

RESEARCH COMMUNICATION

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 BB_{CA}₊₁BW 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^a and BRE^b), 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 promot-

ers 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₊₁NWYY 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₊₁NWYY (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>.

© 2017 Vo ngoc et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <http://genesdev.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

(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

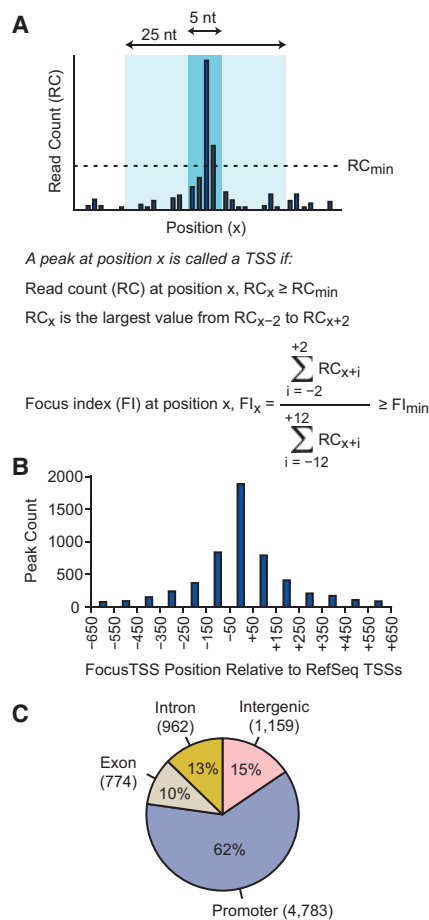


Figure 1. Identification of focused TSSs in 5'-GRO-seq data with FocusTSS. (A) The peak-calling scheme in FocusTSS is based on the properties of the transcription PIC. In the PIC, the polymerase is able to initiate transcription in a window of ~5 nt. Thus, FocusTSS selects peaks based on the concentration 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.

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.

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

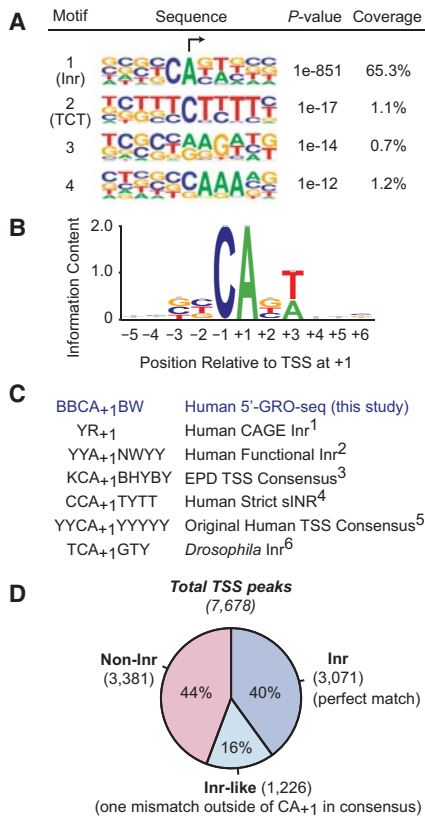


Figure 2. The BBCA₊₁BW consensus for the human initiator (Inr) is present in a majority of focused TSSs. (A) The Inr is the most abundant overrepresented sequence near the TSS. Motif discovery analysis of the -5 to +6 region [relative to the +1 TSS] was performed with 7678 focused TSSs in MCF-7 cells. The prevalence (coverage) and *P*-values of the top four sequence motifs are shown. Motif 1 (BBCA₊₁BW, 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₊₁BW were used to generate the logo. (C) Comparison of the new Inr consensus (BBCA₊₁BW) 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₊₁BW 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₊₁BW 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₊₁BW 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₊₁BW 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₊₁BW Inr consensus is distinct from other versions of the human Inr. The widely used functional Inr consensus (YYA₊₁NWYY) (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₊₁BW 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₊₁BW Inr relative to TSSs with lower FI values (Supplemental Fig. S7A,B). Likewise, promoters with a perfect match to the BBCA₊₁BW 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₊₁BW Inr motif.

Variants of the degenerate BBCA₊₁BW hexanucleotide at focused TSSs

We next considered the possibility that some of the 54 variants of the BBCA₊₁BW 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₊₁BW 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₊₁TA and TGCA₊₁BW sequences (Supplemental Fig. S8). Thus, nearly all of the 54 variants of the BBCA₊₁BW 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

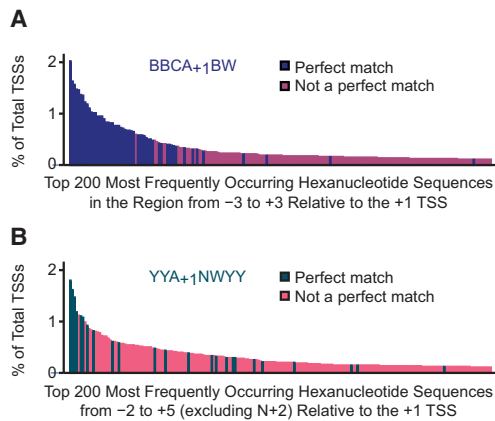


Figure 3. The BBCA₊₁BW Inr consensus generally represents the most frequently occurring sequences at the TSS. (A) Distribution of BBCA₊₁BW Inr sequences among the most frequently occurring hexanucleotides in the region from -3 to +3 relative to the +1 TSS. The plot shows the frequencies (percentage of 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₊₁BW Inr consensus. All 54 versions of BBCA₊₁BW are in the top 200 sequences. The less frequently occurring perfect match outliers ($\leq 0.2\%$ frequency) are variants with the BBCA₊₁TA sequence (Supplemental Fig. S8). (B) The YYA₊₁NWYY 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₊₁NWYY consensus.

(YYA₊₁NWYY), 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₊₁BW Inr, which is strongly represented at the genome-wide level among the most commonly occurring focused TSSs.

Functional analysis of the BBCA₊₁BW sequences in the basal transcription process

Next, we investigated the function of the BBCA₊₁BW 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₊₁BW 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₊₁BW 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₊₁BW consensus result in a reduction or an alteration of the activity of the core promoter. In contrast, mutations outside of the BBCA₊₁BW 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₊₁BW consensus at positions ranging from -3 to +3 and then generated single-nucleotide substitutions that convert the nonconsensus sequences to the BBCA₊₁BW 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₊₁BW 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₊₁BW 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₊₁TA and TGCA₊₁BW sequences (Fig. 3; Supplemental Fig. S8). Moreover, the TATA box and TATA-like sequences are less common in BBCA₊₁BW 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.

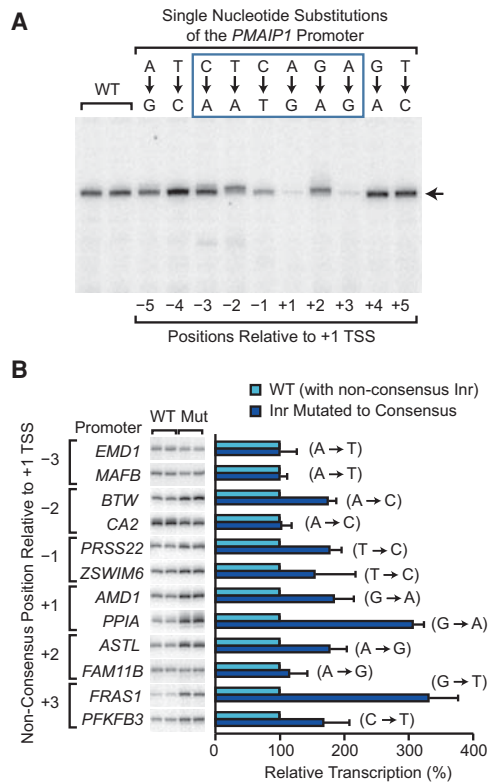


Figure 4. The BCCA₊₁BW Inr sequence is essential for efficient and accurate transcription initiation. The core promoter regions from -50 to +51 relative to the +1 TSS (for DNA sequences, see Supplemental Fig. S13) of the indicated human genes were used in these experiments. (A) Alterations in the Inr 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 BCCA₊₁BW 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₊₁BW Inr consensus. The exclu-

sion 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₊₁BW sequence (Supplemental Fig. S12). This revealed that 45% of TSSs contain a perfect match to BCA₊₁BW and that an additional 13% of TSSs contain only a single mismatch to BCA₊₁BW outside of the central CA₊₁ dinucleotide. Moreover, the 18 variants of the BCA₊₁BW sequence include the 17 most frequently occurring pentanucleotide sequences at focused TSSs, and the overrepresentation of pentanucleotides that perfectly match BCA₊₁BW is striking (Supplemental Fig. S12C,D). Thus, the simplified BCA₊₁BW sequence is an excellent version of the human Inr.

In conclusion, the BCCA₊₁BW 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

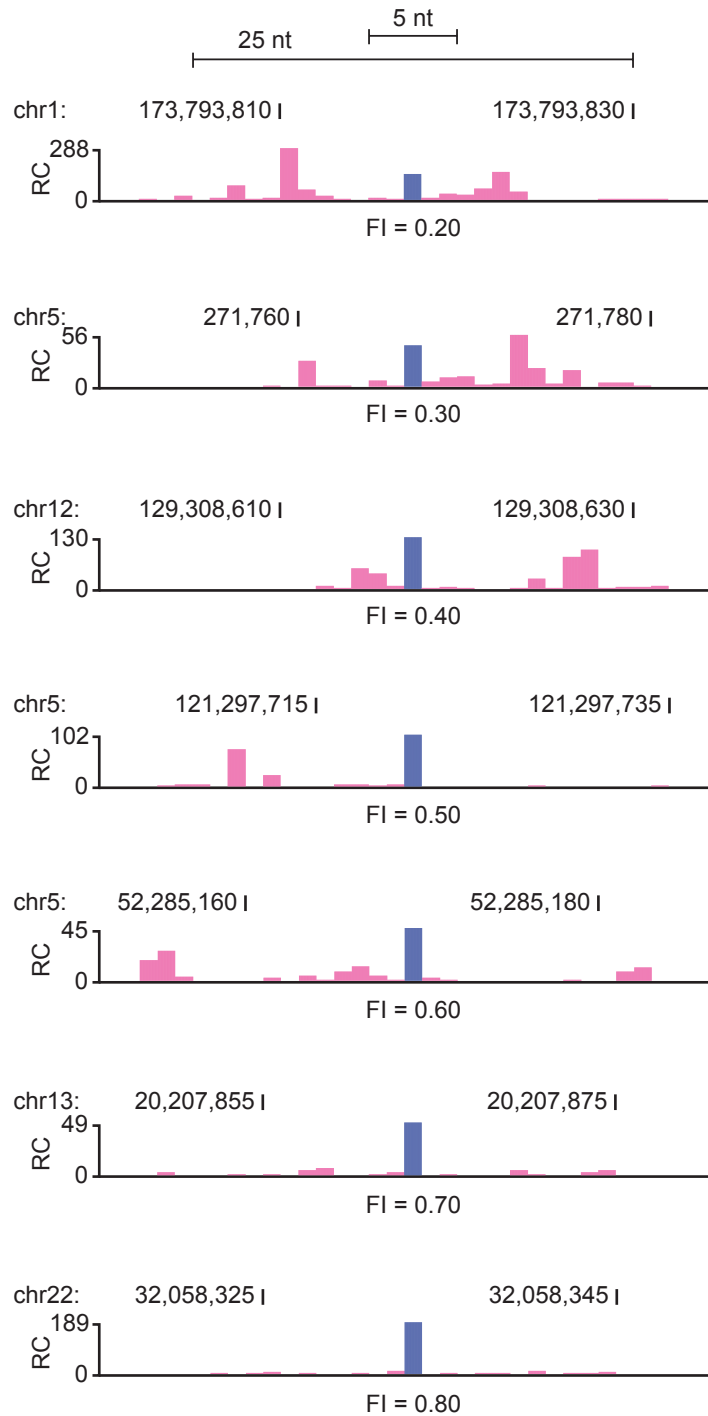
Bucher P. 1990. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol* 212: 563–578.

- 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, Wasyluk 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.
- Gershenson 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.
- Hetzl 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, Zenie-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.

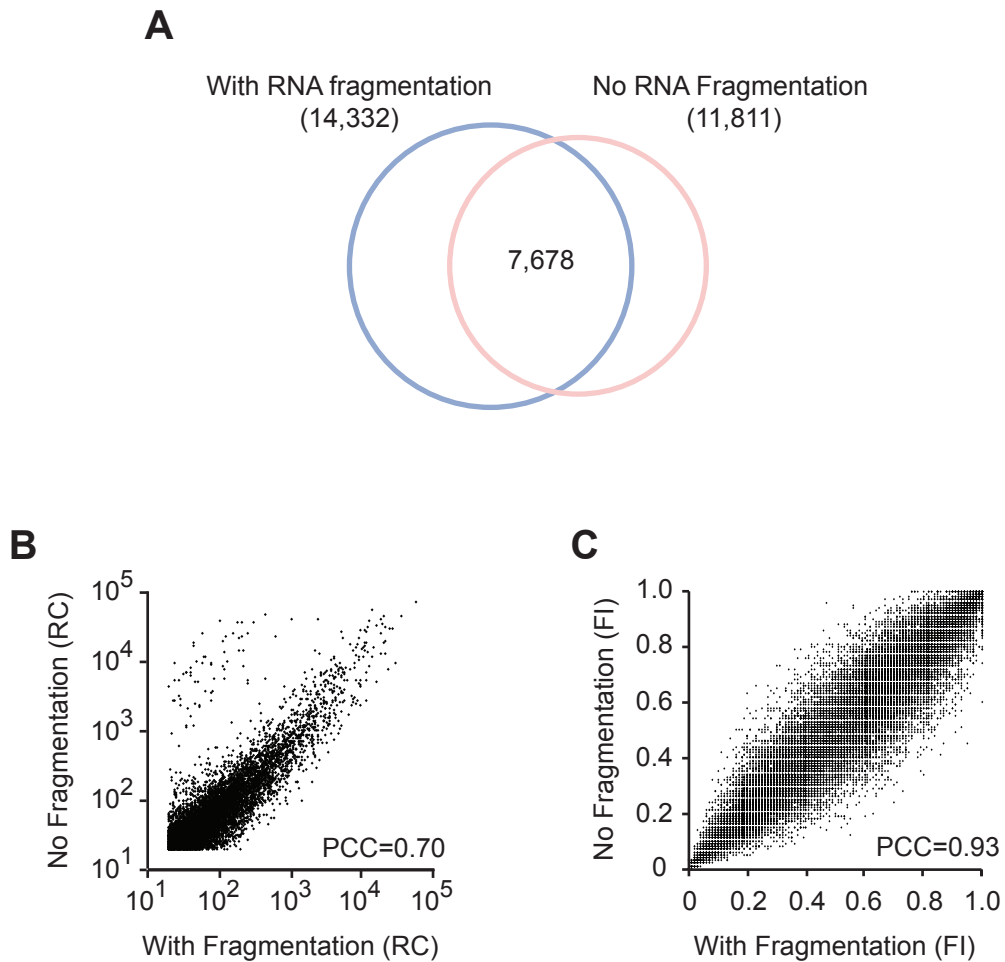
SUPPLEMENTAL MATERIAL

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



Supplemental Figure S1. Examples of TSS peaks with different values of the focus index (FI). FocusTSS was used to identify TSS peaks with FI values that range from 0.2 to 0.8 ($RC_{\min} = 20$). The dark blue bars represent the peaks called by FocusTSS with the indicated FI values. The hg19 coordinates are also shown.








Supplemental Figure S2. FocusTSS peaks are related whether or not a Zn(II)-mediated fragmentation step is included in the preparation of the RNA for 5'-GRO-seq analysis. FocusTSS peaks were called with $RC_{\min} = 20$ (approximately 1 read per million) and $FI_{\min} = 0.67$ in (A) and (B) and with $RC_{\min} = 20$ and $FI_{\min} = 0$ in (C). PCC (Pearson correlation coefficient) values were calculated based on peaks that were present in both datasets. (A) There are many shared TSS peaks between the two TSS datasets that were generated either with or without the Zn(II)-catalyzed RNA fragmentation step. The TSS peaks for each dataset and the number of shared TSSs are shown. (B) Scatter plot of the correlation between the read count (RC) values between the two experiments. (C) Scatter plot of the correlation between the focus index (FI) values between the two experiments.

Positions Relative to the +1 TSS





	-3	-2	-1	+1	+2	+3
A	0.07	0.05	0.00	1.00	0.02	0.43
C	0.36	0.35	0.98	0.00	0.28	0.06
G	0.37	0.27	0.00	0.00	0.39	0.00
T	0.21	0.33	0.02	0.00	0.31	0.50
	G	C	C	A	G	T
	C	T			T	A
	T	G			C	

Supplemental Figure S3. Frequency matrix for the human Inr consensus in Fig. 2A. This table shows the base frequencies of the Inr motif identified with HOMER (Heinz et al. 2010) at each position in the -3 to +3 region relative to the +1 TSS.

A

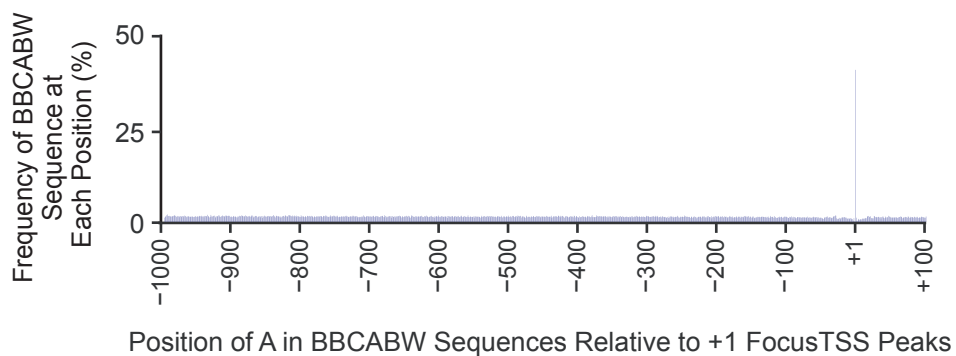
RC _{min}	FI _{min}	Motif	P-value	Coverage
20	0.67		1e-851	65.3%
20	0.50		1e-1345	62.1%
20	0.75		1e-652	60.3%
10	0.67		1e-1218	62.4%
50	0.67		1e-385	66.2%

B

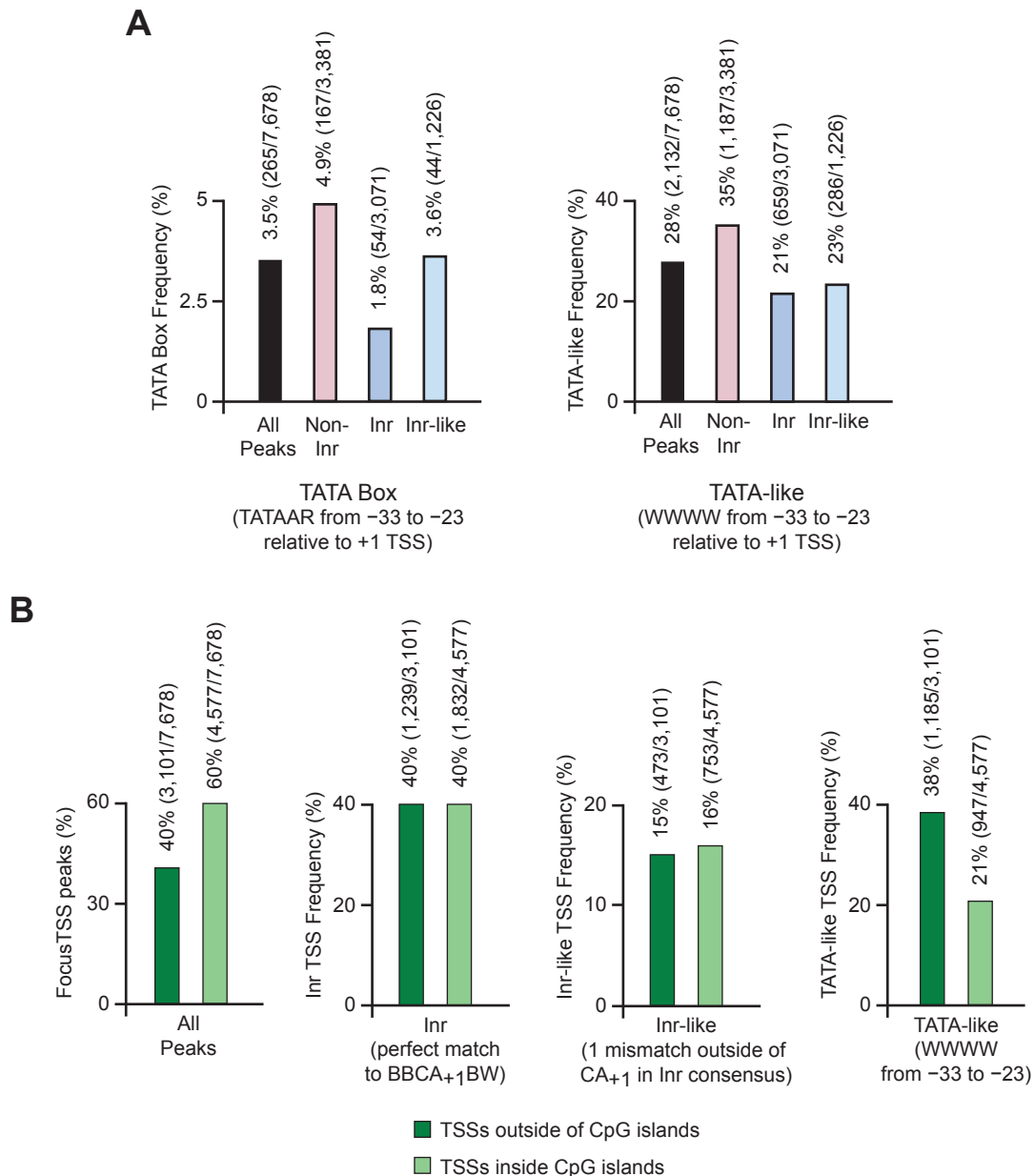
Cell line	Motif	P-value	Coverage
MCF-7		1e-851	65.3%
GM12878 [‡]		1e-2860	49.0%
K562 [‡]		1e-2697	51.2%
HeLa S3 [§]		1e-386	46.6%

[‡]Core et al. (2014)
[§]Duttke et al. (2015)

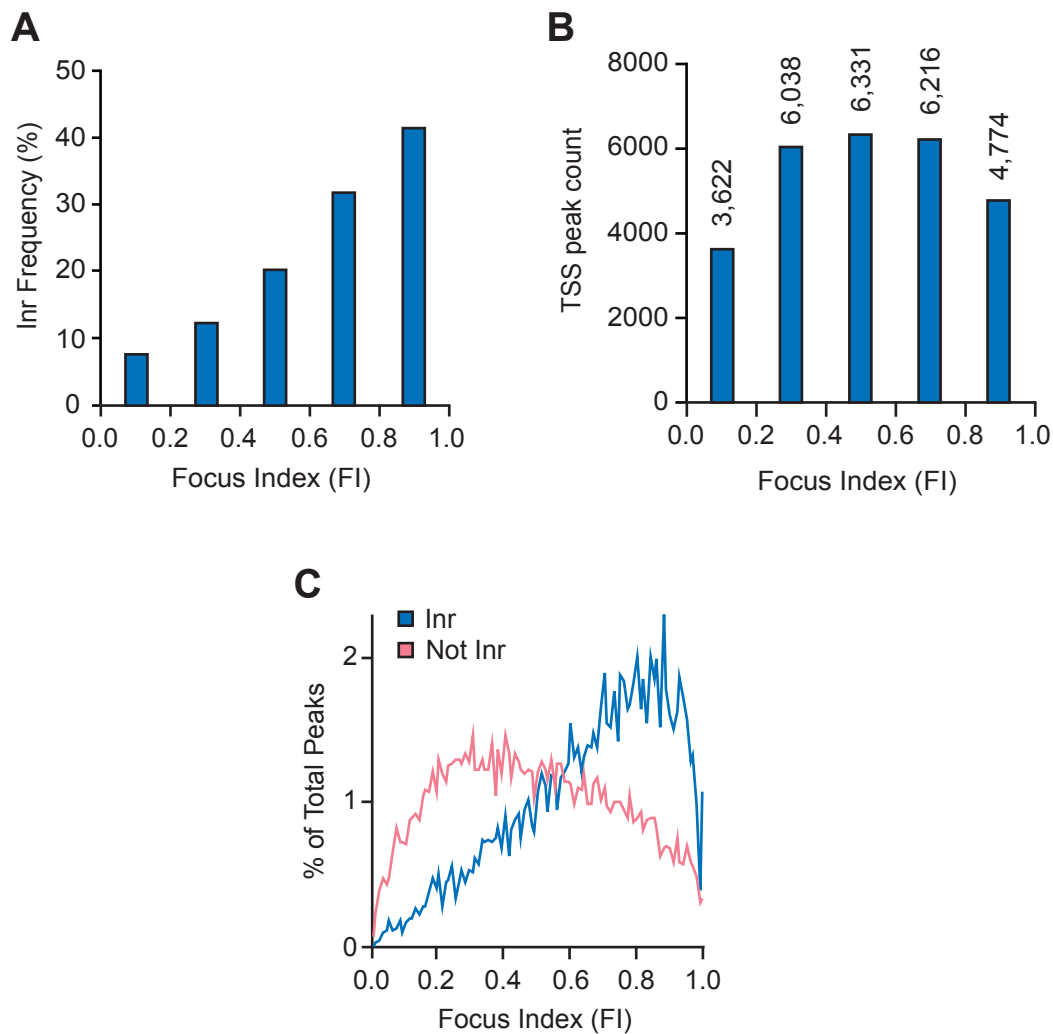
Supplemental Figure S4. The human Inr consensus remains largely the same under different conditions. The sequence motifs were obtained by analyzing FocusTSS peaks with HOMER (Heinz et al. 2010). (A) Inr consensus motifs from FocusTSS peaks obtained with different values of RC_{min} and FI_{min}. (B) Inr consensus motifs with 5'-GRO-seq or GRO-cap data from different human cell lines with RC_{min} = 20 and FI_{min} = 0.67.



Supplemental Figure S5. Distribution of the BBCABW sequence relative to +1 FocusTSS peaks. For each nucleotide from -1000 to $+100$ relative to the $+1$ FocusTSS peaks, the frequency of occurrence of the BBCABW sequence motif is shown. The BBCABW motifs were positioned by the location of the A nucleotide in the motif. This plot shows a major peak at $+1$ and a random distribution of BBCABW motifs at other positions. 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



Supplemental Figure S6. Analysis of the BBCA+1BW Inr, the TATA box, and CpG islands. **(A)** The TATA box and TATA-like sequences are less common in Inr and Inr-like promoters than in Non-Inr promoters. The frequency of occurrence of a consensus TATA box [TATAAR, identified by using HOMER (Heinz et al. 2010), 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) is shown for Inr, Inr-like, and Non-Inr promoters, as categorized in Fig. 2D. The percentages and absolute numbers of TSSs (in parentheses) are indicated. **(B)** TSSs with the BBCA+1BW Inr are not enriched in CpG islands. The left panel shows the percentages of focused TSSs that are located inside versus outside CpG islands in the human hg19 track of the UCSC Genome Browser (<https://genome.ucsc.edu>; Wu et al. 2010). The middle-left panel shows the percentages of consensus Inr TSSs inside versus outside CpG islands. The middle-right panel shows the percentages of Inr-like TSSs (with one mismatch outside of CA₊₁ in the BBCA+1BW Inr consensus) inside versus outside of CpG islands. The right panel shows the percentages of TATA-like-associated TSSs inside versus outside CpG islands. The percentages and absolute numbers of TSSs (in parentheses) are indicated.



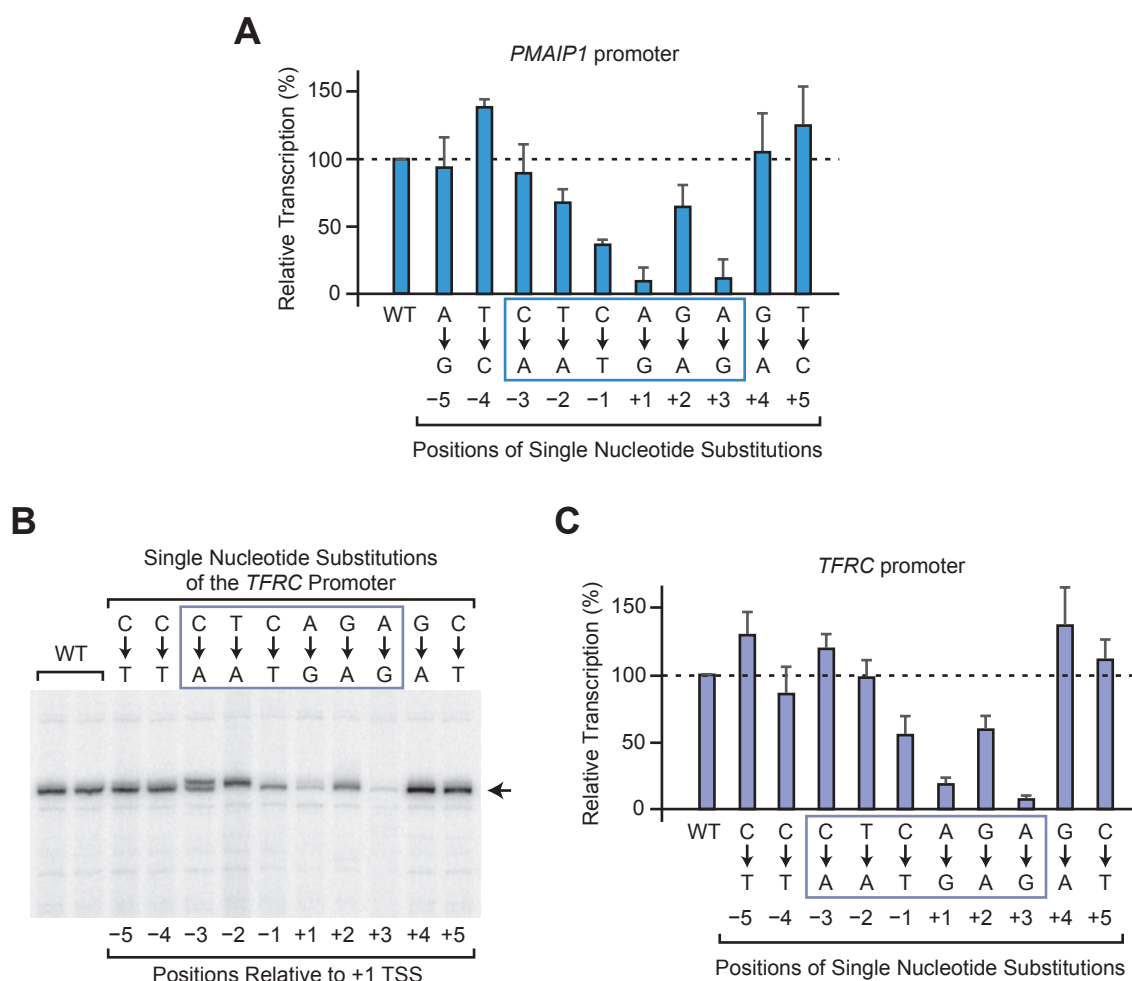
Supplemental Figure S7. The frequency of occurrence of the Inr correlates with the focus index (FI). (A) The Inr is more frequent in promoters of high FI. TSS peaks were called with FI_{\min} set at 0 and $RC_{\min} = 20$, and separated into five bins according to their FI value. The Inr frequency for each bin is shown. (B) A substantial fraction of TSSs have FI values of 0.6 or higher. The number of TSS peaks in each of the five bins shown in (A) are indicated in the graph with the exact number shown above each bar. (C) Distribution of the frequency of the FI in the TSS region (-3 to $+3$) with or without a perfect match to the Inr. TSS peaks identified as in (A) were segregated into Inr (perfect match to the $BBCA_{+1}BW$ Inr consensus) and Not Inr (not a perfect match to the Inr) groups, and then each group was subdivided into 100 bins of FI (from 0.00 to 1.00 in increments of 0.01). The figure shows the frequency of the Inr (% of total TSS peaks) versus the FI value.

Rank	Sequence	%	Rank	Sequence	%	Rank	Sequence	%	Rank	Sequence	%
1	GCCATT	2.0	51	CCCATA	0.4	101	GCCAGC	0.2	151	ACCACT	0.1
2	CTCAGT	1.6	52	ACCAGT	0.4	102	GCTAGA	0.2	152	TCCATC	0.1
3	CTCATT	1.6	53	ACCAGA	0.3	103	ACCATA	0.2	153	ATCATA	0.1
4	GCCAGA	1.5	54	GCTGAG	0.3	104	ACCACA	0.2	154	GGCGAG	0.1
5	CCCAGA	1.5	55	GTCATA	0.3	105	TTCCTT	0.2	155	CGCAGG	0.1
6	GGCAGA	1.4	56	AGCACA	0.3	106	CCCATC	0.2	156	GGCAAG	0.1
7	GGCAGT	1.3	57	GACAGA	0.3	107	ATCATT	0.2	157	GTCATC	0.1
8	CTCAGA	1.2	58	ACCATT	0.3	108	TTTACA	0.2	158	TTCAAT	0.1
9	GCCAGT	1.2	59	TTTATA	0.3	109	TTTACT	0.2	159	TTCAAA	0.1
10	CTCACT	1.1	60	GACATT	0.3	110	AGCATA	0.2	160	GCTACA	0.1
11	GGCATT	1.0	61	TGCACA	0.3	111	TCCAGG	0.2	161	TCCGAG	0.1
12	CCCAGT	1.0	62	GGGAGA	0.3	112	GGTACT	0.2	162	GCCGAG	0.1
13	CGCAGA	1.0	63	GGCAGC	0.3	113	CCCAAG	0.2	163	GTC AAT	0.1
14	GCCACT	0.9	64	TGCACT	0.3	114	GGGAGG	0.2	164	GGTGTA	0.1
15	TTCATT	0.9	65	CTCAGG	0.3	115	GGCCTT	0.2	165	GGTGTG	0.1
16	CCCATT	0.9	66	AGCACT	0.3	116	CCTAAG	0.2	166	TCTAGT	0.1
17	GTCACT	0.9	67	CTTAGT	0.3	117	GCTATT	0.2	167	TGCGCA	0.1
18	CCCCT	0.8	68	GGCAGG	0.3	118	CACAGA	0.2	168	GGTAAG	0.1
19	GTCATT	0.8	69	GGTAGT	0.3	119	GGTGTT	0.2	169	CGCGCA	0.1
20	CGCAGT	0.8	70	GACACA	0.3	120	GGCATC	0.2	170	TGCAGG	0.1
21	TCCAGT	0.8	71	GACAGT	0.2	121	ATCAGT	0.2	171	TGCAGC	0.1
22	TTCAGT	0.8	72	ATCACA	0.2	122	CTCAAT	0.2	172	GTTATA	0.1
23	TCCATT	0.8	73	CCTAGT	0.2	123	CCTAGA	0.2	173	TCCAAT	0.1
24	GTCAGT	0.8	74	CCCAGC	0.2	124	TGCATA	0.2	174	CGCAGC	0.1
25	GGCACA	0.7	75	CCCAGG	0.2	125	CCTACT	0.2	175	GCCAAG	0.1
26	TCCAGA	0.7	76	TTTAGT	0.2	126	CCTATT	0.2	176	GGCGTT	0.1
27	GTCACA	0.7	77	GTTATT	0.2	127	CTTAAT	0.2	177	AACACA	0.1
28	CGCACT	0.7	78	GGCAAT	0.2	128	CACACT	0.2	178	CCCAGG	0.1
29	CTCACA	0.7	79	GGTATT	0.2	129	CTCATC	0.2	179	TTTGTA	0.1
30	TTCATT	0.7	80	GGCGCA	0.2	130	GCTACT	0.2	180	GTTAGA	0.1
31	GCCACA	0.6	81	GGGACT	0.2	131	GTTCTT	0.1	181	CGTATT	0.1
32	AGCAGA	0.6	82	GCTAGT	0.2	132	GGTGAG	0.1	182	TCTAAG	0.1
33	TGCAGT	0.6	83	GGCATA	0.2	133	GCCGTT	0.1	183	GGCCTC	0.1
34	GGCACT	0.6	84	CTCAGC	0.2	134	CCCATG	0.1	184	TCTATC	0.1
35	CCCACA	0.6	85	GCCAAT	0.2	135	GACATA	0.1	185	GCCACC	0.1
36	TCCACT	0.6	86	TTTATT	0.2	136	TTCCTC	0.1	186	GCTAGG	0.1
37	CTCATA	0.5	87	GGGAGT	0.2	137	CGCATC	0.1	187	TTCATC	0.1
38	TGCAGA	0.5	88	GCACT	0.2	138	GCTGTT	0.1	188	CTCAAG	0.1
39	GCCATA	0.5	89	GCCAGG	0.2	139	CTCACC	0.1	189	CTCCTT	0.1
40	TTCAGA	0.5	90	CTTAGA	0.2	140	CCTGTT	0.1	190	TCTGTT	0.1
41	GCCATC	0.5	91	GGCATG	0.2	141	CACACA	0.1	191	CGGAAG	0.1
42	AGCATT	0.4	92	ATCAGA	0.2	142	GGTACA	0.1	192	CGCATA	0.1
43	CGCATT	0.4	93	CGCATG	0.2	143	CCTGAG	0.1	193	CCTACA	0.1
44	CGCACA	0.4	94	TCCATA	0.2	144	AACAGA	0.1	194	TTTAGA	0.1
45	GCCATG	0.4	95	GGTCTT	0.2	145	CTTATT	0.1	195	GTCACC	0.1
46	AGCAGT	0.4	96	GCCCTT	0.2	146	AGCAGG	0.1	196	TACAGT	0.1
47	GTCAGA	0.4	97	GTTAGT	0.2	147	ATCACT	0.1	197	CACATT	0.1
48	TCCACA	0.4	98	CCCAAT	0.2	148	ACCATC	0.1	198	GGGAAG	0.1
49	TGCATT	0.4	99	GTTACA	0.2	149	TTTAAT	0.1	199	GGTGCA	0.1
50	TTCACA	0.4	100	GTTACT	0.2	150	GACAAG	0.1	200	GGTCTC	0.1

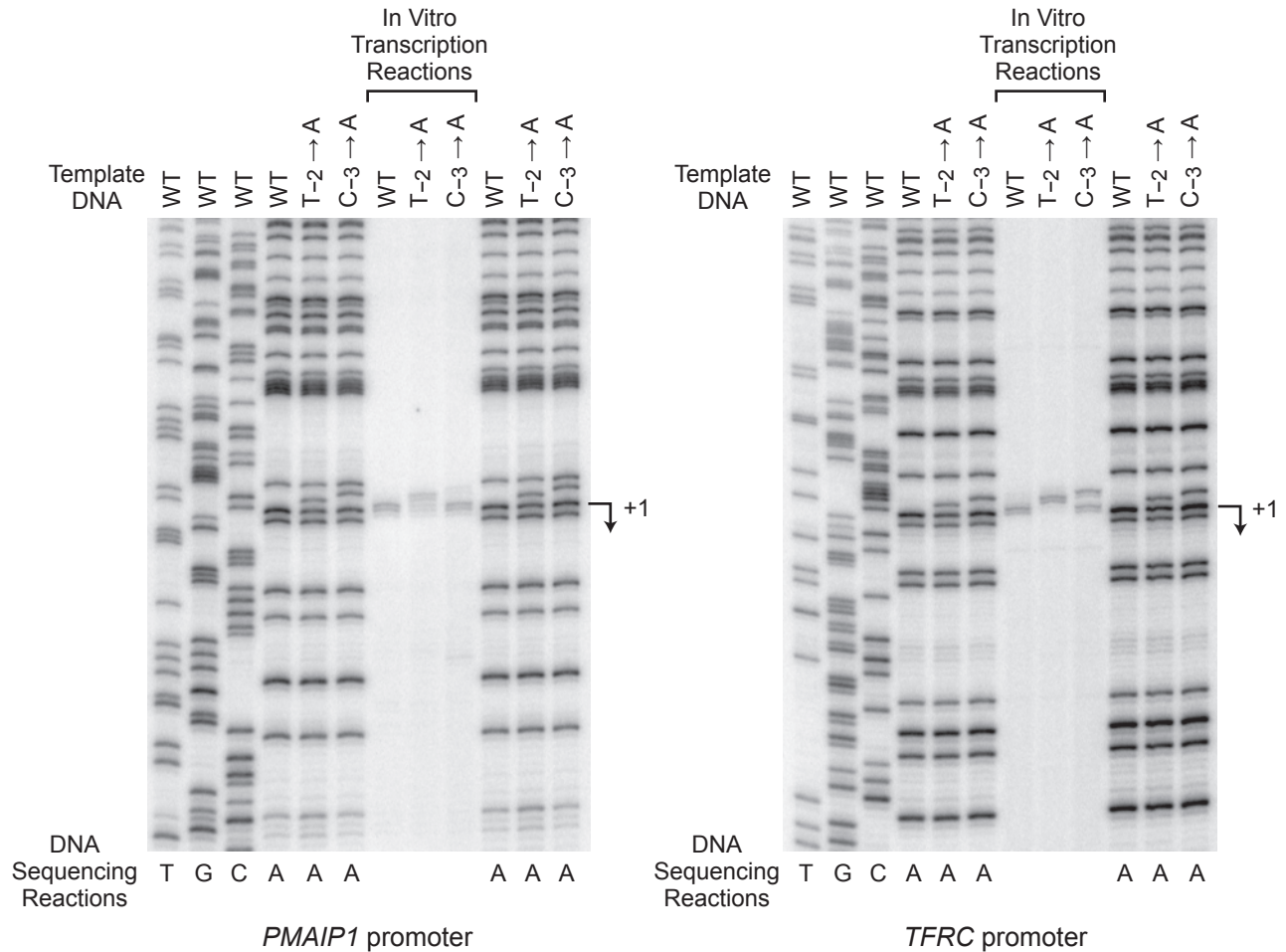
Supplemental Figure S8. The 200 most abundant hexanucleotide sequences in the Inr region (−3 to +3 relative to the +1 TSS). All 54 versions of the BB_CA₊1BW consensus are included in this list and are highlighted in blue and green type. The green sequences are the less common BB_CA₊1TA variants of the Inr consensus. The TG_CA₊1BW sequence is also underrepresented.

Rank	Sequence	%	Rank	Sequence	%	Rank	Sequence	%	Rank	Sequence	%
1*	CCANTTT	1.8	51*	TCANTTA	0.4	101*	TCANTAG	0.2	151	CTANAGG	0.1
2*	TCANTTC	1.6	52*	CCANTAC	0.4	102	GTANTTT	0.2	152	GCANGGC	0.1
3*	CCANTTC	1.5	53*	GCANATC	0.4	103	CCANGTC	0.2	153	CTANTCA	0.1
4*	CCANTCC	1.2	54*	CCANTCA	0.4	104	ACANTCG	0.2	154	GCANCCCT	0.1
5*	GCANTTC	1.1	55*	CCANATG	0.4	105	GGANTTC	0.2	155	GGANAGG	0.1
6*	TCANTTT	1.1	56*	TCANAGG	0.4	106	TTANAAA	0.2	156	CCANCAG	0.1
7*	TCANTCC	1.1	57	TTANTTT	0.4	107	CTANTTT	0.2	157	GCANGTC	0.1
8*	CCANTCG	1.0	58*	CCANTGA	0.4	108	TCCNTTT	0.2	158	TCANCTC	0.1
9*	TCANTCT	0.9	59*	TCANAGT	0.4	109*	CCANATA	0.2	159	GCGNAGA	0.1
10*	GCANTCG	0.9	60*	TCANATG	0.4	110*	GCANTTA	0.2	160	GGGNTTT	0.1
11*	TCANTCG	0.8	61*	GCANTGT	0.4	111	GTANTTG	0.2	161*	GCANTAT	0.1
12*	CCANTCT	0.8	62*	CCANAAT	0.4	112*	TCANACA	0.2	162	TTANTTG	0.1
13*	TCANTTG	0.8	63*	CCANACG	0.4	113*	GCANATG	0.2	163*	ACANTCT	0.1
14*	GCANTCC	0.8	64*	TCANAAC	0.3	114	TCGNTTT	0.2	164*	ACANTCC	0.1
15*	TCANTGC	0.8	65*	TCANTGA	0.3	115	TTCNCTT	0.2	165*	GCANATA	0.1
16*	GCANTTG	0.7	66*	CCANTAG	0.3	116	CCANCCG	0.2	166	CTANAAA	0.1
17*	GCANTGC	0.7	67	CCANGTT	0.3	117	CCANCCCT	0.2	167*	ACANAGC	0.1
18*	GCANTTT	0.7	68*	TCANACC	0.3	118*	TCANTAT	0.2	168	ACANGTG	0.1
19*	CCANAGC	0.7	69*	GCANAAT	0.3	119*	ACANAGA	0.2	169*	CCANTAT	0.1
20*	CCANTTG	0.6	70*	TCANATC	0.3	120	TTANTCA	0.2	170	TCTNTTT	0.1
21*	CCANATT	0.6	71	ACANTTT	0.3	121	TTANAGG	0.2	171	CTGNNGTT	0.1
22*	CCANAAC	0.6	72*	GCANATT	0.3	122*	TCANATA	0.2	172	TTGNAGT	0.1
23*	GCANAAG	0.6	73*	GCANACT	0.3	123*	GCANTAA	0.2	173	TCANGTG	0.1
24*	CCANACT	0.6	74*	CCANACA	0.3	124*	GCANTAC	0.2	174	GTGNTTT	0.1
25*	TCANTCA	0.6	75*	CCANATC	0.3	125	TCANGCG	0.2	175	CTGNNGAA	0.1
26*	GCANTCT	0.6	76*	CCANTTA	0.3	126	TCANGAA	0.1	176	CCGNAAA	0.1
27*	CCANAGT	0.6	77*	TCANAAT	0.3	127	CCANCTG	0.1	177	GTANTGG	0.1
28*	GKANAGA	0.5	78*	CCANACC	0.3	128*	CCANTAA	0.1	178	TTANATT	0.1
29*	GKANAGG	0.5	79	CTANTTC	0.3	129	TTANTAA	0.1	179	GCGNACT	0.1
30*	CCANTGG	0.5	80*	GCANTGA	0.3	130	CTGNNGTC	0.1	180	CCANCCC	0.1
31*	GKANAAA	0.5	81*	TCANACG	0.3	131	CCGNTTT	0.1	181	GTANATT	0.1
32*	CCANAAA	0.5	82*	GCANTCA	0.3	132	CCGNTTC	0.1	182	GGANTCT	0.1
33*	TCANAAA	0.5	83*	GKANACA	0.3	133	GCANGAG	0.1	183	GGANTCC	0.1
34*	CCANAGG	0.5	84	GCANCCG	0.3	134	TTANTCT	0.1	184	TCANCGG	0.1
35*	CCANAAG	0.5	85*	GCANTAG	0.3	135*	ACANAAT	0.1	185	GGANAGA	0.1
36*	GCANAGT	0.5	86	GCANGCG	0.3	136*	ACANAAA	0.1	186	TCANCTT	0.1
37*	GCANAGC	0.5	87	CTGNTTT	0.3	137	CTANTCT	0.1	187	CCANCTC	0.1
38*	CCANTGT	0.5	88	TTANTTC	0.3	138	GCANCTG	0.1	188	TTANTTA	0.1
39*	CCANTGC	0.5	89*	GCANACC	0.3	139	GCANCAG	0.1	189	TCANCCG	0.1
40	CCANCTT	0.5	90*	GCANAAC	0.2	140	GCANGTG	0.1	190	CCANGAG	0.1
41*	TCANACT	0.5	91	GCGNAGT	0.2	141	GTANTTC	0.1	191	TCCNTTC	0.1
42*	TCANAGC	0.5	92	TTANTCC	0.2	142	CCANGTG	0.1	192	CTGNNGGA	0.1
43*	TCANAAG	0.4	93	CCANGGC	0.2	143	CTANTTG	0.1	193	TCANGCC	0.1
44*	TCANTGT	0.4	94	ACANTTC	0.2	144	CCANCGC	0.1	194	ACANATT	0.1
45*	TCANAGA	0.4	95*	TCANTAA	0.2	145	CCANGCC	0.1	195	GCANGAC	0.1
46*	TCANATT	0.4	96*	TCANTAC	0.2	146	CTANTCG	0.1	196	TTANTCG	0.1
47*	TCANTGG	0.4	97	TCANCTG	0.2	147	TTANAAG	0.1	197*	ACANACA	0.1
48*	GCANTGG	0.4	98	CCANGAC	0.2	148	GCANGCT	0.1	198*	ACANACT	0.1
49*	CCANAGA	0.4	99	GCCNTTT	0.2	149	CCGNNGAC	0.1	199	CCGNNGCT	0.1
50*	GCANACG	0.4	100	GTCNTTT	0.2	150	TCANGTC	0.1	200	TTCNTTT	0.1

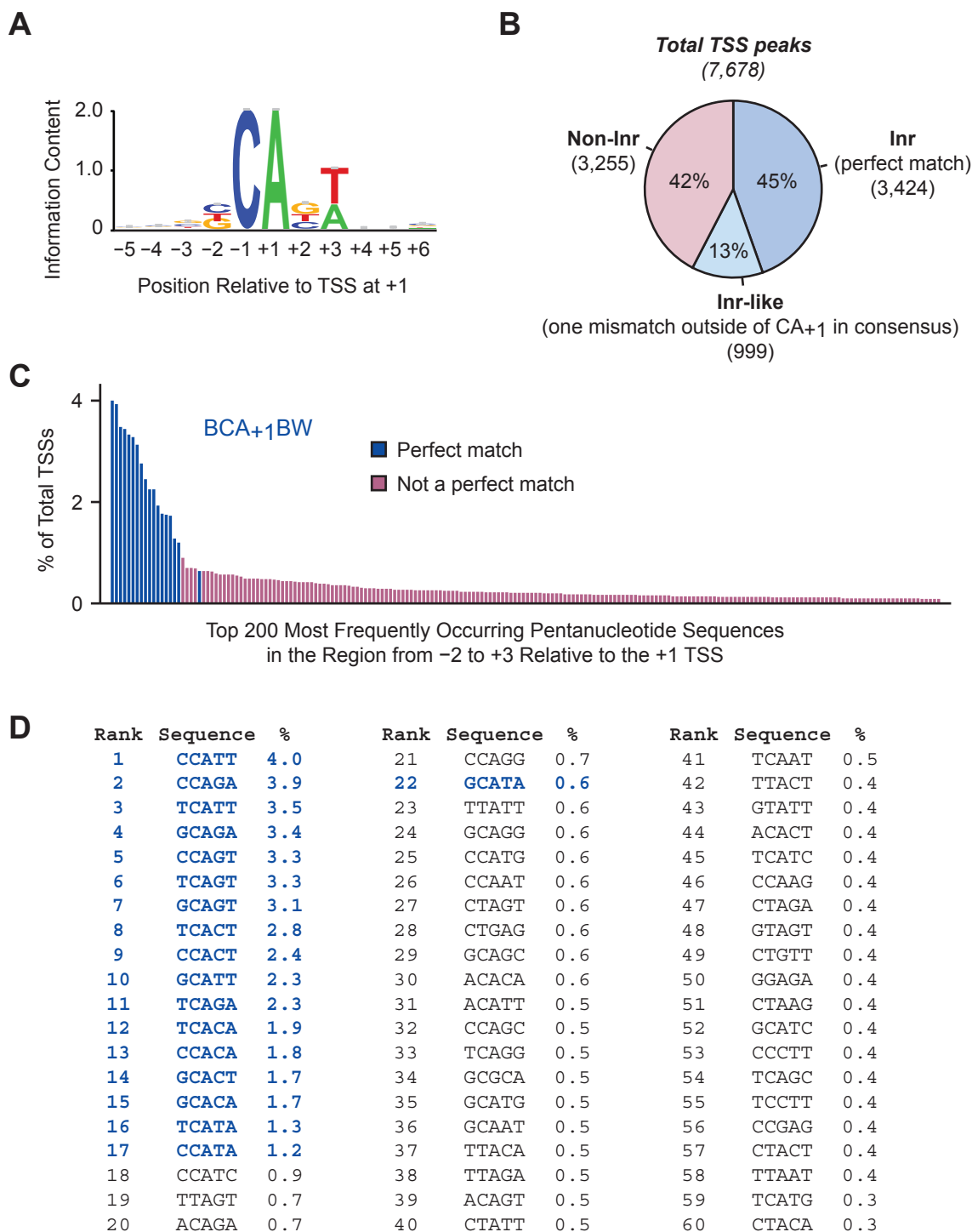
Supplemental Figure S9. The 200 most abundant nonrandom hexanucleotide sequences in the Inr region from -2 to +5 (excluding the N₊₂ position) relative to the +1 TSS. Variants of the YYA₊₁NWYY consensus are highlighted in blue type. This list includes 24 of the 32 variants of the YYA₊₁NWYY consensus. The asterisks indicate sequences that match the BCCA₊₁BW Inr consensus at the -2, -1, +1, and +3 positions.



Supplemental Figure S10. The $BBCA_{+1}BW$ Inr consensus sequence is essential for efficient and accurate transcription initiation. The core promoter region from -50 to $+51$ relative to the $+1$ TSS (for DNA sequences, see Supplemental Fig. S13) of the *PMAIP1* and *TFRC* genes were used in these experiments. (A) Alterations in the *PMAIP1* Inr sequence impair transcriptional strength. Quantitation of the transcription levels from at least four experiments performed as in Fig. 4A of the main text. Mutations in the $BBCA_{+1}BW$ Inr consensus are within the blue box. The data are the mean (relative to WT) \pm sd. (B) Alterations in the *TFRC* Inr sequence impair transcriptional strength and start site selection. The consensus Inr sequence in the *TFRC* 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. The horizontal arrow indicates the $+1$ TSS. The data are representative of at least four independent experiments. Mutations in the $BBCA_{+1}BW$ Inr consensus are within the purple box. (C) Quantitation of the transcription levels from at least four experiments performed as in (B). The data are the mean (relative to WT) \pm sd.



Supplemental Figure S11. Mutation of the B₋₂ or B₋₃ position of the Inr to an A results in a shift in the start site selection. The indicated wild-type and mutant versions of the *PMAIP1* and *TFRC* promoters were subjected to in vitro transcription and primer extension analysis. The reverse transcription products were separated by denaturing polyacrylamide gel electrophoresis in parallel with DNA sequencing ladders that were generated with the same 5'-end labeled oligonucleotide that was used in the primer extension reactions. The +1 position of the Inr is indicated.



Supplemental Figure S12. The BCA₊₁BW sequence is a simplified representation of the human Inr at focused TSSs. (A) Sequence logo of the BCA₊₁BW Inr consensus. The 3,424 focused TSS peaks with a perfect match to BCA₊₁BW were used to generate the logo. (B) The BCA₊₁BW Inr consensus occurs frequently in focused promoters. 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 numbers of TSSs in each group are shown in parentheses. (C) The BCA₊₁BW Inr consensus generally represents the most frequently occurring sequences at the TSS. The graph shows the 200 most frequently occurring pentanucleotides from -2 to +3 relative to the +1 TSS. (D) List of the 60 most abundant pentanucleotides from -2 to +3 relative to the +1 TSS. The 18 versions of the BCA₊₁BW consensus (shown in blue type) are found in the 22 most commonly occurring pentanucleotides.

+1
└─┘

PMAIP1 WT CCAGGGAAGTTCTCACTGGACAAAAGCGTGGTCTCTGGCGCGGGGATCTCAGAGTTTCCCGGGCACTCACCGTGTGTAGTTGGCATCTCCGCGCGTCCGGA
PMAIP1 mut -5 CCAGGGAAGTTCTCACTGGACAAAAGCGTGGTCTCTGGCGCGGGGATCTCAGAGTTTCCCGGGCACTCACCGTGTGTAGTTGGCATCTCCGCGCGTCCGGA
PMAIP1 mut -4 CCAGGGAAGTTCTCACTGGACAAAAGCGTGGTCTCTGGCGCGGGGACCTCAGAGTTTCCCGGGCACTCACCGTGTGTAGTTGGCATCTCCGCGCGTCCGGA
PMAIP1 mut -3 CCAGGGAAGTTCTCACTGGACAAAAGCGTGGTCTCTGGCGCGGGGATATCAGAGTTTCCCGGGCACTCACCGTGTGTAGTTGGCATCTCCGCGCGTCCGGA
PMAIP1 mut -2 CCAGGGAAGTTCTCACTGGACAAAAGCGTGGTCTCTGGCGCGGGATCAGAGTTTCCCGGGCACTCACCGTGTGTAGTTGGCATCTCCGCGCGTCCGGA
PMAIP1 mut -1 CCAGGGAAGTTCTCACTGGACAAAAGCGTGGTCTCTGGCGCGGGGATCTTAGAGTTTCCCGGGCACTCACCGTGTGTAGTTGGCATCTCCGCGCGTCCGGA
PMAIP1 mut +1 CCAGGGAAGTTCTCACTGGACAAAAGCGTGGTCTCTGGCGCGGGGATCTCGAGTTTCCCGGGCACTCACCGTGTGTAGTTGGCATCTCCGCGCGTCCGGA
PMAIP1 mut +2 CCAGGGAAGTTCTCACTGGACAAAAGCGTGGTCTCTGGCGCGGGGATCTCAAAGTTTCCCGGGCACTCACCGTGTGTAGTTGGCATCTCCGCGCGTCCGGA
PMAIP1 mut +3 CCAGGGAAGTTCTCACTGGACAAAAGCGTGGTCTCTGGCGCGGGGATCTCAGGGTTTCCCGGGCACTCACCGTGTGTAGTTGGCATCTCCGCGCGTCCGGA
PMAIP1 mut +4 CCAGGGAAGTTCTCACTGGACAAAAGCGTGGTCTCTGGCGCGGGGATCTCAGAGTTTCCCGGGCACTCACCGTGTGTAGTTGGCATCTCCGCGCGTCCGGA
PMAIP1 mut +5 CCAGGGAAGTTCTCACTGGACAAAAGCGTGGTCTCTGGCGCGGGGATCTCAGAGCTTCCCGGGCACTCACCGTGTGTAGTTGGCATCTCCGCGCGTCCGGA
TFRC WT CGGGGGCGGGCCAGGCTATAAACCGCCGGTTAGGGGCGCCATCCCTCAGAGCGTCGGGATATCGGGTGGCGGCTCGGGACGGAGGACCGCTAGTGTG
TFRC mut -5 CGGGGGCGGGCCAGGCTATAAACCGCCGGTTAGGGGCGCCATCTCTCAGAGCGTCGGGATATCGGGTGGCGGCTCGGGACGGAGGACCGCTAGTGTG
TFRC mut -4 CGGGGGCGGGCCAGGCTATAAACCGCCGGTTAGGGGCGCCATCCTCTCAGAGCGTCGGGATATCGGGTGGCGGCTCGGGACGGAGGACCGCTAGTGTG
TFRC mut -3 CGGGGGCGGGCCAGGCTATAAACCGCCGGTTAGGGGCGCCATCCCTCAGAGCGTCGGGATATCGGGTGGCGGCTCGGGACGGAGGACCGCTAGTGTG
TFRC mut -2 CGGGGGCGGGCCAGGCTATAAACCGCCGGTTAGGGGCGCCATCCCCACAGAGCGTCGGGATATCGGGTGGCGGCTCGGGACGGAGGACCGCTAGTGTG
TFRC mut -1 CGGGGGCGGGCCAGGCTATAAACCGCCGGTTAGGGGCGCCATCCCTTAGAGCGTCGGGATATCGGGTGGCGGCTCGGGACGGAGGACCGCTAGTGTG
TFRC mut +1 CGGGGGCGGGCCAGGCTATAAACCGCCGGTTAGGGGCGCCATCCCTCAGAGCGTCGGGATATCGGGTGGCGGCTCGGGACGGAGGACCGCTAGTGTG
TFRC mut +2 CGGGGGCGGGCCAGGCTATAAACCGCCGGTTAGGGGCGCCATCCCTCAAAGCGTCGGGATATCGGGTGGCGGCTCGGGACGGAGGACCGCTAGTGTG
TFRC mut +3 CGGGGGCGGGCCAGGCTATAAACCGCCGGTTAGGGGCGCCATCCCTCAGGGCGTCGGGATATCGGGTGGCGGCTCGGGACGGAGGACCGCTAGTGTG
TFRC mut +4 CGGGGGCGGGCCAGGCTATAAACCGCCGGTTAGGGGCGCCATCCCTCAGAACGTCGGGATATCGGGTGGCGGCTCGGGACGGAGGACCGCTAGTGTG
TFRC mut +5 CGGGGGCGGGCCAGGCTATAAACCGCCGGTTAGGGGCGCCATCCCTCAGAGGTCGGGATATCGGGTGGCGGCTCGGGACGGAGGACCGCTAGTGTG
PFKFB3 WT TCTGCGGCCAGCCCGGACTCTTTAAAGGCCGGCGGTGCGCGGGGCATCCAGCCAGCCGAGAGGAGGCGAGCAGCAGGGCCCTGGTGGCGAGAGCGCGGC
PFKFB3 mut +3 TCTGCGGCCAGCCCGGACTCTTTAAAGGCCGGCGGTGCGCGGGGCATCCAGTCAAGCCGAGAGGAGGCGAGCAGCAGGGCCCTGGTGGCGAGAGCGCGGC
FRAS1 WT CGTATGGTGCCAAGCGAACTTTAAAGAGCTGCTTCGGACAAACCAGAGCCAGTATTTCCACTGTGCGGGACCCGGGATCGGAAGGGTCTAGCCCGAGGGAA
FRAS1 mut +3 CGTATGGTGCCAAGCGAACTTTAAAGAGCTGCTTCGGACAAACCAGAGCCAGTATTTCCACTGTGCGGGACCCGGGATCGGAAGGGTCTAGCCCGAGGGAA
FAM11B WT GGGAAAAGTTTCACTGAGAGATATAAAGAGCAGTCTTTCCAGCACCCTGCAAATCCAGAGCGGCGGGCACTGACGGGCACTTGCACCGTGTGGACAGACTC
FAM11B mut +2 GGGAAAAGTTTCACTGAGAGATATAAAGAGCAGTCTTTCCAGCACCCTGCAAGTCCAGAGCGGCGGGCACTGACGGGCACTTGCACCGTGTGGACAGACTC
ASTL WT GGGAGGTGGAGCAGCTGCTATTTAAGAGGGGGTGGTGGTCCGGTCTGCAATTAGGTTACTGTGCTTTGCTGGGGCTTGGTCTTGTGTTGCTGAAGGGGCA
ASTL mut +2 GGGAGGTGGAGCAGCTGCTATTTAAGAGGGGGTGGTGGTCCGGTCTGCAAGTTAGGTTACTGTGCTTTGCTGGGGCTTGGTCTTGTGTTGCTGAAGGGGCA
PPIA WT CGGGCGGGGCGGAACGTGGTATAAAGGGGCGGGAGGCCAGGCTCGTGCCGTTTTGACAGCGCCACCAGGAGGAAAACCGTGTACTATTAGCCATGGTCA
PPIA mut +1 CGGGCGGGGCGGAACGTGGTATAAAGGGGCGGGAGGCCAGGCTCGTGCCATTTTGCAGAGCGCCACCAGGAGGAAAACCGTGTACTATTAGCCATGGTCA
AMD1 WT CTTTTGGGGGAGCCGGGATATAAAGGGCGGTGCTCAGCAGCGCTCTCCTTACACAGTATGGCCGGCGACATTAGCTAGCGCTCGCTCTACTCTCTCT
AMD1 mut +1 CTTTTGGGGGAGCCGGGATATAAAGGGCGGTGCTCAGCAGCGCTCTCACTTACACAGTATGGCCGGCGACATTAGCTAGCGCTCGCTCTACTCTCTCT
ZSWIM6 WT CGCGCTTCTAGTGCCGTTTTATAGGGTCCCGCACTTCCGCTGTGCGGTTAGAAGCGGCGGTCATGGCGGAGCGCGACAGCAGCCCTCTCCCGGAAA
ZSWIM6 mut -1 CGCGCTTCTAGTGCCGTTTTATAGGGTCCCGCACTTCCGCTGTGCGGTCAGAAGCGGCGGTCATGGCGGAGCGCGGACAGCAGCCTCTCCCGGAAA
PRSS22 WT ACGGCATTCGCGCTCCAGGATAAAAACCTGGGGCGACCTGACAGGAACTACACACCCTGACCCGATCGCCCTGGGTCTCTCGAGCCTGTGCCTGCTCC
PRSS22 mut -1 ACGGCATTCGCGCTCCAGGATAAAAACCTGGGGCGACCTGACAGGAACTACACACCCTGACCCGATCGCCCTGGGTCTCTCGAGCCTGTGCCTGCTCC
CA2 WT ACGAAGTTGGCGGGAGCCTATAAAGCTGGTGCAGGCGGACCCCGGACACACAGTGCAGGCGCCCAAGCCGCGCCGAGATCGGTGCCGATTCCTGC
CA2 mut -2 ACGAAGTTGGCGGGAGCCTATAAAGCTGGTGCAGGCGGACCCCGGACACACAGTGCAGGCGCCCAAGCCGCGCCGAGATCGGTGCCGATTCCTGC
BTW WT TGTGTGCAGGAGGAGGGGGATAAATAGGAGGCTCCTCCTCCGCGGACATTACAGGAGCGGCGGCTCCCGCCTGGGTGTTTCCCTGCCTTGTAGC
BTW mut -2 TGTGTGCAGGAGGAGGGGGATAAATAGGAGGCTCCTCCTCCGCGGACATTACAGGAGCGGCGGCTCCCGCCTGGGTGTTTCCCTGCCTTGTAGC
EMD1 WT GCCACTGAGGGACCGACCCATAAAGGCCGCTCCGCGAGGGGTGCGCAGCATTCCGACAGAGGGCGCTTCGACGGGCTGGGCTGTGCGCCTGCGCAGTGTGG
EMD1 mut -3 GCCACTGAGGGACCGACCCATAAAGGCCGCTCCGCGAGGGGTGCGCAGCATTCCGACAGAGGGCGCTTCGACGGGCTGGGCTGTGCGCCTGCGCAGTGTGG
MAFB WT CCCAGTGACATCAGGAGGCGATAAAGGCTCGGGCGCCCGGATCCAGCACAGCTGCACCCGAGCTGCAGGCGGCTGCAGGCGAGAGAGCGTAAGAGC
MAFB mut -3 CCCAGTGACATCAGGAGGCGATAAAGGCTCGGGCGCCCGGATCCAGCACAGCTGCACCCGAGCTGCAGGCGGCTGCAGGCGAGAGAGCGTAAGAGC

Supplemental Figure S13. Sequences of the core promoters used for in vitro transcription experiments. Sequences from -50 to +51 relative to the +1 TSS were synthesized with overhangs that allow cloning into the Xba I (upstream) and Pst I (downstream) sites of pUC119T vector. The location of the +1 TSS is indicated.

Supplemental Materials and Methods

Cell culture

MCF-7 cells (ATCC) were maintained in DMEM (ATCC) supplemented with 10% FBS (ATCC), 15 µg/mL bovine insulin (Sigma-Aldrich), 50 U/mL penicillin (ThermoFisher), and 50 µg/mL streptomycin (ThermoFisher).

5'-GRO-seq

Two 5'-GRO-seq experiments were carried out by variation of the procedure described in Duttke et al. (2015) and Hetzel et al. (2016). We were curious with regard to whether the RNA fragmentation step would affect the 5'-GRO-seq data, and we therefore performed one experiment with the RNA fragmentation and other experiment without the RNA fragmentation. As shown in Supplemental Fig. S2, we obtained similar results from the two experiments, and thus used both datasets for our analyses. For each experiment, nuclear run-on reactions were performed with 10^7 MCF-7 nuclei in the presence of BrUTP (Sigma Aldrich) for labeling of the RNA. Trizol LS (Fisher Scientific) was used to terminate the reaction and to extract the RNA.

For 5'-GRO-seq with fragmentation, RNA was hydrolyzed by Zn(II)-mediated fragmentation (Ambion). BrU-labeled transcripts were immunoprecipitated with anti-BrdU agarose beads (Santa Cruz Biotech) and extracted with Trizol LS. The resulting RNA was incubated with T4 polynucleotide kinase (NEB) in low pH buffer followed by treatment with calf intestinal alkaline phosphatase (CIP; NEB) to achieve 5' and 3' dephosphorylation. A second BrU enrichment was performed, and 5'-monophosphate-RNAs were removed with Terminator 5'-phosphate-dependent exonuclease (Epicentre), as described in Hetzel et al. (2016). The RNA was subjected to a second round of dephosphorylation with CIP (NEB). Decapping and library preparation were performed as described in Hetzel et al. (2016).

For 5'-GRO-seq without RNA fragmentation, the run-on step was followed by RNA extraction with Trizol. BrU-labeled transcripts were immunoprecipitated with anti-BrdU agarose

beads. RNA was dephosphorylated with CIP (NEB), and subjected to a second round of BrU enrichment that was followed by a second round of CIP (NEB). Decapping and library preparation were performed as described in Hetzel et al. (2016).

It might be noted that these experiments involve the ligation of an adapter (GTTCAGAGTTCTACrArGUrCrCrGrArCrGrAUrC) to the 5' end of the transcripts. The ligation step has been found to exhibit sequence bias (Jayaprakash et al. 2011). Hence, it might alter the relative amounts of different transcripts, but would not affect the accuracy of the mapping of the 5' ends of the transcripts.

5'-GRO-seq read processing

Adapter sequences were removed from 5'-GRO-seq reads with homerTools trim (Heinz et al. 2010), and the resulting reads were mapped to the hg19 human genome by using Bowtie2 with the default settings (Langmead and Salzberg 2012). High quality reads (MAPQ \geq 10) were selected. This yielded 15,594,012 reads for 5'-GRO-seq without fragmentation, and 23,121,822 reads for 5'-GRO-seq with fragmentation. The location of the 5' end of each read was then retrieved and pile-ups were calculated at these genomic positions.

Use of FocusTSS for the identification of focused transcription start sites

FocusTSS is a Python program (Focus_TSS.py; available in Supplemental Material) that was developed to identify focused TSSs in 5'-GRO-seq data. [A version of FocusTSS that accepts decimal values (Focus_TSSdecimal.py) is also available.] As described in the text and in Fig. 1A, FocusTSS is based on the properties of transcription preinitiation complexes. FocusTSS identifies peaks that have a user-specified minimum read count (RC_{min} ; in this study, typically 20 reads, which is approximately 1 RPM) and are local maxima in a ± 2 nt window. Then, for each peak, the focus index (FI) is calculated as the ratio of the combined 5'-GRO-seq reads in a narrow user-specified (typically, 5 nt) window that is centered on the peak divided by the combined reads in a wider user-specified (typically, 25 nt) window that is centered on the peak.

If the FI is at least as large as a user-specified minimal FI (FI_{\min} ; typically, 0.67), then the peak is called as a focused TSS.

In our analysis of the MCF-7 cell data, we used FocusTSS separately on each of the two 5'-GRO-seq datasets. For each set of focused TSSs, peaks intersecting with regions of RepeatMasker were discarded. Then, peaks found at the same genomic location in both datasets were considered to be common in the two experiments.

As a matter of clarification for Supplemental Figures S1 and S7, we mention here that the peaks used to generate the graphs are common in both (with and without fragmentation) datasets, and that the specific FI and RC values that are shown in the figures correspond to data from the 5'-GRO-seq experiment with Zn(II)-mediated fragmentation.

Peak annotation, motif discovery and generation of the sequence logo

Hypergeometric Optimization of Motif EnRichment (HOMER) (Heinz et al. 2010) was used for motif discovery and for FocusTSS annotation.

For motif discovery, the findMotifs.pl HOMER tool was used to search for 11-nt motifs in the sequences corresponding to the -5 to +6 region of FocusTSS peaks. The DNA stretch spanning positions -1205 to -1195 of the same peaks were used as background sequences.

For TSS annotation, peaks marked by the annotatepeaks.pl HOMER tool as "Exon (coding)", "3'UTR", "5'UTR", "Non-coding", and "TTS" were placed in the "Exon" category, and the "Intergenic", "Intron", and "Promoter-TSS" (which we renamed as "Promoter") groups were not changed.

WebLogo 3 (Crooks et al. 2004; Schneider et al. 1990) was used to generate the sequence logo of the Inr from the 3,071 FocusTSS peaks with a perfect match to BBCA+1BW.

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; the sequences are shown in Supplemental

Fig. S13) in the Xba I and Pst I sites of the pUC119T vector, which contains an RNA polymerase III-specific terminator that blocks any potential transcription by RNA polymerase III (Duttke 2014). Constructs containing mutated promoters were generated by using the Q5® Site-Directed Mutagenesis Kit (NEB) according to the manufacturer's instructions.

In vitro transcription reactions were performed essentially as previously described (Theisen et al. 2013). Briefly, 500 ng of DNA template was pre-incubated with HeLa nuclear extract for PIC assembly at 30 °C for 1 hour in 46 µL of transcription buffer [20 mM HEPES-K⁺ (pH 7.6), 50 mM KCl, 6 mM MgCl₂, 2.5% (w/v) polyvinyl glycol (compound), 0.5 mM DTT, 3 mM ATP, 0.02 mM EDTA, and 2% (v/v) glycerol]. rNTPs [4 µL; 0.4 mM final concentration in each rNTP] were added to initiate transcription. The reaction was incubated at 30 °C for 20 min and terminated by the addition of 150 µL of Stop Mix [20 mM EDTA, 200 mM NaCl, 1% (w/v) SDS, 0.3 mg/mL glycogen]. Proteinase K (5 µL, 2.5 mg/mL) was added, and the mixture was incubated at 30 °C for 15 min. The nucleic acids were isolated by phenol-chloroform extraction and ethanol precipitation, and then subjected to primer extension analysis with 5'-³²P-labeled M13 reverse sequencing primer (5'-AGCGGATAACAATTTACACAGGA). The primer extension products were resolved on a polyacrylamide-urea gel and quantified by using a Typhoon imager (GE Health Sciences). All in vitro transcription experiments were performed independently at least three times to ensure reproducibility of the data.

Databases and accession numbers

The 5'-GRO-seq data from the MCF-7 cells are available from Gene Expression Omnibus (GEO; accession number, GSE90035).

GRO-cap files from Core et al. (2014) (GSM1480321, GSM1480325), and 5'-GRO-seq data from Duttke et al. (2015) (GSM1558745) were obtained from the Gene Expression Omnibus (GEO) website.

Supplemental References

- 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.
- Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res* **14**: 1188–1190.
- Duttke SHC. 2014. RNA polymerase III accurately initiates transcription from RNA polymerase II promoters in vitro. *J Biol Chem* **289**: 20396–20404.
- 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.
- 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.
- Hetzl J, Duttke SH, Benner C, Chory J. 2016. Nascent RNA sequencing reveals distinct features in plant transcription. *Proc Natl Acad Sci USA* **113**: 12316-12321.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.
- Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R. 2011. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Res* **39**: e141.
- Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18**: 6097–6100.
- 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.

Wu H, Caffo B, Jaffee HA, Irizarry RA, Feinberg AP. 2010. Redefining CpG islands using hidden Markov models. *Biostatistics* **11**: 499-514.

OUTLOOK

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 $\sim 50\%$ of human core promoters contain an Inr (Gershenson 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>.

© 2017 Kugel and Goodrich This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <http://genesdev.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

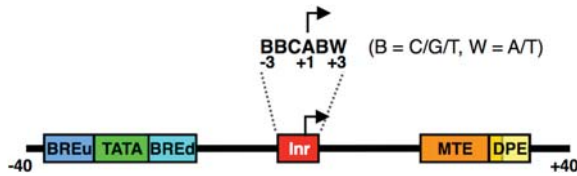


Figure 1. Relative locations of select human core promoter elements and the Inr consensus sequence found in promoters with focused TSSs. The promoter elements depicted include BREu (the upstream TFIIB recognition element), TATA (the TATA box), BREd (the downstream TFIIB recognition element), Inr (new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

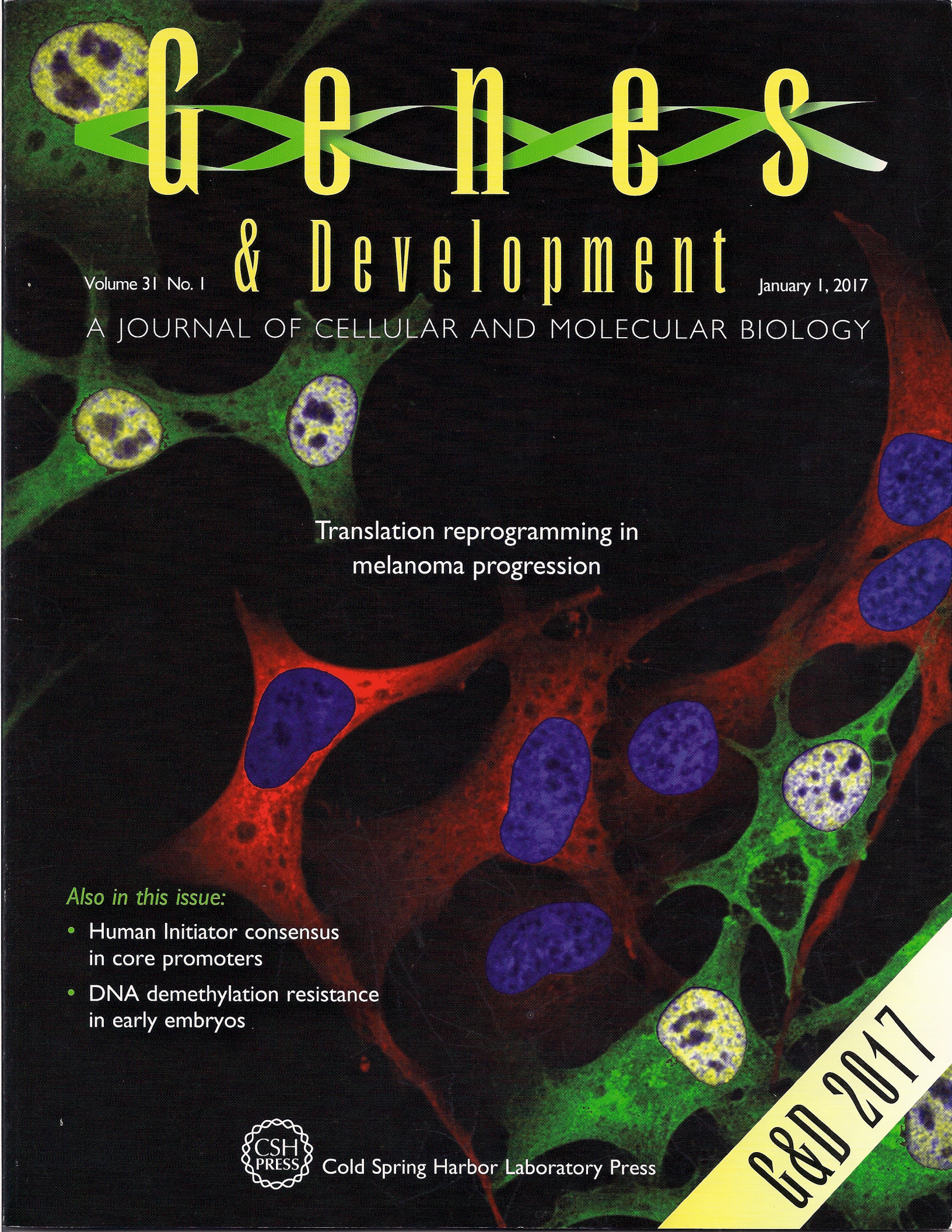
core promoters were chosen that each differed from consensus at one position; these positions were mutated to create the Inr consensus. Mutating positions -1 to $+3$ toward consensus increased transcriptional activity, and, again, the mutations at positions $+1$ and $+3$ had the greatest effect.

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.
- Gershenson 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.



Genes

& Development

Volume 31 No. 1

January 1, 2017

A JOURNAL OF CELLULAR AND MOLECULAR BIOLOGY

Translation reprogramming in
melanoma progression

Also in this issue:

- Human Initiator consensus in core promoters
- DNA demethylation resistance in early embryos



Cold Spring Harbor Laboratory Press

G&D 2017