SUPPLEMENTAL RESULTS

CAGE peaks in 3' UTRs

The major class of CAGE peaks in annotated 3' UTRs of protein-coding genes is highly significant and requires further study. In mammals, cap-trapped and polyadenylated RNAs associated with CAGE peaks and overlapping 3' UTRs of many protein-coding transcripts have been described and verified by RACE (Carninci et al. 2006). Similarly, we observe that in the *Drosophila* embryo, 60% of protein-coding genes have a significant CAGE peak within the 3' UTR, and that number is 80% for genes with 5' UTR CAGE peaks. These peaks tend to be proportionally weaker than the 5' CAGE peaks of the same transcript. The strength of CAGE peaks correlates much more strongly with expression level (RPKM) measured by poly(A)+ RNA-seq in 3' UTRs than it does in 5' UTRs (log linear, $r \sim 0.73$ in 3'UTRs, vs. $r \sim 0.38$ for CAGE peaks in 5' UTRs), but this may be explained simply by the fact that 3' UTRs tend to be unspliced and more constitutively represented in mRNAs than 5' UTRs, and hence it is easier to correctly associate each CAGE peak with the appropriate exon for the purposes of computing expression level. Our statistically based alignment analysis shows that these 3' CAGE peaks are not an alignment artifact due to the lower sequence complexity in 3' UTRs.

We identified cap-trapped 5' ESTs from the RE cDNA library for 14 CAGE peaks within 3' UTRs, a sample of convenience and a small minority of the peaks in this class. Full-insert sequencing of the corresponding cDNAs defined long polyadenylated transcripts, several of which are polyadenylated at a site downstream of the annotated polyadenylation site of the parental mRNA transcript. Note that the RE ESTs were selected against short transcripts by the cDNA library construction process, so the unannotated polyadenylation sites discovered in these cDNA sequences are biased toward longer, downstream sites. Thus, 3' UTR CAGE peaks are associated with long transcripts or transcript fragments that are co-linear with 3' UTRs and can have sufficient length and stability to be represented in conventional cDNA libraries. The origin of this class of CAGE peaks and transcripts is a subject for future study. They may represent transcript fragments resulting from cleavage by microRNAs. Since there is no evidence of these transcripts in nuclear-fraction short RNA (see (Nechaev et al. 2010) and Results), these fragments might be re-capped by a recently described cytoplasmic capping complex (Otsuka et al. 2009).

SUPPLEMENTAL DISCUSSION

Statistical approaches improve analysis of CAGE data

The analysis of CAGE tags presents two primary challenges. First, the tags are short (27 nt). To map these tags to the genome, we used a statistical program called *StatMap* (Supplemental Methods). *StatMap* is an exhaustive approach to mapping that enables the explicit modeling of CAGE as a sampling process, and in addition to mapped locations for each tag, it returns an estimate of the underlying probability at each basepair that, if one more tag were sequenced, its 5' end would map to that location. *StatMap* can then

use these probabilities to refine the mappings for tags that align to multiple genomic locations, which are in turn used to iteratively refine the estimated probabilities. That is, *StatMap* models dependence between genomic locations induced by sequence similarity. These features allowed us to infer that 16M CAGE tags map unambiguously, in that they map to a location to which they are at least 100-fold more likely to map than to any other location in the genome. This is a more useful and stronger condition than unique and exact mappings, which are often reported in the literature for short reads.

The second challenge, as has been previously reported (van Bakel et al. 2010), is that CAGE tags identify a diverse population of RNA elements, including the 5' ends of capped transcripts, some uncapped transcripts including rRNAs which are very abundant, re-capped transcript fragments, and uncapped transcripts resulting from less-than-100% efficiency in the cap-trapping protocol. While nearly the entire population of elements detected by the CAGE assay may be of interest, in this study we were primarily interested in CAGE tags that mapped to the 5' ends of transcripts, and which therefore identify TSSs within promoter regions. To enrich for this population of CAGE tags, we used a statistical approach. We modeled the distribution of mapped CAGE tags across the genome as linearly proportional to the distribution of mapped RNA-seq tags from a developmental time course of embryo total RNA samples(Graveley et al. 2010), and masked the CAGE tags wherever this linear model could account for the local density. This two-step analysis removed 18% of mapped CAGE tags and enriched for tags in promoter elements. However, many CAGE tags still mapped to mitochondrial transcripts, coding exons and 3' UTRs. This was not entirely surprising. Re-capping sites, for instance, will also be enriched by our analysis if the re-capping event is frequent and occurs at a consistent position in the transcript.

RACE primer design caveats

RACE primer design depends on prior knowledge of transcript structure. In a small minority (ca. 50) of RACE experiments, primer design appears to have been mis-targeted. In one set of cases, the inner gene-specific primer sequence overlaps or maps upstream of a large CAGE peak, so that the RACE experiment artifactually amplifies the 5' portion of the true TSS distribution. In another set of cases, the inner primer appears to be too distant (> 500 bp) from a large embryonic CAGE peak to amplify efficiently and sequence RACE reads corresponding to it. Improved transcriptome annotation from RNA-seq analysis and correlation with unsupported CAGE tags will facilitate more accurate targeting of RACE experiments in future work.

SUPLEMENTAL TABLES

Supplemental Table 1. Summary of RE EST, CAGE and RACE datasets.

	Average Length (bp)	Number of Sequences	Number of aligned sequences	Number of spliced sequences
5' RE EST	453	66,169	61,437	47,286
CAGE	27	41,804,261	20,191,962	NA
RACE	149	1,775,191	1,501,415	409,605

Supplemental Table 2. Summary of core promoter motif enrichment in promoters.

Common name	Ohler name	Fitzgerald name	Average frequency in peak promoters	Average frequency in broad promoters	enrichment type	enrichment class	enrichment position	pvalue*
PauseButton			0.87	0.63	positional	P	+23	8.01E-033
TATA	ohler3	DMp1	0.18	0.19	positional	P	-32	1.26E-011
INR	ohler4	DMp2	2.95	2.59	positional	P	-1	5.07E-027
INR		DMp3	14.33	14.21	positional	P	-1	7.53E-005
DPE	ohler9	DMp4	0.19	0.14	positional	P	+25	1.29E-011
DPE1		DMp5	0.69	0.59	positional	P	+26	3.61E-013
		DMv1	0.64	0.65	overrepresentation	B	upstream	0
	ohler8	DMv2	10.61	10.69	-	-	-	-
	ohler7	DMv3	2.22	2.36	overrepresentation	B	upstream	4.06E-08
	ohler1	DMv4	0.96	0.97	-	-	-	-
	ohler6	DMv5	0.11	0.18	overrepresentation	B	upstream	6.27E-011
GAGA		NDM1	0.49	0.29	overrepresentation	P	upstream	6.32E-012
		NDM2	8.87	7.11	-	-	-	-
		NDM3	0.38	0.35	-	-	-	-
DRE	ohler2	NDM4	15.02	16.15	-	-	-	-
Ebox	ohler5	NDM5	0.19	0.22	overrepresentation	B	upstream	7.45E-004

*****p-values are from Wilcoxon rank sum test between peaked and broad promoter classes with Bonferonni correction for multiple testing.

SUPPLEMENTAL METHODS

CAGE data production

First-strand cDNA was synthesized from 50 μg total RNA using 10 μg of primer RT-N15-EcoP (5'- AAGGTCTATCAGCAGNNNNNNNNNNNNNNC -3') and 2 μg Oligo-dT16-VN (5'- AAGGTCTATCAGCAGTTTTTTTTTTTTTTTVN -3') using M-MLV Reverse Transcriptase, RNase H Minus, Point Mutant (Promega) in the presence of trehalose-sorbitol (0.14M-0.7M) as previously described(Carninci et al. 1998). The RT reaction was purified by GFX-CTAB purification(Salimullah et al. 2009). Subsequently, a biotin molecule was added to the CAP site by chemical biotinylation, treated with RNase I, and subsequently mRNA/cDNA complexes were isolated such that the cDNAs reached the cap-site, by capturing with streptavidin sepharose beads (GE Healthcare)(Carninci et al. 1996). The beads were washed 10 times with Binding & Wash buffer (0.5 M NaCl and 50 mM EDTA), once with B&W buffer containing 0.05% SDS, and twice with B&W buffer containing 0.1% Tween20.

First strand cDNA was recovered from the beads by denaturation with 50 mM NaOH, followed by neutralization by addition of Tris pH 7.0 at 100 mM final concentration. The cDNA was cleaned up and the volume was reduced using a Microcon YM 100 filter (Millipore). The specific linker 5'SOL41#1 (10 ng), containing ID tags ACA and a recognition site for the class II restriction enzyme EcoP15I (upper oligonucleotide: 5'- Biotin CCACCGACAGGTTCAGAGTTCTACAGACACAGCAGNNNNNN -3'; lower oligonucleotide: 5'- Phosphate CTGCTGTGTCTGTAGAACTCTGAACCTGTCGGTGGNH2 -3') was ligated to the 3' end of the first-strand cDNA with a DNA ligation kit version 2.1 (Takara Bio Inc).

Second-strand cDNA was synthesized using 10 ng of the primer 2nd SOL (5'- Biotin CCACCGACAGGTTCAGAGTTCTACAG -3') and 5 units of La Taq (TaKaRa). The resulting double-stranded cDNA was cleaved with 0.1 units of EcoP15I (NEB) in the presence of 100 μM sinefungin (Sigma) and incubated at 37° for 3 hours. The enzyme was inactivated by addition of $MgCl_2$ to 10 mM and incubation at 65° for 20 minutes.

Next, 10 ng of a second linker 3'SOLX (upper oligonucleotide: 5'- Phosphate NNTCGTATGCCGTCTTCTGCTTG -3'; lower oligonucleotide: 5'- CAAGCAGAAGACGGCATACGA -3') was ligated to the restriction digestion products. Ligation was performed overnight at 16° C with T4 DNA ligase (NEB). Ligation products (with biotin at the 5' ends) were separated from linker dimers and restriction fragments from the 3'-end of the cDNA using streptavidin sepharose beads (GE Healthcare). After binding of biotinylated DNA, beads were washed 10 times with B&W buffer (0.5 M NaCl and 50 mM EDTA), once with B&W buffer containing 0.05% SDS, and twice with B&W buffer containing 0.1% Tween20. The washed beads were resuspended in 20 μl of H_2O.

DNA fragments were amplified by PCR with two linker-specific primers, SOLX41F_34 (5'- AATGATACGGCGACCACCGACAGGTTCAGAGTTC -3') and SOLX_R (5'- CAAGCAGAAGACGGCATACGA -3'). PCR was performed in 1xHF buffer

(Finnzymes; 50 mM $MgCl_2$, 33% DMSO), 0.2 mM dNTPs, 0.5 µM of each primer, 1 µl of DNA template and 1 unit of Phusion polymerase (Finnzymes) in a total volume of 50 µl. The template was denatured at 98° for 30 sec and amplified by 19 cycles of PCR (10 sec at 98°, 10 sec at 60°). The resulting PCR products from 15 separate reactions were pooled, digested with proteinase K, purified by ethanol precipitation, and resuspended in 12.5 µl of TE.

The PCR products were separated on a 12% PAGE gel, and the 96-bp band was excised from the gel, crushed, and incubated with 1 ml of elution buffer (0.5 M ammonium acetate, 10 mM magnesium acetate, 1 mM EDTA, pH 8.0, 0.1% SDS) overnight at room temperature. The material was filtered by centrifugation in MicroSpin Empty Columns (GE Healthcare). Finally, the sample was extracted with phenol-chloroform, ethanol precipitated, and resuspended in 15 µl of H_2O. The DNA concentration was measured using an Agilent 2100 Bioanalyzer and adjusted to a final concentration of 10 nM for sequencing operations. A 3.0 pM aliquote of the CAGE library was sequenced on the Illumina GAI platform using standard protocols and the sequence primer 5'-CGGCGACCACCGACAGGTTCAGAGTTCTACAG -3'. Multiple lanes of data were collected to produce a total of 42 million reads.

The *StatMap* algorithm and CAGE data analysis

Statmap is a statistical framework for mapping short reads back to a (potentially uncertain) reference genome. It is a tool to find every genomic location from which each read may have come, given user-defined thresholds. Furthermore, given a model of the process by which genomic DNA was selected for sequencing, *Statmap* models the joint distribution of reads and genomic loci, resulting in a probability distribution on the space of mappings consistent with the observed data. In order to do this, *Statmap* takes into account three key classes of uncertainty:

1) Mapping uncertainty: genomes often contain multiple instances of identical or nearly identical sequences of length on the scale of read length. Hence, the genomic location from whence a particular read originated often cannot be ascertained from the sequence of the read alone.

2) Sequencing uncertainty: reads may contain many sequencing errors in the form of incorrectly called nucleotides. We assume, out of computational necessity, that these errors are independent. We do not yet model insertions or deletions. We assume that the mutation probabilities encoded by the sequencing platform of interest in the FASTQ file constitute accurate estimates of these probabilities.

3) Genome uncertainty: *Statmap* maps back to a distribution at each base-pair of the reference genome. As in 2), above, we assume independence at each position. This assumption is presently an important computational convenience. We note that GPU parallelization may ameliorate this problem and permit more complex error distributions.

We model each read as a noisy sample from a population of DNA sequences. Each sequence has an associated genomic location, typically a chromosome-index pair. To

allow for genomic repeats, we do not insist that each sequence in the population is unique. Furthermore, to allow for alternate genomes, we do not insist that the set of genomic locations is unique. However, we do require that the sequence-location pairs are unique, but this is true by definition for all extant reference genomes.

Formally, consider the sequence-location pair (s,g). Neither s nor g is observed. Our data are a set of reads R. Each r in R came from some underlying sequence s, where s was the sequence actually fed into the sequencer. We estimate the probability that s was fed into the sequencer, given r was the observed read, which we hereafter write Pr[r|s], from the FASTQ file error model. We estimate Pr[r|s] under positional independence as the product of the probabilities of point mutations taking the given sequence s into the observed read r.

We will use the following notation:

Pr[r|s] – probability of observing a read, r, given that it came from a sequence s
This quantifies our uncertainty about the sequencing technology. Given s, we can estimate this from the FASTQ provided error model.

Pr[g] – probability that a randomly selected read was generated by the sequence starting at g. That is, the population portion of reads originating at g.
As such, the sum over all g must be 1. Typically, the goal of the assay is to uncover this quantity.

$q_\theta[s \mid g]$ – estimate of the probability that the sequence s generating a read comes from g
This is our model. For instance, in an RNA-seq experiment, if g is in a single-isoform exon, the sequence s should be entirely contained in the exon, the value of $q_\theta[s \mid g]$ should be high for all such s, and zero otherwise. Note that the sum over all s of $q_\theta[s \mid g]$ is 1 for each g.

To motivate the following, assume for now that there is no sequencing error. That is, we assume Pr[r|s] to be one if s is identical to r, and zero otherwise. Furthermore, assume that the reference genome is known; mathematically, Pr[s|g] is 1 if and only if s is identical to the sequence at g. Then we can express the probability of observing a read in the form of an integral equation:

EQ 1) $$Pr[s] = \sum_g q_\theta[s \mid g] Pr[g]$$

As described in (Vardi and Lee 1993), this can be solved via an iterative method. Write, $\pi_j = \hat{P}r[g_j]$ and $p_k = \hat{P}r[r_k]$, the estimated probabilities of g_j and $r_k = s_k$. That is, we initialize π_j to be uniform and update our estimates until convergence via:

EQ 2) $$\pi_j^{NEW} = \pi_j^{OLD} \sum_k \frac{q_\theta[s_k \mid g_j]\hat{p}_k}{\sum_l q_\theta[s_k \mid g_l]\pi_l^{OLD}}$$

When sequencing error and genomic uncertainty are present, let

$f_\theta[r_k \mid g_j] = \sum_m q_\theta[s_m \mid g_j] Pr[r_k \mid s_m]$. Then, with *f* replacing *q*, equation 2 becomes:

EQ 3) $$\pi_j^{NEW} = \pi_j^{OLD} \sum_k \frac{f_\theta[s_k \mid g_j]\hat{p}_k}{\sum_l f_\theta[s_k \mid g_l]\pi_l^{OLD}}$$

Statmap in principle computes Pr[*r*|*s*] for each *s* of the same length as *r*. In practice, the space of all sequences is too large to be computationally tractable, and we are generally interested only in the very small subset of sequences that occur in the reference genome, or are near in the sense of hamming distance to those that occur. To account for this, *Statmap* has two essential operational parameters, and they are thresholds that govern the algorithms exploration of sequence space. They are as follows:

1) MIN_BND = *x*: minimum bound on Pr[*r*|*s*]: compute Pr[*r*|*s*] only when greater than *x*

2) MIN_RAT = *y*: minimum bound on ratio $\frac{Pr[r \mid s]}{max_k(Pr[r \mid s_k])}$: compute Pr[*s*|*r*] only when its relative magnitude compared to the s_k that maximizes the likelihood of *r* is greater than *y*.

Additional parameters may enter through the kernel $q_\theta[s \mid g]$. The number and type of parameters depends on the assay and the particular problem or question at hand. This kernel is implemented by the user, via the Statmap Python API, which is described in detail in the code available from the ENCODE consortium at encodestatistics.org.

The formulation used for CAGE is as follows. We mapped the 42M reads using MIN_BND = 1E-10 and MIN_RAT = 1E-2. We specified $q_\theta[s \mid g]$ to be a 100 bp Dirichlet smoothing window (http://en.wikipedia.org/wiki/Dirichlet). We modeled potential untemplated dG residues on the 5' ends of CAGE reads as a Bernoulli process. To initialize our iterative mapping procedure, we estimated the Bernoulli parameter as uniform on mononucleotides. We conducted 200 iterative updates until the mean change in the updated value across the genome was less than 1e-9. Questions about this analysis should be directed to N.B. (npboley@gmail.com) and J.B.B. (ben@me.berkeley.edu).

CAGE tag filtering

We modeled the CAGE assay background as a mixture of signal linearly proportional to expression level and uniform, unstranded signal that we treat as random noise. To classify the tags, we first estimated the mixture parameter. Then, we used data from stranded RNA-seq of total RNA from a timecourse of *D. melanogaster* embryonic development (Graveley et al. 2010) to calculate the probability that, under our model, each tag was drawn from the background distribution. If the probability of a read having come from the background distribution was less than 1e-3, we labeled it as signal. This was done as follows: we modeled each mapped tag as having been drawn from a mixture of two multinomial distributions. The bin frequencies of the first, the signal distribution, are the parameters of interest. The bin frequencies of the second, the noise, are assumed to be the same as those from which the mapped RNA-seq tags was drawn. To fit the mixture parameter and the first multinomial bin frequencies, we initialized the mixture parameter, λ, to the value which minimized the sum over all basepair positions of $|N\lambda p_i \quad C_i|$ where N is the total number of reads, p_i is the empirical RNA-seq bin probability at position i, and C_i is the expectation of the number of CAGE reads at position i. Then, for each position, we estimated the marginal probability that the reads at that position were drawn from the true distribution. Formally, we assumed that each position i had a bin frequency p_i , with a $Beta(1+\lambda C_i, 1+(1-\lambda)C_i)$ distribution. Then, we estimated the probability as:

$$\int_0^1 Beta(p_i, 1+\lambda C_i, 1+(1-\lambda)C_i)Bin(C_i, p_i, N)dp_i .$$

If this estimate fell below 1e-3, we labeled the reads as signal. After having done this for each position, we re-estimated the mixture parameter by N_n/N. Then, we re-classified each read with the new mixture parameter. Further iterations of these estimates yielded the same classifications, and the procedure was insensitive to our initial estimate of λ.

CAGE tag clustering

An iterative hierarchical clustering approach using the complete divisive method was applied to arrive at clusters with a maximum span of 300 bp. To remove outliers, small "singleton" clusters, clusters of at most five tags where the 5' ends of all member tags map to a single bp, were filtered whenever no other tags mapped within 100 bp. Lastly, clusters less than 50 bp apart were merged. Changing this threshold from 50 to 200 bp resulted in changes to less than 1% of clusters, whereas a threshold of fewer than 50 bp failed to merge clusters associated with the same transcript (determined by overlap with mapped, associated RACE reads or RE ESTs).

RACE data production

Total RNA (1.25 µg) was treated with calf intestinal alkaline phosphotase (CIP; NEB) to remove the 5' phosphate from uncapped transripts. The reaction was carried out at 37° for one hour and was followed by phenol-chloroform extraction and ethanol percipitation.

Dephosphorolated RNA was treated with tobacco acid pyrophosphatase (TAP; NEB) to remove the 5' cap structure. The reaction was carried out at 37° for one hour and was followed by phenol-chloroform extraction and ethanol percipitation. A 5' RNA adapter (Ambion) was ligated to the TAP-treated RNA using T4 RNA ligase (NEB). The reaction was carried out at 37° for two hours and was followed by phenol-chloroform extraction and ethanol percipitation. Reverse transcription was perfomed on the oligo-ligated RNA using M-MLV reverse transcriptase (NEB). The reaction was incubated at 25° for 30 minutes and then at 42° for one hour. The resulting first-strand cDNA was treated with RNAse H (Invitrogen) at 37° for 20 minutes. The final volume of the cDNA reaction was 100 µl, sufficient for 100 RACE experiments.

Nested RACE PCR reactions were carried out on the first-strand cDNA (1 µl) using HotFirePol DNA polymerase (Solis Biodyne). The first round of PCR was performed using a common primer within the 5' adapter sequence (Ambion) and a gene-specific primer approximately 250 bases downstream of the annotated 5' end. PCR conditions were as follows: 95°, 15 min.; 20 cycles of 95°, 30 sec., 56°, 30 sec., 72°, 1 min,. followed by 72°, 15 min. The second round of PCR was performed using a nested common primer (Ambion) and a nested gene-specific primer approximately 150 bases downstream of the annoated 5' end. Second-round PCR conditions were as follows: 95°, 15 min.; 20 cycles of 95°, 30 sec., 60°, 30 sec., 72°, 1 min,. followed by 72°, 15 min. For 1453 experiments (see below), the number of cycles ine the seconnd-round PCR reaction was increased from 20 to 25 to produce sufficient product.

Gel electrophoresis was preformed on a 5 µl aliquot of inner PCR product to assess yield and product length(s). Gels were stained with ethidium bromide, and images were captured using a BioRad GelDoc system and analyzed using BioRad Quantity One software. In cases where no PCR band was visible, the second-round PCR reaction was repeated using 25 cycles. Individual PCR products were pooled to create a molar-normalized mixtures of 1,440 to 2,677 poducts using a Tecan Genesis 2 robot. The sample pools were then concentrated to approximatley 50 ng/µl using Millipore Amicon Ultra filters.

The concentrated pools were processed through the Roche 454 Genome Sequencer FLX library construction protocol, including size fractionation by PAGE to remove library adapter dimers.

RACE pools 1 to 4 were sequenced on Roche 454 FLX instruments at the Washington Univeristy Genome Sequencing Center. RACE pool 5 was sequences on a Roche 454 Titanium instrument at the QB3 sequencing center at the University of California, Berkeley.

RACE pool 1 consisted of 1440 experiments, pool 2 consisted of 1920 experiments, pool 3 consisted of 2677 experiments and pool 4 consisted of 2446 experiments. RACE pool 5 consisted of 2787 experiments including 2361 rework experiments and 426 new experiments. Of the pool 5 rework experiments, 1453 had no PCR bands in previous attempts and were amplified using 25 cycles in the second-round PCR step. An additional

908 experiments had PCR bands but did not return sequences in previous pools.

RACE data analysis

RACE reads were processed to trim low-quality regions and to identify and trim the RLM-RACE adapter sequence. Only sequences with a significant match to the 3' end of the 3'-most 38 bases of the adapter sequence were retained. The adapter-trimmed sequence was quality trimmed by searching for the longest sequence that had an estimated probability of error of 10% or less; at least 35 high-quality bases were retained in a database, compared to the Release 5 genome sequence using BLASTN (Altschul et al. 1997), then aligned using *sim4* (Florea et al. 1998) to a 200 kb region of the genome centered on the best scoring BLAST high-scoring segment pair. Only sequences that were aligned to the genome beginning at the most 5' nucleotide were used in subsequent analyses.

Based on these alignments, reads were associated with a targeted transcript from the corresponding pool of RACE products. Each such read was presumed to represent a transcription initiation event. Because the protocol relies on PCR amplification, each read does not necessarily represent an independent event. A read was associated with a particular transcript and RACE reaction by searching for interrogated transcripts on the same strand as the aligned sequence and within 5 kb of the 3' end of the aligned sequence. If there was only one transcript within this region that was a target in the RACE pool, it was assigned to the read. If there were multiple targeted transcripts within the region, then the sequence was assigned to the transcript and RACE reaction with the highest scoring BLAST hit to the inner transcript-specific PCR primer if applicable, or to the highest scoring BLAST hit to a targeted transcript.

Associated RACE reads were clustered to define promoter regions. The unique locations of all reads associated with a given target were collected, and the Euclidian distances between all pairs of distinct start locations were calculated. An iterative hierarchical clustering approach using the "complete" agglomeration method was applied to find the minimum number of clusters for which the number of TSS in a cluster is greater than three and the cluster span is less than 300 bp. To remove outliers, tags that mapped farther away than 1.5 times the inter-quartile range were excluded from a promoter region.

Promoter classification

To determine whether sequencing depth was sufficient to classify a promoter as peaked or broad, we sampled tags and computed the fraction of sites we could rediscover within a promoter. At each stage k in the sampling, we limited the sampling to clusters that had greater than k reads. We estimated that peaked promoters require approximately 20 products per cluster to rediscover 80% of the unique tags in that cluster, while broad promoters require approximately 100 tags to rediscover 80% of the unique TSS in that cluster (Supplementary Fig. 5).

To assess confidence, we ensured that the classification would remain the same under a

bootstrapped sample of the tag distribution and if we added a single tag in the worst possible location. That is, for each broad promoter, we took *N* bootstrap samples from the tag distribution and, for each sample, added a single tag at the location with the highest tag count. For a peaked promoter, we bootstrapped from the tags and added a single tag to a uniform, random position within the promoter region. If less than 99% of the bootstrap samples changed classification, then we classified the region as 'unknown' (U). Under this model, 1,484 (83%) promoters retained their 'peaked' classification. Similarly, 6005 (87%) broad promoters retained their broad classification. This is consistent with the analysis above: greater tag depths are necessary to confidently classify broad promoters.

Motif analysis

The match score, or probability of observing a motif along a given sequence, was calculated as the sum of nucleotide frequencies given by the PSSM, and has been shown to be correlated to the equilibrium occupancy of a TF to a given sequence in a simple thermodynamic model(Schneider et al. 1986; Berg and von Hippel 1987). Aggregate plots of motif occurrence were made by aligning promoter sequences at the tallest TSS peak within each promoter and averaging motif occurrences across all positions within the aligned sequences.

Defining promoters

We modeled the joint distributions of the mapped tag counts at each bp as follows:

$$CAGE_{i+\alpha_{CAGE}} \sim Bin(n_{CAGE}, p_i)$$
$$RACE_{i+\alpha_{RACE}} \sim Bin(n_{RACE}, p_i)$$
$$EST_{i+\alpha_{EST}} \sim Bin(n_{EST}, p_i)$$

where the distributions of tag counts are uniformly "shifted", or "offset" versions of some underlying "summary" or "consensus" multinomial. We estimated the optimal shifts non-parametrically under cross correlation at each shift from -5 to +5 bp. We also studied cross correlation at each promoter for all shifts from -50 to +50 bp, but the extended width of the analysis improved the maximal correlation of the pairs of assays in only 10% of cases, and these cases corresponded to marginal improvements. We performed kernel density smoothing under a uniform kernel of width determined by the above estimate of the optimal shift on each of the three tag distributions from the three assays. We combined the tag distributions into a single, consensus PDF by computing the variance normalized estimate of p_i, above, at each position in the promoter. Variance normalization here prevents one assay with much greater tag counts (usually CAGE)

from washing out the other two.

SUPPLEMENTAL FIGURE LEGENDS

Supplemental Figure 1. Peaked and broad TSS distributions defined by RE ESTs. The 5'-end alignments of the indicated ESTs are shown. In parentheses are indicated the number of redundant, near-identical EST reads (same number of splice-sites, exon boundaries are within 5% of the figure range). (a) RE ESTs associated with the peaked promoter of *dhd*. (b) RE ESTs associated with the broad promoter of *CG4769*.

Supplemental Figure 2. CAGE peaks map to the mitochondrial genome and the rDNA repeat. (a) The entire *D. melanogaster* mitochondrial genome is shown with CAGE peaks in the top track and mitochondrial genes in the bottom track. (b) The alignment of 1.4M CAGE reads to the *D. melanogaster* rDNA tandem repeat sequence (Tautz et al. 1988) (GenBank acc. no. M21017) is shown. Peaks map to each of the processed rRNAs: 18S, 5.8S, 2S, and 28S. The locations of the mature rRNAs within the rDNA repeat are indicated by the green bars. The interdigitated black bars correspond to the internal transcribed spacers (ITS). The peak at zero on the x-axis corresponds to the Pol I transcription start site (arrow) for the long primary transcript. A small fraction (5%) of reads map to the antisense strand, which is shown in red, with the y-axis scaled differently for visibility.

Supplemental Figure 3. Promoter classification. (a) The relationship between width of promoter clusters and shape index is plotted. Promoter widths (histogram projection of x-axis) are bimodally distributed with a trough at approximately 20 nt. The shape index (histogram projection of y-axis) is continuously distributed. Promoters with SI > -1 (above the dotted line) were classified as peaked; promoters with SI <= -1 were classified as broad. (b) The average fraction of distinct start sites rediscovered by sampling tags within peaked and broad promoters, under a multinomial model of TSS distribution. Approximately 20 and 100 tags are necessary to discover 80% of the TSS locations within peaked and broad promoters, respectively.

Supplemental Figure 4. Embryo RACE versus Adult RACE. The relative offsets of TSS locations in embryo RACE and TSS locations in adult RACE is indicated.

Supplemental Figure 5. Positional histograms of 16 core promoter motifs mapped on both strands. Peaked promoters are positionally enriched for the pause button, TATA, INR (DMp2 and DMp3) and DPE (DMp4 and DMp5) motifs. Broad promoters are generally enriched for the DMv1, DMv3, Dmv5, NDm1 and NDM5 motifs.

Supplemental Figure 6. Analysis of CAGE peaks in 3' UTRs. Forty percent of CAGE peaks overlap 3' UTRs. These peaks are positionally enriched for the PB motif at -10 nt relative to the primary TSS location in the CAGE peak.

SUPPLEMENTAL REFERENCES

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Berg OG, von Hippel PH. 1987. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol* **193:** 723-750.

Carninci P, Kvam C, Kitamura A, Ohsumi T, Okazaki Y, Itoh M, Kamiya M, Shibata K, Sasaki N, Izawa M et al. 1996. High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* **37:** 327-336.

Carninci P, Nishiyama Y, Westover A, Itoh M, Nagaoka S, Sasaki N, Okazaki Y, Muramatsu M, Hayashizaki Y. 1998. Thermostabilization and thermoactivation of thermolabile enzymes by trehalose and its application for the synthesis of full length cDNA. *Proc Natl Acad Sci U S A* **95:** 520-524.

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

Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* **8:** 967-974.

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin J, Yang L, Artieri C, van Baren MJ, Booth BW, Brown JB et al. 2010. The developmental transcriptome of *Drosophila melanogaster*. *Nature* **submitted**.

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.

Otsuka Y, Kedersha NL, Schoenberg DR. 2009. Identification of a cytoplasmic complex that adds a cap onto 5'-monophosphate RNA. *Mol Cell Biol* **29:** 2155-2167.

Salimullah M, Kato S, Murata M, Kawazu C, Plessy C, Carninci P. 2009. Tunable fractionation of nucleic acids. *Biotechniques* **47:** 1041-1043.

Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. 1986. Information content of binding sites on nucleotide sequences. *J Mol Biol* **188:** 415-431.

Tautz D, Hancock JM, Webb DA, Tautz C, Dover GA. 1988. Complete sequences of the rRNA genes of *Drosophila melanogaster*. *Mol Biol Evol* **5:** 366-376.

van Bakel H, Nislow C, Blencowe BJ, Hughes TR. 2010. Most "dark matter" transcripts are associated with known genes. *PLoS Biol* **8:** e1000371.

Vardi Y, Lee D. 1993. From Image Deblurring to Optimal Investment Portfolios: Maximum Likelihood Solutions for Positive Linear Problems. *JRSS, B* **55:** 569-612.