Published in final edited form as:

Nature. 2014 August 28; 512(7515): 393–399. doi:10.1038/nature12962.

Diversity and dynamics of the *Drosophila* transcriptome

James B. Brown[a,b,§], **Nathan Boley**[a,§], **Robert Eisman**[c,§], **Gemma E. May**[d,§], **Marcus H. Stoiber**[a], **Michael O. Duff**[d], **Ben W. Booth**[b], **Jiayu Wen**[g], **Soo Park**[b], **Ana Maria Suzuki**[i], **Kenneth H. Wan**[b], **Charles Yu**[b], **Dayu Zhang**[e], **Joseph W. Carlson**[b], **Lucy Cherbas**[c], **Brian D. Eads**[c], **David Miller**[c], **Keithanne Mockaitis**[c], **Johnny Roberts**[e], **Carrie A. Davis**[f], **Erwin Frise**[b], **Ann S. Hammonds**[b], **Sara Olson**[d], **Sol Shenker**[g], **David Sturgill**[h], **Anastasia A. Samsonova**[j], **Richard Weiszmann**[b], **Garret Robinson**[a], **Juan Hernandez**[a], **Justen Andrews**[c], **Peter J. Bickel**[a], **Piero Carninci**[i], **Peter Cherbas**[c,e], **Thomas R. Gingeras**[f], **Roger A. Hoskins**[b], **Thomas C. Kaufman**[c], **Eric C. Lai**[g], **Brian Oliver**[h], **Norbert Perrimon**[j], **Brenton R. Graveley**[d], and **Susan E. Celniker**[b]

[a]Department of Statistics, University of California Berkeley, Berkeley, CA

[b]Department of Genome Dynamics, Lawrence Berkeley National Laboratory, Berkeley, CA

[c]Department of Biology, Indiana University, 1001 E. 3rd Street, Bloomington, IN

[d]Department of Genetics and Developmental Biology, Institute for Systems Genomics, University of Connecticut Health Center, 400 Farmington Avenue, Farmington, CT

[e]Center for Genomics and Bioinformatics, Indiana University, 1001 E. 3rd Street, Bloomington, IN

[f]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

[g]Sloan-Kettering Institute, 1017C Rockefeller Research Labs 1275 York Avenue, Box 252 New York, NY 10065

[h]Section of Developmental Genomics, Laboratory of Cellular and Developmental Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes or Health, Bethesda MD

Correspondence to: James B. Brown; Brenton R. Graveley; Susan E. Celniker.

[§]These authors contributed equally and should be considered co-first authors

Author Contributions J.A., T.R.G., B.R.G., R.A.H., T.C.K., and S.E.C. designed the project. J.A., P.Ch., T.R.G., B.R.G., R.A.H., J.B.B., B.O. and S.E.C. managed the project. R.E. designed treatment protocols and prepared biological samples. T.C.K., J.A. and L.C. oversaw biological sample production. B.D.E., D. M. and J.R. prepared biological samples. D.Z. and B.E. prepared RNA samples. J.A. oversaw RNA sample production. G.E.M., S.O., and L.Y. prepared Illumina RNA-Seq libraries. A.M.S. prepared CAGE libraries. P.Ca. oversaw production of CAGE libraries. C.A.D., G.E.M., S.O., L.Y., S.P., and K.H.W. performed Illumina sequencing. B.R.G. and S.E.C. managed Illumina sequencing production. RAH conceived the polyA+seq method. RW and RAH developed the polyA+seq protocol and produced the libraries. K.M. performed 454 sequencing. C.Y., S.P., and K.H.W. performed cDNA library screens and full-insert cDNA sequencing. S.E.C. oversaw cDNA production. E.F. and N.B. installed and administered computer infrastructure for data storage and analysis. J.B.B., N.B., M.H.S., M.O.D., B.W.B., D.S., J.W.C., S.S., J.W., A.A.S., N.P., E.C.L., P.J.B., and B.R.G. developed analysis methods. J.B.B., N.B., M.H.S., M.O.D., B.W.B., A.H., E.F., R.A.H., S.S., D.S., L.C., G.R., J.H., J.W., A.A.S., E.C.L., K.H.W., B.R.G., and S.E.C. analyzed data. N.B., J.B.B. M.H.S., K.H.W., and S.E.C. generated annotations. D.S. and B.O. analyzed species validation data. S.S., J.W. and E.C.L. analyzed 3'UTR and antisense data. A.H., E.F., and S.E.C. analyzed image data. M.H.S. analyzed proteomics data. M.H.S., S.S., D.S., B.O., E.C.L., T.C.K., R.E., R.A.H., and P.Ch. contributed to the text. A.H. assisted with manuscript preparation. J.B.B., 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.

[i]Omics Science Center, RIKEN Yokohama Institute, Yokohama, 230-0045 Kanagawa, Japan

[j]Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA

[k]Howard Hughes Medical Institute, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA

Abstract

Animal transcriptomes are dynamic, each cell type, tissue and organ system expressing an ensemble of transcript isoforms that give rise to substantial diversity. We identified new genes, transcripts, and proteins using poly(A)+ RNA sequence from *Drosophila melanogaster* cultured cell lines, dissected organ systems, and environmental perturbations. We found a small set of mostly neural-specific genes has the potential to encode thousands of transcripts each through extensive alternative promoter usage and RNA splicing. The magnitudes of splicing changes are larger between tissues than between developmental stages, and most sex-specific splicing is gonad-specific. Gonads express hundreds of previously unknown coding and long noncoding RNAs (lncRNAs) some of which are antisense to protein-coding genes and produce short regulatory RNAs. Furthermore, previously identified pervasive intergenic transcription occurs primarily within newly identified introns. The fly transcriptome is substantially more complex than previously recognized arising from combinatorial usage of promoters, splice sites, and polyadenylation sites.

INTRODUCTION

Next-generation RNA sequencing has permitted the mapping of transcribed regions of the genomes of a variety of organisms[1,2]. These studies demonstrated that large fractions of metazoan genomes are transcribed and cataloged individual elements of transcriptomes, including promoters[3], polyadenylation sites[4,5], exons and introns[6]. However, the complexity of the transcriptome arises from the combinatorial incorporation of these elements into mature transcript isoforms. Studies that inferred transcript isoforms from short read sequence data focused on a small subset of isoforms, filtered using stringent criteria[7,8]. Studies using cDNA or EST data to infer transcript isoforms have not had sufficient sampling depth to explore the diversity of RNA products at the majority of genomic loci[9]. While the human genome has been the focus of intensive manual annotation[10], analysis of strand-specific RNA-seq data from human cell lines reveals over 100,000 splice junctions not incorporated into transcript models[11]. Thus, a large gap exists between genome annotations and the emerging transcriptomes observed in next-generation sequence data. In *Drosophila*, we previously described a non-strand-specific poly(A)+ RNA-seq analysis of a developmental time course through the life cycle[6] and CAGE analysis of the embryo[12], which discovered thousands of unannotated exons, introns and promoters, and expanded coverage of the genome by identified transcribed regions, but not necessarily transcript models. Here, we describe an expansive poly(A)+ transcript set modeled by integrative analysis of transcription start sites (CAGE and 5' RACE), splice sites and exons (RNA-seq), and polyadenylation sites (3' ESTs, cDNAs and RNA-seq). We analyzed poly(A)+ RNA data from a diverse set of developmental stages[6], dissected organ systems and environmental perturbations, most of which is new and strand-specific. Our data provide higher spatiotemporal resolution and allow for deeper exploration of the *Drosophila* transcriptome than was previously possible. Our analysis reveals a transcriptome of high complexity that is expressed in discrete, tissue- and condition-specific mRNA and ncRNA transcript isoforms that span the majority of the genome and provides valuable insight into metazoan biology.

RESULTS

A dense landscape of discrete poly(A)+ transcripts

To broadly sample the transcriptome, we performed strand-specific, paired-end sequencing of poly(A)+ RNA in biological duplicate from 29 dissected tissue samples including the nervous, digestive, reproductive, endocrine, epidermal, and muscle organ systems of larvae, pupae and adults. To detect RNAs not observed under standard conditions we sequenced poly(A)+ RNA in biological duplicate from 21 whole-animal samples treated with environmental perturbations. Adults were challenged with heat-shock, cold-shock, and exposure to heavy metals (cadmium, copper and zinc), the drug caffeine, and the herbicide paraquat. To determine if exposing larvae resulted in novel RNA expression we treated them with heavy metals, caffeine, ethanol and rotenone. Lastly, we sequenced poly(A)+ RNA from 21 previously described[13] and three ovary-derived cell lines (Supplementary Methods). In total, we produced 12.4B strand-specific read-pairs and over a terabase of sequence data, providing 44,000 fold coverage of the poly(A)+ transcriptome.

Reads were aligned to the *Drosophila* genome as described[6], and full-length transcript models were assembled using our custom pipeline, GRIT[14] (Supplementary Methods). GRIT uses RNA-seq, p(A)+seq, CAGE, RACE[12], ESTs[15], and full-length cDNAs[16] to generate gene and transcript models. We integrated these models with our own and community manual curation datasets to obtain an annotation (Supplementary Material, section 12) consisting of 304,788 transcripts and 17,564 genes (Fig. 1a and Supplementary Fig. 1), of which 14,692 are protein-coding (Supplementary Data File 1). Ninety percent of genes produce at most 10 transcript and five protein isoforms, while 1% of genes have highly complex patterns of alternative splicing, promoter usage, and polyadenylation, and may each be processed into hundreds of transcripts (Fig. 1a, example 1b). Our gene models span 72% of the euchromatin, an increase from 65% in FlyBase 5.12 (FB5.12), the reference annotation at the beginning of the modENCODE project (Supplementary Table 1 compares annotations 2008–2013). There were 64 euchromatic gene-free regions longer than 50kb in FB5.12, and 25 remaining in FB5.45. Our annotation includes new gene models in each of these regions. Newly identified genes (1468 total) are expressed in spatially- and temporally-restricted patterns (Supplementary Fig. 2), and 536 reside in previously uncharacterized gene-free regions. Others map to well-characterized regions, including the *ovo* locus, where we discovered a new ovary-specific, poly(A)+ transcript (*Mgn94020*), extending from the second promoter of *ovo* on the opposite strand and spanning 107kb (Fig. 1c). Exons of 36 new genes overlap molecularly defined mutations with associated phenotypes (GSC *p*-value~0.0002), suggesting potential functions (Supplementary Table 2). For instance, the lethal P-element insertions *l(3)L3051* and *l(3)L4111*[17] map to promoters of *Mgn095159* and *Mgn95009*, respectively, suggesting these may be essential genes. Nearly

60% of the intergenic transcription we previously reported[6] is now incorporated into gene models.

Transcript Diversity

Over half of spliced genes (7412; 56%) encode two or more transcript isoforms with alternative first exons (AFEs). The majority of such genes produce AFEs through coordinated alternative splicing and promoter usage (59%, 4389 genes, hypergeometric *p*-value<1e–16); however, a substantial number of genes utilize one, but not both mechanisms (Fig. 2a). Only 1058 spliced genes have AFEs that alter coding capacity and increase the complexity of the predicted proteome. Some genes, such as *G protein β-subunit 13F* (*Gβ13F*, Fig. 2b, Supplementary Fig. 3) have exceptionally complex 5'UTRs, but encode a single protein.

We measured splicing efficiency using the “percent spliced in” (ψ) index – the fraction of isoforms that contain the exon[6]. Introns flanked by coding sequence are retained at an average ψ=0.7, whereas introns flanked by non-coding sequence are retained >5-fold more often, with an average ψ=3.8 (*p*<1e–16 subsampling/2-sample t-test), and is most frequent in 5'UTRs (mean ψ=5.1, Fig. 2c).

Despite the depth of our RNA-seq, these data show that 42% of genes encode only a single transcript isoform, and 55% encode a single protein isoform (Supplementary Methods). In mammals, it has been estimated that 95% of genes produce multiple transcript isoforms[18,19], (estimates for protein-coding capacity have not been reported).

The majority of transcriptome complexity is attributable to forty-seven genes that have the capacity to encode >1000 transcript isoforms each (Supplementary Table 3), and account for 50% of all transcripts (Fig. 3a). Furthermore, 27% of transcripts encoded by these genes were detected exclusively in samples enriched for neuronal tissue, and another 56% only in the embryo (83% total). To determine their tissue specificities we conducted embryonic *in situ* expression assays (Fig. 3b) and found that 18 of 35 are detected only in neural tissue (51% vs. 10% genome-wide, hypergeometric *p*-value<1e–16, Supplementary Table 4). Of these genes, 48% have 3'UTR extensions in embryonic neural tissue[20] (5% genome-wide, *p*<1e–16). Furthermore, 44% are targets of RNA editing (4% genome-wide[6], *p*<1e–16, with 18 of 21 validated[21]), and 21% have 3'UTR extensions and RNA editing sites (10 of 65 genome-wide, *p*<1e–100). The capacity to encode thousands of transcripts is largely specific to the nervous system and coincides with other classes of rare, neural-specific RNA processing.

Tissue- and sex-specific splicing

To examine the dynamics of splicing, we calculated switch scores, or Δψ, for each splicing event by comparing the maximal and minimal ψ values across all samples, and in subsets including just the developmental and tissue samples. In contrast to the median ψ values, the distribution of Δψ values is strikingly different between the developmental and tissue samples. Among the developmental samples, 38% of events have a Δψ≥50% while between the tissue samples 63% of events have a Δψ≥50%. This difference is even more pronounced at higher Δψ thresholds – only 6% of events have a Δψ≥80% between the developmental samples while 31% of events have a Δψ≥80% between the tissue samples. Thus, most splicing events are highly tissue-specific. Of the 17,447 alternative splicing events analyzed (Supplementary Materials, section 19), we find that 56.6% changed significantly (Δψ>20%, Bayes Factor>20). Clustering revealed groups of splicing events that are coordinately regulated in a tissue-specific manner. For example, 1147 splicing events are specifically included in heads and excluded in testes or ovaries, while 797 splicing events are excluded in heads but included in testes or ovaries (Fig. 4a).

We identified hundreds of sex-specific splicing events from adult male and female RNA-seq data[6]. To further explore sex-specific splicing, we compared the splicing patterns in male and female heads enriched for brain tissues. There were striking differences in gene expression levels, however, only seven splicing events were consistently differentially spliced at each time point after eclosion (average Δψ>20%), and these largely corresponded to genes in the known sex-determination pathway (Supplementary Material). We find few examples of head sex-specific splicing. This is in contrast to previous studies, which have come to conflicting conclusions and used either microarrays analyzing only a subset of splicing events or single read 36bp RNA-Seq[22,23] with an order of magnitude fewer reads[24].

We identified 575 alternative splicing events that are differentially spliced in whole male and female animals (Δψ>20%) and analyzed the tissue-specific splicing patterns of each event (Fig. 4b). We found that 186 of the 321 male-biased splicing events were most strongly included in testes or accessory glands, and 157 of 254 female-biased exons were ovary-enriched. Consistent with the extensive transcriptional differences observed in testes compared to other tissues, the genes containing male-specific exons are enriched in functions related to transcription. In contrast, the female-specific exon containing genes are enriched in functions involved in signaling and splicing (http://reactome.org[25], Supplementary Table 6). Together, these results indicate that the majority of sex-specific splicing is due to tissue-specific splicing in tissues present only in males or females.

Long non-coding RNAs

A growing set of candidate long non-coding RNAs (lncRNAs) have been identified in *Drosophila*[6,26,27]. In FB5.45 there were 392 annotated lncRNAs, and it has been suggested that as many as 1119 lncRNAs may be transcribed in the fly[28]. However, this number was based on transcribed regions, not transcript models, and utilized non-stranded RNA-seq data[28]. We find 3880 genes produce transcripts with ORFs encoding fewer than 100 amino acids (aa). Of these, 795 encode conserved proteins (Methods) longer than 20aa. For example, a single exon gene on the opposite strand and in the last intron of the early developmental growth factor *spätzle* encodes a 42aa protein that is highly conserved across all sequenced *Drosophila* species. We identified 1875 candidate lncRNA genes producing 3085 transcripts, 2990 of which have no overlap with protein-coding genes on the same strand (Supplementary Data File 2). Some of these putative lncRNAs may encode short polypeptides, e.g. the gene *tarsal-less* encodes three 11aa ORFs with important developmental functions[29]. We determined protein conservation scores for each ORF between 20 and 100aa (Supplementary Table 6). Of the 1119 predicted lncRNAs[28], we provide full-length transcript models for 246 transcribed loci; the remainder were expressed at levels beneath thresholds used in this study. This is not surprising, the expression patterns of lncRNAs are more restricted than those of protein-coding genes: the average lncRNA is expressed (BPKM >1) in 1.5 developmental and 3.2 tissue samples, compared to 6.6 and 17 for protein-coding genes, respectively. Many lncRNAs (563 or 30%) have peak expression in testes, and 125 are detectable only in testes. Similarly restricted expression patterns have been reported for lncRNAs in humans and other mammals[30,31].

Interestingly, all newly annotated genes overlapping molecularly defined mutations with phenotypes are lncRNAs (Supplementary Table 2). For instance, the mutation D114.3 is a regulatory allele of *spineless (ss)* that maps 4 kb upstream of *ss*[32] and within the promoter of *Mgn4221*. Similarly, *Mgn00541* corresponds to a described, but unannotated 2.0 kb transcript overlapping the regulatory mutant allele *ci[57]* of *cubitus interruptus*[33]. It remains to be determined whether these mutations are a result of the loss of function of newly annotated transcripts or *cis*-acting regulatory elements (e.g. enhancers) or both.

Antisense transcription

Drosophila antisense transcription has been reported[34], but the catalog of antisense transcription has been largely limited to mRNA-mRNA overlaps. We identify non-coding antisense transcript models for 402 lncRNA loci that are antisense to mRNA transcripts of 422 protein-coding genes (e.g. *prd*, Fig. 5a), and 36 lncRNAs form “sense-antisense gene-chains” overlapping more than one protein-coding locus, as observed in mammals[30,35]. In *Drosophila*, 21% of lncRNAs are antisense to mRNAs, whereas in human 15% of annotated lncRNAs are antisense to mRNAs (GENCODE v10). We assembled antisense transcript models for 5057 genes (29%, compared to previous estimates of 15%[34]). For 67% of these loci, antisense expression is observable in at least one cell line, indicating that sense/antisense transcripts may be present in the same cells. LncRNA-mediated antisense accounts for a small minority of antisense transcription – 94% of antisense loci correspond to overlapping protein-coding mRNAs transcribed on opposite strands, and of these, 323 loci (667 genes) share overlapping CDSs. The majority of antisense is due to overlapping UTRs: 1389 genes have overlapping 5'UTRs (divergent transcription), 3430 have overlapping 3'UTRs (convergent transcription), and 540 have both, meaning that, as with many lncRNAs, they form gene-chains across contiguously transcribed regions. A subset of antisense gene-pairs overlap almost completely (>90%), which we term reciprocal transcription. There are 13 such loci (Supplementary Fig. 5) and seven are male-specific (none are female-specific).

The mRNA/lncRNA sense-antisense pairs tend to be more positively correlated in their expression than mRNA/mRNA pairs, (mean r~0.16 vs. 0.13, KS 2-sample one-sided test p<1e–9), and while this mean effect is subtle, the trend is clearly visible in the quantiles (95th% lnc/mRNA 0.729 vs. m/mRNA 0.634, Supplementary Fig. 6a). This effect is stronger when the analysis is restricted to cell line samples (Supplementary Fig. 6b).

Even in homogenous cell cultures, evidence for sense-antisense transcription does not guarantee that both transcripts exist within individual cells: transcription could originate from exclusive events occurring in different cells. Cis-natural antisense transcripts (cis-

NATs) are a substantial source of endogenous siRNAs[36], and their existence directly reflects the existence of precursor dsRNA. Cis-NAT-siRNA production typically involves convergent transcription units that overlap on their 3' ends, but other documented loci generate siRNAs across internal exons, introns or 5'UTRs[37,38,39]. Analysis of head, ovary and testis RNAs showed that 328 unique sense/antisense gene pair regions generated 21nt RNAs indicative of siRNA production (Supplementary Table 8), and these were significantly enriched (Supplementary Figure 7a, Supplementary Methods) for pairs showing positively correlated expression between sense and antisense levels across tissues (*p*~2e–5), embryo developmental stages (*p*~4e–3), conditions (*p*~9e–4), and across all samples (*p*~3e–5). The tissue distribution of these cis-NAT-siRNAs showed a bias for testis expression (Supplementary Fig. 7b), with 4-fold greater number relative to ovaries (*p*~2e–17, binomial test) and 7-fold relative to heads (*p*~4e–25) and expression levels of siRNAs were substantially higher in testes than other tissues (Supplementary Fig. 7c).

Over 80% of cis-NAT-siRNAs were derived from 3'-convergent gene pairs. Abundant siRNAs emanate from an overlap of the *gryzun* and *CG14967* 3'UTRs (Supplementary Fig. 5). The remainders were distributed amongst CDSs, introns, and 5'UTRs. We identified abundant, testis-enriched, siRNA production from a 5'-divergent overlap of *Cyt-c-d* and *CG31808* (Fig. 5b) and from the entire CDS of *dUTPase* and its antisense noncoding transcript *Mgn99994*.

Environmental stress reveals new genes, transcripts and common response pathways

Whole-animal perturbations each exhibited condition-specific effects, e.g. the metallothionein genes were induced by heavy metals (Fig. 6a), but not by other treatments (Supplementary Table 9). The genome-wide transcriptional response to cadmium (Cd) exposure involves small changes in expression level at thousands of genes (48 hours after exposure), but only a small group of genes change >20-fold, and this group includes six lncRNAs (the third most strongly induced gene is *CR44138*, Fig. 6a, Supplementary Fig. 8a). Four newly modeled lncRNAs are differentially expressed (1% FDR) in at least one treatment, and constitute novel eco-responsive genes. Furthermore, 57 genes and 5259 transcripts (of 811 genes) were detected exclusively in these treatment samples. Although no two perturbations revealed identical transcriptional landscapes, we find a homogeneous response to environmental stressors (Fig. 6b, Supplementary Fig. 8b). The direction of regulation for most genes is consistent across all treatments; very few are up-regulated in one condition and down-regulated in another. Classes of strongly up-regulated genes included those annotated with the GO term “Response to Stimulus, GO:0050896” (most enriched, *p*-value<1e–16, Supplementary Fig. 8c), and those that encode lysozymes (>10-fold), cytochrome P450s, and mitochrondrial components mt:ATPase6, mt:CoI, mt:CoIII (>5-fold). Genes encoding egg-shell, yolk, and seminal fluid proteins are strongly down-regulated in response to every treatment except “Cold2” and “Heat Shock” (Supplementary Fig. 8d). For these two stressors, samples were collected 30 minutes after exposure, corresponding to an “early response test” showing suppression of germ cell production is not immediate.

DISCUSSION

The majority of transcriptional complexity in *Drosophila* occurs in tissues of the nervous system, and particularly in the functionally differentiating central and peripheral nervous systems. A subset of ultra-complex genes encodes more than half of detected transcript isoforms and these are dramatically enriched for RNA editing events and 3'UTR extensions, both phenomena largely specific to the nervous system. Our study indicates that the total information output of an animal transcriptome may be heavily weighted by the needs of the developing nervous system.

The improved depth of sampling and spatiotemporal resolution resulted in the identification of more than 1200 new genes not discovered in our previous study of *Drosophila* development[6]. A large fraction of the new genes are testes-specific, and many of these are antisense RNAs, as previously described in mammals[30]. Some new lncRNAs, such as *Mgn94020* (Fig. 1), form sense/antisense gene-chains that bring distant protein-coding genes into transcriptional relationships, another phenomenon previously described only in mammals[40]. Whenever *Mgn94020* is detectably transcribed, the genes on the opposite strand in its introns are not, suggesting that its transcription may serve a regulatory function independent of the RNA transcribed. The presence of short RNAs at many regions of antisense transcription indicates that sense and antisense transcripts are present in the same cells at the same times. Many of these *Drosophila* antisense transcripts correspond to “positionally equivalent”[30] antisense transcripts in human. In the two species we found antisense lncRNAs opposite to orthologous protein-coding genes. The apparent positional equivalence of fly and human antisense transcription at genes like *Monocarboxylate transporter 1* (*Mct1*), *even-skipped* (*EVX1*), *CTCF* (*CTCF*), *Adenosine receptor* (*ADORA2A*), and many others[10,31] across 600 million years of evolution suggests a conserved regulatory mechanism basal to sexual reproduction in metazoans.

Perturbation experiments identified new genes and transcripts, but perhaps more importantly, a general response to stress that is broader than the heat shock pathway. A similar study conducted on marsh fishes in the wake of the Deep Water Horizon incident in the Gulf of Mexico[41] demonstrated that the killifish response to chronic hydrocarbon exposure included induction of lyzosome genes, P450 cytochromes, and mitochondrial components, and the down-regulation of genes encoding egg-shell and yolk proteins[41]. This overlap of expressional responses by gene families across phyla suggests a conserved metazoan stress response involving enhanced metabolism and the suppression of genes involved in reproduction.

We defined an extensive catalog of putative lncRNAs. However, many genes are known to encode poorly conserved, short polypeptides, including genes specific to the male gonad and accessory gland. Analysis of ribosome profiling initially indicated that a number of mammalian lncRNAs may be translated[43], but this observation has been difficult to validate by proteomics[44], and further analysis has suggested that although lncRNAs have signatures of ribosome occupancy, they are not translated[45]. Therefore, while we refer to these RNAs as “non-coding”, additional data are needed to determine if they produce small polypeptides.

Our observations raise many questions. Why do genes encoding RNA binding proteins exhibit extraordinary splicing complexity, often within their 5'UTRs? The splicing factor *pUf68* encodes more than 100 alternatively spliced 5'UTR variants, but encodes a single protein. The notion that splicing factors may regulate one another to generate complex patterns of splicing is consistent with recent computational models[45]. What is the role of complex splicing during the development of the nervous system? To answer the questions that come with increasingly complete transcriptomes in higher organisms, it will be necessary to study gene regulation downstream of transcription initiation, including the regulation of splicing, localization and translation.

METHODS SUMMARY

Animal Staging, Collection and RNA extraction

Tissues were dissected from Oregon R larval, pupal and adult staged animals synchronized with appropriate age indicators. Pupal and adult animals were treated with a number of environmental stresses. RNA was isolated using TRIzol (Invitrogen), DNased, and purified on a RNAeasy column (Qiagen). poly(A)+ RNA was prepared from an aliquot of each total RNA sample using an Oligotex kit (Qiagen).

RNA-Seq

Libraries were generated and sequenced on an Illumina Genome Analyzer IIx or HiSeq 2000 using paired-end chemistry and 76 or 100bp cycles. 454 sequencing used poly(A)+ RNA from Oregon R adult males and females and mixed-staged y^{1} cn^{1} bw^{1} sp^{1} embryos. Sequences are available from the Short Read Archive and the modENCODE website (http://www.modencode.org/). CAGE[46] was sequenced on a Illumina Genome Analyzer IIx with 36bp reads. Poly(A)+seq was generated using a custom protocol (Supplementary Methods).

Analysis

RNA-seq, CAGE, and polyA+ reads were mapped and filtered[12]. GRIT was used to identify transcript models[14]. Expression levels for genes and exons were computed in BPKM[6]. GSC *p*-values were computed[47]. values were calculated with MISO[48]. Differential expression analysis conducted with a custom method (Supplementary Methods) and with DEseq[49]. RPS-BLAST was used to conduct the conserved domain search with version v3.08 of the NCBI CDD (Supplementary Methods). Orthology analysis between human and fly was conducted using DIOPT (http://www.flyrnai.org/cgi-bin/DRSC_orthologs.pl). Phenotypic alleles were downloaded from FlyBase r5.50, and were selected as any allele localized to the genome with a disease phenotype.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank the members of the modENCODE transcription consortium, especially Jane Landolin and Jeremy Sandler for their early contributions to these studies. We also thank Anshul Kundaje and Haiyan Huang for helpful discussions. This work was funded by a contract from the National Human Genome Research Institute modENCODE Project, contract U01 HG004271 and U54 HG006944, to S.E.C. (Principal Investigator) and P.C., T.R.G., R.A.H., and B.R.G. (co-Principal Investigators) with additional support from R01 GM076655 (S.E.C.) both under Department of Energy contract no. DE-AC02-05CH11231. J.B.B.'s work was supported by NHGRI K99 HG006698. Work in P.J.B.'s group was supported by the modENCODE DAC sub award 5710003102, 1U01HG007031-01 and the ENCODE DAC 5U01HG004695-04. Work in Bloomington was supported in part by the Indiana METACyt Initiative of Indiana University, funded by an award from the Lilly Endowment, Inc. Work in E.C.L.'s group was supported by U01-HG004261 and RC2-HG005639.

References

1. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008; 5:621–628. [PubMed: 18516045]
2. Nagalakshmi U, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008; 320:1344–1349. [PubMed: 18451266]
3. Takahashi H, Kato S, Murata M, Carninci P. CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks. Methods Mol Biol. 2012; 786:181–200. doi: 10.1007/978-1-61779-292-2_11. [PubMed: 21938627]
4. Mangone M, et al. The landscape of C. elegans 3'UTRs. Science. 2010; 329:432–435. [PubMed: 20522740]
5. Jan CH, Friedman RC, Ruby JG, Bartel DP. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. Nature. 2011; 469:97–101. [PubMed: 21085120]
6. Graveley BR, et al. The developmental transcriptome of Drosophila melanogaster. Nature. 2011; 471:473–479. [PubMed: 21179090]
7. Trapnell C, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012; 7:562–578. [PubMed: 22383036]
8. Collins JE, White S, Searle SM, Stemple DL. Incorporating RNA-seq data into the zebrafish Ensembl genebuild. Genome Res. 2012; 22:2067–2078. [PubMed: 22798491]
9. Carninci P, et al. Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. Genome Res. 2003; 13:1273–1289. [PubMed: 12819125]
10. Harrow J, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012; 22:1760–1774. [PubMed: 22955987]
11. Djebali S, et al. Landscape of transcription in human cells. Nature. 2012; 489:101–108. [PubMed: 22955620]
12. Hoskins RA, et al. Genome-wide analysis of promoter architecture in Drosophila melanogaster. Genome Res. 2011; 21:182–192. [PubMed: 21177961]
13. Cherbas L. The Transcriptional Diversity of 25 Drosophila Cell Lines. Genome Res. 2010
14. Boley N, et al. Genome guided transcript construction from integrative analysis of RNA sequence data. Nature Biotechnology. 2013
15. Celniker SE, Rubin GM. The Drosophila melanogaster genome. Annu Rev Genomics Hum Genet. 2003; 4:89–117. [PubMed: 14527298]
16. Stapleton M, et al. The Drosophila Gene Collection: Identification of Putative Full-Length cDNAs for 70% of D. melanogaster Genes. Genome Res. 2002; 12:1294–1300. [PubMed: 12176937]
17. Spradling AC, et al. The Berkeley Drosophila Genome Project gene disruption project: Single P-element insertions mutating 25% of vital Drosophila genes. Genetics. 1999; 153:135–177. [PubMed: 10471706]
18. Wang ET, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008; 456:470–476. [PubMed: 18978772]
19. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008; 40:1413–1415. [PubMed: 18978789]
20. Smibert P, et al. Global patterns of tissue-specific alternative polyadenylation in Drosophila. Cell Rep. 2012; 1:277–289. doi:10.1016/j.celrep.2012.01.001. [PubMed: 22685694]
21. St Laurent G, et al. Genome-wide analysis of A-to-I RNA editing by single-molecule sequencing in Drosophila. Nat Struct Mol Biol. 2013

22. Telonis-Scott M, Kopp A, Wayne ML, Nuzhdin SV, McIntyre LM. Sex-specific splicing in Drosophila: widespread occurrence, tissue specificity and evolutionary conservation. Genetics. 2009; 181:421–434. [PubMed: 19015538]

23. Hartmann B, et al. Distinct regulatory programs establish widespread sex-specific alternative splicing in Drosophila melanogaster. RNA. 2011; 17:453–468. [PubMed: 21233220]

24. Chang PL, Dunham JP, Nuzhdin SV, Arbeitman MN. Somatic sex-specific transcriptome differences in Drosophila revealed by whole transcriptome sequencing. BMC Genomics. 2011; 12:364. [PubMed: 21756339]

25. Matthews L, et al. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res. 2009; 37:D619–622. [PubMed: 18981052]

26. Lipshitz HD, Peattie DA, Hogness DS. Novel transcripts from the *Ultrabithorax* domain of the Bithorax Complex. Genes and Development. 1987; 1:307–322. [PubMed: 3119423]

27. Tupy JL, et al. Identification of putative noncoding polyadenylated transcripts in Drosophila melanogaster. Proc Natl Acad Sci U S A. 2005; 102:5495–5500. [PubMed: 15809421]

28. Young RS, et al. Identification and properties of 1,119 candidate lincRNA loci in the Drosophila melanogaster genome. Genome Biol Evol. 2012; 4:427–442. [PubMed: 22403033]

29. Kondo T, et al. Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. Nat Cell Biol. 2007; 9:660–665. [PubMed: 17486114]

30. Katayama S, et al. Antisense transcription in the mammalian transcriptome. Science. 2005; 309:1564–1566. [PubMed: 16141073]

31. Derrien T, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 2012; 22:1775–1789. [PubMed: 22955988]

32. Duncan DM, Burgess EA, Duncan I. Control of distal antennal identity and tarsal development in Drosophila by spineless-aristapedia, a homolog of the mammalian dioxin receptor. Genes Dev. 1998; 12:1290–1303. [PubMed: 9573046]

33. Schwartz C, Locke J, Nishida C, Kornberg TB. Analysis of cubitus interruptus regulation in Drosophila embryos and imaginal disks. Development. 1995; 121:1625–1635. [PubMed: 7600980]

34. Misra S, et al. Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. Genome Biology. 2002; 3 research0083.

35. Lipovich L, et al. Activity-dependent human brain coding/noncoding gene regulatory networks. Genetics. 2012; 192:1133–1148. [PubMed: 22960213]

36. Okamura K, Lai EC. Endogenous small interfering RNAs in animals. Nat Rev Mol Cell Biol. 2008; 9:673–678. [PubMed: 18719707]

37. Okamura K, Balla S, Martin R, Liu N, Lai EC. Two distinct mechanisms generate endogenous siRNAs from bidirectional transcription in Drosophila melanogaster. Nat Struct Mol Biol. 2008; 15:581–590. [PubMed: 18500351]

38. Czech B, et al. An endogenous small interfering RNA pathway in Drosophila. Nature. 2008; 453:798–802. [PubMed: 18463631]

39. Ghildiyal M, et al. Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. Science. 2008; 320:1077–1081. [PubMed: 18403677]

40. Engstrom PG, et al. Complex Loci in human and mouse genomes. PLoS Genet. 2006; 2:e47. doi: 10.1371/journal.pgen.0020047. [PubMed: 16683030]

41. Whitehead A, et al. Genomic and physiological footprint of the Deepwater Horizon oil spill on resident marsh fishes. Proc Natl Acad Sci U S A. 2012; 109:20298–20302. [PubMed: 21949382]

42. Ingolia NT, Ghaemmaghami S, et al. Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009; 324:218–223. [PubMed: 19213877]

43. Banfai B, et al. Long noncoding RNAs are rarely translated in two human cell lines. Genome Res. 2012; 22:1646–1657. [PubMed: 22955977]

44. Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell. 2013; 154:240–251. [PubMed: 23810193]

45. Kosti I, Radivojac P, Mandel-Gutfreund Y. An integrated regulatory network reveals pervasive cross-regulation among transcription and splicing factors. PLoS Comput Biol. 2012; 8

46. Takahashi H, Lassmann T, Murata M, Carninci P. 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. Nat Protoc. 2012; 7:542–561. [PubMed: 22362160]

47. Bickel PJ, Boley N, Brown JB, Huang H, Zhang NR. Subsampling methods for genomic inference. Ann. Appl. Stat. 2010; 4:1660–1697.

48. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods. 2010; 7:1009–1015. [PubMed: 21057496]

49. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010; 11:R106. [PubMed: 20979621]

50. Yepiskoposyan H, et al. Transcriptome response to heavy metal stress in Drosophila reveals a new zinc transporter that confers resistance to zinc. Nucleic Acids Res. 2006; 34:4866–4877. [PubMed: 16973896]

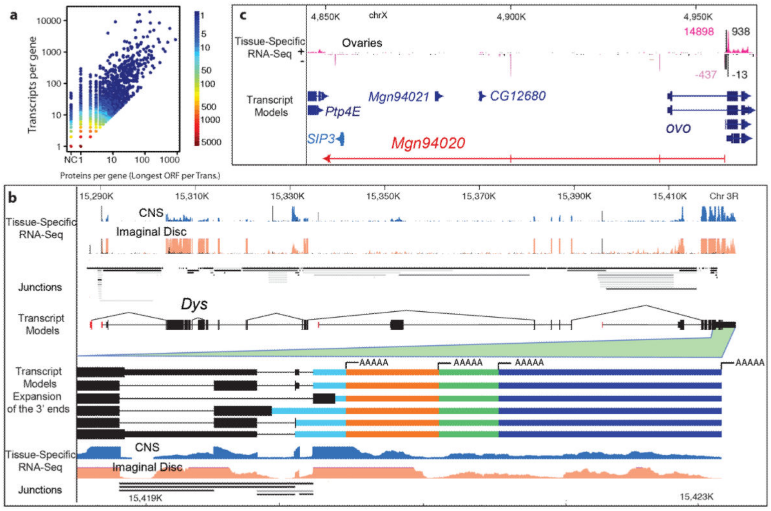

Figure 1. Overview of the annotation

a, Scatterplot showing the per gene correlation between number of proteins and number of transcripts. The genes *Dscam* and *para* are omitted as extreme outliers both encoding >10,000 unique proteins. **b**, *Dystrophin (Dys)* produces 72 transcripts and encodes 32 proteins. Highlighted is alternative splicing and polyadenylation at the 3' end. Shown: CAGE (black), RNA-seq (tan, blue), splice junctions (shaded gray as a function of usage). **c**, An internal promoter of *ovo* is bidirectional in ovaries and produces a lncRNA (430bp, red) bridging two gene deserts. CAGE (black), RNA-seq (pink), counts are read-depth (minus-strand given as negative).

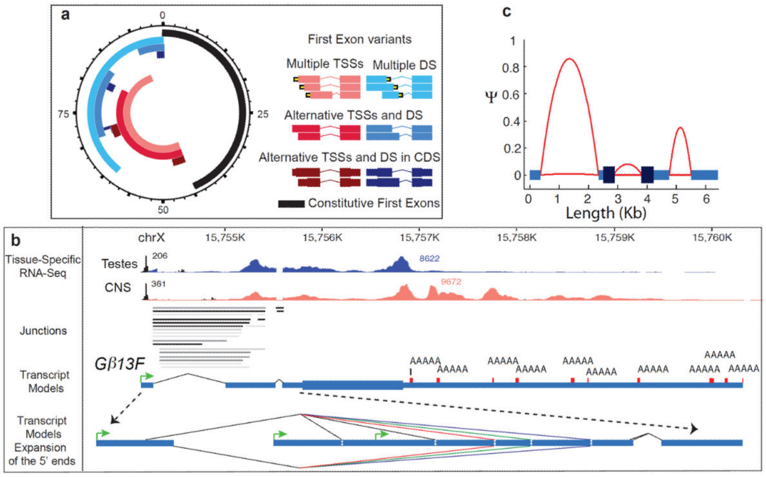

Figure 2. Splicing complexity across the gene body

a, Alternative first exons occur in two main configurations: multiple transcription start sites (TSS, pink) and multiple donor sites (DS, light blue). A subset of the genes in the multiple TSS category produce transcripts with different TSSs and shared DSs (red), and a subset of the genes in the DS category produce transcripts with a shared TSS and different DSs (blue). Some genes in the multiple TSS category directly affect the encoded protein (maroon), and similarly for DS (dark blue). Overlap of configurations is radially proportional (units indicate percentage of all spliced genes). **b**, Poly(A)+ testes (blue) and CNS (orange) stranded RNA-seq of *G*β*13F* showing complex processing and splicing of the 5'UTR. An expansion of the 5'UTR showing some of the complexity. Transcription of the gene initiates from one of three different promoters (green arrows) terminates at one of ten possible polyA + addition sites (from adult head poly(A)+seq, red) and generates 235 transcripts. The first exon has two alternative splice acceptors that splice to one of eleven different donor sites. Only five donor sites are shown due to the proximity of splice sites. Four splice donors are represented by the single red line differing by 12, 5 and 19bp respectively. Three splice donors are represented by the single green line differing by 12 and 11bp. Two splice donors are represented by the single purple line differing by 7bp. These splice variants are combined with four proximal internal splices (Supplementary Fig. 3a) to generate the full complement of transcripts. **c**, Intron retention rates (ψ) across the gene body. The genome-wide mean lengths of exons and introns are connected by red parabolic arcs, which illustrate the upper and lower quartiles of intron retention (across all samples) for introns retained at or above 20 ψ in at least one sample.

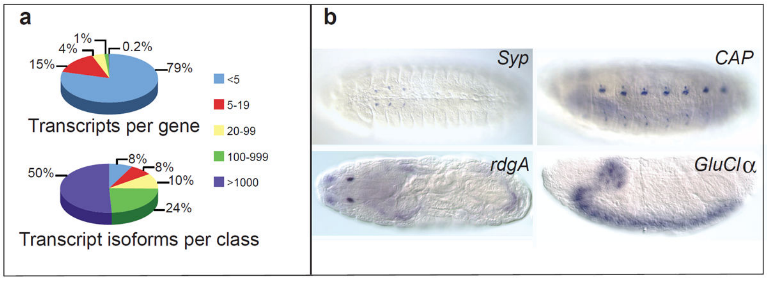

Figure 3. Complex splicing patterns are largely limited to neural tissues

a, A small minority of genes (47, 0.2%) encode the majority of transcripts. **b**, *In situ* RNA staining of constitutive exons of four genes with highly complex splicing patterns in the embryo. Syncrip (Syp), Cap, Retinal degeneration A (rdgA) and GluClalpha show specific late embryonic neural expression in the ventral midline neurons; dorsal/lateral and ventral sensory complexes; Bolwig's organ or larval eye; and central nervous system respectively.

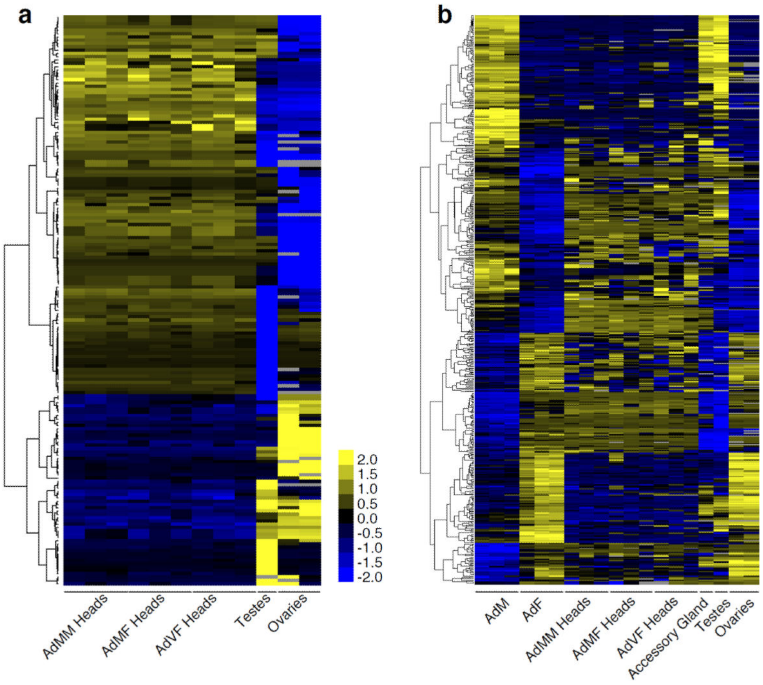

Figure 4. Sex-specific splicing is largely tissue-specific splicing

a, Clusters of tissue-specific splicing events. The scale bar indicates *z*-scores of ψ. **b**, Sex-specific splicing events in whole animals are primarily testes- or ovary-specific splicing events.

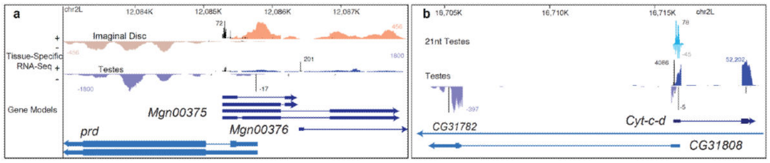

Figure 5. Examples of antisense transcription

a, 5'/5' bidirectional antisense transcription at the *prd* locus. Short RNA sequencing does not reveal substantial siRNA (i.e. 21 nt-dominant small RNA) signal in this region (data not shown). **b**, A 5'/5' antisense region that produces substantial small RNA signal on both strands.

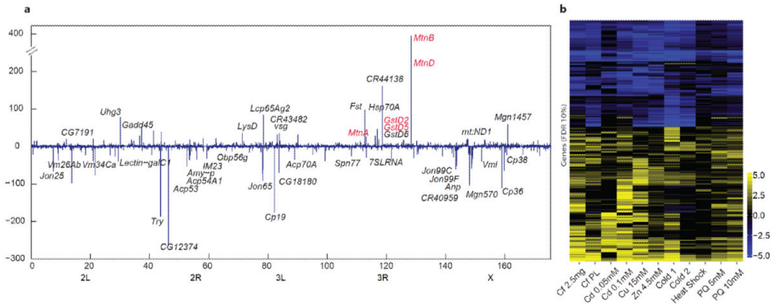

Figure 6. Effects of environmental perturbations on the *Drosophila* transcriptome

Adults were treated with caffeine (Cf), Cd, Cu, Zn, cold, heat, and paraquat (PQ). **a**, A genome-wide map of genes that are up or down regulated as a function of Cd treatment. Labeled genes are those that showed a 20-fold (<10% FDR) change in response (linear scale). Genes highlighted in red are those identified in larvae[50]. Some genes are omitted for readability, the complete figure and list of omitted genes are given in Supplementary Fig. 8a. **b**, Heat map showing the fold change of genes with an FDR<10% (differential expression) in at least one sample (log2 scale).