GENOME RESEARCH

A Genomics Analysis of RNA polymerase II modification and chromatin architecture related to 3' end RNA polyadenylation

Zheng Lian, Alexander Karpikov, Jin Lian, Milind C Mahajan, Stephen Hartman, Mark Gerstein, Michael Snyder and Sherman M Weissman

Genome Res. published online May 16, 2008;
Access the most recent version at doi:10.1101/gr.075804.107

A Genomics Analysis of RNA polymerase II modification and chromatin architecture related to 3' end RNA polyadenylation

Lian, Z.,[1†] Karpikov, A.,[2†] Lian, J.,[1] Mahajan, M.,[1] Hartman, S.,[2] Gerstein, M.,[2] Snyder, M.[3],Weissman, S.M.[1#]

Running head: Chromatin and Transcription Termination
Keywords: polyadenylation , RNA polymerase II modification, chromatin architecture

Corresponding Author [#]
1. Department of Genetics, Yale University School of Medicine, New Haven, CT. 2. Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT. 3. Department of Molecular Cellular and Developmental Biology, Yale University, New Haven, CT
† These authors contributed equally to this work

Abstract: Genomic analyses have been applied extensively to analyze the process of transcription initiation in mammalian cells, but much less to the events associated with transcript 3' end formation and transcription termination**.** We have used a novel approach to prepare 3' end fragment libraries from polyadenylated RNA of several cell types, and globally mapped the position of the poly(A) addition site using oligonucleotide arrays tiling one percent of the human genome. This approach revealed more 3' ends than had been previously annotated. The distribution of these ends relative to DNA sites bound by RNA polymerase II and the distributions with those for di- and trimethylated lysine 4 and lysine 36 of histone 3, was compared by ChIP-chip analysis. We found that a substantial fraction of unannotated 3' ends of RNA are intronic and reside antisense to the embedding gene. Poly(A) ends of annotated messages lie at a variable distance averaging approximately two kb upstream of the end of RNA polymerase binding (termination). Near the sites of RNA polymerase termination, as well as in some internal sites, there is an accumulation of both unphosphorylated and carboxy-terminal domain (CTD) serine 2 phosphorylated large subunit of polymerase II, suggesting pausing of the polymerase and perhaps dephosphorylation prior to release. Lysine 36 trimethylation occurs across the body of many transcribed genes, sometimes alternating with stretches of DNA in which lysine36 dimethylation is relatively more prominent. Lysine 36 methylation often decreases beginning at or near the site of polyadenylation, sometimes disappearing before disappearance of phosphorylated RNA polymerase II and release of RNA polymerase from DNA. Our results suggest that transcription termination may involve the separable events of loss of histone3 lysine 36 methylation and later release of RNA polymerase. The latter is often associated with polymerase pausing before release. Thus, overall our study reveals extensive sites of poly(A) addition across the human genome and provides insights into the events that may occur during 3' end formation.

INTRODUCTION

Identification of the regions of the human genome that encode transcripts is essential for a complete functional understanding of the function of the genome. Studies over the last few years have found that many more regions are transcribed into RNA than can be accounted for by genes encoding known or predicted proteins (reviewed in (Rozowsky et al. 2006; Kapranov et al. 2007a)), and non-coding RNAs that serve a number of functions have been identified (reviewed in (Mattick and Makunin 2006; Shamovsky and Nudler 2006; Carninci and Hayashizaki 2007; Kapranov et al. 2007b; Taft et al. 2007)). Examples include the XIST RNA that is involved in X chromosome silencing, RNAs transcribed from portions of imprinted regions and functionally related to imprinting, precursors for small regulatory RNAs, RNA that can directly regulate transcription factors such as the steroid receptor, inter-genic transcripts that appear to regulate the expression of adjacent coding genes such as the Hox genes, and cytoplasmic antisense RNAs from introns that may modulate the levels of expression of protein coding genes. However the function of most non-coding RNAs is not known and a substantial portion of these RNAs are intra-nuclear (Furuno et al. 2006; Gingeras 2007).

Our current understanding of the extent of transcriptionally active DNA has come primarily from massive application of established technology for cDNA and EST sequencing (Maeda et al. 2006) and more recently from newer technologies. These latter technologies include approaches for the display and sequence analysis of short sequences adjacent to sites of oligo(dT) primed cDNA synthesis (Wei et al. 2004) and/or to cap sites at the 5’ end of mRNAs (Maruyama and Sugano 1994; Choi and Hagedorn 2003; Kodzius et al. 2006; Ng et al. 2006; Denoeud et al. 2007), as well as developments in the field of microarray analysis (Rinn et al. 2003; Kapronov et al., 2002; Bertone et al., 2004).

Studies employing genomic tiling arrays have been quite informative regarding the occurrence and distribution of transcriptionally active regions in large portions of the human genome. Early arrays consisted of PCR products derived from non-repetitive portions of the genome. An early application of this approach was the study of the transcriptional activity of chromosome 22. This study showed the presence of substantial amounts of intergenic transcription as well as accumulation of transcripts from within introns, often in an anti-sense direction (Rinn et al. 2003). However, with advances in technology, the PCR product arrays have been replaced by microarrays containing very large numbers of oligonucleotides covering non-repetitive regions of large portions of the genome such as entire chromosomes (Kapranov et al. 2002; Cheng et al. 2005; Kapranov et al. 2005) or the regions studied intensively by the ENCODE consortium(2004). Whole genome oligonucleotide tiling arrays have also been applied to transcript identification (Bertone et al. 2004; Cheng et al., 2005), the advent of high density oligonucleotide microarrays is expected to make the cost of whole genome scanning generally affordable in the future.

One of the most extensively applied approaches for identifying the 3’ ends of transcripts involves generating short sequence tags from the ends of RNA by the addition of oligonucleotides that allow restriction site cleavage 21 bases from the 3’ end (Saha et al. 2002). This leads to short sequence tags that can be concatemerized and sequenced. Extensive sequencing is required in order to obtain enough tag sequences to identify and quantify less abundant RNA species and the wide application of these approaches requires advances in economy and scale of sequencing that are only now becoming feasible. In addition, the short sequence tags may be challenging to align to unique regions of the genome, particularly if they are derived from repeat containing regions, and are rather short to be used for analysis with genome tiling microarrays.

The relationship between polyadenylation signals and transcription termination in higher cells is complex (reviewed in (Buratowski 2005)). Studies of nascent transcripts in a few selected highly transcribed loci have shown that transcription may proceed well beyond the site of poly(A) addition, supporting a model in which elongated transcripts are cleaved at poly(A) addition sites (reviewed in (West et al. 2006)), and the remaining 3' ends of the transcripts degraded by mechanisms that are still under study. The presence of an accessible polyadenylation signal (AAUAAA in this case) was found to be necessary for efficient transcription termination (Nag et al. 2006) and transcription termination without precursor RNA cleavage has been observed in vivo in Drosophila (Osheim et al. 2002) as well as in vitro. However, our overall understanding of the relationship between poly(A) addition and transcription termination is limited, particularly for genes expressed at a low level.

Chromatin immunoprecipitation experiments can directly or indirectly indicate the presence of elongating RNA polymerase and potentially be used to map transcription units relative to the positions of poly(A) addition. The largest subunit of RNA polymerase II (polII) contains a carboxyl terminal domain consisting of many repeats (52 in humans) of the heptapeptide YSPTSPS that is variably phosphorylated (Buratowski 2003; Phatnani and Greenleaf 2006). PolII initially associates with the promoter in a form devoid of CTD phosphorylation. Early elongating forms of polII are phosphorylated at serine 5 of the CTD principally by kinase activity of the general transcription factor TFIIH. The later stages of transcript elongation are associated with CTD serine 2 phosphorylation, mediated at least partly by CDK9. Therefore the presence of polII phosphorylated on CTD S2 is a mark of progression of RNA polymerase II and presumably transcription, along the DNA template.

In addition to direct detection of RNA polymerase II on DNA, specific histone modifications are associated with the presence of elongating RNA polymerase II. Histone 3 lysine 4 is di- and trimethylated around the site of transcription initiation, mediated by the Set1 histone methylase (Barski et al. 2007; Heintzman et al. 2007). Subsequent phosphorylation of CTD on serine 2 recruits another histone methylase, Set2 (Kizer et al. 2005; Morillon et al. 2005). Set2 adds methyl groups to form trimethylated lysine 36 on histone 3. This mark has been suggested to be part of a process that ensures deacetylation of chromatin in the wake of polII and thereby prevent internal reinitiation of transcription (see discussion in (Sims et al. 2004; Li et al. 2007)).

In the present report we have used a method for preparing collections of fragments from the 3' ends of cDNA fragments that are of sufficient length for array hybridization. These fragments were analyzed with ENCODE genomic tiling arrays to map the sites of polyadenylation of stable RNAs in five cell types. We have compared these results with chromatin immunoprecipitation experiments using genomic tiling arrays and antibodies against various forms of the largest subunit of RNA polymerase II and against di- and trimethylated lysines 4 and 36 of histone 3. The results confirm that a large fraction of poly(A) sites of mRNA do not correspond to the 3' end of known mRNAs, and do not represent sites at which RNA polymerase II is removed from the DNA template. RNA polymerase CTD serine 5 phosphorylation is largely concentrated at promoter proximal regions that also show histone 3 lysine 4 di- and trimethylation. Serine 2 phosphorylation occurs across the body of most genes and initiates more or less concordantly with lysine 36 di- and trimethylation within genes. However serine 2 phosphorylated and unphosphorylated RNA polymerases tend to accumulate at sites averaging about 2 kilobases beyond the site of transcript cleavage and polyadenylation, and often well beyond the area of detectable histone 3 lysine 36 tri methylation. The results suggest a relatively complex mechanism of transcription termination that involves reduction of histone 3 lysine 36 trimethylation prior to RNA polymerase dephosphorylation and release from DNA.

RESULTS

We had previously established a procedure for the selective amplification of the 3' ends of cDNAs and used it to study patterns of mRNA expression in human cells (Prashar and Weissman 1996; Subrahmanyam et al. 2001). In this procedure cDNA synthesis was primed with an oligo(dT) primer that had a PCR primer binding sequence attached to its 5' end. The cDNA was then cut with a restriction enzyme and ligated to a Y-shaped adapter. Amplification was accomplished by PCR using one primer complementary to the oligo(dT) primer site and an second primer that had the same sequence as one arm of the Y-shaped adapter. In these studies we sequenced over one thousand individual bands excised from gels and found that nearly all were derived from the 3' ends of mRNA species. The only exceptions were those in which the oligo(dT) had primed from an internal adenine (A) rich sequences in RNA.

We have now modified this procedure to further improve its specificity by using a biotinylated primer to selectively capture the desired 3' end fragments prior to further amplification and to prepare libraries of 3' end fragments from polyadenylated RNA (Fig. 1). The resulting library of 3' end fragments was used with genomic tiling arrays to map sites of poly(A) addition after normalization against total cDNA signals. The 3' end cDNA fragments were prepared by cutting total cDNA with either Sau3AI or NlaIII restriction enzymes. An advantage of this approach is that the signal obtained from such fragments generally begins at a position corresponding to the restriction site and ends at the site of poly(A) addition, so that the orientation of the first strand cDNA can often be deduced from the signal on genomic tiling arrays (Fig. 2).

To establish our procedures, we first prepared 3'end libraries from HeLa cells using Sau3AI and analyzed these samples using Affymetrix U133 oligonucleotide chips that were designed for measuring mRNA levels. To compare the sensitivity of the 3' end tiling chip display with that of commercial oligonucleotide chips designed for measuring cDNA levels, we prepared 3'end libraries from HeLa cells and analyzed these on two types of arrays: a) genomic tiling arrays covering the ENCODE regions of the genome, and b) Affymetrix U133 chips. The Affymetrix arrays revealed that 107 mRNAs were present in this region from HeLa cells. Of these 96 had a definite poly(A) signal from the ENCODE array data, as observed manually on an IGB browser (http://www.affymetrix.com/support/developer/tools/download_igb.affx) display of Sau3AI generated 3' end fragments. An additional 7 showed weak or diffuse signals and only 4 gave no signals. Statistically this is somewhat more signals than would be expected for a random distribution of Sau3AI sites, since a fragment less than about 35 base pairs in length would not give a significant signal on the arrays.

A second group of two samples was prepared in which the cDNA was cut with NlaIII rather than Sau3AI. Two of the genes that did not show a 3' end RNA signal with Sau3AI cut fragments now showed a 3' end. The absence of 3' end signals for two mRNAs that were reported positive by the Affymetrix chips could stem from a variety of causes including limits of sensitivity, diffuse or far downstream 3' ends or false positive calls by the Affymetrix chip. Conversely of 23 genes represented on the U133 chip and reported as absent, 15 showed a poly(A) signal and two others had weak signals, suggesting that, overall, the display is more sensitive than the total cDNA analysis on 25 mer oligonucleotides. Of 42 genes not represented on the U133 chip 30 showed a definite poly(A) tail and 2 showed a weak poly(A) signal. As the threshold for calling a positive signal was lowered a greater fraction of the calls were from regions not adjacent to the 3' ends of known genes (Fig. 3A). However the number of signals associated with known genes also increased throughout the range as thresholds were lowered

(Fig. 3B). More signals were observed associated with known genes than the number of such genes themselves, due to the use of multiple minor 3' ends by many genes.

We performed similar analysis of the ENCODE region for triplicate RNA samples prepared from the NB4 promyelocytic cell line, from NB4 cells induced to neutrophil differentiation by treatment with retinoic acid, from normal human neutrophils, and from GM06990 lymphoblastoid cells, as well as a single analysis from K562 erythroleukemic cells. To deal with this large amount of data, we bioinformatically detected 3' end fragments at a range of thresholds of sensitivity.

For each cell type between 40 and 60% of the signals intersected regions showing a signal from the other cell types in this study, when using a threshold of 0.04 for calling 3' ends. This threshold was chosen because it gave a false discovery rate of less than 5% for every cell type (Fig. 3C).Somewhat unexpectedly, the remaining 3' ends often appeared to be cell type specific (Fig. 3B and Table 1.) These did not reflect variability in the method as comparison of two randomly chosen HeLa experiments showed over 98% agreement in the location of 3' ends detected with the same threshold. As shown in Fig. 3B the cell type specific 3' end signals spanned the full range of intensities with signals shared between cell types only modestly more abundant at high than at low intensities, so that the cell type specificity was not due only to low intensity signals. Further, attempts to compare two cell types using different thresholds for one cell than the other did not result in more than 50% overlap, over the threshold range of 0.04-0.05 (Fig.4).

Approximately one hundred of the HeLa or NB4 regions that gave 3' end signals but were not within 2.5 kilobases downstream of the annotated 3' ends of known or predicted genes were chosen randomly (Table 2). These signals were manually compared against the genomic sequence embedding them. As shown in Table 2, 58 of the signals were preceded by an AATAAA or ATTAAA polyadenylation sequence, 12 signals were probably attributable to A (or T) rich sequences in the genomic DNA that are presumably in transcripts extending through these sequences and represent either copied from incompletely spliced mRNA, intronic excised introns that had not been degraded, or longer transcripts passing through this region. In 3 cases an A rich sequence was preceded by a potential poly(A) addition signal lying about 20 bases upstream. This suggests that these sequences represent fragments of re-inserted cDNA copies of other genes. In some cases there was no obvious reason for the signal in terms of poly(A) addition sites or runs. These could represent polyadenylation at atypical or non-canonical poly(A) addition signals (Venkataraman et al. 2005), regions spliced into distant fragments that were linked to poly(A) , or transcripts terminating in repetitive regions not represented on the microarray.

Overall almost 60% of the cases examined were good candidates to represent the 3' polyadenylated ends of unannotated RNAs. Twelve 3' ends lay in regions upstream and more than 2.5 kb downstream of any annotated gene. These might represent 3' ends of entirely unannotated transcripts or very long 3' extensions of known transcripts (Moucadel et al. 2007). Of the 3' ends within introns that could be evaluated, 26 of the poly(A) signals that could be oriented arose from transcripts in the same orientation as the embedding gene while 20 were in the antisense orientation. This compares well with an earlier estimate that 10 of 25 intronic transcripts detected from a PCR fragment arrays representing chromosome 22 were in the antisense orientation with respect to the embedding gene (Rinn et al. 2003).

As a further check of the method, we selected nineteen 3' end HeLa signals that did not lie within 2.5 kb of any 3' end annotated in RefSeq. For each such signal we designed a pair of

PCR primers with one upstream and the other downstream of the estimated position of the poly(A) site, and performed 3' RACE on HeLa cell cDNA. In 18 of the 19 cases a band was seen with one or the other of these primers that was substantially clearer and more prominent than any band seen with the other primer of the pair. These bands were sequenced and in 17 cases gave satisfactory reads. Of these sixteen bands showed a poly(A) signal (either AAUAAA or, in one case, ATTAAA) in the DNA approximately 20 bases upstream of the poly(A). In the remaining case, there was a poly(A) tract encoded in the genomic DNA, and oligo(dT) priming presumably occurred from a transcript embedding this genomic sequence. These results indicated a somewhat higher percentage of 3' end signals were associated with poly(A) signal hexanucleotides in the DNA than might have been anticipated from the manual estimates.

Methylation of histone 3 at lysine 36 is mediated by the enzyme Set2/NDS1. This is a histone methylase that directly binds to the polII S2 phosphorylated C-terminal domain of the largest subunit of RNA polymerase II (Li et al. 2007). This has been suggested to be part of a mechanism that recruits the histone deacetylase RPD3 to remove histone acetylation behind the progressing RNA polymerase, and therefore prevents random transcription re-initiations within the body of the gene (Cuthbert et al. 2004; Carrozza et al. 2005; Joshi and Struhl 2005; Keogh et al. 2005). To the extent that this mechanism accounts for histone 3 lysine 36 methylation, the methylation pattern might be expected to reflect the presence of elongating RNA polymerase.

In a few cases of actively transcribed mammalian genes, it has been experimentally confirmed that transcription extends well downstream of the polyadenylation site. To study the relationship between polyadenylation sites, chromatin modifications, and termination of transcription on a global basis we performed ChIP-chip experiments across the ENCODE regions with material from HeLa, NB4, and K562 cells. These studies were performed, using various antibodies directed against specific histone modifications or different forms of RNA polymerase II as described in the Methods section.

Strong signals of lysine 36 methylation were detected across many actively transcribed genes, beginning downstream of the transcription initiation site and extending generally up to or beyond the polyadenylation site. Reduction in the signal often occurred near the site of polyadenylation. In a number of cases it was not possible to determine where the signal ended for the reasons mentioned above. In the remaining genes the signal for histone 3 lysine 36 trimethylaiton commonly declined and then ended at distances, ranging from -2.5 to +6 (average +1.8) kilobases downstream of the poly(A) addition site. The decline in signal beyond the poly(A) site was not always continuous and often appeared to occur in several steps.

Fig. 5 presents a fairly typical pattern for an actively transcribed gene. As lysine 4 di- and trimethylation signals were strong around sites of initiation and often biphasic with a gap at the actual site of initiation(Barski et al. 2007). This gap is presumably due to a combination of binding by RNA polymerase II and other factors, resulting in nucleosome displacement. Dimethylation of histone 3 lysine 4 extended somewhat further downstream than tri-methylation in many cases (Fig. 5).

There were a number of genes that showed substantial variation from the common pattern of lysine36 methylation. A few genes or genetic areas showed very low levels of lysine 36 methylation but instead histone 3 lysine 4 dimethylation (Fig. 6, 7) spread beyond the promoter into or across the body of a gene. Some, often short genes or genes transcribed at relatively low levels, had minimal or no detectable lysine 36 tri-methylation (Fig. 7).

To test whether the reduction of trimethylated lysine 36 beyond the polyadenylation site was due to partial removal of methyl groups, we also performed chromatin immunoprecipitation with an antibody directed to dimethyl lysine 36 of histone 3. We observed a signal of dimethyl lysine 36 at the position where the trimethylation waned. Unexpectedly there was often a strong signal of dimethylation of lysine 36 upstream and around the initiation site for transcription (Figs.7, 8). Some genes, especially large ones, showed broad stretches of dimethylation of lysine 36, extending tens of kilobases, preceding or following another broad zone in which tri-methylation of lysine 36 was more predominant (for example, Fig. 8), as Also, there were blocks of dimethylated lysine 36 that did not overlap with known transcripts.

Measurement of DNA associated RNA polymerase would in principle, provide a direct assessment of the relationship between sites of poly(A) addition and sites of transcription termination or at least the release of DNA from the RNA polymerase. However, RNA polymerase II undergoes a series of phosphorylations during mRNA transcription. The enzyme initially binds to promoters in an unphosphorylated state, then becomes phosphorylated first on serine 5 of the YSPTSPS repeat motif, then subsequently on serine 2. To more directly study the relationship between the ends of RNA polymerase II binding regions and mRNA polyadenylation sites, we employed antibodies reported to interact selectively with unphosphorylated CTD of the large subunit, with CTD phosphorylated on S2, CTD phosphorylated on S5 or total RNA polymerase regardless of its phosphorylation state (Table 3). We explored the use of four different antibodies reported to detect different phosphorylation states of RNA polymerase II (Table 3). Monoclonal antibody 8wg16 is directed against the unphosphorylated form of polII CTD. Antibody pol II pho S5 reacts with CTD phosphorylated at serine 5. An antibody to polII pho2 that has been used in earlier studies has been reported to react with serine 2 phosphorylated CTD but more recent studies have shown that it can also react with serine 5 phosphorylated CTD and especially with the doubly phosphorylated CTD (Phatnani and Greenleaf 2006). We used a newer antibody (reported to be specific for serine 2 phosphorylation). Antibody 4H8 is stated to react with all forms of CTD but the chromatin immunoprecipitation studies reported here suggest that it may inefficiently recognize S2 phosphorylated polymerase during transcript elongation.

Unphosphorylated RNA polymerase II typically showed a peak of DNA binding that overlaps the site of transcription initiation found at the center of the prominent biphasic peaks of histone 3 lysine 4 di- and trimethylation that occur flanking the transcription initiation site. In addition in a few actively transcribed genes unphosphorylated polymerase II signals occurred throughout the gene body and in many cases there was a peak of signal coinciding with the 3' termination site of RNA polymerase binding, as estimated by comparison of patterns with each of the antibodies. Furthermore, there was a curious unevenness or lumpiness of signals across many genes that did not necessarily correlate closely with the positions of known exons. Similar signals were seen in each of three biologic samples and in analyses performed on different cell types and by either of two investigators in our group (J.L. and S.H.) using either the same antibody or a second antibody directed against unphosphorylated RNA polymerase CTD. The signals for S5 phosphorylated RNA polymerase II were usually confined to a region of 1-2 kilobases overlapping and extending slightly downward from the transcription initiation site. Of note, in most genes there was no significant signal from this antibody across the bulk of the transcribed region. The signals for S2 RNA polymerase were distributed in a complementary fashion, being weak or absent at the initiation site and stronger within a couple of kilobases downstream from the initiation site. When present significant signals extended (Fig. 5, for example), for an average distance of about 2 kilobases downstream of the poly(A) addition site but with considerable variability between genes. Conversely there were several genes in which the histone modification could be detected in the absence of detectable signals for phosphorylated RNA polymerase. As with the signals for unphosphorylated RNA polymerase, the intensity of the signals for S2 phosphorylated RNA polymerase was irregular across the transcribed regions of genes, the pattern of intensity did not always match that of histone 3 lysine 36 methylation, and there was frequently a prominently increased signal that began well after the poly(A) addition site and immediately preceded the region where all polII signals disappeared. The latter is strongly suggestive of RNA polymerase pausing prior to release (Birse et al. 1997). The alternative possibility that this signal represents a second initiation site for RNA polymerase II is unlikely because there was no mark of initiation of transcription such as trimethyl or dimethyl histone 3 lysine 4, S5 phosphorylated RNA polymerase II, or excess of unphosphorylated RNA polymerase II.

Except at the site of transcription initiation there was commonly a correlation between signals for S2 phosphorylated and unphosphorylated RNA polymerase. One caveat in interpreting the data is that, because the CTD contains 52 copies of the heptapeptide motif in which phosphorylation occurs, we cannot distinguish between single molecules of RNA polymerase large subunit with partially phosphorylated CTDs and the presence of two populations of molecules, one completely phosphorylated and the other unphosphorylated. This makes it particularly difficult to judge whether or to what extent dephosphorylation necessarily occurs prior to RNA polymerase release.

To study the relationship between histone3 lysine 36 methylation and the presence of polymerase II at the 3' end of transcription units, we first had to exclude genes where: 1. There was no signal for RNA polymerase II and/or for lysine 36 methylation. 2. There were large blocks of repetitive sequences downstream of the poly(A) addition site. 3. There was another gene whose 3' end was within 4 kb or whose 5' end was within 3kb of the gene of interest. 4. The polyadenylation site was within 2 kilobases of the end of the ENCODE region.

These criteria were used to eliminate genes that could not be evaluated. Forty-five genes were not excluded and these were manually inspected. For 35 of these genes it was possible to compare the furthest extent of lysine 36 methylation to the presence of RNA polymerase II. For 23 of these genes the polymerase clearly extended beyond (3' to) the region marked by K36 trimethylation. In ten of the remaining genes either the result was ambiguous and in twelve genes lysine 36 trimethylation and RNA polymerase II binding disappeared at the same position. No case showed lysine 36 methylation extending downstream beyond regions associated with polymerase II.

DISCUSSION

Characterization of the structure and function of RNA transcripts in mammalian cells has become a large and active area of research. Approaches that have been used include hybridization of RNA to genomic tiling arrays, and extensive sequencing of ESTs, or of short 3' end tags. The extent of the transcriptionally active genome seems to regularly expand, the more sensitive the detection method or the more exhaustive the analysis.

The approach described in the present report has considerable sensitivity, as it detects evidence for more transcripts from known genes than are reported as present by the relatively sparse 35 mer oligonucleotide arrays used commonly for cDNA detection and quantitation. With a fairly stringent threshold we detected a large number of transcripts with 3' ends in introns or extragenic regions. Most of these had hallmarks of sites of initiation of reverse transcription by oligo(dT), including appropriately located upstream poly(A) addition signals and/or A-rich tracks in DNA at one edge of the positive signal. The putative ends lacking these signatures could have derived from cross-hybridization, fragments linked to poly(A) fragments through splicing, poly(A) sites located within repetitive sequences not represented on the arrays, polyadenylation at variant or alternative poly(A) signals (Beaudoing et al. 2000; Venkataraman et al. 2005; Cheng et al. 2006), or mis-priming during PCR amplification. Polyadenylation has also been suggested to occur on RNA fragments derived from transcription extending beyond the polyadenylation site on known genes, but most of these would fall within the 2.5 kb region downstream of known polyadenylation sites that were scored as being associated with known genes in our analyses. It is curious that a substantial fraction of the 3' ends that did not match known genes were apparently cell type specific.

Poly(A) sites corresponding to the 3' ends of transcripts of known genes are generally presumed to correspond to one or more splice variants of the coding regions of the gene (Le Texier et al. 2006). However recent evidence shows that a much more complex situation may exist. Hybridization of 5' RACE products to genomic tiling arrays have indicated the occurrence of many transcripts whose 5' end lies far upstream of known promoters for a gene (Kapranov et al. 2005). In one extensively studied and informative situation spliced transcripts of the beta globin locus have been identified that originate in sequences far upstream of the coding portions of the gene (Xiang et al. 2006). Analysis of some of these transcripts has shown that they can arise from a repetitive element lying more than 200 kb upstream of the embryonic epsilon globin gene, and be spliced into multiple forms. These apparently non-coding RNAs are spliced into the second exon of the globin gene and therefore give rise to transcripts that will have identical 3' end sequences to those of the globin coding mRNA.

Polyadenylated non-coding RNA transcripts have become increasingly interesting as their prevalence becomes appreciated (Goodrich and Kugel 2006; Mattick and Makunin 2006; Pang et al. 2006; Kapranov et al. 2007b; Kapranov et al. 2007a; Pauler et al. 2007; Prasanth and Spector 2007; Roma et al. 2007). These transcripts include precursors of small regulatory RNAs but also are related to a number of other processes. Some may act in cis to regulate X chromosome silencing or autosomal imprinting whereas others act in trans, either as activators or inhibitors of certain transcription factors. Transcription extending from upstream and across known genes may inhibit expression of the gene without regard to the structure of the non-coding RNA (De Gobbi et al. 2006).

Antisense transcripts arising within known genes could represent intronic genes but antisense intronic transcription is a more general phenomenon (De Gobbi et al. 2006; Hayashizaki and Carninci 2006; Kapranov et al. 2007a). For example Rinn et al used synthetic oligonucleotide probes directed against previously undescribed intronic transcripts that had been detected with a PCR fragment array tiling chromosome 22. Ten of the 25 oligonucleotide pairs detected antisense intronic transcription (Rinn et al. 2003). In the present experiments, of those intronic transcripts whose termini lay downstream of a polyadenylation signal, about 40% were oriented in an antisense direction compared to the embedding gene. These antisense RNAs may be important in gene regulation, such as in the parental origin-specific silencing of imprinted genes (Pauler et al. 2007). In another particularly informative case, Tenen and colleagues found two antisense transcripts originating in introns of the gene for the transcription factor PU.1, a transcription factor important for determining lineage differentiation in hematopoietic development. These transcripts were found in the cytoplasm and down regulation of the transcripts with siRNA led to upregulation of PU.1 mRNA (Ebraliddze and Tenen 2006). Studies such as these suggest that antisense intronic transcripts may represent a rather common mechanism for regulating the level of expression of specific genes.

Intronic transcripts with 3' sequences in the sense orientation with respect to known genes are of uncertain significance. They could hypothetically arise from incompletely spliced mRNA or retained introns that include A rich sequences. Alternatively these transcripts could represent novel RNAs that extend across all or portions of known genes. Transcription from upstream sites has been observed to either be required for (Ho et al. 2006; Zhao et al. 2006) or to inhibit transcription of nearby genes (Kelley and Kuroda 2000; Yang and Kuroda 2007) so that such RNAs might also have modulatory roles in the control of gene expression.

The current model for changes in RNA polymerase II CTD phosphorylation during transcriptions is that RNA polymerase II with an unphosphorylated CTD is bound to the initiation complex. The RNA polymerase then is phosphorylated on CTD serine 5 by kinases associated with the general transcription factor TFIIH. As RNA polymerase traverses the gene it becomes phosphorylated at serine 2 by CDK9 or related kinases. This phosphorylated RNA polymerase then binds Set2, an enzyme that methylates lysine 36 of histone 3. RNA polymerase continues transcription for some distance beyond the poly(A) addition signal before it is released from the DNA. The present data provides more specificity to several aspects of this model as well as showing phenomenological heterogeneity between genes.

The pattern of histone 3 lysine 36 trimethylation in the majority of genes is relatively simple. This modification begins some distance downstream of the transcription initiation site and continues across the gene template. The modification decreases in what is sometimes more of a stepwise than a continuous fashion near or beyond the end of the gene, often near the site of cleavage and polyadenylation, and some distance upstream of the furthest regions where RNA polymerase S2 phosphorylation can be detected. This is consistent with models for multiple transcription termination signals in the DNA that are not closely linked to the site of polyadenylation (West et al. 2006), in addition to effects of the polyadenylation signals on RNA polymerase pausing or termination. They also indicate that either demethylation becomes more active or Set2 less active prior to removal of the RNA polymerase from the DNA template.

In some genes, principally some of the less actively transcribed genes, histone 3 lysine 36 methylation is seen without RNA polymerase S2 phosphorylation. This presumably represents the longer persistence of histone methylation in comparison to the dwell time of phosphorylated RNA polymerase. On the other hand, there were very few examples where substantial S2 phosphorylation of RNA polymerase was detected in the absence of histone 3 lysine 36 methylation in the body of the gene, consistent with the direct relationship between the phosphorylation and the association of Set2 with DNA. In one class of exceptional genes histone 3 lysine 4 dimethylation was found across much or all of the body of the gene. This is reminiscent of studies in yeast in which inactivation of a kinase thought to be responsible for S2 phosphorylation of RNA polymerase II resulted in spreading of histone 3 lysine 4 di/trimethylation into the body of genes.

There is a certain analogy between the distribution of di- versus trimethylation of histone 3 lysine 4 and lysine 36, although they are largely non-overlapping in their location. In both cases the trimethylated form of the modified histone is more narrowly distributed. In particular histone 3 trimethylated lysine 4 is generally limited to the region proximate to the transcription initiation site methylase(s) (Santos-Rosa et al. 2002; Schneider et al. 2004; Bernstein et al. 2005; Barski et al. 2007; Koch et al. 2007). Of possible relevance, specific demethylases for histone 3 tri-methylated lysine 4 have been identified recently (Liang et al. 2007; Secombe et al. 2007).

Lysine 36 trimethylation is generally limited to the body of the gene, presumably reflecting recruitment of Set2 by CTD serine 2 phosphorylated RNA polymerase. On the other hand, both dimethyl lysine 4 and dimethyl lysine 36 also occur in large blocks of chromatin not necessarily connected to sites where trimethylation occurs and the functional significance of these blocks is currently under investigation.

Interpretation of the presence and phosphorylation state of RNA polymerase are somewhat complicated by potential limitations in the specificity of the antibodies. Nevertheless there is a clear distinction in patterns detected with the antibodies directed against S2, S5 and unphosphorylated RNA polymerase, with most of the regions detected by the S2 antibody not showing any reaction with S5. The S5 antibody generally showed a clear peak at or near the transcription start site where reactivity with S2 antibody was low or absent. S5 signals were commonly absent over most of the body of the gene. When present, they were often weak and did not correlate with the signals from S2 phosphorylated RNA polymerase or unphosphorylated RNA polymerase. This could occur because of dephosphorylation of S5 in the body of the gene, because of presence of S5 phosphorylation at only a limited subset of potential sites, or because of blocking of the doubly phosphorylated sites on the enzyme by tight association with other proteins, such as Set2. While the first is the simplest explanation for the complete lack of reactivity of progressing RNA polymerase to S5 antibody, other possibilities cannot be excluded. The antibody directed against dephosphorylated RNA polymerase is at least partially specific as demonstrated by the lack of correlation with S2 detection at the origin and incomplete correlation with S5 detection in the body of the gene. The pattern of intensities of dephosphorylated or total RNA polymerase detection across genes sometimes showed regions of strong signal sometimes alternating with regions of extremely weak signal. Similar patterns may be seen with the S2 directed antibody so that the pattern does not merely reflect a change in the ratio of phosphorylated to dephosphorylated enzyme. Presumably these observations reflect differences in the rate of progression of RNA polymerase through the gene. While regions of RNA polymerase density sometimes correlate with the presence of exons (Brodsky et al. 2005) this is often not the case in the present studies. RNA polymerase slowing or pausing has been implicated both in the generation of alternate spliced forms of mRNA and in transcription termination. The accumulation of RNA polymerase at some internal regions of genes is prominent and the basis and significance of the differences in progression rate of the RNA polymerase remains a subject for further investigation.

Another recurrent feature of the distribution of different phosphorylation states of RNA polymerase is the accumulation of dephosphorylated RNA polymerase at the 3' end of the transcription unit, often concurrent with the accumulation of S2 RNA polymerase. This suggests both a pile-up of RNA polymerase molecules preceding the point of release and also the possibility that RNA polymerase may become dephosphorylated prior to or concurrent with release from the DNA. A third observation with respect to dephosphorylated RNA polymerase is that, in some very active genes, strong signals for this form of the enzyme are detected across the entire body of the gene, relative to the signals for S2 phosphorylated enzyme. Perhaps when the gene is densely populated with RNA polymerases, the unphosphorylated enzyme can progress through the body without modification, or dephosphorylation occurs secondary to crowding of RNA polymerase molecules on DNA. Finally, the description of CTD as either phosphorylated at serine 2, serine 5 or both serine 2 and 5 is a major simplification of the potential for alternative sites or levels of phosphorylation and ignores potential conformational changes in the CTD (Phatnani and Greenleaf 2006).

In summary, sensitive detection of polyadenylation sites in cellular RNA has confirmed the existence of a number of antisense and intergenic transcripts as well as more transcripts from known genes than are detectable by some of the standard approaches. Combining this data with ChIP-chip data has provided a more detailed picture of the relationship between RNA polymerase occupation of DNA and phosphorylation in relation to the 3' ends of mature transcripts as well as revealing gene specific variations in the general pattern. Termination of transcription is a complex process involving several non-synchronous steps in addition to recognition of transcribed polyadenylation signals in RNA (Nag et al. 2006). Histone 3 lysine36 tri-methylation did not always parallel the levels of S2 phosphorylated RNA polymerase II and often diminished near the site of polyadenylation and disappeared before S2 phosphorylated RNA polymerase II CTD disappeared.

Near the 3' end of transcription units there frequently was an increase concentration of both S2 phosphorylated and unphosphorylated RNA polymerase suggesting that polymerase progression slowed or paused prior to release from the DNA, and that dephosphorylation of the RNA polymerase CTD may precede polymerase release from the DNA.

METHODS

1. Sample preparation

Four cell lines as well as normal human neutrophils were used in this study. In general, three biological replicates were used for the analyses. Cervical Adenocarcinoma (HeLa-S3) total RNA was ordered from Ambion Cell culture Ambion (AM7852, An Applied Biosystems Business**.** 2130 Woodward St. Austin, TX 78744-1832 USA). The acute promyelocytic leukemia (APL) cell line NB4 (Lanotte et al. 1991) was used directly or cultured with all-trans retinoic acid (RA) for 48 hours to differentiate cells to neutrophils (Khanna-Gupta et al. 1994). The human B-Lymphocyte cell line GM06990 was obtained from Coriell Cell Repositories (403 Haddon Avenue Camden, New Jersey 08103). Human erythroblast cell line K562 (Migliaccio et al. 2002) was obtained from American Type Culture Collection (ATCC; catalog # CCL-243^{TM}; P.O. Box 1549, Manassas, VA 20108). The cell lines were cultured by their respective standard protocols. Human primary neutrophil cells (Tsukahara et al. 2003) were prepared from healthy donors as previously described (Subrahmanyam et al. 2001). The total RNA was extracted and purified with Trizol reagents (Invitrogen Corporation 1600 Faraday Avenue PO Box 6482 Carlsbad, California 92008) and QIA quick PCR Purification Kit (27220 Turnberry Lane Suite 200 Valencia, CA 91355). An aliquot of the RNA obtained from each cell preparation was analyzed by electrophoresis on a formaldehyde agarose gel, to ensure sample quality and integrity based on the relative intensity of the 28S and 18S ribosomal RNA bands (requiring a ratio of 1.7:1).

2. cDNA synthesis

To 10 μg of total RNA in 10 μL of water in a 0.5 ml microcentrifuge tube (no stick, USA Scientific Plastics), we added 1μL (200 ng) of 2-base anchored oligo(dT) primer with a heel: 5'TAGAAGCCGAGACGTCGGTCG-T(18) NN-3' N=A, C, G, T. The contents were mixed on ice, heated to 65°C for 5 min, and chilled on ice for 5 min. This denaturation and annealing was repeated and the tubes held on ice. The first strand cDNA synthesis reaction was set up as follows: 5X first strand buffer 4 µL, 0.1M DTT 1 µL, RNase inhibitor (40 U/µL) 1 µL were mixed and warmed to 45°C. The cDNA synthesis was initiated by adding 1 µL (200 units) of SUPERSCRIPT™ II, RNase H- Reverse Transcriptase (Cat. No. 18064-022, Invitrogen) and incubation continued (at this stage the final reaction volume was 20 µL) at 45°C for 1 hour. This step was performed in a humidified incubator instead of a water bath to avoid evaporation.

After first strand synthesis the tube was chilled on ice and centrifuged briefly to collect all the contents and the second strand synthesis reaction was set up on ice in the same tube as follows: 20 µL of first strand reaction, 91 µL of water, 30 µL of 5X second strand buffer, 3 µL of 10 mM dNTPs, 4 µL E.coli DNA Polymerase I (10 U/µL) (Cat. No. 18010-017), 1 µL of E.coli DNA Ligase (10 U/µL), (Cat. No. 18052-019), and 1 µL of RNase H (3U/µL). The total volume of the reaction was 150 µL. The tubes were incubated at 16°C for 2 hours.

The reaction was stopped by adding 10 µL of 0.5M EDTA pH8.0. The mixture was extracted once with phenol: chloroform (1:1 v/v) and once with chloroform (presaturated with nuclease free water). The cDNA was precipitated by adding 0.5 volume of 7.5 M ammonium acetate and 2.5 volumes of ethanol (-20°C). At this stage the sample could be left overnight at -20°C. Prior to precipitation of the cDNA, 1 µL (20 ug) of glycogen was added as a carrier.

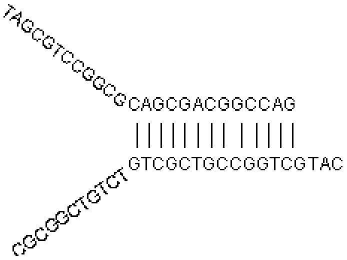
The cDNA was precipitated by centrifugation in an Eppendorf centrifuge at top speed for 15 min. Without disturbing the pellet, the ethanol was carefully removed. The pellet was washed with 70% ethanol and centrifuged again for 15 min. The supernatant was removed and the pellet dried at room temperature. The cDNA was dissolved in 20 µL of water or TE buffer.

3. cDNA restriction enzyme cutting

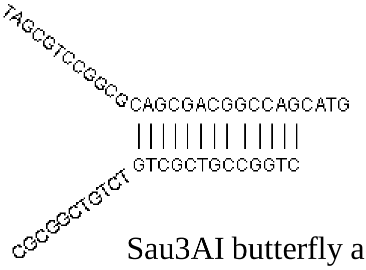
To the 20 µL of cDNA prepared described above, 3 µL of Sau3AI buffer (NEB), 0.3 µL of 10X BSA, 3.7 µL water and 3 µL of Sau3AI (10 unit/µL) were added and the sample was gently mixed and incubated at 37°C for 2 hours. The enzyme was heat inactivated at 65°C for 20 min and the tubes kept on ice.

4. Butterfly adaptor ligation

The enzyme cut fragments were ligated to synthetic Y-shaped adaptors (butterfly adapters) with a complementary overhang as described: we added 4 µL of the butterfly-adaptor with an overhang complementary to the ends of the restriction enzyme digested sample (30 µL), then added 4 µL of 10X T4 DNA ligase buffer and 2 µL of T4 DNA ligase, mixed well and incubated the sample at 16°C overnight. The enzyme was inactivated at 65°C for 15 min. Below are the sequences and structures of the butterfly adapters for NlaIII and Sau3AI fragments.



NlaIII butterfly adaptor



Sau3AI butterfly adaptor

5. PCR amplification of 3’ end fragments (Fig. 1)

The ligated cDNA was used as the template for the 3’ end fragment enrichment at this stage. The PCR mixture (50 µL) consisted of 2 µL of the template, 5 µL of 10X PCR buffer (100 mM Tris HCl, pH 8.3/500 mM KCl), 2 µL of 15 mM MgCl2, 200 µM dNTPs, 500 nM each 5' and 3' PCR primers, and 1 unit of Platinum® Taq DNA Polymerase High Fidelity system (Cat# 11304 , Invitrogen) and was heated for 2 min to 94°C in the first PCR cycle. PCR consisted of 28-30 cycles of 94°C for 30 sec, 56°C for 2 min, and 72°C for 30 sec.
The following sets of primers were used for PCR amplification of the adapter ligated 3'-end cDNAs: Biotin-TAGAAGCCGAGACGTCGGTCG was used as 3' primer;
while TAGCGTCCGG CGCAGCGAC served as the 5' primer.

6. Purification of PCR products

3’ end PCR amplified products prepared with a biotinylated primer were captured on magnetic porous glass (MPG) particles coated with streptavidin (CPG, Lincoln Park, glass (MPG). 0.1 mg (100 µL) of beads were used per reaction. Beads coated with streptavidin (CPG, Lincoln Park), were blocked by adding 1/10 volume of 40 mg/mL DNA-free tRNA and incubating on ice for 1 hour with occasional gentle vortexing. Just before nucleic acid capture, the beads were separated using a magnetic stand and the supernatant was removed by pipetting. All subsequent capture, washing, and release procedures were performed with the help of a magnetic stand. After the blocking step, the binding reaction was performed by adding the biotinylated-PCR products (50 uL) and 8 µL of 3M NaCl to the sample followed by addition of the beads. The reaction was carried out at room temperature for 30 min with occasional gentle mixing to avoid bead sedimentation. After removal of unbound cDNA, the beads were washed 3 times with washing buffer (20% glycerol, 10 mM NaCl, 0.2 mM EDTA, 10 mM Tris–HCl, pH 7.5) 3 times with nuclease-free water containing 50 mg/mL tRNA, and 3 times with nuclease-free water.

To remove the single stranded cDNA, alkaline hydrolysis was performed in the presence of 50 µL of Tris–formate buffer, pH 9.0 (obtained by combining 100 mM Tris base with 16.6 mM formic acid and 0.016 mM EDTA, final concentrations) at 65°C for 10 min. After the incubation, 100 µL of 0.5 M Tris-EDTA (pH 7.4) were added before the separation of the liquid phase (containing cDNA) from the beads. The cycle of alkaline elution was repeated three times. The alkaline-treated fractions were removed and pooled together and single-stranded cDNA was subsequently precipitated with ethanol under standard conditions.

7. PCR amplification of 3’ end fragments

The 3’ end enrichment fragments released from streptavidin beads were PCR amplified to obtain sufficient material for array analysis, by the PCR protocol described in step 5.

8. Encode arrays

PCR products from each sample were hybridized to. NimbleGen's ENCODE microarray. This array is available via the Nimblegen standard service model to life science researchers. The single array contains more than 384,000 unique 50mer probes selected from 30 megabases of human sequence data specified by the ENCODE Project Consortium (2007). These probes are spaced apart every 38 bases on the average, thus creating a 12-base overlap between probes. No probes were included for interspersed repetitive DNA, thus there are inevitable gaps in genome tiling paths on the array. Data is presented in comparison to human genome build HG17.

9. Antibodies, nuclear extracts, and immunoprecipitations

The antibodies used in the present study are listed in table 3. For each ChIP-chip assay, 1X 10^8 cells were cross-linked with formaldehyde at a final concentration of 1% for 10 min followed by addition of glycine in PBS at a final concentration of 125 mM. Cells were collected by centrifugation and washed twice in cold 1x PBS, and nuclear-enriched extracts were prepared as described (Bernstein et al. 2005). The lysate was sonicated with a Branson 250 Sonifier to shear the chromatin (Output 20%, 100% duty cycle, five 30-sec pulses), and the samples were clarified by centrifugation. Factor-DNA complexes were immunoprecipitated with their unique antibodies overnight at 4°C. Each immunoprecipitation sample was incubated with protein A-agarose (Upstate Biotechnology) for 1 h at 4°C followed by three washes with RIPA buffer and one wash with 1x PBS. The antibody-DNA complexes were eluted from the beads by addition of 1% SDS, 1x TE (10 mM Tris-Cl at pH 7.6, 1 mM EDTA at pH 8), incubation for 10 min at 65°C, addition of 0.67% SDS in 1x TE, incubation for another 10 min at 65°C, and finally gentle vortexing at room temperature for 10 min. The beads were removed by centrifugation, and the supernatants were incubated overnight at 65°C to reverse the cross-linking. To purify the DNA, proteinase K solution (400 µg/mL proteinase K, 1x TE) was added, and the samples were incubated for 2 h at 45°C, followed by a phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation to recover the DNA. Immunoprecipitated DNA was analyzed by PCR for the presence of enriched factor binding at target sequences. Reactions used 2x Taq Mastermix (Qiagen) under the following reaction conditions: 5 min at 94°C, 30 cycles of 30 sec at 94°C, 30 sec at 53°C, 30 sec at 72°C, and 10 min at 72°C. PCR products were analyzed by gel electrophoresis.
DNA samples to be hybridized to microarrays were labeled by random priming with nonamer oligonucleotides attached to Cy3 or Cy5 dyes. Control samples for 3’ end cDNA were prepared by labeling total cDNA with Cy3 and controls for the chromatin immunoprecipitation experiments were total genomic DNA prepared from chromatin cross-inked and precipitated by the same procedure as this test sample but with non-specific IgG rather than factor specific antibodies. These controls were also labeled with Cy3. Test samples were labeled with Cy5 and applied to the same chip as the Cy3 labeled control sample.

10. Bioinformatic Analysis

Nimblegen Encode tiling arrays with probe length equal to 50 bp and with probe separation of 38 bp were used in the current study for analysis of 3’ ends as well as for chromatin immunoprecipitation experiments. Signals from Cy5 and Cy3 channels of the microarray scans were averaged within the moving window of width of 114 bp. containing 3 probes.
Locally weighted linear regression (Loess) was applied to log-transformed data after the averaging of the signals. Cy5 and Cy3 dyes perform differently at different average signal intensities but Loess regression compensates for these intensity-dependent effects. Normalization between replicates was performed using quantile normalization. The signal intensities in the Cy5 and Cy3 channels of the replicates were quantile normalized against Cy5 and Cy3 channels of an

arbitrarily chosen array. This procedure forced the signals in each channel to have identical distributions.

After quantile normalization data from replicates was combined by averaging the signals of the replicates. Signal intensity $y_i=\log(R_i)-\log(G_i)$ was assigned to each probe, where $\log(R_i)$ and $\log(G_i)$ are Cy5 and Cy3 channels intensities after performing the aforementioned transformations. Contiguous segments (bars) due to the signal coming from the enriched regions were obtained by joining probes with intensities y_i above the threshold separated by less than a certain distance (max-gap of 114 b.p (Kampa et al. 2004)). Only segments whose length was greater than a particular size (min-run=114 b.p) were selected. Analysis of chromatin immunoprecipitation experiments utilized similar procedures with quantile normalization and a window of 1000 base pairs(Zhang et al. 2007). The false discovery rate (FDR) in our experiments was estimated in the following way. For each dataset the genomic locations of the probes on the microarray were randomly shuffled. The max-gap and min-run procedures described above were applied to the randomized data. The FDR was computed as: $\mathrm{FDR(threshold)} = \mathrm{N1(threshold)/N2(threshold)}$, where N1 is the number of discovered blocks for the randomized data and N2 is the number of blocks for the non-randomized data. FDR as a function of threshold for the different cell lines datasets is shown in Fig. 3C. GENCODE annotation was used for the all the analysis in this study. And all our manual analyses are based on the REFSEQ gene set.

ACKNOWLEDGEMENTS

This work was supported by CEGS NIH grant No.1P50HG02357-01 and UMASS NIH grant No. R01 DK54369. We appreciate Mrs. Janet Hernandez who did great editorial work for this manuscript.

.

References:

(2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447(7146): 799-816.

Ahn SH, Kim M, Buratowski S (2004) Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and 3' end processing. Molecular cell 13(1): 67-76.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE et al. (2007) High-resolution profiling of histone methylations in the human genome. Cell 129(4): 823-837.

Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D (2000) Patterns of variant polyadenylation signal usage in human genes. Genome Res 10(7): 1001-1010.

Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK et al. (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. Cell 120(2): 169-181.

Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE et al. (2004) Global identification of human transcribed sequences with genome tiling arrays. Science (New York, NY 306(5705): 2242-2246.

Birse CE, Lee BA, Hansen K, Proudfoot NJ (1997) Transcriptional termination signals for RNA polymerase II in fission yeast. The EMBO journal 16(12): 3633-3643.

Brodsky AS, Meyer CA, Swinburne IA, Hall G, Keenan BJ et al. (2005) Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells. Genome biology 6(8): R64.

Buratowski S (2003) The CTD code. Nature structural biology 10(9): 679-680.

Buratowski S (2005) Connections between mRNA 3' end processing and transcription termination. Curr Opin Cell Biol 17(3): 257-261.

Carninci P, Hayashizaki Y (2007) Noncoding RNA transcription beyond annotated genes. Current opinion in genetics & development.

Carrozza MJ, Li B, Florens L, Suganuma T, Swanson SK et al. (2005) Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. Cell 123(4): 581-592.

Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S et al. (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science (New York, NY 308(5725): 1149-1154.

Cheng Y, Miura RM, Tian B (2006) Prediction of mRNA polyadenylation sites by support vector machine. Bioinformatics (Oxford, England) 22(19): 2320-2325.

Choi YH, Hagedorn CH (2003) Purifying mRNAs with a high-affinity eIF4E mutant identifies the short 3' poly(A) end phenotype. Proceedings of the National Academy of Sciences of the United States of America 100(12): 7033-7038.

Cuthbert GL, Daujat S, Snowden AW, Erdjument-Bromage H, Hagiwara T et al. (2004) Histone deimination antagonizes arginine methylation. Cell 118(5): 545-553.

De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ et al. (2006) A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. Science (New York, NY 312(5777): 1215-1217.

Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R et al. (2007) Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions. Genome Res 17(6): 746-759.

Ebraliddze A, Tenen DG (2006) Regulation of the PU.1 Gene by Sense and Functionanl Antisense RNAs Generated through the Same Chromatin Architecture. Blood 108 (Supplement): 234a.

Furuno M, Pang KC, Ninomiya N, Fukuda S, Frith MC et al. (2006) Clusters of internally primed transcripts reveal novel long noncoding RNAs. PLoS Genet 2(4): e37.

Gingeras TR (2007) Origin of phenotypes: genes and transcripts. Genome Res 17(6): 682-690.

Goodrich JA, Kugel JF (2006) Non-coding-RNA regulators of RNA polymerase II transcription. Nat Rev Mol Cell Biol 7(8): 612-616.

Hayashizaki Y, Carninci P (2006) Genome Network and FANTOM3: assessing the complexity of the transcriptome. PLoS Genet 2(4): e63.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature genetics 39(3): 311-318.

Ho Y, Elefant F, Liebhaber SA, Cooke NE (2006) Locus control region transcription plays an active role in long-range gene activation. Molecular cell 23(3): 365-375.

Joshi AA, Struhl K (2005) Eaf3 chromodomain interaction with methylated H3-K36 links histone deacetylation to Pol II elongation. Molecular cell 20(6): 971-978.

Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S et al. (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res 14(3): 331-342.

Kapranov P, Willingham AT, Gingeras TR (2007a) Genome-wide transcription and the implications for genomic organization. Nat Rev Genet 8(6): 413-423.

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL et al. (2002) Large-scale transcriptional activity in chromosomes 21 and 22. Science (New York, NY 296(5569): 916-919.

Kapranov P, Drenkow J, Cheng J, Long J, Helt G et al. (2005) Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. Genome Res 15(7): 987-997.

Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R et al. (2007b) RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science (New York, NY 316(5830): 1484-1488.

Kelley RL, Kuroda MI (2000) Noncoding RNA genes in dosage compensation and imprinting. Cell 103(1): 9-12.

Keogh MC, Kurdistani SK, Morris SA, Ahn SH, Podolny V et al. (2005) Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. Cell 123(4): 593-605.

Khanna-Gupta A, Kolibaba K, Zibello TA, Berliner N (1994) NB4 cells show bilineage potential and an aberrant pattern of neutrophil secondary granule protein gene expression. Blood 84(1): 294-302.

Kizer KO, Phatnani HP, Shibata Y, Hall H, Greenleaf AL et al. (2005) A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3 K36 methylation with transcript elongation. Molecular and cellular biology 25(8): 3305-3316.

Koch CM, Andrews RM, Flicek P, Dillon SC, Karaoz U et al. (2007) The landscape of histone modifications across 1% of the human genome in five human cell lines. Genome Res 17(6): 691-707.

Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S et al. (2006) CAGE: cap analysis of gene expression. Nature methods 3(3): 211-222.

Lanotte M, Martin-Thouvenin V, Najman S, Balerini P, Valensi F et al. (1991) NB4, a maturation inducible cell line with t(15;17) marker isolated from a human acute promyelocytic leukemia (M3). Blood 77(5): 1080-1086.

Le Texier V, Riethoven JJ, Kumanduri V, Gopalakrishnan C, Lopez F et al. (2006) AltTrans: transcript pattern variants annotated for both alternative splicing and alternative polyadenylation. BMC bioinformatics 7: 169.

Li B, Gogol M, Carey M, Lee D, Seidel C et al. (2007) Combined action of PHD and chromo domains directs the Rpd3S HDAC to transcribed chromatin. Science (New York, NY 316(5827): 1050-1054.

Liang G, Klose RJ, Gardner KE, Zhang Y (2007) Yeast Jhd2p is a histone H3 Lys4 trimethyl demethylase. Nature structural & molecular biology 14(3): 243-245.

Maeda N, Kasukawa T, Oyama R, Gough J, Frith M et al. (2006) Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. PLoS Genet 2(4): e62.

Maruyama K, Sugano S (1994) Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. Gene 138(1-2): 171-174.

Mattick JS, Makunin IV (2006) Non-coding RNA. Human molecular genetics 15 Spec No 1: R17-29.

Migliaccio G, Di Pietro R, di Giacomo V, Di Baldassarre A, Migliaccio AR et al. (2002) In vitro mass production of human erythroid cells from the blood of normal donors and of thalassemic patients. Blood cells, molecules & diseases 28(2): 169-180.

Morillon A, Karabetsou N, Nair A, Mellor J (2005) Dynamic lysine methylation on histone H3 defines the regulatory phase of gene transcription. Molecular cell 18(6): 723-734.

Moucadel V, Lopez F, Ara T, Benech P, Gautheret D (2007) Beyond the 3' end: experimental validation of extended transcript isoforms. Nucleic acids research 35(6): 1947-1957.

Nag A, Narsinh K, Kazerouninia A, Martinson HG (2006) The conserved AAUAAA hexamer of the poly(A) signal can act alone to trigger a stable decrease in RNA polymerase II transcription velocity. RNA (New York, NY 12(8): 1534-1544.

Ng P, Tan JJ, Ooi HS, Lee YL, Chiu KP et al. (2006) Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. Nucleic acids research 34(12): e84.

Osheim YN, Sikes ML, Beyer AL (2002) EM visualization of Pol II genes in Drosophila: most genes terminate without prior 3' end cleavage of nascent transcripts. Chromosoma 111(1): 1-12.

Pang KC, Frith MC, Mattick JS (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet 22(1): 1-5.

Pauler FM, Koerner MV, Barlow DP (2007) Silencing by imprinted noncoding RNAs: is transcription the answer? Trends Genet 23(6): 284-292.

Phatnani HP, Greenleaf AL (2006) Phosphorylation and functions of the RNA polymerase II CTD. Genes & development 20(21): 2922-2936.

Prasanth KV, Spector DL (2007) Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. Genes & development 21(1): 11-42.

Prashar Y, Weissman SM (1996) Analysis of differential gene expression by display of 3' end restriction fragments of cDNAs. Proceedings of the National Academy of Sciences of the United States of America 93(2): 659-663.

Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM et al. (2003) The transcriptional activity of human Chromosome 22. Genes & development 17(4): 529-540.

Roma G, Cobellis G, Claudiani P, Maione F, Cruz P et al. (2007) A novel view of the transcriptome revealed from gene trapping in mouse embryonic stem cells. Genome Res.

Rozowsky J, Wu J, Lian Z, Nagalakshmi U, Korbel JO et al. (2006) Novel transcribed regions in the human genome. Cold Spring Harbor symposia on quantitative biology 71: 111-116.

Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ et al. (2002) Using the transcriptome to annotate the genome. Nature biotechnology 20(5): 508-512.

Santos-Rosa H, Schneider R, Bannister AJ, Sherriff J, Bernstein BE et al. (2002) Active genes are tri-methylated at K4 of histone H3. Nature 419(6905): 407-411.

Schneider R, Bannister AJ, Myers FA, Thorne AW, Crane-Robinson C et al. (2004) Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. Nature cell biology 6(1): 73-77.

Secombe J, Li L, Carlos L, Eisenman RN (2007) The Trithorax group protein Lid is a trimethyl histone H3K4 demethylase required for dMyc-induced cell growth. Genes & development 21(5): 537-551.

Shamovsky I, Nudler E (2006) Gene control by large noncoding RNAs. Sci STKE 2006(355): pe40.

Sims RJ, 3rd, Belotserkovskaya R, Reinberg D (2004) Elongation by RNA polymerase II: the short and long of it. Genes & development 18(20): 2437-2468.

Subrahmanyam YV, Yamaga S, Prashar Y, Lee HH, Hoe NP et al. (2001) RNA expression patterns change dramatically in human neutrophils exposed to bacteria. Blood 97(8): 2457-2468.

Taft RJ, Pheasant M, Mattick JS (2007) The relationship between non-protein-coding DNA and eukaryotic complexity. Bioessays 29(3): 288-299.

Tsukahara Y, Lian Z, Zhang X, Whitney C, Kluger Y et al. (2003) Gene expression in human neutrophils during activation and priming by bacterial lipopolysaccharide. Journal of cellular biochemistry 89(4): 848-861.

Venkataraman K, Brown KM, Gilmartin GM (2005) Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. Genes & development 19(11): 1315-1327.

Wei CL, Ng P, Chiu KP, Wong CH, Ang CC et al. (2004) 5' Long serial analysis of gene expression (LongSAGE) and 3' LongSAGE for transcriptome characterization and genome annotation. Proceedings of the National Academy of Sciences of the United States of America 101(32): 11701-11706.

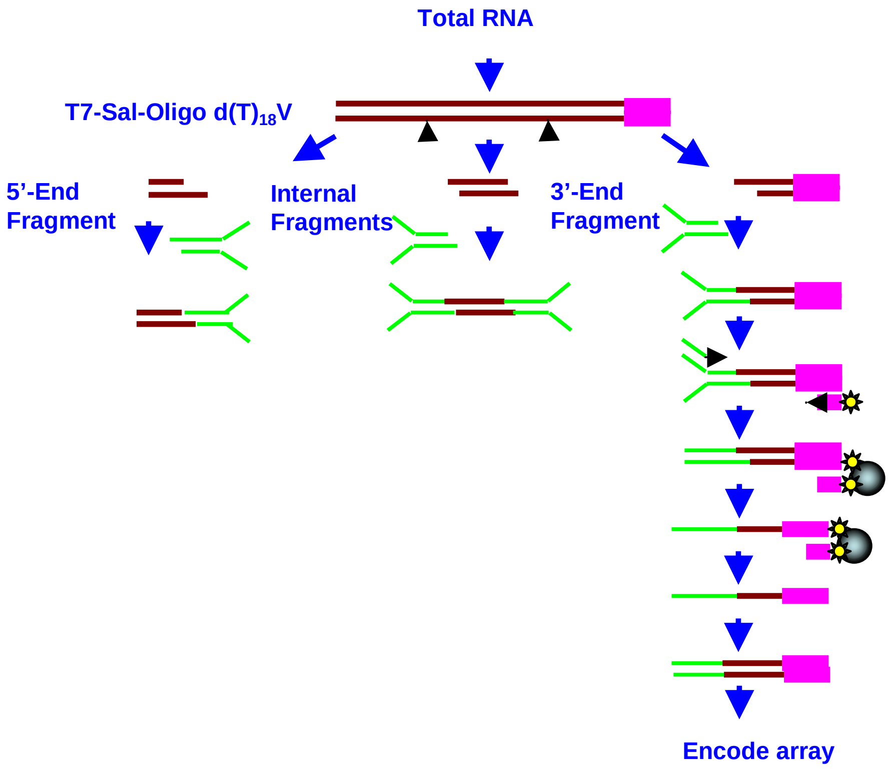
West S, Zaret K, Proudfoot NJ (2006) Transcriptional termination sequences in the mouse serum albumin gene. RNA (New York, NY 12(4): 655-665.

Xiang P, Fang X, Yin W, Barkess G, Li Q (2006) Non-coding transcripts far upstream of the epsilon-globin gene are distinctly expressed in human primary tissues and erythroleukemia cell lines. Biochemical and biophysical research communications 344(2): 623-630.

Yang PK, Kuroda MI (2007) Noncoding RNAs and intranuclear positioning in monoallelic gene expression. Cell 128(4): 777-786.

Zhang ZD, Rozowsky J, Lam HY, Du J, Snyder M et al. (2007) Tilescope: online analysis pipeline for high-density tiling microarray data. Genome biology 8(5): R81.

Zhao H, Kim A, Song SH, Dean A (2006) Enhancer blocking by chicken beta-globin 5'-HS4: role of enhancer strength and insulator nucleosome depletion. The Journal of biological chemistry 281(41): 30573-30580.



(1) cDNA Synthesis

(2) ds cDNA
Restriction Enzyme Digestion

(3) Butterfly-Adaptor
Ligation

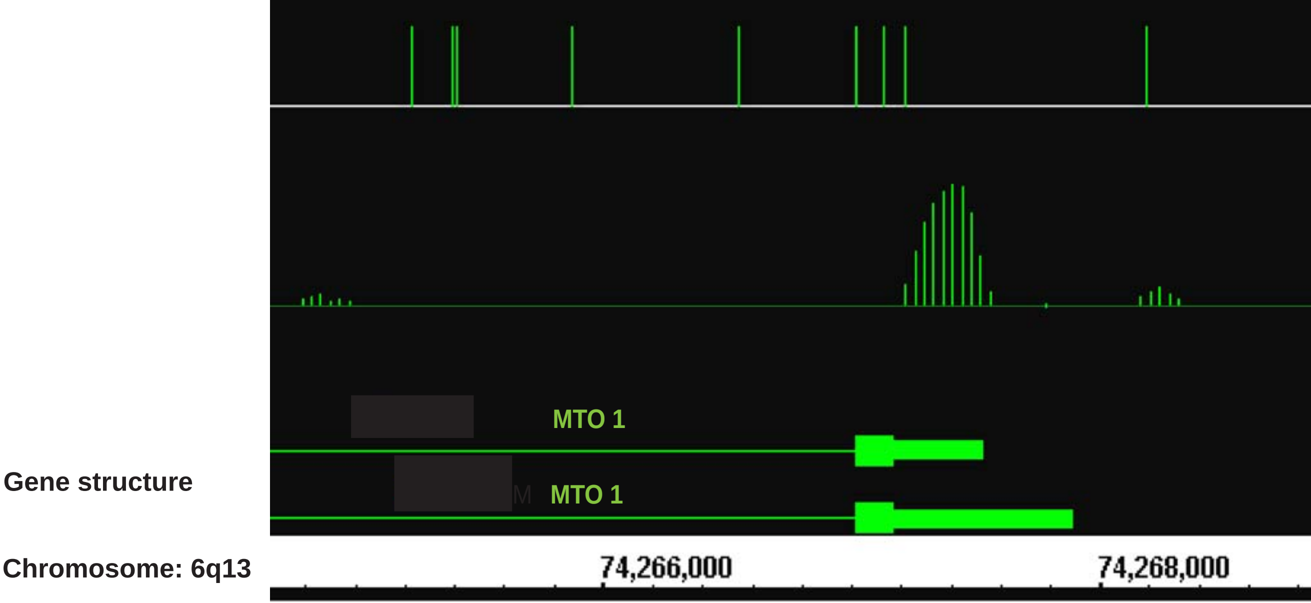
(4) PCR with Biotinylated oligo
dT

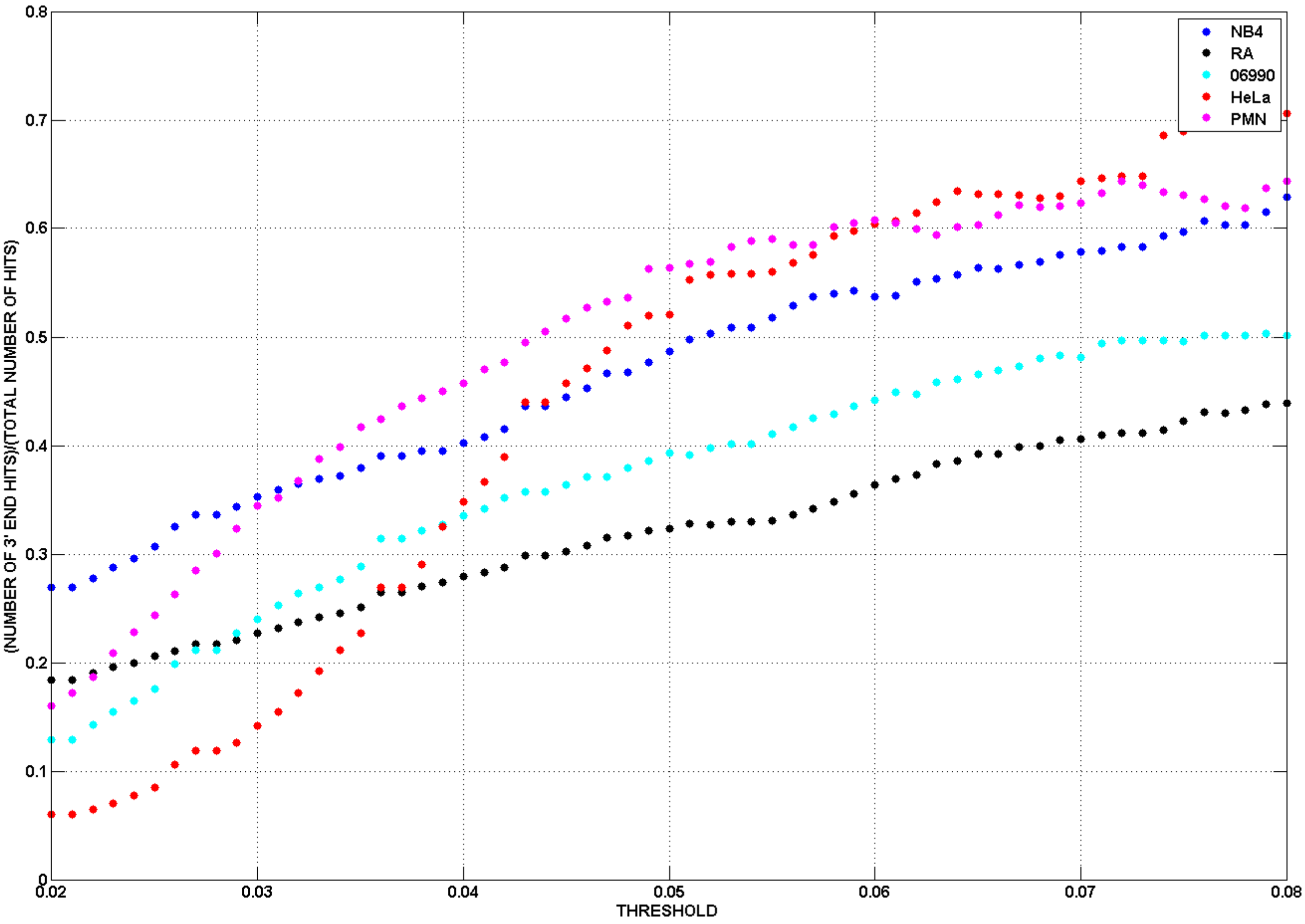
(5)Streptavidin bead
purification

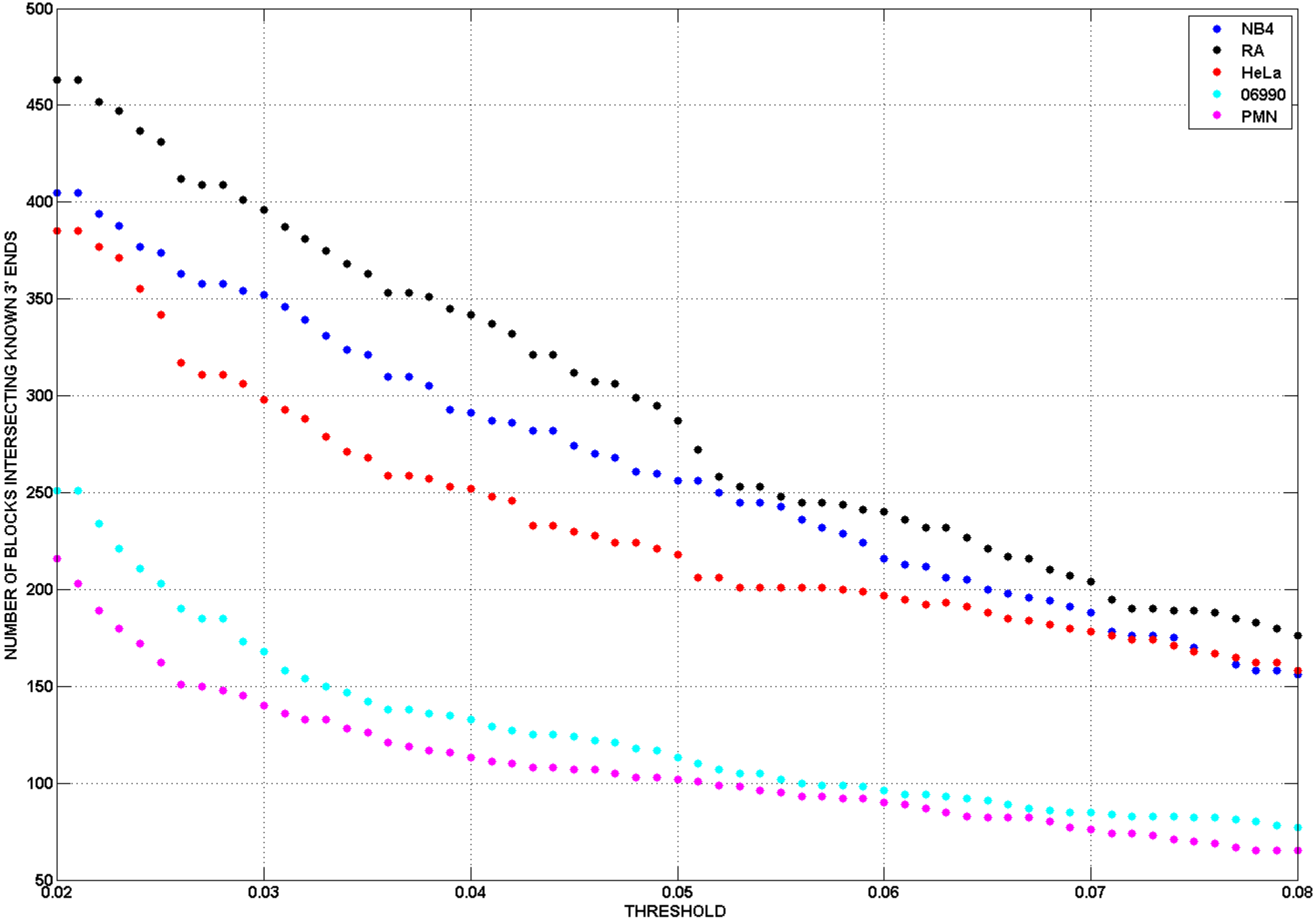
(6)Alkali denature dsDNA from
beads

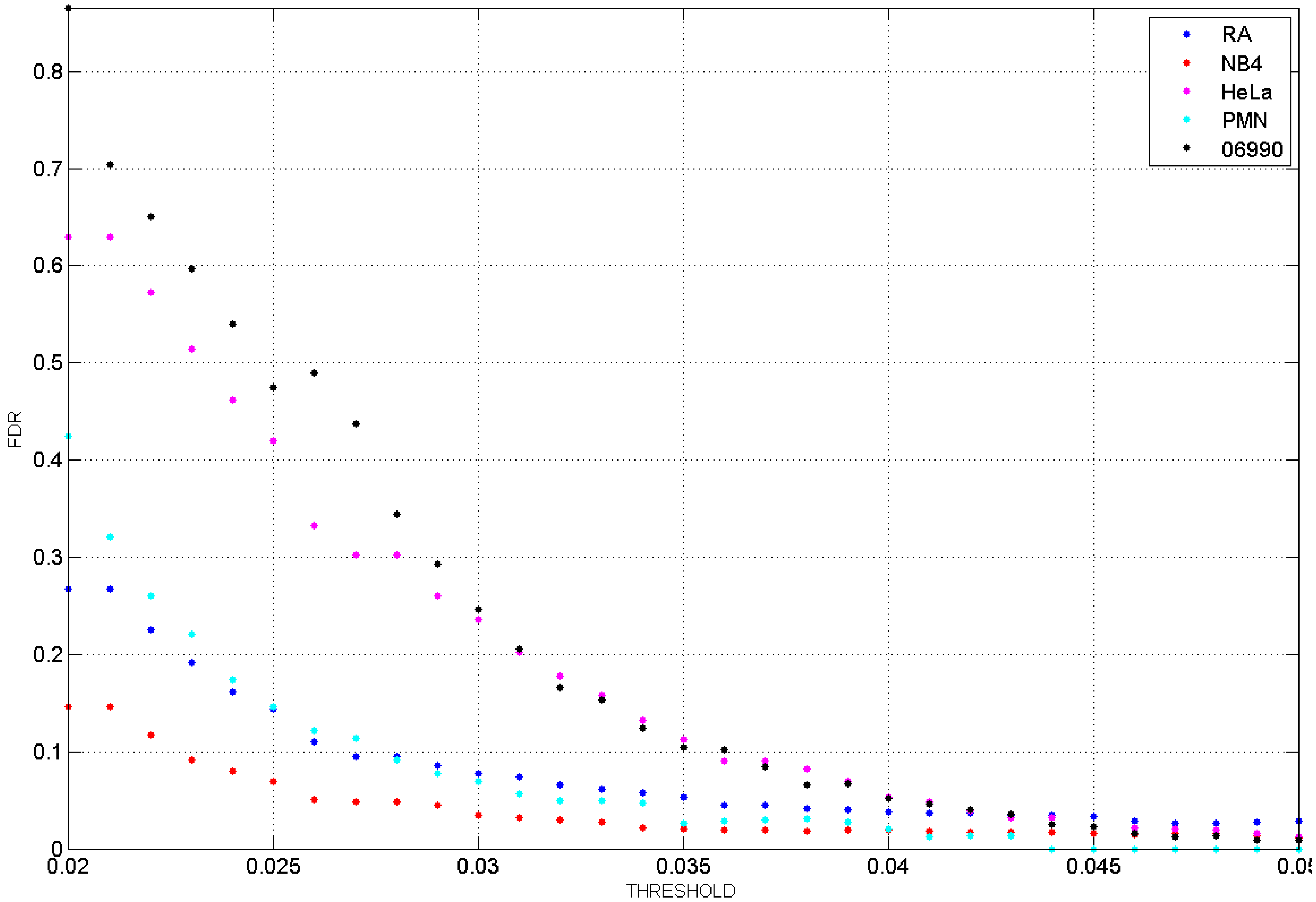
(7)Collect none biotinylated
strand of cDNA

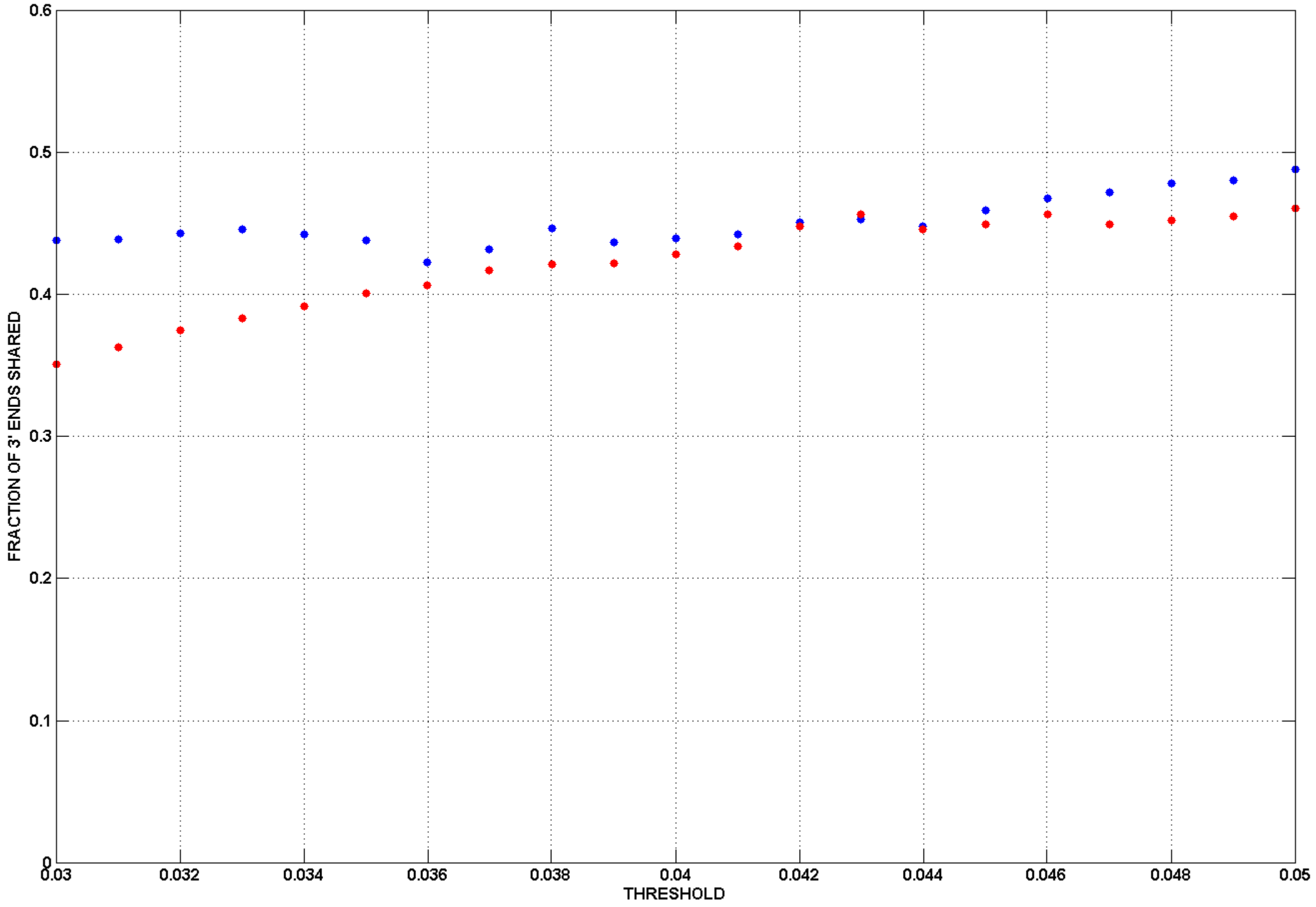
(8) PCR amplification

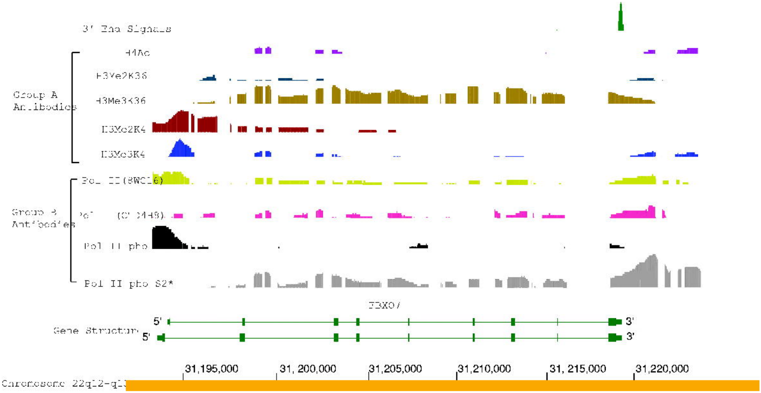

[redacted]
MTO 1
Gene structure
[redacted]
MTO 1
Chromosome: 6q13
74,266,000
74,268,000

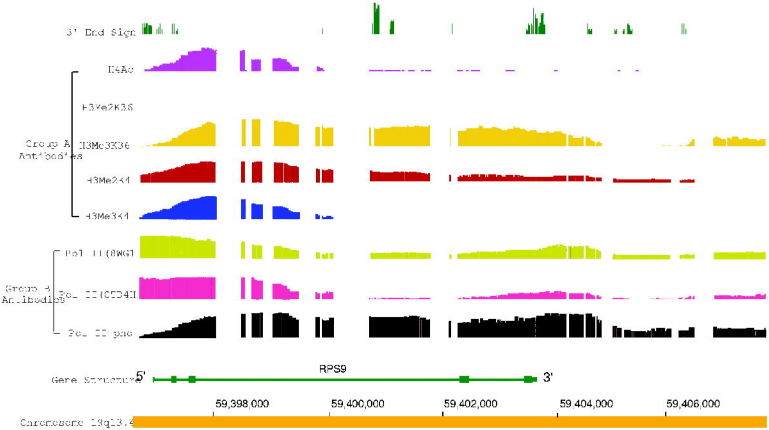

NB4
RA
06990
HeLa
PMN
(NUMBER OF 3' END HITS)/(TOTAL NUMBER OF HITS)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.02
0.03
0.04
0.05
0.06
0.07
0.08
THRESHOLD

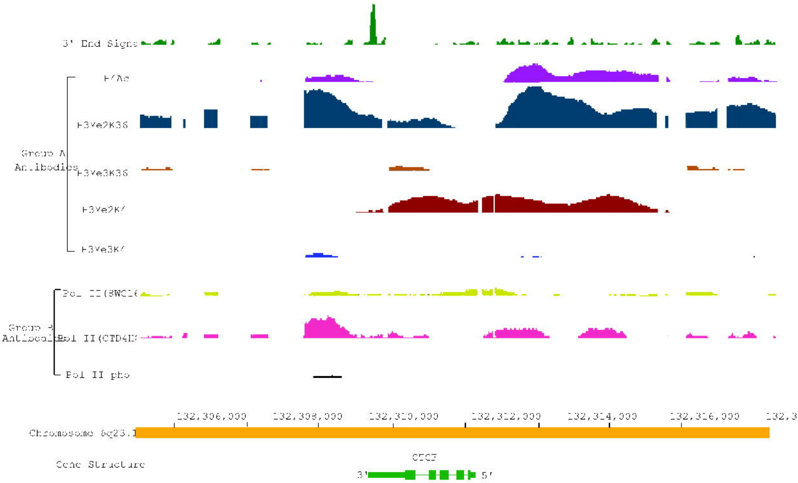

NB4
RA
HeLa
06990
PMN
NUMBER OF BLOCKS INTERSECTING KNOWN 3' ENDS
THRESHOLD
500
450
400
350
300
250
200
150
100
50
0.02
0.03
0.04
0.05
0.06
0.07
0.08

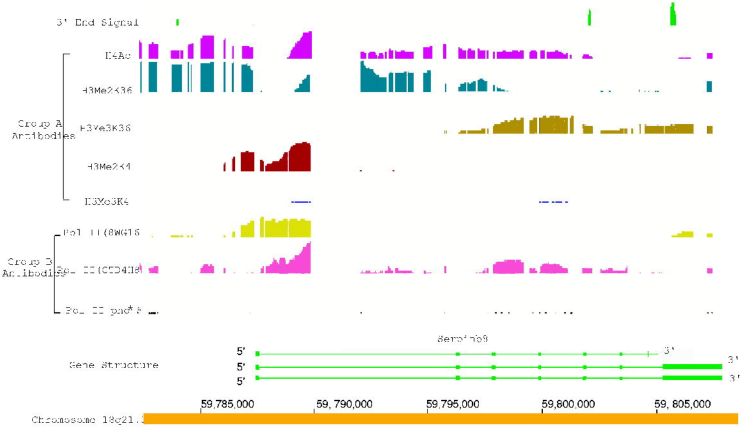

FDR
THRESHOLD
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.02
0.025
0.03
0.035
0.04
0.045
0.05
RA
NB4
HeLa
PMN
06990


0.6
0.5
0.4
0.3
0.2
0.1
0
FRACTION OF 3' ENDS SHARED
0.03
0.032
0.034
0.036
0.038
0.04
0.042
0.044
0.046
0.048
0.05
THRESHOLD


3' End Signals
H4Ac
H3Me2K36
H3Me3K36
H3Me2K4
H3Me3K4
Group A Antibodies
Pol II (8WG16)
Pol II (CTD4H8)
Pol II pho
Pol II pho S2*
Group B Antibodies
FBXO7
Gene Structure
5'
3'
5'
3'
31,195,000
31, 200,000
31,205,000
31,210,000
31, 215,000
31,220,000
Chromosome 22q12-q13


3' End Sign
H4Ac
H3Me2K36
Group A Antibodies
H3Me3K36
H3Me2K4
H3Me3K4
Pol II(8WG1
Group B Antibodies
Pol II(CTD4H
Pol II pho
Gene Structure
5'
RPS9
3'
59,398,000
59,400,000
59,402,000
59,404,000
59,406,000
Chromosome 19q13.4


3' End Signa
H4Ac
H3Me2K36
Group A
Antibodies
H3Me3K36
H3Me2K4
H3Me3K4
Pol II (8WG16
Group B
Antibodies
Pol II (CTD4H
Pol II pho
132,306,000
132,308,000
132,310,000
132,312,000
132,314,000
132,316,000
132,3
Chromosome 6q23.1
Gene Structure
CTGF
3'
5'


3' End Signal
H4Ac
H3Me2K36
Group A
Antibodies
H3Me3K36
H3Me2K4
H3Me3K4
Pol II(8WG16
Group B
Antibodies
Pol II(CTD4H8
Pol II pho*S
Serpinb9
5'
3'
Gene Structure
5'
3'
5'
3'
59,785,000
59, 790,000
59,795,000
59,800,000
59, 805,000
Chromosome 18q21.3