Biased chromatin signatures around polyadenylation sites and exons

Citation: Spies, Noah et al. "Biased Chromatin Signatures Around Polyadenylation Sites and Exons." Molecular Cell 36.2 (2009): 245–254.

As Published: http://dx.doi.org/10.1016/j.molcel.2009.10.008

Publisher: Elsevier

Persistent URL: http://hdl.handle.net/1721.1/72553

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Published in final edited form as:
Mol Cell. 2009 October 23; 36(2): 245–254. doi:10.1016/j.molcel.2009.10.008.

Biased Chromatin Signatures Around Polyadenylation Sites and Exons

Noah Spies[1,2,*], **Cydney B. Nielsen**[1,*,3], **Richard A. Padgett**[4], and **Christopher B. Burge**[1,5,6]

[1] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142

[2] Whitehead Institute for Biomedical Research, Cambridge, MA 02142

[4] Cleveland Clinic Foundation, Cleveland, OH 44195

[5] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02142

Summary

Core RNA processing reactions in eukaryotic cells occur cotranscriptionally in a chromatin context, but the relationship between chromatin structure and pre-mRNA processing is poorly understood. We observed strong nucleosome depletion around human polyadenylation sites (PAS), and nucleosome enrichment just downstream of PAS. In genes with multiple alternative PAS, higher downstream nucleosome affinity was associated with higher PAS usage, independently of known PAS motifs that function at the RNA level. Conversely, exons were associated with distinct peaks in nucleosome density. Exons flanked by long introns or weak splice sites exhibited stronger nucleosome enrichment, and incorporation of nucleosome density data improved splicing simulation accuracy. Certain histone modifications, including H3K36me3 and H3K27me2, were specifically enriched on exons, suggesting active marking of exon locations at the chromatin level. Together, these findings provide evidence for extensive functional connections between chromatin structure and RNA processing.

Introduction

In multicellular organisms, most primary RNA transcripts undergo extensive processing. Both pre-mRNA splicing and cleavage/polyadenylation are usually initiated or completed cotranscriptionally, and several mechanistic links between transcription and RNA processing are known. Upon phosphorylation of the C-terminal domain (CTD) of RNA polymerase II (pol II) shortly after transcription initiation, 5′ end capping enzymes are recruited to the nascent transcript (Moore and Proudfoot, 2009). Factors central to pre-mRNA splicing are loaded onto the CTD (Kornblihtt et al., 2004), and some splicing factors, including the U1 and U2 snRNPs and SR proteins, are deposited on nascent pre-mRNAs as the 5′ and 3′ splice sites are transcribed (Gornemann et al., 2005; Lin et al., 2008). Similarly, cleavage and polyadenylation factors associate with the phosphorylated CTD and recognize the polyadenylation signal, often before pol II termination (Moore and Proudfoot, 2009). The kinetics of transcription can influence

6To whom correspondence should be addressed. cburge@mit.edu/P: (617) 258-5997/F: (617) 452-2936.
*These authors contributed equally
3Current address: Michael Smith Genome Sciences Centre, Vancouver, Canada V5Z 4S6

both pre-mRNA splicing (de la Mata et al., 2003; Howe et al., 2003) and cleavage and polyadenylation; for example, pol II pausing in the 3′ region of a gene has been shown to favor use of the more 5′ among two alternative PAS (Peterson et al., 2002).

Pre-mRNA splicing requires extremely precise identification of the correct 5′ and 3′ splice sites, frequently from amongst many kilobases of intronic sequence containing a large excess of potential splice sites that are not used. Additional *cis*-regulatory RNA sequence elements, including exonic splicing enhancers (ESEs) and silencers (ESSs), assist in the accurate identification of splice sites, generally through recruitment of factors of the serine-arginine rich (SR) protein and heterogeneous nuclear ribonucleoprotein (hnRNP) classes. Despite fairly extensive characterization of such elements, splicing simulators that incorporate these elements are still only able to correctly identify approximately half to two-thirds of exons from the sequence alone (Wang et al., 2004), underscoring that the complete set of rules for recognition of exons by the splicing machinery remains to be determined.

Transcription is influenced by chromatin structure and by histone modifications such as methylation and acetylation; both nucleosome positioning and modification status are in turn influenced by the process of transcription (Li et al., 2007). The pol II CTD functions not only to recruit RNA processing factors but also chromatin modifying factors. For example, the enzyme responsible for trimethylation of histone H3 lysine 36 (H3K36me3) is recruited to the Ser2-phosphorylated CTD, thereby establishing a pattern of this modification that is biased towards the downstream regions of expressed genes (Li et al., 2007).

A handful of recent studies have identified links between histone modifications and RNA processing (Kolasinska-Zwierz et al., 2009; Loomis et al., 2009; Schor et al., 2009; Sims et al., 2007) but whether aspects of nucleosome positioning and modification influence RNA processing generally (or vice versa) is not known. Here we show that specific chromatin signatures are associated with exons and with sites of cleavage and polyadenylation, and correlate with the strength or usage of these RNA elements, establishing a framework for interaction between chromatin structure and RNA processing.

Results

Nucleosomes are strongly enriched on exons

We observed that nucleosomes are significantly enriched on DNA encoding internal exons compared to flanking introns (Fig. 1) through analysis of high-throughput nucleosome chromatin immunoprecipitation and sequencing (ChIP-Seq) data from human T cells (Schones et al., 2008). The magnitude of the enrichment on exons (1.41-fold enrichment for nucleosomes above background; 95% confidence interval: [1.408, 1.425], by resampling) rivals or exceeds that observed at the +1 nucleosome peak near the transcription start site (TSS) when plotted using the same dataset and methods (Fig. 1G). We also observed similar enrichment of nucleosomes on exons in published data from the Japanese killifish (data not shown; (Sasaki et al., 2009)).

Biased exon composition explains nucleosome enrichment

We hypothesized that this enrichment might be explained at least partially by sequence features specific to exons, such as splice site- or ESE-related motifs. We analyzed sets of “decoy” 3′ splice site (3′ss) and decoy 5′ splice site (5′ss) sequences in introns, i.e. sequences that match the 3′ss or 5′ss consensus as well as authentic splice sites, but are not observed to be used in splicing. Nucleosome density was only slightly enriched in the vicinity of decoy 3′ss and 5′ss (Fig. 1D, F), suggesting that the enrichment on exons cannot be explained simply by effects of oligonucleotides that form parts of the splice site consensus motifs.

An alternative possibility was that the exonic bias of nucleosomes might be attributable to the distinctive oligonucleotide composition of exons (Baldi et al., 1996; Denisov et al., 1997). To explore this possibility, exon-sized stretches of nucleotides in intergenic regions or introns were identified that scored as high on exonic character as authentic exons but lacked evidence of splicing nearby and were flanked by regions of typically intronic character; we refer to these stretches as "exonic composition regions" (ECRs). Here, exonic or intronic character was assessed using homogeneous 5^{th}-order Markov models (Burge and Karlin, 1997) that captured the distinctive hexanucleotide (6mer) compositions of human exons and introns, but did not consider reading frame or splice site motifs. Notably, these ECRs exhibited strong enrichment for nucleosome density comparable in magnitude to that observed in authentic exons (Fig. 1E). Nucleosome enrichment was similar for ECRs located in annotated intergenic regions or intronic regions, and remained when controlling for the mappability of genomic positions and for the biased 5′ nucleotide content of ChIP-Seq reads (Fig. S1). We conclude from these observations that nucleosomes are preferentially localized to exons, and that the biased oligonucleotide content of exons can explain at least a major part of this effect. That oligonucleotide content could create such a strong bias in nucleosome position is supported by recent studies indicating that intrinsic DNA sequence preferences play a central role in determining nucleosome organization *in vivo* (Kaplan et al., 2009).

Specific histone marks are enriched on exons

In addition to nucleosome positions, the patterns of methylation marks on specific residues of the component histones can also play important roles in regulation of gene expression. These roles include demarcation of functional genomic regions and recruitment of protein factors to DNA, including both transcription and RNA processing factors (Sims et al., 2007). Based on published genome-wide histone methylation data in human T cells (Barski et al., 2007), the enrichment of specific methylation marks on exons was assessed by calculating the ratio of ChIP-Seq read density in exons to that in the flanking introns. Because many histone methylation marks show increasing or decreasing densities from beginning to end of genes, flanking intronic read density was estimated based on the average of regions located equidistantly 5′ and 3′ of each exon. Using this measure, all methylated forms of histones except H3K9me3 were significantly enriched on exons relative to flanking intronic regions (Fig. 2A; $p<0.01$ after Bonferroni correction for multiple testing, bootstrap sampling test). ChIP-Seq data for the chromatin insulator factor CTCF from the same study did not exhibit a bias toward exons relative to introns (Fig. 2A, F). Because CTCF is not associated with nucleosomes, these data serve as a type of negative control, indicating that the exonic biases observed are not simply some sort of artifact of the ChIP-Seq protocol.

Most of the observed overall average ~1.3-fold enrichment of histone marks on exons can be attributed to the increased nucleosome density on exons observed above. However, three marks in particular were enriched in exons by 1.5-fold or more, significantly exceeding the average enrichment of nucleosomes (and of histone marks overall) on exons. These marks included not only the classical transcription elongation mark H3K36me3, whose enrichment on exon-associated nucleosomes has been previously noted (Kolasinska-Zwierz et al., 2009), but also H3K27me2, which are less associated with transcription elongation. H3K27me2 has generally been associated with repressed rather than active chromatin (Barski et al., 2007). Since many of these marks show distinctive patterns within gene bodies, we investigated whether their enrichment on exons was dependent on position relative to the TSS. An overall increase in nucleosome enrichment at larger distances from the TSS was observed (Fig. S2). The H3K4me3 mark showed characteristic enrichment for both exons and introns located near the TSS, but exon:intron ratios for this mark and for other position-biased marks generally increased with distance from the TSS (Fig. S3).

The pronounced enrichment of these marks on exons suggested potential connections between RNA processing and histone methylation. For example, co-transcriptional recognition of exons at the RNA level might in some way influence methylation of specific histone residues in exon-associated nucleosomes or vice versa. Comparing histone marks in subsets of genes that were either expressed or not expressed in human T cells (based on mRNA microarray data), we observed that most histone marks exhibited similar levels of enrichment in exons independent of transcriptional activity (Fig. 2B, Fig. S4). These data suggested the possibility that differential marking of exons may not require transcription and RNA processing, but may contribute to recognition and even definition of exons at the RNA level, e.g., through direct recognition of histone marks by RNA processing factors or by factors that modify or interact with RNA processing factors. It is also possible that the observed differential marking of exons was established at an earlier stage in cellular differentiation during which these genes were expressed. In contrast, a few marks, most notably H3K27me2, were significantly less exon-enriched in genes with high expression in human T cells. One mark, H3K9me3, was unusual in being under-represented rather than over-represented in exons (Fig. 2A), suggesting that this repression-associated mark might have a different relationship to RNA processing than other marks.

ChIP-Seq reads for RNA pol II were marginally enriched on exons, with somewhat higher enrichment observed in highly expressed genes (Fig. 2), but enrichment was not significant in the data from Schones and coworkers. Pol II enrichment, if it occurs, could result from slowing of the polymerase due to presence of increased nucleosome density or specific histone marks, or to recognition of splicing-related motifs in the nascent transcript by splicing factors associated with the pol II CTD.

Isolated exons have stronger nucleosome enrichment

The specificity of exon recognition by the pre-mRNA splicing machinery is not completely understood (Wang et al., 2004). While the core splice site motifs and known splicing regulatory elements located in exons and introns play central roles in splicing specificity, these motifs do not appear sufficient to define exon locations with high accuracy. The insufficiency of known motifs is particularly acute for mammalian genes with long, multi-kilobase introns, where more information is required to distinguish authentic exons and splice sites from the larger pool of decoys (Lim and Burge, 2001; Wang et al., 2004). Notably, nucleosome enrichment was significantly greater for “isolated” exons flanked by long introns compared to “clustered” exons flanked by short introns, with both lower intronic nucleosome density and a sharper peak of exonic density observed for isolated exons (Fig. 3).

A subset of histone methylation marks also showed significantly higher enrichment in isolated exons, including both of the marks most highly enriched globally in exons –H3K27me2 and H3K36me3 – as well as H3K4me3, H3K27me1 and H3K36me1, but not the insulator element CTCF (Fig. 3A). Since the information requirements for accurate splicing of longer transcripts containing isolated exons are intrinsically higher, the increased enrichment of nucleosomes and of specific exon-associated histone marks on isolated exons represents a source of information encoded in the chromatin that would be particularly useful for ensuring accurate pre-mRNA splicing if it could be read out by the splicing machinery. Of course, the potential of marks that are extremely rare on actively expressed genes to contribute to the overall specificity of pre-mRNA splicing is less than for similarly exon-enriched marks that are abundant in expressed genes. Recognition of one of these enriched marks, H3K4me3, is known to facilitate pre-mRNA splicing, likely mediated through the CHD1 protein, which interacts both with H3K4me3 and with components of the spliceosome (Sims et al., 2007). Interestingly, H3K4me3 enrichment on isolated exons was significantly more pronounced in highly expressed genes, although considerable variance was observed when comparing highly and

low expressed isolated exons with clustered exons (Fig. S5). Previously, lower density of H3K36me3 was reported in alternative exons relative to constitutive exons (Kolasinska-Zwierz et al., 2009). However, in our analyses using larger datasets of alternative exons, significant differences in the density of histone marks in alternative relative to constitutively spliced exons were not detected (Methods).

Weak splice site exons have stronger nucleosome enrichment

Sequence features that enhance recognition of exons, including both ESEs and intronic splicing enhancers (ISEs), are common in and adjacent to constitutively spliced exons, and are particularly enriched when core splice site motifs are weaker, i.e. have below average match to the consensus (Fairbrother et al., 2002; Murray et al., 2008; Xiao et al., 2009). Considering the relationship between splice site strength and nucleosome density, a significant negative correlation between 3′ss strength and exonic nucleosome enrichment was observed (Fig. 3C; $p<0.001$, comparing strongest and weakest splice site strength bins, bootstrap sampling test). The inverse correlation with 3′ss strength persisted after controlling for splice site distance to the TSS (Fig. S6), flanking intron length, and exonic oligonucleotide composition, indicating that the association is largely independent of these variables. An inverse relationship was also observed between exonic nucleosome enrichment and 5′ss strength, though this relationship was less pronounced than for the 3′ss (Fig. S7). This inverse relationship was also apparent for H3K36me3 and H3K27me2 (Fig. S8). Thus, as for isolated versus clustered exons, a more pronounced nucleosome enrichment signal was observed for the subset of exons expected to have the greatest requirements for splicing enhancement.

Nucleosome locations enhance splicing simulation accuracy

The patterns of nucleosome enrichment on exons observed above suggested the hypothesis that nucleosome positions might contribute to recognition of exons in pre-mRNA splicing. Under this hypothesis, inclusion of nucleosome position information should improve the accuracy of algorithms that seek to simulate splicing specificity, such as ExonScan (Wang et al., 2004). For this purpose, log-odds scores were derived for specific ranges of exonic nucleosome density in a training set of 1000 genes based on the nucleosome data of Schones and coworkers (Schones et al., 2008). Application of this scoring model using empirical nucleosome densities in a separate set of ~12,800 genes yielded modest but highly significant improvements in the prediction of exon locations (Table 1). This improvement occurred whether nucleosome scoring was incorporated into models involving scoring of 5′ss and 3′ss motifs only, or using the full model that included also scoring of ESEs, ESSs, and ISEs. The latter result indicated that exonic nucleosome density provides additional information useful for exon recognition beyond that present in known splicing motifs.

Polyadenylation sites are strongly depleted of nucleosomes

Previous work has suggested connections between transcript termination, chromatin structure and histone modification (Lian et al., 2008). Additionally, a nucleosome-depleted region has been observed near the PAS in yeast (Mavrich et al., 2008). We observed a sharp dip in nucleosome signal around human PAS, extending roughly 100 bp upstream and downstream of the canonical polyadenylation signal 6mer, AATAAA (Fig. 1H). Differences in nucleosome binding affinity have been reported for distinct genomic sequences and, in particular, poly (dA:dT) stretches have low nucleosome affinity as a result of their resistance to curvature (Drew and Travers, 1985; Peckham et al., 2007; Satchwell et al., 1986). Nucleosome density plots centered at control AATAAA 6mers in intergenic regions supported that this 6mer by itself has a nucleosome positioning effect, with a dip in nucleosome density observed at the AATAAA sequence flanked by increased nucleosome density ~100 bp upstream and downstream (Fig. 1H). Controls based on other common variants of the poly(A) signal 6mer

yielded similar patterns (data not shown). However, authentic PAS differed from the controls in that the reduction in nucleosome density near the 6mer was much stronger – stronger even than the "nucleosome-free" region observed near the TSS (Fig. 1G) – and differed from the TSS distribution in that clear phasing of adjacent nucleosomes was not observed. These differences may result in part from additional sequence effects of the U-rich downstream sequence element (DSE) and/or other regulatory elements of cleavage and polyadenylation (Hu et al., 2005). Alternatively, it is conceivable that the differences could result from the presence of nucleosome-excluding DNA binding proteins if such factors commonly bound near the PAS. Both high- and low-expressed genes exhibited pronounced nucleosome depletion near the PAS, with only moderately weaker depletion in inactive genes (not shown), suggesting that the primary mechanisms responsible for PAS-associated nucleosome depletion are not dependent on expression.

Higher downstream nucleosome affinity is associated with higher PAS usage

Several thousand human genes express mRNAs with multiple distinct 3′ untranslated regions (UTRs) through regulated usage of "tandem PAS", i.e. distinct PAS located at some distance apart without intervening splicing (Wang et al., 2008). To investigate the possibility that PAS recognition and nucleosome positioning might be functionally related, the individual PAS in such pairs were designated as high usage or low usage based on available transcript data (Fig. S9). Strikingly, high usage sites displayed a significantly stronger reduction in nucleosome density immediately surrounding the PAS, and stronger nucleosome enrichment from approximately +75 to +375 downstream of the PAS ($P < 10^{-10}$ and $P < 10^{-7}$, respectively; Fig. 4A). These differences were evident even after controlling for the strength of core poly(A) sequence elements that function at the RNA level. To assess the potential contributions of intrinsic nucleosome affinity to the observed biases in nucleosome positioning relative to alternative explanations such as chromatin remodeling, a sequence-based model of nucleosome affinity was developed (Methods). This model yielded a distribution of nucleosome affinity scores (NAS) that qualitatively matched the observed distribution of nucleosome density around transcription start sites (Fig. S10). When applied to regions around tandem PAS, this model predicted a somewhat more pronounced dip in nucleosome affinity around high usage PAS than around low usage sites, and significantly stronger nucleosome affinity downstream of high usage than low usage PAS ($P < 10^{-23}$; Fig. 4B). These observations, matching the ChIP-Seq data in both aspects, indicated that sequences surrounding high usage PAS differ from those near low usage PAS in their inherent nucleosome affinity.

Discussion

Here, we have shown that the major sites of pre-mRNA processing in human genes, including both exons and the PAS, differ substantially from background levels of nucleosome density. Furthermore, more highly used alternative PAS had both higher downstream nucleosome density and higher intrinsic nucleosome affinity than less highly used alternative sites. These differences suggest that nucleosome positioning might directly influence PAS usage, e.g., through effects on the kinetics of polymerase elongation in the vicinity of the PAS, or mediated through interactions between nucleosome-associated proteins and the cleavage and polyadenylation machinery, components of which are associated with pol II (Nag et al., 2007). This possibility could be tested by inserting well-characterized nucleosome positioning elements near PAS and assessing the effects on PAS activity. It is also possible that sequence elements not included in standard core PAS scoring influence both PAS usage and nucleosome affinity. The largely expression-independent depletion of nucleosomes near the PAS does not support the alternative interpretation that components of the cleavage and polyadenylation machinery commonly alter nucleosome positions. The density of histone mark data was too low in the vicinity of the PAS to be informative about whether or not histones near sites of

cleavage and poladenylation exhibit a distinctive modification signature, but the biased distribution of histone marks observed on exons motivates investigation of this possibility. In any event, these data indicate that differences in empirical nucleosome density and/or in NAS have significant potential to predict PAS usage and alternative 3′ UTR expression.

Enrichment of nucleosomes and specific histone marks on exons has been noted in three papers published very recently (Andersson et al., 2009; Schwartz et al., 2009; Tilgner et al., 2009). While some of these works noted that the sequence composition of exons is biased in a direction that tends to favor nucleosome occupancy, our analysis of exonic composition regions in introns and intergenic regions demonstrates not only that exonic composition favors nucleosome occupancy but that the biases in oligonucleotide content of exons are sufficient to account for the magnitude of nucleosomal enrichment observed on exons (Fig. 1). The importance of this finding is that it supports models in which the biased DNA sequence composition positions nucleosomes on exons (where they could potentially modulate splicing activity) independently of transcription or RNA processing. However, this observation does not preclude the existence of additional nucleosome positioning constraints for subsets of exons, particularly those exons with weak splice sites or long flanking introns.

Here, splicing simulation algorithms were used to demonstrate that empirical nucleosome density significantly improves the accuracy of exon identification. The increase in accuracy was observed when scoring only splice site sequences. But, interestingly, the increase was also observed when known ESE, ESS, and ISE sequences were scored as well. This observation thus provides direct evidence that nucleosome positioning contains information not present in known *cis*-acting RNA elements involved in splicing.

Two types of models (not mutually exclusive) could plausibly account for the observed improvements in splicing simulation resulting from nucleosome scoring. First, the set of exonic motifs that have ESE activity at the RNA level might (coincidentally) also have high inherent nucleosome affinity at the DNA level. Under this scenario, in order to account for the improvement in accuracy observed relative to splicing models that include scoring of known ESEs, the set of ESE sequences with high nucleosome affinities would need to include a number of ESEs that have not been previously described. Second, nucleosomes might directly influence splicing, e.g., mediated through effects of nucleosomes on the kinetics of pol II transcription, or through interactions between nucleosome-associated or nucleosome-modifying proteins on the one hand and RNA splicing factors on the other (Moore and Proudfoot, 2009). This possibility could be tested through assessment of effects on splicing following manipulation of nucleosome positions in the vicinity of exons.

Tilgner and coworkers noted that exons with strong splice sites show the least nucleosome enrichment (Tilgner et al., 2009). Our results show that this inverse relationship persists even after controlling for exonic composition biases (as well as other factors; see Methods), suggesting the existence of additional influences on nucleosome positions. Several interesting possibilities could explain this result. First, intronic sequences may exist which help modulate nucleosome density in the region of exons, particularly those with weak splice sites. Second, sequences not fully captured in our Markov model of exonic nucleotide content might serve to recruit chromatin remodeling factors. The SWI/SNF chromatin remodeling complex has been reported to regulate alternative splicing (Batsche et al., 2006), although the fact that its chromatin remodeling activity appears dispensable for this regulation complicates discussion of a potential role in marking exons with weak splice sites.

Nucleosome density was inversely correlated not only with splice site strength but also with proximity to neighboring exons. This is the sort of pattern that would be expected if nucleosome occupancy enhanced exon recognition in splicing. Most vertebrate exons are recognized by

exon definition, involving recognition of pairs of splice sites across exons, a mechanism that is favored by presence of long introns (Robberson et al., 1990). Thus, nucleosome enrichment on exons might specifically facilitate recognition of exons by exon definition mechanisms, perhaps by influencing the activity of SR proteins associated with the CTD of pol II (Das et al., 2007). Because splicing can occur independently of transcription in vitro, chromatin is clearly not essential for splicing. However, transcription-coupled splicing occurs far more efficiently (Das et al., 2006) and our results suggest the chromatin structure itself may contribute to these differences.

Previous research has shown that specific changes at the chromatin level can locally affect splicing factor recruitment (Loomis et al., 2009) and splicing regulation (Allo et al., 2009; Batsche et al., 2006; Schor et al., 2009; Tyagi et al., 2009). Beyond connections to nucleosome positioning, several histone modifications were observed to differ from background nucleosome levels in exons, raising the intriguing possibility that these or other modifications directly or indirectly regulate splicing on a global scale. The depth of ChIP-Seq data presently available for individual histone marks did not seem sufficient to rigorously test potential contributions of these marks to splicing by splicing simulation analyses; this issue could be explored through manipulations of histones or histone modifying enzymes. Histone modifications represent a reversible but stable form of chemical marking that could potentially be used either to enhance the fidelity of splicing or to toggle between distinct patterns of alternative splicing, e.g., in a program of cellular differentiation. Involvement of a long-lasting mark such as histone modification in splicing control could help in situations where long-term maintenance of expression of a specific alternative isoform might be desirable, e.g., in the context of immune memory or definition of cellular identity (Wojtowicz et al., 2007).

Because chromatin structure impacts mutation rates, the biased distribution of nucleosomes relative to exons has important evolutionary implications. Recently, a pattern has been observed in which positions with higher nucleosome occupancy had higher rates of substitutions but lower rates of insertions and deletions than adjacent positions with lower nucleosome density in the Japanese killifish (Sasaki et al., 2009). Thus, the association of nucleosomes with exons is expected to exert a protective effect on coding regions, lowering the rate of potentially reading frame-disrupting insertions/deletions relative to less disruptive substitution mutations.

Experimental Procedures

ChIP-Seq Datasets

We analyzed two previously published ChIP-Seq datasets: histone methylation marks in human T cells (Barski et al., 2007) and nucleosome positioning data in human T cells (Schones et al., 2008). We chose only internal exons and ensured flanking introns were at least 500 bp long. For a given genomic position, we calculated the read coverage as the number of reads mapping upstream (on the + strand) at −73 bp and downstream (on the − strand) at +73 bp, corresponding to average nucleosome dyad positions. Because reads were only mapped to unique positions in the genome, we computed densities as a ratio of reads per unique genomic position. To reduce the impact of potential PCR amplification biases, the read count for any specific read sequence was truncated at 10. Nucleosome density was smoothed using a sliding window of size 25 or 50 bp.

Exon Analyses

Certain nucleotides were overrepresented in the first few bases at the 5′ ends of sequencing reads (see Fig. S1). These biases are likely to result primarily from aspects of MNase digestion (Johnson et al., 2006) or other technical factors. To control for this technical bias, read counts

were normalized as follows. Overrepresentation of each 5′ pentamer in a library was estimated as the ratio of the number of occurrences at the 5′ ends of all reads to the average number of occurrences of that pentamer at positions 15–25 downstream, and read counts were normalized accordingly. This control moderated the sharp peaks at the 3′ss and 5′ss but had little overall effect on the nucleosome densities around exons. Results were largely unchanged when reads that mapped to the most homogeneous positions around 5′ and 3′ splice sites, including the conserved 5′ splice site GT and 3′ splice site AG dinucleotides (Fig. 2), were removed.

ECRs were derived from intronic regions lacking exons within 2 kb or intergenic regions with no cDNA/EST coverage that were at least 1 kb from the nearest gene annotation. We randomly chose 20,000 5′ss and 20,000 3′ss and matched 9-mer sequences from those splice sites to intronic or intergenic regions to define decoy sites. To define pseudo-exons based on nucleotide content, we generated a 5^{th}-order Markov model to score exonic vs intronic sequence composition, and identified intergenic regions which closely matched authentic exons in length and 6mer composition. ECRs defined based on exonic 5mer, 4mer, 3mer or even 2mer composition exhibited peaks of nucleosome density qualitatively similar to those observed in Figure 1e (not shown). This observation indicates that the nucleotide and dinucleotide content of exons is sufficient to explain, at least qualitatively, why nucleosomes are biased toward exons. The distribution of histone marks in exons was explored by comparing read densities within exons to read densities in the flanking introns. The entire exon was included as well as the most proximal 10 bp of the upstream and downstream introns. Intronic densities were calculated from the regions 200 to 300 bp upstream of the 3′ss and 200 to 300 bp downstream of the 5′ss. The choice of both upstream and downstream intronic regions helped control for changes in density of some histone marks along the length of transcripts. Average read densities were determined as the total read counts across 69,000 exons divided by the total number of unique positions. Confidence intervals and p-values were produced by bootstrap sampling. In Figure 2b, genes were ranked by microarray expression signal (resting T cell data from Schones and coworkers). The top 10% and bottom 10% were defined as highly and lowly expressed genes, respectively.

Internal exons with both flanking introns of size between 500 and 1000 bp were defined as clustered exons, and those with both flanking introns of size at least 5 kb were defined as isolated exons. We analyzed a like number of isolated and clustered exons, sampled to match exon length between the two sets.

Splice site strength was scored using the maximum entropy-based log-odds scoring method (Yeo and Burge, 2004), and all 5′ss and 3′ss scores were required to be non-negative. Exons were divided into 5 equally sized bins based on 3′ss score, and exons were sampled from each bin to match flanking intron size and average exonic nucleotide composition. Exon:intron ratios, confidence intervals and p-values were calculated as in Figure 2.

Alternative splicing analysis

Nearly 600 sets of adjacent exon triples were identified, where the first and third exons are constitutively spliced and the middle exon is skipped in a subset of ESTs. Similar to Kolasinska-Zwierz and coworkers (Kolasinska-Zwierz et al., 2009), we calculated read densities for the central skipped exons, normalized to the read densities of the adjacent constitutive exons. No significant difference was observed for any histone mark when compared to a like number of control triples each consisting of three adjacent, constitutively spliced exons, matched for length and oligonucleotide composition (data not shown).

Splicing simulation analyses

The ExonScan algorithm (Wang et al., 2004) was modified to perform exon predictions using nucleosome density information. A training set of 1000 randomly chosen genes was used to estimate log-odds scores distinguishing correctly and incorrectly predicted exons based on their nucleosome densities. This model was then applied to a set of 12,585 known genes with no evidence of alternative splicing or alternative overlapping transcripts in Refseq (Pruitt et al., 2005). As there is a significant amount of noise in the exonic nucleosome read counts (because of their small average size), the nucleosome scoring model was applied only to exons with at least 50 mappable genomic positions. To estimate the significance of improvements in exon prediction based on nucleosome densities, we compared accuracy when nucleosome density was scored normally to simulations in which the scores of nucleosome density in random genic regions of the same length were assigned to exons. Receiver operator characteristics were calculated using the R package ROCR [www.r-project.org].

Poly(A) Analyses

Genome-wide sequence alignments of available cDNAs and ESTs were obtained from the University of Santa Cruz Genome Browser Database. Uniquely mapping cDNAs and ESTs were filtered for evidence of a non-genomically derived poly(A) tail and a canonical or variant poly(A) signal (Beaudoing and Gautheret, 2001) in the −1 to −40 region upstream of the aligned poly(A) site (Fig. S9). The resulting set was then mapped to a comprehensive and non-redundant set of Refseq transcripts (Pruitt et al., 2005) and clustered to create a database of polyadenylation sites. Sites with usage were defined as those supported by greater than 70% of the gene's mapped polyadenylated ESTs, whereas low usage sites were defined as those having less than 30% of the supporting ESTs.

Weight matrix models of core poly(A) motifs described by Hu and coworkers (Hu et al., 2005) were obtained as a part of their Polya_svm distribution, http://exon.umdnj.edu/polya_svm/. The output of polya_svm.pl run in matching-element-mode was parsed to obtain scores for each poly(A) cis-element. The core poly(A) motif score was then reported as the sum of the score for the CUE2 element, corresponding to the poly(A) signal, and the average score for the CDE1-CDE4 elements, corresponding to the U-rich downstream signals.

Nucleosome affinity scores

A total of ~84 million Illumina read starts, representing 75% of the perfectly and uniquely mapping reads in the Barski and coworkers (Barski et al., 2007) data set, were chosen at random for the nucleosome training set. An equally sized background set was obtained by randomly sampling a position within +/− 500 bp of each of the read starts in the nucleosome training set (excluding sites mapped by other read starts in the nucleosome training set). Using these data, a 5th-order Markov model was trained for every position *n* in the nucleosome occupied region (or control region), such that we obtained $P(X_n = x \mid X_{n-1} = x_{n-1},\ldots, X_{n-5} = x_{n-5})$ for every $n = 1,\ldots,146$, and for every combination of x, x_{n-1}, x_{n-2}, … = A, C, G, T. Due to the aforementioned 5′ nucleotide bias in the sequencing data, positions 1 to 15 were subsequently excluded from the model. Nucleosome affinity scores were calculated as the $\log_2$ ratio of P (seq|nucleosome model) to P(seq|background model). Scores were plotted at a 73 bp offset to reflect the center of the corresponding nucleosome.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank G. Frampton, C. Lin, R. Sandberg, P. Sharp, and R. Young and members of the Burge lab for helpful comments and discussions. This work was supported by a fellowship from the NSERC (C. B. N.), by an NIH training grant (N. S.) and by grants from the NIH (C. B. B.).

References

Allo M, Buggiano V, Fededa JP, Petrillo E, Schor I, de la Mata M, Agirre E, Plass M, Eyras E, Elela SA, et al. Control of alternative splicing through siRNA-mediated transcriptional gene silencing. Nat Struct Mol Biol 2009;16:717–724. [PubMed: 19543290]

Andersson R, Enroth S, Rada-Iglesias A, Wadelius C, Komorowski J. Nucleosomes are well positioned in exons and carry characteristic histone modifications. Genome Res 2009;19:1732–1741. [PubMed: 19687145]

Baldi P, Brunak S, Chauvin Y, Krogh A. Naturally occurring nucleosome positioning signals in human exons and introns. J Mol Biol 1996;263:503–510. [PubMed: 8918932]

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. High-resolution profiling of histone methylations in the human genome. Cell 2007;129:823–837. [PubMed: 17512414]

Batsche E, Yaniv M, Muchardt C. The human SWI/SNF subunit Brm is a regulator of alternative splicing. Nat Struct Mol Biol 2006;13:22–29. [PubMed: 16341228]

Beaudoing E, Gautheret D. Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. Genome Res 2001;11:1520–1526. [PubMed: 11544195]

Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol 1997;268:78–94. [PubMed: 9149143]

Das R, Dufu K, Romney B, Feldt M, Elenko M, Reed R. Functional coupling of RNAP II transcription to spliceosome assembly. Genes Dev 2006;20:1100–1109. [PubMed: 16651655]

Das R, Yu J, Zhang Z, Gygi MP, Krainer AR, Gygi SP, Reed R. SR proteins function in coupling RNAP II transcription to pre-mRNA splicing. Mol Cell 2007;26:867–881. [PubMed: 17588520]

de la Mata M, Alonso CR, Kadener S, Fededa JP, Blaustein M, Pelisch F, Cramer P, Bentley D, Kornblihtt AR. A slow RNA polymerase II affects alternative splicing in vivo. Mol Cell 2003;12:525–532. [PubMed: 14536091]

Denisov DA, Shpigelman ES, Trifonov EN. Protective nucleosome centering at splice sites as suggested by sequence-directed mapping of the nucleosomes. Gene 1997;205:145–149. [PubMed: 9461388]

Drew HR, Travers AA. DNA bending and its relation to nucleosome positioning. J Mol Biol 1985;186:773–790. [PubMed: 3912515]

Fairbrother WG, Yeh RF, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. Science 2002;297:1007–1013. [PubMed: 12114529]

Gornemann J, Kotovic KM, Hujer K, Neugebauer KM. Cotranscriptional spliceosome assembly occurs in a stepwise fashion and requires the cap binding complex. Mol Cell 2005;19:53–63. [PubMed: 15989964]

Howe KJ, Kane CM, Ares M Jr. Perturbation of transcription elongation influences the fidelity of internal exon inclusion in Saccharomyces cerevisiae. RNA 2003;9:993–1006. [PubMed: 12869710]

Hu J, Lutz CS, Wilusz J, Tian B. Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. RNA 2005;11:1485–1493. [PubMed: 16131587]

Johnson SM, Tan FJ, McCullough HL, Riordan DP, Fire AZ. Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin. Genome Res 2006;16:1505–1516. [PubMed: 17038564]

Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, Segal E. The DNA-encoded nucleosome organization of a eukaryotic genome. Nature 2009;458:362–366. [PubMed: 19092803]

Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J. Differential chromatin marking of introns and expressed exons by H3K36me3. Nat Genet 2009;41:376–381. [PubMed: 19182803]

Kornblihtt AR, de la Mata M, Fededa JP, Munoz MJ, Nogues G. Multiple links between transcription and splicing. RNA 2004;10:1489–1498. [PubMed: 15383674]

Li B, Carey M, Workman JL. The role of chromatin during transcription. Cell 2007;128:707–719. [PubMed: 17320508]

Lian Z, Karpikov A, Lian J, Mahajan MC, Hartman S, Gerstein M, Snyder M, Weissman SM. A genomic analysis of RNA polymerase II modification and chromatin architecture related to 3′ end RNA polyadenylation. Genome Res 2008;18:1224–1237. [PubMed: 18487515]

Lim LP, Burge CB. A computational analysis of sequence features involved in recognition of short introns. Proc Natl Acad Sci U S A 2001;98:11193–11198. [PubMed: 11572975]

Lin S, Coutinho-Mansfield G, Wang D, Pandit S, Fu XD. The splicing factor SC35 has an active role in transcriptional elongation. Nat Struct Mol Biol 2008;15:819–826. [PubMed: 18641664]

Loomis RJ, Naoe Y, Parker JB, Savic V, Bozovsky MR, Macfarlan T, Manley JL, Chakravarti D. Chromatin binding of SRp20 and ASF/SF2 and dissociation from mitotic chromosomes is modulated by histone H3 serine 10 phosphorylation. Mol Cell 2009;33:450–461. [PubMed: 19250906]

Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, Qi J, Schuster SC, Albert I, Pugh BF. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Res 2008;18:1073–1083. [PubMed: 18550805]

Moore MJ, Proudfoot NJ. Pre-mRNA processing reaches back to transcription and ahead to translation. Cell 2009;136:688–700. [PubMed: 19239889]

Murray JI, Voelker RB, Henscheid KL, Warf MB, Berglund JA. Identification of motifs that function in the splicing of non-canonical introns. Genome Biol 2008;9:R97. [PubMed: 18549497]

Nag A, Narsinh K, Martinson HG. The poly(A)-dependent transcriptional pause is mediated by CPSF acting on the body of the polymerase. Nat Struct Mol Biol 2007;14:662–669. [PubMed: 17572685]

Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, Struhl K, Weng Z. Nucleosome positioning signals in genomic DNA. Genome Res 2007;17:1170–1177. [PubMed: 17620451]

Peterson ML, Bertolino S, Davis F. An RNA polymerase pause site is associated with the immunoglobulin mus poly(A) site. Mol Cell Biol 2002;22:5606–5615. [PubMed: 12101252]

Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 2005;33:D501–504. [PubMed: 15608248]

Robberson BL, Cote GJ, Berget SM. Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol Cell Biol 1990;10:84–94. [PubMed: 2136768]

Sasaki S, Mello CC, Shimada A, Nakatani Y, Hashimoto S, Ogawa M, Matsushima K, Gu SG, Kasahara M, Ahsan B, et al. Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. Science 2009;323:401–404. [PubMed: 19074313]

Satchwell SC, Drew HR, Travers AA. Sequence periodicities in chicken nucleosome core DNA. J Mol Biol 1986;191:659–675. [PubMed: 3806678]

Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K. Dynamic regulation of nucleosome positioning in the human genome. Cell 2008;132:887–898. [PubMed: 18329373]

Schor IE, Rascovan N, Pelisch F, Allo M, Kornblihtt AR. Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing. Proc Natl Acad Sci U S A 2009;106:4325–4330. [PubMed: 19251664]

Schwartz S, Meshorer E, Ast G. Chromatin organization marks exon-intron structure. Nat Struct Mol Biol 2009;16:990–995. [PubMed: 19684600]

Sims RJ 3rd, Millhouse S, Chen CF, Lewis BA, Erdjument-Bromage H, Tempst P, Manley JL, Reinberg D. Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. Mol Cell 2007;28:665–676. [PubMed: 18042460]

Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcarcel J, Guigo R. Nucleosome positioning as a determinant of exon recognition. Nat Struct Mol Biol 2009;16:996–1001. [PubMed: 19684599]

Tyagi A, Ryme J, Brodin D, Ostlund Farrants AK, Visa N. SWI/SNF associates with nascent pre-mRNPs and regulates alternative pre-mRNA processing. PLoS Genet 2009;5:e1000470. [PubMed: 19424417]

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. Nature 2008;456:470–476. [PubMed: 18978772]

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. Systematic identification and analysis of exonic splicing silencers. Cell 2004;119:831–845. [PubMed: 15607979]

Wojtowicz WM, Wu W, Andre I, Qian B, Baker D, Zipursky SL. A vast repertoire of Dscam binding specificities arises from modular interactions of variable Ig domains. Cell 2007;130:1134–1145. [PubMed: 17889655]

Xiao X, Wang Z, Jang M, Nutiu R, Wang ET, Burge CB. Splice site strength-dependent activity and genetic buffering by poly-G runs. Nat Struct Mol Biol. 2009

Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol 2004;11:377–394. [PubMed: 15285897]

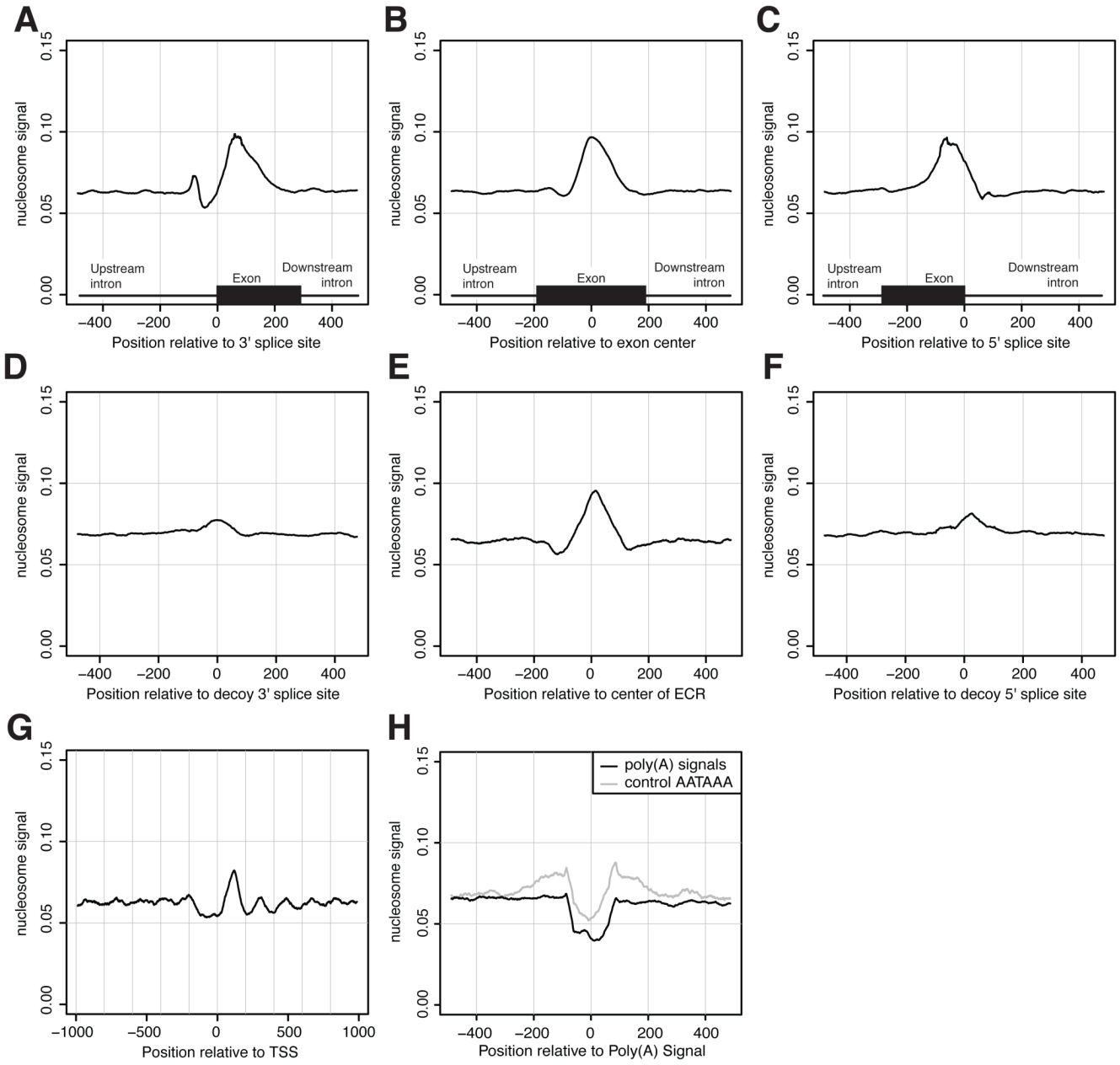

Figure 1. Nucleosome enrichment and depletion in the vicinity of core sites of RNA processing and controls

Nucleosome read signal, centered on (A) 3′ss, (B) exon centers and (C) 5′ss, with approximate exon sizes indicated by black box below. Nucleosome signal relative to (D) sequence-matched decoy 3′ss, (E) regions of exonic nucleotide composition and (F) decoy 5′ss. For reference, we have plotted nucleosome signal on the same y-axis for (G) transcription start sites of expressed genes and (H) PAS.

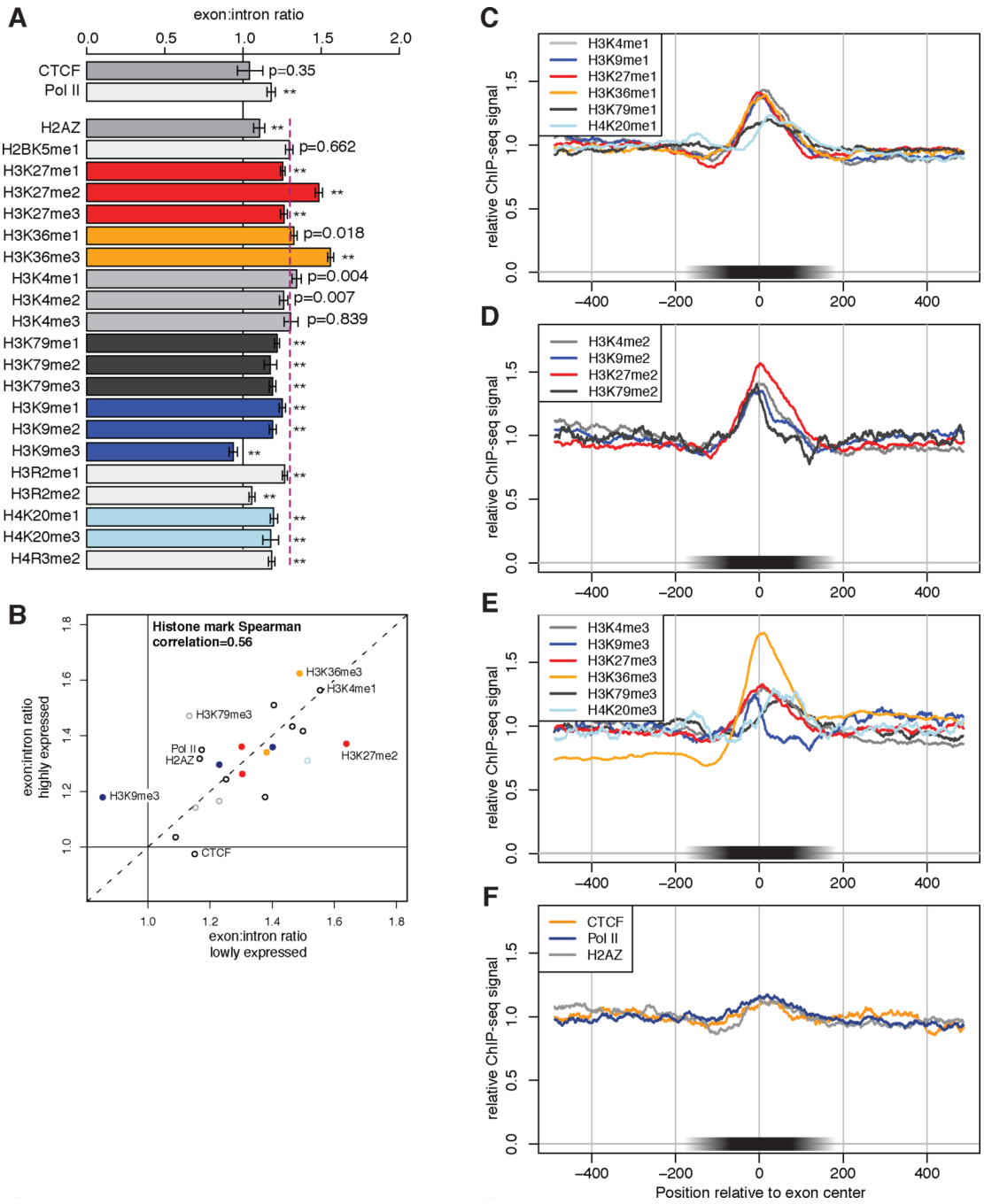

Figure 2. Exon-biased distribution of specific histone H3 methylation marks

(A) ChIP enrichment for exons, relative to flanking intronic regions (see methods), compared to 1.0 (CTCF and Pol II) or histone overall average of 1.3 (purple dashed line). Error bars are 95% confidence intervals (resampling). ** indicates p<0.01 after correction for multiple testing (resample test, Bonferroni-corrected). (B) Histone marks are similarly enriched in highly and lowly expressed genes. Profiles centered on exons for (C) mono-methyl histone marks, (D) di-methyl histone marks and (E) tri-methyl histone marks and Pol II, H2AZ and the negative control CTCF (F). (C)–(F) are normalized to average library ChIP signal across the displayed region.

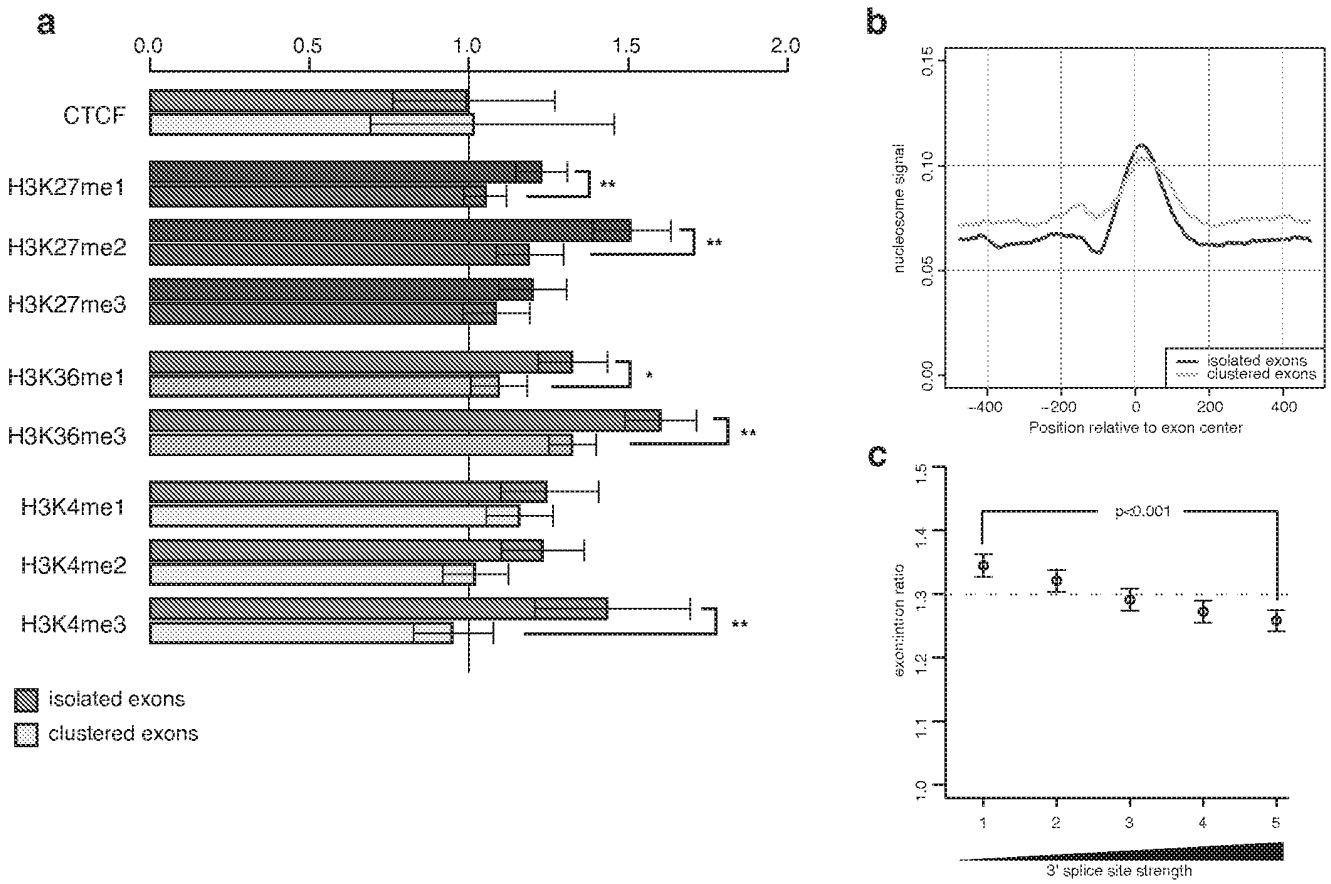

Figure 3. Increased exonic bias of specific histone H3 methylation marks in exons with long flanking introns or weaker 3′ss motifs. (A) Exon enrichment, relative to flanking introns for isolated exons (flanking introns > 5 kb, top bar of each pair) and clustered exons (flanking introns between 0.5 kb and 1.0 kb). Error bars are 95% confidence intervals. * indicates $p<0.05$ and ** indicates $p<0.01$ after Bonferroni correction for multiple testing (resample test). (B) Nucleosome signal profile for exons with short and with long flanking introns. (C) Nucleosome enrichment on exons is inversely correlated with 3′ splice site strength.

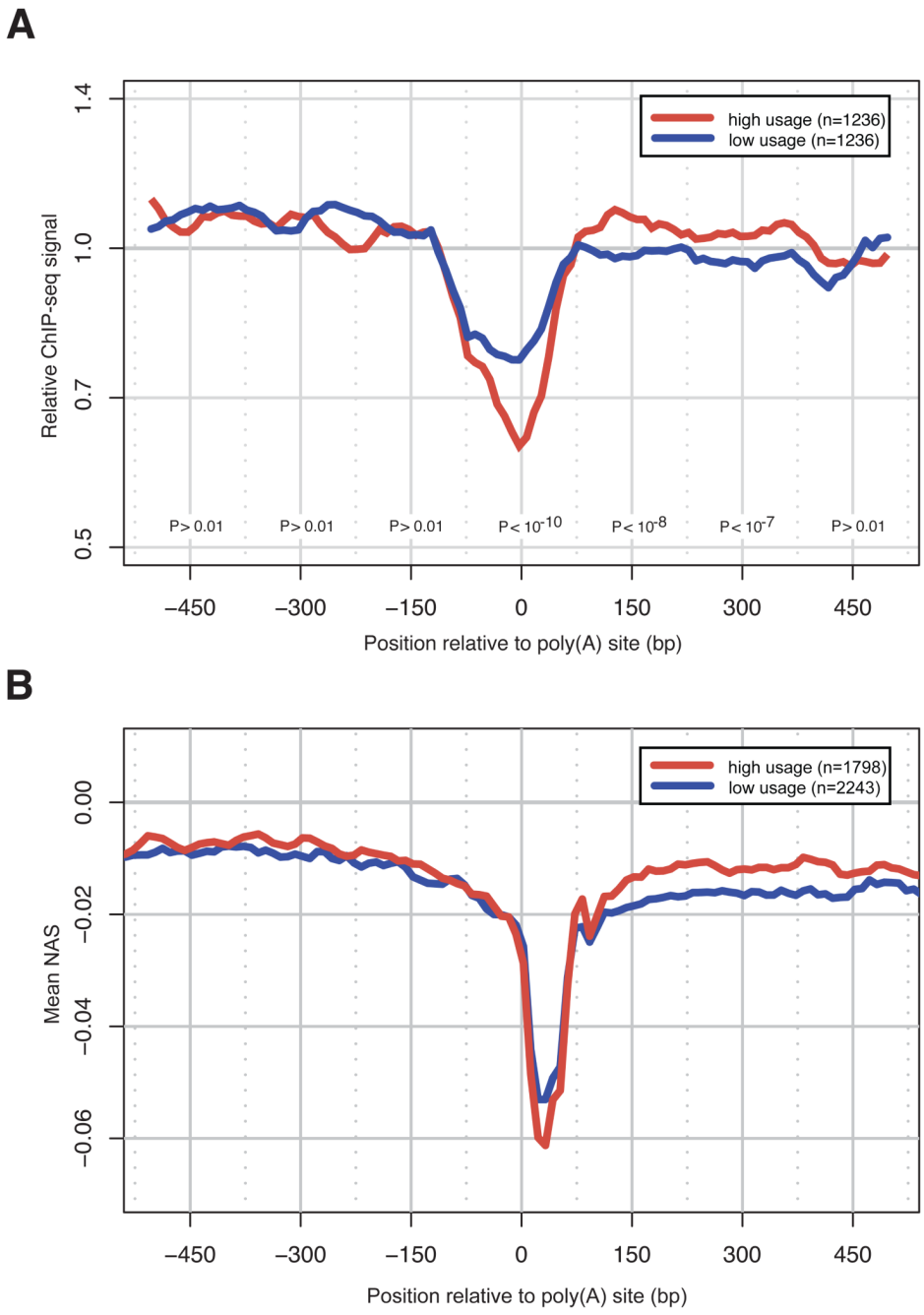

Figure 4. Nucleosome depletion and downstream nucleosome enrichment at high usage PAS

(A) Mean nucleosome density around human PAS of low (blue) or high (red) usage, normalized to average ChIP signal. (B) Mean NAS for positions around human PAS of low or high usage. Wilcoxon rank sum test p-values shown for 150 bp windows centered on indicated positions.

Table 1

Nucleosome position information improves splicing simulation accuracy.

Exon sequence features scored						Effects of nucleosome scoring		
5′ss	3′ss	ESE	ESS	ISE	no. exons correct	Net change in TP–FP	Change in AUC	p-value
+	+	−	−	−	43,567	+475	0.78%	< 0.002
+	+	+	−	−	53,880	+3,099	0.45%	< 0.002
+	+	−	+	−	54,145	+3,197	0.50%	< 0.002
+	+	+	+	−	61,998	+1,718	0.25%	< 0.002
+	+	+	+	+	63,329	+1,389	0.19%	< 0.002

P-value was estimated as the fraction of times out of 500 permutations of the nucleosome data that higher accuracy was observed. TP–FP is calculated as the difference between the number of predicted exons with at least one splice site correct (true positives) and the number of incorrect predictions (false positives). AUC, area under the curve (see Methods).