Supplementary Figure 1. Distribution of 5’ reads relative to annotated FlyBase TSSs

All aligned 5’ reads were included in the analysis except those mapped to intronic regions; these reads were excluded because the distribution is based on the distance within the transcript from the TSS, not the overall genomic distance. For intergenic reads, the distance was determined based on the nearest downstream annotated TSS. The tag frequency for each group (binned for every 10 bp) is shown. The results showed a clear peak at the annotated TSSs with a long tail extending into regions downstream of TSSs. This is in agreement with previous observations\(^1\) which indicated that Flybase annotations are generally based on the longest known transcript rather than the most frequent one (i.e. the mode of the cluster, as used in this study).

\(^1\) Rach, E.A., Yuan, H.Y., Majoros, W.H., Tomancak, P. & Ohler, U. Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. Genome biology 10, R73 (2009).
Supplementary Figure 2. Distribution of the distance between 5’ and 3’ reads at the transcript level

The distance between 5’ and 3’ tags in each paired read was determined based on annotated transcripts in FlyBase. The distances were binned at a 10-bp resolution (X axis), and the abundance of each group is shown (Y axis). The median distance between the 5’ and 3’ reads at the transcript level was 279 bp.
Supplementary Figure 3. Distribution of TSS tag counts in annotated genes

The count of uniquely mapped 5’ TSSs in each gene was computed. The results were log$_2$ transformed and plotted. On average, there are 256 tags per gene (log$_2$ count = ~ 8), indicating that our data set has deep coverage for annotated genes, including relatively rare transcripts.
Supplementary Figure 4. Comparison between the PEAT results and MachiBase

MachiBase is based on 5’ CAGE tags only, and the comparison of the read count per gene between the two datasets was therefore based on the number of 5’ reads in the TSS proximal regions, including those that are ≤ 250nt upstream of an annotated TSS, overlap the TSS, or are found within the 5’ UTR. The correlation coefficient ($R = 0.53$) between PEAT library and MachiBase is based on all genes that had at least one tag in each dataset. The relatively low correlation might be in part due to the considerably larger number of reads per gene in the PEAT dataset compared to those available from MachiBase.
Supplementary Figure 5. High reproducibility of the PEAT method

Correlation coefficient ($R$) between two replicates was computed based on the number of 3’ reads mapped to the transcribed region of individual genes. (a) Employs all read pairs mapped to a single gene. (b) The read pairs were further collapsed to non-redundant reads, i.e. multiple reads with identical 5’ and 3’ tags were counted once. Under both conditions, the two replicates showed very high correlation, suggesting the PEAT method is highly reproducible.

The results also demonstrated that little bias was introduced during library construction since the non-redundant reads show consistent results as well. Thus, all mapped reads, rather than non-redundant reads, were used in the analyses of TSS clustering and initiation pattern identification. This allowed us to improve the overall coverage and accuracy.
Supplementary Figure 6. Comparison between the PEAT and microarray-based approach

Microarray dataset (Y axis) was obtained from GEO: GSE11880. The data set contained a mix of samples from *Drosophila* embryos of stages 0-11. Expression level of each gene was computed by averaging the signal minus background values from three separate arrays. To determine the expression level from the PEAT results (X axis), we computed the number of non-redundant read pairs by consolidating those with 3’ tags mapped in the same transcribed region. 10,101 genes were included in the analysis that were present in the array and had at least one read-pair mapped to them.
Supplementary Figure 7. Validation of novel TSSs by oligo-capping

10 novel TSSs were selected for validation by RT-PCR assay. To ensure that the final amplification products are derived from capped transcripts, junction primers were used, each of which contains a partial 5' linker sequence (resulting from oligo-capping) and a gene-specific sequence at the mode of a putative TSS. The reverse primer was designed to be 100-200bp downstream of the candidate novel TSS. Amplification specificity was controlled by comparing cDNA fragments generated from RNA samples with (+ linker) and without (-linker) oligo-capping. Specific bands with the expected size were detected in 7 out the 10 cases, confirming novel TSSs at the distant sites.

Order of target genes (FlyBase) in Supplementary Figure 7 and 8:

|   |   |   |   |   |
|---|---|---|---|---|
| 1 | FBgn0028537 |   |   |   |
| 2 | FBgn0033068 |   |   |   |
| 3 | FBgn0037410 |   |   |   |
| 4 | FBgn0033113 |   |   |   |
| 5 | FBgn0033688 |   |   |   |
| 6 | FBgn0260442 |   |   |   |
| 7 | FBgn0004228 |   |   |   |
| 8 | FBgn0085320 |   |   |   |
| 9 | FBgn0030136 |   |   |   |
| 10| FBgn0024366 |   |   |   |
Supplementary Figure 8. Validation of novel TSSs by cap-trapping

As an independent approach, cap-trapping was used to validate the 10 novel TSSs shown in Supplementary Figure 8. As negative control, total RNAs were pretreated with TAP to remove the cap. The resulting RNAs (- 5’cap) were processed side-by-side with the RNAs with 5’ cap (+ 5’ cap) along the cap-trapping procedure (see Online Methods for detail). To ensure that the final amplification products are derived from capped transcripts, junction primers were used, each of which contains a partial 5’ linker sequence (resulting from cap-trapping) and a gene-specific sequence at the mode of a putative TSS. The reverse primer was designed to be 100-200bp downstream of the candidate novel TSS. Specific bands with the expected size were detected in 8 out the 10 cases, confirming novel TSSs at the distant sites.
Supplementary Figure 9. Motif prevalence at preferred locations

Motif searches were carried out with PATSER in the promoter sequences for each initiation pattern. The analyses were conducted at different P-value stringency thresholds: 0.001, 0.005 and 0.01. For each known promoter motif, the results are presented as the percentage of the promoters with at least one respective motif found at the preferred location. As expected, the motif prevalence increases as the threshold stringency decreases. Results at the most stringent condition (0.001) are shown in Figure 3C of the main text.
Supplementary Figure 10. A validated example of internally capped transcripts

Analogous to Figure 4, the 5’ TSS reads obtained from PEAT and the FlyBase annotation are shown for the another example, the Elf gene locus (upper panel). The areas shaded in grey represent the called read clusters ($\geq$ 100 reads/cluster). Two gene-specific primers (green arrows) were designed; the forward primer contains a partial sequence of a 5’ linker, which was added to all capped transcripts by oligo-capping. PCR amplification was conducted with first-strand cDNA generated from total RNAs (0-24hr embryos) with or without oligo-capping (+ or – linker). The latter served as a negative control to ensure all amplification products are generated from capped transcripts. Correct amplification products were detected as resolved by gel electrophoresis (lower left panel). The PCR products were further confirmed by Sanger sequencing and the junction region (5’ linker and its immediate downstream sequence) is shown.
Supplementary Figure 11. Validation of 5’ capped reads in CDS by oligo-capping

12 cases of internally capped transcripts were selected for further validation. RT-PCR was performed and a similar strategy was used as shown in Figure 4 and Supplementary Figure 10. For each primer pair, the distance between two gene-specific primers was between 100-200 bp. Two RT-PCR reactions were performed for each locus using RNA samples with or without oligo-capping. The PCR products were resolved by gel electrophoresis. Correct bands with expected sizes were detected in 11 out of the 12 cases. For the single failed case (sample 11), a weak band was also observed (< 100 bp) and is likely due to nonspecific amplification.

Candidate genes selected for validation in Supplementary Figure 11 and 12:

| 1: FBgn0014269 | 2: FBgn0020443 | 3: FBgn0029629 | 4: FBgn0030341 |
|----------------|----------------|----------------|----------------|
| 5: FBgn0031769 | 6: FBgn0035121 | 7: FBgn0037301 | 8: FBgn0037707 |
| 9: FBgn0003870 | 10: FBgn0024841| 11: FBgn0026188| 12: FBgn0051729|
Supplementary Figure 12. Validation of 5' capped reads in CDS by cap-trapping

As an independent approach, cap-trapping was used again to validate the 12 cases of internally capped transcripts shown in Supplementary Figure 11. The same experimental procedure as described in Supplementary Figure 7 was used. Correct bands with expected sizes were detected in 10 out of the 12 cases. 10 candidates were validated by both oligo-capping and cap-trapping methods.
Supplementary Figure 13. Comparison of the 5’ TSS reads mapped to the CDS and initiation regions

TSS reads or clusters mapped in close proximity to annotated genes were broadly divided into two categories. A TSS read (or cluster) was classified as “CDS” if it mapped to a coding region. Likewise, the TSS reads were defined as “initiation region” if they overlapped with a TSS, fall within ≤ 250 nt upstream of a TSS or within the 5’ UTR. Direct comparison between the two TSS categories was made by using either (a) the total number of 5’ reads or (b) the largest cluster of each category. To reduce potential complications due to stochastic noise, only genes with a TSS cluster containing more than 100 reads in either the “CDS” group or the “initiation region” group, or both, were considered in (b).
Supplementary Figure 14. Distribution of internally capped TSS clusters across exons

The internal read clusters are defined as those containing more than 100 reads and are mapped to a coding region. The mode of each cluster was used to determine its relative location within an exon. The results were normalized by the exon length and further divided among three promoter classes (NP, BP and WP). To accommodate reads spanning across the exon junction, we first aligned the paired-reads to the genome, and then the transcriptome when the 5’ read could be mapped to the genome and the 3’ read was not. Details are described in the Online Methods section.
Supplementary Figure 15. TBP and TRF2 are associated with different promoter classes and, or core promoter motifs

(a-c) Number of clusters overlapping binding sites for TBP, TRF2, or both TBP and TRF2 were counted and normalized to the number of occurrences per 1,000 read clusters.

(d) Relative frequencies of factors bound to the different shape classes, for those clusters that have at least one of the factors bound.
Supplementary Figure 16. Pol II binding in association with transcription factors

The percentage of Pol II bound sites was calculated by dividing the number of sites bound by Pol II to the total number of sites bound by each transcription factor (TF).
Supplementary Tables

Supplementary Table 1. Comparison with a previous *Drosophila* 5’ CAGE study

|                              | PEAT*                  | MachiBase**             |
|------------------------------|------------------------|-------------------------|
| Total Alignable Reads        | 15,728,918             | 5,619,701               |
|                              | (two channels)         | (one channel)           |
| Reads mapped to unique location | 11,900,619             | 3,512,967               |
| % of uniquely mapped reads   | 75.7%                  | 62.5%                   |
| Reads mapped to multiple location | 1,645,576             | 234,519                 |
| % of reads mapped to multiple locations | 10.4%       | 4.2%                    |
| Non-redundant 5’ reads       | 1,168,474              | 306,829                 |
| Genes Represented by at Least 1 Read | 11,418                | 10,196                  |
| Genes With An Identified Read Cluster Consisting of More Than 10 5’ Reads*** | 8,577              | 4,799                   |
| Genes With An Identified Read Cluster Consisting of More Than 50 5’ Reads*** | 5,563              | 2,406                   |
| Genes With An Identified Read Cluster Consisting of More Than 100 5’ Reads*** | 4,007              | 1,644                   |
| % clusters within coding region containing ≥ 100 Reads | 25%                  | 19%                     |
| % clusters of class NP with ≥ 100 Reads | 33%                  | 37%                     |
| % clusters of class BP with ≥ 100 Reads | 18%                  | 23%                     |
| % clusters of class WP with ≥ 100 Reads | 49%                  | 40%                     |

* Data of Embryo (0-24h)
** Data of Embryo as defined by Ahsan *et al*.
*** Clusters within 250 nt upstream of a currently annotated transcript were included

1 Ahsan, B. *et al*. MachiBase: a *Drosophila* melanogaster 5’-end mRNA transcription database. Nucleic Acids Res 37, D49-53 (2009).
### Supplementary Table 2. Improved mapping of raw data by paired reads

|                                | Replicate 1 | Replicate 2 | Combined |
|--------------------------------|-------------|-------------|----------|
| Pairs with linker sequence     | 8,258,735   | 7,470,183   | 15,728,918 |
| 5’ reads that would have mapped to multiple locations | 303,737     | 273,020     | 576,757   |
| % of 5’ reads that would have mapped to multiple locations | 3.68%       | 3.65%       | 3.67%     |
| 5’ reads that would have mapped to a different location | 28,343      | 25,176      | 53,519    |
| % of 5’ reads that would have mapped to different location | 0.34%       | 0.34%       | 0.34%     |
**Supplementary Table 3. Candidate distal TSSs selected for validation**

| Gene       | Strand | Chromosome | TSS Location | Tested | Confirmed by oligo-capping | Confirmed by cap-trapping | TSS Cluster Shape |
|------------|--------|------------|--------------|--------|-----------------------------|---------------------------|-------------------|
| FBgn0028537 | -      | 2L         | 14741837     | Yes    | +                           |                           | NP                |
| FBgn0033113 | +      | 2R         | 2768500      | Yes    | -                           | -                         | NP                |
| FBgn0033068 | -      | 2R         | 2083769*     | Yes    | +                           | +                         | NP                |
| FBgn0033688 | -      | 2R         | 8050517      | Yes    | +                           | +                         | BP                |
| FBgn0260442 | +      | 3L         | 8553290      | Yes    | -                           | +                         | NP                |
| FBgn0004228 | -      | 3L         | 15512716     | Yes    | -                           | +                         | NP                |
| FBgn0037410 | +      | 3R         | 2035098      | Yes    | +                           | +                         | NP                |
| FBgn0085320 | -      | 3R         | 22251501     | Yes    | +                           | +                         | NP                |
| FBgn0030136 | +      | X          | 9319746      | Yes    | +                           | -                         | NP                |
| FBgn0024366 | -      | X          | 1232255      | Yes    | +                           | +                         | NP                |

*Also Identified by Manak et al.¹.

¹ Manak, J.R. et al. Biological function of unannotated transcription during the early development of Drosophila melanogaster. Nature genetics 38, 1151-1158 (2006).
Supplementary Table 4. Three classes of capped clusters in coding region

|                      | Narrow with Peak (NP) | Broad with Peak (BP) | Weak Peak (WP) |
|----------------------|-----------------------|----------------------|----------------|
| # of clusters        | 73                    | 125                  | 1,162          |
| # of clusters with at least 1 TFBS in a preferred location* | 2                     | 3                    | 67             |

* Preferred binding location for any given transcription factor can be seen in Figure 3 of the main text. A p-value cutoff of 0.001 was used for the motif search.
Supplementary Table 5. Candidate internal capping selected for validation

| Gene      | Strand | Chromosome | Cluster Mode | Location | Confirmed by oligo-capping | Confirmed by cap-trapping |
|-----------|--------|------------|--------------|----------|----------------------------|----------------------------|
| FBgn0003870 | +      | 3R         |              | 27553415 | +                          | +                          |
| FBgn0014269 | -      | 2R         |              | 14858076  | +                          | +                          |
| FBgn0020443 | +      | 2L         |              | 12438946  | +                          | +                          |
| FBgn0024841 | +      | 3R         |              | 25549175  | +                          | +                          |
| FBgn0026188 | +      | 3R         |              | 5271105   | -                          | -                          |
| FBgn0029629 | -      | X          |              | 2503629   | +                          | +                          |
| FBgn0030341 | -      | X          |              | 11746055  | +                          | +                          |
| FBgn0031769 | +      | 2L         |              | 6052196   | +                          | +                          |
| FBgn0035121 | -      | 3L         |              | 259680    | +                          | +                          |
| FBgn0037301 | +      | 3R         |              | 1058801   | +                          | +                          |
| FBgn0037707 | +      | 3R         |              | 5379226   | +                          | +                          |
| FBgn0051729 | +      | 2L         |              | 13299058  | +                          | -                          |
Supplementary Table 6. Frequency of TSS clusters bound by TBP, TRF2, or both in promoters with or without TATA-box

| TSS Pattern       | TF     | TATA | No TATA |
|-------------------|--------|------|---------|
| **Narrow with Peak** |        |      |         |
| TBP               | 76%    | 24%  |         |
| TRF2              | 11%    | 89%  |         |
| TBP,TRF2          | 0%     | 100% |         |
| **Broad with Peak** |        |      |         |
| TBP               | 63%    | 37%  |         |
| TRF2              | 19%    | 81%  |         |
| TBP,TRF2          | 15%    | 85%  |         |
| **Weak Peak**     |        |      |         |
| TBP               | 37%    | 63%  |         |
| TRF2              | 13%    | 87%  |         |
| TBP,TRF2          | 0%     | 100% |         |
Supplementary Table 7. Read cluster specific background markov models used in identifying core promoter motifs

|                                | A    | C    | G    | T    |
|--------------------------------|------|------|------|------|
| Narrow with Peak               | 0.262| 0.242| 0.236| 0.260|
| Broad with Peak                | 0.280| 0.222| 0.216| 0.281|
| Weak Peak                      | 0.298| 0.214| 0.204| 0.285|
| Coding Region Read Clusters    | 0.260| 0.261| 0.256| 0.223|
| Random Intergenic Sites Set #1 | 0.292| 0.203| 0.207| 0.297|
| Random Intergenic Sites Set #2 | 0.295| 0.205| 0.205| 0.295|
| Random Intergenic Sites Set #3 | 0.299| 0.202| 0.202| 0.297|
Supplementary Table 8. Core promoter motif position weight matrices used in the PATSER search

| Motif 1 | DRE | TATA | INR |
|---------|-----|------|-----|
| A      | 84  | 137  | 45  |
| C      | 75  | 35   | 49  |
| G      | 105 | 52   | 89  |
| T      | 47  | 53   | 68  |
| A      | 111 | 78   | 47  |
| C      | 101 | 52   | 59  |
| G      | 73  | 104  | 110 |
| T      | 26  | 43   | 35  |
| A      | 76  | 73   | 33  |
| C      | 68  | 108  | 95  |
| G      | 42  | 24   | 114 |
| T      | 125 | 72   | 9  |
| A      | 138 | 61   | 33  |
| C      | 76  | 32   | 5  |
| G      | 63  | 1    | 241 |
| T      | 34  | 183  | 72  |
| A      | 4   | 263  | 263 |
| C      | 173 | 0    | 0   |
| G      | 7   | 0    | 0   |
| T      | 127 | 0    | 277 |
| A      | 0   | 266  | 0   |
| C      | 0   | 0    | 0   |
| G      | 309 | 0    | 277 |
| T      | 2    | 0    | 0   |
| A      | 0   | 106  | 106 |
| C      | 32  | 14   | 10  |
| G      | 18  | 30   | 53  |
| T      | 47  | 40   | 48  |
| A      | 3   | 52   | 1   |
| C      | 0   | 8    | 48  |
| G      | 1   | 11   | 58  |
| T      | 1    | 8    | 1   |
| A      | 9   | 44   | 67  |
| C      | 0   | 17   | 76  |
| G      | 3   | 17   | 129 |
| T      | 5   | 31   | 76  |
| A      | 1   | 47   | 112 |
| C      | 0   | 15   | 49  |
| G      | 0   | 17   | 17   |
| T      | 2    | 31   | 17   |
| A      | 8   | 23   | 83  |
| C      | 0   | 8    | 1    |
| G      | 48  | 0    | 101  |
| T      | 52  | 3    | 1    |
| A      | 0   | 36   | 98  |
| C      | 23  | 0    | 0    |
| G      | 2   | 106  | 0    |
| T      | 18  | 0    | 0    |
| A      | 68  | 38   | 68   |
| C      | 13  | 38   | 68   |
| G      | 191| 38   | 68   |
| T      | 39  | 77   | 68   |

| Motif 6 | DPE | Motif 7 | MTE |
|---------|-----|---------|-----|
| A      | 6   | 30    | 0   |
| C      | 0   | 14    | 18  |
| G      | 32  | 10    | 53  |
| T      | 18  | 52    | 112 |
| A      | 3   | 8     | 2   |
| C      | 52  | 11    | 58  |
| G      | 0   | 30    | 58  |
| T      | 58  | 31    | 27  |
| A      | 9   | 11    | 52  |
| C      | 0   | 17    | 8   |
| G      | 44  | 17    | 76  |
| T      | 47  | 7     | 112 |
| A      | 1   | 32    | 40  |
| C      | 40  | 14    | 8   |
| G      | 0   | 10    | 26  |
| T      | 0   | 17    | 0   |
| A      | 106 | 106   | 0   |
| C      | 0   | 0     | 0   |
| G      | 0   | 0     | 107 |
| T      | 0   | 0     | 107 |
| A      | 25  | 0     | 54  |
| C      | 0   | 0     | 52  |
| G      | 82  | 0     | 19  |
| T      | 0   | 0     | 72  |
| A      | 7   | 0     | 0   |
| C      | 0   | 0     | 197 |
| G      | 0   | 0     | 0   |
| T      | 100 | 0     | 0   |
| A      | 1   | 0     | 0   |
| C      | 1   | 105   | 0   |
| G      | 105 | 23    | 0   |
| T      | 67  | 73    | 133 |
| A      | 11  | 6     | 4   |
| C      | 40  | 12    | 60  |
| G      | 12  | 37    | 18  |
| T      | 7    | 31    | 29  |
| A      | 46  | 23    | 85  |
| C      | 31  | 31    | 7   |
| G      | 7   | 7     | 4   |
| T      | 185 | 185   | 24  |
Supplementary Table 9. Primer pairs for validation of 5' capped reads in CDS:

| Gene name in FlyBase | Forward junction primer | Reverse GSP primer |
|----------------------|-------------------------|-------------------|
| FBgn0003870          | GCTCGAGTCCAACCATATACTAAGGAT | ACCTCGGTGAGGCCCTTTGATG |
| FBgn0014269          | TCGAGTCCAACCCGAAAAGGAT   | TCCTCGTCCAACCATATCGGTTGA |
| FBgn0020443          | CTCGAGTCCAACCCGAAAAGGAT   | TGCACTGCAATCAGAGAGAA |
| FBgn0024841          | GCTCGAGTCCAACGATGATTCGCTGTTG | TGCCTATCTGCTGATGTTG |
| FBgn0026188          | CTCGAGTCCAACCCGAAAAGGAT   | GGTCTCTGCGAGATCTGTC |
| FBgn0029629          | TCGAGTCCAACCCGAAAAGGAT   | TGCACTGCAATCAGAGAGAA |
| FBgn0030341          | CTCGAGTCCAACCGAAGGAA    | ATGCTGCACACCGAGGAG |
| FBgn0031769          | GCTCGAGTCCAACCCGAAAAGGAT | GGTCTCTGCGAGATCTGTC |
| FBgn0035121          | CTCGAGTCCAACCGAAGGAA    | ATGCTGCACACCGAGGAG |
| FBgn0037301          | CTCGAGTCCAACCGAAGGAA    | ATGCTGCACACCGAGGAG |
| FBgn0037707          | GCTCGAGTCCAACCGAAGGAA    | ATGCTGCACACCGAGGAG |
| FBgn0051729          | GCTCGAGTCCAACCGAAGGAA    | ATGCTGCACACCGAGGAG |

Primer pairs for validation of novel TSSs

| Gene name in FlyBase | Forward junction primer | Reverse GSP primer |
|----------------------|-------------------------|-------------------|
| FBgn0028537          | CTCGAGTCCAACATCATACGAGCA | CCGCGAATCCAAACGTGGT |
| FBgn0033113          | CTCGAGTCCAACATCATACGAGCA | CCGCGAATCCAAACGTGGT |
| FBgn0033068          | CTCGAGTCCAACATCATACGAGCA | CCGCGAATCCAAACGTGGT |
| FBgn0033688          | GCTCGAGTCCAACATCATACGAGCA | CCGCGAATCCAAACGTGGT |
| FBgn0260442          | CTCGAGTCCAACATCATACGAGCA | CCGCGAATCCAAACGTGGT |
| FBgn004228           | CTCGAGTCCAACATCATACGAGCA | CCGCGAATCCAAACGTGGT |
| FBgn0037410          | CTCGAGTCCAACATCATACGAGCA | CCGCGAATCCAAACGTGGT |
| FBgn0085320          | GCTCGAGTCCAACATCATACGAGCA | CCGCGAATCCAAACGTGGT |
| FBgn0030136          | CTCGAGTCCAACATCATACGAGCA | CCGCGAATCCAAACGTGGT |
| FBgn0024366          | CTCGAGTCCAACATCATACGAGCA | CCGCGAATCCAAACGTGGT |
Supplementary Results

Identification of novel transcription start sites

Paired-reads can also facilitate the direct link of novel TSSs to their respective genes. This becomes particularly important when TSSs are located distal from an annotated transcript. Previous technologies required the assumption that a potential novel TSS belongs to the nearest gene\(^1\), or had to rely on expression correlation between TSSs and the downstream gene to which they may belong\(^2\). Because the median distance between the 5’ and 3’ tags is 279 nt, identification of distal TSS by the PEAT method is not expected to be exhaustive. Of 342,943 read pairs where the 5’ read fell more than 250 nt upstream of an annotated TSS, 58,415 had the corresponding 3’ read mapped to the transcribed region of the downstream gene. From the clusters defined by these read-pairs, we selected 10 novel TSSs, meeting stringent criteria for experimental validation (Supplementary Table 3). Two independent approaches (oligo-capping and cap-trapping) were used; 7 and 8 out of 10 candidates were validated, respectively (Supplementary Fig. 7 and 8). Since the reverse primers were designed to be within 100-200 nt of the TSS, the actual positive rate could be even higher considering that some of the failed sites might be caused by a splicing event downstream of the TSSs. Despite the high validation rate, we did not expect a substantial number of novel first exons for known protein-coding genes, as had been proposed by a previous tiling array study\(^2\). Distal TSS tags were rare in general (0.5% of the total reads) and the majority of them did not form clusters. Indeed, comparison with the tiling array study only resulted in 57 clusters with \(\geq 10\) reads being identified near one of their novel exons. Only 8 of these 57 clusters contained more than 100 reads, one of which was included in the experimentally validated set. One possibility is that most of the distal TSSs found in tiling arrays may be the result of non-polyadenylated RNAs, which would not be represented in our dataset due to poly(A)+ selection.

TBP and TRF2 binding profiles distinguish different initiation patterns

Analyses of core promoter sequences in *Drosophila melanogaster* have identified a set of overrepresented motifs\(^3, 4\). These elements include the canonical position specific motifs

26
shared among metazoans, including the TATA box, which is found ~30 nt upstream of the TSS, the initiator element (INR), which is located directly at the point of transcription initiation, and motifs downstream of the TSS, including the downstream promoter element (DPE). TATA boxes are recognized by the TATA-box binding protein (TBP), a component of the basal factor TFIIID. Other enriched motifs are fly specific and include the DRE motif, which has less positional bias than the canonical elements, and is associated with core promoters bound by the TBP-related factor 2 (TRF2), which substitutes for TBP in remodeled basal complexes.

Previous studies have therefore demonstrated that alternative components of Pol II complexes in *Drosophila* have different binding preferences and functions. Given the differences we observe in initiation patterns and the varied motif occurrences between them, we further investigated whether distinct binding factors might be associated with the three initiation patterns defined here. As mentioned above, TBP and TRF2 are two transcription factors known to be components of different basal complexes in *Drosophila*. We assessed the binding of these factors using previously published ChIP-chip data obtained from *Drosophila* S2 embryonic cell lines, which provided sets of regions to which either TBP, TRF2, or a combination of both was bound. Due to the limited resolution of ChIP-chip assays, a subset of 1,138 NP, 660 BP, and 1,849 WP clusters separated by 500nt from any other cluster were evaluated.

Given the different experimental conditions, we did not expect perfect agreement between the ChIP-chip and our TSS data. Yet, we found that NP promoters had over 3-fold the amount of TBP and TBP&TRF2 binding to them than WP promoters, while WP promoters had greater levels of TRF2 binding than NP promoters. BP promoters showed TF binding patterns similar to both NP and WP promoter types, as they have an intermediate level of TBP binding, a high amount of TRF2 binding, and high levels of both TFs binding. This further supports that these promoters are a potential hybrid of both the NP and WP classes, which agrees with both the definition of the initiation pattern and their motif composition. Across all initiation patterns, the ChIP-chip data showed higher numbers of promoters to be bound by TRF2 than by TBP. In agreement with potential experiment-wide differences, higher occupancy of Pol II was observed for regions bound by
TRF2 than bound by TBP (Supplementary Fig. 16). As such, the associations between NP promoters with TBP binding and WP promoters with TRF2 binding were clearer when the relative frequency of each ChIP was compared separately among the clusters within each initiation pattern. NP promoters had the highest percentage of TBP binding (27%), followed by BP promoters (9%) and WP promoters (5%) (Supplementary Fig. 15d). Conversely, a higher frequency of TRF2 binding was observed in WP promoters (88%) followed by BP promoters (61%) and then NP promoters (38%). Differences in TBP binding also correlated with the frequency of a TATA box, which is the target sequence motif of TBP (Supplementary Table 6).

Taken together, NP promoters show an enrichment of TATA boxes, and the functional relevance of this is confirmed by the preferential binding of TBP, a subunit of the TFIID transcription initiation complex which is known to directly bind to the TATA box motif\textsuperscript{11}. In contrast, the WP promoters, which lack bias towards a specific location for initiation, were found to be enriched for the DRE motif, which is preferentially bound by TRF2-associated with DREF, subcomponents of an alternative initiation complex\textsuperscript{10}. The presence of WP promoters in \textit{Drosophila} is an interesting observation by itself as broad promoters in mammals are enriched for CpG islands\textsuperscript{1}, a genomic feature not present in the fly. However, mammalian studies have so far not uncovered that different basal complexes bind to peaked and broad promoters; our findings here suggest the possibility that distinct complexes, similar to TRF2, may be associated with CpG islands in mammals.

Finally, binding of TBP and, or TRF2 was strongly under-represented in internally capped clusters (Supplementary Fig. 15a-c), providing additional support to the motif analysis which had shown that the surrounding sequences were depleted of the known \textit{Drosophila} core promoter elements.
References

1. Carninci, P. et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**, 626-635 (2006).

2. Manak, J.R. et al. Biological function of unannotated transcription during the early development of Drosophila melanogaster. *Nat Genet* **38**, 1151-1158 (2006).

3. FitzGerald, P.C., Sturgill, D., Shyakhtenko, A., Oliver, B. & Vinson, C. Comparative genomics of Drosophila and human core promoters. *Genome Biol* **7**, R53 (2006).

4. Ohler, U., Liao, G.C., Niemann, H. & Rubin, G.M. Computational analysis of core promoters in the Drosophila genome. *Genome Biol* **3**, RESEARCH0087 (2002).

5. Rabenstein, M.D., Zhou, S., Lis, J.T. & Tjian, R. TATA box-binding protein (TBP)-related factor 2 (TRF2), a third member of the TBP family. *Proc Natl Acad Sci U S A* **96**, 4791-4796 (1999).

6. Butler, J.E. & Kadonaga, J.T. Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes Dev* **15**, 2515-2519 (2001).

7. Holmes, M.C. & Tjian, R. Promoter-selective properties of the TBP-related factor TRF1. *Science* **288**, 867-870 (2000).

8. Isogai, Y., Keles, S., Prestel, M., Hochheimer, A. & Tjian, R. Transcription of histone gene cluster by differential core-promoter factors. *Genes Dev* **21**, 2936-2949 (2007).

9. Lee, D.H. et al. Functional characterization of core promoter elements: the downstream core element is recognized by TAF1. *Mol Cell Biol* **25**, 9674-9686 (2005).

10. Hochheimer, A., Zhou, S., Zheng, S., Holmes, M.C. & Tjian, R. TRF2 associates with DREF and directs promoter-selective gene expression in Drosophila. *Nature* **420**, 439-445 (2002).

11. Dynlacht, B.D., Hoey, T. & Tjian, R. Isolation of coactivators associated with the TATA-binding protein that mediate transcriptional activation. *Cell* **66**, 563-576 (1991).