Computational identification and experimental characterization of preferred downstream positions in human core promoters

René Dreos[1,¶, #a], Nati Malachi[2,¶], Anna Sloutskin[2,¶], Philipp Bucher[1,3,*] and Tamar Juven-Gershon[2,*]

[1] Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

[2] The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, Israel

[3] School of Life Sciences, Swiss Federal Institute of Technology, Lausanne, Switzerland

[#a] Current address: Center for Integrative Genomics, University of Lausanne, CH-1015 Lausanne, Switzerland.

* Corresponding authors

E-mail: tamar.gershon@biu.ac.il (TJG)

E-mail: philipp.bucher@sib.swiss (PB)

¶ These authors contributed equally to this work.

Abstract

Metazoan core promoters, which direct the initiation of transcription by RNA polymerase II (Pol II), may contain short sequence motifs termed core promoter elements/motifs (*e.g.* the TATA box, initiator (Inr) and downstream core promoter element (DPE)), which recruit Pol II via the general transcription machinery. The DPE was discovered and extensively characterized in *Drosophila*, where it is strictly dependent on both the presence of an Inr and the precise spacing from it. Since the *Drosophila* DPE is recognized by the human transcription machinery, it is most likely that some human promoters contain a downstream element that is similar, though not necessarily identical, to the *Drosophila* DPE. However, only a couple of human promoters were shown to contain a functional DPE, and attempts to computationally detect human DPE-containing promoters have mostly been unsuccessful. Using a newly-designed motif discovery strategy based on Expectation-Maximization probabilistic partitioning algorithms, we discovered preferred downstream positions (PDP) in human promoters that resemble the *Drosophila* DPE. Available chromatin accessibility footprints revealed that *Drosophila* and human Inr+DPE promoter classes are not only highly structured, but also similar to each other, particularly in the proximal downstream region. Clustering of the corresponding sequence motifs using a neighbor-joining algorithm strongly suggests that canonical Inr+DPE promoters could be common to metazoan species. Using reporter assays we demonstrate the contribution of the identified downstream positions to the function of multiple human promoters. Furthermore, we show that alteration of the spacing between the Inr and PDP by two nucleotides results in reduced promoter activity, suggesting a strict spacing dependency of the newly discovered human PDP on the Inr. Taken together, our strategy identified novel functional downstream positions

within human core promoters, supporting the existence of DPE-like motifs in human promoters.

Author summary

Transcription of genes by the RNA polymerase II enzyme initiates at a genomic region termed the core promoter. The core promoter is a regulatory region that may contain diverse short DNA sequence motifs/elements that confer specific properties to it. Interestingly, core promoter motifs can be located both upstream and downstream of the transcription start site. Variable compositions of core promoter elements have been identified. The initiator (Inr) motif and the downstream core promoter element (DPE) is a combination of elements that has been identified and extensively characterized in fruit flies. Although a few Inr+DPE -containing human promoters have been identified, the presence of transcriptionally important downstream core promoter positions within human promoters has been a matter of controversy in the literature. Here, using a newly-designed motif discovery strategy, we discovered preferred downstream positions in human promoters that resemble fruit fly DPE. Clustering of the corresponding sequence motifs in eight additional species indicated that such promoters could be common to multicellular non-plant organisms. Importantly, functional characterization of the newly discovered preferred downstream positions supports the existence of Inr+DPE-containing promoters in human genes.

Introduction

Regulation of eukaryotic gene expression is critical for diverse biological processes, including embryonic development, differentiation, cell cycle progression and

apoptosis. Cellular signals that regulate gene expression affect many different factors and co-regulators, but the ultimate decision whether or not to initiate transcription occurs at the core promoter. The core promoter, which lies at the heart of transcription, is generally defined as the minimal region that directs the accurate initiation of transcription by RNA polymerase II (Pol II) [1-5].

There are three major modes of transcription initiation patterns: focused, dispersed and mixed [1-3, 5-9]. Focused (also termed “sharp”) promoters encompass from −40 to +40 relative to the transcription start site (TSS; referred to as +1), and contain a single predominant TSS or a few TSSs within a narrow region of several nucleotides. Focused transcription initiation is associated with spatiotemporally regulated genes. Because of the biological significance of regulated genes, focused initiation is the most studied mode of transcription initiation. Dispersed (also termed “broad”) promoters contain multiple weak start sites that span over 50 to 100 nucleotides. Dispersed transcription initiation is associated with constitutive or housekeeping genes. Mixed (also termed “broad with peak”) promoters combine the abovementioned modes by exhibiting a dispersed initiation pattern with a single strong transcription start site.

Interestingly, although the core promoter was previously regarded as a universal component of the transcription machinery, it is nowadays clear that core promoters differ both in their architecture and function [1, 3, 5, 10-12]. In addition, the core promoter composition was demonstrated to affect transcriptional output, thus demonstrating the regulatory role of the promoter sequence itself [13-16].

Metazoan focused core promoters may contain short DNA sequences termed core promoter elements/motifs. These motifs, such as the TFIID-bound elements TATA box, initiator (Inr), downstream core promoter element (DPE), motif ten

element (MTE) and the Bridge configuration, function as recognition sites for the basal transcription machinery that recruits Pol II and have a positional bias (reviewed in [1-5, 17, 18]). The function of the DPE, MTE and Bridge downstream motifs is exclusively dependent on a strictly-spaced functional Inr motif [19-22].

The DPE, MTE and Bridge motifs were discovered and extensively characterized in *Drosophila melanogaster* promoters [16, 19-32]. Although the conservation of the DPE and MTE from *Drosophila* to humans was demonstrated, only a few human promoters were shown to be dependent on a functional DPE strictly located at positions +28 to +32, relative to the A_{+1} of the Inr [20, 33, 34], and one review article even postulated that the DPE may be unique to *Drosophila* [3]. Nevertheless, as fruit flies are evolutionarily distant from humans, it is very likely that some human promoters contain a downstream core promoter element that is similar, but not identical to, *Drosophila* DPE.

TFIID is the first basal transcription factor that binds the core promoter and recruits Pol II and other basal transcription factors to initiate transcription [1, 4, 35-38]. The TAF1 and TAF2 subunits of TFIID subunits were previously implicated in binding the downstream core promoter region [39]. Remarkably, the downstream region of the super core promoter (SCP), a synthetic promoter that includes the TATA box, Inr, MTE and DPE [14], exhibits a robust transcriptional output in multiple human cell lines [14, 40], as compared to other commercially-available potent promoters. Mutating any of these 4 elements significantly reduces TFIID binding and the transcriptional output of the SCP [14, 41]. This observation strongly suggests that the transcription machinery in human cells recognizes downstream positions conforming to the *Drosophila*-defined DPE and MTE motif sequences. Moreover, based on recent cryo-electron microscopy (cryo-EM), it was suggested that the SCP

is bound by the TAF1, TAF2 and TAF7 subunits of human TFIID [42]. These findings imply that distinct human core promoters are recognized by the transcription machinery in human cells via specific nucleotides in the downstream core promoter region.

To identify preferred downstream positions in focused human core promoters, we designed a motif discovery strategy, using probabilistic partitioning algorithms, based on Expectation-Maximization model optimization.

This algorithm was applied to human and *Drosophila* core promoter regions comprising the base pairs from -10 to +40 relative to the TSS. Interestingly, we identified downstream overrepresented positions that resemble the *Drosophila* DPE motif. Available chromatin accessibility (ATAC-seq) footprints reveal that *Drosophila* and human Inr+DPE promoter classes resemble each other, especially in the proximal downstream region. Clustering analysis of the identified sequence motifs in ten species using a neighbor-joining algorithm indicated that canonical Inr+DPE -containing promoters could be common to metazoan species. Using dual-luciferase reporter assays we demonstrate the contribution of the identified downstream positions to the function of several human promoters. Furthermore, we show that the spacing between the preferred downstream positions and the Inr motif is important for human core promoter activity, as demonstrated for *Drosophila* promoters. Taken together, our motif discovery strategy identified novel functional downstream positions in human core promoters, supporting the existence of DPE-like motifs in the downstream region of human promoters that may serve as recognition sites for human TFIID.

Results

Evidence for preferred downstream positions that resemble the DPE, in human promoters

The DPE motif is readily identified in *Drosophila* [16, 19, 20, 22, 23, 25, 26, 28, 30-32], and there is unquestionable evidence that *Drosophila* DPE motifs are recognized by the human transcription machinery *in vitro* and in multiple human cell lines [14, 20, 33, 34, 41]. Nevertheless, attempts to computationally identify a corresponding sequence motif in human promoters have been controversial [3]. Applying the basic probabilistic partitioning algorithm illustrated in Fig 1A, we can easily identify a DPE motif in *Drosophila* promoters (Fig 2). Partitioning *Drosophila* promoters into three subclasses, we obtained one class containing both a canonical Inr and DPE motif (Class 1), a second one containing only an Inr motif (Class 2), and a third one containing a weak non-canonical Inr motif featuring G and A at about equal frequency at the TSS, which is preferentially flanked by T's on both sides (Class 3). Applying the same algorithm to human promoters, the results were somewhat different from the results of the run on *Drosophila* promoters: we identified a class containing a strong canonical Inr motif (Class 1) and another one containing a surprisingly similar weak non-canonical Inr motif (Class 2). A third identified class had almost no conserved base positions, except a weak preference for a purine at the TSS (Class 3).

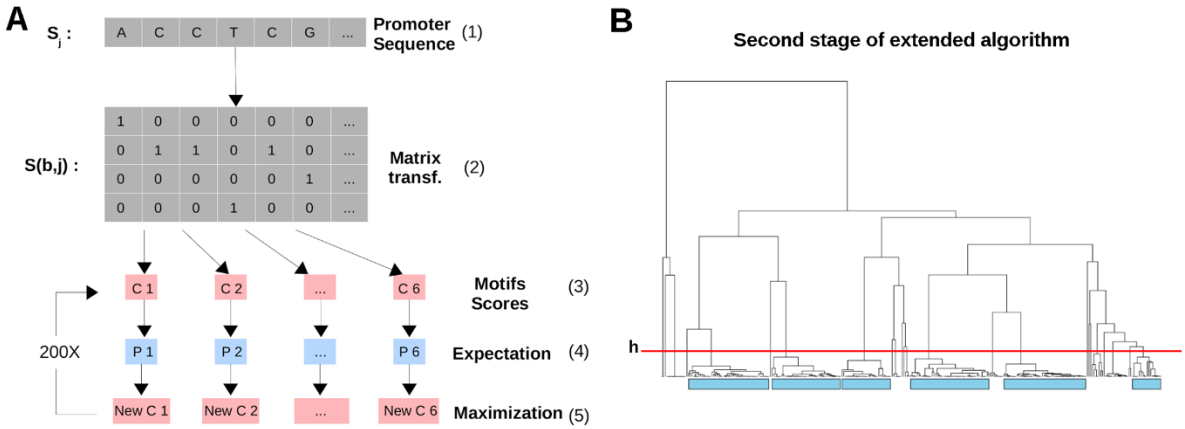


Fig 1. EM algorithms implemented in this study. (A) Diagram of the simple probabilistic partitioning basic algorithm. A promoter sequence S_i (1) is transformed into a binary matrix *S(b,j)* (2) following guidelines in [43] where each row represents one of the four bases *b* (A, C, G and T). Each element of the matrix *S(b,j)* has a value of 1 if the corresponding base is present at position *j* in the sequence. The matrix is then scored against *K* number of motifs (in this example K=6, C1 to C6) (3) to generate a probability score for each motif (P1 to P6) (4). In the first cycle, the motifs are generated using a random seeding strategy where the sequence probabilities follow a beta distribution. Next, each motif consensus is updated using the promoter sequences in conjunction with their probabilities (New C1 to New C6) (5). This cycle is repeated a number of times (in this example 200 times) to obtain the final motifs. (B) Probabilistic partitioning extended algorithm. All steps in A are repeated a number of times (in this example 50 times) to generate 300 motifs. These are then clustered hierarchically. The resulting tree is cut at a specific height (*h*, here at distance equal to 0.5) and the K nodes comprising the largest amount of motifs (identified by cyan rectangles) are retained and averaged to generate the final motifs.

It is important to remember in this context that at least three human promoters, namely IRF1, CALM2 and TAF7 (TAFII55), were experimentally shown to have functional DPE motifs [20, 33, 34]. In line with this, human class 1 promoters seemed to contain a very weak preference for nucleotides in positions +28, +29, which prompted us to develop a more refined algorithm. One potential limitation of the basic probabilistic partitioning algorithm is that it appears to have a tendency to split the input sequences into classes of similar sizes, as can be inferred from the frequencies presented in Fig 2. If we hypothesize that the DPE motif occurs only in a very small subclass of human promoters, the corresponding sequence motif may simply be hidden in one or several of the abundant subclasses shown in Fig 2. To

test this hypothesis, we modified the basic algorithm to favor the discovery of low frequency classes with highly skewed base composition (Fig 1B, Methods section).

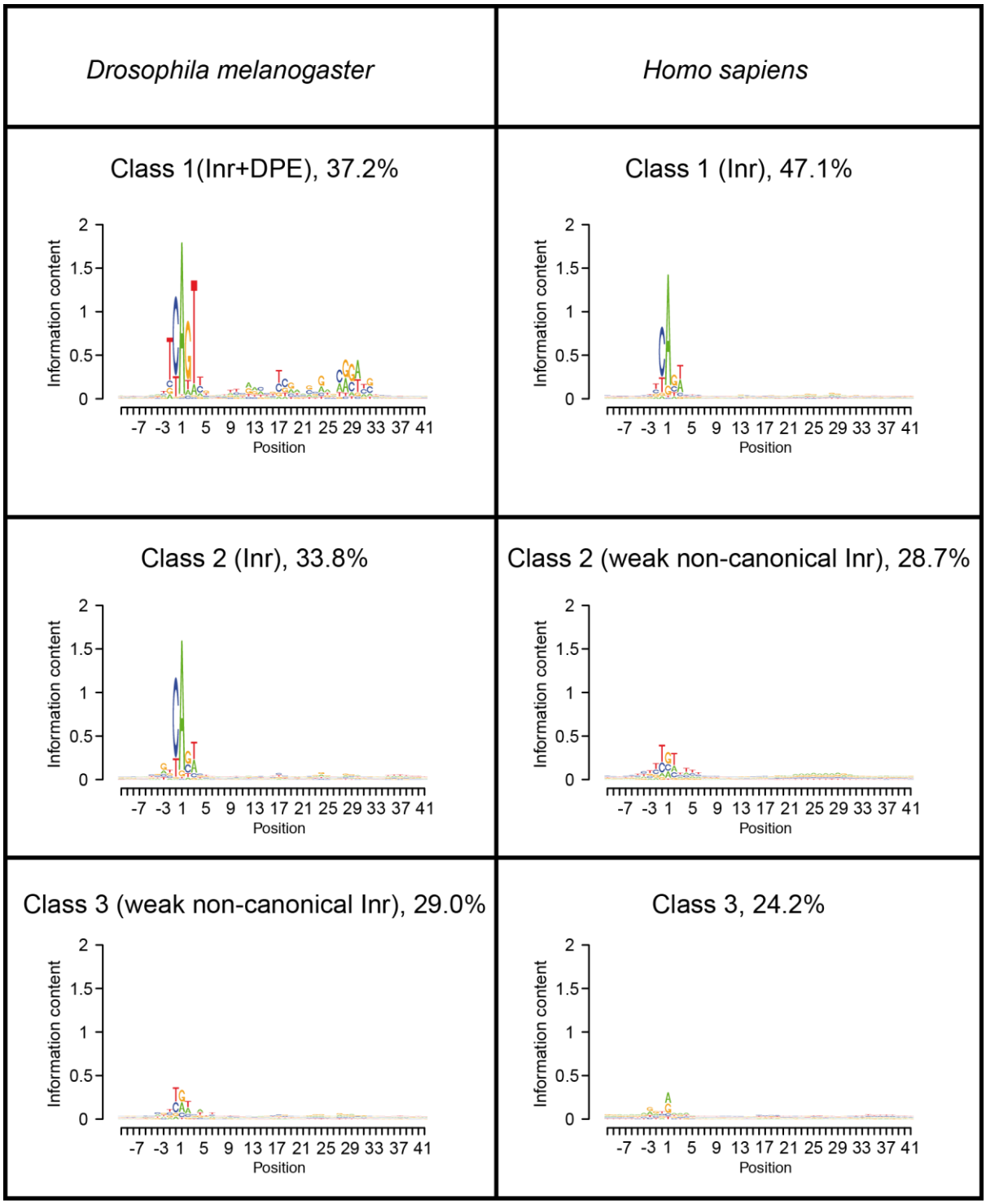


Fig 2. Partitioning of promoter sequences using the basic probabilistic partitioning EM algorithm. Three major classes with distinct core promoter compositions were identified within *Drosophila melanogaster* and human promoters. For each class, its frequency among the examined promoters is indicated.

Applying the new algorithm to human promoter sequences, a stable partitioning was achieved with 6 classes (Fig 3). The vast majority of promoters (87.9%) fall into a major class showing a very weak initiator motif, essentially consisting of a purine at the TSS preceded by a pyrimidine, previously termed a YR_{+1} initiator [3, 7, 44]. This class is reminiscent of class 1 obtained with the basic partitioning algorithm. The second most frequent class (4.4%) contains another known element, the TCT motif [45], which is found in promoters of ribosomal protein genes and other genes related to translation. The third most frequent class (3.0%) very much resembles the Inr+DPE class found in *Drosophila*. In particular, positions 28-32 relative to the A_{+1} of the Inr, show almost identical base preferences between the two species. The remaining three classes show the same trinucleotide-repeat pattern $(GCN)_n$ in three different frames relative to an initiator motif consisting mostly of a purine at the TSS preceded by a pyrimidine. To our knowledge, this is a new pattern of unknown function.

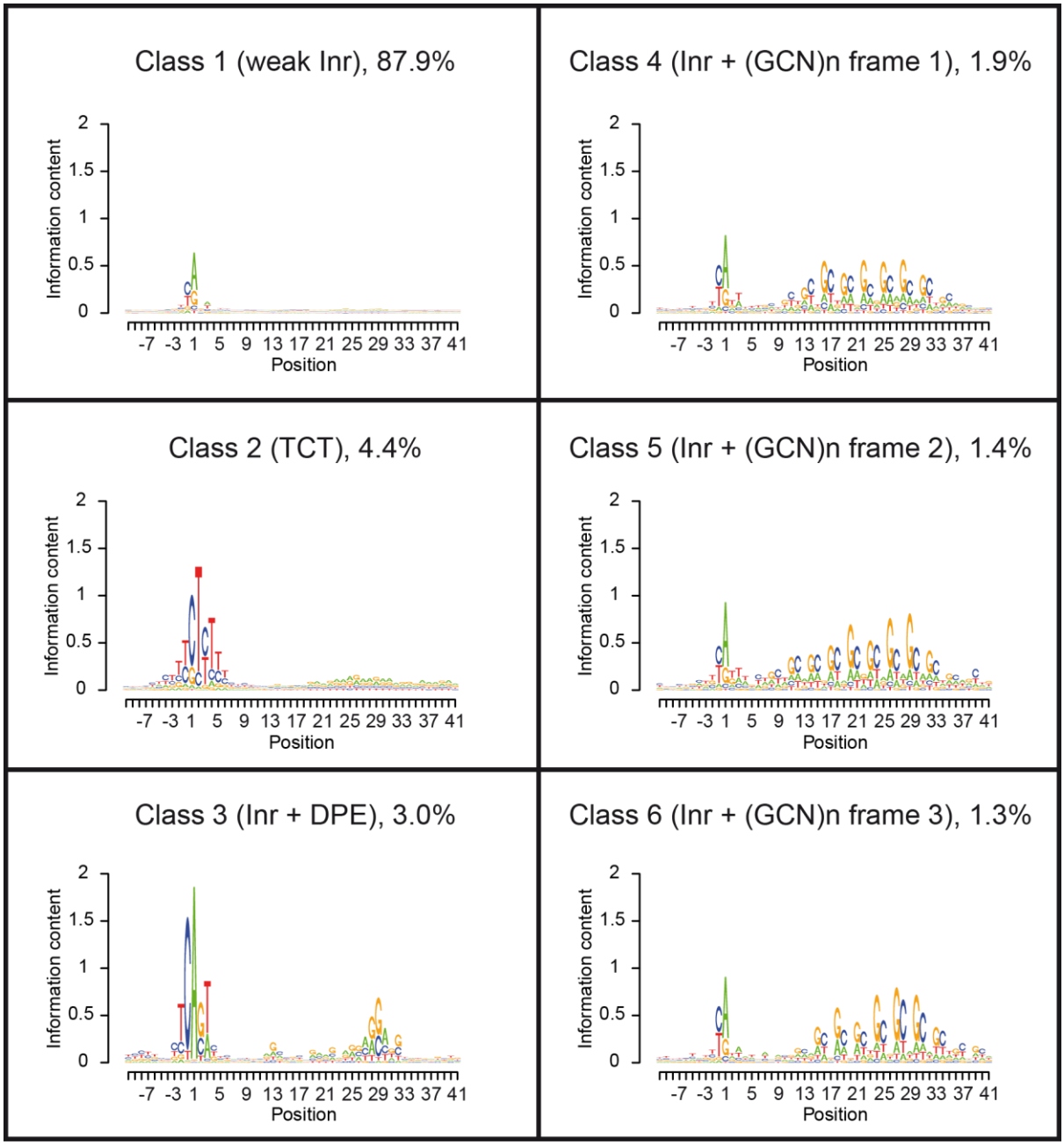


Fig 3. Partitioning of human promoter sequences using the newly developed extended EM algorithm. Six most frequent classes were identified within human promoters. The third most frequent class, which very much resembles the Inr+DPE class found in *Drosophila*, accounts for 3% of human core promoters. For each class, its frequency among the examined promoters is indicated.

In order to identify low frequency classes, which could have been missed with the basic algorithm, the new algorithm was also applied to *Drosophila* promoters partitioning them into 6 classes (S1 Fig). Based on its abundance and motif pattern, we speculate that the majority class (88.9%) obtained by this run is a mixture of all three classes obtained with the basic algorithm (Fig 2). Class 4 (2.5%) shows an extended Inr motif, GGTCACACT, but little base conservation in the downstream region. The other four classes are variants of the Inr+DPE class. In contrast to our expectations, no TCT and no trinucleotide repeat-containing classes were discovered. In summary, with regards to rare promoter classes, six-fold partitioning of human and *Drosophila* promoters highlights differences rather than commonalities between the two species.

To assess the robustness of the newly identified promoter classes, we performed bootstrapping. The complete promoter set was randomly resampled 10 times using the "sampling with replacement" method. The resampled promoter sets were then analyzed with the extended partitioning algorithm. To minimize the risk that a class is missed by chance, we retained the 10 rather than 6 most frequently found classes from each bootstrapping round. To quantify reproducibility, we recorded for each class in Fig 3 the Pearson correlation coefficient with the most similar subclass from each round (S2 Fig). The results are highly reassuring. Five of the six newly identified promoter classes (including Inr+DPE) are reproduced by all resampled data sets with a high correlation coefficient ($r > 0.8$). For class 6 (a GCN-repeat class), one (out of 10) of the bootstrapping rounds demonstrated low correlation with the newly identified class ($r = 0.26$).

The new computationally identified human Inr+DPE class closely resembles its *Drosophila* counterpart in terms of its DNA accessibility

The questions whether a *Drosophila*-like DPE element exists in human (or in any other species) could also be debated from a biochemical perspective. In this case, one would have to show that the human DPE discovered computationally in this study undergoes similar protein-DNA interactions as its well-characterized *Drosophila* counterpart. One way to approach this question is by looking at chromatin accessibility footprints. ATAC-seq assays provide detailed information about protein-DNA contacts at single base resolution. Even though it does not reveal the identity of the interacting proteins, it has an advantage over ChIP-seq that it can distinguish between direct and indirect binding mechanisms. This is important in this study's context, because the proposed interaction partners of the human downstream promoter elements are part of a larger complex, TFIID, which could be recruited to a core promoter via other sequence elements, *e.g.* a TATA-box.

We evaluated ATAC-seq footprints for the most frequent promoter classes identified in *Drosophila* and human with regard to their capacity to discriminate between the computationally derived promoter classes (Fig 4). Notably, compared to the other classes, the DPE-containing classes are highly structured in the +10 to +35 downstream regions. This suggests tight contacts with a specific protein surface, which do not occur in promoters lacking a DPE. Unsurprisingly, the ATAC-seq footprint of the human TCT class looks different from all other classes, especially at positions very close to the TSS.

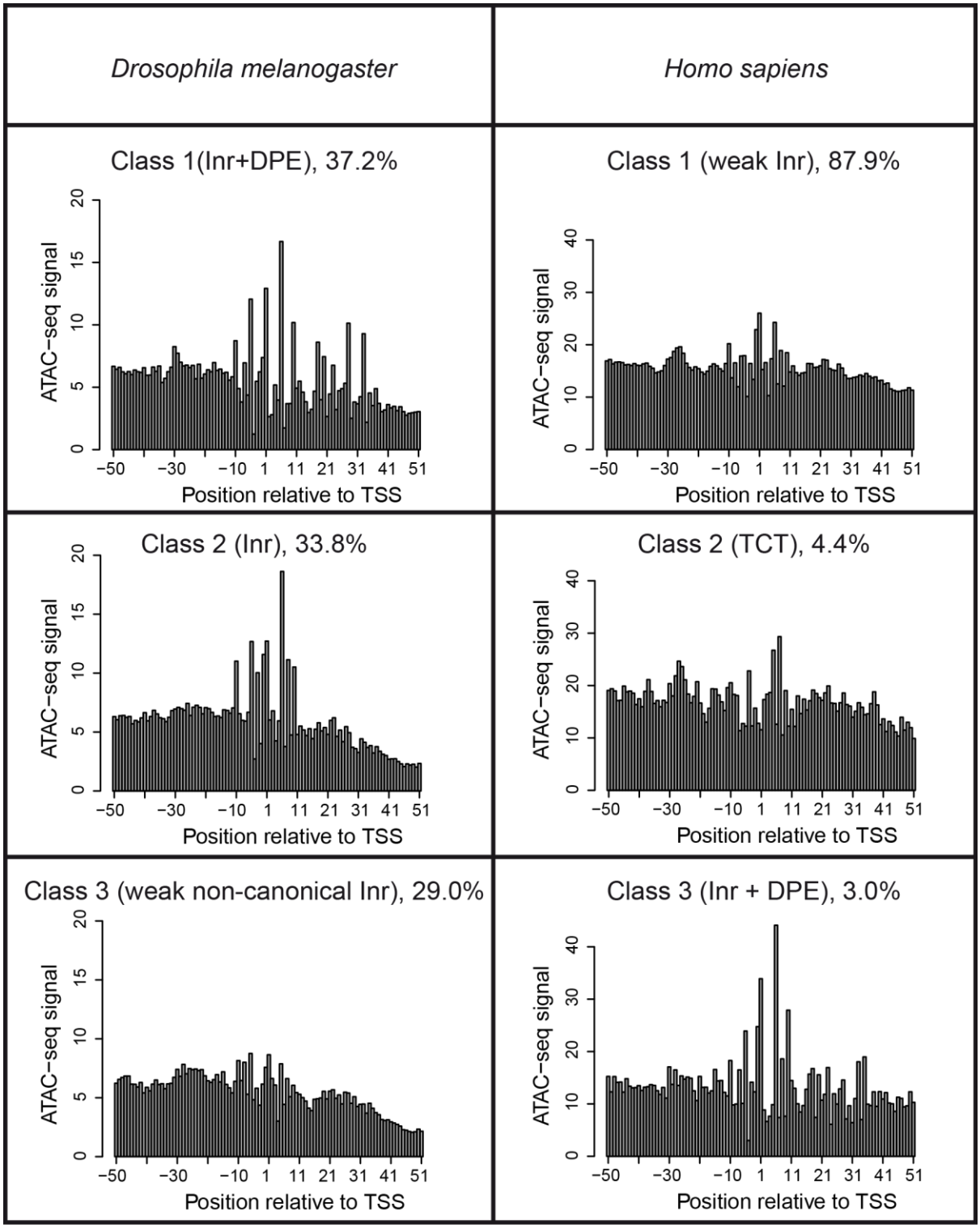


Fig 4. ATAC-seq footprints of different promoter classes. Single-base resolution ATAC-seq footprints are shown for the six most frequent promoter classes presented in Figs 2 and 3. The ATAC-seq signal displayed on the vertical axis is expressed as fold enrichment over genome-wide background. These numbers tend to be high because promoters are among the most accessible regions of the genome.

The ATAC-seq footprints of the Inr+DPE promoter classes from the two species are not only highly structured but also similar to each other, in particular in the proximal downstream region (see detailed views in Fig 5A and 5B). In both species, local maxima appear at positions 1, 6, 10, 18, 20, 23, 28 and 33, while local minima appear at positions 7, 19, 21, 24, 29 and 34. Furthermore, a U-shaped valley is seen between positions 12 and 17.

To support these intuition-guided assessments in a more objective manner, we computed correlation coefficients of ATAC-seq footprints for all positions in the proximal downstream promoter regions for all pairs combinations of promoter classes (Fig 5C). Indeed, the two Inr+DPE classes show the highest correlation (r=0.89). Classes with a canonical or recognizable Inr (dm6_c1, dm6_c2, hg19_c1, hg19_c3) also show positive correlations among themselves, whereas the human TCT class (hg19_c2) negatively correlates with all but one class. In summary, our results confirm that the newly discovered human Inr+DPE class, identified by computational sequence analysis in a completely experiment-blind manner, closely resembles its *Drosophila* counterpart in terms of direct protein-DNA contacts.

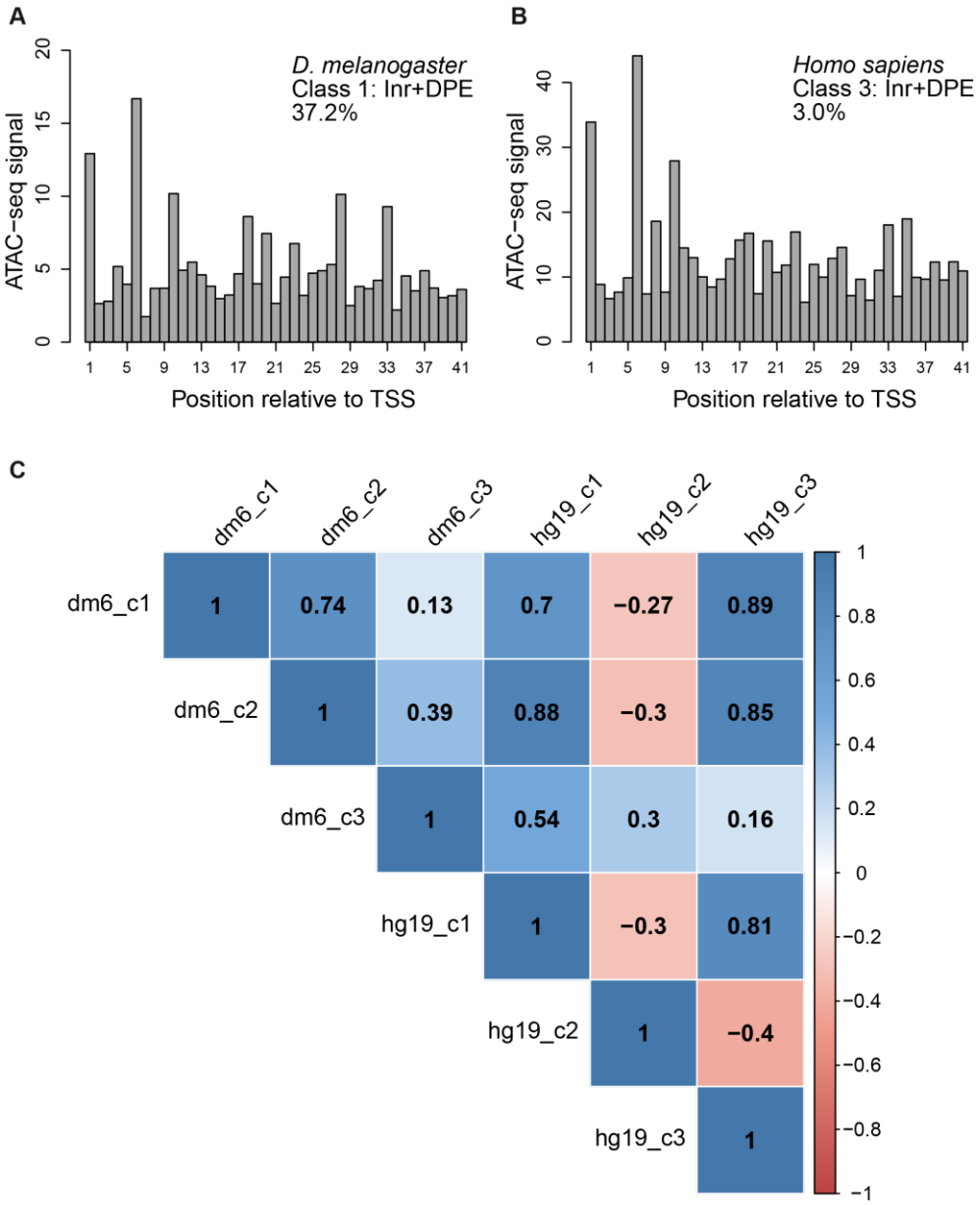


Fig 5. Comparisons of single-base resolution footprints for proximal promoter downstream regions of *Drosophila* and human promoter classes. Single-base resolution footprints for proximal promoter downstream regions of *Drosophila* (A) and human (B) Inr+DPE promoter classes. (C) Correlation similarity matrix of promoter class-specific ATAC-seq footprints. Shown are Pearson correlation coefficients computed from the ATAC-seq footprints for promoter regions +1 to +41.

DPE-like motifs in other species

The finding that the Inr+DPE element was present in human promoters, opened the intriguing hypothesis that it could be more widespread and might occur in other

316 species, perhaps even beyond the metazoan kingdom. To this end, we applied the
317 EM partitioning algorithm with K=6 to promoters from eight additional species,
318 including plants (*A. thaliana* and *Z. mays*) and fungi (*S. cerevisiae* and *S. pombe*).
319 To visualize the relationship between all promoter classes obtained in this way
320 (including those from human and *Drosophila*), we clustered the corresponding
321 sequence motifs using neighbor-joining. The resulting tree (Fig 6) was composed of
322 three distinct domains: two clades (colored blue and red) and a middle ground
323 (green) comprising multiple branches originating from nodes close to the tree center.
324 To relate these domains to the motifs shown in Figs 2 and 3, we included consensus
325 logos for each domain, which were obtained by averaging over the base probabilities
326 of all motifs from each domain. Clearly, the sequence logo of the red sub-tree
327 resembles the *Drosophila* Inr+DPE promoter class shown in Fig 2. The green and
328 blue domains corresponded to CA and TG variants of the basic YR initiator motif,
329 respectively. We noted that *D. melanogaster* Inr+DPE motifs ((Dm).m1, m3, m4, m5,
330 m6) have close neighbors from all metazoan species (*H. sapiens* (Hs).m3; *M.*
331 *musculus* (Mm).m2; *D. rerio* (Dr).m2, m3, m5, and m6; *A. mellifera* (Am).m2, m3,
332 m4, m5 and m6; *C. elegans* (Ce).m4 and m6), while none of them are from species
333 outside the metazoan kingdom, like plant and yeast. Taken together, the
334 aforementioned observations strongly suggest that canonical Inr+DPE promoters
335 could in fact be common to all metazoan species, and absent outside the metazoan
336 kingdom.
337

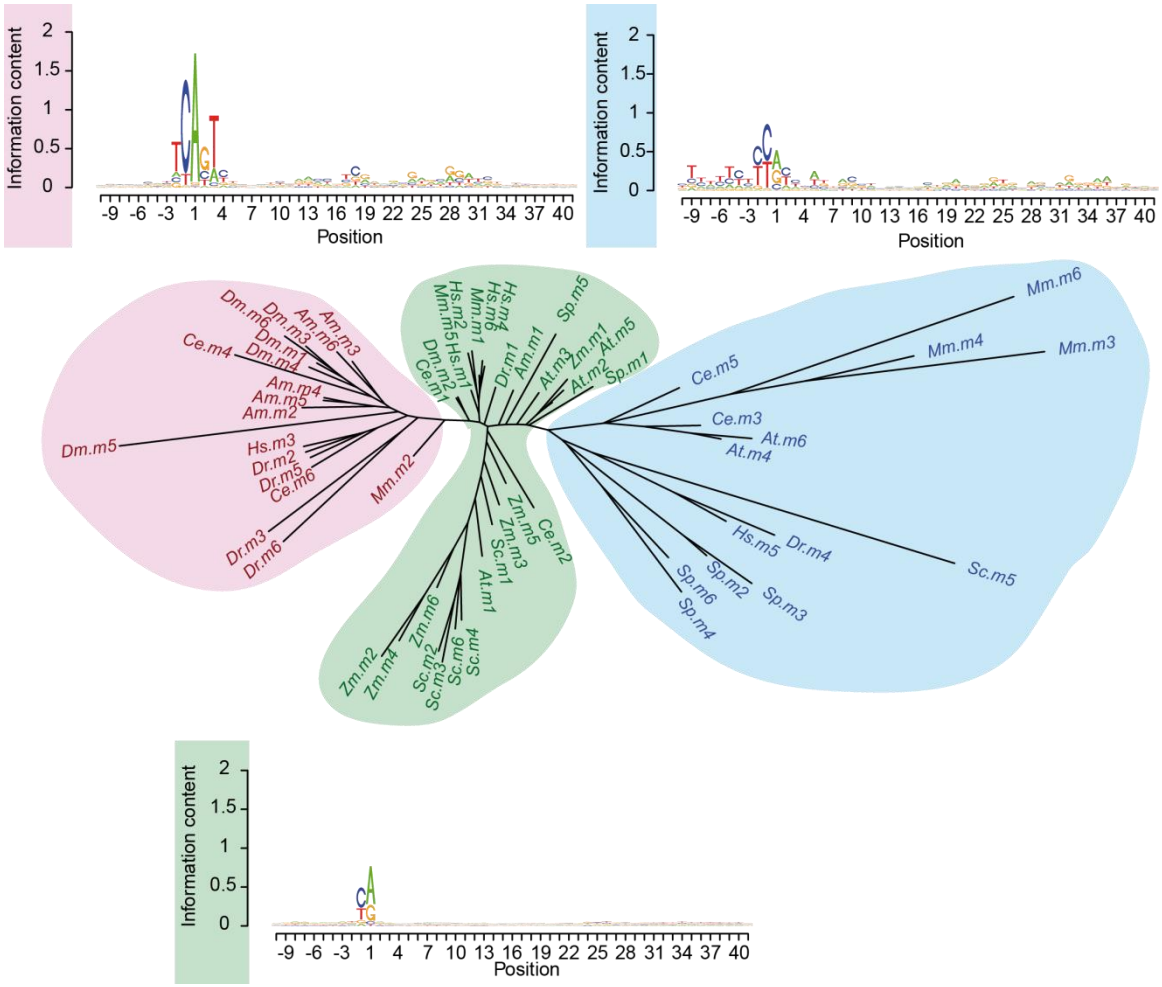


Fig 6. Neighbor joining tree of motifs found in the promoter region of 10 species. Global NJ tree obtained by clustering 6 motifs (identified using the presented EM algorithm, see Method for detail) in 10 species (*H. sapiens*; *M. musculus*; *D. rerio*; *C. elegans*; *D. melanogaster*; *A. mellifera*; *A. thaliana*; *Z. mays*; *S. cerevisie*; *S. pombe*). The tree is composed of two main clades (highlighted in red and blue) and a middle ground (green) containing several small branches originating from nodes close to the center. The consensus sequence of each clade is plotted alongside it. The Inr+DPE cluster (red) does not contain plants nor fungi motifs, highlighting the idea that the Inr+DPE element is present only in metazoa. The green and blue branches are variations of the basic YR motif.

The identified downstream positions are functional in HEK293 cells

In order to experimentally test whether the identified downstream positions are indeed functional, we analyzed a list of 20 potentially functional human core promoters. To narrow down the selection to several promoters, we applied the ElemeNT algorithm [46] to detect possible initiator and DPE motif, based on PWM's

356 constructed using experimental work in *Drosophila* [46]. We also verified that the
357 promoters lack a TATA-box upstream of the examined region, and ensured that the
358 initiation type is sharp (S3 Fig). Finally, two candidate core promoters were chosen
359 for experimental analysis, namely LRCH4 (Leucine Rich Repeats And Calponin
360 Homology Domain Containing 4) and ANP32E (Acidic Nuclear Phosphoprotein 32
361 Family Member E).

362 Notably, the prominent positions in the newly identified human downstream motif
363 (Fig 3, class 3) are G nucleotides at positions +28 and +29 (relative to the A_{+1}
364 position of the relevant initiator motif). Moreover, a sequence bias at +24(G) (relative
365 to the A_{+1} of the Inr) was previously observed and experimentally shown to
366 contribute to the function of *Drosophila* DPE-containing promoters [28]. Thus, we
367 focused on 3 preferred downstream positions (+24, +28 and +29 relative to the A_{+1}
368 position of the relevant initiator motif), mutating each of them from G to T nucleotide
369 (mPDP version; exact sequences provided in Table 1). These substitutions were
370 based on prior knowledge regarding functional downstream positions in *Drosophila*
371 *melanogaster* promoters [23, 46].

372
373
374
375

Table 1. Sequences used for testing activity of identified downstream positions.

Name	Cloned promoter sequence	DPE score (ElemeNT)	Class 3 score (EM)
LRCH4	cggtcccg*tcagtca*ggcagcg ggagccgccgg**G**agc**GG**atggc ggcggc	0.2494	11.24
ANP32E	atggaggc*tcagtct*ctgagca gccattgaagg**G**gaa**GG**aactg cgggtg	0.0278	13.58
CKS2	tgcggtcg*ttagtct*ccggcga gttgttgcctg**G**gct**GG**acgtg gttttgt	0.8182	7.22
CELF1	ggggtgtt*ctgctct*ggcggca gcggcagcggc**G**gcg**GG**acgcg gaggctc	0.2425	-0.20
CTSA	catgactt*ccagtcc*ccgggcg cctcctggaga**G**caa**GG**acgcg ggggagc	0.2425	8.27

Mutated positions are marked in bold and UPPERCASE (G>T substitutions). Initiator and DPE elements, as detected by the ElemeNT algorithm, are italicized or underlined, respectively.

We have generated both WT and mPDP constructs (Table 1), and tested them using dual-luciferase assays in HEK293 cells (Fig 7A). Strikingly, the substitution of the 3 positions was sufficient to reduce LRCH4 and ANP32E reporter levels to either 0.6 or 0.75-fold relative to the WT promoter, respectively.

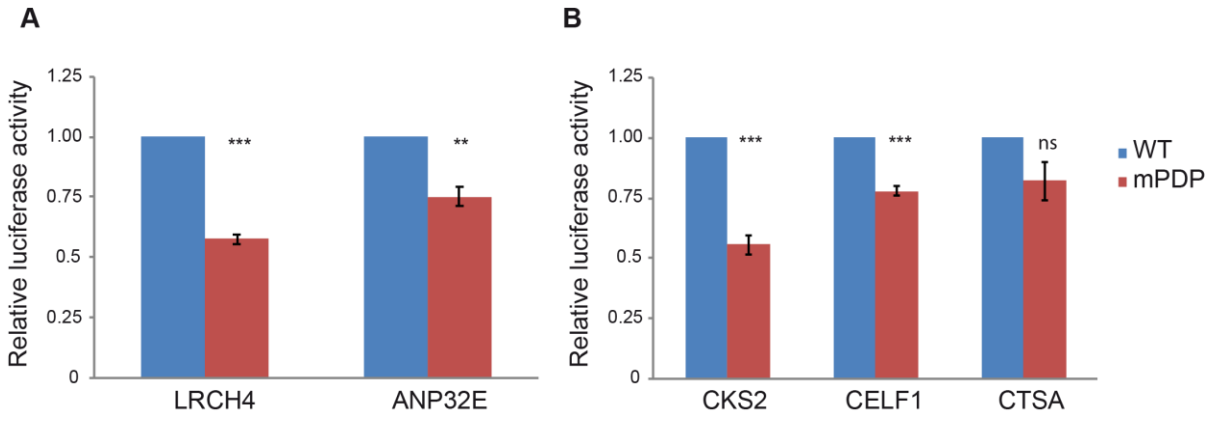


Fig 7. The preferred downstream core promoter positions (PDP) are functional in HEK293 cells. Results indicate the fold change in the WT versus mPDP version of the relevant promoter, tested by dual-luciferase assays in HEK293 cells. Each experiment was performed in triplicates, results represent 4-6 independent experiments ±SEM. ***p<0.001, **p<0.01, ns- not significant, calculated using

Student's t-test. Two candidate genes, LRCH4 and ANP32E (A), were first chosen based on their core promoter composition and conservation, as discussed in the Results section. (B) As the reduction in the LRCH4 reporter activity was more pronounced than that of the ANP32E gene, the characteristics of LRCH4 promoter were used as a reference (see Results section for the exact criteria), and the promoters of CKS2, CELF1 and CTSA were chosen for experimental examination.

We next sought to examine additional candidates, to gain a better understanding of the preferred downstream positions. Since the reduction in the LRCH4 reporter activity was more pronounced than that of the ANP32E gene (p-value 0.016) (as may have been expected based on the ElemeNT score, Table 1), we used the characteristics of LRCH4 as a reference. To this end, we started from a broader list of potentially-functional promoters. The resulting list was analyzed using ElemeNT, with the DPE score required to be >0.2 and accompanied by a Bridge element, similarly to LRCH4. The absence of a TATA-box was verified as well. As we analyzed minimal promoters (-10 to +40) (*i.e.*, resulting in relatively low expression), and the expression of the LRCH4 gene in HEK293 cells is 61 (based on CAGE data generated by FANTOM5 consortium), an expression cutoff of >61 was applied as a criterion to select candidate promoters that would likely be expressed in our experimental system. Moreover, transcription initiation pattern (sharp or broad) was manually determined using the EPDnew website for each examined gene, based on the distribution of CAGE tags around the reported transcription start site.

Using the above guidelines, we chose 3 additional unrelated promoters to be tested, namely, CKS2 (CDC28 Protein Kinase Regulatory Subunit 2), CELF1 (CUGBP Elav-Like Family Member 1) and CTSA (Cathepsin A). Using dual-luciferase reporter assays in HEK293 cells, we discovered that CKS2 and CELF1 reporter activities were reduced to either 0.6 or 0.8-fold relative to the WT promoter, respectively (Fig 7B). However, the luciferase reported activity of the mPDP version

of CTSA was not significantly lower than the WT version. Notably, this may result from the transcription initiation pattern of CTSA, which was slightly less focused than LRCH4, ANP32E, CKS2 and CELF1 (S3 Fig). Taken together, using the described EM algorithm and reporter assays in HEK293 cells, we identified a preference for conserved downstream positions within natural human core promoters with sharp transcription initiation patterns, and demonstrated that they are functional.

The identified downstream positions are strictly dependent on the spacing from the Inr

In order to test whether the identified downstream positions are canonical core promoter elements that, similarly to the *Drosophila* DPE, are strictly dependent on the spacing from the Inr, we generated multiple mutants of the of LRCH4 and CKS2 promoters, in which two nucleotides were either deleted or added (m2 or p2, respectively) in positions 10 or 18 relative to the A_{+1} position of the TSSs. Using dual-luciferase reporter assays in HEK293 cells, we detected significantly reduced activities of LRCH4 promoters in which 2 nucleotides were either deleted or added at positions 10 or 18 (Fig 8). Although deletion of 2 nucleotides in position 10 of the CKS2 promoter did not result in reduced activity, significantly reduced activities were detected in CKS2 promoters in which 2 nucleotides were either deleted or added at position 18, and when 2 nucleotides were added in position 10. By and large, the effects of these addition/deletion mutations argue in favor of a spacing dependency of the newly discovered PDP on the Inr, and against the possibility that these PDP merely serve as a binding site for a sequence-specific transcription factor that is not typically associated with core promoters.

We also examined whether two consecutive G nucleotides outside the PDP could result in reduced activities, similar to the observed mPDP activities. To this end, we mutated 2 consecutive G nucleotides to T nucleotides (mGG) in the vicinity of the PDP in the LRCH4 (at +35-36) and CKS2 (at +34-35) promoters. Interestingly, the mGG version of the LRCH4 promoter displayed reduced activity, whereas the mGG version of the CKS2 promoter did not display a similar reduction. Thus, the specific context of core promoter elements may have variable effects, as previously demonstrated (see [47], for example).

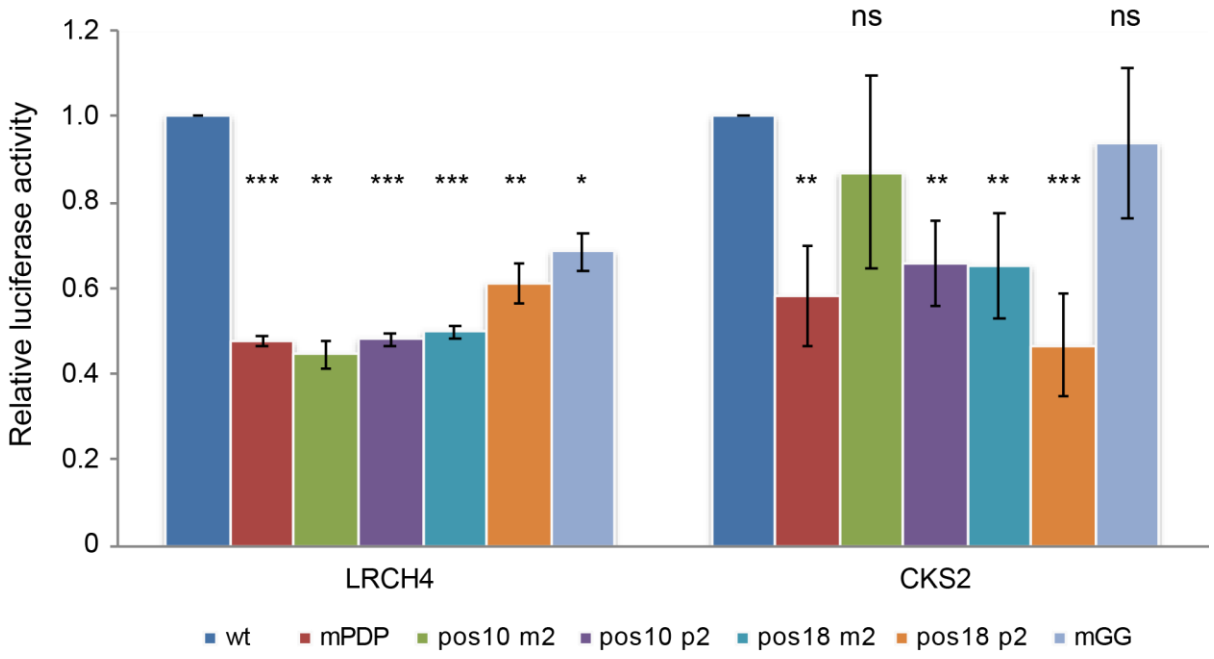


Fig 8. The activities of the LRCH4 and CKS2 in HEK293 cells are dependent on the spacing between the Inr and the PDP. Results indicate the fold change in the WT versus the mutant versions (mPDP, deletion or addition (m2 or p2, respectively) of 2 nucleotides in positions 10 or 18 relative to the A_{+1} position of the TSSs, or mutation of 2 consecutive G nucleotides in the vicinity of the PDP to T) of the indicated promoters, tested by dual-luciferase assays in HEK293 cells. Each experiment was performed in triplicates, results represent 3-4 independent experiments ±SEM. ***p<0.001, **p<0.01, ns- not significant, calculated using Student's t-test.

Discussion

The presence of downstream core promoter positions within human promoters that are transcriptionally important has been a matter of controversy in the literature. Although the DPE was originally reported as conserved from *Drosophila melanogaster* to humans [20], and additional studies identified functional downstream core promoter motifs in human promoters [33, 48], one publication suggests that the DPE motif is *Drosophila melanogaster*-specific [3], whereas another bioinformatics analysis indicated that ~25% of human promoters contain a sequence that matches the consensus of *Drosophila* DPE [49]. It should be noted, however, that the latter study did not account for the strict spacing dependency between the DPE and the Inr.

Nonetheless, ample evidence exists showing that the downstream region is an important regulator of transcriptional output in humans. The super core promoter (SCP), containing the TATA-box, initiator, MTE and DPE core promoter motifs, exhibits a robust transcriptional output in human cells, as compared to other commercially-available potent promoters [14, 40]. Mutating any of these elements significantly reduced the transcriptional output of the promoter [14], suggesting that the transcription machinery in human cells recognizes the DPE. Moreover, human TFIID is associated with the downstream core promoter area of the SCP [41, 42, 50], and both TFIID subunits TAF1 [42, 50] and TAF2 [42, 50, 51] bind the downstream core promoter region.

The aim of our study was to search for a DPE-like core promoter motif in human promoters. In line with previous studies [3, 7, 44], our analysis showed that the majority of human promoters contain a YR_{+1} initiator (Fig 3, class 1). Importantly, using the extended EM algorithm, we discovered a novel class of human promoters

containing an Inr and a downstream sequence motif that resembles the *Drosophila* DPE (Fig 3, class 3). Unlike *Drosophila* DPE-containing promoters that account for more than a third of promoters (Fig 2, class 1), human class 3 promoters account for 3% and were not enriched for developmental processes or for biological regulation.

Interestingly, we did not identify an enrichment of human Inr and MTE (motif 10 element)-containing promoters. The MTE motif was first inferred from computational analysis of *Drosophila* promoter sequences [30]. The motif was originally defined by an algorithm allowing for extensive distance variation relative to the TSS. Its functional significance was later demonstrated in both *Drosophila* and human gene expression systems, using promoters from both species [21]. In that study, the MTE is presented as a core promoter element with a consensus sequence CSARCSSAACGS that occurs between positions +18 to +29, overlapping with the DPE motif by two base pairs. Similar to the DPE motif, it was reported that the MTE function is strictly dependent upon a functional Inr, and is involved in interaction with TFIID [21, 22]. Furthermore, although it was defined as a distinct element, a synergy between the MTE and the DPE was demonstrated. The *Drosophila* class 1 sequence logo that was detected using our algorithm supports C at position +18, R at +22, and CGS at +27-29. We further note an additional conserved Y at position +17, just preceding the reported MTE region.

Further examination of the downstream region revealed additional TFIID-interacting subregions, comprised of +18-22 and +30-33, termed Bridge [22]. The Bridge element was demonstrated to support, but not fully-restore, DPE-dependent transcription [23]. It was recently proposed that the downstream core promoter region might be a single functional unit (resembling the “Ohler-defined DPE”, [30]) [52]. We compared our sequence motifs for the *Drosophila* and human Inr+DPE

promoter classes to the "functional" MTE motifs of two *Drosophila* promoters (*Tollo* and CG10479) derived by single-base mutational analysis (Theisen et al. 2010). To make the motifs visually comparable, we converted mutational analysis data for each promoter into a corresponding sequence logo by dividing the relative transcriptional activities of each base at a given positions by the sum of the transcriptional activities at the same position. We further computed Pearson correlation coefficients for all logo pairs, in order to assess similarity in a more objective manner (Fig 9). By visual inspection we note a good agreement between the functional MTE motif of the CG10479 promoter and our computationally derived motif for the *Drosophila* Inr+DPE promoter class. This intuitive judgment is supported by a high Pearson correlation coefficient of 0.83. The functional MTE motif for *Tollo s*hows more divergence with regard to both the CG10479 functional motif and the computationally derived Inr+DPE motif. Not surprisingly, both functional motifs show better correlation with the *Drosophila* than with the human Inr+DPE motif. We further note a high correlation coefficient of 0.74 for the two computationally derived motifs, suggesting that the two species share conserved sequence determinants not only within the canonical Inr and DPE motifs, but also in the region between them.

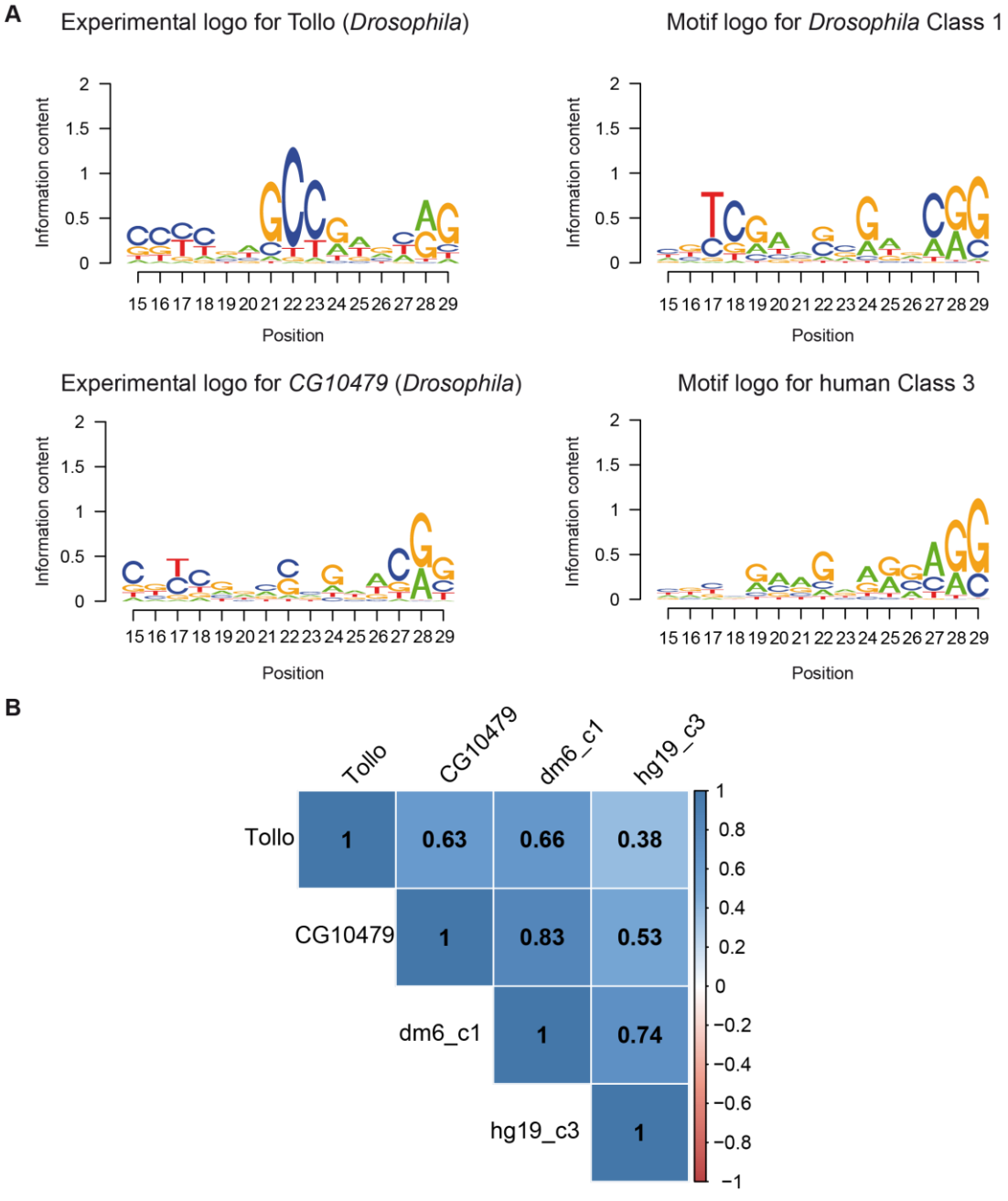


Fig 9. Comparison of experimental logos with sequence motif logos. (A) Experimental logos are based on exhaustive single-base mutational analysis of the +15 to +29 region of two *Drosophila* promoters [22]. Relative expression values were rescaled such as to sum up to one at each position. The sequence motif logos were extracted from to logos shown in Figs 2 and 3. All logos have been over-skewed with an exponent of 2 to highlight differences between them. (B) Correlation plot showing Pearson correlation coefficients computed from the base probabilities underlying the logos. Note the high correlation of the *CG10479* experimental logo with the *Drosophila* motif logo.

A critical reader could question our results by arguing that we modified our extended partitioning algorithm to obtain the desired result. A general problem with partitioning and other unsupervised machine learning approaches is that the result cannot be assessed in terms of accuracy. We thus can only argue that the classification we obtain with the extended partitioning algorithm is biologically plausible or meaningful. The nature of the other four simultaneously discovered low-frequency promoter classes gives us assurance in this respect. Class 2 perfectly matches a previously reported promoter class, characterized by the presence of a TCT motif and its association with genes involved in translation. The other three minority classes strikingly resemble each other in that they contain the same trinucleotide repeats in three different frames relative to the TSS. These highly unusual properties make it unlikely that these classes are collateral noise of an algorithm specifically designed and fine-tuned to discover another promoter class.

The weak TGT motifs observed with the basic algorithm (Fig 2, *Drosophila* class 3 and human class 2), which are reminiscent of the previously described TGT motif [53], were not detected using the extended EM algorithm (Fig 3). Notably, weak motifs in general, may reflect the presence of additional or a mixture of sub-classes of promoters.

The newly discovered human promoter classes 4-6 are characterized by the same tri-nucleotide motifs (GCN)n in three different frames. The reason why these promoters were put into different classes is because we used an algorithm that does not allow for limited shifting of sequences relative to each other. Trinucleotide repeats in 5'UTRs are suggestive of a function in translation. Specifically, we conjecture that they may be part of regulatory upstream open reading frames (see

569 [54] for review). If true, we would expect that the repeats be preceded by in-frame
570 ATG codons. To test this hypothesis, we tabulated the frequencies of ATG at
571 proximal promoter downstream positions (S1 Table). Indeed, in each class, we
572 observed a strong, 3bp periodic bias in the positional distribution of ATG codons,
573 compatible with translation of the CGN repeats into poly-alanine. A regulatory
574 function of these repeats involving translation thus seems plausible.
575 Importantly, we demonstrate the contribution of the 3 G nucleotides, located at
576 positions +24, +28 and +29 relative to the A_{+1} position, to the function of four natural
577 human promoters. Using luciferase reporters driven by minimal promoter constructs
578 (-10 to +40) in HEK293 cells, we demonstrated that changing G nucleotides at these
579 positions to T significantly reduces the transcriptional output to 0.6-0.8 fold, as
580 compared to the WT promoters. This is a substantial effect on enzymatic reporter
581 activities, considering the fact that only 3 nucleotides in a non-Inr region of the
582 minimal promoters were substituted. Remarkably, the reduced reporter activities of
583 promoters in which the spacing between the Inr and the DPE was altered by addition
584 or deletion of 2 nucleotides, largely suggest that, similarly to the *Drosophila* DPE, the
585 newly discovered PDP depends on spacing from the Inr. It also disfavors the
586 possibility that the PDP serves as a binding site for a sequence-specific transcription
587 factor that is not normally associated with core promoters.
588 During the preparation of the manuscript, we became aware of a comprehensive
589 work from the Kadonaga lab [55], which used machine learning to generate
590 predictive models to analyze human Pol II core promoters and identified a
591 downstream promoter region (DPR) spanning from +17 to +35, which contributes to
592 the transcriptional output of a fraction of human promoters. Reassuringly, the
593 positions identified in our study highly match specific positions within the DPR

identified by the Kadonaga lab, which supports the concept of a single functional downstream unit [30, 52]. Moreover, different approaches to identify the important downstream positions were taken; while we started from bioinformatics analysis and then tested naturally-occurring minimal promoters, the Kadonaga lab has first used massively parallel reporter assays (MPRA) of an extensive library composed of randomized version of the downstream region, using a specific promoter backbone. Moreover, the experiments were performed in two different cell lines, using different readout as the outcome, either the indirect luciferase reporter activity (this study) or the RNA output itself, using either RNAseq or primer extension analysis [55]. Surprisingly, the two independent approaches identified functional downstream positions/region within the ANP32E promoter. Moreover, we ran the support vector regression (SVRb) model that was generated using *in vitro* transcription [55] on the +17 to +35 sequences of the wt and mutant promoters identified using the EM algorithm (S2 Table). Overall, our computational model was successful in making similar predictions (correlation coefficient ~0.75) as the SVRb model that used experimentally-based training data. Thus, both independently-performed studies complement each other, strengthening the notion that the downstream core promoter region contributes to transcriptional regulation of human promoters. Our mutational analysis highlights the importance of three specific nucleotides for the transcriptional output, as well the strict spacing requirement between the preferred downstream positions and the Inr motif, reminiscent of the *Drosophila* DPE.

To conclude, specific positions within the downstream core promoter region of human promoters are important for the transcriptional outcome; thus transcriptional regulation of human promoters via the downstream region is an important regulatory mechanism, likely conserved among metazoans but absent in other eukaryotes.

Methods

Promoter sets

The promoter sets and the corresponding dominant TSS positions were taken from EPDnew [56]: version 5 for *H. sapiens* and *D. melanogaster*; version 2 for *M. musculus*, *A. thaliana* and *S. cerevisiae*; version 1 for all other organisms studied. EPDnew promoter collections have been validated by hundreds of high-throughput sequencing experiments (i.e. CAGE), giving a very high confidence in identifying the correct transcription start site. For each gene, the promoter that was validated by the largest number of experiments was selected as the representative. This gave very high confidence for the positions of the initiation sites, and reduced the probability of selecting promoters used only in particular cell lines and/or conditions. Moreover, to reduce possible sequence bias by coding sequences, promoters that had translation start sites within the first 40 bases were discarded.

Probabilistic partitioning basic algorithm

In its basic structure, the algorithm is identical to the Expectation-Maximization (EM) algorithm presented in [57], which was originally designed for partitioning sets of genomic regions based on ChIP-seq data and represented as count data (integer) vectors and is described in Fig 1A. The adaptation to sequence data requires some modifications described below.

In the following, we adhere to the notation used in Stormo's review on specificity models of protein-DNA interactions [43]. Sequences of length N denoted S_i are represented as binary matrices with four rows corresponding to the bases A, C, G and T, and N columns corresponding to successive positions in the sequence. A matrix element $S_i(b,j)$ has a value of 1, if base b occurs at the jth position of

sequence *i*, and a value of zero otherwise. A class C_k is represented by a matrix of the same dimensions as the sequences, plus its occurrence probability p_k. A matrix element $C_k(b,j)$ contains the probability that base *b* occurs at the *j*th position of a sequence belonging to class *k.* The probability of sequence S_i given class C_k is then given by:

$$P(S_i \mid C_k) = \prod_{b,j} C_k(b,j)^{S_i(b,j)} \quad (1)$$

The formula for computing the probability of class C_k given sequence S_i remains unchanged:

$$P(C_k \mid S_i) = \frac{p_k \cdot P(S_i \mid C_k)}{\sum_{k'} p_{k'} \cdot P(S_i \mid C_{k'})} \quad (2)$$

Using these probabilities, the base probability matrix for class C_k is updated in 2 steps:

$$C^*_k(b,j) = \frac{\sum_i P(C_k \mid S_i) S_i(b,j)}{q_b} Z_{kj}^{-1} \qquad C_k(b,j) = \frac{C^*_k(b,j) + w}{1 + 4w} \quad (3a, b)$$

Here, q_b denotes the frequency of base *b* in the input sequence set, and z_{kj} is a column specific normalization constant chosen such that the column *j* of base probability matrix C_k sums to one. The first equation defines the MAP (maximum a posteriori probability) estimation of the base probability matrix for each class *k*. The second equation adds a small correction term to the MAP estimations that prevents probabilities from converging to zero. Note however, that the algorithms returns c_j^* as

the final results after the last iteration. A small correction term *x* is also added the re-estimated class probabilities:

$$p_k = \frac{\left(\frac{1}{N}\right)\left(\sum_i P(C_k \mid S_i)\right) + x}{1 + Kx} \quad (4)$$

The algorithm is initiated by a random seeding strategy. The probabilities of individual sequences of belonging to specific classes are sampled from a Beta distribution

$$P(C_k \mid S_i) \sim \frac{\mathrm{Beta}(\alpha, \beta)}{Z_i} \quad (5)$$

with shape parameters α=0.01 and β=1. Z_i is a sequence-specific normalization constant chosen such that the class probabilities for sequence *i* sum to one. The classes themselves are assigned equal probabilities p_k=1/*K*. After initializing these probabilities, the EM algorithm starts with equation 3.

Probabilistic partitioning extended algorithm

The extended partitioning algorithm (Fig 1B) features two innovations: (i) a two-state clustering strategy and (ii) a new, so-called "over-skewing" parameter σ. The two extensions are independent of each other, *i.e.* two-stage clustering can be used without over-skewing, and vice-versa. Two-stage clustering serves to increase the reproducibility of the results when initiating the algorithm with different random seeds. Over-skewing causes the algorithm to prefer partitionings with classes of

679 highly unequal sizes, typically a majority class plus a number of small classes with
680 highly skewed base compositions.
681 With the two-stage clustering strategy, the basic EM algorithm is applied *n* times
682 to produce $n \times K$ subclasses. Each subclass is characterized as a base probability
683 matrix henceforth referred to as a "motif". During the second stage, the motifs from
684 the first stage are hierarchically clustered and subsequently partitioned into motif
685 groups using a fixed height *h.* The *K* largest motif groups are retained, and a
686 consensus base probability matrix C_k is computed for each group by averaging over
687 all its members. Likewise, the p_k is computed as the average over the occurrence
688 probabilities of all motifs belonging to group *k.* Hierarchical clustering was carried out
689 with the R functions *dist* and *hclust*, using "Euclidean" as a distance measure, and
690 "complete" as a clustering method. Tree partitioning was carried out with the R
691 function *cutree*.

692
693
694 **ATAC-seq analysis**
695
696 Average ATAC-seq footprints for promoter classes were produced with public data
697 from human lymphoblastoid cell line GM12878 [58], and from *Drosophila* wild type
698 eye-antennal imaginal disc [59], see supplementary material for GEO accession
699 numbers and download URLs. We used processed versions of the data, i.e. read
700 alignment files, available from the MGA repository [60]. Aggregation plots for the
701 promoter classes shown in Figs 4 and 5A, B were generated via the web interface of
702 the ChIP-Cor tool [61] using the following parameters: Reference feature *oriented*,
703 target feature *any*, *centering* 4, window width 1, count cut-off 10, normalization
704 *global*.

Neighbor joining analysis

Promoter sets of 10 organisms (*H. sapiens*, *M. musculus*, *D. rerio*, *C. elegans*, *D. melanogaster*, *A. mellifera*, *A. thaliana*, *Z. mays*, *S. cerevisiae*, *S. pombe*) were analyzed with the newly developed algorithm. In the first step (Figure 1A), 200 iterations were applied by the probabilistic partitioning to generate 6 motifs. This procedure was independently repeated 50 times to generate 300 motifs for each specie (see Figure 1A for reference). The motifs were then hierarchically clustered, and the resulting tree was cut to obtain 10 clusters (Figure 1B). The 6 nodes with the highest number of motifs were then chosen and averaged to generate the final motifs. These motif collections were further clustered with Euclidean distance (functions 'dist', from package 'stats') and plotted using a Neighbor Joining tree (function 'nj' from package 'ape' [45]). The frequency matrices of motifs belonging to each of the 3 branches were averaged to generate the branch consensus.

Plasmid construction

For cloning the minimal promoters of the selected genes into a reporter plasmid, double-stranded oligonucleotides (IDT) comprising core promoter sequences from –10 to +40/+41 were inserted into the KpnI and SpeI sites of a pGL3-Basic plasmid with a modified polylinker. For each promoter, both WT and mutated preferred downstream positions (mPDP) (G>T at position +24, +28 and +29 relative to the relevant A_{+1} position) versions were cloned. Primers used are listed in S3 Table. All generated constructs were verified by sequencing (Hy Labs).

728 **Cell culture, transient transfections and reporter gene assay**

729 Human Embryonic Kidney (HEK) 293 cells were cultured in DMEM high-glucose
730 (Biological Industries) supplemented with 10% FBS, 0.1% penicillin-streptomycin,
731 and 1% L-Glutamine, and grown at 37°C with 5% CO_2.

732 For dual luciferase assays, 1-2x10^6 cells were plated per 60mm dish one day prior
733 to transfection. Cells were transfected using the calcium phosphate method with a
734 total of 3µg DNA (2.5µg firefly luciferase plasmid, 100ng of Thymidine Kinase-*Renilla*
735 luciferase plasmid, and 400ng of pBlueScript plasmid) per 60mm dish. Prior to the
736 transfection, the medium was changed to contain 25µM Chloroquine, and replaced
737 with fresh medium 6-8 hours following the transfection. Cells were harvested 48
738 hours post-transfection and assayed for dual-Luciferase activities as specified by the
739 manufacturer (Promega). To correct for variations in transfection efficiency, the firefly
740 luciferase activity of each sample was normalized to the corresponding *Renilla*
741 luciferase activity. Each transfection was performed in triplicates, and each graph
742 represents an average of 4 to 6 independent experiments ± SEM. Student's two-
743 sided t-test was applied in order to determine the statistical significance of the
744 observed difference.

745

746 ## Acknowledgments

747 We are indebted to Jim Kadonaga for his generous support and invaluable
748 suggestions, and for sharing unpublished data. We thank Jim Kadonaga, Diana
749 Ideses, Yehuda M. Danino, Orit Adato, Hadar Krap and Hodaya Komemi for critical
750 reading of the manuscript.

751

752

References

1. Danino YM, Even D, Ideses D, Juven-Gershon T. The core promoter: At the heart of gene expression. Biochimica et biophysica acta. 2015;1849(8):1116-1131. doi: 10.1016/j.bbagrm.2015.04.003. PubMed PMID: 25934543.
2. Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol. 2018. Epub 2018/06/28. doi: 10.1038/s41580-018-0028-8. PubMed PMID: 29946135.
3. Lenhard B, Sandelin A, Carninci P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nature reviews Genetics. 2012;13(4):233-245. doi: 10.1038/nrg3163. PubMed PMID: 22392219.
4. Thomas MC, Chiang CM. The general transcription machinery and general cofactors. Critical Reviews in Biochemistry and Molecular Biology. 2006;41(3):105-178. doi: 10.1080/10409230600648736. PubMed PMID: WOS:000237768300001.
5. Vo Ngoc L, Wang YL, Kassavetis GA, Kadonaga JT. The punctilious RNA polymerase II core promoter. Genes Dev. 2017;31(13):1289-1301. Epub 2017/08/16. doi: 10.1101/gad.303149.117. PubMed PMID: 28808065.
6. Bajic VB, Tan SL, Christoffels A, Schonbach C, Lipovich L, Yang L, et al. Mice and men: their promoter properties. PLoS Genet. 2006;2(4):e54. Epub 2006/05/10. doi: 10.1371/journal.pgen.0020054. PubMed PMID: 16683032.
7. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet. 2006;38(6):626-635. Epub 2006/04/29. doi: 10.1038/ng1789. PubMed PMID: 16645617.
8. Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. Genome Res. 2006;16(1):1-10. doi: 10.1101/gr.4222606. PubMed PMID: 16344566.
9. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, et al. A high-resolution map of active promoters in the human genome. Nature. 2005;436(7052):876-880. Epub 2005/07/01. doi: nature03877 [pii] 10.1038/nature03877. PubMed PMID: 15988478.
10. Dreos R, Ambrosini G, Bucher P. Influence of Rotational Nucleosome Positioning on Transcription Start Site Selection in Animal Promoters. PLoS Comput Biol. 2016;12(10):e1005144. Epub 2016/10/08. doi: 10.1371/journal.pcbi.1005144. PubMed PMID: 27716823.
11. Dikstein R. The unexpected traits associated with core promoter elements. Transcription. 2011;2(5):201-206. Epub 2012/01/11. doi: 10.4161/trns.2.5.17271 17271 [pii]. PubMed PMID: 22231114.
12. Heintzman ND, Ren B. The gateway to transcription: identifying, characterizing and understanding promoters in the eukaryotic genome. Cell Mol Life Sci. 2007;64(4):386-400. Epub 2006/12/16. doi: 10.1007/s00018-006-6295-0. PubMed PMID: 17171231.
13. Butler JE, Kadonaga JT. Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. Genes Dev. 2001;15(19):2515-2519. Epub 2001/10/03. doi: 10.1101/gad.924301. PubMed PMID: 11581157.
14. Juven-Gershon T, Cheng S, Kadonaga JT. Rational design of a super core promoter that enhances gene expression. Nature methods. 2006;3(11):917-922. doi: 10.1038/nmeth937. PubMed PMID: 17124735.

15. Zabidi MA, Arnold CD, Schernhuber K, Pagani M, Rath M, Frank O, et al. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. Nature. 2015;518(7540):556-559. Epub 2014/12/18. doi: 10.1038/nature13994. PubMed PMID: 25517091.
16. Zehavi Y, Sloutskin A, Kuznetsov O, Juven-Gershon T. The core promoter composition establishes a new dimension in developmental gene networks. Nucleus. 2014;5(4):298-303. Epub 2014/12/09. doi: 10.4161/nucl.29838. PubMed PMID: 25482118.
17. Ohler U, Wassarman DA. Promoting developmental transcription. Development. 2010;137(1):15-26. doi: 10.1242/dev.035493. PubMed PMID: 20023156.
18. Deng W, Roberts SG. TFIIB and the regulation of transcription by RNA polymerase II. Chromosoma. 2007;116(5):417-429. Epub 2007/06/27. doi: 10.1007/s00412-007-0113-9. PubMed PMID: 17593382.
19. Burke TW, Kadonaga JT. Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. Genes Dev. 1996;10(6):711-724. Epub 1996/03/15. PubMed PMID: 8598298.
20. Burke TW, Kadonaga JT. The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. Genes Dev. 1997;11(22):3020-3031. doi: 10.1101/gad.11.22.3020. PubMed PMID: 9367984.
21. Lim CY, Santoso B, Boulay T, Dong E, Ohler U, Kadonaga JT. The MTE, a new core promoter element for transcription by RNA polymerase II. Genes Dev. 2004;18(13):1606-1617. doi: 10.1101/gad.1193404. PubMed PMID: 15231738.
22. Theisen JW, Lim CY, Kadonaga JT. Three key subregions contribute to the function of the downstream RNA polymerase II core promoter. Mol Cell Biol. 2010;30(14):3471-3479. doi: 10.1128/MCB.00053-10. PubMed PMID: 20457814.
23. Shir-Shapira H, Sloutskin A, Adato O, Ovadia-Shochat A, Ideses D, Zehavi Y, et al. Identification of evolutionarily conserved downstream core promoter elements required for the transcriptional regulation of Fushi tarazu target genes. PloS one. 2019;14(4):e0215695. doi: 10.1371/journal.pone.0215695. PubMed PMID: 30998799.
24. Hsu JY, Juven-Gershon T, Marr MT, 2nd, Wright KJ, Tjian R, Kadonaga JT. TBP, Mot1, and NC2 establish a regulatory circuit that controls DPE-dependent versus TATA-dependent transcription. Genes Dev. 2008;22(17):2353-2358. Epub 2008/08/16. doi: 10.1101/gad.1681808. PubMed PMID: 18703680.
25. Juven-Gershon T, Hsu JY, Kadonaga JT. Caudal, a key developmental regulator, is a DPE-specific transcriptional factor. Genes Dev. 2008;22(20):2823-2830. Epub 2008/10/17. doi: 10.1101/gad.1698108. PubMed PMID: 18923080.
26. Zehavi Y, Kuznetsov O, Ovadia-Shochat A, Juven-Gershon T. Core promoter functions in the regulation of gene expression of Drosophila dorsal target genes. J Biol Chem. 2014;289(17):11993-12004. Epub 2014/03/19. doi: 10.1074/jbc.M114.550251. PubMed PMID: 24634215.
27. Kedmi A, Zehavi Y, Glick Y, Orenstein Y, Ideses D, Wachtel C, et al. Drosophila TRF2 is a preferential core promoter regulator. Genes Dev. 2014;28(19):2163-2174. Epub 2014/09/17. doi: 10.1101/gad.245670.114. PubMed PMID: 25223897.
28. Kutach AK, Kadonaga JT. The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. Mol Cell Biol.

2000;20(13):4754-4764. Epub 2000/06/10. doi: 10.1128/mcb.20.13.4754-4764.2000. PubMed PMID: 10848601.

29. Willy PJ, Kobayashi R, Kadonaga JT. A basal transcription factor that activates or represses transcription. Science. 2000;290(5493):982-985. Epub 2000/11/04. doi: 10.1126/science.290.5493.982. PubMed PMID: 11062130.
30. Ohler U, Liao GC, Niemann H, Rubin GM. Computational analysis of core promoters in the Drosophila genome. Genome Biol. 2002;3(12):RESEARCH0087. Epub 2003/01/23. doi: 10.1186/gb-2002-3-12-research0087. PubMed PMID: 12537576.
31. Rach EA, Yuan HY, Majoros WH, Tomancak P, Ohler U. Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. Genome Biol. 2009;10(7):R73. Epub 2009/07/11. doi: 10.1186/gb-2009-10-7-r73 gb-2009-10-7-r73 [pii]. PubMed PMID: 19589141.
32. Ohler U. Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction. Nucleic acids research. 2006;34(20):5943-5950. Epub 2006/10/28. doi: 10.1093/nar/gkl608. PubMed PMID: 17068082.
33. Duttke SH. RNA polymerase III accurately initiates transcription from RNA polymerase II promoters in vitro. J Biol Chem. 2014;289(29):20396-20404. Epub 2014/06/12. doi: 10.1074/jbc.M114.563254. PubMed PMID: 24917680.
34. Zhou T, Chiang CM. The intronless and TATA-less human TAF(II)55 gene contains a functional initiator and a downstream promoter element. J Biol Chem. 2001;276(27):25503-25511. Epub 2001/05/08. doi: 10.1074/jbc.M102875200. PubMed PMID: 11340078.
35. Bhuiyan T, Timmers HTM. Promoter Recognition: Putting TFIID on the Spot. Trends Cell Biol. 2019;29(9):752-763. Epub 2019/07/14. doi: 10.1016/j.tcb.2019.06.004. PubMed PMID: 31300188.
36. Cramer P. Organization and regulation of gene transcription. Nature. 2019;573(7772):45-54. Epub 2019/08/30. doi: 10.1038/s41586-019-1517-4. PubMed PMID: 31462772.
37. Patel AB, Greber BJ, Nogales E. Recent insights into the structure of TFIID, its assembly, and its binding to core promoter. Curr Opin Struct Biol. 2019;61:17-24. Epub 2019/11/22. doi: 10.1016/j.sbi.2019.10.001. PubMed PMID: 31751889.
38. Roeder RG. 50+ years of eukaryotic transcription: an expanding universe of factors and mechanisms. Nat Struct Mol Biol. 2019;26(9):783-791. Epub 2019/08/24. doi: 10.1038/s41594-019-0287-x. PubMed PMID: 31439941.
39. Verrijzer CP, Chen JL, Yokomori K, Tjian R. Binding of TAFs to core elements directs promoter selectivity by RNA polymerase II. Cell. 1995;81(7):1115-1125. Epub 1995/06/30. doi: 10.1016/s0092-8674(05)80016-9. PubMed PMID: 7600579.
40. Even DY, Kedmi A, Basch-Barzilay S, Ideses D, Tikotzki R, Shir-Shapira H, et al. Engineered Promoters for Potent Transient Overexpression. PloS one. 2016;11(2):e0148918. doi: 10.1371/journal.pone.0148918. PubMed PMID: 26872062.
41. Cianfrocco MA, Kassavetis GA, Grob P, Fang J, Juven-Gershon T, Kadonaga JT, et al. Human TFIID binds to core promoter DNA in a reorganized structural state. Cell. 2013;152(1-2):120-131. doi: 10.1016/j.cell.2012.12.005. PubMed PMID: 23332750.

42. Louder RK, He Y, Lopez-Blanco JR, Fang J, Chacon P, Nogales E. Structure of promoter-bound TFIID and model of human pre-initiation complex assembly. Nature. 2016;531(7596):604-609. Epub 2016/03/24. doi: 10.1038/nature17394. PubMed PMID: 27007846.
43. Stormo GD. Modeling the specificity of protein-DNA interactions. Quant Biol. 2013;1(2):115-130. Epub 2014/07/22. doi: 10.1007/s40484-013-0012-4. PubMed PMID: 25045190.
44. Frith MC, Valen E, Krogh A, Hayashizaki Y, Carninci P, Sandelin A. A code for transcription initiation in mammalian genomes. Genome Res. 2008;18(1):1-12. Epub 2007/11/23. doi: 10.1101/gr.6831208. PubMed PMID: 18032727.
45. Parry TJ, Theisen JW, Hsu JY, Wang YL, Corcoran DL, Eustice M, et al. The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. Genes Dev. 2010;24(18):2013-2018. Epub 2010/08/31. doi: 10.1101/gad.1951110. PubMed PMID: 20801935.
46. Sloutskin A, Danino YM, Orenstein Y, Zehavi Y, Doniger T, Shamir R, et al. ElemeNT: a computational tool for detecting core promoter elements. Transcription. 2015;6(3):41-50. doi: 10.1080/21541264.2015.1067286. PubMed PMID: 26226151.
47. Lubliner S, Regev I, Lotan-Pompan M, Edelheit S, Weinberger A, Segal E. Core promoter sequence in yeast is a major determinant of expression level. Genome Res. 2015;25(7):1008-1017. Epub 2015/05/15. doi: 10.1101/gr.188193.114. PubMed PMID: 25969468.
48. Zhou T, Chiang CM. Sp1 and AP2 regulate but do not constitute TATA-less human TAF(II)55 core promoter activity. Nucleic acids research. 2002;30(19):4145-4157. Epub 2002/10/05. doi: 10.1093/nar/gkf537. PubMed PMID: 12364593.
49. Gershenzon NI, Ioshikhes IP. Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. Bioinformatics. 2005;21(8):1295-1300. Epub 2004/12/02. doi: bti172 [pii] 10.1093/bioinformatics/bti172. PubMed PMID: 15572469.
50. Patel AB, Louder RK, Greber BJ, Grunberg S, Luo J, Fang J, et al. Structure of human TFIID and mechanism of TBP loading onto promoter DNA. Science. 2018;362(6421). doi: 10.1126/science.aau8872. PubMed PMID: 30442764.
51. Shao W, Zeitlinger J. Paused RNA polymerase II inhibits new transcriptional initiation. Nat Genet. 2017;49(7):1045-1051. Epub 2017/05/16. doi: 10.1038/ng.3867. PubMed PMID: 28504701.
52. Vo Ngoc L, Kassavetis GA, Kadonaga JT. The RNA Polymerase II Core Promoter in Drosophila. Genetics. 2019;212(1):13-24. doi: 10.1534/genetics.119.302021. PubMed PMID: 31053615.
53. Hetzel J, Duttke SH, Benner C, Chory J. Nascent RNA sequencing reveals distinct features in plant transcription. Proc Natl Acad Sci U S A. 2016;113(43):12316-12321. Epub 2016/10/30. doi: 10.1073/pnas.1603217113. PubMed PMID: 27729530.
54. Barbosa C, Peixeiro I, Romao L. Gene expression regulation by upstream open reading frames and human disease. PLoS Genet. 2013;9(8):e1003529. Epub 2013/08/21. doi: 10.1371/journal.pgen.1003529. PubMed PMID: 23950723.
55. Vo Ngoc L, Huang CY, Cassidy CJ, Medrano C, Kadonaga JT. Identification of the human DPR core promoter element using machine learning. Nature. 2020. Epub 2020/09/11. doi: 10.1038/s41586-020-2689-7. PubMed PMID: 32908305.

56. Dreos R, Ambrosini G, Groux R, Cavin Perier R, Bucher P. The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. Nucleic acids research. 2017;45(D1):D51-D55. Epub 2016/12/03. doi: 10.1093/nar/gkw1069. PubMed PMID: 27899657.
57. Nair NU, Kumar S, Moret BM, Bucher P. Probabilistic partitioning methods to find significant patterns in ChIP-Seq data. Bioinformatics. 2014;30(17):2406-2413. Epub 2014/05/09. doi: 10.1093/bioinformatics/btu318. PubMed PMID: 24812341.
58. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature methods. 2013;10(12):1213-1218. Epub 2013/10/08. doi: 10.1038/nmeth.2688. PubMed PMID: 24097267.
59. Davie K, Jacobs J, Atkins M, Potier D, Christiaens V, Halder G, et al. Discovery of transcription factors and regulatory regions driving in vivo tumor development by ATAC-seq and FAIRE-seq open chromatin profiling. PLoS Genet. 2015;11(2):e1004994. Epub 2015/02/14. doi: 10.1371/journal.pgen.1004994. PubMed PMID: 25679813.
60. Dreos R, Ambrosini G, Groux R, Perier RC, Bucher P. MGA repository: a curated data resource for ChIP-seq and other genome annotated data. Nucleic acids research. 2018;46(D1):D175-D180. Epub 2017/10/27. doi: 10.1093/nar/gkx995. PubMed PMID: 29069466.
61. Ambrosini G, Dreos R, Kumar S, Bucher P. The ChIP-Seq tools and web server: a resource for analyzing ChIP-seq and other types of genomic data. BMC Genomics. 2016;17(1):938. Epub 2016/11/20. doi: 10.1186/s12864-016-3288-8. PubMed PMID: 27863463.
62. Vo Ngoc L, Huang CY, Cassidy CJ, Medrano C, Kadonaga JT. Identification of the human DPR core promoter element using machine learning. Nature. 2020;585(7825):459-463. Epub 2020/09/11. doi: 10.1038/s41586-020-2689-7. PubMed PMID: 32908305.

Supporting information

Public data used

ATAC-seq data for human lymphoblastoid cell line GM12878:

Source data: GEO series GSE47753, samples GSM1155957, GSM1155958, GSM1155959, GSM1155960

Processed data: MGA series buenrostro13, sample GM12878|ATACseq|50K|short ftp://ccg.epfl.ch/mga/hg19/buenrostro13/GM12878_50K.oriented.sga

ATAC-seq data for *Drosophila* wild type eye-antennal imaginal disc:

Source data: GEO series GSE59078, sample GSM1426261

Processed data: MGA series dm6/davie15/, sample WT|FAIRE|Control ftp://ccg.epfl.ch/mga/dm6/davie15/GSM1426261.sga

S1 Table. Three bp periodic distributions of ATG in human promoter classes 4-6.

Position	Class 4	Class 5	Class 6
+9	0	4	8
+10	12	0	4
+11	3	5	0
+12	0	1	5
+13	13	1	0
+14	5	8	0
+15	0	3	10
+16	9	0	6
+17	3	8	0
+9, +12, +15	0	8	23
+10, +13, +16	34	1	10
+11. +14, +17	11	21	0

S2 Table. The EM algorithm (this study) makes similar predictions as the SVRb model [62].

Human promoter	Sequence (+17 to +35)	Class 3 (EM) log score	SVRb score
LRCH4	CCGCCGGGAGCGGATGGCG	5.19	6.15
LRCH4_mPDP	CCGCCGGTAGCTTATGGCG	0.61	0.98
LRCH4_mGG	CCGCCGGGAGCGGATGGCT	5.20	7.16
LRCH4_pos10_m2	GCCGGGAGCGGATGGCGGC	-4.55	1.19
LRCH4_pos10_p2	AGCCGCCGGGAGCGGATGG	-0.79	0.99
LRCH4_pos18_m2	CCCGGGAGCGGATGGCGGC	-4.24	0.55
LRCH4_pos18_p2	CCTCGCCGGGAGCGGATGG	-0.16	0.12
CKS2	TTGCCTGGGCTGGACGTGG	2.95	7.97
CKS2_mPDP	TTGCCTGTGCTTTACGTGG	-1.63	1.52
CKS2_mGG	TTGCCTGGGCTGGACGTTT	2.96	8.83
CKS2_pos10_m2	GCCTGGGCTGGACGTGGTT	-2.41	1.69
CKS2_pos10_p2	TGTTGCCTGGGCTGGACGT	-8.95	0.67
CKS2_pos18_m2	TCCTGGGCTGGACGTGGTT	-2.26	1.38
CKS2_pos18_p2	TTTCGCCTGGGCTGGACGT	-8.47	1.23
ANP32E	TTGAAGGGGAAGGAACTGC	5.51	12.89
ANP32E_mPDP	TTGAAGGTGAATTAACTGC	0.94	3.17
CELF1	CAGCGGCGGCGGGACGCGG	2.81	5.44
CELF1_mPDP	CAGCGGCTGCGTTACGCGG	-1.76	1.56
CTSA	CTGGAGAGCAAGGACGCGG	3.81	8.45
CTSA_mPDP	CTGGAGATCAATTACGCGG	-0.77	1.36

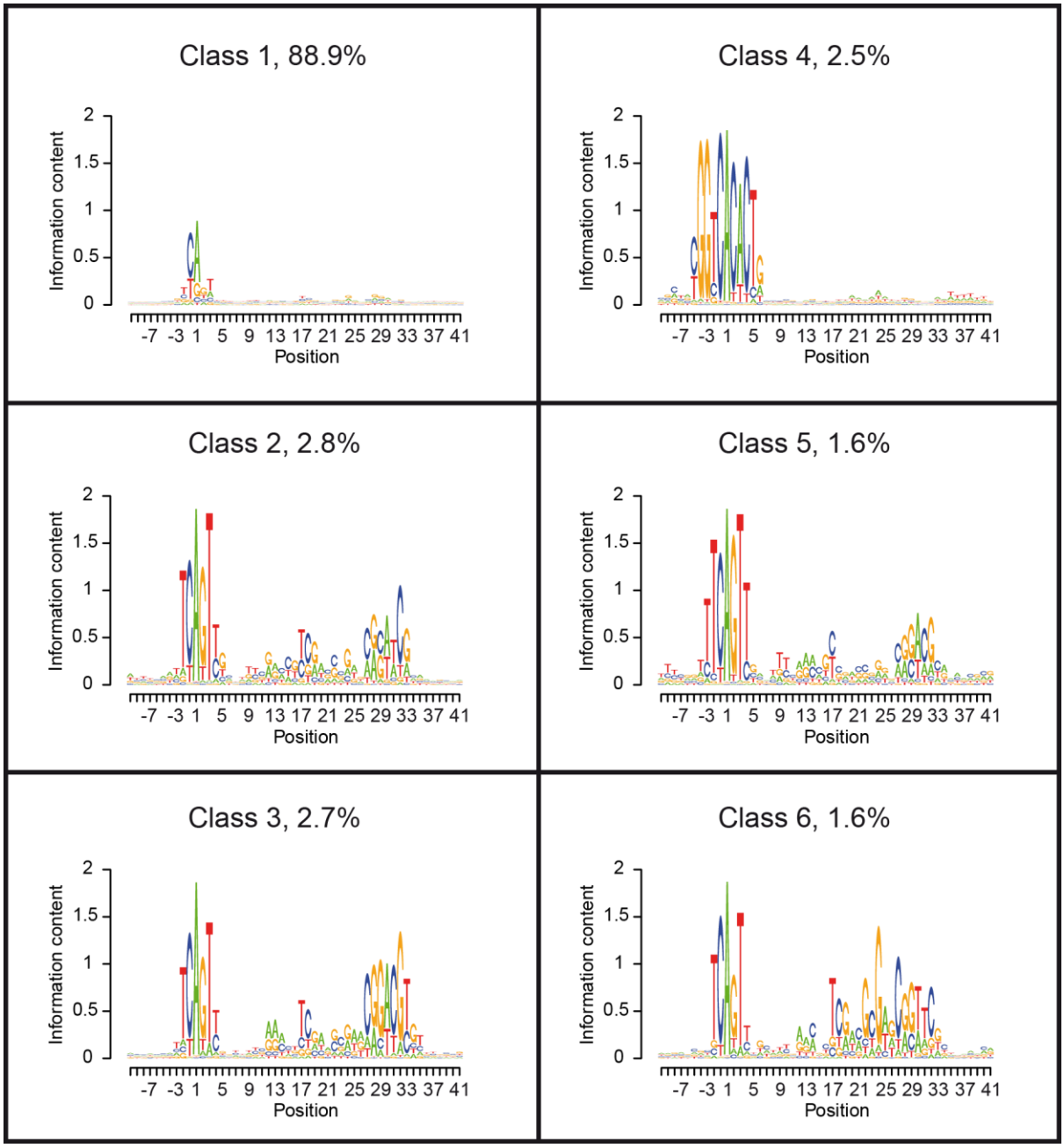
Columns 3 (EM algorithm, class 3 log score) and 4 (SVRb**)** correlate with a coefficient of about 0.75.

S3 Table. Primers used to generate the examined promoters.

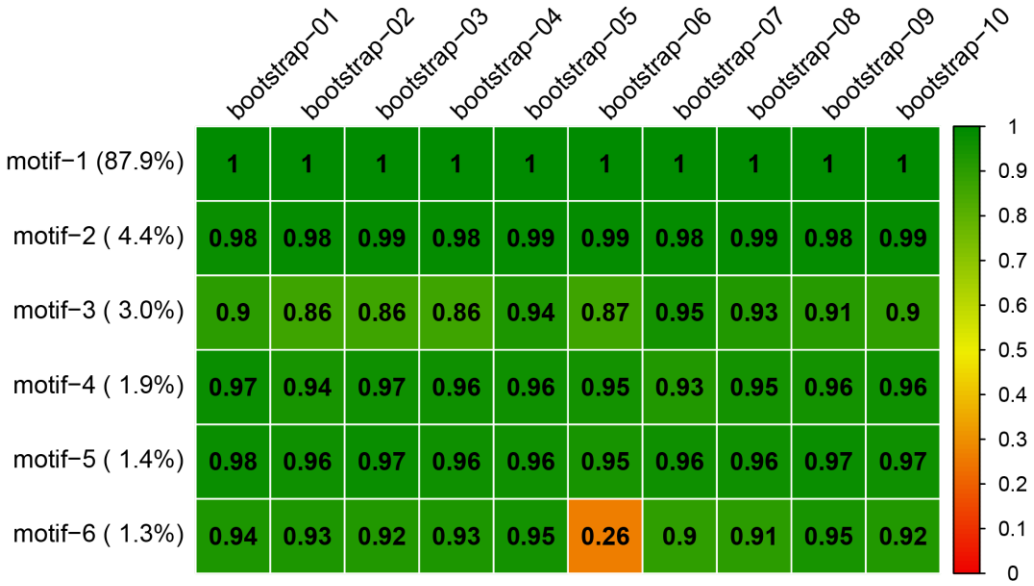
Primer name	sequence (5' to 3')
LRCH4_wt_Top	CCGGTCCCGTCAGTCAGGCAGCGGGAGCCGCCGGGAGCGGATGGCGGCGGCA
LRCH4_wt_bottom	CTAGTGCCGCCGCCATCCGCTCCCGGCGGCTCCCGCTGCCTGACTGACGGGACCGGGTAC
LRCH4_mPDP_Top	CCGGTCCCGTCAGTCAGGCAGCGGGAGCCGCCGGtAGCttATGGCGGCGGCA
LRCH4_mPDP_bottom	CTAGTGCCGCCGCCATaaGCTaCCGGCGGCTCCCGCTGCCTGACTGACGGGACCGGGTAC
LRCH4_pos10_m2_Top	CCGGTCCCGTCAGTCAGGCAGGGAGCCGCCGGGAGCGGATGGCGGCGGCA
LRCH4_pos10_m2_bottom	CTAGTGCCGCCGCCATCCGCTCCCGGCGGCTCCCTGCCTGACTGACGGGACCGGGTAC
LRCH4_pos10_p2_Top	CCGGTCCCGTCAGTCAGGCAG**tc**CGGGAGCCGCCGGGAGCGGATGGCGGCGGCA
LRCH4_pos10_p2_bottom	CTAGTGCCGCCGCCATCCGCTCCCGGCGGCTCCCG**ga**CTGCCTGACTGACGGGACCGGGTAC
LRCH4_pos18_m2_Top	CCGGTCCCGTCAGTCAGGCAGCGGGAGCCCGGGAGCGGATGGCGGCGGCA
LRCH4_pos18_m2_bottom	CTAGTGCCGCCGCCATCCGCTCCCGGGCTCCCGCTGCCTGACTGACGGGACCGGGTAC
LRCH4_pos18_p2_Top	CCGGTCCCGTCAGTCAGGCAGCGGGAGCC**tc**GCCGGGAGCGGATGGCGGCGGCA
LRCH4_pos18_p2_bottom	CTAGTGCCGCCGCCATCCGCTCCCGGC**ga**GGCTCCCGCTGCCTGACTGACGGGACCGGGTAC
LRCH4_mGG_Top	CCGGTCCCGTCAGTCAGGCAGCGGGAGCCGCCGGGAGCGGATGGC**tt**CGGCA
LRCH4_mGG_bottom	CTAGTGCCG**aa**GCCATCCGCTCCCGGCGGCTCCCGCTGCCTGACTGACGGGACCGGGTAC
ANP32E_wt_Top	CATGGAGGCTCAGTCTCTGAGCAGCCATTGAAGGGGAAGGAACTGCGGGTGA
ANP32E_wt_bottom	CTAGTCACCCGCAGTTCCTTCCCCTTCAATGGCTGCTCAGAGACTGAGCCTCCATGGTAC
ANP32E_mPDP_Top	CATGGAGGCTCAGTCTCTGAGCAGCCATTGAAGGtGAAttAACTGCGGGTGA
ANP32E_mPDP_bottom	CTAGTCACCCGCAGTTaaTTCaCCTTCAATGGCTGCTCAGAGACTGAGCCTCCATGGTAC
CKS2_wt_Top	CTGCGGTCGTTAGTCTCCGGCGAGTTGTTGCCTGGGCTGGACGTGGTTTTGTA
CKS2_wt_bottom	CTAGTACAAAACCACGTCCAGCCCAGGCAACAACTCGCCGGAGACTAACGACCGCAGGTAC
CKS2_mPDP_Top	CTGCGGTCGTTAGTCTCCGGCGAGTTGTTGCCTGtGCTttACGTGGTTTTGTA
CKS2_mPDP_bottom	CTAGTACAAAACCACGTaaAGCaCAGGCAACAACTCGCCGGAGACTAACGACCGCAGGTAC
CKS2_pos10_m2_Top	CTGCGGTCGTTAGTCTCCGGAGTTGTTGCCTGGGCTGGACGTGGTTTTGTA
CKS2_pos10_m2_bottom	CTAGTACAAAACCACGTCCAGCCCAGGCAACAACTCCGGAGACTAACGACCGCAGGTAC
CKS2_pos10_p2_Top	CTGCGGTCGTTAGTCTCCGGC**tc**GAGTTGTTGCCTGGGCTGGACGTGGTTTTGTA

CKS2_pos10_p2_bottom	CTAGTACAAAACCACGTCCAGCCCAGGCAACAACTC**ga**GCCGGAGACTAACGACCGCAGGTAC
CKS2_pos18_m2_Top	CTGCGGTCGTTAGTCTCCGGCGAGTTGTCCTGGGCTGGACGTGGTTTTGTA
CKS2_pos18_m2_bottom	CTAGTACAAAACCACGTCCAGCCCAGGACAACTCGCCGGAGACTAACGACCGCAGGTAC
CKS2_pos18_p2_Top	CTGCGGTCGTTAGTCTCCGGCGAGTTGTT**tc**GCCTGGGCTGGACGTGGTTTTGTA
CKS2_pos18_p2_bottom	CTAGTACAAAACCACGTCCAGCCCAGGC**ga**AACAACTCGCCGGAGACTAACGACCGCAGGTAC
CKS2_mGG_Top	CTGCGGTCGTTAGTCTCCGGCGAGTTGTTGCCTGGGCTGGACGT**tt**TTTTGTA
CKS2_mGG_bottom	CTAGTACAAAA**aa**ACGTCCAGCCCAGGCAACAACTCGCCGGAGACTAACGACCGCAGGTAC
CELF1_wt_Top	CGGGGTGTTCTGCTCTGGCGGCAGCGGCAGCGGCGGCGGGACGCGGAGGCTCA
CELF1_wt_bottom	CTAGTGAGCCTCCGCGTCCCGCCGCCGCTGCCGCTGCCGCCAGAGCAGAACACCCCGGTAC
CELF1_mPDP_Top	CGGGGTGTTCTGCTCTGGCGGCAGCGGCAGCGGCtGCGttACGCGGAGGCTCA
CELF1_mPDP_bottom	CTAGTGAGCCTCCGCGTaaCGCaGCCGCTGCCGCTGCCGCCAGAGCAGAACACCCCGGTAC
CTSA_wt_Top	CCATGACTTCCAGTCCCCGGGCGCCTCCTGGAGAGCAAGGACGCGGGGGAGCA
CTSA_wt_bottom	CTAGTGCTCCCCCGCGTCCTTGCTCTCCAGGAGGCGCCCGGGGACTGGAAGTCATGGGTAC
CTSA_mPDP_Top	CCATGACTTCCAGTCCCCGGGCGCCTCCTGGAGAtCAAttACGCGGGGGAGCA
CTSA_mPDP_bottom	CTAGTGCTCCCCCGCGTaaTTGaTCTCCAGGAGGCGCCCGGGGACTGGAAGTCATGGGTAC

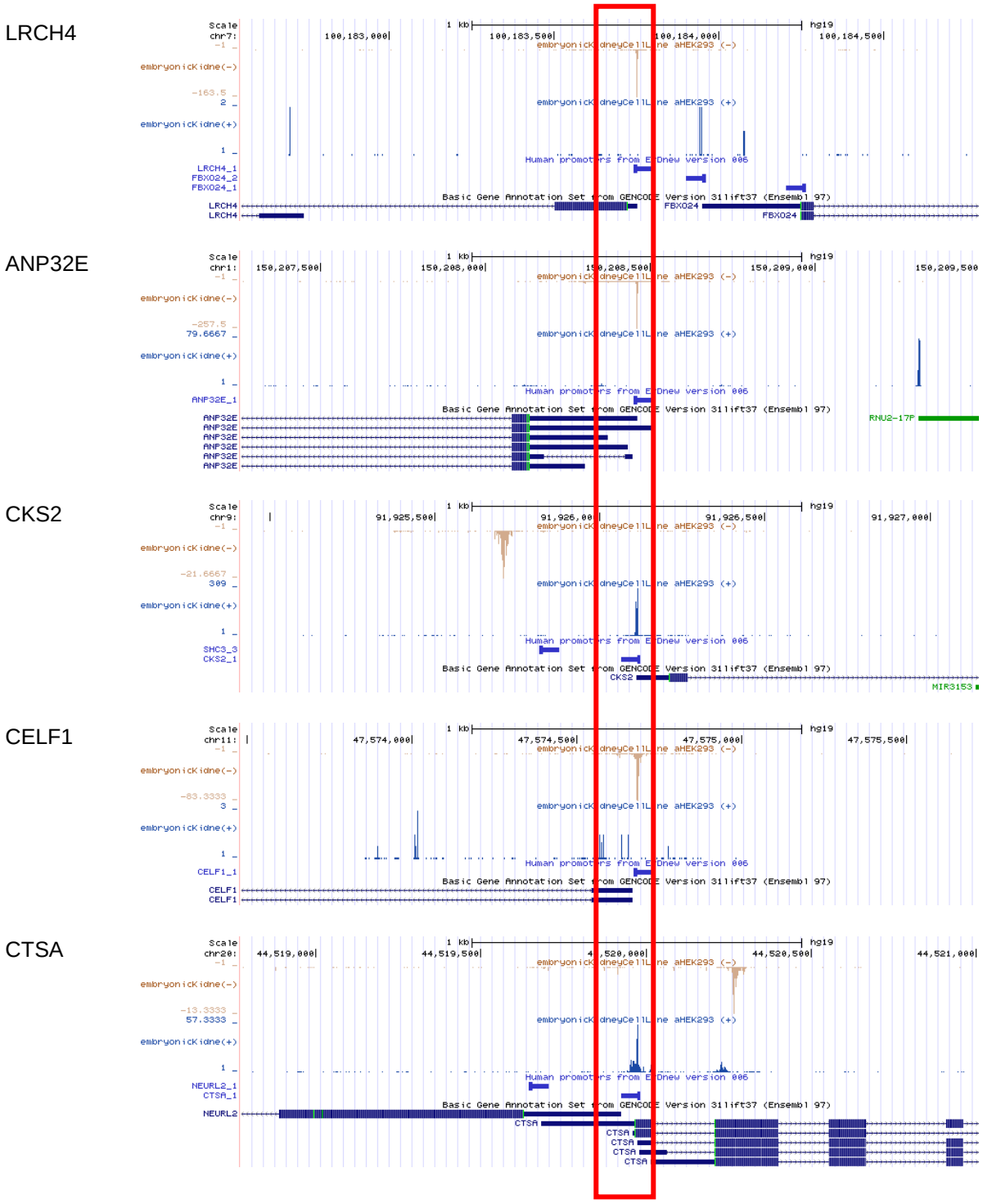
Primers comprising minimal core promoter sequences containing additional nucleotides in order to be ligated (following the annealing of top and bottom oligonucleotides) into a pGL3-Basic vector with a modified polylinker, digested with KpnI and SpeI restriction enzymes. Nucleotides added in the p2 promoters or mutated in the mGG promoters are depicted in bold lowercase letters.



S1 Fig. Partitioning of *Drosophila* promoter sequences with the extended EM algorithm.



S2 Fig. Bootstrap analysis of human promoter classes. The complete promoter sequence collection was resampled 10 times. The extended partitioning algorithm was applied to the bootstrapped data sets retaining the 10 most frequently found classes. The heatmap reflects the similarity (expressed as Pearson correlation coefficients) of the newly identified motifs with the corresponding most similar motifs found in each bootstrapping round.



S3 Fig. EPDnew screenshots of the analyzed promoters, used to define promoter shape. FANTOM5-generated CAGE tags distribution of individual promoters in HEK-293 cells was manually examined using the EPDnew viewer, in order to determine their transcription initiation pattern.