The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters

Long Vo ngoc, California Jack Cassidy, Cassidy Yunjing Huang, Sascha H.C. Duttke, and James T. Kadonaga

Section of Molecular Biology, University of California at San Diego, La Jolla, California 92093, USA

DNA sequence signals in the core promoter, such as the initiator (Inr), direct transcription initiation by RNA polymerase II. Here we show that the human Inr has the consensus of BBCA$_{+1}$BW at focused promoters in which transcription initiates at a single site or a narrow cluster of sites. The analysis of 7678 focused transcription start sites revealed 40% with a perfect match to the Inr and 16% with a single mismatch outside of the CA$_{+1}$ core. TATA-like sequences are underrepresented in Inr promoters. This consensus is a key component of the DNA sequence rules that specify transcription initiation in humans.

Supplemental material is available for this article.

Received November 14, 2016; revised version accepted December 19, 2016.

The multifarious signals that lead to the initiation of transcription ultimately converge at the core promoter, which is sometimes referred to as the gateway to transcription (for reviews, see Smale and Kadonaga 2003; Goodrich and Tjian 2010; Kadonaga 2012; Danino et al. 2015). The core promoter is the stretch of DNA—which typically is from about −40 to +40 nucleotides (nt) relative to the +1 transcription start site (TSS)—that directs the initiation of transcription. Core promoters are diverse in terms of their composition and function, and their activities are driven by the presence or absence of DNA sequence motifs such as the TATA box, initiator (Inr), TFIIB recognition elements (BRE^{u} and BRE^{d}), polypyrimidine initiator (TCT), motif ten element (MTE), and downstream core promoter element (DPE). There are no universal core promoter motifs. Specific core promoter elements can be important for enhancer–promoter specificity (for example, see Butler and Kadonaga 2001; Juven-Gershon et al. 2008) as well as the regulation of gene networks (for example, see Juven-Gershon et al. 2008; Parry et al. 2010; Duttke et al. 2014; Wang et al. 2014).

The long-term goal of this study is to gain a more specific understanding of the human core promoter. It has been estimated, for instance, that <25% of human core promoters contain the well-known TATA box or a TATA-like sequence (Gershenzon and Ioshikhes 2005; Carninci et al. 2006; Yang et al. 2007). In fact, it appears that the Inr is the most common core promoter element in humans. For example, ~48%–49% of human promoters were found to have a sequence in the TSS region (from −5 to +6 relative to the +1 TSS) that is related to the 8-nt "cap signal" (i.e., Inr) position-weight matrix (based on 502 eukaryotic promoters) (Bucher 1990; Gershenzon and Ioshikhes 2005). In addition, it has been found that ~46% of human promoters contain the YYA$_{+1}$NWYY Inr consensus within −80 to +80 nt relative to the TSS (Yang et al. 2007). These observations were interesting, but the precise sequence, abundance, and positioning of the human Inr remained to be determined.

The Inr is an extensively studied core promoter element. The presence of a distinct sequence motif that encompasses the TSS was initially described by Corden et al. (1980), and the function of this sequence, which was termed the "initiator," was incisively articulated by Smale and Baltimore (1989). Biochemical studies revealed that the Inr is recognized by the TAF1 and TAF2 subunits of TFIID (Kaufmann and Smale 1994; Purnell et al. 1994; Verrijzer et al. 1995; Chalkley and Verrijzer 1999). The mutational analysis of the human Inr led to the widely used functional Inr consensus of YYA$_{+1}$NWYY (Javahery et al. 1994; Lo and Smale 1996). However, the genome-wide mapping of the 5′ ends of steady-state transcripts by the cap analysis gene expression (CAGE) method yielded the human Inr consensus of YR$_{+1}$ (Carninci et al. 2006; Frith et al. 2008), which is also commonly used. Hence, the nature of the human Inr is unresolved.

We therefore sought to investigate the human Inr consensus. It is important to have the most accurate as possible representation of the Inr consensus for further studies of transcriptional regulation in humans. This is essential for not only the analysis of the Inr itself but also the identification and analysis of other core promoter elements that act in conjunction with the Inr. Recent advances have enabled the genome-wide mapping of the 5′ ends of nascent transcripts and have thus provided the opportunity to obtain new insights into TSSs and core promoters in humans. In this context, we examined the consensus, occurrence, and characteristics of the human Inr at focused promoters in which transcription initiates at a single site or in a narrow cluster of sites.

Results and Discussion

Identification of focused TSSs in human MCF-7 cells with FocusTSS

To investigate the human Inr, we sought to generate a data set of focused TSSs that represent specifically positioned RNA polymerase II transcription preinitiation complexes (PICs). We therefore generated two independent 5′-GRO-seq (5′ end-selected global run on followed by sequencing)

[*Keywords*: RNA polymerase II; initiator; core promoter; transcription start site; focused transcription]
Corresponding author: jkadonaga@ucsd.edu
Article published online ahead of print. Article and publication date are online at http://www.genesdev.org/cgi/doi/10.1101/gad.293837.116.

(Lam et al. 2013) libraries with human MCF-7 breast carcinoma cells. The 5′-GRO-seq method detects the 5′ ends of nascent transcripts and is related to GRO-cap (Kruesi et al. 2013; Core et al. 2014). These methods capture minimally processed nascent transcripts and are thus well suited for the mapping of the 5′ ends of transcripts.

To identify TSSs, we developed a peak-calling algorithm, termed FocusTSS, which is based on the properties of the PIC. After assembly of the PIC at the promoter, the RNA polymerase II can initiate transcription at a single site or in a narrow cluster of sites (see, e.g., Kadonaga 1990). We thus designed FocusTSS to reflect this property of the PIC. As outlined in Figure 1A, it initially identifies peaks that have at least a minimal read count (RC_{min}) and are larger than other peaks in their immediate (±2-nt) vicinity. For each peak, it then determines whether the combined reads in a narrow 5-nt window centered on that peak are at least a minimal proportion (the minimal focus index [FI_{min}]) of the combined reads in a wider 25-nt window that is centered on that peak. The FI reflects the extent to which transcription is focused at a single PIC. Examples of peaks with different FI values are in Supplemental Figure S1.

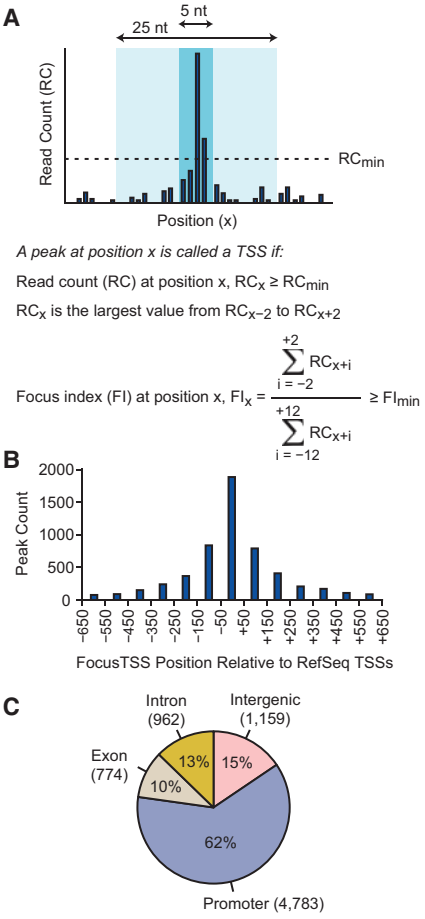

Figure 1. Identification of focused TSSs in 5′-GRO-seq data with FocusTSS. (*A*) The peak-calling scheme in FocusTSS is based on the properties of the transcription PIC. In the PIC, the polymerase is able to initiate transcription in a window of ~5 nt. Thus, FocusTSS selects peaks based on the concentration of reads in a 5-nt window relative to the total reads in a larger 25-nt window. The formula used for peak calling is shown with a visual representation of the parameters. In our data from MCF-7 cells, we typically used a RC_{min} of 20 (approximately one read per million) and a FI_{min} of 0.67. (*B*) FocusTSS peaks are generally close to annotated RefSeq TSSs. FocusTSS peaks (7678) were called with RC_{min} of 20 and FI_{min} of 0.67, and the peak count was calculated for each bin within the indicated range of distances to the closest annotated TSS. (*C*) The majority of FocusTSS peaks is located in promoter regions. The 7678 FocusTSS peaks were classified according to their location in genomic elements. Most TSSs were located near or within annotated genes. Promoters were defined as the region from −1000 nt to +100 nt relative to the closest annotated TSS. The numbers of TSSs in each group are shown in parentheses.

Hence, FocusTSS identifies isolated and focused TSSs that appear to derive from a specifically positioned PIC. For the purposes of this study, which is the analysis of the human Inr sequence, it is useful to have clearly separated and defined TSSs. For other applications, it is possible to vary parameters such as the window sizes, RC_{min}, and FI_{min}.

In our analysis of the human TSS data, we selected FocusTSS peaks with RC_{min} of 20 (approximately one read per million) and FI_{min} of 0.67. With these criteria, the two independent 5′-GRO-seq data sets yielded 7678 shared peaks with similar properties (Supplemental Fig. S2). The 7678 FocusTSS peaks are found mainly near RefSeq-annotated TSSs for protein-coding and noncoding genes (Fig. 1B). Most (75%; 5753 out of 7678) of the FocusTSS peaks are within 1 kb of a RefSeq TSS. In addition, the FocusTSS peaks are predominantly located in promoter regions (from −1000 to +100 relative to the RefSeq TSS) (Fig. 1C). (Because the 5′-GRO-seq method detects nascent transcripts, many of the nonpromoter TSSs may be associated with short-lived species such as enhancer RNAs [eRNAs].) Hence, by the use of 5′-GRO-seq in conjunction with FocusTSS, we generated a data set of thousands of human focused TSSs that could be used for the analysis of core promoters.

A new Inr consensus is frequently used in focused human promoters

To identify overrepresented sequences in the immediate vicinity of the TSS, we analyzed our focused TSS data set with the HOMER motif discovery tool (Heinz et al. 2010). This yielded an Inr-like sequence (motif 1) (the frequency matrix is shown in Supplemental Fig. S3), the TCT motif (motif 2) (Parry et al. 2010; Wang et al. 2014), and two other sequences (Fig. 2A). The Inr-like sequence is the most abundant sequence in the vicinity of the TSS and has the consensus of $BBCA_{+1}BW$ (where B = C/G/T and W = A/T) from −3 to +3 relative to the +1 TSS (Fig. 2B). Given the prevalence of this sequence as well as its resemblance to various versions of the Inr in *Drosophila* and humans (Fig. 2C), it appears that $BBCA_{+1}BW$ is the consensus of the human Inr in focused promoters.

We further tested the range of conditions under which this consensus might be observed. To this end, we found that variation of RC_{min} from 10 to 50 and FI_{min} from 0.50 to 0.75 resulted in $BBCA_{+1}BW$ (Supplemental Fig. S4A). In addition, we performed FocusTSS and HOMER analyses of 5′-GRO-seq or GRO-cap data sets from three other human cell lines (HeLa, GM12878, and K562) and obtained the same $BBCA_{+1}BW$ consensus (Supplemental Fig. S4B). Thus, the $BBCA_{+1}BW$ Inr consensus is widely observed in different conditions and cells.

Out of the 7678 focused TSS peaks in our MCF-7 data set, there are 3071 peaks (40%) with a perfect match to

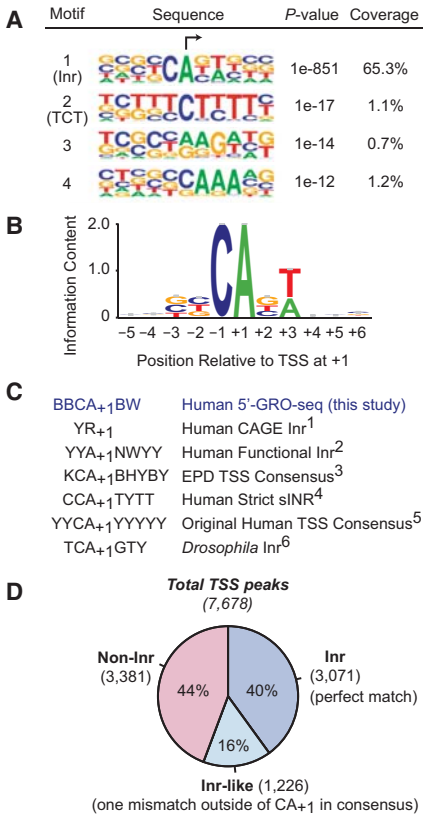

Figure 2. The BBCA$_{+1}$BW consensus for the human initiator (Inr) is present in a majority of focused TSSs. (*A*) The Inr is the most abundant overrepresented sequence near the TSS. Motif discovery analysis of the −5 to +6 region (relative to the +1 TSS) was performed with 7678 focused TSSs in MCF-7 cells. The prevalence (coverage) and *P*-values of the top four sequence motifs are shown. Motif 1 (BBCA$_{+1}$BW, where B = C/G/T and W = A/T) is the Inr, and motif 2 is the TCT motif (Parry et al. 2010). The arrow indicates the position of the TSS. (*B*) Sequence logo of the human Inr at focused TSSs. The sequences of the 3071 FocusTSS peaks with a perfect match to BBCA$_{+1}$BW were used to generate the logo. (*C*) Comparison of the new Inr consensus (BBCA$_{+1}$BW) with some previously described Inr consensus sequences. (1) Human genome-wide CAGE (Carninci et al. 2006; Frith et al. 2008). (2) Functional consensus based on mutational analysis of the human Inr (Javahery et al. 1994; Lo and Smale 1996). (3) Single-nucleotide representation of position-weight matrix of the TSS consensus based on the Eukaryotic Promoter Database (EPD) (Bucher 1990). (4) A rare "strict Inr" in humans (Yarden et al. 2009). (5) The original consensus of the human Inr (Corden et al. 1980). (6) The Inr consensus in *Drosophila* (Ohler et al. 2002; FitzGerald et al. 2006). (*D*) The BBCA$_{+1}$BW Inr occurs frequently in focused promoters in humans. FocusTSS peaks were divided into three groups: perfect match (Inr), one mismatch outside of the central CA$_{+1}$ (Inr-like), and all other sequences (non-Inr). The number of TSSs in each group are shown in parentheses.

the BBCA$_{+1}$BW Inr consensus and 1226 Inr-like peaks (16%) that have only one mismatch outside of the central CA$_{+1}$ in the consensus (Fig. 2D). Hence, the new Inr consensus is frequently observed in human promoters.

Moreover, the BBCABW sequence is strongly enriched at the +1 position of the FocusTSS peaks and is otherwise distributed randomly (Supplemental Fig. S5). This is consistent with the model that the Inr does not usually function by itself but rather acts in conjunction with other sequence motifs to give a fully active core promoter.

We wondered whether the TATA box or TATA-like sequences are enriched or depleted in promoters with Inr or Inr-like motifs. To address this question, we examined the frequency of occurrence of either a consensus TATA box (TATAAR, as identified by HOMER, from −33 to −23 relative to the +1 TSS) or a degenerate TATA-like sequence (WWWW from −33 to −23 relative to the +1 TSS, where W = A/T) in the Inr, Inr-like, or non-Inr promoters shown in Figure 2D. This analysis revealed that both consensus and degenerate TATA sequences were less common in Inr and Inr-like promoters than in non-Inr promoters (Supplemental Fig. S6A). For instance, the degenerate TATA-like sequence was observed in ~21% and 23% of the Inr and Inr-like promoters, respectively, relative to ~35% in non-Inr promoters (Supplemental Fig. S6A). It is possible that promoters with an Inr are less dependent on a TATA box and vice versa.

We also examined whether the consensus BBCA$_{+1}$BW Inr is preferentially found within CpG islands. Approximately 60% of focused TSSs are found in CpG islands, but there is no apparent enrichment or depletion of BBCA$_{+1}$BW Inr TSSs or Inr-like TSSs in CpG islands (Supplemental Fig. S6B). In contrast, focused TSSs that are associated with TATA-like sequences are depleted in CpG islands (Supplemental Fig. S6B).

As seen in Figure 2C, the new BBCA$_{+1}$BW Inr consensus is distinct from other versions of the human Inr. The widely used functional Inr consensus (YYA$_{+1}$NWYY) (Javahery et al. 1994; Lo and Smale 1996) was based on the mutational analysis of the Inr. Another commonly used version of the human Inr (YR$_{+1}$) was obtained from genome-wide CAGE data (Carninci et al. 2006; Frith et al. 2008). The differences between the YR$_{+1}$ consensus and the BBCA$_{+1}$BW consensus may be due in part to the analysis of steady-state transcripts in the CAGE experiments and nascent transcripts in the 5′-GRO-seq and GRO-cap experiments. Another potential factor is the use of FocusTSS to identify focused start sites. Notably, we observed that TSSs with higher FI values are enriched for the BBCA$_{+1}$BW Inr relative to TSSs with lower FI values (Supplemental Fig. S7A,B). Likewise, promoters with a perfect match to the BBCA$_{+1}$BW Inr have higher FI values than promoters that do not contain a perfect match to the motif (Supplemental Fig. S7C). Thus, the selection of focused TSSs with FocusTSS enriches for promoters with the BBCA$_{+1}$BW Inr motif.

Variants of the degenerate BBCA$_{+1}$BW hexanucleotide at focused TSSs

We next considered the possibility that some of the 54 variants of the BBCA$_{+1}$BW consensus are overrepresented or underrepresented at promoters. To address this issue, we determined the frequency of occurrence of each of the 4096 possible hexanucleotide sequences from −3 to +3 (relative to the +1 TSS) in our data set of 7678 TSSs. This revealed that 46 of the 51 most abundant hexanucleotides are a perfect match to the BBCA$_{+1}$BW consensus (Fig. 3A; Supplemental Fig. S8). Notably, there is not a specific subset of variants that is highly overrepresented. However, there is some underrepresentation of BBCA$_{+1}$TA and TGCA$_{+1}$BW sequences (Supplemental Fig. S8). Thus, nearly all of the 54 variants of the BBCA$_{+1}$BW Inr are among the most commonly used hexanucleotides at focused TSSs.

For comparison, we carried out the same analysis with the 32 variants of the functional Inr consensus

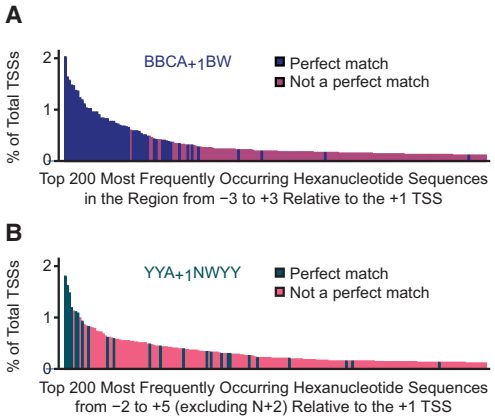

Figure 3. The $BBCA_{+1}BW$ Inr consensus generally represents the most frequently occurring sequences at the TSS. (*A*) Distribution of $BBCA_{+1}BW$ Inr sequences among the most frequently occurring hexanucleotides in the region from −3 to +3 relative to the +1 TSS. The plot shows the frequencies (percentage of total TSSs) of the top 200 (out of a possible 4096) occurring hexanucleotide sequences that are either a perfect match (blue) or not a perfect match (red) to the $BBCA_{+1}BW$ Inr consensus. All 54 versions of $BBCA_{+1}BW$ are in the top 200 sequences. The less frequently occurring perfect match outliers (≤0.2% frequency) are variants with the $BBCA_{+1}TA$ sequence (Supplemental Fig. S8). (*B*) The $YYA_{+1}NWYY$ functional Inr consensus is somewhat broadly distributed among the most commonly occurring hexanucleotide sequences from −2 to +5 (excluding the random N_{+2}). The plot includes 24 out of the 32 variants of the $YYA_{+1}NWYY$ consensus.

($YYA_{+1}NWYY$), which has six nonrandom positions from −2 to +5 relative to the +1 TSS (Fig. 3B; Supplemental Fig. S9). This revealed that eight of the 12 most common sequences are a match to the functional Inr; however, the other 24 variants of this consensus are not concentrated among the most commonly occurring sequences. Therefore, although the functional Inr consensus, which was elucidated >20 years ago, is an excellent representation of the Inr, the emergence of new technologies has now allowed the determination of the $BBCA_{+1}BW$ Inr, which is strongly represented at the genome-wide level among the most commonly occurring focused TSSs.

Functional analysis of the $BBCA_{+1}BW$ sequences in the basal transcription process

Next, we investigated the function of the $BBCA_{+1}BW$ Inr by in vitro transcription analysis of human core promoters in their natural context from −50 to +51 relative to the +1 TSS. In the first set of experiments, we examined the *PMAIP1* and *TFRC* promoters, both of which contain a consensus $BBCA_{+1}BW$ Inr. We tested a series of single-nucleotide substitution mutations for each position from −5 to +5. Outside of the Inr (positions −5, −4, +4, and +5), we used transition mutations, whereas inside the Inr, we mutated the nucleotides to nonconsensus bases (Fig. 4A; Supplemental Fig. S10).

These studies indicated that the sequences from −1 to +3, particularly the +1 and +3 positions, are important for core promoter activity. Moreover, we observed that CA_{+1}, as in the $BBCA_{+1}BW$ consensus, mediates higher levels of transcription than CG_{+1} or TA_{+1}, which match the YR_{+1} consensus. In addition, B_{-3} and B_{-2} (where B = not A) appear to contribute to focused initiation at A_{+1}, as we observed increased levels of transcription initiation at −3 and −2 when those positions are mutated to A (Fig. 4A; Supplemental Figs. S10, S11). Hence, single-nucleotide mutations that disrupt the $BBCA_{+1}BW$ consensus result in a reduction or an alteration of the activity of the core promoter. In contrast, mutations outside of the $BBCA_{+1}BW$ Inr consensus had little effect on core promoter function (Fig. 4A; Supplemental Fig. S10).

We additionally tested the effect of mutation of nonconsensus Inr sequences to the consensus sequence. To this end, we selected 12 naturally occurring core promoters that contain a single mismatch to the $BBCA_{+1}BW$ consensus at positions ranging from −3 to +3 and then generated single-nucleotide substitutions that convert the nonconsensus sequences to the $BBCA_{+1}BW$ Inr consensus (Fig. 4B). These experiments revealed that conversion of the nonconsensus sequences to the Inr consensus generally led to an increase in transcriptional activity, with the largest effects observed at the +1 and +3 positions.

Altogether, the mutational analyses indicate that transcription initiates optimally from the $BBCA_{+1}BW$ Inr consensus and that the region from −1 to +3 is most important for the efficiency of transcription. These results reflect the nucleotide distributions that were observed in the Inr region (Fig. 2; Supplemental Fig. 3) and are consistent with the findings of Smale and colleagues (Javahery et al. 1994; Lo and Smale 1996) in their analysis of the functional Inr consensus. In some promoter contexts, the lack of an A nucleotide at positions −2 and −3 appears to suppress transcription initiation at those sites and thus support more focused transcription from the +1 TSS. It is also notable that CA_{+1} more specifically reflects the active Inr element than the more general YR_{+1} consensus.

It can further be seen that C_{-1} and A_{+1} are more prominent in the Inr consensus than W_{+3}, whereas A_{+1} and W_{+3} are more important for transcriptional activity than C_{-1}. In addition, all of the 40 most frequently occurring hexanucleotides at the Inr region include C_{-1}, A_{+1}, and W_{+3} (Supplemental Fig. S8). These findings collectively suggest that there is an additional constraint for the use of C_{-1} that extends beyond its role in contributing to promoter strength. As an example, such a constraint might be the need to avoid inadvertent binding by a sequence-specific factor with a related and/or overlapping recognition sequence.

The human Inr, a distinct and abundant element that is precisely positioned at focused TSSs

In this study, we identified and characterized the $BBCA_{+1}BW$ Inr consensus sequence, which is positioned precisely at more than half of focused human TSSs (Fig. 2). Of the 54 variants of this consensus, none are highly overrepresented; there is, however, some underrepresentation of $BBCA_{+1}TA$ and $TGCA_{+1}BW$ sequences (Fig. 3; Supplemental Fig. S8). Moreover, the TATA box and TATA-like sequences are less common in $BBCA_{+1}BW$ Inr and Inr-like promoters than in non-Inr promoters (Supplemental Fig. S6).

The articulation of the Inr element is essential for the understanding of the mechanisms of transcription in humans. This new consensus can now be used as a foundation for the analysis of the other sequences and associated factors that regulate gene activity.

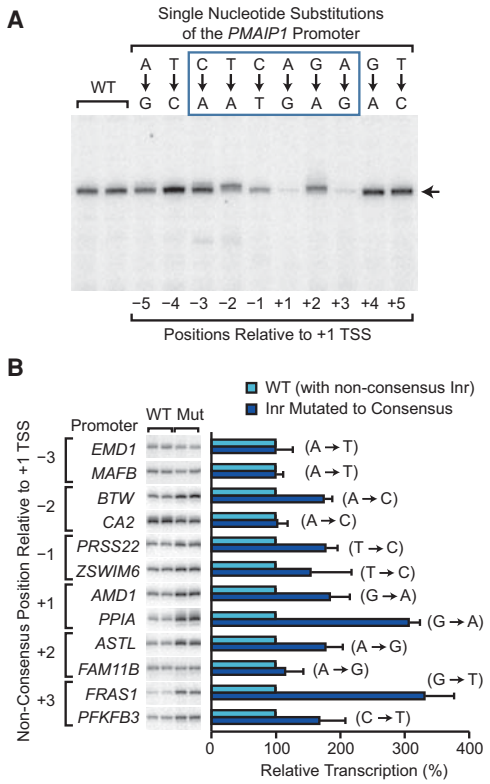

Figure 4. The BBCA$_{+1}$BW Inr sequence is essential for efficient and accurate transcription initiation. The core promoter regions from −50 to +51 relative to the +1 TSS (for DNA sequences, see Supplemental Fig. S13) of the indicated human genes were used in these experiments. (*A*) Alterations in the Inr sequence impair transcription strength and start site selection. The consensus Inr sequence in the *PMAIP1* promoter was mutated by using the indicated single-nucleotide substitutions. The wild-type (WT) and mutant constructs were subjected to in vitro transcription and primer extension analysis. Mutations in the BBCA$_{+1}$BW Inr are inside the blue box. The horizontal arrow indicates the +1 TSS. Quantitation of the transcription levels from at least four independent experiments is shown in Supplemental Figure S10A. (*B*) Mutation of nonconsensus Inr sequences to the consensus generally results in higher levels of transcription. Nonconsensus wild-type promoters (WT) or mutant promoters that were altered at a single nucleotide to match the consensus (Mut) were subjected to in vitro transcription analysis. The single-nucleotide substitutions are indicated in parentheses. The autoradiograms show representative results, and the quantitative data from three or more experiments are shown as the mean (relative to wild type) ± SD.

Importantly, it should be noted that this study has been restricted to the analysis of focused TSSs, which have a clearly isolated site (or narrow 5-nt region) at which transcription initiates. The analysis of focused TSSs has minimized ambiguity with regard to the sequences that direct transcription and has thus facilitated the elucidation of the Inr consensus. In addition, our MCF-7 data set yielded 7678 focused TSSs (Figs. 1, 2), which represent thousands of protein-coding genes and noncoding transcripts. Nevertheless, our analysis of focused promoters does exclude nonfocused promoters (also known as dispersed or broad promoters). Some nonfocused promoters may be tandemly arranged focused core promoters, whereas others may direct dispersed transcription by an entirely different mechanism.

At a practical level, we also considered the merits of a slightly simplified BCA$_{+1}$BW Inr consensus. The exclusion of B$_{-3}$ from the consensus was considered because the B$_{-3}$ position exhibits the lowest amount of sequence conservation relative to the other positions (e.g., see Supplemental Fig. S4A), and mutation of B$_{-3}$ has little effect on the overall strength of transcription (Fig. 4; Supplemental Fig. S10). We therefore carried out an analysis of the BCA$_{+1}$BW sequence (Supplemental Fig. S12). This revealed that 45% of TSSs contain a perfect match to BCA$_{+1}$BW and that an additional 13% of TSSs contain only a single mismatch to BCA$_{+1}$BW outside of the central CA$_{+1}$ dinucleotide. Moreover, the 18 variants of the BCA$_{+1}$BW sequence include the 17 most frequently occurring pentanucleotide sequences at focused TSSs, and the overrepresentation of pentanucleotides that perfectly match BCA$_{+1}$BW is striking (Supplemental Fig. S12C,D). Thus, the simplified BCA$_{+1}$BW sequence is an excellent version of the human Inr.

In conclusion, the BBCA$_{+1}$BW Inr and Inr-like sequences (with only one mismatch outside of CA$_{+1}$) are found at precisely the same location in more than half of focused human TSSs (Fig. 2D; Supplemental Figs. 10D, 11B) and are much more abundant than the TATA box or TATA-like sequences (Supplemental Fig. S6). This revised Inr consensus should serve as a useful and reliable beacon for the study of transcription in humans.

Materials and methods

5′-GRO-seq

Two 5′-GRO-seq experiments were carried out with MCF-7 cells essentially as described in Duttke et al. (2015) and Hetzel et al. (2016). The detailed procedure is provided in the Supplemental Material. The 5′-GRO-seq data are available from Gene Expression Omnibus (GEO; accession number, GSE90035).

FocusTSS

FocusTSS is a Python program (Focus_TSS.py) and is available in the Supplemental Material. The design and use of FocusTSS is described in Figure 1A as well as in the Supplemental Material.

In vitro transcription assays

The plasmids used in the in vitro transcription assays were constructed by insertion of core promoter sequences (−50 to +51 relative to the TSS) in the XbaI and PstI sites of the pUC119T vector. Transcription reactions were performed essentially as described previously (Theisen et al. 2013). The specific reaction conditions are indicated in the Supplemental Material. All in vitro transcription experiments were performed independently at least three times to ensure reproducibility of the data.

Acknowledgments

We thank E. Peter Geiduschek, George Kassavetis, and Yuan-Liang Wang for critical reading of the manuscript, and Chris Benner for advice on the computational analysis. Contributions by Thomas Boulay, Scott Iwashita, and Timothy Bretz to earlier functional studies of the human Inr are also acknowledged. This work was supported by National Institutes of Health grants R01 GM041249, R21 HG008781, and R35 GM118060 to J.T.K.

References

Bucher P. 1990. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol* **212:** 563–578.

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

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

Chalkley GE, Verrijzer CP. 1999. DNA binding site selection by RNA polymerase II TAFs: a TAF(II)250–TAF(II)150 complex recognizes the initiator. *EMBO J* **18:** 4835–4485.

Corden J, Wasylyk B, Buchwalder A, Sassone-Corsi P, Kedinger C, Chambon P. 1980. Promoter sequences of eukaryotic protein-coding genes. *Science* **209:** 1406–1414.

Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. 2014. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet* **46:** 1311–1320.

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.

Duttke SHC, Doolittle RF, Wang YL, Kadonaga JT. 2014. TRF2 and the evolution of the bilateria. *Genes Dev.* **28:** 2071–2076.

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

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

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.

Goodrich JA, Tjian R. 2010. Unexpected roles for core promoter recognition factors in cell-type-specific transcription and gene regulation. *Nat Rev Genet* **11:** 549–558.

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.

Hetzel J, Duttke SH, Benner C, Chory J. 2016. Nascent RNA sequencing reveals distinct features in plant transcription. *Proc Natl Acad Sci* **113:** 12316–12321.

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.

Juven-Gershon T, Hsu JY, Kadonaga JT. 2008. Caudal, a key developmental regulator, is a DPE-specific transcription factor. *Genes Dev* **22:** 2823–2830.

Kadonaga JT. 1990. Assembly and disassembly of the *Drosophila* RNA polymerase II complex during transcription. *J Biol Chem* **265:** 2624–2631.

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

Kaufmann J, Smale ST. 1994. Direct recognition of initiator elements by a component of the transcription factor IID complex. *Genes Dev* **8:** 821–829.

Kruesi WS, Core LJ, Waters CT, Lis JT, Meyer BJ. 2013. Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation. *Elife* **2:** e00808.

Lam MTY, Cho H, Lesch HP, Gosselin D, Heinz S, Tanaka-Oishi Y, Benner C, Kaikkonen MU, Kim AS, Kosaka M, et al. 2013. Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. *Nature* **498:** 511–515.

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

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

Parry TJ, Theisen JWM, Hsu JY, Wang YL, Corcoran DL, Eustice M, Ohler U, Kadonaga JT. 2010. The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes Dev* **24:** 2013–2018.

Purnell BA, Emanuel PA, Gilmour DS. 1994. TFIID sequence recognition of the initiator and sequences farther downstream in *Drosophila* class II genes. *Genes Dev* **8:** 830–842.

Smale ST, Baltimore D. 1989. The 'initiator' as a transcription control element. *Cell* **57:** 103–113.

Smale ST, Kadonaga JT. 2003. The RNA polymerase II core promoter. *Annu Rev Biochem* **72:** 449–479.

Theisen JWM, Gucwa JS, Yusufzai T, Khuong MT, Kadonaga JT. 2013. Biochemical analysis of histone deacetylase-independent transcriptional repression by MeCP2. *J Biol Chem* **288:** 7096–7104.

Verrijzer CP, Chen JL, Yokomori K, Tjian R. 1995. Binding of TAFs to core elements directs promoter selectivity by RNA polymerase II. *Cell* **81:** 1115–1125.

Wang YL, Duttke SHC, Chen K, Johnston J, Kassavetis GA, Zeitlinger J, Kadonaga JT. 2014. TRF2, but not TBP, mediates the transcription of ribosomal protein genes. *Genes Dev* **28:** 1550–1555.

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.

Yarden G, Elfakess R, Gazit K, Dikstein R. 2009. Characterization of sINR, a strict version of the Initiator core promoter element. *Nucleic Acids Res* **37:** 4234–4246.