Title

Analysis of the Drosophila and human DPR elements reveals a distinct human variant whose specificity can be enhanced by machine learning.

Journal

Genes & Development, 37(9-10)

ISSN

0890-9369

Authors

Vo Ngoc, Long
Rhyne, Torrey E
Kadonaga, James T

Publication Date

2023-05-01

DOI

10.1101/gad.350572.123

RESEARCH COMMUNICATION

Analysis of the *Drosophila* and Human DPR Elements Reveals a Distinct Human Variant Whose Specificity Can Be Enhanced by Machine Learning

Long Vo ngoc,[1,2,3] Torrey E. Rhyne,[1,3] and James T. Kadonaga[1,4]

[Key words: Transcription, RNA Polymerase II, Core Promoter, Gene Expression, *Drosophila*]

Running Head: Enhancing DNA element specificity by using AI

[1]Department of Molecular Biology, University of California, San Diego, La Jolla, California 92093, USA; [2]Current address: Velia Therapeutics, San Diego, California 92130.

[3]These authors contributed equally to this work
[4]Corresponding author: jkadonaga@ucsd.edu

Abstract

The RNA polymerase II core promoter is the site of convergence of the signals that lead to the initiation of transcription. Here, we perform a comparative analysis of the downstream core promoter region (DPR) in *Drosophila* and humans by using machine learning. These studies revealed a distinct human-specific version of the DPR and led to the use of machine learning models for the identification of synthetic extreme DPR motifs with specificity for human transcription factors relative to *Drosophila* factors, and vice versa. More generally, machine learning models could similarly be used to design synthetic DNA elements with customized functional properties.

The initiation of transcription by RNA polymerase II is an important step in the expression of genes (see, for example: Cramer 2019, Galouzis and Furlong 2022, Haberle et al. 2018, Roeder 2019, Schier and Taatjes 2020, Sloutskin et al. 2021, Vo ngoc et al. 2019, Zeitlinger 2020). Transcription initiates at the core promoter, which comprises the stretch of DNA from about −40 to +40 relative to the +1 transcription start site (TSS). The core promoter is often referred to as the "gateway to transcription", as it is the site of convergence of the signals that direct the initiation of transcription. Core promoter activity is driven by DNA sequence elements such as the TATA box, initiator (Inr), and downstream promoter region (DPR), which is the revised name for the combined motif ten element (MTE; Lim et al. 2004) and downstream core promoter element (DPE; Burke and Kadonaga 1996) from +17 to +35 relative to the +1 TSS (Vo ngoc et al. 2020). There are no universal core promoter elements. Moreover, individual core promoter motifs are involved in transcriptional regulatory functions such as enhancer-core promoter specificity (see, for example, Ohtsuki et al. 1998, Butler and Kadonaga 2001, Zabidi et al. 2015, Galouzis and Furlong 2022).

A key challenge in the study of the core promoter has been the ability to predict the existence and activity of particular core promoter elements. In the past, the presence of core promoter motifs has been generally assessed by the similarity of the DNA sequences to consensus sequences, such as TATAAA for the TATA box. This method can lead to the incorrect assignments of motifs and also does not provide a quantitative prediction of the transcription strength of the putative elements.

To address this problem, we employed a two-step approach in which we first determine the transcription strengths of each of hundreds of thousands of DNA sequence variants of a core promoter motif (by high-throughput analysis of randomized promoter elements, HARPE) and then use the resulting data to create a support vector regression (SVR, a form of machine learning and artificial intelligence) model of the element (Vo ngoc et al. 2020). The SVR model for a core promoter element provides an objective, data-based, quantitative prediction of the

transcription strength of any test DNA sequence. Thus far, we have generated SVR models for the human TATA box and the human DPR (Vo ngoc et al. 2020).

In this study, we carried out a comparative analysis of the DPR motif in humans and in *Drosophila*. To accomplish this objective, we generated high-throughput HARPE data and a machine learning model for the *Drosophila* DPR. We felt that this work would yield new insights into the DPR from the perspectives of a protostome and a deuterostome, which have an estimated species divergence time of about 700 million years (Kumar et al. 2022). In addition, the DPE has been found to occur frequently in *Drosophila* promoters (see, for example, Kutach and Kadonaga 2000, Ohler et al. 2002, FitzGerald et al. 2006, Chen et al. 2014); hence, the analysis of the *Drosophila* DPR would provide useful information for this important model organism. We thus embarked on the HARPE and SVR analyses of the *Drosophila* DPR and compared its properties to those of its human counterpart. These studies unexpectedly revealed a human-specific DPR variant, which further led to the use of SVR models for the design of synthetic DPR motifs with customized functional properties.

Results & Discussion

The SVRd model provides a reliable prediction of DPR activity in Drosophila

To generate an SVR model for the *Drosophila* DPR, we carried out HARPE (Fig. 1A) and SVR analyses by the method of Vo ngoc et al. (2020). To measure basal transcription activity from the core promoter, we used the *Drosophila* embryo extract system developed by Soeller et al. (1988), which has been found to mediate accurate initiation of transcription in vitro that is essentially identical to that seen in vivo in embryos for a wide range of promoters (see, for example: Biggin and Tjian 1988, Kerrigan et al. 1990, Kutach and Kadonaga 2000, Lim et al. 2004, Perkins et al. 1988). The HARPE data from two independent replicates were found to be reproducible (PCC = 0.97; Supplemental Fig. S1). Only a small fraction of the randomized sequences exhibited DPR activity (Fig. 1B), and HOMER analysis (Heinz et al. 2010) of the top

0.1% most active sequences yielded a motif with a strong resemblance to the *Drosophila* DPE consensus (RGWYS from +28 to +32 relative to the +1 TSS; Kutach and Kadonaga 2000) (Fig. 1C).

We then used the HARPE data (200,000 sequence variants, each with an experimentally determined transcription strength) to train and to optimize an SVR model, which we term SVRd, for SVR model of the *Drosophila* DPR (Supplemental Fig. S2). There is a strong correlation (PCC = 0.89, rho = 0.91) between the observed transcription strengths of 7,115 independent (*i.e.*, not used in the training of SVRd) test DPR sequences and their predicted DPR activities (Fig. 1D). Thus, SVRd provides a strong prediction of DPR activity in *Drosophila*.

To characterize the effectiveness of SVRd, we carried out a performance assessment and found that the SVRd score of 1.5 is the best threshold for distinguishing active (SVRd ≥ 1.5) versus inactive (SVRd < 1.5) DPR elements (Supplemental Fig. S3). Moreover, DPR sequences with SVRd scores ≥ 1.5 exhibit at least 11-fold higher activity than the median inactive sequence (Supplemental Fig. S4). We therefore consider DPR sequences with SVRd scores ≥ 1.5 to be active. Strikingly, by this measure, about 68% of focused natural *Drosophila* promoters in embryos contain active DPR motifs, whereas, in contrast, only 19% of random sequences with the same G/C content are predicted to function as active DPR elements (Fig. 1E). Similarly, we observed that about 68% of focused promoters in *Drosophila* S2 cells are predicted to have an active DPR (Supplemental Fig. S5A). The DPR thus appears to be a widely used core promoter element in *Drosophila*.

We also compared the performance of SVRd with that of the DPE consensus sequence. We found that a perfect match to the DPE consensus is a poor predictor of DPR activity, whereas the SVRd model provides excellent assessments of the activities of the same sequences (Supplemental Fig. S6). Hence, the SVRd model is superior to the DPE consensus sequence for the prediction of DPR activity in *Drosophila*. Importantly, the SVRd model predicts the quantitative strength of each test DPR sequence.

A human-specific DPR variant is used in humans but not in Drosophila

With the *Drosophila* HARPE data and the SVRd model, we compared the properties of the *Drosophila* DPR with those of the human DPR (Vo ngoc et al. 2020). [Note: in this study, the SVRb model in Vo ngoc et al. (2020) is termed SVRh, for SVR model of the human DPR.] The direct comparison of the observed transcription strengths (as assessed by HARPE) of DPR sequence variants in *Drosophila* versus humans revealed many general variants that are active in both organisms as well as a distinct set of "human-specific" (*i.e.*, specific for humans relative to *Drosophila*) variants that are more active in humans than in *Drosophila* (Fig. 2A). HOMER analysis indicated that the general variants contain the canonical DPR element with the DPE motif (RGWYS) in the standard position (from +28 to +32 relative to the +1 TSS), whereas the human-specific variants contain the DPE motif shifted 1 nt upstream (to +27 to +31 relative to the +1 TSS) (Fig. 2B).

We also examined the occurrence of human-specific variants in natural human and *Drosophila* promoters. In humans, about 25% (799 out of 3,161) of predicted active DPR sequences [which corresponds to about 7% (799 out of 11,932) of all focused promoters] are the human-specific variants (Fig. 2C), whereas in *Drosophila*, only about 1% (29 out of 3,070) of predicted active DPR sequences [which corresponds to about 0.6% (29 out of 4,489) of all focused promoters] are similar to the human-specific variants (Fig. 2D; also see Supplemental Fig. S5).

The −1 spacing is the basis of the human specificity of the DPR variants

We then sought to gain a better understanding of the human-specific DPR variants. We first tested whether these variants are important for core promoter function in humans. To this end, we analyzed five natural human-specific variants with the −1 DPR spacing, and found that they are important for transcriptional activity in the context of their entire core promoter regions from

−36 to +50 relative to the +1 TSS (Supplemental Fig. S7). Hence, the human-specific DPR variants are functionally important in natural promoters.

Next, we investigated whether the lack of activity of the human-specific variants in *Drosophila* is due to the −1 spacing of the DPR relative to the canonical position. Although the human-specific variants possess the −1 spacing of the DPR (Fig. 2B), we did not know whether or not the human specificity is due to this altered spacing. To address this question, we subjected five canonical and five human-specific variants to in vitro transcription analysis with either human or *Drosophila* factors (Fig. 3; Supplemental Fig. S7A). To enable a direct comparison of the activities of the different DPR sequences, we placed each test DPR sequence into the SCP1m promoter, which lacks a TATA box but contains an Inr element (Juven-Gershon et al. 2006). We therefore tested the activity of each DPR sequence in the same promoter context.

First, we observed that the human transcription factors are able to function with both the canonical and the human-specific DPR variants, whereas the *Drosophila* factors function with the canonical DPR elements but not with the human-specific −1 variants (Fig. 3A,B; Supplemental Fig. S8A). We then tested whether the *Drosophila* factors could function with the human-specific variants if they were shifted 1 nt downstream to the canonical position. These experiments revealed that the *Drosophila* factors could indeed function with the +1 nt-shifted human-specific DPR elements (Fig. 3C,D; Supplemental Fig. S8B). Thus, the inability of the *Drosophila* factors to transcribe the human-specific −1 DPR variants can be rescued by shifting the element by 1 nt downstream to the canonical position. These transcriptional effects were also predicted somewhat accurately with the SVRh and SVRd models (Fig. 3C,D).

It remained possible, however, that the *Drosophila* factors might be able to transcribe some −1 DPR variants. To address this point, we analyzed our HARPE dataset of 437,002 DPR sequence variants (Fig. 1A,B) and found that those with the canonical +28 positioning of the RGWYS DPE motif (Fig. 1C) possess much higher transcription strengths than variants with the

−1 positioning of this motif at +27 (Supplemental Fig. S9A,B). Furthermore, in the analysis of natural *Drosophila* promoters, the RGWYS motif is much more commonly found at the +28 position than at the +27 position (Supplemental Figs. S5C, S9C). We therefore conclude that the DPR in the noncanonical −1 position is an active core promoter element in humans but not in *Drosophila*.

Use of SVR models to identify synthetic extreme DNA sequence elements

The identification of canonical and human-specific variants of the DPR led us to explore new applications for the SVR models. We were first interested in using the machine learning models to enhance the human to *Drosophila* specificity of the −1 DPR variants. To this end, we substantially expanded the range of DPR sequence candidates (~100-fold relative to the HARPE library, as in Fig. 1A) by generating 50 million random 19-nt sequences and determining their predicted DPR scores with SVRd and SVRh (Supplemental Fig. S10). In this analysis, the top 0.001% variants with the highest SVRh:SVRd score ratios yielded a distinct HOMER motif in which the 1 nt upstream shift of the −1 DPR can be clearly seen (Fig. 4A).

Then, to test the transcriptional properties of the predicted extreme DPR elements, we selected four synthetic sequences, which we termed E1 to E4 (Fig. 4B, Supplemental Fig. S11A), with high SVRh:SVRd ratios, and found that they are highly active with human factors and possess little or no detectable activity with *Drosophila* factors (Fig. 4C, Supplemental Fig. S11B). The E2, E3, and E4 variants are particularly human-specific, with human:*Drosophila* transcription ratios (with respect to the SCP1m DPR reference) of at least 75. Hence, these findings suggest that machine learning analysis of a wide range of sequence variants can be used to identify DNA elements with specific functional properties.

We next examined whether it is possible to perform the reciprocal experiment – that is, use the SVR models to identify *Drosophila*-specific DPR sequences. In this regard, we identified four synthetic DPR sequences, termed E5 to E8 (Fig. 4B, Supplemental Fig. 11A), with high

SVRd:SVRh ratios, and found that they exhibit stronger activity with *Drosophila* factors than with human factors (Fig. 4C, Supplemental Fig. 11C). The E8 variant exhibited about 17-fold higher transcription activity with *Drosophila* factors relative to human factors (with respect to the SCP1m DPR reference). Thus, although there is not a distinct *Drosophila*-specific class of DPR elements, it is possible to identify synthetic DPR sequences that have stronger transcription activity with *Drosophila* relative to human factors, as assessed under the same conditions that were employed to generate the data for the machine learning models. These results further support the conclusion that the generation and use of machine learning models can be used to identify synthetic DNA sequence motifs with customized properties.

Summary and perspectives

In this study, we performed a comparative analysis of the DPR core promoter motif in *Drosophila* and humans. This work led to the identification of a human-specific DPR variant in which the DPR is located 1 nt upstream of the canonical DPR (Figs. 2, 3, 4A). The human-specific −1 DPR appears to be used in about 25% of DPR elements in natural human promoters (Fig. 2C). Strikingly, the DPR is predicted to be present in about two-thirds of natural *Drosophila* promoters in embryos and in cells (Fig. 1E, Supplemental S5A). Moreover, even though key promoter characteristics such as CpG islands are present in humans but not in *Drosophila* (Deaton and Bird 2011), the predicted optimal canonical DPR has remained mostly unchanged over the estimated 700 million years of species divergence time from *Drosophila* to humans (Supplemental Fig. S12). In this manner, the analysis of the DPR in both *Drosophila* and humans has led to new insights that would not have been obtained in the study of the DPR in either organism alone. The altered spacing in the human-specific DPR reveals differences in the transcription machinery in humans relative to *Drosophila*. It remains to be determined, however, whether the expanded range of function of the human transcription factors is used to achieve new modes of regulation.

The existence of the human-specific DPR motif inspired us to explore the use of the SVR models to predict extreme versions of the DPR that exhibit specificity for transcription with human factors relative to *Drosophila* factors, and vice versa. In this work, we used the SVR models to expand the scope of the analysis to 50 million DPR sequence variants (Fig. 4, Supplemental Figs. S10, S11, S12). In the 50 million variants, the extreme human- or *Drosophila*-specific DPR motifs that were predicted by the SVR models were found to be excellent candidates, as assessed by transcriptional analyses (Fig. 4C, Supplemental Fig. S11). The SVR model predictions were good but not quantitatively perfect, possibly due to the extreme or fringe nature of the candidates, but they did yield DPR motifs that were much more active with human transcription factors relative to *Drosophila* factors, and vice versa. It is also expected that the accuracy of the machine learning models will continue to improve in the future.

Thus, these experiments provide a demonstration of the use of machine learning models for the identification of DNA sequence motifs with custom-tailored functions. For example, SVR models could be made for a promoter element that stimulates transcription in Condition 1 (SVR1) as well as in Condition 2 (SVR2). Then, by using an analogous approach as in this work, DNA sequence motifs that activate transcription in Condition 1 but not in Condition 2, and vice versa, could be identified. Hence, in this manner, the use of machine learning models for the study of DNA sequence elements can extend beyond the analysis of natural DNA sequence elements to the prediction and identification of synthetic sequence variants with specifically-desired properties.

Materials and methods

HARPE method

The HARPE plasmid libraries for the DPR were described in Vo ngoc et al. (2020). Sample and data processing were performed as in Vo ngoc et al. (2020), with the exception that the in vitro transcription reactions were carried out with *Drosophila* nuclear extracts for 30 min by the method of Wampler et al. (1990). Additional information is provided in the Supplemental Materials and methods. Sequencing of the PCR amplicons was performed on an Illumina Novaseq 6000 at the IGM Genomics Center, University of California, San Diego, La Jolla, CA (Moores Cancer Center, supported by NIH grant P30 CA023100 and NIH SIG grant S10 OD026929). The genome-wide data have been deposited at the Gene Expression Omnibus (GEO; accession number GSE225570), and will be released upon acceptance of the paper for publication.

Transcription of individual test sequences

The plasmids that were used for testing individual DPR sequences were constructed with the Q5® Site-Directed Mutagenesis Kit (New England Biolabs) as recommended by the manufacturer. Transcription reactions were performed as described in Vo ngoc et al. (2020) for human factors and in Wampler et al. (1990) for *Drosophila* factors. The transcripts were subjected to primer extension analysis, and the reverse transcription products were resolved by 6% polyacrylamide–8 M urea gel electrophoresis and quantified by using a Typhoon imager (GE Health Sciences) and Amersham™ Typhoon™ control software v1.1. Quantification of radiolabeled samples was performed with ImageJ version 2.1.0. All experiments with individual promoter constructs were performed independently at least three times to ensure reproducibility of the data. The quantitated data from the transcription reactions are in Supplemental Table S1. The sequences of the core promoters and DPR elements used in this study are given in Supplemental Table S2.

Competing Interest Statement

The authors declare no competing interests.

Acknowledgments

We thank George Kassavetis, Grisel Cruz-Becerra, and Jack Cassidy for critical reading of the manuscript. J.T.K. is the Amylin Chair in the Life Sciences. This work was supported by NIH grant R35 GM118060 to J.T.K.

Author Contributions

L.V.n. and J.T.K. initially conceived the project and oversaw the overall execution of this work. L.V.n. and T.E.R. performed the laboratory experiments as well as the computational analyses. L.V.n., T.E.R., and J.T.K. prepared the figures and wrote the manuscript.

References

Biggin MD, Tjian R. 1988. Transcription factors that activate the *Ultrabithorax* promoter in developmentally staged extracts. *Cell* **53**: 699–711.

Burke TW, Kadonaga JT. 1996. *Drosophila* TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev* **10**: 711–724.

Butler JE, Kadonaga JT. 2001. Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes Dev* **15**: 2515–2519.

Chen ZX, Sturgill D, Qu J, Jiang H, Park S, *et al.* 2014. Comparative validation of the *D. melanogaster* modENCODE transcriptome annotation. *Genome Res* **24**: 1209–1223.

Cramer P. 2019. Organization and regulation of gene transcription. *Nature* **573:** 45–54.

Deaton AM, Bird A. 2011. CpG islands and the regulation of transcription. *Genes Dev* **25**: 1010–1022.

FitzGerald PC, Sturgill D, Shyakhtenko A, Oliver B, Vinson C. 2006. Comparative genomics of *Drosophila* and human core promoters. *Genome Biol* **7**: R53.

Galouzis CC, Furlong EEM. 2022. Regulating specificity in enhancer-promoter communication. *Curr Opin Cell Biol* **75**: 102065.

Haberle V., Stark A. 2018. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol* **19**: 621–637.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**: 576–589.

Juven-Gershon T, Cheng S, Kadonaga JT. 2006. Rational design of a super core promoter that enhances gene expression. *Nat Methods* **3**: 917–922.

Kerrigan LA, Croston GE, Lira LM, Kadonaga JT. 1991. Sequence-specific transcriptional antirepression of the *Drosophila Krüppel* gene by the GAGA factor. *J Biol Chem* **266**: 574–582.

Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, Stecher G, Hedges SB. 2022. TimeTree 5: an expanded resource for species divergence times. *Mol Biol Evol* **39**: msac174.

Kutach AK, Kadonaga JT. 2000. The downstream promoter element DPE appears to be as widely used as the TATA box in *Drosophila* core promoters. *Mol Cell Biol* **20**: 4754–4764.

Lim CY, Santoso B, Boulay T, Dong E, Ohler U, Kadonaga JT. 2004. The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev* **18**: 1606–1617.

Ohler U, Liao GC, Niemann H, Rubin GM. 2002. Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol* **3**: RESEARCH0087.

Ohtsuki S, Levine M, Cai HN. 1998. Different core promoters possess distinct regulatory activities in the *Drosophila* embryo. *Genes Dev.* **12**: 547–556.

Perkins KK, Dailey GM, Tjian R. 1988. In vitro analysis of the *Antennapedia* P2 promoter: identification of a new *Drosophila* transcription factor. *Genes Dev* **2**: 1615–1626.

Roeder RG. 2019. 50+ years of eukaryotic transcription: an expanding universe of factors and mechanisms. *Nat Struct Mol Biol* **26**: 783–791.

Schier AC, Taatjes DJ. 2020. Structure and mechanism of the RNA polymerase II transcription machinery. *Genes Dev* **34**: 465–488.

Sloutskin A, Shir-Shapira H, Freiman RN, Juven-Gershon T. 2021. The core promoter is a regulatory hub for developmental gene expression. *Front Cell Dev Biol* **9**: 666508.

Soeller WS, Poole SJ, Kornberg T. 1988. In vitro transcription of the *Drosophila engrailed* gene. *Genes Dev* **2**: 68–81.

Theisen JW, Lim CY, Kadonaga JT. 2010. Three key subregions contribute to the function of the downstream RNA polymerase II core promoter. *Mol Cell Biol* **30**: 3471–3479.

Vo ngoc L, Kassavetis GA, Kadonaga JT. 2019. The RNA polymerase II core promoter in *Drosophila*. *Genetics* **212**: 13–24.

Vo ngoc L, Huang CY, Cassidy CJ, Medrano C, Kadonaga JT. 2020. Identification of the human DPR core promoter element using machine learning. *Nature* **585**: 459–463.

Wampler SL, Tyree CM, Kadonaga JT. 1990. Fractionation of the general RNA polymerase II transcription factors from *Drosophila* embryos. *J Biol Chem* **265**: 21223–21231.

Zabidi MA, Arnold CD, Schernhuber K, Pagani M, Rath M, Frank O, Stark A. 2015. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**: 556–559.

Zeitlinger J. 2020. Seven myths of how transcription factors read the cis-regulatory code. *Curr Opin Syst Biol* **23**: 22–31.

Figure Legends

Figure 1. HARPE and SVR analyses of the DPR in *Drosophila melanogaster*. (*A*) Use of the HARPE method for the analysis of the *Drosophila* DPR. The TATA-less SCP1m promoter backbone, which contains two GC boxes upstream of the mutated TATA box, is identical to that used in the analysis of the human DPR (Vo ngoc et al. 2020). The DPR region is randomized from +17 to +35 relative to the +1 TSS. (*B*) Most DPR sequence variants exhibit low transcriptional activity, but a small fraction of the variants are highly active. The graph depicts the transcription strength of each of the 437,002 DPR variants, which are ranked along the x-axis in order of decreasing activity. The data are the average values from two independent biological replicates. (*C*) HOMER analysis of the 0.1% most active DPR variants reveals a distinct motif that contains a DPE-like sequence. This figure displays the web logo of the top HOMER motif obtained from the data in *B*. The DPE-like sequence (RGWYS) is from +28 to +32. (*D*) The SVRd model of the *Drosophila* DPR accurately predicts the transcription strengths of DPR sequence variants. The SVRd machine learning model was generated by training with 200,000 variants in the HARPE dataset, and it provides a numerical prediction for the activity of any potential DPR sequence. SVRd was then tested with 7,115 independent sequences (*i.e.*, not used in the training of SVRd) from the HARPE dataset. For each of the independent test sequences, the predicted SVRd score was compared with the observed transcription strength. The value of the SVRd score is not identical to the transcription strength. PCC, Pearson's correlation coefficient; rho, Spearman's rank correlation coefficient. (*E*) Approximately two-thirds of natural *Drosophila* promoters are predicted to contain an active DPR. The graph shows the cumulative frequencies of SVRd scores of sequences in the DPR region (+17 to +35) in 4,489 natural *Drosophila* promoters that are active in *Drosophila* embryos. This analysis revealed that approximately 68% of *Drosophila* promoters in embryos are predicted to have an active DPR (SVRd score ≥ 1.5; Supplemental Figs. S3 and S4). In contrast, only about 19% of 500,000 random 19-nt sequences with the same overall G/C content (51.3%) as *Drosophila* core

promoters are predicted to have DPR activity. To examine the variability of the % DPR usage relative to the degree of focus in the TSSs, we determined the % DPR usage in promoters in which the minimum focus index (FImin; Vo ngoc et al. 2017) varies from 0.65 to 0.85, and observed a range of 61% to 71% DPR usage in embryos.

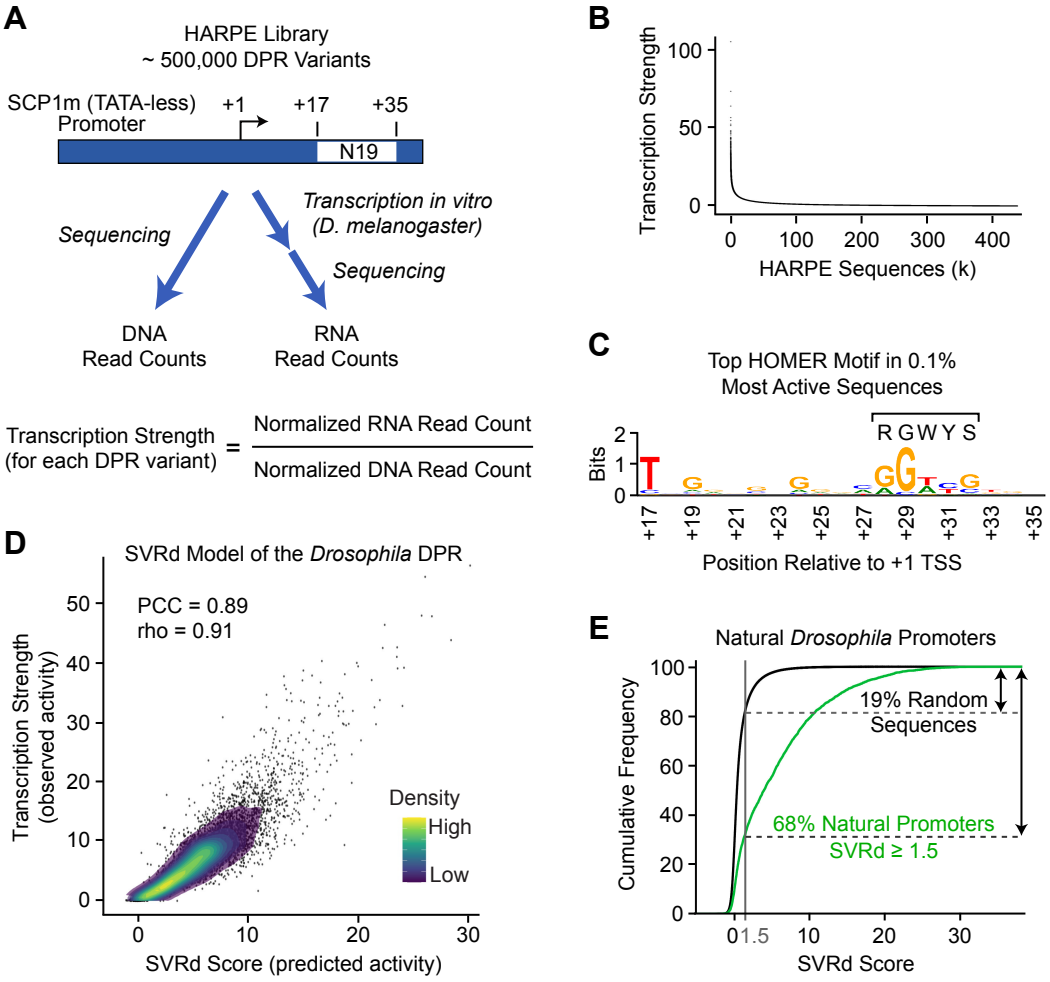
Figure 2. Identification of a species-specific DPR variant that is present in humans but not in *Drosophila*. (*A*) Comparison of the observed transcription strengths, as assessed in HARPE assays, of 437,002 DPR sequence variants in humans versus *Drosophila*. General DPR variants with high activity in both humans and *Drosophila* are depicted in blue. Human-specific DPR variants with high activity in humans and low activity in *Drosophila* are denoted in red. All other variants are shown in grey. We did not observe a distinct class of *Drosophila*-specific DPR variants. The dashed light violet lines depict 5x mean activities of the DPR variants in humans (vertical line) and in *Drosophila* (horizontal line). The black diagonal dashed line demarcates the general variants and the human-specific variants. (*B*) The human-specific DPR variants are positioned 1 bp upstream of the general/canonical DPR variants. All canonical DPR variants (blue dots in *A*) as well as all human-specific DPR variants (red dots in *A*) were analyzed with HOMER. The web logo of the top HOMER motif for each of the classes of variants is shown. (*C*) The human-specific −1 DPR variant appears to be present in about 25% of the 3161 predicted active human DPR elements (SVRh ≥ 2; Vo ngoc et al. 2020) in 11,932 natural human promoters. (*D*) The human-specific −1 DPR variant appears to be present in about 1% of 3,070 predicted active DPR elements (SVRd ≥ 1.5; dashed light line; Supplemental Figs. S3 and S4) in 4,489 natural *Drosophila* promoters. In both *C* and *D*, the black diagonal dashed line demarcates the canonical DPR sequences (blue dots) and the human-specific −1 DPR variants (red dots).

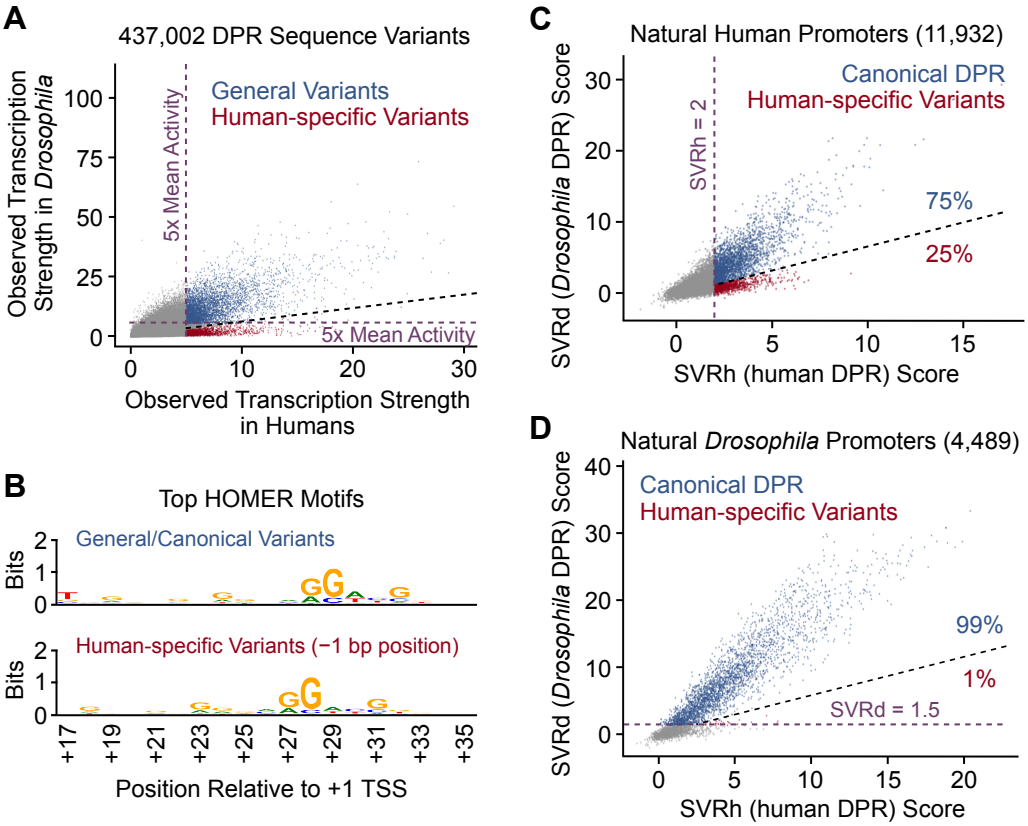
Figure 3. The −1 DPR element is active with human transcription factors but not with *Drosophila* transcription factors. In these experiments, five canonical DPR elements and five

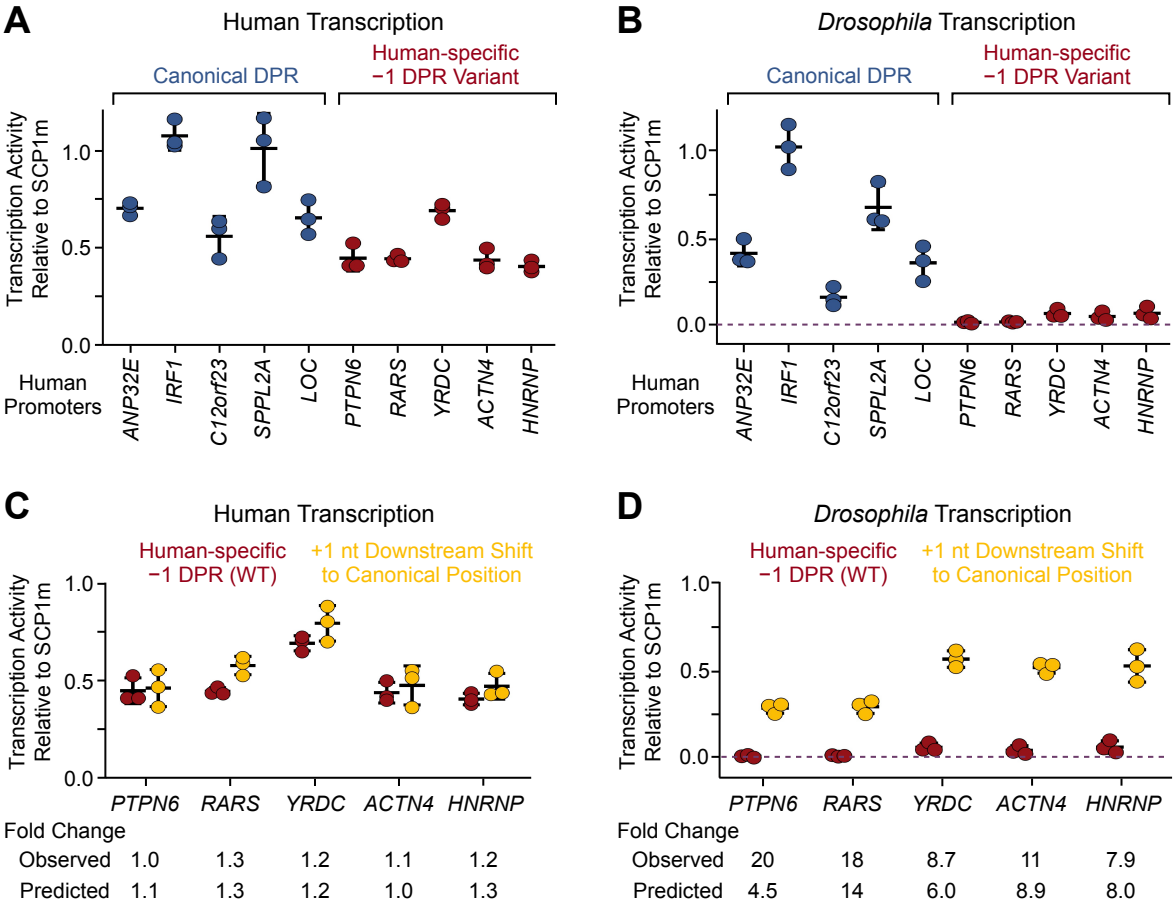
human-specific −1 DPR variants (Supplemental Fig. S7A) were analyzed by replacing the DPR motif in the TATA-less SCP1m reference promoter with DPR sequences from the indicated human genes. The *LOC100505495* and *HNRNPA2B1* genes are abbreviated as *LOC* and *HNRNP*. The resulting promoter constructs were subjected to in vitro transcription analysis with either human or *Drosophila* nuclear extracts. The indicated DPR activities are relative to that of the strong DPR in the SCP1m promoter. The data are shown as the mean ± standard deviation with n = 3 biologically independent samples. Autoradiograms of representative experiments are shown in Supplemental Figs. S8A and S8B, and the quantitated results from each experiment are given in Supplemental Table S1. It is also relevant to note that all of the test DPR sequences have been found to be active in their natural promoter contexts by in vitro transcription analysis (Supplemental Fig. S7B; Vo ngoc et al. 2020). (*A*) The canonical and human-specific −1 DPR elements both have strong transcription activity with human transcription factors. (*B*) Five different human-specific −1 DPR elements exhibit little or no activity with *Drosophila* transcription factors. (*C*) The translocation of −1 DPR variant sequences to the canonical DPR position has little or no effect upon their activity with human transcription factors. The observed and predicted (with SVRh) fold changes in activity (downstream-shifted DPR relative to wild-type −1 DPR) are indicated. (*D*) The translocation of −1 DPR variant sequences to the canonical DPR position results in a substantial increase in activity with *Drosophila* transcription factors. The observed and predicted (with SVRd) fold changes in activity (downstream-shifted DPR relative to wild-type −1 DPR) are indicated.

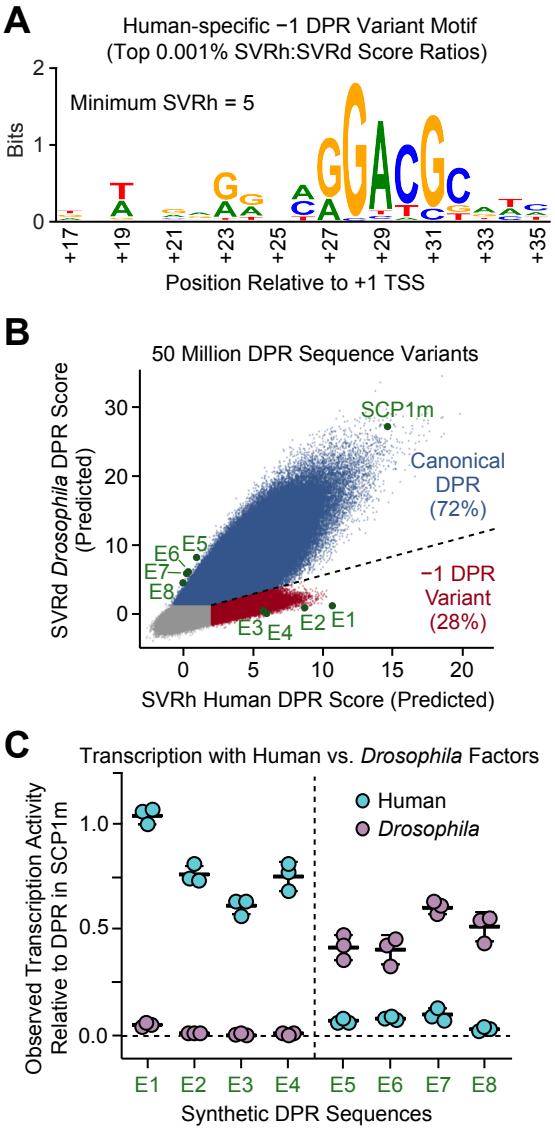
Figure 4. Use of machine-learning to generate DPR variants with specificity for transcription with human factors relative to *Drosophila* factors, and vice versa. (*A*) SVR analysis of 50 million random 19 nt sequences with SVRh and SVRd reveals a distinct preferred motif for the human-specific −1 DPR variant. The upper panel is the HOMER web logo for the human-specific −1 DPR variant, as assessed with the top 500 sequences with SVRh ≥ 5 that have the highest SVRh:SVRd score ratios. (*B*) Identification of DPR variants that are predicted to be specific for

transcription in humans relative to *Drosophila*, and vice versa. Four DPR sequences, termed E1 to E4, have high SVRh:SVRd score ratios, whereas four other sequences, termed E5 to E8, have high SVRd:SVRh score ratios. The specific sequences and their predicted SVRd and SVRh scores are given in Supplemental Table S2. The position of the DPR in SCP1m is also indicated. (*C*) The synthetic E1 to E4 DPR sequences exhibit specificity for transcription with human factors relative to *Drosophila* factors, whereas the E5 to E8 DPR motifs have a distinct preference for *Drosophila* factors relative to human factors. The synthetic DPR sequences in *B* were analyzed in the SCP1m promoter backbone, and their transcriptional activities with human factors and with *Drosophila* factors with were compared to that of the reference DPR in SCP1m (Juven-Gershon et al. 2006). Autoradiograms of representative experiments are shown in Supplemental Fig. S11. The quantitated results from each experiment are given in Supplemental Table S1.


A
HARPE Library
~ 500,000 DPR Variants
SCP1m (TATA-less) Promoter
+1
+17
+35
N19
Sequencing
Transcription in vitro (D. melanogaster)
Sequencing
DNA Read Counts
RNA Read Counts
Transcription Strength (for each DPR variant) = Normalized RNA Read Count / Normalized DNA Read Count
B
Transcription Strength
100
50
0
0
100
200
300
400
HARPE Sequences (k)
C
Top HOMER Motif in 0.1% Most Active Sequences
R G W Y S
Bits
2
1
0
+17
+19
+21
+23
+25
+27
+29
+31
+33
+35
Position Relative to +1 TSS
D
SVRd Model of the Drosophila DPR
PCC = 0.89
rho = 0.91
Transcription Strength (observed activity)
50
40
30
20
10
0
Density
High
Low
0
10
20
30
SVRd Score (predicted activity)
E
Natural Drosophila Promoters
Cumulative Frequency
100
80
60
40
20
0
19% Random Sequences
68% Natural Promoters
SVRd ≥ 1.5
0
1.5
10
20
30
SVRd Score


A
437,002 DPR Sequence Variants
General Variants
Human-specific Variants
5x Mean Activity
5x Mean Activity
Observed Transcription Strength in Drosophila
Observed Transcription Strength in Humans
B
Top HOMER Motifs
General/Canonical Variants
Human-specific Variants (−1 bp position)
Bits
Position Relative to +1 TSS
C
Natural Human Promoters (11,932)
Canonical DPR
Human-specific Variants
SVRh = 2
75%
25%
SVRd (Drosophila DPR) Score
SVRh (human DPR) Score
D
Natural Drosophila Promoters (4,489)
Canonical DPR
Human-specific Variants
99%
1%
SVRd = 1.5
SVRd (Drosophila DPR) Score
SVRh (human DPR) Score

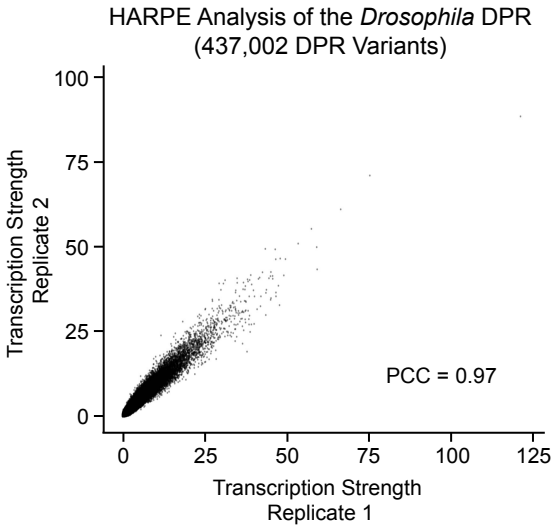

A
Human Transcription
Canonical DPR
Human-specific
−1 DPR Variant
Transcription Activity
Relative to SCP1m
1.0
0.5
0.0
Human
Promoters
ANP32E
IRF1
C12orf23
SPPL2A
LOC
PTPN6
RARS
YRDC
ACTN4
HNRNP
B
Drosophila Transcription
Canonical DPR
Human-specific
−1 DPR Variant
Transcription Activity
Relative to SCP1m
1.0
0.5
0.0
Human
Promoters
ANP32E
IRF1
C12orf23
SPPL2A
LOC
PTPN6
RARS
YRDC
ACTN4
HNRNP
C
Human Transcription
Human-specific
−1 DPR (WT)
+1 nt Downstream Shift
to Canonical Position
Transcription Activity
Relative to SCP1m
1.0
0.5
0.0
PTPN6
RARS
YRDC
ACTN4
HNRNP
Fold Change
Observed 1.0 1.3 1.2 1.1 1.2
Predicted 1.1 1.3 1.2 1.0 1.3
D
Drosophila Transcription
Human-specific
−1 DPR (WT)
+1 nt Downstream Shift
to Canonical Position
Transcription Activity
Relative to SCP1m
1.0
0.5
0.0
PTPN6
RARS
YRDC
ACTN4
HNRNP
Fold Change
Observed 20 18 8.7 11 7.9
Predicted 4.5 14 6.0 8.9 8.0


A
Human-specific −1 DPR Variant Motif
(Top 0.001% SVRh:SVRd Score Ratios)
Minimum SVRh = 5
Bits
Position Relative to +1 TSS
B
50 Million DPR Sequence Variants
SVRd Drosophila DPR Score (Predicted)
SVRh Human DPR Score (Predicted)
SCP1m
Canonical DPR (72%)
−1 DPR Variant (28%)
E1
E2
E3
E4
E5
E6
E7
E8
C
Transcription with Human vs. Drosophila Factors
Observed Transcription Activity Relative to DPR in SCP1m
Human
Drosophila
Synthetic DPR Sequences

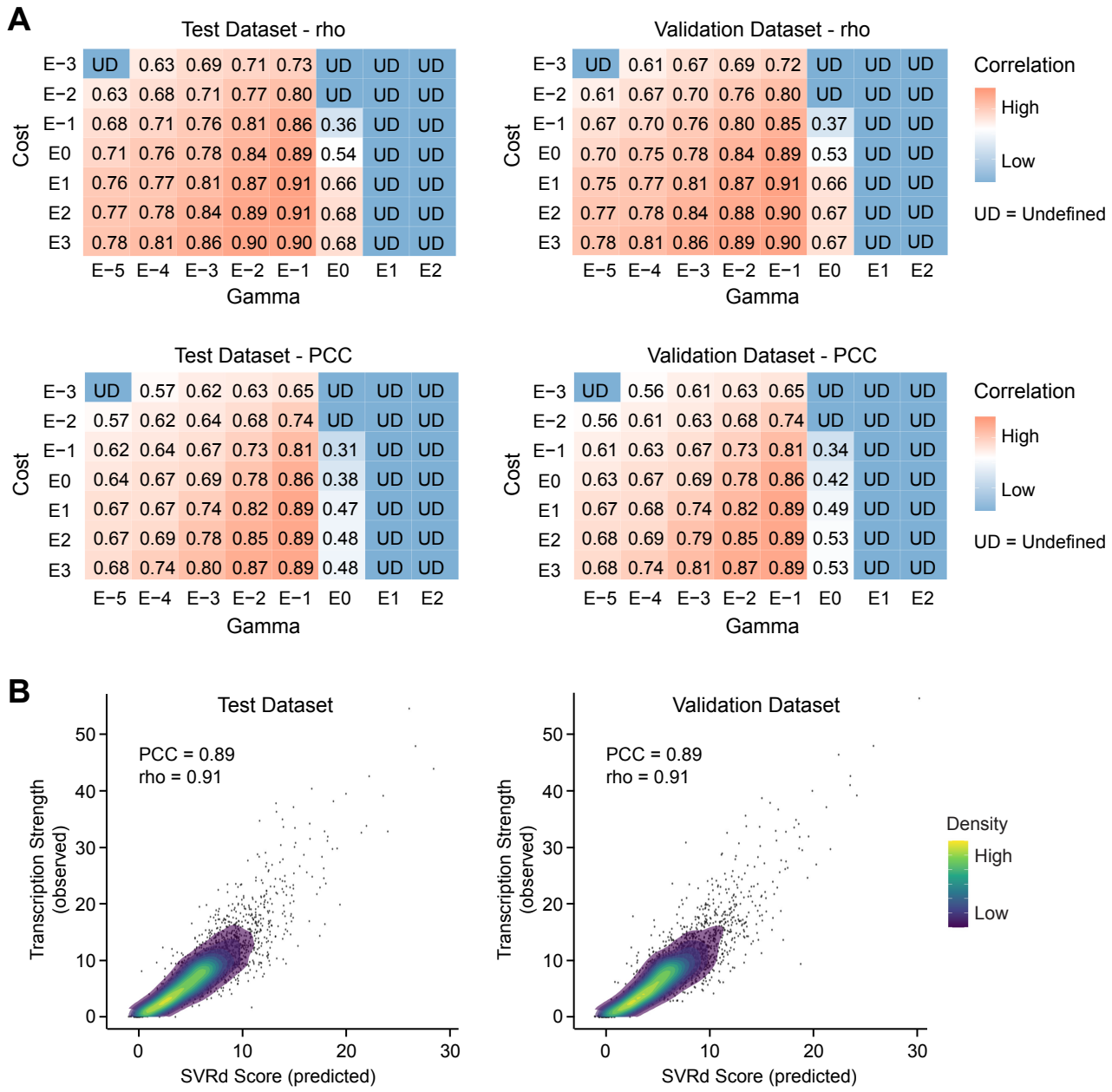
SUPPLEMENTAL MATERIAL

Analysis of the *Drosophila* and Human DPR Elements Reveals a Distinct Human Variant Whose Specificity Can Be Enhanced by Machine Learning

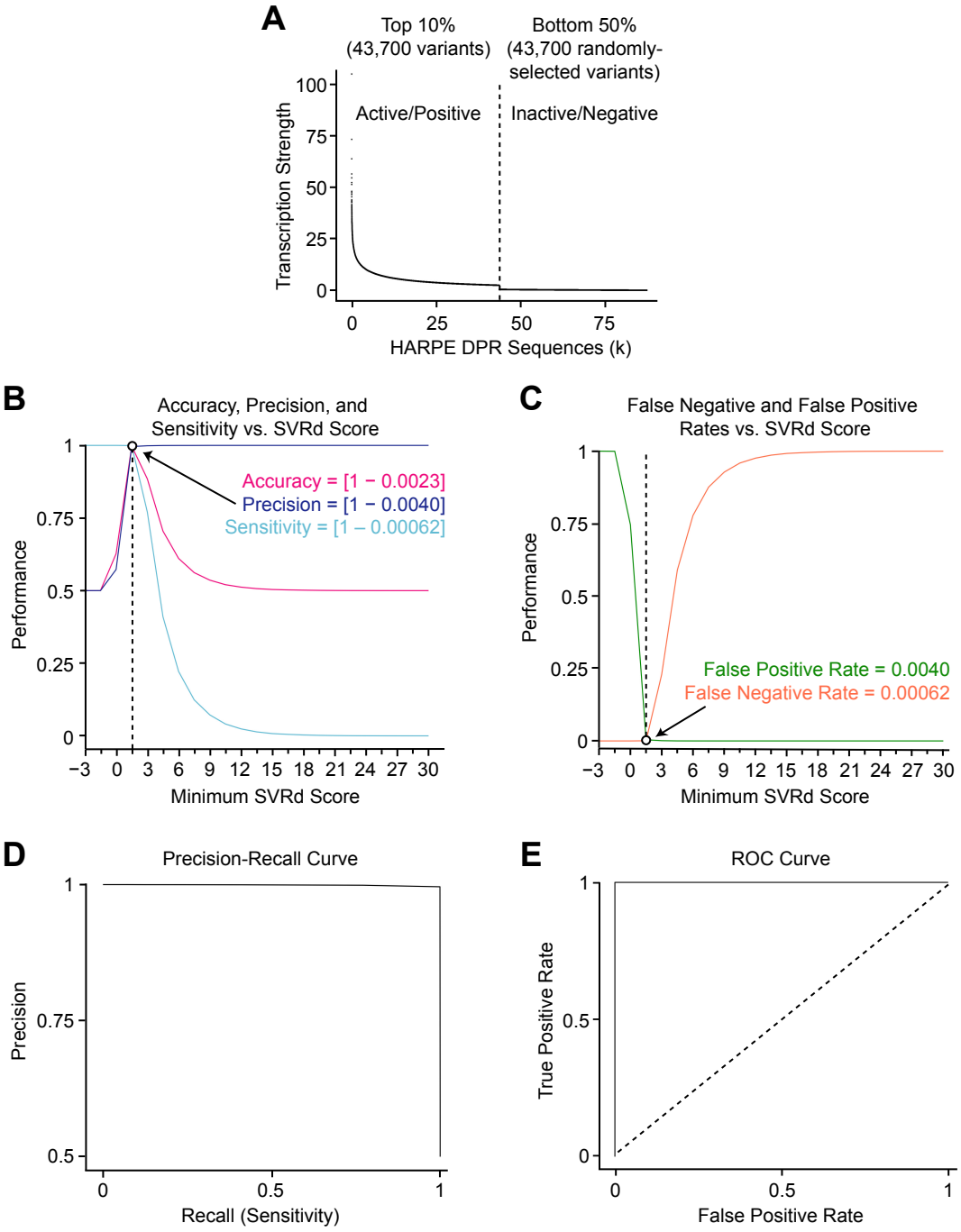
Long Vo ngoc, Torrey E. Rhyne, and James T. Kadonaga



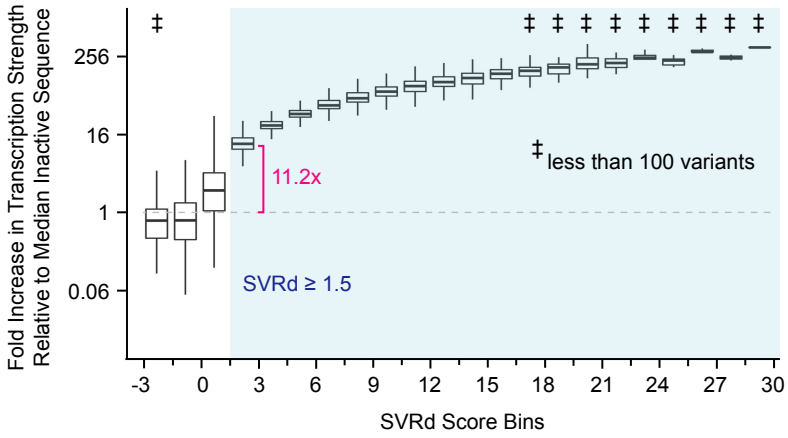
Supplemental Figure S1. Reproducibility of HARPE analysis of the *Drosophila* DPR, which spans from +17 to +35 relative to the +1 transcription start site. HARPE was performed as depicted in Fig. 1A and as described in Vo ngoc et al. (2020). The graph shows the results from two independent biological replicates. PCC, Pearson's correlation coefficient with two-tailed P-value < 2.2×10^{-18}.



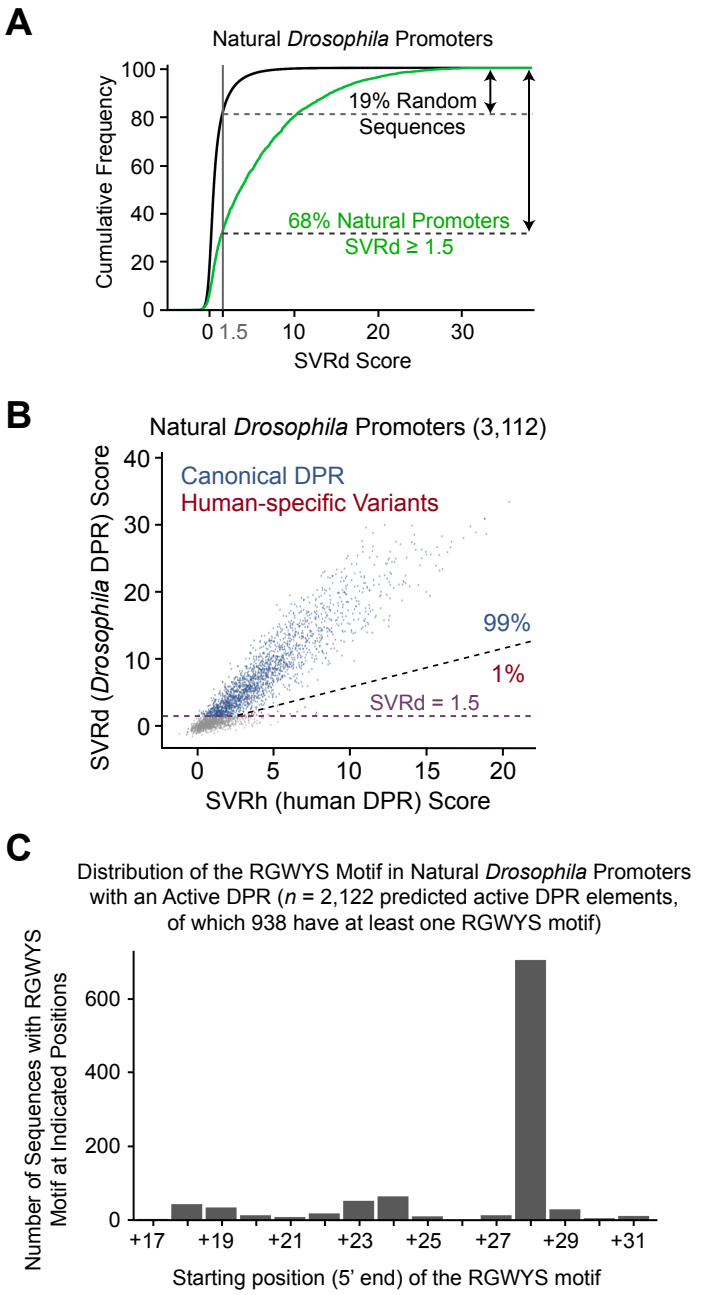
Supplemental Figure S2. SVRd DPR model optimization and cross validation. (*A*) Grid search results with different hyperparameter values for the cost of misclassification (cost) and individual training example influence (gamma). Each matrix displays the values of Spearman's rank correlation coefficient (rho; upper panels) or Pearson's correlation coefficient (PCC; lower panels) for SVR models that were generated with the indicated values of cost and gamma with the test dataset (left panels) or the validation dataset (right panels), which are separate halves of the 7,115 independent test sequences (which were not used to train the SVR models) shown in Fig. 1D. Undefined (UD) correlation is observed when the prediction of a model is constant regardless of the sequence. The SVRd model was trained as described in the Materials and methods. The hyperparameter values that were selected for SVRd are cost = 10 = E1 and gamma = 0.1 = E−1. (*B*) Comparison of the observed transcription strengths (HARPE data) and the predicted transcription strengths (assessed with the SVRd model generated with cost = 10 and gamma = 0.1) with the test dataset (left) as well as the validation dataset (right).



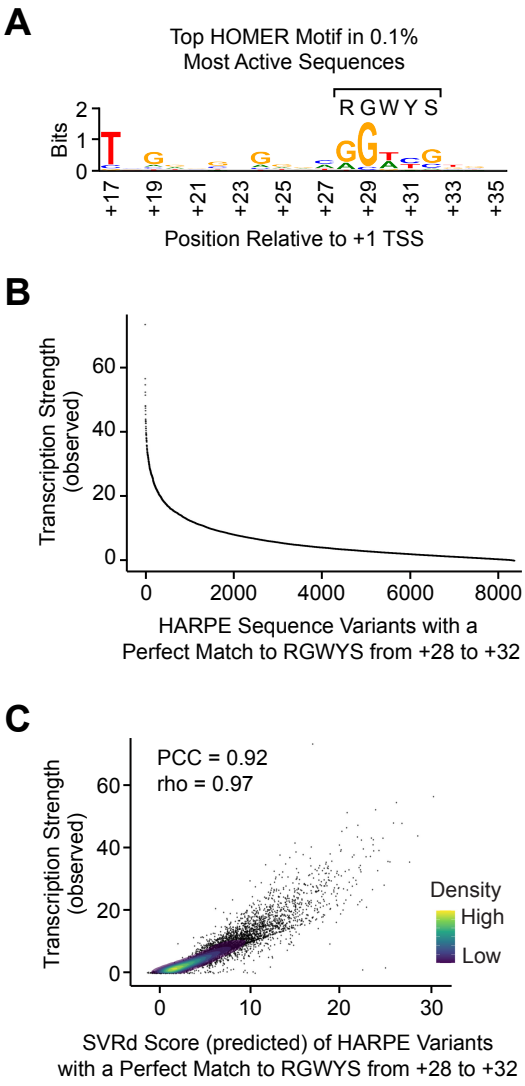
Supplemental Figure S3. Performance assessment reveals that the SVRd score of 1.5 is the best threshold for distinguishing active (SVRd ≥ 1.5) versus inactive (SVRd < 1.5) DPR elements. (*A*) Selection of the HARPE sequence variants for the performance assessment. The top 10% sequence variants were designated as active/positive for transcription, and an equal (randomly selected) number of the bottom 50% of sequence variants were designated as inactive/negative for transcription. These sequences were then used in the performance assessment. Intermediate variants that were between the top and bottom groups were not included. The transcription strengths (average of two HARPE replicates; $n = 2$ biologically independent samples) of all selected sequences are shown. (*B*, *C*) Performance measures relative to the minimum SVRd score required for a positive prediction. Performance was computed by counting true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Accuracy [(TP+TN) / (TP+FP+TN+FN)] reflects how often SVRd predictions are correct. Precision [TP / (TP + FP)] is the proportion of positive predictions that are correct. Sensitivity or recall or true positive rate [TP / (TP + FN)] is the proportion of transcriptionally active variants that are correctly predicted as positives. False positive rate [FP / (FP + TN)] is the probability for an inactive sequence to be incorrectly predicted as positive. False negative rate [FN / (FN + TP)] = (1 – Sensitivity) is the probability for an active sequence to be incorrectly predicted as negative. Performance is optimal at SVRd score ≥ 1.5. (*D*) Precision-recall (PR) curve. (*E*) Receiver operating characteristic (ROC) curve.



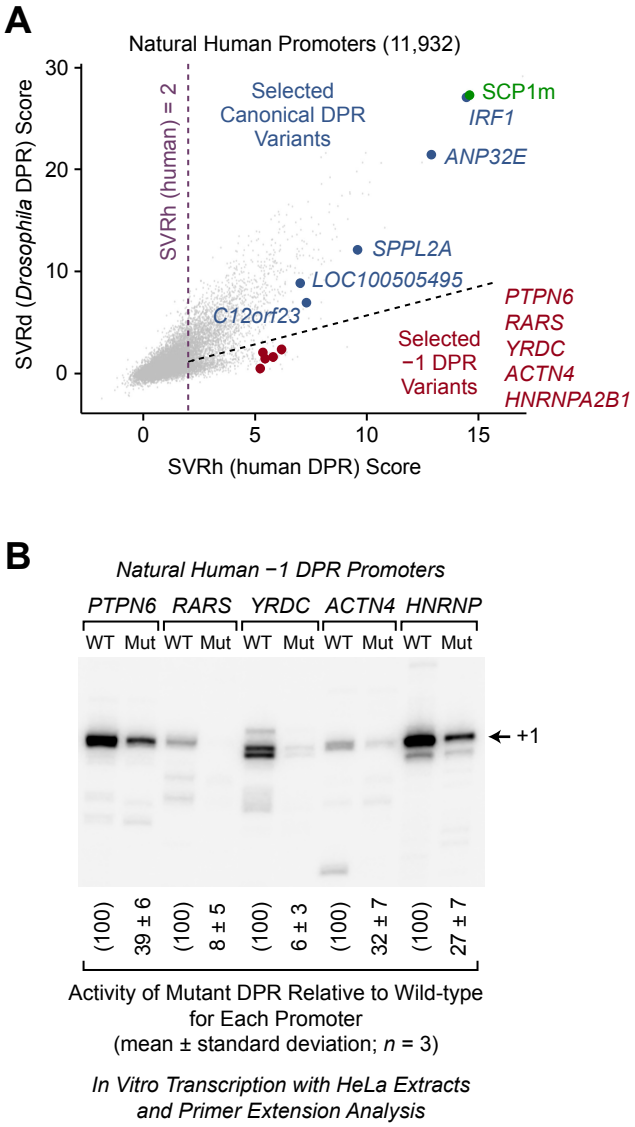
Supplemental Figure S4. DPR sequences with SVRd score ≥ 1.5 are at least 11-fold more active than the median inactive DPR sequence. This figures shows a box-plot diagram of the transcription strengths of all HARPE sequence variants placed into bins of the indicated SVR score ranges. For example, the highlighted bin contains SVRd scores from 1.5 to 3.0. Sequence variants with SVRd scores ≥ 1.5 (light blue shaded region) are typically at least about 11 times more active than an inactive sequence. The thick horizontal lines are the medians, and the lower and upper hinges are the first and third quartiles, respectively. Whiskers extend from the hinges to the largest or lowest value no further than 1.5 * IQR from the hinge. Data beyond the end of the whiskers (outlying points) are omitted from the box-plot. Variants with transcription strength = 0 were removed to allow log-scale display of the diagram. The horizontal dashed grey line denotes the median transcription strength of inactive sequences.



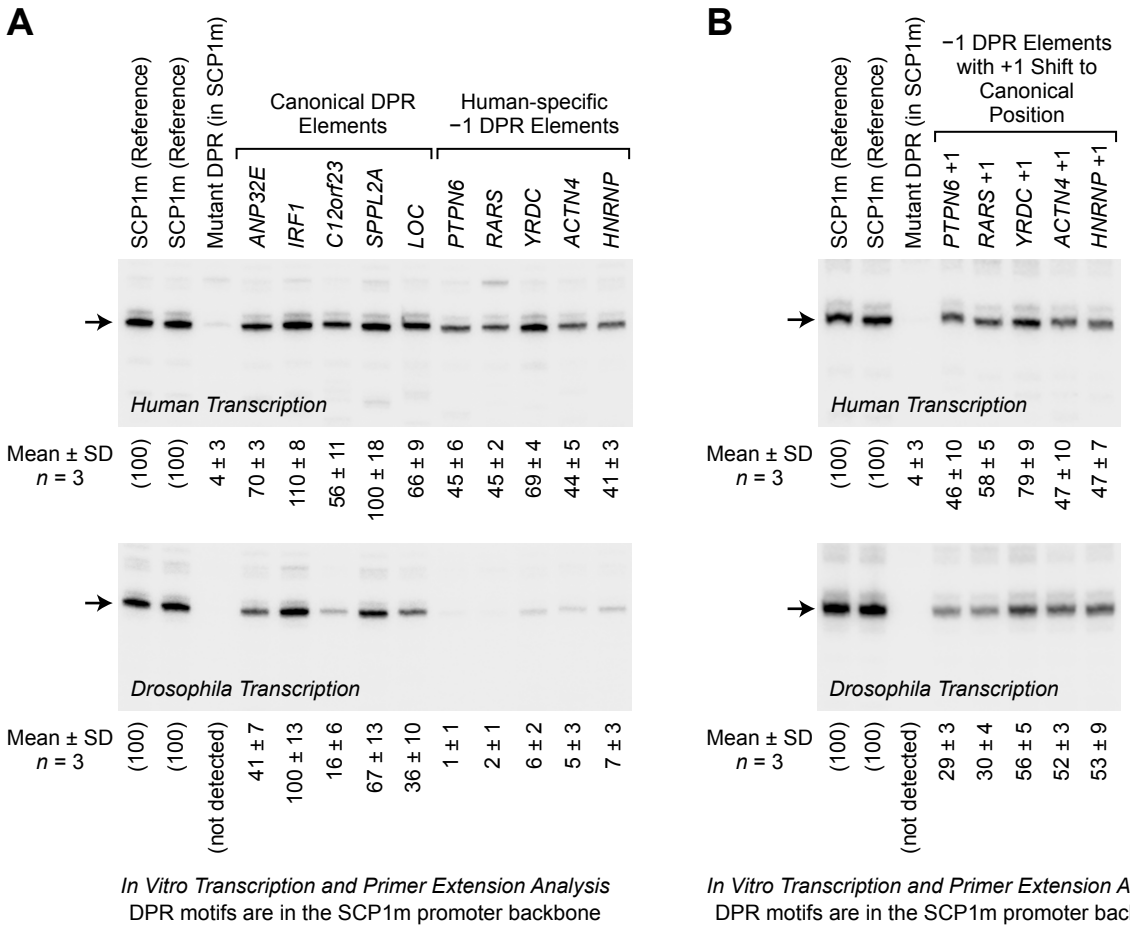
Supplemental Figure S5. Analysis of *Drosophila* nascent TSS data from 5'-GRO-seq experiments that were performed with *Drosophila* S2 cells (GEO accession number GSE68677). Focused TSSs in the nascent *Drosophila* TSS data in GSE68677 were identified as described in Vo ngoc et al. (2017) by using Focus_TSS.py with a focus index (FI) ≥ 0.67. This analysis yielded 3,112 focused promoters. (*A*) Approximately two-thirds of natural *Drosophila* promoters are predicted to contain an active DPR. The graph shows the cumulative frequencies of SVRd scores of sequences in the DPR region (+17 to +35) in 3,112 natural *Drosophila* promoters that are active in *Drosophila* S2 cells. This analysis revealed that approximately 68% (2,122 out of 3,112) of *Drosophila* promoters in S2 cells are predicted to have an active DPR (SVRd score ≥ 1.5; Supplemental Figs. S3 and S4). In contrast, only about 19% of 500,000 random 19 nt sequences with the same overall G/C content (52.6%) as *Drosophila* core promoters are predicted to have DPR activity. To examine the variability of the % DPR usage relative to the degree of focus in the TSSs, we determined the % DPR usage in promoters in which the minimum FI varies from 0.65 to 0.85, and observed a range of 67% to 74% DPR usage in S2 cells. This figure is related to main Fig. 1E. (*B*) The human-specific −1 DPR variant appears to be present in about 1% of 2,122 predicted active DPR elements (SVRd ≥ 1.5; dashed light line; Supplemental Figs. S3 and S4) in 3,112 natural *Drosophila* promoters. The black diagonal dashed line demarcates the canonical DPR sequences (blue dots) and the human-specific −1 DPR variants (red dots). This figure is related to main Fig. 2D. (*C*) The RGWYS motif in 2,122 predicted active DPR elements (of which 938 have at least one RGWYS motif) in natural *Drosophila* promoters is found most commonly from +28 to +32 (705 out of 938) and very rarely from +27 to +31 (15 out of 938). This figure is related to Supplemental Fig. S9C.



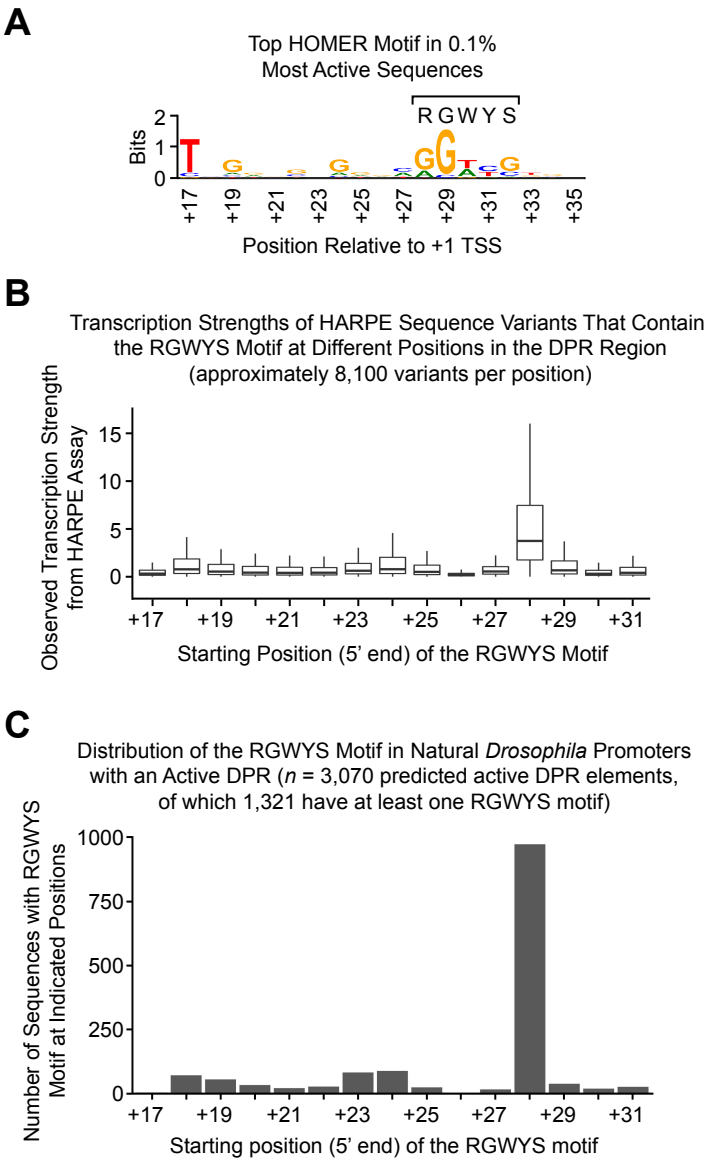
Supplemental Figure S6. SVR analysis incorporates information that is not encapsulated in a consensus sequence. (*A*) HOMER analysis of the 0.1% most active *Drosophila* DPR sequences reveals an RGWYS consensus sequence from +28 to +32 relative to the +1 TSS. This figure is identical to Fig. 1C of the main text. (*B*) HARPE variants with a perfect match to the RGWYS consensus sequence exhibit transcription strengths that range from highly active to inactive. Shown are the 8,351 out of 437,002 DPR variants that have a perfect match to RGWYS from +28 to +32. The variants are ranked along the x-axis in order of decreasing activity. (*C*) SVRd accurately predicts the range of transcription strengths of the same sequence variants, as shown in *B*, with a perfect match to the RGWYS consensus. PCC, Pearson's correlation coefficient; rho, Spearman's rank correlation coefficient.



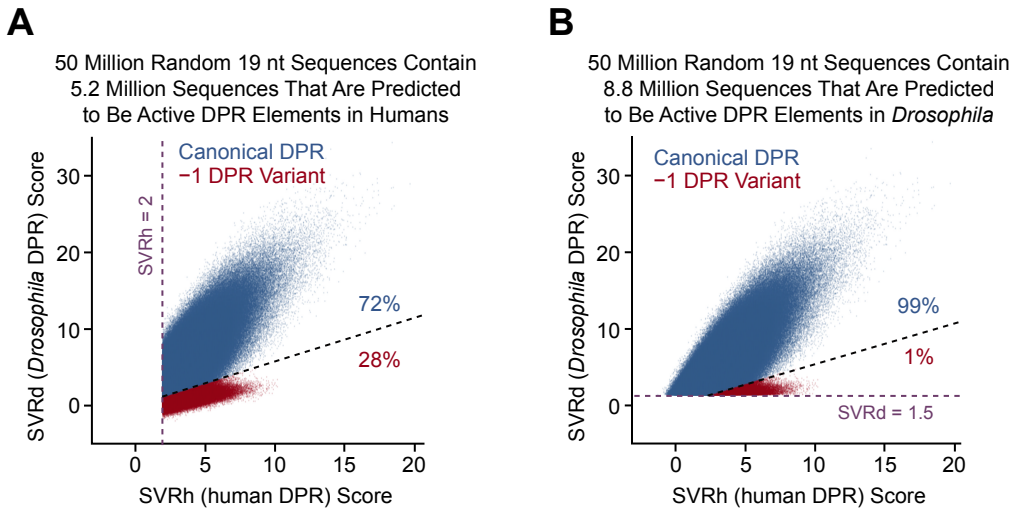
Supplemental Figure S7. Individual human-specific –1 DPR motifs are active in the context of natural human promoters. (*A*) Five human genes with predicted canonical DPR elements (blue dots) and five human genes with predicted –1 DPR variants (red dots) are highlighted. The synthetic optimized DPR in the SCP1m promoter (Juven-Gershon et al. 2006) is also shown. The diagram shows the predicted strengths of the DPR motifs in humans (SVRh) and in *Drosophila* (SVRd). (*B*) The human-specific –1 DPR motifs are functionally important in the context of natural human promoters from –36 to +50 relative to the +1 TSS. The wild-type (WT) and mutant (Mut) versions of the five –1 DPR promoters shown in *A* were subjected to in vitro transcription analysis with HeLa extracts. In the mutant promoters, the DPR regions (+17 to +35 relative to the +1 TSS) were replaced with an inactive mutant DPR sequence (+17 TATAGCCTAGGCTCCTTGC +35; SVRh = 0.3). The transcription experiments indicated that the –1 DPR motifs are important for the activity of these promoters. Three biologically independent series of transcription reactions were performed, and the data are given as the mean ± standard deviation. A representative autoradiogram is shown. The *HNRNPA2B1* gene is abbreviated as *HNRNP*. The quantitated data from all experiments are given in Supplemental Table S1.



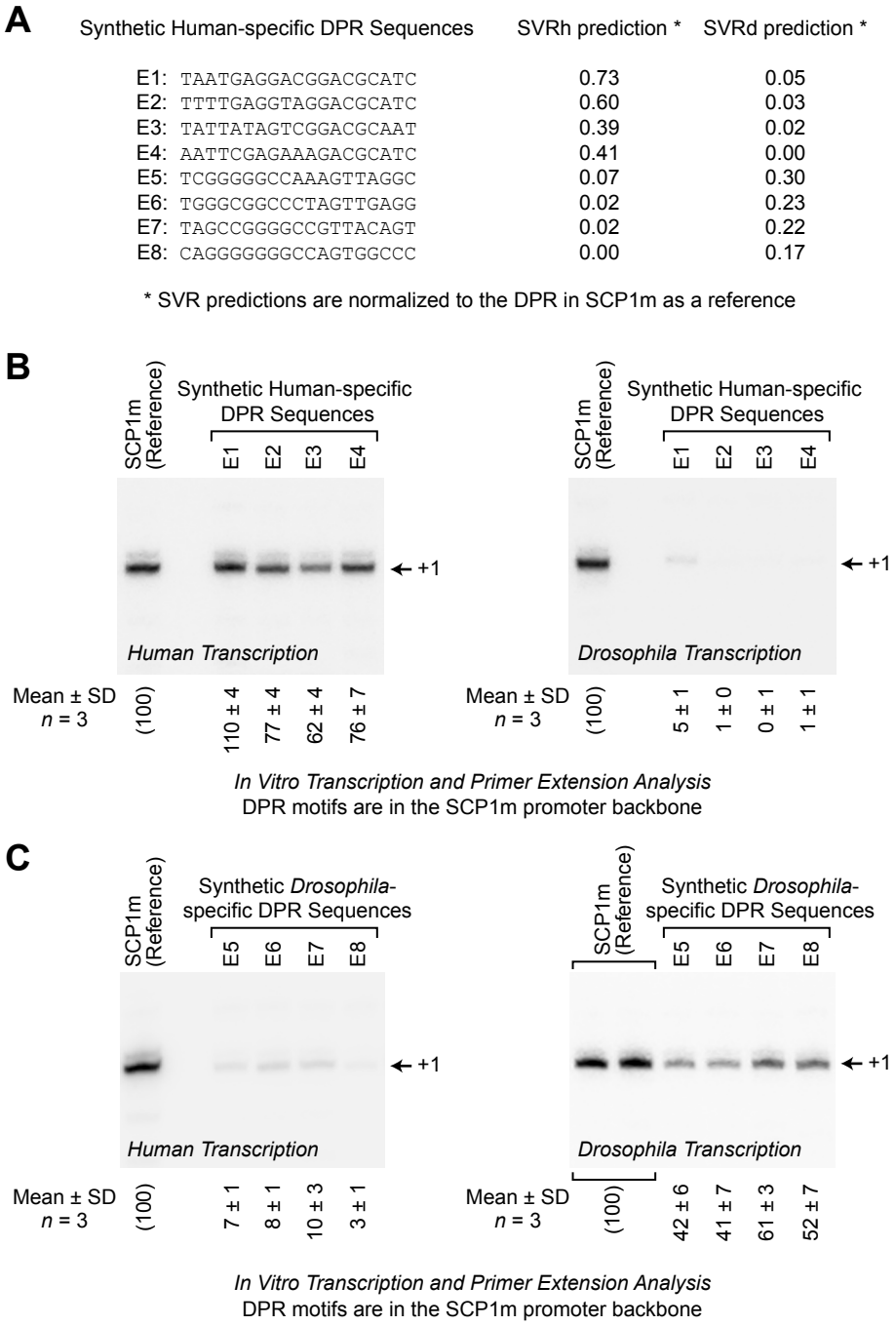
Supplemental Figure S8. The –1 DPR elements function specifically in humans. Five predicted canonical DPR elements (blue dots) and five predicted –1 DPR variants (red dots) from natural human genes were selected for functional transcription analyses (Supplemental Fig. 7A). The SCP1m promoter, which contains an optimized synthetic DPR and lacks a TATA box (Juven-Gershon et al. 2006), was used as a reference. (*A*) The human-specific –1 DPR element functions with human transcription factors but not with *Drosophila* transcription factors. Representative autoradiograms are shown. The data from three independent experiments are given as the mean ± standard deviation and in the graphs shown in Figs. 3A and 3B. (*B*) The translocation of –1 DPR motifs to the canonical DPR position results in a substantial increase in activity with *Drosophila* transcription factors. Representative autoradiograms are shown. The data from three independent experiments are given as the mean ± standard deviation and in the graphs shown in Figs. 3C and 3D. The *LOC10050549* and *HNRNPA2B1* genes are abbreviated as *LOC* and *HNRNP*. The quantitated data from all experiments are given in Supplemental Table S1. The DPR sequences that were used in each of these experiments are given in Supplemental Table S2.



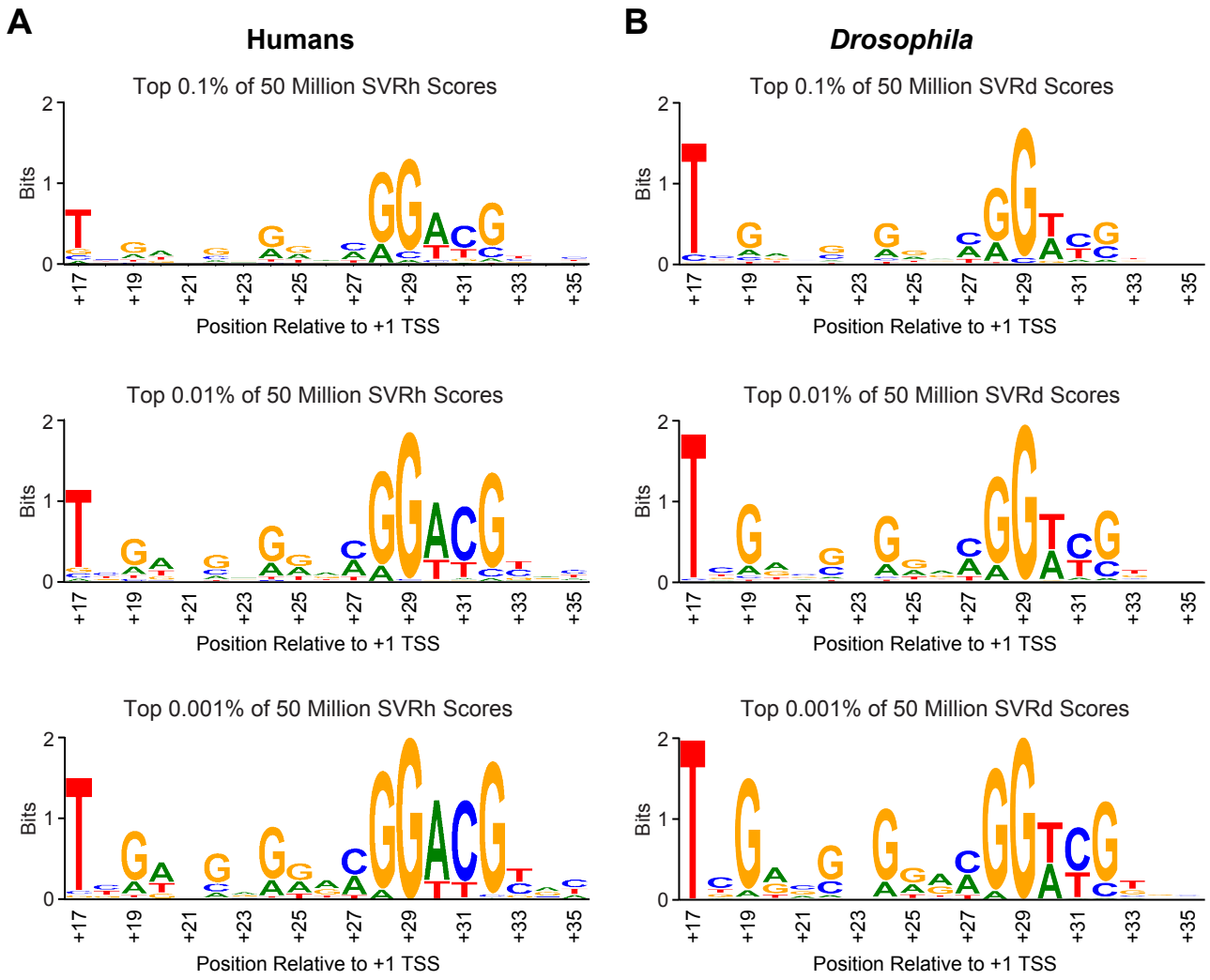
Supplemental Figure S9. The DPE-like RGWYS motif in the *Drosophila* DPR exhibits a strong preference for the canonical +28 to +32 position relative to the −1 nt shifted +27 to +31 position. (*A*) HOMER analysis of the top 0.1% most active DPR sequences in *Drosophila* reveals a DPE-like RGWYS motif from +28 to +32 relative to the +1 TSS. This figure is identical to main Fig. 1C and is included as a reference. (*B*) Analysis of the observed transcription strengths of DNA sequence variants reveals a strong preference for the RGWYS motif at position +28 to +32. This figure shows box-plot diagrams of the observed transcription strengths for all variants in the HARPE dataset that contain the RGWYS motif at each of the indicated positions. The horizontal lines are the medians, and the lower and upper hinges are the first and third quartiles. Each upper (or lower) whisker extends from the upper (or lower) hinge to the largest (or smallest) value no further than 1.5 x IQR from the hinge. Data beyond the end of the whiskers (outlying points) are omitted from the box plot. (*C*) The RGWYS motif in 3,070 predicted active DPR elements (of which 1,321 have at least one RGWYS motif) in natural *Drosophila* promoters (GEO accession number GSE203135; Delos Santos et al. 2022) is found most commonly from +28 to +32 (971 out of 1,321) and very rarely from +27 to +31 (16 out of 1,321).



Supplemental Figure S10. Analysis of 50 million random DNA sequences with the SVR models for the human and *Drosophila* DPR elements. The DPR scores of 50 million random 19 nt DNA sequences were determined for humans (SVRh) and for *Drosophila* (SVRd). The lines that demarcate the canonical DPR elements and the −1 DPR variants are the same as those used in Fig. 2. (*A*) 5.2 million out of the 50 million random 19 nt sequences are predicted to be active DPR motifs in humans, as assessed by an SVRh score of at least 2. About 72% of the predicted active DPR sequences are canonical motifs, whereas about 28% of the predicted active DPR sequences are the −1 DPR variants. Predicted inactive sequences are not shown. (*B*) 8.8 million out of the 50 million random 19 nt sequences are predicted to be active DPR motifs in *Drosophila*, as assessed by an SVRd score of at least 1.5. About 99% of the predicted active DPR sequences are canonical motifs, whereas about 1% of the predicted active DPR sequences are the −1 DPR variants. Predicted inactive sequences are not shown.



Supplemental Figure S11. Identification of synthetic DPR sequences with specificity for transcription with human factors relative to *Drosophila* factors, and vice versa. (*A*) DNA sequences and predicted SVR scores of synthetic DPR motifs with high SVRh:SVRd scores or high SVRd:SVRh scores. E1 was designed based on the HOMER motif of the top SVRh:SVRd scoring sequences, as shown in Fig. 4A. The other DPR sequences were identified in the analysis of 50 million random DPR sequences, as depicted in Fig. 4B. (*B*) The E1 to E4 high SVRh:SVRd DPR elements are much more active with human transcription factors than with *Drosophila* factors. (*C*) The E5 to E8 high SVRd:SVRh DPR elements are more active with *Drosophila* transcription factors than with human factors. The synthetic E1 to E8 DPR sequences in *B* and *C* were analyzed in the SCP1m promoter backbone, and their transcriptional activities with human factors and with *Drosophila* factors with were compared to that of the reference DPR in SCP1m (Juven-Gershon et al. 2006). Representative autoradiograms are shown. The data from three independent experiments are given as the mean ± standard deviation and in the graph shown in Fig. 4C. The quantitated data from all experiments are given in Supplemental Table S1.



Supplemental Figure S12. SVR analysis of 50 million random 19 nt sequences with SVRh and SVRd reveals conservation of the canonical DPR element in humans and *Drosophila*. (*A*) HOMER web logos for the human DPR, as assessed with the top 0.1%, 0.01%, and 0.001% SVRh scores in 50 million random sequences. (*B*) HOMER web logos for the *Drosophila* DPR, as assessed with the top 0.1%, 0.01%, and 0.001% SVRd scores in 50 million random sequences.

Supplemental Table S1: Quantitation of in vitro transcription data

Figure 3A / S8A: Human Transcription with Canonical vs. −1 DPR Variants in SCP1m Promoter Backbone*

Promoter	Rep 1	Rep 2	Rep 3	Average	SD	SVRh prediction
mDPR	0.01	0.06	0.04	0.04	0.03	0.02
ANP32E	0.67	0.71	0.73	0.70	0.03	0.88
IRF1	1.02	1.04	1.16	1.1	0.08	0.99
C12orf23	0.44	0.6	0.64	0.56	0.11	0.5
SPPL2A	0.81	1.17	1.05	1.0	0.18	0.66
LOC100505495	0.57	0.65	0.75	0.66	0.09	0.48
PTPN6	0.41	0.41	0.52	0.45	0.06	0.37
RARS	0.44	0.43	0.47	0.45	0.02	0.36
YRDC	0.7	0.65	0.72	0.69	0.04	0.42
ACTN4	0.42	0.4	0.5	0.44	0.05	0.4
HNRNPA2B1	0.38	0.44	0.4	0.41	0.03	0.37

Figure 3B / S8A: Drosophila Transcription with Canonical vs. −1 DPR Variants in SCP1m Promoter Backbone*

Promoter	Rep 1	Rep 2	Rep 3	Average	SD	SVRd prediction
mDPR	**	**	**	**	**	0.01
ANP32E	0.37	0.36	0.49	0.41	0.07	0.79
IRF1	1.02	0.89	1.15	1.0	0.13	0.99
C12orf23	0.14	0.11	0.22	0.16	0.06	0.26
SPPL2A	0.60	0.59	0.82	0.67	0.13	0.45
LOC100505495	0.25	0.37	0.45	0.36	0.10	0.33
PTPN6	0.01	0.01	0.02	0.01	0.01	0.08
RARS	0.02	0.02	0.01	0.02	0.01	0.02
YRDC	0.05	0.05	0.09	0.06	0.02	0.09
ACTN4	0.04	0.03	0.08	0.05	0.03	0.06
HNRNPA2B1	0.06	0.04	0.1	0.07	0.03	0.06

Figure 3C / S8B: Effect of +1 nt Downstream Shift of −1 DPR Variants with Human Transcription Factors*

Promoter	Rep 1	Rep 2	Rep 3	Average	SD	SVRh prediction
PTPN6	0.41	0.41	0.52	0.45	0.06	0.37
PTPN6 +1nt DPR	0.36	0.55	0.47	0.46	0.10	0.39
RARS	0.44	0.43	0.47	0.45	0.02	0.36
RARS +1nt DPR	0.59	0.62	0.53	0.58	0.05	0.45
YRDC	0.7	0.65	0.72	0.69	0.04	0.42
YRDC +1nt DPR	0.7	0.88	0.8	0.79	0.09	0.49
ACTN4	0.42	0.4	0.5	0.44	0.05	0.4
ACTN4 +1nt DPR	0.36	0.55	0.51	0.47	0.10	0.41
HNRNPA2B1	0.38	0.44	0.4	0.41	0.03	0.37
HNRNPA2B1 +1nt DPR	0.43	0.55	0.44	0.47	0.07	0.47

Figure 3D / S8B: Effect of +1 nt Downstream Shift of −1 DPR Variants with Drosophila Transcription Factors*

Promoter	Rep 1	Rep 2	Rep 3	Average	SD	SVRd prediction
PTPN6	0.01	0.01	0.02	0.01	0.01	0.08
PTPN6 +1nt DPR	0.3	0.31	0.25	0.29	0.03	0.34
RARS	0.02	0.02	0.01	0.02	0.01	0.02
RARS +1nt DPR	0.31	0.33	0.25	0.30	0.04	0.28
YRDC	0.05	0.05	0.09	0.06	0.02	0.09
YRDC +1nt DPR	0.56	0.61	0.52	0.56	0.05	0.53
ACTN4	0.04	0.03	0.08	0.05	0.03	0.06
ACTN4 +1nt DPR	0.54	0.53	0.48	0.52	0.03	0.54
HNRNPA2B1	0.06	0.04	0.1	0.07	0.03	0.06
HNRNPA2B1 +1nt DPR	0.62	0.52	0.44	0.53	0.09	0.44

Figure 4C / S12B: Human transcription of extreme human-specific candidate sequences *

Sequence	Rep 1	Rep 2	Rep 3	Average	SD	SVRh prediction
E1:TAATGAGGACGGACGCATC	1.07	1.08	1.01	1.1	0.04	0.73
E2:TTTTGAGGTAGGACGCATC	0.75	0.74	0.82	0.77	0.04	0.60
E3:TATTATAGTCGGACGCAAT	0.64	0.64	0.57	0.62	0.04	0.39
E4:AATTCGAGAAAGACGCATC	0.78	0.69	0.82	0.76	0.07	0.41

Figure 4C / S12B: Drosophila transcription of extreme human-specific candidate sequences *

Sequence	Rep 1	Rep 2	Rep 3	Average	SD	SVRd prediction
E1:TAATGAGGACGGACGCATC	0.04	0.05	0.06	0.05	0.01	0.05
E2:TTTTGAGGTAGGACGCATC	0.01	0.01	0.01	0.01	0.00	0.03
E3:TATTATAGTCGGACGCAAT	0.00	0.00	0.01	0.00	0.01	0.02
E4:AATTCGAGAAAGACGCATC	0.01	0.00	0.01	0.01	0.01	0.00

Figure 4C / S12C: Human transcription of extreme Drosophila-specific candidate sequences *

Sequence	Rep 1	Rep 2	Rep 3	Average	SD	SVRh prediction
E5:TCGGGGGCCAAAGTTAGGC	0.06	0.06	0.08	0.07	0.01	0.07
E6:TGGGCGGCCCTAGTTGAGG	0.09	0.08	0.07	0.08	0.01	0.02
E7:TAGCCGGGGCCGTTACAGT	0.09	0.13	0.07	0.10	0.03	0.02
E8:CAGGGGGGGCCAGTGGCCC	0.02	0.04	0.03	0.03	0.01	0.00

Figure 4C / S12C: Drosophila transcription of extreme Drosophila-specific candidate sequences *

Sequence	Rep 1	Rep 2	Rep 3	Average	SD	SVRd prediction
E5:TCGGGGGCCAAAGTTAGGC	0.36	0.48	0.43	0.42	0.06	0.30
E6:TGGGCGGCCCTAGTTGAGG	0.33	0.43	0.46	0.41	0.07	0.23
E7:TAGCCGGGGCCGTTACAGT	0.58	0.64	0.62	0.61	0.03	0.22
E8:CAGGGGGGGCCAGTGGCCC	0.55	0.56	0.44	0.52	0.07	0.17

Supplemental Figure S7B: Human Transcription of Mutant DPR Sequence in Natural Promoter Backgrounds***

Promoter	Rep 1	Rep 2	Rep 3	Average	SD
PTPN6 mDPR	0.40	0.32	0.44	0.39	0.06
RARS mDPR	0.08	0.13	0.04	0.08	0.05
YRDC mDPR	0.05	0.03	0.09	0.06	0.03
ACTN4 mDPR	0.30	0.27	0.40	0.32	0.07
HNRNPA2B1 mDPR	0.26	0.21	0.35	0.27	0.07

* All values and SVR predictions are normalized to SCP1m as a reference

** The SCP1m mDPR transcription signal is too weak to quantify in *Drosophila*

*** All values are normalized to the cognate wild-type natural promoters

Supplemental Table S2: DNA sequences in promoter constructs

Figure 3 / S8: Canonical DPR, human-specific -1 DPR, and +1 nt shifted -1 DPR sequences

Promoter	DPR Sequence (+17 to +35 relative to +1 TSS)
SCP1m (reference)	TCGAGCCGAGCAGACGTGC
mutant DPR	TATAGCCTAGGCTCCTTGC
ANP32E	TTGAAGGGGAAGGAACTGC
IRF1	TAGTCGAGGCAAGACGTGC
C12orf23	ACATCCTGAGAGGACGCCT
SPPL2A	GGGAACCGAGCAGACGCTC
LOC100505495	TGAAACCAACAGCACGCTC
PTPN6	GGATCGAGGAGGAAGTGGC
RARS	GCTGATGGGAGGATGGACG
YRDC	GGGCCTGGGCGGATGTCTC
ACTN4	AGGCGGGAGCGGACAGGCT
HNRNPA2B1	GGTCCCGTGCGGAGGTGCT
PTPN6 +1nt shifted	CGGATCGAGGAGGAAGTGG
RARS +1nt shifted	CGCTGATGGGAGGATGGAC
YRDC +1nt shifted	CGGGCCTGGGCGGATGTCT
ACTN4 +1nt shifted	CAGGCGGGAGCGGACAGGC
HNRNPA2B1 +1nt shifted	CGGTCCCGTGCGGAGGTGC

Figure 4 / S12: Synthetic extreme human-specific DPR sequence candidates

Candidate	Synthetic DPR sequence (+17 to +35 relative to TSS)
E1	TAATGAGGACGGACGCATC
E2	TTTTGAGGTAGGACGCATC
E3	TATTATAGTCGGACGCAAT
E4	AATTCGAGAAAGACGCATC
E5	TCGGGGGCCAAAGTTAGGC
E6	TGGGCGGCCCTAGTTGAGG
E7	TAGCCGGGGCCGTTACAGT
E8	CAGGGGGGGCCAGTGGCCC

Figure S7: Natural human promoters

Promoter	Promoter Sequence (-36 to +50 relative to +1 TSS)
PTPN6	GGAGACTATTAGTCCAGGGTTTGTCCCTGCAGTGCCATTGGCCTGGCAGGCAGGATCGAGGAGGAAGTGGCTGATTACTGAGCGGT
RARS	GCTTCCGGGAGAGGCTGACCGTTTCCGCTTCCGTCCACTTGGCGAGTGAGACGCTGATGGGAGGATGGACGTACTGGTGTCTGAGT
YRDC	CCGAGTCTCCTGGACCGGAAGCTGGCTGGGAGCGTCACTTCCTCCCGGAAGCGGGCCTGGGCGGATGTCTCCGGCGCGTCGGTGCA
ACTN4	AGCAGCTGAAGCGGCGGTAGCGGCGGCGGCTCGGGCAGAGGGGCGGGAGCTGAGGCGGGAGCGGACAGGCTGGTGGGCGAGCGAGA
HNRNPA2B1	AGGTTCTAGAAAAGCGGCGGCAGCGGCTCTAGCGGCAGTAGCAGCAGCGCCGGGTCCCGTGCGGAGGTGCTCCTCGCAGAGTTGTT

Supplemental Materials and Methods

HARPE procedure and data analysis

The HARPE procedure and data processing were performed for the DPR as described in Vo ngoc et al. (2020), with the exception that the in vitro transcription reactions were carried out with *Drosophila* nuclear extracts by the method of Wampler et al. (1990). The TATA-less SCP1m promoter backbone, which contains two GC boxes upstream of the mutated TATA box, is identical to that used in the analysis of the human DPR (Vo ngoc et al. 2020). The randomized DPR region (+17 to +35 relative to the +1 TSS) is also the same as that used in Vo ngoc et al. (2020).

The procedure was carried out as follows. The HARPE plasmid library was subjected to in vitro transcription (six standard reactions, 300 ng DNA each) with *Drosophila* extracts for 30 minutes. The transcription reactions were performed independently two times (*i.e.*, two sets of six reactions were performed independently) to ensure reproducibility of the data. For each set of six reactions, the combined RNA transcripts were extracted with Trizol™ LS (Thermo Fisher Scientific), and the contaminating plasmid DNA was removed with the TURBO DNA-free™ Kit (Thermo Fisher Scientific). After ethanol precipitation, the RNA was subjected to reverse transcription with SuperScript™ III Reverse Transcriptase (Thermo Fisher Scientific) and treatment with RNase H (New England Biolabs). The cDNA was extracted with phenol-chloroform-isoamyl alcohol, precipitated with ethanol, and size-selected on a 6% (w/v) polyacrylamide-8 M urea gel with radiolabeled size markers. The HARPE plasmid DNAs and the size-selected cDNAs were used as templates to generate DNA amplicons for Illumina sequencing by using custom forward oligonucleotides containing the Illumina P5 and Read1-primer sequences preceding the sequence corresponding to nucleotides +1 to +16 (relative to the +1 TSS) of the SCP1m promoter cassette. Reverse primers were selected from the NEBNext® Multiplex Oligos for Illumina® kits (New England Biolabs), which match the Illumina Read2-primer sequence present on the HARPE plasmid and corresponding cDNA. NGS PCR

amplicons were size-selected on native 6% (w/v) polyacrylamide gels prior to Illumina sequencing.

Single-read sequences (SR75) were required to have a perfect match to the 10 nt sequences immediately upstream and downstream of the randomized region, which was required to be exactly 19 nt. All reads that matched this pattern were deemed usable and trimmed for sequences outside of the randomized region. When present, highly abundant reads in the randomized box that correspond to the original promoter sequence or to invariant sequences from other constructs were discarded, as they likely originated from inaccurate indexing of other multiplexed samples. DNA and cDNA read counts were then computed for each variant. For the DNA datasets, we used only sequences with a minimum read count of 0.75 reads per million (RPM) so that low confidence variants would not be included in the analysis. RNA (cDNA) dataset sequences were then matched to the corresponding DNA dataset, which was used as a reference. Transcription strength was then defined as RNA tag count (in RPMs) divided by DNA tag count (in RPMs).

Data processing

Motif discovery was performed with Hypergeometric Optimization of Motif EnRichment (HOMER)25. findMotifs.pl was used to search for 19-nt motifs among the most transcribed HARPE sequences and the highest SVR-scoring random sequences. The top sequences were used to generate sequence logos with WebLogo 3. Variants randomly selected from all tested sequences were used as background.

All calculations were performed in the R environment (version 4.1.2) with R packages ggplot2 v3.3.6, tidyr v1.2.1, dplyr v1.0.10 and rlist v0.4.6.2, or with Microsoft Excel. All replicate measurements were taken from different samples.

The transcription start sites (TSSs) of the natural promoters were determined by using the Focus_TSS.py program (Vo ngoc et al. 2017) and minimum focus index (FI) values ranging from 0.65 to 0.85. The *Drosophila* embryo TSS data (GEO accession number GSE203135;

Delos Santos et al. 2022) were obtained by using Focus_TSS.py with FI ≥ 0.8. The TSS data from *Drosophila* S2 cells (GEO accession number GSE68677) were obtained by using Focus_TSS.py with FI ≥ 0.67. The human TSS data from HeLa cells (GEO accession number GSE63872; Duttke et al. 2015) were obtained by using Focus_TSS.py with FI ≥ 0.67.

Generation and use of the SVR models

Machine learning analyses were performed by using functions of the R package e1071 (version 1.7-2; https://CRAN.R-project.org/package=e1071). For SVR training, we used the default Radial Basis Function (RBF) kernel, which yielded the best results among those tested. Nucleotide variables for HARPE variants were computed as four categories (A, C, G, T), which are known as factors in R. These factors were used as the input features, and transcription strength was used as the output variable. We started with 437,002 sequences, and set aside 7,115 test sequences (with the full range of transcription strengths). With the remaining sequences, we trained SVRd with 100,000 of the most transcribed (Best) variants and 100,000 of randomly-selected Non-Best variants. Grid search was performed for the hyperparameters C (cost) and gamma, and cross validation was carried out with two independent test sets (two halves of the 7,115 test sequences) that were not used for the training (Supplemental Fig. S2).

We used SVRd for the *Drosophila* DPR and SVRb (Vo ngoc et al. 2020; GEO GSE139635; in this work, we designated SVRb as SVRh) for the human DPR. The SVR models can be used to predict transcription strength with R by using the predict() function included in CRAN package e1071. Models are imported with readRDS(). Query sequence data must be formatted as follows. The variable names are V1 to V19 for DPR models (corresponding to the 19 positions from +17 to +35). Query sequences are split with one nucleotide per column and one sequence per row. Each column must have at least one A, one C, one G, and one T to ensure that all variables are read as 4 categories (A, C, G, T). Prediction using an SVR model and a query sequence will return an output “SVR score” that is related to the transcription strength and set on an arbitrary scale.

To facilitate the use of the models, we previously provided an R script named SVRpredict.R, which requires R with CRAN packages e1071 and docopt (Vo ngoc et al. 2020; GEO GSE139635). SVRpredict.R inputs an SVR model file as well as a query sequence file (one sequence per line), and outputs a new file with each sequence and its associated predicted transcription strength in an added column (SVR_score). In this study, we now provide an updated SVRpredict.R script that allows additional information to be included in the input file (the columns must be tab-delimited, and the column containing the sequences must be specified).

Supplemental References

Delos Santos NP, Duttke S, Heinz S, Benner C. 2022. MEPP: more transparent motif enrichment by profiling positional correlations. *NAR Genom Bioinform* **4**: lqac075

Duttke SHC, Lacadie SA, Ibrahim MM, Glass CK, Corcoran DL, Benner C, Heinz S, Kadonaga JT, Ohler U. 2015. Human promoters are intrinsically directional. *Mol Cell* **57**: 674–684.

Vo ngoc L, Cassidy CJ, Huang CY, Duttke SHC, Kadonaga JT. 2017. The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters. *Genes Dev* **31**: 6–11.

Vo ngoc L, Huang CY, Cassidy CJ, Medrano C, Kadonaga JT. 2020. Identification of the human DPR core promoter element using machine learning. *Nature* **585**: 459–463.

Wampler SL, Tyree CM, Kadonaga JT. 1990. Fractionation of the general RNA polymerase II transcription factors from *Drosophila* embryos. *J Biol Chem* **265**: 21223–21231.