ARTICLE

doi:10.1038/nature09715

The developmental transcriptome of *Drosophila melanogaster*

Brenton R. Graveley[1]*, Angela N. Brooks[2]*, Joseph W. Carlson[3]*, Michael O. Duff[1]*, Jane M. Landolin[3]*, Li Yang[1]*, Carlo G. Artieri[4], Marijke J. van Baren[5], Nathan Boley[6], Benjamin W. Booth[3], James B. Brown[6], Lucy Cherbas[7], Carrie A. Davis[8], Alex Dobin[8], Renhua Li[4], Wei Lin[8], John H. Malone[4], Nicolas R. Mattiuzzo[4], David Miller[9], David Sturgill[4], Brian B. Tuch[10,11], Chris Zaleski[8], Dayu Zhang[7], Marco Blanchette[12,13], Sandrine Dudoit[14], Brian Eads[9], Richard E. Green[15], Ann Hammonds[3], Lichun Jiang[4], Phil Kapranov[8], Laura Langton[5], Norbert Perrimon[16], Jeremy E. Sandler[3], Kenneth H. Wan[3], Aarron Willingham[17], Yu Zhang[4], Yi Zou[7], Justen Andrews[9], Peter J. Bickel[6], Steven E. Brenner[2,17], Michael R. Brent[5], Peter Cherbas[7,9], Thomas R. Gingeras[8,18], Roger A. Hoskins[3], Thomas C. Kaufman[9], Brian Oliver[4] & Susan E. Celniker[3]

Drosophila melanogaster **is one of the most well studied genetic model organisms; nonetheless, its genome still contains unannotated coding and non-coding genes, transcripts, exons and RNA editing sites. Full discovery and annotation are pre-requisites for understanding how the regulation of transcription, splicing and RNA editing directs the development of this complex organism. Here we used RNA-Seq, tiling microarrays and cDNA sequencing to explore the transcriptome in 30 distinct developmental stages. We identified 111,195 new elements, including thousands of genes, coding and non-coding transcripts, exons, splicing and editing events, and inferred protein isoforms that previously eluded discovery using established experimental, prediction and conservation-based approaches. These data substantially expand the number of known transcribed elements in the *Drosophila* genome and provide a high-resolution view of transcriptome dynamics throughout development.**

Drosophila melanogaster is an important non-mammalian model system that has had a critical role in basic biological discoveries, such as identifying chromosomes as the carriers of genetic information[1] and uncovering the role of genes in development[2,3]. Because it shares a substantial genic content with humans[4], *Drosophila* is increasingly used as a translational model for human development, homeostasis and disease[5].

High-quality maps are needed for all functional genomic elements. Previous studies demonstrated that a rich collection of genes is deployed during the life cycle of the fly[6–8]. Although expression profiling using microarrays has revealed the expression of ~13,000 annotated genes, it is difficult to map splice junctions and individual base modifications generated by RNA editing[9] using such approaches. Single-base resolution is essential to define precisely the elements that comprise the *Drosophila* transcriptome.

Estimates of the number of transcript isoforms are less accurate than estimates of the number of genes. Whereas ~20% of *Drosophila* genes are annotated as encoding alternatively spliced pre-mRNAs, splice-junction microarray experiments indicate that this number is at least 40% (ref. 7). Determining the diversity of mRNAs generated by alternative promoters, alternative splicing and RNA editing will substantially increase the inferred protein repertoire. Non-coding RNA genes (ncRNAs) including short interfering RNAs (siRNAs) and microRNAS (miRNAs) (reviewed in ref. 10), and longer ncRNAs such as *bxd* (ref. 11) and *rox* (ref. 12), have important roles in gene regulation, whereas others such as small nucleolar RNAs (snoRNAs) and small nuclear RNAs (snRNAs) are important components of macromolecular machines such as the ribosome and spliceosome. The transcription and processing of these ncRNAs must also be fully documented and mapped.

As part of the modENCODE project to annotate the functional elements of the *D. melanogaster* and *Caenorhabditis elegans* genomes[13–15], we used RNA-Seq and tiling microarrays to sample the *Drosophila* transcriptome at unprecedented depth throughout development from early embryo to ageing male and female adults. We report on a high-resolution view of the discovery, structure and dynamic expression of the *D. melanogaster* transcriptome.

Strategy for characterization of the transcriptome

To discover new transcribed features (Supplementary Table 1) and comprehensively characterize their expression dynamics throughout development, we conducted complementary tiling microarray and RNA-Seq experiments using RNA isolated from 30 whole-animal samples representing 27 distinct stages of development (Supplementary Table 2). These included 12 embryonic samples collected at 2-h intervals for 24 h, six larval, six pupal and three sexed adult stages at 1, 5

[1]Department of Genetics and Developmental Biology, University of Connecticut Health Center, 263 Farmington Avenue, Farmington, Connecticut 06030-6403, USA. [2]Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA. [3]Department of Genome Dynamics, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [4]Section of Developmental Genomics, Laboratory of Cellular and Developmental Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA. [5]Center for Genome Sciences and Department of Computer Science, Washington University, St Louis, Missouri 63108, USA. [6]Department of Statistics, University of California, Berkeley, California, 94720 USA. [7]Center for Genomics and Bioinformatics, Indiana University, 1001 E. 3rd Street, Bloomington, Indiana 47405-7005, USA. [8]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA. [9]Department of Biology, Indiana University, 1001 E. 3rd Street, Bloomington, Indiana 47405-7005, USA. [10]Genetic Systems Division, Research and Development, Life Technologies, Foster City, California 94404, USA. [11]Genome Analysis Unit, Amgen, South San Francisco, California 94080, USA. [12]Stowers Institute for Medical Research, 1000 East 50th street, Kansas City, Missouri 64110, USA. [13]Department of Pathology and Laboratory Medicine, Kansas University Medical Center, 3901 Rainbow Boulevard, Kansas City, Kansas 66160, USA. [14]Division of Biostatistics, School of Public Health, University of California, Berkeley, California 94720, USA. [15]Department of Biomolecular Engineering. University of California, Santa Cruz. Santa Cruz, California 95064, USA. [16]Department of Genetics and Howard Hughes Medical Institute, Harvard Medical School, Boston, Massachusetts 02115, USA. [17]Department of Plant and Microbial Biology, University of California, Berkeley, California 94720, USA. [18]Affymetrix, Santa Clara, California 95051, USA.

*These authors contributed equally to this work.

©2011 Macmillan Publishers Limited. All rights reserved

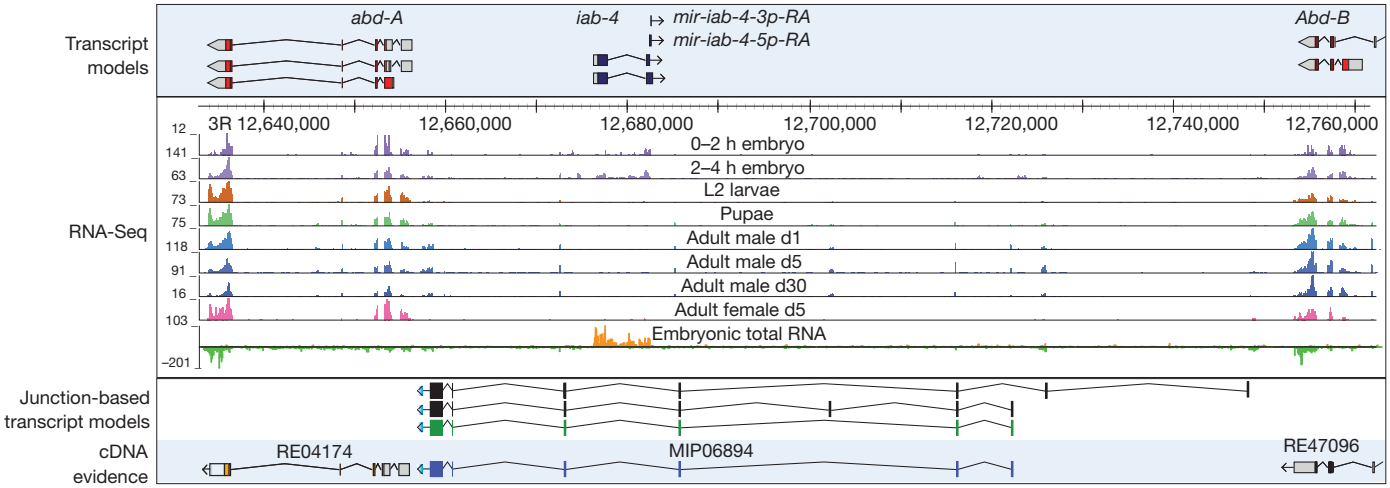

Figure 1 | Discovery of new RNAs in the Bithorax complex. Genomic organization and experimental evidence for new transcripts located between the HOX genes, *abd-A* and *Abd-B*, based on short poly(A)$^+$ RNA and total RNA-Seq expression profiles. The numbers to the left of each track indicate the maximal number of reads for that sample. Three manually curated junction-based transcript models are shown; the green transcript model was fully validated by a cDNA, MIP06894.

and 30 days after eclosion. We used 38-base-pair (bp) resolution genome tiling microarrays to analyse total RNA from all 30 biological samples and poly(A)$^+$ mRNA from the 12 embryonic samples (Supplementary Fig. 1). To attain single-nucleotide resolution and to facilitate the analysis of alternative splicing and RNA editing, we performed non-strand-specific poly(A)$^+$ RNA-Seq from all 30 samples generating a combination of single and paired-end ~75-bp reads on the Illumina Genome Analyser IIx platform (short poly(A)$^+$ RNA-Seq) (Supplementary Table 3 and Supplementary Fig. 2). To identify primary transcripts and non-coding RNAs, the 12 embryonic time points were also interrogated with strand-specific 50-bp sequence reads from partially rRNA-depleted total RNA on the Applied Biosystems SOLiD platform (Supplementary Table 4 and Supplementary Fig. 3). To improve connectivity, mixed-stage embryos, adult males and adult females were used to generate ~250-bp reads on the Roche 454 platform (non-strand-specific long poly(A)$^+$ RNA-Seq) (Supplementary Table 5). In total, we generated 176,962,906,041 bp of mapped sequence representing 1,266-fold coverage of the genome and 5,902-fold coverage of the annotated *D. melanogaster* transcriptome.

Discovery of new transcribed regions

We identified 1,938 new transcribed regions (NTRs) not linked to any annotated gene models. Herein, 'transcripts' refer to RNA molecules synthesized from a genomic locus whereas 'genes' refer to one or more transcripts that share exons in their mature spliced form. modENCODE cDNAs fully support 13% of the NTRs (Supplementary Fig. 4) and partially support 23%. Most NTRs (84%) are detected by poly(A)$^+$ RNA-Seq, 44% by total RNA-Seq and 42% by tiling array. Approximately half of the NTRs are conserved in the distantly related *Drosophila pseudoobscura* and *Drosophila mojavensis* (Supplementary Fig. 4b) and 30% of these are detected by poly(A)$^+$ RNA-Seq data from *D. pseudoobscura* or *D. mojavensis* adult heads (Supplementary Fig. 4c, d, Supplementary Table 6 and Supplementary Methods). The NTRs probably eluded previous detection because they are expressed at low levels, in temporally restricted patterns, and are enriched for single-exon genes. The new multi-exon gene models (48%) have fewer, shorter and less conserved exons than annotated genes.

Nearly one-third of the NTRs have a predicted open reading frame (ORF) greater than 100 amino acids. The remaining NTRs could encode small peptides but many are likely to be non-coding RNAs. A small fraction (9%) of NTRs are heterochromatic; most of these (232) have sequence similarity (greater than 100-nucleotide match and greater than 60% identity) to transposable elements (TEs) and represent transcribed TEs or TE fragments. It remains to be determined whether these regions have any function, although recent studies describe TE-associated regions that have acquired functions[16,17].

Even in the well-studied Bithorax complex[2] we found an NTR. Known genetic breakpoints in the infra-abdominal regions *iab-3* to *iab-8*, which lie between the homeotic genes *abdominal A* (*abd-A*) and *Abdominal B* (*Abd-B*), disrupt normal male development and affect fertility[18,19]. Within this region are regulatory elements[20] and evidence for long non-coding RNAs that have eluded detection for over 20 years[21–23]. We used the RNA-Seq data to infer the structures of at least three overlapping transcripts and validated one form (Fig. 1). The RNAs are expressed in embryos and adult males but not females. On the basis of the presumed role of this new gene and spatial expression in the embryonic gonad (data not shown), we have named it *male specific abdominal* (*msa*). The cDNA contains short ORFs that are conserved in the *melanogaster* subgroup and could encode male-specific peptides. Whether they function as regulatory and/or as peptide-encoding RNAs is an important question for understanding development and segmental morphological diversity.

Discovery of small ncRNAs

We identified 37 unannotated intron-encoded and two unannotated intergenic small ncRNAs (<300 nucleotides) with an average fragments per kilobase of transcript per million fragments mapped (FPKM)[24] >20 from total embryonic RNA-Seq (Fig. 2 and Supplementary Table 7). Most of these ncRNAs are highly conserved in *Drosophila* sibling species[25]. We found published but unannotated ncRNAs: a U4atac snRNA[26] and four small Cajal-body-specific RNAs (scaRNAs)[27]. Of the remaining 34 ncRNAs, three are box C/D-like snoRNAs, 28 are box H/ACA-like snRNAs, one is a scaRNA-like RNA, and two are unclassified. One-third of these are located in the introns of genes encoding RNA-binding proteins, the majority of which are involved in pre-mRNA splicing (*xl6*, *SC35*, *tra2*, *Dek*, *Prp8*, *Tudor-SN*, and *pUf68*).

Discovery of microRNA primary transcripts

MicroRNAs are processed from primary microRNA transcripts (pri-miRNAs) and are either independently transcribed or embedded in the introns of protein-coding genes. We identified 23 putative independently transcribed pri-miRNAs from the total embryonic RNA-Seq and tiling array data that encode 37 annotated miRNAs (Supplementary Table 8). Only two primary transcripts were previously annotated (*bft* and *iab-4*). The pri-miRNAs range from 1 to 18 kb and terminate

©2011 Macmillan Publishers Limited. All rights reserved

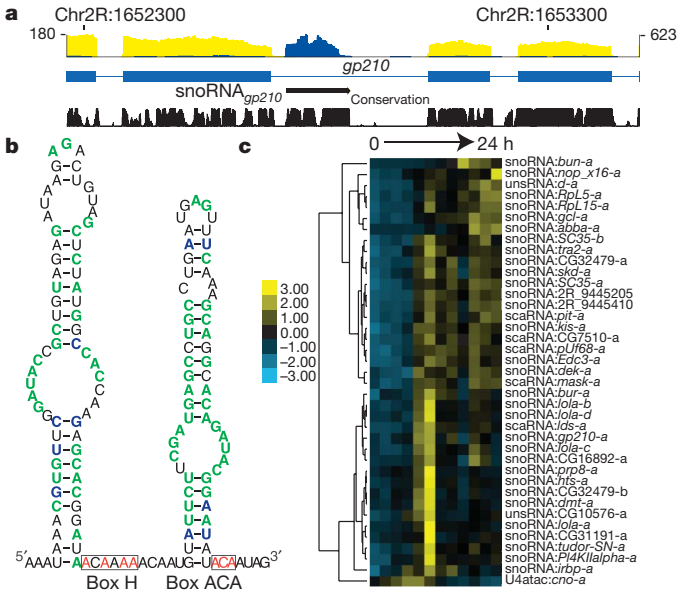

Figure 2 | Discovery of small non-coding RNAs. **a**, Poly(A)$^+$ (yellow) and total RNA (blue) data from 10–12-h embryos are shown for the *gp210* gene which hosts a representative new snoRNA. The maximal number of reads in the poly(A)$^+$ and total RNA-Seq data are shown on the left and right of the track, respectively. **b**, The predicted RNA secondary structure of snoRNA$_{gp210}$ is characteristic of a H/ACA-box snoRNA. Nucleotides that are 100% conserved in sequence or base-pairing are indicated in green and blue, respectively. **c**, Embryonic expression of the new small RNAs. The scale bar indicates FPKM *Z*-scores. unsRNA, unclassified small RNA.

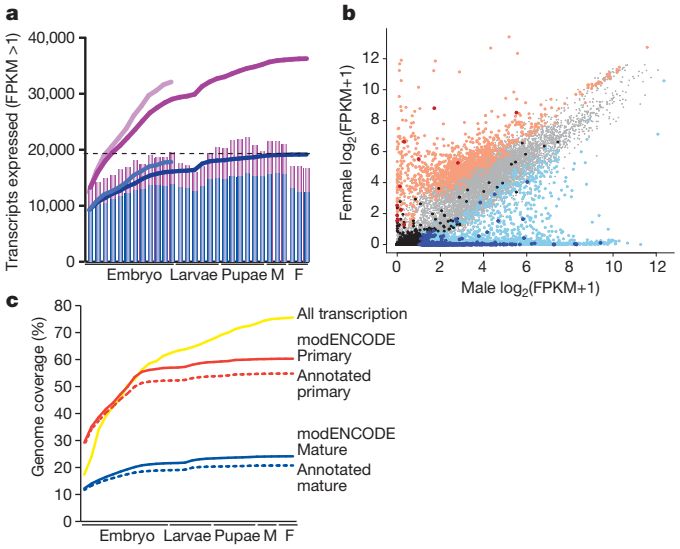

Figure 3 | Dynamics of gene expression. **a**, Transcripts expressed (FPKM >1) in the short poly(A)$^+$ RNA-Seq data: FlyBase 5.12, blue; modENCODE, purple. The bar graphs indicate the number of transcripts expressed in each sample (Supplementary Table 1); the lines indicate the cumulative number of expressed transcripts. The lighter blue and purple lines indicate the cumulative number of transcripts expressed in the embryonic total RNA-Seq samples. The horizontal dotted lines indicate the number of expressed previously annotated transcripts. F, female; M, male. **b**, Scatter plot of sex-biased gene expression. Light red, female-biased annotated ($n = 960$); dark red, female-biased NTRs ($n = 12$); light blue, male-biased annotated ($n = 2{,}401$); dark blue, male-biased NTRs ($n = 431$); light grey, unbiased annotated ($n = 8{,}217$); black, unbiased NTRs ($n = 136$). **c**, Genome coverage. For each developmental sample, the short poly(A)$^+$ reads were used to estimate the percentage of the genome covered using a cutoff of two reads. The mature and primary transcripts were inferred for the previously FlyBase 5.12 (dotted lines) and modENCODE (solid lines) gene models.

at the mature miRNA (*pre-mir-315*, Supplementary Fig. 5a). Twelve of the 23 precursors have cap analysis of gene expression (CAGE) peaks that map at their initiation sites[28]. pri-miRNA expression is dynamic in embryonic development (Supplementary Fig. 5b).

Overview of the *Drosophila* transcriptome

We calculated expression levels of annotated genes, transcripts and NTRs (Supplementary Table 9) in the short poly(A)$^+$ RNA-Seq and tiling array data sets. From the RNA-Seq data we detected expression of 14,862 genes (Supplementary Fig. 7a) and 36,274 transcripts (Fig. 3a) with an FPKM >1 (Supplementary Tables 9–18) of which 67% of genes and 58% of transcripts were also observed in the array data (score >300) (Supplementary Fig. 6 and Supplementary Tables 19 and 20). This includes the confirmation of 87% of annotated genes and transcripts and the discovery of 17,745 new transcripts. In addition, from the total RNA-Seq data we detected expression of 12,854 genes and 32,139 transcripts with an FPKM >1 (Supplementary Tables 12, 13, 21 and 22) of which 77% of genes and 89% of transcripts were also observed in the array data. Of the genes and transcripts observed exclusively in the total RNA-Seq data, 519 genes and 1,005 transcripts (primarily noncoding) were previously annotated and 122 genes and 1,422 transcripts are new discoveries. The genes and transcripts not detected in any data set include small genes (<200 bp), members of multi-copy gene families such as ribosomal RNAs, paralogues (expected owing to our mapping parameters), genes known to be expressed at low levels or in small numbers of cells (for example, gustatory and odorant receptor genes), and non-polyadenylated transcripts.

Expression dynamics

We examined the dynamics of gene expression throughout development using the short poly(A)$^+$ RNA-Seq data. The numbers of expressed genes (FPKM >1) (Supplementary Fig. 7a) and transcripts (Fig. 3a) gradually increases, from 7,045 (0-2 h embryos) to 12,000 (adult males). Adult males express ~3,000 more genes than adult females, consistent with the known transcriptional complexity of the testis[29]. We observed that 40% of expressed genes are constitutively expressed in 30 samples (Supplementary Fig. 7b). We also observed developmentally regulated expression of TEs (Supplementary Materials and Supplementary Fig. 8).

We observed pronounced expression changes in over 1,500 genes in the first two third instar larval samples (Supplementary Fig. 7a, c). Expression of 1,199 genes increased at least tenfold, and 421 genes decreased at least tenfold (Supplementary Table 23). Nearly all of the upregulated genes are expressed for the first time during the third instar stage and most are poorly characterized genes.

The earliest known event in metamorphosis is the 'mid-3rd transition'[30], identified by the synchronous changes in the transcription of a number of well studied genes, *Ecdysone-induced protein 28/29kD* and *Fat body protein 1* (reviewed in ref. 31), and the switch from proximal to distal promoters of *Alcohol dehydrogenase*[32]. These markers coincide with the surge reported here. The mid-3rd transition has no morphological or behavioural correlates and is associated with a pulse of the steroid hormone ecdysone[33] acting through a non-standard receptor[34]. Whether the onset of testis development is a consequence of the mid-3rd transition, or whether the two events are functionally related, remains to be investigated.

Over 29% of protein-coding genes showed significant sex-biased expression in adults (false discovery rate <0.1%), with more male-biased (1,829) or male-specific genes (572) than female-biased (945) or female-specific genes (15) (Supplementary Tables 24 and 25, and Fig. 3b). Known female (*ovo* and *otu*) and male (*dj*) sex-biased genes were expressed as expected. We found that 74% of the NTRs expressed in adults were significantly male-biased whereas only 2.1% were significantly female-biased.

©2011 Macmillan Publishers Limited. All rights reserved

Table 1 | Classification of alternative splicing events

Splicing event	Diagram	FlyBase r5.12	modENCODE	New events	Short poly(A)$^+$ RNA-Seq	Significantly changing
Cassette exons		793	2,717	2,014	2,369	1,539
Alternative 5′ splice sites		843	5,192	4,599	4,583	3,142
Alternative 3′ splice sites		879	6,253	5,505	5,579	3,242
Mutually exclusive exons		229	251	123	228	226
Coordinate cassette exons		301	1,227	979	992	467
Alternative first exons		1,767	4,936	3,442	4,473	3,996
Alternative last exons		227	604	432	553	471
Retained/unprocessed introns		1,434	2,679 (5,667)	1,275 (4,263)	2,439 (35,641)	868 (8,998)
Total		6,437	23,859 (26,847)	18,369 (21,478)	21,216 (54,418)	13,951 (22,081)

The number of retained/unprocessed introns in parentheses indicates the total number identified, whereas the number not in parentheses indicates the subset of identified events that have been validated by cDNA sequences or FlyBase 5.12 annotations.

Genome coverage

Mature mRNAs are encoded by 20% of the *D. melanogaster* genome and primary transcripts by 60% (Fig. 3c). An additional 15% of the genome (~75% total) is detected when considering all of the short poly(A)$^+$ RNA-Seq data. However, as greater than 99% of the reads map within the bounds of the transcript models, the reads that map to intergenic regions constitute a small minority of our data. Thus, although pervasive transcription of mammalian genomes has been observed in microarray studies[35], we found little evidence of such 'dark matter'[36] (that is, pervasive transcription) in *D. melanogaster*.

Discovery and dynamics of alternative splicing

To characterize constitutive and alternative splicing, we identified 71,316 splice junctions, of which 22,965 were new discoveries. Of the new splice junctions, 26% were supported by multiple experimental data types and 74% by only one data type, (Supplementary Fig. 9a) primarily short poly(A)$^+$ RNA-Seq. Of the 20,751 new junctions from the short poly(A)$^+$ RNA-Seq data, 7,833 were incorporated into new transcript models or transcribed regions (NTRs). The remaining new junctions have yet to be incorporated into transcript models.

We also identified a total of 102,026 exons (Supplementary Table 26). Of the 52,914 representing new and revised exons, 65% were validated by capture and sequencing of cDNAs and 2,586 were supported by RNA-Seq data from *D. mojavensis* and *D. pseudoobscura*. Of the new exons, 3,392 were identified from the new splice junctions but have yet to be incorporated into transcript models.

To examine splicing dynamics throughout development, we categorized all splicing events into the common types of alternative splicing events (Table 1). We identified a total of 23,859 splicing events, of which 18,369 were new or recategorized, a threefold increase from annotated splicing events. An additional 2,988 retained/unprocessed introns were identified that were supported by only one experimental data type. In all, 7,473 genes contain at least one alternative splicing event, which is 60.7% of the 12,295 expressed multi-exon genes—also a threefold increase in the fraction of genes with alternatively spliced transcripts. Although smaller than the fraction of human genes with alternatively spliced transcripts (95%)[37,38], a larger proportion of *Drosophila* genes encode alternative transcripts than was previously known.

Of the new alternative exons, 8,226 were previously annotated as constitutive. As observed[39], annotated cassette exons, and their flanking introns, are more highly conserved than annotated constitutive exons (Fig. 4a). The newly discovered cassette exons are more highly conserved than the new constitutive exons, although both classes are less conserved than the corresponding class of annotated exons. New cassette exons that were previously annotated as constitutive exons are the most highly conserved set of exons (Fig. 4a). Annotated and new cassette exons show a strong tendency to preserve reading frame (Supplementary Fig. 9b), indicating that these transcripts increase protein diversity. Both annotated and new cassette exons tend to be shorter than their constitutive counterparts, although both sets of new exons tend to be shorter than annotated exons.

To assess the extent of splicing variation we calculated the 'per cent spliced in' or Ψ (ref. 38) for each splicing event in each sample as well as the switch score ($\Delta\Psi$) by determining the difference between the highest and lowest Ψ values across development ($\Delta\Psi = \Psi_{max} - \Psi_{min}$). This revealed a very smooth distribution of $\Delta\Psi$ among all events, indicating that the splicing of most exons is fairly constant whereas only a minority change markedly (Supplementary Fig. 9c and Supplementary Table 27). Only 831 splicing events have a $\Delta\Psi$ value >90. Further statistical

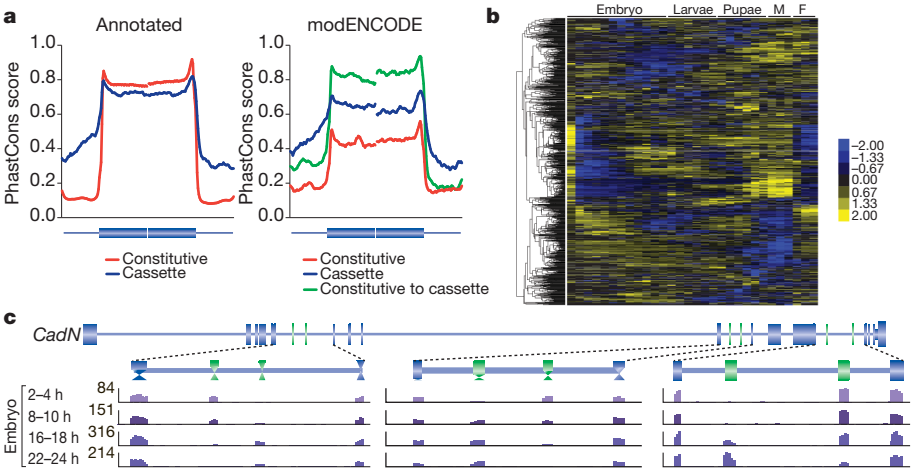

Figure 4 | Developmentally regulated splicing events. **a**, Conservation of internal constitutive and cassette exons >50 nucleotides that were annotated or new discoveries. (Annotated constitutive, $n = 26{,}127$; annotated cassette, $n = 438$; modENCODE cassette, $n = 173$; modENCODE constitutive, $n = 306$; FlyBase 5.12 constitutive to modENCODE cassette, $n = 304$.) **b**, Clusters of regulated cassette exon events during development. The scale bar indicates *Z*-scores of Ψ. **c**, Regulated alternative splicing in *CadN* during embryogenesis. The maximal number of reads in the poly(A)$^+$ RNA-Seq data are indicated for each track.

©2011 Macmillan Publishers Limited. All rights reserved

analyses (see Supplementary Methods) identified 13,951 (66%) alternative splicing events that change significantly throughout development (Supplementary Table 28).

Hierarchical clustering of cassette exon events revealed the dynamic nature of splicing throughout development (Fig. 4b), as exemplified by *Cadherin-N* (*CadN*), a gene with three sets of mutually exclusive exons (Fig. 4c). In each set, one exon is preferentially included in early embryos, the other in late embryos, with a smooth transition between the two. Our analysis also identified groups of exons that have coordinated splicing patterns (Fig. 4b). A set of 55 genes contain exons that are preferentially included in early embryos, late larvae, early pupae and females but skipped in all other stages. Gene Ontology (GO) analysis of these genes indicates that many encode proteins involved in epithelial cell-to-cell junctions. GO analysis of genes that contain exons preferentially included during late pupal and adult stages indicates that many encode proteins that are part of neuronal synapses.

Sex-biased alternative splicing

Sex determination in *Drosophila* is mediated by a cascade of regulated alternative splicing events involving *Sex lethal* (*Sxl*), *transformer* (*tra*), *male-specific lethal 2* (*msl-2*), *doublesex* (*dsx*) and *fruitless* (*fru*) that specify nearly all physical and behavioural dimorphisms between males and females as well as X chromosome dosage compensation[40]. Our RNA-Seq data confirm sex-biased ($\Delta\Psi = |\Psi_{male} - \Psi_{female}|$) splicing of *Sxl* ($\Delta\Psi = 89.6$), *tra* ($\Delta\Psi = 39.2$), *dsx* ($\Delta\Psi = 59.7$) and *fru* ($\Delta\Psi = 100$).

In addition to the canonical sex-determination cascade, we identified 119 strongly sex-biased splicing events ($\Delta\Psi > 70$) (Supplementary Fig. 9d). One striking example is *Reps*, which was annotated as containing six constitutive exons. RNA-Seq data indicate that exon five is a sex-biased alternative cassette exon ($\Delta\Psi = 73.39$) (Supplementary Fig. 10). This highly conserved exon is included in males and skipped in females. The intron upstream of this cassette exon contains conserved SXL binding sites, indicating that it is regulated by SXL and is a candidate sex differentiation gene.

Discovery of RNA editing sites

Previous studies identified 127 sites in 55 *Drosophila* genes that undergo A-to-I RNA editing[41]. This post-transcriptional modification is catalysed by dADAR, which is expressed at increasing levels throughout development and is thought to target products involved in nervous system function. We analysed the poly(A)$^+$ RNA-Seq data to identify exonic nucleotide positions consistent with A-to-I editing and defined 972 edited positions within transcripts of 597 genes, including previously described edited sites in the transcripts of 36 genes (Supplementary Table 29). These genes include those required for rapid neurotransmission and other widely ranging functions. For most sites, the frequency of editing increases throughout development and does not correlate with overall expression levels (Fig. 5a). Editing typically begins in late pupal stages, although we find transcripts that seem to be edited in late embryogenesis. Consistent with earlier studies[42], exons containing editing sites are more highly conserved than unedited exons. The majority of the edited positions (630) alter amino acid coding, the others are either silent (201) or within untranslated regions (141). For example, the transcripts of *quiver* (*qvr*) are edited at six positions, four that result in amino acid changes (Fig. 5b). *qvr* encodes a potassium channel subunit that modulates the function of the voltage-gated Shaker (SH) potassium channel. *Sh* transcripts are also edited at multiple positions[43]. The combinatorial editing of both proteins probably has an important role in modulating action potentials in the arthropod nervous system and may have implications for the regulation of sleep[44]. Expressed sequence tags, long poly(A)$^+$ RNA-Seq and cDNAs cross-validate nearly one-quarter (214) of the newly discovered sites.

Computational analysis identified three potential editing-associated sequence motifs (Fig. 5a). We observe 381 sites with one or more motifs in close proximity to the edited nucleotide (Supplementary Table 30). Motif C, although less common than motifs A and B, is more strongly associated with the editing site. Most (93%) instances of motif C occur on the sense strand of the transcript and the A at the 3′ end of the motif is the edited nucleotide. This motif is over-represented in editing events that occur early in development.

Discussion

Our interrogation of the transcriptome of *D. melanogaster* throughout development has considerably expanded the number of building blocks used to make a fly. Specifically, we identified nearly 2,000 NTRs, increased the number of alternative splicing events by threefold and the number of RNA editing sites by an order of magnitude. The resulting view of the transcriptome at single-base resolution markedly improves our understanding of expression dynamics throughout the *Drosophila* life cycle and has substantial biological implications.

The *D. melanogaster*, *C. elegans* and human genomes are organized quite differently. Specifically, 20%, 45% and 2.5% of the *D. melanogaster*, *C. elegans* and human genomes, respectively, encode exons or mature transcripts. Primary transcripts comprise a larger fraction of each genome—60%, 82% and 37%. This highlights the fact that primary transcripts and introns are much shorter in *D. melanogaster* and *C. elegans* than in human and that the *D. melanogaster* and *C. elegans* genomes are more compact than the human genome.

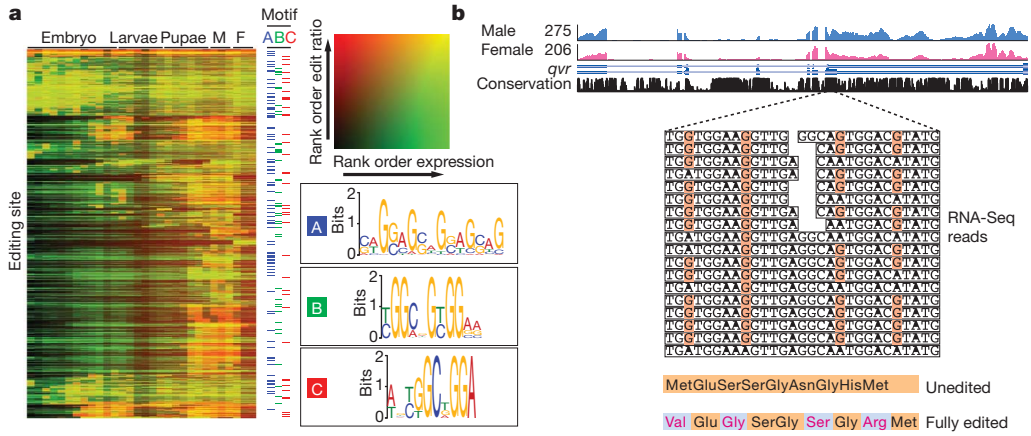

Figure 5 | Discovery of RNA editing events. **a**, Rows represent edited sites. Rank-ordered expression levels (number of reads) are shown in green and the rank-ordered editing ratios are shown in red. Pictogram representations of editing motifs A, B and C are shown. **b**, RNA editing of *qvr*. Male and female expression and conservation tracks are shown above RNA-Seq reads from adult females that align to the edited positions (orange). Conceptual translation of the unedited and fully edited transcripts result in four amino acid changes (red) at the C terminus of QVR.

©2011 Macmillan Publishers Limited. All rights reserved

The existence of unannotated genes was indicated by microarray studies[8,45] and conservation among Drosophilid genomes[25]. However, the NTRs that we identified were not identified by comparative sequence analysis[46] as they are less conserved than most previously known genes. This emphasizes the importance of using both comparative analyses and transcriptome profiling for genome annotation.

Despite the depth of our sequencing, the annotation of the *D. melanogaster* transcriptome is not finished. We failed to detect expression of 1,488 annotated genes including members of gene families to which short reads can not be uniquely mapped and genes expressed at low levels or in spatially and temporally restricted patterns. Moreover, although we substantially increased the fraction of genes that encode alternatively spliced or edited transcripts, we again failed to detect several annotated RNA processing events. Study of more temporally and spatially restricted samples will allow deeper exploration of the *Drosophila* transcriptome, and almost certainly result in the discovery of yet additional features. Furthermore, functional studies of the new and previously unstudied elements will provide valuable insight into metazoan development.

METHODS SUMMARY

Animal staging, collection and RNA extraction. Isogenic (y^1; *cn* bw^1 sp^1) embryos were collected at 2-h intervals for 24 h. Collection of later staged animals started with synchronized embryos and included resynchronizing with appropriate age indicators. Six larval, six pupal and three adult sexed stages, 1, 5 and 30 days, were collected. RNA was isolated using TRIzol (Invitrogen), DNased and purified on an RNAeasy column (Qiagen). poly(A)$^+$ RNA was prepared from an aliquot of each total RNA sample using an Oligotex kit (Qiagen).

Tiling arrays. RNAs from three biological replicates of each sample were independently hybridized on 38-bp arrays (Affymetrix GeneChip *Drosophila* Tiling 2.0R array) as described[47].

RNA-Seq. Libraries were generated and sequenced on an Illumina Genome Analyser IIx using single or paired-end chemistry and 76-bp cycles. SOLiD sequencing used total RNA treated with the RiboMinus Eukaryote Kit (Invitrogen). Samples were fragmented, adaptors ligated (Ambion) and sequenced for 50 bases using SOLiD V3 chemistry. 454 sequencing used poly(A)$^+$ RNA from Oregon R adult males and females and mixed-staged y^1; *cn* bw^1 sp^1 embryos. Sequences are available from the Short Read Archive and the modENCODE website (http://www.modencode.org/).

Targeted RT–PCR and cDNA isolation and sequencing. Standard procedures were used for RT–PCR and targeted cDNA isolation and sequencing.

Analysis. Cufflinks[24] was used to identify new transcript models and to calculate expression levels for annotated and predicted transcript models. MFold[48] was used to predict secondary structures from the new snoRNA-like RNAs. JuncBASE[49] identified alternative splicing events and calculated per cent spliced in (Ψ)[38]. Editing sites were identified by comparing aligned reads to the reference genome.

Received 3 September; accepted 2 December 2010.
Published online 22 December 2010.

1. Morgan, T. H. Sex limited inheritance in *Drosophila*. *Science* **32,** 120–122 (1910).
2. Lewis, E. B. A gene complex controlling segmentation in *Drosophila*. *Nature* **276,** 565–570 (1978).
3. Nüsslein-Volhard, C. & Wieschaus, E. Mutations affecting segment number and polarity in *Drosophila*. *Nature* **287,** 795–801 (1980).
4. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287,** 2204–2215 (2000).
5. Spradling, A. C. Learning the common language of genetics. *Genetics* **174,** 1–3 (2006).
6. Arbeitman, M. N. *et al.* Gene expression during the life cycle of *Drosophila melanogaster*. *Science* **297,** 2270–2275 (2002).
7. Stolc, V. *et al.* A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* **306,** 655–660 (2004).
8. Manak, J. R. *et al.* Biological function of unannotated transcription during the early development of *Drosophila melanogaster*. *Nature Genet.* **38,** 1151–1158 (2006).
9. Bass, B. L. *RNA Editing* (Oxford Univ. Press, 2001).
10. Rana, T. M. Illuminating the silence: understanding the structure and function of small RNAs. *Nature Rev. Mol. Cell Biol.* **8,** 23–36 (2007).
11. Lipshitz, H. D., Peattie, D. A. & Hogness, D. S. Novel transcripts from the *Ultrabithorax* domain of the Bithorax Complex. *Genes Dev.* **1,** 307–322 (1987).
12. Meller, V. H., Wu, K. H., Roman, G., Kuroda, M. I. & Davis, R. L. roX1 RNA paints the X chromosome of male *Drosophila* and is regulated by the dosage compensation system. *Cell* **88,** 445–457 (1997).
13. Celniker, S. E. *et al.* Unlocking the secrets of the genome. *Nature* **459,** 927–930 (2009).
14. The modENCODE Consortium *et al.* Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* doi:10.1126/science.1198374 (in the press).
15. Gerstein, M. B. *et al.* Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* doi:10.1126/science.1196914 (in the press).
16. Bejerano, G. *et al.* A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441,** 87–90 (2006).
17. Xie, X., Kamal, M. & Lander, E. S. A family of conserved noncoding elements derived from an ancient transposable element. *Proc. Natl Acad. Sci. USA* **103,** 11659–11664 (2006).
18. Karch, F. *et al.* The abdominal region of the Bithorax Complex. *Cell* **43,** 81–96 (1985).
19. Celniker, S. E., Sharma, S., Keelan, D. & Lewis, E. B. The molecular genetics of the bithorax complex of *Drosophila* cis-regulation in the *Abdominal-B* domain. *EMBO J.* **9,** 4277–4286 (1990).
20. Ho, M. C. *et al.* Functional evolution of cis-regulatory modules at a homeotic gene in *Drosophila*. *PLoS Genet.* **5,** e1000709 (2009).
21. Sanchez-Herrero, E. & Akam, M. Spatially ordered transcription of regulatory DNA in the bithorax complex of *Drosophila*. *Development* **107,** 321–329 (1989).
22. Bae, E., Calhoun, V. C., Levine, M., Lewis, E. B. & Drewell, R. A. Characterization of the intergenic RNA profile at abdominal-A and Abdominal-B in the *Drosophila* bithorax complex. *Proc. Natl Acad. Sci. USA* **99,** 16847–16852 (2002).
23. Bender, W. MicroRNAs in the *Drosophila* bithorax complex. *Genes Dev.* **22,** 14–19 (2008).
24. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnol.* **28,** 511–515 (2010).
25. Clark, A. G. *et al.* Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450,** 203–218 (2007).
26. Padgett, R. A. & Shukla, G. C. A revised model for U4atac/U6atac snRNA base pairing. *RNA* **8,** 125–128 (2002).
27. Tycowski, K. T., Shu, M. D., Kukoyi, A. & Steitz, J. A. A conserved WD40 protein binds the Cajal body localization signal of scaRNP particles. *Mol. Cell* **34,** 47–57 (2009).
28. Hoskins, R. A. *et al.* Genome-wide analysis of promoter architecture in *Drosophila melanogaster*. *Genome Res.* doi:10.1101/gr.112466.110 (in the press).
29. Parisi, M. *et al.* A survey of ovary-, testis-, and soma-biased gene expression in *Drosophila melanogaster* adults. *Genome Biol.* **5,** R40 (2004).
30. Andres, A. J. & Cherbas, P. Tissue-specific ecdysone responses: regulation of the *Drosophila* genes Eip28/29 and Eip40 during larval development. *Development* **116,** 865–876 (1992).
31. Andres, A. J., Fletcher, J. C., Karim, F. D. & Thummel, C. S. Molecular analysis of the initiation of insect metamorphosis: a comparative study of *Drosophila* ecdysteroid-regulated transcription. *Dev. Biol.* **160,** 388–404 (1993).
32. Lockett, T. J. & Ashburner, M. Temporal and spatial utilization of the alcohol dehydrogenase gene promoters during the development of *Drosophila melanogaster*. *Dev. Biol.* **134,** 430–437 (1989).
33. Warren, J. T. *et al.* Discrete pulses of molting hormone, 20-hydroxyecdysone, during late larval development of *Drosophila melanogaster*: correlations with changes in gene activity. *Dev. Dyn.* **235,** 315–326 (2006).
34. Costantino, B. F. *et al.* A novel ecdysone receptor mediates steroid-regulated developmental events during the mid-third instar of *Drosophila*. *PLoS Genet.* **4,** e1000102 (2008).
35. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
36. van Bakel, H., Nislow, C., Blencowe, B. J. & Hughes, T. R. Most "dark matter" transcripts are associated with known genes. *PLoS Biol.* **8,** e1000371 (2010).
37. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genet.* **40,** 1413–1415 (2008).
38. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456,** 470–476 (2008).
39. Philipps, D. L., Park, J. W. & Graveley, B. R. A computational and experimental approach toward a priori identification of alternatively spliced exons. *RNA* **10,** 1838–1844 (2004).
40. Sanchez, L. Sex-determining mechanisms in insects. *Int. J. Dev. Biol.* **52,** 837–856 (2008).
41. Stapleton, M., Carlson, J. W. & Celniker, S. E. RNA editing in *Drosophila melanogaster*: New targets and functional consequences. *RNA* **12,** 1922–1932 (2006).
42. Jepson, J. E. & Reenan, R. A. Genetic approaches to studying adenosine-to-inosine RNA editing. *Methods Enzymol.* **424,** 265–287 (2007).
43. Hoopengardner, B., Bhalla, T., Staber, C. & Reenan, R. Nervous system targets of RNA editing identified by comparative genomics. *Science* **301,** 832–836 (2003).
44. Wang, J. W. & Wu, C. F. Modulation of the frequency response of Shaker potassium channels by the quiver peptide suggesting a novel extracellular interaction mechanism. *J. Neurogenet.* **24,** 67–74 (2010).
45. Hild, M. *et al.* An integrated gene annotation and transcriptional profiling approach towards the full gene content of the *Drosophila* genome. *Genome Biol.* **5,** R3 (2003).
46. Stark, A. *et al.* Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450,** 219–232 (2007).
47. Cherbas, L. The transcriptional diversity of 25 *Drosophila* cell lines. *Genome Res.* doi:10.1101/gr.112961.110 (in the press).
48. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31,** 3406–3415 (2003).

©2011 Macmillan Publishers Limited. All rights reserved

49. Brooks, A. N. *et al.* Conservation of an RNA regulatory map between *Drosophila* and mammals. *Genome Res.* doi:10.1101/gr.108662.110 (in the press).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Trapnell and L. Pachter for discussions and assistance with Cufflinks, and E. Clough for comments and feedback. A.N.B. was partially supported by an NSF graduate fellowship. This work was funded by an award from the National Human Genome Research INstitute modENCODE Project (U01 HB004271) to S.E.C. (Principal Investigator) and M.R.B., P.C., T.R.G., B.R.G. and N.P. (co-Principal Investigators) under Department of Energy contract no. DE-AC02-05CH11231, and by the National Institute of Diabetes and Digestive and Kidney Diseases Intramural Research Program (B.O.).

Author Contributions J.A., M.R.B., P.C., T.R.G., B.R.G., R.A.H., T.C.K., B.O., N.P. and S.E.C. designed the project. J.A., S.E.B., M.R.B., P.C., T.R.G., B.R.G., R.A.H., B.O. and S.E.C. managed the project. D.M. prepared biological samples. T.C.K. oversaw biological sample production. D.Z. and B.E. prepared RNA samples. J.A. oversaw RNA sample production. W.L. and A.W. analysed array data. P.K. managed array data production. L.Y. prepared Illumina RNA-Seq libraries. C.A.D., L.L., J.E.S., K.H.W. and L.Y. performed Illumina sequencing. J.M.L., B.R.G. and S.E.C. managed Illumina sequencing production. M.B. and R.E.G. performed 454 sequencing of adults. R.A.H. managed production of the embryonic SOLiD and 454 sequencing. C.A.D. managed data transfers. C.Z. managed databases and formatted array and sequence data for submission. C.G.A., P.J.B., S.E.B., A.N.B., S.D., M.O.D., B.R.G. and D.S. developed analysis methods. C.G.A., J.B.B., N.B., B.W.B., S.E.B., A.N.B., J.W.C., S.E.C., L.C., P.C., C.A.D., A.D., M.O.D., B.R.G., R.L., J.H.M., N.R.M., D.S. and Yi.Z. analysed data. B.B.T. aligned the SOLiD data. M.J.V. and J.M.L. generated annotations. C.G.A., D.S. and J.H.M. analysed species validation data. L.J., C.G.A., D.S. and N.R.M. performed species RNA-Seq quality control. Yu.Z. and J.H.M. oversaw sequencing and gathered species samples. C.G.A., A.N.B., J.W.C., L.C., P.C., A.H., D.S., J.M.L., R.L. N.R.M., J.H.M. and B.O. contributed to the text. A.H. assisted with manuscript preparation. B.R.G. and S.E.C. wrote the paper with input from all authors. All authors discussed the results and commented on the manuscript.

Author Information All sequence data have been deposited in the SRA, cDNA sequences have been deposited in GenBank, and array data deposited in GEO (see Supplementary Table 35 for all accession numbers). All data is also available at http://www.modencode.org. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to B.R.G. (graveley@neuron.uchc.edu) or S.E.C. (celniker@fruitfly.org).

©2011 Macmillan Publishers Limited. All rights reserved