NF-Y controls fidelity of transcription initiation at gene promoters through maintenance of the nucleosome-depleted region

Andrew J Oldfield[1,6,*], Telmo Henriques[1,3], Adam B. Burkholder[2], Damien Paulet[4,5], Senthilkumar Cinghu[1], Pengyi Yang[1,7], Benjamin S. Scruggs[1], Christopher A. Lavender[2], Dhirendra Kumar[1], Brian Bennett[2], Eric Rivals[4,5], Karen Adelman[1,3,*], Raja Jothi[1,*]

[1]Epigenetics and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC, USA

[2]Integrative Bioinformatics Support Group, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC, USA

[3]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA, USA

[4]Department of Computer Science, LIRMM, CNRS et Université de Montpellier, Montpellier, France

[5]Institut de Biologie Computationnelle (IBC), Université de Montpellier, Montpellier, France

[6]Current address: Institute of Human Genetics, CNRS, University of Montpellier, Montpellier, France.

[7]Current address: Charles Perkins Centre and School of Mathematics and Statistics, University of Sydney, Sydney, Australia

*Correspondence: andrew.oldfield@igh.cnrs.fr, karen_adelman@hms.harvard.edu, jothi@nih.gov

SUMMARY

Faithful transcription initiation is critical for accurate gene expression, yet the mechanisms underlying specific transcription start site (TSS) selection in mammals remain unclear. Here, we show that the histone-fold domain protein NF-Y, a ubiquitously expressed transcription factor, controls the fidelity of transcription initiation at gene promoters. We report that NF-Y maintains the region upstream of TSSs in a nucleosome-depleted state while simultaneously protecting this accessible region against aberrant and/or ectopic transcription initiation. We find that loss of NF-Y binding in mammalian cells disrupts the promoter chromatin landscape, leading to nucleosomal encroachment over the canonical TSS. Importantly, this chromatin rearrangement is accompanied by upstream relocation of the transcription pre-initiation complex and ectopic transcription initiation. Further, this phenomenon generates aberrant extended transcripts that undergo translation, disrupting gene expression profiles. These results demonstrate that NF-Y is a central player in TSS selection in metazoans and highlight the deleterious consequences of inaccurate transcription initiation.

HIGHLIGHTS

- NF-Y binding upstream of TSSs maintains accessible promoter chromatin architecture
- Loss of NF-Y binding at promoters leads to ectopic transcription initiation
- Aberrant mRNA species generated in NF-Y-depleted cells undergo translation
- Inaccurate TSS selection upon NF-Y depletion perturbs gene and protein expression

KEYWORDS

Transcription, TSS selection, Transcription initiation, Nucleosome-depleted region, NF-Y, CCAAT

INTRODUCTION

While the sequence, structure, and binding partners of gene promoters have been intensely scrutinized for nearly half a century (Epstein and Beckwith, 1968), how the cell discerns when and where to initiate transcription is still not fully understood (Valen and Sandelin, 2011). Recent studies have established basic rules regarding spatial arrangements of cis-regulatory elements, ordered recruitment of general transcription factors (GTFs) for transcription pre-initiation complex (PIC) formation, and the role of chromatin in defining the promoter environment (Core et al., 2014; Danino et al., 2015; Lai and Pugh, 2017; Muller and Tora, 2014; Roy and Singer, 2015; Vo Ngoc et al., 2017). One key determinant of active gene promoters is the requirement for an accessible transcription initiation/start site (TSS), characterized by a nucleosome-depleted region (NDR) flanked by two well-positioned nucleosomes (the −1 and +1 nucleosomes) (Scruggs et al., 2015).

Within the NDR, several core promoter elements such as the TATA box and the Initiator (Inr) element exhibit a positional bias relative to the TSS (Carninci et al., 2006; Smale and Kadonaga, 2003) and play important roles in TSS selection. However, the core promoter elements vary from one promoter to the next and can be either absent or present multiple times within a single NDR, suggesting that these elements are not the sole determinants of TSS selection. Thus, how the RNA Polymerase II (Pol II) chooses one transcription initiation site over another remains unclear. In yeast, mutational studies have identified several GTFs and other accessory factors with key roles in TSS selection (Ghazy et al., 2004; Kasahara et al., 2011; Pinto et al., 1994; Rhee and Pugh, 2012; Venters and Pugh, 2009; Wang and Donze, 2016). However, despite greater specificity in TSS selection in metazoans, such accessory factors have yet to be described in higher eukaryotes.

NF-Y, also known as the CCAAT-binding factor CBF, is a highly conserved and ubiquitously expressed heterotrimeric transcription factor (TF) composed of the NF-YA, NF-YB and NF-YC subunits, all three of which are necessary for stable DNA binding of the complex (Dolfini et al., 2012; Maity, 2017; Maity and de Crombrugghe, 1998; Nardini et al., 2013). NF-YA, which harbors both DNA-binding and transactivation domains, makes sequence-specific DNA contacts, whereas the histone-fold domain (HFD) containing NF-YB and NF-YC interact with DNA via non-specific HFD-DNA contacts (Nardini et al., 2013). The structure and DNA-binding mode of NF-YB/NF-YC HFDs are similar to those of the core histones H2A/H2B, TATA-binding protein (TBP)-associated factors (TAFs), the TBP/TATA-binding negative cofactor 2 (NC2α/β), and the CHRAC15/CHRAC17 subunits of the nucleosome remodeling complex CHRAC (Nardini et al., 2013).

NF-Y has an established role in gene regulation through cell type-invariant promoter-proximal binding (Fleming et al., 2013; Oldfield et al., 2014). However, the mechanisms through which NF-Y influences gene expression remain unclear. Several lines of evidence have pointed towards a possible role in the recruitment of chromatin modifiers and/or the PIC to the promoters of its target genes (for a review, see (Dolfini et al., 2012)). Previously, we reported a critical role for NF-Y in facilitating a permissive chromatin conformation at cell type-specific distal enhancers to enable master transcription factor binding (Oldfield et al., 2014). Here, we investigate the role NF-Y plays at gene promoters, what effect it may have on chromatin accessibility and recruitment of the transcription machinery, and how this might impact gene expression.

Through genome-wide studies in mouse embryonic stem cells (ESCs), we find that NF-Y is essential for the maintenance of the NDR at gene promoters. Depletion of the NF-YA protein leads to the accumulation of ectopic nucleosomes over the TSS, reducing promoter accessibility. Interestingly, under these conditions, we find that the PIC can relocate to a previously NF-Y-occupied upstream site, from where it commences ectopic transcription initiation. Remarkably, the resulting ectopic transcript can create novel mRNA isoforms and, in a large number of cases, lead to abnormal translation. Overall, we establish NF-Y as the first TF implicated in TSS selection in metazoans and demonstrate its importance in safeguarding the integrity of the NDR at gene promoters.

RESULTS

NF-Y promotes chromatin accessibility at gene promoters

Our previous characterization of NF-Y binding sites in ESCs revealed that a majority of NF-Y binding sites are located within 500 base pairs (bp) of annotated TSSs of protein-coding genes (Oldfield et al., 2014). Further analysis of our NF-Y ChIP-Seq data revealed a positional bias, with nearly all of these binding sites located immediately upstream (median distance 94 bp) of the TSS (**Figures 1A and 1B)** and a positive correlation between CCAAT motif occurrence and NF-YA occupancy (**Figures 1B, 1C, S1A, and S1B**). To investigate whether NF-Y promotes chromatin accessibility at proximal promoters, as it does at distal enhancers (Oldfield et al., 2014), we used small interfering RNA (siRNA) to knockdown (KD) *NF-YA*, the DNA-binding subunit, in ESCs (**Figures S1C and S1D**) and assessed DNase I hypersensitivity at candidate promoters either bound or not bound by NF-Y. Quantitative assessment of the relative “openness” of the probed regions revealed that depletion of NF-YA results in a significant reduction in DNA accessibility at promoters bound by NF-Y but not at promoters

without NF-Y binding (**Figures 1D, 1E, S1E and S1F**). These data suggest that NF-Y helps maintain accessible chromatin at promoters.

NF-Y binding protects NDRs from nucleosome encroachment

To further explore the role of NF-Y binding at promoters, we performed micrococcal nuclease (MNase) digestion followed by high-throughput sequencing (MNase-Seq). Our data confirmed that NF-Y binding sites genome-wide are depleted of nucleosomes, with a strong nucleosome phasing observed around NF-Y binding sites (**Figures 1F and S2A**). Focusing on NF-Y-bound promoter regions, we noted a well-positioned +1 nucleosome immediately downstream of TSSs and a strong anti-correlation between NF-Y occupancy and nucleosome depletion (**Figures 1B and 1F**). To determine whether NF-Y binding plays a direct role in the reduction of nucleosome occupancy at its binding sites, we performed MNase-Seq in NF-YA depleted cells.

Examination of the nucleosomal landscape at candidate NF-Y target promoters in NF-Y depleted cells revealed a striking gain of a nucleosome(s) precisely within what was previously a well-defined NDR (**Figure 1G**). This observation was confirmed by both MNase-qPCR and histone H3 ChIP-qPCR analyses (**Figures 1H and S2E**). Genome-wide, ectopic recruitment of nucleosomes upon *NF-YA* KD is observed specifically at NF-Y bound promoters (**Figures 1I and S2B**), but not at non-NF-Y bound promoters (**Figures S2C, and S2D)**, suggesting a direct effect. Additionally, our data indicate that the +1 nucleosome, although showing a general loss in recruitment intensity (**Figures 1G and 1I**), does not show a change in positioning upon *NF-YA* KD (**Figure S2F**), suggesting that the ectopic nucleosome

results either from a *de novo* recruitment or from the stabilization of an unstable nucleosome (Henikoff et al., 2009; Kubik et al., 2017; Weiner et al., 2010).

Interestingly, the peak of ectopic nucleosomes observed in NF-YA depleted cells overlaps the TSS of NF-Y bound genes (**Figure S2B**), suggesting that NF-Y could protect the TSS from inhibitory nucleosome binding. Moreover, this ectopic positioning correlates precisely with sequence-based predictions of nucleosome binding preference at these regions (Kaplan et al., 2009) (**Figure S2G**). Altogether, these results suggest that NF-Y binding at gene promoters plays a key role in protecting the NDR and the TSS from nucleosome encroachment and that in its absence, ectopic nucleosomes are able to bind within this region.

NF-Y binding impacts PIC positioning

Nucleosomes characteristically provide a refractory chromatin environment for the binding of TFs. In fact, the TATA-binding protein (TBP), and hence the entire Pol II transcription machinery, is unable to bind nucleosomal DNA (Imbalzano et al., 1994). Consequently, upon observing the appearance of ectopic nucleosome binding over the TSS and within the NDR of NF-Y bound promoters upon *NF-YA* KD, we sought to investigate whether this outcome affects binding of the transcription machinery.

To this end, we performed ChIP-exonuclease followed by high-throughput sequencing (ChIP-Exo) to precisely map Pol II binding in control and *NF-YA* KD cells. NF-YA depletion led to a stark loss of Pol II occupancy at its canonical binding sites and over the entire promoter region of NF-Y bound genes genome-wide (**Figures 2A and S3A**), even though global levels of

Pol II were not affected (**Figure S3B**). Notably, this reduction in Pol II occupancy was observed at the promoters of both the NF-Y-bound protein-coding genes and their cognate upstream anti-sense transcription (**Figure 2A**). Interestingly, however, we observed a new Pol II binding site immediately upstream of the annotated canonical TSS (**Figure 2A**).

Because NF-Y is known to interact with the Pol II-recruiting TBP (Bellorini et al., 1997) and several TBP-associated factors (TAFs) (Dolfini et al., 2012), essential components of the TFIID complex that recruits Pol II, we decided to investigate whether TBP enrichment was also affected upon *NF-YA* KD. Indeed, loss of NF-Y binding led to a significant diminishment and/or upstream shift of TBP's binding pattern (**Figure S3C**), a consequence that could have a dramatic impact on TSS localization. These results therefore confirm NF-Y's role in TFIID and Pol II recruitment and raise the possibility that NF-Y affects PIC positioning and thus TSS selection.

NF-Y binding directs TSS selection

Having established that NF-YA depletion leads to altered localization of the transcription pre-initiation complex, we performed RNA-seq to assess the impact on global gene expression. Loss of NF-YA led to both up- and down-regulation of mRNA levels, consistent with NF-Y having roles beyond that of a canonical transcription activator (**Figure S3D**). Interestingly, RNA-Seq experiments in NF-YA depleted cells also revealed ectopic RNA emanating from sites upstream of the canonical TSS (**Figures 2B and 2C**). Importantly, this same-strand, upstream transcription is specific to NF-Y bound promoters **(Figure 2C and 2D)** and predominantly takes place within the 500bp region upstream of the canonical TSS (**Figures 2C**).

In order to obtain a high-resolution view of Pol II activity and to map TSS utilization at base-pair resolution, we performed Start-Seq (Scruggs et al., 2015) a high-throughput sequencing method that captures capped RNA species from their 5'-ends. Start-Seq faithfully mapped the canonical TSSs in the control cells, confirming its utility in capturing transcription initiation sites (**Figure 2B**). More importantly, analysis of Start-Seq data in NF-YA depleted cells revealed clear upstream shifts in TSS usage specifically at NF-Y bound promoters (**Figures 2B, 2E, and 2F**). The differences in Start-Seq read counts upstream of the TSS between control and *NF-YA* KD cells directly correlate with NF-YA binding intensity (**Figure 2G**), suggesting a role for NF-Y in TSS selection. Indeed, comparison of NF-Y-bound promoters that exhibit TSS shifts upon *NF-YA* KD against those that do not revealed stronger NF-YA occupancy at genes with TSS shifts (**Figure S3E**). Moreover, the loss of NF-YA binding at these genes upon *NF-YA* KD results in substantial increase in nucleosome deposition over canonical TSSs compared to their non-shifting counterparts (**Figure S3F**).

All our *NF-YA* KD studies were performed 48 hr after siRNA transfection, when NF-YA is depleted but the cells appear normal with no obvious differentiation phenotype (Oldfield et al., 2014). But, to rule out the possibility that the observed changes are due to ESC differentiation, we assessed ectopic TSS usage in ESCs undergoing retinoic acid (RA)-induced differentiation. Four days of RA-induced differentiation did not elicit the use of ectopic TSSs (**Figure 2H**), indicating that the observed phenomenon is not due to global cellular differentiation effects that may occur upon *NF-YA* KD. Knockdown of *NF-YC*, or all 3 *NF-Y* subunits produced similar results (**Figure 2H**).

Surprisingly, NF-Y-mediated TSS selection is restricted to TSSs of protein-coding genes, as the TSSs of upstream and downstream antisense RNAs (Core et al., 2008; Preker et al., 2008; Seila et al., 2008) at NF-Y-bound promoters exhibit no such shift (median shifts of 0 bp; **Figure S3G**).

Intriguingly, RT-PCR experiments revealed that the transcription originating from ectopic TSSs generate stable multi-exonic transcripts (**Figures 2I**). This finding was confirmed using individual paired-end reads from the RNA-Seq data (**Figure 2I**). Remarkably, in certain cases, such as *Hsp90b1*, instead of simply extending the 5' end of the canonical mRNA, the ectopic transcript incorporates a new splicing donor site, resulting in transcripts that skip the canonical TSS, 5'UTR and first exon, giving rise to a new isoform (* product in **Figure 2I**). Collectively, these findings suggest that NF-Y binding impacts TSS selection at promoters of protein-coding genes, and that ectopic initiation creates aberrant mRNA species.

DNA sequence plays an important role in *NF-YA* KD-induced effects

To further our understanding of how ectopic nucleosome and TSS positioning are established, we examined changes in Start-Seq signal within a ±1 Kb window surrounding the canonical TSS. To ensure that only significantly shifted TSSs were called, we used a stringent criterion to exclude transcription initiation events within ±25bp of the canonical TSS. Of the 3,056 genes whose promoter is bound by NF-Y, 538 exhibit ectopic TSS usage (**Figure 3A**), including 462 with a TSS shift upstream of the canonical TSS. Ectopic TSSs are located at a median distance of 115bp upstream of the canonical TSS (**Figure 3B**).

In some cases (~22% of NF-Y-bound and TSS-shifted genes), such as with the *Ezh2* gene, the ectopic TSS used in response to *NF-YA* KD corresponds to a previously described alternative TSS (**Figure 3C**), indicating that DNA sequence-based elements play a role in defining sites of ectopic transcription initiation. To expand this observation, we examined the underlying DNA sequence at TSS-shifted promoters. *De novo* motif analysis (±5 bp) surrounding the canonical and ectopic TSSs, as determined using Start-Seq, revealed the typical YR initiator dinucleotide (Carninci et al., 2006), indicative of a sequence recognition mechanism determining the location of ectopic TSSs (**Figure 3D**). The canonical transcription initiation site at the TSS of mammalian mRNA transcripts is characterised either by a full Inr motif (YYANWYY or $^{C}/_{T}$ $^{C}/_{T}$ A N $^{A}/_{T}$ $^{C}/_{T}$ $^{C}/_{T}$) (Bucher, 1990) or by the YR ($^{C}/_{T}$ $^{A}/_{G}$) consensus. Given the similarity between the mammalian Inr sequence and NF-Y's CCAAT binding motif, it is conceivable that in the absence of NF-Y, the PIC may mistake the CCAAT motif for an Inr sequence. To investigate this possibility, we performed a *de novo* motif search around the ectopic TSS sites (±25 bp) and found that the CCAAT motif was by far the most enriched (P-value = 3.0e-341; 2nd was the GC-box, P-value = 2.5e-036). We next investigated the distance between the ectopic TSS and its nearest CCAAT motif and found that the median distance between these two features is 24.5bp (**Figure 3E**). These analyses indicate that at a specific subset of promoters, the displaced PIC can indeed mistake the vacated NF-Y binding site for a transcription initiation site (**Figure S3H**), supporting the conclusion that NF-Y binding prevents ectopic initiation from upstream loci.

Next, we investigated sequence conservation of the regions containing ectopic TSSs. Comparison of PhyloP conservation scores, calculated from genomic sequence alignment across placental mammals (Siepel et al., 2005), of NF-Y bound promoters revealed

significantly higher conservation of the region upstream of the canonical TSS compared to that of an equivalent number of randomly chosen gene promoters (**Figure 3F**), indicating a possible bias in sequence conservation of the promoters targeted by NF-Y. The conservation at NF-Y bound and TSS-shifted promoters is even higher than that at NF-Y bound promoters in general (**Figure 3F**), along with a higher enrichment of CCAAT motifs (**Figure 3G**) and NF-Y occupancy (**Figure S3E**), possibly revealing the importance of NF-Y binding to these promoter sequences to facilitate selection of the correct TSS. Focusing our analysis specifically on ectopic TSSs, we were able to confirm a high degree of sequence conservation starting at the ectopic TSS and continuing downstream towards the canonical TSS within these traditionally non-coding regions (**Figure 3H**). Given the cross-species conservation of the ectopic TSSs, the sequence and structural conservation of NF-Y (Nardini et al., 2013), and the cell type-invariant promoter-proximal binding of NF-Y in mouse (Oldfield et al., 2014) and human cells (**Figure S3I**), we propose that NF-Y's role in the organization of the NDR and in TSS selection is likely conserved in other species.

RNAs from ectopic TSSs in NF-Y depleted cells undergo translation

Considering the importance of the 5'UTR in the regulation of translation (Dikstein, 2012), and the fact that close to 70% of genes showing a TSS shift upon *NF-YA* KD possess an AUG (translation start codon) within the ectopically transcribed region (**Table S1**), we explored the potential impact of these aberrant transcripts on translation output. To do so, we performed Ribo-Seq, a ribosome-profiling experiment (Ingolia et al., 2009), on control and NF-YA depleted cells. By sequencing only the ribosome-protected fraction of the transcriptome, Ribo-Seq allows us to determine which RNAs are being actively translated at a given time. Typical of Ribo-Seq, triplet phasing was observed beginning at the annotated

translation start site (**Figure S4A**), whereas RNA-Seq presented a flat, uniform distribution (**Figure S4A**). To determine if the transcripts originating from ectopic TSSs in *NF-YA* KD cells are also undergoing translation, we investigated the differences in Ribo-Seq read coverage within the region between the canonical and the ectopic TSSs. At the individual gene level, we can clearly detect the ribosome-protected RNA originating from the region upstream of canonical TSSs of NF-Y bound genes that exhibit ectopic transcription initiation (**Figures 4A and S4B**). Furthermore, of the 429 NF-Y bound genes with ectopic TSSs that had sufficient Ribo-Seq coverage between the canonical and the ectopic TSSs, 92% showed significantly higher levels of ribosome-protected RNA in *NF-YA* KD cells compared to control KD cells (**Figures 4B and 4C**). To ensure that the ribosome-protected RNA, transcribed from the region between the ectopic TSS and canonical TSS, is undergoing translation and is not an artefact, we re-analysed Ribo-Seq data for read coverage phasing after individually determining which ectopically transcribed AUG was most likely to be used as a translation start site for each gene. We found a significant enrichment of triplet periodicity in the Ribo-Seq read coverage beginning at the putative ectopic translation start site, in NF-YA KD cells compared to control cells (**Figure 4D**), indicating that the ectopically transcribed regions indeed undergo translation.

In an attempt to evaluate whether these translated upstream open reading frames (ORFs) generate fusion or variant forms of endogenous protein, we performed western-blot analysis, using commercially available antibodies for candidate proteins, in cells depleted of NF-Y and treated with the proteasome inhibitor MG-132 (to stabilize any unstable fusion proteins, **Figure S4C and S4D**). Using these reagents, we did not detect any obvious aberrant fusion/variant protein of distinct molecular weight, suggesting that, for the

candidates tested, the ectopic ORFs do not generally alter canonical protein composition. Alternatively, we cannot rule out that truncated protein products are unrecognizable by the antibodies used. Nevertheless, these experiments revealed something interesting: for a number of NF-Y bound and TSS-shifted genes, there is a noticeable discordance between the manners with which the RNA and protein levels change in response to *NF-YA* KD (**Figure S4D**), which was surprising considering that changes in transcript levels generally result in proportional changes in protein abundance (Spearman correlation of ~0.8) (Jovanovic et al., 2015; Liu et al., 2017). Recent studies have shown that upstream ORFs (uORFs) can have either a favourable or a deleterious effect on downstream mRNA translation (Chen et al., 2017; Cheng et al., 2018; Chia et al., 2017; Johnstone et al., 2016; Wethmar, 2014) and suggested that the abundance of transcript is less important than the difference in translatability of the canonical versus ectopic transcript (Cheng et al., 2018). Thus, in situations wherein an upstream extension of the transcript leads to altered protein production, we suggest that this could be due to introduction of an uORF. To summarize, we establish here that transcripts originating from ectopic TSSs undergo translation, and highlight the variable effects this can have on the translation of the canonical ORF within these transcripts.

DISCUSSION

Driven by the prevalence of the CCAAT motif(s) within core promoters, NF-Y's function as a regulator of gene expression has almost exclusively been studied in relation to its promoter-proximal binding. Yet, the exact mechanism by which it exerts control over gene expression remains poorly understood. Through comprehensive genome-wide studies in ESCs, we have

uncovered a previously unidentified role for NF-Y in safeguarding the integrity of the NDR structure, PIC localization, and TSS selection at protein-coding genes.

NF-Y binding is incompatible with nucleosome presence

NF-Y can access its target DNA motif, the CCAAT box, in a heterochromatic environment (Fleming et al., 2013). Furthermore, NF-Y's unique DNA-binding mode, which induces an ~80° bend in the DNA, may allow and/or promote binding of other TFs, whose recognition sequences become more accessible (Oldfield et al., 2014). Supporting this thesis, DNase experiments have shown that NF-Y is essential for the maintenance of an accessible chromatin (Lu et al., 2016; Oldfield et al., 2014; Sherwood et al., 2014). These attributes have led us and others to propose that NF-Y is a "pioneer factor" (Fleming et al., 2013; Lu et al., 2016; Nardini et al., 2013; Oldfield et al., 2014; Sherwood et al., 2014; Tao et al., 2017).

Through comparison of NF-YA ChIP-Seq and MNase-Seq data, we have shown mutual exclusivity between NF-Y and nucleosome occupancy genome-wide (**Figures 1B, 1F, and S2A**). Given that the structure and DNA-binding mode of NF-YB/NF-YC HFDs are similar to those of the core histones H2A/H2B (Liberati et al., 1999; Nardini et al., 2013; Romier et al., 2003) our findings suggest steric incompatibility between NF-Y and nucleosomes. This conclusion is supported by the observation that upon *NF-YA* KD, nucleosomes bind within the NDRs left vacant by NF-Y, positioning them in a manner that strongly reflects DNA sequence preferences (**Figure S2G**).

NF-Y safeguards the integrity of the NDR

The presence of a well-defined NDR within active gene promoters is essential for access by GTFs and PIC assembly, and thus correct transcription initiation. While NF-Y binding does not seem to impact positioning of the +1 or -1 nucleosomes that demarcate the NDR, we find that NF-Y is essential for maintaining a nucleosome-depleted NDR. Although we cannot rule out the possibility that NF-Y recruits an ATP-dependent chromatin remodeler to orchestrate nucleosome removal, given NF-Y's capacity to disrupt the compaction of the chromatin, it is tempting to speculate that NF-Y, with its sequence-specific binding ability, could be acting as an ATP-independent chromatin remodeler. This idea is supported by previous findings that show NF-Y's capacity to displace nucleosomes in an *in vitro* context (Coustry et al., 2001; Motta et al., 1999). Overall, it seems that proteins with tertiary structures similar to core histones can independently preclude nucleosome occupancy. It will be interesting to see if all proteins containing histone-fold domains have a similar effect on nucleosome binding, as has been shown to be the case with subunits of the CHRAC complex (Kukimoto et al., 2004).

NF-Y's direct and indirect role in PIC positioning

NF-Y has been shown to play a role in the recruitment of the pre-initiation complex through direct interactions with TBP and several TAFs (Bellorini et al., 1997; Frontini et al., 2002). We have shown here that NF-Y also plays an indirect role in the recruitment of PIC-associated proteins, since its binding to promoter-proximal regions is necessary for the maintenance of an open chromatin structure over the TSS, allowing Pol II and TBP to effectively bind. Both these direct and indirect mechanisms are likely responsible for the global loss of Pol II and TBP binding observed at NF-Y bound promoters upon *NF-YA* KD.

In the case of NF-Y's indirect impact on PIC recruitment, we observed an associated upstream shift in TSS location upon *NF-YA* KD. Yet, it is interesting to note that only a portion of NF-Y bound genes exhibit this TSS shift. Besides the strength and stability of NF-Y binding and the number of binding events, this likely reflects involvement of additional factors. Firstly, the distance between the vacated NF-Y binding site and the canonical TSS will impact where an ectopic nucleosome can assemble within the NDR. Indeed, for NF-Y binding sites located several hundred base pairs upstream of the TSS (as in the case of sites shown at the top of the heatmap in **Figure 1B**), the ectopic binding of a nucleosome at this distant region is unlikely to affect PIC recruitment at the TSS, hence the diminished amount of ectopic RNA observed at or near these distal sites (top portion of **Figure 2C**). Secondly, efficient utilization of an ectopic TSS requires sequence features within the exposed DNA that are amenable to proper PIC binding, such as previously described alternative start sites (as is the case for *Ezh2*; **Figure 3C**) or initiator motifs. Indeed, despite our discovery that a majority of the ectopic TSSs are observed to occur at or very near vacated CCAAT boxes (within 25 bp), this is not a strict rule and some promoters might not have secondary sites with high enough affinity to recruit the transcription machinery.

NF-Y plays an important role in safeguarding transcriptome identity

Our findings establish NF-Y as the first TF implicated in TSS selection in metazoans. NF-Y binding at promoters seems to have at least three roles: (1) direct PIC recruitment to the promoter region through its interactions with TBP and the TAFs, (2) prevent ectopic nucleosome binding within the NDR through its nucleosome-like structural properties and DNA-binding mode, and (3) occlude alternative transcription initiation sites to ensure correct TSS usage. The shift in TSS usage upon *NF-YA* KD appears to stem from the two latter roles,

whereby, upon loss of NF-Y binding, a potential transcription initiation site is uncovered while, simultaneously, a nucleosome prohibits optimal PIC binding to the canonical TSS, forcing the PIC to relocate.

As might be expected, the usage of an ectopic, upstream TSS has variable consequences on the steady-state levels of resulting mRNAs and protein products. The strength of the ectopic initiation site along with the regulatory potential of the additional upstream mRNA sequences (region between the ectopic and the canonical TSS) undoubtedly impact transcription levels and mRNA stability. Additionally, the TSS employed in NF-YA depleted cells could represent a known alternative start site (as in the case of *Ezh2*, **Figure 3C**), or generate a previously uncharacterized mRNA isoform (as with *Hsp90b1*, **Figure 2I**). Furthermore, the upstream extension of mRNA to include a novel ORF can cause abnormal translation (as with *Khsrp* and *C7orf50*, **Figure 4A and S4B**, respectively). All of these sequelae can affect the quantity and quality of mRNAs and proteins, which could have significant, yet unpredictable, consequences on cell survival or function.

Notably, NF-Y predominantly binds the CG-rich promoters of essential genes (cell-cycle, transcription, DNA repair, etc.), whose accurate expression is vital in most cell types (Fleming et al., 2013; Oldfield et al., 2014). In fact, ~80% of all promoter-proximal NF-Y binding sites overlap CpG islands (Oldfield et al., 2014). Unlike CG-poor promoters, which often correspond to tissue-specific genes and initiate transcription from a well-defined site, CG-rich promoters contain a broad array of transcription initiation sites and often associate with housekeeping genes (Carninci et al., 2006; Rach et al., 2011; Valen and Sandelin, 2011). The

requirement for NF-Y at promoters of essential genes could reflect both the role of NF-Y in PIC recruitment, and enforcement of appropriate TSS usage.

We suggest that this dual role for NF-Y may explain why promoter-proximal NF-Y binding is so well conserved across mouse and human cell types. Intriguingly, studies in *Saccharomyces pombe* have shown that deletion of *Php5*, an NF-YC orthologue, leads to a ~250bp upstream shift in TSS of the gluconeogenesis gene *Fbp1* (Asada et al., 2015). Moreover, in *Saccharomyces cervisiae*, out of the 46 TFs studied, the binding sites for NF-YA homolog Hap2 were shown to have the biggest difference in predicted nucleosome occupancy between Hap2-bound (lower nucleosome occupancy) and non-Hap2 bound (higher occupancy) sites (Segal et al., 2006). This opens the exciting possibility of NF-Y's role in promoter chromatin organization being conserved throughout the eukaryotic kingdom.

Effects of uORFs within ectopic transcripts on protein expression

A consequence of altered TSS selection upon *NF-YA* KD is that ribosomes can scan over any ectopic mRNA. Ribosomes typically initiate translation upon encountering the AUG start codon, although other codons have been shown to induce translation initiation (Dever, 2012; Lobanov et al., 2010; Starck et al., 2012). In our study, we found that nearly three quarters of genes showing a TSS shift upon *NF-YA* KD possess at least one ATG triplet between the ectopic and the canonical TSS. Importantly, we found evidence for translation initiation from such sites (**Figure 4D**). Given that uORFs can modulate downstream translation and thus act as potent regulators of translation and protein expression (Calvo et al., 2009; Chen et al., 2017; Cheng et al., 2018; Chia et al., 2017; Johnstone et al., 2016), it is conceivable that translation initiation from non-canonical start codon(s) within uORFs alters

the reading frame and/or protein length; alternatively, it may affect the efficiency with which ribosomes translate the rest of the transcript (Vilela and McCarthy, 2003).

In summary, our studies describe NF-Y's mechanistic role at promoters, where it is necessary for both maintenance of the NDR's structural architecture and correct positioning of the transcriptional machinery, therefore influencing TSS selection. Furthermore, our results strongly suggest that the sites of NF-Y binding and the +1 nucleosome demarcate the 5' and 3' boundaries, respectively, of the region available for PIC assembly, thereby directing the transcription machinery to the correct TSS while occluding alternative TSSs and other sites of sub-optimal transcription initiation. It will be interesting to explore whether other histone-fold domain proteins, with similar structural and DNA-binding properties analogous to NF-Y, may function in a similar manner.

SUPPLEMENTAL INFORMATION

Supplemental information includes four figures and two tables.

AUTHOR CONTRIBUTIONS

A.J.O., K.A., and R.J. conceived the study. A.J.O., T.H., and S.C. performed the experiments. A.J.O., A.B.B., D.P., P.Y., B.S.S., C.A.L., D.K., B.B., and E.R. analyzed the data. A.J.O., T.H., K.A., and R.J. wrote the manuscript. All authors reviewed, edited, and approved the final version of the manuscript.

ACKNOWLEDGEMENTS

We thank R. Luco, E. Nora, and P. Navarro for critical comments on the manuscript. We thank A.E. Conway and D. Fargo for useful discussion, J.P. Villemin and E. Agirre for guidance on bioinformatic analyses and F. Aubignat for assistance with figure illustration. We thank the NIEHS Epigenomics core, especially J. Malphurs, for their support. D.P. and E.R. thank funding from the Institut de Biologie Computationnelle (ANR-11-BINF-0002) Institut Français de Bioinformatique (ANR-11-INBS-0013), and France Génomique D.P. and E.R. also thank the ATGC bioinformatics platform in Montpellier. This work was supported by the Intramural Research Program of the NIH, National Institutes of Environmental Health Sciences (R.J.: 1Z1AES102625, K.A.: 1ZIAES101987).

CONFLICT OF INTEREST

The authors report they have no conflicts of interest to declare.

REFERENCES

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol *11*, R106.

Asada, R., Takemata, N., Hoffman, C.S., Ohta, K., and Hirota, K. (2015). Antagonistic controls of chromatin and mRNA start site selection by Tup family corepressors and the CCAAT-binding factor. Mol Cell Biol *35*, 847-855.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res *37*, W202-208.

Bellorini, M., Lee, D.K., Dantonel, J.C., Zemzoumi, K., Roeder, R.G., Tora, L., and Mantovani, R. (1997). CCAAT binding NF-Y-TBP interactions: NF-YB and NF-YC require short domains adjacent to their histone fold motifs for association with TBP basic residues. Nucleic Acids Res *25*, 2174-2181.

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.

Calvo, S.E., Pagliarini, D.J., and Mootha, V.K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci U S A *106*, 7507-7512.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A., Taylor, M.S., Engstrom, P.G., Frith, M.C.*, et al.* (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet *38*, 626-635.

Chen, J., Tresenrider, A., Chia, M., McSwiggen, D.T., Spedale, G., Jorgensen, V., Liao, H., van Werven, F.J., and Unal, E. (2017). Kinetochore inactivation by expression of a repressive mRNA. Elife *6*.

Cheng, Z., Otto, G.M., Powers, E.N., Keskin, A., Mertins, P., Carr, S.A., Jovanovic, M., and Brar, G.A. (2018). Pervasive, Coordinated Protein-Level Changes Driven by Transcript Isoform Switching during Meiosis. Cell *172*, 910-923 e916.

Chia, M., Tresenrider, A., Chen, J., Spedale, G., Jorgensen, V., Unal, E., and van Werven, F.J. (2017). Transcription of a 5' extended mRNA isoform directs dynamic chromatin changes and interference of a downstream promoter. Elife *6*.

Cinghu, S., Yang, P., Kosak, J.P., Conway, A.E., Kumar, D., Oldfield, A.J., Adelman, K., and Jothi, R. (2017). Intragenic Enhancers Attenuate Host Gene Expression. Mol Cell *68*, 104-117 e106.

Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A., and Lis, J.T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat Genet *46*, 1311-1320.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845-1848.

Coustry, F., Hu, Q., de Crombrugghe, B., and Maity, S.N. (2001). CBF/NF-Y functions both in nucleosomal disruption and transcription activation of the chromatin-assembled topoisomerase IIalpha promoter. Transcription activation by CBF/NF-Y in chromatin is dependent on the promoter structure. J Biol Chem *276*, 40621-40630.

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

Dever, T.E. (2012). Molecular biology. A new start for protein synthesis. Science *336*, 1645-1646.

Dikstein, R. (2012). Transcription and translation in a package deal: the TISU paradigm. Gene *491*, 1-4.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Dolfini, D., Gatta, R., and Mantovani, R. (2012). NF-Y and the transcriptional activation of CCAAT promoters. Crit Rev Biochem Mol Biol *47*, 29-49.

Epstein, W., and Beckwith, J.R. (1968). Regulation of gene expression. Ann. Rev. Biochem. *37*, 411-436.

Fleming, J.D., Pavesi, G., Benatti, P., Imbriano, C., Mantovani, R., and Struhl, K. (2013). NF-Y coassociates with FOS at promoters, enhancers, repetitive elements, and inactive chromatin regions, and is stereo-positioned with growth-controlling transcription factors. Genome Res *23*, 1195-1209.

Frontini, M., Imbriano, C., diSilvio, A., Bell, B., Bogni, A., Romier, C., Moras, D., Tora, L., Davidson, I., and Mantovani, R. (2002). NF-Y recruitment of TFIID, multiple interactions with histone fold TAF(II)s. J Biol Chem *277*, 5841-5848.

Ghazy, M.A., Brodie, S.A., Ammerman, M.L., Ziegler, L.M., and Ponticelli, A.S. (2004). Amino acid substitutions in yeast TFIIF confer upstream shifts in transcription initiation and altered interaction with RNA polymerase II. Mol Cell Biol *24*, 10975-10985.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell *38*, 576-589.

Henikoff, S., Henikoff, J.G., Sakai, A., Loeb, G.B., and Ahmad, K. (2009). Genome-wide profiling of salt fractions maps physical properties of chromatin. Genome Res *19*, 460-469.

Henriques, T., Gilchrist, D.A., Nechaev, S., Bern, M., Muse, G.W., Burkholder, A., Fargo, D.C., and Adelman, K. (2013). Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals. Mol Cell *52*, 517-528.

Imbalzano, A.N., Kwon, H., Green, M.R., and Kingston, R.E. (1994). Facilitated binding of TATA-binding protein to nucleosomal DNA. Nature *370*, 481-485.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218-223.

Johnstone, T.G., Bazzini, A.A., and Giraldez, A.J. (2016). Upstream ORFs are prevalent translational repressors in vertebrates. EMBO J *35*, 706-723.

Jovanovic, M., Rooney, M.S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., Rodriguez, E.H., Fields, A.P., Schwartz, S., Raychowdhury, R.*, et al.* (2015). Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. Science *347*, 1259038.

Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J.*, et al.* (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. Nature *458*, 362-366.

Kasahara, K., Ohyama, Y., and Kokubo, T. (2011). Hmo1 directs pre-initiation complex assembly to an appropriate site on its target gene promoters by masking a nucleosome-free region. Nucleic Acids Res *39*, 4136-4150.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol *14*, R36.

Kubik, S., Bruzzone, M.J., and Shore, D. (2017). Establishing nucleosome architecture and stability at promoters: Roles of pioneer transcription factors and the RSC chromatin remodeler. Bioessays *39*.

Kukimoto, I., Elderkin, S., Grimaldi, M., Oelgeschlager, T., and Varga-Weisz, P.D. (2004). The histone-fold protein complex CHRAC-15/17 enhances nucleosome sliding and assembly mediated by ACF. Mol Cell *13*, 265-277.

Lai, W.K.M., and Pugh, B.F. (2017). Understanding nucleosome dynamics and their links to gene expression and DNA replication. Nat Rev Mol Cell Biol *18*, 548-562.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Liberati, C., di Silvio, A., Ottolenghi, S., and Mantovani, R. (1999). NF-Y binding to twin CCAAT boxes: role of Q-rich domains and histone fold helices. J Mol Biol *285*, 1441-1455.

Liu, Y., Gonzalez-Porta, M., Santos, S., Brazma, A., Marioni, J.C., Aebersold, R., Venkitaraman, A.R., and Wickramasinghe, V.O. (2017). Impact of Alternative Splicing on the Human Proteome. Cell Rep *20*, 1229-1241.

Lobanov, A.V., Turanov, A.A., Hatfield, D.L., and Gladyshev, V.N. (2010). Dual functions of codons in the genetic code. Crit Rev Biochem Mol Biol *45*, 257-265.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550.

Lu, F., Liu, Y., Inoue, A., Suzuki, T., Zhao, K., and Zhang, Y. (2016). Establishing Chromatin Regulatory Landscape during Mouse Preimplantation Development. Cell *165*, 1375-1388.

Luna-Zurita, L., Stirnimann, C.U., Glatt, S., Kaynak, B.L., Thomas, S., Baudin, F., Samee, M.A., He, D., Small, E.M., Mileikovsky, M.*, et al.* (2016). Complex Interdependence Regulates Heterotypic Transcription Factor Distribution and Coordinates Cardiogenesis. Cell *164*, 999-1014.

Maity, S.N. (2017). NF-Y (CBF) regulation in specific cell types and mouse models. Biochim Biophys Acta *1860*, 598-603.

Maity, S.N., and de Crombrugghe, B. (1998). Role of the CCAAT-binding protein CBF/NF-Y in transcription. Trends Biochem Sci *23*, 174-178.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011 *17*.

Motta, M.C., Caretti, G., Badaracco, G.F., and Mantovani, R. (1999). Interactions of the CCAAT-binding trimer NF-Y with nucleosomes. J Biol Chem *274*, 1326-1333.

Muller, F., and Tora, L. (2014). Chromatin and DNA sequences in defining promoters for transcription initiation. Biochim Biophys Acta *1839*, 118-128.

Nardini, M., Gnesutta, N., Donati, G., Gatta, R., Forni, C., Fossati, A., Vonrhein, C., Moras, D., Romier, C., Bolognesi, M.*, et al.* (2013). Sequence-specific transcription factor NF-Y displays histone-like DNA binding and H2B-like ubiquitination. Cell *152*, 132-143.

Nechaev, S., Fargo, D.C., dos Santos, G., Liu, L., Gao, Y., and Adelman, K. (2010). Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. Science *327*, 335-338.

Oldfield, A.J., Yang, P., Conway, A.E., Cinghu, S., Freudenberg, J.M., Yellaboina, S., and Jothi, R. (2014). Histone-fold domain protein NF-Y promotes chromatin accessibility for cell type-specific master transcription factors. Mol Cell *55*, 708-722.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2016). Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference. bioRxiv.

Pinto, I., Wu, W.H., Na, J.G., and Hampsey, M. (1994). Characterization of sua7 mutations defines a domain of TFIIB involved in transcription start site selection in yeast. J Biol Chem *269*, 30569-30573.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res *20*, 110-121.

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. Science *322*, 1851-1854.

Rach, E.A., Winter, D.R., Benjamin, A.M., Corcoran, D.L., Ni, T., Zhu, J., and Ohler, U. (2011). Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. PLoS Genet *7*, e1001274.

Rhee, H.S., and Pugh, B.F. (2012). Genome-wide structure and organization of eukaryotic pre-initiation complexes. Nature *483*, 295-301.

Romier, C., Cocchiarella, F., Mantovani, R., and Moras, D. (2003). The NF-YB/NF-YC structure gives insight into DNA binding and transcription regulation by CCAAT factor NF-Y. J Biol Chem *278*, 1336-1345.

Roy, A.L., and Singer, D.S. (2015). Core promoters in transcription: old problem, new insights. Trends Biochem Sci *40*, 165-171.

Scruggs, B.S., Gilchrist, D.A., Nechaev, S., Muse, G.W., Burkholder, A., Fargo, D.C., and Adelman, K. (2015). Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. Mol Cell *58*, 1101-1112.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I.K., Wang, J.P., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772-778.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent transcription from active promoters. Science *322*, 1849-1851.

Sherwood, R.I., Hashimoto, T., O'Donnell, C.W., Lewis, S., Barkal, A.A., van Hoff, J.P., Karun, V., Jaakkola, T., and Gifford, D.K. (2014). Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nat Biotechnol *32*, 171-178.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S.*, et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res *15*, 1034-1050.

Smale, S.T., and Kadonaga, J.T. (2003). The RNA polymerase II core promoter. Annu Rev Biochem *72*, 449-479.

Starck, S.R., Jiang, V., Pavon-Eternod, M., Prasad, S., McCarthy, B., Pan, T., and Shastri, N. (2012). Leucine-tRNA initiates at CUG start codons for protein synthesis and presentation by MHC class I. Science *336*, 1719-1723.

Tao, Z., Shen, L., Gu, X., Wang, Y., Yu, H., and He, Y. (2017). Embryonic epigenetic reprogramming by a pioneer transcription factor in plants. Nature *551*, 124-128.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc *7*, 562-578.

Valen, E., and Sandelin, A. (2011). Genomic and chromatin signals underlying transcription start-site selection. Trends Genet *27*, 475-485.

Venters, B.J., and Pugh, B.F. (2009). A canonical promoter organization of the transcription machinery and its regulators in the Saccharomyces genome. Genome Res *19*, 360-371.

Vilela, C., and McCarthy, J.E. (2003). Regulation of fungal gene expression via short open reading frames in the mRNA 5'untranslated region. Mol Microbiol *49*, 859-867.

Vo Ngoc, L., Wang, Y.L., Kassavetis, G.A., and Kadonaga, J.T. (2017). dThe punctilious RNA polymerase II core promoter. Genes Dev *31*, 1289-1301.

Wang, Q., and Donze, D. (2016). Transcription factor Reb1 is required for proper transcriptional start site usage at the divergently transcribed TFC6-ESC2 locus in Saccharomyces cerevisiae. Gene.

Weiner, A., Hughes, A., Yassour, M., Rando, O.J., and Friedman, N. (2010). High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. Genome Res *20*, 90-100.

Wethmar, K. (2014). The regulatory potential of upstream open reading frames in eukaryotic gene expression. Wiley Interdiscip Rev RNA *5*, 765-778.

Williams, L.H., Fromm, G., Gokey, N.G., Henriques, T., Muse, G.W., Burkholder, A., Fargo, D.C., Hu, G., and Adelman, K. (2015). Pausing of RNA polymerase II regulates mammalian developmental potential through control of signaling networks. Mol Cell *58*, 311-322.

FIGURE LEGENDS

Figure 1. NF-Y binding at gene promoters protects the nucleosome-depleted region (NDR) from nucleosome encroachment

(A) Average NF-YA occupancy (y-axis), as measured using ChIP-Seq, near transcription start sites (TSSs) of genes with (black, *n*=3,191) or without (grey, *n*=21,195) promoter-proximal NF-Y binding in mouse ESCs. RPM, reads per million mapped reads.

(B) NF-YA occupancy near TSSs of genes with promoter-proximal NF-Y binding.

(C) CCAAT motif occurrence near TSSs of genes with promoter-proximal NF-Y binding.

(D, E) DNase I hypersensitivity and qPCR analysis of promoters with (D; *n* = 6) or without (E; *n* = 4) NF-Y binding in control (blue) and *NF-YA* KD (red) ESCs. Error bars, SEM. Data for individual promoters shown in Figure S1E and S1F.

(F) Nucleosome occupancy, as measured using MNase-Seq, near TSSs of genes with promoter-proximal NF-Y binding in control ESCs.

(G) Genome browser shots of candidate NF-Y target genes showing nucleosome occupancy in control (blue) or *NF-YA* KD (red) ESCs. Gene structure is shown at the bottom. Arrows highlight nucleosomal gain in *NF-YA* KD cells over what was previously a well-defined NDR in control KD cells.

(H) MNase-qPCR validation of MNase-Seq data shown in Figure 1H. Error bars, SEM of three biological replicates. *P-value < 0.04 (Student's t-test, two-sided)

(I) Relative change in nucleosome occupancy (red, gain; blue, loss) near TSSs of genes with promoter-proximal NF-Y binding in *NF-YA* KD vs control KD cells.

See also Figures S1 and S2.

Figure 2. NF-Y binding influences PIC positioning and TSS selection

(A) Genome browser shots of NF-Y target genes showing RNA Polymerase II (Pol II) occupancy, as measured using ChIP-Exo, in control (blue) or *NF-YA* KD (red) ESCs. Gene structure is shown at the bottom. Arrows highlight regions with Pol II binding in *NF-YA* KD cells but not in control KD cells.

(B) Genome browser shots of NF-Y target genes showing gene expression (RNA-Seq) and transcription initiation-associated RNA enrichment (Start-Seq) in control and *NF-YA* KD ESCs. ChIP-Seq track showing NF-YA occupancy in ESCs is shown at the top. Arrows highlight regions with ectopic RNA or transcription initiation in *NF-YA* KD cells. Gene structure is shown at the bottom along with PCR amplicons used for RT-qPCR analysis in Figure 2I.

(C) Relative change in RNA-Seq signal (red, gain; blue, loss) near TSSs of genes with NF-Y binding (*n* = 3,056) in *NF-YA* KD vs control KD cells.

(D) Box plot showing relative change in RNA-Seq signal (in *NF-YA* KD vs control KD ESCs) within promoters (1 Kb upstream of TSS) of genes with (black) or without (gray) NF-Y binding. *P-value = 6.45E-14 (Wilcoxon rank-sum test, two-sided)

(E) Relative change in Start-Seq signal (red, gain; blue, loss) near TSSs of genes with NF-Y binding (*n* = 3,056) in *NF-YA* KD vs control KD cells.

(F) Box plot showing relative change in RNA-Seq signal (in *NF-YA* KD vs control KD ESCs) within promoters (-1 Kb to -25 bp) of genes with (black) or without (gray) NF-Y binding. *P-value = 4.76E-26 (Wilcoxon rank-sum test, two-sided)

(G) Box plot showing maximum difference observed in Start-Seq read count, between control and *NF-YA* KD cells, within the region upstream of TSS (-900 to -25bp). Read count differences using a 10 bp sliding window. Genes with NF-Y binding were binned

into six groups based on NF-YA ChIP-Seq read count within -900 to +100 bp of TSS. *P-value < 0.05, **P-value < 0.0009, ***P-value = 4.47E-07, #P-value = 1.16E-18, ##P-value=2.49E-18 (Wilcoxon rank-sum test, two-sided)

(H) Genome browser shot showing RNA-Seq signal at the *Hsp90b1* gene in control (blue) or *NF-YA* KD (red) ESCs. Arrow highlights region with ectopic RNA in *NF-YA* KD cells. A representative selection of individual RNA-Seq reads is shown beneath. Red and blue rectangles highlight the ectopic and endogenous splice sites, respectively. Schematic of the RT-PCR results shown at the bottom represent the different PCR amplification products shown in the right panel. PCR amplification was performed using the "ectopic" primer pair shown underneath the gene structure. * denotes new isoform fragment (bypassing the canonical 5'UTR, TSS and the 1st exon), ** denotes mRNA fragment with prolonged 5' fragment (uses the canonical 1st exon), *** denotes pre-mRNA fragment, and # denotes non-specific/unknown fragment.

(I) RT-qPCR analysis of relative gene expression using the "total" and "ectopic" primer pairs shown in Figure 2B. Data normalized to *Actin*, *HAZ* and *TBP*. Error bars, SEM of three to five biological replicates.

See also Figure S3.

Figure 3. Characterization of ectopic TSSs

(A) Relative change in Start-Seq signal (red, gain; blue, loss) near TSSs of genes with promoter-proximal NF-Y binding that exhibit an ectopic TSS (TSS shift, *n* = 538) in *NF-YA* KD vs control KD cells.

(B) Violin plot showing the distribution of TSS shift distance, as measured using Start-Seq, in *NF-YA* KD vs control KD ESCs. Median shift distance = 115bp.

(C) Genome browser shot of RNA-Seq signal at the *Ezh2* gene showing an isoform switch in *NF-YA* KD (red) vs control (blue) ESCs. A representative selection of individual RNA-Seq reads is shown beneath. Red and blue rectangles highlight the ectopic and endogenous splice sites, respectively.

(D) Consensus sequence motif enriched within the 10 bp region centered on canonical TSSs (top) or ectopic TSSs (bottom) of genes with promoter-proximal NF-Y binding, derived from *de novo* motif search.

(E) Frequency distribution of the distance between ectopic TSS and the nearest CCAAT motif of expressed NF-Y bound genes (*n* = 538; first 100bp are shown, representing ~75% of events).

(F) *Left*: Average sequence conservation in mammals, as computed using PhyloP (Pollard et al., 2010), of the 2 Kb region centered on canonical TSS of genes with promoter-proximal NF-Y binding (grey, *n* = 3,056), those that bind NF-Y and exhibit TSS shift (black, *n* = 538), or those chosen at random (purple, *n* = 538). *Right*: Box plot showing PhyloP conservation score for the 1 Kb region upstream of canonical TSS. *P-value=0.01372, **P-value=1.69E-10, ***P-value=3.77E-12 (Wilcoxon rank-sum test, two-sided)

(G) Box plot showing the number of CCAAT motifs within 1 Kb region upstream of canonical TSS for genes with promoter-proximal NF-Y binding that exhibit TSS shifts (black) or not (grey) in *NF-YA* KD ESCs. *P-value = 6.67E-36 (Wilcoxon rank-sum test, two-sided)

(H) Average sequence conservation in mammals, as computed using PhyloP (Pollard et al., 2010), of the 2 Kb region centered on ectopic TSS of genes with promoter-proximal NF-Y binding that exhibit an ectopic TSS (*n* = 538).

Figure 4. Transcripts originating from ectopic TSSs in NF-Y depleted cells undergo translation

(A) Genome browser shot of NF-Y target gene *Khsrp* showing ribosome-protected RNA expression, as measured using Ribo-Seq, in control and *NF-YA* KD ESCs. Also shown are tracks for NF-YA ChIP-Seq, and RNA-Seq and Start-Seq in control and *NF-YA* KD ESCs.

(B) Scatter plot showing the Ribo-Seq expression (coverage), in control (x-axis) and *NF-YA* KD (y-axis) ESCs, of the region between the ectopic (shifted) TSS and 25 bp upstream of the canonical TSS of genes with promoter-proximal NF-Y binding that exhibit a TSS shift (red). For comparison purposes, Ribo-Seq expression of the region between 115 bp (median TSS shift distance) and 25 bp upstream of TSS of genes with no promoter-proximal NF-Y binding is shown (black). RPKM, reads per kilobase per million mapped reads.

(C) Box plot showing the Ribo-Seq expression, in control (blue) and *NF-YA* KD (red) ESCs, of the region between the ectopic TSS and 25 bp upstream of the canonical TSS of genes with promoter-proximal NF-Y binding that exhibit an ectopic TSS. For comparison purposes, Ribo-Seq expression of the region between 115 bp (median TSS shift distance) and 25 bp upstream of TSS of genes with no promoter-proximal NF-Y binding is shown. ***P-value = 8.21E-39 (Wilcoxon rank-sum test, two-sided).

(D) Ribo-Seq read coverage, centered on the most likely ectopic start codon (ATG), in control and *NF-YA* KD ESCs. Only genes with promoter-proximal NF-Y binding that exhibit an ectopic TSS in *NF-YA* KD ESCs were used (*n* = 429).

See also Figure S4.

STAR METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Mouse ES Cell Culture
- METHOD DETAILS
  - Transient Transfection
  - Chromatin Immunoprecipitation
  - ChIP-Exo
  - ChIP-Exo Data Analysis
  - DNase I Hypersensitivity
  - Quantitative RT-PCR
  - Western Blot
  - MNase-Seq
  - MNase-Seq Data Analysis
  - RNA-Seq
  - RNA-Seq Data Analysis
  - Start-Seq
  - Start-Seq Data Analysis
  - Observed TSS Calling
  - Ectopic TSS Calling
  - Ribo-Seq
  - Ribo-Seq Data Analysis
  - Motif Analysis
  - Sequence Conservation and Predicted Nucleosome Occupancy Data
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND SOFTWARE AVAILABILITY

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-NF-YA, Mouse monoclonal	Santa Cruz Biotechnology	G-2, sc-17753X
Anti-Ran, Mouse monoclonal	BD Bioscience	610341
Anti-Rbp3	Custom	Williams et al., 2015
Anti-Histone H3, Rabbit polyclonal	Abcam	ab1791
Anti-RNA PolII, Mouse monoclonal	Covance	8WG16, MMS-126R
Anti-Tbp, Mouse monoclonal	Abcam	ab51841
Anti-Ppp2r5d	Bethyl	A301-100A
Anti-Tmx2	Abcam	ab105675
Anti-Atp5g1	Abcam	ab180149
Anti-Nono, Mouse monoclonal	Santa Cruz Biotechnology	A-11, sc-166702
Anti-Shc1, Rabbit polyclonal	Bethyl	A302-019A
Anti-Orc6, Mouse monoclonal	Santa Cruz Biotechnology	D4, sc-390490
Anti-Khsrp, Rabbit polyclonal	Bethyl	A302-021A
Anti-C7orf50, Rabbit polyclonal	Bethyl	A305-091A
Anti-Gapdh, Rabbit polyclonal	Santa Cruz Biotechnology	FL-335, sc-25778
Chemicals, Peptides, and Recombinant Proteins		
2-Mercaptoethanol	Sigma	M3148
Ampure XP beads	Beckman Coulter	A63880
CaCl2	Sigma	C1016
Cyclohexamide	Sigma	01810-1G
DMEM	Thermo Fisher Scientific	11965-084
DNase	Worthington	LS006344
DTT	Roche	10708984001
Dynabeads (protein G)	Pierce	88847
Dynabeads (protein A)	Pierce	88846
EDTA	Gibco	15575-038
EGTA	Bioworld	40520008-2
EmbryoMax nucleosides	Millipore	ES-008-D
EmeraldAmp GT PCR Master Mix	Takara	RR310A
Exonuclease Lambda	NEB	M0262S
Exonuclease RecJ	NEB	M0264S
Fetal Bovine Serum (FBS)	Gemini	100-125
Formaldehyde	Sigma	F8775
Gelatin	Sigma	G1890
Glycine	Sigma	G8898
GlycoBlue	Ambion	AM9515
Hepes	Sigma	H3375
IGEPAL CA-630	Sigma	I-3021
KCl	Sigma	P9541
Klenow large fragment DNA polymerase	NEB	M0210S
LiCl	Sigma	62476
LIF	Millipore	ESG1107
Lipofectamine 2000	Invitrogen	1168019
MgCl2	Sigma	M2670
MNase	Worthington	LS004797

NaCl, 5M	Sigma	S5150
NaHCO3	Sigma	S6014
NF-YA FlexiTube siRNA	Qiagen	SI01327193
NF-YB siRNA	Invitrogen	MSS247473
NF-YC FlexiTube siRNA	Qiagen	SI05348217
N-Lauroylsarcosine	MP Biomedicals	190110
Negative Control siRNA	Dharmacon	D-001810-02-50
Next High-Fidelity 2x PCR Master-mix	NEB	M0541S
Non-Essential Amino Acids Solution	Thermo Fisher Scientific	11140050
PBS	Made in house	
Phi29 polymerase	NEB	M0269S
PMSF	Sigma	P7626
Protease Inhibitors	Roche	4693159001
Proteinase K	Invitrogen	25530049
QIAzol Lysis Reagent	Qiagen	79306
RNase A	Thermo Scientific	EN0531
RNase cocktail	Ambion	AM2286
SDS, 20%	Fisher Scientific	BP166
Sodium Deoxycholate	Sigma	30970
Spermidine	Sigma-Aldrich	S0266
Spermine	Sigma-Aldrich	S3256
SsoFast EvaGreen supermix	Bio-Rad	1725201
Sucrose	MP Biomedicals	152584
Superase-In	Ambion	AM2694
T4 DNA polymerase	NEB	M0203S
T4 PNK	NEB	M0201S
Tris HCL, pH 8.0	Sigma	T2663
Triton X-100	Sigma	X100
Critical Commercial Assays		
iScript cDNA Synthesis Kit	Bio-Rad	1708891
QIAquick Gel Extraction Kit	Qiagen	28706
Rneasy Mini Kit	Qiagen	74104
NuGen Ovation Ultralow Library System v2	NuGEN	0344NB-08
TruSeq stranded total RNA library prep gold	Illumina	20020598
TruSeq small RNA library prep kit	Illumina	RS-200-0012
TruSeq Ribo Profile Mammalian library prep kit	Illumina	RPHMR12126
Experimental Models: Cell Lines		
Mouse ESCs (E14Tg2a)	ATCC	CRL-1821
Oligonucleotides		
Primers used for DNase hypersensitivity assays, ChIP-qPCR, MNase-qPCR and RT-qPCR	This paper	Table S2
Publicly Available Datasets Used For Analyses		
Mouse ESCs, NF-YA ChIP-Seq	Oldfield et al., 2014	GEO GSE34520

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for reagents may be directed to and will be fulfilled by the Lead Contact, Dr. Raja Jothi (jothi@nih.gov).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Mouse ESC Cell Lines, Culture and RNAi

Mouse ESCs (E14Tg2a, ATCC) were maintained on gelatin-coated plates in the ESGRO complete plus clonal grade medium (Millipore), as previously described (Cinghu et al., 2017; Oldfield et al., 2014). For experiments, ESCs were cultured on gelatin-coated plates in M15 medium: DMEM (Invitrogen) supplemented with 15% FBS, 10 µM 2-mercaptoethanol, 0.1 mM nonessential amino acids (Invitrogen), 1x EmbryoMax nucleosides (Millipore), 1 U/ml of ESGRO mLIF (Millipore).

METHOD DETAILS

Transient Transfection

For transfections, ESCs were cultured in M15 medium and transfected with 50 nM siRNA using Lipofectamine 2000 (Invitrogen) at day 0 and collected after 48 h. Gene-specific siRNAs used: *NF-YA* (Qiagen, SI01327), *NF-YB* (Invitrogen, MSS247473), *NF-YC* (Qiagen, SI05348217), non-targeting control (Dharmacon, D- 001810-02-50).

Chromatin Immunoprecipitation

ChIP was performed as previously described (Oldfield et al., 2014). Briefly, mouse ESCs ($1x10^7$) were cross-linked with 1% formaldehyde in DMEM for 10 min, and the reaction was quenched by the addition of glycine at a final concentration of 125 mM for 5 min. Cells were

washed twice with PBS, and resuspended in 1 ml of lysis buffer A [50 mM Hepes pH 7.5; 140 mM NaCl; 1 mM EDTA; 10% Glycerol; 0.5% IGEPAL CA-630; 0.25% Triton X-100; 1x Complete protease inhibitor mixture (Roche), 200 nM PMSF]. After 10 min on ice, the cells were pelleted and resuspended in 1 ml of lysis buffer B [10 mM Tris-HCl pH 8.0; 200 mM NaCl; 1 mM EDTA; 0.5 mM EGTA; 1x protease inhibitors, 200 nM PMSF]. After 10 min at room temperature, cells were sonicated in lysis buffer C [10 mM Tris-HCl pH 8.0; 100 mM NaCl; 1 mM EDTA; 0.5 mM EGTA; 0.1% sodium deoxycholate; 0.5% N-lauroylsarcosine; 1x protease inhibitors, 200 nM PMSF] using Diagenode Bioruptor for 16 cycles (30 sec ON; 50 sec OFF) to obtain ~200–500 bp fragments. Cell debris were pre-cleared by centrifugation at 14,000 rpm for 20 min, and 25 µg of chromatin was incubated with either NF-YA (Santa Cruz, G-2, sc-17753X), histone H3 (Abcam, ab1791), RNA Pol II (Covance, MMS-126R) or TBP (Abcam, ab51841) antibodies overnight at 4˚C. Protein A/G-conjugated magnetic beads (Pierce Biotech) were added the next day for 2 hours. Subsequent washing and reverse cross-linking were performed as previously described (Heard et al., 2001). ChIP enrichment for a primer-set was evaluated using quantitative PCR as percentage of input, and normalized to a negative primer-set.

ChIP-Exo

ChIP-Exo was performed as previously described and references therein (Luna-Zurita et al., 2016). Briefly, ChIP was done according to the previously described method (Oldfield et al., 2014) until the post-immunoprecipitation washes. Between each of the following steps, 2 RIPA buffer (50 mM HEPES pH 7.6, 1 mM EDTA, 0.7% NA-Deoxycholate, 1% NP-40, 0.5 M LiCl) and 2 TE buffer washes were done. After the final TE wash following overnight immunoprecipitation, the DNA was end-repaired for 30 min at 30˚C (using T4 DNA

polymerase, Klenow fragment DNA polymerase and T4 PNK, New England Biolabs). A P7 adapter was then ligated at 25˚C for 60 min. This was followed by a nick-repair step (30˚C for 20 min) using Phi29 polymerase (New England Bioloabs). The first exonuclease step was performed for 30 min. at 37˚C using Lambda exonuclease (New England Bioloabs). The second exonuclease digestion was done with RecJ exonuclease (New England Bioloabs) at 37˚C for 30 min. After the last TE wash, the DNA was isolated by first eluting it for 30 min. at 65˚C (50 mM Tris-HCl pH 8.0, 10 mM EDTA, 1% SDS). The samples were then treated at 37˚C for 30 min with RNase-A and finally incubated overnight at 65˚C after the addition of 1 µl Proteinase K (20 mg/ml) and 12 µl of 5 M NaCl. The next day, the 2nd strand was synthesized using Phi29 polymerase. After an ampure bead purification (Beckman Coulter) the P5 adapter was ligated (25˚C for 60 min. then 10 min at 65˚C. Following a second ampure bead purification (using only 0.9x instead of the usual 1.8x). The DNA was then PCR amplified using the Next High-Fidelity 2x PCR Master-mix (New England Biolabs) for between 12 and 16 cycles depending on the sample. After a final ampure bead purification (0.9x instead of 1.8x), samples were purified on a 2% agarose gel, from which the librairies (migrating at between 200 and 400 bp) were cut out and eluted using the Qiagen gel extraction kit.

ChIP-Exo Data Analysis

Though paired-end sequencing was employed, only the end 1 reads, those subject to exonuclease digestion, were aligned to the mouse (mm9) genome using Bowtie (Langmead et al., 2009) retaining only those that map to unique location (-m1, -v2, -X10000, --best). Mean distance between opposite strand 5’ ends was determined using Homer 4.8.3 (Heinz et al., 2010). Custom scripts were utilized to generate strand-merged bedGraphs files, shifting each read half the calculated distance in the 3’ direction. Duplicate fragments were

filtered using custom scripts. Replicates were merged for each condition and normalized to 1x depth coverage.

DNase I Hypersensitivity

DNase I hypersensitivity experiments were performed as previously described (Oldfield et al., 2014). Briefly, mouse ESCs treated with non-targeting control siRNA or NF-YA siRNA were collected 48 hours post-transfection in cold PBS. Nuclei isolation and DNase I digestion were performed as previously described (Burch and Weintraub, 1983), with minor modifications. Nuclei were isolated by incubation of 10^7 cells for 10 min on ice with 5 ml RSB buffer [10 mM Tris–HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.15 mM spermine and 0.5 mM spermidine, 1 mM PMSF, 0.5% IGEPAL], and pelleted by centrifugation at 300 g and 4°C for 10 min. Nuclei were then resuspended in 1 ml DNase reaction buffer [40 mM Tris-HCl pH 7.4, 10 mM NaCl, 6 mM $MgCl_2$, 1 mM $CaCl_2$, 0.15 mM Spermine, 0.5 mM Spermidine] and counted. Additional resuspension buffer was used to generate equal concentrations of nuclei between samples. Nuclei from $5x10^5$ cells were aliquoted into microcentrifuge tubes and incubated at 37°C for 5 min with varying amounts of DNase (0 to 75 U, Worthington). Digestion was stopped by addition of an equal volume of termination buffer [10 mM Tris pH 7.4, 50 mM NaCl, 100 mM EDTA, 2% SDS, 10 µg/ml RNAse cocktail]. The nuclei were then incubated at 55°C for 15 min, followed by addition of 2 µl of 20 mg/ml Proteinase K. Reaction mixtures were incubated overnight at 55°C, followed by a phenol–chloroform extraction of the DNA. The DNA was then precipitated and resuspended in 100 µl H2O.

Quantitative RT-PCR

Quantitative RT-PCR was performed as previously described (Oldfield et al., 2014). Briefly,

Total RNAs were prepared from cells using Qiazol lysis reagent (Qiagen), and cDNAs were generated using the iScript kit (Bio-Rad) according to the manufacturer's instructions. Quantitative PCRs were performed on the Bio-rad CFX-96 or CFX-384 Real-Time PCR System using the Bio-rad SsoFast EvaGreen supermix. Three or more biological replicates were performed for each experiment. Data are normalized to *Actin*, *Haz* and *TBP* expression, and plotted as mean +/- S.E.M.

Western-blot

Western-blots were performed as previously described (Oldfield et al., 2014). Briefly, Cell pellets, lysed in RIPA buffer (25 mM Tris-HCl, pH 7.4, 150 mM NaCl, 1% NP-40, 1% Sodium deoxycholate) with protease inhibitors, were sonicated using Bioruptor (Diagenode) for three cycles (30 sec ON; 50 sec OFF). The lysate was boiled with SDS-PAGE sample buffer, loaded onto NuPAGE gel, and transferred to 0.22 µM PVDF membranes. The membranes were pre-wet in 100% methanol and rinsed with ultrapure water before being washed for 5 min in 1x PBS. The membranes were then blocked with Odyssey blocking buffer for 1 h at room temperature with gentle shaking. Each membrane was treated with appropriate primary and secondary (IRDye®) antibodies. RNA Pol II and NF-YA levels were measured using an anti-Rpb3 antibody (Williams et al., 2015) and Santa Cruz, G-2, sc-17753X, respectively with Ran (BD Bioscience, 610341) as a loading control. The following antibodies were used in **Figure S4D**: Ppp2r5d (Bethyl, A301-100A), Tmx2 (Abcam, ab105675), Atp5g1 (Abcam, ab180149), Nono (Santa cruz biotechnology, sc-166702), C7orf50 (Bethyl, A305-091A), Orc6 (Santa cruz biotechnology, sc-390490), Shc1 (Bethyl, A302-019A), Khsrp (Bethyl, A302-021A) and Gapdh (Santa cruz biotechnology, sc-25778). The membranes were then washed in PBS (0.1% Tween 20), rinsed with PBS and scanned and quantified on an Odyssey

imaging system.

MNase-Seq

ES cells were crosslinked for 45 sec with 1% formaldehyde followed by glycine (125 mM) quenching for 5 min. Cells were washed once with ice-cold PBS and gently resuspended in 300 µl RSB (10 mM TrisHCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 1 mM PMSF, 0.15 mM Spermine, 0.5 mM Spermidine). 7ml of RSB + 0.5% IGEPAL were slowly added and the solution was incubated on ice for 10 min. Nuclei were pelleted at 300 X G for 10 min at 4˚C and resuspended in 400 µl of digest buffer (15 mM Tris pH 8, 60 mM KCl, 15 mM NaCl, 1 mM $CaCl_2$, 0.25 M sucrose, 0.5 mM DTT), and incubated with 40 U of MNase (Worthington) for 5 min. 400 µl stop solution (1% SDS, 0.1 M sodium bicarbonate, 20 mM EDTA) was added and digested nuclei were incubated at 65°C for 90 min. 24 µl Tris pH 7.6, 3 µl Proteinase K (Life Technologies), and 3 µl GlycoBlue (Life Technologies) were added and digestions were incubated overnight at 65°C. DNA was extracted by phenol/chloroform extraction and mono-nucleosome fragments were gel-purified from fractions digested to ~70% mono-nucleosome, ~20% di-nucleosome, and ~10% tri-nucleosome sized fragments. MNase-Seq libraries were prepared using NuGen's Ovation Ultralow Library System V2.

MNase-Seq Data Analysis

MNase-Seq read pairs for all samples were aligned to the mouse (mm9) genome using Bowtie (Langmead et al., 2009), retaining only uniquely mappable pairs (-m1, -v2, -X10000, --best). Fragments shorter than 120 nt and larger than 180 nt were filtered, as were all duplicate fragments, using custom scripts. Replicates were merged for each condition, and normalized per 10 million uniquely mappable, non-duplicate fragments. BedGraph files

containing single-nucleotide resolution fragment centers were generated to facilitate metagene analyses and creation of heatmaps, while whole-fragment coverage bedGraphs were generated for visualization purposes.

RNA-Seq

Total RNA was extracted with Qiazol lysis reagent (Qiagen) treatment and ethanol precipitation. The samples were then treated with DNAseI Amplification grade (Invitrogen) and stranded libraries were prepared using the TruSeq stranded RNA kit (Illumina) with RiboZero depletion (Gold kit).

RNA-Seq Data Analysis

Reads were mapped to the mouse (mm9) genome using TopHat v2.1.0 (Kim et al., 2013). In order to get the transcripts GTF from our samples, Cufflinks(Trapnell et al., 2012) was run with the following options, -g (mm9 GTF from ENSEMBL, version 67, provided as guide). We generated transcriptome assemblies for each of these samples separately and then use Cuffmerge (Trapnell et al., 2012) to combine all the annotations. We used Deseq2 (Love et al., 2014) with default parameters for all differential expression analyses with gene count data from Salmon quantification(Patro et al., 2016).

Start-Seq

RNA for Start-Seq experiments were prepared from control or *NF-YA* KD cells as previously described (Williams et al., 2015). Briefly, mouse ESCs were grown as described for RNA-Seq and capped-RNAs were isolated essentially as described in (Nechaev et al., 2010). In brief, approximately $2x10^7$ ESCs were trypsinized and collected by centrifugation. After washing

with ice-cold 1x PBS, cells were swelled in 10 ml of Swelling Buffer (10 mM Tris [pH 7.5], 10 mM NaCl, 2 mM $MgCl_2$, 3 mM CaCl, 0.3 M sucrose, 0.5% Igepal, 5 mM dithiothreitol, 1 mM PMSF, protease inhibitors and SUPERase-IN RNAse inhibitor (Ambion)) by incubating for 15 min on ice followed by 14 strokes with a loose pestle. The dounced cells were spun for 5 minutes at 500x g, the supernatant (cytoplasm) was discarded, the pellet resuspended in 30 ml of Swelling Buffer and spun as above. The supernatant was discarded and the nuclei pellet was resuspended in 1 ml of Swelling Buffer, aliquoted and stored at -80˚C. Libraries were prepared according to the TruSeq Small RNA Kit (Illumina). To normalize samples, 15 synthetic capped RNAs were spiked into the Trizol preparation at a specific quantity per 10^6 cells, as in (Henriques et al., 2013).

Start-Seq Data Analysis

Start-Seq reads were trimmed for adapter sequence using cutadapt 1.2.1 (Martin, 2011); pairs with either mate trimmed shorter than 20 nt were discarded. A single additional nucleotide was removed from the 3' end of each read to facilitate mapping of fully overlapping pairs. Remaining pairs were filtered for rRNA and tRNA by aligning to indices containing each using Bowtie 0.12.8 (-v2, -X1000, --best, --un, --max), and retaining unmapped pairs. Following this, a similar alignment was performed to an index containing the sequence of spike-in RNAs only (-m1, -v2, -X1000, --best, --un, --max), and finally, the remaining unmapped reads were aligned to the mouse (mm9) genome utilizing the same parameters, retaining only uniquely mappable pairs. Strand-specific bedGraph files containing the combined raw counts of short-capped RNA 5' ends for all control replicates were generated to facilitate observed TSS calling. For all other purposes, 5' end counts were normalized per 10 million mappable reads, then based on depth-normalized counts aligning

to spike-in RNAs. Spike normalization factors were determined as the slope of the linear regression of each sample's depth normalized spike-in read counts versus the single sample with the lowest total count. Control and NF-YA knock-down bedGraph files were generated from these spike-normalized counts by taking the mean of all replicates, genome-wide, at single-nucleotide resolution.

Observed TSS Calling

Observed TSSs were identified as in (Scruggs et al., 2015), based on the control Start-Seq data, using mm9 RefSeq annotations downloaded from the UCSC genome browser (January 2015). Briefly, the position with the highest read count within 1000 nt of an annotated TSS, or that with the highest count within the 200 nt window of highest read density was selected, depending on proximity. When insufficient Pol II ChIP-Seq signal existed in the 501 nt window centered on the selected locus, relative to a comparable window about the annotated TSS (a ratio less than 2:3), the observed TSS was shifted to the location with the highest Start-Seq read count within 250 nt of the annotated TSS. When fewer than 5 reads were mapped to the selected locus, the annotated TSS was maintained. Groups of transcripts with identical observed TSS were filtered, maintaining a single representative with the shortest annotated to observed TSS distance. Groups of observed TSSs within 200 nt of one another were similarly reduced by first removing any RIKEN cDNAs, predicted genes, or observed TSSs moved to the annotation due to lack of Start-Seq reads. Following this, a single observed TSS was selected based on observed to annotated proximity. In this manner, observed TSSs were called for 24,498 transcripts; of these, 16,483 were selected based on Start-Seq data, while for 8,015 the annotated position was maintained. NFY-bound promoters were then identified as those with an NF-YA ChIP-Seq peak intersecting the

observed TSS -900 to +100 nt window.

Ectopic TSS Calling

Ectopic TSSs were identified through the comparison of NF-YA knock-down Start-Seq read counts to control using DESeq (Anders and Huber, 2010). Counts were determined for all samples in 10 nt bins tiling the region -995 to +995 nt, relative to each observed TSS. Bins closer to an upstream or downstream TSS than their own were excluded, as were those in the observed TSS -25 to +24 nt region. Normalization was performed based on size factors calculated within DESeq from spike-in RNAs alone, ensuring these values were equivalent across all samples. All bins with a positive log2 fold change and adjusted p-value less than 0.1 were identified. If more than one bin associated with a single observed TSS was selected, that with the lowest adjusted p-value was retained. Within each of these bins, the position with the total Start-Seq read count across all NF-YA knock-down samples was selected as the ectopic TSS. In cases where multiple sites exist with identical counts, that closest to the observed TSS was selected.

Ribo-Seq

Approximately 8-9 x 10^6 ESCs (per sample) were treated with cycloheximide (0.1 mg/ml) for 1 minute prior to trypsinization and cell lysis. Control and *NF-YA* KD cells were used as input material for the TruSeq Ribo Profile Mammalian Library Prep Kit (Illumina) following the manufacture's protocol.

Ribo-Seq Data Analysis

Total RNA-Seq and Ribosome-protected-RNA-Seq (Ribo-Seq) read pairs were trimmed with

cutadapt (Martin, 2011); fragments shorter than 15 nt were discarded. Read pairs were filtered for rRNA and tRNA by aligning to respective indices using Bowtie 0.12.8 (-v2, -X1000, --best, --un, --max), and retaining unmapped pairs. The remaining read pairs were aligned to the mouse (mm9) genome using STAR v2.6.0c (Dobin et al., 2013). The read counts intersecting CDSs were determined per sample. We then determined the normalization factors from the aligned counts using DESeq2 v1.18.1 (Love et al., 2014). Using STAR's bedGraph output (read pairs with unique alignment), both sequencing runs were merged using unionBedGraphs (a component of bedtools 2.25.0). We then applied the previously calculated normalization factors to the merged bedGraphs. To obtain bigwig format files, both normalized biological replicates were merged using unionBedGraphs and then converted to bigWig. To infer position of ribosomes from Ribo-Seq reads, we used the 5' end of each fragment. We then added 12 nt in order to locate the highest peak on the A of ATG. The same process was applied to reads from total RNA-Seq for comparison, to check that RNA-Seq read positions do not exhibit a 3n periodicity like Ribo-Seq reads do. To study the coverage in shifted regions we used a two-step process. First, we searched for all the ATGs in those regions. Then, the ATG with the highest coverage within the downstream 60 nt was selected. We finally plotted the cumulative coverage, taking the A of the selected ATG as the anchor. Genes whose cumulative Ribo-seq signal, within the shifted region, represented > 5% of total Ribo-Seq reads were removed from the phasing analysis (n = 3).

Motif Analysis

Around observed TSSs bound by NF-YA, those not bound by NF-YA, those bound by NF-YA with an associated ectopic TSS, as well as bound and non-bound ectopic TSSs themselves, *de novo* motif discovery was performed in the TSS +/- 50 region using Meme (Bailey et al.,

2009) (-dna -mod zoops -nmotifs 25 -minw 6 -maxw 20 -revcomp). For each ectopic TSS, the closest sense or anti-sense CCAAT motifs, and their distances, were identified via simple string searches. Empirical cumulative distribution functions were determined separately for those bound and not bound by NF-YA, as a function of distance, and were compared via one-tailed Kolmogorov-Smirnov test against a negative binomial distribution where the probability of success is based on observing "CCAAT" or it's reverse complement by chance, given the background nucleotide frequencies present in these sequences.

Sequence Conservation and Predicted Nucleosome Occupancy Data

Per-nucleotide predicted nucleosome occupancy for the mouse (mm9) genome was provided for download by authors of (Kaplan et al., 2009). Per-nucleotide phyloP conservation scores, based on 20 placental mammals, were downloaded from the UCSC Genome Browser. Both data sets were converted to bedGraph format using custom scripts to facilitate generation of metagene analyses and heatmaps.

QUANTIFICATION AND STATISTICAL ANALYSIS

See Methods Details for details of quantification and statistical analysis.

DATA AVAILABILITY

MNase-Seq, RNA-Seq, Start-Seq, Ribo-Seq, and RNA Pol II ChIP-Exo data generated for this study have been deposited in the GEO repository under the accession number GSE115110

Figure 1

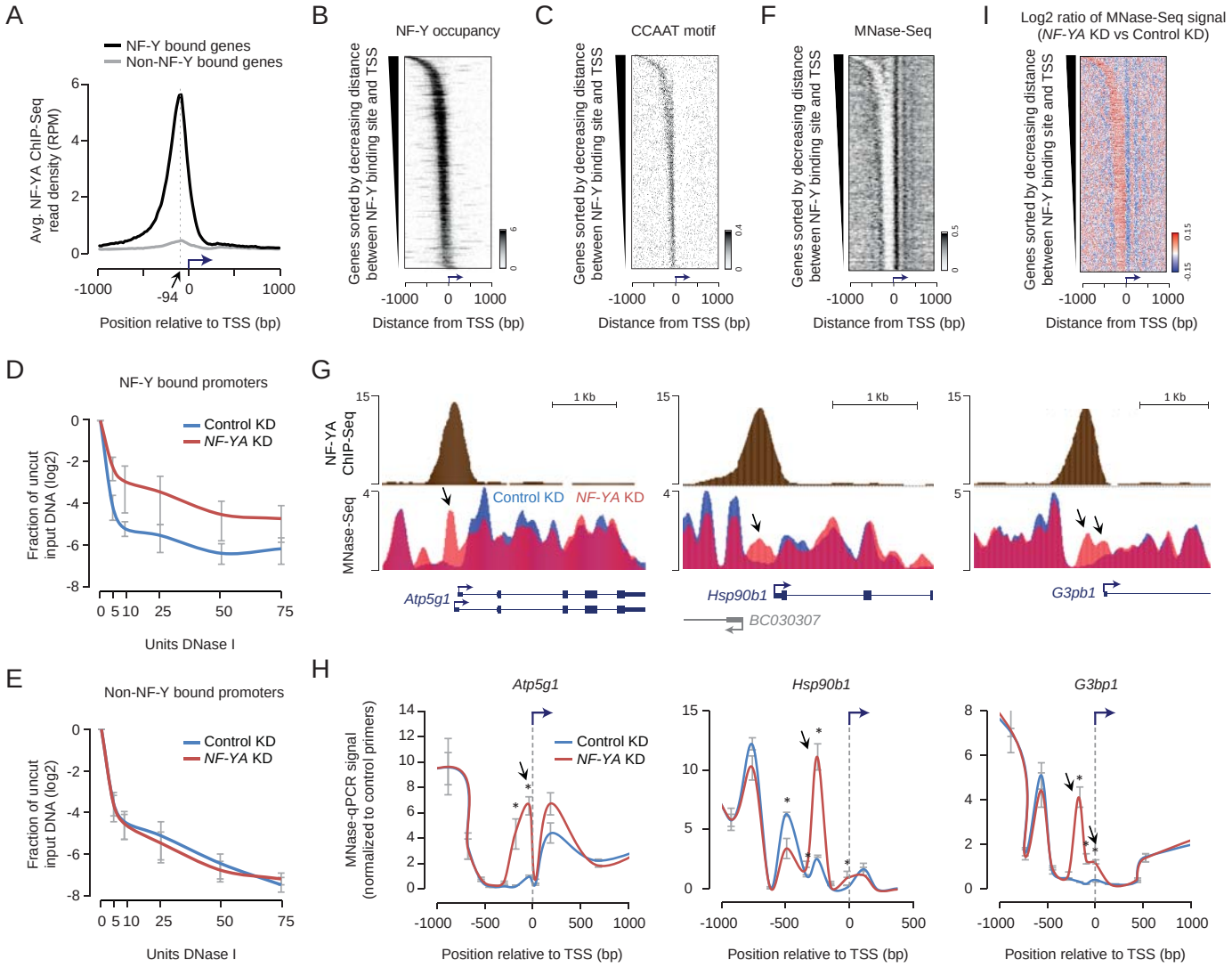

A
NF-Y bound genes
Non-NF-Y bound genes
Avg. NF-YA ChIP-Seq read density (RPM)
-94
Position relative to TSS (bp)
B
NF-Y occupancy
Genes sorted by decreasing distance between NF-Y binding site and TSS
Distance from TSS (bp)
C
CCAAT motif
Genes sorted by decreasing distance between NF-Y binding site and TSS
Distance from TSS (bp)
F
MNase-Seq
Genes sorted by decreasing distance between NF-Y binding site and TSS
Distance from TSS (bp)
I
Log2 ratio of MNase-Seq signal (NF-YA KD vs Control KD)
Genes sorted by decreasing distance between NF-Y binding site and TSS
Distance from TSS (bp)
D
NF-Y bound promoters
Control KD
NF-YA KD
Fraction of uncut input DNA (log2)
Units DNase I
G
NF-YA ChIP-Seq
MNase-Seq
1 Kb
Control KD
NF-YA KD
Atp5g1
Hsp90b1
BC030307
G3pb1
E
Non-NF-Y bound promoters
Control KD
NF-YA KD
Fraction of uncut input DNA (log2)
Units DNase I
H
Atp5g1
Hsp90b1
G3bp1
MNase-qPCR signal (normalized to control primers)
Control KD
NF-YA KD
Position relative to TSS (bp)

Figure 2

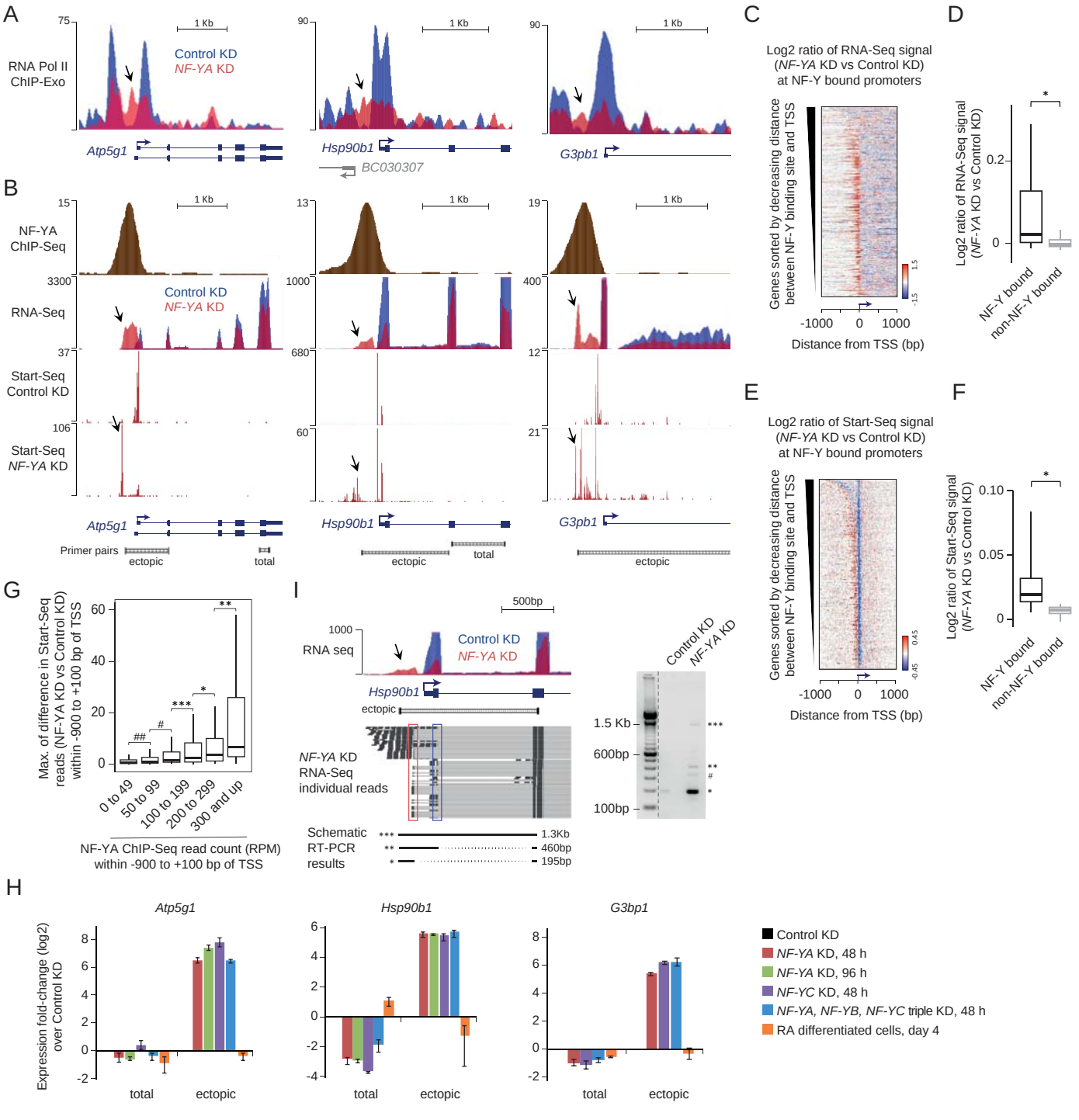

A
RNA Pol II ChIP-Exo
Control KD
NF-YA KD
1 Kb
Atp5g1
Hsp90b1
BC030307
G3pb1
B
NF-YA ChIP-Seq
RNA-Seq
Start-Seq Control KD
Start-Seq NF-YA KD
Primer pairs
ectopic
total
C
Log2 ratio of RNA-Seq signal (NF-YA KD vs Control KD) at NF-Y bound promoters
Genes sorted by decreasing distance between NF-Y binding site and TSS
Distance from TSS (bp)
D
Log2 ratio of RNA-Seq signal (NF-YA KD vs Control KD)
NF-Y bound
non-NF-Y bound
E
Log2 ratio of Start-Seq signal (NF-YA KD vs Control KD) at NF-Y bound promoters
F
Log2 ratio of Start-Seq signal (NF-YA KD vs Control KD)
G
Max. of difference in Start-Seq reads (NF-YA KD vs Control KD) within -900 to +100 bp of TSS
0 to 49
50 to 99
100 to 199
200 to 299
300 and up
NF-YA ChIP-Seq read count (RPM) within -900 to +100 bp of TSS
I
500bp
RNA seq
NF-YA KD RNA-Seq individual reads
Schematic RT-PCR results
1.3Kb
460bp
195bp
1.5 Kb
600bp
100bp
H
Atp5g1
Hsp90b1
G3bp1
Expression fold-change (log2) over Control KD
total
ectopic
Control KD
NF-YA KD, 48 h
NF-YA KD, 96 h
NF-YC KD, 48 h
NF-YA, NF-YB, NF-YC triple KD, 48 h
RA differentiated cells, day 4

Figure 3

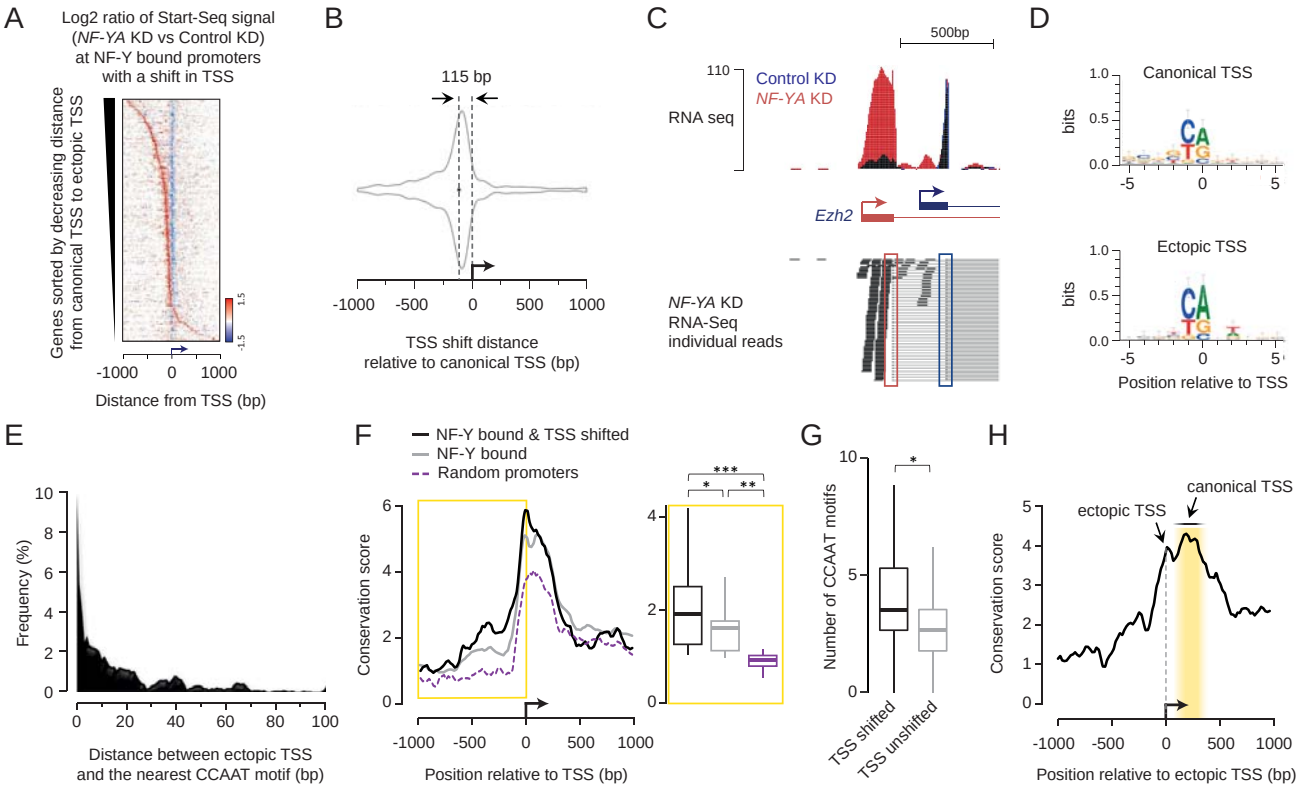
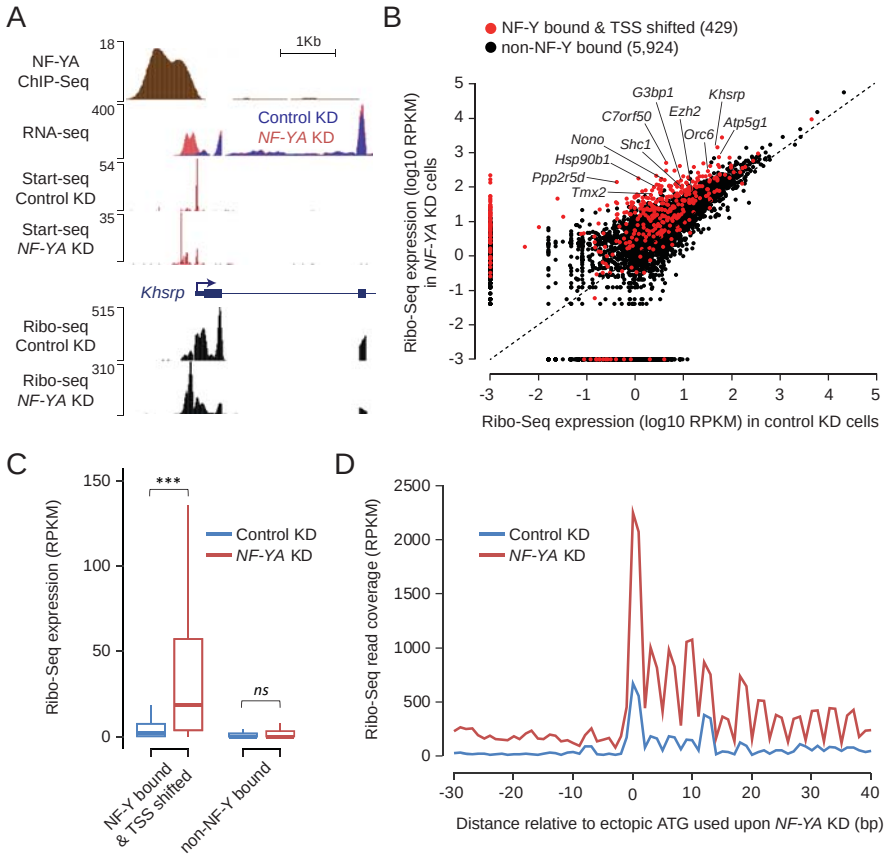

A
Log2 ratio of Start-Seq signal
(NF-YA KD vs Control KD)
at NF-Y bound promoters
with a shift in TSS
Genes sorted by decreasing distance
from canonical TSS to ectopic TSS
1.5
-1.5
-1000
0
1000
Distance from TSS (bp)
B
115 bp
-1000
-500
0
500
1000
TSS shift distance
relative to canonical TSS (bp)
C
500bp
110
Control KD
NF-YA KD
RNA seq
Ezh2
NF-YA KD
RNA-Seq
individual reads
D
Canonical TSS
bits
1.0
0.5
0.0
-5
0
5
Ectopic TSS
Position relative to TSS
E
Frequency (%)
10
8
6
4
2
0
0
20
40
60
80
100
Distance between ectopic TSS
and the nearest CCAAT motif (bp)
F
NF-Y bound & TSS shifted
NF-Y bound
Random promoters
Conservation score
6
4
2
0
-1000
-500
0
500
1000
Position relative to TSS (bp)
***
*
**
G
Number of CCAAT motifs
10
5
0
*
TSS shifted
TSS unshifted
H
ectopic TSS
canonical TSS
Conservation score
5
4
3
2
1
0
-1000
-500
0
500
1000
Position relative to ectopic TSS (bp)

Figure 4


A
NF-YA ChIP-Seq
18
1Kb
RNA-seq
400
Control KD
NF-YA KD
Start-seq Control KD
54
Start-seq NF-YA KD
35
Khsrp
Ribo-seq Control KD
515
Ribo-seq NF-YA KD
310
B
NF-Y bound & TSS shifted (429)
non-NF-Y bound (5,924)
Ribo-Seq expression (log10 RPKM) in NF-YA KD cells
Ribo-Seq expression (log10 RPKM) in control KD cells
G3bp1
Khsrp
C7orf50
Ezh2
Atp5g1
Nono
Shc1
Orc6
Hsp90b1
Ppp2r5d
Tmx2
C
Ribo-Seq expression (RPKM)
***
ns
Control KD
NF-YA KD
NF-Y bound & TSS shifted
non-NF-Y bound
D
Ribo-Seq read coverage (RPKM)
Control KD
NF-YA KD
Distance relative to ectopic ATG used upon NF-YA KD (bp)

SUPPLEMENTAL INFORMATION

NF-Y controls fidelity of transcription initiation at gene promoters through maintenance of the nucleosome-depleted region

Andrew J Oldfield, Telmo Henriques, Adam B. Burkholder, Damien Paulet, Senthilkumar Cinghu, Pengyi Yang, Benjamin S. Scruggs, Christopher A. Lavender, Dhirendra Kumar, Brian Bennett, Eric Rivals, Karen Adelman, Raja Jothi

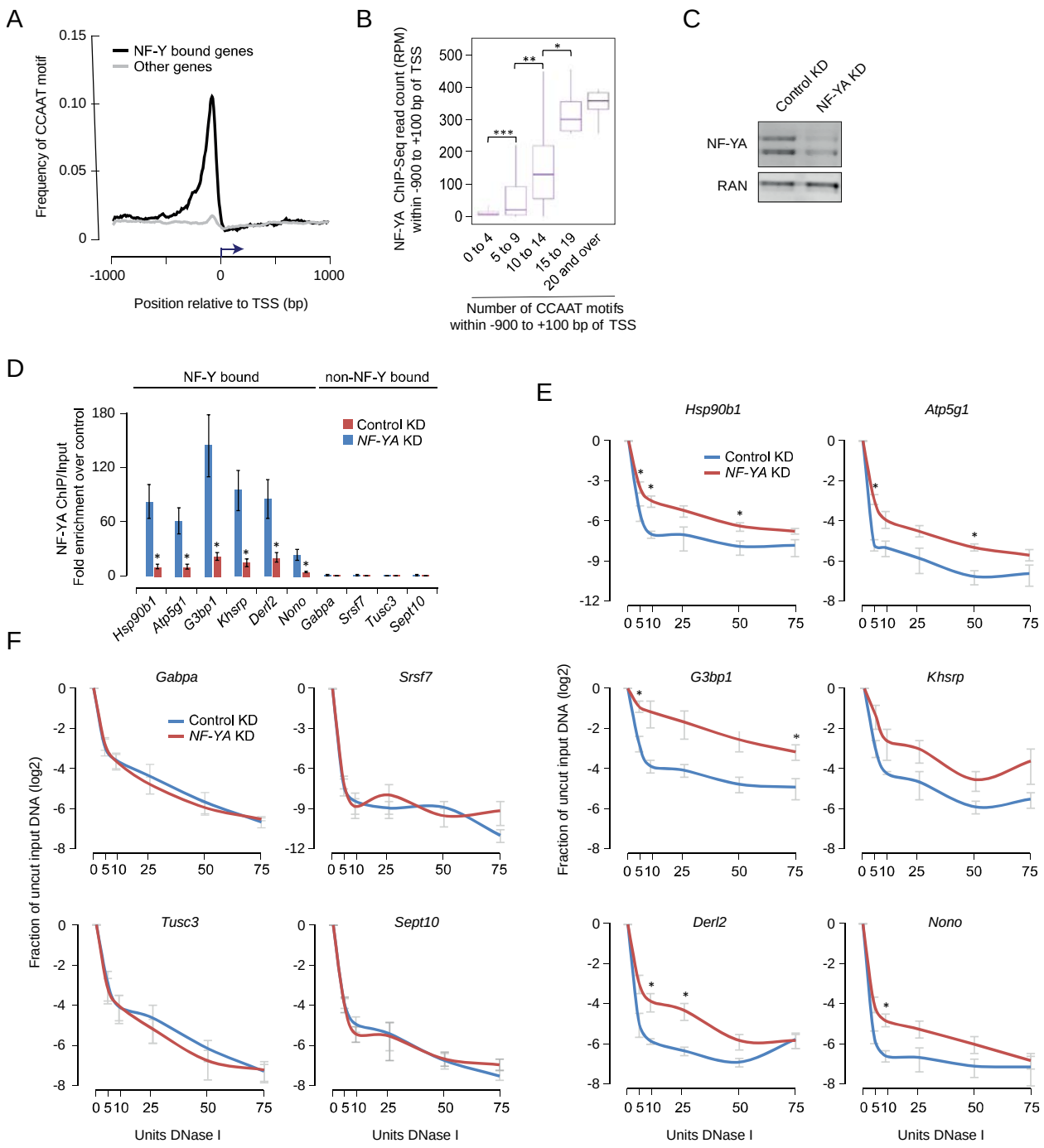


Figure S1 (related to Figure 1). NF-Y binding is required to maintain accessible chromatin at its target promoters

(A) Frequency of CCAAT motif occurrence near TSSs of genes with (*n* = 3,056, black) or without *n* = 21,195, grey) promoter-proximal NF-Y binding in ESCs.

(B) Box plot showing the relationship between CCAAT motif occurrence within the promoter (-900 bp to +100 bp region relative to TSS, x-axis) and NF-YA ChIP-Seq read density (y-axis) for all annotated RefSeq genes. *P-value < 0.0006, **P-value = 1.61E-21, ***P-value = 1.67E-74 (Wilcoxon rank-sum test, two-sided)

(C) Western-blot analysis of NF-YA in control or *NF-YA* KD ESCs, 48 hr after siRNA transfection. Ran used as a loading control.

(D) ChIP-qPCR analysis of NF-Y binding sites in control or *NF-YA* KD ESCs. Error bars, SEM of three to five Biological replicates.

(E, F) DNase I hypersensitivity and qPCR analysis of promoters with (E) or without (F) NF-Y binding in control (blue) and *NF-YA* KD (red) ESCs. Error bars, SEM of three biological replicates. *P-value < 0.05 (Student's t-test, two-sided).

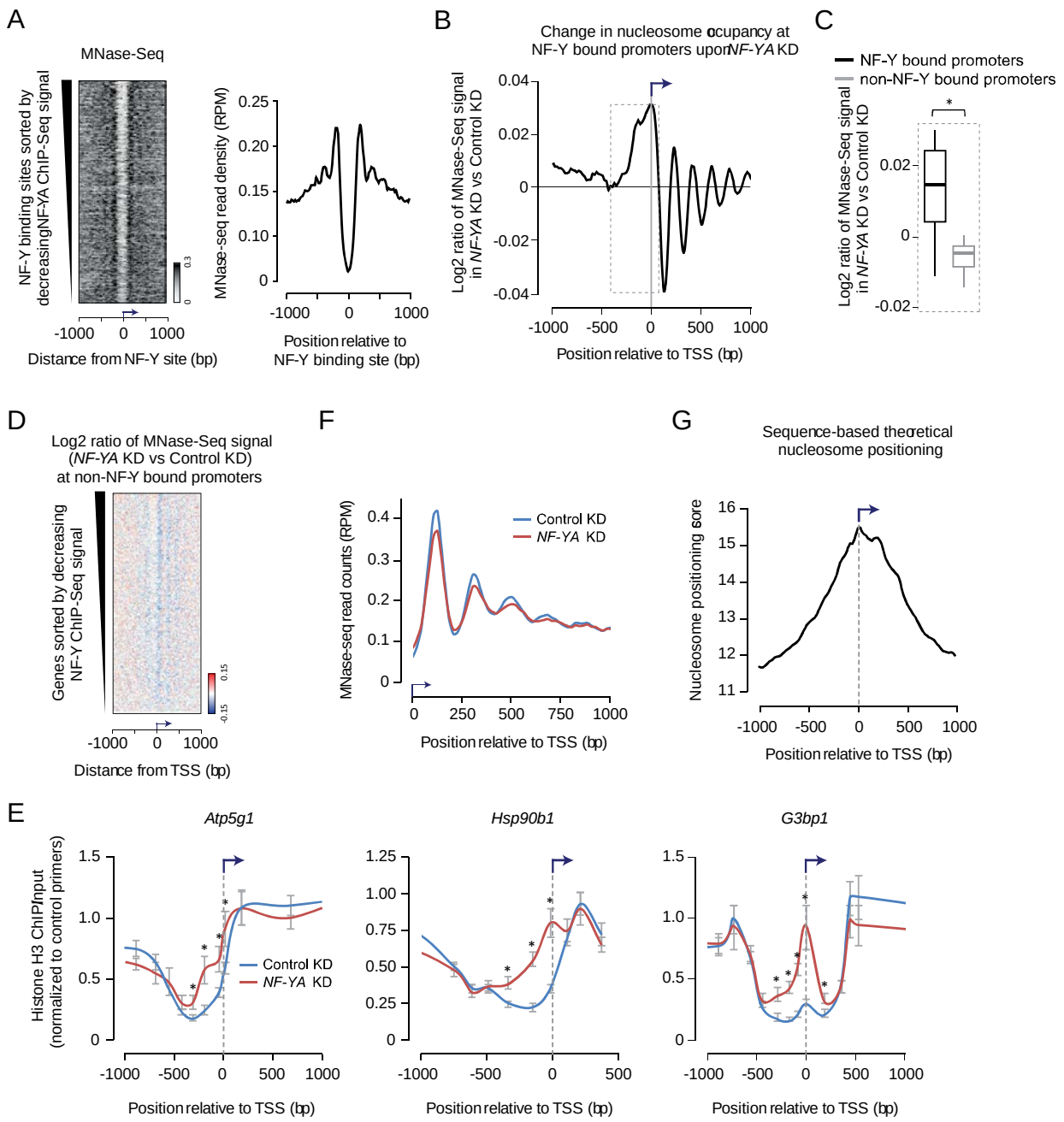


Figure S2 (related to Figure 1). NF-Y binding protects nucleosome-depleted region from nucleosome encroachment

(A) *Left*: Nucleosome occupancy, as measured using MNase-Seq, at all NF-Y binding sites (*n* = 5,359) in ESCs. *Right*: Average nucleosome occupancy (y-axis) at NF-Y binding sites. RPM, reads per million mapped reads.

(B) Relative change (gain or loss) in nucleosome occupancy (y-axis) near TSSs of genes with promoter-proximal NF-Y binding (*n* = 3,056) in *NF-YA* KD vs control KD ESCs.

(C) Box plot showing relative change in nucleosome occupancy (in *NF-YA* KD vs control KD ESCs) within the NDRs (-400 to +80 bp) of genes with (black) or without (gray) promoter-proximal NF-Y binding. *P=9.88E-09 (Wilcoxon rank-sum test, two-sided)

(D) Relative change in nucleosome occupancy (gain, red; loss, blue) near TSSs of genes without promoter-proximal NF-Y binding in *NF-YA* KD vs control KD ESCs.

(E) ChIP-qPCR analysis of Histone H3 occupancy at candidate gene promoters with NF-Y binding in control or *NF-YA* KD ESCs. Error bars, SEM of three replicates.

(F) Average nucleosome occupancy (y-axis) downstream of TSSs for genes with promoter-proximal NF-Y binding in control (blue) or *NF-YA* KD (red) ESCs.

(G) Density plot showing average theoretical nucleosome positioning, predicted based on DNA sequence composition, near TSSs of genes with promoter-proximal NF-Y binding in ESCs.

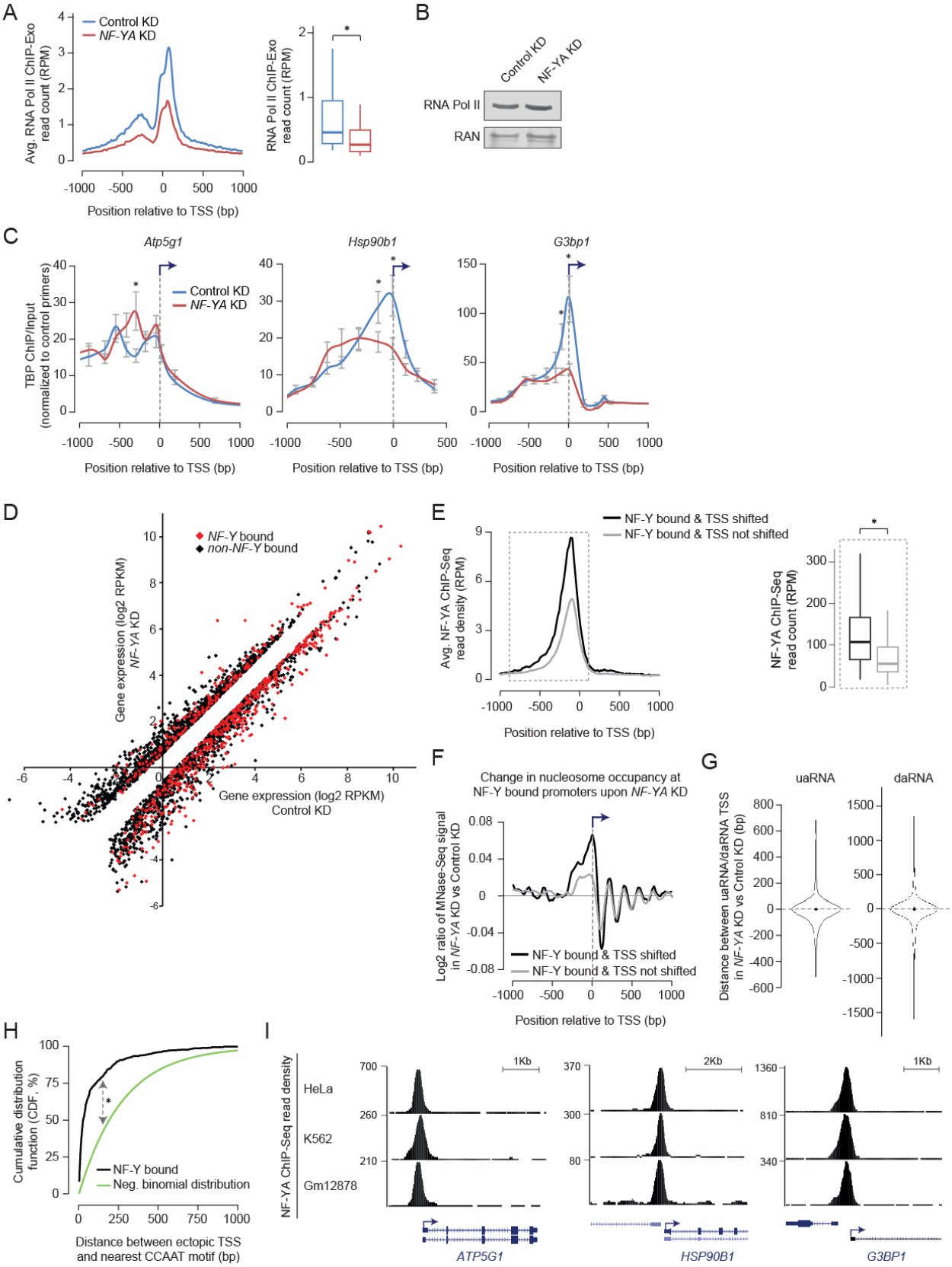

A
Control KD
NF-YA KD
Avg. RNA Pol II ChIP-Exo read count (RPM)
Position relative to TSS (bp)
RNA Pol II ChIP-Exo read count (RPM)
B
Control KD
NF-YA KD
RNA Pol II
RAN
C
Atp5g1
Hsp90b1
G3bp1
TBP ChIP/Input (normalized to control primers)
Control KD
NF-YA KD
Position relative to TSS (bp)
D
NF-Y bound
non-NF-Y bound
Gene expression (log2 RPKM) NF-YA KD
Gene expression (log2 RPKM) Control KD
E
NF-Y bound & TSS shifted
NF-Y bound & TSS not shifted
Avg. NF-YA ChIP-Seq read density (RPM)
Position relative to TSS (bp)
NF-YA ChIP-Seq read count (RPM)
F
Change in nucleosome occupancy at NF-Y bound promoters upon NF-YA KD
Log2 ratio of MNase-Seq signal in NF-YA KD vs Control KD
NF-Y bound & TSS shifted
NF-Y bound & TSS not shifted
Position relative to TSS (bp)
G
uaRNA
daRNA
Distance between uaRNA/daRNA TSS in NF-YA KD vs Control KD (bp)
H
Cumulative distribution function (CDF, %)
NF-Y bound
Neg. binomial distribution
Distance between ectopic TSS and nearest CCAAT motif (bp)
I
NF-YA ChIP-Seq read density
HeLa
K562
Gm12878
1Kb
2Kb
1Kb
ATP5G1
HSP90B1
G3BP1

Figure S3 (related to Figure 2). NF-Y binding influences PIC positioning and TSS selection

(A) *Left*: Average RNA Polymerase II (Pol II) occupancy (y-axis), as measure using ChIP-Exo, near TSSs of genes with promoter-proximal NF-Y binding (*n*=3,191) in control (blue) or *NF-YA* KD (red) ESCs. RPM, reads per million mapped reads. *Right*: Box plot showing Pol II ChIP-Exo read counts with within +/-1 Kb region of TSSs in *NF-YA* KD or control KD ESCs. *P=1.31E-23 (Wilcoxon rank-sum test, two-sided)

(B) Western blot analysis of Pol II in control or *NF-YA* KD ESCs, 48 hr after siRNA transfection. Ran used as a loading control.

(C) ChIP-qPCR analysis of TBP occupancy at candidate gene promoters with promoter-proximal NF-Y binding in control (blue) or *NF-YA* KD (red) ESCs. Error bars, SEM of three replicates.

(D) Scatter plot showing the (RNA-Seq) expression of NF-Y bound (red, *n* = 660) and non-NF-Y bound (black, *n* = 2,389) genes in control (x-axis) and NF-YA KD (y-axis) ESCs. Only genes whose expression is significantly different (>1.5-fold) in NF-YA KD vs control cells are shown.

(E) *Left*: Average NF-YA occupancy (y-axis), as measured using ChIP-Seq, near TSSs of genes with promoter-proximal NF-Y binding that exhibit (black, *n* = 538) or do not exhibit (grey, *n* = 2,518) an ectopic TSS (TSS shift) in *NF-YA* KD vs control KD ESCs. RPM, reads per million mapped reads. *Right*: Box plot showing NF-YA ChIP-Seq read counts within the highlighted promoter region (-900 bp to +100 bp). *P-Value = =1.65E-57 (Wilcoxon rank-sum test, two-sided)

(F) Relative change in nucleosome occupancy (gain, red; loss, blue) near TSSs of genes with promoter-proximal NF-Y binding that exhibit (black, *n* = 538) or do not exhibit (grey, *n* = 2,518) an ectopic TSS (TSS shift) in *NF-YA* KD vs control KD ESCs.

(G) Violin plot showing the distribution of distance by which the TSS of upstream (or downstream) anti-sense RNA (uaRNA or daRNA, respectively) shifted its position, as measured using Start-Seq, in *NF-YA* KD vs control KD ESCs. Median indicated by dotted grey line.

(H) Cumulative distribution function (CDF) of the distance between ectopic TSS and the nearest CCAAT motif for NF-Y bound genes (black). Also shown is the CDF of negative binomial distribution computed based on randomized DNA sequence with similar nucleotide composition (green). *P-value = 2.2E-16 (Kolmogorov-Smirnov test)

(I) Genome browser shots of genes, which bind NF-Y within promoter-proximal regions in mouse ESCs (see Figure 2), showing promoter-proximal NF-Y binding in human normal (Gm12878) and cancer cells (HeLa and K562). Human NF-YA ChIP-Seq data from ENCODE (GSE31477) was used.

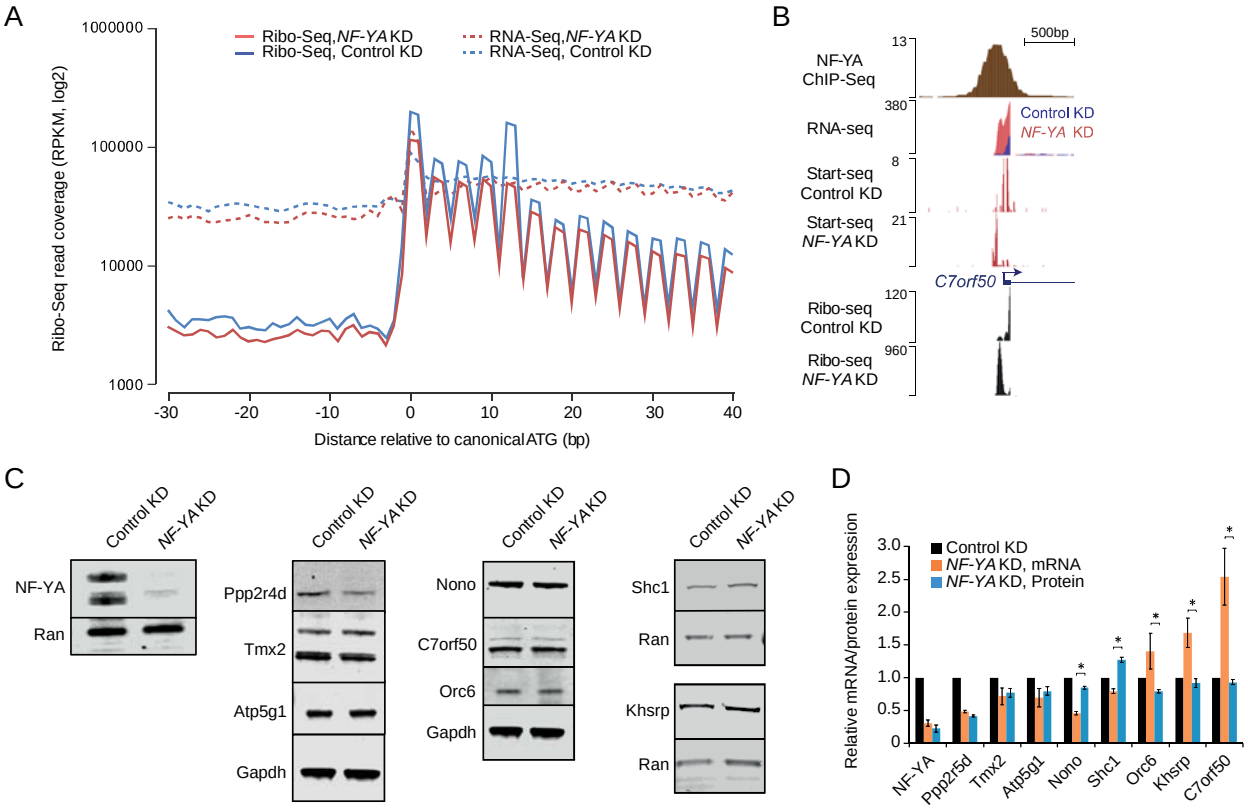


Figure S4 (related to Figure 4). Transcripts originating from ectopic TSSs in NF-Y depleted cells undergo translation

(A) Ribo-seq and RNA-Seq read coverage, centered on the canonical start codon (ATG), for all genes in control and *NF-YA* KD ESCs. Triplet phasing, beginning at the canonical start codon, is observed for Ribo-Seq but not RNA-Seq data. RPKM, reads per kilobase per million mapped reads.

(B) Genome browser shot of NF-Y target gene *C7orf50* (3110082I17Rik) showing ribosome-protected RNA expression, as measured using Ribo-Seq, in control (blue) and *NF-YA* KD (red) ESCs. Also shown are tracks for RNA-Seq and Start-Seq in control and *NF-YA* KD ESCs.

(C) Western-blot analysis of NF-YA, Ppp2r5d, Tmx2, Atp5g1, Nono, C7orf50, Orc6, Shc1 and Khsrp in control and NF-YA knock-down (KD) ESCs 48h after siRNA transfection. Ran or Gapdh used as loading controls. Representative images are shown.

(D) Relative changes in gene expression (mRNA, orange) and protein levels (blue) in *NF-YA* KD vs control ESCs. Protein levels were determined using Western-blot analysis (see Figure S4C), quantified by Licor Image Studio® software. mRNA data normalized to *Actin*, *HAZ* and *TBP*. Protein data normalized to Ran or Gapdh. Error bars, SEM of three to five biological replicates. *P-value < 0.00002 (Student's t-test, two-sided)

(E) Scatter plot showing the (RNA-Seq) expression of NF-Y bound (red) and non-NF-Y bound (black) genes in control (x-axis) and NF-YA KD (y-axis) ESCs.