BIOINFORMATICS DISCOVERY NOTE

Vol. 29 no. 14 2013, pages 1713–1717
doi:10.1093/bioinformatics/btt288

Genome analysis

Advance Access publication June 5, 2013

Alternative polyadenylation sites reveal distinct chromatin accessibility and histone modification in human cell lines

Che-yu Lee and Liang Chen*

Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA

Associate Editor: Inanc Birol

ABSTRACT

Motivation: In addition to alternative splicing, alternative polyadenylation has also been identified as a critical and prevalent regulatory mechanism in human gene expression. However, the mechanism of alternative polyadenylation selection and the involved factors is still largely unknown.

Results: We use the ENCODE data to scan DNA functional elements, including chromatin accessibility and histone modification, around transcript cleavage sites. Our results demonstrate that polyadenylation sites tend to be less sensitive to DNase I. However, these polyadenylation sites have preference in nucleosome-depleted regions, indicating the involvement of chromatin higher-order structure rather than nucleosomes in the resultant lower chromatin accessibility. More interestingly, for genes using two polyadenylation sites, the distal sites show even lower chromatin accessibility compared with the proximal sites or the unique sites of genes using only one polyadenylation site. We also observe that the histone modification mark, histone H3 lysine 36 tri-methylation (H3K36Me3), exhibits different patterns around the cleavage sites of genes using multiple polyadenylation sites from those of genes using a single polyadenylation site. Surprisingly, the H3K36Me3 levels are comparable among the alternative polyadenylation sites themselves. In summary, polyadenylation and alternative polyadenylation are closely related to functional elements on the DNA level.

Contact: liang.chen@usc.edu

Received on February 22, 2013; revised on May 15, 2013; accepted on May 16, 2013

1 INTRODUCTION

Post-transcriptional RNA processing, including 5′ capping, intron splicing, RNA editing and polyadenylation [poly (A)], has been recognized as important regulation steps of gene expression. Alternative polyadenylation (APA) is a widespread mechanism in eukaryotic cells to control gene expression by producing transcript isoforms with different poly (A) sites. It has been reported that ~30% of human genes can select APA sites in their 3′-untranslated regions (UTRs) (Lin *et al.*, 2012). Even though most of the genes with APA produce identical proteins, APA can also alter gene expression by changing microRNA-binding regions (Mayr and Bartel, 2009), transportation and protein production rates (Lutz, 2008). Recent research provides evidence that genes have preference for isoforms with shorter 3′-UTRs in embryonic or cancer cells because these isoforms have higher translational efficiency than longer isoforms (Ji *et al.*, 2009; Lin *et al.*, 2012; Shepard *et al.*, 2011). The study of polyadenylation in five mammals demonstrated that the usage of poly (A) sites is more conserved in the same tissue across different species than within a species (Derti *et al.*, 2012). All of these reflect the functional and evolutional importance of APA regulation.

In general, poly (A) sites contain two core elements, the conserved AAUAAA hexamer in the upstream and the GU-/U-rich region in the downstream. The AAUAAA, referred as polyadenylation signal (PAS), is located ~10–35 nucleotides upstream of the cleavage site and acts as the binding motif for the cleavage and polyadenylation specificity factor (CPSF). The GU- or U-rich region, located 20–40 nucleotides downstream of poly (A) sites, is recognized by the cleavage stimulatory factor (CstF). CstF is recruited by CPSF and assembled into a protein complex to process polyadenylation (Edmonds, 2002). The high evolutional conservation of these two motifs indicates their critical roles in polyadenylation.

To investigate other factors related to polyadenylation and the alternative selection of poly (A) sites, we combined different types of data from the ENCODE project (Dunham *et al.*, 2012), including RNA-PET data, DNase-seq data, MNase-seq data and ChIP-seq data. We observed the low chromatin accessibility around poly (A) sites, especially around the distal poly (A) sites of genes using two cleavage sites. This low chromatin accessibility is not because of the high density of nucleosomes. Conversely, poly (A) sites prefer nucleosome-depleted regions. In addition, the higher level of tri-methylated histone H3 at lysine 36 (H3K36me3) was observed around APA sites compared with unique poly (A) sites, suggesting the role of histone modification in simultaneously deploying multiple poly (A) sites for a single gene. However, no significant difference was detected among the multiple poly (A) sites themselves.

2 METHODS

The poly (A) sites were identified from the RNA-PET data in the nucleus of cells from the ENCODE project (Djebali *et al.*, 2012). Note that RNA-PET has been used to map the 5′ cap and the 3′ poly (A) signatures of individual transcripts (Fullwood *et al.*, 2009). Here, we specifically targeted on the nuclear compartment because the cleavage of transcripts from DNA occurs in the nucleus. Ensembl gene annotations (http://genome.ucsc.edu/) were applied to obtain the largest potential genomic region of each gene. Poly (A) sites within 20 bp were considered as the

*To whom correspondence should be addressed.

same one. To capture novel poly (A) locations beyond the known gene regions, we extended each gene region to at most 3500 bp without overlapping other genes. After inferring the poly (A) sites from the RNA-PET data, the genes were classified into several groups based on the number of poly (A) sites they used for the considered cell line. In this article, we mainly focused on the comparison between genes using a single poly (A) site and genes using two poly (A) sites. HeLa-S3 cells, for example, had 1560 genes using two poly (A) sites and 3607 genes using only one poly (A) site. According to our procedures, if a gene could choose an alternative poly (A) site in a different cell line but only exhibited a single poly (A) site in the considered cell line, we counted it as a gene using only one poly (A) site. Similarly, we required that both poly (A) sites were used in the considered cell line to claim a gene using two poly (A) sites. Randomly chosen positions from the ending sites of middle exons and body regions of exons were treated as control groups. Additionally, we considered random positions in intergenic regions or pseudogenes. We applied the same examinations to six different human cell lines including HeLa-S3, HepG2, HUVEC, GM12878, NHEK and K562.

The chromatin accessibility was estimated by the DNase I hypersensitive experiments performed at Duke University in the ENCODE project. Because of the low coverage of DNase-seq data, we constructed a window of 80 bp and calculated the average DNase signal. We focused on the 160 bp regions consisting of two 80 bp windows for both the upstream and the downstream of poly (A) sites. Because of the possible relationship between nucleosome occupancy and chromatin accessibility, we investigated the nucleosome positions around poly (A) sites with the MNase-seq data that were only available in the GM12878 and K562 cell lines from the ENCODE project.

The Chip-seq data of H3K36Me3 marks from University of Washington in the ENCODE project were combined with the RNA-PET data to investigate the histone modification marks around poly (A) sites with the 80 bp-window strategy aforementioned. Genes using three APA sites were also examined to confirm our conclusions. Additionally, we examined other modification data from the ENCODE project such as H3K27me3 and H3K4me3 marks. However, no evidence of possible association between these markers and APA sites was observed. All the statistical tests were performed with the SAS programming.

3 RESULTS

Based on the Ensembl gene annotations, there were 57 892 non-overlapping genes. For each of the six considered human cell lines, we inferred the used poly (A) sites of each gene using the RNA-PET reads and classified the genes into different groups. Then the DNase-seq data reflecting the chromatin accessibility were considered on the poly (A) sites. Using the K562 cell line as an example, the DNase signals demonstrated lower chromatin accessibility around poly (A) sites than those around randomly chosen ending positions of middle exons or those around random positions inside an exon body ($P<0.0001$, Wilcoxon tests; average DNase signals were 0.010–0.015 versus 0.017–0.023) (Fig. 1A). Intriguingly, the distal cleavage sites of genes using two poly (A) sites displayed significantly lower signals than other poly (A) sites ($P=0.0001$–0.0078, Wilcoxon signed-rank tests or Wilcoxon tests; average DNase signals were 0.010–0.012 versus 0.013–0.015). However, the difference between the proximal poly (A) sites and the unique poly (A) sites was insignificant ($P=0.56$–0.88, Wilcoxon tests), although the latter but not the former was the final terminus of gene transcription. As shown in Figure 1B, similar patterns were observed in other cell lines. However, the signals in the HUVEC and the NHEK cell lines were weaker, and this could be due to the different data qualities.

Intergenic regions and pseudogenes are tightly associated with lower levels of DNase hypersensitivity. The lower chromatin accessibility around the distal cleavage sites could be due to the higher level of contamination with sites from intergenic regions or pseudogenes. Our study of random positions from intergenic regions and pseudogenes confirms the lower hypersensitivity in these regions than other groups ($P<0.037$, Wilcoxon tests; Fig. 1A). We then investigated the original RNA-PET signals for our poly (A) sites. If the distal cleavage sites include more false sites from intergenic regions or pseudogenes, weaker RNA-PET signals are expected for this group. However, we found no significant signal strength difference between the distal and the proximal poly (A) sites ($P>0.246$, Wilcoxon signed-rank tests for all six cell lines). Distal poly (A) sites had even higher RNA-PET signals than unique poly (A) sites in three cell lines (HepG2, GM12878 and NHEK; $P<0.012$, Wilcoxon tests) and no significant difference was observed in the other three cell lines. All these indicate that the quality of poly (A) site discovery is comparable among different groups, and the lower levels of chromatin accessibility around distal poly (A) sites are not because of the higher contamination levels of false sites from intergenic regions or pseudogenes. Additionally, the verification experiments of the RNA-PET data in the ENCODE project showed that >99% of the PETs represented full-length transcripts (Dunham *et al.*, 2012). This indicates the low false discovery rate in our identification of poly (A) sites.

To confirm the association between the lower chromatin accessibility and the transcript cleavage, we designed another analysis. Genes with two potential poly (A) sites across all the cell lines were selected. For each site, the DNase signals from the six cell lines were divided into two groups according to whether the cell line used the site for polyadenylation. The average signals were calculated for each group of the site. Then the Wilcoxon signed-rank test was performed on the average signals of all the proximal or the distal poly (A) sites, respectively. The significantly lower accessibility was observed for cell lines using the sites for polyadenylation compared with cell lines without using the sites for polyadenylation [$P=2.96\times10^{-7}$ or 2.66×10^{-6} for proximal or distal poly (A) sites].

Because nucleosome occupancy is closely related to chromatin accessibility, we further investigated whether the lower chromatin accessibility around poly (A) sites was due to the densely occupied nucleosomes or because of other higher-order chromatin structures. Some studies showed that in *Saccharomyces cerevisiae*, the high % AT content correlates with low nucleosome occupancy especially in homopolymeric runs of poly (A) and poly (T) occurring in the 5′ or 3′ nucleosome-free region (Jansen and Verstrepen, 2011). The nucleosomes depletion around poly (A) sites in human T cells was also observed (Spies *et al.*, 2009). To check the nucleosome occupancy around poly (A) sites in our considered cell lines, we obtained the MNase-seq signals. Randomly chosen exonic or exon ending sites were processed as control groups as mentioned previously. Using the K562 cell line as an example, Figure 2 clearly shows that poly (A) sites, no matter whether they are unique or APA sites, have lower nucleosome occupancy than random exonic or exon ending sites ($P=0.0001$–0.0055, Wilcoxon tests; average

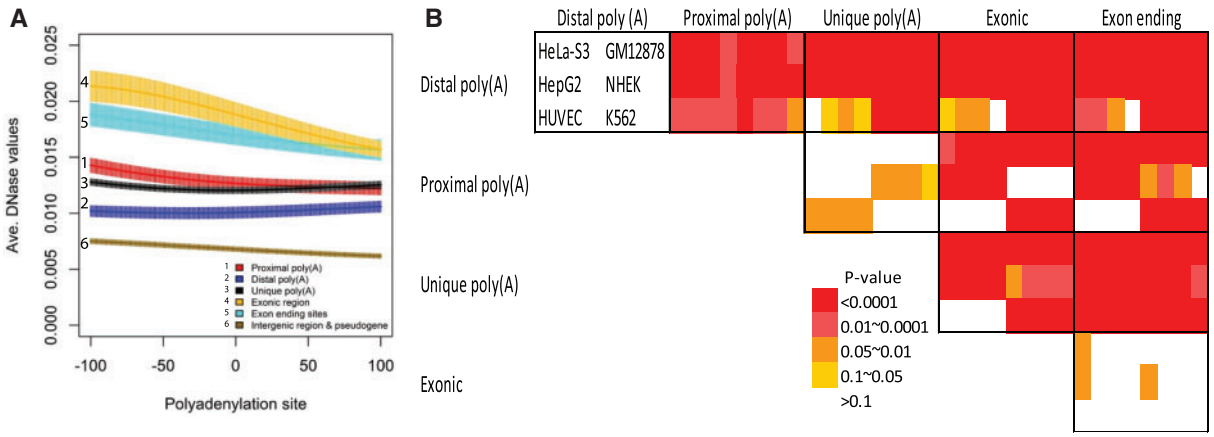

Fig. 1. Chromatin accessibility differences among poly (A) sites. (**A**) DNase-seq signals around different positions in the K562 cell line. The lines show the average DNase-seq density around proximal poly (A) sites (1), distal poly (A) sites (2), unique poly (A) sites (3), random exonic sites (4), random exon ending sites (5), random positions in intergenic regions or pseudogenes (6). Position 0 represents the considered sites, and the 100 bp upstream and downstream regions are shown. The vertical lines represent the standard errors for each position. (**B**) Wilcoxon rank test results for DNase-seq signals around different sites in the six cell lines. Each rectangular block displays the results for the six cell lines shown in the top left block. Thus, each block is divided into six smaller rectangular regions. For each cell line, four different regions, including the −160 to −80 bp upstream, −80 to 0 bp upstream, 0–80 bp downstream and 80–160 bp downstream regions of the considered sites, were tested separately, and the results are shown from the left to the right in the smaller rectangles. Thus, each smaller rectangle is further divided into four segments. The colors were coded to demonstrate the significance level of the difference between the two groups. For the comparison between distal poly (A) sites and proximal poly (A) sites, Wilcoxon signed-rank tests were performed instead

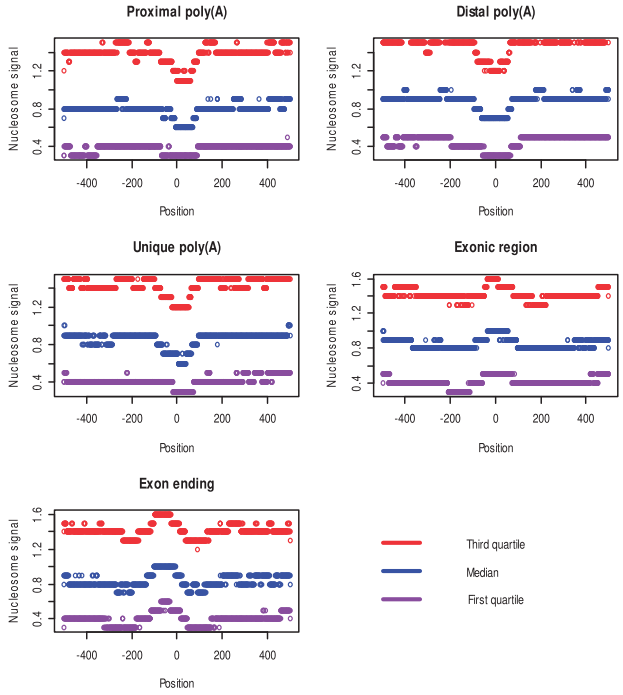

Fig. 2. Nucleosome distribution around poly (A) sites in the K562 cell line. Three lines represent the third quartile, the median and the first quartile of the MNase-seq values at each position. For all the three different poly (A) groups, the cleavage sites display fewer nucleosomes than randomly chosen exonic or exon ending positions

signals were 0.85–1.05 versus 1.05–1.20). The MNase-seq data were also available for the GM12878 cell line. Similar results were observed ($P < 0.0001$, Wilcoxon tests). The results indicate that the lower chromatin accessibility around poly (A) sites is not due to the high density of nucleosomes, and it could be due to other higher-order chromatin structures. Additionally, nucleosomes were well positioned in exons compared with introns (the 'Exon ending' panel in Fig. 2). Previous research has discovered that nucleosomes are preferentially positioned in exons, and the organization is evolutionary conserved (Schwartz *et al.*, 2009).

H3K36me3 has been reported as a determinant of pre-mRNA alternative splicing (Soojin Kim, 2011). To further study the possible relationship between APA and H3K36me3 modification, the H3K36me3 signals around poly (A) sites were collected from the ChIP-seq data of the ENCODE project, and the average read coverage within a 80 bp window was calculated. Significantly differences were observed in the H3K36me3 signals between unique poly (A) sites and APA sites after applying Wilcoxon rank tests ($P < 0.0178$ with the vast majority <0.0001, Wilcoxon tests, all six cell lines). More specifically, the poly (A) sites of genes using multiple cleavage sites in a cell line demonstrated higher H3K36me3 levels. The average signals are 3.5 for multi-poly (A) genes and 3.1 for unique poly (A) genes. As H3K36me3 modification may depend on nucleosomes, we removed the regions with weak MNase-seq signals (i.e. the average density signal equal to zero), the conclusion is the same ($P < 0.0018$, Wilcoxon tests, the GM12878 and K562 cell lines). Similar analysis was performed for gene using three APA sites to confirm our conclusion that the higher H3K36me3 level was observed in APA sites. The proximal, middle and distal APA sites all exhibited significantly higher H3K36me3 signals than unique poly (A) sites ($P < 0.0030$ with the vast majority <0.0001, Wilcoxon tests, all six cell lines). Interestingly, the multiple cleavage sites of the same gene demonstrated similar H3K36me3 signals among themselves ($P > 0.0523$, Wilcoxon

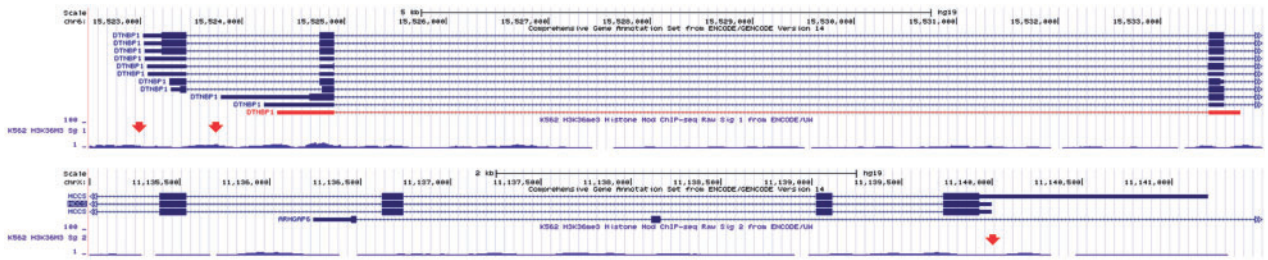

Fig. 3. An example of H3K36me3 signals in the K562 cell line. The big arrows indicate the APA sites of DTNBP1 and the uniquely deployed poly (A) site of HCCS in the considered cell line. The APA sites tend to display higher H3K36me3 modification than the unique poly (A) site

signed-rank tests, all six cell lines). An example of H3K36me3 around poly (A) sites is shown in Figure 3. Gene DTNBP1 used two poly (A) sites in the K562 cell line, but gene HCCS only used one poly (A) site in this cell line. The figure clearly shows the higher H3K36me3 levels of the APA sites than those of the unique poly (A) site.

4 DISCUSSION

We discovered the low chromatin accessibility around poly (A) sites across six human cell lines, especially for the distal poly (A) sites of genes using multiple poly (A) sites.

The phenomenon of less number of nucleosomes located around poly (A) sites is consistent with previous reports. Specifically, regions around poly (A) sites contained higher AT % than those random exonic or exon ending sites (e.g. $P < 2.2 \times 10^{-16}$, Wilcoxon tests, NHEK cell line). Therefore, we excluded the possibility that the lower chromatin accessibility around poly (A) sites was resulted from higher nucleosome occupancy. Interestingly, the proximal instead of the distal poly (A) sites exhibited similar chromatin accessibility with the unique poly (A) sites, although both the distal and the unique poly (A) sites were the final termini of gene transcription.

H3K36me3 modification was found to peak within actively transcribed genes especially in exonic regions and, therefore, has been considered as an indicator for transcriptional regions of genes (Bannister, 2004). Accordingly, the unique poly (A) sites of genes with only one cleavage site, as the transcription termini, were expected to have lower H3K36me3 compared with the proximal poly (A) sites of APA genes. Surprisingly, there was no significant difference between the distal and the proximal poly (A) sites of the same gene. These APA sites all tended to have higher H3K36me3 than the unique poly (A) sites (Fig. 3), indicating the distinct role of H3K36me3 in genes using APA.

Previous study reported that H3K36me3 was more enriched at poly (A) sites of short isoforms than those of long isoforms (Lin *et al.*, 2012). This discrepancy may be due to the distinct criteria for the selection of proximal and distal poly (A) sites. We required the usage of both sites in the considered cell line. On the contrary, they pooled the poly (A) sites from multiple breast tissues to claim a short or long isoform and did not require the usage of both poly (A) sites when they compared histone modification signals in a specific cell line. Thus, some of the poly (A) sites of their short or long isoforms can be the unique sites according to our definitions. Instead of using the windows, they focused on the distributions of the distance between the poly (A) site and the nearest H3K36me3 mark. When we applied the Kolmogorov–Smirnov test to our selected sites in our considered cell lines, there was still no significant difference for the distance distributions among the APA sites of the same genes ($P > 0.84$ for all the six cell lines).

For the H3k36me3 study, to confirm there were no additional poly (A) sites in the downstream, we searched for the longest genomic ranges of genes *ab initio* instead of relying on the known gene annotation, as we showed in Section 3. Thus, after identifying the initial and terminal sites of transcripts from the RNA-PET reads, overlapped transcripts were clustered and considered as the same genes. Similar results were observed.

We also examined the possible associations between poly (A) site signals and DNase hypersensitivity, nucleosome occupancy or H3K36me3 marks. The range of significant Spearman's rank correlation coefficients was −0.019 to 0.142 for DNase hypersensitivity, −0.063 to 0.065 for nucleosome occupancy and −0.078 to 0.166 for H3K36me3. Thus, the association patterns were inconsistent among different cell lines, and the correlation magnitude was small. As we mentioned before, the RNA-PET data in the ENCODE project are of high quality, indicating the low false discovery rate even for poly (A) sites with low RNA-PET signals. Furthermore, the level of RNA-PET signals is confounded by gene expression, and gene expression is also tightly associated with DNase hypersensitivity, nucleosome occupancy or H3K36me3 marks.

The human poly (A) sites identification had been implemented by machine-learning methods using sequence and structural patterns around poly (A) sites (Chang *et al.*, 2011). The accuracy rate for their model or polya_svm (Cheng *et al.*, 2006) was ~60–70% when applying coding sequences as the negative set. Here, we applied the Support Vector Machine method libsvm (Chang and Lin, 2011) and input DNase sensitivity, nucleosome occupancy signals and histone marks as features. The accuracy was 65–68% for poly (A) sites prediction, 56–57% for APA prediction and 54–56% for proximal or distal poly (A) sites prediction. The results suggest that our discovered DNA functional elements provide comparable information with the sequence and structural patterns when predicting poly (A) sites. In the future, we can develop a complicate model to use all these features to further improve the prediction of poly (A) sites.

In conclusion, we observed distinct patterns of DNase I sensitivity and H3K36me3 around the multiple cleavage sites simultaneously used by a gene. The detailed mechanism needs more specific biological examinations.

Funding: National Institute of General Medical Sciences (R01GM097230); National Human Genome Research Institute (P50HG002790).

Conflict of Interest: none declared.

REFERENCES

Bannister,A.J. (2004) Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. *J. Biol. Chem.*, **280**, 17732–17736.

Chang,C.C. and Lin,C.J. (2011) LIBSVM: a library for support vector machines. *Acm T Intell. Syst. Tech.*, **2**.

Chang,T.H. *et al.* (2011) Characterization and prediction of mRNA polyadenylation sites in human genes. *Med. Biol. Eng. Comput.*, **49**, 463–472.

Cheng,Y. *et al.* (2006) Prediction of mRNA polyadenylation sites by support vector machine. *Bioinformatics*, **22**, 2320–2325.

Derti,A. *et al.* (2012) A quantitative atlas of polyadenylation in five mammals. *Genome Res.*, **22**, 1173–1183.

Djebali,S. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.

Dunham,I. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Edmonds,M. (2002) A history of poly a sequences: from formation to factors to function. *Prog. Nucleic Acid Res. Mol. Biol.*, **71**, 285–389.

Fullwood,M.J. *et al.* (2009) Next-generation DNA sequencing of paired-end tags (pet) for transcriptome and genome analyses. *Genome Res.*, **19**, 521–532.

Jansen,A. and Verstrepen,K.J. (2011) Nucleosome positioning in *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.*, **75**, 301–320.

Ji,Z. *et al.* (2009) Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc. Natl Acad. Sci. USA*, **106**, 7028–7033.

Lin,Y. *et al.* (2012) An in-depth map of polyadenylation sites in cancer. *Nucleic Acids Res.*, **40**, 8460–8471.

Lutz,C.S. (2008) Alternative polyadenylation: a twist on mRNA 3′ end formation. *ACS Chem. Biol.*, **3**, 609–617.

Mayr,C. and Bartel,D.P. (2009) Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, **138**, 673–684.

Schwartz,S. *et al.* (2009) Chromatin organization marks exon-intron structure. *Nat. Struct. Mol. Biol.*, **16**, 990–995.

Shepard,P.J. *et al.* (2011) Complex and dynamic landscape of RNA polyadenylation revealed by pas-seq. *RNA*, **17**, 761–772.

Soojin Kim,H.K. *et al.* (2011) Pre-mRNA splicing is a determinant of histone H3K36 methylation. *Proc. Natl Acad. Sci. USA*, **108**, 13564–13569.

Spies,N. *et al.* (2009) Biased chromatin signatures around polyadenylation sites and exons. *Mol. Cell*, **36**, 245–254.