Universality of core promoter elements?

ARISING FROM B. J. Venters & B. F. Pugh *Nature* **502,** 53–58 (2013); doi:10.1038/nature12535

How cells locate the regions to initiate transcription is an open question, because core promoter elements (CPEs) are found in only a small fraction of core promoters[1–4]. A recent study[5] measured 159,117 DNA binding regions of transcription factor IIB (TFIIB) by ChIP-exo (chromatin immunoprecipitation with lambda exonuclease digestion followed by high-throughput sequencing) in human cells, found four degenerate CPEs—upstream and downstream TFIIB recognition elements (BRE_u and BRE_d), TATA and initiator element (INR)—in nearly all of them, and concluded that these regions represent sites of transcription initiation marked by universal CPEs. We show that the claimed universality of CPEs is explained by the low specificities of the patterns used and that the same match frequencies are obtained with two negative controls (randomized sequences and scrambled patterns). Our analyses also cast doubt on the biological significance of most of the 150,753 non-messenger-RNA-associated ChIP-exo peaks, 72% of which lie within repetitive regions. There is a Retraction accompanying this Brief Communication Arising by Venters, B. J. & Pugh, B. F. *Nature* **511,** http://dx.doi.org/10.1038/nature13588 (2014).

Short sequence motifs such as the INR consensus YYANWYY may occur frequently by chance, in particular when allowing up to three mismatched positions. The probability of observing an exact match at any one position of a random sequence with 60% GC content is $P_Y{}^4 P_A P_W = P_{(C\,or\,T)}{}^4 P_A P_{(A\,or\,T)} = 0.5^4 \times 0.2 \times 0.4 = 0.005$. Hence the probability of seeing no match within 60 possible start positions (search space 60) is $(1–0.005)^{60} = 0.74$, and the probability of observing at least one match is $1–0.74 = 0.26$. This and similar estimates for one to three mismatches were in strong disagreement with the negative controls in figures 2c and 3e of ref. 5. We therefore checked the reported results using two negative control procedures.

We first analysed ChIP-exo peaks near annotated transcription start sites of mRNAs. Although the match frequencies of CPEs around ChIP-exo peaks reported in ref. 5 agree with our results, the three negative controls indeed match far too infrequently. We could reproduce one of the controls (60% GC random sequences) by assuming a wrong search space size of 1 instead of 161 (TATA), 60 (INR), or 40 (BRE_u and BRE_d), respectively. Our negative controls closely follow the match frequencies of the four CPEs that had been observed around ChIP-exo peaks (Fig. 1a). Therefore, the frequent occurrence of the four CPEs in the ChIP-exo peaks is fully explained by their very low specificity (owing to allowing up to three mismatches).

We next investigated the positional enrichment of CPE pattern matches around mRNA-associated ChIP-exo peaks (figures 2d and 3d in ref. 5). For TATAWAWR with up to three mismatches we expect about 5.3 chance matches per sequence on average. By selecting only the motif

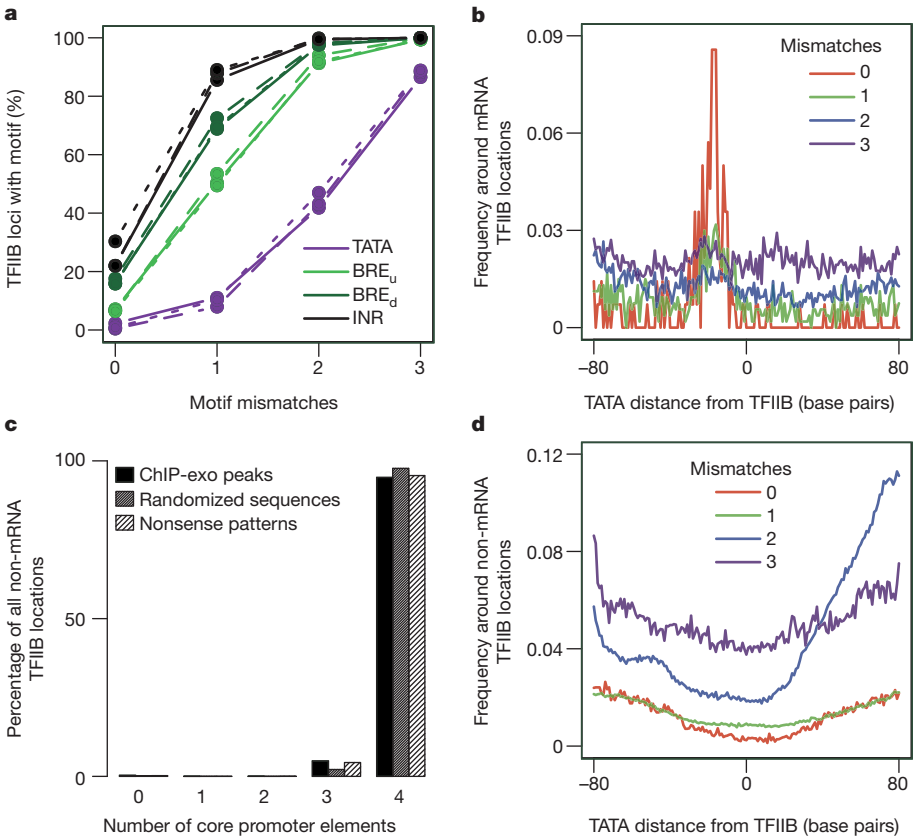

Figure 1 | TATA, INR, BRE_u and BRE_d are not enriched in regions around TFIIB ChIP-exo peaks. **a**, Match frequencies of CPE patterns in regions around mRNA-associated TFIIB peaks (solid lines) coincide with two negative controls (dashed lines). **b**, Unsmoothed positional distributions of matches to TATAWAWR around mRNA-associated ChIP-exo peaks, normalized by the number of sequences with corresponding motif matches. Motifs with two or three mismatches are not enriched. **c**, The fractions of the 150,753 non-mRNA ChIP-exo peak regions with zero to four CPE pattern matches are reproduced by negative controls. **d**, Same as in **b**, but for regions around non-mRNA-associated ChIP-exo peaks.

match closest to its assumed location (as in ref. 5), we could reproduce the artefactual peaks of TFIIB locations around the motif matches. The peaks disappear when all of the motifs are taken into account. We analysed the local enrichment around the mRNA-associated TFIIB peaks. As expected[3,4], TATAWAWR with zero or one mismatches shows enrichment between 30 to 10 nucleotides upstream of TFIIB peaks (Fig. 1b). However, no enrichment is detectable for pattern matches with more than one mismatched position. These are therefore unlikely to be biologically meaningful.

We repeated the previous analyses on the 150,753 non-mRNA-associated TFIIB peaks and again the negative controls closely resembled the true CPE pattern matches. Next, we investigated the claim that the vast majority of regions around the ChIP-exo peaks contain at least three of the four CPEs. The two negative-control procedures explained the observed CPE match frequencies around ChIP-exo peaks (Fig. 1c). We tested the predictive power of the 'core promoter consensus pattern'[5] but found no enrichment around the TFIIB peaks.

Regarding the averaged transcriptional activity and active histone marks reported around non-mRNA-associated TFIIB peaks[5], we note that these could stem from the contributions of relatively few highly expressed non-coding transcripts. Also, TATAWAWR with up to one mismatch is not positionally enriched around non-mRNA-associated peaks (Fig. 1d), even though such enrichment is observed in mRNA-associated peaks (Fig. 1b). We obtained similar results for the 10% strongest non-mRNA-associated ChIP-exo peaks. Finally, 72% of non-mRNA-associated peaks lie within repetitive regions, including 47% Alu repeats, which are transcribed by Pol III and are not expected to bind TFIIB[6]. Hence, the evidence for the biological role of most of the non-mRNA-associated ChIP-exo peaks is inconclusive.

Methods

As negative controls, we permuted the nucleotides of the sequences around TFIIB peaks, ensuring identical nucleotide composition. Second, we generated nonsense patterns from CPE motifs by alphabetically sorting their IUPAC letters (for example, AAARTTWW). In Fig. 1b, we ignored sequences with matches with up to $k-1$ mismatched positions when recording matches for patterns with k mismatched positions, as done in ref. 5. In Fig. 1d, negative positions are upstream (5′) of the ChIP-exo peak on the Watson or Crick strand, and positive positions are downstream (3′). The asymmetry is due to a high proportion of sequence repeats among ChIP-exo peaks.

Matthias Siebert[1] & Johannes Söding[1,2]

[1]Gene Center Munich and Department of Biochemistry, Center for Integrated Protein Science Munich (CIPS[M]), Ludwig-Maximilians-Universität München, Feodor-Lynen-Strasse 25, 81377 Munich, Germany.

[2]Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany.

email: soeding@mpibpc.mpg.de

Received 6 December 2013; accepted 12 June 2014.

1. Ohler, U., Liao, G. C., Niemann, H. & Rubin, G. M. Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol.* **3,** RESEARCH0087 (2002).
2. Kadonaga, J. T. Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip. Rev. Dev. Biol.* **1,** 40–51 (2012).
3. Lenhard, B., Sandelin, A. & Carnici, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Rev. Genet.* **13,** 233–245 (2012).
4. Hartmann, H., Guthöhrlein, E. W., Siebert, M., Luehr, S. & Söding, J. P-value-based regulatory motif discovery using positional weight matrices. *Genome Res.* **23,** 181–194 (2013).
5. Venters, B. J. & Pugh, B. F. Genomic organization of human transcription initiation complexes. *Nature* **502,** 53–58 (2013).
6. Deininger, P. L. & Batzer, M. A. Mammalian retroelements. *Genome Res.* **12,** 1455–1465 (2002).

Author Contributions M.S. performed research and J.S. guided research; both M.S. and J.S. wrote the manuscript.

Competing Financial Interests Declared none.

doi:10.1038/nature13587