Transcription start site profiling uncovers divergent transcription and enhancer-associated RNAs in *Drosophila melanogaster*

Michael P. Meers[1,2,3], Karen Adelman[4], Robert J. Duronio[1,2,3], Brian D. Strahl[1,5], Daniel J. McKay[1,2,3] and A. Gregory Matera*[1,2,3]

[1]Curriculum in Genetics and Molecular Biology, [2]Integrative Program for Biological and Genome Sciences, [3]Departments of Biology and Genetics, [5]Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC 27599

[4]Dept. of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115

*Corresponding Author: matera@unc.edu

Running Title: The initiation landscape in *Drosophila* larvae

Keywords: Bioinformatics, Transcription initiation

Abstract

High-resolution transcription start site (TSS) mapping in *D. melanogaster* embryos and cell lines has revealed a rich and detailed landscape of both *cis*- and *trans*-regulatory elements and factors. However, TSS profiling has not been investigated in an orthogonal *in vivo* setting. Here, we present a comprehensive dataset that links TSS dynamics with nucleosome occupancy and gene expression at unprecedented sequencing depth in the wandering third instar larva, a developmental stage characterized by large-scale shifts in transcriptional programs in preparation for metamorphosis. The data recapitulate major regulatory classes of TSSs, based on peak width, promoter-proximal polymerase pausing, and *cis*-regulatory element density. We confirm the paucity of divergent transcription units in *D. melanogaster*, but also identify notable exceptions. Furthermore, we identify thousands of novel initiation events occurring at unannotated TSSs that can be classified into functional categories by their local density of histone modifications. Interestingly, a sub-class of these unannotated TSSs overlaps with functionally validated enhancer elements, consistent with a regulatory role for "enhancer RNAs" in defining transcriptional programs important for animal development. We conclude that high-depth TSS mapping is a powerful strategy for identifying and characterizing low-abundance and/or low stability RNAs.

Introduction

Transcription initiation constitutes the first step in gene expression, and thus its fidelity is of utmost importance for proper regulation of gene expression (Sainsbury *et al.* 2015). Initiation begins when the pre-initiation complex (PIC) assembles on exposed DNA at a promoter region upstream of the transcription start site (TSS) (Buratowski et al. 1989). Through the action of both active and passive mechanisms, promoters are disproportionately depleted of nucleosomes, and are thus available for PIC assembly. These mechanisms include binding of specialized transcription factors (Mitchell and Tjian 1989), activity of nucleosome remodelers (Lorch et al. 1999), and sequence-dependent likelihood of nucleosome assembly (Segal et al. 2006; Kaplan et al. 2009). The interplay of these factors is important for generating transcripts with temporal and spatial specificity (Lee and Young 2000), and for suppressing initiation

from cryptic or developmentally inappropriate sites that may otherwise be competent for initiation (Carrozza et al. 2005; Keogh et al. 2005). Regulation of initiation has important implications for cell differentiation, where activation of developmentally significant "master regulator" genes can alter gene expression regimes that define cellular morphology and identity. For instance, the expression of a handful of transcription factors associated with pluripotency is sufficient to transform differentiated cells into induced pluripotent stem cells (Takahashi and Yamanaka 2006).

Transcription initiation can be regulated at several levels. Prior to RNA polymerase II (RNA pol II) engaging with DNA at the TSS, nucleosome depletion directly upstream of the TSS facilitates assembly of the PIC and other general transcription factors. This stereotypical 'minus-1' nucleosome depleted region (NDR) is conserved across eukaryotes (Mavrich et al. 2008; Ozsolak et al. 2007), and is highly correlated with transcription initiation activity. Factors that alter the likelihood that a NDR occurs will also alter the propensity of RNA pol II to initiate at that site. Similarly, transcription factor binding to *cis* elements in the promoter results in displacement of nucleosomes. Additional descriptive characteristics of transcription initiation activity, such as the breadth or distribution of initiating polymerases across a given domain (Carninci et al. 2006; Hoskins et al. 2011), correlate with gene expression outcomes. However, it is not known whether these factors play a role in proper regulation of gene expression.

Furthermore, transcription initiation has been shown to occur in divergent directions, with unclear consequences for gene expression (Kim et al. 2010; Djebali et al. 2012). In most cases transcripts that are produced in the antisense direction relative to an annotated gene are rapidly degraded (Preker et al. 2008). Divergent transcription initiation is a common feature in mammals (Scruggs et al. 2015), and is observed across annotated TSSs and enhancer regions (Core et al. 2014). However, it is still unclear whether bidirectional transcription is functionally relevant to gene expression, particularly because certain cell types, including *D. melanogaster* S2 cells, appear to be largely devoid of divergent initiation (Nechaev et al. 2010).

A final initiation-related regulatory step occurs after PIC assembly, when RNA pol II transcribes ~50-100 nt into the gene body before it is subject to promoter proximal pausing. Pausing can act as a regulatory step to help integrate signals or it can prepare promoters for rapid activation (Muse et al. 2007; Henriques et al. 2013). Although the dynamics of polymerase pausing are well understood in cell culture (Muse et al. 2007; Nechaev et al. 2010; Henriques et al. 2013), to date there have been few studies that have comprehensively characterized pausing *in vivo* (Saunders et al. 2013).

At potential sites of transcription initiation outside of annotated TSSs, in most cases surveillance and degradation by the nuclear exosome occurs rapidly (Andersen et al. 2013; Lubas et al. 2015). This degradation is likely important because initiation at non-canonical or cryptic promoters can interfere with coding transcripts or create a deleterious load of non-functional ones, including dsRNAs (Lopez et al. 2016). In general, sites of initiation unassociated with annotated gene promoters have a high propensity for nucleosome occupancy, and are energetically unfavorable for assembly of the PIC (Segal et al. 2006; Kaplan et al. 2009). However, in the budding yeast *Saccharomyces cerevisiae*, genetic perturbations that cause cryptic initiation in coding regions are tolerated (Carrozza et al. 2005; Keogh et al. 2005; Kaplan et al. 2003). Furthermore, there is evidence that transcription from unannotated promoters may also serve beneficial functions (Verdel et al. 2004), particularly at enhancer regions (Li et al. 2016), which have been shown to produce enhancer RNAs (eRNAs) that may play regulatory roles (Kim et al. 2010; Djebali et al. 2012; Scruggs et al. 2015). Whereas cryptic and unnanotated transcription has been extensively characterized and described in *S. cerevisiae* (e.g. (Vera and Dowell 2016)), it is less well characterized in metazoans.

Here, we present a detailed characterization of matched Start-seq (Nechaev et al. 2010), ATAC-seq (Buenrostro et al. 2013), and nuclear RNA-seq datasets in *D. melanogaster* 3rd instar larvae (Meers et al. 2017). From these data, we were able to annotate larval TSSs with nucleotide resolution, and analyze connections between local cis-regulatory motifs, TSS shape, pausing activity, and divergent transcription. Additionally, we identified thousands of unannotated initiation events, and used existing datasets for histone post-translational modifications (PTMs) and validated

enhancer regions to impute their functions. Our findings are among the first to detail the global initiation patterns in a developing organism, uncovering a vast number of new initiation events that define likely enhancer RNAs and transcripts critical for animal development.

Results

Start-seq signal correlates with nucleosome depletion, gene expression, and promoter proximal pausing

To characterize the genome-wide landscape of gene expression, transcription initiation, and chromatin accessibility in third instar *Drosophila melanogaster* larvae, we carried out rRNA-depleted total nuclear RNA-seq, Start-seq, which quantifies short, capped, nascent RNAs that represent newly initiated species (Nechaev et al. 2010; Henriques et al. 2013), and ATAC-seq, which quantifies transposase-accessible open chromatin (Buenrostro et al. 2013), as previously described (Meers et al. 2017). For every annotated gene, we assigned the dominant Start-seq peak most likely to represent its bona-fide TSS from its most frequently used start site in order to cross-compare open chromatin, initiation, and gene expression values within each gene (Fig. 1A). As shown in Figure 1B, ATAC-seq signal is highest in the 150 nt upstream and 50 nt downstream of the TSS, corresponding to the expected location of a promoter-proximal NDR (Mavrich et al. 2008). Additionally, Start-seq signal accumulates robustly and almost exclusively within the ~50nt directly downstream of the assigned TSSs, consistent with expected signal distributions from previously reported Start-seq analyses (Nechaev et al. 2010). Importantly, the first nucleotide in the 5' read of each Start-seq read pair acts as a proxy for the first transcribed nucleotide in the nascent mRNA chain (Nechaev et al. 2010), enabling bona-fide TSS mapping at base-pair resolution.

Nucleosomes are barriers to transcription factor binding and PIC assembly (Ozsolak et al. 2007). Accordingly, the extent of chromatin accessibility has been shown to correlate with the level of gene expression (Ozsolak et al. 2007; Mavrich et al. 2008), and thus should correlate well with the level of transcription initiation. To evaluate these expected relationships on a gene-specific level, we used the most frequently used start site for each gene in the genome to assign a discrete value for chromatin

accessibility, transcription initiation, and nuclear RNA-seq gene expression level, and performed correlation comparisons between each pair of values across all genes. Although Start-seq intensity generally correlates well with overall gene expression, ATAC-seq levels correlate poorly with Start-seq (Fig. S2A). Curiously, both Start-seq and ATAC-seq correlate more strongly with nuclear RNA-seq than with each other (Fig. S1A), indicating a more complex relationship between transcription initiation and nucleosome depletion. Discrete partitioning of genes into quintiles based on gene expression values derived from RNA-seq signal (1st = lowest expression, 5th = highest) further confirms that the relationships between nucleosome depletion, transcription initiation, and gene expression are imperfectly correlated. For instance, the highest gene expression quintile is characterized by reduced ATAC-seq enrichment as compared to the second highest quintile, despite it having the highest enrichment in Start-seq signal (Fig. S1B). These data demonstrate that open chromatin and transcription initiation do not directly track with each other.

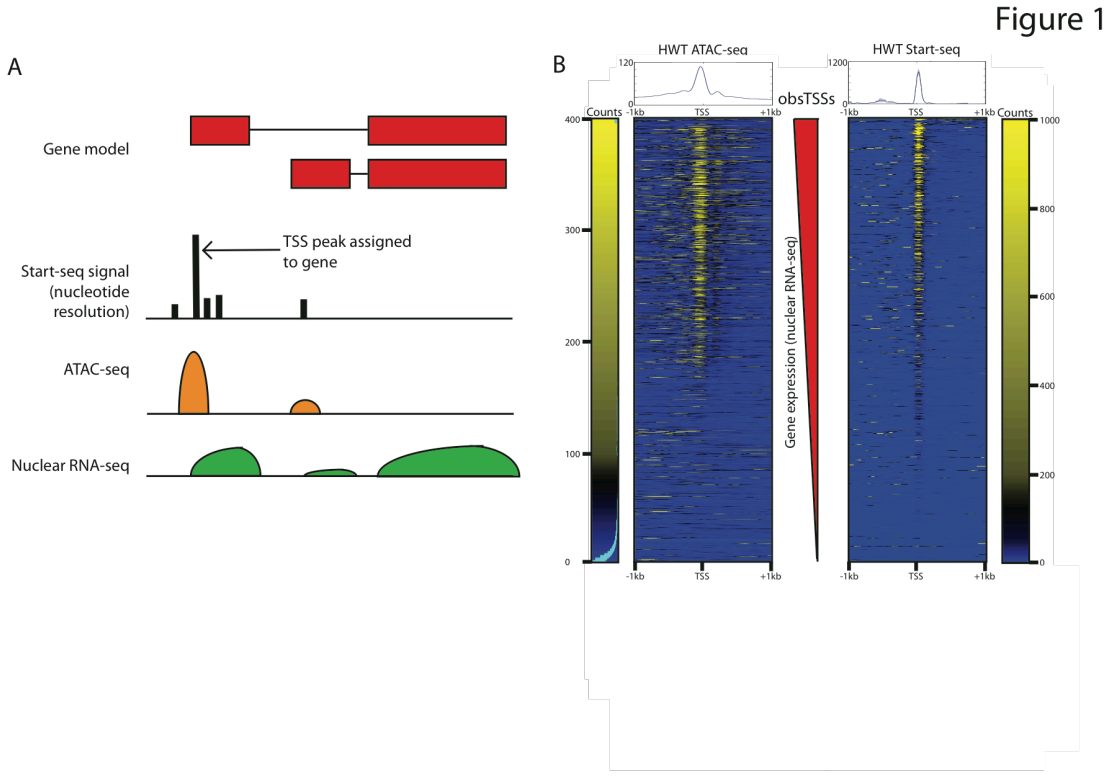

Figure 1: A) Schematic describing assignment and linkage of Start-seq, ATAC-seq, and nuclear RNA-seq within a single gene. B) Heatmap for ATAC-seq (left) and Start-seq (right) signal mapping at annotated transcription start sites (obsTSSs), ordered by increasing nuclear RNA-seq signal.

We hypothesized that discrepancies between ATAC-seq, Start-seq, and nuclear RNA-seq could be due to the influence of promoter proximal polymerase pausing on the relationship between nucleosome depletion and transcription initiation (Gilchrist et al. 2010). Specifically, we sought to test whether polymerase pausing might increase the extent of nucleosome depletion within a NDR, as inferred from MNase-seq data in S2 cells (Gilchrist et al. 2010). To evaluate the relationship between differential pausing and chromatin accessibility, we derived 'pausing index' (PI) values for each gene by determining the ratio of TSS Start-seq signal vs. gene body nuclear RNA-seq signal. Whereas Start-seq and nuclear RNA-seq levels are correlated (Fig. S1A), a scatterplot of those values for each promoter identifies significant variability from the regression line, indicating a wide range of pausing propensities (Fig. 2A). Moreover, PI can predictably stratify classes of genes that are expected to be more (or less) paused on average, based on previous studies (Adelman et al. 2009). For example, many housekeeping genes exhibit very low PI values, consistent with their ubiquitous and temporally consistent expression (example in Fig. S2A). In contrast, immune response and transcription factor genes, which in many cases are subject to rapid temporal and signal-responsive regulation that is achieved by pausing, display high PI values (Fig. 2B, example in Fig. S2B). Furthermore, enrichment of the "Pause Button" *cis*-regulatory motif, which is characteristic of many paused promoters (Hendrix et al. 2008), is positively correlated with PI quartile (Fig. S2C). We conclude that the PI metric (as calculated here) is a biologically relevant measure of gene-specific pausing propensity.

To test the relationship between pausing and chromatin accessibility, we partitioned genes into quartiles based on their PI values, and quantified ATAC-seq nucleosome depletion and predicted nucleosome occupancy in a window surrounding TSSs in those quartiles. We found that genes in the most highly paused quartile have the highest ATAC-seq signal at the minus-1 nucleosome position, despite also having the highest

Figure 2

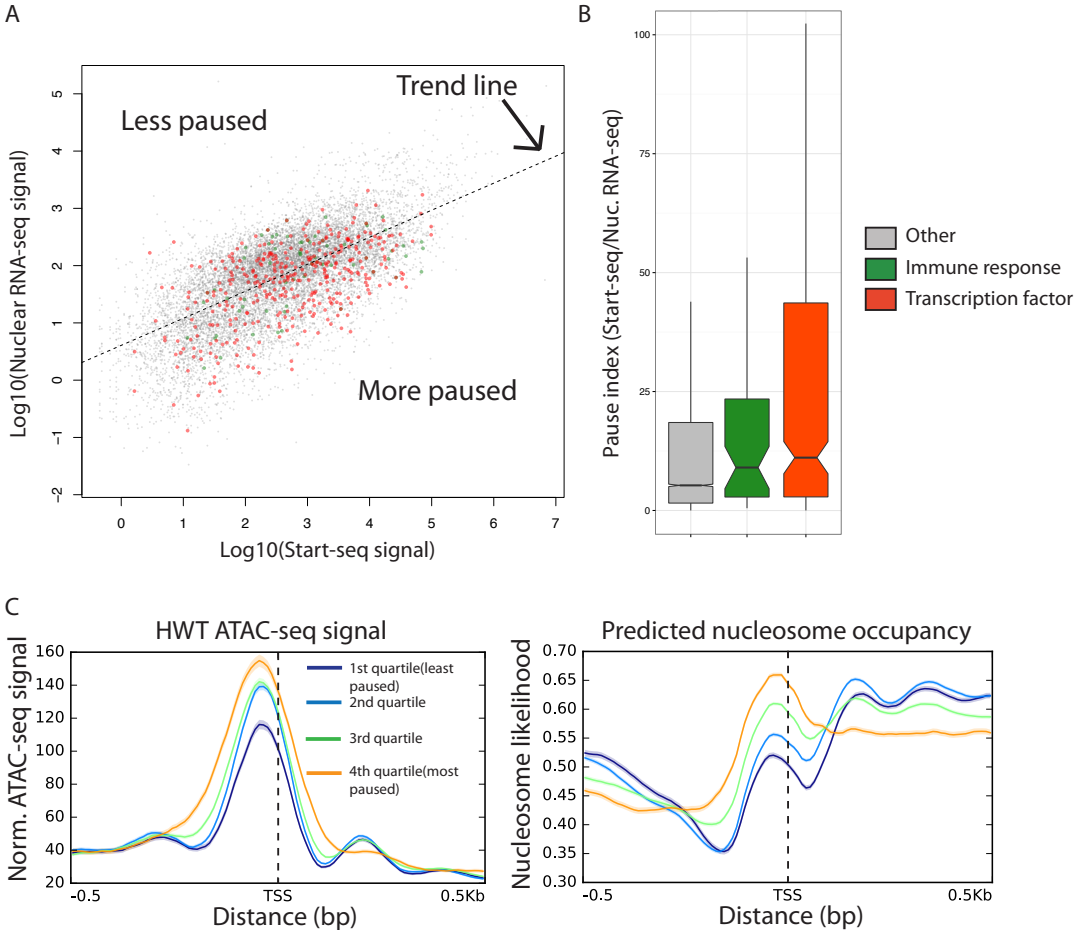

Figure 2: A) Scatterplot of Start-seq (x-axis) vs. nuclear RNA-seq (y-axis) signal. B) Pause index (Start-seq/RNA-seq) for obsTSSs from different classes of genes, including immune response and transcription factor genes (more paused than average). C) ATAC-seq signal and predicted nucleosome occupancy at obsTSSs stratified by pausing index (PI). High-PI obsTSSs have higher chromatin accessibility at the minus-1 nucleosome free region, despite also having higher predicted nucleosome occupancy in that region.

predicted nucleosome occupancy (Fig. 2C), which is consistent with previous observations (Gilchrist et al. 2010). Notably, the less-paused genes have a well-phased plus-1 nucleosome, further indicating that PI predicts the expected variability in nucleosome phasing based on pausing (Gilchrist et al. 2010; Rach et al. 2011). Using a direct assay for open chromatin in third instar larval nuclei, we conclude that polymerase pausing positively correlates with nucleosome depletion at the minus-1 nucleosome. These findings support the idea (Gilchrist et al. 2010) that pausing may play an active role in maintaining NDRs.

Start-seq signal clusters into spatially restricted groups of peaks at annotated TSSs

As observed in S2 cells (Nechaev et al. 2010), Start-seq signal often manifests as single-nucleotide peaks that are grouped in a spatially restricted region, such that TSSs can be described as "clusters" of initiation events at a handful of nucleotides near the 5' end of a gene. To illustrate these clusters, we grouped individual +1 Start-seq nucleotides into likely TSS clusters, and assigned them to annotated gene promoters. This procedure yielded 21,830 TSS clusters that matched stringent statistical criteria, 18,070 of which mapped to promoters annotated previously in the dm5.57 update of the *D. melanogaster* genome build. We termed these clusters observed promoter TSSs (obsTSSs). The remaining 3,123 high-confidence TSSs that failed to map to an annotated promoter region were considered novel unannotated TSSs (nuTSSs). Out of 12,514 "Integrated promoters" previously detected from DEEP-CAGE sequencing of *D. melanogaster* embryos (Hoskins et al. 2011), 6,847 (52.9%) overlapped with one or more of the 12,952 obsTSS clusters detected in our study, a significant fraction considering major differences between the two studies in developmental staging, library preparation, and peak detection methods (Fig. S1C).

Previous studies of *Drosophila* embryos and embryonic cell lines have revealed the presence of different TSS "shapes," as defined by the breadth of the distribution of initiation signals within a given TSS (Ni et al. 2010; Nechaev et al. 2010; Hoskins et al. 2011). Studies in mammalian cells have shown that TSS shape characteristics, particularly "sharp" and "broad" classifications, correlate with different sequence motifs enriched at the associated promoters, and different transcriptional outcomes

from the corresponding genes (Carninci et al. 2006; Rye et al. 2014). To measure TSS shape in wandering 3rd instar larvae, we measured cluster width and the fraction of total Start-seq signal in the cluster contained in the 5 nt surrounding the highest peak in the cluster. Strikingly, maximum cluster width was substantially less than those values obtained by both Hoskins *et al.* (Hoskins et al. 2011) and Ni *et al.* (Ni et al. 2010), and more in line with S2 cell data from Nechaev *et al.* (Nechaev et al. 2010) who found that initiation was more highly focused (Fig. S3A). Specifically, of TSSs that were broader than a single nucleotide, ~47% were 6 nt or fewer in width, ~75% were 10 or fewer, and ~94% were narrower than 20 nt. When we categorized TSSs as being "peaked" (<12 nt in width), "broad" (>12 nt, with > 50% signal in highest peak) and "weak" (>12 nt, < 50%), we found that 81.8% of TSSs broader than 1 nt would be classified as peaked, 8.0% as broad, and 10.2% as weak. Again, these numbers are dramatically distinct from the 32.6% peaked, 18% broad, and 49.4% weak, as described by Ni *et al.* (Ni et al. 2010). We acknowledge that the other groups employed different library preparation methods, and also relied on smoothing-density estimates for signal quantitation in peak clusters, and therefore their results may not be directly comparable. However, given the significant depth to which we sequenced our libraries, and the expectation that all peak shape modalities should be represented at that depth, we conclude that *D. melanogaster* TSSs are indeed largely "sharp" and focused, in contrast with previous findings that broad TSSs are well represented in the fruit fly transcriptome. This finding is consistent with the absence of CpG island promoters in *Drosophila*, a feature that is characteristic of non-focused TSSs in mammals (Carninci et al. 2006).

Cis-regulatory sequence motifs are enriched in patterns around obsTSSs that correlate with peak shape and polymerase pausing

Given our confirmation in larvae that *D. melanogaster* TSSs do not conform to the typical sharp/broad duality seen in mammals, we reasoned that TSS peak shape might correlate differently with sequence motifs and gene expression outcomes in flies. Therefore, we sought to generate a numeric shape metric that would allow us to elucidate relationships between TSS shape and other aspects of gene expression. To do so, we adapted the approach taken by Hoskins *et al.* (Hoskins et al. 2011) to assign a Shape Index (SI) value to each cluster (see Supplementary Methods), where higher SI

values represent “sharper” peaks (i.e. the majority of TSS signal occurring within a few nucleotides), and lower SI values “broader” peaks (i.e. signal was spread more evenly across a wider locus). Because we found TSSs to be universally sharper than previously reported (and therefore more likely to have high SI values), most promoters had an SI value between -1 and +2 (Fig. S3B). Interestingly, shape index was mildly, but positively, correlated with pause index (Fig. S3B). This finding is consistent with the idea that high SI promoters are highly enriched for the Pause button (PB) motif (Hoskins et al. 2011), and argues that SI remains a useful metric despite the narrower distribution found in our study.

To determine whether peak shape or pausing index is associated with particular sequence motifs, we searched the regions flanking obsTSSs for a suite of motifs that were previously shown to be enriched at *Drosophila* promoters (FitzGerald et al. 2006), and then clustered them based on motif enrichment (Fig. 3A). Unsupervised clustering analysis partitioned obsTSSs into three bins: a cluster characterized by the enrichment of GAGA, initiator element (INR), downstream promoter element (DPE), and PB (Cluster 1); a cluster with reduced frequency of the aforementioned motifs and a strong enrichment of TATA (Cluster 2); and a third cluster that was enriched for elements such as DRE and E-box and devoid of the other aforementioned motifs (Cluster 3). Cluster 3 had both the lowest PI and lowest SI among the three clusters (Fig. 3B, green), and is very similar to a previously identified class of low SI promoters that lack the PB motif (Hoskins et al. 2011). Cluster 1 has the highest PI (Fig. 3B), and corresponds to a class of high SI TSSs (Hoskins et al. 2011) enriched for many of the same motifs (GAGA, INR, PB). Similar to our observations in Fig. S2D (highest PI quartile), the promoters in Cluster 1 exhibited equally robust ATAC-seq signal to Cluster 3 despite a higher nucleosome occupancy expectation (Fig. 3C). Interestingly, whereas both higher PI and SI distinguish Clusters 1 and 2 from Cluster 3, #1 and #2 share similar SI values, further separating #3 as a functionally distinct class of broad TSSs with unique sequence elements.

Interestingly, the motifs enriched in Cluster 1 (and to a lesser extent Cluster 2) correlate positively with each other (Fig. S3C, dashed lines), indicating that multiple motifs tend to co-occur near the same TSS, whereas the motifs enriched in Cluster 3 are

uncorrelated and tend to be mutually exclusive (Fig. S3C). This finding suggests that sharp, highly paused TSSs are more sequence-dependent than their broader, less paused counterparts. Consistent with this hypothesis, Cluster 1 had the highest information content in its consensus sequence, implying a role for sequence in defining the characteristics of sharp TSSs (Fig S3D). Further, 67% of promoters that can be considered "sharp" in embryos remain so in larvae, whereas only 27% of peaks considered "broad" remain that way in larvae, suggesting that intrinsic cis-regulatory information contributes to sharp, but not to broad promoters. By using stages of *Drosophila* development that have not been previously analyzed, we show that

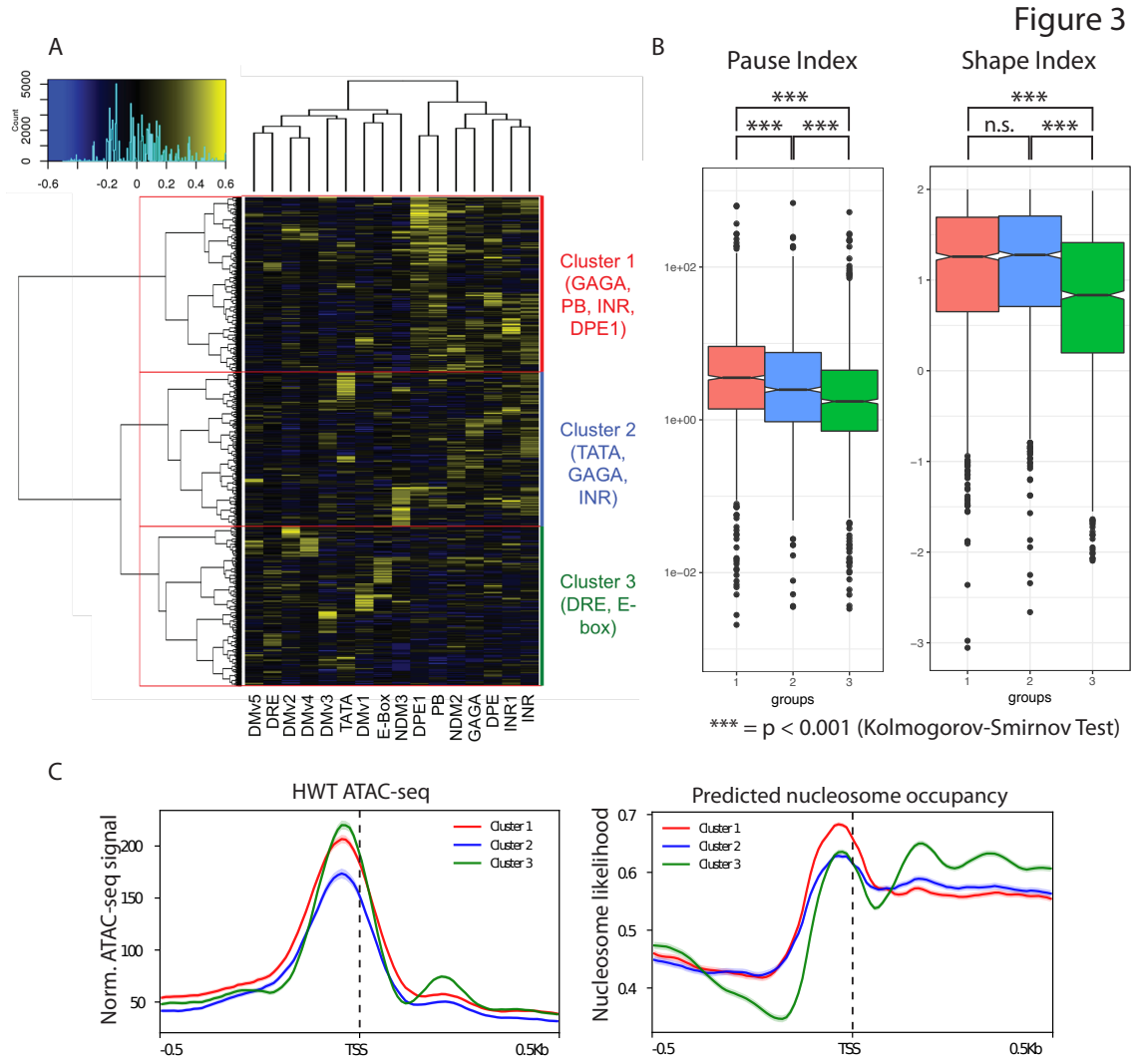

Figure 3: A) Heatmap describing enrichment of 16 motifs associated with *D. melanogaster* promoters (columns) in each TSS (rows) with more than 100 start-seq reads mapping to its dominant peak. Row clustering dendrogram partitions genes into three groups (outlined in red, described at right). B) Boxplots describing distributions of Pause Index (left) and Shape Index (right) values in TSSs belonging to Cluster 1 (red), Cluster 2 (Blue), or Clutser 3 (Green). P-values generated by Kolmogorov-Smirnov Test. C) ATAC-seq signal (left) and predicted nucleosome occupancy (right) in a 1 kb window around obsTSSs belonging to Cluster 1, 2, or 3.

transcription factor and other cis-regulatory motifs are reliably correlated with TSS shape and polymerase pausing. Taken together with previous work (Carninci et al. 2006; Ni et al. 2010; Hoskins et al. 2011), our data provide strong support for functional connections between sequence motifs, promoter-proximal pausing and TSS shape across multiple developmental points in *Drosophila*.

Divergent promoters in *D. melanogaster* larvae

"Divergent" promoters, regions from which transcription proceeds from a coupled set of core promoter elements oriented in opposite directions, have been widely reported in mammalian systems (Kim et al. 2010; Djebali et al. 2012; Scruggs et al. 2015). Antisense transcription from non-coding TSSs is thought to have regulatory consequences for expression of the sense-oriented protein-coding gene (Scruggs et al. 2015). However, there is very little evidence of the same phenomenon in *D. melanogaster* (Nechaev et al. 2010), though to date it has not been analyzed *in vivo* in an organismal context. Using our high-depth 3rd instar larval Start-seq dataset, we searched for divergent transcription units. For a given TSS, we mapped the fraction of Start-seq signal accumulating in sense and antisense directions relative to each site in question. We found that both obsTSSs and nuTSSs exhibited highly sense-oriented signal (Fig. S4A). We also quantified sense-oriented reads as a proportion of the total reads mapping in a 200 nt window on either side of each TSS. Although obsTSSs were statistically more enriched for sense-oriented reads than were nuTSSs, the mean was greater than 90% sense-oriented for both cohorts, indicating a high degree of unidirectionality at both *Drosophila* coding and non-coding TSSs (Fig. S4B).

To identify divergent TSSs, we aligned all high-confidence TSSs whose nearest neighboring TSS was oriented in the opposite direction, then ordered them based on genomic distance between each pair, and plotted ATAC-seq signal in order to identify single NDRs housing the two TSSs. A threshold distance of roughly 200 nt between the TSS pair yielded 537 pairs of TSSs for which a single continuous ATAC-seq NDR overlapped both TSSs (Fig. 4A). These 537 pairs contained 1,023 distinct TSSs, or ~4.8% of the all the high-confidence TSSs queried (Fig. S4C, example in Fig. 4B), as compared with greater than 75% of active promoters observed with divergent transcription in mammalian systems (Scruggs et al. 2015). Despite the dearth of

Figure 4

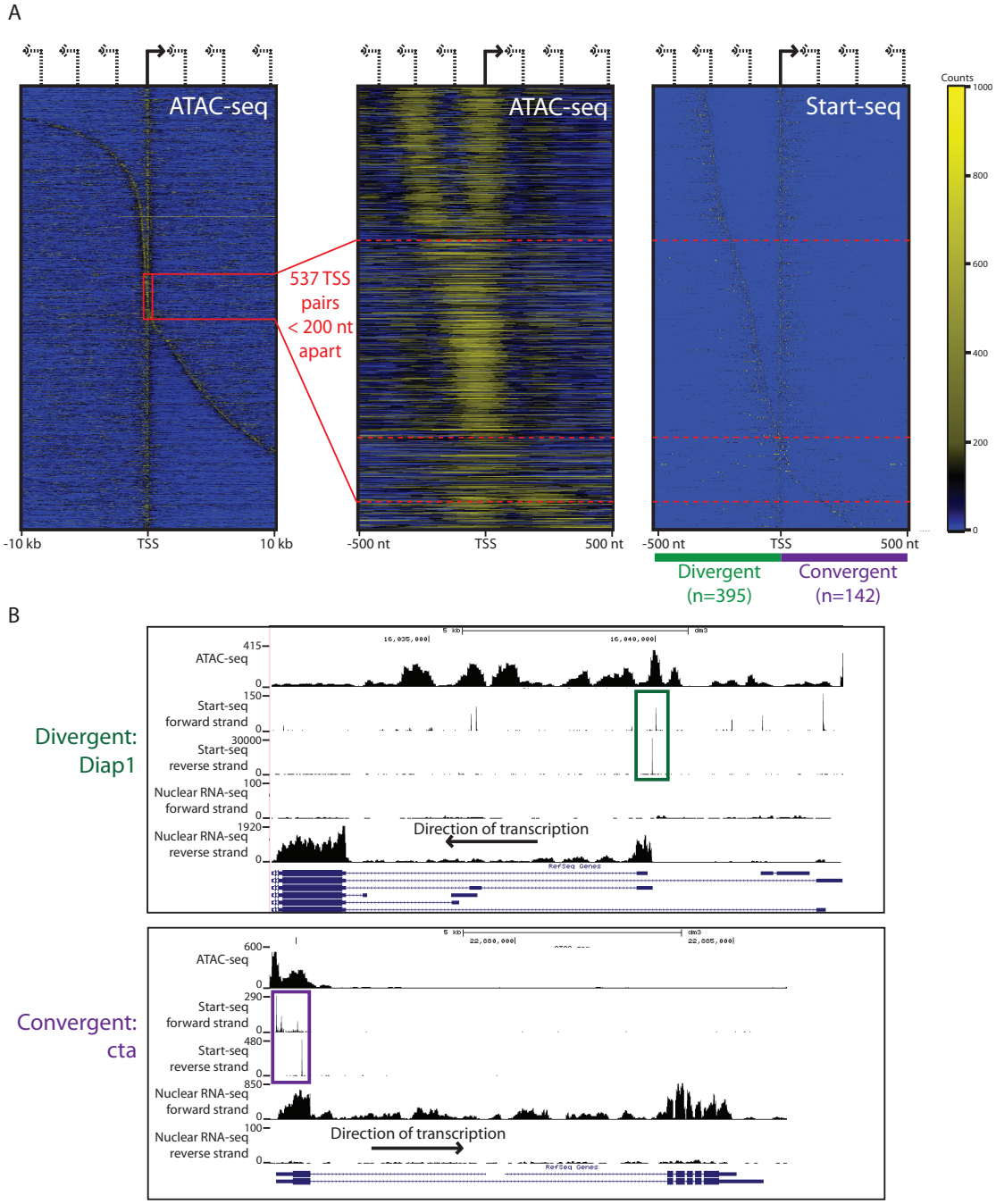

Figure 4: A) Panel 1 (left): Heatmap of ATAC-seq signal mapping in a 20 kb window around TSSs whose nearest neighbor TSS is oriented in the opposite direction. TSSs are ordered by distance between TSS pair. Panel 2 (center): Heatmap of ATAC-seq signal mapping in a 1 kb window around TSS pairs separated by less than 300 nt. TSS pairs separated by less than 200 nt (highlighted by red box in panel 1) are partitioned into divergent or convergent TSS pairs by red dashed lines. Panel 3 (right): Heatmap of Start-seq signal for TSSs exhibited in panel 2. B) Representative browser window examples of divergent (top) and convergent (bottom) obsTSS pairs.

paired TSSs genome-wide, of those that were paired, 444 (43.4%) were obsTSSs paired with another obsTSS from separate, divergent coding genes, which we refer to as bidirectional promoters (Fig. S4C). This is in contrast with the human transcriptome, in which only a small proportion of divergent transcription is represented by bidirectional promoters (Trinklein et al. 2004).

Divergent transcription might aid in recruiting transcription initiation machinery and maintaining a robust NDR at active promoters. To determine whether nucleosome depletion is increased at sites of divergent initiation in *Drosophila*, we quantified ATAC-seq reads in a 200 nt window around obsTSSs participating in bidirectional, divergent, or non-divergent initiation, and compared it to Start-seq levels (see Fig. S2A). We found that bidirectional obsTSSs had significantly more ATAC-seq signal than divergent or non-divergent nuTSSs, despite the expectation that ATAC-seq would correlate with the lower level of Start-seq signal at bidirectional obsTSSs (Fig. S4D). However, it is known that bidirectional promoters are often separated by the BEAF32 insulator in *Drosophila* (Yang et al. 2012), and indeed bidirectional promoters were enriched for BEAF32 ChIP-seq signal relative to divergent and non-divergent TSSs (Fig. S4E). Therefore, we could not rule out the possibility that increased ATAC-seq signal at bidirectional promoters may be due to displacement of nucleosomes by BEAF32. Importantly, divergent and non-divergent obsTSSs exhibited similar levels of BEAF32 ChIP-seq signal (Fig. S4E), and negligible differences in ATAC-seq signal (Fig. S4D), indicating that divergent transcription is generally insufficient to enforce a more robust NDR than would be expected by initiation activity in *D. melanogaster*. This observation is consistent with the finding that RNA pol II and H3K4me3 ChIP-seq accumulation is similar between directional and divergent TSSs in S2 cells (Core et al. 2012).

Strikingly, 142 of the TSS pairs we identified were not divergent, but rather were oriented towards each other (Fig. 4A, example in Fig. 4B). We termed these "convergent" pairs, and they included 86 obsTSSs converged on by nuTSSs, and 16 pairs of obsTSSs that converged and productively elongated in both directions. In general, the distance between convergent pairs of TSSs was much larger than that of divergent pairs, indicating selection against convergent transcription in close genomic proximity (Fig. S4F). These results are consistent with the characteristics of convergent

initiation pairs detected in mammalian cell culture (Mayer et al. 2015), and confirm the presence of convergent transcription in an *in vivo* context. However, similarly to other studies, we cannot rule out the possibility that convergent transcripts originate from distinct cell populations. We conclude that, within a broader regime of unidirectionality, several *D. melanogaster* TSSs represent striking exceptions.

Novel unannotated TSSs (nuTSSs) are widespread and can be partitioned into predicted functional categories based on local histone modifications

Owing to the depth of our Start-seq libraries (>100M mappable reads combined), we were able to identify bona-fide Start-seq signal at thousands of locations across the genome that did not correspond to an annotated TSS. To systematically analyze these locations, we applied several metrics to all peaks that did not fall in a TSS cluster that matched to existing observed TSS (obsTSS), which we dubbed novel unannotated TSSs (nuTSSs). We identified a total of 11,916 distinct nuTSSs, including 3,123 that met an average 9 read false-discovery rate (FDR) threshold within every biological replicate. In general, nuTSSs exhibit NDRs comparable in shape to those found at obsTSSs, despite residing at loci with a higher intrinsic likelihood of nucleosome occupancy (Fig S5A). nuTSSs are spread throughout the genome, though they cluster predominantly at locations within or proximal to annotated coding genes (Fig. S5B).

A handful of well characterized histone post-translational modifications (PTMs) co-localize with bona-fide TSSs, and therefore we surmised that enrichment of particular histone PTMs at nuTSSs might provide an indication of nuTSS functions. Therefore, we measured the enrichment of a battery of histone PTMs at nuTSSs, and conducted unsupervised hierarchical clustering. Out of several modEncode ChIP-seq tracks (www.modencode.org) taken at a matched stage of development, we found that the most informative set of PTMs included H3K4me1, H3K4me3, H3K27ac, and H3K36me3. Hierarchical clustering based on these four marks resulted in seven categories (Fig. 5A). We characterized them as follows: a “Featureless” group (Cluster 5) lacking significant enrichment in any of the four marks, “TSS-like” groups (Clusters 1, 2, and 6) with enrichment for H3K4me3 characteristic of annotated coding gene start sites (Liang et al. 2004), two “Coding” cohorts (Clusters 3 and 4) characterized by varying levels of enrichment for H3K36me3 as is expected in gene bodies bodies

Figure 5

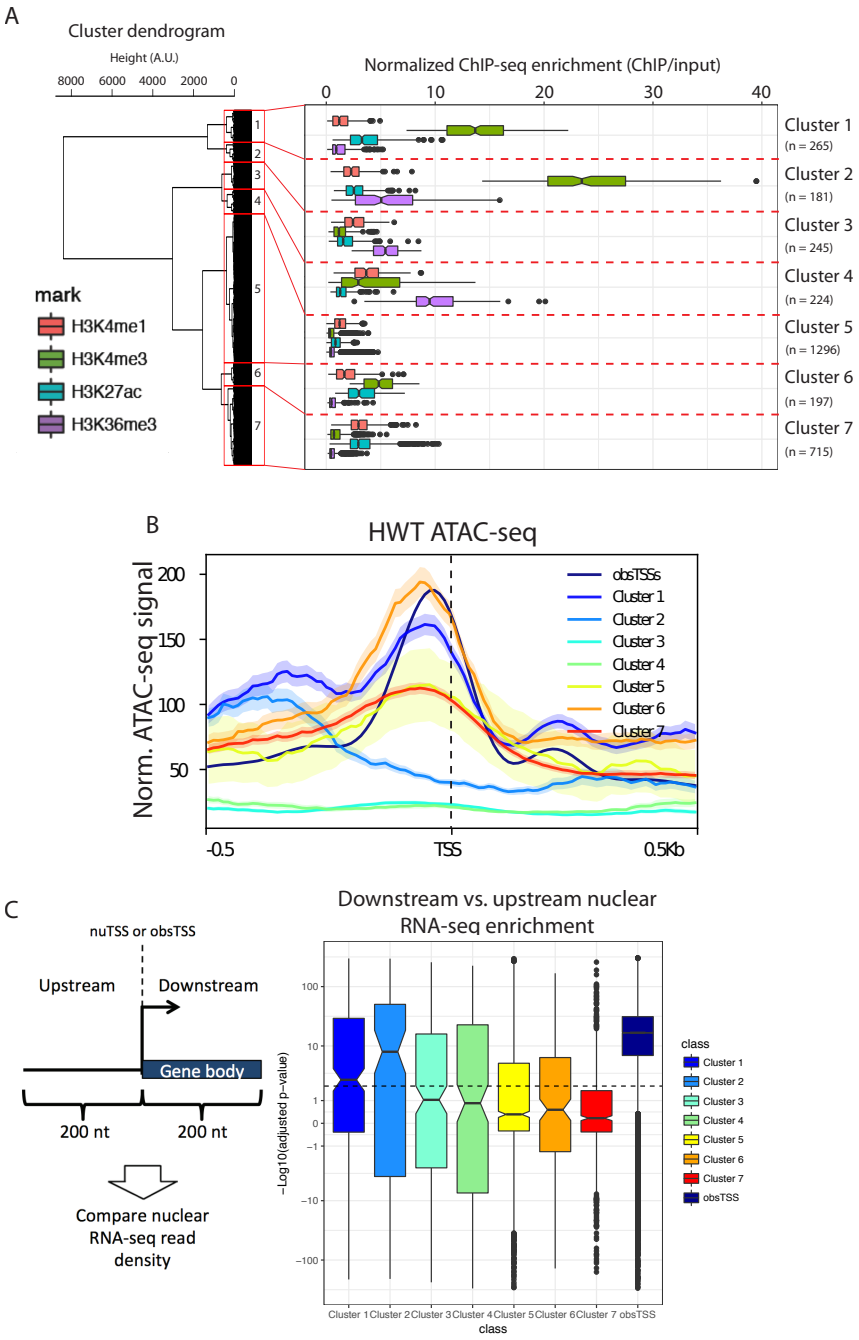

Figure 5: A) Clusters of nuTSSs based on enrichment of H3K4me1 (red), H3K4me3 (green), H3K27ac (teal), and H3K36me3 (purple). The number of nuTSSs in each cluster is indicated at the right of the plot. B) ATAC-seq signal at obsTSSs (dark blue) and nuTSSs from different histone PTM-based clusters. C) Nuclear RNA-seq reads were mapped to regions 200 nt upstream and downstream of each TSS, and enrichment of downstream vs. upstream reads was analyzed as a proxy for elongation. At right: boxplot of –Log10-transformed adjusted p-values for downstream signal enrichment over upstream within each nuTSS cluster (downstream-enriched nuTSSs above 0, upstream-enriched nuTSSs below 0).

and an "Enhancer-like" group (Cluster 7) with enrichment of H3K4me1 and H3K27ac marks characteristic of enhancer regions (Creyghton et al. 2010; Rada-Iglesias et al. 2011; Zentner et al. 2011). We further validated these functional classifications by observing that the "active" cohorts (TSS-like and Enhancer-like) were generally accompanied by strong ATAC-seq signal comparable to that of obsTSSs, whereas the other cohorts were depleted of ATAC-seq signal (Fig. 5B). Strikingly, the majority of nuTSSs clustered into "Featureless" (1296/3123, 41.5%) or "Enhancer-like" (715/3123, 22.9%) cohorts that lacked H3K4me3, indicating the prevalence of transcription initiation events that may serve functions other than transcription of as-yet unannotated genes. Together, these findings strongly suggest that nuTSSs localize within functionally relevant chromatin contexts across the genome.

Most nuTSSs do not produce stable transcripts

Although high-throughput sequencing methods have enabled extensive and detailed annotation of global transcriptomes, studies that employ ever higher depth and sensitivity are continuing to uncover previously undiscovered RNAs. We reasoned that the high depth of our Start-seq libraries may have identified initiation sites for unannotated genes that undergo productive elongation and produce mature, stable transcripts, particularly for nuTSSs associated with PTM-based clusters enriched for H3K4me3. To determine whether this was the case, we mapped nuclear RNA-seq reads in 200 nt windows upstream and downstream of all nuTSSs, and measured the balance of signal on either side, reasoning that productive elongation would be identified by an overrepresentation of downstream reads. We confirmed this hypothesis by performing the same test on obsTSS, and found that they are universally enriched for RNA-seq reads in the downstream region over the upstream (Fig 5C, Fig. S5C). In contrast, nearly all nuTSSs showed no significant enrichment of downstream signal (Fig. S5C). This trend held for most of the histone PTM-defined nuTSS clusters, where all but Clusters 1 and 2 had a mean adjusted p-value of downstream signal enrichment that was below the minimum threshold for significance. (Fig. 5C, dashed line). This observation suggests that nuTSSs generally are not converted into mature, stable transcripts.

Because Clusters 1 and 2 were most biased towards downstream vs. upstream read density, and were both highly enriched for H3K4me3, we investigated them more closely to determine whether their constituents represented unannotated "canonical" TSSs that resulted in elongating mRNAs. Upon closer examination of Cluster 2 nuTSSs, we found that they were devoid of local nucleosome depletion immediately surrounding the nuTSS, but curiously exhibited a strong ATAC-seq peak 250-500 nt upstream of the nuTSS (Fig. 5B). This corresponded precisely with the fact that almost all Cluster 2 nuTSSs are found 200-500 nt downstream of an obsTSS (Fig. S5B), leading us to conclude that Cluster 2 nuTSSs likely represented spurious products resulting from inefficient degradation of uncapped RNAs. Meanwhile, Cluster 1 uncovered distinct cases wherein the updated dm6 *D. melanogaster* genome build annotated new first exons that were overlapped by Cluster 1 nuTSSs derived from an earlier build (example in Fig. S5D). When we intersected our nuTSSs with 5'UTR regions converted from the most recent genome build, 37 nuTSSs in Cluster 1 overlapped, suggesting several Cluster 1 nuTSSs in fact correspond with coding gene initiation. Overall, we find that nuTSSs do not elongate into stable transcripts, with few exceptions corresponding to newly identified coding gene start sites.

nuTSSs enriched for enhancer-associated chromatin marks overlap with functionally validated, tissue-specific enhancers

Recent studies in mammalian model systems have reported the presence of short-lived transcripts (eRNAs) originating from developmentally-regulated enhancers (Kim et al. 2010; Djebali et al. 2012; Scruggs et al. 2015). These findings are evocative of regulatory non-coding RNAs (ncRNA) at enhancer regions in *Drosophila* (Lipshitz et al. 1987). As mentioned above, classification of nuTSSs by histone PTMs showed that Cluster 7 is distinguished by local enrichment of H3K4me1 and H3K27ac, both of which are hallmarks of active enhancers (Fig. 5A). We therefore termed nuTSSs belonging to Cluster 7 "Enhancer-like nuTSSs" (E-nuTSSs). As with previously reported eRNAs, E-nuTSSs are associated with robust NDRs (Fig. 5B), suggesting assembly of the PIC similar to "canonical" promoters. E-nuTSSs predominantly appear in intronic sequence (~74%), whereas only ~12% occur in coding sequence (Fig. S6A), which is suggestive of low sequence conservation and perhaps more recent cis-element evolution. In agreement with this interpretation, the information content of the E-nuTSS consensus

sequence is low, and comparable to that of lowest information cohort of obsTSSs that we analyzed (Fig. S6B).

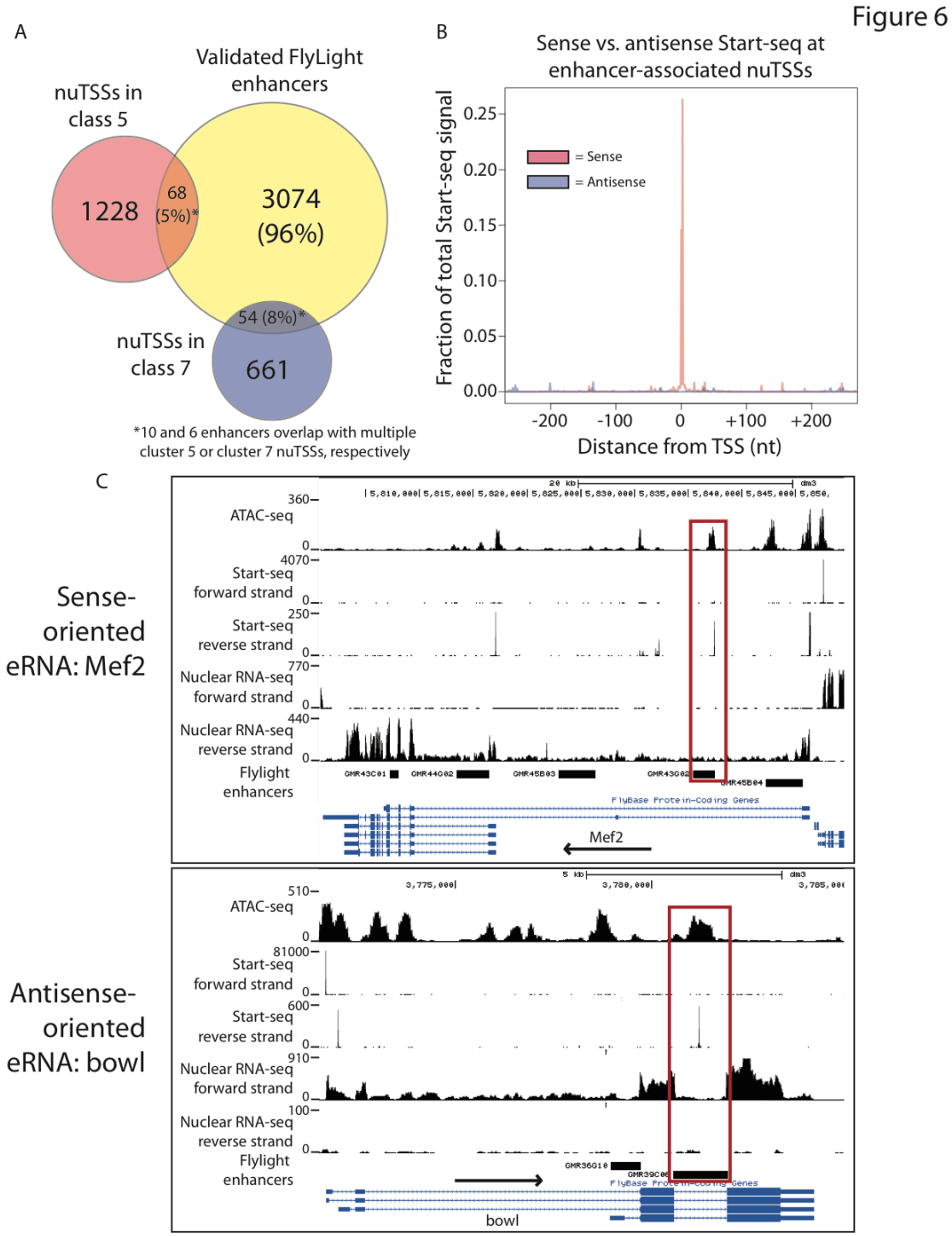

Figure 6: A) Overlap between Flylight enhancers validated for expression in larval CNS or imaginal discs (yellow, 3179) and nuTSSs from Cluster 5 (red, 1296) or cluster 7 (blue, 715). B) Metaplot of sense (red) or antisense (blue) Start-seq signal mapping in a 500 nt window around enhancer-associated nuTSSs. C) Representative browser window examples of enhancer-associated nuTSSs oriented in the sense (top) or antisense (bottom) direction relative to the gene containing the enhancer.

If E-nuTSSs represent bona-fide eRNAs, they would dramatically expand the ensemble of known regulatory RNAs in *D. melanogaster*, a model system wherein eRNAs have not been characterized systematically. To determine whether E-nuTSSs overlap with functional enhancer regions, we curated *D. melanogaster* 3rd instar larval enhancers from the FlyLight collection (Jenett et al. 2012; Jory et al. 2012). These enhancers have been functionally validated via GAL4-UAS based screening, and we chose 3,179 of the 7,113 total enhancers in the collection on the basis that they were shown to promote expression of a fluorescent reporter protein in either larval imaginal discs or in the larval CNS. Among the 3,123 high-confidence nuTSSs interrogated, we found that 135 unique high-confidence nuTSSs overlapped directly with one of the 3,179 validated enhancer regions, and that 116 enhancers contained at least one nuTSS. These numbers represented ~4.4% and ~3.6% of the populations queried, respectively (Fig. 6A). Importantly, 90% (122/135) of the nuTSSs that overlap with an enhancer belonged to either Cluster 5 (68, 5% of all Cluster 5 nuTSSs) or Cluster 7 (54, 7.6%). When Cluster 7 regions were randomized throughout the genome prior to measuring overlap with enhancer regions *in silico*, fewer of the resultant shuffled nuTSSs overlapped with enhancers than did Cluster 7 nuTSSs in 94% percent of 5000 random trials (Fig. S6C). Thus histone PTM-derived clustering analysis is useful for identifying functional TSSs.

The low percentage of total enhancers detected by nuTSSs indicates that nuTSSs alone are unlikely to be useful as a tool for predicting enhancers as compared with other methods. For instance, open chromatin data has been used to detect *D. melanogaster* tissue-specific enhancers (McKay and Lieb 2013). We therefore used 38,696 non obsTSS-overlapping ATAC-seq peaks to measure overlap with enhancers, and found that 3,537 peaks (~9.1%) overlapped with 1,761 known enhancers (55% of total enhancer set) (Fig. S6D). We conclude that *D. melanogaster* tissue-specific enhancers are generally characterized by nucleosome depletion, and (at present sequencing depth) only a small fraction of these NDRs are associated with eRNAs.

nuTSSs associate with broadly expressing enhancers and represent unidirectional eRNAs

We hypothesized that eRNAs might only occur at the strongest enhancer regions (i.e. those active in the largest number of cells), so we examined the tissues in which nuTSS-containing enhancer regions were reported to express. Relative to the 3,063 enhancers not containing a high-confidence nuTSS, nuTSS-containing enhancers were enriched for expression in all five imaginal disc categories (leg, wing/haltere, eye, antennal, and genital), and in the optic lobe, all of which represent large, broad-based cell populations (Fig. S7A). Similarly, nuTSS-containing enhancers were depleted for expression in brain-, thoracic-, and subesophageal-specific neurons and neural lineages, all of which are small subcellular populations requiring highly specific regulatory elements (Fig. S7A). Furthermore, 26% (30/116) of nuTSS-associated enhancers express in 5 or more distinct tissues as compared with 8.3% (253/3063) of other enhancers, further suggesting that eRNAs are associated primarily with broadly expressing enhancers in *D. melanogaster*. From this hypothesis, we reasoned that lowering the threshold for nuTSS detection might uncover more enhancer regions that are not expressed as broadly. Therefore, we intersected the remaining 8,793 nuTSSs that met a minimum statistical threshold for detection, but not our more stringent threshold of at least 9 reads across all biological replicates, with the base set of 3,179 enhancers. The low-stringency nuTSSs overlapped with 53 of the 116 previously detected enhancers and 456 previously undetected enhancers, resulting in 572 total enhancers overlapping directly with a nuTSS (~18% of the base set). Of the newly detected enhancers, only ~15.1% (69/456) expressed in five or more distinct tissues, indicating that lower-usage eRNAs are more likely to overlap with enhancers that are active in fewer cells.

Although eRNAs in mammalian cells are typically divergently transcribed (Kim et al. 2010; Djebali et al. 2012; Scruggs et al. 2015; Dorighi et al. 2017), we find no evidence of divergent transcription from validated enhancers that overlap with a nuTSS, indicating that eRNAs in *D. melanogaster* conform to the unidirectional character of obsTSSs (Fig. 6B, Fig. S7B). Of the 135 high-confidence nuTSSs that overlap with validated enhancers, only 9 of them participate in a divergent transcription pairing. Intriguingly, unidirectional eRNAs can be oriented in either the sense or antisense

direction relative to their resident gene (75 sense vs. 51 antisense, examples in Fig. 6C). In summary, we conclude that eRNAs in *Drosophila* are unidirectional and correlated with enhancer strength.

Discussion

Fidelity of transcription initiation is crucial for proper regulation of gene expression. There are several characteristics of transcription from annotated promoters that are thought to correlate with aspects of downstream gene expression, including promoter nucleosome depletion (Ozsolak et al. 2007), interaction of initiation complexes with cis-regulatory motifs, start site "shape" (Carninci et al. 2006), and promoter-proximal pausing (Muse et al. 2007; Gilchrist et al. 2010; Nechaev et al. 2010). To analyze transcription initiation in *D. melanogaster*, we performed matched ATAC-seq, Start-seq, and nuclear RNA-seq in larvae. For accurate developmental staging, we selected animals displaying the wandering behavior that is characteristic of the late 3rd instar. We elucidated regulatory trends for initiation at annotated coding genes genome wide that are largely in agreement with previous studies, and also uncovered myriad sites of unannotated transcription.

Relationship between peak shape and transcriptional outcomes

Previous TSS mapping studies in *Drosophila* (Nechaev et al. 2010; Ni et al. 2010; Hoskins et al. 2011) and other organisms (Carninci et al. 2006) noted that the overall shape of a TSS domain had potential functional implications. We find that *D. melanogaster* larval TSS peaks are typically very sharp and focused. We also find that broad TSS tend to occur at highly expressed, lowly paused genes, whereas sharp peaks are highly paused and enriched for a host of cis-regulatory elements. Importantly, the average width of TSSs we detect from larvae is in contrast with TSSs from embryos, in which broader TSSs are more prevalent (Ni et al. 2010; Hoskins et al. 2011). We primarily attribute this difference to the library preparation techniques used in the various studies (Start-seq (this study); PEAT (Ni et al. 2010); and CAGE in prior work (Hoskins et al. 2011)). For instance, the PEAT and CAGE methods only capture 5' ends that are associated with mature or actively elongating transcripts, and therefore might be biased towards a subset of TSSs that complete transcription, whereas the Start-seq

method captures all newly initiated transcripts within the first 120 nt of their elongation (Nechaev et al. 2010). Despite the fact that the average width of our obsTSSs is considerably smaller than the minimum width detected using other methods (e.g. Hoskins et al. 2011), we nevertheless were able to stratify TSSs into functional categories based on the width and distribution of reads within a given domain or peak. Thus, it is likely that *D. melanogaster* TSSs are similarly "sharp" across all developmental stages and cell types.

Interestingly, the broad TSSs that we infer to lack stably paused Pol II could be considered analogous to similarly broad mammalian housekeeping promoters that are enriched for CpG islands (Rye et al. 2014). Given the absence of CpG island promoters in *Drosophila*, this finding suggests a convergent evolutionary force that promotes a broad modality of transcription initiation specifically at ubiquitously expressed genes. Although direct comparisons of peak width across sequencing platforms are difficult, most of the peaks that are considered "sharp" in embryos are also considered sharp in larvae, whereas the shape of broad peaks is less well conserved. Combined with the more consistent sequence context of sharp promoters (Fig. S3D), it is possible that sharp peaks were selected during evolution to behave as such because of the necessity of promoter proximal pausing-related regulation of their expression, whereas broad peaks require no such constraints, and are instead driven by their strong propensity for nucleosome depletion. Whether or not TSS peak shape is a functional characteristic upon which natural selection can act is speculative and will require further study.

The role of promoter directionality in *D. melanogaster*

Previous studies have found that *D. melanogaster* promoters are highly unidirectional (Nechaev et al. 2010), despite voluminous data that argues for intrinsic bidirectionality of promoters in mammalian systems (Preker et al. 2008; Kim et al. 2010; Djebali et al. 2012; Scruggs et al. 2015). However, the unidirectional character of *Drosophila* TSSs has not been evaluated carefully in endogenous tissues. Here we find that similar to previous studies, TSSs lack antisense transcription initiation for the vast majority of promoters in wandering 3rd instar larvae. However, we also uncover hundreds of new cases of divergent transcription, which we conservatively define as transcription initiating in two directions from the same contiguous nucleosome free region (NDR).

Notably, more than half of the divergent TSS pairs we identified correspond to a bidirectional promoter pair in which annotated genes are transcribed in opposite directions from the same promoter region. There is evidence of this phenomenon in the human transcriptome (Trinklein et al. 2004; Wakano et al. 2012), but it represents a small proportion of all divergent TSSs.

It has been suggested that divergent transcription may serve to strengthen the recruitment of transcription factors and other initiation machinery to the site of the sense-directed gene, thereby increasing its expression (Duttke et al. 2015). Though we detect much higher ATAC-seq signal at bidirectional obsTSSs than at unidirectional obsTSSs (Fig. S4E), it is unclear whether this is due to a synergistic effect of coordinated recruitment of transcriptional machinery, whether it reflects the fact that more cells are initiating from one TSS over the other, or whether the complex effects of insulator binding upon observed ATAC-seq signal are at play.

Identification and characterization of enhancer RNAs

We showed that around 18% of validated larval enhancers from the Janelia FlyLight collection (Jenett et al. 2012; Jory et al. 2012), also overlapped with a nuTSS peak identified in our Start-seq experiments. To our knowledge, no exhaustive post-hoc functional validation of a set of predicted enhancers has even been undertaken. Hence, it is unclear whether our findings are comparable to other genome-wide approaches used to identify enhancers that have used enrichment of histone post-translational modifications (PTMs) (Creyghton et al. 2010; Zentner et al. 2011), transcription factor binding sites (Hallikas et al. 2006), or regions of open chromatin or DNase hypersensitivity (McKay and Lieb 2013) to identify enhancers.

From our ATAC-seq data, we find that NDRs are more successful at detecting validated enhancers than nuTSSs alone. From a practical perspective, using nucleotide-resolution nuTSSs to identify developmental enhancers may have the further benefit of improving the resolution of enhancer identification. Though enhancer length is variable, and the functional fraction of a given enhancer is undoubtedly longer than the spatially restricted region identified by nuTSSs, single-nucleotide resolution provides a helpful starting point for defining minimal sequence requirements for activation. It

remains to be seen whether sequencing nuTSSs to higher depth could reliably identify spatially restricted enhancers. Future studies that carefully benchmark enhancer detection methods with experimentally validated enhancers will be instructive in this regard. Furthermore, it is possible that incorporating nuTSSs with other existing methods of enhancer detection, such as nucleosome depleted regions, may improve our ability to identify novel enhancers, particularly those that are active broadly in several tissues within a complex mixture of cells. Indeed, our annotation of nuTSSs with their enrichment for enhancer-associated histone PTMs already partially achieves this goal.

Assaying eRNAs in *D. melanogaster*

From the perspective of *D. melanogaster* eRNA function, it is important to note that ~96% of validated enhancers lacked high-confidence eRNAs. This may be due to the small number of validated enhancers we were able to investigate, or the highly stage- and cell-type specific expression of enhancer RNAs, that might not be captured in our study. It is unclear whether to conclude from these data that eRNAs are neither a universal nor a necessary feature of developmental enhancers, or whether the depth of sequencing was not sufficient, or the mixture of cell types too heterogeneous, to detect eRNAs from the remainder of enhancers. Nonetheless, we found that an additional 456 validated enhancer regions overlapped with identified nuTSSs that did not meet our most stringent threshold (for a total of 572, 18% of the enhancer regions queried), suggesting that, in the absence of technical constraints to our methods, more eRNAs might be detected.

Strikingly, we found that nuTSSs present within annotated enhancers are strongly unidirectional. In contrast, previous reports showed that, in S2 cells, putative eRNAs associated with computationally predicted intronic enhancers were significantly more divergent than active promoters (Core et al. 2012). Because the eRNAs we identified are associated with validated enhancer regions and are not enriched for downstream relative to upstream nuclear RNA-seq signal, we can be confident that our results are not significantly confounded by unannotated promoters. Therefore, either unidirectionality is a true characteristic of *D. melanogaster* eRNAs or perhaps the decay of one of the two presumptive eRNAs is much more rapid than the other. Nevertheless,

the data point to a novel modality for eRNA genesis and function, perhaps distinct from the existing hypothesis that eRNAs originate from "underdeveloped" promoter regions that have not yet accumulated the cis-regulatory elements necessary to discriminate against antisense transcription (Li et al. 2016).

Materials and methods

RNA library preparation and sequencing. For all libraries, nuclei were isolated from whole 3rd instar *D. melanogaster* larvae as previously described (Meers et al. 2017). For Nuclear RNA-seq and Start-seq, RNA was extracted from isolated nuclei using TRIzol reagent (Thermo Fisher). Start-seq libraries were prepared from nuclear RNA as previously described (Henriques et al. 2013; Nechaev et al. 2010), and were sequenced on a NextSeq500 generating paired-end, 26 nt reads. For nuclear RNA-seq, Total nuclear RNA was used as input to Ribo-zero Stranded RNA-seq library preparation (Illumina). Four biological replicates were prepared for Start-seq and nuclear RNA-seq. Libraries were sequenced on a HiSeq2000 generating paired-end, 50nt reads (Illumina).

ATAC-seq library preparation and sequencing. ATAC-seq libraries were prepared as previously described (Meers et al. 2017). For each replicate, nuclei from 10 whole 3rd instar larvae were isolated as per Start-seq and nuclear pellets were gently homogenized with wide-bore pipette tips in 50 uL ATAC-seq lysis buffer (10 mM Tris·Cl, pH 7.4, 10 mM NaCl, 3 mM $MgCl^2$, 0.1% (v/v) Igepal CA-630). Homogenate was directly used as input to the Nextera DNA library preparation kit (NEB) for tagmentation of chromatinized DNA, as described in Buenrostro *et. al.* (Buenrostro et al. 2013). Three biological replicates were prepared. Libraries were sequenced on a HiSeq2000 generating single-end, 50 nt reads (Illumina).

Bioinformatic analysis. All raw data (fastq files) from ATAC-seq, Start-seq, and nuclear RNA-seq are available in the Gene Expression Omnibus (GEO) archive at NCBI under the following accession number: GSE96922. All ChIP-seq data were downloaded from modEncode (www.modencode.org). In all cases where possible, data were derived from the 3rd instar larval time point as determined by modEncode developmental staging procedures. GEO accession numbers for modEncode data used

in this study are as follows: H3K4me1: GSM1147325-28; H3K4me3: GSM1200083-86; H3K27ac: GSM1200071-74; H3K36me3: GSM1147189-92; H3: GSM1147289-92; BEAF32: GSM1256853-56. All histone PTM ChIP-seq data was normalized to H3 ChIP-seq data using the Deeptools bigwigCompare utility. Predicted nucleosome occupancy data was obtained from genome-wide nucleosome prediction tracks in *D. melanogaster* generated by the Eran Segal laboratory (https://genie.weizmann.ac.il/software/nucleo_genomes.html). For ChIP-seq, ATAC-seq, and predicted nucleosome occupancy, metagene plots were generated using the Deeptools package (Ramírez et al. 2014). All browser screenshots were captured from the UCSC Genome Browser (Kent et al. 2002).

Start-seq peak assignment. Reads from start-seq FASTQ files were clipped to the first 26 nt to remove adapter sequence, then mapped to the *D. melanogaster* dm3 genome build with Bowtie2 (Langmead and Salzberg 2012). We used a custom script to quantify mapped Start-seq read density at individual nucleotides. Briefly, we parsed the SAM alignment ouput files from Bowtie2 by bitwise flag to select only first-in-pair reads (representing bona-fide initiation sites), then assigned the position (chromosome and nucleotide) of the first nucleotide in each read to a hash table, then combined the results from each replicate into a counts table. To normalize read depth, we first calculated the number of reads mapping to a set of spike-in control transcripts using the bedtools multicov utility (Quinlan and Hall 2010). For the number of raw reads (R) mapping to each spike-in transcript t in each replicate i (R_{ti}), we normalized R_{ti} to the geometric mean of all $\{R_{ti} .. R_{tn}\}$ to generate the normalized transcript value N_{ti}. Finally, for each replicate we calculated a replicate normalization score S_i by calculating the geometric mean of all $\{N_{ti} .. N_{ni}\}$. For initial analysis in Figures 1 and 2, Start-seq reads were assigned to *D. melanogaster* TSSs defined from a previous study (Nechaev et al. 2010).

To generate Start-seq peak clusters likely to belong to the same TSS, we first calculated a false-discovery rate (FDR) cutoff for bona-fide Start-seq peak detection at 9 normalized reads per nucleotide per biological replicate, based on sequencing depth. Then we clustered nucleotides meeting this threshold as follows: For each nucleotide n_i, an edge was established with a neighboring nucleotide n_j if it occurred within 5 nt of n_i. Then clusters were formed by including all nucleotides n that occurred between terminal nucleotides upstream (n_u) and downstream (n_d) that were bound to the

cluster by only a single edge, and thus terminated the "chain." For each cluster, the cluster "summit" was identified as the nucleotide containing the most mapped reads, and secondary and tertiary peaks were identified as containing the second- and third-most reads, respectively of any nucleotide in the cluster, if applicable. From the summit, we calculated the proportion of reads in the cluster contained within 2 nt on either side of the summit.

To assign Start-seq peak clusters to observed TSSs (obsTSSs) or novel TSSs (nuTSSs), we searched for overlap between our peak clusters and a list of coordinates for the first exons from every transcript in the dm3 gene annontation, and assigned those that overlapped as obsTSSs. All other clusters that did not map to a defined first exon were assigned as nuTSSs. Promoter regions defined by DEEP-CAGE in *D. melanogaster* embryos were obtained from Hoskins *et al.* (2011), and "integrated promoters" that overlapped with obsTSS peaks were detected using the bedtools intersect function (Quinlan and Hall 2010).

Start-seq peak shape and pausing index. To calculate peak shape index (SI), we adapted a formula from Hoskins *et al.* (Hoskins et al. 2011):

$$SI = 2 + \sum_{i}^{N} p_i \log_2 p_i$$

where N = the set of single nucleotides i in the peak cluster, and p_i = the proportion of total reads in the cluster mapping to nucleotide i. To calculate peak pausing index (PI), we divided the normalized start-seq signal mapping to the obsTSS peak cluster by the nuclear RNA-seq RPKM calculated for its corresponding gene.

Analysis of promoter motif enrichment. To discover motifs proximal to obsTSSs, we first sourced transcription factor motifs from FitzGerald *et al.* (FitzGerald et al. 2006). We then determined the expected distribution of those motifs relative to a TSS, and for every obsTSS we used the bedtools getfasta function to generate FASTA sequences that were 50 nt in length and roughly restricted to the expected localization of each motif. For instance, the TATA box motif is expected to occur ~32 nt upstream of the TSS, so the FASTA file used to test for the presence of TATA captured all nucleotides from -50 to 0 relative to the TSS. Then we used the 15 sequences present in FitzGerald *et al.* (FitzGerald et al. 2006), plus the "Pause Button" motif sequence (Hendrix et al. 2008), and their corresponding restricted FASTA files for all obsTSSs, to execute the "homer2 find" function from the Homer motif analysis software (Heinz et al. 2010), allowing for

up to 4 mismatches. For each obsTSS X and each motif y, a single motif score X_y was determined by selecting the highest log odds score among all $\{X_y1..X_yn\}$ detected in the assigned FASTA sequence. Those obsTSSs for which a sequence with no more than 4 mismatches to a motif was not detected were assigned a score X_y equal to the lowest log odds score for the all obsTSSs tested for the motif in question. X_y values were converted to z-scores using the following formula:

$$Z_{Xy} = \frac{|X_y - \mu_y|}{\partial_y}$$

where Z_{Xy} = z-score for obsTSS X and motif y, μ_y = mean log odds score for motif y, and ∂_y = standard deviation of log odds scores for motif y. The resultant z-scores were visualized using the heatmap.2 utility in the gplots R package (https://CRAN.R-project.org/package=gplots), with column clustering. Correlation coefficients for motifs were generated using the "cor" function in the R base package, and were visualized using the heatmap.2 utility.

Divergent transcription analysis. To find instances of divergent transcription, pairs of neighboring TSSs oriented in opposite directions were ordered by distance of the reverse-strand TSS from the forward-strand TSS (negative numbers denote upstream, positive denote downstream), ATAC-seq signal was plotted in a 20kb window surrounding the forward-strand TSS using the Deeptools "computeMatrix reference-point" utility with default parameters (Ramírez et al. 2014), and signal was visualized using the heatmap.2 utility. From this analysis, divergent transcription was visually defined as the maximal distance between paired TSSs for which a conitiguous ATAC-seq enriched region could be detected, or roughly 200 nt. For divergent pairs, Start-seq signal was visualized as described above, using computeMatrix reference-point with the following flags: -binSize 1, -afterRegionStartLength 1, -missingDataAsZero. Enrichment of ATAC-seq signal around bidirectional, divergent, and non-divergent TSSs was determined by using the bedtools multicov tool (Quinlan and Hall 2010) to map ATAC-seq signal within a 400 nt window surrounding each TSS.

nuTSS clustering and elongation analysis. To cluster nuTSSs by histone post-translational modification (PTM) density, PTM enrichment for H3K4me1, H3K4me3, H3K27ac, and H3K36me3 ChIP-seq signal at each nuTSS was calculated by mapping ChIP and input reads to a 200 nt window on either side of the nuTSS using the bedtools multicov tool (Quinlan and Hall 2010), then dividing ChIP reads by input reads.

Optimal cluster number was discovered by calculating within-group sum of square distances for each cluster solution between 2 and 15 clusters, and 5 clusters were found to simultaneously minimize distance and cluster number. We assigned clusters using the hclust method in R.

To analyze transcription elongation from nuTSSs, we quantified the number of strand-specific nuclear RNA-seq reads mapping within either 200 nt upstream or 200 nt downstream of the nuTSS. Then, the enrichment of downstream over upstream reads was calculated using DESeq2 (Love et al. 2014). Though the expected paucity of reads mapping upstream of TSSs likely increases the threshold for significance across all TSSs, we reasoned that taking an approach that assigns significance partially based on the number of reads assigned to the feature in question would help in identify lowly elongating coding nuTSS transcripts oriented in the sense direction relative to their resident genes, since more reads would accumulate in both upstream and downstream regions in those cases.

Enhancer region analysis. Functionally validated enhancers were obtained from the FlyLight database ((Jenett et al. 2012), http://flweb.janelia.org/cgi-bin/flew.cgi) by querying based on anatomical expression in the larval CNS or in imaginal discs, and selecting all enhancers with validated expression in any one of those tissues. Each enhancer was annotated with the tissues in which it was reported to express according to the FlyLight database. Genomic coordinates of enhancer regions were obtained from the Bloomington Stock Center website (http://flystocks.bio.indiana.edu/bloomhome.htm). Overlap between nuTSSs and enhancer regions was evaluated using the bedtools intersect tool (Quinlan and Hall 2010).

Data Access

All data used in this study is publicly available in the Gene Expression Omnibus (GEO) database at NCBI. Please see methods section for relevant accession numbers.

Acknowledgements

MPM was supported by an NIH predoctoral fellowship, F31-CA177088. This work was supported by the NIH Epigenomics Roadmap Project, R01-DA036897 (to AGM, BDS,

DJM and RJD), and by the Intramural Research Program of the NIH (Z01-ES101987), National Institute of Environmental Health Sciences (to KA).

Author Contributions

MPM and AGM conceptualized the study and experiments. MPM designed methodology and performed formal analysis, validation, data curation, software design, and writing of the first draft. AGM, BDS, DJM, RJD and KA contributed to funding acquisition and review and editing.

References

Adelman K, Kennedy MA, Nechaev S, Gilchrist DA, Muse GW, Chinenov Y, Rogatsky I. 2009. Immediate mediators of the inflammatory response are poised for gene activation through RNA polymerase II stalling. *Proc Natl Acad Sci U S A* **106**: 18207–18212.

Andersen PR, Domanski M, Kristiansen MS, Storvall H, Ntini E, Verheggen C, Schein A, Bunkenborg J, Poser I, Hallais M, et al. 2013. The human cap-binding complex is functionally connected to the nuclear RNA exosome. *Nat Struct Mol Biol* **20**: 1367–1376.

Bannister AJ, Schneider R, Myers FA, Thorne AW, Crane-Robinson C, Kouzarides T. 2005. Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. *J Biol Chem* **280**: 17732–17736.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**: 1213–1218.

Buratowski S, Hahn S, Guarente L, Sharp PA. 1989. Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell* **56**: 549–561.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engström PG, Frith MC, et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**: 626–635.

Carrozza MJ, Li B, Florens L, Suganuma T, Swanson SK, Lee KK, Shia W-J, Anderson S, Yates J, Washburn MP, et al. 2005. Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell* **123**: 581–592.

Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. 2014. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet* **46**: 1311–1320.

Core LJ, Waterfall JJ, Gilchrist DA, Fargo DC, Kwak H, Adelman K, Lis JT. 2012. Defining the status of RNA polymerase at promoters. *Cell Rep* **2**: 1025–1035.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**: 21931–21936.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* **489**: 101–108.

Dorighi KM, Swigut T, Henriques T, Bhanu NV, Scruggs BS, Nady N, Still CD, Garcia BA, Adelman K, Wysocka J. 2017. Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Mol Cell* **66**: 568–576.e4.

Duttke SHC, Lacadie SA, Ibrahim MM, Glass CK, Corcoran DL, Benner C, Heinz S, Kadonaga JT, Ohler U. 2015. Human promoters are intrinsically directional. *Mol Cell* **57**: 674–684.

FitzGerald PC, Sturgill D, Shyakhtenko A, Oliver B, Vinson C. 2006. Comparative genomics of Drosophila and human core promoters. *Genome Biol* **7**: R53.

Gilchrist DA, Dos Santos G, Fargo DC, Xie B, Gao Y, Li L, Adelman K. 2010. Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell* **143**: 540–551.

Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, Taipale J. 2006. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**: 47–59.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**: 576–589.

Hendrix DA, Hong J-W, Zeitlinger J, Rokhsar DS, Levine MS. 2008. Promoter elements associated with RNA Pol II stalling in the Drosophila embryo. *Proc Natl Acad Sci U S A* **105**: 7762–7767.

Henriques T, Gilchrist DA, Nechaev S, Bern M, Muse GW, Burkholder A, Fargo DC, Adelman K. 2013. Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals. *Mol Cell* **52**: 517–528.

Hoskins RA, Landolin JM, Brown JB, Sandler JE, Takahashi H, Lassmann T, Yu C, Booth BW, Zhang D, Wan KH, et al. 2011. Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Res* **21**: 182–192.

Jenett A, Rubin GM, Ngo T-TB, Shepherd D, Murphy C, Dionne H, Pfeiffer BD, Cavallaro A, Hall D, Jeter J, et al. 2012. A GAL4-driver line resource for Drosophila neurobiology. *Cell Rep* **2**: 991–1001.

Jory A, Estella C, Giorgianni MW, Slattery M, Laverty TR, Rubin GM, Mann RS. 2012. A survey of 6,300 genomic fragments for cis-regulatory activity in the imaginal discs of Drosophila melanogaster. *Cell Rep* **2**: 1014–1024.

Kaplan CD, Laprade L, Winston F. 2003. Transcription elongation factors repress transcription initiation from cryptic sites. *Science* **301**: 1096–1099.

Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, et al. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**: 362–366.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12**: 996–1006.

Keogh M-C, Kurdistani SK, Morris SA, Ahn SH, Podolny V, Collins SR, Schuldiner M, Chin K, Punna T, Thompson NJ, et al. 2005. Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. *Cell* **123**: 593–605.

Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**: 182–187.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Lee TI, Young RA. 2000. Transcription of eukaryotic protein-coding genes. *Annu Rev Genet* **34**: 77–137.

Li W, Notani D, Rosenfeld MG. 2016. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat Rev Genet* **17**: 207–223.

Liang G, Lin JCY, Wei V, Yoo C, Cheng JC, Nguyen CT, Weisenberger DJ, Egger G, Takai D, Gonzales FA, et al. 2004. Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. *Proc Natl Acad Sci U S A* **101**: 7357–7362.

Lipshitz HD, Peattie DA, Hogness DS. 1987. Novel transcripts from the Ultrabithorax domain of the bithorax complex. *Genes Dev* **1**: 307–322.

Lopez P, Wagner K-D, Hofman P, Van Obberghen E. 2016. RNA Activation of the Vascular Endothelial Growth Factor Gene (VEGF) Promoter by Double-Stranded RNA and Hypoxia: Role of Noncoding VEGF Promoter Transcripts. *Mol Cell Biol* **36**: 1480–1493.

Lorch Y, Zhang M, Kornberg RD. 1999. Histone octamer transfer by a chromatin-remodeling complex. *Cell* **96**: 389–392.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Lubas M, Andersen PR, Schein A, Dziembowski A, Kudla G, Jensen TH. 2015. The human nuclear exosome targeting complex is loaded onto newly synthesized RNA to direct early ribonucleolysis. *Cell Rep* **10**: 178–192.

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

Mayer A, di Iulio J, Maleri S, Eser U, Vierstra J, Reynolds A, Sandstrom R, Stamatoyannopoulos JA, Churchman LS. 2015. Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell* **161**: 541–554.

McKay DJ, Lieb JD. 2013. A common set of DNA regulatory elements shapes Drosophila appendages. *Dev Cell* **27**: 306–318.

Meers MP, Henriques T, Lavender CA, McKay DJ, Strahl BD, Duronio RJ, Adelman K, Matera AG. 2017. Histone gene replacement reveals a post-transcriptional role for H3K36 in maintaining metazoan transcriptome fidelity. *eLife* **6**.

Mitchell PJ, Tjian R. 1989. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245**: 371–378.

Muse GW, Gilchrist DA, Nechaev S, Shah R, Parker JS, Grissom SF, Zeitlinger J, Adelman K. 2007. RNA polymerase is poised for activation across the genome. *Nat Genet* **39**: 1507–1511.

Nechaev S, Fargo DC, dos Santos G, Liu L, Gao Y, Adelman K. 2010. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science* **327**: 335–338.

Ni T, Corcoran DL, Rach EA, Song S, Spana EP, Gao Y, Ohler U, Zhu J. 2010. A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nat Methods* **7**: 521–527.

Ozsolak F, Song JS, Liu XS, Fisher DE. 2007. High-throughput mapping of the chromatin structure of human promoters. *Nat Biotechnol* **25**: 244–248.

Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, Mapendano CK, Schierup MH, Jensen TH. 2008. RNA exosome depletion reveals transcription upstream of active human promoters. *Science* **322**: 1851–1854.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma Oxf Engl* **26**: 841–842.

Rach EA, Winter DR, Benjamin AM, Corcoran DL, Ni T, Zhu J, Ohler U. 2011. Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLoS Genet* **7**: e1001274.

Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**: 279–283.

Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. 2014. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42**: W187-191.

Rye M, Sandve GK, Daub CO, Kawaji H, Carninci P, Forrest ARR, Drabløs F, FANTOM consortium. 2014. Chromatin states reveal functional associations for globally defined transcription start sites in four human cell lines. *BMC Genomics* **15**: 120.

Saunders A, Core LJ, Sutcliffe C, Lis JT, Ashe HL. 2013. Extensive polymerase pausing during Drosophila axis patterning enables high-level and pliable transcription. *Genes Dev* **27**: 1146–1158.

Scruggs BS, Gilchrist DA, Nechaev S, Muse GW, Burkholder A, Fargo DC, Adelman K. 2015. Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Mol Cell* **58**: 1101–1112.

Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, Wang J-PZ, Widom J. 2006. A genomic code for nucleosome positioning. *Nature* **442**: 772–778.

Takahashi K, Yamanaka S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**: 663–676.

Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM. 2004. An abundance of bidirectional promoters in the human genome. *Genome Res* **14**: 62–66.

Vera JM, Dowell RD. 2016. Survey of cryptic unstable transcripts in yeast. *BMC Genomics* **17**: 305.

Verdel A, Jia S, Gerber S, Sugiyama T, Gygi S, Grewal SIS, Moazed D. 2004. RNAi-mediated targeting of heterochromatin by the RITS complex. *Science* **303**: 672–676.

Wakano C, Byun JS, Di L-J, Gardner K. 2012. The dual lives of bidirectional promoters. *Biochim Biophys Acta* **1819**: 688–693.

Yang J, Ramos E, Corces VG. 2012. The BEAF-32 insulator coordinates genome organization and function during the evolution of Drosophila species. *Genome Res* **22**: 2199–2207.

Zentner GE, Tesar PJ, Scacheri PC. 2011. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res* **21**: 1273–1283.