Nucleosome Distribution near the 3' Ends of Genes in the Human Genome

Huan Huang,1,2 Hongde Liu,1,2 and Xiao Sun1,2,†

1 State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, China
2 School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China

Received May 20, 2013; Accepted July 9, 2013; Online Publication, October 7, 2013
[doi:10.1271/bbb.130399]

By systematic analysis of high-throughput sequencing datasets from the human genome, we found that protein-coding genes have a specific chromatin structure near transcription termination sites relative to non-coding genes, one related to polyadenylation. Nucleosome was depleted near the site of cleavage/polyadenylation (polyA site) regardless of its relative position in the gene. DNA sequence plays an important role in nucleosome distribution, and conservative sequence elements and the protein binding to them are major determinants in causing nucleosome depletion near polyA sites. Furthermore, nucleosome occupancy was regulated by gene transcription and RNA polymerase II (RNAPII) occupancy. Our results reveal influences on nucleosome occupancy near polyadenylation sites and constitute evidence indicating that nucleosome distribution regulates 3' end processing of protein-coding genes.

Key words: nucleosome distribution; transcription termination; polyadenylation; protein-coding gene

In eukaryotes, three RNA polymerases transcribe different types of genes. While initiation and elongation of transcription have been well studied, termination was obscure until recently, especially for RNA polymerase II (RNAPII) transcription, which is responsible for protein-coding genes and many non-coding genes. RNA polymerase III (RNAPIII) and RNA polymerase I (RNAPI) transcribe other non-coding genes, and their terminations appear to be simpler than RNAPII termination. Generally speaking, the three transcription machineries use different strategies for transcription termination, although they share some common features. Some researchers have found that changes in chromatin are important for termination and suggest that RNAPII requires a specific chromatin structure to terminate efficiently.1

The formation of 3' end of precursor messenger RNA (pre-mRNA) is an essential step in the procedure of eukaryotic gene expression, and it directly impacts many other steps in the gene expression pathway, such as transcription termination, mRNA stability, and export.2-4 Inappropriate 3' end formation of human mRNAs can have a tremendous impact on health and disease,5 although their molecular modes of action remain unknown. 3' end processing of mammalian cells involves two tightly coupled steps, cleavage and polyadenylation, and requires a polyadenylation signal (PAS, AAUAAA, or something similar, generally found about 10–30 bases upstream of the site of cleavage/polyadenylation or polyA site) and a downstream sequence element (DSE, U, or a G/U rich sequence, generally found about 14–70 bases downstream of the polyA site).6 Some researchers have found that the strength of the polyA site correlates with efficient pausing-dependent termination,7,8 and the relative positions of PAS and DSE are important in determining the polyA sites and cleavage efficiency.9

Recent evidence indicates that 3' end processing is coupled to transcription,10,11 and that RNAPII plays a critical role in coordinating co-transcriptional pre-mRNA processing by the largest subunit, mediating interactions with processing factors.12 More than half of genes with multiple polyA sites are alternatively polyadenylated in the human genome,13 and polyA sites can locate in different regions of genes:3'UTR and introns or exons. Some researchers have suggested that nucleosomes have a potential effect on pre-mRNA 3' end formation and termination.14,15 It has been reported that nucleosome occupancy dropped precipitously near the polyadenylation site in many species. But the influences on nucleosome occupancy and the relationship between nucleosome organization and polyadenylation remain unclear. Our group have done a lot of work on nucleosome occupancy near key point sites, including splicing sites16 and transcription factor binding sites,17 that suggests nucleosome-regulated gene expression and the processing of pre-mRNA. We have also studied nucleosome regulation of the use of polyadenylation sites (Huan Huang, Hongde Liu, and Xiao Sun, unpublished results). In present work, by processing high-throughput experimental datasets of the human genome, we obtained further evidence that nucleosome distribution near the 3' end of a protein-coding gene is related to polyadenylation, which depends on the DNA sequence, conservative sequence elements, and protein binding and is regulated by gene transcription and by RNAPII occupancy.

Materials and Methods

Gene sets. In order to explore the relationship between chromatin structure and RNA polymerase, genes from the human genome were...
divided into three groups: (i) protein-coding genes, which often require a polyadenylation site and a downstream terminator sequence (23,870 genes), (ii) non-coding genes transcribed by RNAPII (2,068 genes including snRNAs, snoRNAs, and miRNAs precursors), and (iii) non-coding genes transcribed by RNAPI or RNAPIII (3,127 genes including rRNAs, tRNAs, and U6 spliceosomal snRNA genes). The genes were downloaded from RefSeq Genes and RNA Genes of hg18 database (UCSC, http://genome.ucsc.edu/cgi-bin/hgTables), and transcription termination site (TTS) of the gene came from gene annotation information of hg18 database (UCSC).

Pol(A) site dataset. The genomic coordinates of the Pol(A) sites as well as the locations of the PAS elements of the human genome were obtained from the PolyA_DB2 database. We used the Batch Coordinate Conversion (liftOver) tool of the UCSC Genome Bioinformatics resource to remap the Pol(A) sites from NCBI/hg17 to NCBI/hg18 (http://genome.ucsc.edu/cgi-bin/hgLiftOver). All pol(A) sites not mapped uniquely to the reference gene were removed. The number of pol(A) sites in 3’UTR is 30,032. There are also 7,882 pol(A) sites in introns and 2,772 pol(A) sites in exons.

The canonical PAS consists of the AAUAAA motif, and most base substitutions in this sequence, except for the AAUAAA variant, have been found to have significantly reduced cleavage and polyadenylation efficiency. According to the conservation of PAS applying the expectation maximization approach, Akhtar et al. differentiated poly(A) sites into three classes: (i) PAS-strong poly(A) sites with the AAUAAA or AUUAAA motif, (ii) PAS-weak poly(A) sites with the remaining 10 forms of the PAS-motif: AGUAAA, UAUAUA, CAUAUA, GAUUAU, AUAAUA, AAUAAC, ACAUUA, AAGAAA, and AAUGAA, and (iii) PAS-less poly(A) sites that do not show any of these 12 forms. By this criterion, 26,029 PAS-strong poly(A) sites, 8,071 PAS-weak poly(A) sites, and 6,632 PAS-less poly(A) sites were obtained.

DSE also plays important roles in cleavage and polyadenylation efficiency. By detecting 10–80 bp downstream poly(A) sites with a sliding window, we found 33,256 poly(A) sites with U-rich sequences that consisted of five nucleotides, at least four of which were uridine, and 7,476 U-less sequences. Some poly(A) sites (n = 1,334) had a GU-rich element that included that sequence YGGUUGYY (Y = pyrimidine).

Nucleosome level and RNAPII occupancy. Sequencing datasets for nucleosome positions in CD4+ T cells, granulocytes, and in vitro, which were generated by high-throughput SOLiD technology, were obtained from NCBI, accession no. GSE25133. In addition, sequencing datasets for nucleosome positions in the GM12878 ENCODE cell line (NCBI, accession no. GSE35586) and activated human CD4+ T cells, and fetal fibroblast cell line IMR90 were obtained as supplement datasets. The tag coordinate bed file of RNAPII in human CD4+ T cells using Solexa sequencing technology was also used.

The nucleosome score was calculated as described previously. They represent average numbers of sequenced reads uniquely mapping to the sense strand 80 bases upstream and to the antisense strand 80 bases downstream with a step size of 10 bp. Nucleosome occupancy was calculated in alignment with the TTS or the poly(A) site, and the data were smoothed using a moving average filter. The moving average span was 5. The Wilcoxon rank sum test was employed to gauge the significance (p). RNAPII occupancy was calculated similarly to nucleosome occupancy, but the score represents the average number of sequenced reads uniquely mapped to the sense strand 300 bases upstream and to the antisense strand 300 bases downstream of one locus.

The predicted nucleosome score was calculated by means of nucleosome-prediction software developed by the Segal Laboratory (http://geniec.weizmann.ac.il/pubs/nucleosomes08/), which is based solely on DNA sequences.

Analysis of gene expression levels using mRNA-seq data. mRNA-seq data on nuclear extracts from human CD4+ T cells and granulocytes were obtained from NCBI’s Gene Expression Omnibus, accession no. GSE25133. We quantified transcriptional levels in reads per kilobase of the exon model per million mapped reads (RPKM), calculated following Valouev et al. 2011,23 with a modification that adjusted for transcript length because of the lack of mappings across splice junctions, calculated according to the formula \( L' = L - 50 \times (E - 1) \), where \( L \) is the transcript length, and \( E \) is the number of exons in the gene.

Results and Discussion

Different nucleosome distribution patterns near transcription termination sites of various genes

Some researchers have found that the chromatin structure near the transcription termination site of a gene is specific and that nucleosome occupancy dropped precipitously. In order to explore further the relationship between chromatin structure and transcription termination, nucleosome distributions in various types of genes were assessed by examining nucleosome occupancy profiles across a 2,000-nucleotide (nt) window surrounding the TTSs (Fig. 1, Supplemental Fig. 1; see Biosci. Biotechnol. Biochem. Web site). Three types of genes (see Materials and Methods) showed differential nucleosome distributions near the TTS. Strikingly, protein-coding genes displayed a significantly nucleosome-depleted region (NDR) upstream of TTSs and stronger nucleosome occupancy immediately downstream of TTSs than non-coding genes in various cells and in vitro, suggesting that nucleosome occupancy around TTSs is not associated with the type of cells, and is determined by the DNA sequence to some extent. The peak of nucleosome occupancy immediately downstream of TTSs in vivo was more marked than that in vitro, suggesting that in vivo environment plays an important role in positioning the nucleosome downstream of NDR. In addition, the profiles of nucleosome distribution had clearly peaks upstream of TTSs and had lower nucleosome levels downstream of the TTS of the non-coding genes. As the average length of the non-coding genes was 115 bp, we deduced that the nucleosomes upstream of TTSs occupy the non-coding gene body to protect this key region. Nucleosome distributions in protein-coding genes and non-coding genes transcribed by RNAPII show clear differences, suggesting that RNAPII is not a sufficient condition for nucleosome distribution around TTSs, and nucleosome

![Fig. 1. Nucleosome Occupancy Surrounding Transcription Termination Sites (TTSs) of Genes Transcribed by Various Polymerases in Vivo and in Vitro.](image-url)

Solid black curve represents protein-coding genes transcribed by RNAPII; black dashed curve represents non-coding genes transcribed by RNAPII; gray curve represents non-coding genes transcribed by RNAPI or by RNAPIII.
distributions might be associated with different transcription termination mechanisms for protein-coding genes and non-coding genes, although the details of the mechanisms remain unclear.

RefSeq genes are manually curated based on high-quality full-length cDNA sequences and aligned onto the genome to generate the RefGene table, and hence most of the protein-coding gene end coordinates in the RefGene table are exactly at the genome position where the polyA tail in the full-length cDNA cannot be aligned, that is, the polyA site of the full-length cDNA. Therefore, the TTSs are all major polyA sites. We conjecture that nucleosome distribution in protein-coding genes is related to polyadenylation. Some researchers have also found chromatin remodeling factors in mammalian termination, including the region downstream of the polyA site,34,35 suggesting that RNAPII requires a specific chromatin structure to polyadenylate and terminate efficiently. In protein-coding genes, nucleosomes positioned downstream of the TTSs might act as RNAPII transcription pause sites and, by slowing down elongation speed, allow the RNAPII more time to process 3′ end events. The termination of non-coding genes appear simpler than protein-coding genes termination. Non-coding genes have special genetic structures or RNAPs themself has special subunits to promote the termination of transcription. For example, RNAPIII has subunits C37/C53 to reduce the elongation rate of RNAPIII and to allow for an increased pausing time at the terminator elements.

Nucleosome distribution near the polyA site in different regions of genes

As for protein-coding genes, more than half of the mRNAs in the human genome are alternatively polyadenylated, and the position of alternative polyA sites can be in different regions of the genes.3) To assess the specificity of nucleosome distribution around polyA sites according to the relative position of the polyA site in a given gene, polyA sites were divided into three classes: in 3′ UTR, in intron, and in exon. Consistently with previous reports,14,29,33 deep troughs of nucleosome occupancy profiles near the polyA sites were also observed in vivo and in vitro, suggesting that this is general and not associated with the state of cells, and partly based on the DNA sequence (Fig. 2, Supplemental Fig. 2). The nucleosome level near the polyA sites in the 3′ UTR was lowest in vivo but not in vitro, indicating that other determinants affect it. In addition, the results revealed a marked difference in nucleosome occupancy between polyA sites in exons and polyA sites in introns and 3′ UTRs. The overall level of nucleosomes near polyA sites in the exons was higher than the others, and this is potentially related to exon recognition,16) and had a trough relative to the center of the exons (Supplemental Fig. 3A). In the introns, the profile of nucleosome distribution was flat around the centers of the introns, but there were two peaks and a trough around the polyA sites in the introns (Supplemental Fig. 3B).

In order to go deeper into the determinants of nucleosome positioning, the distance between the polyA sites and the two ends of the exons or introns was measured (Supplemental Fig. 3C and D). About 35% of the polyA sites in the exons were close to the 3′ end (<50 bp), and about 28% of the polyA sites in the exons were far from the 3′ end of the exons (>500 bp). The distance from the polyA sites to the 5′ end of the exons was various. In the introns, nearly half of the polyA sites were close to the 5′ end of the introns (<1,000 bp), and the others were various. These results indicate that the influence of the special position of the polyA site in exon/intron on nucleosome occupancy was weak and that the principal element was polyadenylation itself. The profiles of nucleosome distribution near polyA sites within different regions were discrepant in different cell types, but there were also significant similar patterns that might play important roles in 3′ end processing. Further investigation indicated that nucleosomes were depleted near polyA sites regardless of their relative positions in the genes, and that the DNA sequence plays an important role in nucleosome distribution, indicating that the chromatin structure regulates polyadenylation or that polyadenylation affects nucleosome occupancy.

Nucleosome occupancy near polyA sites was related to conservative sequence elements

In order to determine the influence of conservative sequence elements on nucleosome occupancy, the nucleosome level was also examined across a 2,000-nt window surrounding polyA sites with contrasting conservation of sequence elements (Fig. 3, Supplemental Fig. 4). It was observed that nucleosome depletion near
the polyA site was strongly positively correlated with the conservation of PAS in vivo, but in vitro the difference as to nucleosome occupancy was not obvious as between the PAS-strong and the PAS-weak polyA sites, which had similar nucleosome disfavoring sequences and more strongly depleted nucleosomes than the PAS-less polyA sites. There was also a slightly negative correlation between nucleosome occupancy and the conservation of DSE in vivo (Fig. 4, Supplemental Fig. 5). In vitro, the difference as to nucleosome occupancy was also not obvious as between the U-rich and the GU-rich polyA sites, which had more strongly depleted nucleosomes than the U-less polyA sites.

The precise site of cleavage within a region is defined by the PAS, and is determined by the first nucleotide upstream of the cleavage site in the order of preference A > U > C > G. Nucleosome occupancy was examined surrounding polyA sites with different sequence preferences for the first nucleotide upstream of the cleavage site (Supplemental Fig. 6). Although the cleavage site measurement was not very accurate and there was not any obvious consistency between nucleosome occupancy and sequence preference, nucleosome occupancy near the polyA sites with weak hydrogen bonding local sequences (A/U) was weaker than that with strong hydrogen bonding local sequences (C/G), and showed negative correlations with sequence preference at the cleavage site. This suggests that nucleosomes depleted near polyA sites are related to the hydrogen bonding strength of the first nucleotide upstream of the cleavage site to some extent.

It has been found that DNA sequences rich in AT disfavor core histones. In order to determine further the effect of DNA sequences, GC content was calculated around polyA sites (Supplemental Fig. 7). The lowest GC content about 20 bp upstream of the polyA site appeared near the position of PAS, and near the site of cleavage/polyadenylation there was a peak of the GC content profile. Then the GC content is lower near DSE. There was a low GC content near polyA sites, consistently with nucleosome depletion. These results suggest that GC content plays a crucial role in nucleosome occupancy. We conjectured that nucleosome disfavoring sequence-PAS plays an important role in causing nucleosome depletion near polyA sites, and that DSE and the first nucleotide upstream of the cleavage site are also auxiliary determinants, since the more conservative sequence elements are, the more proteins bind easily to them. Protein binding to conservative sequence elements and chromatin remodelers might also regulate nucleosome levels by preventing histones from binding to DNA in vivo.

**Nucleosome occupancy near polyA sites was related to gene transcription and RNAPII occupancy**

Some research suggests that polyA site usage is coupled to transcriptional activity, which has an additional impact on nucleosome levels. In order to explore the effect of gene transcription on nucleosome occupancy, the nucleosome levels around polyA sites of genes with different expression level were analyzed (Fig. 5A and B). According to gene expression values (RPKM), polyA sites were divided into two groups: polyA sites of highly and lowly expressed genes. The highly expressed genes (RPKM > 10) had lower nucleosome levels near polyA sites than the lowly expressed genes. The highly expressed genes (RPKM < 0.1) consistently with previous studies of proximal and distal polyA sites. In addition, RNAPII occupancy was examined across a 2,000-nt window surrounding polyA sites of highly and lowly expressed genes in CD4+ T cells (Fig. 5C). The polyA sites of high expressed genes had lower nucleosome occupancy and higher RNAPII enrichment than the lowly expressed genes, and RNAPII occupancy near the PAS was lower than downstream and upstream of the polyA sites, suggesting that RNAPII usually pauses before and after transcription passes the polyA sites and that nucleosomes positioned upstream and downstream of polyA sites serve as speed-bumps, slowing down transcription rates to improve polyA site recognition. Our results indicate that nucleosome occupancy near polyA sites was inversely correlated with gene transcription, and that RNAPII is a determining factor of nucleosome occupancy.
Conclusion

We found nucleosome distribution around TTSs of different types of genes and polyA sites of protein-coding genes in the human genome, and several lines of evidence indicating that nucleosome occupancy regulates the 3' end processing of protein-coding genes. The largely different profiles of nucleosome distribution as between protein-coding genes and non-coding genes revealed in our study may be responsible for a required specific chromatin structure for the protein-coding gene to terminate efficiently. Further investigation indicated that nucleosomes are also depleted near polyA sites regardless of their relative positions in the gene and that the DNA sequence plays an important role in nucleosome distribution, indicating that chromatin structure regulates polyadenylation, or vice versa. Conservative sequence elements and chromatin remodelers can enhance the barrier to the positioning of nucleosomes, while DSE and the first nucleotide upstream of the cleavage site were auxiliary determinants. Protein binding to conservative sequence elements and chromatin remodelers can cause nucleosome depletion near polyA sites, while DSE causes the formation of nucleosome arrays. Gene transcription and RNAPII occupancy also had an effect on nucleosome positioning.

Taken together, our results indicate that nucleosome distribution near the 3' end of protein-coding genes depends on the DNA sequence, especially for polyA sites, conservative sequence elements, and protein binding, producing barriers, and that these are determinates of nucleosome occupancy. Furthermore, nucleosome occupancy was regulated by gene transcription and RNAPII occupancy. Nucleosome occupancy near TTSs and polyA sites can regulate transcription termination and the polyadenylation of protein-coding genes by slowing down the transcription speed and recruiting 3' processing factors.

Acknowledgments

This work was supported by grants from the National Basic Research Program of China (no. 2012CB316501), the National Natural Science Foundation of China (61073141), and the Leading Science Project of Southeast University.

References