The role of sequence, gene orientation, and intergenic distance in chromatin structure and function

By

Sasha A. Langley

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Molecular and Cellular Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Gary Karpen, Chair
Professor Mike Eisen
Professor Jasper Rine
Professor Renee Sung

Spring 2010

Abstract

The role of sequence, gene orientation, and intergenic distance in chromatin structure by

Sasha A.Langley

Doctor of Philosophy in Molecular and Cellular Biology

University of California, Berkeley

Professor Gary Karpen, Chair

Remarkable conservation of patterns of histone modifications and variants has been observed at active genes [1,2]. We show that genome-wide average distributions of chromatin marks relative to transcription start sites (TSSs) in early *Drosophila melanogaster* embryos are largely consistent with previously observed patterns in other organisms. This work describes qualitatively different chromatin domains and draws with functional inferences about these regional differences in modification patterns. We find 'plateaus' of the histone variant, H2Av spanning transcriptionally inactive loci, as observed in mammalian cells [3]. These mixed cell populations from embryos exhibit distributions of chromatin marks similar to bivalent domains, specialized regions which extend across and beyong genes involved in developmental regulation and cell differentiation in mammalian stem cells; 5' ends of genes in these regions lack H2Av, regardless of the transcriptional state of the genes.

In spite of the agreement of our initial studies with similar work in other organisms or tissues, analysis of chromatin patterns for genes grouped with respect to promoter orientation and distance from the adjacent upstream gene identifies differences in the positions and enrichment levels of active mark peaks, as well as levels of gene activity. Analysis of published data from human CD4+ cells and *Saccharomyces cerevisiae* reveals that many of the relationships between promoter context and the distributions of active marks surrounding TSSs are conserved among widely divergent eukaryotes. We propose that short-range, distance-dependent synergistic interactions between neighboring promoters impact both chromatin state and gene activity.

The functional consequences of DNA sequence variation for the specificity and stability of nucleosome associations are largely unknown but open to new avenues of investigation based on emerging DNA sequencing technologies. We present preliminary results on interaction between molecular evolution and nucleosome positioning in Drosophila. Base composition surrounding nucleosomes isolated from *melanogaster* embryos supports some level sequence-encoded higher order structure with a periodicity of ~200bp. Analysis of the interspecific divergence shows that the rates of change in these regions support an equilibrium model maintaining these patterns. We note a substantially accelerated GC->AT rate on the *melanogaster* lineage. Dinucleotide periodicities across nucleosomal fragments are quite similar between *melanogaster* and *simulans*, and we find good support for conserved positional changes in in *simulans* relative to nucleosomes isolated from *melanogaster*.

Table of Contents

Chapter 1

Chromatin domains in stage 5 *D. melanogaster* embryos

Background

Chromatin structure and control of DNA accessibility are fundamental components of eukaryotic chromosome organization and function, as well as gene regulation. Post-translational modifications of histones, such as methylation or acetylation of lysine residues, as well as histone variants, are associated with a broad range of chromatin-mediated processes, including heterochromatin formation, DNA repair, centromere formation, and developmental differentiation of cells [1,2]. Patterns of modifications and substitution of histone variants for canonical histones influence epigenetic states through their effects on nucleosome stability, chromatin fiber dynamics, and their ability to serve as docking sites for other structural proteins and modifiers [1,2].

Fine-scale mapping and combinatorial analysis of the euchromatic genomic locations of histones, their variants, and modifications in *S. cerevisiae* and higher eukaryotes have established a tight relationship between epigenetic states and transcriptional status. Methylation of histone H3 at Lysine 4 (H3K4me), present in regions of open or active chromatin, functions in RNA polymerase II (Pol II)-dependent transcription and occurs in a 5'-to-3' tri- to di- to monomethyl gradient across transcribed genes [1,2]. In addition to the relationship between K4 methylation and gene expression, H3K4me2 has been observed in centromeric chromatin interspersed between regions containing the centromere-specific histone H3 variant CENP-A, a feature conserved between mammals, plants, and flies [3].

Histone H3 methylations also play a conserved role in developmental regulation in many multicellular eukaryotes. In stem cells, bivalent domains, characterized by the presence of both H3K4me3 and trimethylation of Histone H3 Lysine 27 (H3K27me3), necessary for the binding of Polycomb group (PcG) proteins, span genes involved in differentiation and development [4-8]. These large expanses of overlapping of K4 and K27 trimethylation enrichment represent a basal chromatin state associated with pluripotency [4-8]. Induced differentiation of cultured mammalian stem cells demonstrated that bivalent domains lose either K4 trimethylation upon silencing or of K27 when cell fate requires gene activation [4-8]. Hox gene expression in Drosophila is promoted by the displacement of the H3K4 demethylase RBP2/Lid [9-11]. In mammals, the H3K4 methyl binding protein WDR5 recruits methyltransferases required for gene activation, suggesting that this modification must be maintained for proper expression [12]. Distinctive differences in distribution patterns of H3K4me3 in bivalent domains and surrounding other euchromatic transcription start sites (TSSs) illustrates that modifications can play different roles in different contexts.

The Drosophila histone H2A variant, H2Av, also has diverse, context-dependent functions. Phenotypic analysis of *H2Av* mutants has implicated the variant in silencing of transgene arrays, genetic interaction with PcG genes, and PEV suppression [13]. H2Av mutants show reduced H4K12 acetylation, H3K9 methylation, and Heterochromatin Protein 1 (HP1) recruitment to heterochromatin, suggesting a the variant is involved in heterochromatin establishment [13]. In addition, the Ser137 phosphorylated form of H2Av marks double strand breaks (DSB) and plays a role in repair of these lesions in Drosophila [14].

In *S. cerevisiae*, the H2Av ortholog, H2A.Z, plays roles in both regulating gene expression and boundary maintenance [15-17]. This variant flanks the Nucleosome Free Region (NFR) at transcription start sites (TSS), and acetylated H2A.Z antagonizes the spread of silencing factors in telomeric regions [16]. In a type of epigenetic memory, H2A.Z mediates the localization of recently repressed genes at the nuclear periphery and promotes reactivation of repressed loci [18]. In *Arabidopsis*, H2A.Z is required for activating a repressor of the transition to flowering [19]. Sequencing of chromatin immunoprecipitation (ChIP) products in human cells demonstrated a correlation of this variant with expression, as well as low enrichment over the silent MYO1B locus [20]. These results imply a critical role for this H2A variant in mediating expression and, particularly, in facilitating the transition between short-term silencing and activation.

H2Av and mammalian H2A.Z have been localized to both pericentric heterochromatin and the euchromatin [20-25]. Heterochromatic association is dynamic and coupled to differentiation during embryonic development of the mouse [23]. Nucleosomes containing H2A.Z and H3K4me2 are interspersed between subdomains of CENP-A at centromeres [22]. In mammals, the heterochromatization of the X and Y chromosomes following meiosis is accompanied by deposition of H2A.Z-containing nucleosomes [26]. RNAi targeting H2A.Z in mammalian (murine) cell culture leads to defects in chromosome segregation and a decrease in HP1-alpha binding in both euchromatic and pericentric regions [24]. Localization in multiple chromatin domains and the functional importance of H2A.Z and H2Av in several fundamental biological processes may explain some of the pleiotropy seen in mutants.

Heterochromatin is, in part, defined by post-translational histone modifications associated with gene silencing. The canonical epigenetic mark of heterochromatin Histone H3 Lysine 9 dimethyl (H3K9me2) is added by Su(var)3-9 and other methyltransferases and recruits HP1 [27,28]. H3K9 methylation first becomes visible at cycle 14 in Drosophila embryos, and is enriched over repeated DNA elements [29-31]. Although it is generally associated with gene silencing and PEV, heterochromatin in Drosophila and other organisms contains a significant number of genes, many of which are essential for viability, as well as clusters of the ribosomal RNA genes (rDNA), which are the most highly transcribed genes in the genome [32,33]. Surprisingly, many of these genes require a heterochromatic environment for proper expression [34]. This raises important questions about potential differences in the mechanisms of gene activation in heterochromatin. However, much less is known about the fine-scale modification and variant patterns in these regions due to high repeat content and the paucity of unique sequence necessary for ChIP-array or ChIP-seq mapping.

We have performed Native Chromatin Immunoprecipitation (NChIP)-array mapping of H3K9Me2, H3K4Me2, H3K4Me3, H3K27me3 and the histone variant H2Av in chromatin purified from *Drosophila* stage 5 embryos (blastoderm). The major advantages of performing such chromatin ‘landscape’ studies in Drosophila are the availability of significant amounts of assembled and annotated heterochromatic sequence, completely assembled euchromatin, and an expansive annotation of genes and other functional elements.

Our analyses suggest a potential role for H2Av in Drosophila in gene silencing and a distribution similar to that observed in humans [20]. These data also argue against recent reports that Drosophila embryos lack bivalent domains, although the results are not conclusive, due to the mixed cell types in stage 5 embryos [35]. We identify two functionally distinct classes of genes with unique patterns of chromatin marks (“epigenetic profiles”) at transcription start sites and across genes. Finally, we observe unexpected similarity in the patterns of epigenetic marks associated with transcription in the heterochromatin. This whole genome characterization provides a resource for further studies of differentiated organs and tissues, and a point of comparison for the analyses of mutants that effect chromatin and its developmental remodeling.

Results

Two distinct chromosomal domains across euchromatic genes

On a gross scale, our data show the division of chromosomes into two types of transcriptional territories. In addition to large domains of overlapping H3K4me3 and K27me3, we see, as expected, H3K4me2 and K4me3 surrounding clusters of other euchromatic genes (Fig. 1A). In these regions, we also observe enrichment for H2Av, which shows a strong association with H3K4 methylation outside of H3K4 and K27me3 domains. In the euchromatin, the genomic sequence outside of these two types of chromatin territories displays relatively little enrichment for the modifications considered here or for the variant, H2Av. H3K9Me2 is limited to the pericentric heterochromatin in the stage 5 embryo, where it broadly occupies genic and intergenic regions, with the exception of those regions bound by Pol II (Fig. 1A).

The overlap of H3K4me3 and H3K27me3 in wide plateaus across potential bivalent domains is illustrated by the presence of a group of strongly correlative points along the diagonal when the gene average Normalized Log Ratios (NLRs) across genes are plotted against each other (Fig. 1B). The majority of genes, however, form a large off-diagonal cluster of H3K4me3-enriched loci with very low levels of H3K27me3. Although some of these may represent activated bivalent domain genes, most of them belong to the other stereotypical class of euchromatic genes distinguished by the presence of H2Av. The NLR across transcriptional units for H2Av and H3K4me2 and me3 cluster along the diagonal, illustrating the tendency of these marks to occur together across most (active) genes. H3K4me2 and me3 show the tightest association, as expected, and H3K27me3 and H2Av are enriched somewhat exclusively of each other (Fig. 1B). In order to assess the broad relationships of these modifications and H2Av to transcription, we plotted gene averages against those for Pol II. These plots highlight the general association of H3K4me2 and me3 with transcription and the expected inverse relationship of the repressive mark, H3K27me3, to polymerase enrichment (Fig. 1B). The comparison of H2Av and Pol II intensities across genes, however, suggests a more complex connection between the variant and transcription (Fig. 1B).

Transcription-associated enrichment of H3K4 methylation and H2Av at transcription start sites (TSSs)

We next examined the average patterns of these marks at annotated euchromatic genes. Transcription-linked H3K4 methylation is conserved across a broad spectrum of organisms. Averaged NLRs suggested that H3K4 methylation is enriched immediately upstream and downstream of transcription start at genes actively transcribed in stage 5 embryos (Fig. 1C), consistent with previous array data that covered a subset of the Drosophila genome [36]. In addition to the presence of K4 methylation, we observe that H2Av occupies the 5’ ends of actively transcribed genes. Unlike the stereotypic poised TSS in *S. cerevisiae* flanked by single H2A.Z-containing nucleosomes, the distribution of H2Av in Drosophila embryos extends further up and downstream of TSS and is largely coincident with K4 methylation [17,37,38]. Averaged transcription termination sites (TTSs) show low levels of up and downstream enrichment for K4 methylation and H2Av (Fig. 1C). The downstream enrichment may be the result of clustered distribution of active genes or local depletion or destabilization of nucleosomes at the 3' end of genes. The average pattern at TTSs is characterized by a sharp local drop in K4 methylation and a peak of phosphorylated Pol II binding.

H3K4 methylation is correlated with gene activity, and H3K27me3 is a repressive modification, but the relationship of H2Av (and H2A.Z) to expression is less clear cut. To compare levels of enrichment and gene expression, we separated genes into quartiles based on expression level in early embryos and plotting average intensities at transcription initiation sites (Fig. 2) [39]. H3K4me2 (Fig. 2A), H3K4me3 (Fig. 2C), and H2Av (Fig. 2B) show a transcription level-dependent enrichment at the 5' ends of genes. In contrast, H3K27me3 (Fig. 2D), shows the inverse profile, with higher NLR intensities associated with lower transcription levels on average. At stage 5, H3K9me2 enrichment is limited to heterochromatic regions and shows no association with euchromatic genes (results not shown).

These results confirm the enrichment of H3K4 methylation surrounding TSSs observed previously in many organisms. H2Av patterns suggest that this variant may play a role in the regulation of transcription in Drosophila distinct from those described in *S. cerevisiae*, and similar to that of H2A.Z in mammals [20]. These results also distinguish H2Av from another variant, H3.3, which is peaked at 5' ends but also enriched across active genes in a broad pattern which extends across the entire transcription unit, similar to the enrichment of Pol II [40,41].

Transcriptionally inactive genes display coding region plateaus of H2Av but not H3K4 methylation.

In *S. cerevisiae*, H2A.Z is associated with promoters of repressed genes, and with active genes when acetylated at K14 [15,38,42]. In addition to the localization of H2Av with methylated H3K4 at the 5' regions of active genes, we observe a separate class of genes that display a "plateau" of H2Av across the locus, in the absence of K4 methylation (Fig. 3A). In order to identify genes enriched for H2Av and lacking K4 methylation, we computed the average NLR for H2Av and H3K4me3 across the largest transcript for all genes, and selected genes based on enrichment of H2Av and depletion of H3K4me3 (see methods).

A comparison of the distributions of expression levels revealed that H2Av "plateau" genes are less transcriptionally active than all genes (Fig. 3C). Additionally, Pol II is not enriched at the majority of these genes (Fig. 3D), and their TSSs and TTSs lack the 5' peaks of H3K4me2 and me3 (Fig. 3B) observed at active genes (Fig. 1C). These observations suggest that H2Av can mark transcriptionally inactive or repressed genes. H2Av "plateau" genes are functionally diverse, but show significant overrepresentation of genes involved in sensory perception or encoding products not active in early embryonic development (data not shown).

These H2Av enriched loci are located in clusters of actively transcribed genes. Furthermore, they are frequently within close proximity to genes bound by polymerase; these upstream Pol II signals and K4 methylation in the TSS and TTS profiles for H2Av plateau genes (Fig. 3C) are not observed at all genes or active genes (Fig. 1C). Genes flanking H2Av plateau loci can have their TSS or TTS most proximal to the plateau gene. Upstream genes fall into two categories, with either the TSS (Fig 4A) or TTS (Fig. 4B) adjacent to the H2Av plateau gene's 5' region. Neighbors with TSSs upstream of H2Av plateau genes lack the double peaks of H3K4 di- and tri-methylation that flank TSSs (Fig. 4C) in the averaged euchromatic profile or that of active genes (Fig. 1C). Instead, this group shows a single peak of K4 methylation downstream of the TSS. In contrast, the TTS profile of genes with the 3' end adjacent to H2Av promoters (Fig. 4D) is similar to that of all euchromatic genes, with the exception of an overall increase in H2Av and a small increase in H3K4 methylation and Pol II at the 3' end. Average Pol II density across genes upstream of H2Av plateau loci are higher than those for all genes (data not shown), suggesting that, while they are not themselves expressed, H2Av plateau genes are more likely to be located near other actively transcribed genes. We propose that H2Av functions in the epigenetic silencing of genes located in clusters of active genes and that presence of H2Av plateaus may represent structural changes of the chromatin in these regions.

Domains enriched for H3K4 and K27 trimethylation lack H2Av, independent of transcriptional status

As observed in other organisms, we see broad, overlapping distributions of H3K4 and K27 trimethylation across genes involved in development and differentiation [4-8]. At this early embryonic stage, many genes are expected to remain in the poised, undifferentiated state. However, due to the mixed population of cells in the embryo, such regions of overlap could represent a mixture of differentiated and possibly bivalent states or true bivalent domains common to the majority of cells.

Genes associated with bivalent domains are often developmental regulators, such as Gap and pair-rule proteins, involved in segmentation and embryo patterning. Many of these transcription factors are regulated by other members of the same class. We noted that H2Av and H3K27me3 were not enriched over the same genes (Fig. 1A,B). In order to test if H2Av itself is generally excluded from activated bivalent domain genes, we used BDNTP (http://bdtnp.lbl.gov/Fly-Net/) mapping data from stage 5 embryos, and examined the average modification profiles for the set of genes closest to peaks of binding for several transcription factors [43]. Several of these, including identified targets of the Giant transcription factor (Fig. 5D) show a similar absence of H2Av in 5' regions and of lack of K4 methylation upstream of transcription start sites.

The set of putative Giant-regulated genes includes genes bound and unbound by Pol II and/or TFIIB in stage 5 embryos (data from the BDNPT) [43]. These expression differences occur in a position-dependent developmentally regulated manner across the embryo. A comparison of target genes that are bound by polymerase or TFIIB (Fig. 5E) to those that are not (Fig. 5F) shows that average H3K4me2 and me3 intensities are increased significantly across transcribed regions of genes bound by transcription-associated proteins, while H2Av shows no enrichment for either class.

By separating binding site associated genes into quartiles, we found that this characteristic profile becomes more evident with increasing score (provided by BDNTP) of the binding site prediction (data not shown). This suggests that binding of these factors is highest in bivalent domains, but that they may also participate in the regulation of genes outside those domains.

Modification profiles surrounding transcription factor binding sites show variation in H3K4 and K27 methylation and an absence of H2Av.

Bivalent domains extend well beyond genic sequences into putative and known regulatory regions. Although, the NLR profiles surrounding identified binding sites for several developmental regulators at stage 5 show variability, they share distinctive features not seen in other euchromatic intergenic regions (data from the BDNTP) [43]. For many factors, H3K4me3 is peaked at binding sites, on average, demonstrating preferential occupancy of these regions by H3K4me3 marked nucleosomes (Fig. 5A,B). Viewed individually, some data points cluster directly over binding sites with much higher than average intensities and others show an obvious negative enrichment for H3K4me3 directly over the binding site (data not shown). These may represent binding sites where histones with trimethylated H3K4 are either occupying the site in the majority of cells or H3K4triMe nucleosomes are excluded in most cells due to transcription factor binding.

In spite of its presence across putative bivalent domains in our data, for the factors for which data is available, only Caudal showed elevated H3K27me3 surrounding binding sites at stage 5. H3K4me2 also shows a modest increase around binding sites. On average, H2Av does not occupy regions of TF binding for any of the factors examined, confirming the observation that this variant is absent from bivalent domains.

Gypsy8 transposable elements are potential heterochromatic insulators

Heterochromatin is composed primarily of satellites, fragments of transposable elements, and other repetitive sequences, interspersed with short blocks of unique coding sequence [32]. Cytological heterochromatin first becomes visible in the *Drosophila* embryo at the apical pole of early blastoderm nuclei [31,44]. Following cellularization, H3K9me2 and other heterochromatic markers begin to colocalize in this region [45]. Our data therefore represent the very early stages of heterochromatin establishment in the embryo.

Consistent with cytological findings, H3K9me2 is variably but broadly distributed across the pericentric heterochromatin at stage 5 and is largely absent in the euchromatin (Fig. 6A,B). Unlike heterochromatin on the autosomes, the heterochromatic regions X chromosome included on the arrays do not appear to have enrichment for H3K9me2. Previous cytological studies indicated that H3K9me2 distribution overlapped with that of H2Av or H3K4me3 in the heterochromatin [13]. Our fine-scale mapping suggests that gene boundaries and transcriptional status primarily determine the presence of H2Av and K4 methylation in the known unique heterochromatic sequences of embryonic cells.

All classes of heterochromatic transposable elements considered (those with sufficient surrounding unique sequence and high enough copy number) are, on average, in regions enriched for H3K9me2 and lacking signal for H3K4me2 or me3, H3K27me3, or H2Av (Fig. 6C shows averages for Gypsy elements, other classes not shown), with the exception of the Gyspy8 subgroup (Fig. 6D). These elements often coincide with small islands (<5Kb) of H3K4me3 and H3K27me3 located within larger blocks of H3K9me2. Heterochromatic Gypsy elements as a whole do not exhibit this pattern at stage 5 (Fig. 6C). These "islands" often flank genes or lie in introns. A subset of these regions, not associated with Gypsy8 elements, surrounds 4 small hypothetical genes with homology only to other closely related Drosophilids. Only two of these genes are supported by cDNAs, both of which contain large introns, and none are bound by Pol II at stage 5. In spite of the presence of overlapping H3K4 and K27 trimethylation, these domains may not share a functional relationship with such domains in the euchromatin. They may instead may function as insulators. Mutations of mod(mdg4), a member of the Su(Hw) insulator complex which binds Gypsy insertions, are known to genetically interact with PcG and TrxG mutations [46]. No direct interaction between these complexes is known, but our data suggest that heterochromatic insertions of Gypsy8 may recruit PcG and TrxG proteins in the early embryo [46].

Distinctive chromatin patterns of H3K9me2 and H2Av at heterochromatic genes

The presence of essential genes in heterochromatin raises questions about how they are expressed, despite being embedded in a supposedly 'repressive' part of the genome. Some features of their expression are shared with euchromatic genes, most fundamentally the reliance on Pol II. Due to the inclusion of predicted genes in the annotations and the complex structure of genes in the heterochromatin, we selected a set of genes based on polymerase enrichment to generate an average profile of TSS. The TSS and TTS patterns for this subset (Fig. 7A,B) are surprisingly similar to those observed at euchromatic genes (Fig. 1C). Upstream and downstream peaks of H2Av and H3K4me2 and me3 surround heterochromatic TSSs bound by Pol II, and average patterns in the 3' region of these genes do not differ significantly from the average TTS profile for euchromatic genes.

Although these general similarities exist between the modification patterns of heterochromatic and euchromatic genes, transcription in the heterochromatin involves unique factors, such as Hp1 [28,34]. The substantial variation in size, structure, transcriptional status, and local gene density within this group of genes may explain the range of modifications and H2Av patterns observed in our data. H3K9me2 spans most loci not bound by Pol II, regardless of size. However, when polymerase is present, small genes often lack H3K9me2 signal. Large genes, common to the heterochromatin, show (persistent?) enrichment for K9 methylation in 3' regions of genes with Pol II peaks in 5' regions. The pattern of polymerase binding and H3K9me2 enrichment suggest a possible exclusivity or displacement in regions of high Pol II density, with H3K9me2 generally increasing where Pol II signal declines. For some heterochromatic loci, the distribution of H2Av at the 5' end appears shifted upstream and extends further than is typically found in the euchromatin (Fig. 7C). This upstream enrichment overlaps with H3K9me2, and the two marks show much higher correlations in 5' regions of genes bound by Pol II (Fig. 7D). Given that these data represent the early stages of its formation, these patterns may change significantly with the further establishment of the epigenetic components of heterochromatin.

Discussion

Alterations in nucleosome stability at the level of histone-histone and histone-DNA interactions are thought regulate gene expression through modulation of higher order chromatin conformation. One mechanism of transforming chromatin fiber dynamics is the replacement of canonical histones with variants. Biochemical analysis of the effects of H2A.Z on nucleosome stability suggests that nucleosomes containing both H2A.Z and H3.3 are less stable than those containing H3.3 and the canonical Histone, H2A [47,48]. The observation that, unlike H3.3, H2Av is peaked at TSSs and not distributed across the entire length of active genes implies that nucleosome stability across genes may be variable [47,48]. Localization of H2Av nucleosomes around the TSS, where H3.3 is also enriched, may facilitate access of transcriptional activators progression of the transcriptional machinery [41].

The impact of H2A.Z incorporation on chromatin dynamics also has intriguing implications for short-term gene silencing. *In vitro* studies have shown that H2A.Z facilitates intranucleosomal interactions and formation of nucleosomal arrays [49]. However, the inhibitory effect the variant has on internucleosomal interactions suggests that H2Av-containing fibers could be, in the absence of other chromatin factors, more amenable to unwinding [49]. H2Av deposition across silent loci may therefore make it easier to both "close" and "reopen" genes. Changes at the nucleosome surface upon substitution of H2A with H2A.Z promote Hp1-alpha-mediated chromatin compaction [50]. Perhaps one of the five HP1 paralogs in Drosophila interacts with euchromatic H2Av plateau regions to induce chromatin fiber folding and gene silencing [28].

H2Av is absent from the wide domains of H3K4- and K27me3 in the early embryo, which often span 10s of Kb. While the variant is present at active and inactive genes in other domains, H2Av shows no enrichment at developmental genes, regardless of their transcriptional status. This lack of the pervasive substitution of H2Av for H2A at these genes suggests that, beyond regulation by PcG and TrxG proteins, unique mechanisms may distinguish homeotic gene expression from that of other euchromatic genes. Whether the absence of this H2Av results from exclusion of the Swr1/SRCAP remodelers responsible for H2A.Z deposition or differences in the dynamics of H2A/H2B displacement during transcription remains an open question [51-54]. Many of these genes, which encode transcription factors, are transcribed at low levels and are relatively isolated. In this case, there may be fewer constraints on nucleosome packing and H2A/H2B eviction levels may be quite low.

The patterns of broad H3K4me3 and H3K27me3 enrichment we see in Drosophila embryos are in agreement with previous mapping of PcG proteins and H3K27 methylation in Drosophila S2 cells [55]. While these domains appear similar to bivalent domains observed in undifferentiated mammalian cells, our maps are a composite of different cells in the embryo and may not represent true bivalent domains [4-8]. However, these results suggest that the regulation of transcription factors and other determinants of cell fate and pluripotency may be maintained similarly in Drosophila embryos and mammalian ES cells.

While heterochromatic Gypsy 8 elements are marked with bivalent domain modifications, the functional significance of these domains is likely to be different from those in the euchromatin. Su(Hw) binding of Gypsy elements has been implicated in regulation of higher-order chromatin structure through its insulator function [46,56]. These regions may be involved in the organization of heterochromatin and its localization within the nucleus, recruitment of genes to transcriptionally favorable domains, or perhaps they serve as standard anti-silencing insulators.

Stage 5 embryos represent the onset of heterochromatin establishment. At this time H3K9me2 presence is primarily limited to intergenic regions of the pericentric heterochromatin and inactive heterochromatic genes. The patterns of K4 methylation and H2Av surrounding active heterochromatic genes roughly parallel those in seen in the euchromatin. Studies in mammalian cell culture report relatively little H2A.Z in the heterochromatin [57]. Further, H2A.Z nucleosomes from the heterochromatin show lower levels of H3K9 methylation and increased K4 methylation compared to average heterochromatic levels [57]. This is consistent with the patterns of enrichment we observe surrounding genes. In addition to peaks of H2Av downstream of the TSS, we see a notable enrichment of the variant upstream of many heterochromatic genes bound by Pol II. Incorporation of H2Av 5' of genes suggest that, in these regions, the variant may play an anti-silencing role analogous to that seen in H2A.Z in *S. cerevisiae* [16].

Given our results, the genetic role of *H2Av* as an upstream component of heterochromatin establishment remains intriguing and unresolved [13]. Does heterochromatin establishment require transcription and transcription-associated changes in chromatin which depend on H2Av? Alternatively, does the absence of H2Av lead to higher levels of transcription, which antagonize establishment? Or, finally, is this yet another example of the pleiotropic effects of *H2Av* mutations?

Figure 1

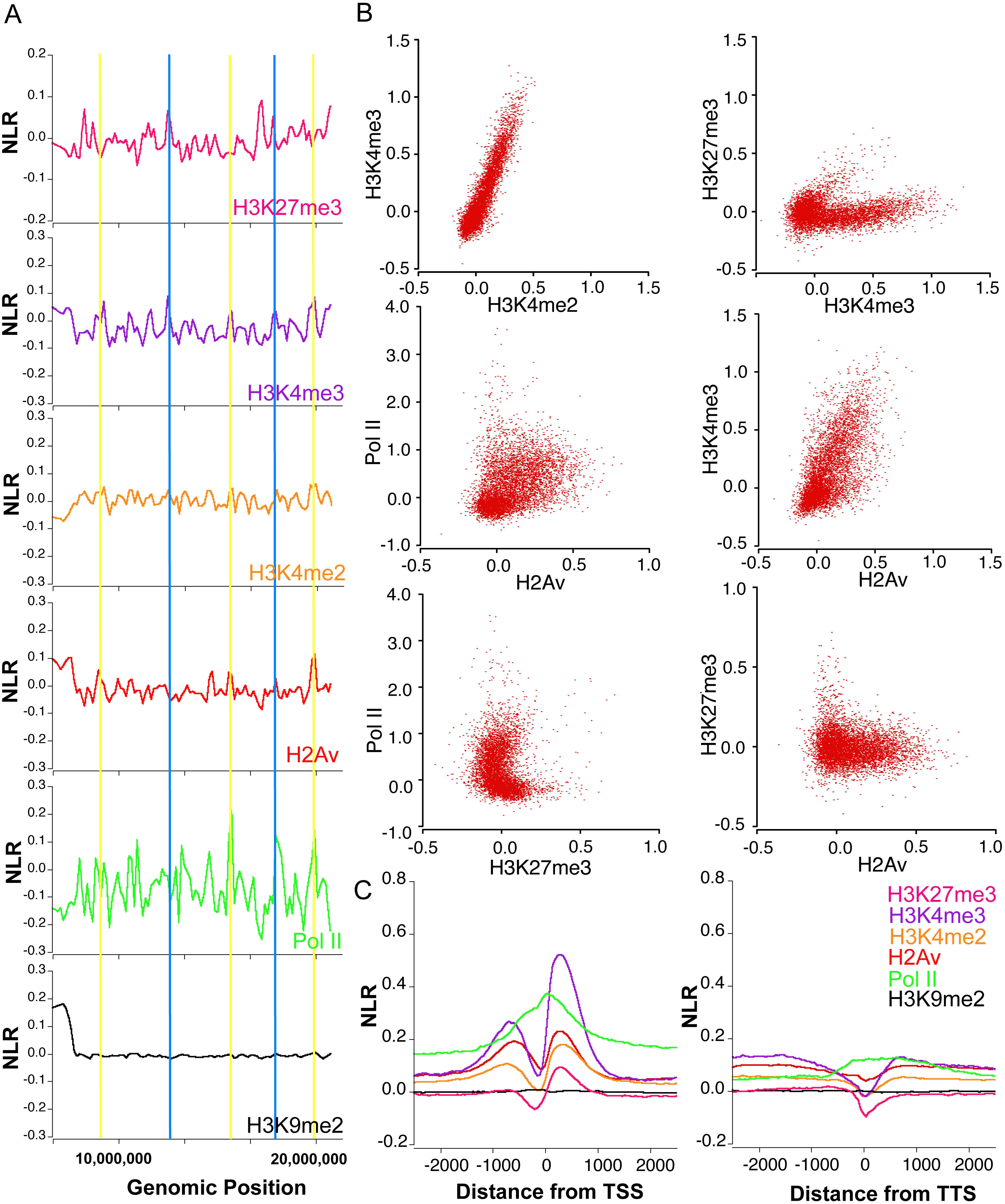


Figure 1. A) Smoothed 5kb window averages (minimum 25 probes) across 2R. Blue lines indicate examples of H3K4me3 and H3K27me3 enriched domains. Yellow lines indicate examples of regions enriched for H2Av and H3K4me2 and me3. B) Average NLR across individual genes (+500bp up and downstream) for target modifications, H2Av, and phosphorylated Pol II [43] are plotted against each other. C) Average NLR of modifications and H2Av surrounding the TSS and TTS for all non-overlapping genes.

Figure 2

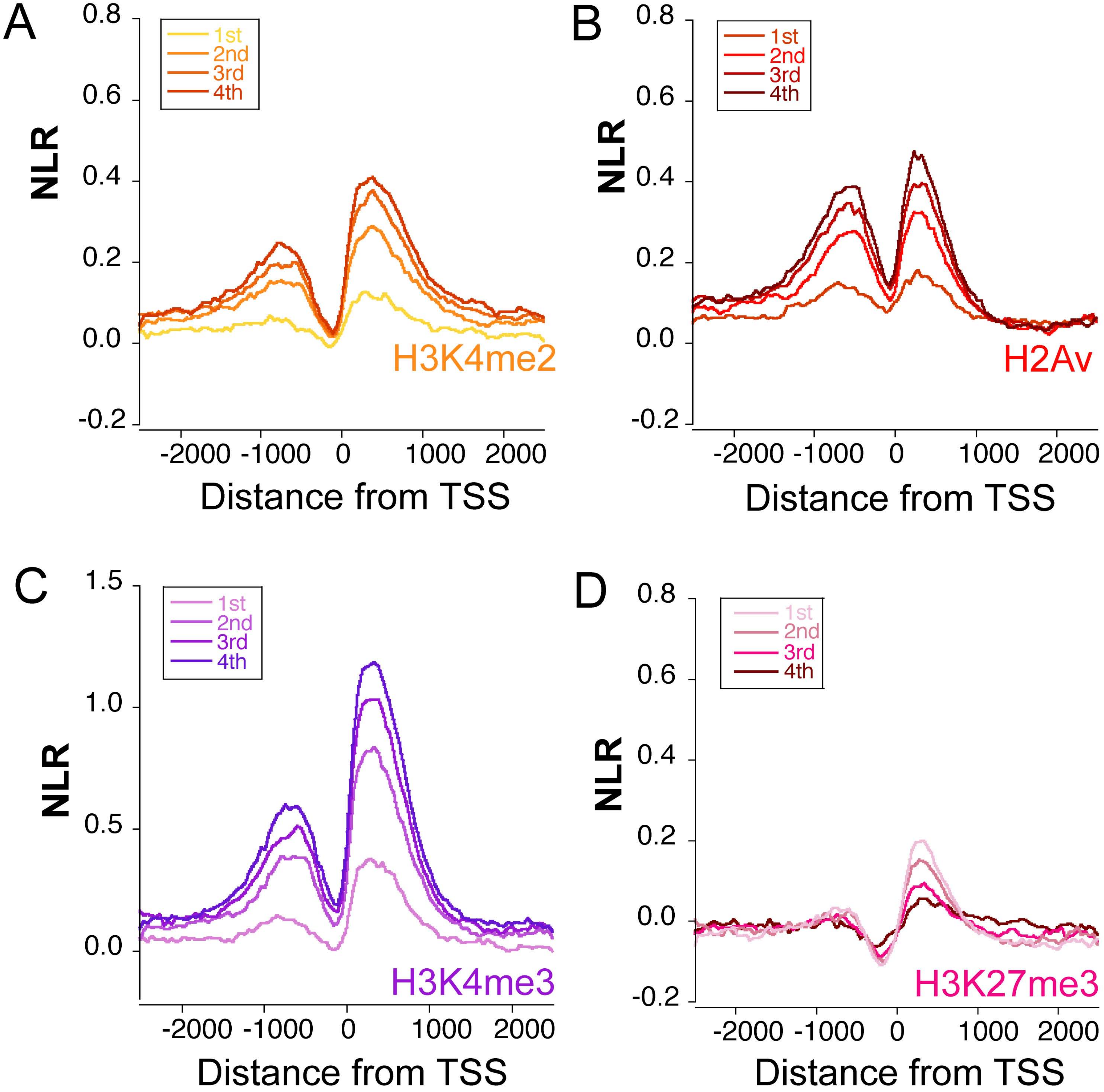


Figure 2. Modifications and H2Av enrichments are associated with gene expression. Genes were separated into quartiles based on expression levels at stage 5 [38]. The average NLR surrounding TSS for each quartile is plotted for A) H3K4me2, B) H2Av, C) H3K4me3, and D) H3K27me3.

Figure 3

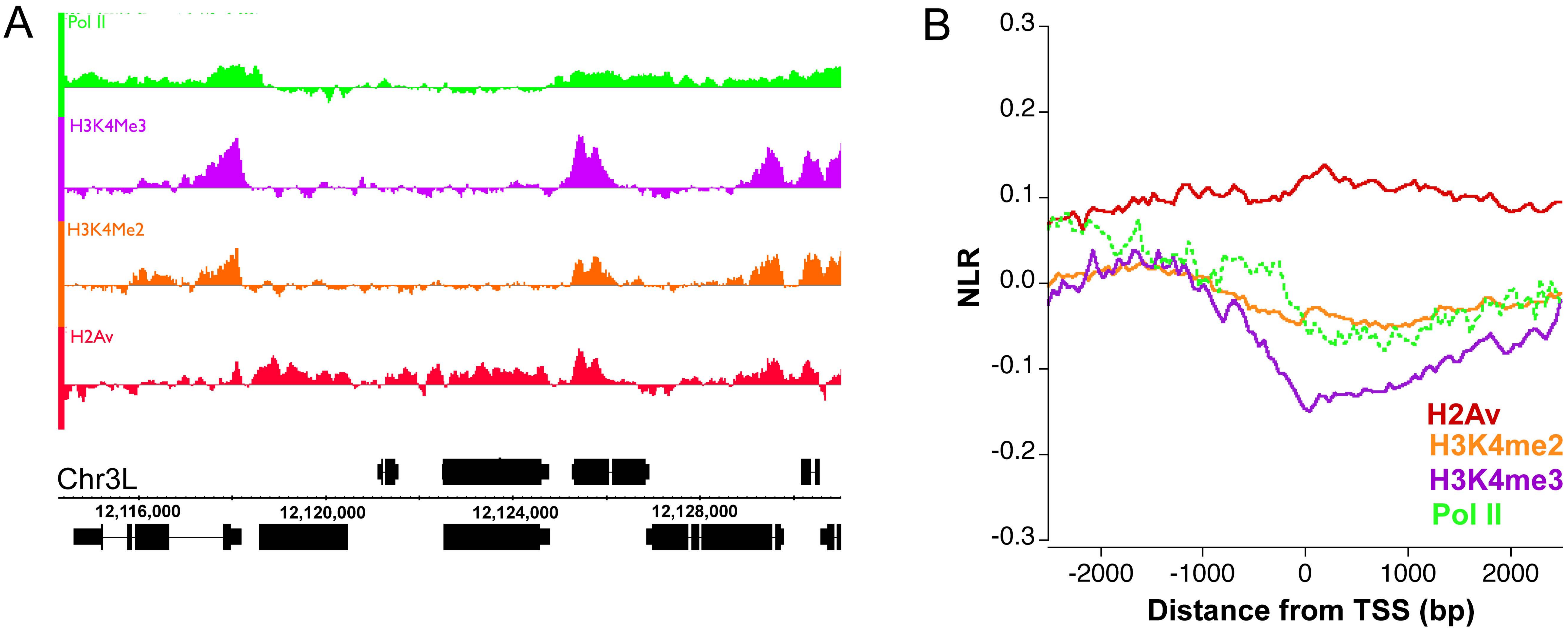


Figure 3. A) H2Av "plateau" enrichment in a region with other actively transcribed genes. B) Average NLR for H2K4me2, H3K4me3, H2Av, and Pol II surrounding TSSs of a set of 800 genes enriched for H2Av and lacking H3K4me3. C) Distribution of expression levels for all genes (3,499 non-overlapping genes included in study) and H2Av "plateau" genes [38]. D) Distribution of average Pol II across transcribed regions for all non-overlapping euchromatic genes and H2Av "plateau" genes.

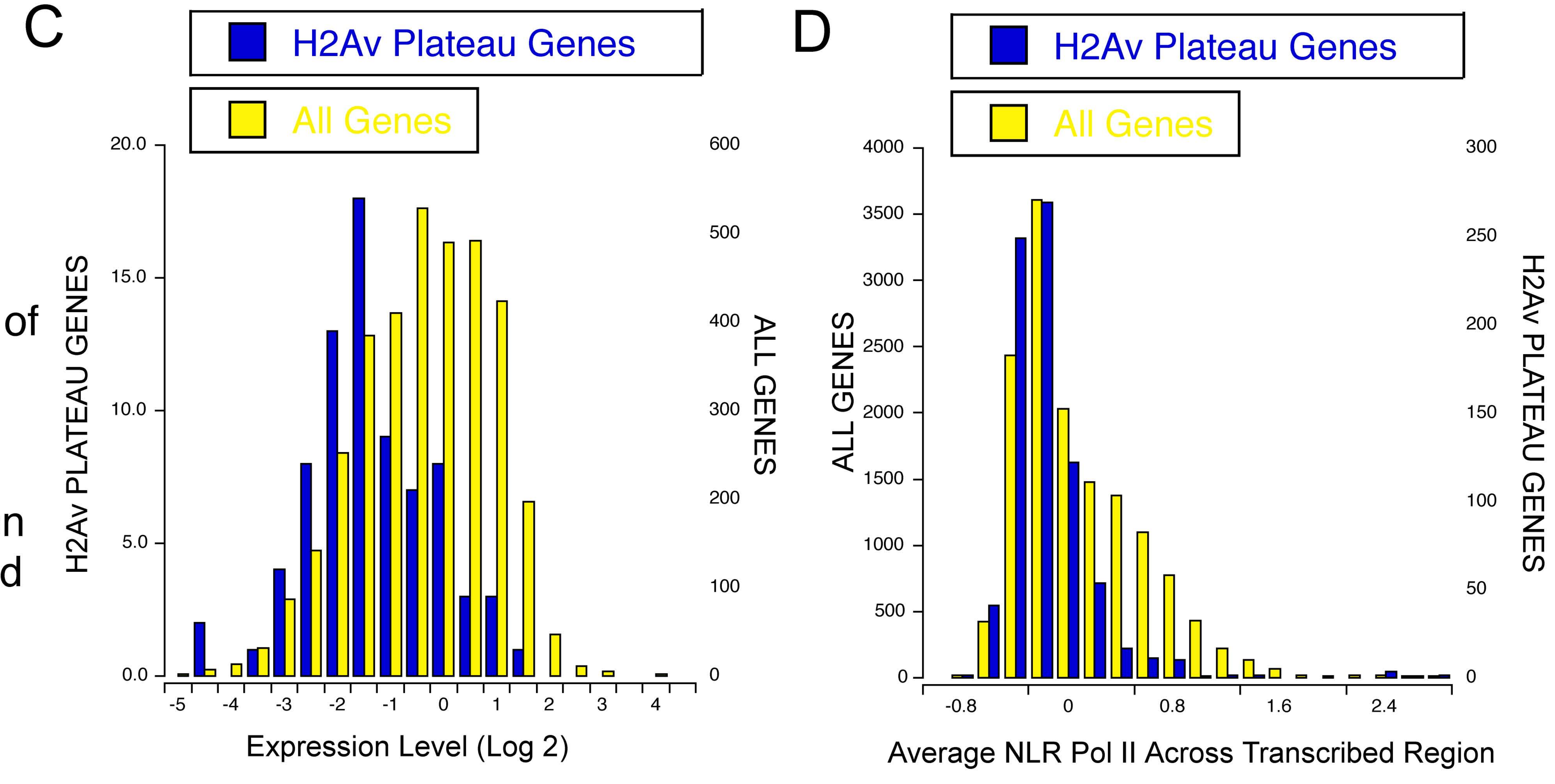


Figure 4

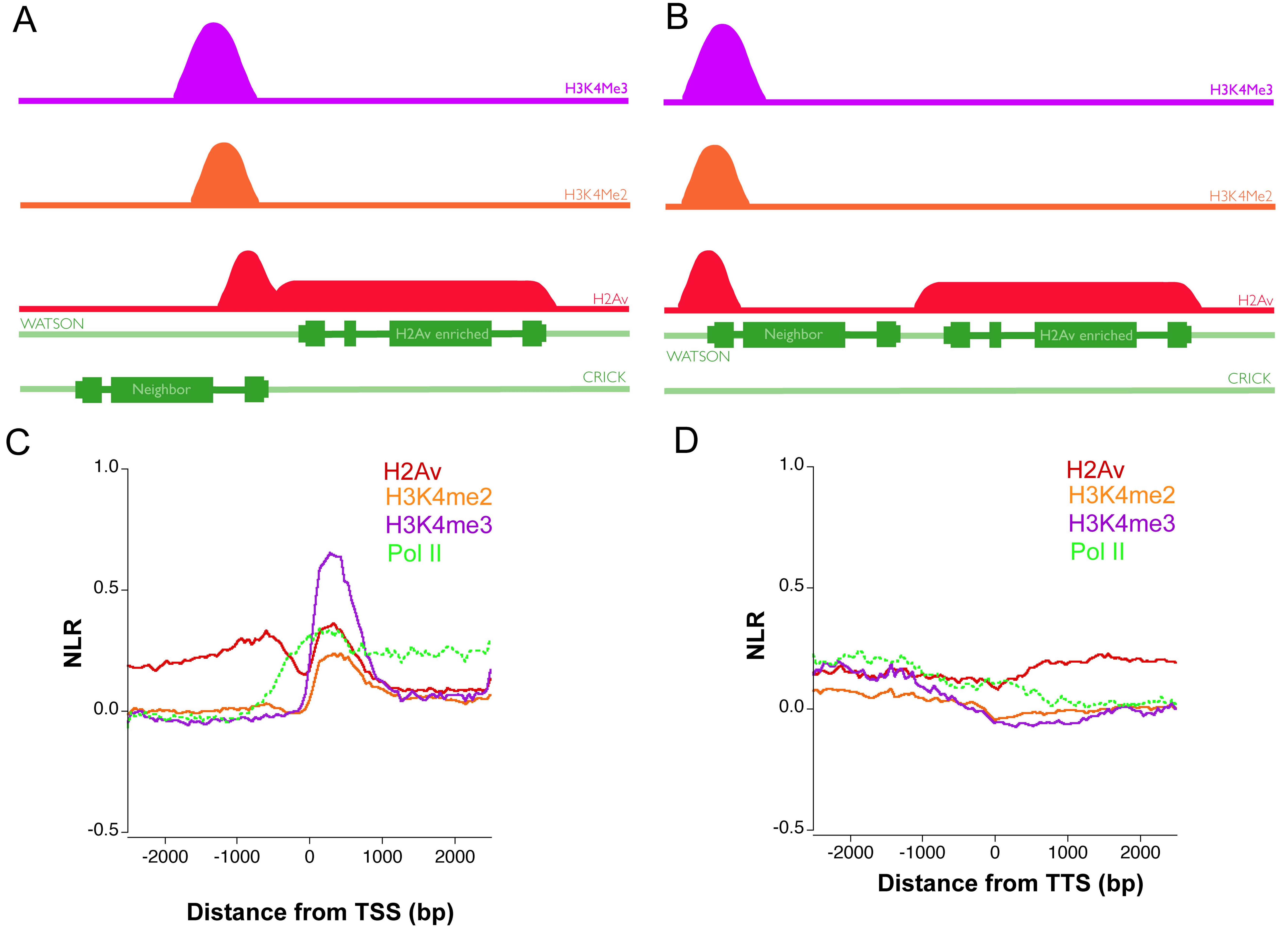


Figure 4. Genes upstream of H2Av plateau loci can be oriented with TSS (A) upstream or TTS (B) upstream. C) Average TSS profile for H3K4me2, H3K4me3, H2Av and Pol II for genes with TSS upstream of identified H2Av plateau genes. D) Average TTS profile for H3K4me2, H3K4me3, H2Av and Pol II for genes with TTS upstream of identified H2Av plateau genes

Figure 5

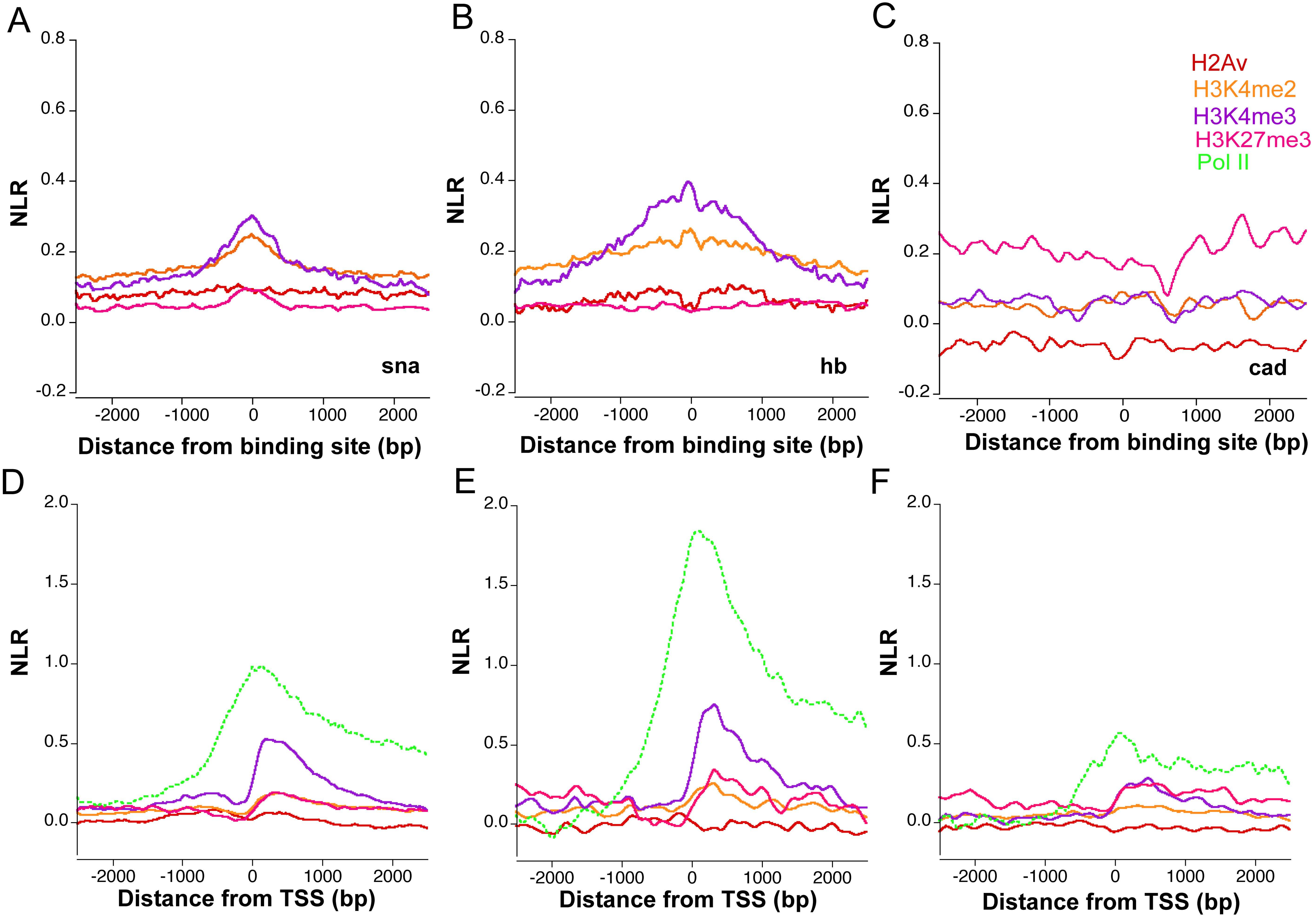


Figure 5. Averaged H3K4me2 and me3, H3K27me3, and H2Av profiles surrounding (A) sna(2), (B) hb(1), and (C) cad binding sites identified by the BDNTP [43]. Average intensities for H3K4me2 and me3, H3K27me3, H2Av and Pol II surrounding the TSS of genes closest to peaks of Giant binding (D), those bound (E), and not bound by Pol II and/or TFIIB (F) in stage 5 embryos (as identified by BDNTP [43]).

Figure 6

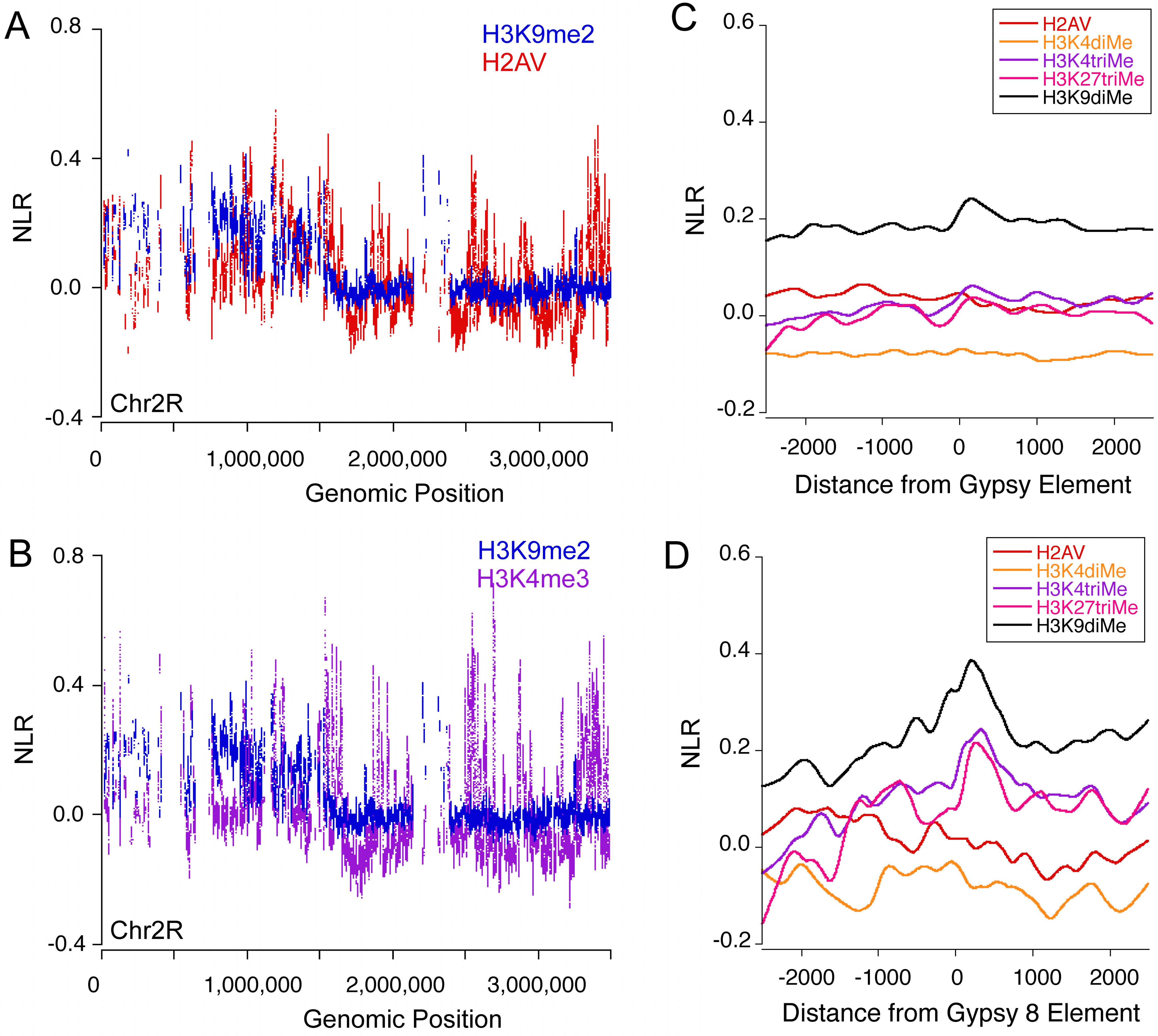


Figure 6. 5kb sliding window average NLR across the heterochromatin-euchromatin boundary at the base of chr2R. Averages were taken for all windows with > 25 probes and values were assigned to the mean genomic position of all probes in the window. H3K9me2 is ploted against (A) H2Av and (B) H3K4me3. Average NLR profile surrounding (C) Gypsy elements in the heterochromatin and (D) Gypsy8 elements in the heterochromatin

Figure 7

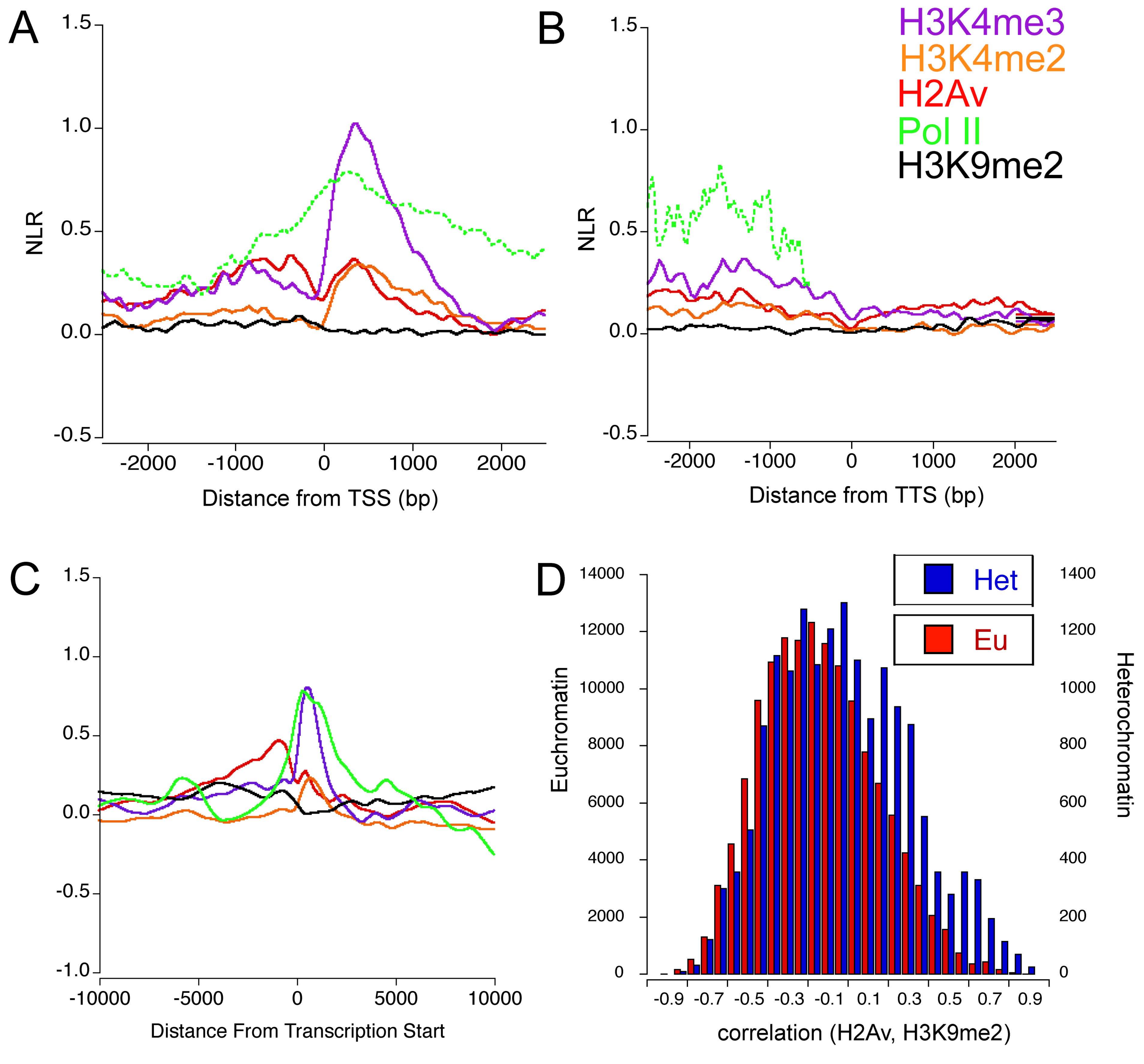


Figure 7. The average NLR for ChIP targets surrounding (A) TSS and (B) TTS of a set of heterochromatic genes bound by Pol II (genes in upper quartile of average Pol II density). C) Average NLR for ChIP targets surrounding heterochromatic genes enriched for upstream H2Av (Note change in X axis scale). Euchromatic (red) and heterochromatic (blue) distributions of windowed correlations of H3K9me2 and H2Av in the 5Kb surrounding TSSs of genes bound by Pol II.

Chapter 2

Conserved influence of gene orientation and intergenic distance on chromatin patterns surrounding promoters

Background

The composition and organization of chromatin at promoter regions plays essential roles in gene expression and silencing. Nucleosomes, the primary level of chromatin organization, must be excluded from promoter regions, either constitutively or transiently, for transcription factors to bind their target sequences and to allow the RNA Polymerase II (Pol II) machinery to initiate transcription [58-60]. Histone modifications and variants can either inhibit or facilitate this process through interactions with DNA and chromatin-associated proteins.

Correlated expression has been observed for small and large gene clusters in diverse eukaryotes [61]. Genes are often organized into large expression "neighborhoods," on the scale of tens of genes [61-66]. Although groupings of functionally related, coregulated genes exist, most clusters do not show enrichment for specific functional categories. At a more local level, coexpression of adjacent gene pairs has also emerged as a commonality across species [61,64,67-72].

The mechanisms underlying coordinated gene expression are not well understood. One hypothesis is that shared cis-regulatory elements may link transcription of adjacent genes [71,73]. Alternatively, expression of clustered genes may create an open chromatin environment, locally increasing Pol II binding and activity [62,68,74,75]. The key distinction is that the chromatin-based model predicts a synergistic effect of promoter density and proximity on gene expression. This model also predicts that the distance between genes will show a direct relationship with active chromatin marks, and that this association will be stronger for divergently transcribed genes, where the promoters of a gene pair are juxtaposed. In contrast, a model where shared cis-regulatory elements drive coexpression predicts only that adjacent genes will tend to be expressed under the same conditions or in the same tissues, not that proximity of two coregulated promoters will influence gene activity or chromatin state.

For correlated gene pairs, both orientation and promoter density play roles in coexpression. In *S. cerevisiae*, divergently (<- ->) transcribed genes show higher levels of co-expression than pairs with tandem (-> ->) orientation [72]. However, after correction for inter-promoter distance, similar correlations are observed for both orientation classes, suggesting that distance between promoters influences coexpression of neighboring genes [72]. In support of this, genes with shorter intergenic regions are also more highly correlated [76]. This appears to be the case in *Arabidopsis* as well, where intergenic distance is negatively correlated with coexpression [66]. The scale of this local effect is relatively small. Correlated expression in *S. cerevisiae* decreases over 1-5kb, with the sharpest drop in the 1-2kb range [72]. In *C. elegans*, correlated expression across tissues, also higher for divergent gene pairs than for tandem genes, displays a steep drop over the 1kb range [77].

In humans, shared regulatory systems are thought to drive the co-expression of some adjacent genes [78,79]. Human genomes contain a distinct class of bidirectionally transcribed (<- ->), highly expressed genes, unique to the mammalian lineage [78,80-82]. These genes often lack TATA boxes, are CG rich, and are thought to have shared *cis*-regulatory elements

[69,78,80-83]. Close to half of all bidirectional pairs are coexpressed, and they exhibit higher levels of Pol II enrichment at their promoters than other genes [83]. In liver-derived cultured human cells, 47% of genes showing enrichment for H3 acetylation (H3ac), an active chromatin mark, are bidirectional [84]. Although there is evidence that some of these genes are uniquely and coordinately regulated at the sequence level, the mechanisms that regulate their coexpression remain unresolved [78,82]. Similarly close (<1kb apart), divergently transcribed gene pairs in Drosophila often lack TATA boxes and show higher levels of correlated expression than similarly spaced tandem (->->) or convergent (<-->) gene pairs [85].

In spite of long-standing models suggesting chromatin involvement in coexpression of gene pairs, the influence of neighboring genes on promoter chromatin states has not been examined. Epigenomic data provide an opportunity to investigate the relationship between genomic organization, gene expression, and enrichment of chromatin marks and histone variants [86,87]. Stereotypical chromatin signatures at gene promoters have emerged from such analyses. Most *S. cerevisiae* promoters have a nucleosome depleted region (NDR) of varying length flanked by two nucleosomes containing the H2A variant, H2A.Z, referred to as '+1' and '-1' nucleosomes (relative to TSSs) [17,38,88-94]. These regions are surrounded by a striking upstream and downstream phasing of nucleosome occupancy [90,93]. ChIP-sequencing in human cells has revealed a wider upstream and downstream distribution of H2A.Z surrounding the TSS [20]. The Drosophila homolog, H2Av, shows a similarly broad enrichment, though phased H2Av-containing nucleosomes are reportedly limited to regions downstream of the TSS [95]. Average TSS patterns have established other chromatin signatures associated with gene expression across species, such as methylation of histone H3 at Lysine 4 (H3K4me). Tri-methylated H3K4 (H3K4me3) is found in peaks upstream and downstream of promoters, whose enrichments are positively correlated with expression levels [20,36,37,86,96]. In human liver cells, the upstream enrichment of H3ac, found surrounding the TSSs of active genes in averages from many species, is primarily associated with bidirectional gene pairs [84]. However, two peaks of H3ac are evident in TSS averages for highly active genes that are not bidirectionally transcribed, leaving the issue of neighboring promoter influence on upstream chromatin states unresolved [84].

We have investigated the relationship between genome organization, gene activity and chromatin structure by examining the patterns of epigenetic marks at promoter regions in Drosophila. We generated Native ChIP (N-ChIP)-array data from stage 5 Drosophila embryos for H3K4me2, H3K4me3, H3K9me2, H2Av, and H3K27me3 (Fig. 1A). N-ChIP is performed using micrococcal nuclease (MNase) digested, unfixed chromatin (mononucleosomal fragments) as input for the ChIP reaction. This provides fine-scale mapping of histone modifications and variants. Published data on the distribution of the elongating (ser5-phospho) form of Pol II from the same developmental time point were incorporated into our analysis [43].

Results

Gene orientation impacts enrichment levels and patterns of active chromatin marks surrounding promoters in Drosophila embryos

In our initial examination of average chromatin patterns surrounding TSSs for all non-overlapping genes, we observed elevated levels of H2Av and the active marks H3K4me2 and me3 immediately upstream and downstream of the TSSs (Fig. 1B). These peaks flanking the promoter regions were consistent with previous observations in diverse eukaryotes, and are currently considered to be ‘signatures’ of active genes. However, browser-based inspection of a subset of genes identified significant differences in comparison to the genome-wide averages (Fig. 1A). This led us to examine the potential impact of gene orientation and distance to the nearest upstream gene on histone modifications and variants surrounding the TSS.

We separated non-overlapping genes into classes based on promoter orientation; individual genes were classified as divergent or tandem, depending on the orientation of the nearest upstream neighbor. A skew toward the 0-1kb intergenic distance class is evident in distributions for both orientations, but is most striking for divergent genes (Fig. 2). Genes in the ≤1kb class have subtly different distributions ($P < 10^{-6}$, two-sided Wilcoxon rank sum test (WRST)), with tandemly oriented genes having shorter mean length (3,106 bp for tandem, 3,499 bp for divergent) and greater standard deviation (6,122 bp for tandem, 5,526 bp for divergent) (Fig. 3). For both tandem and divergent genes in the ≤1kb class, average gene length increases with intergenic distance. While significant, the scale of these differences is relatively small, and it is likely that the biological impact with regard to chromatin patterns at the promoter is minor. The distributions of gene lengths for tandem and divergent genes whose promoters are >1kb from upstream neighbors are not significantly different ($P = 0.99$, two-sided WRST) (Fig. 3).

Genes grouped by orientation showed distinct profiles in comparison to each other and to the genome-wide averages (Fig. 1B). Surprisingly, divergent genes display higher enrichments for the ‘active marks’ H2Av, H3K4me2 and H3K4me3, and the upstream peaks are more pronounced in comparison to genome-wide average profiles (Fig. 1B). Tandem genes show significantly lower enrichments relative to both genome wide averages and divergent genes (Fig. 1B). Separation of genes based on orientation also eliminates upstream H2Av, H3K4me2 and me3 peaks for tandem genes, resulting in TSS averages that better reflect chromatin patterns at individual genes.

An analysis of ser5-phospho-Pol II ChIP-chip and expression data from early embryos demonstrates that gene expression levels are also impacted by gene orientation and distance, consistent with the patterns for active chromatin marks. Both tandem and divergent genes separated by ≤1kb show higher average Pol II levels than those with longer (>1kb) intergenic distances ($P < 10^{-6}$ for both, one-sided WRST), and close divergent genes have higher levels than observed at close tandem genes ($P < 10^{-6}$, one-sided WRST) (Fig. 4E, Fig. 5). Pol II densities for divergent genes >1kb from neighbors are also higher than >1kb tandem genes ($P = 0.044$, one-sided WRST). Although close genes may contribute to enrichment levels of neighboring genes in the analysis of ChIP data generated from crosslinked chromatin, the observed enrichment for Pol II at divergent genes is consistent with the observed higher enrichments for active chromatin marks from the N-ChIP experiments. Finally, analysis of tiling array expression data for 2-4hr embryos provided by modENCODE (http://www.modencode.org/) confirms that close divergent genes (<1kb) are expressed at higher levels than both tandem genes and divergent genes >1kb apart (Fig. 6).

Although tandem genes on average show lower expression and lower enrichment for active chromatin marks, many individual tandem genes are highly active. If the chromatin pattern disparities between tandem and divergent genes were the result of differences in expression levels, then we would expect chromatin patterns for highly expressed genes from both categories to be similar. To address this question, we randomly sampled 200 genes from the upper expression quartile for close tandem genes. For each gene, we then selected (without replacement) the best expression level match from the divergent ≤1kb class, producing groups of equivalent size with highly similar expression levels (Fig. 7A). The resulting averages surrounding TSSs show that, even at high expression levels, tandem genes do not have upstream peaks for active marks (Fig. 7B).

Divergent genes show a distance-dependent decrease in active chromatin marks and gene expression in Drosophila

To confirm that upstream peaks of H2Av and H3K4 methylation at divergent genes were attributable to upstream genes, we binned genes based on distance to their closest upstream neighbors (in 200bp increments). Indeed, as predicted by our initial results, upstream peaks shift proportionally with increasing distance to the neighboring gene, whereas the downstream peaks do not change position (initiating at the '+1' nucleosome in all classifications) (Fig. 4A,C). Plots surrounding individual genes ranked by distance to the upstream neighbor clearly confirm that the upstream enrichment of active marks shifts with distance to the neighboring gene (Fig. 4B,D). We conclude that upstream peaks, for the marks examined here, represent chromatin enrichments initiating at the +1 nucleosome of the upstream promoter. Reanalysis of published H2Av ChIP-seq data from 0-12hr Drosophila embryos supports these results, and reveals upstream and downstream phasing for close divergent genes (Fig. 8A,B) [95]. The distance dependence of upstream peaks, combined with their absence from tandem gene promoters (Fig. 1B), confirms that the upstream peaks observed in genome-wide averages reflect the chromatin organization of close, active divergent gene pairs in Drosophila embryos.

For the approximately half of all divergent genes that are within 1 kb of their 5' neighbor, high average enrichments of H2Av, Pol II and H3K4 methylation decline strongly with distance (Fig. 4A-D). Although tandem genes exhibit much less enrichment for these marks surrounding their TSSs, they also show a consistent decline with distance. Additionally, average Pol II densities (Fig. 5A, B) and RNA transcript levels (Fig. 6B,C) drop as the distance to upstream neighbors increases, particularly for divergent genes. This gradual decrease demonstrates that close promoters are more likely to be enriched for active marks.

If intergenic distance influences gene activity, it should impact both members of a gene pair, particularly for divergent pairs. We find that, along with overall levels, the correlation of Pol II density for gene pairs drops with increasing intergenic distance from ~0.3 to 0 for both divergent and tandem gene pairs across the same range (0-1kb), suggesting a link between gene spacing and coexpression (Supplementary Fig. 9A). Overall, the striking drop in 'active' chromatin signatures, expression levels, and correlation of average Pol II density across such a narrow range (1kb) suggest that the distance between promoters influences their activity and chromatin state. We propose that synergism occurs between adjacent promoters that is sensitive to relatively short, nucleosome-length differences in distance.

Roughly 50% of all non-overlapping Drosophila genes are separated by >1kb (Fig. 2) and show no or low average enrichment for H2Av and H3K4me3 (Fig. 4A-D), or Pol II and H3K4me2 (data not shown). This large group of genes does not contribute significantly to average patterns of active chromatin marks surrounding TSSs in embryos, except to reduce peak heights. Although average values are low for these genes, many individual genes are highly expressed. To determine if the chromatin environments of highly expressed genes are different when isolated versus clustered, we selected tandem or divergent genes (>1kb from neighbors) with expression levels similar to 200 highly expressed close tandem genes (Fig. 7A). Averages surrounding these genes are characterized by roughly equivalent Pol II enrichment, but, in addition to the absence of upstream peaks, their downstream enrichments of H3K4me3, H3K4me2, and H2Av are strikingly lower (Fig. 7B).

Synergistic activation of nearby genes could be advantageous, *e.g.* for genes expressed in many cell types or co-expressed in the same cell type. However, independence from the impact of neighboring gene chromatin states may be necessary for genes with more restricted expression patterns during development. We next asked if the large differences in expression levels based on gene orientation and distance are associated with distinct gene functions. Analysis of GO slim categories (Supplementary Table 1) reveals that genes in the ≤1kb divergent class are overrepresented for many general housekeeping categories, including ATP binding and general Pol II transcription factors, which are likely to be expressed in many cell types. In contrast, tandem and divergent genes separated by >1kb from neighbors are significantly enriched for transcription factors involved in development. We note that these genes are often located in regions characterized by moderate, broad enrichment of the 'active' H3K4me3 and 'silent' H3K27me3 marks (Supplementary Data). These modifications are the hallmark of 'bivalent domains', a specialized chromatin state associated with pluripotency in other organisms, recently reported absent in Drosophila embryos [4-7,20,35,97]. Due to the mixed population of cells in whole embryos, additional work is necessary to determine if the observed enrichments in our data represent true bivalent domains. Independent of the unresolved issue surrounding bivalent domains in flies, we find that gene orientation and distance classes show interesting differences in enrichment for functional categories. Our results suggest that developmental regulators are relatively isolated and that genes potentially associated with proliferation and growth are often in close divergent pairs.

Gene orientation and distance influence chromatin patterns and nucleosome occupancy at *S. cerevisiae* promoters

We next explored whether these strong associations between promoter orientation, distance and hallmarks of gene activity were conserved or unique to Drosophila embryos. Specifically, we reanalyzed published ChIP-array and ChIP-seq data from *S. cerevisiae* [17,37,93,98], and ChIP-seq data from human CD4+ cells [20] by classifying genes based on the orientation of and distance to their closest upstream neighbor. The large differences in genome size, organization, and proportion of expressed genes in these species and cell types allowed us to address the influence of local chromatin environments on average TSS patterns in very different contexts.

In *S. cerevisiae*, genes classified into divergent and tandem subsets (Fig. 2) showed clear differences in H3K4me3 and H2A.Z promoter patterns, consistent in many ways with those observed in Drosophila. Levels of H3K4me3 decrease with distance for genes within 1kb of their upstream neighbors (Fig. 10A), and H2A.Z is specifically enriched at close, divergent genes (Fig. 10C,D). As seen for divergent gene pairs in Drosophila, upstream peaks for both marks shift proportionally with the distance to upstream neighbors (Fig. 10A-D). We note that the H2A.Z pattern for divergent genes ≤200bp from the upstream neighbor most closely resembles the stereotypical patterns derived from genome-wide averages. However, for this class, the '–1' peak is actually downstream of the TSS of the neighboring upstream gene (Fig. 10, Fig. 11, Fig. 12). This is consistent with our conclusion from the Drosophila data that peaks 5' to divergent gene TSSs are primarily associated with the upstream promoter. Heat maps of enrichments surrounding divergent genes show that the distance between the two H2A.Z nucleosomes increases with intergenic distance, resulting in larger NDRs (Fig. 10D). As this distance exceeds

~400bp, roughly equivalent to the length of a nucleosome (146bp) and two NDRs (131bp each), intergenic H2A.Z enrichment is evident between divergent promoters [99]. This is also reflected in TSS averages where additional H2A.Z peaks are evident as the distance between adjacent promoters becomes large enough to accommodate additional nucleosomes (Fig. 10C, Fig. 11). For tandem genes, ChIP-seq data reveal that the average '–1' H2A.Z peak is smaller than the '+1' peak (Fig 10C,D, Supplementary Fig. 11). Contrary to what we observe for divergent genes, the position of the -1 peak at promoters of tandem genes does not shift with increasing distance to the upstream neighbor.

Both divergent and tandem genes separated by more than ~400bp from their neighbors continue to show evidence of upstream H2A.Z enrichment (Fig. 10A-D). This pattern may, in part, reflect the bidirectional transcription recently identified at individual promoters in both mammalian cells and *S. cerevisiae* [99-102]. Average RNA intensities from both the wild type S96 strain and nuclear exosome mutant cells (*rrp6*Δ*)* [99] enriched for cryptic unstable transcripts (CUTs) show peaks of antisense transcription initiating ~200-250bp upstream of aligned tandem gene TSSs, regardless of distance to the neighboring gene, consistent with antisense transcription initiating at upstream H2A.Z nucleosomes (Fig. 13B,C). Surprisingly, in average profiles, the evidence for intergenic transcription at divergent genes first appears for genes >400bp apart, the same class where intergenic H2A.Z signal appears. Divergent genes in the *rrp6*Δ strain display increased levels of RNA coming from both the sense and antisense strands, as distance to the upstream neighbor increases (Fig. 13C). These results suggest that, in addition to the unexpected prevalence of bidirectional transcription, the precision of transcriptional initiation may depend on promoter context. Initiation between genes, rather than at defined TSSs, appears to be more common as distance between genes increases.

The functional significance of H2A.Z nucleosomes flanking the NDR is not fully understood, but clearly gene orientation and distance influence the presence and positioning of such nucleosomes in promoter regions. Our analysis of published array data [93] also reveals marked variation in total nucleosome occupancy surrounding TSSs (Fig. 10E). In previous studies, cluster analysis of promoters based on patterns of H2A.Z enrichment or nucleosome occupancy identified groups of genes with similarly organized promoter regions. These classes were characterized by varying NDR widths and nucleosome (or H2AZ) occupancy patterns, and often by interesting functional associations, such as transcriptional plasticity and stress response [93,103]. We note that the patterns associated with many of these classes are also identified by our analysis (Fig. 10C-E, Fig. 14), suggesting that promoter context alone can explain much of the observed variation in NDR size and nucleosome positioning associated with these classes. Further, separation based on distance and orientation reveals that nucleosome phasing upstream of the TSS is primarily a property of divergent genes (Fig 10E). Thus, as in Drosophila [95] (Fig. 8), nucleosome phasing in *S. cerevisiae* is strongly associated with gene bodies, initiating at the TSS.

In agreement with previously observed low nucleosome densities in promoter and intergenic regions [90,92-94], nucleosome occupancy is generally lower between genes in our plots (Fig. 10E, Fig. 14, Fig. 15). Surprisingly, the nucleosome density in intergenic regions increases with the distance between genes, even within NDRs. This property may be encoded at the sequence level. Separation of published sequence-based analyses shows that both the average model score for intergenic regions and the predicted nucleosome occupancies increase with distance (Fig. 15A,B) [104]. However, *in vivo* and predicted occupancies do differ in some respects (152C). In particular, while tandem genes generally have lower average scores from the model and lower predicted nucleosome occupancies in intergenic regions, the *in vivo* averages are actually lower for divergent genes in most cases. Model scores and predicted occupancies in Drosophila follow similar trends, but with higher average values for divergent genes (Fig. 15D,E).

Although results from *S. cerevisiae* share many of the features seen in Drosophila embryos, most significantly the decline in levels of active chromatin marks as distance between genes increases, enrichment levels of H3K4me3 do not differ greatly between divergent and tandem genes. Consistent with this, expression data [105] do not indicate that divergent genes are more highly expressed in these cells, nor do we see strong evidence for an inverse correlation with intergenic distance (Fig. 13A). The compact organization and high transcriptional activity of the *S. cerevisiae* genome when growing at log phase may diminish the impact of the proposed synergism at the level of individual promoter pairs on expression.

Intergenic distance and orientation impact chromatin patterns at promoters of human CD4+ cells

The average inter-gene distance in the human genome is much larger than in *S. cerevisiae* or Drosophila (Fig. 2). Nevertheless, we observed that TSS averages in sequencing data from human CD4+ cells display many of the signatures present in those from more compact genomes, most notably the upstream peaks of H3K4 methylation [20]. Previous results regarding H3ac suggested that, as in *S. cerevisiae* and Drosophila, orientation may impact chromatin state in humans [84]. To determine if this pattern was predominantly associated with close divergent genes in humans, as in Drosophila and *S. cerevisiae,* we applied the same distance and orientation analysis to human ChIP-seq data from CD4+ cells [20] using the UCSC Known Genes collection.

Results for this human cell line are largely consistent with what we observe in other species. TSS averages for CD4+ cells show striking orientation and distance-dependent differences in levels and patterns of epigenetic marks. As in Drosophila and *S. cerevisiae,* strong upstream H3K4me3 peaks are prominent only for divergent genes ≤1kb apart (Fig. 16A). Furthermore, intensities surrounding individual divergent promoters show that upstream enrichment of H2A.Z and H3K4me3 shifts proportionally with distance to the neighboring gene (Fig. 16B,D, Fig. 17). Genes with tandem upstream neighbors have substantially lower levels of H3K4 methylation and H2A.Z (Fig. 16A-D). For the divergent genes, enrichment for these marks declines sharply with distance for intergenic distances less than 1kb, as seen in both Drosophila and *S. cerevisiae*. Results using the Ensembl gene set are more dramatic and show the same overall trends (Fig. 18). Pol II shows similar patterns (Fig. 19), and, though not as dramatic as the difference observed in Drosophila, CD4+ cell expression data [20] confirms that genes within 1kb of their neighbors are more highly expressed than those with greater intergenic distances (Fig. 16E) ($P < 10^{-6}$ for both divergent and tandem genes, one-sided WRST), consistent with the observed decrease in average patterns of epigenetic marks. For close pairs, at odds with the results from Pol II ChIP-seq (Fig. 19), we do not find that divergent genes are more highly expressed than tandem ($P < 0.36$, one-sided WRST). However, the available expression data from CD4+ cells covers a limited number of genes. More comprehensive expression analysis is required to determine if the differences in epigenetic marks for tandem and divergent genes ≤1kb from neighbors are reflective of differences in expression levels, as in Drosophila.

Two differences in TSS patterns between CD4+ cell ChIP-seq data and Drosophila embryo array data (and, to a lesser extent, that from *S. cerevisiae*) cannot be resolved by the current analysis. First, divergent genes >400bp apart show greater intergenic enrichment for H2A.Z and H3K4me3 in CD4+ cells (Fig. 16A-D). These regions may be particularly sensitive to MNase digestion, generating an overrepresentation in the ChIP-seq data (see Supplemental Methods), or the conditions used in these ChIP experiments may have retained higher levels of relatively unstable H3.3 or H2A.Z-containing nucleosomes [48,106]. Regions between human bidirectional genes are known to be GC rich, so technical bias during sequencing could also play a role. In the future, input controls are necessary to determine if this is the case. However, potential biological differences between transcription and inter-promoter chromatin states in Drosophila, *S. cerevisiae*, and human cells could also be responsible for such differences.

Second, we see persistent low levels of enrichment of active modifications upstream of TSSs for tandem genes in CD4+ cells (Fig. 16A-D). This is not the case for H3K4 methylation in our Drosophila embryo data (Fig. 4A,B) or the ChIP-array data from *S. cerevisiae* (Fig 10A,B). For tandem genes in CD4+ cells, upstream peaks do not show a shift with distance to the 3' end of the upstream neighbor. This 5' enrichment is also present at divergent human genes >1kb apart. Recent reports of short antisense transcripts at active promoters in mammalian cells provide a possible explanation for this observation [100,101]. Like genes with bidirectional promoters, active tandem genes and distant divergent genes in the human genome are likely to be characterized by some level of divergent transcription and associated chromatin changes. However, we do not see evidence of this in *S. cerevisiae*, which also shows pervasive antisense transcription [99,102]. Whether these distinctions between Drosophila, *S. cerevisiae*, and CD4+ cells are the result of differences in methodologies, or are biologically relevant, will require further work.

Discussion

These analyses demonstrate that classification of genes by distance and orientation relative to their upstream neighbors can expose distinct chromatin patterns. When averaging across all genes or looking at a single locus, upstream signals from neighboring genes have often been filtered from the analysis. In fact, such data represents biologically, functionally relevant information that can contribute to models of chromatin and nucleosome organization surrounding TSSs. Through separation of data based on gene orientation and distance, we show that upstream peaks in TSS averages of active marks are primarily associated with neighboring divergently transcribed genes. In Drosophila and CD4+ cells, close, divergent genes are characterized by high relative expression levels and the greatest enrichment of active marks and H2A.Z(v). H2A.Z in *S. cerevisiae* shows a similar relationship to promoter context, with close, divergent genes showing highest average enrichments. We conclude that local interactions between neighboring genes have significant influence over enrichment of active marks, particularly within the 1kb range.

Aggregation of genomic data allows the detection of common features at higher resolution. However, without separation of genes into classes based on orientation and distance, our conception of what a "typical" TSS looks like may be heavily biased toward dominant classes. Additionally, for a given genome, differences in the number of genes in specific distance and orientation classes can also produce strong biases in the average profiles (Fig. 2). For Drosophila, where more than half the genes are ≤1kb apart, average patterns are strongly skewed towards those for close divergent genes. Removing such dominant classes reveals masked patterns, and produces averages that better reflect the chromatin states of individual promoters.

The observation that upstream genes influence average chromatin profiles also has biological relevance to our understanding of the interplay between genome organization and epigenetic regulation of gene expression. The discovery of expression “neighborhoods” inspired models of open chromatin spanning several kbs and clusters of mostly active genes [61-66]. Here we demonstrate a more local effect, where distance to the upstream gene is inversely associated with enrichment of active chromatin marks. In many species, correlation of expression of gene pairs also decreases with increasing intergenic distance [66,72,76,77]. Though some human divergent gene pairs share *cis*-regulatory elements [78,79,82], studies in *S. cerevisiae* have not found evidence to suggest that common regulatory elements are responsible for coexpression of neighboring genes [107]. It has been proposed instead that physical proximity leads to coexpression of gene pairs through chromatin remodeling [74].

The graded chromatin differences we observe based on distance suggest such a link between the activity of neighboring genes at the epigenetic level, especially for distances within the 1kb range. To account for our observations, we propose that primary signals for gene activation are amplified through distance-dependent synergistic connections between Pol II, epigenetic modifications of active chromatin, and the factors they recruit. Primary signals include transcription factor binding and Pol II recruitment, which may be more efficient when promoters are clustered, particularly when they are directly adjacent. Distance-dependent amplification of these signals could be mediated by enhanced recruitment of complexes involved in chromatin remodeling, promote enrichment for active marks, or histone replacement. A model of promoter synergism is supported by expression data, particularly in Drosophila embryos, where transcript levels for divergent genes show a strong association with short intergenic distances. The same mechanism may act in large clusters of active genes [63], or even in three dimensional associations of distant genes [108]. Distance-dependent synergism mediated through chromatin does not preclude contributions from cis-regulatory elements. In addition to sequence-based recruitment of primary signals, some divergent genes may share a distinct set of transcriptional activators that confer higher activity. However, a model based solely on *cis*-regulatory elements directing coexpression of neighboring genes does not predict *higher* levels of active marks for close genes, only that the expression levels of two neighboring genes should be correlated.

In Drosophila embryos and human CD4+ cells, we see a significant difference in levels of active marks for divergent and tandem genes. Divergent genes may be transcriptionally favored in some contexts, particularly in genomes with large amounts of non-coding DNA, due to their ability to cooperatively recruit the same factors. Widespread anti-sense transcription has recently been observed at mammalian and *S. cerevisiae* promoters [99-102]. Although bidirectional transcription at promoters has not been identified in Drosophila, it is tempting to speculate that bidirectionality of Pol II initiation could contribute to the proposed synergistic activation of divergent gene pairs. Short antisense transcription may generate a more permissive chromatin state or alter DNA topology; this could impact expression from an individual promoter, and could also contribute to the synergistic interactions with upstream genes described here. Further studies are required to test these and other hypotheses concerning the mechanistic basis for the observed distance effects on chromatin patterns and the proposed local synergism.

Given the metabolic cost of remodeling associated with transcription, in a synergistic activation model there is an obvious energetic favorability provided by linking highly transcribed genes. In Drosophila, genes ≤1kb apart show significant enrichment for GO categories involved in basic cellular processes (Supplementary Table 4). Evidence of promoter synergism is greatly diminished for genes greater than 1kb apart, suggesting that synergistic chromatin interactions act strongly only over a few nucleosomes. Although close to half of all Drosophila genes fall into the >1kb class, the average enrichments surrounding their TSSs exhibit strikingly low densities of Pol II and low levels of active epigenetic markers. Our model predicts that genes expressed in a limited number of cells or tissues will be more isolated, reducing the requirement for active silencing in inappropriate cell types. Indeed, genes >1kb from upstream neighbors show significant overrepresentation of GO categories such as "structural constituent of the cuticle" and "olfactory receptor activity" (Supplementary Table 4).

Upstream space between genes may, in part, reduce the crosstalk and any local synergism between neighbors at the chromatin level. In particular, genes requiring precise temporal or spatial expression during development may be organized in the genome in a manner that ensures regulatory independence from neighboring genes. Transcription factors at the top level of developmental hierarchies require complex regulatory elements and strict transcriptional control, thus it is not surprising that intergenic distance is greater for such genes. The genomic isolation of developmental transcription factor loci, along with PcG-mediated silencing [109], may be necessary to prevent synergistic interaction with other genes and improper activation. This distinction between isolated regulatory genes and clustered housekeeping or widely expressed genes that do not require as specialized transcriptional regulation may be a common feature of metazoan genomes. Conversely, silent genes present in active clusters may utilize special mechanisms to counteract promoter synergism. Interestingly, we observe 'plateaus' of H2Av enrichment across genes not bound by Pol II when they are embedded in clusters of active genes, but not when silent genes are present in regions lacking expressed genes (Fig. 1A). These patterns have been reported for H2AZ in CD4+ cells as well [20] and may represent a special type of chromatin composition required to counteract synergistic activation in gene clusters.

In summary, our results establish a basic link between genomic organization and chromatin structure across widely divergent species, demonstrating the need to incorporate gene orientation and intergenic distances into genome-wide analyses. Gene organization is linked to expression levels in Drosophila and humans, but not in the more compact *S. cerevisiae* genome. Further investigations are required to examine the underlying mechanisms and functional consequences of promoter interactions on gene regulation, as well as the interplay between the proposed local synergism, gene organization, and genome evolution.

Figure 1

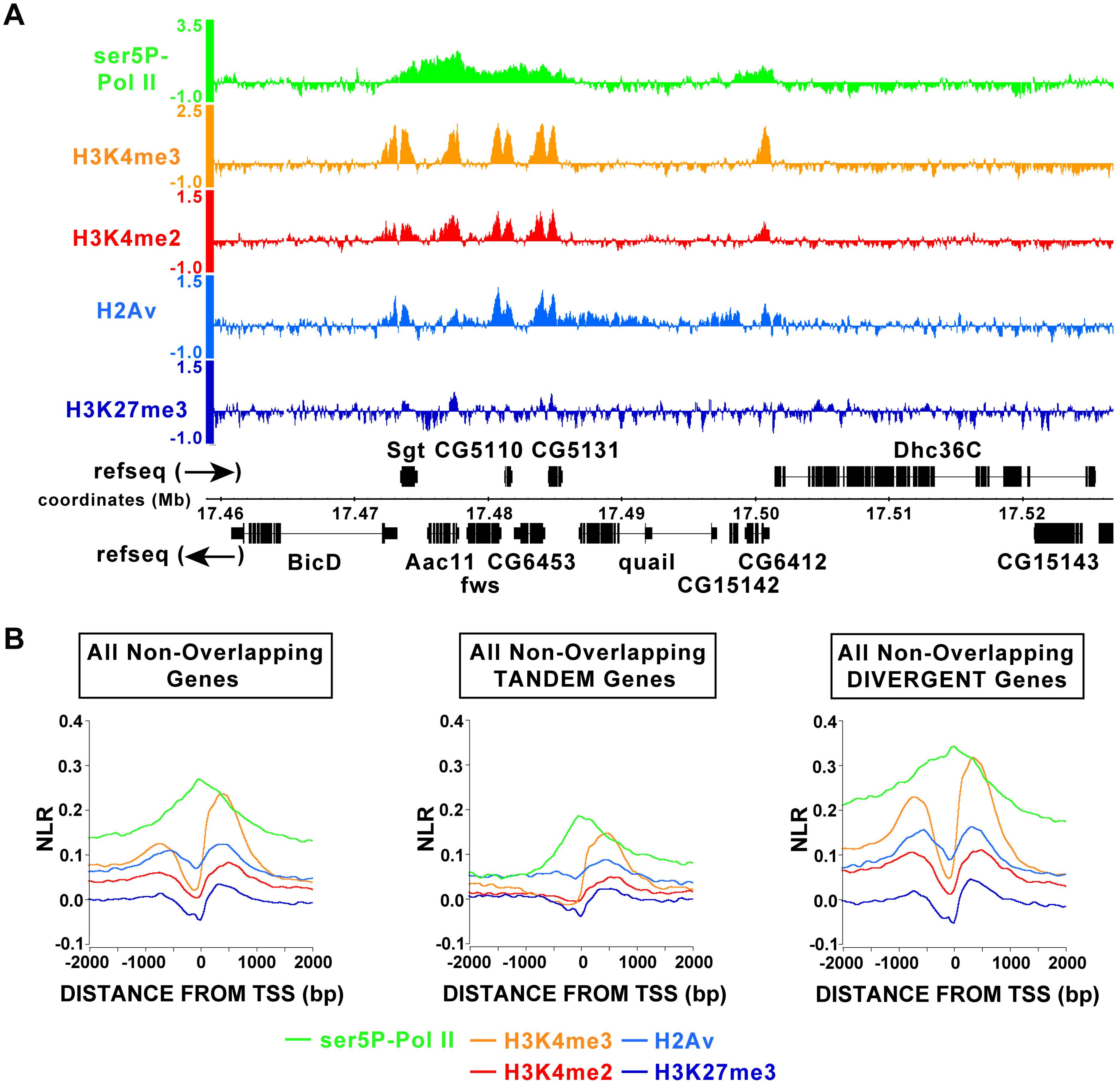


Figure 1. Chromatin patterns at TSSs of *Drosophila* embryos are correlated with gene orientation. A. Browser image from chromosome arm 2L represents a typical genomic region containing active genes. Tandem genes bound by ser5P Pol II, such as *AAc11*, do not show upstream enrichment for H2Av, H3K4me2, or H3K4me3. When intergenic distances are small, paired peaks characterize active divergent genes. Although the *Bicaudal D* gene (*BicD*), shows low ser5P Pol II enrichment, it has similar H2Av, H3K4me2, and H3K4me3 enrichment levels as its active divergent upstream neighbor, *small glutamine-rich tetratricopeptide containing protein* (*Sgt*), indicating that these marks may also be present at genes that are not significantly bound by polymerase. This also suggests that the relationship between active marks and ser5 Pol II binding is not linear. A subset of genes with no ser5P Pol II binding located in active clusters, such as *quail*, show plateaus of H2Av enrichment across the locus. B. Average enrichments surrounding aligned TSSs for all non-overlapping genes in *Drosophila* embryos were calculated using normalized, unsmoothed data. Upstream peaks of H3K4me2, H3K4me3, and H2Av present in the profile for all genes are more pronounced for divergent genes and absent from averages surrounding the TSSs of tandem genes. Average levels of ser5P Pol II, H2Av, H3K4me2, and H3K4me3 are also lower for tandem genes.

Figure 2

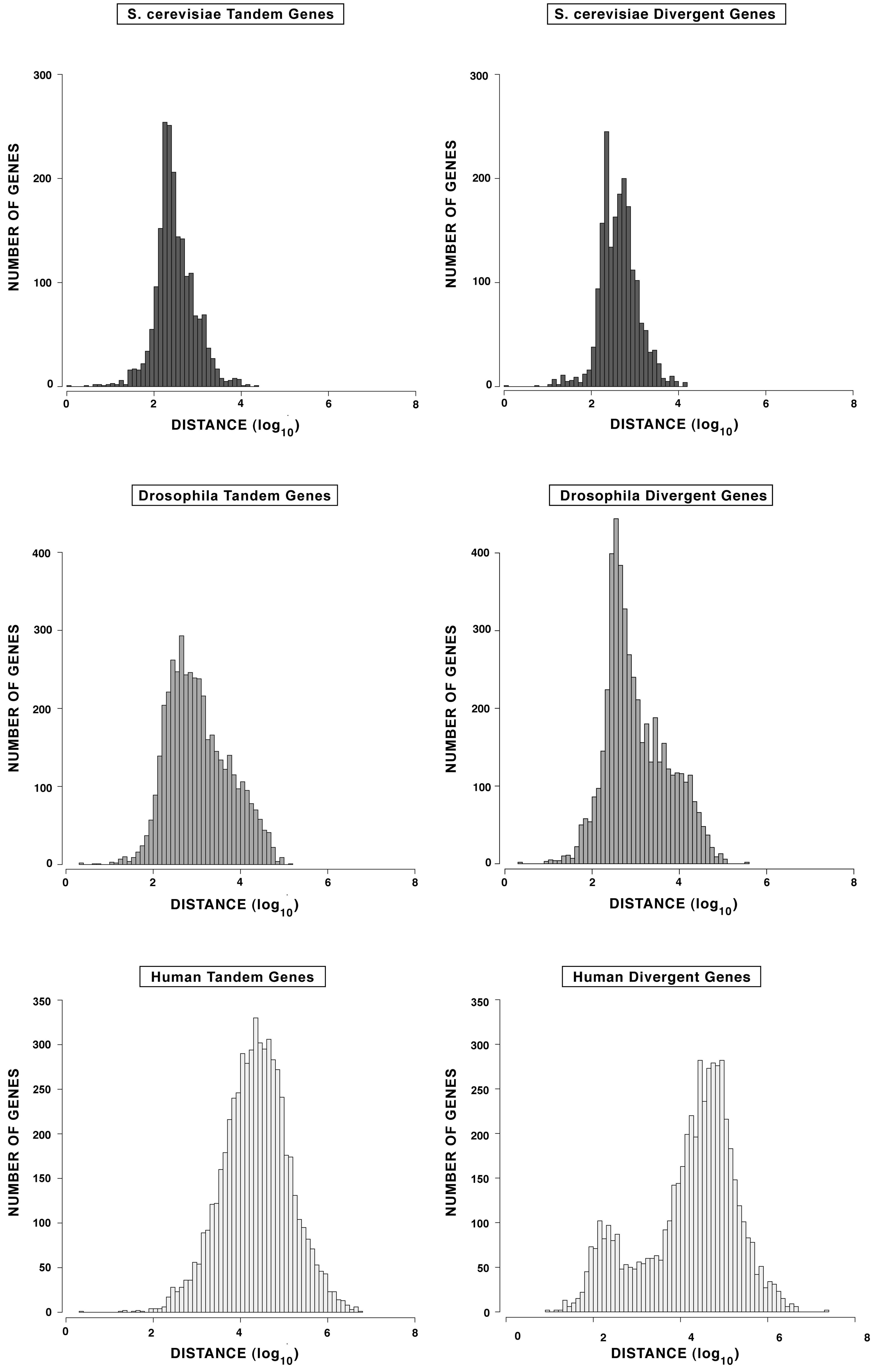
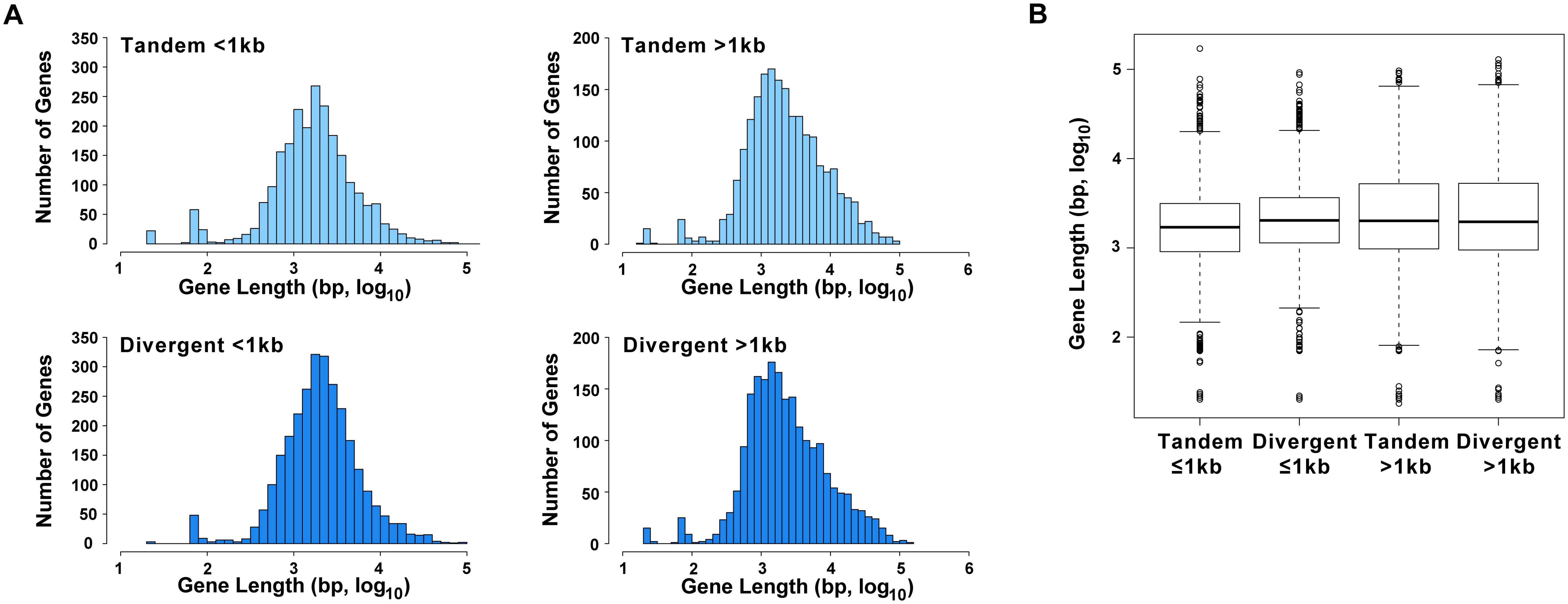


Figure 2. Distributions of distances from TSSs to upstream neighboring genes. Distribution of the $\log_{10}$ distances (bp) of all non-overlapping divergent or tandem genes to their nearest upstream neighbor in the *S. cerevisiae*, *D. melanogaster*, and human genomes. Distances correspond to the length from the TSS to the closest upstream TSS (divergent genes) or TTS (tandem genes). While the majority of genes in *S. cerevisiae* are ≤1kb from upstream neighbors, a skew towards close divergent genes (≤1kb) is clear in Drosophila and human distributions. The distribution of divergent gene distances in *S. cerevisiae* shows a bimodal distribution with peaks centered at ~215bp and ~450bp.

Figure 3



C

	Median gene length (bp)	Mean gene Length (bp)	Std. Err.	Std. Dev.
Tandem 0-200 bp	1516.0	2433.8	155.94	3851.4
Tandem 201-400 bp	1660.0	2877.5	162.53	4394.3
Tandem 401-600 bp	1826.0	4111.1	474.45	10481
Tandem 601-800 bp	1784.0	3207.0	250.76	4336.1
Tandem 801-1,000 bp	2008.5	3351.8	317.54	4794.7
Divergent 0-200 bp	1974.0	2917.2	182.77	4329.1
Divergent 201-400 bp	1935.0	3103.9	128.46	4200.1
Divergent 401-600 bp	2206.0	3833.4	221.34	5577.6
Divergent 601-800 bp	2108.5	4226.7	390.71	7371.9
Divergent 801-1,000 bp	1944.0	4728.3	590.59	8858.9

Figure 3. Distributions of gene lengths for non-overlapping genes in *D. melanogaster*. A. Distributions of gene lengths for non-overlapping tandem and divergent genes ≤1kb or >1kb from upstream neighbors. The distributions for tandem and divergent genes in the >1kb class are not significantly different ($P = 0.9907$, two-sided WRST). Although the distribution of lengths for the less than 1kb classes are significantly different ($P < 10^{-6}$, two-sided WRST), the scale of these differences is small compared to total gene length. Any biological impact with regard to chromatin patterns at the promoter is probably minor. B. Box plot of gene lengths for the same groups shows the medians are similar across all groups. For divergent and tandem genes ≤1kb from neighbors, the means differ by ~400bp and medians differ by <300bp. C. Across 200bp bins (≤1kb) the changes in mean and median gene length show similar trends. Means and medians are slightly lower for tandem genes in all distance bins.

Figure 4

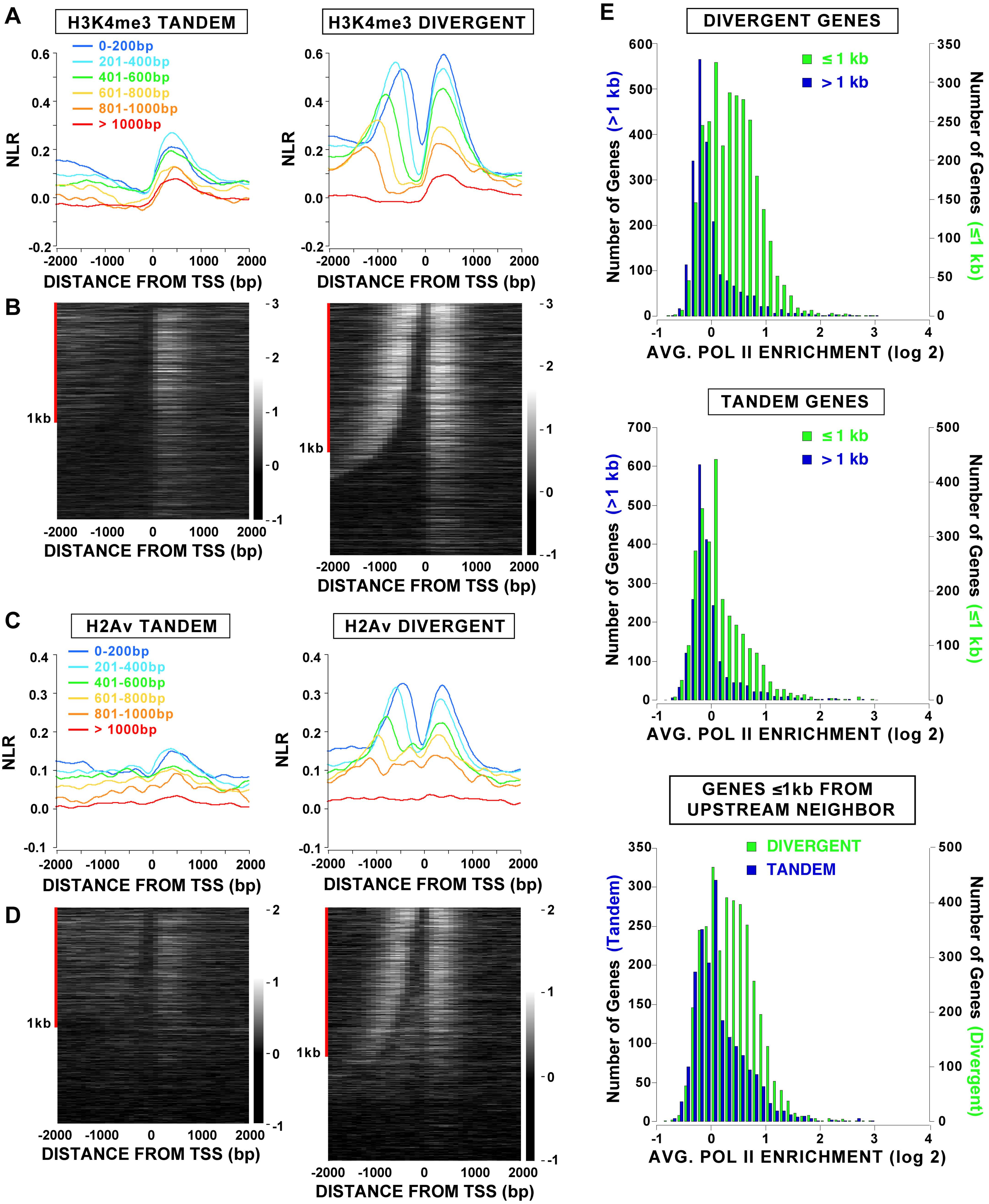


Figure 4. Distance to upstream genes impacts position and levels of active chromatin marks, and Pol II enrichment.–A. Non-overlapping tandem Drosophila genes ≤1kb from their nearest upstream neighbor, based on distance to the closest upstream TSS (divergent genes) or TTS (tandem genes), were binned into 200bp intervals. Average NLRs surrounding the TSSs of binned genes are shown for H3K4me3 (A) and H2Av (C). For comparison, averages for tandem and divergent genes separated by >1kb from upstream neighbors are plotted in red. Heat maps display average H3K4me3 (B) and H2Av (D) enrichment in discrete 100bp windows surrounding the TSS of individual divergent and tandem (Only genes with sufficient array coverage were included; NLR scale shown at right of plots). Genes were ranked by distance before plotting. Red bar indicates genes ≤1kb from the upstream neighbor. E. Distributions of average ser5P Pol II density across genes. We computed the average NLR for the transcribed region of non-ovelapping genes. Genes were then separated based on orientation and distance. Distributions show that both divergent and tandem genes ≤1kb away from upstream neighbors have higher ser5P Pol II enrichment compared to genes separated by >1kb. For genes ≤1kb from neighbors, there is a higher proportion of divergent genes with positive average ser5P Pol II density compared to tandem genes.

Figure 5

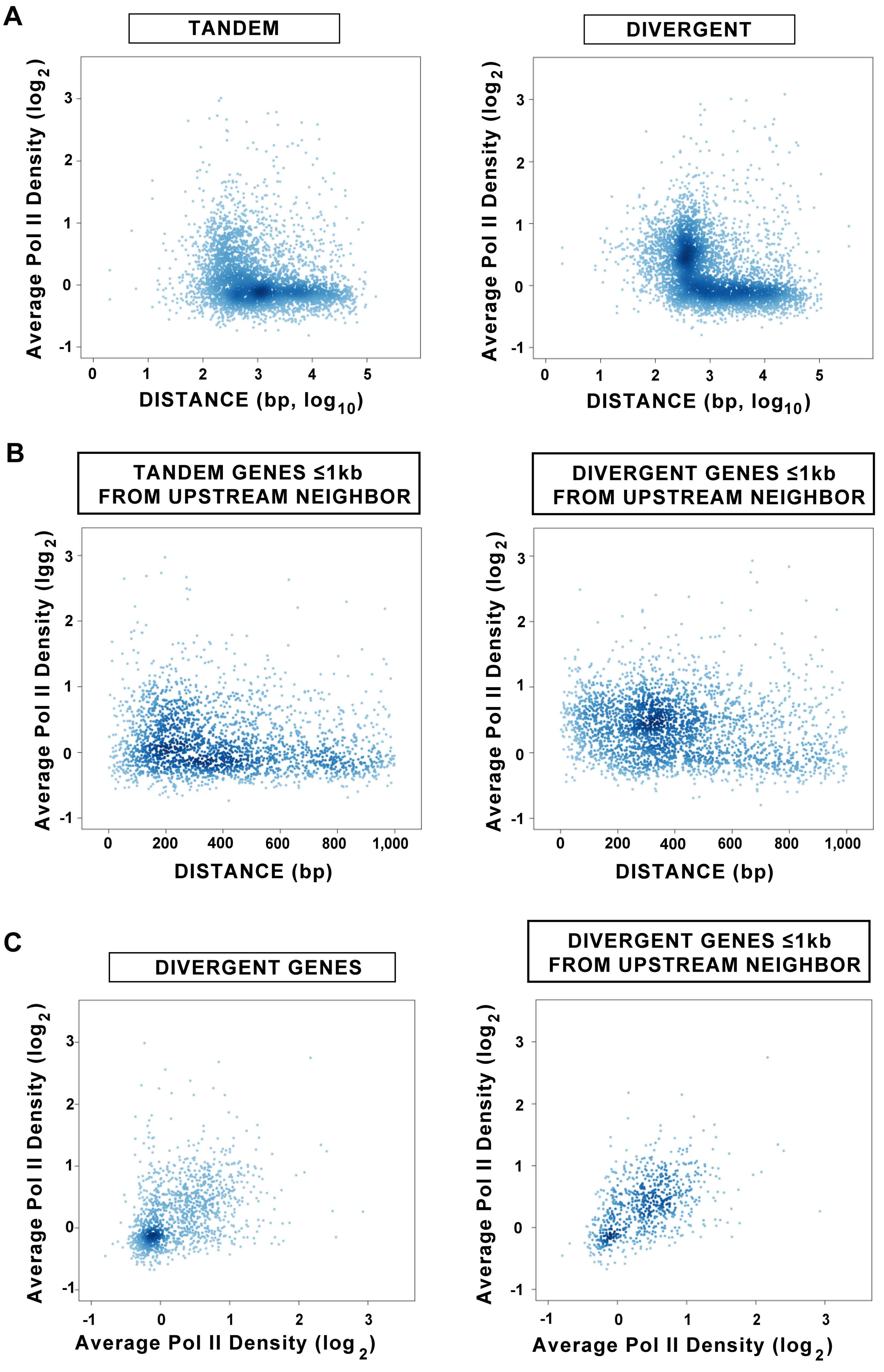


Figure 5. Pol II density in Drosophila embryos indicates higher activity of divergent genes ≤1kb apart. A. Unsmoothed NLRs from ser5P-Pol II ChIP-array data [45] were used to estimate the average Pol II density (average NLR across the transcribed region) for all non-overlapping genes. Average densities were then plotted against log$_{10}$ distance (bp) to the upstream neighbor. Although Pol II density of gene pairs is likely not independent, for average Pol II density and distance to neighbor (not log scale), the Pearson's correlation coefficient (r) = -0.11 for tandem genes and r = -0.14 for divergent genes ($P < 10^{-6}$ for both). B. For genes ≤1kb from the upstream neighbor, average Pol II densities were plotted against the distance (bp) to the upstream neighbor. r = -0.15 for tandem genes (≤1kb) and r = -0.16 for divergent genes (≤1kb) ($P < 10^{-6}$ for both). C. Average Pol II densities for the 1,340 divergent gene pairs where both members do not overlap other genes were plotted against each other. Also shown is the plot for the subset of 729 non-overlapping divergent gene pairs with TSSs ≤1kb apart. For all divergent gene pairs is r = 0.48, whereas for divergent pairs ≤1kb apart, r = 0.51 ($P < 10^{-6}$ for both).

Figure 6

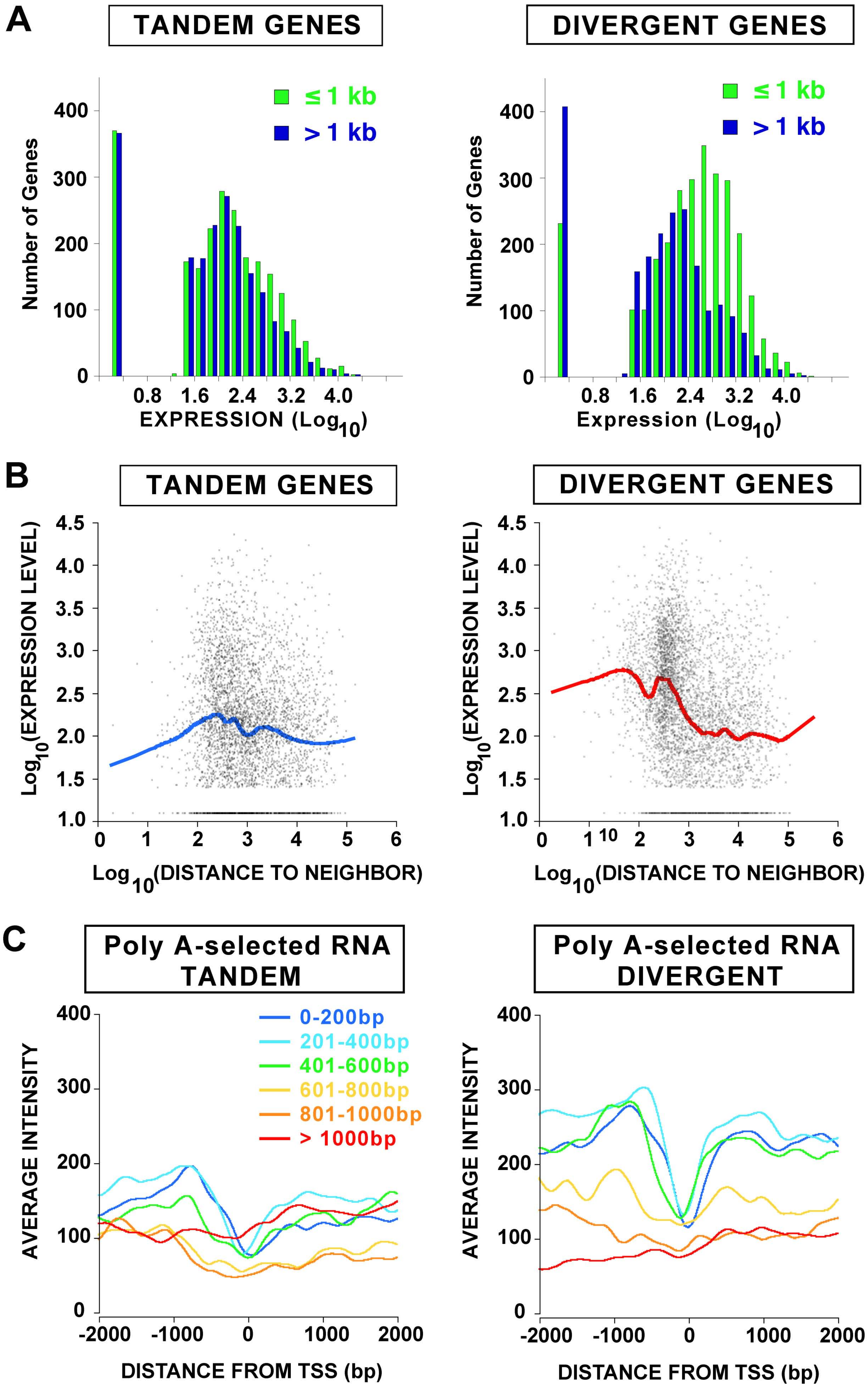


Figure 6. modENCODE (http://www.modencode.org/) RNA tiling array expression data from Drosophila embryos confirm higher expression levels for close divergent genes. A. Average Log_{10}(expression level) from Poly-A-selected RNA from 2-4h Drosophila embryos, separated based on gene orientation and distance (≤1kb or >1kb) to the upstream neighbor, were used to generate distributions. Plots represent the 4,253 tandem and 4,875 divergent non-overlapping genes included in the data set. Transcripts from close divergent genes are more abundant than those from tandem genes in 2-4hr embryos ($P < 10^{-6}$, one-sided WRST) and that, more generally, genes ≤1kb from upstream neighbors tend to have higher expression levels than similarly oriented genes >1kb from neighbors ($P < 10^{-6}$ for both divergent and tandem genes, one-sided WRST) (Supplementary Fig. 4). Divergent genes >1kb are not significantly more expressed than tandem >1kb genes ($P = 0.1822$, two-sided WRST).B. Log_{10}(expression level) plotted against $\log_{10}$(distance_bp) for individual tandem and divergent genes. Divergent genes show higher overall expression levels than tandem genes. Lowess smoothed curves for both orientations show a decrease in expression as intergenic distance increases across the 0-1kb range (0-3 on the x-axis), with a much more pronounced decline for divergent genes. C. Average normalized probe intensities surrounding TSSs for orientation and distance (200bp) classes.

Figure 7

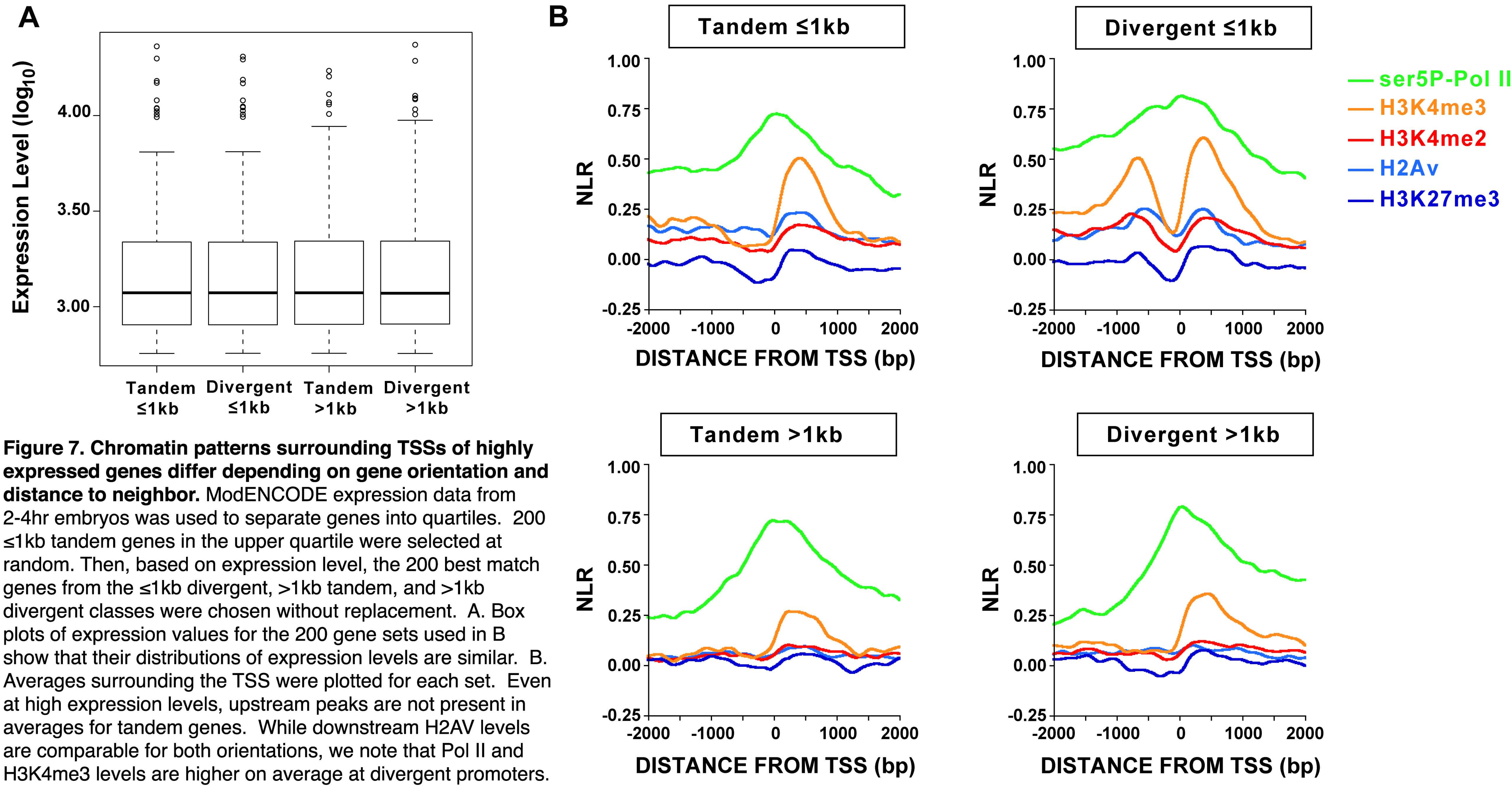


Figure 7. Chromatin patterns surrounding TSSs of highly expressed genes differ depending on gene orientation and distance to neighbor. ModENCODE expression data from 2-4hr embryos was used to separate genes into quartiles. 200 ≤1kb tandem genes in the upper quartile were selected at random. Then, based on expression level, the 200 best match genes from the ≤1kb divergent, >1kb tandem, and >1kb divergent classes were chosen without replacement. A. Box plots of expression values for the 200 gene sets used in B show that their distributions of expression levels are similar. B. Averages surrounding the TSS were plotted for each set. Even at high expression levels, upstream peaks are not present in averages for tandem genes. While downstream H2AV levels are comparable for both orientations, we note that Pol II and H3K4me3 levels are higher on average at divergent promoters. Also, for both tandem and divergent genes, although the Pol II levels are comparable, the levels of H3K4me3 are much lower in the >1kb class.

Figure 8

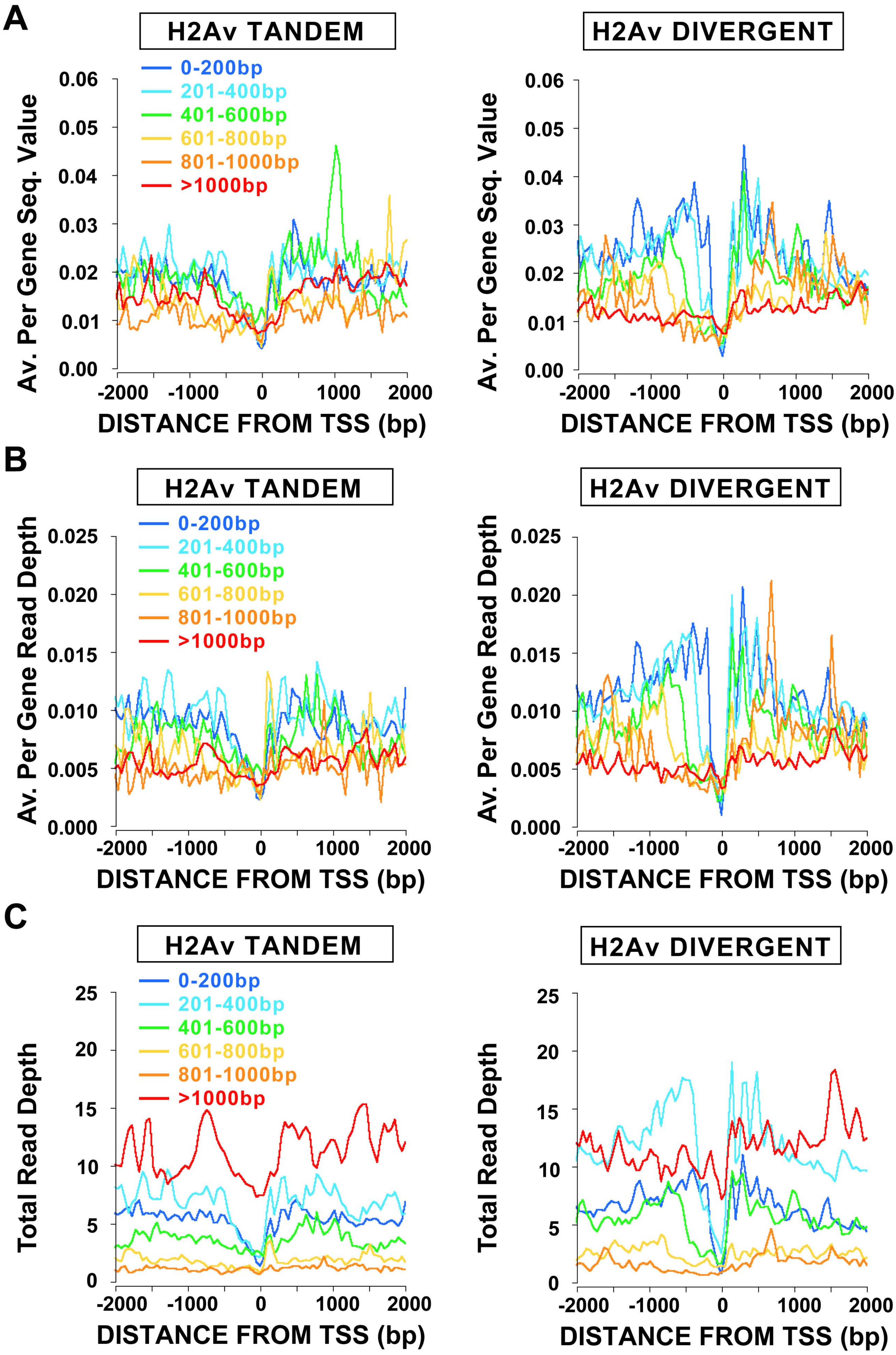


Figure 8. Gene distance and orientation influence chromatin patterns at TSSs in human CD4+ cells. Non-overlapping genes from the UCSC Known Gene set were grouped based on orientation and distance to nearest upstream neighbors. Read depths from [40] surrounding TSSs for tandem (250) and divergent (908) genes ≤1kb were summed for each 200bp distance interval, and divided by the total number of genes in the class to produce a per gene average read depth for H3K4me3 (A) and H2A.Z (C). Heat maps for H3K4me3 (B) and H2A.Z (D) were generated by summing the total reads in discrete 100bp windows surrounding the TSSs of individual tandem (5,744) and divergent (5,342) genes (per 100bp read count scale shown at right of plots). Genes were sorted by distance to upstream neighbor before plotting. Genes ≤1kb from the upstream neighbor are indicated by the red bar. E. Available CD4+ cell expression data from 2,787 divergent and 2,945 tandem non-overlapping genes was separated based on orientation and distance. Distributions of $\log_{10}$ expression values show higher expression for divergent genes ≤1kb away from upstream neighbors as compared to genes separated by >1kb. For genes ≤1kb from neighbors, the distribution expression levels shows a higher proportion of divergent genes with very high expression (>3), as compared to tandem genes.

Figure 9

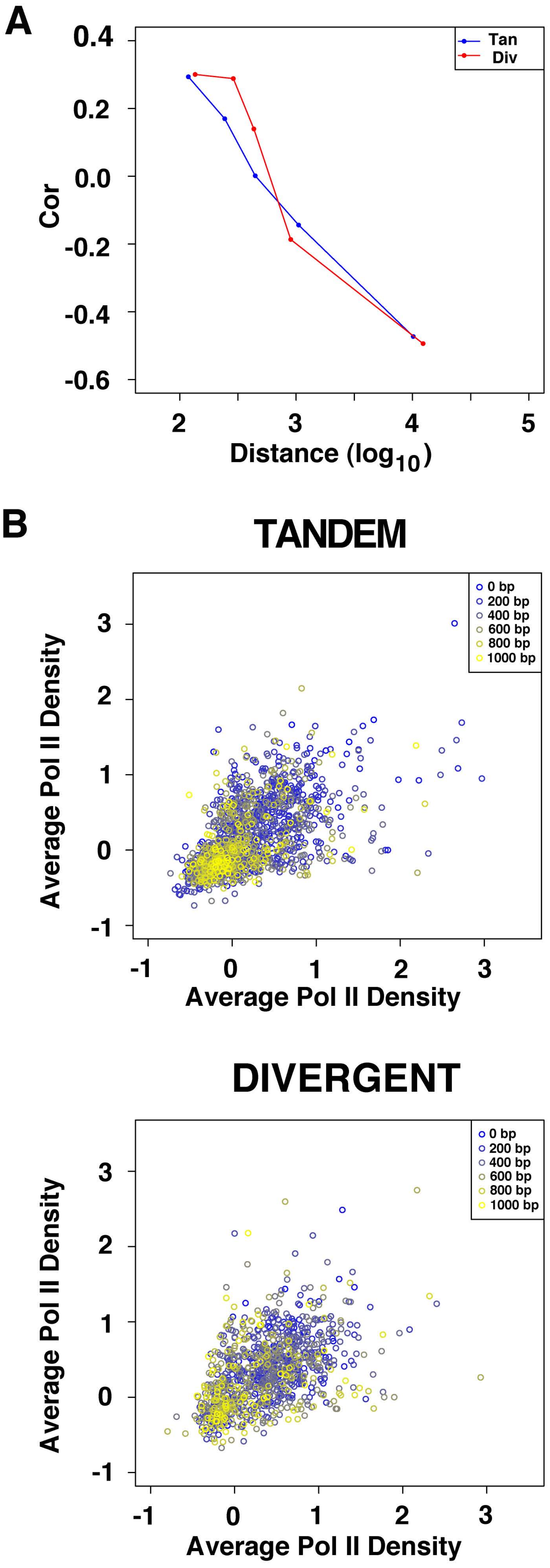


Figure 9. Correlations of average Pol II densities [45] for neighboring tandem and divergent gene pairs decrease with increasing intergenic distance. A. Spearman's rank correlation of average Pol II densities [45] were calculated for all gene pairs where both members do not overlap other genes. Tandem and divergent gene pairs were divided into 5 groups of equal size, and Spearman's rho was plotted at the group mean distance. P-values for all correlations were all <1.6e-05, with the exception of the tandem group with mean distance = 244bp, which was not significant. A total of 1280 for tandem and 1,090 divergent pairs were used in the analysis. Correlation levels are slightly higher for divergent genes. Some of the correlation between tandem pairs results from the fact that the majority of these genes (and thus pairs) have very low enrichment values for active marks (See Supplementary Fig. 7B). Additionally, while close divergent gene pairs tend to be bound by Pol II, they do not necessarily have equivalent levels of enrichment or, by extension, similar expression levels (See Supplementary Fig. 7B). B. Average Pol II densities for gene pairs (≤1kb) with distance between genes mapped along a blue-yellow color spectrum shows that, while averages are generally higher for divergent pairs, there is a great deal of variation. Even in instances where both divergent genes are enriched for Pol II, the levels may be quite different. Levels for tandem genes tend to be lower and fall along the diagonal, particularly as distance nears 1kb, indicating that both genes in a given pair are not bound by Pol II. So, although correlations are similar for both orientations, the distribution of averages for tandem and divergent gene pairs are distinct.

Figure 10

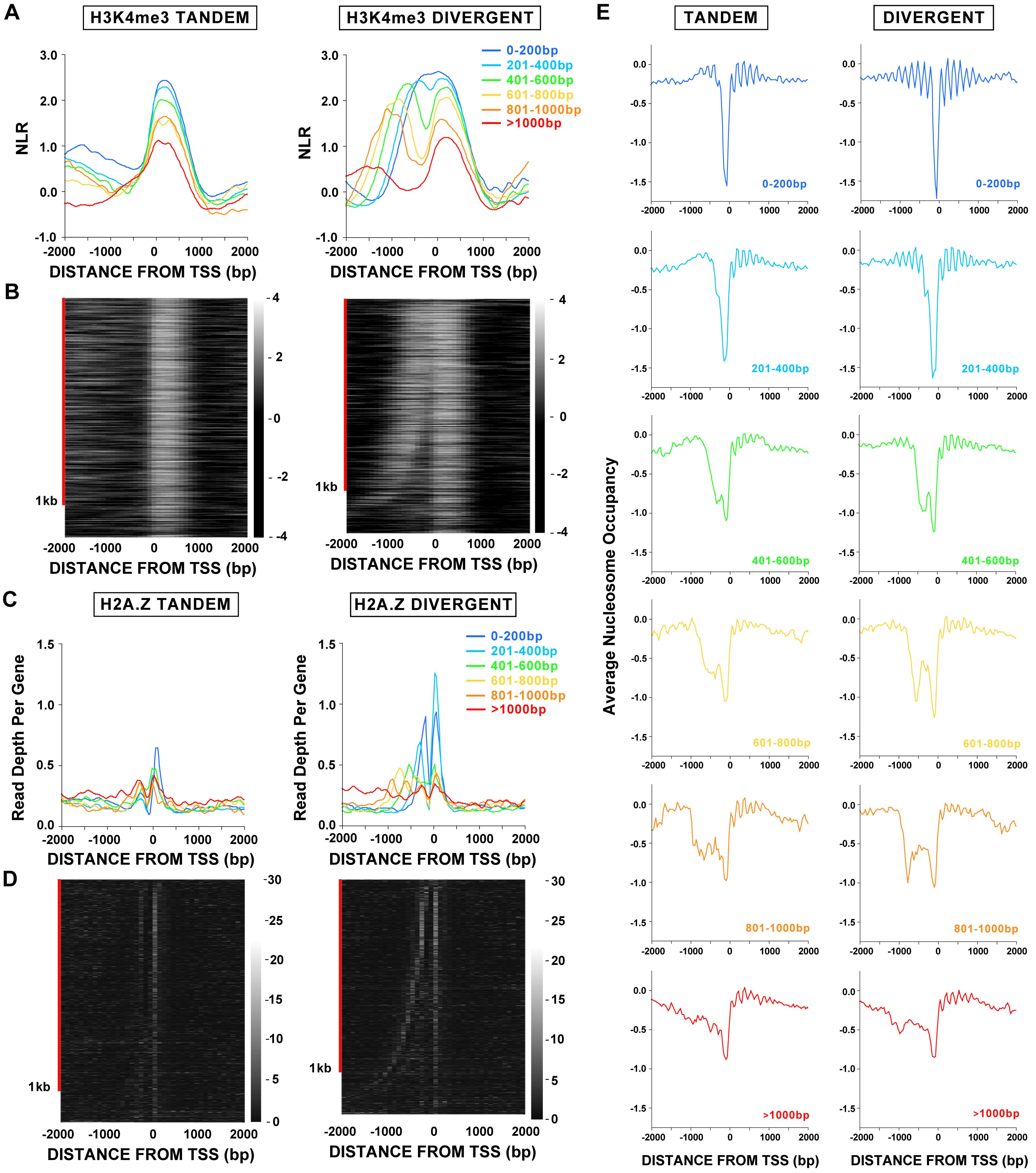


Figure 10. Chromatin patterns at *S. cerevisiae* promoters are influenced by gene orientation and intergenic distance. Average NLR for H3K4me3 (A), data from [52], and per gene corrected read counts from sequenced H2A.Z nucleosomes (C), data from [31], for tandem and divergent non-overlapping genes in *S. cerevisiae.* While divergent genes show the reported '+1' and '-1' H2A.Z nucleosome peaks flanking the TSS, the position of the '-1' nucleosome peak changes depending on distance to TSSs of neighboring upstream genes. Divergent genes also show a distance-dependent decrease in H2A.Z read depth. H3K4me3 levels decrease with distance for both orientations. Heat maps show individual gene average H3K4me3 enrichment (B) or total H2A.Z corrected read counts (D) in discrete 100bp windows surrounding the TSS of tandem and divergent genes (NLR scale and per 100bp corrected read count scale shown at right of plots). Genes were ranked by distance before plotting. Genes ≤1kb from the upstream neighbor are indicated by the red bar. E. Average nucleosome occupancies surrounding TSSs for tandem and divergent distance classes. Size of NDR(s) and intensity and "fuzziness" of nucleosome occupancy surrounding promoters depend strongly on gene orientation and distance. Upstream nucleosome phasing is primarily associated with divergent genes, and phasing is stronger for genes with shorter intergenic distances.

Figure 11

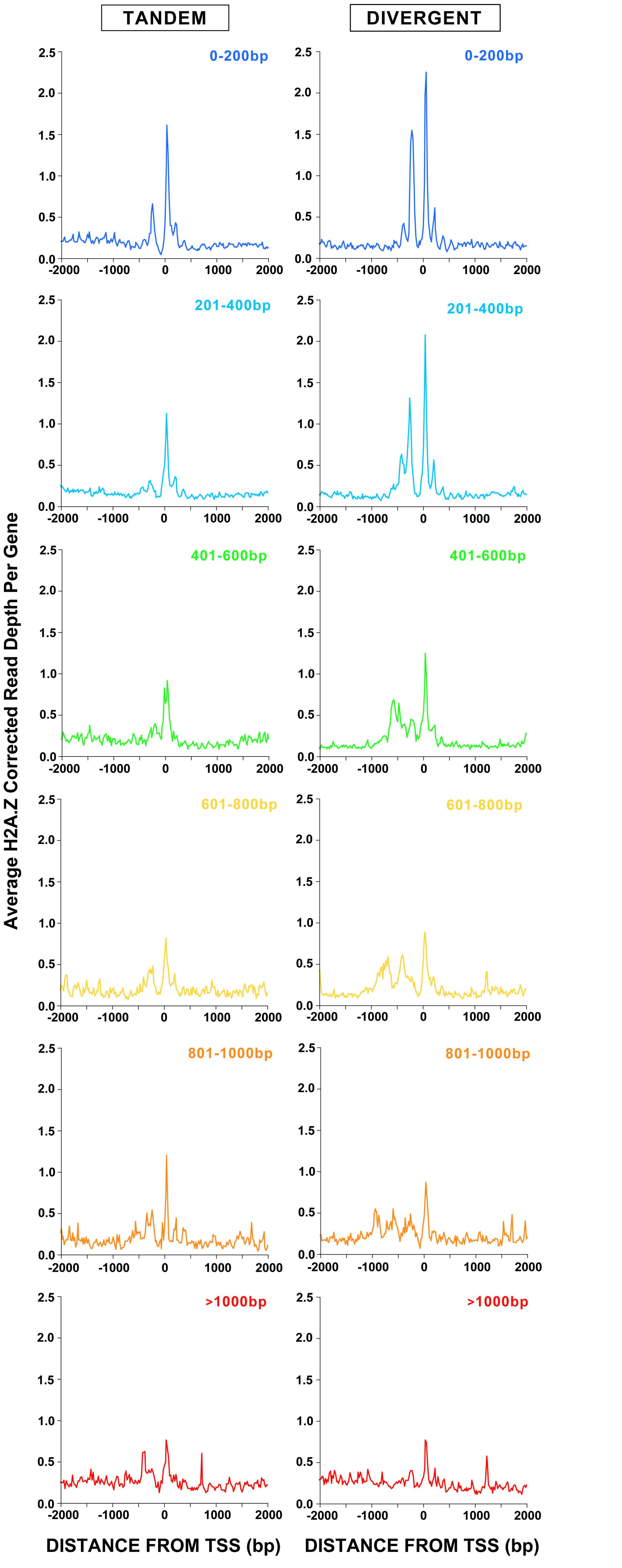


Figure 11. High resolution TSS averages of read depth from H2A.Z nucleosome ChIP-seq in *S. cerevisiae* [31]*.* Unsmoothed average corrected read depth per gene was plotted for discrete 200bp bins for non-overlapping tandem and divergent genes in *S. cerevisiae* [31]. This is the same data shown in Fig. 10C, but plotting unsmoothed averages separately makes patterns easier to resolve. Separation based on distance and orientation produces large differences in H2A.Z nucleosome peak positions and heights.

Figure 12

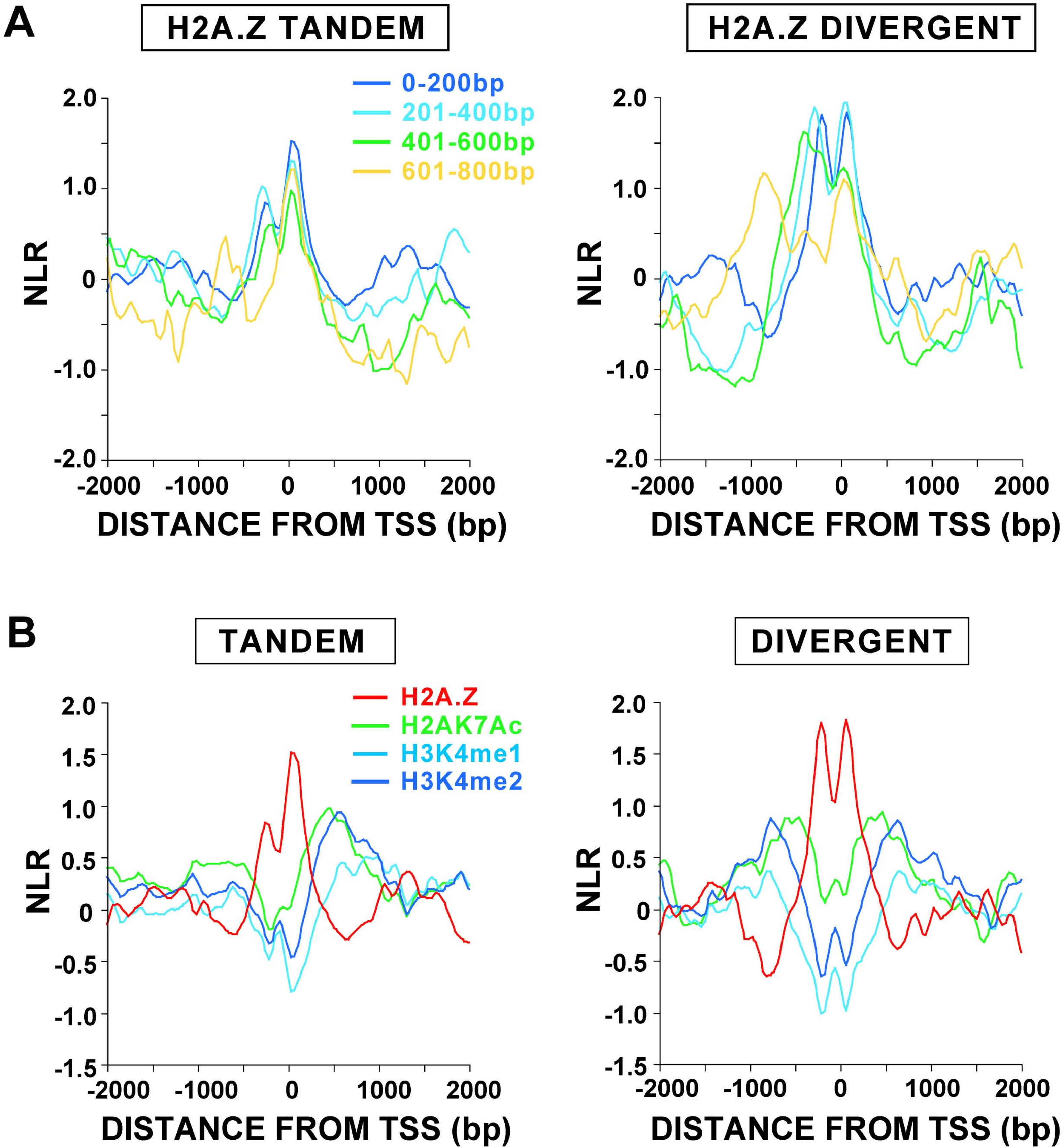


Figure 12. ChIP-array data from *S. cerevisiae* [34,42] supports H2A.Z ChIP-seq results and shows evidence of modified histones in the NFR. A. Average patterns of H2A.Z surrounding aligned TSSs from ChIP-array data largely agree with the results from higher resolution sequencing data. Non-overlapping divergent (1,443) and tandem (1,619) genes ≤800bp from their nearest upstream neighbor with some coverage on the arrays (+/-2kb from the TSS) were binned in 200bp intervals. Average NLRs surrounding the TSS reveal a shifting of upstream H2A.Z peaks proportional to the distance to upstream genes for divergent genes. However, enrichment of H2A.Z is not higher for divergent genes, as observed in *S. cerevisiae* H2A.Z sequencing data. Though the H2A.Z sequencing results are compelling, ChIP-array data with greater coverage or ChIP-seq experiments which utilize input controls will be required to resolve this apparent discrepancy between the existing data sets. B. Average enrichments for H2A.Z, H2AK7Ac, H3K4me1, and H3K4me2 surrounding the TSSs of 367 divergent and 678 tandem genes 0-200bp apart in *S. cerevisiae*. While divergent genes show the reported '+1' and '-1' H2A.Z nucleosome peaks flanking the TSS, tandem genes show a much smaller –1 H2A.Z peak. A small peak of H2A acetylated at K7 lies within the NFR. Though the levels are not above background, H3K4me1 and me2 peaks in the NFR are inversely positioned in relation to H2A.Z enrichment. These patterns suggest that histones occupy the NFR of some genes at some time during the cell cycle and that this occupancy has a positional relationship to H2A.Z.

Figure 13

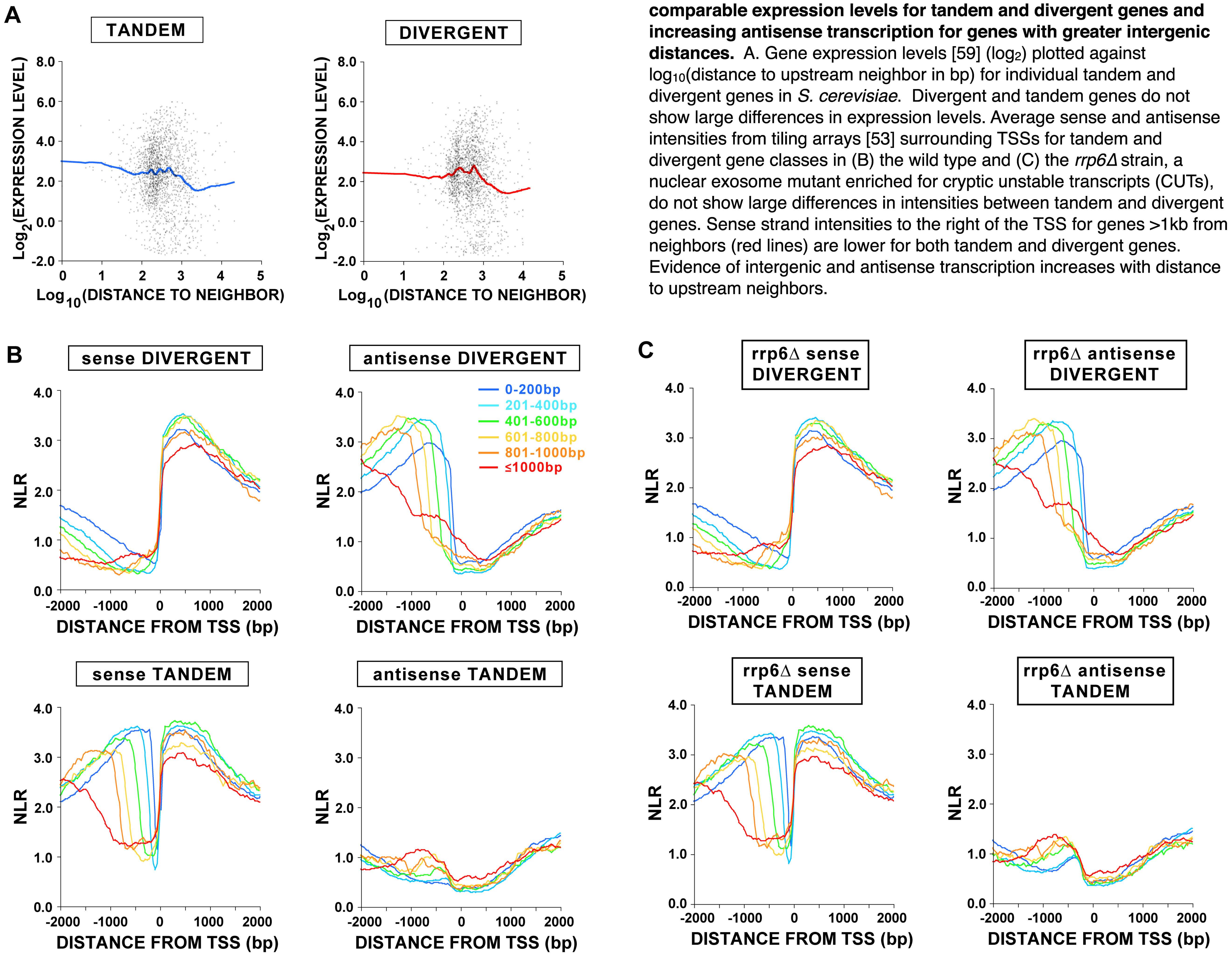


Figure 13. Transcriptome array data [53,59] in *S. cerevisiae* suggests comparable expression levels for tandem and divergent genes and increasing antisense transcription for genes with greater intergenic distances. A. Gene expression levels [59] ($\log_2$) plotted against $\log_{10}$(distance to upstream neighbor in bp) for individual tandem and divergent genes in *S. cerevisiae*. Divergent and tandem genes do not show large differences in expression levels. Average sense and antisense intensities from tiling arrays [53] surrounding TSSs for tandem and divergent gene classes in (B) the wild type and (C) the *rrp6Δ* strain, a nuclear exosome mutant enriched for cryptic unstable transcripts (CUTs), do not show large differences in intensities between tandem and divergent genes. Sense strand intensities to the right of the TSS for genes >1kb from neighbors (red lines) are lower for both tandem and divergent genes. Evidence of intergenic and antisense transcription increases with distance to upstream neighbors.

Figure 14

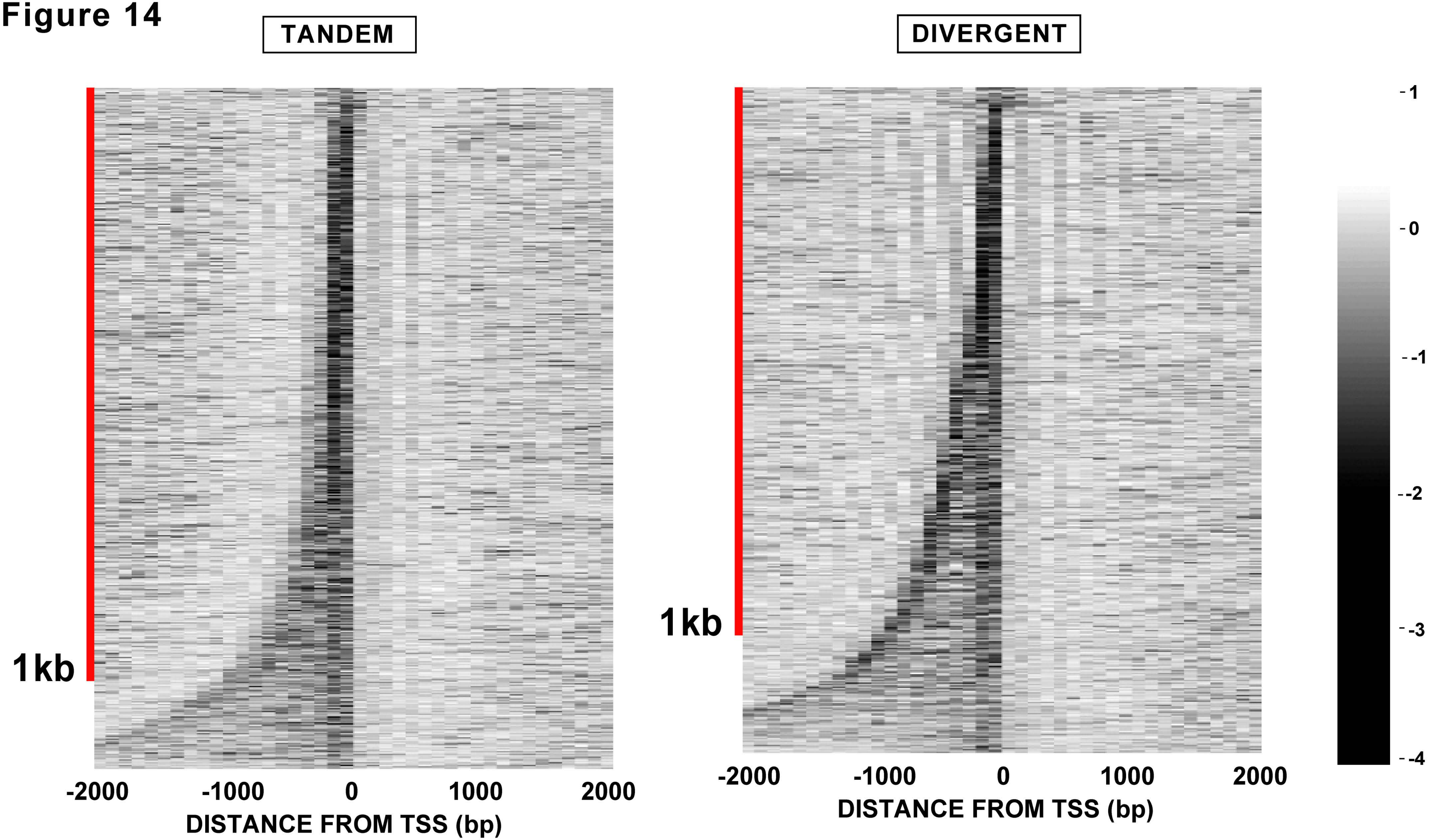


Figure 14. Nucleosome occupancy [38] heat maps from *S. cerevisiae* demonstrate the impact of distance and orientation on NFR size and nucleosome positioning. Heat maps show the average nucleosome occupancy in 100bp windows surrounding the TSS of 1,958 tandem and 1,914 divergent genes. Close genes show narrower NFRs with greater nucleosome depletion. Genes were sorted by distance to upstream neighbor before plotting. Those ≤1kb from the upstream neighbor are indicated by the red bar.

Figure 15

A
B
C
D
E
TANDEM
DIVERGENT
Average Model Score
Average Predicted Nucleosome Occupancy
Average Nucleosome Occupancy
0.0
-1.0
-2.0
-3.0
-4.0
0.8
0.7
0.6
0.5
0.4
-0.5
-1.5
2.0
1.0
0.85
0.75
0.65
0.55
-2000
-1000
0
1000
2000
0-200bp
201-400bp
401-600bp
601-800bp
801-1000bp
>1000bp
DISTANCE FROM TSS (bp)

Figure 15. Sequence-predicted nucleosome occupancies [58] in *S. cerevisiae* and Drosophila vary with intergenic distance and gene orientation and predict *in vivo* occupancies [38] in *S. cerevisiae.* Model scores and nucleosome occupancy predictions as provided by the authors [66] surrounding TSSs were averaged for genes binned based on distance and orientation. A) The average scores (by bp) given by the nucleosome occupancy model in *S. cerevisiae* are lower in the intergenic regions upstream of tandem genes compared to those for divergent genes. B) The predicted nucleosome occupancy in *S. cerevisiae* is in good agreement with the *in vivo* data [38] (reproduced from Fig. 3E for comparison) (C), replotted from Fig. 3E for comparison. D) The average scores (by bp) given by the model in Drosophila and the predicted occupancy (E) show trends similar to those seen in *S. cerevisiae.* However, divergent genes, on average, tend to have higher model scores and predicted occupancies than tandem genes, and close genes (≤400bp) generally have higher scores and predicted occupancies. Interestingly, although there is not strong evidence for -1 nucleosomes in Drosophila, the model produces high scores in the -1 position for many classes.

Figure 16

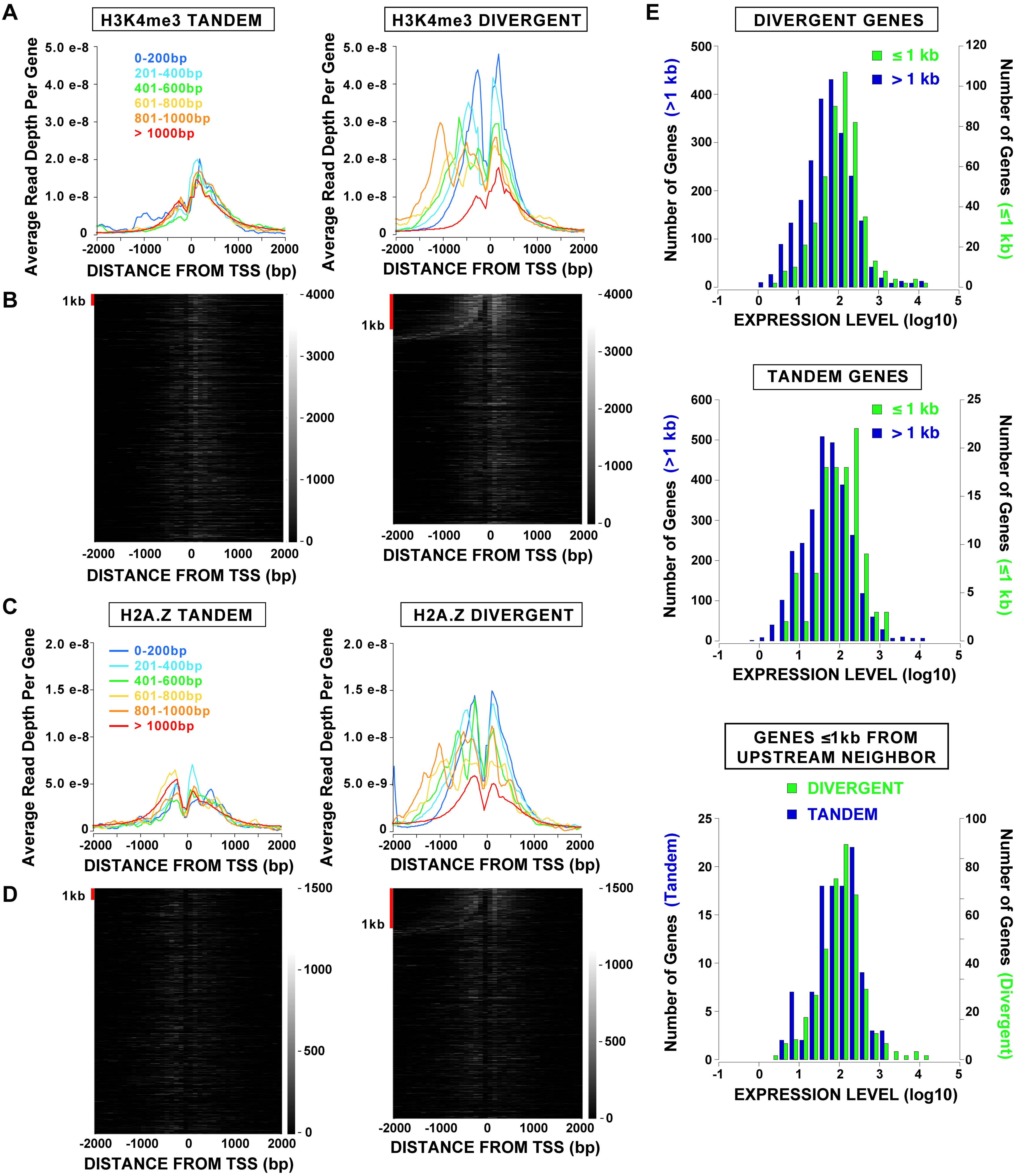


Figure 16. Gene distance and orientation influence chromatin patterns at TSSs in human CD4+ cells. Non-overlapping genes from the UCSC Known Gene set were grouped based on orientation and distance to nearest upstream neighbors. Read depths from [40] surrounding TSSs for tandem (250) and divergent (908) genes ≤1kb were summed for each 200bp distance interval, and divided by the total number of genes in the class to produce a per gene average read depth for H3K4me3 (A) and H2A.Z (C). Heat maps for H3K4me3 (B) and H2A.Z (D) were generated by summing the total reads in discrete 100bp windows surrounding the TSSs of individual tandem (5,744) and divergent (5,342) genes (per 100bp read count scale shown at right of plots). Genes were sorted by distance to upstream neighbor before plotting. Genes ≤1kb from the upstream neighbor are indicated by the red bar. E. Available CD4+ cell expression data from 2,787 divergent and 2,945 tandem non-overlapping genes was separated based on orientation and distance. Distributions of $\log_{10}$ expression values show higher expression for divergent genes ≤1kb away from upstream neighbors as compared to genes separated by >1kb. For genes ≤1kb from neighbors, the distribution expression levels shows a higher proportion of divergent genes with very high expression (>3), as compared to tandem genes.

Figure 17

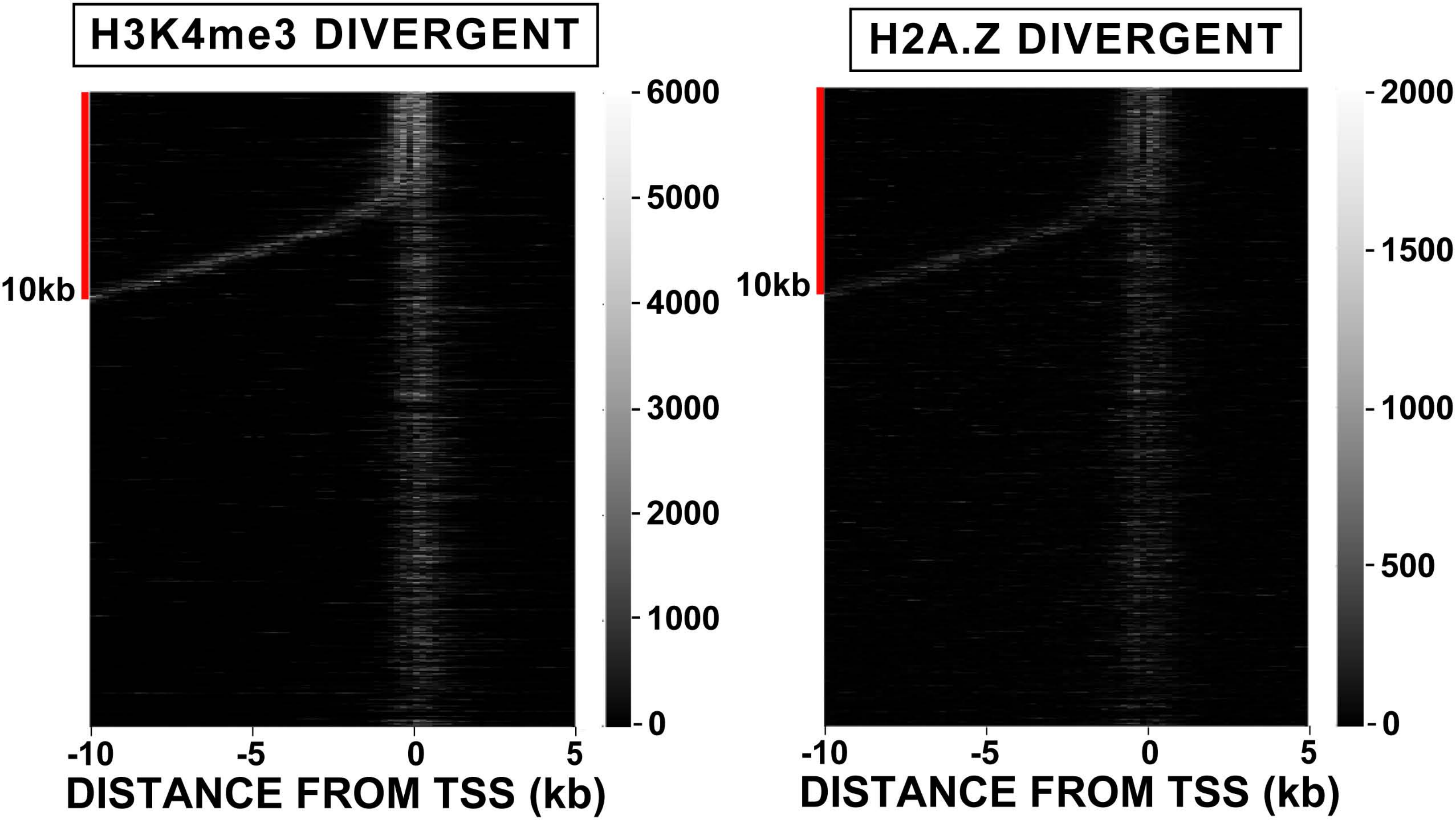


Figure 17. Individual gene heat maps from CD4+ cells [40] show that upstream peaks are primarily associated with neighbors as far as 10kb away. Heat Maps for H3K4me3 and H2A.Z were generated by summing the total reads in discrete 200bp windows surrounding the TSSs of individual divergent (5,342) genes (per 100bp corrected read count scale shown at right of plots). Genes were sorted by distance to upstream neighbor before plotting. Genes ≤10kb from the upstream neighbor are indicated by the red bar.

Figure 18

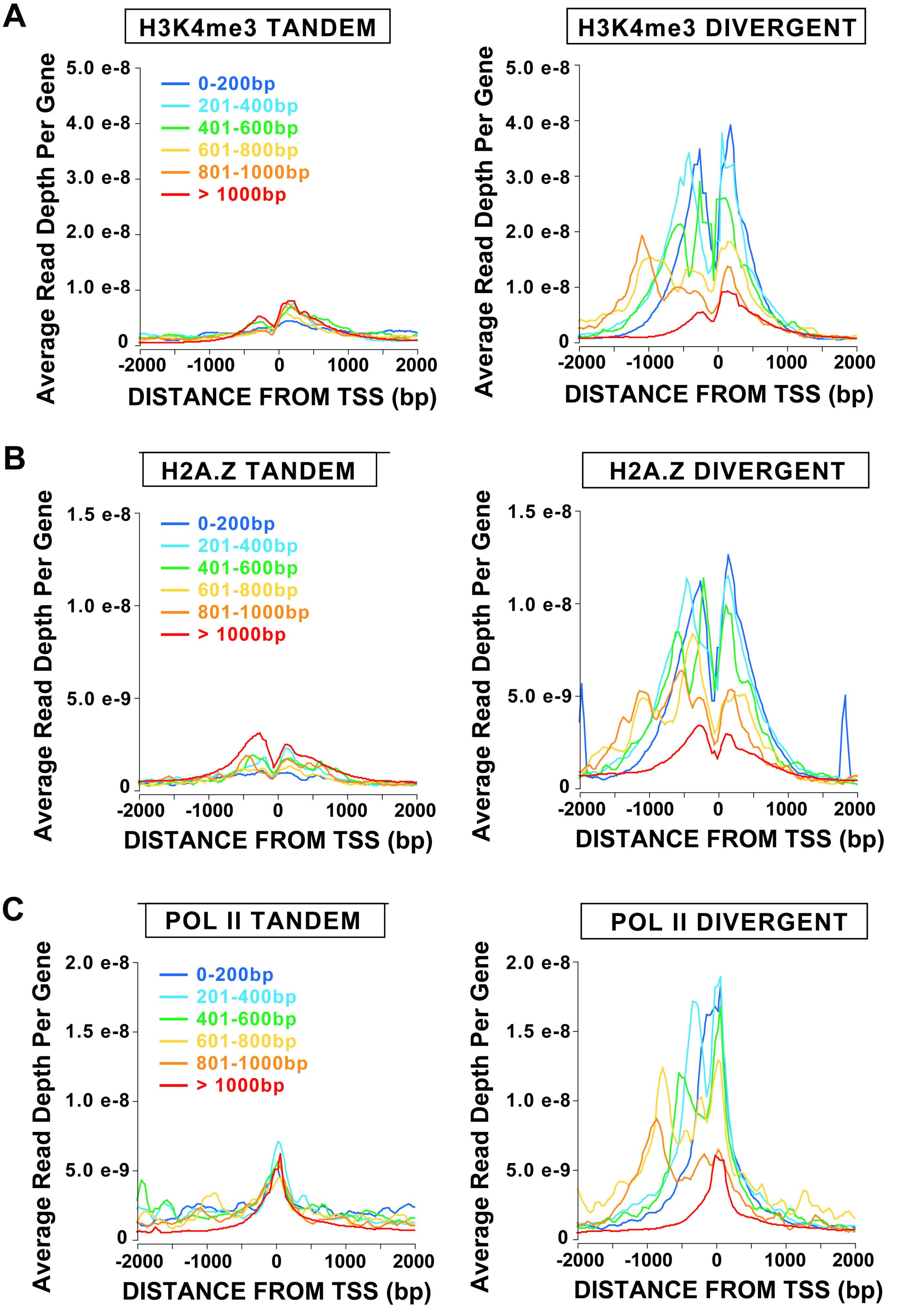


Figure 18. Analysis of chromatin patterns at TSSs in CD4+ cells [40] based on ENSEMBL annotations confirm the influence of gene distance and orientation. Non-overlapping genes from the ENSEMBL gene set were grouped based on orientation and distance to nearest upstream neighbors. Read depths surrounding TSSs for tandem (721) and divergent (1,027) genes ≤1kb were summed for each 200bp distance interval, and divided by the total number of genes in the class to produce a per gene average read depth for H3K4me3 (A), H2A.Z (B) and Pol II (C).

Figure 19

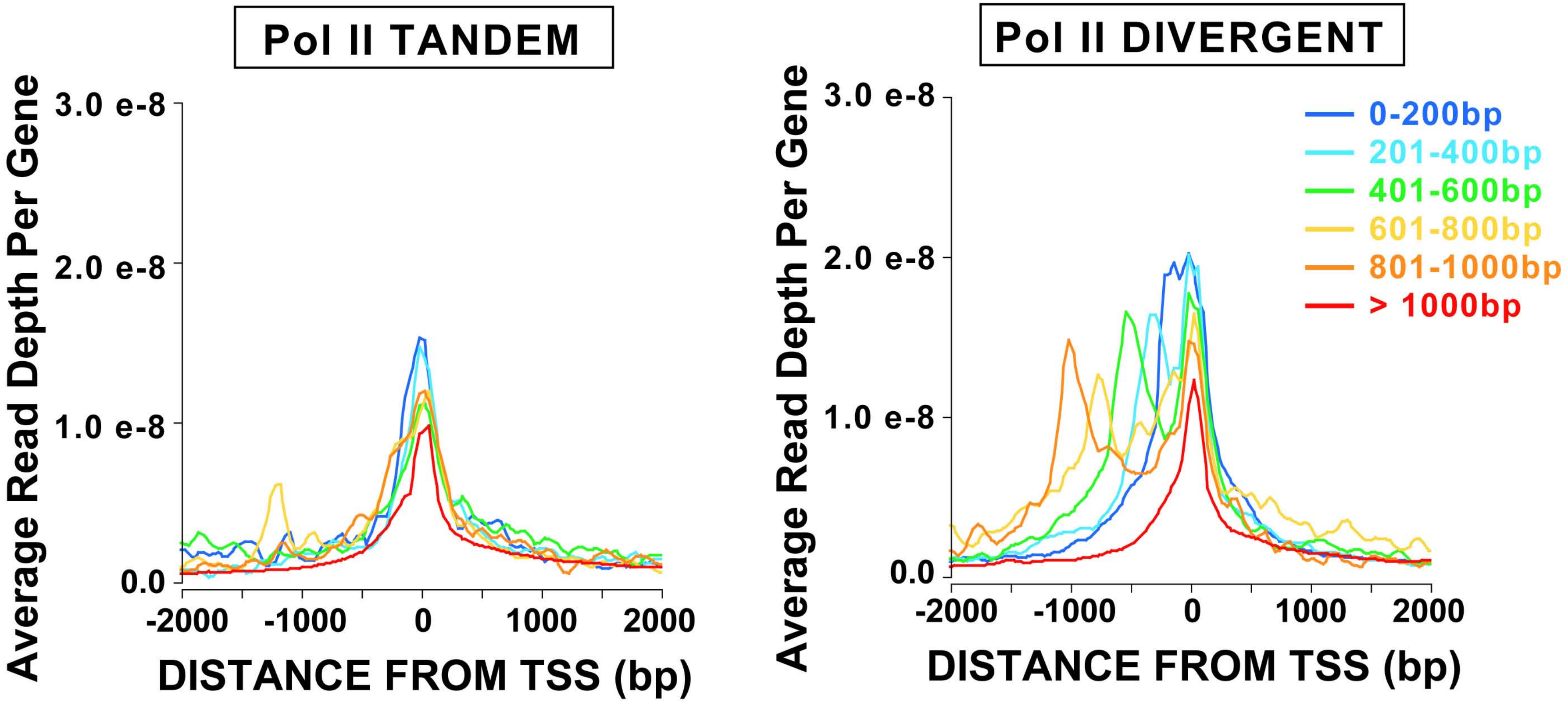


Figure 19. Pol II ChIP-sequencing data [40] from CD4+ cells shows orientation and distance-based enrichment. Non-overlapping genes from the UCSC Known Gene set were grouped based on orientation and distance to nearest upstream neighbor. For 200bp distance intervals, read depths surrounding TSSs for each class were summed and divided by the total number of genes in the class to produce a per gene average read depth for tandem and divergent genes. Divergent genes ≤1kb have higher averages than ≤1kb tandem genes, and the >1kb averages are lower than ≤1kb classes.

Chapter 3

Positional analysis of divergence relative to nucleosome-associated sequences

Background

Chromatin proteins compact, organize, and regulate access to DNA, directing and participating in the fundamental processes of DNA replication, recombination, transcription, repair and chromosome segregation. Facilitating or inhibiting access to DNA relies on the packaging of DNA around the histone octamer through extensive DNA-histone associations [1-3]. Nucleosomal arrays compact further into higher order chromatin fibers, through the binding of histone H1 and its interactions with linker DNA [4]. Although histone variants have evolved novel functions, the 4 canonical histones, which serve as the first order of DNA compaction and organization in the nucleus, are among the most conserved proteins known [1,5,6].

The winding of DNA around nucleosomes is thought to prevent inappropriate access of sequence-specific binding factors. In this way, nucleosome occupancy plays a role in gene regulation through translational positioning, occluding or allowing binding by regulators, particularly in promoters [7-9]. In some instances, chromatin remodeling factors, targeted by other regulatory proteins and histone modifications, are necessary to evict and reposition nucleosomes during gene activation [7,9,10]. However, inflexible poly dA:dT tracts in many promoters have been shown to intrinsically exclude nucleosomes at the sequence level and restrict translational positioning of nucleosomes in these regions [8,11-15]. (Find TF position relative to nucleosome reference)

In addition to exclusion by inhibitory dA:dT tracts, the histone octamer displays sequence-dependent binding affinities, due to large differences in the flexibility of various nucleotide stretches which influence their ability to wrap around the core [3,12,16-18]. Nucleosome-associated sequences are marked by the periodic association of AT and GC-rich motifs in the bends of the major and minor grooves respectively [16]. Reconstitution work has demonstrated that sequences with a 10bp periodicity have the ability to position nucleosomes and that alternating AT and GC base composition is most favorable for nucleosome formation [17-19]. Crystal structure studies support a role for base composition in smooth bending of DNA in the nucleosome [3]. As the DNA wraps, bending in the minor-groove is promoted by changes in base pair orientation associated with GC, GG/CC, and AG/CT dinucleotides [3]. Flexible TA steps were found associated only with regions of major groove curvature [3]. Large scale sequencing of nucleosomal DNA confirmed such 10-11bp periodic associations with certain dinucleotides *in vivo* [16,20,21]. This property is common across species, with TT/AA and TA dinucleotides showing strongest periodicity [16].

Although these patterns were at first construed as a genomic code for nucleosome positioning, *in vitro* studies have only partially upheld the intrinsic role of DNA sequence in nucleosome positioning [12,22,23]. Reconstitution assays have demonstrated depletion in promoter regions, which contain poly dA:dT tracts [12,22,23]. Similar anti-nucleosomal properties are thought to exclude nucleosomes in linker regions. Histone-DNA interactions do govern rotational positioning, as evidenced by the periodic alignment of certain dinucleotides after *in vitro* assembly [22]. The 10-bp periodicity is markedly lower for nucleosomes assembled *in vivo*, suggesting that chromatin remodeling factors may actually decrease the influence of sequence on rotational positioning [22]. While these observations support an intrinsic role for DNA sequence in nucleosome exclusion and rotational positioning, on the whole, translational positioning outside of promoters does not appear to be strongly "encoded" at the sequence level [12,22].

This fact is reflected in the limited success of modeling aimed at predicting nucleosome occupancy from DNA sequence. Some approaches rely on dinucleotide periodicity or nucleosome motifs derived from experimental data, while others have had equal success predicting nucleosome occupancy based only on base compositional differences between bound and linker regions and the inhibitory effects of poly dA:dT tracts [12,13,16,24,25]. Structural approaches based on extrapolation of the biophysical properties of DNA bound to nucleosomes, in particular the deformation of DNA associated with wrapping, offer the benefit of avoiding the sampling bias inherent to empirical studies introduced during micrococcal nuclease (MNase) digestion and nucleosome isolation [26-28].

The predictive ability of these models are weak and varied. Their insufficiency stems in large part from the activity of chromatin remodeling factors and the dynamic and large scale changes in nucleosome positioning during transcription [22,29]. Indeed, the strongest signals for nucleosome positioning are found across the gene bodies of actively transcribed genes, where nucleosomes are phased with varying spacing [20,21,30-32]. The pattern in transcribed regions arises from restrictions on nucleosome packing imposed by the binding and initiation of RNA polymerase II (Pol II) at promoters, termed statistical positioning [22,29,32,33]. These strongly periodic patterns do not represent highly stable nucleosomes; the turnover in these regions is high as histones or entire nucleosomes are displaced as part of the progression of Pol II [10,34,35]. This raises the question of what sequence-encoded positioning signals would be expected in this subset of nucleosomes. Similarly, changes in chromatin associated with development and differentiation or heterochromatinization, through associations of higher order chromatin proteins, could interact with or even completely override local DNA sequence preferences for nucleosome assembly and positioning.

Although DNA sequence does not intrinsically reconstitute *in vivo* nucleosome translational positions, the extensive and ubiquitous association of DNA with histone proteins and the energy required for nucleosome packaging and remodeling suggests the potential for selective forces to act on nucleosome favoring or disfavoring sequences. The strength and functional consequences of such selection would likely differ by region (for example coding versus non-coding) and by function (highly expressed housekeeping genes versus those regulated by Polycomb group, Sir2, or other higher order silencing). For instance, across exons, the phased positioning of nucleosomes may interact with and influence codon usage. Genomic regions which do not undergo extensive remodeling or are not strictly regulated by specialized chromatin proteins may be under reduced selective pressure. The translational positioning of a nucleosome over a sequence that is energetically (or otherwise) less optimal does not necessarily remove the natural selection mediated through functional effects of sequence-nucleosome interactions. It simply confounds it with stronger mechanisms. Labile and strictly positioned nucleosome may have other sequence characteristics, yet to be discovered.

The functional consequences of interactions between sequence and nucleosome occupancy on molecular evolutionary divergence and population genomic polymorphism could well reflect a wide range of mechanisms from the regulation of transcription to DNA replication and DNA repair. The ubiquitous and fundamental role of this particular DNA-protein interaction argues for broad biological relevance and potentially a large and consistent pattern. However, disentangling the impact of disparate selective and mutational mechanisms as well as potential experimental biases depends on careful and quantitative analysis of these rich new data, that is still at an early phase. Already exciting patterns are evident and the prospects of rigorous inference look good. Variation in divergence rates are expected in linker and nucleosomal DNA based on local variation in base composition. Further, the physical association of DNA with the histone ocatmer may protect or otherwise influence mutation or gene conversion rates outside and across the bound region. Such a hypothesis might explain the observation that in coding regions, >1bp indels are found preferentially in linker regions, whereas transitions and transversions are more frequent in nucleosome occupied regions [36].

The repositioned mutation (RM) test has been used to infer natural selection on sequence-dependent nucleosome positioning [37]. In the test, substitutions, insertions, and deletions observed between homologous sequences in nucleosome occupied regions are randomly repositioned. In this way, the test only considers the role of the nucleosome-bound DNA and excludes the linker. New hypothetical sequences are used to assess the effect of random (neutral) changes on nucleosome formation potential, as assessed by various models. Repositioned mutations should disrupt this potential. Purifying or positive selection are inferred when the potential of homologous sequences is either significantly greater or smaller than those generated by the RM process. Although Babbitt et al found evidence for selection for nucleosome positioning using the RM test, it was applied only to limited regions surrounding promoters in *S. cerevisiae* and was strongly influenced by indels [37].

The interpretation of divergence between species across nucleosome-bound regions is greatly aided by population genetic information. The divergence between taxa (usually over millions of years) is a product of dynamics that took place and are taking place in natural populations. A tight and quantitative relationship between observations on these two timescales can often be predicted under the assumption that similar mechanisms have been at work. The frequency spectrum in random samples from natural populations of particular SNPs in specific positions can be tested against the predictions of simple models incorporating genetic drift and others with more elaborate demographic structure. Indeed the patterns of divergence, polymorphism and the frequency spectra can be contrasted among different sites within the nucleosome and linker regions. If the patterns of polymorphism and their relationship to divergence are inconsistent with reasonable models and parameters, it is possible to draw statistical inferences of the broad evolutionary impact of functional consequences of sequence context with nucleosome positioning.

For example, if SNPs at particular positions in nucleosome sequences exhibit a large skew toward rare, “unpreferred” alleles, while at other positions the frequency spectrum is that predicted by selectively neutral genetic drift (or even an excess of common alleles) then the parsimonious interpetation is that these SNP are under purifying selections. Perhaps SNPs with the high frequencies can be interpreted as evolving under a form of balancing selection. The framework for such analyses is well established in population genomics, where it has been honed on the analyses of synonymous and nonsynonymous coding sequences [MK, Tajima’s D, Poisson random fields - Busamante].

We have isolated and sequenced nucleosomal fragments from stage 5 embryos and modified nucleosomes from adult flies. These data can be grouped by regional or functional associations (X v. autosome, high v. low expressed genes, 5’ proximal v 3’ proximal within genes, PcG regulated v. total intergenic, intron v. exon, heterochromatic v. euchromatic). Our interest is in identifying differences in base composition across and the substitutional patterns within and surrounding these categories of nucleosome occupied regions with a goal of identifying the roles of selection and mutational bias in generating the observed differences. The work presented here uses divergence between species to identify distinct patterns of nucleotide substitution and a direct comparison of these on the *melanogaster* lineage to those on the *simulans* lineage. Future work will investigate the patterns of sequence polymorphism (in comparison to divergence) to more rigorously test the predictions of causative models, e.g., selection or mutational bias.

Results

Distributions of fragment size indicate regional differences chromatin accessibility

An open or condensed chromatin state can be measured, to some extent, by accessibility of underlying DNA to nuclease. Exposure of DNA during remodeling makes promoters and other dynamic regions, including *cis*-regulatory elements, sensitive to digestion by certain enzymes, classically DNase I [38-40]. MNase digestion patterns can also reflect underlying dynamics or structural differences in chromatin, as observed in the differential digestion of the transcribed and untranscribed regions of the Drosophila heat shock locus [41].

Digestion of chromatin from stage 5 embryos and adult males produced a large range of fragment sizes, from ~50-200bp (Fig. 1A,B). Salt concentrations impact both stability and release of nucleosomes during isolation [42-44]. Our protocol uses physiological salt levels and no crosslinking. This potentially causes the loss of unstable H3.3-containing nucleosomes [42-44]. The distributions of bulk nucleosomes form both sources are remarkably similar, with peaks at ~60, 100, 135 and 170bp (Fig. 1A,B). This indicates that our chromatin sources are not overdigested and that a large portion of fragments still contain linker DNA. However, the significant number of <147bp fragments indicates that the nuclease activity was not restricted to linker regions and that some degree of overdigestion occurred. Yield of 147bp fragments was increased after size-selection on the embryo sample, but a persistent <50bp peak confirms the bias of the short read sequencing toward shorter fragments (Fig. 1C).

Size distributions for fragments associated with H3K4me3 or H3K27me3 show fewer in the >147bp range, indicating that these regions may be more accessible to the enzyme (Fig. 1D, E). H3K4me3 marks the 5' regions of active genes, which are known to be sensitive to nuclease digestion, and regions regulated by the trithorax group (TrxG) proteins [45-49]. H3K27me3 is generally repressive and is bound by Polycomb during differentiation, inducing silencing of key developmental transcription factors, and occurs in regions overlapping with H3K4me3 in pluripotent mammalian cells [46-48,50].

MNase produces strong sequence-dependent bias at fragment ends and impacts sampling

The bias introduced by MNase digestion presents a significant and confounding challenge to the study of nucleosome-associated sequences and linker regions [20,51]. Biochemical analyses describe the rate of cleavage by the enzyme as 30 times greater 5' of A or T than at G or C [20,51]. This has been substantiated by more recent large scale sequencing projects [13,20]. Our analysis confirms MNase preference for 5‘ cleavage at A or T and a strong cut site bias between AA, TT, and AT dinucleotides (Fig. 2). The frequency of AT at the interior base is above 0.95 and the frequency 5' of the cut is above 0.85. We also note that the base second from the cut site (interior) shows a strong G bias, with a frequency of 0.45 in intergenic 147bp fragments, in agreement with early biochemical studies reporting rapid cleavage of the site CTAG (Fig. 2A,B,C,D,E) [51]. The enzyme generated bias can generally be described as a preference for TTG, ATG and AAG, with cutting between AT bases. Our analysis of nucleotide and dinucleotide frequencies surrounding the cut sites of various fragment size classes (representing over and underdigested material) indicate that the bias is relatively independent of size (data not shown).

The sequence specificity of the enzyme is reportedly the same for naked DNA and reconstituted nucleosomal DNA [51]. During digestion, the nuclease most likely cleaves the DNA at a preferred sequence near the entry/exit point of the nucleosome. This may explain the heterogeneity of fragment sizes we observe. Early studies warn that this method is "prone to serious artifacts" and "In particular it would seem impossible to determine precisely the positions of nucleosomes on DNA using this enzyme," yet these data are often treated as random in nucleosome mapping and modeling studies [51]. The inherent bias in the genomic sampling complicates comparisons across and within species, particularly at nucleosome entry-exit points. Interior sequences are expected to be more independent of the MNase sequence bias [51].

Nucleotide and dinucleotide periodicity across and surrounding nucleosomal fragments

Models of sequence-encoded positioning of nucleosomes include higher GC content across nucleosomes, AT rich linker regions, and a ~10-11bp alternating TT/AA/TA/AT and GC dinucleotide periodicity. If DNA in the nucleus is generally associated with nucleosomes, and linker lengths are not highly variable, then nucleosome favoring and disfavoring (linkers) patterns of nucleotide composition should be visible beyond the nucleosome boundary. We began by examining base composition surrounding 147bp fragments isolated from embryos. Nucleosome sequences are, as expected, higher in GC content. A periodic fluctuation in the AT and GC frequencies with peak to peak distance of ~200bp extends beyond intergenic (Fig. 3A) and CDS (Fig. 3B) -derived nucleosomes. Although less striking, a similar pattern is visible surrounding H3K27me3-marked intergenic nucleosomes (Fig. 3C). Nucleosomes containing H3K27me3 are higher in GC content and are found in slightly more GC rich non-coding regions (Fig. 3C). This type of regional base composition is predicted to interact with translational positioning [12,22-24]. These patterns suggest selection for alternating nucleosome favoring and disfavoring sequences with a linker length of ~50bp. Alternatively this pattern might be attributed to significant variation in mutational or gene conversion bias relative to nucleosome occupancy in *melanogaster*.

On a finer scale, nucleosomal DNA fragments display dinucleotide periodicities thought to direct rotational positioning [16,20-22]. Either this periodicity is a property of the underlying genomic sequence, in which case cleavage site bias might produce similar periodicities in fragments from naked genomic DNA or over- and underdigested chromatin, or it is a property of nucleosome-bound regions (and not linkers or unbound regions). We find periodic dinucleotide distributions across nucleosomal sequences from Drosophila embryos (Fig. 4). Interestingly AA, TT, and GC show continued periodicity outside of nucleosome bound regions (Fig. 4). Such signal has been interpreted as evidence of quantized linker lengths in *S. cerevisiae* [52]. The significantly stronger pattern in our data appears to be a continuation of the internal periodicity across the nucleosome, as it is on the order of 10bp and is in register with those observed across the nucleosome (Fig. 4).

Although we see evidence that dinucleotide periodicities extend beyond the 147bp bounds of the nucleosome-bound regions, we do not find that they are a general property underlying the genomic sequence and potentially generated by enzymatic bias. We also examined dinucleotide frequencies for fragments shorter or longer than 147bp by 10-11bp*n. We find these retain dinucleotide periodicity, whereas those fragments longer or shorter by 5-6bp*n do not (data not shown). This also suggests that overdigestion tends to occur on one strand, as previously reported for H2A.Z nucleosomes from human cells [53]. We conclude that dinucleotide periodicity is not generated by underlying genomic patterns interacting with MNase cleavage preferences.

We also considered the distribution of trinucleotide sequences across intergenic and CDS nucleosomes (Fig. 5). For nucleosomes in coding regions, a large number of GC rich trinucleotides show positional enrichment in the 0-45bp, near the entry/exit site, and external to the cleavage site in putative linker regions. The 0-45bp span on the nucleosome, and its symmetric counterpart at the other side of the dyad axis, is where the DNA interacts with H2A/H2B dimers and the H3 α-N helix and tail [3]. Many of these GC rich trinucleotides are codons known to show codon bias in *melanogaster* (Fig. 5, red stars). Given the difference in context, the sequence bias of the enzyme could introduce new bias in these regions. Perhaps coding regions surrounding the preferred ATG,TTG, and AAG sequences are GC rich. We do not see strong evidence for such a bias beyond +5bp into the nucleosome (Fig. 2A, Fig. 3). The matter could be addressed in the future by determining the GC content surrounding preferred cleavage sites in CDS regions.

Average divergence shows periodicity relative to nucleosome-derived sequences

If population polymorphisms in the sequence characteristics associated with nucleosome translational or rotational positioning have functional consequences, then there should be evidence of natural selection in the patterns of polymorphism and divergence. Of course, we must also entertain the alternative, but equally interesting hypothesis that these interactions lead to variation in the mutation and gene conversion biases. We began by examining patterns of average divergence between *melanogaster* and *D. yakuba* surrounding 147bp fragments (Fig. 6). A clear ~200bp periodicity extends beyond aligned fragments for nucleosomes in intergenic and intronic regions (Fig. 6 A,B). Intergenic sequences associated with H3K27me3 marked nucleosomes are much more highly conserved than average intergenic nucleosome-bound DNA (Fig. 6D). In addition, these nucleosomes show particularly strong flanking divergence periodicity.

These periodicities could reflect variation in mutation or gene conversion rates or functional constraints on nucleosome-associated and linker sequences, or in a mixture of the two of these two mechanisms. Regions of low divergence between *melanogaster* and *yakuba* correspond to regions of low AT and high GC (Fig. 3, Fig. 6). These base compositional fluctuations are less pronounced for H3K27me3 flanking regions where the periodicity of divergence is strongest (Fig. 3C, Fig. 6D). This suggests, at least, that these patterns are not driven solely by mutational biases related to base composition. It does not, however, argue against it it entirely or eliminate other potential sources of mutational bias, such as the influence of nucleosome occupancy itself on mutation rates. It is possible there is a bias specific to H3K27me3 enriched regions or higher order structures associated with this modification. Skewed base frequencies at the cleavage sites create extreme and biased results.

A periodicity in divergence of ~10bp is obvious across 147bp nucleosomal fragments, particularly for intergenic and intron-derived sequences (Fig. 6E,F). This periodicity aligns roughly with peaks of GC enrichment (Fig. 7 A) In contrast, only the 3bp periodicity associated change at the 3rd base of codons is evident across nucleosomes from coding regions (Fig. 6G). Divergence across CDS nucleosomes decreases toward the the entry/exit points from a peak centered at the dyad axis (Fig. 6G). Ignoring the spikes at the edges which are influenced by base compositional bias at cleavage sites, a similar decrease is evident for intergenic sequences. However, intergenic nucleosomes show slightly decreased and no periodicity on average in regions surrounding the dyad axis. Introns do not show a curvature in average divergence relative to nucleosome occupancy (Fig. 6F). The average divergence across H3K27me3 nucleosomes shows less evidence of periodicity. This may be due to lower coverage (Fig. 6H). It could also reflect less reliance on or influence of rotational positioning signals in regions bound by PcG proteins.

Lineage-specific and base-pair specific divergence patterns reveal variation in interactions between nucleosomes and molecular evolution.

The averaging across both the *melanogaster* and *yakuba* lineages obscures lineage-specific differences in rates at individual positions. The ancestral state of say the *melanogaster* lineage can be inferred by comparing the *yakuba* and *erecta* states with that in *simulans*. If either agrees then it is reasonable and accurate to assume that the ancestor of the *melanogaster* lineage was the same, since *melanogaster* and *simulans* have a much more recent common ancestor. Thus “polarized” evolutionary divergence can be identified at sites where *melanogaster* exhibits a different state. A similar algorithm can be applied to estimate the rates of polarized divergence on the *simulans* lineage. Figure 1A presents the observed pattern of polarized divergence within and flanking the nucleosome sequences in intergenic regions. Although the magnitude of divergence is much less than that between *yakuba* and *melanogaster*, a 10bp periodicity with peaks over the GC-rich regions is evident, at least within the nucleosome. And the broader pattern reveals conservation of sequence flanking the two base pairs at the entry/exit site, which exhibits a distinct spike. The average divergence on the *melanogaster* lineage within these 147bp nucleosome fragments is lower than that observed in flanking sequence greater than 40bp distal. These observations are consistent with those for the *melanogaster-yakuba* divergence (see Fig. 6A and 6B) and suggest that this 147bp set of nucleosome associated sequences are flanked by a region of 200-300 bp with slightly elevated divergence.

The comparable, polarized divergence in coding regions exhibits a similar broad pattern, but the immediate flanking base pairs now show more divergence and the proximal interior sites are somewhat more conserved. This difference around the entry/exit site of the intergenic and coding regions no doubt reflects a bias arising from the interaction of sequence specificity of the MNase with distribution of GC rich trinucleotides (see Fig. 5). A 10bp period in the average polarized divergence is not evident in these sequences from coding regions.

Averaging over different substitutional paths (e.g., AT->GC and GC->AT) can also obscure important features. Given the commonly observed 10bp periodicity of the GC rich (and AT rich) sequences in nucleosomes, it is reasonable to consider AT->GC and GC->AT divergence separately. Figure 8B shows substitutions from an GC ancestor to AT in *melanogaster* (GC->AT) are much more periodic than the total divergence (Fig. 8A). Similarly the pattern of AT->GC reflect this same periodicity in divergence, as well as the strong sequence bias of MNase cutting at the putative entry/exit (Fig. 8C). The rate of AT->GC is highest in GC peaked regions

(Fig. 8B) and the rate of GC->AT is higher in regions with higher AT frequencies (Fig. 8C). Although no periodicity is evident in the overall *melanogaster* lineage divergence in CDS regions (Fig. 8D), separation based on the direction of base changes reveals periodic patterns qualitatively very similar to those in intergenic regions (Fig. 8E,F). AT->AT and GC->GC changes show some positional relationship relative to nucleosomal sequences, in that the boundaries are evident, but they do not show periodicity across the nucleosome (Fig. 8G-J).

The distinct patterns of divergence in the few base pairs on either side of the boundary of these nucleosomal sequences undoubtedly reflect ascertainment bias associated with the cutting preference of MNase. Nucleosomal sequences the diverged from GC in the common ancestor with *simulans* to AT in *melanogaster* are obviously more likely to be in this collection of sequences than those that remained GC or diverged from AT to GC. On the other hand the internal base pairs are much less likely to directly influence the MNase cutting at positions 0 and 147. Indeed it seems likely that the periodicity observed arises from interactions with the nucleosome itself. From that it follows that periodicity in the AT->GC and GC->AT divergence most likely derives from mechanisms involving sequence specific interactions with local components of the positioned nucleosome.

Beyond the clear periodicities in the divergence of both GC->AT and AT->GC, it is (as Table 1 shows) the difference in their overall magnitude on the *melanogaster* lineage that is most striking. Comparing the ratios of composition and divergence afford the most direct view. The average ratio of base composition (AT/GC) across intergenic nucleosomes (Fig. 9A) is 1.38, yet the ratio of GC-> AT to AT->GC in these regions is ~2 (Fig. 9B, Table 1). Substantially higher rates of GC->AT than AT->GC are also seen in CDS regions (Table 1), where the ratio of frequencies (AT/GC) is much closer to 1 (Fig. 9C,D). Thus this 2 fold difference in the rates of substitution on the *melanogaster* lineage is not limited to AT rich regions and is most simply interpreted as evidence of a large scale change in mutational (or gene conversion) bias. If so, the base composition in the *melanogaster* genome should be changing relative to its sister species and their most recent common ancestor.

Analysis of nucleosomal fragments from *melanogaster* is complicated by sampling bias stemming from both MNase sequence bias and overall accessibility to the enzyme. Considering the changes on the *simulans* lineage in these same regions is an approach that may be robust against this bias. These regions are assumed to be free of directional bias in *simulans*, as they diverged independently of the *melanogaster* lineage. Polarized divergence in the *simulans* lineage relative to mapped *melanogaster* nucleosomal fragments is lower but shows a very similar pattern, suggesting that nucleosome occupancy and associated sequence changes, whether due to mutational bias or selection, are conserved (Fig. 10). Large scale patterns of directional substitutions (GC->AT and AT->GC) in *simulans* for the regions homologous to the *melanogaster* nucleosomal sequences are largely consistent with those in *melanogaster* (Fig. 11). Since these patterns of divergence on the *simulans* lineage relative to *melanogaster* nucleosomal fragments occurred independently and in the same direction on both lineages, implying that, whether the source is mutational bias relative to nucleosome occupancy or functional selection for sequences that maintain translational positioning, the average positions of nucleosomes are most likely conserved and exert a characteristic force on the patterns of molecular evolution.

At a fine scale, a ~10bp periodicity in divergence on the *simulans* lineage is evident across intergenic regions homologous to those associated with nucleosomes in *melanogaster* (Fig. 12A,B). Peaks in the average divergence on the *simulans* lineage are in register with those in *melanogaster*, centered over local peaks of higher GC content across the nucleosomal fragments (Fig. 12B). Base compositional differences reflect the accelerated GC->AT rate in *melanogaster* (Fig. 12). In CDS regions, divergence in *simulans* does not show large differences from *melanogaster* (Fig. 12C,D). When changes along the lineages are separated by direction (AT->GC and GC->AT), differences between the two lineages are evident (Fig. 13). In intergenic (Fig 13I,J) and CDS (Fig 13M,N) regions, peaks in divergence in *melanogaster* follow local peaks in base composition; GC->AT peaks in register with GC peaks, AT->GC peaks in register with AT peaks. In intergenic regions from *simulans*, peaks in both directions align with local peaks of higher GC content (Fig. 13K,L). This distinguishes putative nucleosomal regions from the patterns observed on a large scale (Fig. 11). GC->AT changes on the *simulans* lineage show decreased periodicity (Fig. 13K), while AT->GC looks quite similar to *melanogaster* (Fig. 13J,L) The increased noise in the *simulans* might be expected nucleosome occupancy has changed subtly between the species. The periodicity in direction of substitutions observed in *melanogaster* CDS regions is weak if not absent in *simulans*, although the curvature of rates across the region suggests these regions are not random and still show a relationship to nucleosome occupancy in *simulans*. Slight changes in positioning (<10bp) could disrupt the this pattern. It is also possible that the increased GC->AT in *melanogaster* generates a clearer signal in this type of change. GC->AT rates are very low in *simulans* CDS regions and show lower dynamic range (Fig. 13O). The rate of 3rd site substitution differs along the two lineages. In *simulans*, rates are substantially higher for AT->GC than GC->AT (Fig. 13O,P). The reverse is true for *melanogaster*, where GC->AT is about 2 fold higher, similar to intergenic regions (Fig. 13M,N).

DISCUSSION

Although the biophysical interactions between the histone octamer and DNA sequence are not generally strong enough to stably position nucleosomes *in vitro* or to predict the occupancy of a given sequence, they must contribute to the overall energetic requirements for assembling, remodeling, and maintaining chromatin [12,22,23]. *In vivo* and *in vitro* studies suggest that Poly dA/dT tracts, GC content, and short alternating AT and GC rich motifs impact nucleosome formation, positioning, and stability [3,11,12,15-18]. If there is “encoding” for nucleosome positions it may not only be for stable but perhaps slidable or unstable nucleosomes as well. The periodic occurrence we observe of nucleosome favoring and disfavoring sequences in the genome implies either a functional advantage of such arrangement of base composition or a significant mutational bias relative to nucleosomal fragments. We note that this periodic base composition and the directional dinculeotide changes necessary to maintain it are found in both the *melanogaster* and *simulans* lineages, in spite of significant differences in substitution.

On the broad scale, increased level of substitution of AT in *melanogaster* does not appear to show a positional relationship to nucleosomes (see Fig. 11). Although overall AT is increased in *melanogaster*, the shape of AT and GC periodicities are fairly similar to *simulans*. The substantial GC->AT in *melanogaster* does not appear to be associated with increases of AT in either linker (as judged by conservation of large scale periodicity) or nucleosomal sequences. Instead, AT frequency is increased in all regions. We do, however, see that GC->AT changes in nucleosome-derived sequences on the *melanogaster* lineage tend to occur at local peaks of AT, which differs from the pattern in *simulans*. Clearly more than one mechanism will be needed to accommodate these observations.

In CDS regions, third site substitutions in *melanogaster* have a positional relationship to occupancy on the nucleosome. Although this could be a mutational bias, these rates follow the known preferences that govern rotational positioning of DNA on the nucleosome. *simulans* regions homologous to nucleosome-derived sequences from *melanogaster* do not display periodicity of base changes in CDS regions. This could be the result of differential nucleosome positioning across these regions in *simulans*, since this periodicity would be sensitive to shifts of only a few base pairs.

Our analyses shows extended dinucleotide periodicity, outside of nucleosome-bound sequences. While some nucleosomes are stably positioned in certain cell types and during specific parts of the cell cycle, the majority of nucleosomes are not static. Any preferences that might shape nucleosome-associated sequences (or mutational biases associated with nucleosome occupancy) could also influence flanking sequences. These local extensions dissipate at ~40bp, which is very close to the linker length of ~50bp suggested by nucleotide frequency periodicity. This extended periodicity may facilitate 10bp sliding or play a role in determining linker length **[52]**. Alternatively, if mutation plays a role in generating dinucleotide patterns, then small scale local variation in translational positioning could generate extended periodicities

GC content and AT->GC changes, though subtle, are higher closer to the DNA entry/exit point of nucleosomes isolated from CDS regions in *melanogaster*. The 0-45bp span on the nucleosome where GC is enriched in CDS fragments and its symmetric counterpart at the other side of the dyad axis are where DNA interacts with H2A/H2B dimers and the H3 α-N helix and tail [3]. In coding regions, variants H2Av (H2A.Z) and H3.3 are substituted for canonical histones [10,35,42]. These variants substantially alter nucleosome stability and structure, increasing turnover of dimers and perhaps nucleosomes in these regions [10,35,42,43]. Regional enrichment of GC may interact with these variants or the extensive modifications of the H3 tail known to accompany transcription. Futher, more GC rich sequences may playing a role in destabilizing nucleosomes or making DNA more flexible and amenable to remodeling and sliding. It would be interesting to see if nucleotide composition across nucleosomes varies between high and low expressed genes.

Differences in GC content and positional differences in rates of substitution across the nucleosome may also play a role in codon bias. Codon bias favors GC bases at synonymous sites and it is most common in shorter, highly expressed genes [54]. These genes are expected to undergo extensive remodeling. Changes in sequence that lead to more labile nucleosomes or more flexible DNA sequence could be strongly favored in this context. Comparative positional mapping of biased and unbiased codons relative nucleosome phasing across genes could begin to determine if nucleosome positioning plays a role in codon bias.

Although the patterns of positional differences in divergence relative to nucleosome occupancy suggest that selection may act to preserve and shape DNA sequence in both intergenic and CDS regions, comparison to polymorphism data is necessary to determine the influence of mutational bias on the patterns we observe. Comparison of rates of divergence to frequency spectrum data across nucleosomes and within linker regions will us allow test hypotheses about the nature of the divergence and polymorphism and thus infer the mechanism(s) behind them. Additionally, while the power of this type of aggregation allows us to detect forces that might be too subtle to be seen at an individual site or in an naive laboratory experiment, a more rigorous investigation into the nature of individual regions is necessary to truly interpret the patterns we have observed thus far.

Figure 1

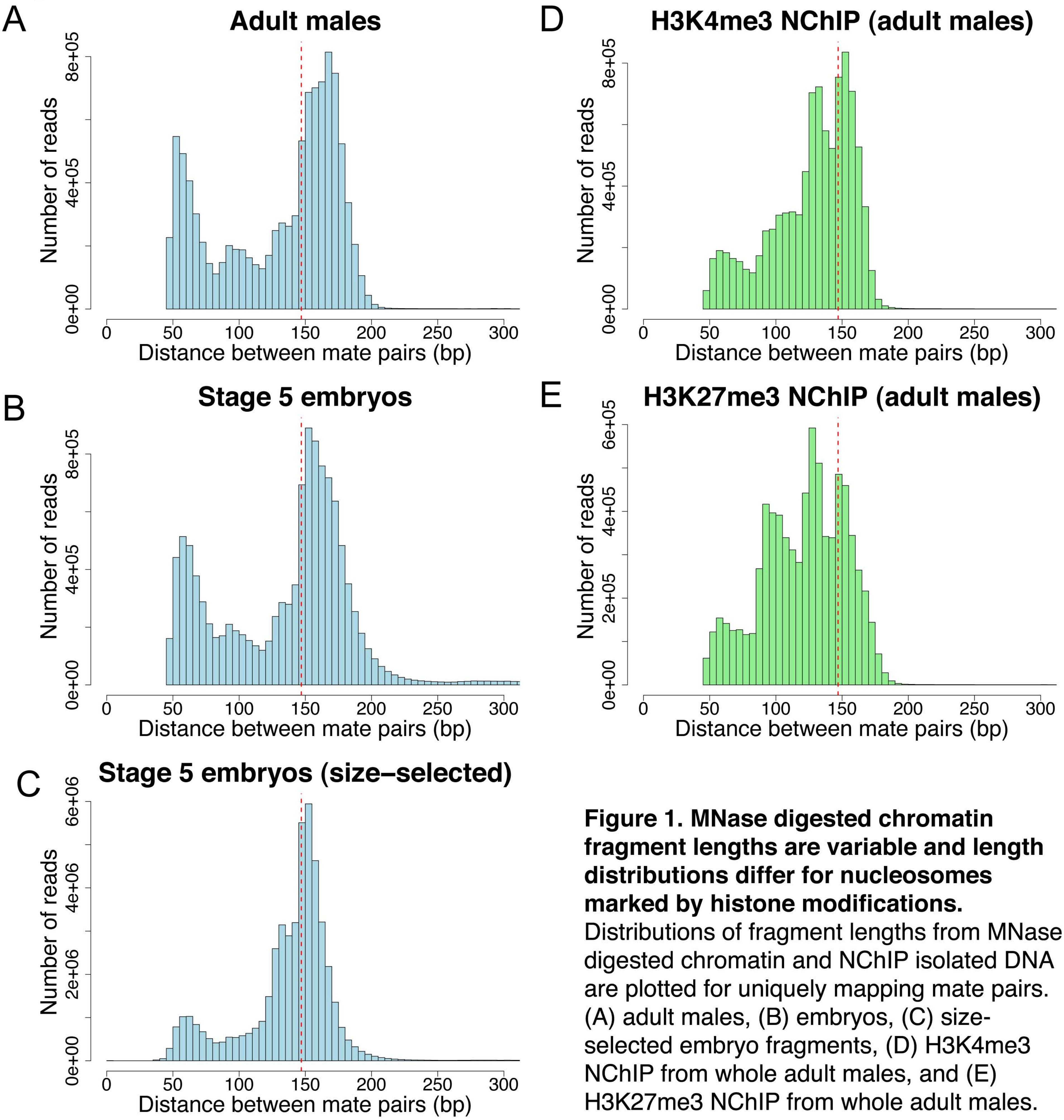


Figure 1. MNase digested chromatin fragment lengths are variable and length distributions differ for nucleosomes marked by histone modifications. Distributions of fragment lengths from MNase digested chromatin and NChIP isolated DNA are plotted for uniquely mapping mate pairs. (A) adult males, (B) embryos, (C) size-selected embryo fragments, (D) H3K4me3 NChIP from whole adult males, and (E) H3K27me3 NChIP from whole adult males.

Figure 2

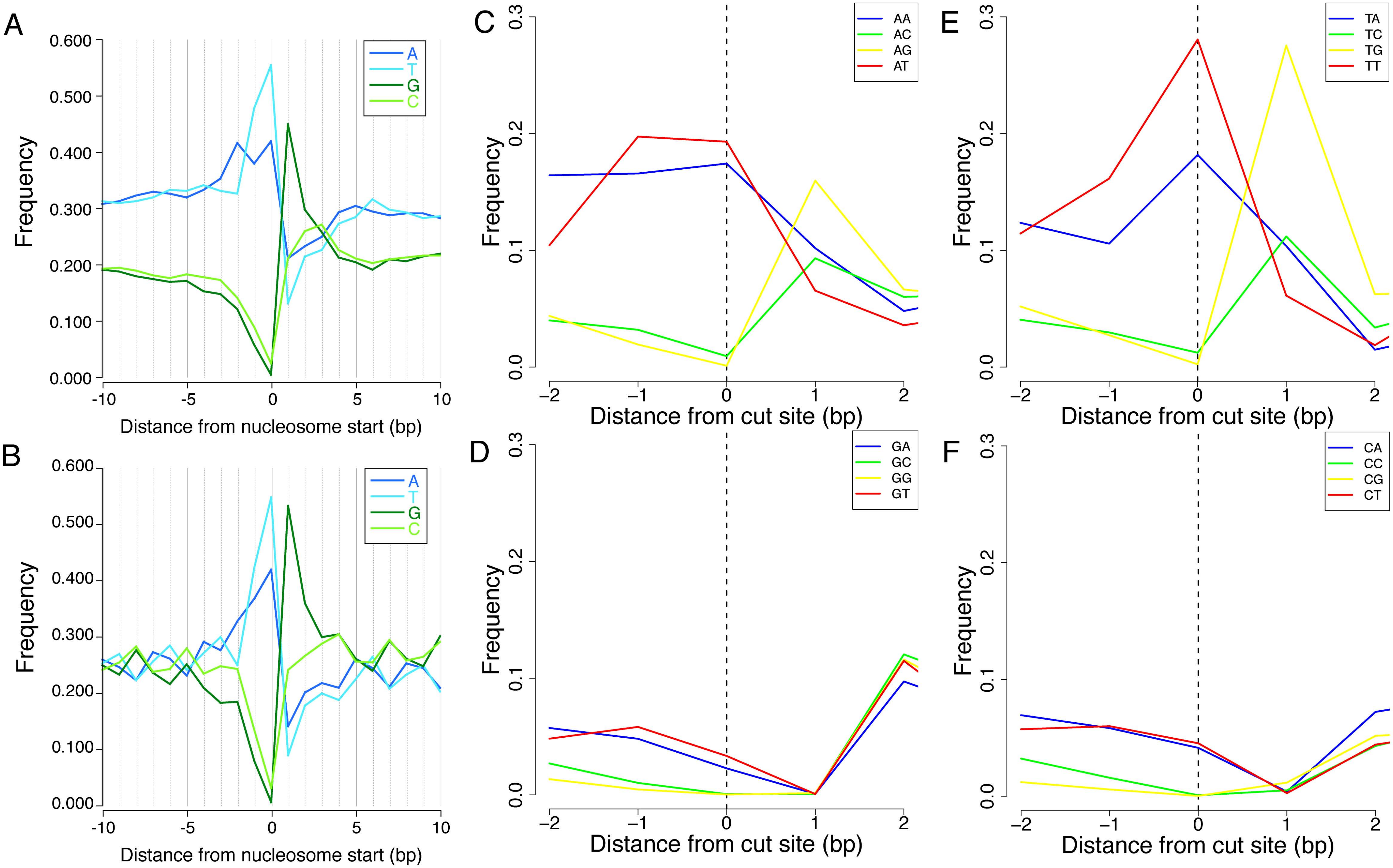


Figure 2. Nucleotide and dinucleotide frequencies surrounding MNase cleavage sites for uniquely mapping euchromatic 147bp fragments from stage 5 embryos. Nucleotide frequencies (5’->3’) surrounding cut sites of (A) intergenic or (B) CDS derived nucleosomes show similar biases in spite of different base compositions in these regions. Dinucleotide frequencies surrounding cut site for intergenic fragments shows a preference for cleavage between TT, AT, and AA and a bias toward G for the next interior base.

Figure 3

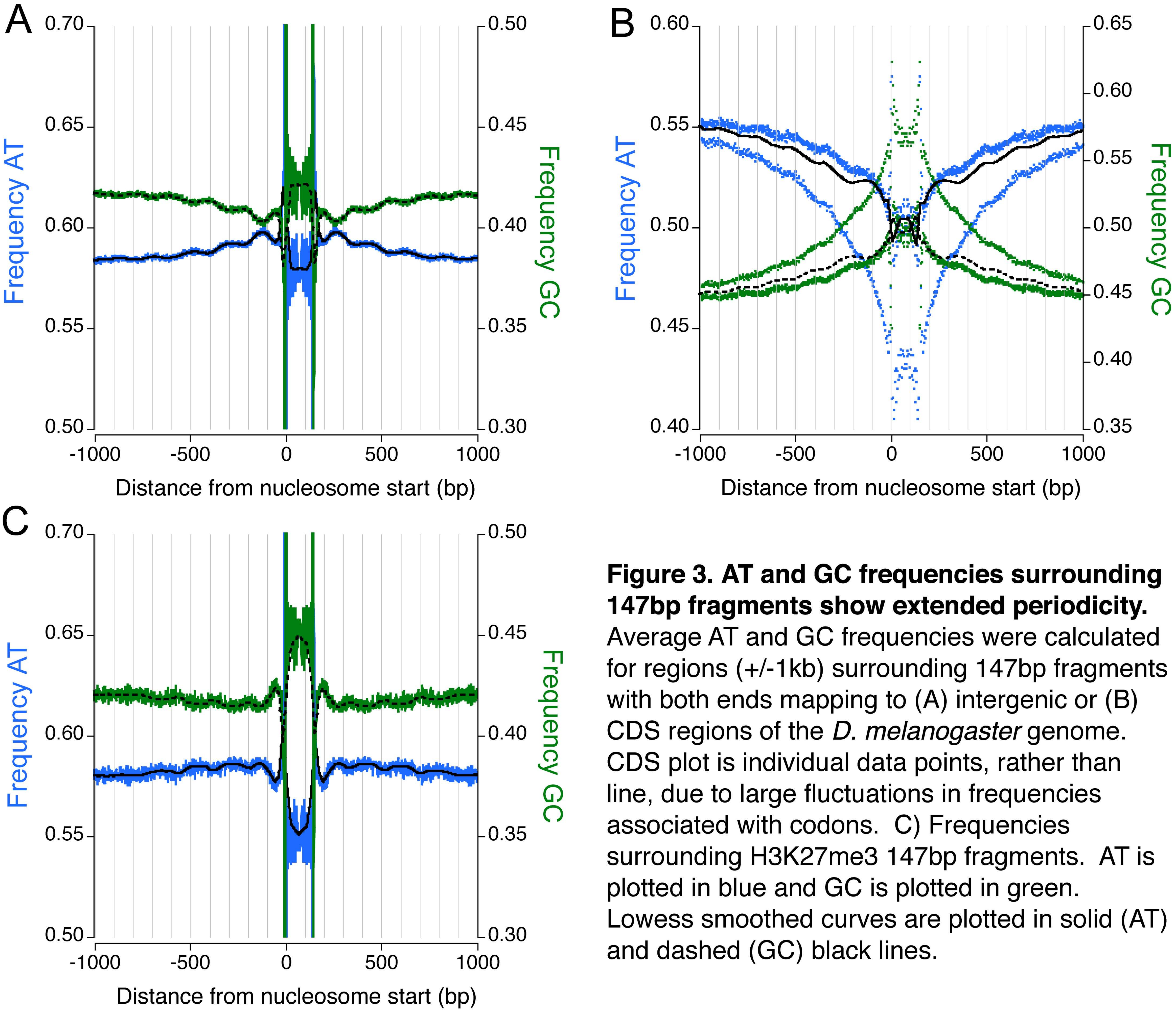


Figure 3. AT and GC frequencies surrounding 147bp fragments show extended periodicity. Average AT and GC frequencies were calculated for regions (+/-1kb) surrounding 147bp fragments with both ends mapping to (A) intergenic or (B) CDS regions of the *D. melanogaster* genome. CDS plot is individual data points, rather than line, due to large fluctuations in frequencies associated with codons. C) Frequencies surrounding H3K27me3 147bp fragments. AT is plotted in blue and GC is plotted in green. Lowess smoothed curves are plotted in solid (AT) and dashed (GC) black lines.

Figure 4

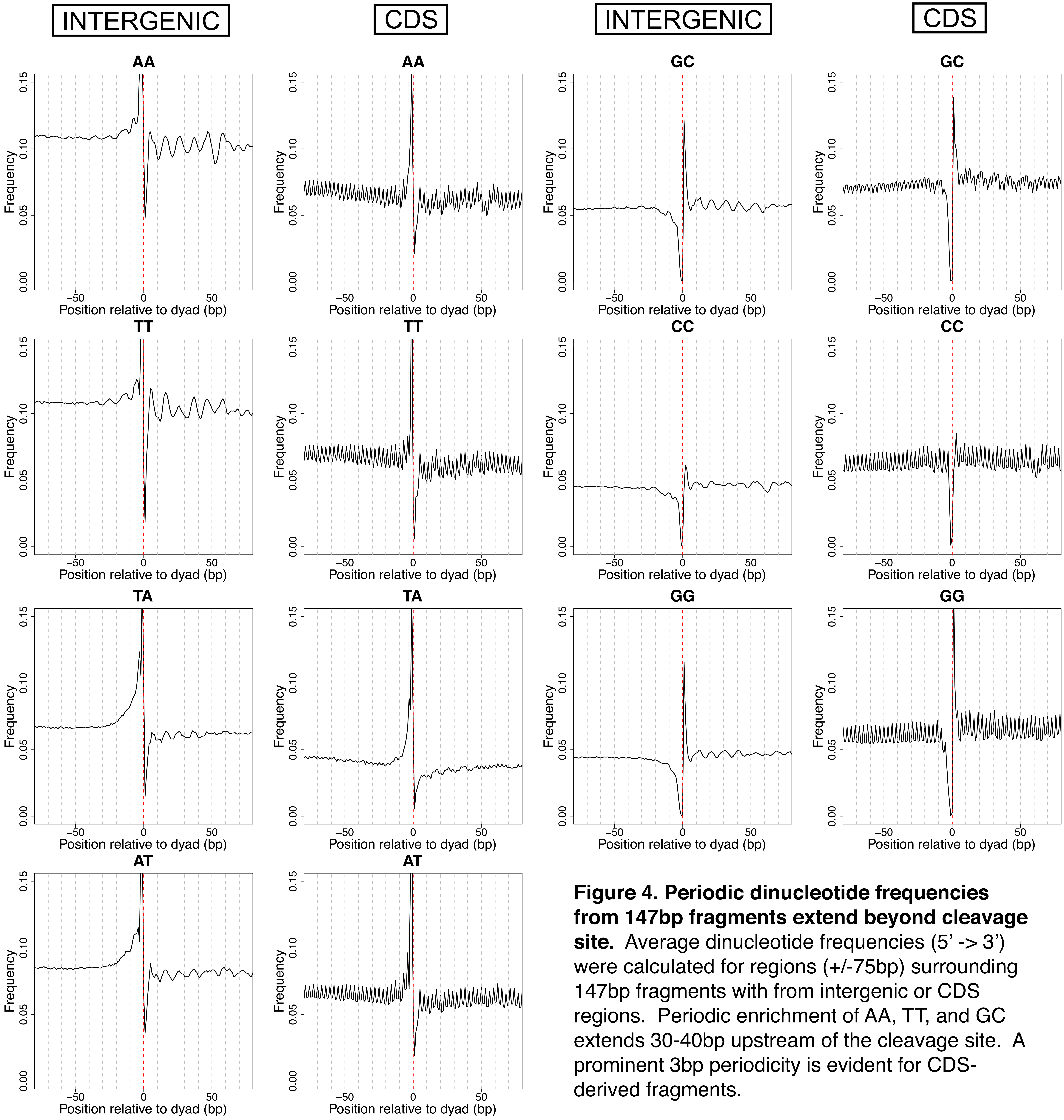


Figure 4. Periodic dinucleotide frequencies from 147bp fragments extend beyond cleavage site. Average dinucleotide frequencies (5’ -> 3’) were calculated for regions (+/-75bp) surrounding 147bp fragments with from intergenic or CDS regions. Periodic enrichment of AA, TT, and GC extends 30-40bp upstream of the cleavage site. A prominent 3bp periodicity is evident for CDS-derived fragments.

Figure 5

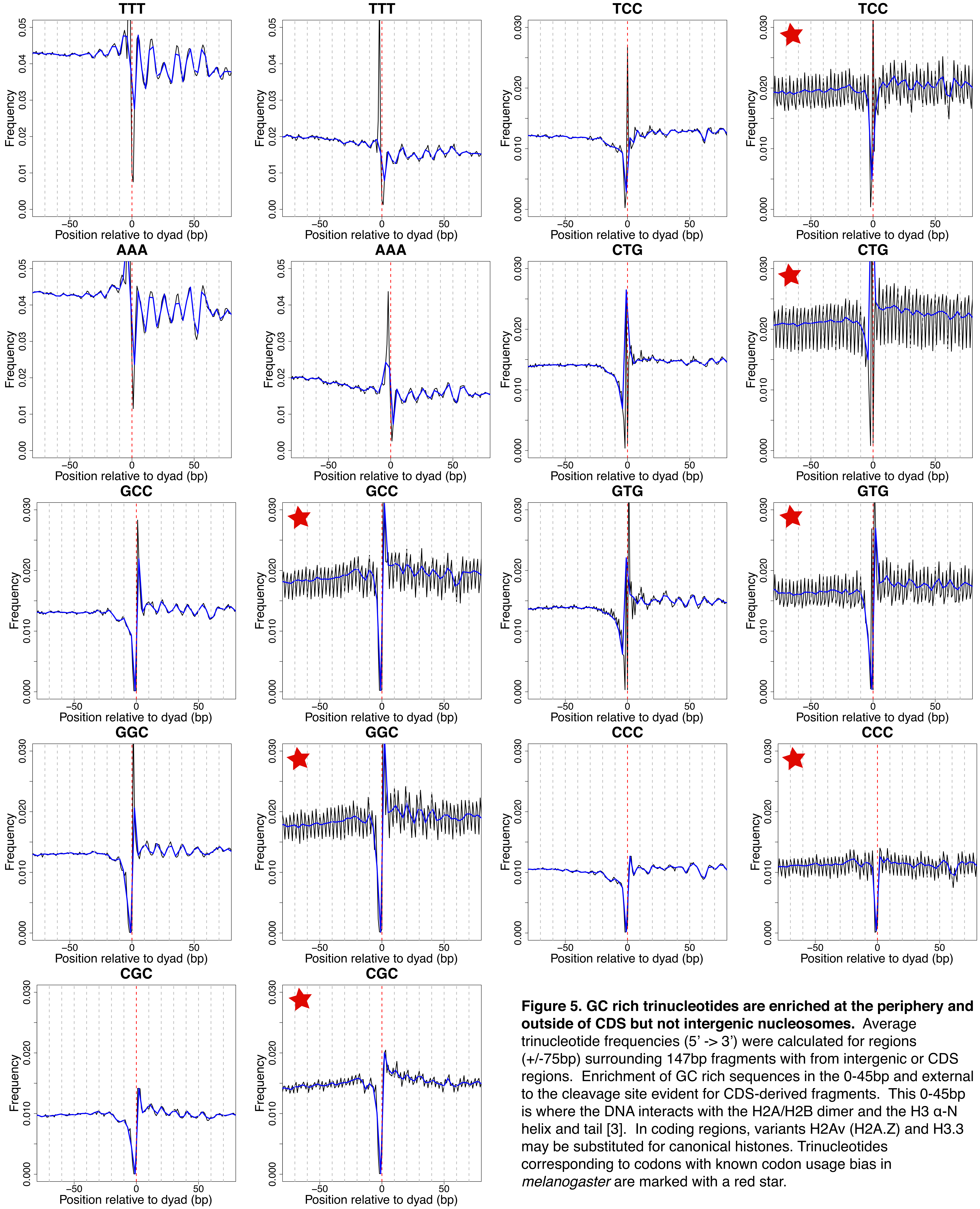


Figure 5. GC rich trinucleotides are enriched at the periphery and outside of CDS but not intergenic nucleosomes. Average trinucleotide frequencies (5' -> 3') were calculated for regions (+/-75bp) surrounding 147bp fragments with from intergenic or CDS regions. Enrichment of GC rich sequences in the 0-45bp and external to the cleavage site evident for CDS-derived fragments. This 0-45bp is where the DNA interacts with the H2A/H2B dimer and the H3 α-N helix and tail [3]. In coding regions, variants H2Av (H2A.Z) and H3.3 may be substituted for canonical histones. Trinucleotides corresponding to codons with known codon usage bias in *melanogaster* are marked with a red star.

Figure 6

INTERGENIC
INTRON
CDS
H3K27me3
A
B
C
D
E
F
G
H
mel_yak divergence
Distance from nucleosome start (bp)
0.070
0.065
0.060
0.055
0.050
0.045
0.040
0.068
0.066
0.064
0.062
0.058
0.056
0.054
0.052
-1000
-500
0
500
1000
-50
50
100
150
200

Figure 6. Average *melanogaster-yakuba* divergence shows large and small scale periodicity. Average divergence was calculated across and +/-1kb surrounding 147bp fragments mapping to (A) intergenic, (B) intron, and (C) CDS regions. Intergenic 147bp fragments from H3K27me3 marked nucleosomes is shown in (D). ~200bp peak to peak periodicity of divergence extends flanking regions. (E-H) Average divergence is plotted across nucleosomal regions +/- 50bp. ~10bp Periodic divergence can be seen in (E and F) intergenic and (F) intronic sequences. Divergence peaks are found at n*10bp, which corresponds to regions of GC enrichment.

Figure 7

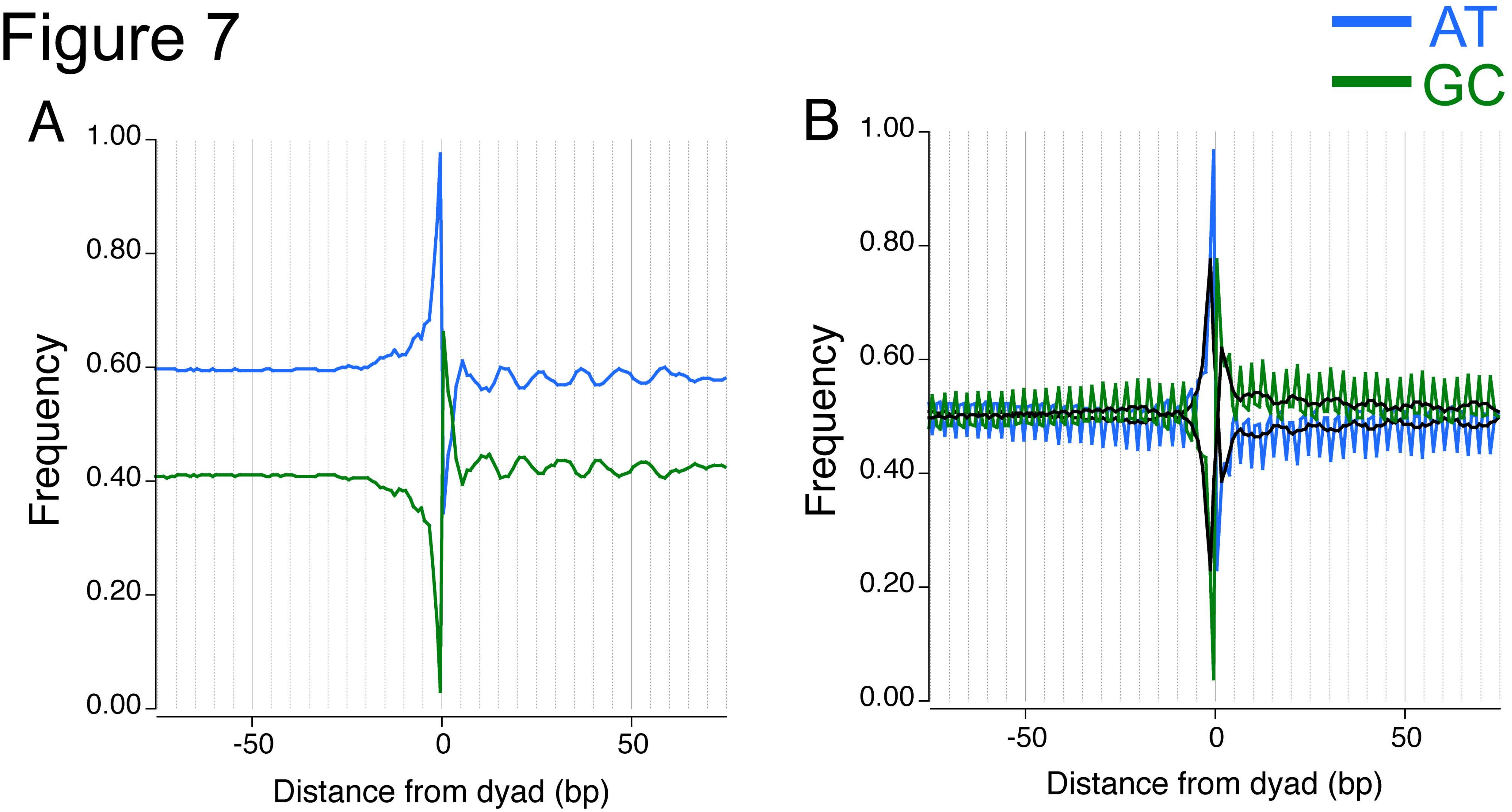


Figure 7. Average AT and GC frequencies across intergenic and CDS nucleosomal sequences. Frequencies of AT and GC across and 75bp upstream of 147bp fragments mapping to (A) intergenic or (B) CDS regions. A 10bp periodicity in AT (blue) or GC (green) enrichment can be seen across nucleosomal regions. In CDS regions, this periodicity is overlying the 3bp periodicity associated with codons. Upstream periodicity is absent, although it is evident in di- and trinucleotide frequencies.

Figure 8

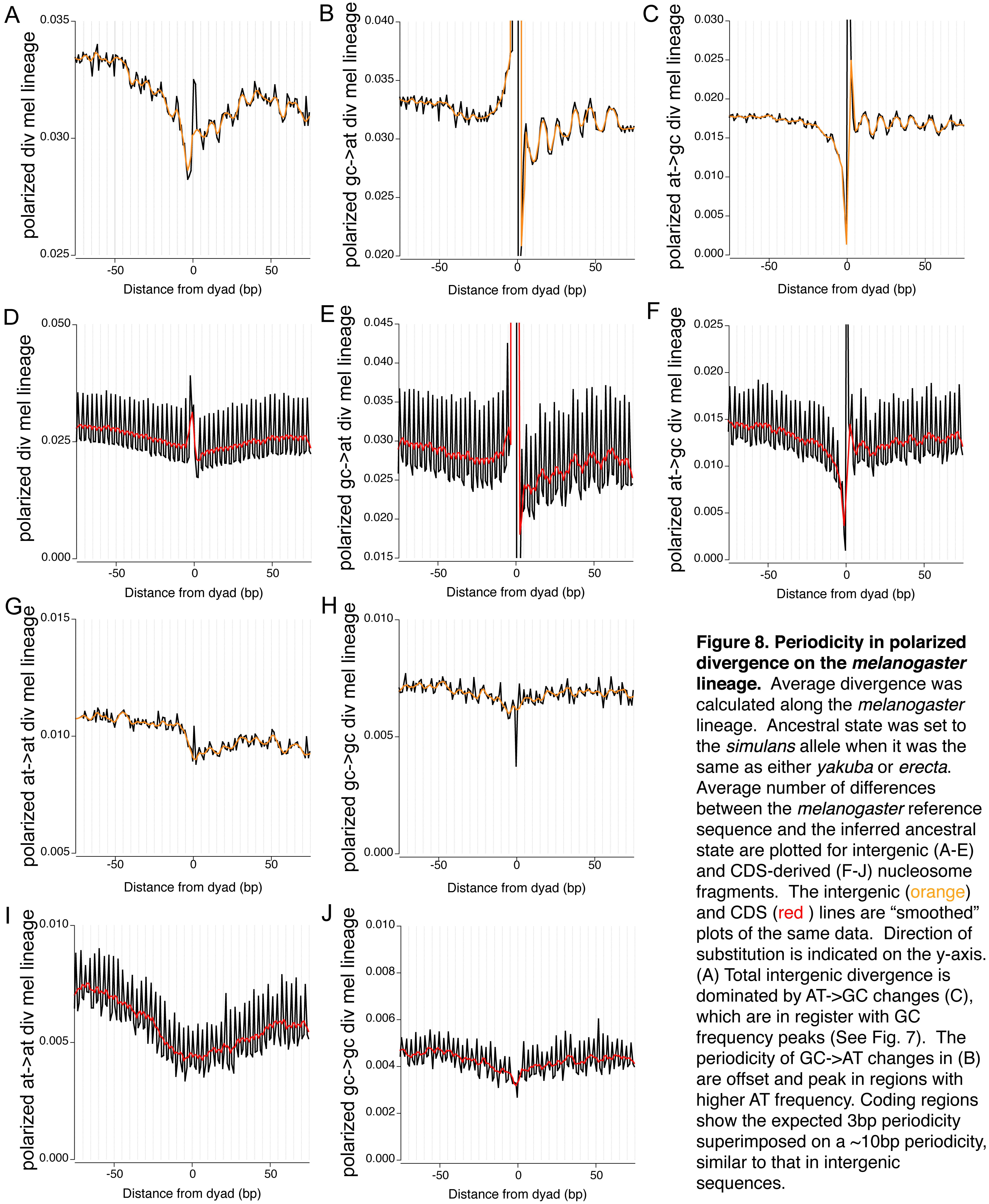


Figure 8. Periodicity in polarized divergence on the *melanogaster* lineage. Average divergence was calculated along the *melanogaster* lineage. Ancestral state was set to the *simulans* allele when it was the same as either *yakuba* or *erecta*. Average number of differences between the *melanogaster* reference sequence and the inferred ancestral state are plotted for intergenic (A-E) and CDS-derived (F-J) nucleosome fragments. The intergenic (orange) and CDS (red) lines are "smoothed" plots of the same data. Direction of substitution is indicated on the y-axis. (A) Total intergenic divergence is dominated by AT->GC changes (C), which are in register with GC frequency peaks (See Fig. 7). The periodicity of GC->AT changes in (B) are offset and peak in regions with higher AT frequency. Coding regions show the expected 3bp periodicity superimposed on a ~10bp periodicity, similar to that in intergenic sequences.

Figure 9

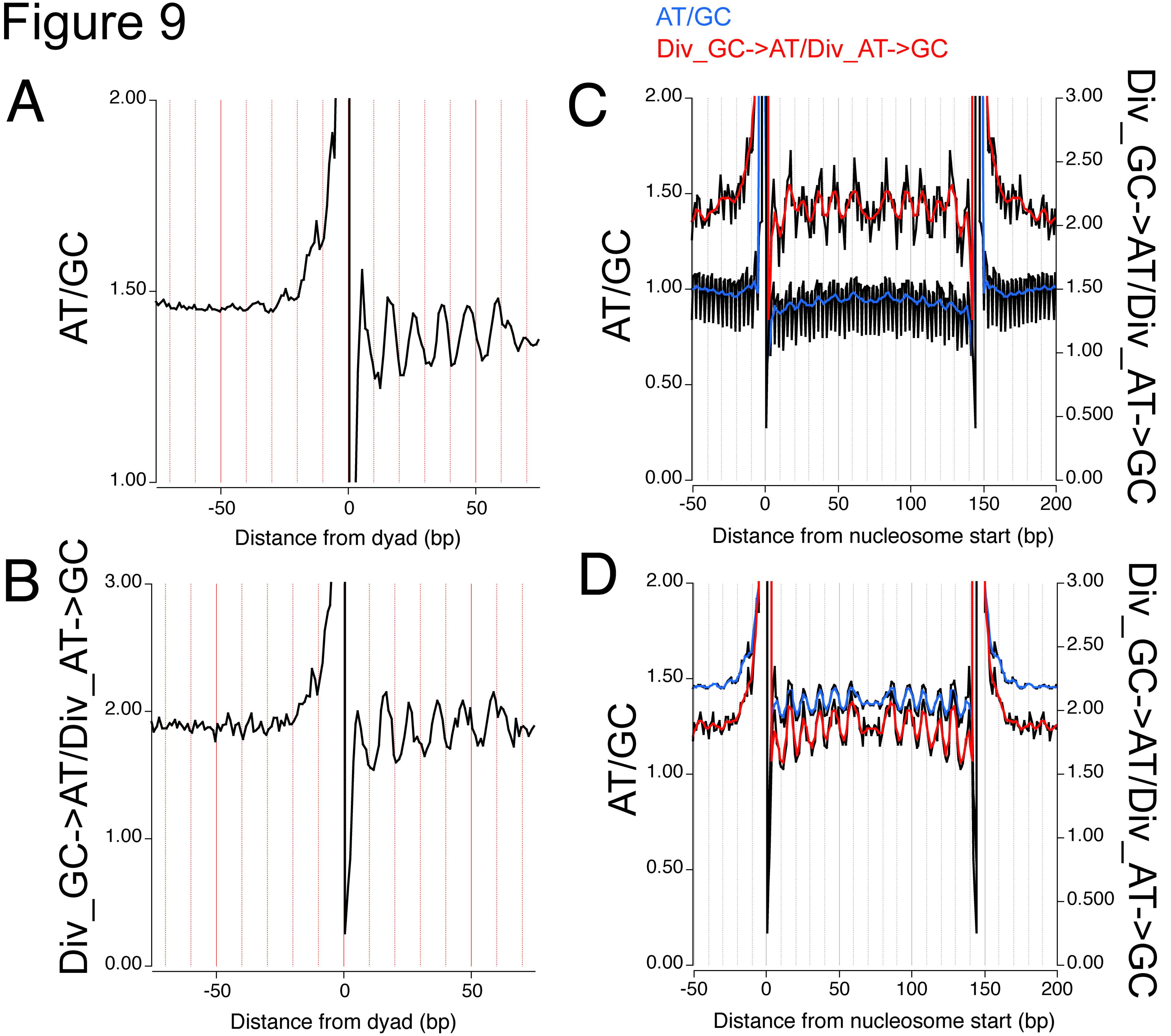


Figure 9. Divergence on the *melanogaster* lineage tracks with but is not proportional to relative nucleotide frequencies. A) Ratio of AT/GC frequencies surrounding intergenic 147bp fragments. B) Ratio of the average polarized GC->AT/AT->GC divergence along the *melanogaster* lineage. Ratios across (C) CDS and (D) intergenic fragments are plotted for comparison.

Figure 10

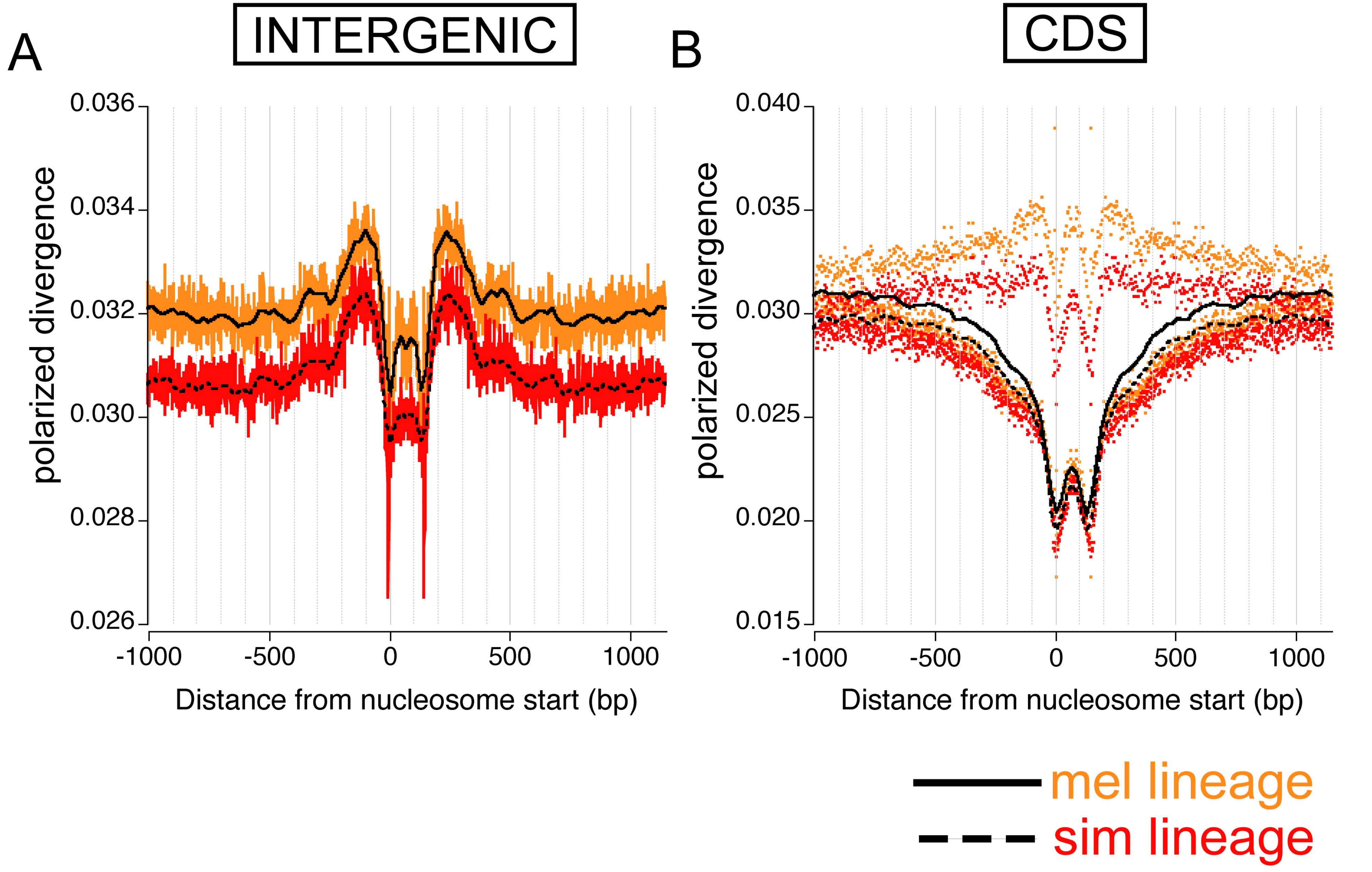


Figure 10. Average polarized divergence on the *simulans* lineage surrounding 147bp nucleosomal DNA fragments isolated from *melanogaster* embryos shows large scale periodicity. For (A) intergenic and (B) CDS fragments, lineage-specific polarized divergence for *melanogaster* (orange) and *simulans* (red) show similar periodic patterns. Data are shown as points in CDS regions due to the large dynamic range of changes in coding regions associated with changes at the 3rd site. Rate of divergence on the *simulans* lineage is lower, particularly in intergenic regions.

Figure 11

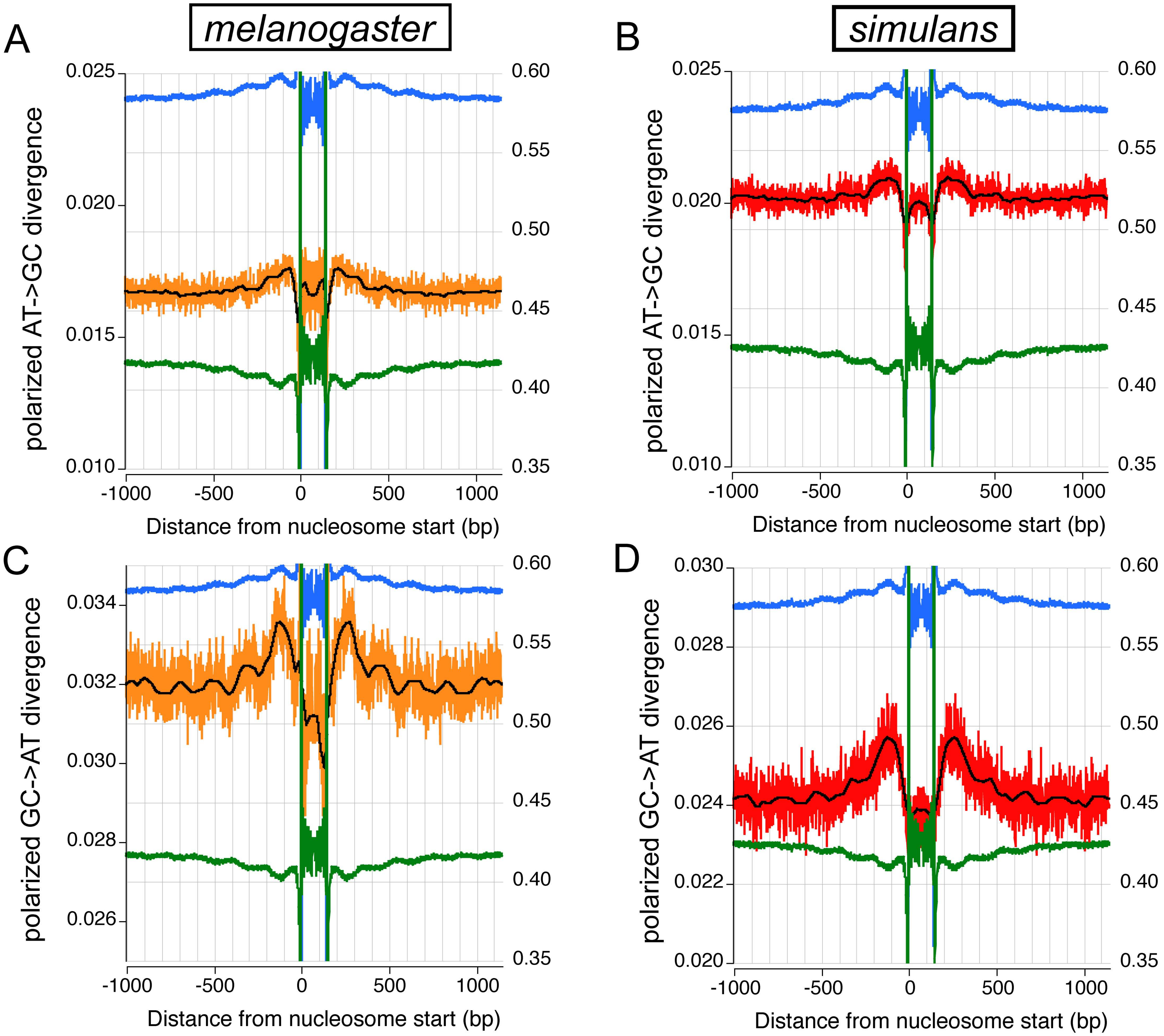


Figure 11. GC->AT changes drive large scale intergenic periodicity on both *simulans* and *melanogaster* lineages relative to *melanogaster* nucleosomal fragments. For intergenic regions surrounding 147bp fragments, AT (blue) content is higher in (A,C) *melanogaster* than (B,D) *simulans*. Polarized AT->GC divergence for (A) *melanogaster* (orange) and (B) *simulans* (red) show only weak periodicity, which may be associated with regions of higher GC, but most strikingly tend to be enriched in regions flanking nucleosomes. Average GC->AT frequencies show a clear periodicity in both (C) *melanogaster* and (D) *simulans*. These peaks of GC->AT align with regions of high AT. Rates of GC->AT are substantially higher on the *melanogaster* lineage.

Figure 12

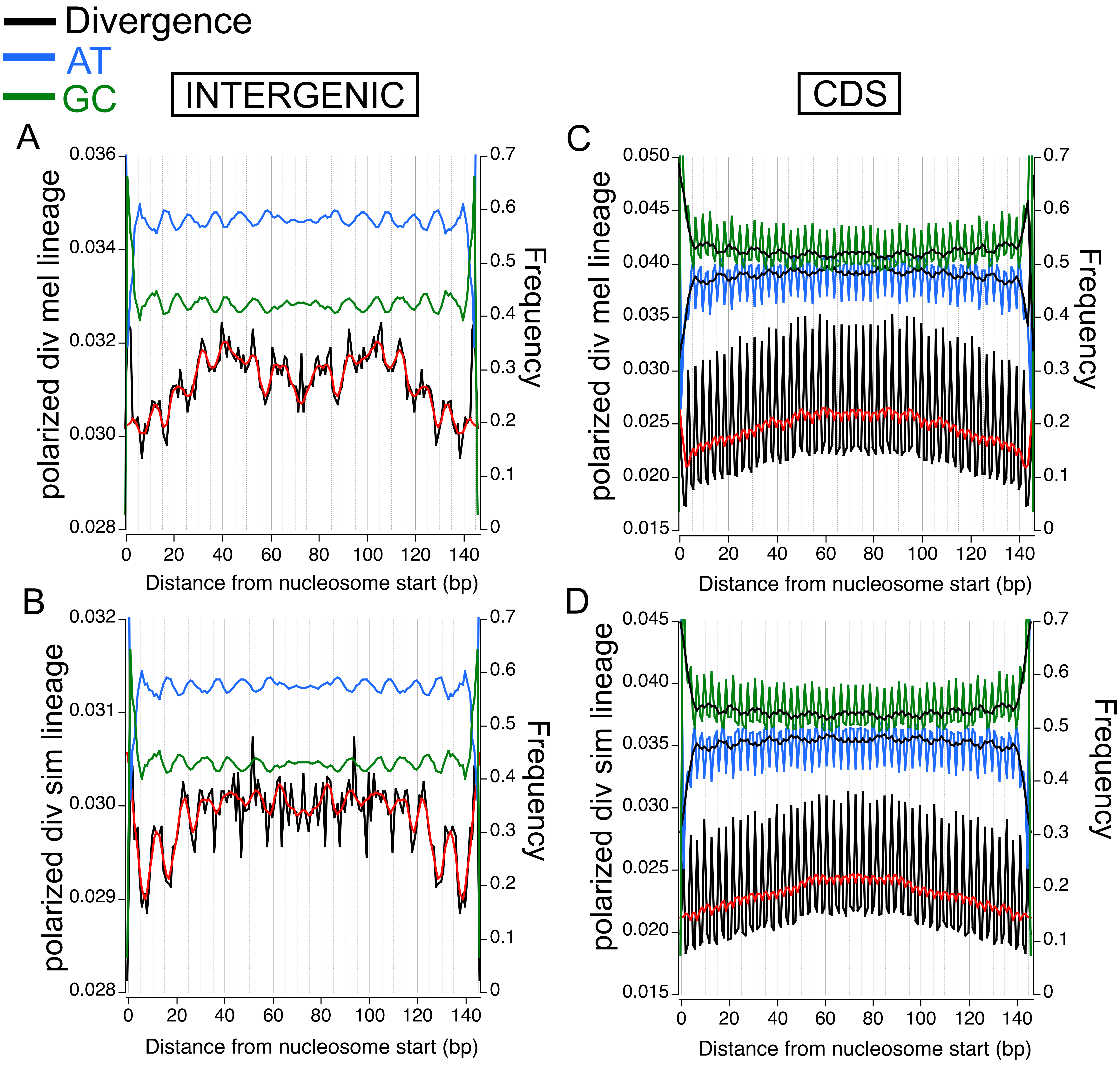


Figure 12. Average divergence on the *simulans* lineage relative to 147bp nucleosomal DNA fragments isolated from *melanogaster* embryos shows ~10bp periodicity. For intergenic fragments, lineage-specific polarized divergence for (A) *melanogaster* and (B) *simulans* show similar periodic patterns and nucleotide frequencies, though *melanogaster* sequences are slightly more AT rich. Peaks of divergence roughly align with peaks in GC frequency across nucleosomal fragments. Positional differences in lineage-specific divergence across CDS nucleosomes are weak in both (C) *melanogaster* and (D) *simulans*, though the general curvature is similar across species.

Figure 13

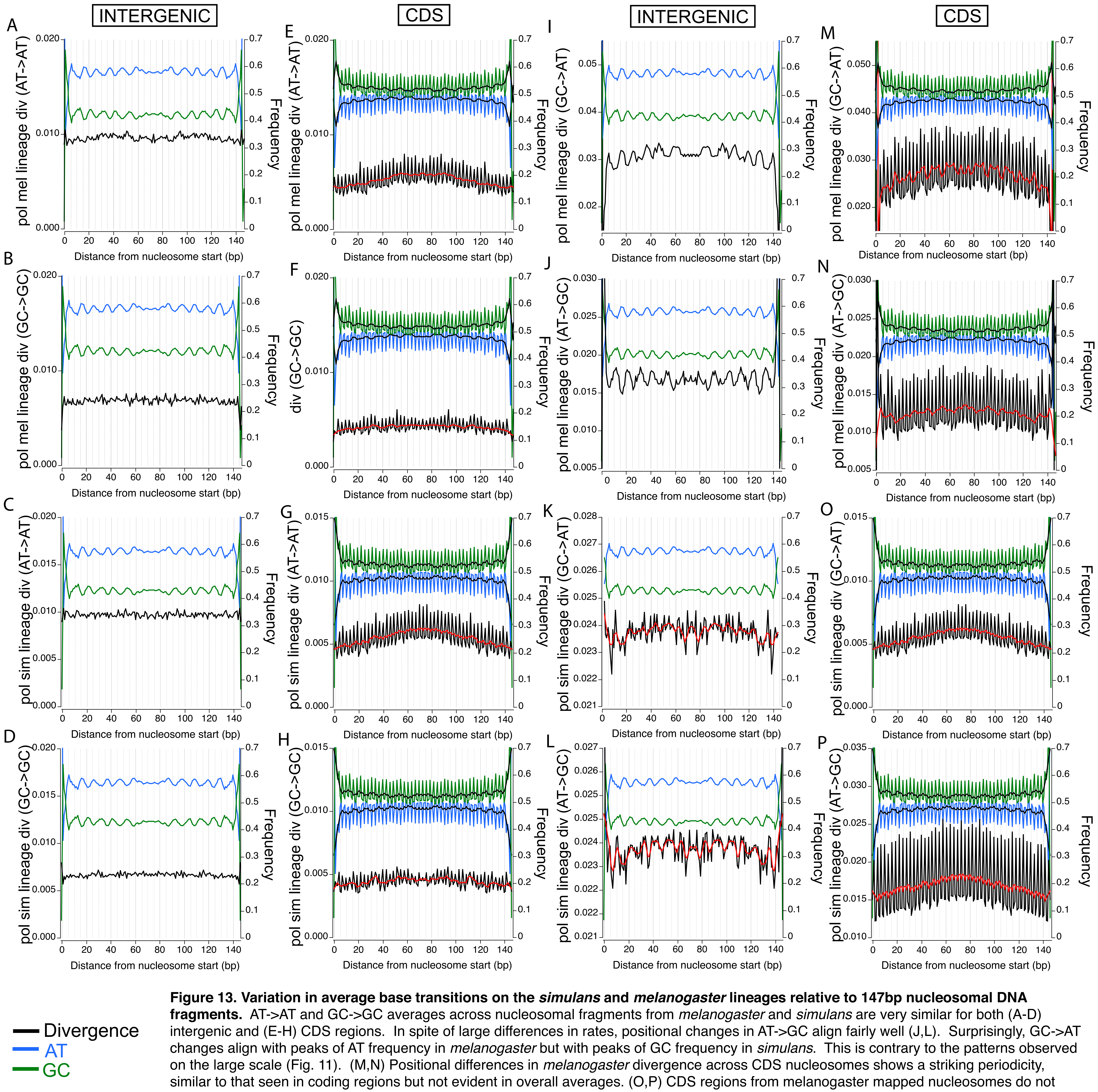


Figure 13. Variation in average base transitions on the *simulans* and *melanogaster* lineages relative to 147bp nucleosomal DNA fragments. AT->AT and GC->GC averages across nucleosomal fragments from *melanogaster* and *simulans* are very similar for both (A-D) intergenic and (E-H) CDS regions. In spite of large differences in rates, positional changes in AT->GC align fairly well (J,L). Surprisingly, GC->AT changes align with peaks of AT frequency in *melanogaster* but with peaks of GC frequency in *simulans*. This is contrary to the patterns observed on the large scale (Fig. 11). (M,N) Positional differences in *melanogaster* divergence across CDS nucleosomes shows a striking periodicity, similar to that seen in coding regions but not evident in overall averages. (O,P) CDS regions from melanogaster mapped nucleosomes do not show this ~10bp periodicity, though there is evidence of a general curvature in divergence which appears to be associated with nucleosome occupancy in these regions.

Chapter 4

Materials and Methods

Chromatin Isolation from embryos and adult flies

2-3hr embryos were dechorionated in 50% bleach for 3 min, then homogenized on ice in SEC buffer (10mM HEPES, 150mM NaCl, 10mM EDTA 10% glycerol, 1mM DTT) with Protease Inhibitors (PI) (0.1mM PMSF and 2X Roche EDTA-free Protease Inhibitor tablets). Whole flies were aged and then frozen in liquid nitrogen. Frozen samples were homogenized in SEC + PI. After lysate filtering and centrifugation, pelleted nuclei from flies or embryos were resuspended in CIB (15 mM Tris pH 7.5, 60 mM KCl, 15 mM NaCl. 0.34M Sucrose, 0.15mM Spermine, 0.5 mM Spermidine) + PI. Centrifugation of nuclei in the sucrose cushion was repeated 3 times. For embryo preps, the resultant pellet was flash frozen and stored at -80C. For NChIP on whole flies, samples were carried directly over into the NChIP protocol without freezing.

Native ChIP

For embryos, pelleted frozen nuclei or were resuspended in CIB + PI, pooled, and chromatin was digested with Micrococcal nuclease. Nuclei from whole flies were digested similarly. MNase treated nuclei were pelleted, resuspended in PBS + 0.1% NP-40 + PI, and incubated at 4° C for 3 hrs to release (primarily) mononucleosomes. A portion of digested chromatin was phenol-chloroform extracted to assess concentration and checked on 1.5% agarose gels for fragment size. For nucleosome mapping in embryos, extracted DNA was used directly to generate libraries for sequencing. For ChIP reactions, input nucleosomal chromatin was adjusted to to 0.05 µg/ul with Incubation Buffer (50mM NaCl, 20mM Tris HCl pH 7.5, 5mM EDTA) + PI and precleared by incubation with 7µl normal rabbit IgG per ml chromatin and, subsequently, with Protein A-Sepharose beads (Sigma).

The cleared lysate was then incubated overnight at 4° C with 150ul Dynabeads (Dynabeads prepared with ~10 µg of the appropriate antibodies per manufacturer's instructions). Antibody sources: H3K9me2 antibodies were obtained from Thomas Jenuwein (Research Institute of Molecular Pathology, Max-Planck Institute (MPI) of Immunobiology, Freiburg, Germany), H2Av antibodies were obtained from Robert Glaser (Wadsworth Center, New York State Department of Health, State University of New York, Albany, NY), H3K4me2 (Upstate Biotech, 07-030), H3K4me3 (Abcam, 8580), H3K27me3 (Lake Placid Biologicals, AR-0171). 30µg of chromatin was used for each ChIP. After incubation, beads were washed once in Buffer A (50mM NaCl, 50mM Tris HCl pH 7.5, 10Mm EDTA), Buffer B (100mM NaCl, 50mM Tris HCl pH 7.5, 10mM EDTA), and Buffer C (150mM NaCl, 50mM Tris HCl pH 7.5, 10mM EDTA) on ice, and then eluted with 1.0% SDS Incubation Buffer. Samples and input control were treated with Proteinase K (1µl 20mg/ml Proteinase K) for 1hr at 50°, and then RNase I (50 µg/ml) for 30 min at 37°C. ChIP and input DNA was extracted with phenol-chloroform and back-extracted to minimize DNA loss. Duplicate experimental ChIP replicates were performed for each antibody.

Amplification and sample preparation

ChIP and input DNA was amplified using the T7 linear amplification method (TLAD) [1]. Briefly, ChIP and input DNA was phosphatased, TdT tailed, and second strand DNA synthesis was performed with Klenow polymerase. Primers containing polyA and a T7 RNA polymerase promoter were ligated to the ends, the dsDNA template was linearly amplified *in vitro* with T7 polymerase. ssDNA probe was produced by RT with random hexamers. After amplification, samples were fragmented with DNase I, biotinylated, and hybridized to Affymetrix *Drosophila* genomic tiling arrays (Version 1). The arrays consist of 25mer probes with a density of approximately one probe every 36bp across the euchromatin, and for unique regions in the heterochromatin.

For short read sequencing data, standard Illumina protocols were used to generate libraries, with the exception that bands were initially cut to include a range of fragment sizes (50-200bp). In a subsequent round to enrich for 147bp fragments, nucleosomal DNA from embryos was size-selected and subject to minimal amplification to avoid PCR bias.

Array processing

Array probes mapping to more than one position in Release 5.4 of the *Drosophila melanogaster* genome were removed using MUMMER with help from Kasper Hansen (Division of Biostatistics, School of Public Health, UC Berkeley, Berkeley, CA). To generate the normalized log ratio (NLR) for each probe, tiling array data from 2 experimental replicates for each ChIP target were normalized such that, for each replicate, the mean of the $\log_2$ (intensity from the ChIP DNA/intensity from the MNase digested input DNA) was zero. Data from the two replicates were averaged. For visual browsing and heatmaps, a sliding 150bp window average was computed with a minimum probe number of 3. Average NLR for probes in a given window was assigned to the average genomic position of all probes in the window. All other analyses were conducted on the normalized, unsmoothed data.

For dot plots of average enrichments of marks, we computed the average NLR for ChIP targets across the largest transcript for genes in Release 5.4 (+500bp upstream and downstream). H2Av "plateau" gene set included 800 genes identified by thresholding average NLR values across transcribed regions (+/-100bp) for H2Av and H3K4triMe (H2Av NLR > 0.1, H3K4triMe NLR < -0.1). Plots of the enrichments surrounding heterochromatin-euchromatin boundary at the base of chr2R were represent 5kb sliding window average NLR for all windows with >25 probes. Average values were assigned to the mean genomic position of all probes in the window.

For correlations of H3K9me2 and H2Av at euchromatic v. heterochromatic genes, we compared average enrichments for 5kb surrounding the TSS of genes bound by Pol II (average NLR > 0). Although individual smoothed data points are not independent, comparative distributions from different genomic regions or gene subsets can show general trends, such as higher correlation across transcriptional units. To avoid spurious correlation that can result from signal generated when both targets are absent from a region (and thus the NLR is background signal over the same MNase digested input), pairwise correlations were based on the set of data points representing the union of all points where NLR > 0 for either target.

Sequencing data processing

36bp paired ends reads were mapped to Release 5.16 of the *D. melanogaster* genome using MAQ. Mate pairs passing quality filtering were used to generate fragment length distributions. From stage 5 embryos, 147,368 147bp fragments mapping internally to CDS regions and 335,306 147bp fragments mapping to intergenic regions were used for downstream analysis presented here. We also examined 40,013 147bp fragments from H3K27me3 NChIP in adult males.

Publicly available chromatin data

Raw stage 5 Drosophila embryo tiling array data for ser5P-Pol II was attained [2] and processed as above. In addition to average distributions surrounding TSSs, the unsmoothed data were used to estimate the average Pol II density across all non-overlapping genes. Chromatin for ser5P-Pol II ChIP experiments was prepared using formaldehyde crosslinking and sonication. Due to the average input chromatin fragment size of 500bp, genes <500bp apart may contribute to the average ser5P-Pol II density of their neighbors in these ChIP experiments. We have therefore used expression studies from early embryos in parallel to validate these results.

Processed *S. cerevisiae* ChIP-array data for histone modifications [3,4] and H2A.Z [5] were used as supplied by the authors. Publicly available high resolution H2A.Z nucleosome prediction values (corrected read counts) from ChIP-Seq data were used as provided [6]. We note that this data set was not normalized to input controls (see footnote 2), and that computational methods used to correct for MNase bias based on H2A.Z nucleosome sequences cannot predict MNase sensitivity differences of bulk nucleosomes. Nucleosome occupancy data [7] was used as provided by the authors.

Publicly available CD4+ ChIP-seq BED files were used to generate a read depth across the human genome for each histone modification or H2AZ [8]. Plus and minus strand reads were combined for these analyses. For averages surrounding TSSs, these data were used to calculate a per gene average read depth for the different orientation and distance classes. As in the case of *S. cerevisiae* ChIP-seq data, these data do not include input chromatin controls (see footnote 2).

Nucleosome model scores and predicted occupancies (default parameters nucleosome concentration = 1, inverse temperature = 0.5) [9] in *S. cerevisiae* and Drosophila were used as supplied by the authors.

Expression Data

Previously published expression data from 2.5-3.5h and 3-4h *Drosophila* embryos were averaged [10]. For plots of average enrichment based on expression level, these data, which cover a subset of 3,499 genes, were ranked and separated into 4 quartiles. Genes were then classed based on orientation and distance to upstream neighbor to generate distributions. 946 parallel and 1,411 divergent non-overlapping genes included in the data set were included in these analyses. For *S. cerevisiae*, we obtained expression data from [11]. Average RNA intensities from wild type and nuclear exosome mutant cells (*rrp6*Δ*)* from [12] were also used. Publicly available human CD4+

cell expression data were provided by Keji Zhao [8]. Non-redundant, non-overlapping UCSC KnownGenes were matched with the $\log_{10}$ expression values for each gene before plotting based on orientation and distance.

Annotation analysis

We used Release 5.4 of the *Drosophila melanogaster* genome to classify euchromatic genes as “divergent” or “parallel” based on orientation relative to the nearest upstream neighboring gene. Genes which overlapped with other genes, including polycistronic and nested transcripts, heterochromatic genes, and those on the 4th chromosome were excluded from the analysis. For genes with more than one annotated transcript, the transcript with the most upstream TSS annotation was used. Distances to the upstream neighbor were computed using the genomic position of the neighboring TSS for divergent genes and the transcription termination site (TTS) for parallel genes. The full list of genes used is included as Supplementary Table 1.

For *S. cerevisiae*, Saccharomyces Genome Database (SGD) annotations were combined with UTRs identified by RNA sequencing [13]. Genes were classed as above, and overlapping genes were excluded from analysis. A total of 1,963 parallel and 1,916 divergent genes were used in the analysis (Supplementary Table 2).

Human analyses were performed using the UCSC KnownGenes and the full ENSEMBL gene set. Duplicate annotations were filtered, and, for genes with multiple annotated transcripts, the transcript with the most upstream TSS was used. Overlapping and nested transcripts, were removed from the set. From the KnownGenes, 5,727 parallel and 5,319 divergent non-overlapping genes were used in the analysis. 8,703 parallel and 7,939 divergent non-overlapping genes were identified from ENSEMBL annotations. The larger gene set from ENSEMBL produced greater differences in profiles for divergent and parallel genes. These results are presented in the Supplemental Fig. 18. Gene sets are included in supplementary tables 3 and 4.

Heatmaps

For *Drosophila*, smoothed ChIP-array data from embryos were used to obtain average values in discrete 200bp windows surrounding the TSSs of individual non-overlapping genes (+/-2kb from TSS). Genes lacking probes in any of the 20 windows were excluded from the plots. Averages for individual genes were sorted based on distance to the upstream neighbor before stack plotting. A total of 3,845 parallel and 4,415 divergent genes were plotted.

CD4+ cell heatmaps were produced using the sum of the read depth in 200bp windows surrounding the TSSs. 5,756 parallel and 5,342 divergent non-redundant, non-overlapping genes from the UCSC KnownGene annotations were plotted.

Correlation

Non-overlapping gene pairs (both members) with average Pol II densities (NLRs) from 0-4 were ranked by intergenic distance and separated into 5 equal groups. A total of 1,280 tandem and 1,090 divergent pairs were used in the analysis. Spearman’s rank correlation of average Pol II densities for pairs was calculated and plotted at the group mean distance. P-values for all correlations were all <1.6e-05, with the exception of the tandem group with mean distance = 244bp, which was not significant.

Base composition and divergence analyses

Syntenic alignments produced by the DPGP (http://www.dpgp.org) were used to extract homologous regions from *melanogaster*, *simulans*, *yakuba*, and *erecta* genomes. Sequences surrounding mapped nucleosomal fragments in *melanogaster* were used to calculate base composition in and di/trinucleotide frequencies. All analyses were performed on both strands (5'->3'), which has the effect of symmetrizing the results. Positional divergence relative to 147bp fragments between *melanogaster* and *yakuba* was calculated as the sum of the average pairwise differences at each position divided by the number of bases considered. Average positional polarized divergence along the *melanogaster* lineage was calculated by comparing the *simulans* allele to *yakuba* and *erecta* and calling the *simulans* base ancestral state if it agreed with one or both. Similarly, divergence on the *simulans* lineage was calculated using *melanogaster* to determine the ancestral state. Changes represent inferred differences specific to the *melanogaster* or *simulans* lineage.

Smoothing

Where necessary, locally weighted Least Squared error (Lowess) smoothing was used to clarify trends in the data . This method is less sensitive to outliers than other smoothing methods and was applied with uniform smoothing factors for individual plots.

Footnotes:

1. For averages surrounding TSSs of divergent genes, the enrichment of each member of a pair is "counted" in the upstream and downstream peaks. The peaks are not fully symmetric because distances in the bins are not uniform, so the upstream peak TSSs are not precisely aligned. The proportional decrease we see in both upstream and downstream peaks for divergent genes with increasing distance does not in itself demonstrate that both genes in a pair for a given bin show lower levels of plotted marks. If chromatin of only one member were less modified with increasing distance, the plot would show a similar trend. However, average ser5-phospho-Pol II densities across genes for adjacent divergent pairs show significant correlation, especially for those ≤1kb apart (Fig. 5D, Fig. 9), suggesting that proportional enrichment of Pol II and hallmarks of gene activity is indeed a feature of close divergent genes in stage 5 Drosophila embryos.

2. ChIP-seq data from *S. cerevisiae* [6] and CD4+ cells [8] are not normalized to total MNase digested nucleosomal input. This may introduce biases in these data, in particular an overrepresentation of nuclease sensitive regions of the genome. MNase selectively cuts in active regions and can cause higher representation of active genes in the mononucleosomal fraction [14,15]. Further, digestion of chromatin to primarily mononucleosomes results in some cutting of nucleosome-associated DNA, which may introduce more complex biases [16,17]. Albert et al. applied a computational method intended to correct for the sequence bias of MNase digestion.

Their model uses the frequencies of bases at the ends of sequenced H2AZ nucleosomal DNA to represent the overall genomic cutting bias of the enzyme. We note that the actual frequencies of sequence bias for MNase must be derived from total digested nucleosomal DNA. Additionally, this correction does not directly address biases associated with chromatin structure or gene expression. We do not, however, expect the these potential biases to qualitatively influence the results presented here. However, further ChIP-seq studies employing controls will be necessary to address the importance of correcting for MNase bias in such experiments.

Bibliography

1. Strahl BD, Allis CD (2000) The language of covalent histone modifications. Nature 403: 41-45.
2. Turner BM (2000) Histone acetylation and an epigenetic code. Bioessays 22: 836-845.
3. Allshire RC, Karpen GH (2008) Epigenetic regulation of centromeric chromatin: old dogs, new tricks? Nat Rev Genet 9: 923-937.
4. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. pp. 315-326.
5. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature. pp. 553-560.
6. Pan G, Tian S, Nie J, Yang C, Ruotti V, et al. (2007) Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. Cell Stem Cell. pp. 299-312.
7. Zhao XD, Han X, Chew JL, Liu J, Chiu KP, et al. (2007) Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. Cell Stem Cell. pp. 286-298.
8. Azuara V, Perry P, Sauer S, Spivakov M, Jorgensen HF, et al. (2006) Chromatin signatures of pluripotent cell lines. Nat Cell Biol 8: 532-538.
9. Christensen J, Agger K, Cloos PA, Pasini D, Rose S, et al. (2007) RBP2 belongs to a family of demethylases, specific for tri-and dimethylated lysine 4 on histone 3. Cell 128: 1063-1076.
10. Klose RJ, Zhang Y (2007) Regulation of histone methylation by demethylimination and demethylation. Nat Rev Mol Cell Biol 8: 307-318.
11. Lee N, Zhang J, Klose RJ, Erdjument-Bromage H, Tempst P, et al. (2007) The trithorax-group protein Lid is a histone H3 trimethyl-Lys4 demethylase. Nat Struct Mol Biol 14: 341-343.
12. Wysocka J, Swigut T, Milne TA, Dou Y, Zhang X, et al. (2005) WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development. Cell 121: 859-872.
13. Swaminathan J, Baxter EM, Corces VG (2005) The role of histone H2Av variant replacement and histone H4 acetylation in the establishment of Drosophila heterochromatin. Genes Dev. pp. 65-76.
14. Kusch T, Florens L, Macdonald WH, Swanson SK, Glaser RL, et al. (2004) Acetylation by Tip60 is required for selective histone variant exchange at DNA lesions. Science 306: 2084-2087.
15. Guillemette B, Gaudreau L (2006) Reuniting the contrasting functions of H2A.Z. Biochem Cell Biol 84: 528-535.
16. Babiarz JE, Halley JE, Rine J (2006) Telomeric heterochromatin boundaries require NuA4-dependent acetylation of histone variant H2A.Z in Saccharomyces cerevisiae. Genes Dev 20: 700-710.
17. Raisner RM, Hartley PD, Meneghini MD, Bao MZ, Liu CL, et al. (2005) Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin. Cell. pp. 233-248.

18. Brickner DG, Cajigas I, Fondufe-Mittendorf Y, Ahmed S, Lee PC, et al. (2007) H2A.Z-mediated localization of genes at the nuclear periphery confers epigenetic memory of previous transcriptional state. PLoS Biol 5: e81.

19. Deal RB, Topp CN, McKinney EC, Meagher RB (2007) Repression of flowering in Arabidopsis requires activation of FLOWERING LOCUS C expression by the histone variant H2A.Z. Plant Cell 19: 74-83.

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

21. Hanai K, Furuhashi H, Yamamoto T, Akasaka K, Hirose S (2008) RSF governs silent chromatin formation via histone H2Av replacement. PLoS Genet. pp. e1000011.

22. Greaves IK, Rangasamy D, Ridgway P, Tremethick DJ (2007) H2A.Z contributes to the unique 3D structure of the centromere. Proc Natl Acad Sci U S A 104: 525-530.

23. Rangasamy D, Berven L, Ridgway P, Tremethick DJ (2003) Pericentric heterochromatin becomes enriched with H2A.Z during early mammalian development. EMBO J 22: 1599-1607.

24. Rangasamy D, Greaves I, Tremethick DJ (2004) RNA interference demonstrates a novel role for H2A.Z in chromosome segregation. Nat Struct Mol Biol 11: 650-655.

25. Leach TJ, Mazzeo M, Chotkowski HL, Madigan JP, Wotring MG, et al. (2000) Histone H2A.Z is widely but nonrandomly distributed in chromosomes of Drosophila melanogaster. J Biol Chem 275: 23267-23272.

26. Greaves IK, Rangasamy D, Devoy M, Marshall Graves JA, Tremethick DJ (2006) The X and Y chromosomes assemble into H2A.Z-containing [corrected] facultative heterochromatin [corrected] following meiosis. Mol Cell Biol 26: 5394-5405.

27. Brower-Toland B, Riddle NC, Jiang H, Huisinga KL, Elgin SC (2009) Multiple SET methyltransferases are required to maintain normal heterochromatin domains in the genome of Drosophila melanogaster. Genetics 181: 1303-1319.

28. Vermaak D, Malik HS (2009) Multiple roles for heterochromatin protein 1 genes in Drosophila. Annu Rev Genet 43: 467-492.

29. Rudolph T, Yonezawa M, Lein S, Heidrich K, Kubicek S, et al. (2007) Heterochromatin formation in Drosophila is initiated through active removal of H3K4 methylation by the LSD1 homolog SU(VAR)3-3. Mol Cell 26: 103-115.

30. Peng JC, Karpen GH (2007) H3K9 methylation and RNA interference regulate nucleolar organization and repeated DNA stability. Nat Cell Biol 9: 25-35.

31. Foe VE, Alberts BM (1983) Studies of nuclear and cytoplasmic behaviour during the five mitotic cycles that precede gastrulation in Drosophila embryogenesis. J Cell Sci 61: 31-70.

32. Smith CD, Shu S, Mungall CJ, Karpen GH (2007) The Release 5.1 annotation of Drosophila melanogaster heterochromatin. Science 316: 1586-1591.

33. Schotta G, Ebert A, Dorn R, Reuter G (2003) Position-effect variegation and the genetic dissection of chromatin regulation in Drosophila. Semin Cell Dev Biol. pp. 67-75.

34. Yasuhara JC, Wakimoto BT (2006) Oxymoron no more: the expanding world of heterochromatic genes. Trends Genet 22: 330-338.

35. Schuettengruber B, Ganapathi M, Leblanc B, Portoso M, Jaschek R, et al. (2009) Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos. PLoS Biol 7: e13.

36. Schübeler D, MacAlpine DM, Scalzo D, Wirbelauer C, Kooperberg C, et al. (2004) The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. Genes Dev. pp. 1263-1271.
37. Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL, et al. (2005) Single-nucleosome mapping of histone modifications in S. cerevisiae. PLoS Biol. pp. e328.
38. Guillemette B, Bataille AR, Gévry N, Adam M, Blanchette M, et al. (2005) Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. PLoS Biol. pp. e384.
39. Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, et al. (2002) Gene expression during the life cycle of Drosophila melanogaster. Science. pp. 2270-2275.
40. Wirbelauer C, Bell O, Schubeler D (2005) Variant histone H3.3 is deposited at sites of nucleosomal displacement throughout transcribed genes while active histone modifications show a promoter-proximal bias. Genes Dev 19: 1761-1766.
41. Mito Y, Henikoff JG, Henikoff S (2005) Genome-scale profiling of histone H3.3 replacement patterns. Nat Genet. pp. 1090-1097.
42. Millar CB, Xu F, Zhang K, Grunstein M (2006) Acetylation of H2AZ Lys 14 is associated with genome-wide gene activity in yeast. Genes Dev. pp. 711-722.
43. Li XY, MacArthur S, Bourgon R, Nix D, Pollard DA, et al. (2008) Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol. pp. e27.
44. Hiraoka Y, Agard DA, Sedat JW (1990) Temporal and spatial coordination of chromosome movement, spindle formation, and nuclear envelope breakdown during prometaphase in Drosophila melanogaster embryos. J Cell Biol 111: 2815-2828.
45. Cleard F, Delattre M, Spierer P (1997) SU(VAR)3-7, a Drosophila heterochromatin-associated protein and companion of HP1 in the genomic silencing of position-effect variegation. EMBO J 16: 5280-5288.
46. Bushey AM, Dorman ER, Corces VG (2008) Chromatin insulators: regulatory mechanisms and epigenetic inheritance. Mol Cell 32: 1-9.
47. Jin C, Zang C, Wei G, Cui K, Peng W, et al. (2009) H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. Nat Genet 41: 941-945.
48. Jin C, Felsenfeld G (2007) Nucleosome stability mediated by histone variants H3.3 and H2A.Z. Genes Dev 21: 1519-1529.
49. Fan JY, Gordon F, Luger K, Hansen JC, Tremethick DJ (2002) The essential histone variant H2A.Z regulates the equilibrium between different chromatin conformational states. Nat Struct Biol 9: 172-176.
50. Fan JY, Rangasamy D, Luger K, Tremethick DJ (2004) H2A.Z alters the nucleosome surface to promote HP1alpha-mediated chromatin fiber folding. Mol Cell 16: 655-661.
51. Kobor MS, Venkatasubrahmanyam S, Meneghini MD, Gin JW, Jennings JL, et al. (2004) A protein complex containing the conserved Swi2/Snf2-related ATPase Swr1p deposits histone variant H2A.Z into euchromatin. PLoS Biol 2: E131.
52. Mizuguchi G, Shen X, Landry J, Wu WH, Sen S, et al. (2004) ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. Science. pp. 343-348.

53. Ruhl DD, Jin J, Cai Y, Swanson S, Florens L, et al. (2006) Purification of a human SRCAP complex that remodels chromatin by incorporating the histone variant H2A.Z into nucleosomes. Biochemistry 45: 5671-5677.
54. Cai Y, Jin J, Florens L, Swanson SK, Kusch T, et al. (2005) The mammalian YL1 protein is a shared subunit of the TRRAP/TIP60 histone acetyltransferase and SRCAP complexes. J Biol Chem 280: 13665-13670.
55. Schwartz YB, Kahn TG, Nix DA, Li XY, Bourgon R, et al. (2006) Genome-wide analysis of Polycomb targets in Drosophila melanogaster. Nat Genet 38: 700-705.
56. Byrd K, Corces VG (2003) Visualization of chromatin domains created by the gypsy insulator of Drosophila. J Cell Biol 162: 565-574.
57. Sarcinella E, Zuzarte PC, Lau PN, Draker R, Cheung P (2007) Monoubiquitylation of H2A.Z distinguishes its association with euchromatin or facultative heterochromatin. Mol Cell Biol 27: 6457-6468.
58. Jiang C, Pugh BF (2009) Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet 10: 161-172.
59. Shivaswamy S, Bhinge A, Zhao Y, Jones S, Hirst M, et al. (2008) Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. PLoS Biol 6: e65.
60. Mito Y, Henikoff JG, Henikoff S (2007) Histone replacement marks the boundaries of cis-regulatory domains. Science 315: 1408-1411.
61. Michalak P (2008) Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics. pp. 243-248.
62. Hurst LD, Pál C, Lercher MJ (2004) The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet. pp. 299-310.
63. Spellman PT, Rubin GM (2002) Evidence for large domains of similarly expressed genes in the Drosophila genome. J Biol. pp. 5.
64. Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, et al. (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science. pp. 1289-1292.
65. Lercher MJ, Hurst LD (2006) Co-expressed yeast genes cluster over a long range but are not regularly spaced. J Mol Biol. pp. 825-831.
66. Williams EJ, Bowles DJ (2004) Coexpression of neighboring genes in the genome of Arabidopsis thaliana. Genome Res 14: 1060-1067.
67. Bortoluzzi S, Rampoldi L, Simionati B, Zimbello R (1998) A Comprehensive, High-Resolution Genomic Transcript Map of Human Skeletal Muscle. Genome Research.
68. Lercher MJ, Urrutia AO, Hurst LD (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet. pp. 180-183.
69. Purmann A, Toedling J, Schueler M, Carninci P, Lehrach H, et al. (2007) Genomic organization of transcriptomes in mammals: Coregulation and cofunctionality. Genomics. pp. 580-587.
70. Sémon M, Duret L (2006) Evolutionary origin and maintenance of coexpressed gene clusters in mammals. Mol Biol Evol. pp. 1715-1723.
71. Kruglyak S, Tang H (2000) Regulation of adjacent yeast genes. Trends Genet. pp. 109-111.
72. Cohen BA, Mitra RD, Hughes JD, Church GM (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat Genet 26: 183-186.

73. Hao P, Yu Y, Zhang X, Tu K, Fan H, et al. (2009) The contribution of cis-regulatory elements to head-to-head gene pairs' co-expression pattern. Sci China C Life Sci 52: 74-79.
74. Batada NN, Urrutia AO, Hurst LD (2007) Chromatin remodelling is a major source of coexpression of linked genes in yeast. Trends Genet. pp. 480-484.
75. Sproul D, Gilbert N, Bickmore WA (2005) The role of chromatin structure in regulating the expression of clustered genes. Nat Rev Genet. pp. 775-781.
76. Hurst LD, Williams EJ, Pal C (2002) Natural selection promotes the conservation of linkage of co-expressed genes. Trends Genet 18: 604-606.
77. Chen N, Stein LD (2006) Conservation and functional significance of gene topology in the genome of Caenorhabditis elegans. Genome Res 16: 606-617.
78. Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI (2004) An Abundance of Bidirectional Promoters in the Human Genome. Genome Research.
79. Koyanagi KO, Hagiwara M, Itoh T, Gojobori T, Imanishi T (2005) Comparative genomics of bidirectional gene pairs and its implications for the evolution of a transcriptional regulation system. Gene. pp. 169-176.
80. Adachi N, Lieber MR (2002) Bidirectional gene organization: a common architectural feature of the human genome. Cell. pp. 807-809.
81. Takai D, Jones PA (2004) Origins of bidirectional promoters: computational analyses of intergenic distance in the human genome. Mol Biol Evol. pp. 463-467.
82. Lin JM, Collins PJ, Trinklein ND, Fu Y, Xi H, et al. (2007) Transcription factor binding and modified histones in human bidirectional promoters. Genome Res 17: 818-827.
83. Yang MQ, Elnitski LL (2008) Diversity of core promoter elements comprising human bidirectional promoters. BMC Genomics 9 Suppl 2: S3.
84. Rada-Iglesias A, Ameur A, Kapranov P, Enroth S, Komorowski J, et al. (2008) Whole-genome maps of USF1 and USF2 binding and histone H3 acetylation reveal new aspects of promoter structure and candidate genes for common human disorders. Genome Res 18: 380-392.
85. Yang L, Yu J (2009) A comparative analysis of divergently-paired genes (DPGs) among Drosophila and vertebrate genomes. BMC Evol Biol 9: 55.
86. Rando OJ (2007) Global patterns of histone modifications. Curr Opin Genet Dev. pp. 94-99.
87. Mendenhall EM, Bernstein BE (2008) Chromatin state maps: new technologies, new insights. Curr Opin Genet Dev. pp. 109-115.
88. Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, et al. (2007) Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. Nature. pp. 572-576.
89. Li B, Pattenden SG, Lee D, Gutiérrez J, Chen J, et al. (2005) Preferential occupancy of histone variant H2AZ at inactive promoters influences local histone modifications and chromatin remodeling. Proc Natl Acad Sci USA. pp. 18385-18390.
90. Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, et al. (2005) Genome-scale identification of nucleosome positions in S. cerevisiae. Science. pp. 626-630.
91. Zhang H, Roberts DN, Cairns BR (2005) Genome-wide dynamics of Htz1, a histone H2A variant that poises repressed/basal promoters for activation through histone loss. Cell. pp. 219-231.
92. Bernstein BE, Liu CL, Humphrey EL, Perlstein EO, Schreiber SL (2004) Global nucleosome occupancy in yeast. Genome Biol 5: R62.

93. Lee W, Tillo D, Bray N, Morse RH, Davis RW, et al. (2007) A high-resolution atlas of nucleosome occupancy in yeast. Nat Genet 39: 1235-1244.
94. Lee CK, Shibata Y, Rao B, Strahl BD, Lieb JD (2004) Evidence for nucleosome depletion at active regulatory regions genome-wide. Nat Genet 36: 900-905.
95. Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, et al. (2008) Nucleosome organization in the Drosophila genome. Nature. pp. 358-362.
96. Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, et al. (2005) Genome-wide map of nucleosome acetylation and methylation in yeast. Cell. pp. 517-527.
97. Roh TY, Cuddapah S, Cui K, Zhao K (2006) The genomic landscape of histone modifications in human T cells. Proc Natl Acad Sci USA. pp. 15782-15787.
98. Kirmizis A, Santos-Rosa H, Penkett CJ, Singer MA, Vermeulen M, et al. (2007) Arginine methylation at histone H3R2 controls deposition of H3K4 trimethylation. Nature 449: 928-932.
99. Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, et al. (2009) Bidirectional promoters generate pervasive transcription in yeast. Nature.
100. Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, et al. (2008) Divergent transcription from active promoters. Science. pp. 1849-1851.
101. Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. pp. 1845-1848.
102. Neil H, Malabat C, d'Aubenton-Carafa Y, Xu Z, Steinmetz LM, et al. (2009) Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. Nature.
103. Tirosh I, Barkai N (2008) Two strategies for gene regulation by promoter nucleosomes. Genome Res 18: 1084-1091.
104. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, et al. (2006) A genomic code for nucleosome positioning. Nature 442: 772-778.
105. Perocchi F, Xu Z, Clauder-Munster S, Steinmetz LM (2007) Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. Nucleic Acids Res 35: e128.
106. Henikoff S, Henikoff JG, Sakai A, Loeb GB, Ahmad K (2009) Genome-wide profiling of salt fractions maps physical properties of chromatin. Genome Res 19: 460-469.
107. Tsai HK, Su CP, Lu MY, Shih CH, Wang D (2007) Co-expression of adjacent genes in yeast cannot be simply attributed to shared regulatory system. BMC Genomics. pp. 352.
108. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326: 289-293.
109. Schwartz YB, Pirrotta V (2008) Polycomb complexes and epigenetic states. Curr Opin Cell Biol 20: 266-273.
110. Kornberg RD, Lorch Y (1999) Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. Cell 98: 285-294.
111. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 389: 251-260.
112. Richmond TJ, Davey CA (2003) The structure of DNA in the nucleosome core. Nature 423: 145-150.
113. Dorigo B, Schalch T, Kulangara A, Duda S, Schroeder RR, et al. (2004) Nucleosome arrays reveal the two-start organization of the chromatin fiber. Science 306: 1571-1573.

114. Cooper JL, Henikoff S (2004) Adaptive evolution of the histone fold domain in centromeric histones. Mol Biol Evol 21: 1712-1718.

115. Malik HS, Henikoff S (2001) Adaptive evolution of Cid, a centromere-specific histone in Drosophila. Genetics 157: 1293-1298.

116. Wyrick JJ, Holstege FC, Jennings EG, Causton HC, Shore D, et al. (1999) Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast. Nature 402: 418-421.

117. Sekinger EA, Moqtaderi Z, Struhl K (2005) Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. Mol Cell 18: 735-748.

118. Adkins MW, Howar SR, Tyler JK (2004) Chromatin disassembly mediated by the histone chaperone Asf1 is essential for transcriptional activation of the yeast PHO5 and PHO8 genes. Mol Cell 14: 657-666.

119. Schwartz BE, Ahmad K (2005) Transcriptional activation triggers deposition and removal of the histone variant H3.3. Genes Dev 19: 804-814.

120. Struhl K (1985) Naturally occurring poly(dA-dT) sequences are upstream promoter elements for constitutive transcription in yeast. Proc Natl Acad Sci U S A 82: 8419-8423.

121. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, et al. (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. Nature 458: 362-366.

122. Field Y, Kaplan N, Fondufe-Mittendorf Y, Moore IK, Sharon E, et al. (2008) Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. PLoS Comput Biol 4: e1000216.

123. Anderson JD, Widom J (2001) Poly(dA-dT) promoter elements increase the equilibrium accessibility of nucleosomal DNA target sites. Mol Cell Biol 21: 3830-3839.

124. Bao Y, White CL, Luger K (2006) Nucleosome core particles containing a poly(dA.dT) sequence element exhibit a locally distorted DNA structure. J Mol Biol 361: 617-624.

125. Anselmi C, Bocchinfuso G, De Santis P, Savino M, Scipioni A (1999) Dual role of DNA intrinsic curvature and flexibility in determining nucleosome stability. J Mol Biol 286: 1293-1301.

126. Shrader TE, Crothers DM (1990) Effects of DNA sequence and histone-histone interactions on nucleosome placement. J Mol Biol 216: 69-84.

127. Ioshikhes I, Bolshoy A, Derenshteyn K, Borodovsky M, Trifonov EN (1996) Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. J Mol Biol 262: 129-139.

128. Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, et al. (2008) Nucleosome organization in the Drosophila genome. Nature 453: 358-362.

129. Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, et al. (2007) Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. Nature 446: 572-576.

130. Zhang Y, Moqtaderi Z, Rattner BP, Euskirchen G, Snyder M, et al. (2009) Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. Nat Struct Mol Biol 16: 847-852.

131. Field Y, Fondufe-Mittendorf Y, Moore IK, Mieczkowski P, Kaplan N, et al. (2009) Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization. Nat Genet 41: 438-445.

132. Tillo D, Hughes TR (2009) G+C content dominates intrinsic nucleosome occupancy. BMC Bioinformatics 10: 442.

133. Ioshikhes IP, Albert I, Zanton SJ, Pugh BF (2006) Nucleosome positions predicted through comparative genomics. Nat Genet 38: 1210-1215.

134. Tolstorukov MY, Colasanti AV, McCandlish DM, Olson WK, Zhurkin VB (2007) A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. J Mol Biol 371: 725-738.

135. Morozov AV, Fortney K, Gaykalova DA, Studitsky VM, Widom J, et al. (2009) Using DNA mechanics to predict in vitro nucleosome positions and formation energies. Nucleic Acids Res 37: 4707-4722.

136. Miele V, Vaillant C, d'Aubenton-Carafa Y, Thermes C, Grange T (2008) DNA physical properties determine nucleosome occupancy from yeast to fly. Nucleic Acids Res 36: 3746-3756.

137. Weiner A, Hughes A, Yassour M, Rando OJ, Friedman N High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. Genome Res 20: 90-100.

138. Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, et al. (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Res 18: 1073-1083.

139. Hodges C, Bintu L, Lubkowska L, Kashlev M, Bustamante C (2009) Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. Science 325: 626-628.

140. Kulaeva OI, Gaykalova DA, Studitsky VM (2007) Transcription through chromatin by RNA polymerase II: histone displacement and exchange. Mutat Res 618: 116-129.

141. Sasaki S, Mello CC, Shimada A, Nakatani Y, Hashimoto S, et al. (2009) Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. Science 323: 401-404.

142. Babbitt GA, Tolstorukov MY, Kim Y The molecular evolution of nucleosome positioning through sequence-dependent deformation of the DNA polymer. J Biomol Struct Dyn 27: 765-780.

143. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652-654.

144. Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. Genetics 132: 1161-1176.

145. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

146. Gross DS, Garrard WT (1988) Nuclease hypersensitive sites in chromatin. Annu Rev Biochem 57: 159-197.

147. Keene MA, Corces V, Lowenhaupt K, Elgin SC (1981) DNase I hypersensitive sites in Drosophila chromatin occur at the 5' ends of regions of transcription. Proc Natl Acad Sci U S A 78: 143-146.

148. Ercan S, Simpson RT (2004) Global chromatin structure of 45,000 base pairs of chromosome III in a- and alpha-cell yeast and during mating-type switching. Mol Cell Biol 24: 10026-10035.

149. Keene MA, Elgin SC (1981) Micrococcal nuclease as a probe of DNA sequence organization and chromatin structure. Cell 27: 57-64.

150. Ruthenburg AJ, Allis CD, Wysocka J (2007) Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. Mol Cell 25: 15-30.

151. Dingwall C, Lomonossoff GP, Laskey RA (1981) High sequence specificity of micrococcal nuclease. Nucleic Acids Res 9: 2659-2673.

152. Wang JP, Fondufe-Mittendorf Y, Xi L, Tsai GF, Segal E, et al. (2008) Preferentially quantized linker DNA lengths in Saccharomyces cerevisiae. PLoS Comput Biol 4: e1000175.

153. Tolstorukov MY, Kharchenko PV, Goldman JA, Kingston RE, Park PJ (2009) Comparative analysis of H2A.Z nucleosome organization in the human and yeast genomes. Genome Res 19: 967-977.

154. Moriyama EN, Powell JR (1998) Gene length and codon usage bias in Drosophila melanogaster, Saccharomyces cerevisiae and Escherichia coli. Nucleic Acids Res 26: 3188-3193.