ElemeNT: a computational tool for detecting core promoter elements

Anna Sloutskin\textsuperscript{a}, Yehuda M Danino\textsuperscript{a}, Yaron Orenstein\textsuperscript{b}, Yonathan Zehavi\textsuperscript{a}, Tirza Doniger\textsuperscript{a}, Ron Shamir\textsuperscript{b} & Tamar Juven-Gershon\textsuperscript{a}

\textsuperscript{a} The Mina and Everard Goodman Faculty of Life Sciences; Bar-Ilan University; Ramat Gan, Israel
\textsuperscript{b} Blavatnik School of Computer Science; Tel-Aviv University; Tel Aviv, Israel

Published online: 30 Jul 2015.

To cite this article: Anna Sloutskin, Yehuda M Danino, Yaron Orenstein, Yonathan Zehavi, Tirza Doniger, Ron Shamir & Tamar Juven-Gershon (2015) ElemeNT: a computational tool for detecting core promoter elements, Transcription, 6:3, 41-50, DOI: 10.1080/21541264.2015.1067286

To link to this article: http://dx.doi.org/10.1080/21541264.2015.1067286

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the “Content”) contained in the publications on our platform. Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Versions of published Taylor & Francis and Routledge Open articles and Taylor & Francis and Routledge Open Select articles posted to institutional or subject repositories or any other third-party website are without warranty from Taylor & Francis of any kind, either expressed or implied, including, but not limited to, warranties of merchantability, fitness for a particular purpose, or non-infringement. Any opinions and views expressed in this article are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor & Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions

It is essential that you check the license status of any given Open and Open Select article to confirm conditions of access and use.
ElemeNT: a computational tool for detecting core promoter elements

Anna Sloutskin1, Yehuda M Danino1, Yaron Orenstein2, Yonathan Zehavi1, Tirza Doniger1, Ron Shamir2, and Tamar Juven-Gershon1,*

1The Mina and Everard Goodman Faculty of Life Sciences; Bar-Ilan University; Ramat Gan, Israel; 2Blavatnik School of Computer Science; Tel-Aviv University; Tel Aviv, Israel

Keywords: BRE, computational tool, core promoter elements/motifs, DPE, initiator, MTE, promoter prediction, RNAP II transcription, TATA box, TCT

Abbreviations: BRE, TFIIB recognition element; BRE4, BRE downstream of the TATA box; BREu, BRE upstream of the TATA box; DCE, downstream core element; DPE, downstream core promoter element; Inr, initiator; MTE, motif 10 element; PWM, position weight matrix; RNAP II, RNA Polymerase II; TBP, TATA box-binding protein; TAFs, TBP-associated factors; TSS, transcription start site.

Core promoter elements play a pivotal role in the transcriptional output, yet they are often detected manually within sequences of interest. Here, we present 2 contributions to the detection and curation of core promoter elements within given sequences. First, the Elements Navigation Tool (ElemeNT) is a user-friendly web-based, interactive tool for prediction and display of putative core promoter elements and their biologically-relevant combinations. Second, the CORE database summarizes ElemeNT-predicted core promoter elements near CAGE and RNA-seq-defined Drosophila melanogaster transcription start sites (TSSs). ElemeNT’s predictions are based on biologically-functional core promoter elements, and can be used to infer core promoter compositions. ElemeNT does not assume prior knowledge of the actual TSS position, and can therefore assist in annotation of any given sequence. These resources, freely accessible at http://lifefaculty.biu.ac.il/gershon-tamar/index.php/resources, facilitate the identification of core promoter elements as active contributors to gene expression.

Introduction

The uniqueness of each cell, as well as the differences between cell types in multicellular organisms, are largely achieved by distinct transcriptional programs. The regulation of transcription initiation is a complex process that is primarily based on the direct interactions between transcription factors and DNA. Transcription initiation occurs at the core promoter region where the RNA Polymerase II (RNAP II) binds, which is often referred to as the ‘gateway to transcription’.1-6 Although it was previously believed that the core promoter is a universal component that works in a similar mechanism for all protein-coding genes, it is nowadays established that core promoters differ in their architecture and function.3,4,7-10 Moreover, distinct core promoter compositions were demonstrated to result in diverse transcriptional outputs.11-15

Transcription initiation is generally thought to occur in either a focused or a dispersed manner with multiple combinations between these modes.4,7 Promoters that exhibit a dispersed initiation pattern typically contain multiple weak transcription start sites (TSSs) within a 50 to 100 bp region and are associated with CpG islands. In vertebrates, dispersed transcription initiation appears to account for the majority of protein-coding genes and is believed to direct the transcription of constitutively-expressed genes. In contrast, focused promoters contain a single predominant TSS or a few TSSs within a narrow region of several nucleotides, and are highly correlated with tightly regulated gene expression.4 The focused core promoter typically spans the region from −40 to +40 relative to the first transcribed nucleotide, which is usually termed “the +1 position.”

The focused core promoter area encompasses distinct DNA sequence motifs, termed core promoter elements or motifs. These elements are recognized by the basal transcription machinery to recruit RNAP II and to form the preinitiation complex.16-18 The TFIIID multi-subunit complex is a key basal transcription factor that recognizes the core promoter in the process of transcription initiation.16-19 Distinct TFIIID subunits, namely TATA box-binding protein (TBP) and TBP-associated factors (TAFs), recognize specific core promoter sequences.2,3,16,20-23

Table 1 and Figure 1 provide a summary of the characteristics of the known core promoter elements of focused promoters. Remarkably, the MTE, DPE and Bridge elements are exclusively dependent on the presence of a functional initiator with a strict spacing.
<table>
<thead>
<tr>
<th>Name</th>
<th>Position (relative to the TSS)</th>
<th>PWM logo representation</th>
<th>Consensus (in IUPAC characters)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>mammalian Initiator</td>
<td>−2 to +5</td>
<td></td>
<td>YYANWWYY</td>
<td>70</td>
</tr>
<tr>
<td>Drosophila Initiator</td>
<td>−2 to +4</td>
<td></td>
<td>TCAKTY</td>
<td></td>
</tr>
<tr>
<td>TATA box</td>
<td>−30/-31 to -23/-24</td>
<td></td>
<td>TATAWAAR</td>
<td>4,71</td>
</tr>
<tr>
<td>BRE&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Immediately upstream of the TATA box</td>
<td>SSRCGCC</td>
<td></td>
<td>45</td>
</tr>
<tr>
<td>BRE&lt;sup&gt;d&lt;/sup&gt;</td>
<td>Immediately downstream of the TATA box</td>
<td>RTDKKKK</td>
<td></td>
<td>44</td>
</tr>
<tr>
<td>DPE (Inr dependent)</td>
<td>+28 to +33</td>
<td></td>
<td>DSWYYV (functional range set)</td>
<td>20,21,24</td>
</tr>
<tr>
<td>MTE (Inr dependent)</td>
<td>+18 to +29</td>
<td></td>
<td>CSARCSSAACGS</td>
<td>25</td>
</tr>
<tr>
<td>Bridge (Inr dependent)</td>
<td>Part I: +18 to +22 Part II: +30 to +33</td>
<td>PART I: CGANC Part II: WYGT</td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>Drosophila TCT</td>
<td>−2 to +6</td>
<td></td>
<td>YYCTTTYY</td>
<td>48</td>
</tr>
<tr>
<td>Human TCT</td>
<td>−1 to +6</td>
<td></td>
<td>YCTYTYY</td>
<td>48</td>
</tr>
<tr>
<td>XCPE1</td>
<td>−8 to +2</td>
<td></td>
<td>DSGYGRASM</td>
<td>51</td>
</tr>
<tr>
<td>XCPE2</td>
<td>−9 to +2</td>
<td></td>
<td>VCYCRTTRCMY</td>
<td>72</td>
</tr>
<tr>
<td>DCE</td>
<td>+6 to +11, +16 to +21, +30 to +34</td>
<td>—</td>
<td>Necessary motifs: CTTC, CTGT, AGC</td>
<td>73,74</td>
</tr>
</tbody>
</table>

The table includes the position (relative to the TSS, +1), motif logo, IUPAC consensus sequence and references for each element.
requirement, and are typically enriched in TATA-less promoters.\textsuperscript{2-4,20,21,23-25}

An important aspect of core promoter elements is their synergistic nature. Although the presence of a specific core promoter element is usually sufficient to influence transcription, different combinations of core promoter elements exist, with some shown to act in concert, and, hence, affect the potency of the transcriptional outcome.\textsuperscript{11,26} It is therefore important to consider all the elements present within the same promoter in order to assess its transcriptional strength.

Prediction of core promoter motifs that affect the transcriptional output, in the absence of experimental validation, is a difficult task. The majority of currently available promoter prediction programs search for over-represented motifs in a given set of promoter sequences (based on annotated TSSs), rather than known core promoter elements.\textsuperscript{27-29} Most of these programs utilize other features, such as transcription factor binding sites, physical properties of the DNA, DNA accessibility, RNAP II occupancy and various epigenetic markers.\textsuperscript{29-35} However, even available programs that aim to identify core promoter elements, such as McPromoter\textsuperscript{36} and Eukaryotic Core Promoter Predictor (YAPP, http://www.bioinformatics.org/yapp/cgi-bin/yapp.cgi), rarely consider the functional constraint of the strict spacing required by the Inr-dependent elements, namely, DPE, MTE, and Bridge.

The selection of promoters that comprise the data set used to predict core promoter elements based on position weight matrices (PWMs) is of pivotal importance, as subtle variations in the sequences may generate completely different PWMs.\textsuperscript{31} Motif finding algorithms, such as XXmotif, can be used to accurately construct a PWM for over-represented motifs within a given set of sequences.\textsuperscript{37,38} Unfortunately, even a perfect model that is only based on sequence features, cannot exclusively account for the observed transcriptional activity, as most of the sequence motifs are short and redundant, and can thus be found in many non-transcriptionally active regions of the genome.\textsuperscript{31} Using experimentally-validated sequences rather than over-represented motifs, can greatly enhance the strength of the prediction program, although it cannot fully guarantee the accuracy of the prediction. Currently, the experimental readout of transcription strength and start sites resulting from mutated promoter sequences is not performed on a high-throughput scale; hence, the currently available experimental results are prone to be biased. Moreover, the known biologically functional sequences may slightly differ from the determined consensus; as a result, a tool for efficient detection of candidate core promoter elements is needed.

Importantly, annotation of individual promoters for the presence of specific core promoter elements can facilitate the discovery of gene groups co-regulated via a common core promoter motif. In a previous study, 205 experimentally-determined \textit{Drosophila} TSSs were manually annotated for the presence of TATA-box, Initiator and DPE to explore their role and function in gene regulation.\textsuperscript{24} This annotation facilitated the discovery that the \textit{Drosophila} Hox gene network is regulated via the DPE.\textsuperscript{39} A more comprehensive analysis of the whole \textit{Drosophila} transcriptome revealed that DPE-containing genes are conserved and highly prevalent among the target genes of Dorsal, a key regulator of dorsal-ventral axis formation.\textsuperscript{12} These examples demonstrate that the comprehensive annotation of core promoter elements in transcripts can greatly advance the understanding of gene expression regulation.

Here we describe 2 contributions in the detection and cura-
tion of core promoter elements within sequences of interest, based on experimentally validated sequences. The Elements Navigation Tool (ElemeNT) is a user-friendly web-based, interactive tool for prediction and display of putative core promoter elements and their biologically-relevant combinations in any given sequence, without a need for prior determination of the TSS. The CORE database utilizes the ElemeNT algorithm to annotate putative core promoter elements near CAGE\textsuperscript{40} and RNA-seq\textsuperscript{41} defined \textit{Drosophila melanogaster} TSSs. Together, both the ElemeNT program and the CORE database present new improved tools to assess the presence of core promoter elements within a given DNA sequence.

\section*{Methods}

\subsection*{Availability}
CORE and ElemeNT are freely accessible at http://lifefaculty.\textsuperscript{42} biu.ac.il/gershon-tamar/index.php/resources. Each resource is described in a separate page, providing both documentation and resources.

\subsection*{The ElemeNT algorithm}
Given a sequence of interest, the algorithm detects in it putative elements whose PWM-similarity to known core promoter elements is above a threshold. For each core promoter element, the user can specify a threshold between 0 and 1 for the presence of the element at a position. Default threshold values were empirically determined for each element, based on known functional sequence elements.

For a PWM matrix $P$ with $k$ columns, the PWM score is calculated for each sub-sequence of length $k$ (k-mer) in the sequence elements. In a previous study, 205 experimentally-determined \textit{Drosophila} TSSs were manually annotated for the presence of TATA-box, Initiator and DPE to explore their role and function in gene regulation.\textsuperscript{24} This annotation facilitated the discovery that the \textit{Drosophila} Hox gene network is regulated via the DPE.\textsuperscript{39} A more comprehensive analysis of the whole \textit{Drosophila} transcriptome revealed that DPE-containing genes are conserved and highly prevalent among the target genes of Dorsal, a key regulator of dorsal-ventral axis formation.\textsuperscript{12} These examples demonstrate that the comprehensive annotation of core promoter elements in transcripts can greatly advance the understanding of gene expression regulation.

Here we describe 2 contributions in the detection and cura-
tion of core promoter elements within sequences of interest, based on experimentally validated sequences. The Elements Navigation Tool (ElemeNT) is a user-friendly web-based, interactive tool for prediction and display of putative core promoter elements and their biologically-relevant combinations in any given sequence, without a need for prior determination of the TSS. The CORE database utilizes the ElemeNT algorithm to annotate putative core promoter elements near CAGE\textsuperscript{40} and RNA-seq\textsuperscript{41} defined \textit{Drosophila melanogaster} TSSs. Together, both the ElemeNT program and the CORE database present new improved tools to assess the presence of core promoter elements within a given DNA sequence.

\section*{Methods}

\subsection*{Availability}
CORE and ElemeNT are freely accessible at http://lifefaculty.\textsuperscript{42} biu.ac.il/gershon-tamar/index.php/resources. Each resource is described in a separate page, providing both documentation and resources.

\subsection*{The ElemeNT algorithm}
Given a sequence of interest, the algorithm detects in it putative elements whose PWM-similarity to known core promoter elements is above a threshold. For each core promoter element, the user can specify a threshold between 0 and 1 for the presence of the element at a position. Default threshold values were empirically determined for each element, based on known functional sequence elements.

For a PWM matrix $P$ with $k$ columns, the PWM score is calculated for each sub-sequence of length $k$ (k-mer) in the sequence elements. In a previous study, 205 experimentally-determined \textit{Drosophila} TSSs were manually annotated for the presence of TATA-box, Initiator and DPE to explore their role and function in gene regulation.\textsuperscript{24} This annotation facilitated the discovery that the \textit{Drosophila} Hox gene network is regulated via the DPE.\textsuperscript{39} A more comprehensive analysis of the whole \textit{Drosophila} transcriptome revealed that DPE-containing genes are conserved and highly prevalent among the target genes of Dorsal, a key regulator of dorsal-ventral axis formation.\textsuperscript{12} These examples demonstrate that the comprehensive annotation of core promoter elements in transcripts can greatly advance the understanding of gene expression regulation.

Here we describe 2 contributions in the detection and cura-
tion of core promoter elements within sequences of interest, based on experimentally validated sequences. The Elements Navigation Tool (ElemeNT) is a user-friendly web-based, interactive tool for prediction and display of putative core promoter elements and their biologically-relevant combinations in any given sequence, without a need for prior determination of the TSS. The CORE database utilizes the ElemeNT algorithm to annotate putative core promoter elements near CAGE\textsuperscript{40} and RNA-seq\textsuperscript{41} defined \textit{Drosophila melanogaster} TSSs. Together, both the ElemeNT program and the CORE database present new improved tools to assess the presence of core promoter elements within a given DNA sequence.

\section*{Methods}

\subsection*{Availability}
CORE and ElemeNT are freely accessible at http://lifefaculty.\textsuperscript{42} biu.ac.il/gershon-tamar/index.php/resources. Each resource is described in a separate page, providing both documentation and resources.

\subsection*{The ElemeNT algorithm}
Given a sequence of interest, the algorithm detects in it putative elements whose PWM-similarity to known core promoter elements is above a threshold. For each core promoter element, the user can specify a threshold between 0 and 1 for the presence of the element at a position. Default threshold values were empirically determined for each element, based on known functional sequence elements.

For a PWM matrix $P$ with $k$ columns, the PWM score is calculated for each sub-sequence of length $k$ (k-mer) in the sequence elements.
**sequence, by multiplying the appropriate values of the PWM for each consecutive position, as follows:**

$$PWM\_SCORE(S_{i+1;i+k}, P) = \prod_{j=i}^{i+k} P(j, S_{i+1})$$

where $S_{i+1;i+k}$ is a k-mer starting at position $i+1$ in sequence $S$ and $P(j, x)$ is the probability for nucleotide $x$ at position $j$ in $P$, normalized so that for a given $j$, $\max\{P(j, x)\} = 1$. The role of this normalization is to guarantee that the final PWM score for every element is between 0 and 1, irrespective of the PWM’s parameters. Each sub-sequence with a score exceeding the specified threshold is termed a ‘hit’. The score is calculated for $0 < i < n - k$, where $n$ is the length of the input sequence $S$, and hits are displayed in a list sorted in descending score order for each element. Consensus match scores, which are the number of nucleotide matches of the hit to the motif’s consensus (Table 1), are also reported for each hit. The flow diagram of the ElemeNT algorithm is depicted in **Figure S1**. The PWMs used, as well as their construction processes, are described in File S1.

**CORE construction**

CORE database construction was based on both CAGE- and RNA-seq-experimentally verified *Drosophila* TSSs. CAGE-based TSSs were determined based on Hoskins et al.40 For each CAGE peak, the reported probability density functions (PDFs) were used to determine the most probable TSS. If two or more positions at >10 bp distance from each other were assigned with the highest TSS probability, each was considered as a separate TSS. The RNA-seq observed TSSs were reported by Nechaev et al.41 For each determined TSS, the sequence encompassing the TSS ±50 bp was used for downstream analysis by the ElemeNT algorithm, using default score cutoff values.

For each core promoter element, the position relative to the TSS and the corresponding score are reported for all hits within the allowed range (±5 bp relative to the predicted position). All listed positions are with respect to the starting nucleotide of the relevant motif. We list the elements used to construct the CORE, with the relative positions and the cutoff scores provided in parenthesis: BREu (-37, 0.05); TATA box (-30, 0.01); BREd (-24, 0.5); *Drosophila* Inr (-2, 0.01); *Drosophila* TCT (-2, 0.1). The Inr-dependent DPE, MTE and Bridge elements were only considered at the precise starting positions Inr+30, Inr+20 and Inr+20, respectively, with cutoff scores of 0.01. A summary of the total numbers of hits of each element within the CAGE and RNA-seq datasets is described in a separate sheet.

**GO terms analysis**

GO terms enrichment was assessed using the PANTHER classification system42 (http://pantherdb.org).43 For each examined element (TATA, Dlnr, DPE, MTE and dTCT), 5 distinct lists were created based on the CORE results- CAGE peaked, CAGE broad, CAGE unclassified, all CAGE tags and RNA-seq. For CAGE data, the classification of promoter types was used as provided with the original data set (see below).40 Each list of genes was analyzed by the PANTHER overrepresentation test (release 20141219) against the *Drosophila* melanogaster reference list, using GO biological
process complete annotation data set (GO ontology database released 2015-04-13). The Bonferroni correction for multiple testing was applied. While enrichment values range between 0.2 and ‘> 5’, only results with fold enrichment ≥ 4 are reported.

Results

The elements navigation tool

In order to facilitate the identification of putative core promoter elements and their biologically relevant combinations within a sequence of interest, we developed the Elements Navigation Tool (ElemeNT). ElemeNT is a web-based, interactive tool for rapid and convenient detection of core promoter elements and their combinations within any given sequence. Core promoter elements have been shown to function at a specific distance from the TSS and to affect transcription (e.g., as examined by mutational analysis). ElemeNT scans the input sequences, applying user-specified parameters, for the presence of core promoter elements that are precisely located relative to the TSS (Fig. 1).

The elements are represented by PWMs, which were constructed based on validated biologically functional sequences (Table 1, File S1). Notably, for some elements, the PWMs differ from the consensus sequences reported in the literature, reflecting differences in the data sources used to generate these models. The elements that can be searched for are: mammalian initiator, *Drosophila* initiator, TATA box, MTE, DPE, Bridge, BRE<sup>u</sup>, BRE<sup>d</sup>, human TCT, *Drosophila* TCT, XCPE1 and XCPE2 (Table 1, Fig. 1). The MTE, DPE and Bridge motifs are only scored at the precise location relative to each detected mammalian/*Drosophila* initiator, based on the known strict spacing requirement that is crucial for these elements to be functional. The TATA box motif is derived from canonical TATA boxes whose 5' T is located at -30 or 31 relative to the TSS. Furthermore, the user can search the sequence for any PWM provided by the user. The scores are normalized to a scale of 0 to 1, to allow standardization and comparison between distinct elements. The ElemeNT algorithm is described in the Methods section, and its flow is illustrated in Figure S1.

The output of the program contains the analyzed sequences, a color display of potential combinations of core promoter elements identified, and a table containing the name of each of the detected elements, alongside its position, the sequence, its PWM score and relevant positions, although the observed peaks were less profound than the distribution peaks (Fig. 4). Notably, the PWMs were constructed based on completely different data sets displayed if the specific TATA box is located 26 to 40 bp upstream of the A+1 of an initiator. In addition, a message is displayed if a BRE<sup>u</sup> or BRE<sup>d</sup> is located in close proximity to the specific TATA box.44-46

To partially assess the performance of the ElemeNT tool, a set of experimentally validated core promoter sequences were analyzed by the tool. The analysis of the *Drosophila* Inr is presented as an example (Fig. S2). Importantly, ElemeNT detected most of the biologically functional *Drosophila* initiator motifs among the dataset at cutoff values around 0.01. As expected, lower threshold values used led to detection of a greater number of correct hits, at a cost of a higher false positive rate. False negative hits were scored as well, based on missed motifs. Previously validated sequence variations in core promoter elements resulted in score values of 0.005-0.01, further supporting the defined default cutoffs (data not shown).24

The CORE database

The ElemeNT algorithm was employed to predict core promoter composition of all *Drosophila melanogaster* transcripts (File S2). TSSs were obtained based on both CAGE<sup>40</sup> and RNA-seq<sup>41</sup> data-determined *Drosophila melanogaster* TSSs. The sequence around each TSS was annotated for the presence of core promoter elements near the expected position relative to the TSS. In addition, we summarized the frequencies of the detected elements among the *Drosophila* transcripts. Importantly, the fraction of promoters containing the distinct elements was similar in the CAGE and RNA-seq data sets. The total analyzed transcripts contained 6-8% TATA box motifs, ~55% Inr, ~17% DPE and ~1% TCT. The CAGE-defined transcripts were previously categorized as peaked, broad and unclassified promoter classes.<sup>40</sup> Inr, TATA box and DPE elements were enriched among peaked promoters, as compared to broad and unclassified subsets (Inr- 71% vs. 48% and 54%, TATA box- 14% vs. 6% and 9%, DPE- 32% vs. 11% and 18%, respectively). In contrast, the rare TCT element was slightly more prevalent among the broad promoters class, compared to peaked and unclassified (1.5% vs. 0.5% and 1%, respectively). These results are in the same range as the proportions reported in the original study<sup>40</sup> – 70% and 35% Inr, 16% and 4% TATA-box for peaked and broad promoters, respectively.

The distribution of elements found among the allowed positions peaked around the expected relative position (Fig. 3). This peak was observed in both CAGE and RNA-seq data, suggesting that the detected elements are biologically functional. Extending the allowed range around the relative position from ±5 bp to ±10 bp did not reveal additional elements (Fig. S3); hence, a ±5 bp range was used for all downstream analyses. Additionally, the distribution of detected elements among the CAGE defined peaked, broad and unclassified promoters did not differ greatly from the overall distribution (Fig. S4). Reassuringly, the average PWM score also peaked at the biologically relevant positions, although the observed peaks were less profound than the distribution peaks (Fig. 4).
obtained by entirely different experimental approaches, as compared to the CAGE and RNA-seq datasets.

We also evaluated the CORE database accuracy by GO term analysis of genes sets that were found to contain the same element. We used PANTHER classification system for this aim.\textsuperscript{42,43} The results (summarized in Table 2, and fully presented in File S3) indicate that distinct GO terms categories are associated with the different core promoter elements. While only a few specific categories were enriched in TATA-containing genes, DPE-containing genes were mostly enriched for development-related gene categories. TCT-containing genes were mostly enriched for translation and ribosomal-related proteins, as well as for structural proteins related to mitosis. Remarkably, the observed enriched categories are in agreement with previous reports, where the DPE was found be associated with developmental genes and TCT with housekeeping and ribosomal genes.\textsuperscript{4,47-49}

**Discussion**

Core promoter elements, located in the immediate vicinity of the TSSs, have a great effect on the transcriptional output.\textsuperscript{4,7} The majority of core promoter elements were identified as DNA sequences that are recognized by components of the preinitiation complex.\textsuperscript{20,44,45,50,51} In addition, overrepresented motifs were discovered in the region around the annotated TSSs.\textsuperscript{52-54} Some of these motifs affected the transcriptional outcome\textsuperscript{25} and some were bound by transcription-regulating proteins.\textsuperscript{55}

The uniqueness of the ElemeNT program, as compared to other
promoter-prediction software, is its major focus on biologically-functional core promoter elements. This is manifested by 2 major principles adapted in the algorithm. The first is the exclusive use of experimentally validated core promoter motifs, rather than overrepresented motifs, to construct the PWMs used. The second is the obligatory presence of an initiator, and the strict spacing for the downstream promoter elements MTE, DPE, and Bridge, which are crucial for the functionality of the downstream

<table>
<thead>
<tr>
<th>TATA peak</th>
<th>CAGE peak</th>
<th>TATA broad</th>
<th>CAGE all tags</th>
<th>RNA-seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAGE peak</td>
<td>chitin-based cuticle development</td>
<td>branch fusion, open tracheal system</td>
<td>heart development</td>
<td>heart development</td>
</tr>
<tr>
<td></td>
<td>cuticle development</td>
<td>tube fusion</td>
<td>circulatory system development</td>
<td>circulatory system development</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cardiocyte differentiation</td>
<td>peripheral nervous system development</td>
<td>peripheral nervous system development</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ventral cord development</td>
<td>digestive system development</td>
<td>digestive system development</td>
</tr>
<tr>
<td></td>
<td></td>
<td>genital disc development</td>
<td>digestive tract development</td>
<td>digestive tract development</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>reproductive system development</td>
<td>reproductive system development</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>reproductive structure development</td>
<td>reproductive structure development</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAGE broad</td>
<td>chitin-based cuticle development</td>
<td>NO ENRICHMENT</td>
<td>negative regulation of molecular function</td>
<td>negative regulation of molecular function</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAGE unclassified</td>
<td>chitin-based cuticle development</td>
<td>stem cell fate commitment</td>
<td>renal system development</td>
<td>renal system development</td>
</tr>
<tr>
<td></td>
<td></td>
<td>regulation of protein localization to nucleus</td>
<td>urogenital system development</td>
<td>urogenital system development</td>
</tr>
<tr>
<td></td>
<td></td>
<td>female meiosis chromosome segregation</td>
<td>pigment metabolic process</td>
<td>pigment metabolic process</td>
</tr>
<tr>
<td></td>
<td></td>
<td>regulation of protein import into nucleus</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAGE all tags</td>
<td>chitin-based cuticle development</td>
<td>NO ENRICHMENT</td>
<td>cardioocyte differentiation</td>
<td>heart development</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>circulatory system development</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cardiovascular system development</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>renal system development</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>urogenital system development</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>skeletal muscle organ development</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>muscle attachment</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Top enriched GO terms categories associated with the analyzed data sets

For each dataset, up to 7 categories that showed significant enrichment ($P < 0.05$ after Bonferroni corrections) are listed. In case there were more than 7, the top 7 according to the $P$-value are shown. The different elements are enriched for distinct biological processes categories. The full list of categories along with their $P$-values is presented in file S3.
elements. These are overlooked by most of the core promoter elements prediction programs.\(^{27,29,32,35,36}\) Moreover, the identification of combinations of elements, which were experimentally demonstrated to result in synergistic effects,\(^{11,25,26}\) may spark new research directions. In contrast to most of the available promoter prediction programs, the web-based ElemeNT is not designed to produce or analyze a genome-scale data, but is rather intended to narrow down a given region of interest, considering the currently available, experimentally-validated information about core promoter motifs themselves.

The determination of actual TSSs, which influence the motifs discovered in their vicinity, is a critical factor in the prediction of core promoter elements. The TSS of the same gene can vary across the developmental stages, tissues, and time points sampled, which presents a great challenge for integration of the data provided by different studies. To date, a wealth of rapidly evolving high-throughput techniques to identify features and sequences that might affect transcription are available; these include PEAT,\(^{56}\) CAGE,\(^{57}\) FAIRE-seq,\(^{58}\) ChIP-seq,\(^{59}\) and GRO-seq.\(^{60}\) The integrated results will be of utmost importance for re-defining TSSs.

We used the ElemeNT algorithm to annotate \textit{Drosophila melanogaster} TSSs defined by either CAGE\(^{40}\) or RNA-seq\(^{41}\) for the different core promoter elements. A major contribution of the CORE database for core promoter elements curation is the ability to easily identify all the core promoter elements associated with a specific \textit{Drosophila} gene, without any previous knowledge.

Generally, CAGE and RNA-seq data showed similar percentages of core promoter elements among the total transcripts. The total frequencies of the TATA box and Inr were in concordance with the numbers reported in the original study.\(^{40}\) However, the original reports on DPE percentages (5% within peaked promoters, 1.5% within broad promoters) are significantly lower than the frequency detected in the CORE database (32% peaked, 11% broad). This discrepancy likely arises from the different approaches taken; while Hoskins et al. searched for a consensus DPE sequence\(^{20}\) within 5 bp of position +26, we have looked for the more biologically relevant functional range set\(^{24}\) located at a precise +26 distance relative to a detected Inr.

Another aspect highlighting the biological relevance of the obtained results is the peak of both the frequency and the average PWM score at the expected positions relative to the TSS (Fig. 3, Fig. 4). The fact that these peaks are clearly evident indicates that both TSS determination and PWM construction have been performed accurately. Further positional constraints apply to the Inr dependent elements—DPE, MTE, and Bridge, as discussed above. Surprisingly, the more strict spacing requirements used in this study yielded a higher proportion of DPE-containing transcripts, thus highlighting the importance of annotation guidelines based on experimentally-validated elements. The TATA box, Inr and DPE elements were enriched among peaked promoters, while the TCT was enriched among the broad promoters class, recapitulating previous observations and highlighting the biological relevance of the obtained results.\(^{32,40,49}\)

In addition, GO terms enrichment differed significantly among the gene groups containing distinct core promoter elements (Table 2, File S3), mostly in agreement with the literature.\(^{4,7,32,40}\) The DPE, which was shown to functionally regulate gene expression of developmental gene networks, namely Hox genes\(^{39}\) and mesodermal genes,\(^{12}\) was found to be enriched among circulatory system developmental genes, consistent with the previous findings.\(^{13}\) The Inr element, which is the most abundant motif and is associated with tightly regulated genes, was not found to be enriched for specific gene groups among the total transcripts group. A possible interpretation is that since the Inr is prevalent among most gene groups no enrichment is detected when examining the whole transcriptome. Focused transcription initiation was previously associated with spatiotemporally regulated tissue-specific genes and with canonical core promoter elements that have a positional bias, such as the TATA box, Initiator, MTE and DPE.\(^{41,62}\) However, broad (dispersed) promoters often contain a distinct set of elements with weaker positional biases (as compared to the focused promoters), as Ohler 1, DNA replication element (DRE), Ohler 6, and Ohler 7\(^{40,62}\) (a detailed discussion is available in refs 4\(^{3}\)).\(^{110}\) When considering separately the CAGE-defined peaked, broad and unclassified promoter classes, a clear enrichment for developmental processes is evident in the peaked and unclassified subsets. This most probably reflects the DPE-containing Inr fraction, highlighting the major contribution of the DPE motif to transcriptional regulation. The TCT element, which was originally reported to be present among translation and ribosomal-related genes,\(^{48}\) was indeed found to be strongly enriched among these gene groups. In addition, structural processes related to mitosis, such as spindle, microtubule, and centrosome related proteins, were enriched. This highlights the importance of the core promoter elements annotation of individual genes, revealing distinct functions associated with a core promoter element.

The algorithm’s performance depends on the accuracy of the constructed models. The redundancy of the core promoter motifs may lead to the identification of sequences that match functionally verified sequences, yet are not functional. Nevertheless, their presence might indicate that the specific genomic locus is transcriptionally active. Based on experience with transcription factors binding motifs,\(^{63}\) sorting out the functionally relevant hits might prove to be a difficult task and will require individual examination. Future improvements of the algorithm will be based on new insights and a better understanding of transcriptional regulation, obtained by ongoing work of major projects and consortia. These are aimed at dissecting the rules governing transcriptional regulation, and include ENCODE,\(^{64}\) modENCODE,\(^{65}\) and FANTOM5,\(^{66}\) as well as other genome-wide studies.\(^{67,68}\) Importantly, the ElemeNT program can assist in the analysis of sequences from organisms whose TSSs have not yet been comprehensively defined. For example, both the TATA box and the BRE motifs are conserved from Archaea to humans\(^{69}\) and many organisms whose transcriptomes have not been annotated, are likely to contain such core promoter elements.

In conclusion, we anticipate that the ElemeNT tool, along with the CORE database, will make the search for specific core promoter elements and their combinations within \textit{Drosophila} transcripts or any sequence of interest, accessible to scientists and
help in elucidating the major role core promoter elements play in gene expression.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Acknowledgments
We thank Marina Socol, Boris Komraz and Dr. Eli Sloutskin for invaluable assistance in ElemeNT development and web execution. We thank Gal Nuta for assisting with optimization of ElemeNT parameters. We thank Dr. Diana Ideses, Dan Even, Adi Kedmi, Hila Shir-Shapira and Gal Nuta for critical reading of the manuscript.

Supplemental Material
Supplemental data for this article can be accessed on the publisher’s website.

References


55. Li J, Gilmour DS. Distinct mechanisms of transcriptional pausing orchestrated by GAGA factor and MED17, a novel transcription factor. The EMBO J 2013; 32:1829-41; PMID:23789876; dx.doi.org/10.1002/emboj.2013.111


60. Fujita TS. CHIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat Rev Genet 2012; 13:840-52; PMID:23090257; dx.doi.org/10.1038/nrg3306


70. Smale ST, Baltimore D. The “initiator” as a transcription control element. Cell 1989; 57:103-13; PMID:24677642; dx.doi.org/10.1006/ajcb.2006.0749


