *Drosophila* (bulk nucleosome mapping by MNase-seq). The AT content is generally lower in nucleosomes compared to linker DNAs. **(A)** The canonical nucleosome pattern usually found in BP promoters. TSSs are depleted from nucleosomes and are associated with a very strongly positioned +1 nucleosome downstream, followed by a regular phasing of nucleosomes. The dinucleotide landscapes show that +1 nucleosome has a strong correlation with its increased upstream AA/TT dinucleotide frequencies. **(B)** The non-canonical nucleosome pattern in NP promoters with disordered nucleosomes around TSSs.

The histone variants H3.3 and H2A.Z usually mark the NDRs at core promoters and other CREs (Jin et al., 2009; Mavrich et al., 2008), suggesting that NDRs may contain dynamic or fragile nucleosomes which still allow transcription machinery to bind instead of being entirely nucleosome-free. Notably, the enrichment of H2A.Z also contributes to reducing the barrier of +1 nucleosome to Pol II and the decreased stalling (Weber et al., 2014). Some pioneer transcription factors are able to bind to the compacted chromatin and help to open the local chromatin by themselves or by recruiting other chromatin/nucleosome remodelers (Fuda et al., 2015; Zaret & Carroll, 2011), thereby making the hidden CREs reachable. In addition, the post-translational modifications of histones including histone H3 lysine 9 acetylation (H3K9ac), histone H3 lysine 27 acetylation (H3K27ac) and tri-methylation of histone H3 lysine 4 (H3K4me3) in the nucleosomes have been found to correlate with the flanking accessible chromatin in active promoter regions (Barski et al., 2007; Négre et al., 2011).

1.6 Approaches for characterization of CREs

The encyclopedia of DNA elements (ENCODE) and model organism ENCODE (modENCODE) projects have generated a broad spectrum of data for systematic annotation of functional genomic elements in human, *D. melanogaster* and *Caenorhabditis elegans* (Brown & Celniker, 2015; Gerstein et al., 2010; Kellis et al., 2014; Roy et al., 2010). They allowed mapping of RNA transcripts, chromatin marks, nucleosome patterns and TF binding sites in different cell types (ENCODE, modENCODE) and tissues or whole organisms across developmental stages (modENCODE). A variety of high-throughput approaches have been developed to characterize TF binding locations and chromatin accessibility in the genome in order to detect the CRE candidates. Chromatin immunoprecipitation sequencing (ChIP-seq) is the most widely used method for identification of TF-bound DNAs or specific histone modifications by antibody recognition (Johnson, Mortazavi, Myers, & Wold, 2007). DNase-seq (Crawford et al., 2006; Song & Crawford, 2010), FAIRE-seq (Gaulton et al., 2010; Giresi, Kim, McDaniell, Iyer, & Lieb, 2007; Simon, Giresi, Davis, & Lieb, 2012) and ATAC-seq (Buenrostro, Giresi, Zaba, Chang, & Greenleaf, 2013) were successively devised for probing the genome-wide DNA accessibility. However, all these approaches only indicate the genomic sites of putative CREs correlative to certain factors (e.g., TFs binding, histone modification marks and open chromatin) rather than the activity of these mapped sites or their real regulatory functions in controlling the gene expression levels. Various high-throughput approaches such as CAGE and global run-on sequencing (GRO-seq) have been applied to study genome-wide endogenous activity of promoter and enhancer (Andersson et al., 2014; Core et al., 2014; Core, Waterfall, & Lis, 2008; Shiraki et al., 2003; D. Wang et al., 2011). Yet, the interplay between promoters and enhancers together with surrounding nucleosome organization and chromatin configuration all affect the endogenous activity. Other computational analyses based on combined sequence data (mostly for promoters) and gene expression data have suggested the regulatory roles of enriched sequence motifs or motif combinations and evaluated the effects of motif strength, position, orientation as well as relative spacing on gene expression (Beer & Tavazoie, 2004; Nguyen & D'haeseleer, 2006; Pilpel, Sudarsanam, & Church, 2001; Segal et al., 2003; Sudarsanam, Pilpel, & Church, 2002), which potentially provide causal hypotheses. Nevertheless, functional validations which require testing mutated variants to ascertain the influence of specific features still lack in all these mentioned studies.

One strategy to study the autonomous function of CREs is to separate them out of their native environment in the genome. Reporter assays were therefore developed for quantitative measurement of the sequence-intrinsic activity of CREs. Luciferase assay is one of the widely used reporter systems which fuses the target regulatory DNA element to a reporter gene such as firefly or renilla luciferase gene and measures its expression through quantifying the released bioluminescence. Compared to other reporter assays like using the green fluorescent protein (GFP), luciferase assay has a better signal-to-noise ratio with broader dynamic range. Besides, firefly luciferase and renilla luciferase can be used together for dual luciferase assays, in which one acts as the reporter and the other acts as the internal control. This allows data normalization to correct variation in cell numbers or transfection efficiencies, for instance. The DNA sequences analyzed in reporter assays are not restricted to only genomic sequences. Mutant or synthetic sequences can be also tested. However, the throughput of traditional reporter assays is mainly limited by the laborious mutagenesis, cloning and construction of the reporter plasmids. Random ligation or random mutagenesis of regulatory sequences were used to create the larger-scale library for reporter assays (Kinney, Murugan, Callan, & Cox, 2010; Ligr, Siddharthan, Cross, & Siggia, 2006). However, the random manipulation cannot tackle specific questions such as systematic dissection of specific features in the sequences.

Massively parallel reporter assays (MPRAs) have been developed to simultaneously test the function of thousands of wild-type or designed sequences at the single-nucleotide resolution in a single experiment. Most of the candidate regulatory elements analyzed in these studies are systematically designed DNA oligonucleotides (≤ 200 nt) synthesized on programmable microarrays. MPRAs take advantage of next-generation sequencing for identification of the tested CREs in the pooled reporter libraries. The method generally falls into two categories according to how reporter gene expression is measured. One kind of methods uses RNA sequencing (RNA-seq) to quantify the barcodes in the 3’ untranslated region (UTR) of the transcripts. By normalizing the counts of transcribed barcodes to the counts of barcode DNA in the original reporter library, the relative activities of the associated CREs can be obtained. This approach also enables using multiple barcodes to tag the same CRE for replicate measurements and is applicable both *in vitro* (Patwardhan et al., 2009) and *in vivo* (Kwasnieski, Mogno, Myers, Corbo, & Cohen, 2012; Melnikov et al., 2012; Patwardhan et al., 2012). The other kind of MRPA method quantifies the protein fluorescence as the readout of reporter gene expression (Sharon et al., 2012). It uses the barcodes upstream of the designed promoter sequences for variant identification instead of barcodes within the RNA which can avoid their influence on expression. The yeast cells carrying plasmids expressing the yellow fluorescence protein (YFP) driven by thousands of designed promoters are sorted into different expression bins according to the YFP intensities by fluorescence-activated cell sorting (FACS). The activity of each tested promoter is measured by its sequencing reads distribution across different expression bins. This technique also allows measurements of promoter effect on reporter expression variability (expression noise) among cells (Sharon et al., 2014).

Furthermore, MPRAs have been utilized to screen the entire genome, such as the STARR-seq which stands for self-transcribing active regulatory region sequencing (Arnold et al., 2013; Muerdter, Boryń, & Arnold, 2015) for annotation of enhancer activity in *Drosophila* genome. Different from conventional MPRAs, STARR-seq requires no barcodes in the construct since it inserts the candidate sequences in the 3’ UTR of the reporter gene. This is based on the fact that enhancers can function independently of their positions (Banerji, Rusconi, & Schaffner, 1981). The tested enhancer sequences are transcribed by themselves and the enhancer activities are measured as their abundance in RNA transcripts. This approach was also used to gain insights into enhancer evolution and enhancer–core-promoter specificity (Arnold et al., 2014; Zabidi et al., 2015).

Although MPRAs have been widely used for high-throughput analysis of CREs, most of them focused on detecting enhancer activity. Only several studies have applied MPRAs for systematic characterization of core promoters, which are the main interest of this thesis. With saturation mutagenesis of mammalian core promoters in an *in vitro* transcription system, mutations disrupting the TATA-Box and INR motif were found to drive the substantial reduction in transcriptional efficiency (Patwardhan et al., 2009). By adapting the fluorescence-based MPRA method, Lubliner and colleagues (Lubliner et al., 2015) analyzed over 7000 native and synthetic yeast core promoter variants. Their results showed that the core promoters make a significant contribution in determining the entire promoter activity and suggested the critical effects of location, orientation and flanking bases in altering TATA-Box function. Additionally, the autonomous promoter strength measurement of random genomic fragments was achieved recently in both fly and human (Arnold et al., 2017; van Arensbergen et al., 2017). As an extension of STARR-seq, the self-transcribing active core promoter sequencing (STAP-seq) was developed for assessing both the basal and enhancer-driven promoter activity (Arnold et al., 2017). Their analysis enabled the definition of enhancer responsiveness and suggested promoters with higher enhancer-driven strength if containing core promoter motifs like INR or TATA-Box. The survey of regulatory elements (SuRE) provided the sufficient coverage to probe the entire human genome with data suggesting that the promoter autonomy is mainly determined by the core promoter region together with further upstream sequences up to several hundred bps (van Arensbergen et al., 2017).

1.7 Transcriptional regulation by ecdysone receptors

Ecdysone is the steroid hormone that regulates molting and metamorphosis in *D. melanogaster*. The pulses of its physiologically active form (20-Hydroxyecdysone) released from the prothoracic glands direct the transitions between major developmental stages, reflecting the central role of ecdysone in defining the developmental timing (Thummel, 2001). Like other lipophilic hormones, ecdysone can diffuse through cell membranes and induce transcriptional responses by binding to its nuclear receptors, which are known as ligand-regulated TFs (King-Jones & Thummel, 2005). The receptor protein that ecdysone binds to is a heterodimer of two nuclear receptors: the ecdysone receptor (EcR) and ultraspiracle (USP) (Koelle et al., 1991; Yao et al., 1993). EcR is the ortholog of the vertebrate farnesoid X receptor (FXR) or liver X receptor (LXR) and cannot bind ecdysone on its own. Its hormone-binding activity depends on the dimerization with USP, which is the ortholog of the vertebrate retinoid X receptor (RXR). The EcR/USP complex recognizes the ecdysone response elements (EcREs) that are usually pseudopalindromic sequences co-localized with functional CREs. When ecdysone is bound (primarily to the ligand-binding pocket of EcR), EcR/USP functions as a transcriptional activator and stimulates expression of ecdysone-responsive genes. In the absence of ecdysone, however, the unliganded EcR/USP can repress transcription by interacting with corepressors (Cherbas, Lee, & Cherbas, 1991; Dobens, Rudolph, & Berger, 1991).

EcREs in *Drosophila* were originally identified by promoter analysis of ecdysone-responsive genes in different cell lines and tissues (Laval, Pourrain, Deutsch, & Lepesant, 1993; Riddihough & Pelham, 1987). However, their functional analyses have been limited to a small scale due to the lack of a genomic mapping. In order to identify the binding sites of EcR/USP across the whole genome, Gauhar and colleagues (Gauhar et al., 2009) applied the DNA adenine methyltransferase identification (DamID) approach to detect EcREs in *Drosophila* Kc167 cells. Their results showed that only 42% of identified binding sites are close to ecdysone target genes in these cells. A larger portion (~ 44%) were found nearby ecdysone-regulated genes in other tissue or cell types involved in metamorphic processes, indicating the EcR/USP binding is mostly not cell-type specific.

Using the STARR-seq to quantitatively analyze the genome-wide ecdysone-responsive enhancer activity, Shlyueva and colleagues (Shlyueva et al., 2014) found induced and repressed enhancers are distinguishable by their sequence motif content. In addition, the motifs for partner TFs of EcR/USP differ between the two tested cell types, which are *Drosophila* Schneider 2 (S2) cells and ovarian somatic cells (OSCs), suggesting their roles in determining the cell-type-specific function of enhancers. Despite this considerable effort in high-throughput analysis of ecdysone-responsive enhancers, the contribution of core promoters to the ecdysone responsiveness remains barely estimated. One study in *Spodoptera frugiperda* cell line Sf9 has shown that the mutations in INR motif, as well as motifs immediately next to TATA-Box or locate in 5’ UTR, have significant effects on reducing the ecdysone inducibility (Jones et al., 2012). Further elucidation of core promoter influence is still needed.

2. Research basis of the thesis

The Söding lab devised the XXmotif (eXhaustive evaluation of matriX motifs), a P-value-based regulatory motif discovery tool using position weight matrices (PWMs) (Hartmann, Guthöhrlein, Siebert, Luehr, & Söding, 2013). A PWM is a two-dimensional matrix with each row corresponding to one of the four nucleotides and each column representing a position within the motif. Each element in the PWM gives the log-likelihood of observing a particular nucleotide at a particular position (Stormo, Schneider, Gold, & Ehrenfeucht, 1982). Assuming each position is independent of each other, a quantitative score can be generated for a given DNA sequence by summing the PWM values of the relevant nucleotides. A threshold is usually defined to separate matching motifs from non-motifs (Stormo, 2000). XXmotif combines the potent PWM representation with an improved statistical model for assessing over-representation of motifs.

Particularly, to analyze the core promoters in *Drosophila*, Hartmann and colleagues firstly defined 19 gene sets based on experimentally derived genome-wide features, including expression strengths and variations throughout developmental stages (Graveley et al., 2011), PolII stalling (Hendrix et al., 2008; Zeitlinger et al., 2007) and TSSs mapping from CAGE data (Hoskins et al., 2011; Ni et al., 2010). They applied XXmotif for the de novo motif search in the core promoter regions of these genes and were able to identify widely known motifs as well as some novel motif candidates with optimized PWMs based on enrichment, localization and conservation (Hartmann, 2012). All identified motifs are: known motifs including INR, MTE/DPE (an overlapping version of the two previously identified motifs MTE and DPE, hereafter referred to as MTEDPE), GAGA, GAGArev, INR2 (widely known as motif 1 or Ohler1), DRE, Ohler7, E-Box1, Ohler6, TATA-Box, R-INR (widely known as TCT motif, here named as ribosomal initiator based on its co-localization with TSSs of ribosomal protein genes), E-Box2; new motifs including CGpal, INR2rev, TTGTT, TTGTTrev, AAG3, ATGAA and RDPE (ribosomal downstream promoter element). A summary of identified motifs and their features is listed in Table S1. In addition, a new motif named as CA-INR was often found co-occurring with TATA-Box, which is a highly conserved derivative of the classical INR motif. CA-INR also has a strong positional preference around TSS and its most representative sequence is GGCATCAGTC with the TSS mostly mapped at its 4th position.

By correlating all identified motifs to the gene sets (Figure 3A), four classes of the core promoter motifs can be defined. Class 1 motifs (INR, MTEDPE, CGpal, GAGA, GAGArev) occur in genes with NP core promoters. The enriched genes are intermediately regulated and show strong correlations to stalled Pol II. Class 2 motifs including TATA-Box and ATGAA also present in NP promoter genes, however, the enriched genes are strongly regulated ones that are either not expressed or most highly expressed in at least one developmental stage. Class 3 motifs (INR2, Ohler6, DRE, Ohler7, E-Box1, TTGTT, TTGTTrev, INR2rev, AAG3) are the ones only found in genes with BP core promoters. The enriched genes are not regulated and similarly expressed in all developmental stages (housekeeping function). Class 4 motifs (R-INR, RDPE) correlate with strongly expressed genes which mainly encode the ribosomal proteins. The motif co-occurrence is in agreement with the four defined classes and also suggests two distinct preferred compositions in the 3rd class (INR2 + Ohler6 pair and DRE + Ohler7 pair; Figure 3B).

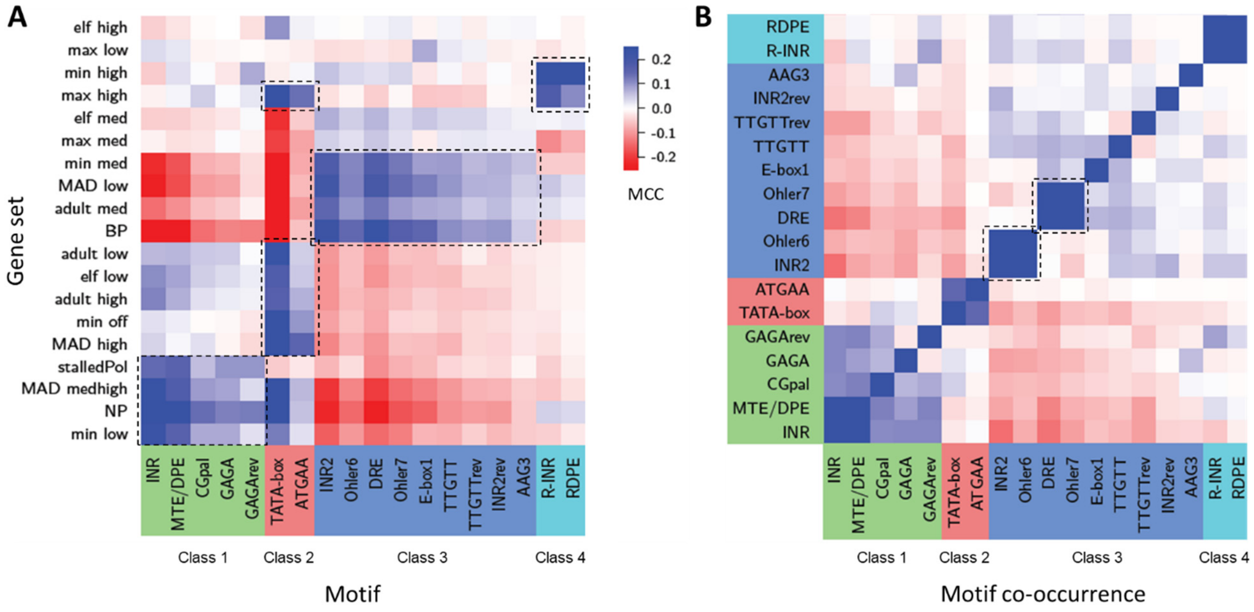


Figure 3. *Drosophila* core promoter motifs occur in four defined classes. (A) Correlation of core promoter motifs to 19 gene sets with different features reveals four distinct motif classes: class 1 motifs enriched in the gene sets of stalledPol, MAD medhigh, NP and min low; class 2 motifs enriched in the gene sets of max high, adult low, elf low, adult high, min off and MAD high; class3 motifs enriched in the gene sets of min med, MAD low, adult med and BP; class 4 motifs enriched in the gene sets of min high and max high (details about different gene sets are listed in Table S2). MCC: Matthews correlation coefficient. **(B)** Correlation of core promoter motifs to each other indicates motif co-occurrence within the same promoter, which agrees with the four defined classes and also suggests two distinct preferred compositions in the 3rd class (INR2 + Ohler6 pair and DRE + Ohler7 pair). Figures are adapted from (Hartmann, 2012).

In light of the findings discussed above, we hypothesize that there are in general four core promoter architectures containing each class of motifs accordingly, reflecting different modes of transcriptional regulation at the core promoter. The simplified architectures with co-occurred motif pairs taken into consideration are illustrated in Figure 4, named as regulated/stalled architecture 1 (Ar.1), highly regulated Ar.2, housekeeping Ar.3.1/Ar.3.2 and ribosomal Ar.4. These four architectures can be further grouped into developmental (Ar.1, Ar.2) and constitutive (Ar.3.1, Ar.3.2, Ar.4) core promoters based on their association with gene functions.

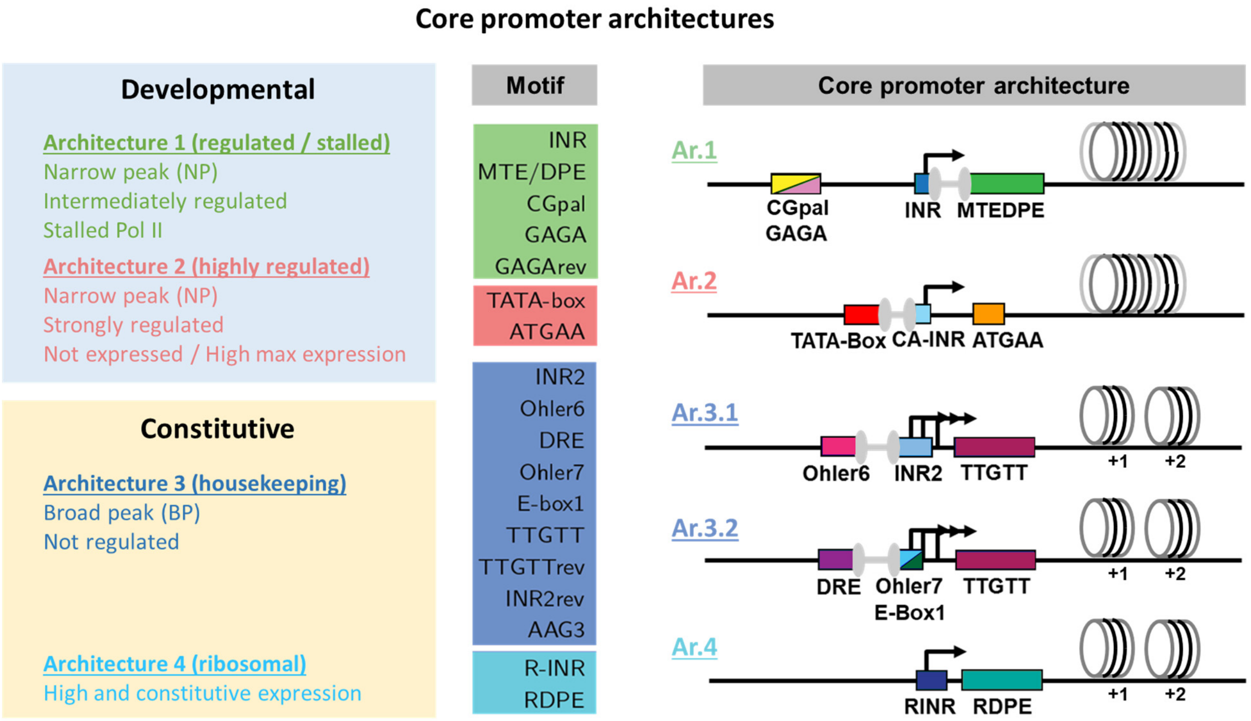


Figure 4. Our hypothesis of *Drosophila* core promoter architectures. Based on the motif classes identified by XXmotif, four core promoter architectures are defined accordingly, reflecting different modes of transcriptional regulation at the core promoter. They are named as regulated/stalled, highly regulated, housekeeping and ribosomal architectures. They can be further grouped into developmental (Ar.1, Ar.2; highlighted in blue) and constitutive (Ar.3.1, Ar.3.2, Ar.4; highlighted in yellow) core promoters based on their association with gene functions.

3. Aim of the thesis

The genome-wide annotation of CREs and evaluation of their properties have provided massive insights into transcriptional regulation. Although the genomic analysis of native sequences suggests causal relationships, the variations in genomic sequences are usually arbitrary, making the sequence attributes for activity changes difficult to be uncovered. We are not able to ascertain the influence of specific features unless we mutationally alter the feature and measure the effect. Facilitated by DNA synthesis technology and next-generation sequencing, high-throughput approaches such as MPRAs have been developed to tackle this problem on a large scale systematically. However, most of the studies focus on enhancers, especially on TF binding sites. Our understanding of other sequence elements and their combinations required for activity is mostly lacking. Despite the pivotal role of core promoter in transcription initiation, how the components and sequence features of the core promoter compute the intrinsic expression levels remains poorly understood. Therefore, this study aims to dissect the core promoter comprehensively and to elucidate the sequence determinants of functional promoters in the well-studied experimental model *D. melanogaster*. To reduce the complexity of the problem, a single cell type, S2 cells, was used in the experiments.

Although sequencing-based approaches have provided a high-throughput solution for functional analysis, their dynamic range is usually small (around two orders of magnitude) and the activity measurements suffer from low accuracy for weak regulatory elements due to their low coverage of reads. This limitation would severely influence our core promoter analysis since they are known to drive basal and modest expression. Additionally, the fluorescence-based MPRAs only get discrete expression measurements because of their bin sorting design, which cannot sense subtle effects. The dual luciferase assay is a robust reporter assay to study gene expression and regulation such as testing promoter activity in a rapid, simple and sensitive way with a broad linear range. Its scale has been limited by the slow and laborious cloning, transfection and luminescence readout. To keep the power of luciferase assay for accurate measurement of core promoter strength and overcome its technological barrier, we integrated the golden gate cloning strategy (BsaI cloning) along with a high-throughput experimental pipeline using automated robot systems for colony picking, reporter plasmids isolation, transient co-transfection and dual luciferase assay. After extensive optimizations of experimental protocols and data normalization, the final method allowed us to measure promoter activity quantitatively in a large scale with high reproducibility, sensitivity and a wide dynamic range.

We then applied this method to measure both basal and induced expressions of thousands of designed promoters, that were synthetic promoter constructs with combined building blocks representing different functional regions (Figure 5). The blocks comprised: the motif-rich core promoter region of 130 bp around TSS with native and perturbative sequences from different core promoter architectures (referred to as block 3-6 in the thesis); a stimulus-response element for binding of the ecdysone receptors to recruit the steroid hormone ecdysone for transcriptional activation (referred to as block 2); and the genomic -1 and +1 nucleosome positioning sequences to mimic the endogenous ±1 nucleosomal context (referred to as block 1 and block 7, respectively). This design intends to test three promoter features separately: core promoter sequence features, especially motifs (the main focus of this thesis); transcriptional response to external stimulus (ecdysone); and effect of genomic ±1 nucleosome sequences around core promoter. The annotation of core promoter architecture and motif content was based on the XXmotif screening results.

To systematically examine the sequence motifs in core promoters, we devised various mutations in wild-type regions, including individual or pairwise knockout (complete replacement with non-functional sequences) of motifs, knockout of all motifs, replacing the original motif with its XXmotif-derived highest frequent genomic sequence (hereafter referred to as consensus), point mutations of motifs, shift of motif positions and substitution with functionally or positionally equivalent motifs from other architectures. We tested not only the widely known motifs like INR and TATA-Box, but also several new motif candidates discovered by XXmotif. The activities of these synthetic promoters with mutated motifs were used to compare with their wild-type strengths and the point mutation results allowed further analysis of the motif specificity. Recent studies also suggest that the sequence motifs alone cannot fully explain the activity variation. Therefore, in our experiments, the motif-surrounding context sequences were also tested. In addition, combinatorial mutations altering both motif strength and motif positioning within core promoter architectures as well as block-wise swap between architectures were implemented for more in-depth analysis which enabled quantitative modeling of promoter activity based on individual sequence features.

II MATERIALS AND METHODS

Key reagents and resources used in this study are listed below.

Reagent or Resource	Source	Identifier
Bacterial and Virus Strains		
E.coli TOP10 Electrocomp cells	U. Gaul lab	N/A
Chemicals, Peptides, and Recombinant Proteins		
HindIII-HF restriction enzyme	NEB	R3104S
BglII restriction enzyme	NEB	R0144S
NheI-HF restriction enzyme	NEB	R3131S
XhoI restriction enzyme	NEB	R0146S
Herculase II fusion DNA polymerase	Agilent Technologies	600677
BsaI restriction enzyme	NEB	R0535L
T4 DNA ligase	Promega	M1801
SOC medium	U. Gaul lab	N/A
Taq/Pfu polymerase mix	U. Gaul lab	N/A
Schneider's *Drosophila* Medium	Bio&Sell	BS 2.43G02J
Fetal Bovine Serum	Biochrom	S 0415
Express Five SFM medium	Invitrogen	10486025
L-Glutamine	Invitrogen	25030024
FuGENE® HD Transfection Reagent	Promega	E2312
20-Hydroxyecdysone	Sigma-Aldrich	H5142
Nile Blue A	Sigma-Aldrich	N5632
Critical Commercial Assays		
QIAquick Gel Extraction Kit	Qiagen	28704
Rapid DNA Ligation Kit	Roche	11635379001
QIAquick PCR Purification Kit	Qiagen	
ONE-Glo™ Luciferase Assay System	Promega	E6120
Renilla-Glo® Luciferase Assay System	Promega	E2720
Zero Blunt TOPO PCR Cloning Kit	Invitrogen	450245
Agencourt AMPure XP magnetic beads	Beckman Coulter	A63880
Wizard MagneSil Tfx™ System	Promega	A2380
Nextera XT Index Kit v2 Set A - Set D for 96 Indexes, 384 Samples	Illumina	FC-131-2001 - 2004
Experimental Models: Cell Lines		
Drosophila S2 cells	U. Gaul lab	N/A
Experimental Models: Organism/Strains		
Fly: *Drosophila melanogaster*	Bloomington Drosophila Stock Center	Stock # 2057
Recombinant DNA		
pGL4.10 Luciferase Vector	Promega	E6651
pGL4.13 Luciferase Vector	Promega	E6681
pGL4.70 Luciferase Vector	Promega	E6881
pKF1	U. Gaul lab	N/A
pUC19	NEB	N3041S

pUG9	U. Gaul lab	N/A
pZQ3	U. Gaul lab	N/A
pZQ5	U. Gaul lab	N/A
Oligonucleotides		
Primers	This study (Eurofins)	N/A
Block 3-6 (native ones in Table S3)	This study (Agilent Technologies)	N/A
Other		
Gene Pulser	Bio-Rad	Model 1652076
Biomek NX^{P} Automated Workstation	Beckman Coulter	Multichannel-96 and Span-8
Incubator Shaker DWP	Inheco	7300009
Biometra TRobot	Analytik Jena	846-050-991
Microplate Print & Apply	Beckman Coulter	148640
Compact Laser Barcode Scanner	Omron Microscan	MS-3
SpectraMax Paradigm Multi-Mode Microplate Reader	Molecular Devices	N/A
SpectraDrop Micro-Volume Microplates	Molecular Devices	N/A
Wasp	Kbiosystems	N/A
VIAFLO Electronic Multichannel Pipette	INTEGRA	4624
ASSIST Pipetting Robot	INTEGRA	4500
Riplate SW 48	Ritter	43001-0062
MegaBlock 96 Well	Sarstedt	82.1972.002
Round 96 Well Storage Plates	4titude	4ti-0116
FrameStar 96 Well Skirted PCR Plate	4titude	4ti-0960/C
Deepwell plate 96/500 µl	Eppendorf	0030501101
Tissue Culture Flask 75 cm^{2}	Corning	430641U
Falcon 96 Well Tissue Culture Plate	Corning	353072
Tissue Culture Dish	Corning	353003
Cell Counter and Analyzer System	Roche	CASY Model TT
AlphaPlate-384	PerkinElmer	6005350

Table 1. Key resources table.

4. The synthetic promoter construct design

We designed synthetic promoter constructs by dividing the promoter region into 7 building blocks (Figure 5): block 3-6 was the motif-rich core promoter region (-80 to +50 bp around the TSS) with native and perturbative sequences from different core promoter architectures to investigate the effects of sequence motifs; block 2 represented the EcREs, which contained the binding sites for the ecdysone receptors to recruit the steroid hormone ecdysone for transcriptional activation; block 1 and block 7 were used for testing the influence of nucleosomal context. The entire lengths for the designed synthetic promoters inserted into the vector backbones were 703 bp with block 7 and 459 bp without block 7.

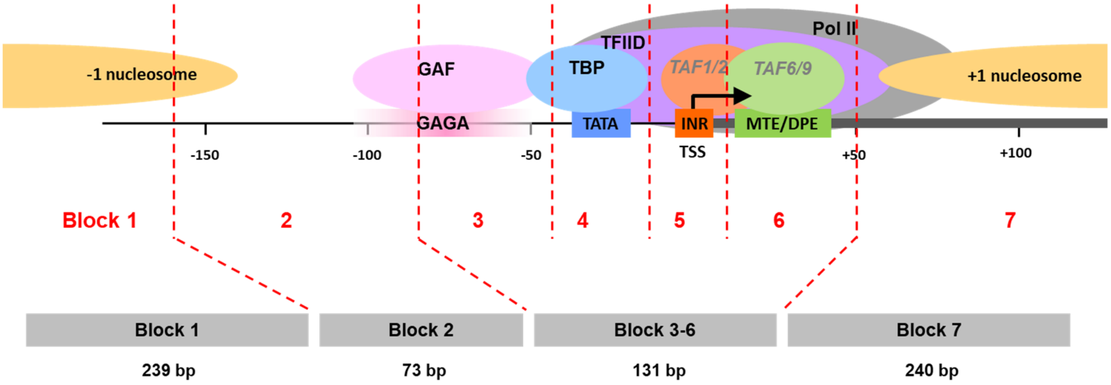


Figure 5. Synthetic promoter design - building blocks. The promoter region was divided into 7 building blocks: block 1 with 239 bp sequence representing the potential -1 nucleosome; block 2 with 73 bp sequence representing the ecdysone receptor binding region; block 3-6 with 131 bp sequence representing the native and perturbative core promoter regions from different architectures; block 7 with 240 bp sequence representing the potential +1 nucleosome.

4.1 Motif-rich core promoter region (block 3-6)

From the four core promoter architectures (including two subclasses Ar.3.1 and Ar.3.2 of the housekeeping Ar.3) and one additional architecture without having any known motif named as architecture 0 (Ar.0), we chose 2-4 native core promoters each with high (- intermediate) - low expressions according to their maximum expression levels in *Drosophila* S2 cells (previous RNA-seq data generated by Dr. Katja Frühauf in our lab; position -80 to +50 relative to TSS which was set to be position 0; block 3: -80 to -35, block 4: -34 to -10, block 5: -9 to +8, block 6: +9 to

+50). In total, we thus selected 19 wild-type core promoters, some of which have mixed architectures due to different motifs co-occurrence (Figure 6; their 131 nt sequences listed in Table S3). The annotation of core promoter motifs in these sequences was carried out by de novo motif search by XXmotif according to previously defined motif features (summarized in Table S4). The corresponding FlyBase gene ID was used for the notation of each native core promoter. In addition, we mutated TSS downstream ATGs in the original sequences to TAGs to remove unwanted translation starts. Various kinds of mutations were designed for these native core promoters, including mutations for motifs within each core promoter (main mutations shown in Figure 7) and block-wise mutations between different core promoters. We also applied the XXmotif algorithm on every designed mutated sequences to check if the mutants we created would lead to undesirable side mutational effects, e.g., the creation of new motifs/TF binding sites or disruption of other motifs (our measurements are also sensitive enough to detect the subtle expression changes caused by those unintended mutations). Finally, all sequences were synthesized by Agilent Technologies (Cleary et al., 2004; LeProust et al., 2010) together with BsaI sites, relevant overhangs and unique primer sequences referred to distinct mutation families, in total 3826 fully designed oligonucleotides (in total ~ 200 nt long for each sequence).

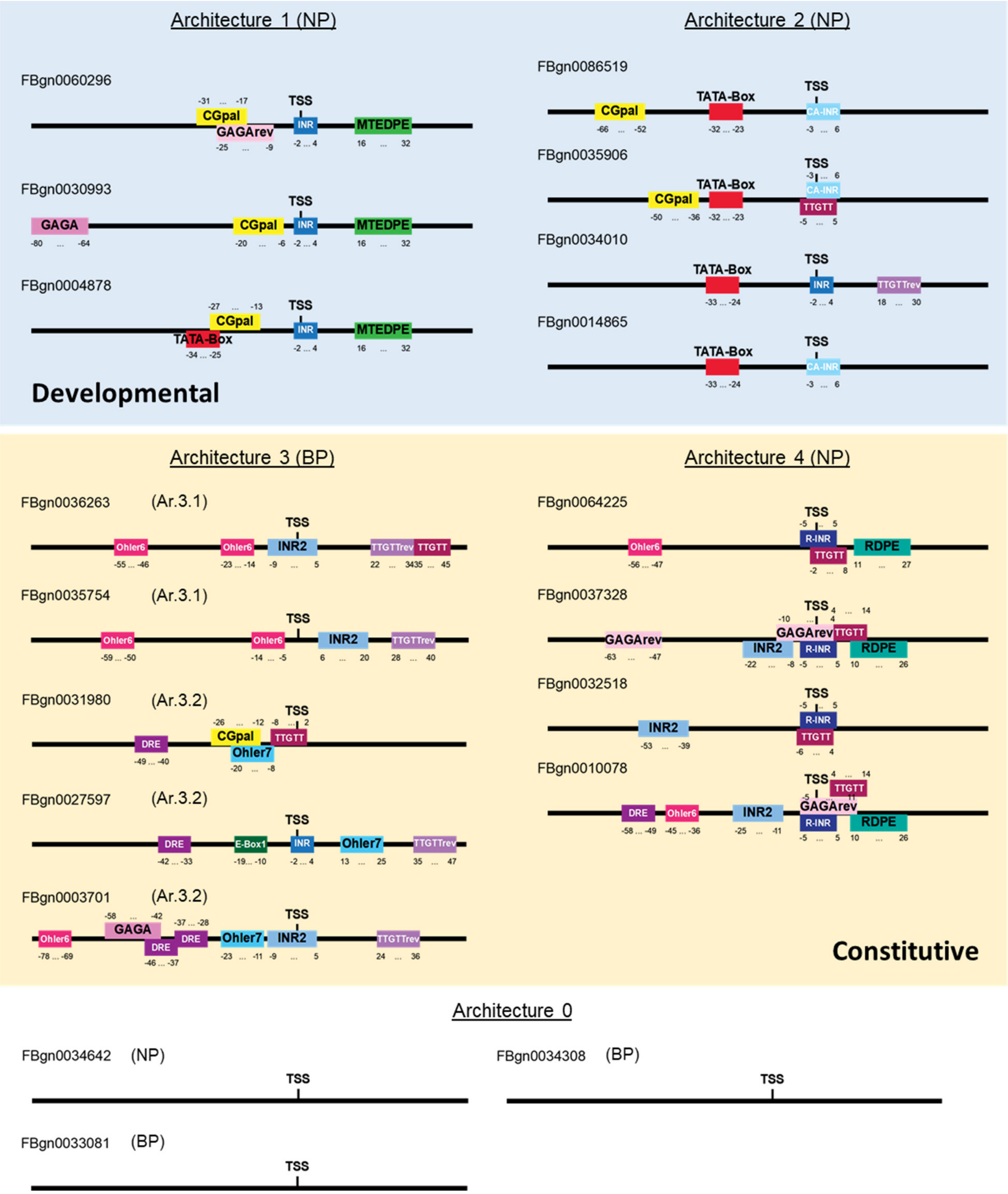


Figure 6. The wild-type core promoters and their motif composition. From the four core promoter architectures Ar.1, Ar.2, Ar.3 (Ar.3.1, Ar.3.2), Ar.4 and one additional architecture with core promoters containing no known motif (Ar.0), 2-4 native sequences were chosen from each architecture (position -80 to +50 relative to TSS; TSS itself at position 0). In total 19 wild-type core promoters with annotated motif positions are shown here. NP, narrow peak; BP, broad peak. Their sequences are listed in Table S3. Developmental and constitutive promoters are highlighted in blue and yellow, respectively.

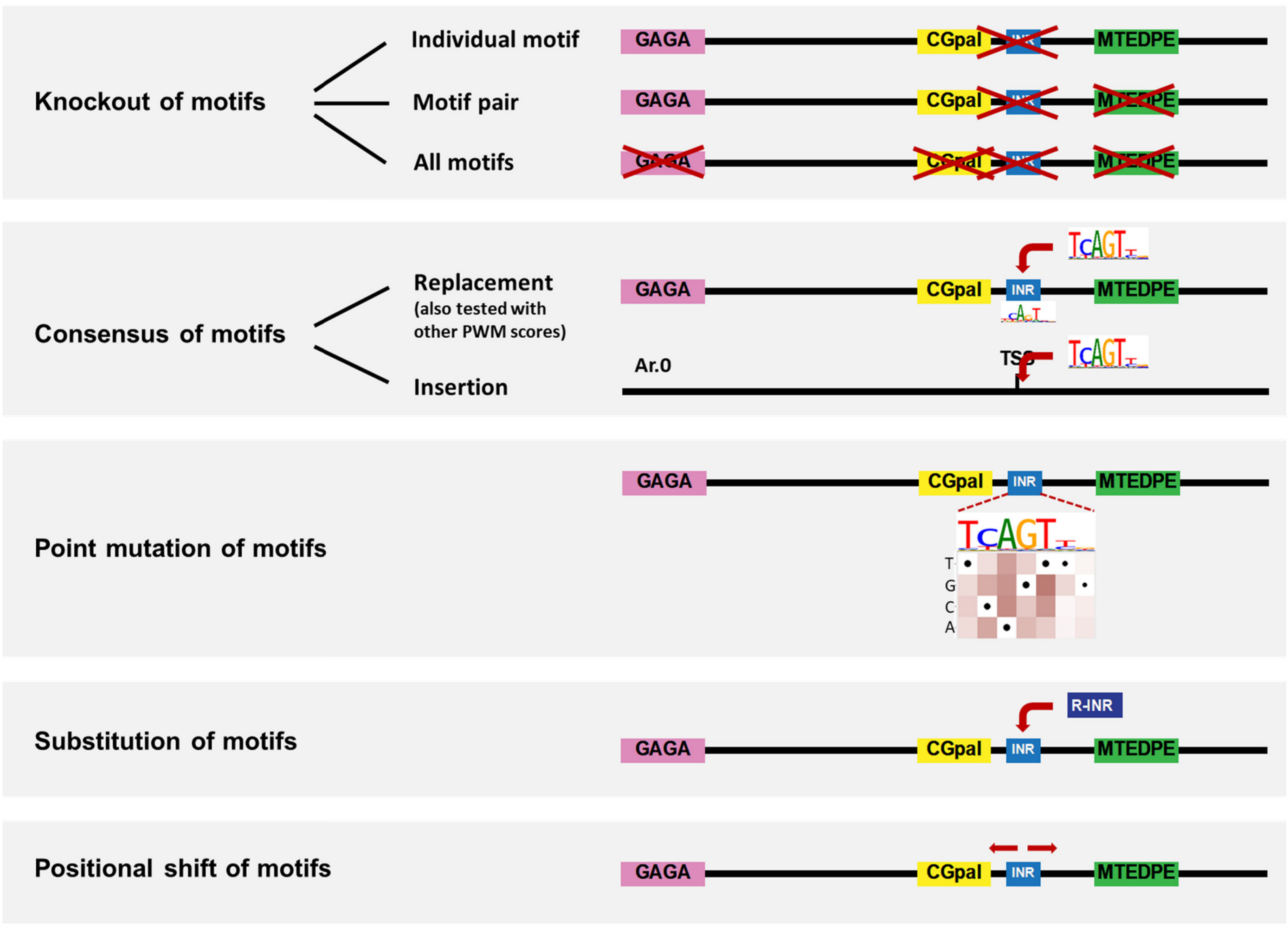


Figure 7. An illustration of the main mutations applied to the wild-type core promoter motifs. Systematically designed mutations for motifs within each core promoter include: knockout of motifs (individual or pairwise knockout of motifs, and knockout of all motifs; 260 sequences designed in total); replacing the original motif with its computationally (XXmotif) derived sequences with different PWM scores (consensus with the highest score) or insertion of the consensus into the Ar.0 sequences (170 sequences); point mutation of motifs (596 sequences); substitution with functionally or positionally equivalent motifs from other architectures (78 sequences); shift of motif positions (164 sequences). The FBgn0030993 motif composition is shown here as an example.

4.1.1 Mutation with different strengths of motifs

4.1.1.1 Knockout of motifs

For knocking out individual motifs in 16 native core promoters (excluding three Ar.0 sequences), two versions of sequences were used as substitutions: random sequences and background sequences. Random sequences were generated by sampling sequences having the same length with the target motifs and checking with the XXmotif derived motif list to make sure no known core promoter motif inside (whose PWM scores lower than the threshold, threshold score of each motif listed in Table S4). These random sequences were not fixed for the same motif in different promoters (every random sequence was different). Background sequence was a fixed sequence from the identical position of the target motif in the Ar.0 core promoter FBgn0034642 (due to the various positions of a certain motif in different promoters, the background sequence might vary). Knockout of all motifs in a given promoter was designed in the same way, using both random and background sequences. Pairwise knockout of motifs only used random sequences for replacing two original motifs at the same time.

4.1.1.2 Consensus replacement of motifs

For the nine main motifs INR, MTEDPE, TATA-Box, INR2, Ohler6, DRE, Ohler7, R-INR and RDPE, we replaced them in native core promoters with the consensus sequences derived from XXmotif. Additionally, these consensus sequences were also inserted into the three Ar.0 core promoters with their start positions at the peaks of the native motif distribution (Table 2; motif distribution shown in the column “Distribution” of Table S1).

Motif	Position to TSS	Motif	Position to TSS	Motif	Position to TSS
INR	-2	MTEDPE	17	TATA-Box	-32
INR2	-9	Ohler6	-32	DRE	-32
Ohler7	-4	R-INR	-5	RDPE	11

Table 2. The motif start positions (relative to TSS) for insertion of consensus sequences into Ar.0 core promoters.

4.1.1.3 Replacing native motifs with their alternatives of various strengths

Alternatives with different PWM scores for the nine main motifs mentioned above were randomly generated with their scores either evenly covered several bins between the threshold and the maximum, or below the threshold.

4.1.1.4 Point mutation of motifs

For 12 motifs INR, MTEDPE, CGpal, TATA-Box, INR2, Ohler6, DRE, Ohler7, R-INR, RDPE, TTGTT and TTGTTrev, we designed all possible single base pair mutations (native target motif was firstly replaced by its consensus sequence and exhaustive point mutations were applied) within a selected native core promoter configuration: INR in FBgn0030993; MTEDPE and CGpal in FBgn0004878; TATA-Box in FBgn0034010; INR2, Ohler6 and TTGTTrev in FBgn0036263; DRE, Ohler7 and TTGTT in FBgn0031980; R-INR and RDPE in FBgn0064225. Additionally, INR, DRE, Ohler7 and R-INR were also checked in an Ar.0 context FBgn0034308 with the insertion of each consensus sequence (as described in Section 4.1.1.2).

4.1.2 Substitution of motifs

The target motif was firstly knocked out with a random sequence which was generated in the same way as described before in Section 4.1.1.1. The motif sequence for substitution was also randomly sampled with a PWM score above the threshold and was always the same for each motif. Three combinations were tested here: INR (7 nt) - INR2 (15 nt) - Ohler7 (13 nt) - R-INR (11 nt); TATA-Box (10 nt) - Ohler6 (10 nt) - DRE (10 nt); MTEDPE (17 nt) - RDPE (17 nt). For INR-like motifs with various lengths, the supposed position for TSS (3rd position in INR, 10th in INR2, 5th in Ohler7 and 6th in R-INR; based on the motif start positions listed in Table 2) was aligned when replacing the sequence.

4.1.3 Positional shift of motifs

Positional shifts were designed for individual motifs and all motifs together in a given core promoter, as well as for sequence context surrounding motifs (motifs kept at the original positions). For strictly positioned motifs like INR, MTEDPE and TATA-Box, shifts of 1, 2, 3, 5, 10 bp either downstream or upstream were applied; for less well-positioned housekeeping core promoter motifs like DRE and Ohler7, larger distances were chosen (±1, ±3, ±5, ±10, ±20 bp).

4.1.4 Other combinatorial mutations

Further combinatorial mutations were designed to the motif-rich core region, including free combinations of mutations both within defined core promoter architectures and between them (termed as intra-architectural motif-wise and inter-architectural block-wise combinatorial mutations; Figure 8A and B). Besides, context sequences surrounding the motifs were also tested by exchanging them between different core promoters (Figure 8C).

For testing these combinatorial mutations, one representative core promoter sequence from each architecture with motifs located within distinct block regions was selected: FBgn0004878 (Ar.1), FBgn0034010 (Ar.2), FBgn0036263 (Ar.3.1), FBgn0031980N (Ar.3.2) and FBgn0064225 (Ar.4). The synthetic promoter FBgn0031980N was derived from the native FBgn0031980 (Ar.3.2) by artificially altered TSS position (shifted by 16 nt upstream) to locate all motifs in the blocks where they occur most frequently based on XXmotif generated distribution. In addition to the five core promoter sequences tested systematically in all three types of combinatorial mutations, several other native sequences were also included (Table 3; FBgn0035754 for intra-architectural mutations; FBgn0014865 and FBgn0086519 for inter-architectural mutations; FBgn0034308 and FBgn0034642 for context exchange).

Intra-architectural mutations	Inter-architectural mutations	Context exchange
FBgn0004878 (Ar.1, NP)	FBgn0004878 (Ar.1, NP)	FBgn0004878 (Ar.1, NP)
FBgn0034010 (Ar.2, NP)	FBgn0034010 (Ar.2, NP)	FBgn0034010 (Ar.2, NP)
FBgn0036263 (Ar.3.1, BP)	FBgn0036263 (Ar.3.1, BP)	FBgn0036263 (Ar.3.1, BP)
FBgn0031980N (Ar.3.2, BP)	FBgn0031980N (Ar.3.2, BP)	FBgn0031980N (Ar.3.2, BP)
FBgn0064225 (Ar.4, NP)	FBgn0064225 (Ar.4, NP)	FBgn0064225 (Ar.4, NP)
FBgn0035754 (Ar.3.1, BP)	FBgn0014865 (Ar.2, NP)	FBgn0034308 (Ar.0, BP)
	FBgn0086519 (Ar.2, NP)	FBgn0034642 (Ar.0, NP)

Table 3. The core promoter sequences selected for combinatorial mutations.

4.1.4.1 Intra-architectural motif-wise combinatorial mutations

Multiple motif-wise mutations for altering both motif strength and motif position within a core promoter sequence were performed here (Figure 8A). The FBgn0035754 (Ar.3.1) was selected because of its strong native activity, which ensures a relatively strong luminescence signal even after severe combinatorial mutations. Single mutations (knockouts, replacing by the consensus or alternatives with different PWM scores and positional shifts) for individual motifs in each core promoter were re-designed in the same way as described before but kept the same in all intra-architectural combinatorial mutations. Shifts of motifs were made within shorter ranges (±1 bp or ±5 bp).

4.1.4.2 Inter-architectural block-wise combinatorial mutations

We applied block-wise swaps between different core promoter sequences here (Figure 8B). Two additional sequences FBgn0014865 and FBgn0086519 were included to provide extra block patterns. In detail, block pieces from 7 native core promoters were selected and freely combined to construct the synthetic block 3-6 regions: four block 3s from FBgn0034010 (background sequence of Ar.2, NP), FBgn0031980N (background sequence of Ar.3.2, BP), FBgn0064225 (Ohler6 existed), FBgn0086519 (CGpal existed); five block 4s from FBgn0004878, FBgn0034010, FBgn0036263, FBgn0031980N, FBgn0064225; four block 5s from FBgn0034010, FBgn0036263, FBgn0031980N, FBgn0064225; six block 6s from FBgn0004878, FBgn0034010, FBgn0036263, FBgn0031980N, FBgn0064225, FBgn0014865 (background sequences of Ar.2, NP).

4.1.4.3 Context exchange

All motifs in a given core promoter were knocked out using the same sequences designed for single knockouts in intra-architectural combinatorial mutations. All motifs from other core promoter sequences were inserted into this context at their native positions (Figure 8C). Two Ar.0 core promoter contexts were also included: FBgn0034308 (BP) and FBgn0034642 (NP).

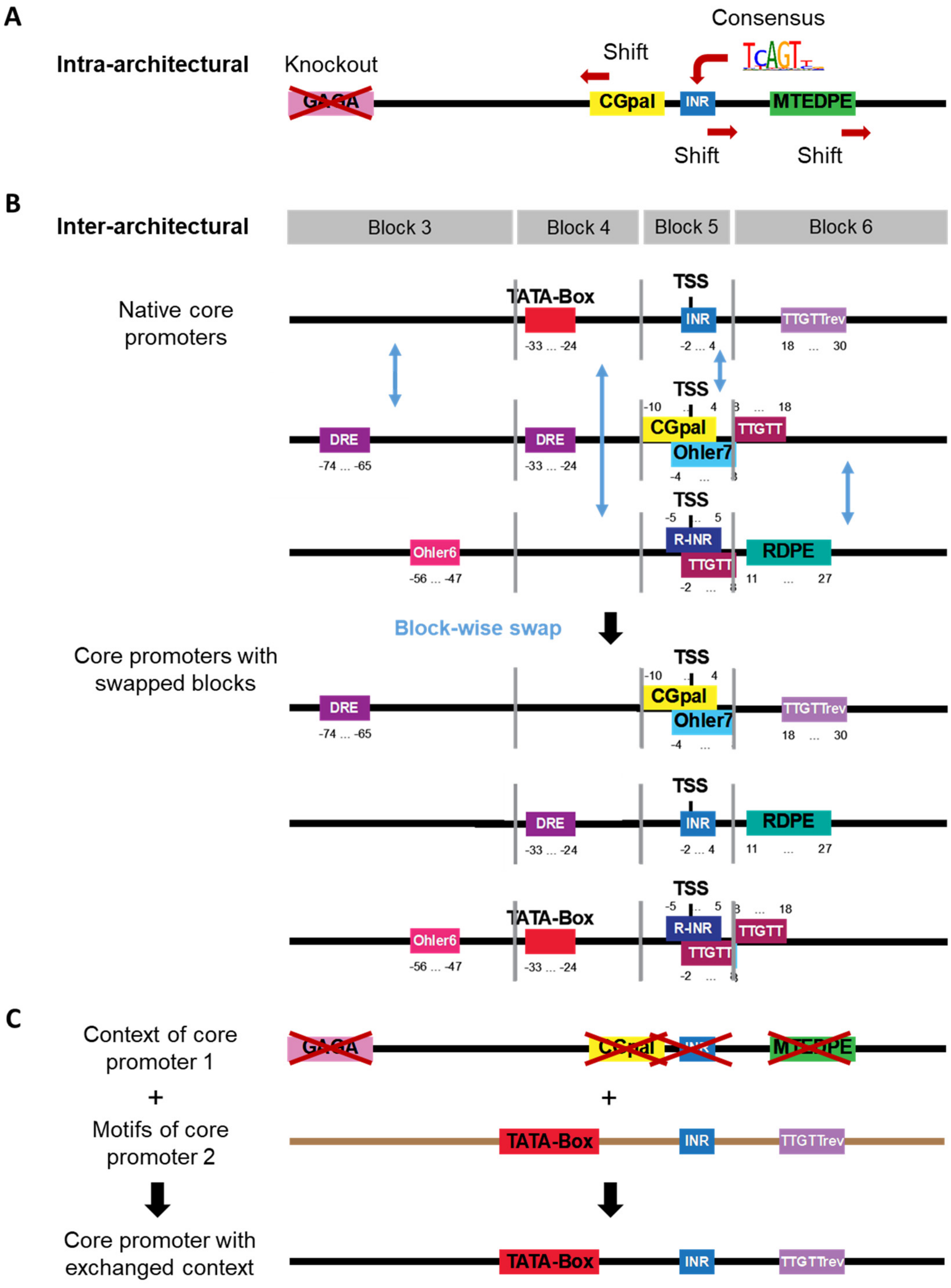


Figure 8. Combinatorial mutations designed to the motif-rich core region. (A) An illustration of the intra-architectural motif-wise combinatorial mutations within the core promoter (2023 sequences designed in total). Both of the motif strength and the motif position are changed. **(B)** An illustration of the inter-architectural block-wise combinatorial mutations between different core promoters (478 sequences designed in total). **(C)** An illustration of the context exchange between different core promoters (30 sequences designed in total).

4.2 Ecdysone receptor binding site (block 2)

The block 2 which contained three EcR/USP heterodimer binding sites with 17 bp spacers in between was synthesized by oligo annealing (5’-gcGGTCTCA*ATGA*agttcattgacctagtgagaattcacagcgagttcattgacctactcaaggcatacatgaagttcattgacct*GGAT*TGAGACCgc-3’, lowercase with underline: EcR/USP binding sites from JASPAR database (Khan et al., 2018); italic: assembly overhangs; uppercase with underline: BsaI restriction sites).

4.3 Nucleosomal context (block 1 and block 7)

After MNase digestion of chromatin, genome-wide nucleosome maps were generated including nucleosome positions and occupancy relative to TSS (especially ±1 nucleosomes) (unpublished data generated in our lab). Accordingly, 12 pairs of block 1 and block 7 representing different potential ±1 nucleosome patterns of 12 genes were selected (Table 4; sequences in Table S5 and S6) and generated by either PCR amplification from the genomic DNA (isolated from sequenced fly strain, stock number 2057 in Bloomington *Drosophila* Stock Center) or oligo synthesis from Life Technologies (for HindIII recognition sites mutated and ATGs mutated sequences). All synthesized sequences of block 1s and block 7s contained BsaI sites and assembly overhangs, and they were stored in TOPO vectors (Zero Blunt TOPO PCR Cloning Kit, Invitrogen).

In the experiments, we tested the block 1 and block 7 in pair with all 19 native core promoter block 3-6s, five out of which were then selected to combine with the free combinations of block 1 and block 7 (one from each architecture with activities covered the entire dynamic range: FBgn0034642 (Ar.0), FBgn0030993 (Ar.1), FBgn0014865 (Ar.2), FBgn0027597 (Ar.3), FBgn0010078 (Ar.4)). We also constructed synthetic promoters containing only block 1s (without block 7) for these five wild-type block 3-6s. One pair block 1.11 and block 7.11 was selected based on its high expression level and used as the fixed nucleosomal context for highly mutated block 3-6s in our study.

Block 1	HindIII mutated	Gene	TSS distribution		Block 7	ATG mutated
1.1	No	FBgn0030993	NP		7.1	Yes
1.2	No	FBgn0034638	NP		7.2	Yes
1.3	No	FBgn0086519	NP		7.3	Yes
1.4	No	FBgn0031886	BP		7.4	Yes
1.5	No	FBgn0035754	BP		7.5	Yes
1.6	Yes	FBgn0028648	BP		7.6	Yes
1.7	No	FBgn0039589	BP		7.7	Yes
1.8	No	FBgn0036263	BP		7.8	Yes
1.9	Yes	FBgn0027597	BP		7.9	Yes
1.10	Yes	FBgn0035060	BP		7.10	Yes
1.11	No	FBgn0033924	BP		7.11	Yes
1.12	Yes	FBgn0010894	BP		7.12	Yes

Table 4. A summary of block 1s and block 7s used as the nucleosomal context in our experiments. 12 pairs of block 1 and block 7 representing different genomic ±1 nucleosome patterns of 12 genes based on the genome-wide MNase digestion of chromatin (unpublished data generated in our lab, sequence details in Table S5 and S6).

5. Reporter and control plasmids for dual luciferase assay

A two-vector system was used in the experiments. Firefly reporter vector backbone was derived from a commercial vector pGL4.13 with luc2 firefly gene (Promega). HindIII and BglII restriction enzymes (NEB) were used to cut out the SV40 early enhancer/promoter region in the original plasmid. To insert BsaI sites and 4 bp overhangs, two dsDNAs with HindIII and BglII sites were generated by oligo annealing: for the constructs containing a block 7, the following sequence was used: gcagatctgc*GAAC*TGAGACCgtcgacgcaaggcctgcaattaatgcagcggccgatcggcatatgGGTCTCA*CCAC*caaagcttcg (only forward sequence; BglII or HindIII restriction sites: lowercase with underline; overhangs: italic; BsaI restriction sites: uppercase with underline); the sequence used for the constructs without block 7 was: gcagatctgc*GAAC*TGAGACCgtcgacgcaaggcctgcaattaatgcagcggccgatcggcatatgGGTCTCA*TCTG*caaagcttcg. After enzymes digestion and gel purification (QIAquick Gel Extraction Kit, Qiagen) of both vector and inserted DNAs, ligation (Rapid DNA Ligation Kit, Roche) was performed to obtain the two final vector backbones (4299 bp), named as BB0 for the constructs without block 7 and BB1 for the constructs containing a block 7.

Renilla control plasmid (3630 bp) was derived from another commercial vector pGL4.70 with the hRluc renilla gene (Promega) by insertion of a moderate-strength P transposase (pTran) promoter between NheI and XhoI sites. The pTran promoter was cloned from a vector created in the lab pKF1 (from Dr. Katja Frühauf in our lab, derived from a P-element sequence, position 34-141 according to (O'Hare & Rubin, 1983)) using primers: 5'-GCGCTAGCAGCCGAAGCTTACCGAAGTATAC-3', 5'-GCCTCGAGCCACGTAAGGGTTAATGTTTTC-3' (underlines: NheI and XhoI restriction sites).

Several inter-plate controls were used in the experiments. The negative control was one commercial vector pUC19 (NEB). There were two positive controls: one was pGL4.10 vector (Promega, with luc2 firefly gene) with pTran promoter inserted between NheI and XhoI sites, termed as pUG9, whose signal was used in data normalization procedure (4350 bp); the other one was a synthetic test plasmid pZQ3 (4691 bp) with moderate promoter activity which contains our firefly reporter backbone BB0 and blocks 1-6 for ecdysone inducibility check: Block 1.3 (sequence listed in Table S5) + Block 2 (sequence in Section 4.2) + Block 3-6 with INR and DPE motifs (sequence: GGCTCCGAATTCGCCCTTTTCCCAGGGCGGCAGAGGCAAAAATTTGCCGA

TCCCAGAGCCAGCCGACTCATTCAAAGCTCCGACTTCGTTGCGTGCACACAGAGTCTCAAGGGCGACCCAGCTTT).

6. Experimental setup and procedures

For carrying out our large-scale systematic analysis, we developed a high-throughput experimental pipeline using automated robot systems (Figure 9, steps highlighted in blue were implemented using automation).

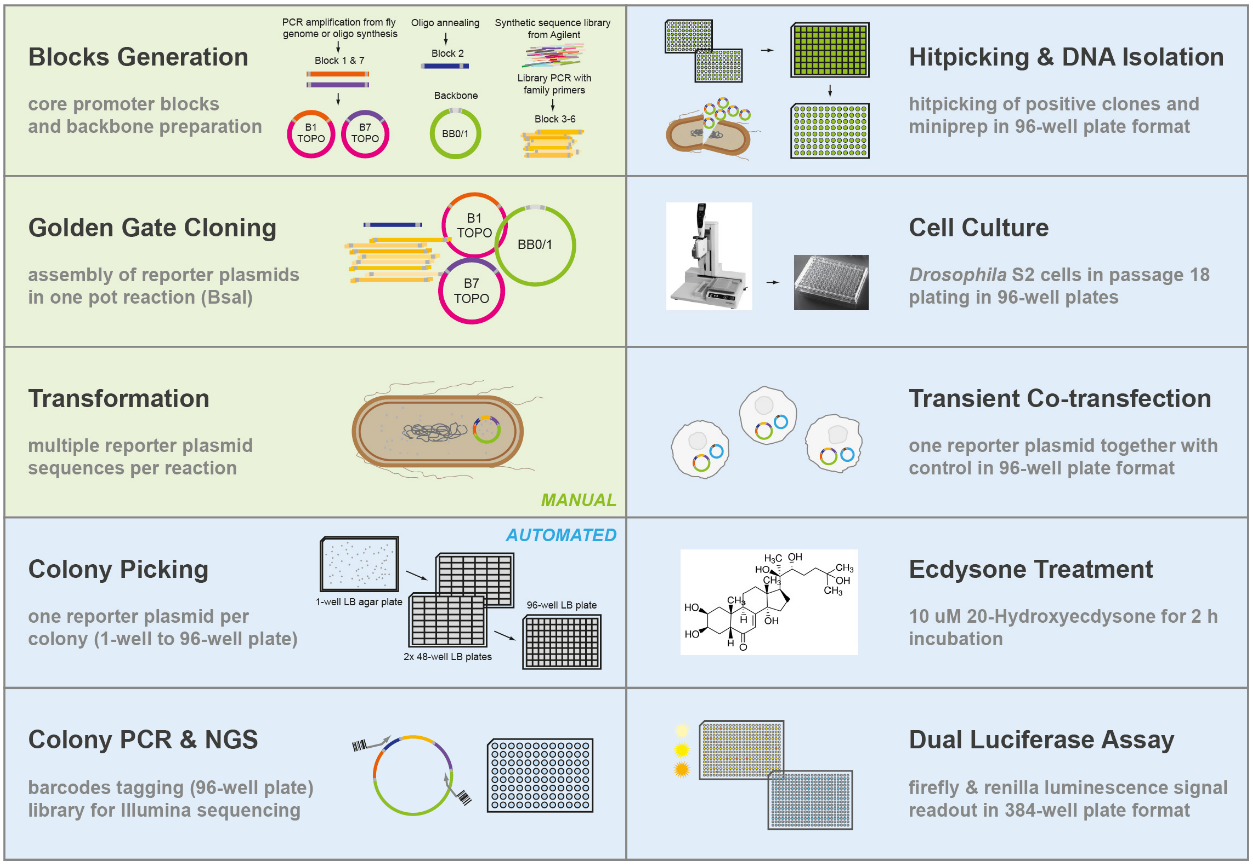

Blocks Generation
core promoter blocks and backbone preparation
PCR amplification from fly genome or oligo synthesis
Block 1 & 7
Oligo annealing
Block 2
Backbone
Synthetic sequence library from Agilent
Library PCR with family primers
Block 3-6
B1 TOPO
B7 TOPO
BB0/1
Hitpicking & DNA Isolation
hitpicking of positive clones and miniprep in 96-well plate format
Golden Gate Cloning
assembly of reporter plasmids in one pot reaction (BsaI)
B1 TOPO
BB0/1
B7 TOPO
Cell Culture
Drosophila S2 cells in passage 18 plating in 96-well plates
Transformation
multiple reporter plasmid sequences per reaction
MANUAL
Transient Co-transfection
one reporter plasmid together with control in 96-well plate format
AUTOMATED
Colony Picking
one reporter plasmid per colony (1-well to 96-well plate)
1-well LB agar plate
2x 48-well LB plates
96-well LB plate
Ecdysone Treatment
10 uM 20-Hydroxyecdysone for 2 h incubation
Colony PCR & NGS
barcodes tagging (96-well plate) library for Illumina sequencing
Dual Luciferase Assay
firefly & renilla luminescence signal readout in 384-well plate format

Figure 9. Overview of the experimental pipeline for synthetic promoter analysis. Different blocks generation, Golden Gate cloning and transformation were performed manually (in green); followed by colony picking, colony PCR for Illumina sequencing library preparation, hitpicking, plasmid isolation, *Drosophila* S2 cells culture, transient co-transfection, ecdysone stimulation and dual luciferase assay, which were carried out using automation (in blue).

After preparation of each construct block (block 1 and block 7: PCR amplification from the fly genome or oligo synthesis; block 2: oligo annealing; block 3-6: PCR amplification from the synthetic library according to mutation families), Golden Gate cloning (BsaI cloning) was applied to join them with the vector backbones sequentially. Then, the newly synthesized reporter plasmids were transformed into electrocompetent *E. coli*, followed by plating an optimal amount of bacteria on one-well plates that facilitating automated colonies picking using the robotic workstation. After bacterial growth in 48-well plates, we rearranged them into 96-well LB plates and prepared the library for next-generation sequencing with two-step PCR using nested barcode primers. Based on the sequencing results, replicates and bad clones were screened out and DNAs from confirmed positive clones were isolated. These firefly reporter plasmids containing all the distinct promoters were then used for transient co-transfection into *Drosophila* S2 cells along with the renilla control plasmid in 96-well plates. After overnight incubation, cells were treated with ecdysone for another 2 hours. Four cell culture plates were coupled into 384-well plates for the final dual luciferase assay readout.

6.1 Automation

We used two independent robot platforms that show the similar basic configuration of pipettor systems (Biomek NX^{P} automated workstations with Multichannel-96 and Span-8 pipetting model, Beckman Coulter). Additional instruments were integrated with the original workstations including incubators (Incubator Shaker DWP, Inheco), thermocyclers (Biometra TRobot, Analytik Jena), barcode printer (Microplate Print & Apply, Beckman Coulter), barcode reader (Compact Laser Barcode Scanner, Omron Microscan), plate reader (SpectraMax Paradigm Multi-Mode Microplate Reader, Molecular Devices), plate sealer (Wasp, Kbiosystems). They were designed for maximum flexibility to perform many different experiments. Specifically, one system is dedicated to bacterial experiments, mainly the cloning-related work: colony picking, colony PCR, hitpicking for positive clones, DNA isolation and concentration measurement. The colony picking is a customized feature of this robotic configuration that the automation specialist Peter Bandilla in our group previously developed and implemented on the system together with the Beckmann Coulter company. The other system is dedicated to *Drosophila* cell assays: transient co-transfection, ecdysone treatment and luciferase assay readout. In addition, an electronic multichannel pipette on an assistant robot (VIAFLO Electronic Multichannel Pipette + ASSIST Pipetting Robot, INTEGRA) was used for automated cell plating into 96-well plates.

6.2 Synthetic library amplification

Block 3-6s for the motif-rich core regions of our synthetic promoter constructs were amplified from the synthetic library (synthesized by Agilent Technologies). The entire oligo pool (lyophilized, 10 pmol) was dissolved in 100 µl Elution buffer (Qiagen) and shaken at room temperature (RT) for 30 min at 450 rpm and 10 min at 950 rpm. 0.5 µl of library DNA was used to amplify the specific sequence family (native sequences or one of distinct mutation families) in a 20 µl PCR reaction, which also included 1.25 µl of both forward and reverse 10 µM customized primers, 4 µl 5× Herculase II reaction buffer, 0.5 µl 10 mM dNTP mix and 0.5 µl Herculase II fusion DNA polymerase (Agilent Technologies). PCR parameters were as follows: 98 °C for 3 min; followed by 15 cycles of 98 °C for 80 s, 54 °C for 30 s, 72 °C for 40 s; 72 °C for 10 min. Each PCR reaction was purified with the QIAquick PCR purification kit (Qiagen) according to the manufacturer's instructions and eluted in 30 µl of nuclease-free water (Qiagen).

6.3 Golden Gate cloning

The Golden Gate cloning uses type II restriction enzymes to cleave DNA outside the recognition site while creating 4 bp overhangs which allow unique and directional fusion of DNA fragments (Engler, Gruetzner, Kandzia, & Marillonnet, 2009; Engler, Kandzia, & Marillonnet, 2008). Here, BsaI restriction enzyme (10,000 U/ml, NEB) and T4 DNA ligase (3 U/µl, Promega) were applied to assemble all of the synthetic promoter blocks sequentially and simultaneously into the firefly reporter vector backbone in a one-pot reaction (Figure 10; 4 bp assembly overhangs were shown explicitly).

For each 20 µl reaction, DNA master mix contained equimolar amount (80 fmol) of each part: block 1 in TOPO vector (3784 bp), block 2 (99 bp), block 3-6 (200 bp), block 7 in TOPO vector (3785 bp, if needed) and backbone (4299 bp) together with 2 µl BsaI, 2 µl T4 DNA ligase and 2 µl 10× ligase buffer. The cloning protocol included 3 steps: (1) 20 cycles of 37 °C for 2 min, 16 °C for 3 min; followed by 50 °C for 5 min and 80 °C for 5 min; (2) After adding 1 µl BsaI, 1 µl T4 DNA ligase, 1 µl 10 mM ATP: 16 °C for 20 min; 15 cycles of 37 °C for 2 min, 16 °C for 3 min; followed by 50 °C for 5 min and 80 °C for 5 min; (3) After adding again 1 µl BsaI: 37 °C for 10 min, 50 °C for 20 min, 80 °C for 10 min and ramp down to 25 °C by 0.1 °C/s.

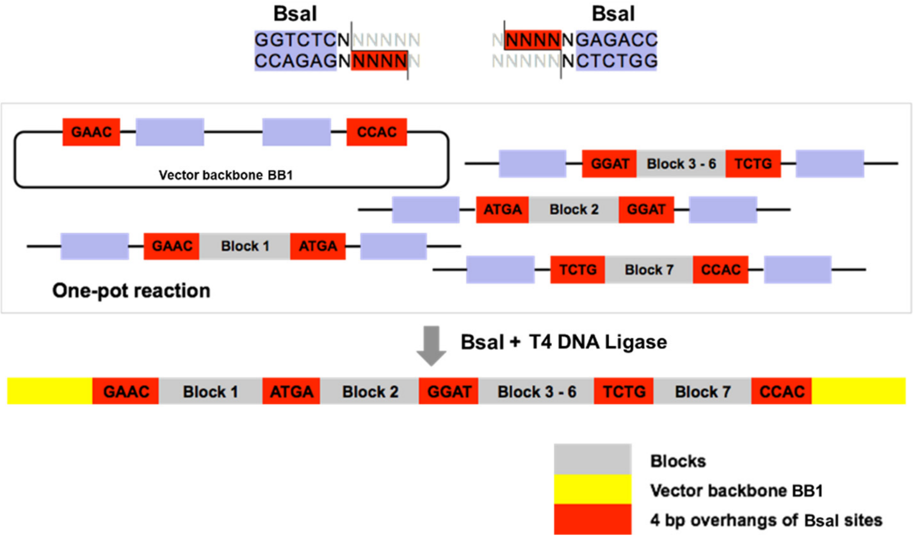


Figure 10. Golden Gate cloning (BsaI cloning) strategy. BsaI and T4 DNA ligase were used to assemble all blocks sequentially and simultaneously into the reporter vector backbone BB1 in a one-pot reaction. The synthetic promoter construct with block 7 is illustrated as the example here. For construct without block 7, the vector backbone BB0 contained directly the TCTG overhang.

6.4 Transformation

After BsaI cloning, 2 µl of the reaction mix was transformed into 40 µl of electrocompetent TOP10 *E. coli* cells (homemade). After electroporation (1.8 kV for 0.1 cm cuvettes, Gene Pulser, Bio-Rad) and 1 ml SOC medium (homemade) addition, cells were incubated for 1 h at 37 °C (shaking at 450 rpm) and plated 100 µl onto prewarmed 1-well LB-agar plates supplemented with 100 µg/ml Ampicillin (appropriate dilution of cells with LB medium for optimal colony density with around 60-70 detectable separated colonies per plate).

6.5 Colony picking

After overnight incubation at 37 °C, the 1-well plates were ready for colony picking. Span-8 pipetting system on the robot was used to automatically pick individual colonies (customized protocol) into two 48-well plates (Riplate SW 48, 5 ml, Riplate) with 2.4 ml LB-Ampicillin medium (Ampicillin concentration: 120 µg/ml). The plates were incubated for 16 h at 37 °C (horizontally shaking at 180 rpm) and rearranged into one 96-well plate (MegaBlock 96 Well, 2.2 ml, Sarstedt). 110 µl/well of bacteria was used to create glycerol stock plate (Round 96 Well Storage Plates, U-bottom, 330 µl, 4titude) and 30 µl/well for PCR plate (FrameStar 96 Well Skirted PCR Plate, 4titude) ready for sequencing library preparation. Since in the previous cloning step, the sequences from the same mutation family were all mixed together, it is technically impossible to recover all of them during the colony picking step. Therefore, we did over-picking of the colonies and were able to recover more than 65% of the designed sequences.

6.6 Next-generation sequencing of the picked clones

Two-step PCR with nested barcode primers was implemented for library preparation. The forward and reverse primers for 1st PCR targeted the sequences in block 2 and vector backbone respectively with specific barcodes (Table S7; block 1 was always known in the BsaI cloning procedure). Their combinations for each sample in a 96-well plate are shown in Table S8. 2 µl/well of bacteria were used to set up a 25 µl PCR reaction containing 1 µl homemade Taq/Pfu polymerase mix, 2.5 µl primer mix (forward and reverse each 500 nM), 1 µl 25 mM $MgCl_2$, 2.5 µl 10× buffer, 1 µl 2.5 mM dNTP. 96-well plate PCRs were performed in the thermocyclers integrated on the robot (96 °C for 7 min; 3 cycles of 94 °C for 30 s, 68 °C for 30 s, 72 °C for 2 min; followed by 3 cycles of 94 °C for 30 s, 64 °C for 30 s, 72 °C for 2 min; 17 cycles of 94 °C for 30 s, 56 °C for 30 s, 72 °C for 2 min).

5 µl/well of the product from each 1st PCR plate was pooled into one specific well of the collection plate (Deepwell plate 96/500 µl, Eppendorf; each well containing all 96 samples from one 1st PCR plate). 3.5 µl/well was then used as template for 2nd PCR in a 50 µl reaction together with 0.5 µl Herculase II fusion DNA polymerase (Agilent Technologies), 10 µl 5× Herculase II reaction buffer, 1.25 µl 10 mM dNTP mix and 5 µl each of Illumina index primers (Nextera XT Index Kit v2, Index 1 (i7) Adapters and Index 2 (i5) Adapters, Illumina). So each well of 2nd PCR plate (each 1st PCR plate samples) got a unique pair of index adapters. PCR was performed as the same protocol for 1st PCR. The final products were pooled and purified using Agencourt AMPure XP magnetic beads (Beckman Coulter) according to the manufacturer's instructions. Next-

generation sequencing (Illumina HiSeq1500) was performed by the LAFUGA sequencing facility at the Gene Center LMU Munich.

6.7 Hitpicking and DNA isolation

Automated hitpicking of positive clones from glycerol stock plates was carried out using our robotic system. 75 µl of the samples in the original plates were reformatted into the final 96-well glycerol stock plates (Round 96 Well Storage Plates, U-bottom, 330 µl, 4titude) and 20 µl were used for reinoculation in 48-well plates (Riplate SW 48, 5ml, Riplate) with 2.4 ml LB-Ampicillin medium (Ampicillin concentration: 120 µg/ml). The plates were incubated for 17 h at 37 °C (horizontally shaking at 180 rpm) and rearranged into one 96-well plate (1.2 ml/well; MegaBlock 96 Well, 2.2ml, Sarstedt). After centrifugation at 5000 g for 15 min, the supernatant was discarded and cell pellets were stored at -20 °C ready for DNA isolation.

Minipreps in 96-well plate format was performed with Wizard MagneSil Tfx^{TM} System (Promega) on the robotic workstation according to the manufacturer's instructions. DNA concentrations were measured using the SpectraMax Microplate Reader integrated on the robot (5 µl DNA samples on the SpectraDrop Micro-Volume Microplates, Molecular Devices).

6.8 Cell culture

Drosophila S2 cells were firstly thawed at passage 12 with Schneider's *Drosophila* Medium (Bio&Sell, supplemented with 10% FBS (Fetal Bovine Serum, Biochrom)) and later cultivated in Express Five SFM medium (protein-free and serum-free, Invitrogen). One bottle of the Express Five medium (1 liter) was supplemented with 90 ml of L-Glutamine (200 mM, Invitrogen). During cultivation, cells were grown at 25 °C without CO_2 in tissue culture flasks (75 cm^2, Corning) and were split into fresh flasks when 90% confluent. The cells in passage 18 were seeded into 96-well plates (Falcon 96 Well Tissue Culture Plates, Corning) with 40,000 cells per well in 100 µl using an electronic multichannel pipette VIAFLO (1250 µl, INTEGRA) on a pipetting robot ASSIST (INTEGRA). The cells 24 h growth rate and viability were monitored in the culture dishes (in duplicate; 100 mm, Corning) with 12×10^6 cells in 14 ml medium. Cell counting and assessment of cell viability were performed using the Cell Counter and Analyzer System (CASY, Roche).

6.9 Transient co-transfection

24 hours after cell plating, transient co-transfection on the robot system was performed using FuGENE® HD Transfection Reagent (Promega) according to the manufacturer's protocol. To avoid multiple freeze-thaw processes, the renilla control plasmid and three inter-plate control plasmids (pUC19, pUG9, pZQ3) were aliquoted in PCR strips sufficient for one transfection experiment. The isolated reporter plasmids and inter-plate control plasmids were transferred into 96-well master mix plates according to the transfection plate layout (Figure 11) together with renilla control plasmids (except for untreated cells (UTCs), reporter plasmid or inter-plate control plasmid : renilla control plasmid ratio = 8 : 1, total DNA amount 0.945 µg per well). Wells indicated with green shadows were filled with various reporter plasmids containing synthetic promoter constructs to be tested.

	1	2	3	4	5	6	7	8	9	10	11	12
A	pUC19	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	UTC
B	pUG9	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
C	pZQ3	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
D	▨	▨	▨	▨	▨	pUG9	▨	▨	▨	▨	▨	▨
E	▨	▨	▨	▨	▨	▨	pUG9	▨	▨	▨	▨	▨
F	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
G	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
H	pUC19	▨	▨	▨	▨	▨	▨	▨	▨	pZQ3	pUG9	UTC

Figure 11. The master mix plate layout for the transient co-transfection in a 96-well plate. The negative controls (pUC19 and UTCs) were located in the four corners. The positive controls pUG9 (4 replicates, the signal used in data normalization) and pZQ3 were distributed in the plate. Other wells (green) were filled with various reporter plasmids containing tested promoter constructs.

2.3 µl/well FuGENE® HD Transfection Reagent was added and the FuGENE® HD-DNA mixture was incubated for 5 min at RT (FuGENE® HD : DNA ratio ~ 2.4 : 1). 10 µl FuGENE®

HD-DNA mixture was then added per well into 96-well cell culture plates. The transient co-transfections were performed in duplicates for cells with and without ecdysone treatment.

6.10 Ecdysone treatment

Cells were incubated for 22 h after transfection, followed by 2 h of ecdysone treatment (final ecdysone concentration: 10 µM; 20-Hydroxyecdysone, Sigma-Aldrich). The other replicate transfected cell plate was treated with the same volume (10 µl/well) of cell culture medium (Express Five medium supplemented with L-Glutamine) and incubated for 2 h.

6.11 Dual luciferase assay

40 µl/well of the medium was removed from each cell culture plate and 20 µl/well of cells were transferred into the final readout plates. For each measurement, samples from four 96-well cell culture plates were joined into two 384-well plates (one for firefly luminescence measurement, the other for renilla luminescence measurement; AlphaPlate-384, PerkinElmer). ONE-Glo^{TM} Luciferase Assay System (Promega) and Renilla-Glo$^{®}$ Luciferase Assay System (Promega) were used respectively (reagent amount: 20 µl/well). There was a common crosstalk issue between two adjacent wells caused by the bleed-through of the stronger luminescence signal to the other. In the optimized protocol, firefly luminescence signal was measured twice with strong signals (> 2×10^5 RLU, relative light unit) identified in the first measurement and removed before the second measurement (samples were pipetted out and a quencher (1 mM Nile Blue A, Sigma-Aldrich) was added instead). This experimental procedure was designed to solve the crosstalk issue and correct overestimated renilla signals (more details described in Section 8.3). Bioluminescence signals were measured using the SpectraMax Microplate Reader (Molecular Devices).

7. Computational methods

7.1 Reads mapping for the sequencing results of the picked clones

This work was performed by Dr. Mark Heron (from the Söding lab, Gene Center LMU Munich). Sequencing reads were demultiplexed based on the Illumina indexes and the designed barcodes in our customized primers. The most enriched sequence (at least 3-fold enrichment against the second most frequent sequence) for each sample was used and trimmed to match the target region of our synthetic promoter construct (part of block 2, blocks 3-6 and block 7). The trimmed reads were mapped to our designed library using the pairwise alignment method.

7.2 Data preprocessing and normalization

Data obtained in dual luciferase assays from each individual luminescence readout were reformed from the 384-well plate format back into four 96-well plate formats according to the transfection layout of specific reporter plasmid samples. For each plate, firefly luciferase expression values (*FF*) of each tested samples were normalized to their renilla luciferase values (*REN*) as well as *FF* values of the inter-plate controls. The 1st firefly measurements (*FF1*) were used as the readout values for samples with strong promoters (*FF1* > 2×10^5 RLU) and the 2nd firefly measurements (*FF2*, signal degradation corrected) were used for other weaker samples.

Background value (*BG*) was calculated as the arithmetic mean of negative control signals (pUC19 and UTCs) got from 2nd firefly measurements (avoiding the potential crosstalk issue; Equation 1). Normalized value of positive control pUG9 ($Norm_{pUG9}$) was defined as the arithmetic mean of its *FF1* signals with *BG* subtracted divided by its *REN* signals (Equation 2).

$$BG_{FF2} = mean\ (\ pUC19_{FF2} + UTC_{FF2}\ ) \qquad (1)$$

$$Norm_{pUG9} = mean\ (\frac{pUG9_{FF1} - BG_{FF2}}{pUG9_{REN}}) \qquad (2)$$

The final normalized luciferase expression value for each tested sample (x_i) was calculated as Equation 3: its FF_i signal ($FF1$ for strong promoters and $FF2$ for others) with BG subtracted was firstly normalized to its REN_i signal and then to the normalized control $Norm_{pUG9}$; the value was then log_2-transformed. This value was used as the estimate of the corresponding synthetic promoter activity.

$$FF_i = \begin{cases} FF1_i\,, & if\ \ FF1_i \leq 2 \times 10^5\ RLU \\ FF2_i\,, & if\ \ FF1_i > 2 \times 10^5\ RLU \end{cases}$$

$$x_i = log_2\left[\frac{1}{Norm_{pUG9}} \times \left(\frac{FF_i - BG_{FF2}}{REN_i}\right)\right] \tag{3}$$

7.3 Outlier identification and filtering

We firstly filtered out samples with irregular renilla signals (REN > 10000 RLU or REN < 300 RLU) and then calculated the median and standard deviation (SD) for normalized luciferase signals of each promoter construct x (> 88% with at least three replicates for both with and without ecdysone stimulation). The score used for defining outliers was calculated as:

$$score = \frac{x_i - median(x)}{median(SD(X))} \tag{4}$$

Here, x_i, as described above, represented the normalized expression value of i^{th} replicate for construct x. $SD(X)$ denoted all SDs for entire synthetic promoter construct library X. The scores with an absolute value of no less than 3 were labeled as outliers and were excluded from further analysis.

7.4 Quantification and statistical analysis

Data analyses were implemented using Origin 8.1 (only for results in Chapter 8) and R with packages for data manipulation and visualization including “dplyr”, “magrittr” and “ggplot2” (as part of the package collection “tidyverse”). Statistical analyses used the simple linear least squares regression to fit the linear trend of the data with usually the 95% confidence interval and the Pearson correlation coefficient (PCC) r shown as well. To compare the paired mean expressions of ATG-containing promoters versus their ATG-less ones in Section 9.1, one-sided Wilcoxon signed rank test was used. To assess the difference of ecdysone responsiveness between constitutive and developmental promoters in Section 9.5, non-parametric Wilcoxon rank-sum test was performed. This test was also used in comparing the expression levels not distributed normally, such as for mutation tests of a certain motif combined in different core promoter configurations. When the expression measurements were from a single core promoter, we applied a two-sample t-test. Except data shown in Section 9.1 and 9.5, arithmetic mean expression of each synthetic promoter construct $\bar{x}$ was mostly used in the results. When assessing the effect of a mutation, the expression fold change relative to the native expression was calculated. All data were separated by ecdysone condition and the expression fold change caused by induction was computed as the difference between expression with ecdysone and without ecdysone treatment.

7.5 Deriving the activity logos based on expression measurements

To generate the activity logos of specific motifs based on the effects of point mutations of their consensus sequences, we used the corresponding expression measurements to calculate the nucleotide probability at each position of a motif as the element in a position probability matrix (PPM):

$$P_j = \frac{\overline{2^{x_j}}}{\sum_{j \in (A,C,G,T)} \overline{2^{x_j}}} \tag{5}$$

where x_j are the expression values (log_2-transformed) of the motif with nucleotide *j* at a particular position. Notably, we assumed that each nucleotide is independent from each other in the motif. The arithmetic mean of the expression in linear scale was used here. If certain point mutation was not recovered (rarely during colony picking or DNA isolation procedure, only for non-important bases), expression value (in linear scale) of 10^{-5} was artificially assigned.

The information content (IC, in bits) at each position *k* of a motif was calculated using the probability values from Equation 5 as:

$$IC_k = 2 + \sum_{j \in (A,C,G,T)} P_j \times log_2(P_j) \qquad (6)$$

The sequence logo was plotted as the height of different letters at position *k* ($h_{j,k}$) given by:

$$h_{j,k} = P_j \times IC_k \qquad (j \in (A,C,G,T)) \qquad (7)$$

7.6 Predicting the expression levels of samples with combinatorial mutations based on individual sequence features

For sample core promoters with intra-architectural combinatorial mutations, a linear regression model was used to predict the expression value of a given promoter based on the combination of single mutations, including all individual mutations changing either motif strength or motif position. For each promoter construct tested here, we scanned for mutation occurrences in all intra-mutated samples and assigned the variables m_i in the model as the qualitative indicators of the mutation existence (Equation 8). All of the coefficients α_i were learned from the model. The PCC between measurements and predictions was computed and the estimated range of our measurement noise (± *3×median(SD(X))* as used in outlier filtering procedure) was also considered when assessing the model performance. Additionally, we also built an additive model to predict the expression based on the combination of every single mutation's effect on the native promoter expression (Equation 9). The intercept α_0 represented the native expression measurement of the given promoter construct and the other coefficients α_i were calculated as the measured expression deviations of each single mutations compared to the native.

Linear regression:

$$Prediction = \sum_{i=0}^{n} \alpha_i m_i \ ; \quad m_i = \begin{cases} 1\ ; \ \ if\ i^{th}\ mutation\ exists \\ 0\ ; \ \ if\ i^{th}\ mutation\ does\ not\ exist \end{cases} \tag{8}$$

Additive model based on measurements:

$$Prediction = \alpha_0 + \sum_{i=1}^{n} \alpha_i m_i \quad (\,\alpha_0 : native\ expression\ value;\ \alpha_i = \overline{x_{m_\imath}} - \alpha_0) \tag{9}$$

For inter-architectural combinatorial mutations, a linear regression was used with the occurrences of distinct block sections (0/1) as the variables in the model.

III RESULTS

8. Establishing the reliable workflow for automated cell transfections and dual luciferase assays

Establishing the final reliable protocols involves various biological and technical optimizations. The main sources of variability in the experiments include the different batches of reporter plasmid minipreps, the variations from the number of cells and the transfection efficiency. Transfer of manual protocols onto automated workstations were the first step to reduce variability, which was conducted with the help of the automation specialist Peter Bandilla in the lab. Previous experimental tests including one-vector vs. two-vector system comparison, selection of different promoters for control renilla luciferase gene hRluc, testing of various firefly luciferase genes (luc2, luc2P, luc2CP) and different combinations of ecdysone receptor binding sites had been done in the lab to set the primary vector system for dual luciferase assays. Experimentally, my contribution was the optimizations of multiple specific procedures of the workflow: cell culture, transient co-transfection, ecdysone incubation and luciferase assay. Experimental conditions were tested using several tested plasmids, including a commercial luciferase reporter vector pGL4.13 (also used to generate the backbones for our synthetic promoter constructs) which contains an SV40 promoter and a firefly luciferase gene luc2, a GFP expression vector (kindly provided by the Förstemann lab, Gene Center LMU Munich), and a synthetic test plasmid pZQ5 with strong promoter activity which contains our firefly reporter backbone BB0 and blocks 1-6 (Block 1.3 + Block 2 + Block 3-6 with TATA-box, INR and two DPE motifs: GGCTCCGAATTCGCCCTTTTCCC AGGGCGGCAGAGGCTATATAAAGCCGATCCCAGAGCCAGCCGACTCATTCAAAGCT CCGACTTCGTTGCGTGCGGTCGGAGTCTCAAGGGCGACCCAGCTTT). The pGL4.13 vector and the GFP vector were used for optimization of the transfection procedure and the pZQ5 plasmid was used for testing the ecdysone induction conditions. For later luciferase assay optimizations, real samples from the synthetic library were used for more comparable results.

8.1 Optimization of cell culture and the transfection procedure

For optimal results, we seek for high well-to-well reproducibility and high transfection efficiency, which can be achieved by low variability in the number of cells and optimization of the transfection procedure. The cell density was kept the same for each well (40,000 cells in 100 µl). In addition, the cells 24 h growth rate and viability were checked and strictly monitored during passaging (details described in Section 6.8). For successful transfection of DNA into cultured cells, the ratio between Fugene® HD transfection reagent and DNA needs to be optimized. In general, Fugene® HD : DNA ratios of 1.5 : 1 to 4 : 1 work well according to the manufacturer's instructions. Other factors including total DNA amount, incubation time, cell culture medium also need to be considered. 24 hours incubation time and the serum-free medium had been tested before and widely used in the lab. In order to get the optimal conditions with small amount of reagent and high pipetting accuracy, different volumes of Fugene® HD reagent (1.5 µl to 3.3 µl, step: 0.2 µl), the total DNA amount (0.945 µg, 1 µg, 1.2 µg, 1.35 µg, 1.5 µg, 2 µg) as well as the mass ratios between the firefly reporter plasmid (pGL4.13) and the renilla control plasmid in our two-vector system (1 : 1, 2 : 1, 5 : 1, 8 : 1, 10 : 1, 15 : 1 and 20 : 1) were systematically tested. We obtained strong firefly signals and reliable renilla control signals (~ 1000 RLU) with as little as 2.3 µl Fugene® HD for 0.945 µg DNA (the Fugene® HD: DNA ratio is around 2.4 : 1) and the two-vector ratio of 8 : 1. These conditions provided a strong signal-to-noise ratio.

Our automated transfection aims to get not only high but also homogeneous transfection efficiency in each well of the final 96-well plate. The transfection reagent we used (Fugene® HD) has low stability in aqueous environment and tendency to absorbed by plastic surfaces. Thus, fast operation and careful handling of the reagent were also necessary to be taken into account when automating the pipetting procedure. Obvious variabilities could be seen from the results of directly adding Fugene® HD into each well as we did in the manual protocol, which was mainly attributed to the long incubation time for early-added samples. Therefore, eight sterile borosilicate tubes were used to temporarily store the Fugene® HD reagent (comparable with the original supplied glass vials) to facilitate the short pipetting process (around 10 min, column-wise pipetting and mixing) using the Span-8 pipettor (Biomek NX^{P} workstation, Beckman Coulter) which minimized the complexing time differences between individual wells and limited Fugene® HD adsorption to the plastic tip surface (Figure 12A). Through this optimization, high transfection efficiency (50-60%)

with good well-to-well reproducibility (expression levels of the test plasmid expressing GFP varied less than 20% among wells) was obtained (Figure 12B).

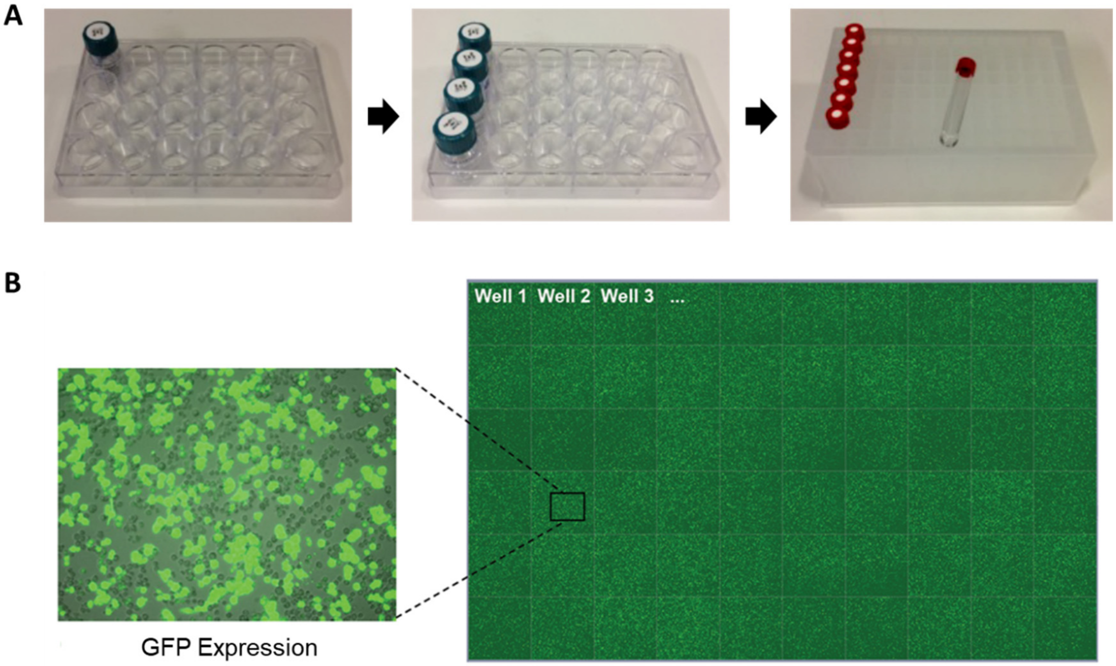


Figure 12. Optimization of the automated transfection procedure. (A) The container for the Fugene® HD transfection reagent was changed from one original supplied glass vial to four different glass vials and in the end to eight sterile borosilicate tubes. **(B)** Confocal tiled images of 60 S2 cells samples (right panel) transfected with a GFP expression plasmid in a 96-well plate. The transfection efficiency was around 50-60% with a high reproducibility between different wells. The left panel is a magnification of a region of one well showing the high number of transfected cells (transmitted light channel in grey).

8.2 Optimization of the ecdysone treatment procedure

Different ecdysone concentrations and incubation time lengths were tested to optimize the stimulation conditions. Twelve ecdysone concentrations (0.1 µM, 0.15 µM, 0.25 µM, 0.4 µM, 0.625 µM, 1 µM, 2 µM, 3.2 µM, 5 µM, 7 µM, 10 µM, 12 µM) were used to check the inducibility of our synthetic promoter in the test plasmid pZQ5. The expression change upon ecdysone stimulus was quite sensitive. With only 0.1 µM ecdysone, we could already get more than 95% of the maximum induction (Figure 13A). In the final protocol, 10 µM was chosen since it provided the highest and the most stable inducibility level as it is within the concentration range of the plateau of expression. We also checked the ecdysone incubation time of 2 hours, 3 hours, 4 hours and 5 hours. Longer incubation time generally gave a higher signal (Figure 13B). However, too high signals could lead to severe luminescence crosstalk between adjacent wells, saturation issues of promoter activity or exceeding the detection range of the detector. The 2 h incubation time was used in the final protocol since it could already drive measurable induction (over 3-fold signal increase for this test promoter pZQ5, shown in Figure 13B) while avoiding the saturation of promoter activity.

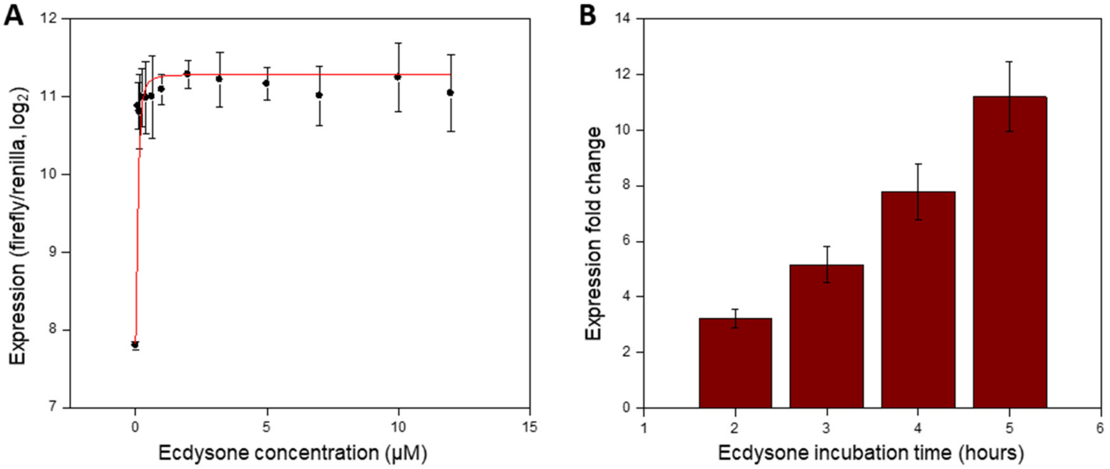


Figure 13. Ecdysone stimulation under different conditions. (A) Expression level (calculated by the ratios of firefly signal and renilla signal) as a function of ecdysone concentration. Twelve different concentrations were used with 5 h incubation time. Error bars: SD, n=3. Fitting with the modified Hill function. **(B)** Expression fold change after ecdysone induction with an incubation time of 2 h, 3 h, 4 h and 5 h, respectively. Expression fold changes were calculated as the ratios of expression levels with 10 µM ecdysone induction and without ecdysone induction. Error bars: SD, n=3.

8.3 Optimization of the dual luciferase assay

Originally, 96-well plates had been used throughout the procedures of cell culture, transient co-transfection and luciferase assay readout. In order to shorten the assay duration and reduce the materials consumption without impacting the cell and transfection quality, we transferred the treated cells (after transfection and ecdysone incubation) from four 96-well plates into one 384-well plate for the bioluminescence signal measurement using the 96-channel pipettor (Biomek

NX^{P} workstation, Beckman Coulter). This enabled us to gain 4-fold higher throughput and save 2/3 luciferase assay reagent.

The choice of the readout plates is also crucial for the luciferase assay as they can be a source of luminescence background signal and exhibit well-to-well crosstalk. The crosstalk is caused by the bleed-through of luminescence signal from the well which contains very strong signal to the neighbor wells. We tested different types of 384-well plates, including various color-coating plates and half-volume plates with square/round-shape wells. The light gray plates with square wells (AlphaPlate-384, PerkinElmer) gave us the best results with the highest signal, low background (< 100 RLU) and very low crosstalk effect. The level of crosstalk was nearly comparable with the crosstalk obtained by using black plates (which cannot be used in our assay due to strong light absorption in the walls). The four nearest neighbors had the direct crosstalk of around 0.05% and the diagonal crosstalk was about ten-fold less (indicated in Figure 14).

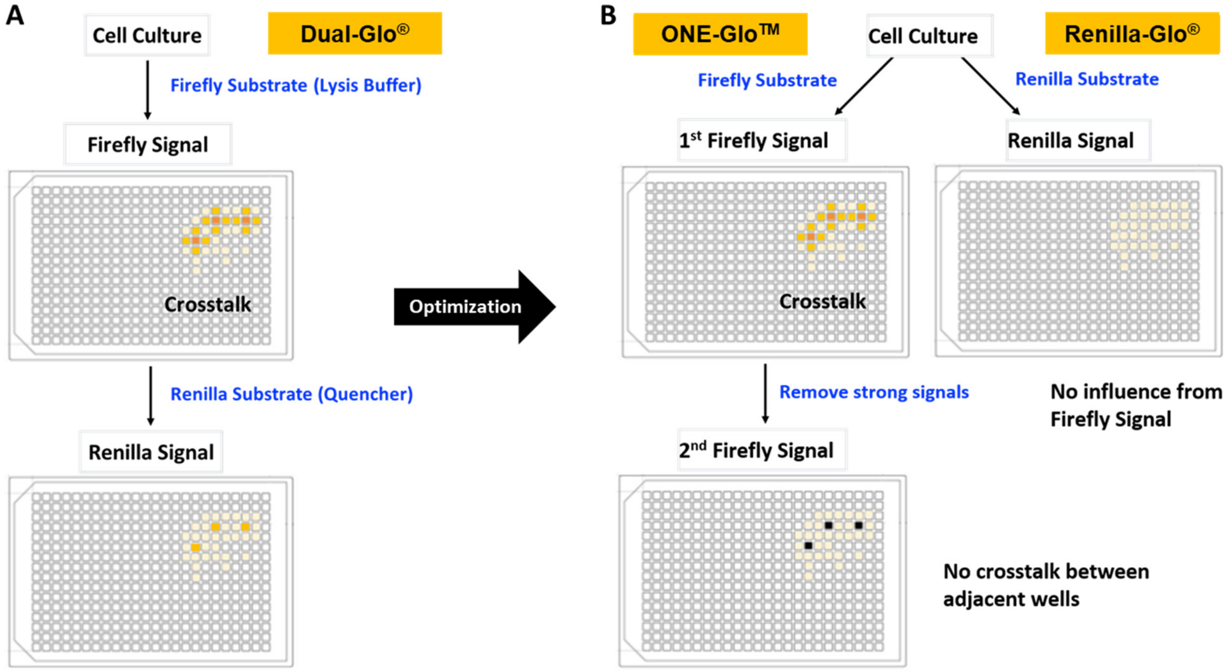


Figure 14. Optimization of the dual luciferase assay procedure. (A) We originally tested the commonly used Dual-Glo® Luciferase Assay System (Promega) which measures the firefly and renilla luminescence signals sequentially in one readout plate. This led to the crosstalk issue of firefly signals between adjacent wells and overestimated renilla signals due to insufficient firefly quenching. **(B)** After our optimization procedure, firefly and renilla measurements were carried out in two separate 384-well plates from the same batch of cells using the ONE-Glo™ and Renilla-Glo® Luciferase Assay Systems (Promega). Firefly signals were measured again with wells exhibiting strong signals (> 2×10^{5} RLU) removed to eliminate the crosstalk issues.

The addition of luciferase reagent and subsequent signal measurement also required optimization. Initially, we used the well-known Dual-Glo® Luciferase Assay System (Promega). With this kit, the firefly substrate containing lysis buffer was first added to the cells and the corresponding luminescence signal was subsequently measured. The control signal was measured immediately after addition of the renilla substrate (including quencher for quenching the firefly signal) to the same plate (Figure 14A). The data obtained using this method presented two issues, however. First, although the crosstalk of 0.05% was negligible for most of the cases, this could still be problematic if extremely strong promoter samples were located directly next to very weak ones. Second, the quenching efficiency of the firefly signals for extremely strong promoters was usually not sufficient and this could lead to overestimated renilla control signals.

We solved both issues by splitting the cells into two different reader plates and measuring the firefly and renilla signals separately. The crosstalk issue was addressed by splitting the firefly measurements into two rounds. After the first measurement, we identified the wells with very strong signals (above 2×10^5 RLU), pipetted them out and added a quencher (Nile Blue) to quench any remaining luminescence signal, and measured the plate a second time (Figure 14B). The 1st measurement was used as the readout for the strong promoters and the 2nd measurement for the others. We modified the data normalization procedure accordingly (more details in Section 7.2), leading to a further improvement of data reproducibility (Figure 15B).

8.4 Evaluation of the assay performance

Dynamic range and reproducibility are important parameters to evaluate the assay performance. We first compared the expression measurements of 11 synthetic promoter constructs prepared manually or with our automated protocol (measured with the same batch of cells in the same readout plate). The measurements provided a broad and linear dynamic range (Figure 15A) that extended over three orders of magnitude. We also found that by using the automated protocol, the coefficient of variation (CV, defined as the ratio of the SD to the mean) was reduced by more than a factor of 2. We also compared the reproducibility of replicate measurements for samples from three different 96-well sample plates measured on different days with or without ecdysone induction (different batches of cells, independent transfection and luciferase assay readout). The results obtained with optimized data normalization procedure (see Section 7.2 for details) showed a considerably lower expression variation with a decrease of around 30% on average compared to standard normalization (Figure 15B).

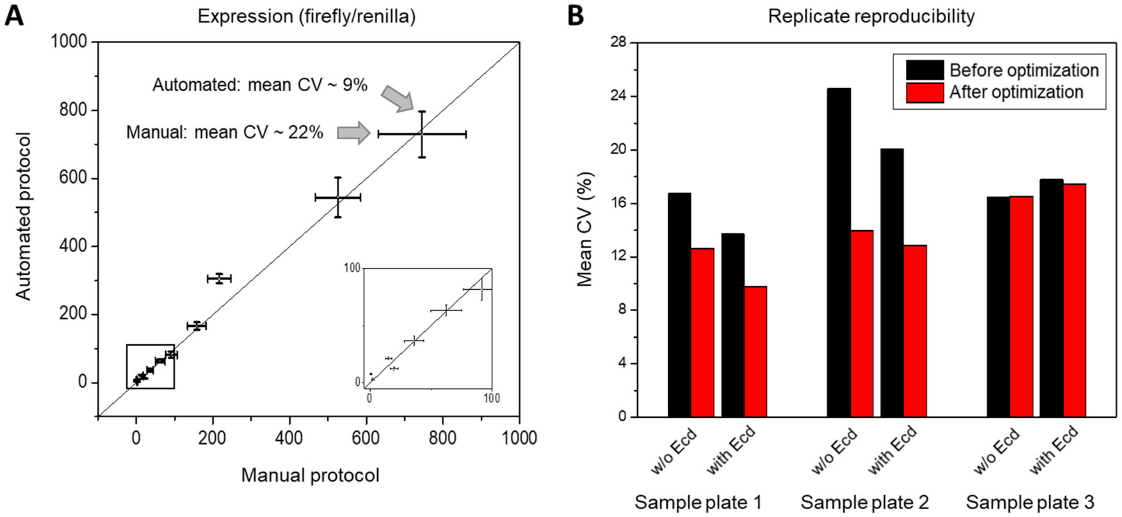


Figure 15. Evaluation of assay reproducibility. (A) Comparison of expression measurements (ratios of firefly and renilla signals) for manual versus automated protocol (11 samples measured in the same plate on the same day). The lower right insert shows a magnification of the region highlighted with the square (region of 0 - 100 RLU). The automated assay led to a much lower mean CV (~ 9%) for the 11 tested promoters here compared to the manual assay (CV ~ 22%). **(B)** Comparison of expression measurements for three 96-well plates containing the same promoter construct samples measured on different days with and without ecdysone induction. After optimizations of data normalization procedure of the luciferase assay readout (details in Section 7.2), the mean reproducibility improved substantially (mean CV ~ 13% after optimizations versus 18% before optimizations). Standard normalization (before optimizations) uses only the ratios of firefly and renilla signals.

Together, the automated workflow established for cell transfections and dual luciferase assays enables highly reproducible measurements of promoter activity. This allowed an accurate and quantitative test of the thousands of designed promoters in a high-throughput manner: for all promoters tested in this work, the dynamic range extended over more than four orders of magnitude with a mean CV of 21% (details in Section 9.1).

9. Dissecting the *Drosophila* promoters based on sequence features

9.1 Synthetic promoter activity measurements cover a wide dynamic range

We designed 3826 synthesized oligonucleotides representing wild-type and perturbative core promoter sequences in *D. melanogaster*. These block 3-6 sequences were assembled with one block 2 and different combinations of block 1 and block 7 sequences, constructing the entire library of synthetic promoters to be tested in our experiments.

In total, more than 16000 colonies were picked and around 3000 distinct promoter constructs were recovered with correct sequences and over 40000 dual luciferase assay measurements (including pre-measurements for protocol development) were performed. After removal of outliers (approximately 5% of the raw data; the method described in Section 7.3), in total 20335 normalized data points were obtained (Figure 16A). For most of the constructs (> 88%), we measured at least three replicates for both with and without ecdysone stimulation. Among the replicates of each promoter construct, over 85% were measured with at least two different reporter plasmids produced from a separate miniprep. Overall, our measurements ranged over more than four orders of magnitude. Reproducibility among replicates was high, with a mean CV of 21% (Figure S1A; median SD of 0.29). The two replicates with the highest CVs for the individual samples after outlier filtering were highly similar (Figure S1B; PCC $r = 0.98$, $p < 2.2\times10^{-16}$). There was also no correlation between the expression level and the CVs (Figure S1A).

We first checked if, as expected, the selected native core promoters covered a wide range of activity. We measured the constructs containing all native promoter sequences (ATGs removed) with the pair of block 1.11 and block 7.11. The measured expression levels showed indeed a broad range that spanned over three orders of magnitude (Figure 16B). Two housekeeping core promoters FBgn0035754 and FBgn0027597 drove the highest expressions, while ribosomal class generally showed an intermediate activity. As expected, the Ar.0 core promoters with no known motif showed the lowest activity. We also checked the activity of the ATG-containing promoters. They exhibited much weaker expression levels (at least 4-fold differences; one-sided Wilcoxon signed rank test $p = 0.018$), showing that their original ATGs do interfere with the normal luciferase expression. However, we found a high correlation between the reporter gene expressions measured from ATG-removed core promoters with their wild-type ATG-containing sequences (PCC $r = 0.87$, $p = 6.2\times10^{-5}$; Figure S2).

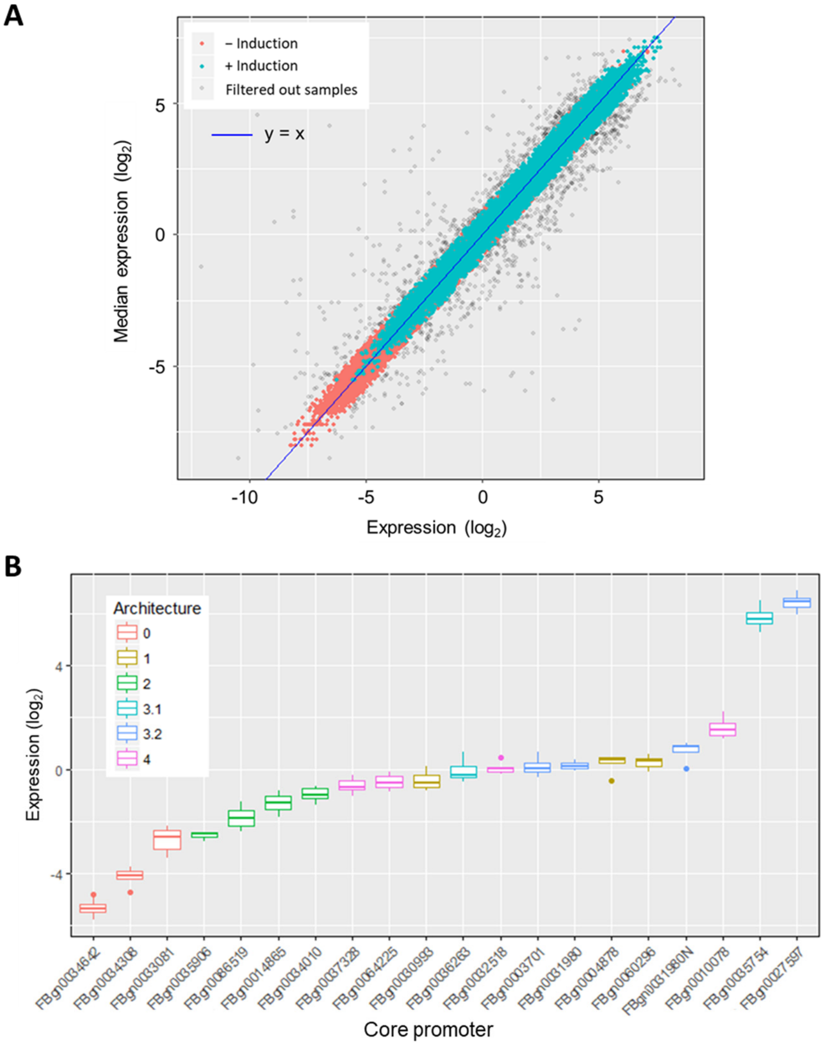


Figure 16. Expression measurements of synthetic promoter constructs ($\log_2$ scale). (A) Replicability of normalized expression levels for all tested promoters. The expression levels covered a wide dynamic range of more than four orders of magnitude. Expressions with and without ecdysone induction are labeled cyan and red, respectively. About 5% of the raw data were filtered out as outliers (gray dots). Blue line: y = x. **(B)** Normalized expression levels of the investigated native core promoters. Their activities also spanned a broad range (over three orders of magnitude; promoter constructs contained block 1.11 and block 7.11 combination as nucleosomal sequences). Each color represents a different class of the core promoter architectures.

In conclusion, these results show that our synthetic promoter activity measurements are highly reproducible and cover a wide dynamic range, with the native expressions spanning a range of more than three orders of magnitude.

9.2 Systematically mutational analysis of core promoter motifs

To systematically investigate the role of sequence motifs in core promoters we applied multiple types of mutations to the block 3-6 region, while keeping constant the pair of surrounding nucleosomal sequences (block 1.11 and block 7.11; details in Section 4.1). This section shows the effects on expression of mutagenized motifs, including (1) knockout of motifs; (2) pairwise knockout of motifs; (3) replacing motifs with their consensus sequences or insertions of motif consensus into Ar.0 core promoters; (4) point mutations of motifs; (5) substitutions with similar motifs from other architectures; and (6) shifts of motif positions.

9.2.1 Knockout of motifs mostly leads to loss of expression and the effects are consistent between different core promoters

To probe the potential influence of replacing sequences for motif knockout, we used three versions of these sequences: background sequences taken from an Ar.0 promoter (ko_bg) and two different random sequences with PWM scores lower than the threshold (ko_random and ko_INTRA; threshold score of each motif is listed in Table S4). The ko_INTRA sequences were designed for individual motif mutations used in intra-architectural combinatorial mutations. Similarly, the ko_bg and ko_random sequences were also applied for knockout of all motifs in a given core promoter. We checked whether different sequences for disrupting the motif have similar effects by comparing the expression levels of all tested constructs with either the background sequences (ko_bg) or the random sequences (ko_random/ko_INTRA) in the case of individual knockout and all-motif knockout (Figure 17A). Indeed, the effects were highly similar with the PCC *r* of 0.94 ($p < 2.2\times10^{-16}$). Therefore, all the knockout data of the same motifs in one promoter but with different replacing sequences were pooled together, and their arithmetic means were used in the analysis. Biased data such as knockout causing the creation of spurious binding sites were filtered out.

For individual knockouts, motifs that overlap with each other could cause bias in the data. In particular, motifs like CGpal and TTGTT often overlap with others due to their composition similarities. For example, CGpal overlaps with GAGArev in FBgn0060296, CGpal overlaps with TATA-Box in FBgn0004878, Ohler7 overlaps with CGpal and TTGTT in FBgn0031980, TTGTT overlaps with CA-INR in FBgn0035906, and TTGTT overlaps with R-INR for all ribosomal promoters (details in Figure 6). Since destroying non-overlapped CGpal and TTGTT nearly did not influence expression (CGpal in FBgn0030993 and FBgn0035906, TTGTT in FBgn0036263; effects are shown in Figure 17B and C), we ignored them when analyzing the other motif-overlapped promoter configurations.

To find out whether the motif knockouts have significant effects on expression, we first compared the expression levels of wild-type configuration with individual/pairwise/all-motif knockouts in two core promoter architectures (developmental and constitutive) with different sets of motifs (Figure 17D and E). As expected, the disruption of well-known motifs like INR and TATA-Box in FBgn0034010 drove substantial activity reductions as well as the Ohler6 in FBgn0064225. However, the initiator for the ribosomal protein genes, surprisingly, showed no significant effect when mutated (in FBgn0064225 as well as in all the other tested ribosomal core promoters; Figure 18B). A similar absence of effect was observed for RDPE motif, while knockout of both these two motifs caused a decrease in expression in FBgn0064225 (~ 2.4-fold reduction). The disruption of all motifs in both promoters led to much weaker expressions (> 30-fold decrease), as expected.

More generally, the complete knockout of motifs for all tested core promoter sequences resulted in a nearly entire loss of function, regardless of the wild-type strengths (Figure 18A). Most of these all-motif disrupted promoters exhibited even lower activity than Ar.0 core promoters containing no known motif (the three promoters with the lowest native expressions in Figure 18A). Compared to native expressions, knocking out individual motifs typically resulted in a reduction, with some motifs, like INR, MTEDPE, CA-INR, TATA-Box, INR2 and DRE, showing strong effects, while others had a weaker or no impact (Figure 18B). Remarkably, these effects were relatively consistent across the different promoters.

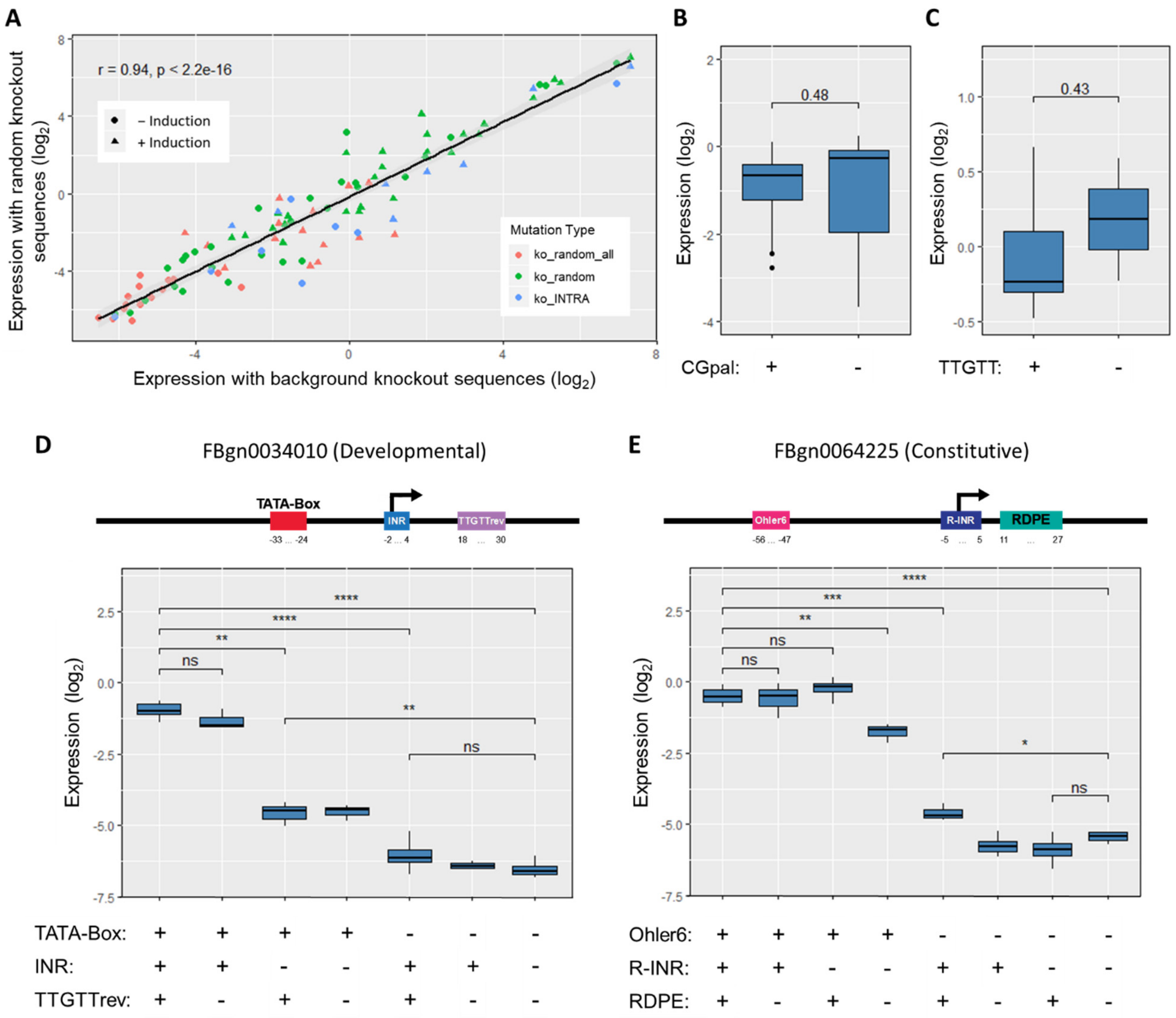


Figure 17. The effect of motif knockout (log$_2$ scale). (A) Comparison of the expression measurements for promoter constructs with motifs disrupted by ko_bg sequences versus ko_random/ko_INTRA sequences. “ko_random_all” represents the measurements for all-motif knockouts. Expressions with and without ecdysone induction are labeled as the triangle and the round, respectively. Black line: linear regression (with 95% confidence interval shown in gray, PCC $r = 0.94$, $p < 2.2\times10^{-16}$). **(B)** Boxplot depicting the effect of CGpal knockout (non-overlapped) in FBgn0030993 and FBgn0035906. There is no significant difference between the expressions with and without a CGpal (all measurements in these two core promoter constructs were pooled together; Wilcoxon rank-sum test p = 0.48). **(C)** Boxplot depicting the effect of TTGTT knockout (non-overlapped) in FBgn000036263. No significant difference was obtained between the expressions with and without a TTGTT (two-sample t-test p = 0.43). **(D-E)** Comparison of normalized expression levels between wild-type configuration and motif knockouts for two types of core promoters (developmental: FBgn0034010; constitutive: FBgn0064225). Upper panel: the schematic depiction of the wild-type motif compositions (TTGTT motif in FBgn0064225 is ignored due to its strong overlap with R-INR). Two-sample t-test: ns, not significant, $p > 0.05$; $*p \leq 0.05$; $**p \leq 0.01$; $***p \leq 0.001$; $****p \leq 0.0001$.

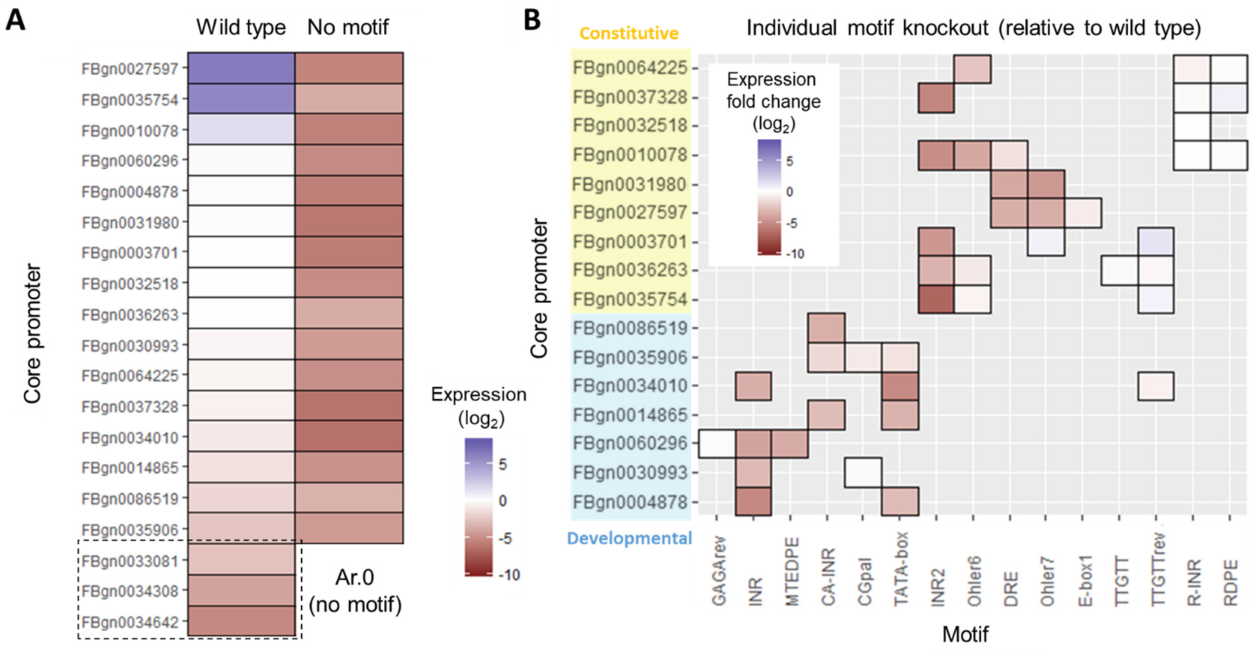


Figure 18. The effect of individual knockout and all-motif knockout ($\log_2$ scale) in all tested core promoters. (A) Heatmap depicting the mean expression levels of promoter constructs with wild-type core promoters and all-motif knockouts, ranked by native expression levels. Ar.0 core promoters without known motifs are marked with a dashed box on the bottom. **(B)** Mean expression fold changes compared to wild-type expressions for individual knockout of motifs in different core promoters. Constitutive and developmental promoters are highlighted in yellow and blue, respectively.

In the case of promoters containing the Ohler6 motif, we saw two distinct effects: in Ar.3.1 core promoters (FBgn0035754 and FBgn0036263), its disruption had no effect on expression (Wilcoxon rank-sum test $p > 0.9$; Figure S3A); in Ar.4 promoters (FBgn0010078 and FBgn0064225), destroying Ohler6 strongly reduced expression (Wilcoxon rank-sum test $p = 3\times10^{-4}$; Figure S3A). This different behavior could be explained by the presence of a second intact Ohler6 motif sequence (mutation not recovered in the experiments) in both FBgn0035754 and FBgn0036263 which were sufficient to lead to relatively strong expression.

The motifs in FBgn0035906 generally showed milder effects probably due to its already low native expression level (the weakest promoter with known motifs in our experiment; Figure 18A). Ohler7 and DRE played less critical roles if the core promoter contains Ohler6 or INR2 (usually in mixed architectures, e.g., DRE in FBgn0010078 and Ohler7 in FBgn0003701). We also found that, although the disruption of TTGTT almost had no influence on the expression level, knocking out its reverse complement TTGTTrev could slightly increase expression in

FBgn0003701 and FBgn0035754 (Figure 18B; > 2-fold increase after disruption of the motif). Hence this motif functioned as a weak repressor. Intriguingly, the core of TTGTTrev (AACAA) is also the central part of the binding site of an adult enhancer factor (AEF-1) in *Drosophila*, which is known to be a short-range transcriptional repressor (Brodu, Mugat, Fichelson, Lepesant, & Antoniewski, 2001; Falb & Maniatis, 1992a, 1992b). Finally, the ribosomal promoter motifs R-INR and RDPE did not lead to a reduction of activity after the disruption in all the four investigated Ar.4 promoters. (Wilcoxon rank-sum test $p > 0.7$ for R-INR and $p = 0.3$ for RDPE; Figure 18B, S3B and C). R-INR sometimes overlaps with a GAGArev motif which also had no impact in our measurements (two-sample t-test $p > 0.1$; Figure S3D).

In conclusion, these results demonstrate that the disruption of all motifs in a given core promoter generally leads to a nearly complete loss of expression. Most individual motif knockouts result in a reduction of expression level compared to the wild-type promoter. The effects after disruption of a specific motif are consistent between different core promoters. However, some motif knockouts like R-INR and RDPE in the ribosomal promoters had nearly no effect on expression.

9.2.2 Pairwise knockout of motifs often show superadditive (synergistic) effects

To further investigate the role of motif interplay on regulating the expression, we compared the results obtained from pairwise knockouts with their individual knockout measurements in different core promoter configurations. The effect of most pairwise knockouts was greater than either of the corresponding individual disruptions. Furthermore, pairwise disruption of motifs often led to superadditive effects (the pairwise effect usually greater than the sum of each individual effect in logarithmic scale; Figure 19), which indicated the synergism between them. Nevertheless, core motifs in developmental promoters such as INR and MTEDPE in Ar.1 promoter FBgn0060296 (Figure 19A), as well as CA-INR and TATA-Box in Ar.2 promoter FBgn0035906 (Figure 19B) were so crucial for promoter activity that knockout of each would result in almost the same effect of disrupting them both (subadditivity).

For promoters FBgn0036263 and FBgn0035754, the pairwise effects showed largely linear additivity (Figure 19C and D). Consistent with the results obtained with individual motif knockout, the Ohler6 hardly influenced the expressions for Ar.3.1 promoters (Figure 18B, S3A, 19C and D);

this is probably due to the presence of the second Ohler6 in the construct which plays a compensating role (the construct with knockout of both Ohler6 motifs was unfortunately not recovered).

The motif pair DRE + Ohler7 or DRE + E-Box1 in Ar. 3.2 promoter FBgn0027597 showed strong synergistic interactions (Figure 19E). Pairwise knockout led to a more substantial loss of expression than the sum of individual effects as well as the paired Ohler7 + E-Box1 effect (superadditive effects of -2.4 and -2.92 compared to -1.39 for the paired knockout Ohler7 + E-Box1). DRE is considered as the most crucial motif in this housekeeping core promoter architecture as it directs a specific TF DREF binding (Hirose F et al., 1993). The strong superadditivity we observed suggests the existence of a compensatory phenomenon for DREF binding involving Ohler7 and/or E-Box1 against potential mutations of the DRE motif. Ohler7 could fully recover the activity when E-Box1 was disrupted, but not vice versa.

For ribosomal core promoters like FBgn0064225, the pairwise knockout of R-INR and RDPE had only a weak negative influence on expression level (a ~ 0.4-fold reduction; Figure 19F), similar to the knockout of the individual motifs. However, other motifs not belonging to our defined ribosomal class, like Ohler6 or INR2 usually found in housekeeping gene core promoters, altered ribosomal promoter activity strongly (Figure 19F and S4).

Taken together, these results demonstrate that the disruption of motif pairs in a given core promoter often result in synergistic effects, except for core motifs in developmental promoters. The effects of Ohler6 removal in different promoter configurations suggest its flexible location requirement. DRE is crucial for housekeeping promoter function and the other two housekeeping motifs Ohler6 along with INR2 also play essential roles in regulating ribosomal gene transcription.

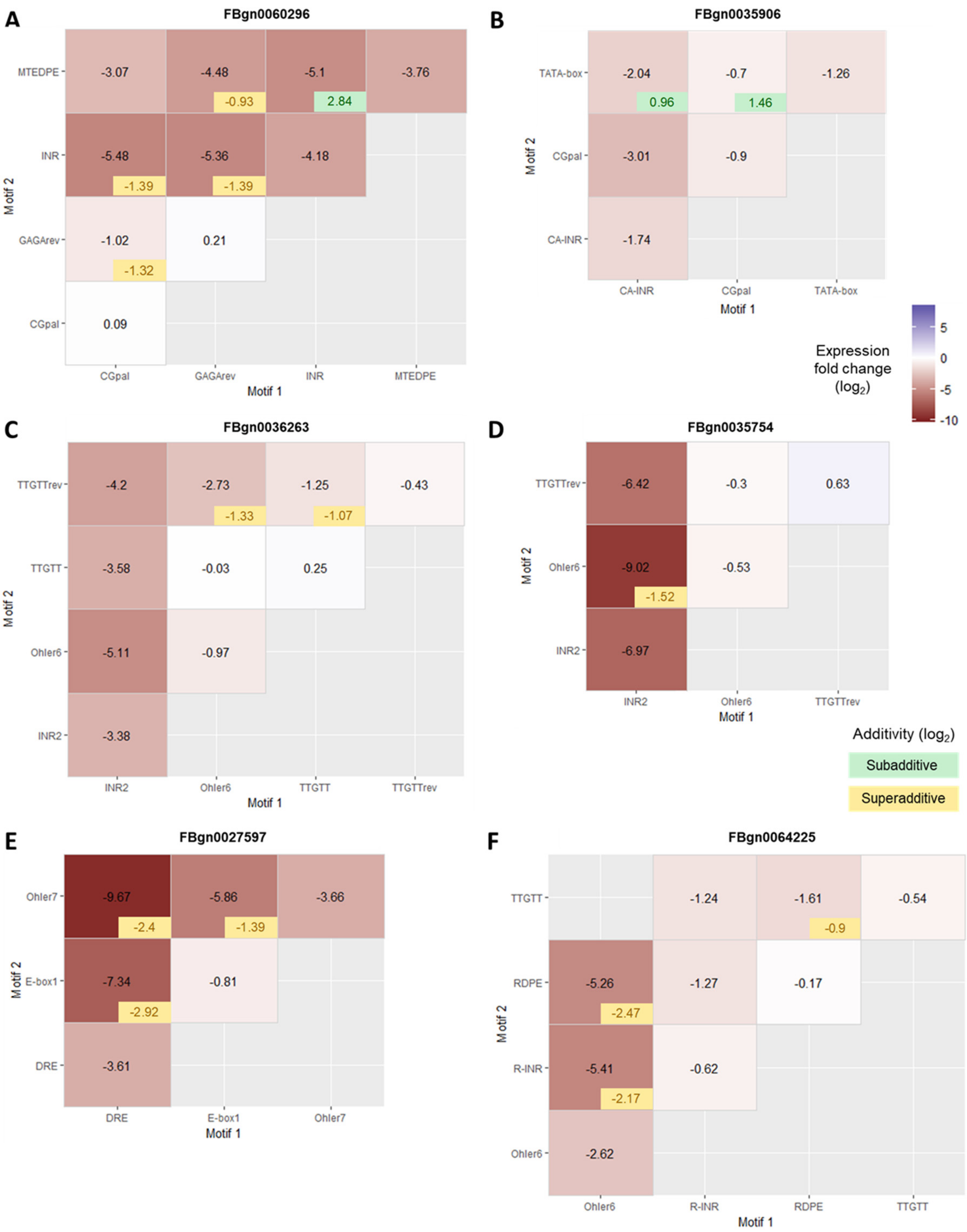


Figure 19. The effect of pairwise motif knockout (log$_2$ scale) in different core promoters. (A-B) Heatmaps of the mean expression fold changes compared to wild-type expressions for pairwise knockout of motifs compared to individual knockouts (diagonals) in developmental core promoters FBgn0060296 (Ar.1; A) and FBgn0035906 (Ar.2; B). **(C-F)** The same for constitutive core promoters FBgn0036263 (Ar.3.1; C), FBgn0035754 (Ar.3.1; D), FBgn0027597 (Ar.3.2; E), and FBgn0064225 (Ar.4; F). Additivity was calculated as the difference between the pairwise effect and the sum of two individual effects (subadditive (green): > 0; superadditive (yellow): < 0; effects > 3×SD shown in the corner of each pairwise effect).

9.2.3 Motif consensus sequences can drive higher expression

In addition to the complete shutdown of the motif function by knockout, we next tested if computationally derived consensus sequences that are preferred in the genome could act positively to increase expression.

Indeed, most of the consensus sequences could drive higher promoter activity, especially the consensus of TATA-Box in FBgn0035906 (more than 15-fold stronger expression; Figure 20A). As an exception, replacing the TTGTTrev motifs with their consensus sequences in three promoters led to a signal reduction, again suggesting its repressive role in the promoter.

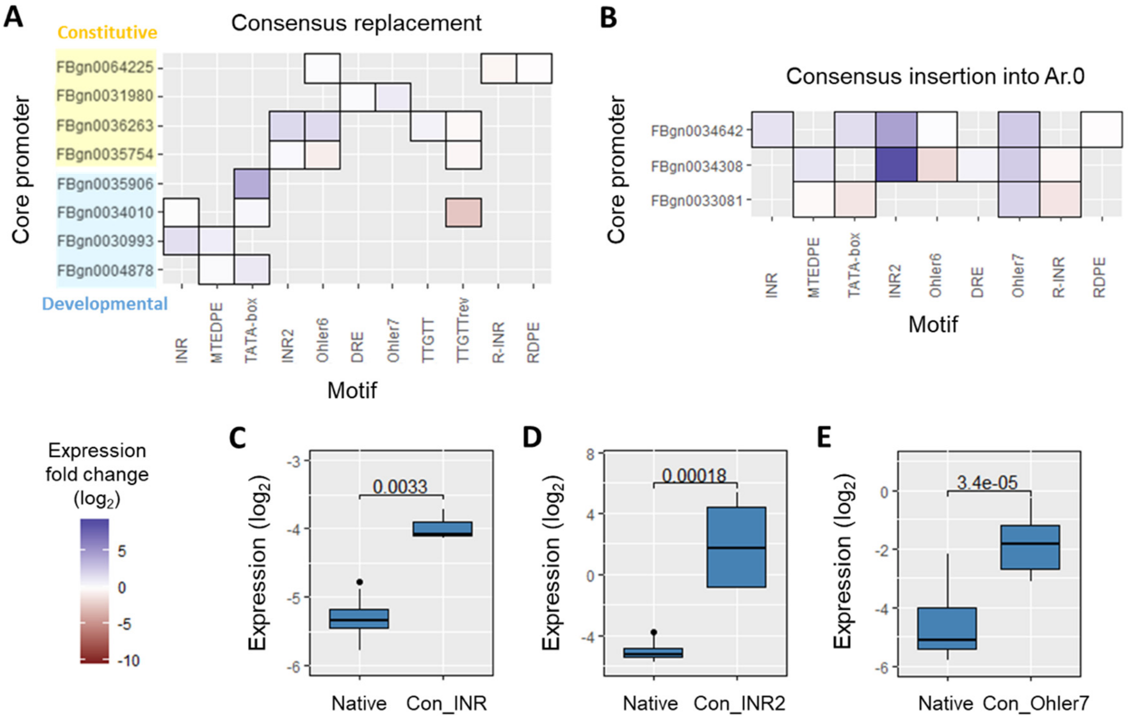


Figure 20. The effect of motif consensus sequences ($\log_2$ scale) in different core promoters. (A) Heatmap depicting the mean expression fold changes compared to wild-type expressions after replacing with motif consensus sequences derived by XXmotif. Constitutive and developmental promoters are highlighted in yellow and blue, respectively. **(B)** The same for mean expression fold changes of consensus insertion into Ar.0 core promoters. **(C-E)** Boxplots depicting the significant effects of INR consensus insertion in FBgn0034642 (two-sample t-test $p = 0.0033$), INR2 consensus insertion in FBgn0034308 and FBgn0034642 (all measurements in these two core promoter constructs were pooled together; Wilcoxon rank-sum test $p = 0.00018$), and Ohler7 consensus insertion in FBgn0033081, FBgn0034308 and FBgn0034642 (all measurements in these three core promoter constructs were pooled together; Wilcoxon rank-sum test $p = 3.4\times10^{-5}$).

Since most replacements of the motifs with their corresponding consensus sequences increased expression level, we questioned whether they were also able to boost the activity of our motif-less promoters (the Ar.0 core promoters FBgn0033081, FBgn0034308 and FBgn0034642) after being added into their architectures (Figure 20B). Indeed, some motifs, particularly INR and INR-like motifs including INR2 and Ohler7, were sufficient to significantly induce expression when inserted in these Ar.0 promoters (Figure 20B-E, > 2-fold increase for INR replacement, ~ 100-fold increase for INR2 and ~ 5-fold increase for Ohler7 on average). In contrast, the other motifs actually did not strengthen or even weaken the expression (Figure 20B), maybe due to the disruption of sequences bound by unknown proteins.

Overall, our results demonstrate positive effects on expression of most computationally-derived motif consensus sequences (except the repressive TTGTTrev). In particular, INR and INR-like motifs (INR2 and Ohler7) can boost Ar.0 promoter expressions. This demonstrates that each of these motifs is sufficient to increase the expression level.

9.2.4 Systematic point mutations enable generation of the expression-based PPMs and activity logos for core promoter motifs

We then quantitated systematically the influence on expression of changing motif binding specificity. We generated all possible single base pair mutations of the motif consensus in a given native promoter configuration (details in Section 4.1.1.4). We recovered nearly all of the variants for motifs including INR, TATA-Box, INR2, DRE and Ohler7. Most of the consensus sequences gave the highest expressions, while other constructs with point-mutated motifs showed a wide range of activities (Figure 21A). Based on these expression measurements, we generated PPMs (Table S9) and thereby activity logos for these motifs, which we used to compare with their XXmotif sequence-based logo (Table 5). Overall, the consensus for each was identical to the one that was computationally derived. Notably, we could capture the core of INR, TATA-Box and DRE, especially the A for the TSS position in the INR motif and the TATA core of the TATA-Box. Interestingly, we observed that all the expression-based activity logos were less specific (as indicated by their lower information content in Table 5) compared to those *in silico* XXmotif defined ones (especially for INR2 and Ohler7). An exception was the CG dinucleotide in the DRE that showed higher information content than the XXmotif generated one, probably suggesting its function as the primary recognition site for DREF binding. This observation is also a hint that the lower information content generally observed is not an artifact of our method.

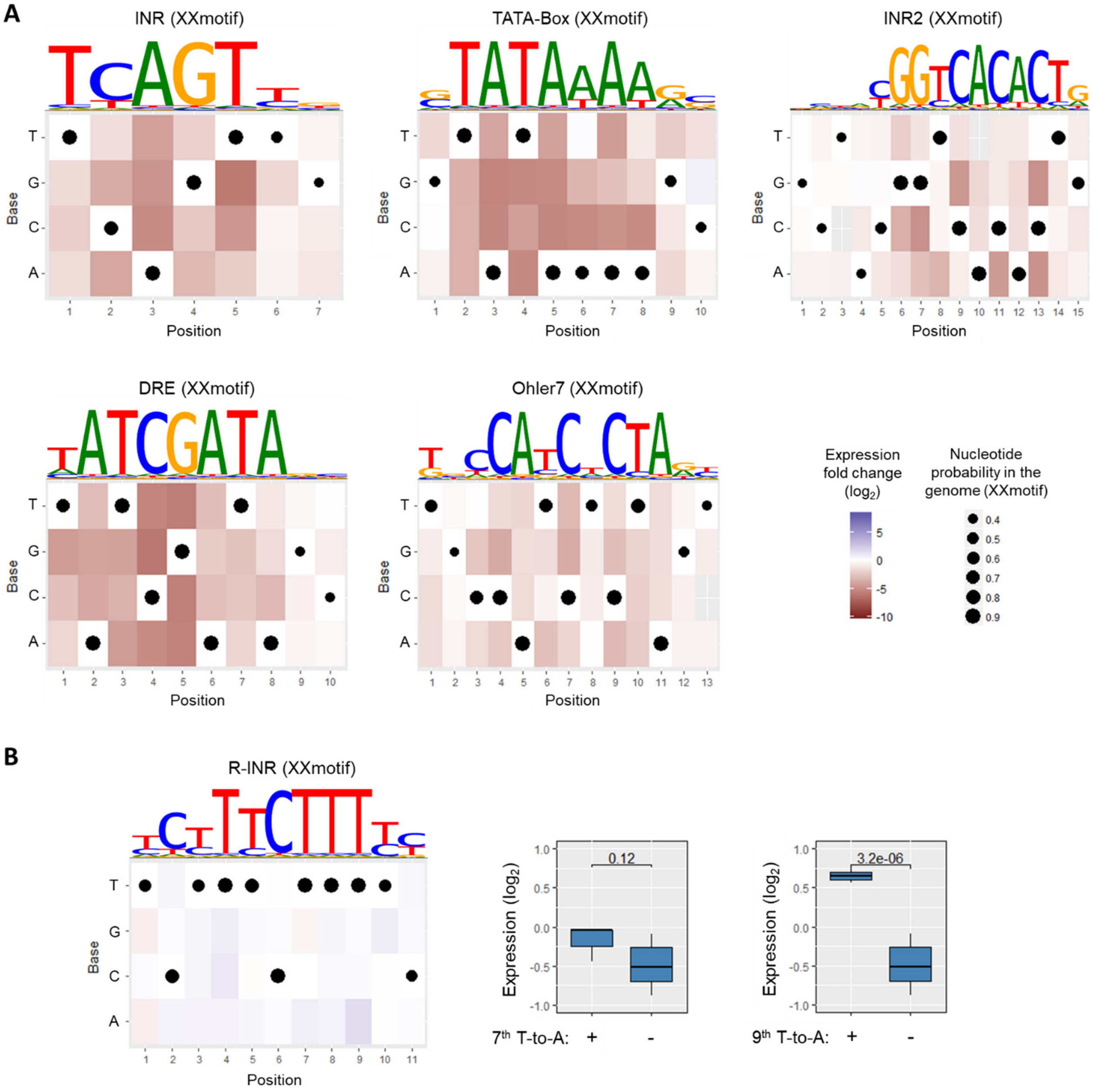


Figure 21. The effect of single base pair mutations of the XXmotif-derived motif consensus (log$_2$ scale). (A) Each column of a heatmap shows the expression fold changes of the point mutations at a specific position in the motif. The XXmotif consensus sequence is highlighted as white blocks with spots which size with the nucleotide probability in the genome defined by XXmotif. **(B)** The effect of point mutations in R-INR. Right panel: boxplots depicting the effects of 7th T-to-A (two-sample t-test p = 0.12, not significant) and 9th T-to-A (two-sample t-test $p = 3.2\times10^{-6}$), respectively.

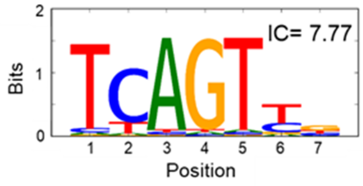
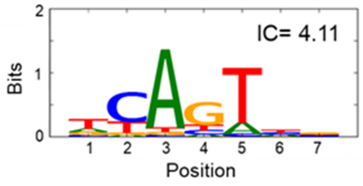
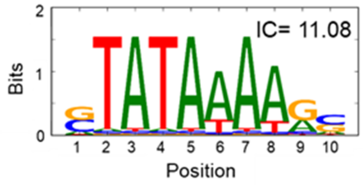
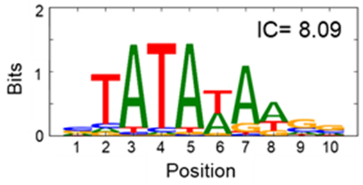
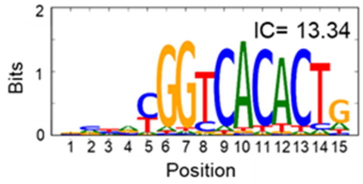
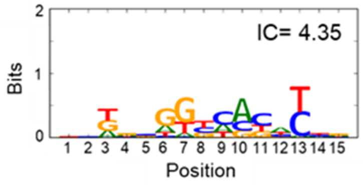
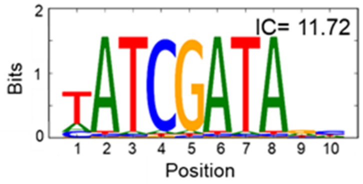
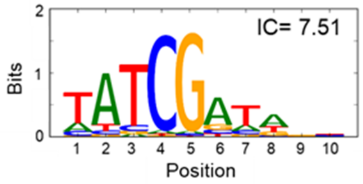
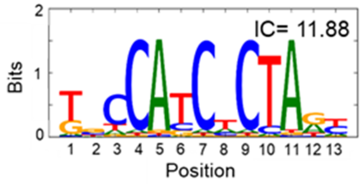
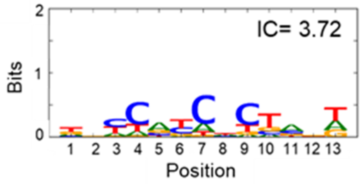
Core promoter motif	XXmotif logos	Expression-based activity logos
INR		
TATA-Box		
INR2		
DRE		
Ohler7		

Table 5. Comparison of the XXmotif logos with the expression-based activity logos for INR, TATA-Box, INR2, DRE and Ohler7. Our logos show an overall lower specificity. IC, information content.

Strikingly, the expression levels for MTEDPE, R-INR and RDPE were nearly not altered after single site variations (although we recovered for these motifs most mutations), as would be expected from their XXmotif specificity logos (Figure 21B and S5). Notably, the various tested mutations like knockout or consensus replacement of R-INR and RDPE in different promoter backgrounds (Figure 18B and Figure 20A) as well as consensus substitution for MTEDPE in FBgn0004878 and FBgn0030993 (PWM scores: 23.55 for consensus vs. 5.66 and 4.8 for native MTEDPEs in each construct; Figure 20A) hardly changed the expression levels. Considering these results, a strong influence of the single base mutations of these three motifs was therefore not expected. This is apparently contradicting with the high PMW scores computed with the XXmotif activity logos. A previous study suggested a single T-to-A substitution at +1 nucleotide relative to TSS in R-INR (which is the 7th nucleotide of R-INR) could convert it into an active INR motif (Parry et al., 2010; the motif was named as TCT there). We indeed saw a slight expression increase for that specific point mutation (Figure 21B; two-sample t-test $p > 0.1$; the PWM score of the putatively created INR was smaller than our threshold for identifying the presence of the motif). Intriguingly, we found that another substitution of 9th T-to-A could construct a functional Ohler7 in this promoter configuration FBgn0064225 (PWM score: 8.6; threshold: 7.3; this 9th nucleotide A in the newly generated Ohler7 was supposed to act as the TSS) which drove a significantly higher expression (Figure 21B; > 2-fold increase on average, two-sample t-test $p = 3.2\times10^{-6}$), which indicated the housekeeping motif Ohler7, like INR2 and Ohler6 discussed in Section 9.2.2, could also function in regulating ribosomal protein gene transcription. Besides, the point mutation effects of Ohler6 were minor again probably due to the compensation of the existed second Ohler6 (Figure S5). The activity logos generated for these above-mentioned motifs including MTEDPE, Ohler6, R-INR, RDPE as well as the motif TTGTT could not capture the XXmotif defined sequence features (Table S10). Motif point mutations checked in the "null" promoter environment could not provide comparable results due to the insufficient number of single mutations recovered and to the higher variability of the data at low expression levels.

Collectively, our highly sensitive measurements of systematic single base pair mutations make it feasible to create the expression-based PPMs and the activity logos for core promoter motifs. Although the computationally identified overrepresented sequence generally represents the best motif, the *in vivo* specificity of the motif as well as each nucleotide in the sequence is not precisely conveyed. Our results suggest the computationally derived activity logos do not accurately capture the *in vivo* strength of the motif.

9.2.5 The positionally or functionally equivalent core promoter motifs from other architectures can hardly function as endogenous sequences

While checking the features of XXmotif-discovered core promoter motifs, we found that certain motifs tend to locate within a similar region relative to TSS (like DRE and Ohler6 at around -100 to -7) or they share similar sequence features such as the “CA”s in INR, INR2 and Ohler7 (Table S1). Therefore, we set to investigate whether positionally or functionally equivalent motifs from other architectures could act similarly as the original motifs and rescue the expressions from knockouts in a given promoter. Three combinations were tested here: INR - INR2 - Ohler7 - R-INR; TATA-Box - Ohler6 - DRE; MTEDPE - RDPE.

For most of the motifs, we saw that substitution could not recover the promoter activity, that is, substitution would lead to same or only slightly higher expression than if the motif was utterly destroyed (Figure 22).

An exception was the INR2, which could almost replace INR in our experiments, thereby showing a rescue effect (Figure 22A and B; Wilcoxon rank-sum test $p = 0.17$ between the native expression and the INR2-substituted expression). Conversely, INR was not able to take the role of INR2 over. Nevertheless, they both increased expression level compared to the native arrangement when substituting R-INR (except INR2 substitution in FBgn0037328, probably due to two INR2 existed but without Ohler6 motif). Other INR-related motifs Ohler7 or R-INR could not function in most of the cases, with a slightly upregulating effect for the Ohler7 substitution in ribosomal promoters.

DRE substitution moderately increased expression of one promoter FBgn0035906 with original TATA-Box construct but had no effect in the case of FBgn0034010. Besides, DRE restored some level of activity when disrupting Ohler6 (the effects were however minor for core promoters containing a second Ohler6: FBgn0035754 and FBgn0036263). Ohler6 rarely had the ability to substitute either TATA-box or DRE. Interestingly, FBgn0010078 was an exception: replacing DRE with Ohler6 resulted in a promoter containing two Ohler6, which presumably constituted a similar configuration of Ohler6 motif pair + INR2 with higher activity, as was the case in the native promoters FBgn0035754 and FBgn0036263.

RDPE could recover mostly of MTEDPE knockout effects in Ar.1 core promoters, most probably because they share partially similar patterns in the sequence. However, this activity restoration still deviated from the MTEDPE wild-type strength (Figure 22A and C; Wilcoxon rank-sum test $p = 0.00058$ between the MTEDPE-disrupted expression and the RDPE-substituted expression; $p = 0.00025$ between the native expression and the RDPE-substituted expression). RDPE disruption, as shown before (Figure 17E and 18B), had almost no influence on ribosomal promoter activity. Consistently, adding MTEDPE had a weak influence.

Taken together, our results show that among all core promoter motifs we tested, only INR2 can largely substitute the function of INR and also surpass R-INR in most of the tested cases. Although some of the motifs locate at similar positions relative to the TSS or share similar sequence features, the original motif is mostly irreplaceable.

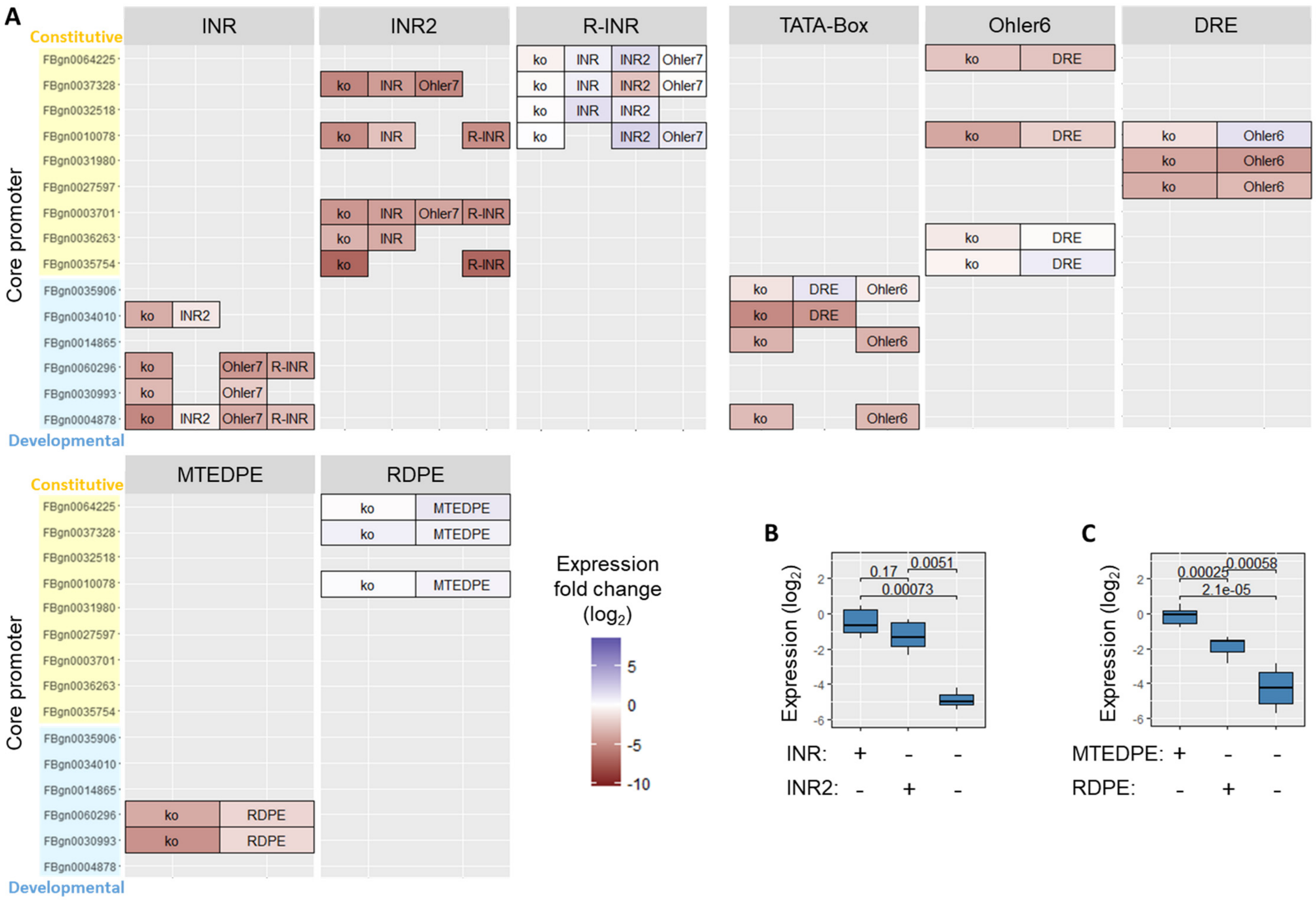


Figure 22. The effect of motif substitutions (log$_2$ scale). (A) Heatmap depicting the mean expression fold changes compared to wild-type expressions for motif knockout and substitution with positionally or functionally equivalent motifs from other architectures. Constitutive and developmental promoters are highlighted in yellow and blue, respectively. **(B)** Boxplot depicting the effects of INR being substituted by INR2 in FBgn0004878 and FBgn0034010 (all measurements in these two core promoter constructs were pooled together; Wilcoxon rank-sum test p = 0.0051 for comparing substitution with knockout (significant) and p = 0.17 for comparing substitution with wild-type (not significant)). **(C)** Boxplot depicting the effects of MTEDPE being substituted by RDPE in FBgn0030993 and FBgn0060296 (all measurements in these two core promoter constructs were pooled together; Wilcoxon rank-sum test p = 0.00058 for comparing substitution with knockout and p = 0.00025 for comparing substitution with wild-type, both effects are significant).

9.2.6 Precise positioning of motifs is an essential feature of core promoter function

The XXmotif analysis has provided evidence for strong positional preferences of some motifs (shown in the columns "Distribution" and "Start Range" in Table S1). To test if these preferences are functionally relevant, we shifted the motifs around their native positions and checked the effects on expressions.

Overall, varying motif positions decreased expression level, regardless of the shift direction. Additionally, our results showed a strong correlation between the promoter activity and the motif positional preference, especially for the region within 10 bp around the original motif location. In the case of strongly positioned motifs (e.g., INR, MTEDPE and TATA-box), even small shifts (< 5 bp) of the motif within an unaltered promoter led to a severe loss of expression, while less well-positioned motifs showed milder effects when shifted (Figure 23A). However, for larger-distance shifts towards TSS, perturbations of other motifs usually could be seen and resulted in substantial reductions of expression. For example, here in promoter FBgn0031980, 20 bp downstream shift of DRE had an influence on the flanks of Ohler7 and ≥ 10 bp downstream shift of Ohler7 could affect TSS position. Note that previous studies found an apparent ~ 10 bp periodicity of expression changes for other non-strongly positioned motifs (Sharon et al., 2012; Weingarten-Gabbay and Segal, 2014). We also did not see any similar pattern for DRE and Ohler7 when shifting them upstream (further away from TSS). This was probably due to their weak positional preferences, or this different native configuration does not require specific spacing. Interestingly, the effects of motif shift in our expression measurements showed similar shapes as the genomic motif distribution within ±20 bp region of the most enriched motif locations (Figure 23B).

In conclusion, the motif position is essential for core promoter function since shifting it usually leads to a decrease of expression. The positional preference of each motif is functionally relevant, especially for the 10 bp region surrounding the original position. In addition, our data are consistent with the motif genomic distributions.

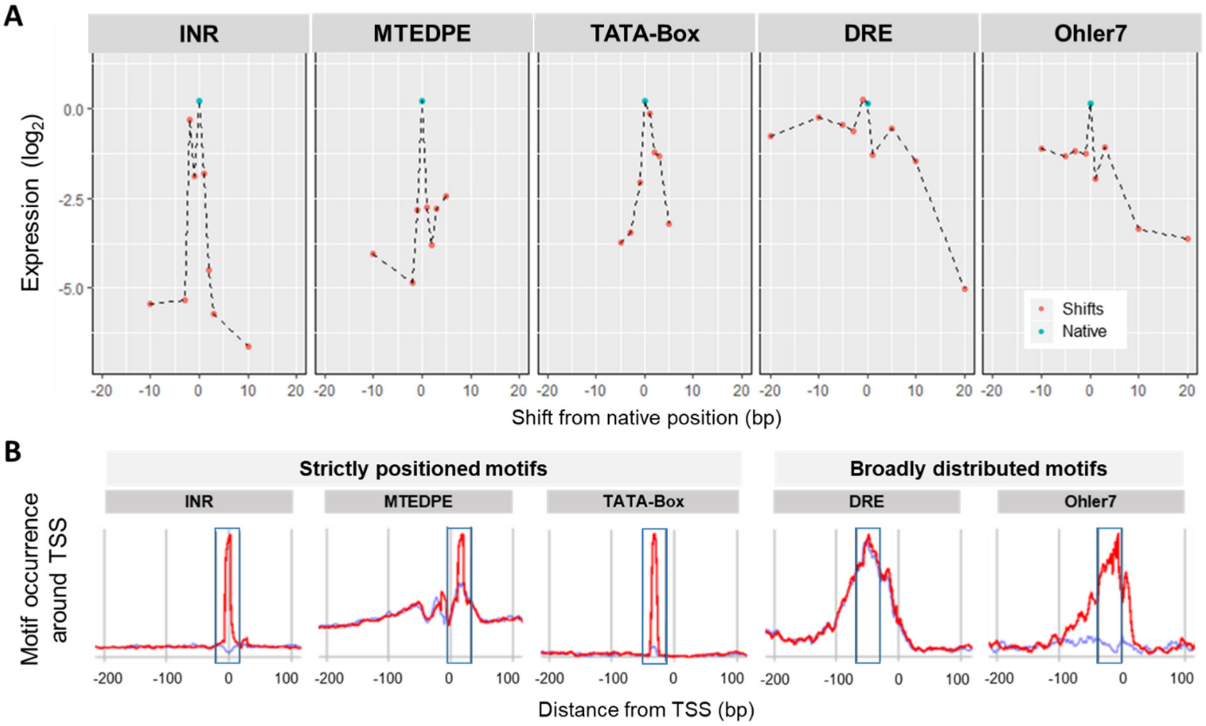


Figure 23. The effect of motif shifts (log$_2$ scale). (A) The expression measurements of native promoters (cyan dots) and positional shifts around the original locations (red dots) of INR, MTEDPE, TATA-Box in FBgn0004878, and DRE, Ohler7 in FBgn0031980. **(B)** The motif occurrence around TSS discovered in the genome-wide analysis by XXmotif. The blue rectangular boxes indicate the -20 to 20 bp region surrounding the original positions of the motifs in the tested core promoters (strictly positioned INR, MTEDPE, TATA-Box in FBgn0004878; broadly distributed DRE, Ohler7 in FBgn0031980). Negative distances correspond to positions upstream of the TSS.

9.3 A linear combination of individual motif features can largely explain the core promoter activity

Our results obtained from the pairwise knockout of motifs revealed subadditive or superadditive effects of individual motif features (in log$_2$ scale; Figure 19). This prompted us to investigate how much of the expression level can be explained by the pure additive contributions of each motif features. Therefore, we applied a linear regression analysis to the promoters with intra-architectural combinatorial mutations. We assigned the variables in the model as the qualitative indicators (0/1) of the individual mutation existence. For all promoters tested, as can be seen in Figure 24, we computed an average correlation of ~ 88% between predicted expressions and experimental expression measurements (average PCC r = 0.88). The data obtained for promoters FBgn0036263 and FBgn0035754 showed the highest correlations (PCC r = 0.91 and 0.94, respectively); this strong linearity was also observed in their pair-wise knockout measurements (Figure 19C and D). The coefficients learned by the models (Table S11) also correlated quite well with expression levels of single mutation samples in the experiments (average correlation PCC *r* = 0.93; Figure S6).

As a more direct test without any fitting procedure, in addition, we also built an additive model (Figure S7) to predict the activity of a given promoter with the intra-architectural combinatorial mutations (based on the measurements of individual motif mutations). The contribution of each feature (both motif strength and position) was assumed to function additively and was derived from the deviation between the expression value of each corresponding motif-mutated sample compared to the native expression. Except for one promoter (FBgn0004878, for which multiple single mutation constructs were not recovered during the cloning procedure), we obtained a comparable mean correlation of 84% (Figure S7).

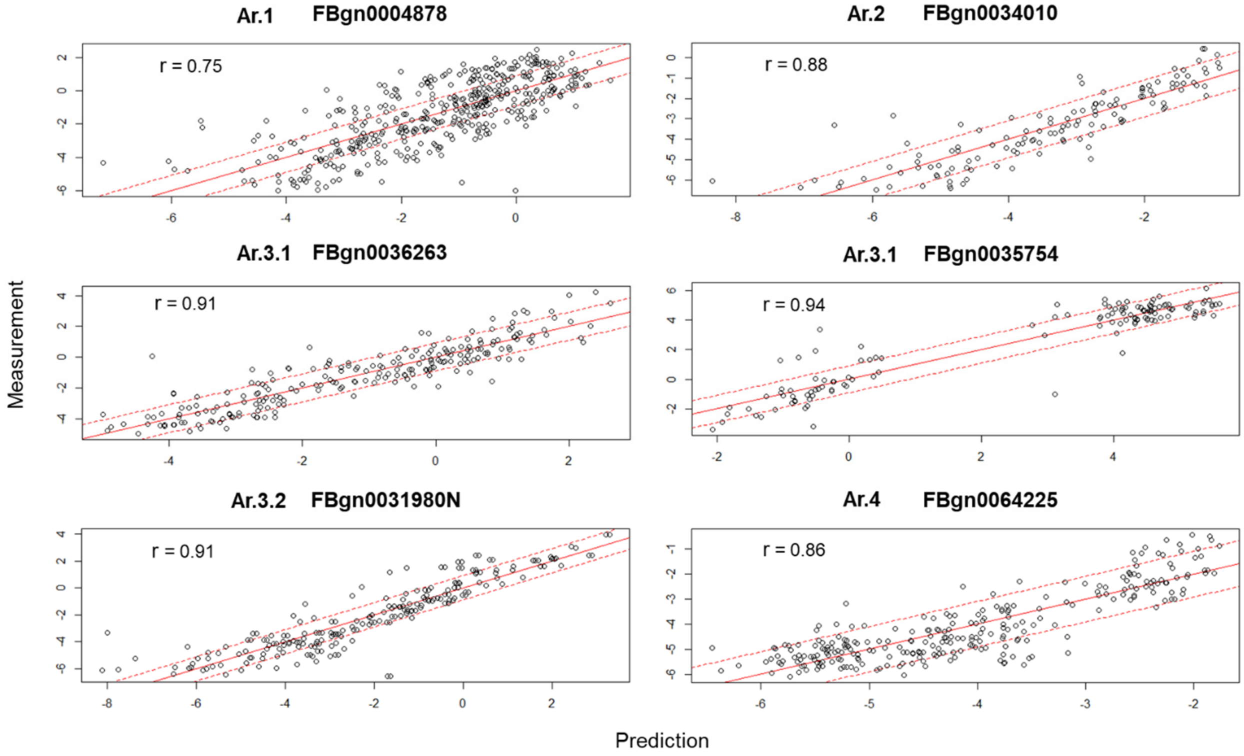


Figure 24. Linear regression applied to predict the synthetic promoter activity based on individual motif features. The measured expressions (on the y-axis) for 6 tested core promoter sequences with combinatorial motif mutations compared to the predicted expressions (on the x-axis) from the linear regression **(**$\log_2$ scale). Red solid line: y = x; red dashed lines: y = x ± 3×SD (SD denoted the experimental noise, that is the median of all SDs for all measured synthetic promoter constructs; also used in outlier filtering procedure described in Section 7.3). The linear regression model can explain on average 77% of the variance in expression (average r^2 = 0.77).

In conclusion, our results suggest the activity of a given synthetic core promoter is largely predicted from the linear combination of individual motif features. The linear regression model can explain on average nearly 77% (average $r^2 = 0.77$) of all the variance in expression. However, deviations are still observed, revealing the complex interplay between the factors involved; those lead to subadditive and superadditive effects.

9.4 Motif context in core promoters also shows influence on expression

In addition to mutations applied to sequence motifs, we also tested the influence on the expression level of the motif context, that is the sequence environment surrounding the motifs in the core promoter region.

We first created promoter variants where either all motifs or motif contexts were shifted together, thus maintaining the relative spacing of motifs, but altering the sequence background in which they located. In general, both cases led to the loss of expression; the amplitude of the effects was comparable or lower relative to the ones obtained from individual motif shifts (Figure S8).

Besides the mutations applied within the native core promoter architecture, we also checked the exchange of motif contexts between different architectures. We saw that overall the motifs preferred their native contexts (Figure 25A). The motifs from FBgn0064225 (Ar.4) resulted in an average more than 10-fold reduction of the expression levels when added in all the other promoter contexts. When inserting motifs from the architectures other than Ar.4 into motif-less core promoters (Ar.0 FBgn0034308 and FBgn0034642 tested here), they could drastically improve the expression with a maximum increase of more than 55 folds (Figure 25A). When comparing the obtained results with the wild-type expressions of the motif-origin promoters (Figure 25B), the context from FBgn0034642 (Ar.0) could recover or even increase the expression of developmental promoters with their native motifs (~ 25% expression increase for FBgn0004878 and > 2-fold increase for FBgn0034010). Similarly, the context from another no-motif core promoter FBgn0034308 (Ar.0) could constitute a better promoter compared to the native FBgn0036263 (a constitutive promoter; with a ~ 2.5-fold increase; Figure 25B). Although we checked if the various context effects could be explained by the type of TSS distribution (NP or BP), we did not see a clear relationship.

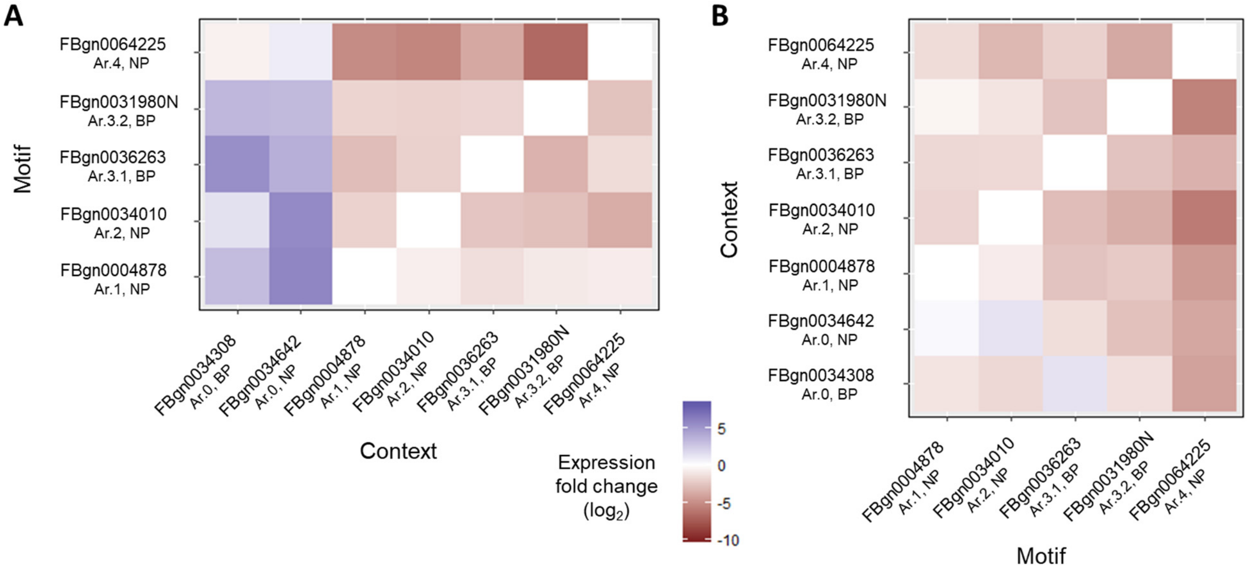


Figure 25. The effect of motif context exchange (log$_2$ scale). (A) Heatmap depicting the mean expression fold changes caused by motifs (y-axis) inserting to different contexts (x-axis), which are the expression changes relative to wild-type expressions of the context-origin promoters FBgn0034308, FBgn0034642, FBgn0004878, FBgn0034010, FBgn0036263, FBgn0031980N and FBgn0064225. **(B)** Heatmap depicting the mean expression fold changes caused by different contexts (y-axis) surrounding motifs (x-axis), which are the expression changes relative to wild-type expressions of the motif-origin promoters FBgn0004878, FBgn0034010, FBgn0036263, FBgn0031980N and FBgn0064225.

Given the effects observed for motif contexts and the strong predictability of core promoter activity based on individual motifs, we wondered whether motifs together with their contexts could behave similarly in defining core promoter function. A linear regression model was also learned here from the results obtained in the inter-architectural block-wise combinatorial mutations (detailed mutation design in Section 4.1.4.2). Remarkably, the predicted values also showed a good correlation with the measured expressions (PCC $r = 0.81$, $p < 2.2\times10^{-16}$; Figure 26), recapitulating the possible additivity for sequence blocks even among various promoter architectures. The coefficients learned from the model revealed the significance of the block features as well (Table S12), although some coefficients were not significant probably due to too sparse data (not all inter-architectural mutated promoter constructs were recovered). Block 5s generally had no impact on the predictions (average $p > 0.6$). The tested block 5s always contain functionally-similar motifs essential for transcription initiation such as INR, INR2, CA-INR, Ohler7 and R-INR. A possible explanation is that those important TSS-related motifs are always retained here. However, one cannot exclude that this block feature is closely correlated to other blocks, leading to a non-significant impact on prediction.

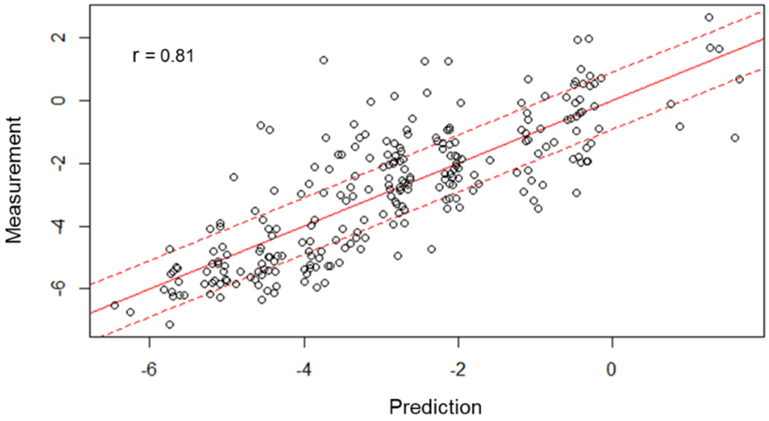


Figure 26. Linear regression analysis for inter-architectural block-wise combinatorial mutations. The measured expressions (on the y-axis) for inter-architectural block-wise combinatorial mutations compared to the predicted expressions (on the x-axis) from the linear regression fit **(**$\log_2$ scale). Red solid line: y = x; red dashed lines: y = x ± 3×SD (SD denoted the experimental noise, that is the median of all SDs for all measured synthetic promoter constructs; also used in outlier filtering procedure described in Section 7.3).

We also found that the contributions of significant block features correlated with the influence on expressions of specific motifs inside these blocks. For instance, block 4 in FBgn0034010, block 4 in FBgn0031980N and block6 in FBgn0004878 were the most significant features ($p < 3\times10^{-7}$; Table S12) found in the model, in which TATA-Box, DRE and MTEDPE motifs locate, respectively. They all contributed positively to the expression levels when swapping with other blocks in the same region (average coefficient > 1.85). Interestingly, the block 3 in FBgn0031980N which contains DRE, however, gave a negative contribution (coefficient = -1.57, p = 0.011). This indicates a possible positional preference for DRE to be located in block 4. The background sequences of block 3 in FBgn0034010 and FBgn0086519 (with a non-functional CGpal) provided significantly negative effects (average coefficient < -2.4). The block 6 in FBgn0034010 with a TTGTTrev played a slightly negative role as well, which is again consistent with the repressive function of this motif (coefficient = -0.7, p = 0.008).

To summarize, our results show that not all information is contained in the motifs. The context sequences surrounding the motifs in core promoters also play an important role in defining the activity. These effects are generally less prominent than the influence of the motifs themselves, which is expected. Importantly, the block sections which contain motifs together with flanks or only their surrounding context sequences largely function in a linear way for setting expression levels.

9.5 Ecdysone responsiveness correlates with the core promoter architecture

Next, we checked the global ecdysone responsiveness (here, the ecdysone responsiveness is defined as the ratio between the induced and uninduced expression level; also referred to as the ecdysone inducibility or the expression fold change caused by the ecdysone induction) for our entire synthetic promoter library. An *a priori* scenario was the possible repressor role of unliganded EcR known before (Cherbas et al., 1991; Dobens et al., 1991). We first performed control experiments which confirmed that the activity of a synthetic promoter containing the block 2 sequence without induction was similar to the activity of the same core promoter sequence but without EcR/USP binding sites in block 2 (data not shown). This suggested that the measurements without ecdysone induction in our experiments represent the basal activity of the tested synthetic promoters.

Overall, the activities of almost all promoter candidates (both native and mutated) tested in our experiments could be increased by ecdysone activation, with a wide inducibility range (more than a 1000-fold difference between the highest and lowest effect). Remarkably, we found out (Figure 27A, Figure S9A and B) that developmental core promoters (Ar.1 and Ar.2) were highly induced with an average > 20-fold activity increase, while constitutive core promoters (Ar.3-housekeeping and Ar.4-ribosomal) showed much weaker responses (around a 4-fold increase on average). Since ecdysone is a developmental stimulus, it is not surprising that it preferably activates developmental core promoters.

Some housekeeping core promoters with already high basal expression levels without ecdysone stimulation ($\log_2$ expressions > 2) exhibited much smaller activations, suggesting saturation of promoter activity (Figure 27A) that cannot be further enhanced. To gain deeper insight, we checked the expression fold changes compared to the basal expression levels of each promoter with native and all mutated core sequences (Figure 27B). With the exception of one group of sequences (derived from FBgn0060296) having increased inducibility with higher expression ($r = 0.51$, $p = 0.012$, most likely due to too few data), we found a generally negative correlation between inducibility and expression level without ecdysone stimulation for all the other promoters (although some showed non-significant effects ($p > 0.05$), especially for Ar.2 promoters). Hence, the higher the expression level, the lower the inducibility. This is consistent with the low activation measured for promoters with high basal expression level. The negative correlation was also found more significant for constitutive core promoters than developmental ones (Wilcoxon rank-sum test $p = 0.0054$; Figure S9C).

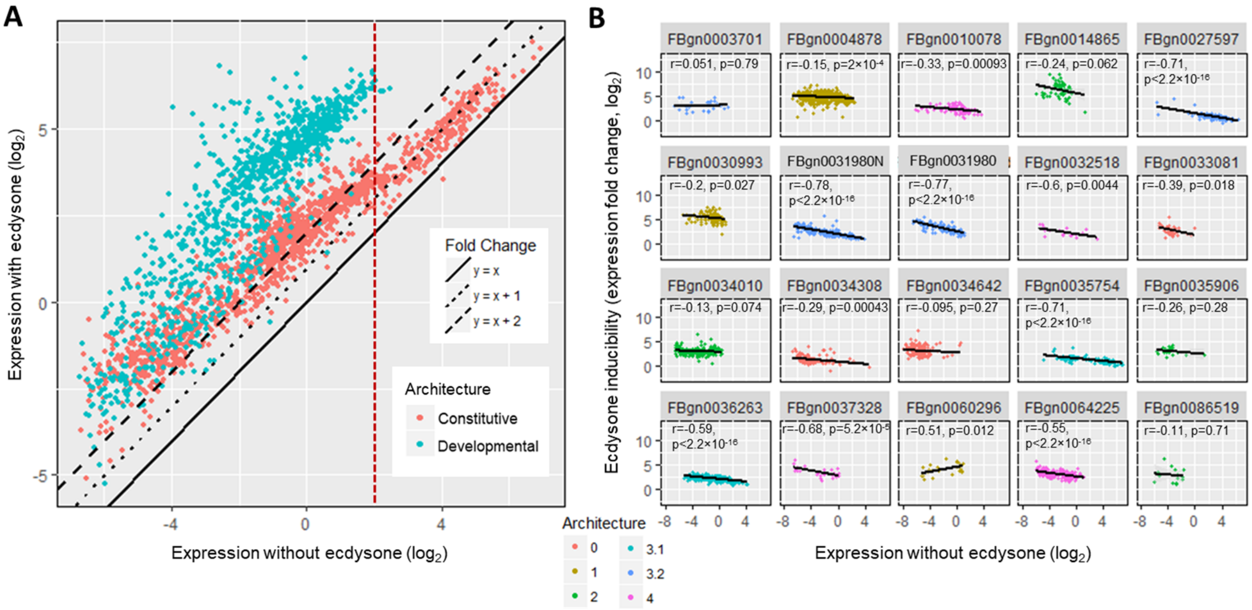


Figure 27. Ecdysone induction effect ($\log_2$ scale) for all tested promoters. (A) Scatterplot depicting the expression measurements with ecdysone induction versus measurements without ecdysone for all tested promoters separated by promoter class. Developmental and constitutive promoters are labeled cyan and red, respectively. Three types of line are used to indicate the expression fold change with no increase (y = x; solid line), 2-fold increase (y = x + 1; dotted line) and 4-fold increase (y = x + 2; dashed line). Red dashed line: $\log_2$ basal expressions = 2. **(B)** Comparison of the expression fold changes (ecdysone inducibility) versus measurements without ecdysone for all tested promoters (grouped by native core promoter sequences). The colors refer to different core promoter architectures. Black line: linear regression (with the 95% confidence interval shown in gray, PCC r and p are shown for each group).

The ecdysone inducibility was generally independent of nearly all motif knockout mutations with an exception of INR (a slightly negative effect of ~ 2.3-fold reduction on average, Wilcoxon rank-sum test $p = 2.1\times10^{-5}$; Figure S9D). Similarly, the motif consensus sequences also could not change dramatically the ecdysone responsiveness (< 20% reduction on average; Figure S9E).

Together, our results demonstrate a correlation between the ecdysone responsiveness and the core promoter architecture. Ecdysone can induce both developmental and constitutive core promoters but drives higher stimulations on developmental ones. There is a negative relationship between the ecdysone inducibility and the basal expression level, that is, the ecdysone inducibility generally decreases with the expression level for a given promoter: the higher the activity, the more difficult it is to boost further expression level. For very strong promoters, inducibility becomes weak, probably due to promoter activity saturation. Finally, the motif disruption rarely influences the ecdysone responsiveness of the core promoter.

9.6 Various potential nucleosomal contexts affect expression

The different block 1s and block 7s were tested as different potential nucleosome surroundings for core promoter sequences in our experiments. Probing the pair-wise block 1s and block 7s showed that the paired block 1.11 and block 7.11 (hereafter termed as B1.11 + B7.11) gave the highest expression (Figure 28A). The nucleosome occupancy of one synthetic promoter construct with this pair (block 1.11 + block 2 + FBgn0035754 + block 7.11) was checked using MNase-seq (performed by Dr. Alessio Renna in our lab). A nucleosome pattern, especially a potential +1 nucleosome, could be detected for this pair of B1.11 + B7.11 which was derived from the genomic location of ±1 nucleosome in a gene FBgn0033924 with a similar +1 nucleosome pattern in MNase digested chromatin samples (Figure 28B). This B1.11 + B7.11 combination was also selected as the fixed nucleosomal context for highly mutated block 3-6s in our later experiments.

We checked the influence of different nucleosomal contexts (block 1 and block 7) on the expression level of five native core promoters. These sequences were selected from the different architectures (Ar.0, Ar.1, Ar.2, Ar.3 and Ar.4; details in Section 4.3), and such that their activities covered the entire dynamic range of our measurements. We created constructs containing these promoters surrounded by all free combinations of the different available blocks 1 and 7 (details in Section 4.3). In total, we tested 127 promoter variants recovered out of 136 possible combinations of different blocks 1 and 7. As expected, we observed lower activities compared to the constructs containing combinations B1.11 + B7.11 (an average signal reduction > 2.5-fold). Smaller variations for B1.X + B7.11 samples compared to B1.11 + B7.X indicated that block 7s might have relatively stronger influences than block 1s on the expression level. The sequence downstream the TSS (B7.X) forming a +1 nucleosome may set a transcriptional obstacle. It may also influence post-transcriptional events as it constitutes the main component of the 5’ UTR region. We computed the GC content of each block 1 and block 7, speculating that as the GC content usually correlates well with nucleosome occupancy, it might correlate with our expression data. We did not see a clear relationship between GC content of the different blocks 1/7 and the expression levels, however (Figure S10A). Furthermore, to test if the presence or absence of block 7s has a strong influence on the expression levels, we also checked core promoter sequences with block 1s variants only, and no block 7 ("w/o B7" in Figure S10A). In these constructs, the length of the 5' UTR was thus reduced from 333 nt to 89 nt. We observed that the expression levels of promoters with or without block 7 sequence, were in the same order of magnitude, whether induced by ecdysone or not (Figure S10B; all constructs contained block 1.11 kept constant, PCC $r = 0.96$, $p = 1.2\times10^{-5}$). The deviations were within the variances we observed with the constructs containing different block 7s. These results suggest that in our synthetic promoter sequences the block 7 presence and its content have an influence on expression level, but the effect is limited compared to core promoter motifs.

After having evaluated the overall effect of different nucleosomal sequences, we next explored potential promoter specificity. Indeed, the two tested constitutive promoters FBgn0010078 and FBgn0027597 exhibited stronger expression variations when altering block 7s (the median SD within the same block 1 is 1.23 compared to the median SD of 0.66 for developmental promoters; Wilcoxon rank-sum test $p = 3.1\times10^{-4}$; Figure S11A). In contrast, both the constitutive and developmental promoters showed similar and milder expression fluctuations upon block 1 variation (the median SDs within the same block 7 for constitutive and developmental promoters are 0.64 and 0.54, respectively; Wilcoxon rank-sum test $p = 0.3$, not significant; Figure S11B). Since previous genome-wide studies showed constitutive promoters tend to have a preferred canonical nucleosome pattern with a strongly positioned +1 nucleosome (Mavrich et al., 2008; Rach et al., 2011), block 7s which were designed in our experiments to act as different potential +1 nucleosomes could have more prominent influences on constitutive promoter activities.

Collectively, our results show that different potential nucleosomal contexts show moderate effects on expression levels with a more significant effect found for sequences potentially forming +1 nucleosomes. Constitutive core promotes are more sensitive to the influence of nucleosomal sequences downstream the TSS.

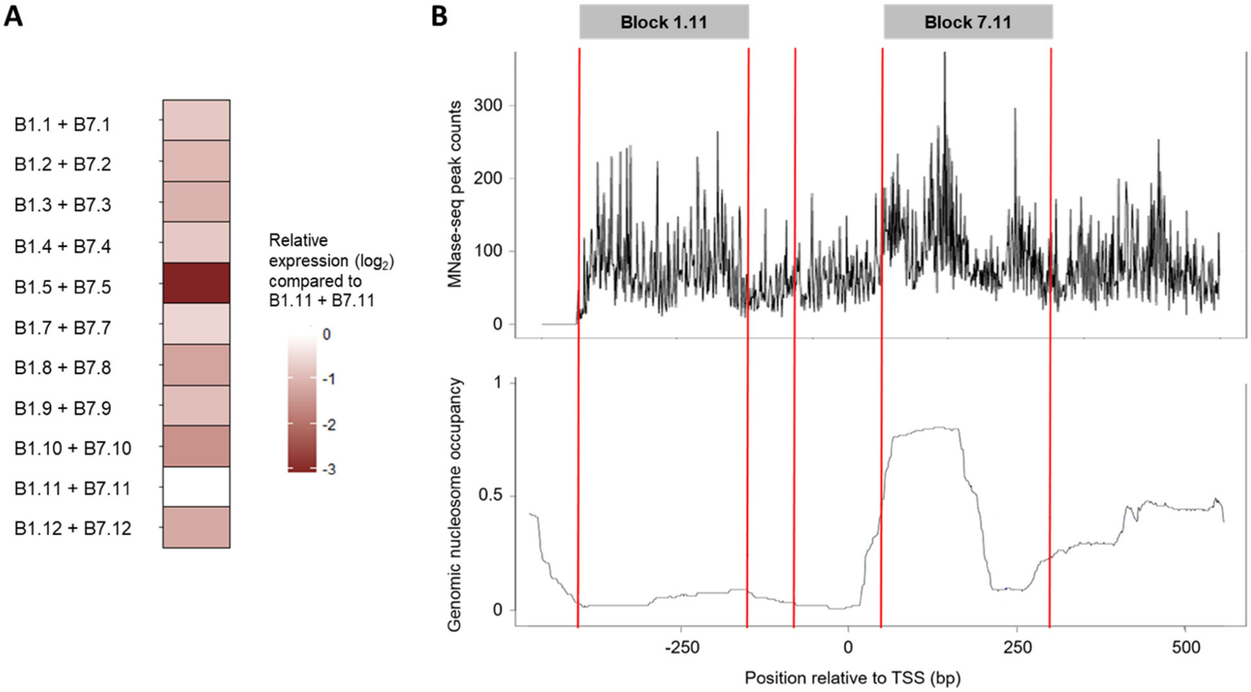


Figure 28. The effect of nucleosomal context on expression (log$_2$ scale) and the nucleosome pattern of the block 1.11 and block 7.11 combination. (A) Heatmap depicting the relative expression measurements of promoter constructs with different pairs of block 1 and block 7 compared to B1.11 + B7.11 expressions. Results were pooled for all tested native core promoters to calculate the average deviation to B1.11 + B7.11 expressions. **(B)** Upper panel: MNase-seq check of the nucleosome occupancy of one synthetic promoter construct (block 1.11 + block 2 + FBgn0035754 + block 7.11). The peak in block 7.11 suggested a potential +1 nucleosome; lower panel: nucleosome occupancy for sequences at similar genomic locations of the gene FBgn0033924, from which block 1.11 and block 7.11 sequences were derived.

IV DISCUSSION AND OUTLOOK

10. Discussion

10.1 The automated workflow established for cell transfections and dual luciferase assays makes the high-throughput quantitative analysis of promoter activity feasible

Minimizing the variabilities becomes crucial if one wants to measure a wide dynamic range of promoter strengths with high reproducibility. In this thesis, I established a reliable workflow for highly accurate promoter activity measurements, mainly involving transient co-transfection, ecdysone treatment and dual luciferase assay, which were validated and successfully implemented on the robotic systems. By testing various transfection reagent FuGENE® HD - DNA ratios and reporter plasmid - control plasmid mass ratios, together with modifications in reagent handling process, I developed a highly efficient transfection procedure for *Drosophila* S2 cells. I also integrated into the automated protocol for ecdysone induction under different conditions (ecdysone concentration and incubation time). To establish the luciferase assay, I screened various types of readout plates in both 96-well format and 384-well format, according to the assay performances and managed to realize time- and cost-efficient luminescence readings whilst eliminating crosstalk and insufficient quenching issues.

In addition to the optimizations mentioned above, several other experimental aspects were also carefully considered. For reporter plasmids isolation, a kit including an endotoxin removal step was used to reduce the possible toxicity to S2 cells after transfection (which might lead to abnormal luciferase signals; (Fan & Wood, 2007). Several inter-plate control plasmids were systematically co-transfected with the renilla control plasmid into cells in fixed wells of 96-well plates for signal monitoring and further correction. These control plasmids including the renilla control plasmid were stored in aliquots to avoid plasmids degradation due to multiple freeze-thaw cycles. To better control the quality of the cells used for transfection and later luciferase assays, the cells 24 h growth rate and viability were checked and strictly monitored during passaging.

For preprocessing of the raw data, I established a computational workflow to normalize the raw luminescence signals and derived the final expression values which represented our synthetic promoter activities. This procedure enabled a strong reduction of variabilities.

Taken together, the adapted protocols on automated workstations, the detailed optimizations for each step and the stringent data normalization procedures made it possible to achieve high-throughput measurements of reporter gene expression driven by thousands of wild-type and synthetic promoters with substantially reduced variations among separate experiments. Notably, this workflow enabled us to transfect 24 plates of sample plasmids in 96-well plate format and measure more than 2000 luciferase assays in one week.

10.2 The strength and limitation of the study

The high-throughput MPRAs have been deployed widely to address the question that how gene expression is regulated by different CREs. Nevertheless, most studies are dedicated to understanding enhancer activities as well as the function of TF binding sites. Few MPRAs were designed for *in vivo* promoter analysis: extensive studies on fully designed yeast proximal promoter regions (Sharon et al., 2012) and yeast core promoter sequences (Lubliner et al., 2015); analysis of autonomous promoter activity of random genome fragments in human (van Arensbergen et al., 2017) and in *Drosophila* (Arnold et al., 2017). In the latter study, Arnold and colleges used the STAP-seq they developed. Unfortunately, their measurements were not sensitive enough to study the basal activity of the putative promoters, but only their enhancer responsiveness.

In this thesis, I provide the first comprehensive dissection of *Drosophila* promoter features including core promoter motifs and their surrounding contexts, ecdysone stimulus responsiveness and potential nucleosome contexts.

Methodologically, our study presents several advantages. First, our luciferase-assay-based method overcomes the limited dynamic range of less than 2-3 orders of magnitude in previous studies with highly sensitive and reproducible results ranging over more than 4 orders of magnitude. Second, compared to the fluorescence-based method clustering expression levels into separate bins (Lubliner et al., 2015; Sharon et al., 2012), our study provides a more quantitative and continuous measurement scale. Third, in contrast to other RNA-seq based MPRAs inserting barcodes or tested CREs in 3’ UTR of the reporter gene, our method will not affect transcript stability with inaccurate reporter expression measurements. Fourth, our systematic mutagenesis of core promoter sequences is based on various wild-type promoter sequences in the genome instead of 1-2 unique backgrounds or entirely artificial sequences. Finally, although the DNA synthesis is still limited by the oligonucleotide length, the Golden Gate cloning strategy (BsaI cloning) used in our approach can combine up to ten fragments in a single reaction for testing long and combinatorial features. Thereby, it also provides the possibility to analyze the transcriptional output with other stimuli by switching the activator region (block 2 sequence in this thesis) with other stimulus-response elements, which can also vary in both number and affinity.

However, the study presented in the thesis still has limitations. All of our measurements were performed using episomal plasmids in transiently transfected cells. Although we inserted the genomic ±1 nucleosome positioning sequences from different genes surrounding the tested core promoter region, they still lack the ability to represent the endogenous chromosomal context and the higher-order genomic structure, which might change the basal expression levels as well as the ecdysone inducibilities. Our method has a moderate throughput that is less than other sequencing-based approaches. The cloning and colony picking procedures also limit our sequence recovery from the designed oligonucleotides.

In summary, we have established the first high-throughput luciferase-assay-based method to quantitatively measure both the basal and induced activity of systematically designed promoters in *Drosophila*. Although our assays are episome-based, our method still provides unique advantages and can be applied for analyzing diverse regulatory sequence features.

10.3 The systematic mutagenesis of core promoters ascertains the effects of specific sequence features on the expression level

The majority of this work focused on interrogating the *Drosophila* core promoter sequence features, with a large amount of effort devoted to the sequence motifs that are well-known or newly discovered in a previous study (Hartmann, 2012). Our results reinforce the conclusions drawn from other smaller-scale studies for the roles of core promoter motifs in determining transcriptional output, also generalizing their effects to more promoter backgrounds. In addition, this work also brings new insights into grammatical rules of *Drosophila* core promoter function.

We demonstrate that the well-known functional motifs like INR, TATA-Box, MTEDPE, INR2 (more widely known as Ohler1 or motif 1), DRE and Ohler7 are crucial for gene expression. Their roles are unique and they cannot be replaced by positionally or functionally similar motifs from other architectures. Only INR2 can largely function as an INR-substitute. Pairwise knockouts mostly elicit more significantly negative effects on transcription and these effects often show superadditivity (in $\log_2$ scale), except for the paramount developmental core promoter motifs including INR, TATA-Box and MTEDPE. As expected, removal of all motifs shuts down the core promoter, resulting in an even lower expression than the basal expression level of the weakest motif-less core promoters. In addition, most of the motif consensus sequences tend to increase core promoter activity. This again emphasizes the importance of the sequence motifs for core promoter function. Importantly, all these findings are consistent between different core promoters.

However, not all well-characterized motifs have a significant effect on expression. This is especially the case with R-INR, more widely known as the TCT motif. It surprisingly makes almost no contribution to the expression although it exists in nearly all ribosomal protein gene promoters in *Drosophila*. In contrast, housekeeping core promoter motifs like INR2 and Ohler6 that co-occur in multiple promoters show stronger influence in our data. It is known that more than half of the ribosomal core promoters contain this INR2 motif (Ma, Zhang, & Li, 2009). A recent study proposed that the INR2 binding protein M1BP can act as an intermediary factor to recruit TRF2 for proper transcription of ribosomal protein genes (Baumann & Gilmour, 2017). Our perturbation analysis of INR2 in various ribosomal promoter backgrounds supports their finding. The results we obtained with Ohler6 also suggest the unknown TF(s) that bind to it may function similarly as M1BP.

Our highly sensitive assay can also accurately capture the expression changes caused by single base pair variations, leading to the measurements of corresponding motif activity. We confirmed that the most-overrepresented sequence of a given motif in the genome still mainly stands for its best functional form, but we also saw differences with the computationally derived matrices: our expression-based activity logos are generally less specific, indicating that computationally obtained motif logos may not capture the true TF binding specificities.

Altering motif positions overall decreases expression. Several studies have suggested the exact spacing is essential for synergism between the core promoter motifs to function as active pairs to recruit GTFs along with Pol II for accurate transcription initiation (Burke & Kadonaga, 1997; Emami et al., 1997; Gershenzon & Ioshikhes, 2005; Gershenzon, Trifonov, & Ioshikhes, 2006; O'Shea-Greenfield & Smale, 1992). Our results are in line with these previous findings for strictly positioned motifs such as INR, MTEDPE and TATA-Box. Their locations and spacings are highly restricted for the effective binding of the TFIID to nucleate the PIC. Other motifs which can function over wide ranges and are not necessary for constituting the major machinery, e.g., DRE, Ohler6 and Ohler7, also show less stringent location requirement and smaller effects on expression as long as they do not disrupt other sequence features.

Importantly, we also demonstrate that not only the core promoter motifs are essential, but also their sequence context. Our results uncover that sequence motifs mostly prefer their native context. Remarkably, although only INR and INR-like motifs including INR2 and Ohler7 can drive higher expression when their consensus inserted into motif-less core promoters, the motif combinations from almost all the other defined architectures can result in the substantial increase of expression level, revealing the importance of motif synergism. We also saw an influence of the sequence context independent from the motifs, which may obey complicated rules. It is however beyond the scope of this study.

Finally, among the four tested novel motif candidates discovered by XXmotif, we identified TTGTTrev and RDPE as having measurable effects on expression after mutation, hereby confirming their biological relevance. TTGTTrev shares a similar function with a negative regulatory element for binding of a transcriptional repressor AEF-1. RDPE is highly correlated with R-INR and can partially replace the function of MTEDPE in developmental architectures. However, we note that the mutations in other newly discovered motifs like TTGTT and CGpal show little effect on expression, suggesting these two computationally derived over-represented sequences lack functional importance. Due to the similarity of TTGTT with R-INR, this motif may act as a redundant version of the weak R-INR motif.

In summary, this first large-scale comprehensive dissection of *Drosophila* core promoter sequence features explicitly demonstrates that motif strength and positioning are the fundamental parameters that determine the core promoter activity and govern the transcriptional output. Besides, our results also support that not only the clearly defined motifs, but also their context sequences are important for core promoter function.

10.4 The motifs and their flanking sequences can act additively to define a significant fraction of the core promoter function

In addition to analyzing individual mutations or positional changes of motif sequences, our study also gives insights into the debated role of motif flanking and context sequences of core promoters. Considering that pairwise motif disruption already suggests certain levels of synergistic effects, the higher-order combinatorial effect of mutant motifs and their context on expression may be more difficult to understand. To dissect this complexity of the mutant combinations, we used a linear regression model to check how much of the core promoter activity can be correlated with individual effects. To our surprise, we found that the expression changes caused by single mutations of sequence motifs joined in a linear fashion can largely predict the output of the free mutant combinations, with about 77% of the variance in the data explained (average $r^2 = 0.77$). Hence, promoter expression levels of the combinatorics can largely be explained by simple linear addition of their individual contributions.

We next extended the sequence features from simply the motifs alone to the larger sequence blocks which contain motifs together with their flanks or only context sequences. Remarkably, the linearity is also found among the data from inter-architectural block substitutions ($r^2 = 0.66$).

In conclusion, a linear combination of individual sequence features including motifs and wider sequence blocks including motif flanking and context sequences can overall account for more than two-thirds of the variance in expression levels as regulated by the core promoter. More detailed models will be required to unravel the nonlinear interactions.

10.5 The developmental and constitutive core promoter architectures differ in their responsiveness of surrounding functional sequences: ecdysone response element and nucleosomal context sequences

Two kinds of core promoter surrounding sequences have been tested in our experiments. One piece of synthetic sequences containing ecdysone receptor binding sites was placed upstream of the core promoter region to probe the steroid hormone ecdysone-responsive core promoter activity. Our data show that the ecdysone responsiveness highly correlates with core promoter architecture. This developmental stimulus ecdysone functions more strongly on developmental core promoters.

Moreover, we uncover a generally negative correlation between the ecdysone responsiveness and the basal expression level. Also, our strongest promoters can be barely activated. Thus, the higher the expression level, the more difficult it is to further boost the signal, giving hints about the existence of a promoter saturation expression level. This negative correlation is also more significant for constitutive core promoters, showing the less efficient activation for them. Interestingly, the disruption of INR in developmental core promoters can lead to a reduction in the ecdysone responsiveness, which is consistent with what was reported in a previous study in *Spodoptera frugiperda* (Jones et al., 2012)

Another tested core promoter surrounding sequences are derived from the genomic ±1 nucleosome positioning sequences in order to test the effect of potential nucleosome binding. The pair selected for the systematic mutant analysis of core promoters were confirmed to form a detectable +1 nucleosome pattern. Compared to motif knockout, we observe moderate effects on expression driven by these different potential nucleosomal backgrounds. We however find greater expression variation for housekeeping and ribosomal core promoters when changing block 7 sequences than developmental core promoters, suggesting the significance of the genomic +1 nucleosome sequences for the function of constitutive core promoters.

Taken together, the different sequence motifs composing distinct core promoter architectures can predict their ecdysone responsiveness: developmental core promoters can get much higher induction. The TSS downstream nucleosome positioning sequences influence more on constitutive core promoter activity.

11. Outlook

The computationally defined sequence motifs tested in our study indeed largely encode the regulatory information in the core promoter. In addition, we discovered that motif-surrounding context sequences of different kinds can also affect regulatory activity. To more systematically explore how these sequence contexts can also drive differential outputs, the nucleotides directly flanking the motifs are naturally the first targets, as similar functions already demonstrated in studies of the TF binding sites in other CREs (Gordân et al., 2013; Maerkl & Quake, 2007; Morin, B., Nichols, L. A. & Holland, 2006; Rajkumar, Dénervaud, & Maerkl, 2013; Schöne et al., 2016). A systematic perturbation of these flanks could provide additional insights into their functional significance and their possible effects on DNA shape, providing the opportunity to extend the current motif definition. The potential experiments can include, for instance, the substitution of flank sequences with other nucleotides as done in our single base pair motif mutations, or the flanks joint with original motifs as an extended module to be shifted either downstream or upstream.

Although our results suggest the motif sequence features and their embedded context can additively define a large portion of the core promoter function, there is still considerable non-linearity found in the data. More complicated models such as machine learning methods can be helpful for exploring higher-order interactions. In addition to sequence motifs, structural properties of DNA surrounding TSSs, including curvature and bendability, DNA melting temperature, helical twist and propeller twist, are thought to influence the landing of transcription machinery on core promoter regions and thereby contribute in core promoter activity (Abeel, Saeys, Bonnet, Rouzé, & Van de Peer, 2008; Dineen, Wilm, Cunningham, & Higgins, 2009; Goñi, Pérez, Torrents, & Orozco, 2007). These physical properties are mostly dependent on the di- or tri-nucleotide content and DNA shape, which can be the extra features incorporated into the models for prediction of the expression levels controlled by core promoters.

In this study, we saw discrepancies between the computationally defined motif PWM logos and the activity logos generated based on our expression measurements of motif point mutations. These newly defined activity-based sequence logos can be presumably regarded as the more accurate representation of the core promoter motifs. Assays to screen more strength-altered motifs for comparing the predicted motif scores and the corresponding expressions can be further validation of these expression-based logos.

Finally, our synthetic promoter measurements were carried out on episomal plasmids in cultured *Drosophila* S2 cells. Therefore, the next goal is to extend our analysis by integrating the tested promoter constructs into the S2 cell genome through site-specific recombination, e.g., using the CRISPR/Cas9 system, for quantitatively probing their activity in the native genomic environment with chromatin context.

APPENDIX A Supplementary Figures

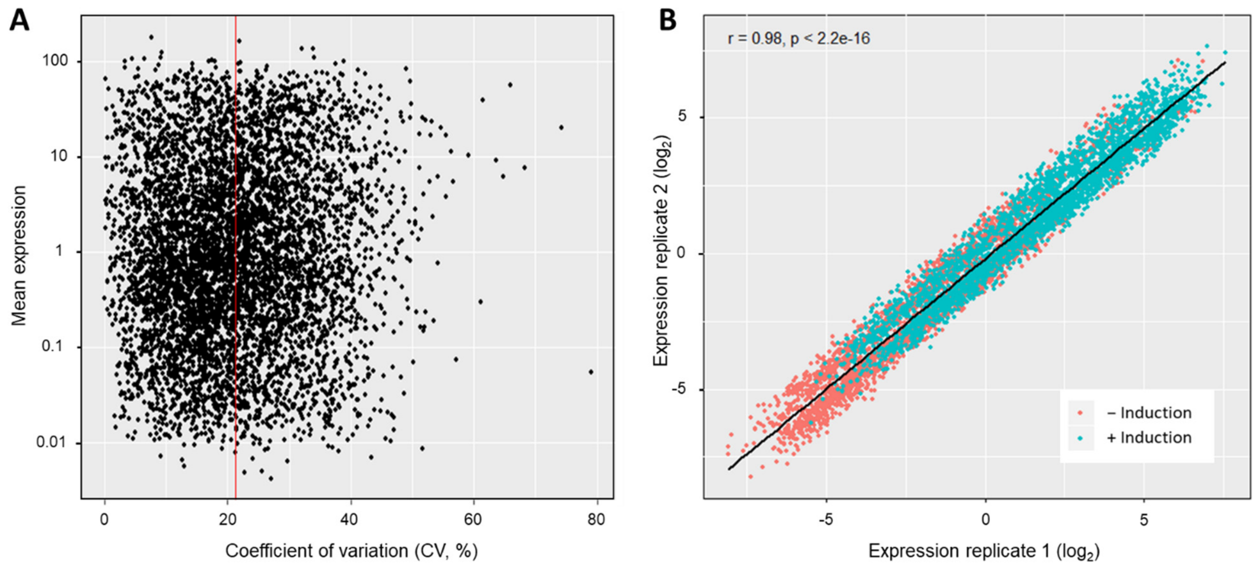


Figure S1. Variation and reproducibility of the expression measurements. (A) Scatterplot depicting the mean expression level for all tested promoters versus the CV. Red line indicates the mean CV of 21%. **(B)** Comparison of the expression measurements ($\log_2$ scale) obtained for two replicates with the most considerable CV of each promoter construct (black line: linear regression, PCC $r = 0.98$, $p < 2.2\times10^{-16}$). Measurements with and without ecdysone induction are labeled cyan and red, respectively.

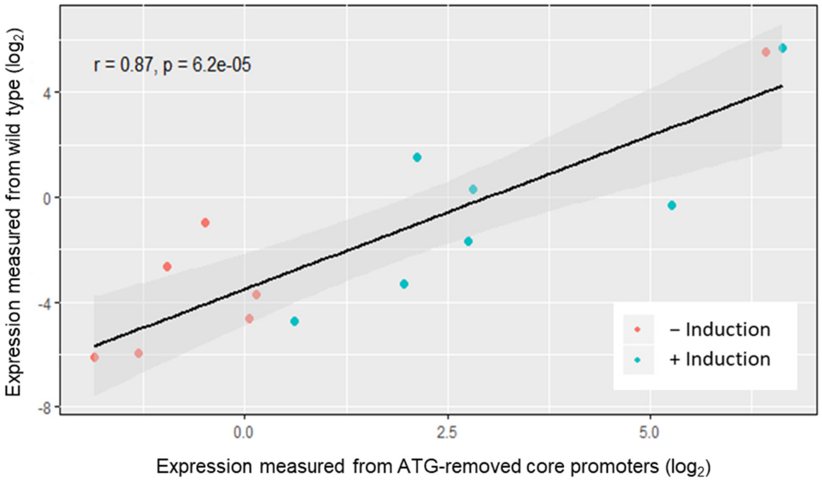


Figure S2. Comparison of expressions measured from ATG-removed core promoters with their wild-type ATG-containing sequences. Black line: linear regression (with 95% confidence interval shown in gray, PCC r = 0.87, p = 6.2×10^{-5}). Measurements with and without ecdysone induction are labeled cyan and red, respectively.

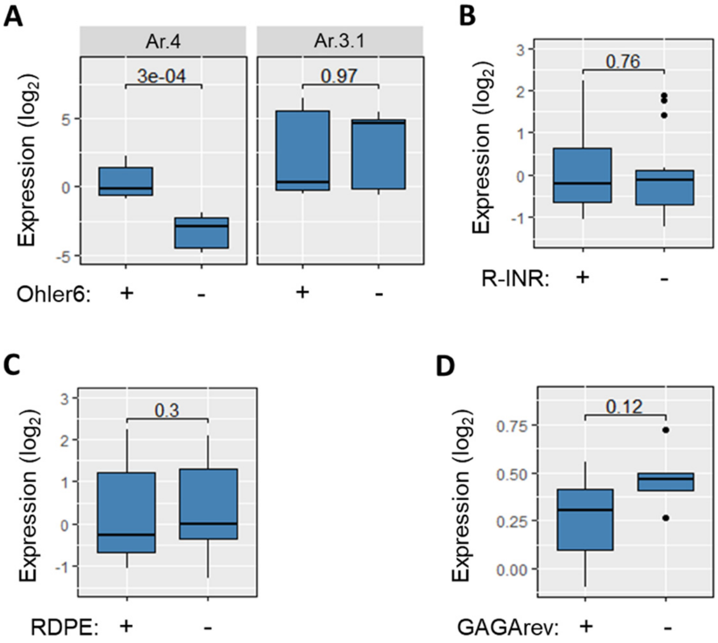


Figure S3. The effect of individual motif knockout (log$_2$ scale). (A) Boxplot depicting the effect of Ohler6 knockout in Ar.4 core promoters (FBgn0010078 and FBgn0064225, all measurements in these two core promoter constructs were pooled together) and in Ar.3.1 core promoters (FBgn0035754 and FBgn0036263, all measurements in these two core promoter constructs were pooled together), respectively. Its knockout effect was highly significant in Ar.4 core promoters (Wilcoxon rank-sum test $p = 3\times10^{-4}$), while it showed no effect in Ar.3 promoters (Wilcoxon rank-sum test $p = 0.97$). **(B)** Boxplot depicting the effect of R-INR knockout in FBgn0010078, FBgn0032518, FBgn0037328 and FBgn0064225. No significant difference was obtained between the expressions with and without an R-INR (all measurements in these four core promoter constructs were pooled together; Wilcoxon rank-sum test $p = 0.76$). **(C)** Boxplot depicting the effect of RDPE knockout in FBgn0010078, FBgn0037328 and FBgn0064225. No significant difference was obtained between the expressions with and without an RDPE (all measurements in these three core promoter constructs were pooled together; Wilcoxon rank-sum test $p = 0.3$). **(D)** Boxplot depicting the effect of GAGArev knockout (non-overlapped) in FBgn0060296. No significant difference was obtained between the expressions with and without a GAGArev (two-sample t-test $p = 0.12$).

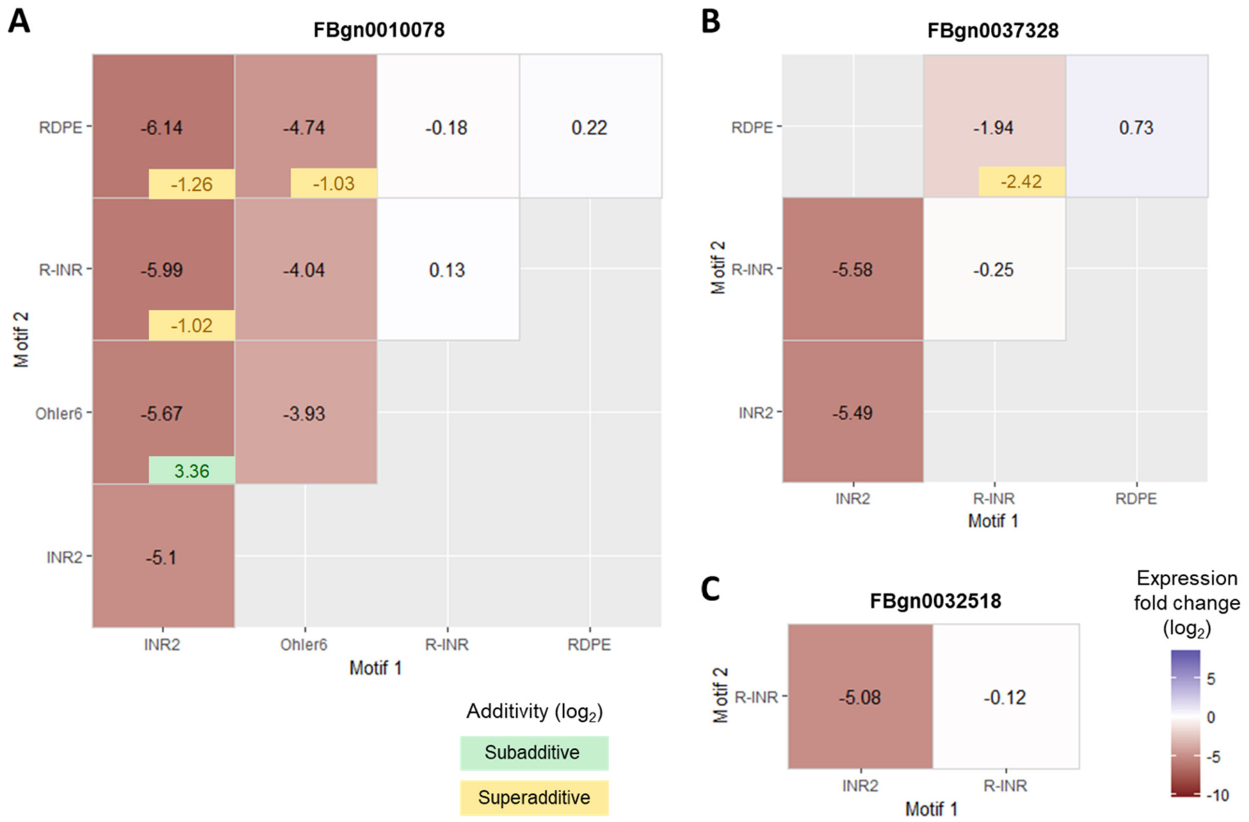


Figure S4. The effect of pairwise motif knockout in ribosomal core promoters (log₂ scale). (A) Heatmap of the mean expression fold changes compared to wild-type expressions for pairwise knockout of motifs compared to individual knockouts (diagonals) in FBgn0010078. **(B)** The same for FBgn0037328 (knockout of INR2 + RDPE pair not recovered). **(C)** Heatmap of the mean expression fold changes compared to wild-type expressions for pairwise knockout of INR2 + R-INR compared to knockout of R-INR only in FBgn0032518. Additivity was calculated as the difference between the pairwise effect and the sum of two individual effects (subadditive (green): > 0; superadditive (yellow): < 0; effects > 3×SD shown in the corner of each pairwise effect).

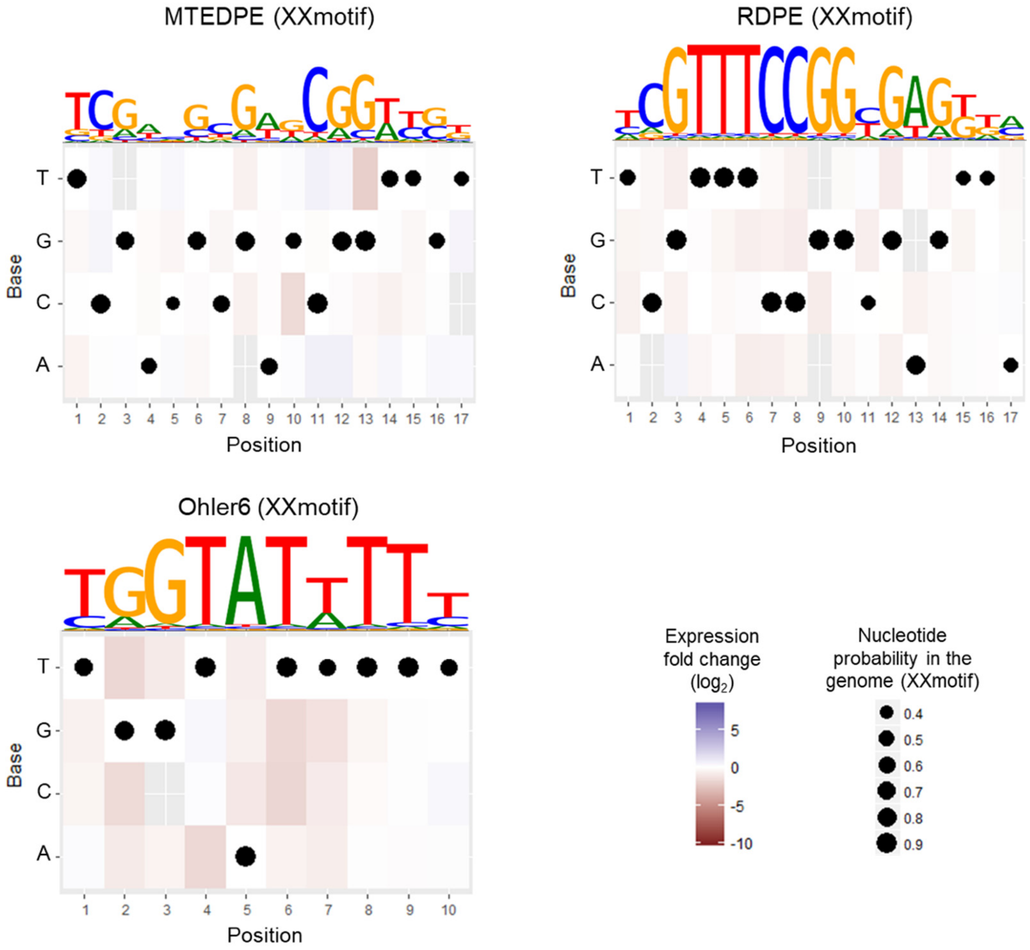


Figure S5. The effect of single base pair mutations of the XXmotif-derived consensus (log$_2$ scale) for MTEDPE, RDPE and Ohler6. Each column of a heatmap shows the expression fold changes of the point mutations at a specific position in the motif. The XXmotif consensus sequence is highlighted as white blocks with spots which size with the nucleotide probability in the genome defined by XXmotif.

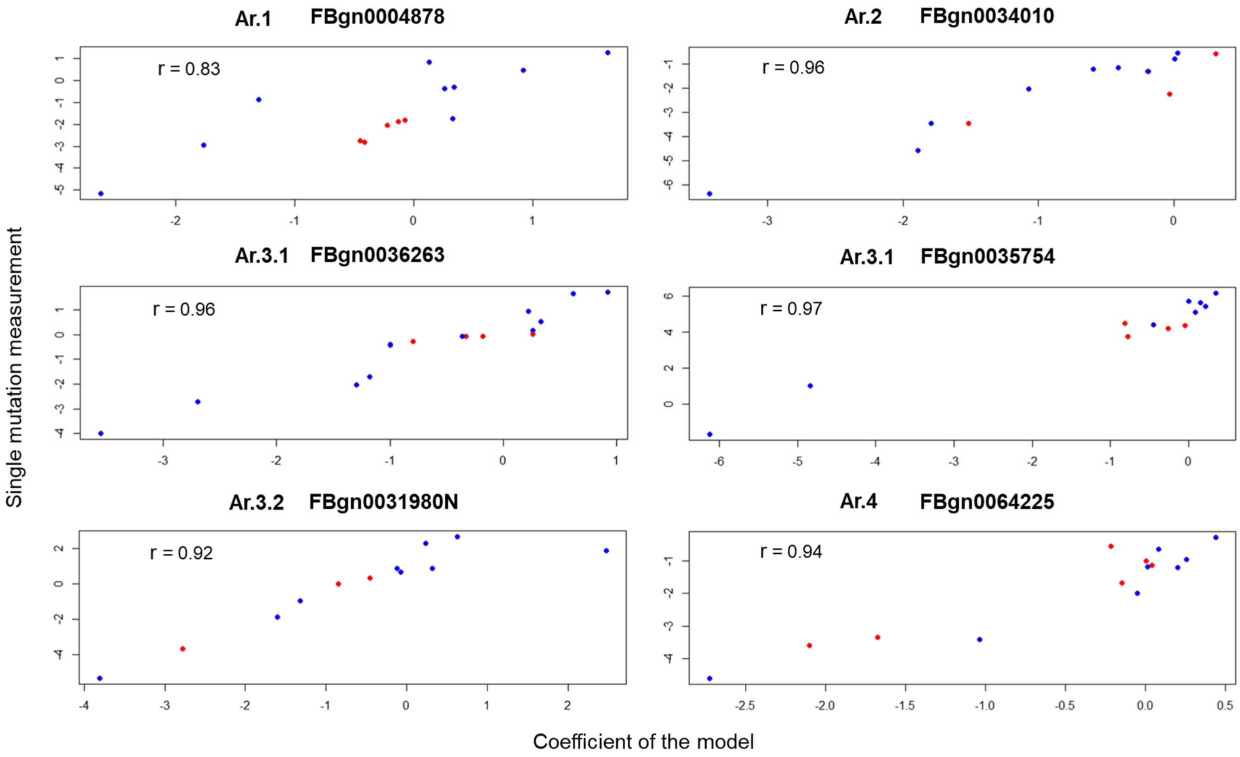


Figure S6. The measured expressions of single mutations (on the y-axis) compared to the coefficients of the linear regression models for the intra-architectural combinatorial mutations (on the x-axis) in 6 tested core promoter sequences ($\log_2$ scale). Single mutations in motif strength and position are labeled blue and red, respectively. The average correlation PCC *r* = 0.93.

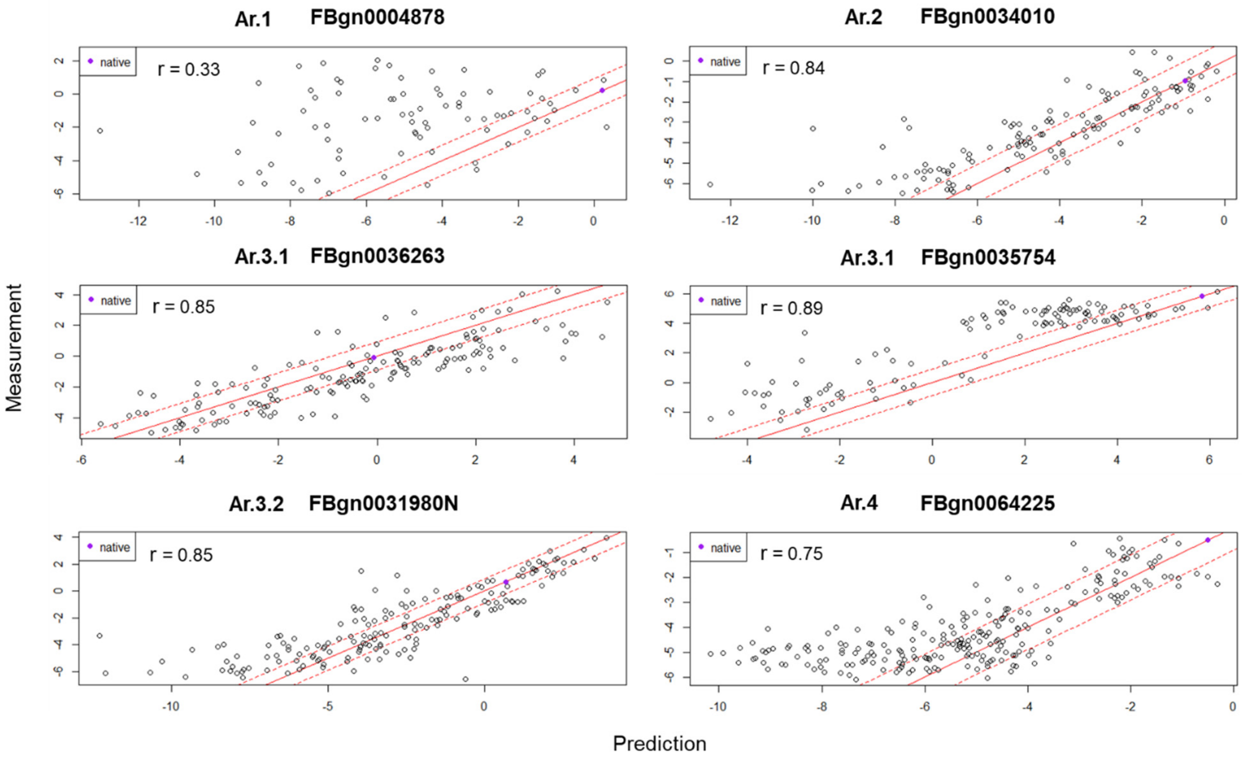


Figure S7. The additive models were able to predict synthetic promoter activity based on individual motif features. The measured expressions (on the y-axis) for 6 tested core promoter sequences with combinatorial motif mutations compared to the predicted expressions (on the x-axis) from the additive model **(**$\log_2$ scale). Red solid line: y = x; red dashed lines: y = x ± 3×SD (SD denoted the experimental noise, that is the median of all SDs for all measured synthetic promoter constructs; also used in outlier filtering procedure described in Section 7.3). Purple dot: the native expression of each promoter.

Except for one tested core promoter FBgn0004878 (PCC r = 0.33, p < 0.003) for which multiple single mutation constructs were not recovered during the cloning procedure (leading to fewer coefficients and much fewer data points used in the model, but originally it has the highest number of motifs and the most complicated motif combinations, thus not comparable with results got from other core promoters), our measurement-based additive model already predicted the expression level with high accuracy (average PCC $r = 0.84$, $p < 2.2\times10^{-16}$). This again indicates that the contribution of the individual sequence features (including both motif strengths and positions) to the core promoter activity is largely linear. However, here we could see considerable discrepancies between the predicted and experimentally derived expression values, consistently to what was observed with the pairwise mutations (Figure 19 and S4). We checked that many of the deviations lay outside our measurement noise. For promoter constructs FBgn0034010, FBgn0031980N and FBgn0064225, we could see a lower boundary of the measured expressions around -5 ($\log_2$), suggesting a minimum level of promoter activity (nearly all values above our detection limit of -5.6 ($\log_2$)). For FBgn0035754 and FBgn0064225, we observed that there was a saturation of the expression levels at 5 to 6 or at -2 to -1 ($\log_2$), respectively (values below the maximal luciferase expression levels measurable with our assay of about 7 ($\log_2$)). A similar trend could be seen for ecdysone-induced measurements, where all low and high saturation plateaus are shifted towards higher values (data not shown). These phenomena are thus biological and not due to any technical artifact from our detection limit. They demonstrate: (a) the existence of a background level for promoter activity independent on the identified motifs and (b) the existence of a maximal expression level corresponding activity saturation of a certain promoter. The other deviations could be attributed to non-linear interactions due to the synergy between the different players of the PIC (as also shown in the pairwise knockout results in Section 9.2.2). Most of the promoter variants expressions shown here were subadditive of the individual mutational effects, while other variants like FBgn0036263 mostly showed weaker activities than the total expression deviation caused by mutations (superadditivity).

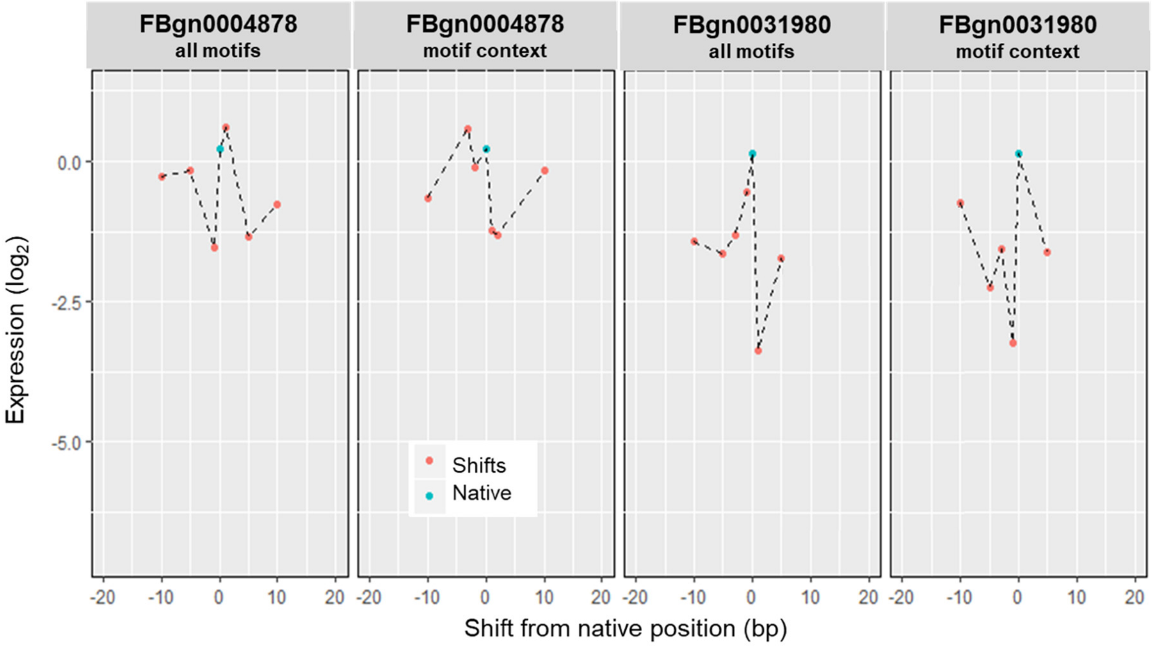


Figure S8. The effect of all-motif shifts and context shifts ($\log_2$ scale). The expression measurements of natives (cyan dots) and positional shifts (red dots) of either all motifs or motif context in FBgn0004878 and FBgn0031980, the two core promoter configurations also used for testing the individual motif shifts.

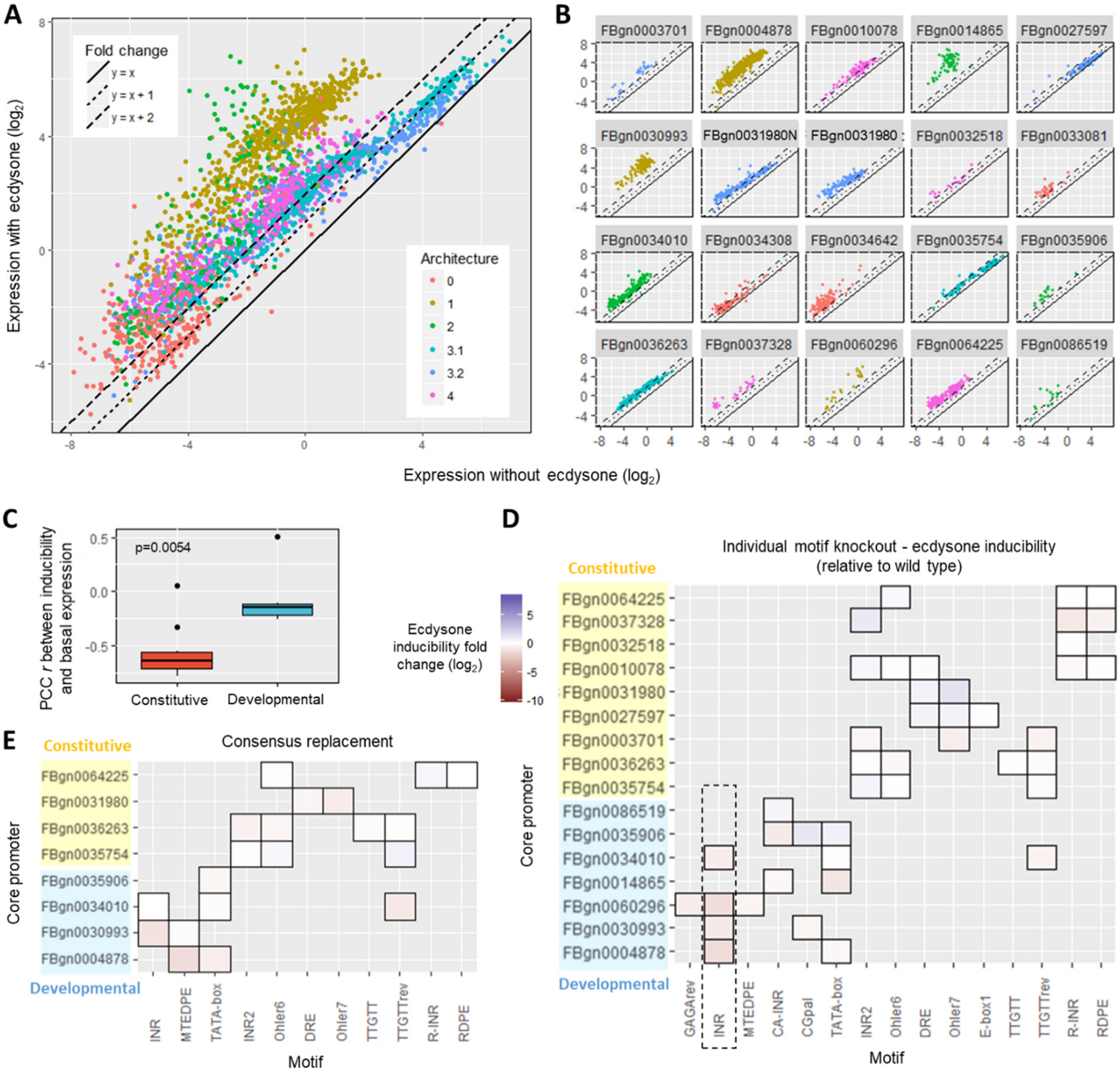


Figure S9. Ecdysone induction effect (log2 scale). (A) Scatterplot depicting the expression measurements with ecdysone induction versus measurements without ecdysone for all tested promoters separated by core promoter architectures. Each color represents one architecture. Three types of line are used to indicate the expression fold change with no increase (y = x; solid line), 2-fold increase (y = x + 1; dotted line) and 4-fold increase (y = x + 2; dashed line). **(B)** The same as (A) but separated by their native core promoter sequences. **(C)** Comparison of the PCC *r*s in Figure 27B grouped by constitutive and developmental core promoters. Wilcoxon rank-sum test *p* = 0.0054. **(D)** Heatmap depicting the ecdysone inducibility fold changes caused by individual knockout of motifs in different core promoters. Disrupted INR had a slightly negative effect on changing the core promoter responsiveness to ecdysone. (~ 2.3-fold reduction on average, Wilcoxon rank-sum test $p = 2.1 \times 10^{-5}$). Constitutive and developmental promoters are highlighted in yellow and blue, respectively. **(E)** Heatmap depicting the ecdysone inducibility fold changes caused by consensus replacement of motifs in different core promoters. Constitutive and developmental promoters are highlighted in yellow and blue, respectively.

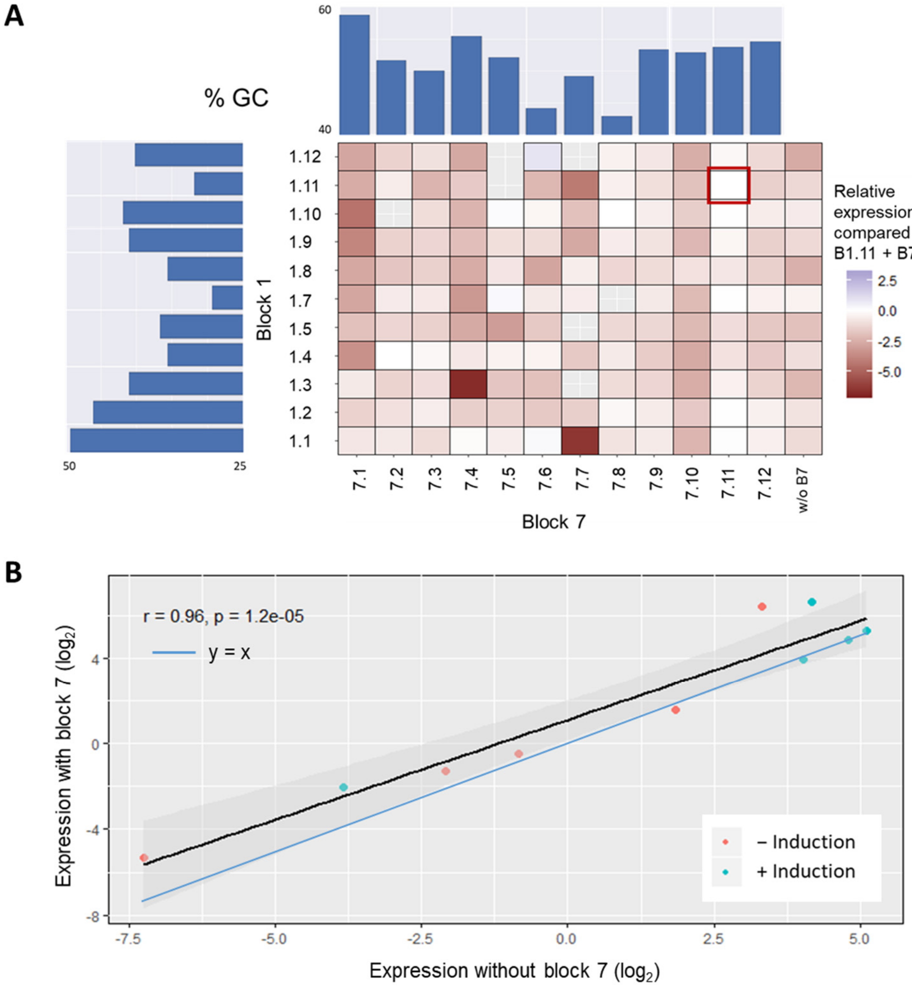


Figure S10. The effect of nucleosomal context on expression ($\log_2$ scale). (A) Heatmap depicting the relative expression measurements of promoter constructs with different free combinations of block 1 and block 7 compared to B1.11 + B7.11 expressions (marked with a red box). Results were pooled for all tested native core promoters to calculate the average deviation to B1.11 + B7.11 expressions. Bar plots represent the GC content of each block 1 and block 7 sequence. Block 7 with column "w/o B7" represents the results got from promoters without block 7 sequence. **(B)** Comparison of the expression measurements for promoter variants with block 7 versus without block 7. Expressions with and without ecdysone induction are labeled cyan and red, respectively. Black line: linear regression (with 95% confidence interval shown in gray, PCC $r = 0.96$, $p = 1.2\times10^{-5}$). Blue line: y = x.

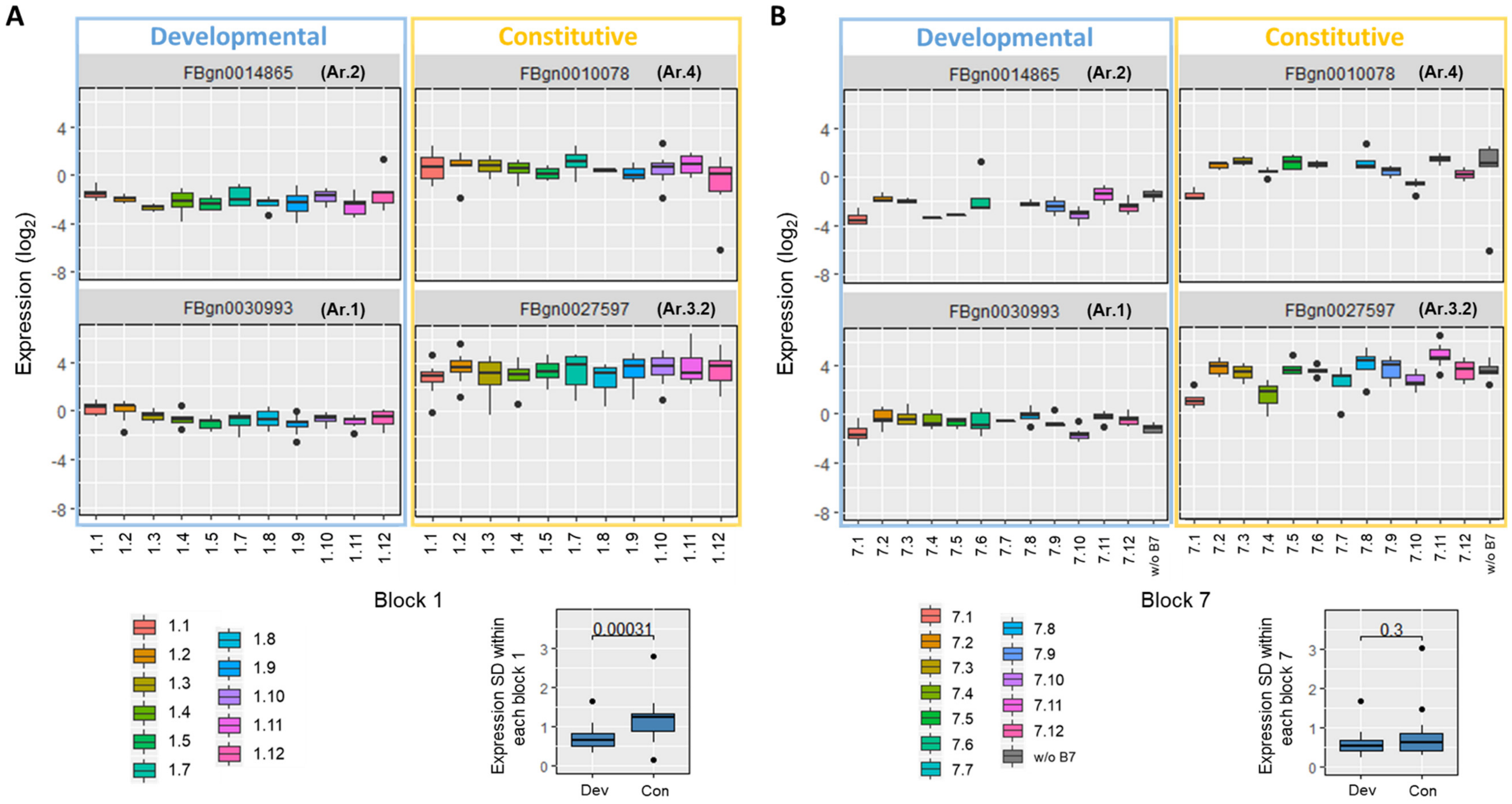
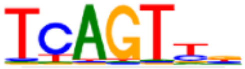
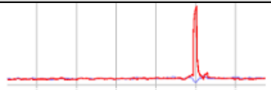
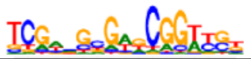
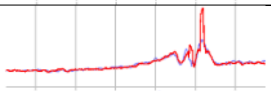
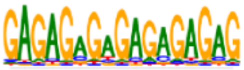
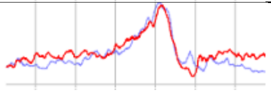
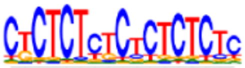
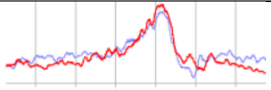
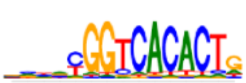
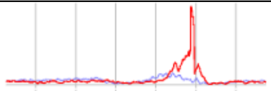
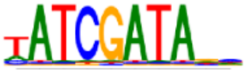
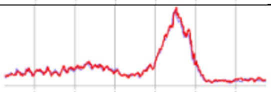
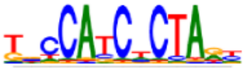
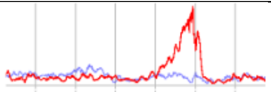
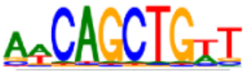
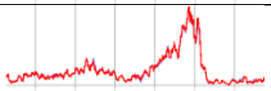
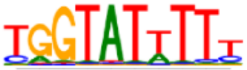
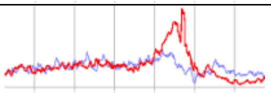
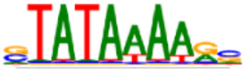
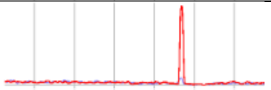
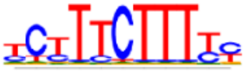
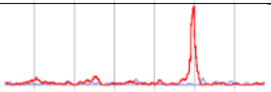
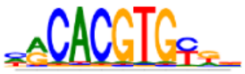
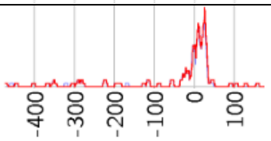
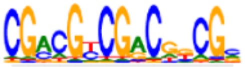
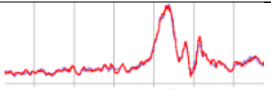
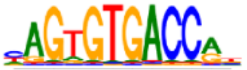
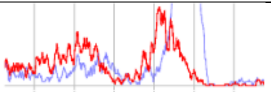
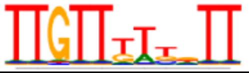
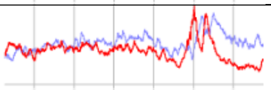


Figure S11. The effect of different block 1s and block 7s on expression in developmental and constitutive promoters (log$_2$ scale). (A) Boxplots depicting block 1 effects for tested core promoters. Effects of different block 7s were merged in each column (within the same block 1): the median SD is 0.66 for developmental promoters compared to the median SD of 1.23 for constitutive promoters; Wilcoxon rank-sum test $p = 3.1\times10^{-4}$, significant. **(B)** Boxplots depicting block 7 effects for tested core promoters. Effects of different block 1s were merged in each column (within the same block 7): the median SD is 0.54 for developmental promoters compared to the median SD of 0.64 for constitutive promoters; Wilcoxon rank-sum test $p = 0.3$, not significant. Block 7 with column "w/o B7" represents the results got from promoters without block 7 sequence. Developmental and constitutive promoters are highlighted in blue and yellow, respectively.

APPENDIX B Supplementary Tables

	Motif Name	Sequence Logo	Distribution	Start Range	Gene Set
Known Motifs	INR*			-2 … -1	NP
	MTE/DPE*			16 … 18	NP
	GAGA*			-100 … -33	NP
	GAGArev*			-100 … 1	NP
	INR2*			-60 … 20	BP
	DRE*			-100 … -7	BP
	Ohler7*			-72 … 14	BP
	E-Box1*			-59 … 32	BP
	Ohler6*			-100 … -10	BP
	TATA-Box*			-35 … -29	MAD high
	R-INR*			-5 … -5	min high
	E-Box2			-22 … 40	elf high
Newly discovered	CGpal*			-100 … -20	NP
	INR2rev			-100 … -2	BP
	TTGTT*			-32 … 40	MAD low

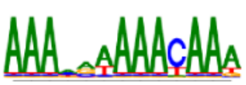
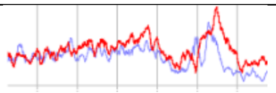
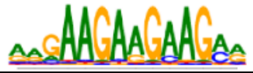
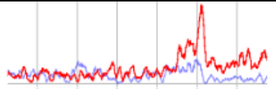
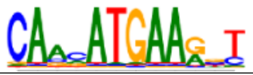
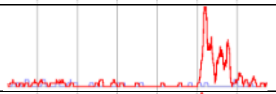
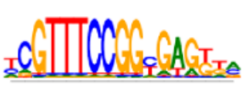
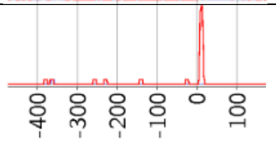
	TTGTTrev[*]			-14 … 38	BP
	AAG3			-57 … 38	min med
	ATGAA			-5 … 39	MAD high
	RDPE[*]			8 … 12	min high

Table S1. Core promoter motifs detected by XXmotif in *D.melanogaster*. Upper panel: known motifs in literature; lower panel: newly discovered motif candidates. "Distribution" depicts the distribution of all assigned motifs within the most enriched gene set ("Gene Set", details in Table S2) smoothed over five nucleotides. "Start Range" is the region that the motif's 1st nucleotide most locates relative to the defined TSS. Adapted from (Hartmann, 2012). Motifs with "*" are the ones used in this thesis.

Gene Set	Definition adapted from (Hartmann, 2012)
NP	TSS cluster width: narrow peak
BP	TSS cluster width: broad peak
MAD low	Mean absolute deviation (MAD) expression: low
MAD medhigh	MAD expression: medium to high (top 40% of genes)
MAD high	MAD expression: high (top 10% of genes)
min off	Minimum gene expression: nearly no expression
min low	Minimum gene expression: low
min med	Minimum gene expression: medium
min high	Minimum gene expression: high
max low	Maximum gene expression: low
max med	Maximum gene expression: medium
max high	Maximum gene expression: high
elf low	Gene expression in embryo, larva or female: low
elf med	Gene expression in embryo, larva or female: medium
elf high	Gene expression in embryo, larva or female: high
adult low	Gene expression in adult: low
adult med	Gene expression in adult: medium
adult high	Gene expression in adult: high
stalledPol	All genes classified as stalled by (Hendrix et al., 2008)

Table S2. Gene sets defined for the core promoter motif classification based on experimentally derived genome-wide features. Adapted from (Hartmann, 2012).

Native Promoter	Sequence (5’-3’)	TSS Distribution
FBgn0003701	TGTAGTATTTGTGTTCTGATATGAAAACTAGAGATATCGATGTTATCGTTAAAGGATTTCCAGCTTTAGCATGGACGGTCACACTGGATCTCAAAATCTGGCGCAAAGCAACAAAAAAGGAAGCGTCGCAG	BP
FBgn0004878	CGGTGGCGAGAGGGTTGCACTTGGGCATGGATTTGCGCAATTTTTGCTATATTAGCCGGAGGCCGCGGAAGAGTTGAAGCAGTTTGAGCCTCGCAGCCGAACTTTGAGGATCGCTGAGACGAGACGCCGTG	NP
FBgn0010078	CTTGGTTATAATTAGGTTATTTTTTCGATATTTTGAGGTATATTTCTACGATAGATCGGCGGTCACATCGTATTTCCCTCCTTTTCGTTTTCGTTTCCGGCGAAGTAAGTATAATAAAATCTCCACGTTTT	NP
FBgn0014865	CTCAAATAAAAAGTCCCCAATCTGCGACTCGTTTGTCTGGGACTGAGCTATAAAAGCCTCACCATCTCAACGCTCAAAGCATCAATCAATTCCCGCCACCGAGCTAAGtaGCAACTTAATCTTGGAGCGAT	NP
FBgn0027597	TTTAAATAGATTTAGCTAGAAAATAGCTGACAGACACATATCGATATATCGCTGCGATAGCCACAGCTGTTCACGCCCGCAGTTTAAGCGtaGTGGCAGCCCTGGTCGGCCACCAAAAAATAAACATTGGA	BP
FBgn0030993	GAGAGAACCAGTGCGCTCTTATCACGTGAGAACGCTTTTGGGCATTCAGTTTGGCTTTTGCGGCGCTGACCGCTGGCGACAGTTTCGAATCCATAGCCGATCGGAGAGCAACGAACGTAGGCCAGAACGGA	NP
FBgn0031980	CATATCAAGTCACAACAATGAAACGAAAACCTATCGATAGCGCATGGCTTGACGGCACGCTGCCATCGCTATGTGATTTCCTTCTTTTTTCGCCTTCACGAAATCAAGtaGGTAAGCGTTTCCCGAAATCG	BP
FBgn0032518	GTATTTTTTAGGTTTTTTCGTCTGCCCGTGGCAGCCACACTAATTTGGCTCAGCTTTTCGTTTCCACTTCCGTTTTCTTTTCTTTTCGTGTTTCTACGCCAGCAAGtaGAAGTACGTTAtaGAAAAGCGTT	NP
FBgn0033081	GACACGAAATCGAGGGATGAAATTGCATGTGATGCAGCCCCTTGAGCAAACAGTGTTGGACAACAGCGCGCGGCATCGGCATTTTTTGGCGGCCATTACAACGAtaGGAAAACTGTGAAGCGTTGCGCACA	BP
FBgn0034010	AGGTATCTGAAAGTCGAGACATAGTTAAGTCCACGCTTACAGATCGGGTATATAAAGAGGCCACTTTCAGAGCGGATTTCAGTTTAATAATTTAAAGCAAAtaGAAACTACTGGTAAGCTAGCTGGGTTAT	NP
FBgn0034308	CGCCATGCGCAGCACTATCCTGCGACTAAGCCAGATCTGACGGGAATAAAGCTGAATCGGCAGCACTGCCGCAAGTATCCACTTTTTCACGGGCAACAAATTGACAAAGAAATTGTAAGATAATTTCCTGT	BP
FBgn0034642	TGGTGCCGATTATCTTATCGCCAAGTGTGGACTGCAAGTTGGGAAAACGAATACATTCATCACCCCGGTCGTTGCTCACTAACTGGGTTTTCGGTGACGCTATTACGGACACGGACCGGCTCTCACCGAAA	NP
FBgn0035754	AGTCTGGCAACCTCTCTGTTACGGTATTTTTACAACGTGGTATTAACAGCGCTCCGGAATACTATACGGTATATTTCAGCAATCGAAGAACGGCCACATTGCGGTGTGGAAAATAAACAAATTGCAATTAT	BP
FBgn0035906	ACAATCGAAATATATTCGATAAATTCACTCGTCCGGCGACTGCGACCACTTTATAAGGTACCGGAATCCCTCTATTTGTTGTCAGTCGATCGGAACTACTTTGCCACCAACATTTCACGTCTTGGGAATTA	NP
FBgn0036263	CTACGTAATATACTAACGCACTTTTAGGTATATTTTTCAAAAATAATATACTGTTCTTGGTATATTGCTCAGGAACGGTCACTCTAGAGAGCCGGCGTAAACAAAGCGATACAATTTGGTTAAATTAATTA	BP
FBgn0037328	ACACTTTCGAGCAACGGCGCGCTGTTTCACTTAACATATCGCGTTTTTGTGGCGTCTAGAGGCAGCCACACTATTTCCTTCTTTTCGCTTTCGTTTCCGGCGAAGTAAGTAAATTAAATTTCTCTGTATAT	NP
FBgn0060296	GTGTGGCCCCTGTTAGCTTTCTGTTAAATTTAAATTTCTGTAAAGTGCCCGCCACTGCGGTCGCTTTCACGGATCAGATTAGTCGTTGTCTGGATATTAACGAGGAAGGTAGTGATCGCGCATTAGTGTCA	NP
FBgn0064225	TTGGCATTTATTATTTTTATTGTAAGGTATTTTTTAGTACATTTGTTTCTTGGATCCAATAAGGCCGCACTATTTTCCTTCTTTTTGCTAGCAATTTCCGGCGAGGTtaGTGTAAAATATTTCTACTCCAC	NP

FBgn0086519	GACTTGAACTTGGGCGCCACCGGCAGAGGTAAGCTGAGCATCAGTATCATATAAAAAGCAGGCAGAAGTTCCAGTTCGATATCAGTTAGCCTTCTCAAGTCTTTCAACAATCAACtaGAAGTTCTTCGTAA	NP

Table S3. The sequences for 19 native core promoters (with TSS downstream ATGs mutated).

Motif Name	Start Position	End Position	Max Score	Threshold Score
CA-INR	-3	-3	16.28	6.5
CGpal	-100	-20	26.46	3.7
DRE	-100	-7	15.78	7.3
E-Box1	-59	32	16.66	11.6
GAGA	-100	-33	31.27	0
GAGArev	-100	1	29.40	2
INR	-2	-1	11.15	4.6
INR2	-60	20	22.42	8.1
MTEDPE	16	18	24.03	0.6
Ohler6	-100	-10	15.94	7.5
Ohler7	-72	14	19.30	7.3
RDPE	8	12	28.38	14
R-INR	-5	-5	16.77	5
TATA-Box	-35	-29	14.73	5.5
TTGTT	-32	40	15.94	1.6
TTGTTrev	-14	38	18.56	2.2

Table S4. A summary of XXmotif-annotated core promoter motif features used in this thesis. Motif 1st nucleotide locates within the range between “Start Position” and “End Position” relative to the defined TSS (also shown in the column “Start Range” in Table S1). “Max Score” is the PWM score of the motif consensus. “Threshold Score” is the minimal score that maximizes the mutual information between a certain motif and all positively correlated gene sets.

Block 1	Sequence (5’-3’)
1.1 +	CGCAGTGCAGTGAATCATCCGTGGTGACCCATGGCTCTCGACTTACAGAGCGGCTCTTGGTGTTTCCCCGGTCGTAGATACACTACACTGAACGAAAATTTACGAGCCGATGCATTTACATTCCATTCCATTACATTCTCTTATATGGCATGTGCTCAATTGCTGTGGAGGATGTACGGACTAGAGATCGCCTCTTTCAGTGGCGGCACACTTGGTTCGTGGTGTTTCAGTGCAGTGCT
1.2 –	TTTCCAGTGTAGAAACGCGTTTATAGATGTTAAATATTAACCACTGATAAGTACAAGCTAATAACAACAATAATCGTAACAACGCGCTCTGGTTTTCTCCGTGTGCACTCAAGAGCGTTGCTGATTGAAGCGGATGAAGCCGAAGCCGATCGGAGTTGGTCAATTTTCTAAACTCTCTCACGGTCTTCAATTGAACGGCACTTCCTCGACTTCCTCCCGTCGCCCCCGCCCTTTCACAC
1.3 –	CGCTCCAAGCTAGACTCAAGAGAGATACGCACCGGAGATACGCAAACCGGTCGTTGGCTGGCTAGCCGTAGCCAAATATTATGCTAATTCGACATTTTTGACAAAAATAATTCGAGAATTAATTATATCTTTACAATGTGTCTGCTAGTTGACACATAGTTAGTTAATGTCTCCGCGTCAAACTCGTCTTCCGTCTGTTTCGGGCATTATTATGGGATTCAATGCGAACTTTAACTGAA
1.4 +	TGTAAAAATTTTATTTTTGGAAAATCAAAAAACTAGTGATAGGGATAGTTAGGTATGTTTATTAGCAGTACAAAACAGTCTTTATTTTCGCTGTGCCACCTTTTTTGGCCAAGTTTTGGAGAAAACCCCGTACGGGCATATCAGATTTTTAGCATCATCTTTTGCAGCGCTGGAGGAAGCCAGAGTTTTCACTCACCAATAAAAAAATGTAACTTAGTAACTTTAGCCAGATTTTCCGTT
1.5 –	TGCTGCAAGTTGCTTGCTGGTTTGTATGTTTTAAGAGTGAAACTGACGGGAGACGCGAACGCGAGGATGCAACAGAGTATTGTAATCTGCCATATTTGAGAAGGTTTGTTAGGTTTGATTTGGCTAATCAACAGAAGAGTTTTATGCTAAAATTGTAACTGCAATTGTAACACACGAGAAATAAACAAGCAAATAATACCTACTAGAGATAACGCTGCGAATTTGTTTTATTTTTGAACT
1.6 –	CAGTTGGCGCGATAGCAGGACTTATCAATTGAATAACAGGACCTTATTATCGACAAAACCTTAGAGCTGCCACGCAATTTAATAAAGTTATCGTTCGTAAAGTTATGTCGAATTTAGATTTAAATTGAAATTGATGAGGTAATATATTTTTAAAATAAAATCCTATCACTTATTGTTGCTTAAACTAAAATTTGTTTCAAGAAAGACTATTATGAGATAGATCTTCGACTAAAAATAAC
1.7 +	TCCCCTCTGCACCATCGTAAATATACGACTTTTATTTATTACTTCATTTTATTTTCATTATTACTATCTTTGAATAAAGAAATTTCAGAATACCAAACAAAATGATTTTCGTTTCATTTTGAATTAAAATTTCACCAGTGGGGGAAAATAACGGTATTTGAACAATAATGGCGTTGCGAATGTAACGTGATTGGCGTAAAATCGCAAATTCCGTTTTATATAGTTGAATGATTTTCAAA
1.8 –	GCTGAAATTCCGATTTTAGGCCGCTGTCCAGTTTACAATTAATGAGATTAATCGTAAATAAACATTTACGGAGAGCACTCTTACTTCTGACACACACGATAATTTGTGGCAGACACATGGAGAATGAAATTTTTTATGAGGAAATCGAATTGCAATCTCCGGCGAACGATTTTGCTCATGCAATTGCAATTACAATTGGTTTTCAATTGTATTTCCTCAAAATCAAAAGAATGTCGAAT
1.9 –	ACCGCAGTTTCCAGCTGGCTTGAAATTTTGGCCGGCGACTCCTCCTCGTTGGCAGTTTTTAGAAAACGAATCGCCTGTGCAGAGGCCAAGGCTCCGGCCGTAAAAGAGGTCAGCAACTGAGCAATCGAAATCATTTGAAATTTGATATTTTTAAATGATTGGAAAGCGAATTTAGATCATGCTTTGAAAGTTCGTTACGATTGCTATGAATTGAGTTTAAATATTTCAAGGCTACATAA
1.10 –	GTGGGCGTGGTGCCCGGCAAGGTGTTGTACTGCCAGTGCGGGGCCCCCAATTGCCGCCTTCGTCTGCTCTAAGTTCTAGCTTAAGTTAGAGATCCATACAGGAAATATACTCATTAAAGAAGAATTAGAATTAAAAATTTAAACTTTAAACATTATCTGTTCCGTTAGGGTAACGGAAACATTGCATTTTTATAAGCTACTGCTGTTCTGCACCGTCCGTTTCAAAGTACACAATTTTC
1.11 –	TTGGGGGAACAGCCTGAAAGTAGGCTACAAAACTCTTGTCTTCAGTACTCTTTATCGTGATTCCCACGACGACTTTCTTGCTTTTACATCATAAACTCAATGTGGTAATAAATTTAAAAATAGTTATACTTTTCTGTCATATTCACCAAAAGCTGGAAATTTGTATTAATTTTAATGTTGATATGAGGTCAAACGCATAAAATAAATGTATAAAAGATGTTTTGCTTACTCCGAAATCA
1.12 +	AATTCGGGCCTGCCCATTCAGACAGCCAGTGACTTGGATGATGCCGCCCACAAGGCTGTGGCAGCCCTTAATTAGGGGAACGATTGAGGAGAGCATGTCTTCCAGAATGAAACGACGCTCATTAGCATTTACAACGGTTGGGCCTTTTAAGTATAAGTTTTTATCGACAATATAACCAAAATATGTTATATTCTATATAAAAACCTTTTTATTTGATTTAAGAAGTACCTTAGCCATCT

Table S5. The sequences of block 1s used in the thesis. “+” in column “Block 1” represents well-positioned -1 nucleosome pattern found in the genome-wide MNase digestion of chromatin (unpublished data generated in our lab); “–” represents not well-positioned -1 nucleosome pattern.

Block 7	Sequence (5’-3’)
7.1 –	AGCATAGGAGCCGCACCAGGATTCGCCCGTGTACGCCAACTACGAAGATTAGCGCAACTCTGGGCCGGCCAGCTCCACGGCGTACTAGGTCAACTAGGGCGCCGCCGGCTAGGCACCCGAGCCGGCACTTCGAGTGCCCGGCACCACGCAGCAGTATCGTGGATTCAAAACTTGTAAGCGGGCGGAAACGGTGTATACTATAGTAGCAGCAAGTAGTAGACCCATCCCGGCCAATATTAG
7.2 –	GTGAAGTGCATTGAAAGCAGAGCGAATCAGAACGAAATTCAAAACGTATCACGCATACGCCCGGTAGTACAGCCGAATATCCCCAAATAGCCGAATTAGTCGCCGAGAAGGAGCTTGCTGTGCTGCCTGCTGGTGAGCAGCATCCTAGTGCTGCACCAGGCTCAGGCCAACATCGAGACAAACGTAGTGATAGACCCCAGTTACTACAGTAAGTGTGCGTAGTGGATCAGGCGGCCATAG
7.3 +	CGCAAATTGATCCTTTATAGATTCTAGCTGATTTGTTATTTACAATCCTTGTAGATTGCTTTCGCCTGCCTGTTGGCCGTCGCCCTCGCCAACGAAGTAGCCATAGTTCTCCGTGCCGAGCAGCAAGTGATAGTTGACGGCTTTGCTTACGCTGTTGAGCTGGACAACTCTGTCATAGTGCAACAGAAGGGTGACCTTAACGGCGAAGAGTGGGTGGTGAAGGGAAGCCAGTCGTGGACA
7.4 +	GGATTGATAGGCGTTATCAGTCGAAATTGAAAGGGTATAGGACCAGGGCAACTGCCTGTAGCCCGACATCAATATCTGCCAAAGCGACTTGGCCAATCCCACCGAGCCCATTGTCACCAAGATCTAGGTGCACTATCTGCGGAGTTTCGGCTTTCGCCTGGAGCCGCCCTATAAGATTGGCACCGAACTCGGTCACTCGTCGCGGGAGGCGCGCGTCTTTCTTATCCGAGTGTGCCGCCA
7.5 +	AAACGATTTGCCGTGATATAGTCGTTTCATTTGAGCACAAAAGAGCGCCTTCTGCTCCTGATCGACGACATTGAGTAGATTGCCAAGGAGTTGATCGAGCAGGCGCACCAGAAGATCTCCAGCACCGAATTGGTGGACCTGTTGGATTTGCTGGTGGCCAAAGTAGAGGAATTTCGCAAATAGCTGGAACTGGCGGAGGAGCAGGCGAAAGTGGAGGAGGCGTAGGACCAACTGCGCGCT
7.6 +	AAAAAATAATAAAATAGTTCACTTAAAATCCATAGCCACCTAGAATTTAGCTTGCGGGAAGCCTGCATAGGAGTTTTGCCTCCGCCGCCAAGGCCGCAAAAGCAAAGCCCGTCAAAACCAGCACCATTTCAGCTGTCGGCGAGTCCATCGCCGCCAAGGGGTAAGCACTTTCTTTAATTTATTTACATAAAAACCAAGTAAATATTCTCCTGATTGTAGCTTCTTGCGTCCCCATAAGCC
7.7 +	CTAAAACACCGACTAAGGACCTCTGAATAGGACTCTACTGCCACTCGCACGCTGCCCTTTGTGTACAAGTATAAAAATCATAAAGGCCAAGACTGCGAATCCCGGCTGGACATCACTTTTCCCTTTGACCAGGAAAATCTGGAGGAGTGTATCACCCAGCTGTAGGCTCCCAACCAGTAGGACCCTTAGTAGCGGTACCTGGACGAGAATATCAGTGAGTATCAGTTAGTGCGGTAGGAG
7.8 +	AAATTAGAAACAAAAGGATCTGAAACGCGCTTACAACATAGTCCTAGCACAGGCTATATCCGGCTCTAAACAGTACCTATTTGCGGGAAATCTCTTTGGCGACATTTTCGTACTTAGGTGAGAATCCTTAATTTGGCGCAATTTCCCAACTAATTGTTTTTCTATTTTGAAGAATAAAAGAACTGGACAAGGGTTCCGAGGAGCCACCTGGCAAACTGAAGATCTTCCCGCAGGGCAGCG
7.9 +	AACTTTATAGAGTTTTATCCAATTTGAGCGTAGCTGCCGGGCGTTGGAGTGTTCGGAACGGGAGAGATAGCCATAGTGCTGGTGCCGCTGCTGCGGGAGAAAGGATTCGAGGTGCGGGCCATTTGGGGCAGGACCCTGAAAGAGGCGAAAGAGACTGCGACCACGCAGATAGTACAATTCCATACGAACGTAATCGACGTAGTCCTGCTGCGGAAGGTAGTGGATCTGGTGTTCATCGTG
7.10 +	ATTCGAGAGTAGCTAAATTAGGGGCCGCATAGCGCCAAGCTCACCTGTAAGTGGCTCCAGCAAGTGTCTCGTCGCAGGTGCACAGGTGCTAGGCAGCAGGACTCCGCCGACGGCTGACTAAGAGCAGCTTATCCGCCCCTTGGGAAGAAAAAATCGATTTTGTGATAACGGGCTTACGCTTACGCGTTTTTGCACTGCTAGTCGGAACCAATTGCCAGTAGACTCCACACCTAGATAGTG
7.11 +	CGATTGCACACGTTGCACTTTTGTTGGCATTGGACGAGGTCGAGTGAAAGCACAGCGCAAGCCCCGAGAATCCCGATTCGTTTGCTTAGTTCAGGGTCAGGCCGCTCCTCCCCAGGATACGAAGCTCCCTGCAAACGACCTTGAAACTCCAAGCAGATACATCGCAATCCGAATCCGAATCGTTCCGTCGGAACCTACGACTCCTACACAACAGTTGTCGTAGTCGCCCTGTGAGCAAGT
7.12 +	TAGAATTTCCCACTCAAAGGTAGCAAAAACCATCCCACTTCGCACGCGAGGATTCCCAGGAGCACCTGCCCACCAGCCACACGCAAAACAGTCACACAGTGAACTAGAAGAGACGCACACTATCCGGAAGTTGTGGCGTTGGCGTCTTCGTGTTCGCCTTTGCCTTCATCGTGATTGCGTTTGCAACGCCCAGTTGGTTGGTCAGTGATTACCGCATCACGGGCGCCAAGCTGGATCGCC

Table S6. The sequences of block 7s used in the thesis. “+” in column “Block 7” represents well-positioned +1 nucleosome pattern found in the genome-wide MNase digestion of chromatin (unpublished data generated in our lab); “–” represents not well-positioned +1 nucleosome pattern.

Primer ID	Sequence
F1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACCCTGACCTACTCAAGGCATACATGAAGT
F2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGTATGACCTACTCAAGGCATACATGAAGT
F3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAGTTGACCTACTCAAGGCATACATGAAGT
F4	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTAGTGACCTACTCAAGGCATACATGAAGT
F5	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGGGTGACCTACTCAAGGCATACATGAAGT
F6	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCATTGACCTACTCAAGGCATACATGAAGT
F7	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTCATGACCTACTCAAGGCATACATGAAGT
F8	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTATCTTGACCTACTCAAGGCATACATGAAGT
R1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACCCTCTTCCATGGTGGCTTTACCAAC
R2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGTATCTTCCATGGTGGCTTTACCAAC
R3	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGAGTTCTTCCATGGTGGCTTTACCAAC
R4	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTAGTCTTCCATGGTGGCTTTACCAAC
R5	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGGGGTCTTCCATGGTGGCTTTACCAAC
R6	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCATCTCTTCCATGGTGGCTTTACCAAC
R7	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTCATTCTTCCATGGTGGCTTTACCAAC
R8	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTATCGTCTTCCATGGTGGCTTTACCAAC
R9	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAAAAGATCTTCCATGGTGGCTTTACCAAC
R10	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTGCTGTCTTCCATGGTGGCTTTACCAAC
R11	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCTGCTTCTTCCATGGTGGCTTTACCAAC
R12	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGCTTTTCTTCCATGGTGGCTTTACCAAC

Table S7. The primers used in 1st PCR for sequencing library preparation. The underscore marks the specific barcode for each primer sequence.

	1	2	3	4	5	6	7	8	9	10	11	12
A	F1/R1	F1/R2	F1/R3	F1/R4	F1/R5	F1/R6	F1/R7	F1/R8	F1/R9	F1/R10	F1/R11	F1/R12
B	F2/R1	F2/R2	F2/R3	F2/R4	F2/R5	F2/R6	F2/R7	F2/R8	F2/R9	F2/R10	F2/R11	F2/R12
C	F3/R1	F3/R2	F3/R3	F3/R4	F3/R5	F3/R6	F3/R7	F3/R8	F3/R9	F3/R10	F3/R11	F3/R12
D	F4/R1	F4/R2	F4/R3	F4/R4	F4/R5	F4/R6	F4/R7	F4/R8	F4/R9	F4/R10	F4/R11	F4/R12
E	F5/R1	F5/R2	F5/R3	F5/R4	F5/R5	F5/R6	F5/R7	F5/R8	F5/R9	F5/R10	F5/R11	F5/R12
F	F6/R1	F6/R2	F6/R3	F6/R4	F6/R5	F6/R6	F6/R7	F6/R8	F6/R9	F6/R10	F6/R11	F6/R12
G	F7/R1	F7/R2	F7/R3	F7/R4	F7/R5	F7/R6	F7/R7	F7/R8	F7/R9	F7/R10	F7/R11	F7/R12
H	F8/R1	F8/R2	F8/R3	F8/R4	F8/R5	F8/R6	F8/R7	F8/R8	F8/R9	F8/R10	F8/R11	F8/R12

Table S8. The primer scheme in a 96-well plate for 1st PCR in sequencing library preparation.

INR:

	1	2	3	4	5	6	7
A	0.178	0.044	0.896	0.074	0.164	0.252	0.169
C	0.118	0.662	0.025	0.118	0.031	0.272	0.201
G	0.170	0.051	0.032	0.653	0.013	0.110	0.368
T	0.534	0.244	0.046	0.156	0.793	0.366	0.261

TATA-Box:

	1	2	3	4	5	6	7	8	9	10
A	0.221	0.067	0.905	0.022	0.909	0.455	0.819	0.562	0.149	0.194
C	0.345	0.069	0.020	0.044	0.020	0.012	0.027	0.015	0.230	0.276
G	0.335	0.066	0.019	0.022	0.023	0.043	0.118	0.150	0.510	0.424
T	0.098	0.798	0.055	0.911	0.048	0.489	0.037	0.273	0.112	0.105

INR2:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	0.195	0.215	0.247	0.286	0.155	0.225	0.094	0.039	0.297	0.581	0.023	0.440	0.017	0.188	0.252
C	0.243	0.301	0.000	0.158	0.322	0.053	0.020	0.292	0.487	0.229	0.483	0.170	0.479	0.270	0.173
G	0.267	0.268	0.366	0.353	0.297	0.586	0.611	0.287	0.020	0.190	0.246	0.138	0.016	0.176	0.366
T	0.295	0.216	0.387	0.203	0.227	0.136	0.276	0.382	0.196	0.000	0.248	0.252	0.488	0.367	0.209

DRE:

	1	2	3	4	5	6	7	8	9	10
A	0.190	0.801	0.039	0.025	0.019	0.685	0.170	0.528	0.236	0.233
C	0.096	0.053	0.075	0.941	0.017	0.094	0.106	0.058	0.197	0.299
G	0.031	0.044	0.045	0.013	0.948	0.132	0.092	0.175	0.319	0.165
T	0.683	0.102	0.841	0.021	0.016	0.090	0.631	0.240	0.248	0.304

Ohler7:

	1	2	3	4	5	6	7	8	9	10	11	12	13
A	0.158	0.212	0.139	0.121	0.505	0.070	0.190	0.310	0.075	0.120	0.486	0.230	0.309
C	0.150	0.247	0.411	0.630	0.167	0.316	0.677	0.185	0.629	0.141	0.175	0.205	0.000
G	0.251	0.296	0.071	0.046	0.183	0.158	0.051	0.176	0.086	0.156	0.175	0.331	0.248
T	0.441	0.246	0.379	0.203	0.145	0.457	0.083	0.329	0.211	0.583	0.164	0.233	0.443

Table S9. The expression-based PPMs of INR, TATA-Box, INR2, DRE and Ohler7 derived from measurements.

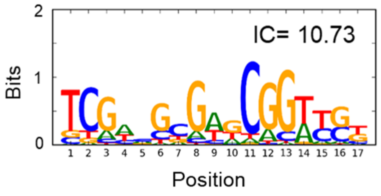
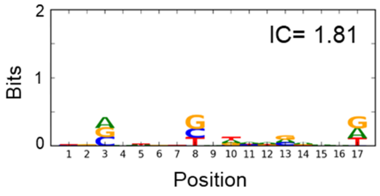
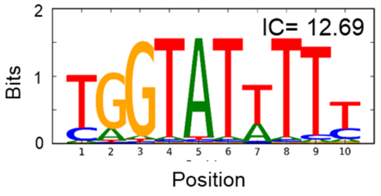
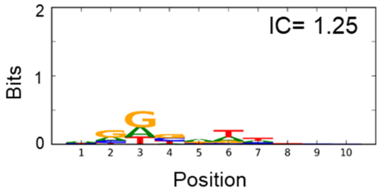
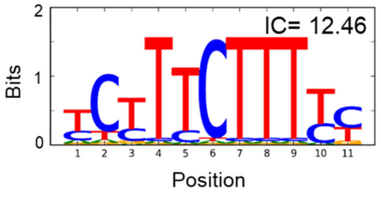
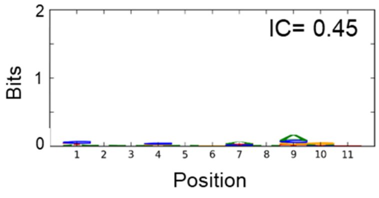
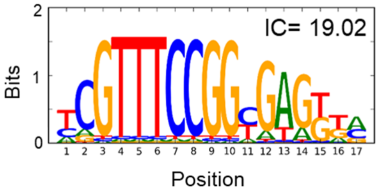
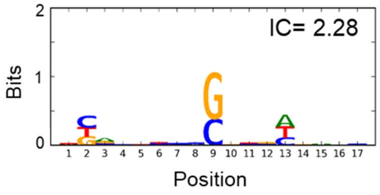
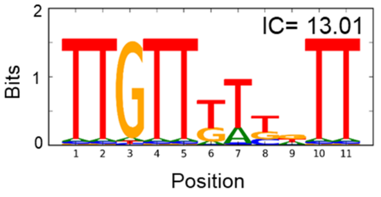
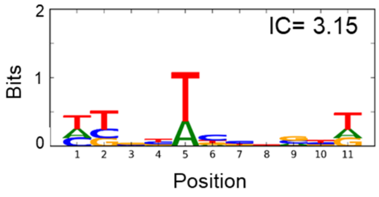
Core promoter motif	XXmotif logos	Expression-based activity logos
MTEDPE		
Ohler6		
R-INR		
RDPE		
TTGTT		

Table S10. Comparison of the XXmotif logos with the expression-based activity logos for MTEDPE, Ohler6, R-INR, RDPE and TTGTT. IC, information content.

FBgn0004878:

Single Mutation	Coefficient	Standard Error	t-statistic	*p*-value	Significance
(Intercept)	-1.88451	0.2978	-6.328	6.57×10^{-10}	***
score_15.1_CGpal	0.25858	0.18557	1.393	0.164256	
ko_-22.4_CGpal	0.33965	0.27234	1.247	0.21306	
con_CGpal	0.32786	0.17473	1.876	0.061329	.
score_7.8_INR	0.12933	0.17047	0.759	0.448465	
ko_-8.7_INR	-2.63238	0.27775	-9.477	$< 2\times10^{-16}$	***
con_INR	0.64404	0.18146	3.549	0.000432	***
score_12.3_MTEDPE	-0.16984	0.19076	-0.89	0.373805	
ko_-21.4_MTEDPE	-1.30418	0.24392	-5.347	1.5×10^{-7}	***
con_MTEDPE	0.91691	0.16561	5.537	5.54×10^{-8}	***
score_10.1_TATA-Box	1.51854	0.17423	8.716	$< 2\times10^{-16}$	***
ko_-13.3_TATA-Box	-1.76603	0.23251	-7.596	2.13×10^{-13}	***
con_TATA-Box	1.62442	0.17715	9.17	$< 2\times10^{-16}$	***
CGpal_shift1	0.13803	0.16998	0.812	0.417249	
CGpal_shift5	0.07406	0.17582	0.421	0.673833	
INR_shift-1	-0.13374	0.17893	-0.747	0.455216	
INR_shift1	-0.07889	0.17317	-0.456	0.64896	
MTEDPE_shift1	-0.45064	0.1806	-2.495	0.012986	*
MTEDPE_shift-1	-0.41735	0.18498	-2.256	0.024592	*
TATA-Box_shift-1	-0.22232	0.14606	-1.522	0.128764	

FBgn0034010:

Single Mutation	Coefficient	Standard Error	t-statistic	*p*-value	Significance
(Intercept)	-1.240392	0.331757	-3.739	0.000292	***
score_7.8_INR	-0.595596	0.226518	-2.629	0.009746	**
ko_-8.7_INR	-1.888393	0.331926	-5.689	1.02×10^{-7}	***
con_INR	0.004973	0.215506	0.023	0.981632	
score_10.1_TATA-Box	-0.408703	0.204139	-2.002	0.047671	*
ko_-13.3_TATA-Box	-3.430717	0.311903	-10.999	$< 2\times10^{-16}$	***
con_TATA-Box	0.026682	0.224845	0.119	0.905749	
score_10.4_TTGTTrev	-1.073766	0.245268	-4.378	2.69×10^{-5}	***
ko_-20.3_TTGTTrev	-0.18831	0.228447	-0.824	0.411504	
con_TTGTTrev	-1.791015	0.235393	-7.609	8.97×10^{-12}	***
INR_shift-1	-0.030037	0.225055	-0.133	0.894062	
INR_shift1	-0.195062	0.225828	-0.864	0.389547	
TATA-Box_shift-1	0.308071	0.223569	1.378	0.170936	
TATA-Box_shift1	-1.516757	0.227746	-6.66	1.04×10^{-9}	***

FBgn0036263:

Single Mutation	Coefficient	Standard Error	t-statistic	*p*-value	Significance
(Intercept)	0.5963	0.2273	2.624	0.00935	**
score_15.3_INR2	-2.6957	0.1595	-16.899	$< 2\times10^{-16}$	***
ko_-15.5_INR2	-3.5522	0.2399	-14.807	$< 2\times10^{-16}$	***
con_INR2	0.9173	0.1556	5.896	1.53×10^{-8}	***
ko_-13.1_Ohler6	-1.1768	0.2105	-5.591	7.19×10^{-8}	***
score_11.7_Ohler6	-0.3652	0.2575	-1.418	0.15765	
con_Ohler6	0.6168	0.1406	4.386	1.85×10^{-5}	***
score_8.8_TTGTT	0.2197	0.1666	1.319	0.1886	
ko_-17.3_TTGTT	0.261	0.1756	1.486	0.13879	
con_TTGTT	0.3307	0.1774	1.863	0.06383	.
score_10.4_TTGTTrev	-0.9963	0.1749	-5.697	4.23×10^{-8}	***
ko_-20.3_TTGTTrev	-1.2938	0.1603	-8.072	5.88×10^{-14}	***
con_TTGTTrev	-0.9978	0.2338	-4.268	3.02×10^{-5}	***
INR2_shift0	-0.3302	0.1582	-2.087	0.03814	*
INR2_shift-1	-0.7968	0.1572	-5.069	8.96×10^{-7}	***
INR2_shift1	NA	NA	NA	NA	
Ohler6_shift-5	0.2577	0.1604	1.606	0.10975	
Ohler6_shift0	-0.1829	0.1668	-1.096	0.27429	
Ohler6_shift1	NA	NA	NA	NA	

FBgn0035754:

Single Mutation	Coefficient	Standard Error	t-statistic	*p*-value	Significance
(Intercept)	5.055516	0.355289	14.229	$< 2\times10^{-16}$	***
score_15.3_INR2	-4.833535	0.241413	-20.022	$< 2\times10^{-16}$	***
ko_-15.5_INR2	-6.124531	0.466532	-13.128	$< 2\times10^{-16}$	***
con_INR2	0.345971	0.236483	1.463	0.1462	
score_11.7_Ohler6	-0.449082	0.233627	-1.922	0.057053	.
ko_-13.1_Ohler6	-1.47371	0.328919	-4.48	1.77×10^{-5}	***
con_Ohler6	0.08765	0.231849	0.378	0.706092	
score_10.4_TTGTTrev	0.149448	0.259499	0.576	0.565801	
ko_-20.3_TTGTTrev	0.000769	0.260306	0.003	0.997648	
con_TTGTTrev	0.220648	0.255184	0.865	0.389024	
INR2_shift-1	-0.776613	0.220134	-3.528	0.000603	***
INR2_shift1	-0.81695	0.233372	-3.501	0.000661	***
Ohler6_shift-5	-0.262869	0.228427	-1.151	0.252211	
Ohler6_shift5	-0.039848	0.231968	-0.172	0.86391	

FBgn0031980N:

Single Mutation	Coefficient	Standard Error	t-statistic	*p*-value	Significance
(Intercept)	0.19013	0.32808	0.58	0.5629	
score_11.5_DRE	-1.32081	0.19134	-6.903	6.74×10^{-11}	***
ko_-11.2_DRE	-3.80757	0.29779	-12.786	$< 2\times10^{-16}$	***
con_DRE	-0.07289	0.21822	-0.334	0.7387	
score_13.3_Ohler7	0.31399	0.22988	1.366	0.1735	
ko_-14.2_Ohler7	-1.60141	0.23023	-6.956	5×10^{-11}	***
con_Ohler7	2.47195	0.22834	10.826	$< 2\times10^{-16}$	***
score_8.8_TTGTT	0.23924	0.23521	1.017	0.3103	
ko_-17.3_TTGTT	-0.12002	0.23812	-0.504	0.6148	
con_TTGTT	0.62655	0.2411	2.599	0.0101	*
CGpal-Ohler7_shift-1	-2.77379	0.15633	-17.744	$< 2\times10^{-16}$	***
CGpal-Ohler7_shift0	NA	NA	NA	NA	
DRE_shift-5	-0.45879	0.20606	-2.227	0.0271	*
DRE_shift5	-0.84371	0.2071	-4.074	6.68×10^{-5}	***
DRE_shift0	NA	NA	NA	NA	

FBgn0064225:

Single Mutation	Coefficient	Standard Error	t-statistic	*p*-value	Significance
(Intercept)	-2.590867	0.133745	-19.372	$< 2\times10^{-16}$	***
score_11.7_Ohler6	-1.035318	0.11917	-8.688	4.47×10^{-16}	***
ko_-13.1_Ohler6	-2.725307	0.168052	-16.217	$< 2\times10^{-16}$	***
con_Ohler6	0.437259	0.117426	3.724	0.000241	***
score_10.9_R.INR	0.012447	0.131273	0.095	0.924533	
ko_-16.6_R.INR	-0.050101	0.124973	-0.401	0.688831	
con_R.INR	0.257863	0.125083	2.062	0.040262	*
score_21.2_RDPE	0.2011	0.118865	1.692	0.091896	.
ko_-23.6_RDPE	-0.587706	0.182573	-3.219	0.001452	**
con_RDPE	0.08332	0.115376	0.722	0.470853	
Ohler6_shift-5	-1.673872	0.115815	-14.453	$< 2\times10^{-16}$	***
Ohler6_shift5	-2.103073	0.120654	-17.431	$< 2\times10^{-16}$	***
RDPE_shift-1	0.004674	0.116335	0.04	0.967985	
RDPE_shift1	-0.218288	0.119299	-1.83	0.068451	.
TTGTT-R-INR_shift-1	0.04224	0.1118	0.378	0.705879	
TTGTT-R-INR_shift1	-0.147618	0.112453	-1.313	0.190458	

Table S11. Coefficients of single mutations learned from the linear regression models for the intra-architectural combinatorial mutations in FBgn0004878, FBgn0034010, FBgn0036263, FBgn0035754, FBgn0031980N and FBgn0064225. Significance codes: .p ≤ 0.1; *p ≤ 0.05; **p ≤ 0.01; ***p ≤ 0.001. NA: not defined because of singularities.

Block	Coefficient	Standard Error	t-statistic	*p*-value	Significance
(Intercept)	-2.65885	1.36776	-1.944	0.053117	.
Block3_FBgn0031980N	-1.57372	0.61621	-2.554	0.011297	*
Block3_FBgn0034010	-2.73639	0.62579	-4.373	1.86×10^{-5}	***
Block3_FBgn0064225	0.15807	0.61497	0.257	0.797373	
Block3_FBgn0086519	-2.20657	0.61994	-3.559	0.000451	***
Block4_FBgn0004878	0.09943	0.26785	0.371	0.710835	
Block4_FBgn0031980N	1.75593	0.27639	6.353	1.11×10^{-9}	***
Block4_FBgn0034010	1.88443	0.27366	6.886	$5.37\times10\text{-}^{11}$	***
Block4_FBgn0036263	0.19417	0.26369	0.736	0.462259	
Block4_FBgn0064225	NA	NA	NA	NA	
Block5_FBgn0031980N	0.35498	1.4742	0.241	0.809927	
Block5_FBgn0034010	0.38067	1.47562	0.258	0.796657	
Block5_FBgn0036263	2.07272	1.47939	1.401	0.162536	
Block5_FBgn0064225	-0.35306	1.48039	-0.238	0.81171	
Block6_FBgn0004878	1.97382	0.37046	5.328	2.35×10^{-7}	***
Block6_FBgn0014865	NA	NA	NA	NA	
Block6_FBgn0031980N	-0.05506	0.25839	-0.213	0.831442	
Block6_FBgn0034010	-0.70079	0.26107	-2.684	0.007793	**
Block6_FBgn0036263	-0.07401	0.29249	-0.253	0.800463	
Block6_FBgn0064225	-0.14806	0.26584	-0.557	0.578106	

Table S12. Coefficients of the linear regression model for the inter-architectural block-wise combinatorial mutations. Significance codes: .p ≤ 0.1; *p ≤ 0.05; **p ≤ 0.01; ***p ≤ 0.001. NA: not defined because of singularities.

APPENDIX C Abbreviations

5’ SAGE	5’ serial analysis of gene expression
AEF	adult enhancer factor
Ar.	architecture
bHLH-zip	basic helix-loop-helix leucine zipper
BP	broad peak
CAGE	cap analysis of gene expression
ChIP-seq	chromatin immunoprecipitation sequencing
CRE	cis-regulatory element
CV	coefficient of variation
D. melanogaster	*Drosophila melanogaster*
DamID	DNA adenine methyltransferase identification
DPE	downstream promoter element
DRE	DNA replication-related element
DSIF	DRB sensitivity-inducing factor
EcR	ecdysone receptor
EcRE	ecdysone response element
ENCODE	encyclopedia of DNA elements
FACS	fluorescence-activated cell sorting
FBS	fetal bovine serum
FXR	farnesoid X receptor
GFP	green fluorescent protein
GRO-seq	global run-on sequencing
GTF	general transcription factor
H3K27ac	histone H3 lysine 27 acetylation
H3K4me3	tri-methylation of histone H3 lysine 4
H3K9ac	histone H3 lysine 9 acetylation

IC	information content
INR	initiator
LXR	liver X receptor
MAD	mean absolute deviation
modENCODE	model organism ENCODE
MPRA	massively parallel reporter assay
MTE	motif ten element
NDR	nucleosome-depleted region
NELF	negative elongation factor
NP	narrow peak
OSC	ovarian somatic cell
PB	pause button
PCC	Pearson correlation coefficient
PIC	pre-initiation complex
Pol II	RNA polymerase II
PPM	position probability matrix
PWM	position weight matrix
RDPE	ribosomal downstream promoter element
RLU	relative light unit
RNA-seq	RNA sequencing
RT	room temperature
RXR	retinoid X receptor
S2 cell	Schneider 2 cell
SD	standard deviation
STAP-seq	self-transcribing active core promoter sequencing
STARR-seq	self-transcribing active regulatory region sequencing
SuRE	survey of regulatory elements
TAF	TBP-associated factor
TBP	TATA-Box-binding protein

TF	transcription factor
TRF2	TBP-related factor 2
TSS	transcription start site
USP	ultraspiracle
UTC	untreated cell
UTR	untranslated region
XXmotif	eXhaustive evaluation of matriX motifs
YFP	yellow fluorescence protein

BIBLIOGRAPHY

Abeel, T., Saeys, Y., Bonnet, E., Rouzé, P., & Van de Peer, Y. (2008). Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Research*, *18*(2), 310–323.

Ahsan, B., Saito, T. L., Hashimoto, S.-I., Muramatsu, K., Tsuda, M., Sasaki, A., … Morishita, S. (2009). MachiBase: a Drosophila melanogaster 5'-end mRNA transcription database. *Nucleic Acids Research*, *37*(Database issue), D49-53.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., … Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, *507*(7493), 455–461.

Arnold, C. D., Gerlach, D., Spies, D., Matts, J. A., Sytnikova, Y. A., Pagani, M., … Stark, A. (2014). Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nature Genetics*, *46*(7), 685–692.

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., & Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, *339*(6123), 1074–1077.

Arnold, C. D., Zabidi, M. A., Pagani, M., Rath, M., Schernhuber, K., Kazmar, T., & Stark, A. (2017). Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. *Nature Biotechnology*, *35*(2), 136–144.

Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, *27*(2), 299–308.

Baptista, T., Grünberg, S., Minoungou, N., Koster, M. J. E., Timmers, H. T. M., Hahn, S., … Tora, L. (2017). SAGA is a general cofactor for RNA polymerase II transcription. *Molecular Cell*, *68*(1), 130–143.e5.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., … Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, *129*(4), 823–837.

Baumann, D. G., & Gilmour, D. S. (2017). A sequence-specific core promoter-binding transcription factor recruits TRF2 to coordinately transcribe ribosomal protein genes.

Nucleic Acids Research, *45*(18), 10481–10491.

Beer, M. A., & Tavazoie, S. (2004). Predicting gene expression from sequence. *Cell*, *117*(2), 185–198.

Brodu, V., Mugat, B., Fichelson, P., Lepesant, J. a, & Antoniewski, C. (2001). A UAS site substitution approach to the in vivo dissection of promoters: interplay between the GATAb activator and the AEF-1 repressor at a Drosophila ecdysone response unit. *Development*, *128*(13), 2593–2602.

Brown, J. B., & Celniker, S. E. (2015). Lessons from modENCODE. *Annual Review of Genomics and Human Genetics*, *16*(1), 31–53.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), 1213–1218.

Burke, T. W., & Kadonaga, J. T. (1997). The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. *Genes & Development*, *11*(22), 3020–3031.

Butler, J. E., & Kadonaga, J. T. (2001). Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes & Development*, *15*(19), 2515–2519.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., … Hayashizaki, Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics*, *38*(6), 626–635.

Chen, R. A.-J., Down, T. A., Stempor, P., Chen, Q. B., Egelhofer, T. A., Hillier, L. W., … Ahringer, J. (2013). The landscape of RNA polymerase II transcription initiation in C. elegans reveals promoter and enhancer architectures. *Genome Research*, *23*(8), 1339–1347.

Cherbas, L., Lee, K., & Cherbas, P. (1991). Identification of ecdysone response elements by analysis of the Drosophila Eip28/29 gene. *Genes & Development*, *5*(1), 120–131.

Cleary, M. A., Kilian, K., Wang, Y., Bradshaw, J., Cavet, G., Ge, W., … Hannon, G. J. (2004). Production of complex nucleic acid libraries using highly parallel in situ oligonucleotide synthesis. *Nature Methods*, *1*(3), 241–248.

Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., & Lis, J. T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature Genetics*, *46*(12), 1311–1320.

Core, L. J., Waterfall, J. J., & Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, *322*(5909), 1845–1848.

Crawford, G. E., Holt, I. E., Whittle, J., Webb, B. D., Tai, D., Davis, S., … Collins, F. S. (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Research*, *16*(1), 123–131.

Deng, W., & Roberts, S. G. E. (2005). A core promoter element downstream of the TATA box that is recognized by TFIIB. *Genes & Development*, *19*(20), 2418–2423.

Dineen, D. G., Wilm, A., Cunningham, P., & Higgins, D. G. (2009). High DNA melting temperature predicts transcription start site location in human and mouse. *Nucleic Acids Research*, *37*(22), 7360–7367.

Dobens, L., Rudolph, K., & Berger, E. M. (1991). Ecdysterone regulatory elements function as both transcriptional activators and repressors. *Molecular and Cellular Biology*, *11*(4), 1846–1853.

Emami, K. H., Jain, A., & Smale, S. T. (1997). Mechanism of synergy between TATA and initiator: synergistic binding of TFIID following a putative TFIIA-induced isomerization. *Genes & Development*, *11*(22), 3007–3019.

Engler, C., Gruetzner, R., Kandzia, R., & Marillonnet, S. (2009). Golden Gate Shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PLoS ONE*, *4*(5), e5553.

Engler, C., Kandzia, R., & Marillonnet, S. (2008). A one pot, one step, precision cloning method with high throughput capability. *PLoS ONE*, *3*(11), e3647.

Engström, P. G., Sui, S. J. H., Drivenes, Ø., Becker, T. S., & Lenhard, B. (2007). Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Research*, *17*(12), 1898–1908.

Falb, D., & Maniatis, T. (1992a). A conserved regulatory unit implicated in tissue-specific gene expression in Drosophila and man. *Genes & Development*, *6*(3), 454–465.

Falb, D., & Maniatis, T. (1992b). Drosophila transcriptional repressor protein that binds specifically to negative control elements in fat body enhancers. *Molecular and Cellular Biology*, *12*(9), 4093–4103.

Fan, F., & Wood, K. V. (2007). Bioluminescent assays for high-throughput screening. *ASSAY and Drug Development Technologies*, *5*(1), 127–136.

FitzGerald, P. C., Sturgill, D., Shyakhtenko, A., Oliver, B., & Vinson, C. (2006). Comparative genomics of Drosophila and human core promoters. *Genome Biology*, *7*(7), R53.

Forrest, A. R. R., Kawaji, H., Rehli, M., Baillie, J. K., Hoon, M. J. L. de, Haberle, V., … Hayashizaki, Y. (2014). A promoter-level mammalian expression atlas. *Nature*, *507*(7493), 462–470.

Fuda, N. J., Guertin, M. J., Sharma, S., Danko, C. G., Martins, A. L., Siepel, A., & Lis, J. T. (2015). GAGA factor maintains nucleosome-free regions and has a role in RNA polymerase II recruitment to promoters. *PLOS Genetics*, *11*(3), e1005108.

Fuda, N. J., & Lis, J. T. (2013). A new player in Pol II pausing. *The EMBO Journal*, *32*(13), 1796–1798.

Gauhar, Z., Sun, L. V, Hua, S., Mason, C. E., Fuchs, F., Li, T.-R., … White, K. P. (2009). Genomic mapping of binding regions for the Ecdysone receptor protein complex. *Genome Research*, *19*(6), 1006–1013.

Gaulton, K. J., Nammo, T., Pasquali, L., Simon, J. M., Giresi, P. G., Fogarty, M. P., … Ferrer, J. (2010). A map of open chromatin in human pancreatic islets. *Nature Genetics*, *42*(3), 255–259.

Gershenzon, N. I., & Ioshikhes, I. P. (2005). Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics*, *21*(8), 1295–1300.

Gershenzon, N. I., Trifonov, E. N., & Ioshikhes, I. P. (2006). The features of Drosophila core promoters revealed by statistical analysis. *BMC Genomics*, *7*, 161.

Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T., … Waterston, R. H. (2010). Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*, *330*(6012), 1775–1787.

Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., & Lieb, J. D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research*, *17*(6), 877–885.

Goñi, J. R., Pérez, A., Torrents, D., & Orozco, M. (2007). Determining promoter location based on DNA structure first-principles calculations. *Genome Biology*, *8*(12), R263.

Goodrich, J. A., & Tjian, R. (2010). Unexpected roles for core promoter recognition factors in cell-type-specific transcription and gene regulation. *Nature Reviews Genetics*, *11*(8), 549–558.

Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., & Bulyk, M. L. (2013). Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Reports*, *3*(4), 1093–1104.

Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., … Celniker, S. E. (2011). The developmental transcriptome of Drosophila melanogaster. *Nature*, *471*(7339), 473–479.

Gressel, S., Schwalb, B., Decker, T. M., Qin, W., Leonhardt, H., Eick, D., & Cramer, P. (2017). CDK9-dependent RNA polymerase II pausing controls transcription initiation. *ELife*, *6*, 1–24.

Gu, W., Lee, H.-C., Chaves, D., Youngman, E. M., Pazour, G. J., Conte, D., & Mello, C. C. (2012). CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as C. elegans piRNA precursors. *Cell*, *151*(7), 1488–1500.

Haberle, V., Li, N., Hadzhiev, Y., Plessy, C., Previti, C., Nepal, C., … Lenhard, B. (2014). Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature*, *507*(7492), 381–385.

Hampsey, M. (1998). Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiology and Molecular Biology Reviews*, *62*(2), 465–503.

Hansen, S. K., Takada, S., Jacobson, R. H., Lis, J. T., & Tjian, R. (1997). Transcription properties of a cell type-specific TATA-binding protein, TRF. *Cell*, *91*(1), 71–83.

Hartmann, H. (2012). Regulatory motif discovery using PWMs and the architecture of eukaryotic core promoters. *PhD Thesis*, LMU München.

Hartmann, H., Guthöhrlein, E. W., Siebert, M., Luehr, S., & Söding, J. (2013). P-value-based regulatory motif discovery using positional weight matrices. *Genome Research*, *23*(1), 181–194.

Hashimoto, S., Suzuki, Y., Kasai, Y., Morohoshi, K., Yamada, T., Sese, J., … Matsushima, K. (2004). 5′-end SAGE for the analysis of transcriptional start sites. *Nature Biotechnology*, *22*(9), 1146–1149.

Hendrix, D. A., Hong, J.-W., Zeitlinger, J., Rokhsar, D. S., & Levine, M. S. (2008). Promoter elements associated with RNA Pol II stalling in the Drosophila embryo. *Proceedings of the National Academy of Sciences*, *105*(22), 7762–7767.

Hirose, F., Yamaguchi, M., Handa, H., Inomata, Y., & Matsukage, A. (1993). 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. *Journal of Biological Chemistry*, *268*(3), 2092–2099.

Hochheimer, A., Zhou, S., Zheng, S., Holmes, M. C., & Tjian, R. (2002a). TRF2 associates with DREF and directs promoter-selective gene expression in Drosophila. *Nature*, *420*(6914), 439–445.

Hochheimer, A., Zhou, S., Zheng, S., Holmes, M. C., & Tjian, R. (2002b). TRF2 associates with DREF and directs promoter-selective gene expression in Drosophila. *Nature*, *420*(6914), 439–445.

Hoskins, R. a., Landolin, J. M., Brown, J. B., Sandler, J. E., Takahashi, H., Lassmann, T., … Celniker, S. E. (2011). Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Research*, *21*(2), 182–192.

Jin, C., Zang, C., Wei, G., Cui, K., Peng, W., Zhao, K., & Felsenfeld, G. (2009). H3.3/H2A.Z double variant–containing nucleosomes mark “nucleosome-free regions” of active promoters and other regulatory regions. *Nature Genetics*, *41*(8), 941–945.

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, *316*(5830), 1497–1502.

Jones, G., Jones, D., Fang, F., Xu, Y., New, D., & Wu, W.-H. (2012). Juvenile hormone action through a defined enhancer motif to modulate ecdysteroid-activation of natural core promoters. *Comparative Biochemistry and Physiology. Part B, Biochemistry & Molecular Biology*, *161*(3), 219–225.

Juven-Gershon, T., Cheng, S., & Kadonaga, J. T. (2006). Rational design of a super core promoter that enhances gene expression. *Nature Methods*, *3*(11), 917–922.

Juven-Gershon, T., Hsu, J.-Y., Theisen, J. W., & Kadonaga, J. T. (2008). The RNA polymerase II core promoter - the gateway to transcription. *Current Opinion in Cell Biology*, *20*(3), 253–259.

Kadonaga, J. T. (2012). Perspectives on the RNA polymerase II core promoter. *Wiley Interdisciplinary Reviews: Developmental Biology*, *1*(1), 40–51.

Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., … Hardison, R. C. (2014). Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences*, *111*(17), 6131–6138.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van der Lee, R., … Mathelier, A. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research*, *46*(D1), D260–D266.

King-Jones, K., & Thummel, C. S. (2005). Nuclear receptors – a perspective from Drosophila. *Nature Reviews Genetics*, *6*(4), 311–323.

Kinney, J. B., Murugan, A., Callan, C. G., & Cox, E. C. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences*, *107*(20), 9158–9163.

Koelle, M. R., Talbot, W. S., Segraves, W. A., Bender, M. T., Cherbas, P., & Hogness, D. S. (1991). The Drosophila EcR gene encodes an ecdysone receptor, a new member of the steroid receptor superfamily. *Cell*, *67*(1), 59–77.

Kopytova, D. V, Krasnov, A. N., Kopantceva, M. R., Nabirochkina, E. N., Nikolenko, J. V, Maksimenko, O., … Georgieva, S. G. (2006). Two isoforms of Drosophila TRF2 are involved in embryonic development, premeiotic chromatin condensation, and proper differentiation of germ cells of both sexes. *Molecular and Cellular Biology*, *26*(20), 7492–7505.

Kornberg, R. D., & Lorch, Y. (1999). Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, *98*(3), 285–294.

Krebs, A. R., Imanci, D., Hoerner, L., Gaidatzis, D., Burger, L., & Schübeler, D. (2017). Genome-wide single-molecule footprinting reveals high RNA polymerase II turnover at paused promoters. *Molecular Cell*, *67*(3), 411–422.e4.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–330.

Kwak, H., Fuda, N. J., Core, L. J., & Lis, J. T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science*, *339*(6122), 950–953.

Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C., & Cohen, B. A. (2012). Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proceedings of the National Academy of Sciences*, *109*(47), 19498–19503.

Lagrange, T., Kapanidis, A. N., Tang, H., Reinberg, D., & Ebright, R. H. (1998). New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes & Development*, *12*(1), 34–44.

Laval, M., Pourrain, F., Deutsch, J., & Lepesant, J.-A. (1993). In vivo functional characterization of an ecdysone response enhancer in the proximal upstream region of the Fbp1 gene of D. melanogaster. *Mechanisms of Development*, *44*(2–3), 123–138.

Lenhard, B., Sandelin, A., & Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics*, *13*(4), 233–245.

LeProust, E. M., Peck, B. J., Spirin, K., McCuen, H. B., Moore, B., Namsaraev, E., & Caruthers, M. H. (2010). Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Research*, *38*(8), 2522–2540.

Levine, M., & Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, *424*(6945), 147–151.

Li, J., & Gilmour, D. S. (2013). Distinct mechanisms of transcriptional pausing orchestrated by GAGA factor and M1BP, a novel transcription factor. *The EMBO Journal*, *32*(13), 1829–1841.

Lifton, R. P., Goldberg, M. L., Karp, R. W., & Hogness, D. S. (1978). The organization of the histone genes in Drosophila melanogaster: functional and evolutionary implications. *Cold Spring Harbor Symposia on Quantitative Biology*, *42*, 1047–1051.

Ligr, M., Siddharthan, R., Cross, F. R., & Siggia, E. D. (2006). Gene expression from random libraries of yeast promoters. *Genetics*, *172*(4), 2113–2122.

Lim, C. Y., Santoso, B., Boulay, T., Dong, E., Ohler, U., & Kadonaga, J. T. (2004). The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes & Development*, *18*(13), 1606–1617.

Lis, J. T. (1998). Promoter-associated pausing in promoter architecture and postinitiation transcriptional regulation. *Cold Spring Harbor Symposia on Quantitative Biology*, *63*, 347–356.

Louder, R. K., He, Y., López-Blanco, J. R., Fang, J., Chacón, P., & Nogales, E. (2016). Structure of promoter-bound TFIID and model of human pre-initiation complex assembly. *Nature*, *531*(7596), 604–609.

Lubliner, S., Regev, I., Lotan-Pompan, M., Edelheit, S., Weinberger, A., & Segal, E. (2015). Core promoter sequence in yeast is a major determinant of expression level. *Genome*

Research, *25*(7), 1008–1017.

Ma, X., Zhang, K., & Li, X. (2009). Evolution of Drosophila ribosomal protein gene core promoters. *Gene*, *432*(1–2), 54–59.

Maerkl, S. J., & Quake, S. R. (2007). A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, *315*(5809), 233–237.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., … Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, *337*(6099), 1190–1195.

Mavrich, T. N., Jiang, C., Ioshikhes, I. P., Li, X., Venters, B. J., Zanton, S. J., … Pugh, B. F. (2008). Nucleosome organization in the Drosophila genome. *Nature*, *453*(7193), 358–362.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., … Mikkelsen, T. S. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology*, *30*(3), 271–277.

Missra, A., & Gilmour, D. S. (2010). Interactions between DSIF (DRB sensitivity inducing factor), NELF (negative elongation factor), and the Drosophila RNA polymerase II transcription elongation complex. *Proceedings of the National Academy of Sciences*, *107*(25), 11301–11306.

Morin, B., Nichols, L. A. & Holland, L. J. (2006). Flanking sequence composition differentially affects the binding and functional characteristics of glucocorticoid receptor homo- and heterodimers. *Biochemistry*, *45*(23), 7299–7306.

Muerdter, F., Boryń, Ł. M., & Arnold, C. D. (2015). STARR-seq – principles and applications. *Genomics*, *106*(3), 145–150.

Muse, G. W., Gilchrist, D. A., Nechaev, S., Shah, R., Parker, J. S., Grissom, S. F., … Adelman, K. (2007). RNA polymerase is poised for activation across the genome. *Nature Genetics*, *39*(12), 1507–1511.

Nechaev, S., Fargo, D. C., dos Santos, G., Liu, L., Gao, Y., & Adelman, K. (2010). Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science*, *327*(5963), 335–338.

Négre, N., Brown, C. D., Ma, L., Bristow, C. A., Miller, S. W., Wagner, U., … White, K. P. (2011). A cis-regulatory map of the Drosophila genome. *Nature*, *471*(7339), 527–531.

Nguyen, D. H., & D'haeseleer, P. (2006). Deciphering principles of transcription regulation in eukaryotic genomes. *Molecular Systems Biology*, *2*(1), 2006.0012.

Ni, T., Corcoran, D. L., Rach, E. A., Song, S., Spana, E. P., Gao, Y., … Zhu, J. (2010). A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nature Methods*, *7*(7), 521–527.

O'Hare, K., & Rubin, G. M. (1983). Structures of P transposable elements and their sites of insertion and excision in the Drosophila melanogaster genome. *Cell*, *34*(1), 25–35.

O'Shea-Greenfield, A., & Smale, S. T. (1992). Roles of TATA and initiator elements in determining the start site location and direction of RNA polymerase II transcription. *Journal of Biological Chemistry*, *267*(2), 1391–1402.

Ohler, U. (2006). Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction. *Nucleic Acids Research*, *34*(20), 5943–5950.

Ohler, U., Liao, G., Niemann, H., & Rubin, G. M. (2002). Computational analysis of core promoters in the Drosophila genome. *Genome Biology*, *3*(12), RESEARCH0087.

Orphanides, G., Lagrange, T., & Reinberg, D. (1996). The general transcription factors of RNA polymerase II. *Genes & Development*, *10*(21), 2657–2683.

Parry, T. J., Theisen, J. W. M., Hsu, J.-Y., Wang, Y.-L., Corcoran, D. L., Eustice, M., … Kadonaga, J. T. (2010). The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes & Development*, *24*(18), 2013–2018.

Patwardhan, R. P., Hiatt, J. B., Witten, D. M., Kim, M. J., Smith, R. P., May, D., … Shendure, J. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. *Nature Biotechnology*, *30*(3), 265–270.

Patwardhan, R. P., Lee, C., Litvin, O., Young, D. L., Pe'er, D., & Shendure, J. (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature Biotechnology*, *27*(12), 1173–1175.

Pilpel, Y., Sudarsanam, P., & Church, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics*, *29*(2), 153–159.

Plaschka, C., Hantsche, M., Dienemann, C., Burzinski, C., Plitzko, J., & Cramer, P. (2016). Transcription initiation complex structures elucidate DNA opening. *Nature*, *533*(7603), 353–358.

Qiu, Y., & Gilmour, D. S. (2017). Identification of regions in the Spt5 subunit of DRB sensitivity-inducing factor (DSIF) that are involved in promoter-proximal pausing. *Journal of Biological Chemistry*, *292*(13), 5555–5570.

Rabenstein, M. D., Zhou, S., Lis, J. T., & Tjian, R. (1999). TATA box-binding protein (TBP)-related factor 2 (TRF2), a third member of the TBP family. *Proceedings of the National Academy of Sciences*, *96*(9), 4791–4796.

Rach, E. A., Winter, D. R., Benjamin, A. M., Corcoran, D. L., Ni, T., Zhu, J., & Ohler, U. (2011). Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLoS Genetics*, *7*(1), e1001274.

Rach, E. A., Yuan, H.-Y., Majoros, W. H., Tomancak, P., & Ohler, U. (2009). Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. *Genome Biology*, *10*(7), R73.

Rajkumar, A. S., Dénervaud, N., & Maerkl, S. J. (2013). Mapping the fine structure of a eukaryotic promoter input-output function. *Nature Genetics*, *45*(10), 1207–1215.

Riddihough, G., & Pelham, H. R. (1987). An ecdysone response element in the Drosophila hsp27 promoter. *The EMBO Journal*, *6*(12), 3729–3734.

Roeder, R. G. (1996). Nuclear RNA polymerases: role of general initiation factors and cofactors in eukaryotic transcription. *Methods in Enzymology*, *273*, 165–171.

Rougvie, A. E., & Lis, J. T. (1988). The RNA polymerase II molecule at the 5′ end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. *Cell*, *54*(6), 795–804.

Roy, S., Ernst, J., Kharchenko, P. V, Kheradpour, P., Negre, N., Eaton, M. L., … Kellis, M. (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, *330*(6012), 1787–1797.

Sainsbury, S., Bernecky, C., & Cramer, P. (2015). Structural basis of transcription initiation by RNA polymerase II. *Nature Reviews Molecular Cell Biology*, *16*(3), 129–143.

Schöne, S., Jurk, M., Helabad, M. B., Dror, I., Lebars, I., Kieffer, B., … Meijsing, S. H. (2016). Sequences flanking the core-binding site modulate glucocorticoid receptor structure and activity. *Nature Communications*, *7*(1), 12621.

Schor, I. E., Degner, J. F., Harnett, D., Cannavò, E., Casale, F. P., Shim, H., … Furlong, E. E. M. (2017). Promoter shape varies across populations and affects promoter evolution and expression noise. *Nature Genetics*, *49*(4), 550–558.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., & Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, *34*(2), 166–176.

Shao, W., & Zeitlinger, J. (2017). Paused RNA polymerase II inhibits new transcriptional initiation. *Nature Genetics*, *49*(7), 1045–1051.

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., … Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology*, *30*(6), 521–530.

Sharon, E., van Dijk, D., Kalma, Y., Keren, L., Manor, O., Yakhini, Z., & Segal, E. (2014). Probing the effect of promoters on noise in gene expression using thousands of designed sequences. *Genome Research*, *24*(10), 1698–1706.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., … Hayashizaki, Y. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences*, *100*(26), 15776–15781.

Shlyueva, D., Stelzer, C., Gerlach, D., Yáñez-Cuna, J. O., Rath, M., Boryń, Ł. M., … Stark, A. (2014). Hormone-responsive enhancer-activity maps reveal predictive motifs, indirect repression, and targeting of closed chromatin. *Molecular Cell*, *54*(1), 180–192.

Simon, J. M., Giresi, P. G., Davis, I. J., & Lieb, J. D. (2012). Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nature Protocols*, *7*(2), 256–267.

Smale, S. T., & Kadonaga, J. T. (2003). The RNA polymerase II core promoter. *Annual Review of Biochemistry*, *72*(1), 449–479.

Song, L., & Crawford, G. E. (2010). DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, *2010*(2), pdb.prot5384.

Spitz, F., & Furlong, E. E. M. (2012). Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, *13*(9), 613–626.

Stormo, G. D. (2000). DNA binding sites: Representation and discovery. *Bioinformatics*, *16*(1), 16–23.

Stormo, G. D., Schneider, T. D., Gold, L., & Ehrenfeucht, A. (1982). Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Research*, *10*(9), 2997–3011.

Sudarsanam, P., Pilpel, Y., & Church, G. M. (2002). Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in Saccharomyces cerevisiae. *Genome Research*, *12*(11), 1723–1731.

Takahashi, H., Lassmann, T., Murata, M., & Carninci, P. (2012). 5′ end–centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nature Protocols*, *7*(3), 542–561.

Thomas, M. C., & Chiang, C.-M. (2006). The general transcription machinery and general cofactors. *Critical Reviews in Biochemistry and Molecular Biology*, *41*(3), 105–178.

Thummel, C. S. (2001). Molecular mechanisms of developmental timing in C. elegans and Drosophila. *Developmental Cell*, *1*(4), 453–465.

Tillo, D., & Hughes, T. R. (2009). G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics*, *10*(1), 442.

Valen, E., Pascarella, G., Chalk, A., Maeda, N., Kojima, M., Kawazu, C., … Carninci, P. (2009). Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Research*, *19*(2), 255–265.

van Arensbergen, J., FitzPatrick, V. D., de Haas, M., Pagie, L., Sluimer, J., Bussemaker, H. J., & van Steensel, B. (2017). Genome-wide mapping of autonomous promoter activity in human cells. *Nature Biotechnology*, *35*(2), 145–153.

Vo Ngoc, L., Cassidy, C. J., Huang, C. Y., Duttke, S. H. C., & Kadonaga, J. T. (2017). The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters. *Genes & Development*, *31*(1), 6–11.

Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., … Fu, X.-D. (2011). Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature*, *474*(7351), 390–394.

Wang, Y.-L., Duttke, S. H. C., Chen, K., Johnston, J., Kassavetis, G. A., Zeitlinger, J., & Kadonaga, J. T. (2014). TRF2, but not TBP, mediates the transcription of ribosomal protein genes. *Genes & Development*, *28*(14), 1550–1555.

Weber, C. M., Ramachandran, S., & Henikoff, S. (2014). Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. *Molecular Cell*, *53*(5), 819–830.

Wei, C.-L., Ng, P., Chiu, K. P., Wong, C. H., Ang, C. C., Lipovich, L., … Ruan, Y. (2004). 5' Long serial analysis of gene expression (LongSAGE) and 3' LongSAGE for transcriptome characterization and genome annotation. *Proceedings of the National Academy of Sciences*, *101*(32), 11701–11706.

Yao, T.-P., Forman, B. M., Jiang, Z., Cherbas, L., Chen, J.-D., McKeown, M., … Evans, R. M. (1993). Functional ecdysone receptor is the product of EcR and Ultraspiracle genes. *Nature*, *366*(6454), 476–479.

Zabidi, M. A., Arnold, C. D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., & Stark, A. (2015). Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, *518*(7540), 556–559.

Zabidi, M. A., & Stark, A. (2016). Regulatory enhancer–core-promoter communication via transcription factors and cofactors. *Trends in Genetics*, *32*(12), 801–814.

Zaret, K. S., & Carroll, J. S. (2011). Pioneer transcription factors: establishing competence for gene expression. *Genes & Development*, *25*(21), 2227–2241.

Zeitlinger, J., Stark, A., Kellis, M., Hong, J.-W., Nechaev, S., Adelman, K., … Young, R. A. (2007). RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nature Genetics*, *39*(12), 1512–1516.

Zhang, Z., & Dietrich, F. S. (2005). Mapping of transcription start sites in Saccharomyces cerevisiae using 5' SAGE. *Nucleic Acids Research*, *33*(9), 2838–2851.