HHS Public Access
Author manuscript
Nat Methods. Author manuscript; available in PMC 2011 October 19.

Published in final edited form as:
Nat Methods. 2010 July ; 7(7): 521–527. doi:10.1038/nmeth.1464.

A paired-end sequencing strategy to map the complex landscape of transcription initiation

Ting Ni[1,2,8,9], **David L Corcoran**[1,8], **Elizabeth A Rach**[1,3], **Shen Song**[1,2], **Eric P Spana**[4], **Yuan Gao**[5], **Uwe Ohler**[1,6,7,*], and **Jun Zhu**[1,2,9,*]

[1]Institute for Genome Sciences & Policy, Duke University Medical Center, 101 Science Drive, Durham, NC 27708, USA

[2]Department of Cell Biology, Duke University Medical Center, 307 Research Drive, Durham, NC 27710, USA

[3]Program in Computational Biology and Bioinformatics, Duke University, 101 Science Drive, Durham, NC 27708, USA

[4]Department of Biology, Duke University, 125 Science Drive, Durham, NC 27708, USA

[5]Center for the Study of Biological Complexity, Virginia Commonwealth University, 401 West Main Street, Richmond, VA 23284, USA

[6]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, 2301 Erwin Road, Durham, NC 27710, USA

[7]Department of Computer Science, Duke University, LSRC Building D101, 450 Research Drive, Durham, NC 27708, USA

Abstract

Users may view, print, copy, download and text and data- mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use: http://www.nature.com/authors/editorial_policies/license.html#terms

[*]Correspondence should be addressed to: uwe.ohler@duke.edu, jun.zhu@nih.gov.

[8]These authors contributed equally to this work.

[9]Current address: Genetics and Development Biology Center, National Heart Lung and Blood Institute, National Institute of Health, 9000 Rockville Pike, Bethesda, MD 20892, USA

AUTHOR CONTRIBUTIONS

U.O and J.Z. oversaw the project. T.N., S.S. and J.Z designed and performed experiments related to PEAT library construction, quality control and various validation assays. E.P.S provided fly stock for collecting embryos. Y.G. performed Illumina sequencing. D.L.C., E.A.R., U.O. performed data analysis. T.N. D.L.C. E.A.R., U.O. and J.Z. wrote the manuscript.

ACCESS NUMBER Raw data has been deposited at the NCBI Short Read Archive (SRX018832).

Note: Supplementary information is available on the Nature Methods website.

COMPETING INTERESTs STATEMENT

The authors declare no competing financial interests.

AOP:

Paired–end reads consisting of 5′ transcription start sites and 3′ downstream sequences from transcripts in *Drosophila melanogaster* reveal distinct initiation patterns at different fly promoters and show that 5′ caps originating in coding regions are added posttranscriptionally.

Issue

Paired-end reads consisting of 5′ transcription start sites and 3′ downstream sequences from transcripts in *Drosophila melanogaster* reveal distinct initiation patterns at different fly promoters and show that 5′ caps originating in coding regions are added posttranscriptionally.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Recent high-throughput sequencing protocols have uncovered the complexity of mammalian transcription by RNA polymerase II, helping to define several initiation patterns in which transcription start sites (TSSs) cluster within both narrow and broad genomic windows. Here, we describe a paired-end sequencing strategy, which enables more robust mapping and characterization of capped transcripts. This strategy was applied to explore the transcription initiation landscape in the *Drosophila melanogaster* embryo. Extending the previous findings in mammals, we found that fly promoters exhibit distinct initiation patterns, which are linked to specific promoter sequence motifs. Furthermore, we identified a large number of 5′ capped transcripts originating from coding exons; analyses support that they are unlikely the result of alternative TSSs, but rather the product of post-transcriptional modifications. Taken together, paired-end TSS analysis is demonstrated to be a powerful method to uncover the transcriptional complexity of eukaryotic genomes.

INTRODUCTION

Transcription by RNA Polymerase II (Pol II) is a critical step in eukaryotic gene expression. To initiate and modulate transcription, factors interact with chromatin and DNA sequence features in regulatory regions. Central to this process is the core promoter region of approximately 100 nucleotides surrounding the transcription start site (TSS) of a gene. Within this region, factors of the basal transcription machinery interact directly with DNA sequence motifs to ensure the proper recruitment of Pol II. Contrary to the simple picture in many textbooks, which often present the basal machinery as invariable and core promoters that generally share the same motifs, many recent studies have demonstrated the diversity in both basal transcription factor complexes and the sequence features to which they bind1, 2. We are still only now beginning to truly understand the diversity at the transcription initiation level, and how it provides for additional regulatory control of gene expression3-6.

Methods to systematically sequence 5′ complete transcripts have provided the breakthrough for genome-wide identification of TSSs7-9. In particular, the capped analysis of gene expression (CAGE) protocol has been used to generate comprehensive mammalian libraries of short sequence tags, which have led to the identification of distinct transcription initiation patterns10-12. In some promoters, transcription initiates from the same exact location, while other transcripts initiate more uniformly across wider genomic windows. Different sequence features have been found to be associated with these patterns, such as an overrepresentation of the canonical TATA box sequence motif in ‘single-peak’ promoters, and CpG islands overlapping ‘broad range’ promoters11. The majority of studies using CAGE protocols have been focused on mouse and human. Thus far there has not been an attempt to investigate, on a similar scale, whether different initiation patterns can also be found in other animals, such as the fruit fly *Drosophila melanogaster,* and whether these are associated with distinct sequence features.

The CAGE protocol has recently been extended to deepCAGE13 which involves a concatemerization step of reads, and the final library has to be sequenced by a platform (such as 454 pyrosequencing) that can generate sufficiently long reads. Their number is on a smaller scale than what may be achieved on other sequencing platforms. In addition,

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

deepCAGE produces a single, typically 20-nt-long, sequence tag from the most 5′ end of the transcript, which may be too short to guarantee a unique and correct alignment to the genome, especially in the presence of sequencing errors. Such challenges can be addressed by longer reads or paired-end reads. The latter strategy is expected to be more advantageous because it can provide additional information on the local transcript structure.

In order to thoroughly characterize the landscape of transcription initiation in eukaryotes we developed the Paired-End Analysis of TSSs (PEAT) strategy, by which each TSS tag (20 nt sequence from the most 5′ end of the transcript) is paired with a 20 nt downstream tag from the same gene. We applied PEAT to analyze capped transcripts of *D. melanogaster* mixed-stage embryos. Our results uncovered that *Drosophila*, like mammals, has multiple initiation patterns, each of which is associated with a distinct set of sequence motifs. Furthermore, we found that ~25% of 5′ capped reads align to the coding region of the *Drosophila* genome. Extending the previous findings in mammals11, 14, we provide strong evidence that these transcripts result from posttranscriptional modification rather than de novo transcription from the coding region. Together, these results demonstrate that PEAT is an improved strategy to map and characterize the landscape of transcription initiation in higher eukaryotes.

RESULTS

A paired-end strategy for deep sequencing of capped RNA

In the PEAT protocol capped transcripts are selectively ligated to a 5′ linker sequence, which contain a MmeI site, using an oligo-capping strategy. Reverse transcription is then carried out with a random hexamer tailed with a second MmeI site. After low-cycle PCR, cDNA products are circularized by bridge ligation followed by exonuclease digestion. The resulting DNA circles are subsequently amplified by rolling circle amplification15, and digested with Mme I to release paired tags, each of which contain a TSS tag and a downstream 3′ tag. After ligation with sequencing adaptors, the final PEAT library is sequenced by a Illumina Genome Analyzer with paired-end capability (Fig. 1a). Compared to conventional single-end TSS mapping strategies7, 16, the PEAT approach improves the alignment yield and accuracy and provides additional information on local transcript structure (such as linking 5′ TSS tags to known genes).

We employed the PEAT strategy to monitor global TSS usage in mixed stage embryos (0-24h) of *D. melanogaster*. We obtained 17.5 million raw paired-reads from two technical replicates. For approximately 90% of the paired-reads, both the TSS and 3′ reads were distinguishable by their built-in linker sequences (Table 1). Of those paired-reads, 76% were mapped to a unique location in the fly genome. An additional 10% of pairs mapped to multiple genomic locations (Fig. 1b), possibly due to transposable elements or other regions with low sequence complexity (data not shown). The majority of 5′ reads were mapped to either a known TSS or its surrounding regions, confirming that our approach captured the very 5′ end of capped transcripts (Fig. 1c, Supplementary Fig. 1). The median distance between the 5′ and 3′ reads at the transcript level was 279nt (Supplementary Fig. 2) and the 3′ reads are mostly mapped to coding regions of annotated genes, indicating the success of the paired-end library construction.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

On average there were 256 tag pairs per gene (Supplementary Fig. 3), demonstrating deep coverage of the genome. 81.5% of genes currently annotated by FlyBase (v5.14) were represented by at least one read-pair, consistent with the notion that eukaryotic genomes are broadly transcribed. Taken together, the mapping yield was considerably higher than that of deepCAGE13. The fraction of aligned tags and coverage of the genome were also dramatically improved from a previous CAGE study of *D. melanogaster*10 (Supplementary Table 1, Supplementary Fig. 4).

The two technical replicates were highly correlated (*R* = 0.98, Supplementary Fig. 5), indicating the reproducibility of PEAT. We next compared our results with a microarray-based expression dataset obtained from fly embryos of a similar broad developmental window (embryonic stages 0-11). With minimal normalization on both the array and sequence data, PEAT and array expression profiles were significantly correlated (*R* = 0.68, Fig. 1d and Supplementary Fig. 6). The result is comparable to the correlation observed between microarray and standard RNA-Seq approaches17. Therefore, the read count of the PEAT method can potentially be used to estimate transcript abundance.

The paired-end strategy clearly allowed for an accurate mapping of the short reads. The addition of the 3′ reads enables ~4% of the 5′ reads to be aligned to a unique genomic location instead of multiple locations. Furthermore, the downstream tags can also correct assignment mistakes caused by sequencing errors. In fact, ~0.3% of the 5′ reads would have been wrongly aligned if the downstream tag had not been provided (Supplementary Table 2). It is expected that such improvements will become more prominent for larger genomes or when the sequencing error is relatively high Paired-reads also facilitated the direct link of novel TSSs to their respective genes. For 342,943 read pairs where the 5′ read fell more than 250 nt upstream of an annotated TSS, the corresponding 3′ read of 17% mapped to the transcribed region of the downstream gene. We successfully validated several individual cases of such distal unannotated TSSs (**Supplementary Results**; Supplementary Figs. 7-8 **and** Supplementary Table 3).

Characterization of read clusters and initiation patterns

High-throughput TSS maps have shown that mammalian promoters exhibit diverse initiation patterns, but it was an open question whether *Drosophila* promoters would show a similar complexity. To this end, we clustered the mapped 5′ reads (Fig. 2a)**,** resulting in 34,664 discrete clusters covering 8,577 genes. More than 5,500 genes had at least one cluster with ≥ 50 reads, and approximately half of these clusters overlapped annotated TSSs (Fig. 2b).

To determine transcription initiation patterns, we focused on 5′ clusters with ≥ 100 reads (5,699 clusters in 4,007 genes). The cutoff was stringent to ensure high-quality assignments of initiation patterns and sequence motifs. The clusters spanned a broad size range, describing a complex multimodal distribution (Fig. 2c) suggesting distinct initiation patterns. In fact, the cluster size distribution could be approximated by two Gaussian distributions, the intersection of which fell at ~25 nt. Read clusters were thus separated into three initiation patterns, Narrow with Peak (NP), Broad with Peak (BP) and Weak Peak (WP), along the two dimensions of cluster size and read distribution within each cluster (Fig. 2d).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Initiation patterns are linked to specific core promoter

In mammals, ‘peak’ and ‘broad’ promoters tend to be associated with TATA box and CpG islands, respectively11. Since the fly genome does not contain CpG islands, it was intriguing to find that broad promoters exist in *Drosophila*. We therefore aimed to determine whether distinct initiation patterns were associated with core promoter motifs previously defined in *Drosophila*18. We extracted 200 nt sequences centered on the mode of each cluster (that is, the most frequent TSS within the cluster). Promoter sequences were aligned for each initiation pattern and the results showed that initiation preferentially occurs at an adenine, immediately preceded by the ‘TC’ di-nucleotide for all 3 initiation patterns (Fig. 3 and Table 2). The (T)CA consensus matched the minimal sequence requirements at the TSS as reported in other eukaryotes from yeast to mammals9, 11, but was only a substring of the fly initiator motif as originally reported19. Thus, even for the broad pattern, defining the reference TSS at the mode was linked to a significant presence of a minimal initiator consensus.

We next evaluated the presence and preferred locations within different initiation patterns of eight sequence motifs reported to be present in the core promoter regions in fly2, including the TATA box and INR motif, the Motif Ten Element (MTE), the Downstream Promoter Element (DPE)20, the DNA-replication related element (DRE)4 and Motif 1,6,7 18, 21 (see **Online Methods**). Strikingly, the results revealed distinct associations between initiation patterns and sequence motifs. The canonical core promoter motifs with previously known location bias (TATA, INR, DPE, MTE) were highly associated with NP promoters. The DPE was enriched at its known location (+26, +30) and at an additional site (−5, −1), which has previously been observed in mammalian data22; the second location likely reflects some overlap in sequence similarity rather than functional DPE occurrences, as the importance of precise spacing has been clearly established20. Mammalian WP promoters have frequently been associated with CpG islands, a feature not present in the fly genome2. Instead, *Drosophila* WP promoters were strongly associated with 3 motifs (Motif1, DRE, Motif7) and showed a moderate enrichment for Motif6 (Fig. 3b-c and Supplementary Fig. 9). BP promoters, which have characteristics of both NP and WP promoters, showed a combination of the most frequent motifs found in the other types. The largest span of motif enrichment was 25 nt for the DRE motif in WP promoters, reflecting the broad initiation pattern in this class. The associations of TATA box and DRE with different classes were also supported by differential binding of factors to the genome as assayed by chromatin immunoprecipitation (**Supplementary Results**). We noted that the INR and Motif1 motifs share a strong conserved ‘TCA’ tri-nucleotide, that is, the minimal initiator consensus described above. Likewise, Motif6 was enriched at the same location as the TATA box and contains a minimal TAT consensus that is shared with the canonical TATA motif, suggesting that Motif6/Motif1 is an alternative to the classic TATA/INR motif pair, an observation that has only become apparent with high-resolution data generated by PEAT. Overall, these results demonstrated that the initiation patterns in fly directly reflected the presence of the specific core promoter motifs.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

5′ capped read clusters in coding regions

In the initial clustering of reads, we observed that 25% of clusters were found within the coding region of an annotated gene, and cluster analysis showed that the majority of them belong to the WP class (Supplementary Table 4). Twelve candidates were selected for validation, and ten were confirmed by two independent methods, oligo-capping and cap-trapping. (Fig. 4, Supplementary Figs. 10-12 and Supplementary Table 5). Therefore, these clusters were not artifacts of the high-throughput protocol and indeed contained a 5′ cap. Supporting this notion, recent studies in mammals have also identified a high prevalence of capped transcripts originating from the coding regions11, 14.

Several mechanisms may underlie the biogenesis of internally capped transcripts. First, they might result from *bona fide* start sites in the coding region. Alternatively, these transcripts may be derived from longer precursors, for which the internal cap is introduced posttranscriptionally by a recapping mechanism14. Multiple lines of evidence from our data support the latter model. First, searching the 200 nt sequences surrounding the coding clusters revealed no overrepresentation of any of the core promoter motifs observed near *bona fide* TSSs (Supplementary Table 4); this was in agreement with our previous observation of a lack of promoter motifs around mammalian coding clusters23. The analysis of ChIP data6 showed frequent binding of TFs (TBP and, or TRF2) at TSS clusters but not at coding clusters (**Supplementary Results**). Together, our data suggested that 5′ capped coding clusters are unlikely initiated by Pol II.

In addition, we found that for 69% of the coding clusters, a larger cluster (with more reads) was identified near the annotated TSS (Supplementary Fig. 13), indicating that internally capped transcripts were often accompanied by more abundant full-length transcripts. Moreover, the locations of the 5′ coding region clusters spread evenly across the exons except for a lack of clusters at the far most 3′ end of the exon (Supplementary Fig. 14), similar to what has been reported in mammals14. The mammalian study relied on TSS reads mapped across exon junctions, which are a tiny fraction of the total reads, to argue that recapping is a posttranscriptional event. Unique to the PEAT dataset, we observed that the downstream paired tags of the coding clusters were predominantly located in well-annotated exons rather than introns (~ 100-fold enrichment), indicating that internally capped transcripts were spliced or at least partially spliced.

We also observed a distinct short sequence motif when aligning the sequences surrounding the mode of coding clusters. While this motif was at first glance reminiscent of the minimal initiator motif found in TSS clusters, it exhibited unique properties. ‘CA’ was the most frequent di-nucleotide at the −1 position in TSS read clusters, while the most prominent di-nucleotide at the mode location within coding region clusters was ‘TC’ (Table 2). Although the molecular mechanism of recapping remains elusive, the distinct motif implied that recapping might depend on specific sequences or protein factors.

DISCUSSION

PEAT distinguishes itself from other paired-end transcriptome sequencing strategies such as GIS-PET24. While GIS-PET also generates a TSS tag paired with a 3′ tag, the downstream

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

tags are designed to query polyadenylation sites. Unlike PEAT, the fixed 3′ tag location could not provide local transcript structures proximal to TSS and is unable to resolve the recapping events. Extending previous observations based on ESTs25, the high-resolution initiation map generated by PEAT allowed the identification of 3 distinct initiation patterns in *Drosophila*. Initiation patterns were linked to the presence of specific core promoter sequence motifs, including an enrichment of the DRE motif in WP promoters. The presence of WP promoters in *Drosophila* is intriguing as broad promoters in mammals are enriched for CpG islands11, a genomic feature not present in the fly. Notably, CpG islands and DRE are associated with housekeeping genes in human11 and fly25 respectively, indicating functional conservation of WP promoters in diverse organisms. Moreover, ChIP data supported the notion that distinct complexes are associated with WP and NP promoters in fly (**Supplementary Results**, Supplementary Figs 15-16 **and** Supplementary Table 6). As their initiation patterns suggest, the class of BP promoters contained a combination of both the motifs seen in the other two classes. However, it is unclear whether this is a consequence of different complexes recognizing the same regulatory region, or if this occurs at different transcripts under the same condition. Our mixed-stage embryonic sample contains both maternal and zygotic transcripts, and vertebrate transcription in oocytes has recently been shown to depend on stage-specific basal transcription initiation complexes26, 27.

Additionally, we provided multiple lines of evidence that internally capped transcripts are likely derived from post-transcriptional processing events in fly, as suggested by a previous mammalian study14. Recapping sites were uniformly distributed across the internal exon except at its extreme 3′ end. This coincides with the exon junction complex (EJC), which is deposited 20-24 nt upstream of splicing junctions28. Since both the early report and our study suggest that internally capped transcripts are likely to be derived from processed (or spliced) transcripts, we speculate that the depletion of the recapping site at the end of the exon may reflect the competition between the EJC and recapping machinery. Further investigations are required to elucidate the biogenesis and functional significance of this novel class of transcripts29.

Lastly, this study focused on initiation sites of long polyadenylated transcripts. This explains why we did not observe promoter-associated non-coding transcripts, which have been reported in other species30, or the short transcripts associated with polymerase stalling31. An earlier study using total RNA detected a large number of transcribed fragments (transfrags) that are well upstream of known TSSs and correlate in expression with the downstream genes32. We showed that such distant TSSs are relatively rare for polyadenylated and capped transcripts, and are unlikely the initiation sites for known downstream transcripts. Although one cannot rule out that the observed differences are due to stage variation (mixed stage library vs. several 2-hr windows), it is suggestive that these transfrags are not polyadenylated or capped, or both; and that they may represent instances of a class of regulatory RNAs (for example, promoter associated long RNAs) in the fly transcriptome. Further efforts are required to profile and characterize different classes of RNA to dissect the complexity and plasticity of eukaryotic transcriptomes.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We are grateful to D. MacAlpine and S. Powell (Duke University Medical Center) for their help in collecting fly embryos. We are indebted to B. Xie (Virginia Commonwealth University) and Y. Bao (University of Virginia) for optimizing paired-end sequencing procedure, and to J. Kadonaga (UC San Diego) for helpful comments on the manuscript. This work was funded by National Institute of Health R01 HG004065 (U.O. and J.Z.) and National Science Foundation MCB0822033 (J.Z. and U.O.).

APPENDIX

ONLINE METHODS

Paired-end library preparation

Mixed stage fly embryos (0-24h) were collected according to a standard protocol32. We used TRIzol reagent (Invitrogen) to extract total RNA. RNeasy Mini kit (QIAGEN) was used for cleanup and on-column DNase I digestion to remove genomic DNA according to the manufacturer's protocol. 150 µg purified RNA was enriched for poly(A)+ RNA with Dynabeads Oligo(dT)25 (Invitrogen) according to a modified protocol. 2 µg of poly(A)+ RNA was BAP/TAP (Bacterial Alkaline Phosphatase/Tobacco Acid Pyrophosphatase) treated and a chimeric linker tagged with a MmeI site was ligated to its 5′ end. The RNeasy MinElute kit (QIAGEN) was used to remove excessive chimeric linkers. Random primers tagged with a MmeI recognition site were used to initiate reverse transcription. First strand cDNAs were amplified using 5 cycles of PCR, and the products were purified with DNA clean & concentrator-5 kit (ZYMO). Circularization was performed with a "collector" oligonucleotide, which converts the PCR product into a single-stranded circular DNA. After Exo I (NEB) and Exo III (NEB) digestion to remove linear DNAs, rolling circle amplification (RCA) was performed to amplify the remaining circular DNAs. The RCA products were digested with MmeI (NEB) to generate a specific 93~95 bp band. The desired product was ligated with two Illumina Paired-End adaptors and amplified with low-cycle PCR. After size-selection and validation by Sanger sequencing, the final library was sequenced using an Illumina GAII with a paired-end module.

Oligo(dT) selection

Dynabeads Oligo(dT)25 from Invitrogen was used to enrich poly(A)+ RNAs. Briefly, 150 µg total RNA was resuspended in 400 µl binding buffer (20 mM Tris-HCl, pH 7.5, 1.0 M LiCl, 2 mM EDTA, 1% LiDS, 0.1% Trion X-100) and heated at 65°C for 2 min to disrupt RNA secondary structures. After snap cool down, 200 µl Dynabeads was added followed by incubation at 50°C for 5 min. We found that incubation at a higher temperature helps remove the non-specific binding of ribosomal RNA. The resulting beads were then washed 3 times with Washing Buffer B (10 mM Tris-HCl, pH 7.5, 0.15 M LiCl, 1 mM EDTA, 0.1% LiDs, 0.1% Triton X-100). The RNA fraction bound to the beads was then eluted with 10 mM Tris-HCl (RNase-free) by heating at 75-80°C for 2 min. The entire Poly(A) selection procedure was repeated one more time. The final RNA sample was further purified by

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

MinElute kit (QIAGEN) to remove lithium salt, which otherwise would affect the activity of Bacterial Alkaline Phosphatase (BAP) in the subsequent step.

BAP/TAP treatment

1-2 μg of poly(A)+ RNA was first dephosphorylated in a 100 μl reaction (2.4 units Bacterial Alkaline Phosphatase (BAP; Takara), 50 mM Tris-HCl buffer (pH 9.0), 1 mM MgCl2, 50 mM NaCl and 100 units of RNasin Ribonuclease inhibitor (Promega)) at 37°C for 40 min. After phenol/chloroform extraction and ethanol precipitation, the resulting RNAs were treated with 20 units of TAP (Tobacco Acid Pyrophosphatase; Epicentre) in a 100 μl reaction (50 mM NaOAc (pH 6.0), 1 mM EDTA, 0.1% β-ME, 0.01% Triton X-100 and 100 unit RNasin) at 37°C for 1 hr. The reaction mixture was then extracted twice with phenol/ chloroform and the RNA fragments were ethanol precipitated for downstream linker ligation.

Linker ligation

A chimeric linker (5′-CTC AAG CTT CTA ACG ATG TAC GCT CGrA rGrUrC rCrArA rC-3′) was ligated to poly(A)+ RNA with BAP/TAP treatment. The ligation was performed in 100 μl reaction including the recovered RNA, 60 pmol PAGE-purified linker (IDT), 200 units of T4 RNA ligase1 (NEB), 50 mM Tris-HCl (pH 7.8), 10 mM MgCl2, 1 mM ATP, 10 mM DTT, 25% PEG8000 and 100 units of RNasin Ribonuclease inhibitor (Promega). The reaction mixture was incubated overnight at room temperature, followed by phenol/ chloroform extraction to remove both protein and PEG8000. Ethanol precipitation was then performed to recover the RNA by adding 1/10 volume of NaOAc (pH 5.2) and 30 μg of GlycoBlue (Ambion).

Reverse Transcription

Random primer with a common sequence (5′-GCG GCT GAA GAC GGC CTA TCC GAC NNN NNN-3′) was used to initiate the reverse transcription. Linker ligated RNAs were reversely transcribed in 40 μl reaction, which contains 20 pmol random primer, 2 nmol dNTP (Bioline), 240 ng actinomycin D, 80 units of RNasin Ribonuclease inhibitor (Promega), 400 units of Superscript III reverse transcriptase (Invitrogen), 50 mM Tris-HCl (pH 8.3), 75 mM KCl, 5 mM MgCl2 and 5 mM DTT. The reaction mixture was incubated at 25°C for 5min and 50°C for 1 hour, followed by heat inactivation at 70°C for 15min.

Circularization and rolling circle amplification

First-strand cDNAs were amplified by 2-5 cycles of PCR with high-fidelity DNA polymerase (Finnzymes) to generate blunt-end dsDNAs. Circularization of dsDNAs was achieved by bridge ligation. A 30 μl reaction was assembled which contains 1x Ampligase buffer (Epicentre), 7.5 units of OptiKinse (USB), 5 mM ATP and 1 mM DTT. The reaction was incubated at 37°C for 30 min, followed by 95 °C for 2 min to inactivate the kinase. 9 pmol “bridge” oligonucleotide (5′-GCC GTC TTC AGC CGC CTCA AGC TTC TAA CGA TGT ACG-3′) and 7.5 units of Ampligase (Epicentre) were then added. Ligation was performed with 5 cycles of 95°C for 30 sec, 68°C for 2 min, 55°C for 1 min and 60°C for 5 min, followed by 5 cycles of 95°C for 30 sec, 65°C for 2 min, 55°C for 1 min and 60°C for

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

5 min. To remove excess oligonucleotides and unligated DNA fragments, we added 3 μl Exonuclease I (NEB) and 0.6 μl Exonuclease III (NEB) to the ligation reaction and the mixture was incubated at 37°C for 45 min followed by 80°C for 20 min. This removes all linear DNA fragments (ssDNAs and dsDNAs) and the remaining circular DNAs were further amplified by rolling circle amplification (RCA). RCA was performed in four 20 μl reactions, each of which contains 2 μl DNA circles, 20 mM dNTP (Epicentre), 4 μg BSA (NEB), 200 pmol of random hexamer (5−-NNNN*N*N-3′; * = phosphothiol group), 10 units phi29 DNA polymerase (NEB), 2 μl DMSO and 1 x phi29 reaction buffer (NEB). The RCA reactions were incubated at 10°C for 10 min, 28°C for 16 hours and heat inactive at 65°C for 10 min.

Library quality control (QC)

Because the linker sequence has a built-in XhoI site, we use XhoI digestion to check the specificity of RCA amplification. Typically, the majority of the RCA products can be digested by XhoI and show an evenly-distributed smear centered around 300-400 bp. As another QC step, Sanger sequencing was also performed to check the quality before Illumina paired-end sequencing. After PCR amplification, a portion of the final library was A-tailed and ligated into T-easy vector (Promega). We followed the standard procedure recommended by the manufacturer and the T7 sequence was used as sequencing primer. In general, ~80-90% of the clones mapped to known TSS or its surrounding regions, consistent with the data generated by Illumina/Solexa sequencing.

Paired-end sequencing and read mapping

Two technical replicates of the embryo library were sequenced as 36mers from each side using Illumina GA II. Before mapping, we filtered low-quality reads and short tags with unidentified linker sequences. The Novoalign short read aligner (v1.05.02; www.novocraft.com; parameters: score difference = 30, report strategy = ’All’) was used to align the paired reads independently to the *D. melanogaster* genome (FlyBase v5.1433). All alignments with up to one mismatch beyond their optimally aligned location for each read were collected. Since 3′ reads might overlap a splice junction, we mapped all those 3′ reads to the transcriptome where the 5′ read of a pair aligned uniquely, and the 3′ read did not map at all to the genome. Similar to the genomic alignment, we collected all 3′ read locations with one mismatch beyond the optimally aligned location. 5′ and 3′ read pairs that mapped in the same orientation within 200,000 nt on the same chromosome were flagged as ‘aligned’. The cumulative Novoalign alignment score for both reads in the pair was used to classify the alignment specificity. Read transcript locations were classified into 6 possible categories based upon FlyBase: annotated TSS, ≤ 250nt upstream of an annotated TSS, 5′ UTR, coding region, 3′UTR, intron, and intergenic region. If a read could be classified into multiple categories because of overlapping transcripts, the read was assigned to one location based on the following priorities: (1) FlyBase annotated TSS, (2) 5′ UTR, (3) ≤ 250 nt upstream of an annotated TSS, (4) coding region and (5) intron.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Trimming of sequence adaptor

The raw data we obtained are paired 35mer reads, each of which consists of a ~20 nt tag (derived from MmeI digestion) followed by a 16nt linker sequence. The linker sequences were trimmed from the reads and subsequently used to identify which end of the transcript the read was from. Although both the 5′ and 3′ linker contain a MmeI site (5′-TCCAAC-3′, 5′-TCCGAC-3′, respectively), the sequences beyond their MmeI sites are completely different, thereby allowing for reliable determination of read directionality. Linker sequences were identified and trimmed such that there were no more than 2 mismatches/ indels between the 3′ end of the read and one of the two complete linker sequences. Read pairs in which either read from a pair failed to meet the linker sequence requirements were discarded from further analysis.

Transcription start site cluster identification

The feature density estimator F-Seq34 (parameters: feature length = 30, fragment size = 0) was applied to the 5′ reads of the uniquely aligned pairs from both replicates in order to create a smoothened estimate of read distributions. A genome-wide background density estimate was calculated by taking the mean of F-Seq values sampled from across the genome, with each chromosome being sampled in proportion to the number of reads aligned to it. Putative read clusters were defined as regions where the F-Seq value was greater than the background estimate. To eliminate lengthy tails in the distributions and create a robust cluster, we re-sized the clusters to the shortest distance that contained 95% of the reads. Clusters with tag numbers exceeding a stringent threshold (typically greater than 100 reads) were then considered as TSS clusters. Clusters were classified into different initiation patterns by the following definitions: NP clusters contained ≥ 50% of the reads within ±2 nt of the mode and span < 25 nt; BP clusters were those that contained ≥ 50% of the reads within ±2 nt of the mode and are ≥ 25 nt in length; all other clusters were classified as WP. TSS cluster locations were determined according to FlyBase, similar to individual reads. If a cluster overlapped an annotated TSS, it was classified as such; otherwise, the classification was based on the mode of the cluster. If the mode fell into multiple categories because of overlapping transcripts, the cluster was classified according to the priorities listed in the previous section. To summarize the terminology, a TSS refers to a genomic location to which at least one 5′ capped sequence tag was aligned; a TSS cluster is a distinct region of TSSs above background; and the initiation pattern describes the distribution of TSSs within a cluster. In all cases, the mode of the cluster is used as the reference TSS for a cluster.

Tag clustering strategy by F-Seq

F-Seq was used to perform the tag clustering34. The ‘fragment size’ parameter refers to the size of the fragment that needs to be clustered and analyzed. It has been shown that the fragment size should be set to ‘1 bp’ (equivalent to a value of 0) for data sets where one end (in our case, the 5′ end) of the sequence represents the point of enrichment. The ‘feature length’ parameter, on the other hand, controls the kernel density estimate bandwidth. The ‘feature length’ was set at 30, which means that the standard deviation of the Gaussian density estimate of a location has a value of 5bp.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Correlation between read counts and microarray expression values

Microarray expression values were collected from the NCBI GEO repository35: dataset accession number GSE11880. The data originated from three arrays containing expression values for wild type *Drosophila melanogaster* mixed embryos of stages 0-11 (GEO accession numbers: GSM300072, GSM300074, and GSM3000). The mean value across all three replicates, after median background subtraction, was used for our analysis. Genes with an average 'signal minus background' value less than 0 were given a log2-transformed expression value of 0. The total number of 3′ reads from the aligned pairs that mapped to the transcribed region of each gene was used for comparison with the microarray data. The Pearson correlation coefficient was calculated across 10,101 genes that had at least one read-pair mapped to them and was present in the microarray data. This analysis was done twice, the first was performed on all read-pairs that mapped to a gene; the second analysis used only non-redundant read pairs.

Identification of novel transcription start sites

We defined a novel transcription start site for a gene as a TSS cluster more than 250nt upstream of the most distally annotated start site according to FlyBase. Candidates for experimental validation were then required to contain a cluster with at least 100 reads, and with at least 80% of its 3′ paired reads mapping to a transcribed region of that gene. Novel 5′ exons identified in a previous analysis of whole-genome tiling expression arrays32 were transferred from Release 4 to Release 5 of the *D. melanogaster* genome. We excluded from the analysis any exons that overlap a transcribed region as defined by FlyBase. Due to the more limited resolution of the tiling arrays, a TSS cluster was considered as overlapping one of the tiling 5′ exons if it fell within 50nt of that exon.

Core promoter motif analysis

We considered the subset of TSS clusters overlapping an annotated TSS, ≤250nt upstream of a TSS, or in the 5′ UTR of a gene. The promoter sequences ±100nt surrounding the mode were extracted. The position weight matrix scanning program PATSER36 was applied to the plus strand of each sequence using pattern-specific background Markov models (Supplementary Table 7). The relative frequency matrices of six previously described core promoter motifs (Motif1, DRE, TATA, INR, Motif6, Motif7)18 as well as shortened non-overlapping matrices for the two motifs DPE and MTE25 were evaluated (Supplementary Table 8). All locations with a p-value ≤ 10^{-3} were deemed motif matches. Motif match counts were then binned into 5nt windows for each initiation pattern. To assess the background level of motif matches, the analysis was repeated on three sets of 1,000 random intergenic sites. The mean value within each bin was calculated. We define the preferred location for a motif as any 5nt window with a mean normalized count equal to or greater than 5-fold enrichment over background.

Experimental validation of novel TSSs and internally capped transcripts

Two independent approaches, oligo-capping and cap-trapping, were used. For the cap-trapping method, total RNA isolated from 0-24hr fly embryo (0-24h) was reversely transcribed with random hexamers and Superscript II reverse transcriptase. The resulting

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

RNA/cDNA hybrids were oxidized with 10 mM NaIO4 in 66 mM NaOAc (pH 4.5) by incubation on ice for 45 min. Biotinylation was then carried out by adding 10 mM biocytin hydrazide (Sigma) and 50 mM sodium citrate (pH 6.1), followed by incubation overnight at room temperature. The cDNA fragments, which are bound to capped RNA transcripts, are enriched by Dynabeads M-270 (Invitrogen), and subsequently ligated to a double-stranded adaptor (5′-AGC TTC TAA CGA TGT ACG CTC GAG TCC AAC NN-3′ and 3′-TCG AAG ATT GCT ACA TGC GAG CTC AGG TTGp-5′) using T4 DNA ligase (NEB). For each candidate transcript to be validated, linker-ligated cDNAs were used as templates; PCR reaction was carried out with a junction primer (which spanning the linker and 5′ gene specific sequence of the TSS cluster mode) and a downstream gene-specific primer (100-200bp distance). As negative control, total RNA was pre-treated with TAP (Tobacco Acid Pyrophosphatase) and processed side-by-side with the RNA sample without TAP treatment (or with 5′ cap structure).

For the validation using oligo-capping strategy, total RNA of the fly embryo (0-24h) was BAP/TAP treated. A chimeric linker was ligated to the released 5′ phosphate group. Reverse transcription was performed following the same procedure as shown in the library construction. A junction primer spanning the linker and 5′ gene specific sequence of the TSS cluster mode, together with a downstream primer (100-200 bp distance, Supplementary Table 9), were used to carry out PCR reaction to validate 5′ sequence immediately downstream of the cap structure. The RNA sample without 5′-linker ligation was used as negative control.

ChIP-chip transcription factor binding analysis

We collected 612 TBP, 1,073 TRF2, and 298 TBP/TRF2 binding sites from Isogai *et al*.6, and converted the release 4 coordinates to release 5 using the FlyBase map coordinate converter37. For comparison, we selected read clusters further than 500 nt from any other cluster; this was necessary because of the limited resolution of the ChIP-chip data. A read cluster was counted as being bound by one or both of the factors if the mode of the cluster was within 50nt of a binding site. Overall, we had 63 clusters bound by TBP, 432 clusters bound by TRF2, and 47 additional clusters bound by both TBP and TRF2. To account for the different coverage of initiation patterns by ChIP-chip, the percentage of TF binding was calculated from counts normalized to the number of occurrences per 1,000 TSSs per 1,000 ChIP-chip binding sites and then divided by the normalized number of promoters with TF binding.

REFERENCES

1. Juven-Gershon T, Kadonaga JT. Regulation of gene expression via the core promoter and the basal transcriptional machinery. Dev. Biol. 2010; 339:225–229. [PubMed: 19682982]
2. Ohler U, Wassarman DA. Promoting developing transcription. Development. 2010; 137:15–26. [PubMed: 20023156]
3. Butler JE, Kadonaga JT. Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. Genes Dev. 2001; 15:2515–2519. [PubMed: 11581157]
4. Hochheimer A, Zhou S, Zheng S, Holmes MC, Tjian R. TRF2 associates with DREF and directs promoter-selective gene expression in Drosophila. Nature. 2002; 420:439–445. [PubMed: 12459787]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

5. Holmes MC, Tjian R. Promoter-selective properties of the TBP-related factor TRF1. Science. 2000; 288:867–870. [PubMed: 10797011]

6. Isogai Y, Keles S, Prestel M, Hochheimer A, Tjian R. Transcription of histone gene cluster by differential core-promoter factors. Genes Dev. 2007; 21:2936–2949. [PubMed: 17978101]

7. Shiraki T, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A. 2003; 100:15776–15781. [PubMed: 14663149]

8. Suzuki Y, Sugano S. Construction of a full-length enriched and a 5′-end enriched cDNA library using the oligo-capping method. Methods Mol Biol. 2003; 221:73–91. [PubMed: 12703735]

9. Zhang Z, Dietrich FS. Mapping of transcription start sites in Saccharomyces cerevisiae using 5′ SAGE. Nucleic Acids Res. 2005; 33:2838–2851. [PubMed: 15905473]

10. Ahsan B, et al. MachiBase: a Drosophila melanogaster 5′-end mRNA transcription database. Nucleic Acids Res. 2009; 37:D49–53. [PubMed: 18842623]

11. Carninci P, et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet. 2006; 38:626–635. [PubMed: 16645617]

12. Suzuki H, et al. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. Nat Genet. 2009; 41:553–562. [PubMed: 19377474]

13. Valen E, et al. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. Genome Res. 2009; 19:255–265. [PubMed: 19074369]

14. Fejes-Toth K, et al. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. Nature. 2009; 457:1028–1032. [PubMed: 19169241]

15. Esteban JA, Salas M, Blanco L. Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. J Biol Chem. 1993; 268:2719–2726. [PubMed: 8428945]

16. Carninci P, et al. The transcriptional landscape of the mammalian genome. Science. 2005; 309:1559–1563. [PubMed: 16141072]

17. Wilhelm BT, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature. 2008; 453:1239–1243. [PubMed: 18488015]

18. Ohler U, Liao GC, Niemann H, Rubin GM. Computational analysis of core promoters in the Drosophila genome. Genome Biol. 2002; 3 RESEARCH0087.

19. Purnell BA, Emanuel PA, Gilmour DS. TFIID sequence recognition of the initiator and sequences farther downstream in Drosophila class II genes. Genes Dev. 1994; 8:830–842. [PubMed: 7926771]

20. Burke TW, Kadonaga JT. Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. Genes Dev. 1996; 10:711–724. [PubMed: 8598298]

21. FitzGerald PC, Sturgill D, Shyakhtenko A, Oliver B, Vinson C. Comparative genomics of Drosophila and human core promoters. Genome Biol. 2006; 7:R53. [PubMed: 16827941]

22. Sandelin A, et al. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nat Rev Genet. 2007; 8:424–436. [PubMed: 17486122]

23. Megraw M, Pereira F, Jensen ST, Ohler U, Hatzigeorgiou AG. A transcription factor affinity-based code for mammalian transcription initiation. Genome Res. 2009; 19:644–656. [PubMed: 19141595]

24. Ng P, et al. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. Nat Methods. 2005; 2:105–111. [PubMed: 15782207]

25. Rach EA, Yuan HY, Majoros WH, Tomancak P, Ohler U. Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. Genome Biol. 2009; 10:R73. [PubMed: 19589141]

26. Akhtar W, Veenstra GJ. TBP2 is a substitute for TBP in Xenopus oocyte transcription. BMC Biol. 2009; 7:45. [PubMed: 19650908]

27. Gazdag E, et al. TBP2 is essential for germ cell development by regulating transcription and chromatin condensation in the oocyte. Genes Dev. 2009; 23:2210–2223. [PubMed: 19759265]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

28. Shibuya T, Tange TO, Sonenberg N, Moore MJ. eIF4AIII binds spliced mRNA in the exon junction complex and is essential for nonsense-mediated decay. Nat Struct Mol Biol. 2004; 11:346–351. [PubMed: 15034551]

29. Schoenberg DR, Maquat LE. Re-capping the message. Trends Biochem Sci. 2009; 34:435–442. [PubMed: 19729311]

30. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008; 322:1845–1848. [PubMed: 19056941]

31. Nechaev S, et al. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. Science. 2010; 327:335–338. [PubMed: 20007866]

32. Manak JR, et al. Biological function of unannotated transcription during the early development of Drosophila melanogaster. Nat Genet. 2006; 38:1151–1158. [PubMed: 16951679]

33. Tweedie S, et al. FlyBase: enhancing Drosophila Gene Ontology annotations. Nucleic Acids Res. 2009; 37:D555–559. [PubMed: 18948289]

34. Boyle AP, Guinney J, Crawford GE, Furey TS. F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics. 2008; 24:2537–2538. [PubMed: 18784119]

35. Barrett T, et al. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res. 2009; 37:D885–890. [PubMed: 18940857]

36. Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics. 1999; 15:563–577. [PubMed: 10487864]

37. Wilson RJ, Goodman JL, Strelets VB. FlyBase: integration and improvements to query tools. Nucleic Acids Res. 2008; 36:D588–593. [PubMed: 18160408]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

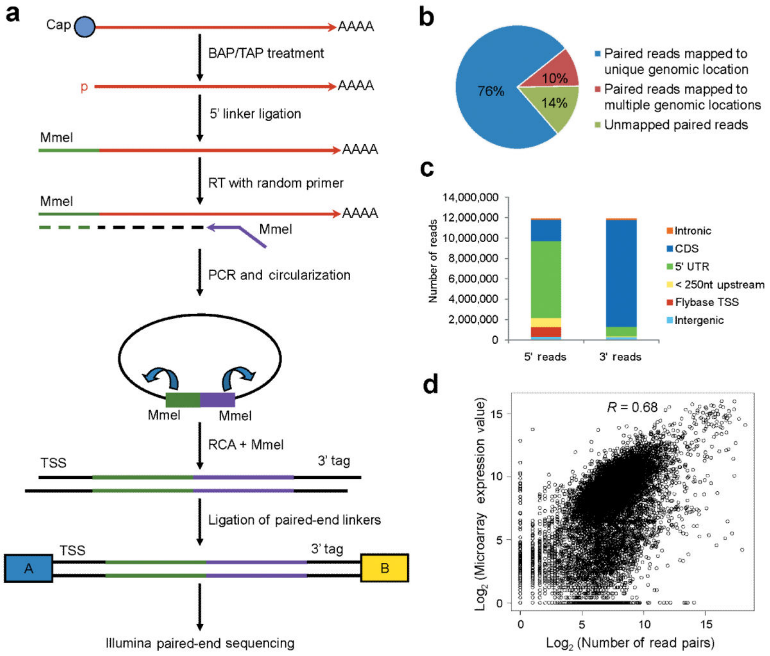

Figure 1. Paired-End Analysis of Transcriptional start sites (PEAT)

(**a**) Schematic outline of the PEAT strategy. The RNA fragment is shown as an arrowed line (red), the two Mme I sites induced at the oligo-capping and reverse transcription (RT) steps are shown in green and purple, respectively. (**b**) Mapping efficiency of the reads that have built-in linker sequences, combined from two technical replicates. (**c**) The distribution of uniquely mapped 5′ and 3′ reads relative to known TSSs and other genomic regions. (**d**) Comparison between PEAT and microarray expression data. 10,101 genes were plotted that had at least 1 mapped read-pair and were included in the microarray data. For the array data, expression level is the mean of simple background subtraction values across 3 replicates from mixed stage 0-11 *D. melanogaster* embryos. To estimate the expression level using paired-end sequencing data, we used the counts of 3′ tags that map to a transcribed region. Correlation coefficient was determined by Pearson correlation.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

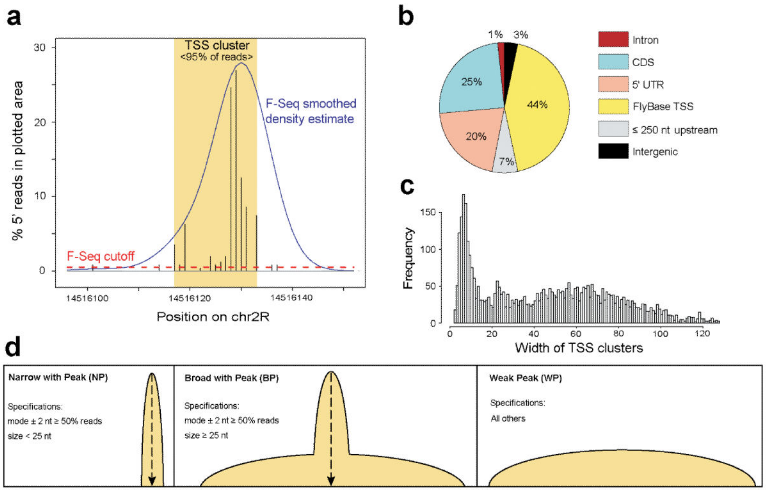

Figure 2. TSS clusters and initiation patterns identified in the *Drosophila* embryo

(**a**) The approach for identifying TSS clusters. A representative example (Chr. 2: 14516000-14516600) is shown. In essence, a smoothed density estimate of 5′ TSS tags was computed (blue line). Cluster boundary was then determined as exceeding a baseline score, estimated on a genomic background (red line). TSS clusters were further condensed to the shortest distance containing 95% of the reads (dark shaded area). (**b**) The genomic locations of all clusters that contain ≥ 100 reads. Clusters overlapping an annotated TSS in FlyBase were classified as FlyBase TSS. For the remaining clusters, classifications were based on the mode of each given cluster and its relative location to annotated transcripts. (**c**) Size distribution of all clusters with ≥ 100 reads. Cluster sizes are similar to previous reports for mammals, with the majority of clusters shorter than 120nt in length. (**d**) Definition of initiation patterns.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

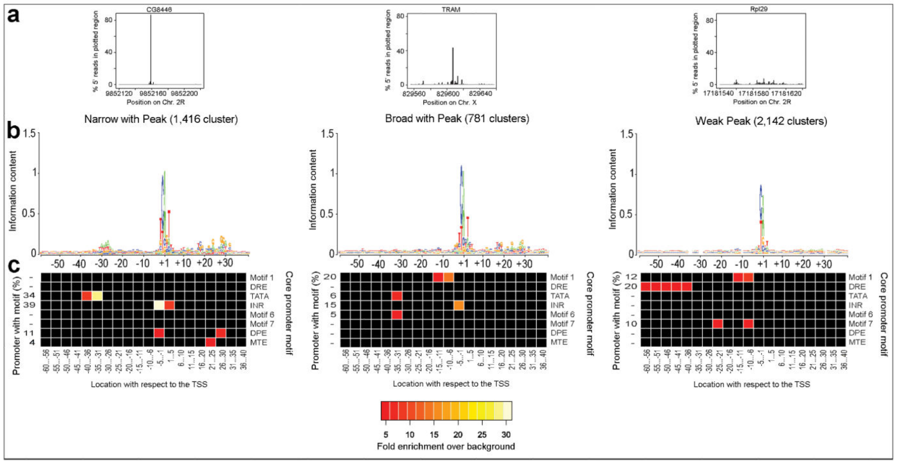

Figure 3. Promoter motifs associated with distinct promoter types

(**a**) The three initiation patterns, NP, BP and WP, are each represented by a candidate locus. The graphs show the relative percentage of 5′ reads that are mapped within a 100nt window. (**b**) Sequence landscape in the promoter region of each pattern. The mode location of each cluster is set as reference point ‘+1’. Sequence logos of 100-nt window are shown. (**c**) The core promoter motifs overrepresented for each initiation pattern. Significant motifs were identified in 200nt core promoter sequences and binned into 5nt intervals; only the 100nt region surrounding the TSS is shown as no motifs were found to be enriched outside of this window. All bins with normalized motif occurrences of 5-fold enriched or above are shown. The percent of sequences containing at least one high-stringency instance of each motif in its preferred location is listed on the left side of the heat map.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

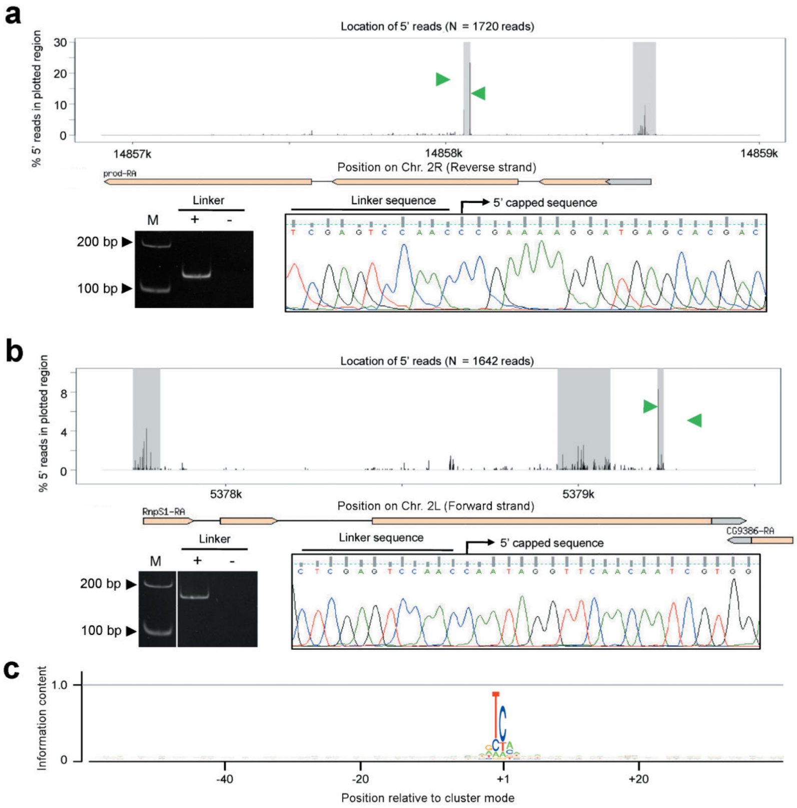

Figure 4. A distinct sequence motif identified for internally capped transcripts

(**a-b**) The gene structures of the *PROD* and *RNPS1* loci indicating exons (thick bar) and introns (thin bar) from FlyBase are shown. A thick grey bar represents the UTR region. Grey areas highlight read clusters (≥ 100 reads/cluster). Green arrows denote primer locations for RT-PCR validation. A junction primer, which spans the linker and 5′ gene specific sequence at the cluster mode, together with a downstream primer (100-200 bp distance) were used to carry out RT-PCR. For each locus, cDNAs derived from RNA samples with (+) or without (−) linker ligation were used as template. The DNA ladder (M) is shown in the left lane. Sanger sequencing results show the correct position of the mode of the called TSS cluster for (**a**) a capped 5′ read cluster in the middle of a coding region; and (**b**) an example of a capped 5′ read cluster near the end of the coding region. (**c**) Sequence logo of a 100 nt window around the mode location (identified as '+1') of all clusters containing more than 100 reads and mapping to a coding region.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Summary of PEAT generated data

	Replicate 1	Replicate 2	Combined
Number of Read-Pairs with Identifiable Linker Sequences	8,258,735	7,470,183	15,728,918
Read-Pairs Mapped to a Unique Genomic Location	6,246,759	5,653,860	11,900,619
Read-Pairs Mapped to Multiple Genomic Locations	862,748	782,828	1,645,576
Non-Redundant Read-Pairs	4,103,558	3,752,136	7,062,714
Non-Redundant Read-Pairs Mapped to a Unique Genomic Location	1,688,228	1,569,274	2,716,981
Genes Represented by at Least 1 Read-Pair	11,111	11,073	11,418
Genes With An Identified Read Cluster Consisting of More Than 10 5′Reads	--	--	8,577
Genes With An Identified Read Cluster Consisting of More Than 50 5′Reads	--	--	5,563
Genes With An Identified Read Cluster Consisting of More Than 100 5′Reads	--	--	4,007

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Table 2

Frequency of consensus di- and tri-nucleotides relative to the TSSs and coding region

	$T^{-2}C^{-1}A^{+1}$	$C^{-1}A^{+1}$	$T^{-}C^{+1}A^{+2}$	$T^{-}C^{+1}$
Narrow with Peak	550 (44%)	858 (68%)	13 (1%)	86 (7%)
Broad with Peak	274 (36%)	483 (64%)	13 (2%)	50 (7%)
Weak Peak	387 (19%)	973 (48%)	46 (2%)	128 (6%)
Coding Region Read Cluster	24 (2%)	108 (8%)	476 (35%)	804 (59%)

Note: The +1 position within each cluster is defined by the mode of that cluster, that is, oblivious to its location in the genome. We here show the analysis comparing coding region clusters to those near the start site of a gene. (Out of 5699 clusters, 426 clusters which fell into either intergenic or intronic regions were not included in the analysis).