Finding the start site: redefining the human initiator element

Jennifer F. Kugel and James A. Goodrich

Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado 80309, USA

Transcription by RNA polymerase II (Pol II) is dictated in part by core promoter elements, which are DNA sequences flanking the transcription start site (TSS) that help direct the proper initiation of transcription. Taking advantage of recent advances in genome-wide sequencing approaches, Vo ngoc and colleagues (pp. 6–11) identified transcripts with focused sites of initiation and found that many were transcribed from promoters containing a new consensus sequence for the human initiator (Inr) core promoter element.

Defining the proper initiation of RNA polymerase II (Pol II) transcription requires a complex interplay of proteins, DNA elements, and RNA that work together to dictate where on the genome transcription begins. This entails the regulated assembly of large multisubunit nucleoprotein complexes containing Pol II and many accessory factors; the platform for forming these large complexes is the core promoter. The core promoter in human genes is the region from −40 to +40 and flanks the transcription start site (TSS) at +1. Although no single core promoter element is contained in all human promoters, many contain one or more of the following core elements (Fig. 1): the TATA box, initiator (Inr), TFIIB recognition elements (BREu and BREd), polypyrimidine initiator (TCT), motif ten element (MTE), and downstream core promoter element (DPE) (for review, see Danino et al. 2015). Of these, the Inr element encompasses the TSS and is thought to be the most common core promoter element, with previous studies estimating that ~50% of human core promoters contain an Inr (Gershenzon and Ioshikhes 2005; Yang et al. 2007). The commonly used consensus sequence for the human Inr, which was derived from mutational analyses, is YYANWYY from −2 to +5 (where, Y = C/T, W = A/T, N = A/C/G/T, and +1 is underlined) (Javahery et al. 1994; Lo and Smale 1996). More recently, analysis of genome-wide CAGE (cap analysis gene expression) data led to the considerably shorter Inr consensus of YR from −1 to +1 (where, R = A/G, and +1 is underlined) (Carninci et al. 2006; Frith et al. 2008). Other studies have also defined somewhat different consensus sequences for the Inr; however, all have an A at +1 in common (for review, see Kadonaga 2012).

Kadonaga and colleagues (Vo ngoc et al. 2017) devised and implemented a novel multistep approach that combines experimental and computational methods to reinvestigate the human Inr consensus sequence. First, they generated two 5′-GRO-seq (5′ end-selected global run-on followed by sequencing) libraries with human MCF-7 cells to identify the 5′ ends of nascent capped transcripts. Second, they developed a peak-calling algorithm named FocusTSS to find transcripts in the 5′-GRO-seq data sets that were initiated at a focused position on the genome, hence identifying clear TSSs to enable analysis of Inr sequences. FocusTSS identified 7678 TSSs that were in both data sets. Third, to identify sequence motifs enriched among the focused TSSs, they used the HOMER motif discovery tool (Heinz et al. 2010), which yielded an Inr-like consensus sequence of BBCABW from −3 to +3 (where, B = C/G/T, W = A/T, and +1 is underlined). Forty percent of the focused TSSs contained a perfect match to the BBCABW consensus Inr. Similar computational analyses performed with data sets from three other human cell lines yielded the same Inr consensus sequence. Interestingly, their analyses also revealed that Inr-containing promoters are less likely to have a TATA box than promoters lacking an Inr and that there is no correlation between the presence of BBCABW Inr elements and CpG islands.

The importance of the sequence at individual positions in the BBCABW consensus Inr sequence was tested using in vitro transcription assays (Vo ngoc et al. 2017). Two native core promoters that each contained a consensus Inr were used, and single-point mutations were made at each position from −3 to +3 that took the sequence away from consensus. The sequences at positions −1 to +3 were the most important for setting levels of basal transcription, with mutations at +1 and +3 showing the largest reductions in transcription levels. In addition, 12 natural

[*Keywords*: RNA polymerase II; initiator; core promoter; transcription start site; focused transcription]

Corresponding authors: james.goodrich@colorado.edu, jennifer.kugel@colorado.edu

Article is online at http://www.genesdev.org/cgi/doi/10.1101/gad.295980.117.

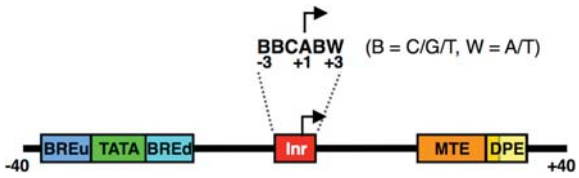

Figure 1. Relative locations of select human core promoter elements and the Inr consensus sequence found in promoters with focused TSSs. The promoter elements depicted include BREu (the upstream TFIIB recognition element), TATA (the TATA box), BREd (the downstream TFIIB recognition element), Inr (new consensus sequence shown), MTE, and DPE.

core promoters were chosen that each differed from consensus at one position; these positions were mutated to create the Inr consensus. Mutating positions −1 to +3 toward consensus increased transcriptional activity, and, again, the mutations at positions +1 and +3 had the greatest effect.

This work provides a substantial step forward in understanding core promoter sequences, establishes a new approach to defining TSSs, and raises many interesting questions that will guide future research. For example, although the Inr is enriched at promoters with focused transcriptional start sites, it is also found randomly distributed throughout the genome. Hence, a consensus Inr alone does not constitute a promoter. The data also showed that promoters with consensus Inr sequences are relatively deficient in TATA boxes. It will be interesting to determine the interplay between other core promoter elements and the Inr at promoters with focused TSSs. Although this work defines a clear correlation between the presence of consensus Inr sequences and focused TSSs, the extent to which the Inr itself causes start sites to be focused remains to be determined. In addition, the role of specific Inr positions in controlling cellular transcription warrants further investigation. For example, C_{-1} and A_{+1} were found most frequently in Inr sequences identified in cells, but mutating C_{-1} away from consensus did not have a strong effect on transcription in vitro. The investigators suggest there is an additional constraint for the use of C_{-1} in cells. Many of the questions raised by this study could be answered by changing the sequences of core promoters in the human genome to determine the effects on the position of the TSS, level of transcription, and occupancy of factors at the core promoter. Finally, this work was limited to the analysis of core promoters with focused TSSs. Although much more complicated, it will be important to extend this new approach to promoters with nonfocused start sites to investigate whether such promoters contain Inr elements. This study illustrates that, despite years of research, much remains to be learned about core promoters and how they set start site positions and levels of transcription at human genes.

References

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engström PG, Frith MC, et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38:** 626–635.

Danino YM, Even D, Ideses D, Juven-Gershon T. 2015. The core promoter: At the heart of gene expression. *Biochim Biophys Acta* **1849:** 1116–1131.

Frith MC, Valen E, Krogh A, Hayashizaki Y, Carninci P, Sandelin A. 2008. A code for transcription initiation in mammalian genomes. *Genome Res* **18:** 1–12.

Gershenzon NI, Ioshikhes IP. 2005. Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics* **21:** 1295–1300.

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

Javahery R, Khachi A, Lo K, Zenzie-Gregory B, Smale ST. 1994. DNA sequence requirements for transcriptional initiator activity in mammalian cells. *Mol Cell Biol* **14:** 116–127.

Kadonaga JT. 2012. Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip Rev Dev Biol* **1:** 40–51.

Lo K, Smale ST. 1996. Generality of a functional initiator consensus sequence. *Gene* **182:** 13–22.

Vo ngoc L, Cassidy CJ, Huang CY, Duttke SHC, Kadonaga JT. 2017. The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters. *Genes Dev* (this issue). doi: 10.1101/gad.293837.116.

Yang C, Bolotin E, Jiang T, Sladek FM, Martinez E. 2007. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene* **389:** 52–65.