Biological Networks: Dynamics, Mechanisms and Responses

by

Marcus Hudak Stoiber

A dissertation submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Biostatistics

and the Designated Emphasis

in

Computational and Genomic Biology

in

the Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Peter J Bickel, Chair

Professor Sandrine Dudoit

Professor Steven E Brenner

Spring 2015

Abstract

Biological Networks: Dynamics, Mechanisms and Responses

by

Marcus Hudak Stoiber

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Peter Bickel, Chair

The study of biological networks has been central to our understanding of life and its complex, dynamic nature. The elucidation of molecular networks began with the discovery and characterization of key cellular processes including metabolism, response to stimuli and control of gene expression. In the last several decades, genomics has emerged as a foundational pursuit within the life sciences. The size of datasets defined in relation to sequenced genomes has grown faster than exponentially, leading to the need for advanced analytical and computational methods. I present here three studies of large RNA-sequencing-based data sets. First, a study of the steady state transcriptional composition for *Drosophila* cell lines, tissues, developmental stages and biological perturbations provide a deeper understanding of spatiotemporally-resolved regulation in *Drosophila*, the first and still central genetic model system. This dataset, at the time of my analysis, was the largest and most complete transcriptional atlas ever composed. It was also the first large strand-specific study of its kind, which presented new opportunities and challenges. Second, a study of the RNA targets of 20 RNA binding proteins provides a map for one layer of post-transcriptional regulation, which contributes to the steady states presented in the first study. Finally, a study of transcriptional responses to the principle developmental hormone in arthropods, ecdysone, across 41 different and physiologically distinct cell lines sheds light on the dynamic, responsive nature of gene-regulatory networks that enable cells to differentiate into the diverse tissues that compose developing and mature organisms. These studies provide foundational knowledge, as well as models for future work in systems biology, as genome-scale studies across larger, more diverse cellular states become increasingly prevalent.

Dedication

To my loving wife, better half and best friend, Alyssa

To my parents, Maggie and Jeff, and my siblings, Julia, Luke and Rachel

To the entire extended Stoiber, Hudak and Smith clans

I could not have accomplished what I have with out all of your enduring support. I love you all so much!

Table of Contents

List of Figures

List of Tables

Introduction

The fields of genetics has a rich history dating back to Mendel's discovery of "units of heredity" in 1865[3]. Major advancements made throughout the 1800's and 1900's led to our current understanding of genetics with a major shift occurring in 1975 when Sanger et al[4] as well as Maxam and Gilbert[5] in 1977 invented a process for DNA sequencing. This technology allowed the interrogation of the exact content of the heritable unit in all organisms, giving rise, over the following decades, to genomics. With the advent of next-generation sequencing technologies throughout the last decade, interrogation of biological networks on a genome-wide scale has become increasingly accessible to the broad scientific community. Now the genomes are interrogated at levels of epigenetics, transcriptional output, three dimensional conformation, and dynamics, to name a few of the many descendant pursuits enabled by high-throughput DNA sequencing technology.

The fruit fly, *Drosophila melanogaster*, provides a tremendously powerful system to further our understanding of genomic biology. The fruit fly has been extensively studied throughout a century of genetic experiments[6]. This enables the rich characterization of discoveries made from the application of genomic and epigenomic profiling technologies in terms of prior genetic knowledge. This strong foundation leads to biological interpretations and functional characterizations of biological networks discovered using next-generation sequencing technologies that are not possible in other metazoan systems. Here I will present three studies leveraging RNA sequencing (RNA-seq) in *Drosophila melanogaster* that each produced novel insights into the regulation and dynamics of gene expression.

In chapter one, I describe the analysis of the largest transcriptional survey of a metazoan[7]. The modENCODE consortium produced 126 RNA samples derived from developmental stages[8], dissected tissues, cell lines and environmental perturbations. This data allows for the study of transcriptional output of *Drosophila* in a manner unbiased compared to studies focused on one or a few single genes. Analysis begins with the annotation of genic regions from the sequencing samples using the GRIT algorithm[9]. Using the GRIT annotation and the diverse survey of expression measurements, I explore the complexity of transcripts produced through the combinatorial usage of alternative transcription start sites, alternative splicing, and polyadenylation sites; the dynamic expression of long non-coding RNAs; and the transcriptional response to environmental perturbations.

In chapter two, I describe the analysis of a 20 RNA binding proteins (RBPs) interrogated using RNA Immunoprecipitation followed by sequencing (RIP-seq), which identifies RNAs bound by bait RNA Binding Proteins (RBP). This study is the largest study of RBP targets at genome-scale and reveals broad characteristics of the post-transcriptional network, as well as insights into the biology of ribonucleoproteins. I find that this network contains high occupancy target (HOT) RNAs, which are enriched for many genes with post-transcriptional regulatory functions. Via integrative analysis with an RNAi screen, I find that

target RNAs tend to undergo alternative splicing when the RBPs that bind them are knocked down. This indicates that significant subsets of the binding events I detected are functional. RBPs tend to bind mRNA and protein products from the same gene, and hence may post transcriptionally regulate their protein interaction partners. Additional characteristics of the RBP network include ribonuclear complexes including non-coding RNAs, specifically bound 3' UTRs, insight into the RNA processing of mRNAs from ultra-complex genes, and enriched intronic binding a subset of RBPs. I developed an algorithm to discover motifs from RIP-seq bound gene sets. I show that most motifs correspond to previously discovered *in vitro* motifs, validating the method, and identify several striking exceptions. This study provides the most comprehensive view of the network of RNA-binding proteins to date, including key factors in the post-transcriptional regulatory machinery.

In chapter three, I describe the transcriptional response of 41 diverse *Drosophila melanogaster* cell lines to the key developmental steroid hormone 20-hydroxyecdysone (20E), ecdysone. Ecdysone triggers many of the major transitional events in *Drosophila* development including molting events. Leveraging the diverse cell states present across 41 cell lines derived from embryonic, larval, and ovarian tissues, we observe a diverse response both in specific transcriptionally responsive genes as well as the total count of genes responsive in a given cell line, which we denote the responsive gene count (RGC). The ecdysone receptor (*EcR*), part of the heterodimer that transduces 20E signaling, shows the strongest correlation with RGC, indicating that *EcR* titer is rate limiting in the ecdysone response. Additionally, the alternative isoforms appear to alternatively regulate essential ecdysone responsive genes. In order to gain a predictive understanding of the diversity of ecdysone response across cellular states, I show that transcription factor (TF) expression along with known TF binding motifs combine to produce significant power to predict the specific set of genes induced in a given cell. However, by modeling the strength of prediction as a function of cell-states surveyed, I show that transcriptional data alone are unlikely to be sufficient to fully elucidate the system. Finally, I analyze an extended time course for three cell lines and show that the response to ecdysone is consistent with a kinetic model of response.

Together these studies provide numerous biological insights and their undertaking has led me to produce several novel analysis pipelines that enable the interpretation of biological "big data". As such data sets become more common with the dropping price of sequencing and development of new high-throughput technologies, the thoughtful analysis of biological big data will become increasingly important. The interpretation of biological data should, in the spirit of descriptive applied statistics in general, provide insightful summaries of vast arrays of data in small collections of useful and easily comprehensible statistics. Interpretation should also enhance our ability to predict and model the system under study, and yield generalizable rules to larger biological contexts. The analyses presented here accomplish these goals and provides hypotheses for future genetic and genomic studies to further our molecular understanding of metazoans.

Acknowledgements

Over the past three years I have received a magnificent amount of support from many individuals. Dr. James “Ben” Brown has been an incredible mentor from before my time as a PhD student up until the present time and has helped me to shape my scientific skills for which I will be forever grateful. I look forward to countless more scientific adventures with Ben. Dr. Peter Bickel, Dr. Haiyan Huang and their entire research group has provided incredible insight into all of my graduate endeavors and the inconceivable intelligence and insight coming from their meetings and talks have shaped the way I understand and continue to learn about statistics and genomics. Dr. Susan Celniker and her research group have provided amazing support and her encyclopedic knowledge of the fruit fly continues to astound me and help to direct my studies in the ever-changing field of model organism genomics especially in the modern age of high-throughput genomics.

I would like to extend a special thanks to my qualifying exam and dissertation committee members, Dr. Peter Bickel, Dr. Sandrine Dudoit, Dr. Steven Brenner and Dr. Elizabeth Purdom, for their assistance and keen interest in my work. Their efforts and time are invaluable in my pursuit of academic and scientific goals.

As a genomic statistician/computational biologist, I must extend a deep debt of gratitude for the collaborators who have provided incredible data leading to the analysis products presented here. In particular, I would like to thank Dr. Peter Cherbas, Dr. Lucy Cherbas, Dr. Brenton R. Graveley, Dr. Spyros Artavanis-Tsakonas as well as the entire ENCODE and modENCODE consortiums.

In starting my genomics career I would also like to thank Dr. Elliott Margulies for giving me my first opportunity in the field of computational biology and genomics and for the many lessons as a beginning student. Those experiences truly shaped the scientist that I am today and I would not have made it this far without such a strong mentor early in my career.

Chapter 1: Statistical Genomics Analysis of an Extensive *Drosophila* Transcriptome Survey

Preface

The contents of this chapter have been adapted from the previously published paper “Diversity and Dynamics of the Drosophila Transcriptome” with permission from primary contributing authors. The contents included here represent analyses conducted by myself.

During 2009-2012, the modENCODE Consortium compiled the largest transcriptional atlas generated for any metazoan organism. The LifeMap included RNA-seq data derived from a developmental time course including 30 life stages[8], 29 dissected tissues, 25 cell lines, and 21 environmental perturbations designed to reveal stress-responsive and adaptive genes not expressed under laboratory conditions[7]. A genome-wide annotation was generated using the GRIT algorithm[9] taking RNA-seq, p(A)+seq, CAGE, RACE[10], ESTs[11], and full-length cDNAs[12] as input. I conducted genomic and statistical analyses of these data and the GRIT annotation to elucidate the biological insights revealed by this extensive life map.

Introduction

Next-generation RNA sequencing has permitted the mapping of transcribed regions of the genomes of a growing variety of organisms[13,14]. These studies have demonstrated that large fractions of metazoan genomes are transcribed and have cataloged the individual elements of transcriptomes, including promoters[15], polyadenylation sites[16,17], exons and introns[8]. However, the complexity of the transcriptome arises from the combinatorial incorporation of these elements into mature transcript isoforms. Studies that have inferred transcript isoforms from short read sequence data have focused on a small subset of isoforms, filtered using stringent criteria[18,19]. Studies that have relied on cDNA or EST data to infer transcript isoforms have not had sufficient depth of sampling to explore the diversity of RNA products at the vast majority of genomic loci[20]. The human genome has been the focus of intensive manual annotation[21], but analysis of strand-specific RNA-seq data from human cell lines reveals over 100,000 splice junctions not incorporated into any transcript model[22]. Hence, a large gap exists between the genome annotations and the emerging picture of transcriptomes observed in next-generation sequence data. Here, we describe a complete transcript set modeled by integrative analysis of promoter data (CAGE and RACE), splice sites and exons (RNA-seq), and polyadenylation sites (poly(A) reads from ESTs and RNA-seq). We analyze RNA from a diverse set of developmental stages, dissected organ systems and environmental perturbations using a strand-specific sequencing strategy. Our data provide higher spatiotemporal resolution and allow for deeper exploration of the *Drosophila* transcriptome than previously possible. Our analysis reveals a transcriptome of unprecedented complexity expressed in discrete, tissue-specific mRNA and ncRNA transcript isoforms that span the majority of the fly genome on both strands and provides valuable insight into metazoan biology.

Results

A Dense Landscape of Discrete poly(A)+ Transcripts

The GRIT annotation[9] consists of 304,788 transcripts and 17,564 genes (Figure 1.1a), of which 14,692 are protein-coding (Methods). 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 (Figure 1.1a, example 1.1b). 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 project[23]. There were 64 euchromatic gene-free regions longer than 50kb in FB5.12, and 25 remaining in FB5.45. The GRIT annotation includes new gene models in eachof these regions. Newly identified genes (1468 total) are expressed in spatially- and temporally-restricted patterns, and 536 reside in

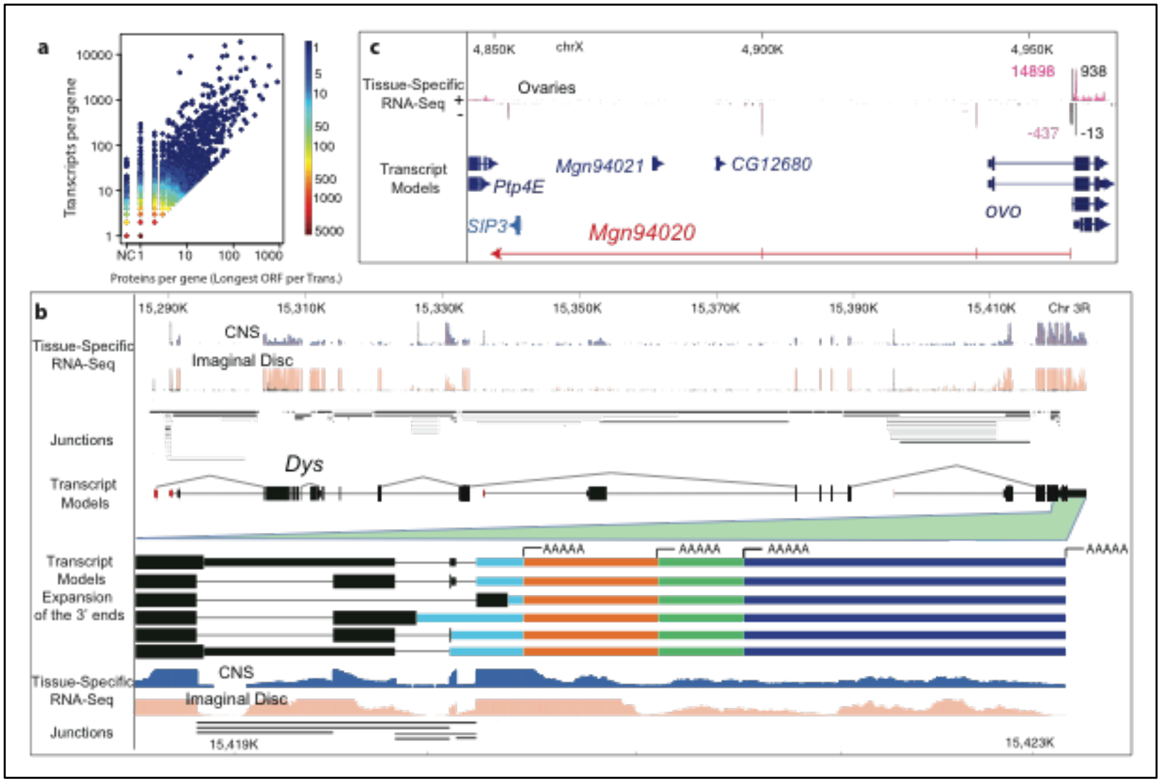


Figure 1.1 Overview of the Annotation

a, Maximum number of transcript isoforms vs. number of unique ORFs per gene (only the longest ORF in each transcript is reported). The genes *Dscam* and *para* are omitted, as they are extreme outliers both encoding more than 10,000 unique ORFs. **b,** A compact visualization of the transcripts encoded by gene *Dys (Dystrophin),* which may encode 72 transcripts and 32 proteins. Highlighted is alternative splicing and polyadenylation at the 3' end of transcripts. **c,** An internal promoter of *ovo* is bidirectional in ovaries and gives rise to a 107kb transcript that bridges two 50kb gene deserts and encodes no long or conserved ORFs. The mature RNA is only 430bp in length.

previously uncharacterized gene-free regions. Others map to well-characterized regions, including the *ovo* locus, where I discovered a new ovary-specific, poly(A)+ transcript (*Mgn94020*), extending from the second promoter of *ovo* on the opposite strand and spanning 107kb (Figure 1.1c). Exons of 36 new genes overlap molecularly defined mutations with associated phenotypes (GSC *p*-value~0.0002), suggesting potential functions. For instance, the lethal P-element insertions *l(3)L3051* and *l(3)L4111*[24] map to promoters of *Mgn095159* and *Mgn95009*, respectively, suggesting these may be essential genes. Nearly 60% of the intergenic transcription reported[8] is now incorporated into gene models.

Transcript Diversity

More than half of protein-coding genes (7940 genes; 54.86%) encode two or more transcript isoforms with alternative first exons (AFEs). There are 31,032 AFEs associated with 5463 genes that reflect alternative and tissue-specific promoter usage but do not affect protein-coding potential. There are 22,530 AFEs of 2477 genes that alter the coding capacity of previously annotated transcripts and increase the complexity of the predicted proteome; the median alteration to the predicted amino acid sequence is 55 N-terminal residues. Spliced genes have an average of 2.6 distinct first exons, but only 1.4 predicted start codons. Alternative pre-mRNA splicing is also enriched near 5’ transcript ends (Figure 1.2a,b). We observe averages of 4.7, 3.7 and 2.4 alternatively spliced isoforms per intron in 5' UTRs, protein coding sequences and 3' UTRs, respectively. We note that splicing in 3' UTRs is comparatively rare: 2765 genes (14% of spliced genes) have spliced 3' UTRs, a value more frequent than seen in gene annotations in mammals, as expected given that nonsense-mediated decay is less influenced by 3' UTR splicing patterns in *Drosophila*[25]. We note that 5' UTR complexity is only weakly correlated with protein-coding sequence complexity: the splice-forms per intron measure yields a log-linear correlation of 0.22 and a Spearman's coefficient of 0.21 (Figure 1.2a).

Genes with novel alternative N-terminal coding sequence include well-studied examples, such as *Prothoracicotropic hormone* (*Ptth*)[26], the neural-secreted hormone that initiates metamorphosis in insects. We find three protein isoforms for *Ptth,* one of which encodes a distinct amino terminus in frame with the conserved hormone domain. Furthermore, a predictive model[27] suggests that this new alternative N-terminal sequences may direct specific subcellular localization of the Ptth protein isoform to the mitochondrial matrix, bypassing its usual pathway to extracellular secretion. This appears to be a representative case of a general phenomenon: 31.22% of alternative start codons are predicted to change sub-cellular localization of the protein, compared to 4.60% of internal cassette exons (*p* < 1e-100 by t-test) and 11.89% of alternative C-terminal coding sequence (*p* < 1e-33 by t-test).

Despite the depth of RNA-seq data (>14B uniquely mapping reads), the data suggest that 59.94% of genes encode no more than a single protein isoform

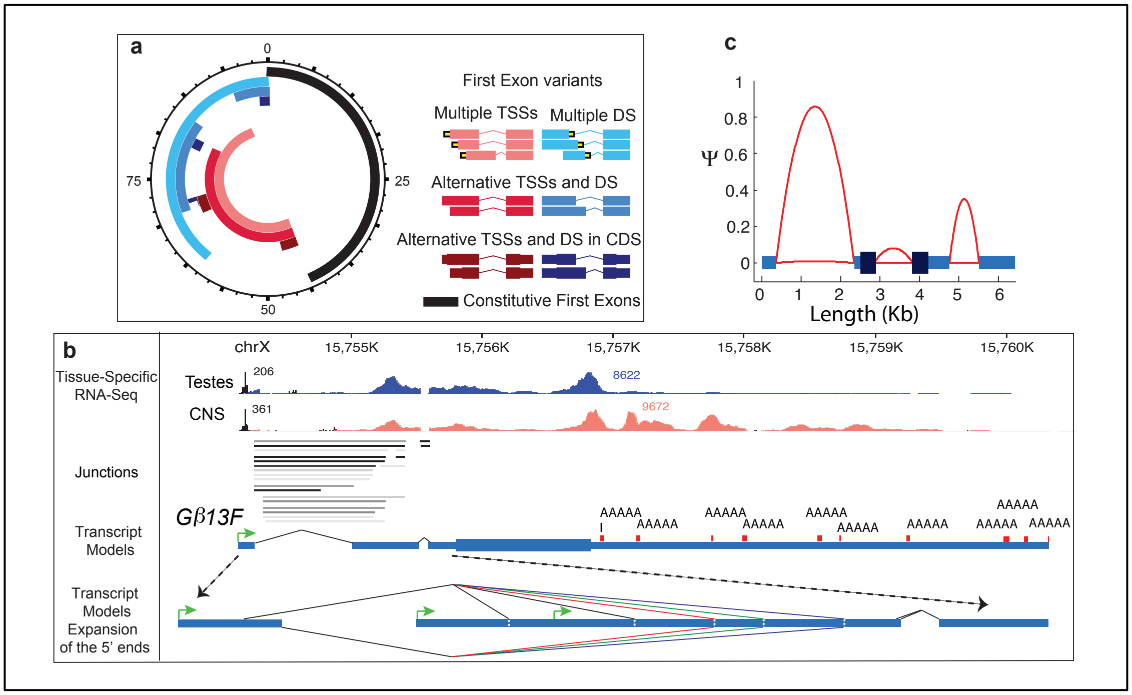


Figure 1.2 Splicing Complexities across the Gene Body
a, Alternative first exons occur in two main configurations: multiple transcription start sites (TSS) and multiple donor sites (DS). A subset of the multiple TSS categories has several TSSs with a shared DS (red transcripts), and similarly for DS (blue transcripts). A further subset of the alternative TSS category directly affect the encoded protein (maroon transcripts), and similarly for DS (dark blue transcripts). Overlap of configurations is radially proportional (units indicate percentage of all spliced genes). b, Complex processing and splicing of the 5' UTR of *Gβ13F*. At the top of the figure are the testes and CNS positively stranded RNA-seq reads followed by the splice junctions (shaded gray as a function of usage), a simplified version of the full-length gene model and 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 and via complex splicing patterns generates 235 transcripts that all produce the same protein. 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 the possible splice sites. Four splice donors are represented by the single red line and map to positions 15,755,148, 67, 72, 84 differing by 12, 5 and 19 bp respectively. Three splice donors are represented by the single green line and map to positions 15,755,256, 68, 79 differing by 12 and 11 bp. Two splice donors are represented by the single purple line 15,755, 409, 16 differing by 7 bp. These splice variants are combined with four different proximal internal splices to generate the full complement of transcripts. Polyadenylation site, shown in red, come from Polya-seq of adult heads. c, Intron retention rates (PSI) across the gene body. Pictured are the genome-wide mean lengths of exons and introns connected by red parabolic arcs. The parabolic arcs illustrate the upper and lower quartiles of intron retention for introns retained at or above 20 PSI in at least one sample, across all samples. We note that, on average, the first intron in the 5' UTR is the longest in a gene, and that splicing in untranslated sequence is less efficient than between CDS exons

(Methods) and 42.29% encode only a single transcript isoform. However, one third of these single protein-encoding genes encode single exon ORFs (19.94% vs. 59.94%). For genes containing ORFs that result from the splicing of multiple exons, 49.11% produce more than one protein isoform. As a point of comparison, in mammals, it has been estimated that 95% of genes produce multiple transcript isoforms[28,29]. There are, however, interesting outliers: seven genes have the potential to produce more than one hundred transcript isoforms, each due to alternative UTR splicing and promoter usage, but encode only a single protein, e.g. *Gbeta13F*. *Dhc98D* has 45 introns, but produces at most two transcripts. Other genes include far fewer splice sites but have the potential to encode hundreds of protein isoforms. For example, *Gug* has 24 introns and may encode up to 170 distinct protein isoforms.

We find that the majority of transcriptome complexity is attributable to a small subset of genes. Forty-seven genes have the capacity to encode more than 1000 transcript isoforms each, and together account for 50% of all transcripts inferred from our data (Figure 1.3). Furthermore, 27% of transcripts encoded by these genes have been detected exclusively in samples enriched for neuronal tissue and another 56% have been detected only in the embryo (83% in total). RNA *in situ* expression assays, for each of these genes in the developing embryo, were conducted, by the Celniker lab, to determine their tissue specificities. I found that 18 out of 33 of these genes expressed in the embryo are detectably expressed only in neural tissue (hypergeometric p-value for enrichment < 1e-16). Hence examining the potential complexity of genes across samples reveals that the capacity to encode more than 1000 transcripts is largely a neuronal phenomenon.

To further characterize genes that express alternatively spliced transcripts, we examined the conserved protein domains encoded by each gene. Among genes with the capacity to produce more than 100 transcripts (304 genes), there are six significantly enriched conserved domains (FDR <0.1%) present in 24 genes. All correspond to RNA binding domains: ELAV/HuD family splicing factor, sex-lethal family splicing factor, glycine-rich RNA-binding protein 4 motif, heterogeneous nuclear ribonucleoprotein R, Q family, and poly-U binding splicing factor, half-pint family. Furthermore, the most enriched Biological Process GO term is synaptic transmission (16 genes, FDR < 7e-14). This cluster contains previously known neural-specific high complexity genes, e.g. *Dscam*[30] and also contains 21% of all genes with neural-specific 3' UTR extensions (hypergeometric p-value < 1e-16).

The set of high-complexity protein-coding genes is not strictly limited to neuronal tissues. There are a small number of exceptionally complex genes expressed in other cell types. For example, *Myosin heavy chain (Mhc)* may encode 438 protein isoforms. Mammalian genomes include many myosin-encoding genes, with separate loci for smooth, cardiac, skeletal, and other distinct muscle tissues[31]. None of these genes in mammals are known to encode more than a dozen protein isoforms. In fly, complex splicing permits a single locus to encode the proteins needed for all muscle tissue. The gene structure is similar to that of *Dscam*, in that five groups of cassette exons of similar length are present, and exactly one cassette exon from each group is included in each transcript.

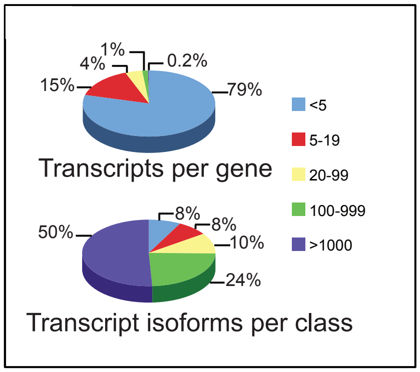


Figure 1.3 Complex Splicing Patterns are Largely Limited to Neural Tissues

Pie charts illustrating the fact that a small minority of genes (48, 0.2%) encodes the majority of all transcripts inferred in our study.

Genes with complex splicing patterns tend to be conserved among *Drosophila* species. Figure 1.4 shows that the number of introns in orthologous genes correlates strongly (r ~ 0.9) between *D. melanogaster* and *D. pseudoobscura*[32]. We see only three strong outliers, including the innate immunity response gene *mustard* (*mtd*), which has undergone a large expansion in *D. melanogaster* that is not seen in *D. pseudoobscura*.

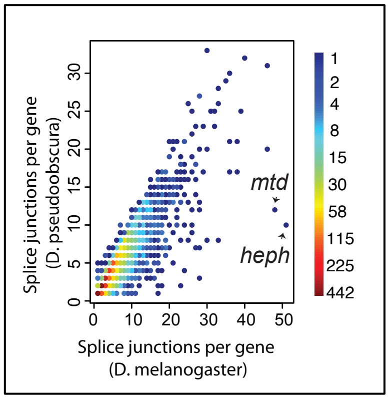


Figure 1.4 Intron Counts are Conserved between Drosophilids[15]

Indicated are two outliers that have undergone species-specific expansion. We attribute the fact that we observe more outliers in the melanogaster lineage to the superior quality of the assembled genome and the increased depth of sequencing data.

Long Non-coding RNAs

A growing set of candidate long non-coding RNAs (lncRNAs) have been identified in *Drosophila*[8,33,34]. 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[35]. However, this number was based on transcribed regions, not transcript models, and utilized non-stranded RNA-seq data[35]. I 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 in the last intron of the early developmental growth factor *spätzle* encodes a 42aa putative ORF that is highly conserved across all sequenced *Drosophila* species. It is not known where, when or if this short ORF is translated. I identified 1875 candidate lncRNA genes producing 3085 transcripts, 2990 of which have no overlap with protein-coding genes on the same strand. Some of these putative lncRNAs may encode short polypeptides, e.g. the gene *tarsal-less* encodes three 11aa ORFs with important developmental functions[36]. I determined protein conservation scores for each ORF between 20 and 100aa. Of the 1119 predicted lncRNAs[35], full-length transcript models for 246 transcribed loci are included in the GRIT annotation; 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[21,37].

Interestingly, all newly annotated genes overlapping molecularly defined mutations with phenotypes are lncRNAs. For instance, the mutation D114.3 is annotated as a loss of function regulatory allele of *spineless (ss)* that maps 4 kb upstream of *ss*[38] and within the promoter of *Mgn4221*. Similarly, *Mgn00541* corresponds to a described, but not annotated 2.0 kb transcript overlapping the annotated regulatory loss of function mutant allele *ci[57]* of *cubitus interruptus*[39]. It remains to be determined whether these mutations are a result of the loss of function of newly annotated transcripts, the possible gain of function of the newly

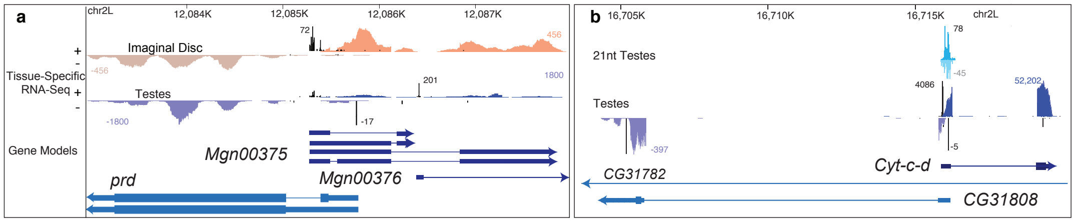


Figure 1.5 Examples of Antisense Transcription

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. **b**, A 5'/5' antisense region that produces substantial small RNA signal on both strands.

annotated transcripts or, as currently annotated, *cis*-acting regulatory elements (e.g. enhancers).

Antisense Transcription

Antisense transcription has been previously reported in *Drosophila*[40], but the catalog of antisense transcription incorporated into gene models has been largely limited to mRNA-mRNA overlaps. I identify non-coding antisense transcript models within the GRIT annotation for 402 lncRNA loci that are antisense to mRNA transcripts of 422 protein-coding genes (e.g. *prd,* Figure 1.5a), and 36 lncRNAs form "sense-antisense gene-chains" overlapping more than one protein-coding locus, as has been previously observed in mammals[37,41]. I note that 21% of lncRNAs in *Drosophila* are antisense to mRNAs, comparable to human where 15% of annotated lncRNAs are antisense to mRNAs (1672 lncRNAs out of 10,840, as of GENCODE v10). In all antisense transcript models for 5057 genes (29%, compared to previous estimates of 15%[40]) have been assembled. For 67% of these loci, the antisense expression is observable in at least one cell line, indicating that the sense and antisense transcripts may be present in the same cells (BPKM > 1 for at least one sense-antisense exon pair). Note that lncRNA-mediated antisense accounts for a small minority of antisense transcription – 94% of antisense loci involve overlapping protein-coding genes. However, only 323 loci (667 genes) encode proteins on both strands. Hence, the vast majority (84%) 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 they, as with many lncRNAs, form gene-chains across contiguously transcribed regions. A small subset of genes with antisense in both UTRs corresponds to a rare transcriptional phenomenon: reciprocal transcription. At these loci, sense and antisense transcripts overlap almost completely (more than 90% reciprocal overlap). There are 13 such examples in *Drosophila*, (e.g. *Polypeptide N-acetylgalactosaminyltransferase 35A* (*Pgant35A*)) seven of which are male-specific (none are female-specific). I note that 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 QQ plot[7]. Furthermore, in both cases positive correlation is more common and more extreme than

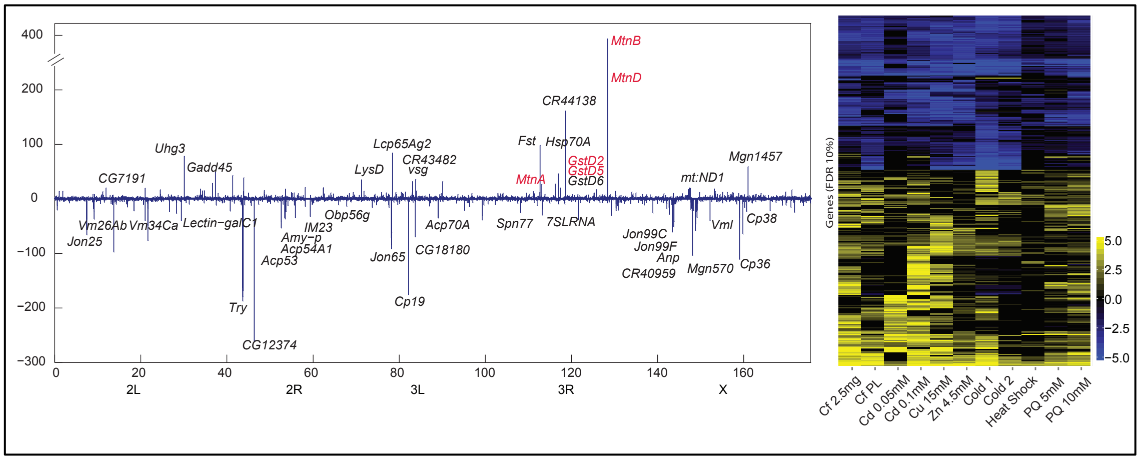


Figure 1.6 Effects of Environmental Perturbations on the *Drosophila* Transcriptome

Adults were treated with the stimulant, caffeine; heavy metals, Cd, Cu and Zn; temperature, cold and heat; and the herbicide, paraquat. **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 previously in larvae[50]. **b**, Heat map showing the fold change of genes with an FDR < 10% (of being differentially expressed) in at least one sample (log2 scale).

negative correlation (lncRNA/mRNA 5th, 95th quantiles -0.125, 0.729, mRNA / mRNA 5th, 95th quantiles -0.169, 0.634), suggesting that negative regulation via the siRNA pathway does not completely process either the sense or antisense transcripts in most cells. Surprisingly, this effect is stronger when the analysis is restricted to cell line samples[7].

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 (Figure 1.6a), but not by other treatments. 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*, Figure 1.6a). 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, I find a homogeneous response to environmental stressors (Figure 1.6b). 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), 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". 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 *D. melanogaster* genome has the potential to encode hundreds of thousands of transcript and protein isoforms via the combinatorial usage of alternative promoters, splice sites, polyadenylation sites, and RNA editing events. The vast majority of splicing complexity occurs in neural tissue, and more than half in differentiating neural and related tissues. A small subset of ultra-complex genes encode more than half of the transcript isoforms that we have inferred, and these are dramatically enriched for RNA editing events; indeed, many are edited at multiple exons, which further amplifies the number of unique proteins these loci may encode. We also observe alternative splicing in UTRs, which do not alter the encoded proteins; genes like *CaMKI* may express thousands of transcript isoforms that differ only in their non-coding UTRs, but fewer than 10 protein isoforms. Our study indicates that the total information output of an animal genome may be heavily weighted by the needs of the developing nervous system, and that transcriptional complexity is a feature of both coding and non-coding sequences.

We identified over 1500 new genomic regions that produce discrete, capped, polyadenylated transcript isoforms in one of the best-annotated animal genomes. Each of these corresponds to a candidate new gene. This underscores the importance of spatiotemporal resolution in comprehensive gene identification. A large fraction of the new genes are testes-specific, many of which are antisense RNAs, as previously described in mammalian systems [37]. Some new non-coding RNAs, such as the lncRNA *Mgn.94020* at the *ovo* locus (Figure 1.1c) have large genomic spans and long-distance splicing events, forming sense-antisense gene chains that bring distal protein-coding genes into direct transcriptional relationships, another previously mammalian-specific phenomenon. The presence of short RNAs at many regions of antisense transcription is suggestive of functional roles for these transcripts, and strongly suggests that the sense and antisense transcripts are present in the same cells and at the same times. The positional conservation of fly and human antisense transcription at genes like *eve* (*EVX1*), *Dcr-2* (*DICER1*), *CTCF* (*CTCF*), *AdoR* (*ADORA2A*), and many others across 600 million years suggests the evolution of a conserved regulatory mechanism basal to sexual reproduction in metazoans. Functional studies are needed to determine if these antisense transcripts play similar roles in mammalian and fly gonads.

One of the largest challenges in the interpretation of transcriptional data is the identification of protein coding vs. non-coding transcripts. We have defined a comprehensive catalog of putative lncRNAs. However, many genes are known to encode poorly conserved, short polypeptides, including genes specific to the male accessory gland, and a number of our candidate lncRNAs may likewise encode short peptides, particularly since a large fraction were discovered in testes and male accessory gland samples. Ribosome profiling has revealed that a number of putative lncRNAs in mammals can be translated[42], but so far these have been difficult to validate with direct proteomics data[43]. Therefore, while we refer to our capped, poly(A)+ RNAs that lack long or conserved ORFs as non-coding, additional data are needed to assess whether these RNAs have functions that are independent of protein translation.

Methods

Predicting Proteins Based on Transcript Models

In each transcript, I automatically annotated the longest ORF as a predicted protein whenever that ORF was at least 100 aa in length. When the longest ORF was between 20 aa and 100 aa, I evaluated each ORF longer than 20 aa as follows: I ran RPS-BLAST using the CDD (as below) and annotated any ORF with a CDD hit E-value of 1e-5 or less as a protein-coding ORF; I ran PhyloCSF (as below) and annotated any ORF with a conservation score of -0.2 or more as a protein-coding ORF. I note that this procedure identified novel conserved ORFs in 277 FB5.45[44] non-coding genes out of 893 such annotated genes, as well as 391 conserved ORFs in novel genes. In all, short conserved ORFs were identified in 27% of genes with no ORF over 100aa. Only 5% of these calls were due to the CDD RPS-BLAST search; PhyloCSF called the remainder of the protein-coding genes. I consider these novel short ORFs "provisional"; extensive validation will be required to determine if they are translated *in vivo.*

Identifying Conserved Domains in Predicted Proteins

I utilized the NCBI Conserve Domain Database (CDD)[45] and the Reverse Psi-BLAST (RPS-BLAST) tool[46] to identify functional domains in predicted proteins, using default settings. I used an E-value threshold of 1e-5 to specify potential hits. The precise executable and settings utilized are detailed below:

- blast standalone execuatable (including RPS-BLAST algorithm): ftp://ftp.ncbi.nih.gov/blast/executables/release/2.2.26/blast-2.2.26-x64-linux.tar.gz
- Conserved Domain detailed definitions and short names: ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd/cddid_all.tbl.gz
- Binary Conserved Domain Database (downloaded 9/1/12): ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd/big_endian/Cdd_BE.tar.gz

The Reverse Position Specific BLAST 2.2.26+ algorithm as part of the NCBI BLAST+ standalone package (version 2.2.26) was used to identify conserved domains within putative conserved domains.

Identifying Conserved ORFs that Lack Known Domains

I utilized the program PhyloCSF[47] to identify novel conserved ORFs that lacked known domains in the CDD database. The inputs to the algorithm are the

14 flies multiple alignments in MAF format (reviewed in [48]) and the set of ORFs called from the GRIT annotation (see above, “Predicting proteins based on transcript models”). The algorithm was run as follows:

- PhyloCSF executable (as of 2012-10-28): http://github.com/mlin/PhyloCSF/tarball/20121028-exe

PhlyoCSF is run in the "AsIs" mode which analyzes only the input ORFs (ORFs are not discovered by PhyloCSF). Based on communication with the Kellis group and their previous experience24 (also, personal communication with Mike Lin), I utilized a conservation score threshold of -0.2 to identify conserved proteins

Defining lncRNA Elements

I define genes that lack any coding transcript given the above definition, and that encode no known small RNA (e.g. tRNAs, miRNAs, etc.) as lncRNAs. I note that this means that our annotation includes non-coding transcripts of coding genes. In Drosophila, there is one gene known to encode four 11aa ORFs[36], and hence it is possible that some of our lncRNAs may yet encode conserved and/or functional short polypeptides. However, PhyloCSF run time is exponential in minimum ORF length between 10 aa and 20 aa, due to an exponential increase in the number of such ORFs present in transcript models. Furthermore, the power of the model is predicated on being able to observe protein-coding structure in multi-species alignments, e.g. third base wobble[47]. This power is dampened in short ORFs, and after extensive manual review I determined that 20 aa was likely close to the limit of detection of the algorithm. This corresponds roughly to the limits of detection of MS/MS in our experience[43], and highlights the difficulty of identifying short protein coding sequencing, and the importance of emerging assays such as Ribo-seq[42].

Computing Gene Expression

I computed gene level expression measurements in BPKM as previously described[8] over the projected gene models. The projected gene models were determined by projecting all overlapping exons for each gene down into non-overlapping exon regions, and then computing the BPKM across the entire region.

Differential Gene Expression Analysis

Differential gene expression analysis was conducted only for our adult treatment samples. Our negative control used for this analysis the wild-type adult fly in gender-balanced mixed populations. Gene-level BPKMs were computed on independent biological replicates. I conducted quantile normalization of the BPKMs across all treatments and the negative control. To compare two conditions, A and B, I selected the replicates in A and B that minimized: $\left|1 - {r_{A,i}}/{r_{B,j}}\right|$ over $i, j \in \{1,2\}$, that is, the pair of replicates that provide the least evidence of differential expression are selected. I call this value corresponding to this minimum the Most Conservative Ratio statistic (MCR). I ignored any gene not expressed above BPKM 1 in either the treatment or the control (represented as a value of identically 0). Additionally, all genes that varied more between replicates within either treatment or control than under the MCR statistic were ignored (represented as a value of identically 1). I formed a rank list of the remaining genes under the MCR statistic. I identified thresholds in two ways. Firstly, I compared our two Cold Shock treatments, Cold 1 and Cold 2. As these treatments differ only slightly, I expected no genes to be differentially expressed between these two samples. No genes had an MCR value >1.77 or <0.33. To ensure a stringent threshold for differential expression, I fit a normal distribution to the log(MCR) values, and computed an FDR value corresponding to identically one false discovery on average per treatment (only one gene falsely discovered, not 1% FDR). This gave a threshold of approximately 5 (or 0.2, reciprocally). Hence, although some genes may be differentially expressed between Cold 1 and Cold 2, under the conservative assumption that in fact there are none, I estimate that I have, on average, one falsely discovered differentially expressed gene per treatment. Secondly, and far more conservatively, I performed permutation tests and studied the distribution of the MCR statistic under random permutations. This has the advantage of providing a sample-by-sample estimate of the FDR.

Multiple Testing and Enrichment

Throughout the manuscript p-values are listed with the test utilized to generate them. These were computed as follows: Statistics computed in R were done with R version 2.15.3[49]. z-scores were computed in R in the usual way, as standardized residuals with empirical unbiased estimates of mean and standard deviation. z-tests were done in R using the function z.test. t-tests were done in R using the function t.test in preference to z.tests for all small sample-size tests of asymptotically Gaussian statistics. Hypergeometric tests were conducted using the R function phyper. Binomial tests were conducted using the R function binom.test. Chi-square tests were conducted in MATLAB (R2011a). Gene Set Enrichment Analysis (GSEA) was conducted as in [50]. Permutation tests were conducted in python with custom scripts. Values were permuted uniformly and empirical p-values were generated. Hence, the number of permutations conducted determines the minimum p-value generated for these tests. False Discovery Rates (FDRs) were computed using Benjamini-Hochberg procedure implemented in custom scripts in Python.

Chapter 2: Extensive Cross-regulation of Post-transcriptional Regulatory Networks in *Drosophila*

Preface

The contents of this chapter are derived from the in submission paper, "Extensive cross-regulation of post-transcriptional regulatory networks in *Drosophila*" with the consent from primary contributing co-authors. The contents presented here represent analyses conducted by myself. Specifically, the Celniker, Artavanis-Tsakonas and Graveley labs conducted data producing experiments. These data producing methods have been omitted from this manuscript and can be found in the above referenced paper. Additionally, supplementary tables have been omitted from this manuscript and can be found in the above referenced paper (available here in pre-print http://brownlab.lbl.gov/marcus.stoiber/preprint_manuscripts/).

Abstract

In eukaryotic cells, RNAs exist as ribonucleoprotein particles (RNPs). Despite the importance of these complexes in many biological processes including splicing, polyadenylation, stability, transportation, localization, and translation, their compositions are largely unknown. We affinity purified 20 distinct RNA binding proteins (RBPs) from cultured *Drosophila melanogaster* cells under native conditions and identified both the RNA and protein compositions of these RNP complexes. We identified "high occupancy target" (HOT) RNAs that interact with the majority of the RBPs we surveyed. HOT RNAs encode components of the nonsense-mediated decay and splicing machinery as well as RNA binding and translation initiation proteins. The RNP complexes contain proteins and mRNAs involved in RNA binding and post-transcriptional regulation. Genes with the capacity to produce hundreds of mRNA isoforms, ultra-complex genes, interact extensively with heterogeneous nuclear ribonuclear proteins (hnRNPs). Our data is consistent with a model in which subsets of RNPs include mRNA and protein products from the same gene, indicating the widespread existence of auto-regulatory RNPs. From the simultaneous acquisition and integrative analysis of protein and RNA constituents of RNPs we identify extensive cross-regulatory and hierarchical interactions in post-transcriptional control.

Introduction

Gene expression involves a complex and often dynamic interplay between proteins and RNA. The synthesis and/or function of almost all known RNAs involve the formation of ribonucleoprotein particles (RNPs)[51]. These RNP complexes range from small (e.g., *Cas9* bound to a guide RNA) to large (e.g., the ribosome or spliceosome). Very few RNP complexes have been characterized in any organism.

The protein components of RNPs can either interact directly with RNA through one or more RNA binding domains, or can be associated indirectly through interaction with another protein that is itself directly bound to RNA[52]. Proteins such as NOVA-1/2, PTBP1, U2AF65 and RBFOX2, as well as others, contain RNA binding domains that directly bind RNA in a largely sequence-specific manner[53,54,55,56]. In contrast, SMN, which is involved in snRNP biogenesis, lacks any known RNA binding domains, and associates with the U snRNAs indirectly. Many assays characterizing protein-RNA interactions utilize UV-crosslinking to ensure that the observed interactions are either direct or occurred in cells prior to lysis[57]. Though powerful, these approaches also have the following limitations. First, many RBPs that interact directly with RNA cannot be crosslinked to RNA due to the configuration of the RNA-protein interaction. Second, even for proteins that can be crosslinked to RNA, the efficiency of crosslinking is low and not every site of interaction is amenable to crosslinking. Finally, these approaches cannot capture indirect interactions including proteins that are part of an RNP that do not directly contact RNA. Thus crosslinking-independent approaches are necessary to capture the larger RNA-protein interaction landscape.

In addition to the diversity of capture approaches used to study RNA-protein interactions, there are differences in the assays used to characterize the interacting molecules. Several groups have used probes to purify specific target RNAs and then identify the associated proteins, though these approaches often require tagging the target RNA, reviewed in McHugh et al. 2014[58]. Hentze[59] and Parker[60] have used oligo-dT to globally purify human and yeast cellular mRNA-protein complexes (mRNPs), respectively, and then identified the bound proteins, but not the associated RNAs. However, very few studies have purified native RNP complexes and characterized both the RNA and protein components.

RNA binding proteins (RBPs) play a crucial role in cellular biology, particularly in higher eukaryotic organisms where ~3% of genes encode proteins that have either known or predicted RNA binding domains[52]. RBPs participate in many essential post-transcriptional functions including pre-mRNA splicing, 3' end formation, RNA localization, RNA turnover and translation. Many RBPs participate in several of these processes[52]. One example of a pleiotropic RBP is the Fragile X Mental Retardation Protein (FMRP), encoded in *Drosophila melanogaster* by *Fmr1*. FMRP forms a complex with components of the RNAi machinery including *Argonaute 2* (*AGO2*), an essential component of the RNA-induced silencing complex (RISC[61]). FMRP also associates with the ribosome to directly block translation by inhibiting tRNA association[62], and in yet another capacity, functions as a translational activator[63]. Other proteins that have been shown to have pleiotropic effects include NOVA[64], MBNL[65] family proteins, and hnRNP H1[66]. It is likely that the participation of RBPs in multiple post-transcriptional processes will be common.

RBPs recognize their RNA targets through RNA binding domains. In *Drosophila*, and most eukaryotes, common classes of RNA binding domains include the RNA-recognition motif (RRM), the K homology domain (KH), the double-stranded RNA-binding motif (dsRBM), and zinc-finger motifs. As with transcription factors, there is no one-to-one mapping between domains and functional roles, and many RBPs with characterized functions appear pleiotropic. Some RBPs have strong sequence specificity for cognate binding sites, including Nova/Pasilla(PS), which binds to YCAY repeats in species from insects to mammals, although the RNA targets regulated by Nova/PS have changed substantially across metazoans[67,68]. The RNAcompete assay has been used to identify *in vitro* binding specificities and relative affinities for a number of RBPs in several species[69]. A number of factors have been studied *in vivo*, but largely within small-scale studies (e.g. [70,71,72,73,74]). An *in vivo* study in yeast[74] surveyed the binding patterns of 40 RBPs and concluded that the targets of different factors fall into distinct functional classes, indicating that specific RBPs participate in defined regulatory pathways. A study of six of the seven Drosophila small ribonucleoprotein particle proteins (Sm proteins) in *Drosophila* showed that the Sm RNA targets fall into three categories: small nuclear RNAs (snRNAs), small Cajal bodies (scaRNAs) and mRNAs[73]. The extent to which *in vitro* binding affinity models are sufficient to explain *in vivo* patterns of binding is unclear. In most cases, it is also largely unknown whether RBPs tend to bind RNA individually as monomers or in larger complexes.

To explore the compositions of RNPs in *Drosophila*, we characterized the RNA and protein components of RNPs purified using 20 distinct proteins as baits. These proteins were chosen based on their known RNA binding domains (e.g., K-Homology domain, RRM) or roles in RNA biology. We group these 20 RBPs into broad functional classes: Exon Junction Complex (EJC), which marks the location of splicing events and provides a link to processing events downstream of splicing; in mammals this includes nonsense-mediated decay (NMD)[75,76] (the release factor, encoded by *Upf1*); Serine-arginine (SR) splicing factors, that although primarily implicated in splicing, have also been shown to participate in other post-transcriptional events[69] (encoded by *B52*, *Rbp1*, *SC35*, *SF2*, *Srp54, tra2*); Spliceosome-associated factors that interact with the canonical spliceosome complex[77,78] (encoded by *snRNP-U1-70K*: abbreviated here as *snRNP70K*, *CG6227, Cbp20, Rm62, U2af50*); heterogeneous nuclear ribonuclear proteins (hnRNPs), a functionally diverse group of proteins that participate in nuclear RNA processing and export[79] (encoded by *elav, ps, mub, msi, Syp*); and lastly pleiotropic proteins including factors with diverse functions in translational regulation and RNA localization (encoded by *Fmr1, qkr58E-1, qkr54B*).

A unique aspect of this study is that RNA and protein are co-purified from the same IP reaction, something that is not possible in CLIP-seq, or other cross-linking-dependent procedures. We utilize RNA-immunoprecipitation (RIP) to identify both the RNA and protein components of ribonuclear complexes. Analysis of the RNAs and proteins associated with these RBPs reveals a densely interconnected network of interactions. Many RBPs associate with the RNA and protein products encoded by the same gene, and therefore may regulate both the protein and RNA components of dozens of RNP complexes. More generally, the RNAs encoding proteins involved in post-transcriptional regulation tend to be bound by most of the factors in our study, forming “high occupancy targets” RNA

(HOT RNAs). Several studies, e.g. [80], have shown that the RNAs encoding RBPs tend to be post-transcriptionally regulated, suggesting that this may occur more often for post-transcriptional regulators than other types of genes. Our data reveals that this tendency may derive from local interactions in the regulatory network, where RBPs interact with, and presumably regulate the mRNAs encoding their protein interaction partners. Hence, via the integrative analysis of matched protein and RNA interaction data, we identify a poorly studied layer of feedback in the hierarchy of gene regulation of metazoan cells.

Results

Identification of the RNA and Protein Components of RNP Complexes

To explore the composition of RNP complexes in *Drosophila*, the Artavanis-Tsakonas lab purified RNP complexes under native conditions (without crosslinking) from cultured cells expressing 20 different epitope-tagged RNA binding proteins and then analyzed the protein components by mass spectrometry and the RNA components by RNA sequencing (Figure 2.1). The

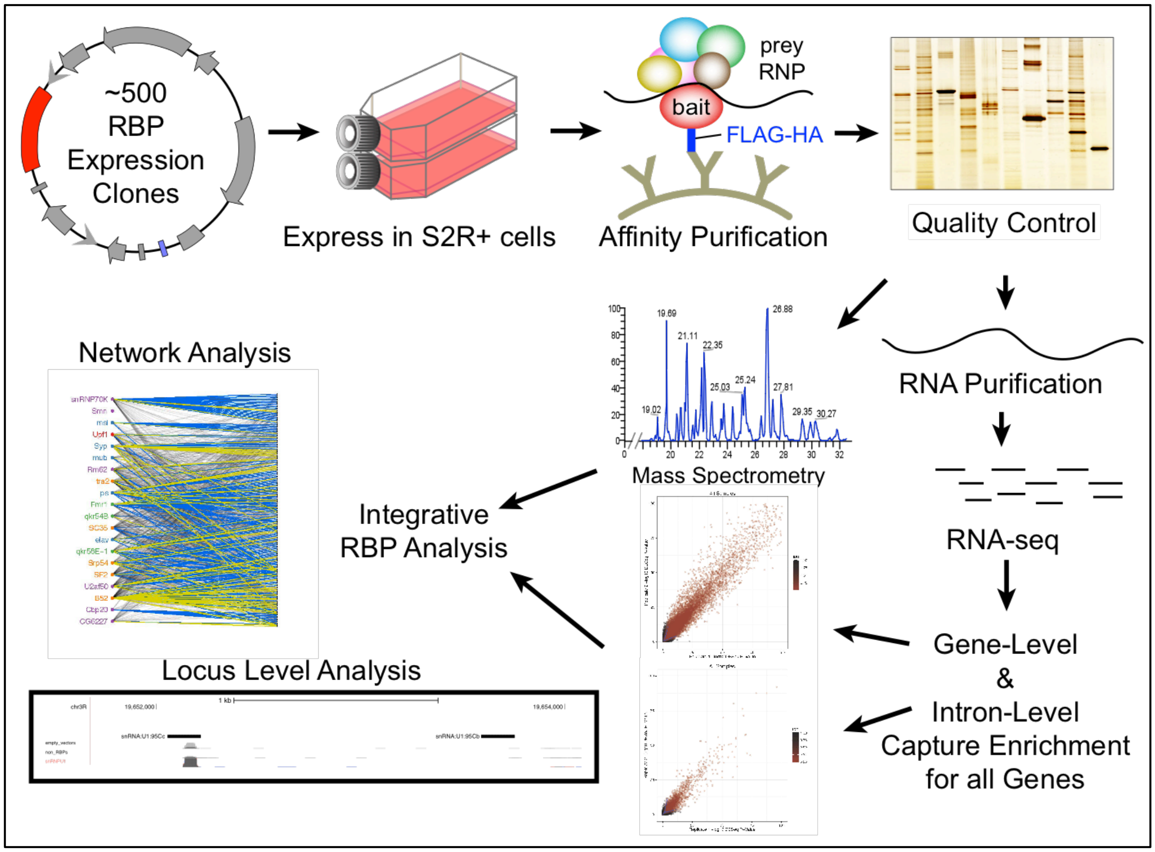


Figure 2.1 Data Production and Processing

The data processing pipeline is described here starting from transfection of RNA binding proteins into S2R+ cells. Immunoprecipitation is then performed to pull down ribonucleoprotein particles. The protein and RNA components of the RNPs are then separated and measured with MS/MS and RNA sequencing. These data are then analyzed together at global and local levels.

proteins selected for these experiments were chosen because they contained KH or RRM type RNA binding domains, DEAD-box RNA helicase domains, or lacked known RNA binding domains, but have important roles in RNA biology. The proteins studied are known to function in splicing, nonsense mediated decay, translation regulation and RNA localization and include members of the SR and hnRNP families of proteins, core components of the spliceosome, and components of the Exon Junction Complex (EJC) (Table 2.1). For each protein we added a C-terminal FLAG-HA epitope to the longest ORF in a vector that allowed inducible expression in transiently transfected cells. This is the same strategy that was developed and demonstrated to be highly effective to produce a *Drosophila* Protein Interaction Map[81].

Protein	Class	RNA Binding Domain(s)	Notes	Number of Identified Targets
tra2	SR-related	1 RRM, 2 RS domains	splicing, sex determination	1126
Rm62	Spliceosome	DEAD-Box Domain	RNAi, splicing, interacts with Fmr1, AGO2 and dcr-1, neurogenesis	1559
B52	SR Protein	2 RRMs, 1 RS Domain	splicing	2384
Cbp20	Spliceosome	1 RRM	Binds to 7mG caps	601
Upf1	NMD, EJC	DEAD-Box Domain, Zinc-binding domains	helicase, NMD	2077
Rbp1	SR protein	1 RRM	splicing	1751
CG17838/Syp	hnRNP	3 RRMs	neurogenesis, R/Q splicing domain	1726
qkr58E-1	Other	1 KH domain	neurogenesis	1724
elav	hnRNP	3 RRMs	neurogenesis	1775
ps	hnRNP	3 KH Domains	splicing	1781
Srp54	SR protein	2 RRMs	splicing	1905
qkr54B	Other	1 KH Domain	poly(A) and poly(U) binding in vitro	1429
msi	hnRNP	2 RRMs	translation repressor	1034
SC35	SR protein	1 RRM, 1 RS Domain	splicing	1272
Fmr1	Other	2 KH Domains	self-binding, protein binding, synapse organization, long-term memory	880
CG6227	Spliceosome	DEAD-Box Domain	splicing, Prp5 Ortholog	639
SF2	SR protein	2 RRMs, 1 RS Domain	splicing	1538
mub	hnRNP	3 KH Domains	splicing	1104
snRNP70K	Spliceosome	1RRM, 1 RS Domain	splicing, component of U1 snRNP, interacts with SMN complex	642
U2af50	Spliceosome	3 RRMs, 1 RS Domain	splicing, part of U2AF heterodimer (with U2AF38)	1866

Table 2.1 RBP Annotations
Literature review and primary class designation of RBPs from this study as well as total number of identified targets from each RBP

Tagged RBPs were transfected into *Drosophila* S2R+ cells in biological duplicate. As controls we used both empty vector and four different non-RNA binding proteins. For these experiments, cell lysates were prepared in the presence of RNase inhibitors to maintain an RNase-free environment to facilitate recovery of intact RNAs. The RNA-protein complexes were purified by immunoprecipitation (IP) and the products of each co-IP were split into two equal fractions. One fraction was used for LC/MS/MS analysis to define the protein composition of the sample and relate the proteins to the DPiM, and the second fraction was depleted of rRNAs and subjected to RNA sequencing to analyze the associated polyadenylated and non-polyadenylated RNAs.

This experimental approach results in the identification of protein-RNA (RIP-seq) and protein-protein (MS/MS) interactions in a single IP reaction. Because we do not crosslink, we pull down stable whole complexes, and therefore our data do not distinguish between direct and indirect interactions or binding events. When we identify interactions of two RBPs with mRNAs from the same gene, we conclude that these two factors share a common target, though the protein-RNA interactions can occur on either the same or different mRNA molecules. However, if we additionally observe protein-protein interactions between these RBPs, we conclude that there is evidence for the existence of an RNP complex that includes the target RNA and the two RBPs. RBPs interact with many RNAs and proteins present in S2R+ cells. Hence, our data is amenable to network analysis techniques that identify community structure. Because we observe whole complexes, not individual pairwise interactions, we expect stable RNPs to yield densely connected “cliques” of associated RNA and protein molecules. Our data is consistent with this model, and described as follows.

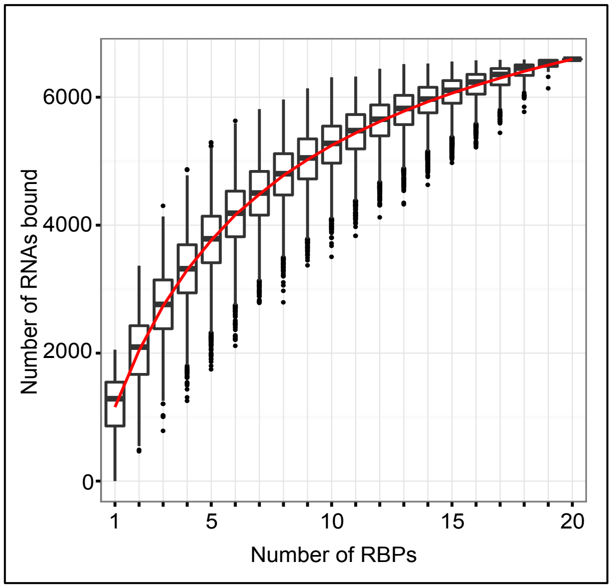


Figure 2.2 Number of Differentially Captured RNAs
Number of differentially captured RNAs for each possible number of RBPs from the study. For number of RBPs that produce more than 5000 possible combinations a random set of 5000 were

To identify RNAs enriched by each RBP, we mapped sequenced reads to the genome and then quantified the capture level (analogous to expression level in a knockout experiment) of each gene (FlyBase r5.57) in both IP and control experiments with DEseq[82] (Methods). We applied two thresholds to the DESeq output: a local Irreproducible Discovery Rate (IDR) of 10% (approximately 3.2% global IDR, Methods) and a Fold Change (FC) of 50% in both biological replicates, corresponding to a local signal to noise ratio of at least 2.0. The IDR, a standard technique for the analysis of IP data developed by the ENCODE

Consortium, is analogous to the false discovery rate (FDR), and leverages biological replicates to measure quantitative reproducibility[83,84]. At this stringent cutoff, we recover an average of 1,231 interacting RNAs per RBP (Figure 2.2). The RBPs we surveyed collectively show statistically significant enrichment of RNA products of 72% of genes expressed in S2R+ cells (Methods) and 40% of all genes in *Drosophila*.

As one way to assess the quality of our data, we examined our results for known RNA-Protein interactions. For example, *snRNA:U1:82Eb* and *snRNA:U1:95Cc* are the two RNAs most strongly enriched by SNRNP70K, an integral component of the U1 snRNP. Consistent with the known interactions between the Cap Binding Complex and U snRNAs[85], CPB20 interacts strongly with the U1, U4, U5, U11 and U12 snRNAs. Moreover, as *Rbp1* is known to cross-regulate *Rbp1-like*[86], we observe a strong interaction of RBP1 protein with R*bp1-like* mRNA. Thus, our dataset recapitulates known protein RNA interactions reported in the literature.

The majority of the factors in our study are involved in splicing regulation. In *Drosophila*, 74% of genes produce spliced transcripts (87% of genes expressed in S2R+ cells). All but one RBP (CBP20, a component of the nuclear cap-binding complex) shows strong enrichment for spliced genes (p-value < 0.005). Hence the preference of most RBPs in this study to bind spliced RNAs supports their functional roles as splicing regulators.

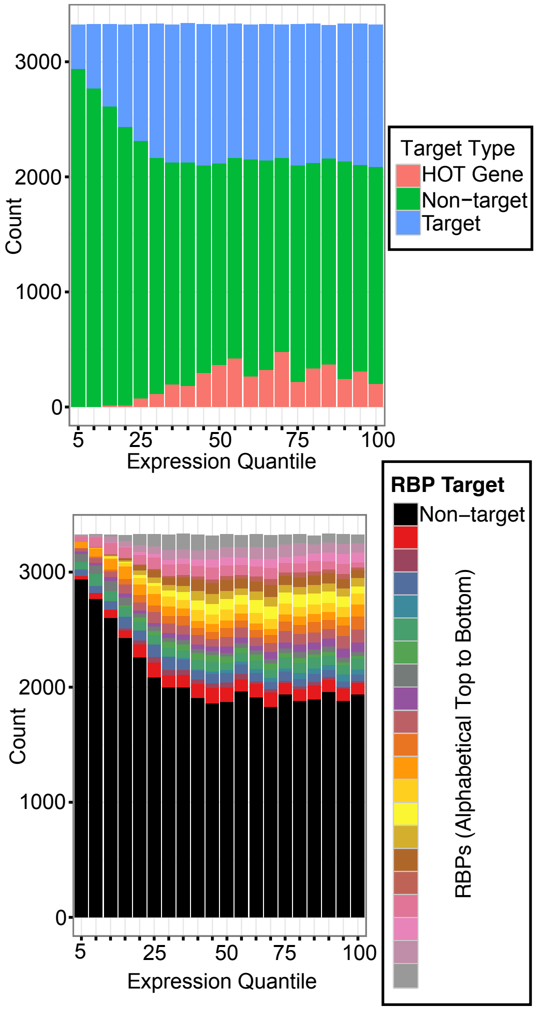


Figure 2.3 Expressions of Target RNAs
Stacked bar plot across quantiles of non-zero expression loci of RBP-gene interactions grouped by a) target/non-target/HOT RNA and b) each RBPs individual hits.

“High Occupancy Target” RNAs are a Feature of Post-Transcriptional Regulation

Most of the RBPs in our study associate with overlapping sets of target RNAs. A total of 74% (141 out of 190) of pairwise intersections of RBP target RNAs across all pairs are larger than expected at random (hypergeometric p-value < 0.01). For example, *Smg5* mRNA, which encodes an RNA binding protein involved in NMD, interacts with RNP complexes containing 15 of the 20

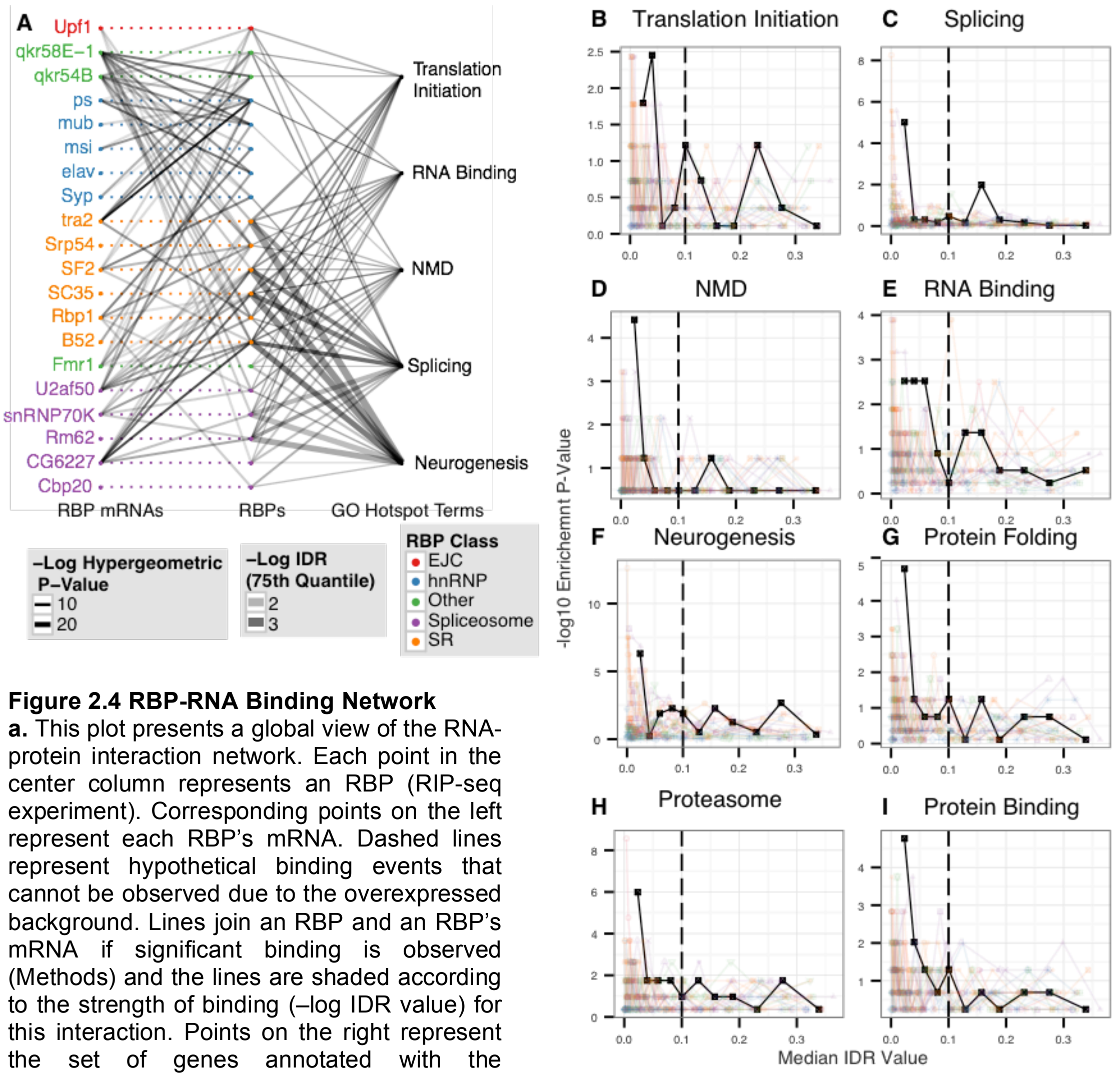


Figure 2.4 RBP-RNA Binding Network
a. This plot presents a global view of the RNA-protein interaction network. Each point in the center column represents an RBP (RIP-seq experiment). Corresponding points on the left represent each RBP's mRNA. Dashed lines represent hypothetical binding events that cannot be observed due to the overexpressed background. Lines join an RBP and an RBP's mRNA if significant binding is observed (Methods) and the lines are shaded according to the strength of binding (–log IDR value) for this interaction. Points on the right represent the set of genes annotated with the corresponding hotspot GO term. Lines are drawn between an RBP and a hotspot GO term if the bound set of RNAs significantly overlaps (p-value < 0.01) the GO term set. The thickness of these lines represents the significance of the overlap between the corresponding sets of RNAs. The shading of these lines indicates the binding strength of this set of bound RNAs (defined as the 75th percentile of the –log IDR values for the bound RNAs). **b-e.** HOT RNAs are driven by the most enriched RNAs. Each plot represents the enrichment for a single hotspot GO term gene set across all experiments. The one solid line represents the median IDR value for each RNA for of all RBPs and each transparent line represents a single RBP. Each point represents 100 RNAs binned by IDR value in increasing order. The y-value for each point represents the –log hypergeometric p-value for the overlap between the 100 bound RNAs and the GO term gene set. Each plot represents the "down-the-rank-list" enrichment for a particular hotspot GO term: **b.** Translation Initiation **c.** Splicing **d.** NMD **e.** RNA Binding **f.** Neurogenesis **g.** Protein Folding **h.** Proteasome **i.** Protein Binding

studied RBPs. Indeed there are six such mRNAs (*CG12065*, *CG3008*, *CG7456*, *Hsp26* and *Hsp27*), which is considerably more than expected under an independence model (probability that the RNA bound by the most RBPs >= 15 is less than 0.001). The RNAs encoded by 282 genes interact with half or more of the RBPs in our study and we will refer to these RNAs as "high occupancy target" (HOT) RNAs. Under a model conservatively conditioned on the assumption that only RNAs associated with at least one RBP are available for binding, this constitutes 282-fold enrichment over expectation (Poisson-binomial p-value $<10^{-15}$). This threshold ensures that HOT RNAs are the targets of a diverse group of RBPs, including multiple binding domains and functional families. We note that the *qkr58E-1* and *qkr54B* mRNAs, which encode two of the RBPs we surveyed, are themselves HOT RNAs. Additionally, we note that the set of HOT RNAs, as well as RIP-seq targets in general, span a wide range of gene expression levels, see Figure 2.3 and are not biased toward highly expressed RNAs.

A number of Biological Process GO terms are strongly enriched in the HOT RNAs, with the strongest being nuclear mRNA splicing (GO:0000398, adjusted p-value <0.001), neurogenesis (GO:0022008, adjusted p-value <0.001) and NMD (GO:0000184, adjusted p-value <0.05). We also observe enrichment for the Molecular Function GO terms RNA Binding (GO:0003723, adjusted p-value <0.05) and translation initiation (GO:0003743, adjusted p-value <0.05). When we rank the target RNAs by their local IDR, using this score as a proxy for a direct measure of binding affinity or fractional occupancy, we find that the most strongly associated RNAs drive the enrichment of the HOT RNA enriched GO terms (Figure 2.4). Collectively, HOT RNAs show strong enrichment for several categories, which include mRNAs of almost a quarter of the genes involved in RNAi (five out of 22, including *Dcr-2* and *AGO2*) and almost half of the genes involved in NMD, despite consisting of only 3% of expressed genes (Figure 2.4). The hnRNP and quaking related RBPs contribute much less significantly to HOT RNA GO term enrichments than SR or splicing related RBPs (rank rum p-value <0.0005). However, at least one hnRNP or quaking related protein targets 92%

RBP	Hypergeometric Test	RNAi within RIP-seq Wilcox Rank-Sum Test	Overlap	Splicing Events Altered by RNAi	RIP-Seq Targets
Srp54	9.57E-17	1.61E-06	103	264	1693
CG6227	4.07E-06	0.0507	18	118	461
Rm62	2.21E-05	0.00212	105	519	1316
mub	3.01E-05	0.188	22	91	962
qkr54B	5.76E-05	0.09831	19	60	1256
Upf1	6.38E-05	0.0540	53	181	1697
B52	0.000112	0.00296	103	343	2052
Rbp1	0.00438	0.112	24	100	1381
elav	0.00638	0.0398	34	135	1615
snRNPU1	0.00649	0.180	35	449	491
Syp	0.00822	0.0953	18	67	1522
SC35	0.0425	0.509	22	138	1082
tra2	0.226	0.586	7	55	1001
Fmr1	0.675	0.492	5	91	697

Table 2.2 Overlap between RNAi and RIP-seq Experiments for Individual RBPs
Hypergeometric p-values quantify the deviation from random overlapping patterns between the target RNAs from RIP-seq and genes that showed differential splicing patterns after RNAi treatment for each RBP. Wilcoxon Rank-Sum p-values represent the preference for genes showing differential splicing after RNAi treatment to show stronger interactions (as measured by IDR value) from RIP-seq experiment on the same RBP.

of HOT RNAs, and thus contribute strongly overall to the biological GO term enrichments of HOT RNAs.

Binding Events Identified by RIP-seq are Functional

To assess potential biological functions of the RNA-protein interactions identified in this study, we compared our RIP-Seq results to RNAi knockdown experiments of 14 of the RBPs included in this study (Srp54, CG6227, Rm62, mub, qkr54B, Upf1, B52, Rbp1, elav, snRNPU170K, Syp, SC35, tra2, and Fmr1) [Brooks et al, submitted]. We observed statistically significant overlaps (max p-value < 0.05) between the splicing events altered upon RNAi knockdown of an RBP and the RIP-Seq targets for the same RBP (Table 2.2). There was lower overlap between targets and affected splicing events for tra2 and Fmr1, though Fmr1 appears to play a role in mRNA localization and has not been reported to directly regulate splicing, and the RNAi depletion efficiency of tra2 was lower than the other RBPs. In conclusion, these overlaps provide overwhelming statistical evidence for the functional importance of the interactions identified by RIP-seq (Fisher combined p-value $<10^{-100}$) and visa-versa for the splicing events identified by RNAi.

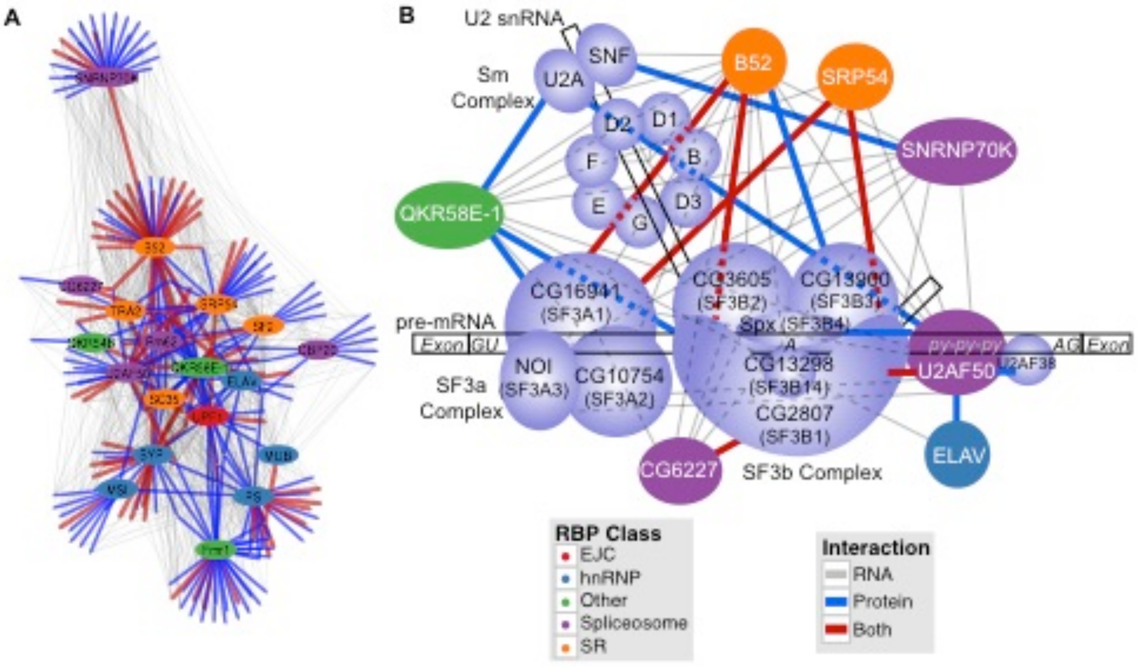


Figure 2.5 RBP-Protein-RNA Interactions

a. Plot represents combined interactions between RBPs and all proteins pulled down in at least one experiment as well as their corresponding transcripts. Edges are drawn where an RBP participates in an interaction with a gene. Grey lines indicate RBP-RNA interactions, blue lines indicate RBP-protein interactions and yellow lines indicate both interactions with the same gene. **b.** Diagram of the U2 snRNP adapted from [2] showing the interactions between core proteins of the U2 snRNP and the RBPs from this study. Only those RBPs involved in protein-protein interactions are presented. The U2 snRNP is composed of U2 snRNA, the SF3a and SF3b splicing complexes as well as the Sm proteins. Lavender-colored proteins are components of the U2 snRNP (along with U2AF50). RBPs from this study are colored according to their primary class designation used consistently throughout the paper. Lines indicate the type of interaction as in Figure 2.6a

RNP Complexes Contain Proteins and their Encoding mRNAs

As mass spectrometry was conducted on each of the IP fractions from which RNA was eluted for sequencing, we can identify the proteins associated with all RBP baits. We obtained at least one confident interacting protein for all but one (RBP1) of the 20 RBPs. We observe an average of 13 proteins associated with each RBP and a total of 198 proteins co-associated with at least one RBP. We confirmed protein-protein interactions for the pairs SC35:QKR58E-2, QKR58E-1:LARK, U2AF50:UPF1 using reciprocal co-IP experiments with an alternative tag. These simultaneously validate the targeted interactions and our protein tagging strategy (Methods). We also compared our results to a database of published interactions (www.droidb.org), and found that 44.6% have been previously reported (>40-fold enrichment, parametric permutation test, p-value < 10^{-16}). The co-associated proteins are strongly enriched for mRNA binding (GO:0003729, p-value <10^{-13}) molecular function despite masking many possible interactions between RBPs targeted in this study due to possible cross-contamination from the MS protocol. The associating proteins are enriched for several terms also enriched in the HOT RNAs including both biological process and cellular component splicing-related terms. We observed highly significant overlap (hypergeometric p-value < 0.01) between the protein and the corresponding RNA targets for three (B52, SYP and CG6227) of the 20 RBPs, and note that there is a strong tendency amongst all RBPs to co-bind proteins and their mRNAs (Fisher's Method P-Value, <10^{-8}), which indicates that proteins may bind their own mRNA. For example, the B52 protein interacts with *CG4849*'s protein and mRNA. Hence, RNP complex members interact with the RNAs encoding interacting proteins. This indicates that post-transcriptional regulation is highly interconnected and cross-regulatory operating at both the transcript and protein levels (Figure 2.5a).

We find significant enrichment within the identified protein interaction partners for genes encoding components of the U2 snRNP and related proteins (GO:0005686, adjusted p-value <10^{-7}). Within this complex, we observe coordinated binding, where RBPs co-immunoprecipitate with proteins and the corresponding mRNAs that encode them (Figure 2.5b), suggesting tight post-transcriptional control of the U2 snRNP complex by constituent and other RBPs. For example, we observe that four RBPs from this study (U2AF50, B52, SRP54 and CG6227/Prp5) interact with both the RNA and protein expressed from *CG2807*, which encodes the ortholog of SF3B1 (SAP155), an integral component of the U2 snRNP complex. Furthermore, QKR58E-1 interacts with the CG2807 protein and SNRNP70K interacts with the *CG2807* RNA. *CG16941*, which encodes the SF3A1/SAP120 subunit of U2 snRNP, is another hub of interactions with protein-protein interactions with QKR58E-1, protein-RNA interactions with SNRNP70K and CG6227, the ortholog of yeast Prp5, and both protein and RNA interactions with the SRP54, U2AF50, and B52 proteins. Finally, we observe that B52 appears to play a central role as it interacts with many U2 snRNP components including the RNAs of four Sm proteins, D1, D2, D3 and F, as well as three proteins, CG16941, CG13900 and CG2807 as mentioned above.

Together these results are consistent with an intricate network of cross-regulatory interactions that control expression of the U2 snRNP components.

In addition to the experiments performed for this study, we investigated relationships to protein complexes and pathways reported in the *Drosophila* Protein interaction Map (DPiM), a protein-protein interaction map generated in the same cell line[81]. The DPiM contains 10 protein complexes containing 12 of the 20 RBPs in this study. For these RBPs, we observe associations with RNAs encoding proteins within the reported complexes for seven out of eight (not including two gene complexes). These interactions include SNRNP70K within DPiM complex 30 (DC30), where we find SNRNP70K binds RNAs encoding three of the seven proteins that compose this complex. DC482 is a complex containing only PS and MSI, which we confirmed, and we observed a strong association between PS and *msi* RNA, but no significant evidence for the reciprocal interaction between MSI and *ps* RNA. DC52 includes QKR54B and SYP, which is replicated in our experiments (using SYP as bait). This complex includes six other proteins, four of which contain RNA binding motifs (CG4612, CG7903, NITO and QKR58E-3) and a fifth that contains an RNA helicase domain and has been implicated in RNAi (CG6701). QKR54B and SYP, as well as QKR58E-1, associate with the transcripts encoding QKR58E-3 and CG6701. Additionally, we see that QKR58E-1 strongly associates with this complex through both protein and RNA interactions. We find reciprocal RNA binding between the pairs of SYP and QKR58E-1 as well as QKR54B and QKR58E-1. Seven gene products interact with two of these three RBPs and four genes interact with all three.

hnRNP/QKRs Associate with Unique Target RNAs and RNAs from Ultra Complex Genes

Recently, a small subset of genes was identified that each generates more than 100 mRNAs via complex alternative splicing, promoter use and polyadenylation and are referred to as "ultra complex genes" or UCGs[7,9]. Most, but not all, UCGs are expressed principally in neural tissue[7]. UCGs are rare in the *Drosophila* transcriptome; 255 are expressed in S2R+ cells. Nonetheless, we find that UCGs are enriched among the RNA targets of RBPs. UCGs are 25% more likely to be associated with at least one RBP than would be expected at random (binomial p-value $<10^{-38}$), and this enrichment is driven largely by hnRNP/QKRs (rank sum p-value: <0.0005). Mice bearing mutations in the orthologs of *qkr54B* and *qkr58E-1* exhibit neural developmental phenotypes[87]. Our results show that the targets of QKR54B and QKR58E-1 are not enriched for genes involved in neurogenesis. However, QKR58E-1 shows amongst the strongest enrichment for UCGs (hypergeometric p-value $<5*10^{-10}$).

In addition to enrichment for UCGs, hnRNP/QKRs tend to associate with unique target RNAs that other RBPs from this study do not target (rank sum p-value <0.001). The most striking examples are ELAV and MSI that have 26% and 19% of their RNA targets associated with no other RBP studied. Additionally, RNAs with low expression (RPKM < 1 in control samples) are 1.7-fold more likely to show binding to hnRNP/QKR (binomial p-value $<10^{-54}$). In particular *Ccn* RNA, which encodes a growth factor implicated in neurogenesis, is detected at very low levels in the control samples (0.27 RPKM) and 20 of the IP samples (max of 0.72 RPKM), yet is enriched greater than 7000-fold by SYP (592 RPKM). This indicates a highly specific and strong association between SYP and this neurogenesis-related mRNA.

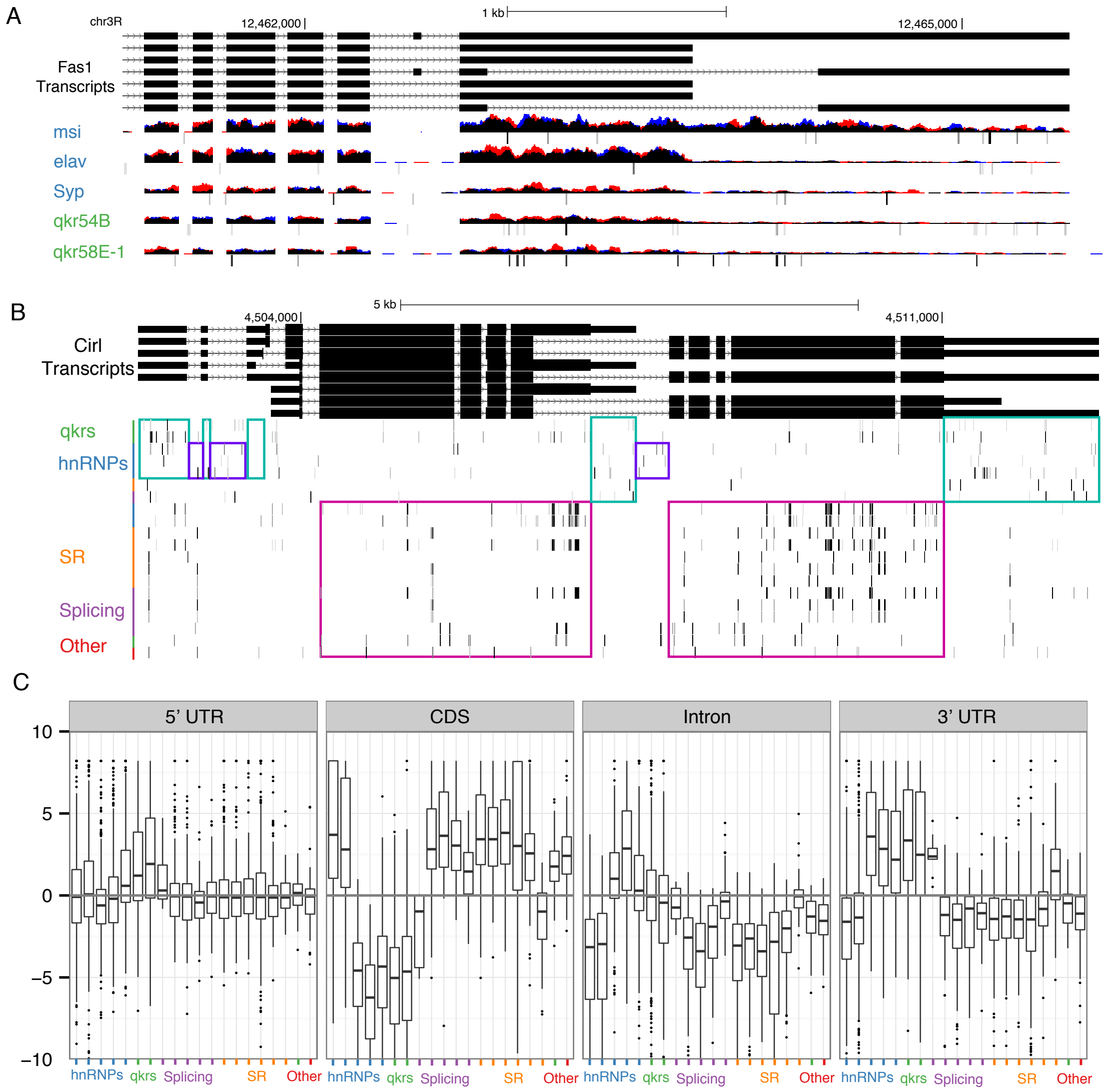


Figure 2.6 Gene Structure Binding
a. The 3' UTR region of the Fas1 locus, where we observe MSI binding specifically to the 3' UTR extended isoform. It has been previously reported that this 3' UTR extension is controlled by *elav*. **b.** *Cirl* is a hotspot RNA in our analysis (bound by, in order of lowest to highest IDR value, SRP54, U2AF50, B52, RBP1, RM62, CG6227, MUB, TRA2, QKR58E-1, PS and SNRNP70K). We note that motif hits group tightly in the gene structure regions. **c.** This plot represents the enrichment of each RBP's motif along gene structure (5' UTR, CDS, 3' UTR). The annotation was collapsed into regions that are only observed as a particular gene structure. Significant motif k-mers (top 1% most likely k-mers given the RBP PWM) are then identified across the transcriptome and overlapped with the gene structure. Each point represents the enrichment of motif-hit proportion within a gene element over the length of the gene structure element at the locus. Note that only enriched loci for each RBP with at least 20 motif hits are plotted. The order of the RBPs is as follows: mub, ps, msi, elav, CG17838, qkr54B, qkr58E-1, Cbp20, Rm62, CG6227, snRNP-U1-70K, U2af50, B52, SF2, SC35, Rbp1, tra2, Srp54, Fmr1 and Upf1.

hnRNP/QKRs Associate with Extended 3’ UTRs

It has been previously published[88] that *elav* is necessary and sufficient to produce several 3’ UTR extensions in *Drosophila,* and that this action is dependent on the direct association of ELAV with target transcripts. We investigate the associations between the RBPs and RNAs expressed from 363 genes with previously reported 3’ UTR extensions that are expressed in S2R+ cells[89]. We find that ELAV associates with 34% of these RNAs containing 3’ UTR extensions (p-value $<10^{-15}$). However, several other hnRNP/QKRs are also strongly associated with RNAs containing 3’ UTR extensions (QKR54B 30%, p-value $<10^{-16}$; QKR58E-1 36%, p-value $<10^{-21}$; MSI 26%, p-value $<10^{-18}$). We find that MSI associates with 52 RNAs containing 3’ UTR extensions that are not detectably associated with ELAV. We manually reviewed each of the eight ELAV targets reported in [88], and found equal or stronger association to 3’ UTR extended isoforms by QKR54B, QKR58E-1 and MSI than with ELAV. One striking example is *Fas1* where MSI associates with isoforms including the extended 3’ UTR, and ELAV associates only with the shorter isoforms (Figure 2.6a). These results indicate that several hnRNP/QKRs in addition to ELAV associate with neural-specific 3’ UTR extensions. These proteins potentially play roles in either poly(A) site selection, RNA localization, RNA stability or translation regulation of the 3’ UTR extended isoforms.

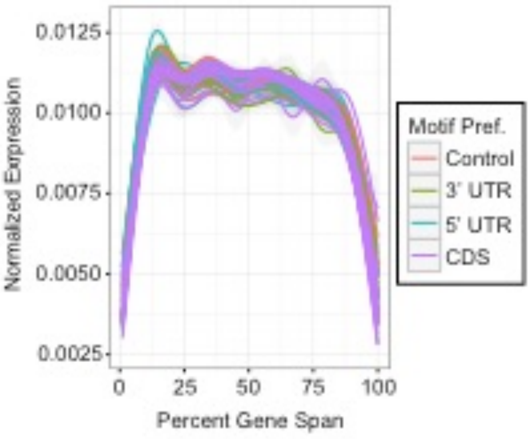


Figure 2.7 Expression across Gene Body Average expression across gene body (from 5' to 3') for single isoform genes with reads covering at least 25% of bins across gene body in all samples.

Gene Region Motif Enrichment

We next sought to identify sequence motifs enriched in the RNA targets associated with each RBP. However, since the approach we used enriches for full transcripts (Figure 2.7), rather than small, RBP-protected fragments, identification of sequence motifs must be performed by considering the entire sequence of the all possible RNAs at a each enriched locus. Extant methods are not available to determine sequence specificity given a set of bound loci within a complex transcriptome, where many genes encode multiple transcripts. We developed a method that identifies enriched sequence signatures within a set of RNAs as compared to all expressed genes, and if statistically significant sequence signatures are found, combines these to produce a sequence motif for each RBP (Methods). We identified motifs for each factor. The RBP encoded by *pasilla* (*ps*) interacts specifically with repeats of YCAY[68], a motif we recover (Figure 2.8). Additionally, we compared our motifs to those discovered using the *in vitro* RNACompete method[90] and found strong correspondence (Methods).

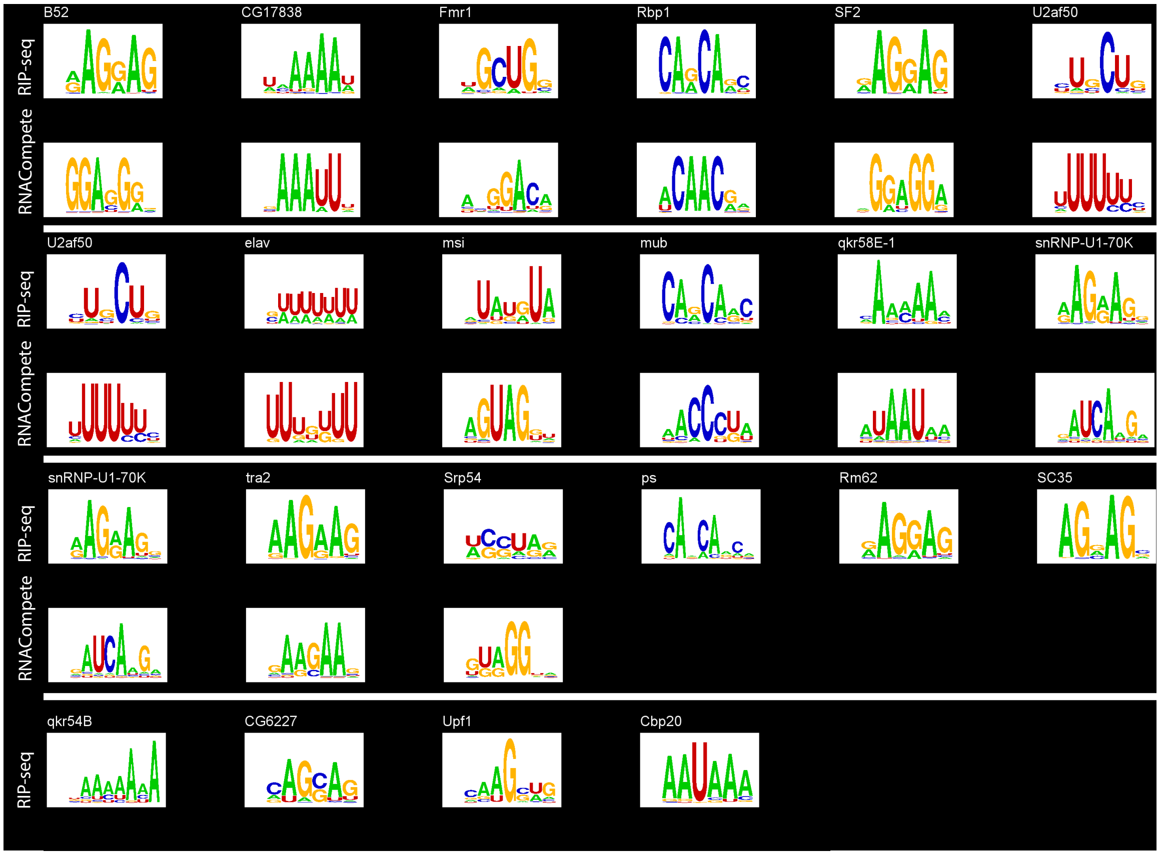


Figure 2.8 RIP Discovered Motifs and Previously Discovered RNACompete Motifs
Both RIP-seq (top row) and in vitro (bottom row, where applicable) discovered motifs for the RBPs in this study

The motif enrichment across target mRNA gene structure suggests that motif analysis may provide insight into the regions of mRNAs bound by particular RBPs. For instance, analysis of the gene *Cirl* (a HOT RNA) reveals a pattern of motif positions consistent with UTR binding for some factors, and CDS binding for others (Figure 2.6b). We asked if any RBPs' motif showed preferential binding in the 5' UTR, CDS, introns and/or 3' UTR and computed the enrichment of motifs across the transcript body (Methods). In general, the motifs for hnRNP/QKRs tend to be present in UTRs more than the spliceosome or SR proteins (Figure 2.6c), though we observe that two hnRNPs, PS and MUB, show enrichment in CDS regions while an SR protein's, SRP54, motif is enriched in the 5' UTR. ELAV targets are strongly enriched for genes with alternative 3' UTRs, as expected, but the strongest enrichment is observed for MSI that shows statistically significant enrichment in over three quarters of genes in *Drosophilia*. We also found significant enrichment within the 5' UTR for QKR54B, QKR58E-1 and SYP. Splicing factors (excluding CBP20) and SR proteins show a mean 2.7 enrichment z-score (p-value <0.01) for motif hits in CDS regions. The EJC release factor UPF1 shows motif enrichment in CDS regions. These motif enrichments were computed solely from exonic sequence, so it is important to consider that these binding preferences may change if intronic sequence is considered. This is particularly true for factors, such as PS, that have known intronic binding function.

Non-coding RNA-RBP Interactions

We sequenced total RNA without a size fractionation step thus recovering unpolyadenylated noncoding RNA targets, which include microRNAs (miRNAs), small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs) and small Cajal body-specific RNAs (scaRNAs) as well as unpolyadenylated long non-coding RNAs. We visually inspected many examples of these targeted RNAs and discovered that the majority of targets are due to the enrichment of the RNA precursors (i.e. retained introns containing noncoding RNAs). However, we also observed several examples of significant enrichment for “intergenic” noncoding RNAs (e.g. *snRNA:U5:14B*, *snRNA:U2:14B* and *snRNA:U2:34ABb*). For example, RBP1, MUB and MSI all significantly enrich for the *snRNA:7SK* RNA, in fact MSI enriches for this RNA over 19-fold. Intriguingly, ELAV displays a very strong (590 fold) interaction with *snRNA:U5:35D* and QKR58E-1 enriches *RNaseP:RNA* over seven-fold. In addition, there are 236 annotated noncoding RNAs (e.g., *CR31044*) that interact with between one and 11 RBPs. For example, *CR31044* which encodes a ~5 kbp RNA that contains *miR-279* and *miR-996* interacts with 11 RBPs, the strongest of which is SYP. Similarly, 10 RBPs interact with *CR43651* which encodes a ~1 kbp RNA hosting miR-14 – the strongest interactor in this case is PS with a 65-fold enrichment. These results identify RBPs that may participate in the biogenesis of specific miRNAs.

We also find that 10 RBPs target one or more small functional RNAs and in total 19 small functional RNAs are targeted by at least one RBP. These include six of 144 expressed snoRNAs, four of nine miRNAs, seven of 18 snRNAs and two of 14 scaRNAs. Of the 10 RBPs, ELAV targets include the most: eight small functional RNAs; no other RBP targets more than four. As mentioned earlier, two U1 snRNAs, *snRNA:U1:82Eb* and *snRNA:U1:95Cc* interact with SNRNP70K, consistent with its known role in the U1 snRNP complex[91], and are among the most enriched RNAs in any IP in this study (84-fold, 81-fold respectively).

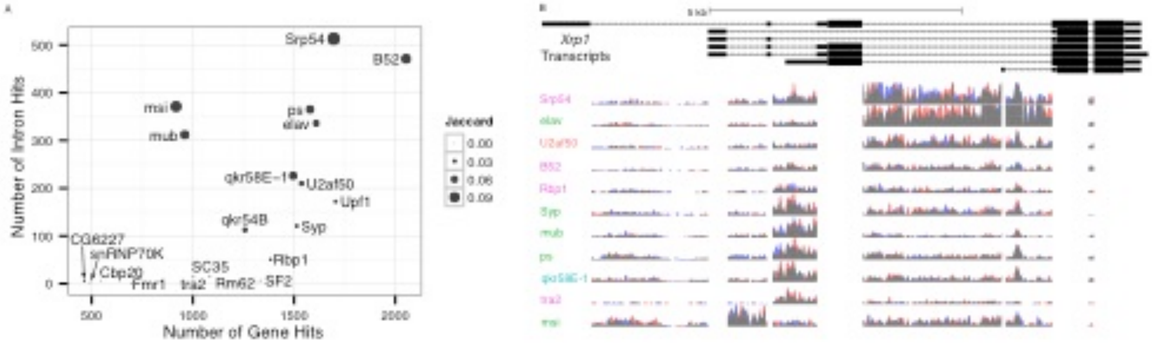


Figure 2.9 Retained Intron Signal in the Data
a. The number of intron and gene level targets is represented for each RBP on the y and x axes, respectively. The amount of overlap between each RBP’s intron and gene level targets, as measured by the Jaccard index at the locus level, is indicated by each point’s size. **b.** The Xrp1 locus is indicative of several genes that produce different cohorts of RBPs binding to different retained introns. The exon regions of reads are removed from this figure. The height of each sequence track is 20 BPKM and the red and blue portions of the tracks indicate the biological replicates.

Enriched Intronic Regions

In addition to investigating the enrichment of particular mRNA transcripts at the gene level, we also queried introns for evidence of enrichment (Methods). We find that while gene level enrichments correlate well (ρ=0.62) with intron enrichment loci across RBPs, there are several factors with many gene level targets that do not show a similar signature at the intron level consistent with intron targeting as a feature of some RBPs and not others (Figure 2.9a). Two factors in particular, B52 and SRP54 are enriched for differentially retained introns with respect to control total RNA samples at almost twice as many loci as any other RBP.

We find individual introns are targets of multiple factors. We also find gene loci with multiple introns that are targets of distinct cohorts of factors. A striking example is the *Xrp1* locus, which encodes a DNA-binding protein and is a HOT RNA at the gene level (targeted by 17 RBPs, Figure 2.9b). *Xrp1* contains seven introns, five portions of which do not overlap other annotated features, and hence are amenable to differential expression analysis (Methods). We find differential intronic enrichment for four of these introns. For example, the second intron is strongly targeted by MSI, but not any other factors, while the fourth intron is preferentially targeted by SRP54 and ELAV. Several other loci with marked differential intron retention include *MRP* and *crol* as well as at the loci of two of the RBPs in this study *ps* and *Syp*.

RBP Functional Groups Bind the mRNAs of Functionally Related Proteins

The co-association of two or more RBPs with a single target RNA occurs broadly throughout the transcriptome. Pairwise intersections provide a natural similarity (and, conversely, dissimilarity) measure between any two RBPs (Methods). Multi-dimensional scaling (MDS), a generalization of principal component analysis, is a powerful technique for visualizing the relationships between data points in high natural dimension[92]. MDS of the co-associations into two dimensions reveals that RBPs from related functional groups (e.g. the SR proteins) bind overlapping sets of target RNAs. This clustering becomes tighter for the three major functional groups when, rather than considering the overlaps in the sets of associated RNAs, the similarity of annotated GO term enrichment profiles (across all biological GO terms) is examined (Methods). As several analyses indicate a strong relationship between the hnRNP class of RBPs and the two quaking related proteins (QKR54B and QKR58E-1), we will refer to these proteins collectively as hnRNP/QKRs.

MDS using GO term profiles indicates that functionally related RBPs associate with mRNAs encoding functionally related proteins. Clusters are tighter under the GO similarity measure than raw transcript overlap. We observe the following ratios between mean within-group distances and mean between-group distances: SR proteins (transcript overlap 0.85, functional overlap 0.69); Spliceosome (transcript overlap 0.98, functional overlap 0.92); hnRNP/QKR (transcript overlap 0.95, functional overlap 0.92). However, these cluster measurement differences do not rise to the level of statistically significant. Both the hnRNP/QKR and SR protein classes functional clustering is driven in part by the HOT RNA enriched term neurogenesis, as defined by the GO terms that provide the largest decreases of within-group distances when removed from the analysis (Methods). Additionally, SR proteins cluster due to mitotic spindle organization (GO:0007052) and translation (GO:0006412) while splicing-related RBPs functionally cluster primarily due to the HOT RNA enriched term splicing as well as telomere capping (GO:0016233).

Comparison of RIP-seq and RNACompete Derived Motifs

Position specific score matrixes (PSSMs) have been determined by RNAcompete, which measures the binding specificity of purified recombinant proteins with a pool of randomized RNA, for 50 *Drosophila* RBPs[90], of which 13 are included in our study. To determine the extent to which RNAcompete PSSMs are sufficient to explain interaction strength in S2R+ cells (quantified by negative log IDR value), we examined the predictive power (Methods) of the RNAcompete motifs, and compared these to motifs derived from our data. We found that in all cases the RNAcompete motif alone was not sufficient to predict RIP-seq binding patterns. Hence, as has been found in numerous studies of transcription factors [93,94], the interactions between RBPs and their target RNAs cannot be strongly predicted by simply using PSSMs scores derived from *in vitro* biochemical assays.

RIP-seq Derived Motifs Cluster within Functional Groups

We quantified the similarity of PSSM binding models for our factors using Kullback-Leibler divergence[95]. MDS analysis reveals relationships between classes of factors. We note that the spliceosome components and SR proteins show the most similar RIP-seq derived motifs, driven primarily by a strong “AGG” submotif. This is in contrast to the RNAcompete motifs, which show significant differences within these classes. We note as well that for U2AF50 in particular, the motif discovered in our experiments diverges significantly from the RNAcompete motif. In the case of U2AF50, the differences between our motif and the RNAcompete motif may be partly due to the fact that we used only exon sequences in our motif discovery. In addition, U2AF50 is known to form a tight heterodimer with U2AF38, and interacts with other proteins, which could impact the binding specificity of U2AF50 in our RIP-seq experiments. Since the RIP-seq derived motifs correspond only to enriched sequence signatures, and may not reflect the direct binding specificity of the factors, the few factors for which the RIP-seq and RNAcompete motifs strongly differ may be due to the detection of sequences associated with other RNA binding proteins or co-factors.

Discussion

We obtained genome-wide RNA-protein and protein-protein interaction profiles for 20 RNA binding proteins. The combined use of next generation sequencing and mass spectrometry on a single immunoprecipitates for each RBP provided new insights into the composition of ribonucleoprotein complexes in metazoans. Validation of RNA-protein interactions by RNA sequencing of

RNAi depleted cells demonstrated the functional importance of these complexes in splicing regulation.

We found that strongly bound RNAs included HOT RNAs that interact with most of the factors in our study. These included many of the genes encoding proteins in the RNAi and NMD pathways, related to neurogenesis, other RNA binding and splicing factors, and components of the proteasome. This is consistent with previous reports[96,97] that genes involved in post-transcriptional regulation tend to be regulated post-transcriptionally. Feedback loops are a central idea in cellular biology, and it is striking that feedback appears to function broadly at the level of an entire regulatory process. Integrative analysis of RBP protein and mRNA interaction profiles revealed ubiquitous interactions with mRNA and protein products of the same gene. Furthermore, we find that RBPs that participate in the same protein complex tend to reciprocally bind the mRNAs of their interaction partners. A striking example of this was presented for the RBPs that interact with the protein components of the U2 snRNP and the RNAs encoding them. Hence, we find that widespread post-transcriptional regulation of post-transcriptional regulators may be an emergent property of local cross-regulation, where RBPs of a complex tend to regulate their interaction partners. Similar patterns have been observed among transcription factors acting in the same pathway, e.g. global cross-regulation within the gap gene network[98]. We find that protein-interaction-associated post-transcriptional regulation is common, and hence constitutes a general layer of feedback in the hierarchy of gene regulation.

The hnRNP/QKR proteins bound a more diverse repertoire of target RNAs than other classes of RBP. We found that hnRNP/QKRs in general were strongly associated with UCGs, genes with many promoters, alternative splicing events, and/or polyadenylation sites. In contrast, SR proteins bound largely overlapping sets of post-transcriptional regulators, with few targets bound by only a single member of this class. This is consistent with their known biochemical redundancy[69], but may in principle also reflect complex regulatory programs requiring the input of multiple SR factors. We find that the hnRNP *Syp* mRNA is itself a target of QKR58E-1, and reciprocally, SYP binds mRNAs of *qkr58E-1*, and we detected protein-protein interactions between SYP and QKR58E-1. The mRNAs of the quaking-related factor *held out wings* (*how*) are targets of both quaking related factors we surveyed, and HOW is a protein interaction partner of SYP. Overall, we find extensive co-regulation and interaction among UCGs and the RBPs that target them.

We found that motifs derived from splicing factors and the EJC component UPF1 tend to be found in CDS regions of target mRNAs indicating potential binding, while hnRNP/QKRs are enriched in UTRs. While several factors, including ELAV, are strongly enriched in 3' UTRs, we found that other hnRNP/QKRs, particularly MSI, show even stronger association with 3' UTR extensions. It was previously reported that ELAV is both necessary and sufficient for these extensions to exist at eight genes[88], but global binding patterns indicate that MSI interacts with the extended 3' UTRs and may play an important role in some aspect of their biology. Thus, ELAV may modulate the biogenesis of extended 3' UTRs, while MSI binds to the extended UTRs.

Our data is consistent with the co-localization of mRNAs of RBPs and the proteins they encode. Furthermore, these associations between interacting proteins and mRNA products from the same genes could be ribosome proximal, or ribosome mediated. It could be that the protein complexes studied here undergo co-translational assembly. This is also an intriguing possibility. Importantly, our assays measure time and space averages across ensembles of homogeneous, but not identical or synchronized cells. Hence, while it is clear that the proteins co-purify and bind the same RNA targets, it may be that these associations occur on different individual RNA molecules that are neither spatially nor temporally localized with the proteins they encode. Additional assays, particularly high content imaging approaches, will be needed to resolve these possibilities, and to elucidate the intriguing biology at the basis of feedback in post-transcriptional regulatory networks.

Methods

Sequencing/Mapping

RNA sequencing libraries were prepared using the Illumina mRNA Sample Preparation kits as described by the manufacturer, but both the poly(A) selection and RNA fragmentation steps were omitted. Libraries were quantitated on an Agilent Bioanalyzer and sequenced on an Illumina HiSeq 2000 to generate single-end 50 bp reads. Library preparation and sequencing were completed by the Graveley lab and are included here for completeness. Reads were mapped to the Drosophila genome using tophat version 1.4.0 guided by the MDv1 annotation[8] with the following settings: --no-novel-juncs, -a 6, -m 2, --min-intron-length 28, -I 200000, -F 0, -g 1, -x 60 and -n 2. Mapped reads are publicly available in the GEO database with accession number **GSE37756.**

Control Filtering/Validation

In order to obtain a confident set of control samples each empty vector control sample was tested for differentially enriched annotated RNAs versus all other empty control samples using the DESeq R package. One empty vector control sample produced a significant number (> 50 loci) of differentially enriched RNAs and was removed from further analysis, in particular from testing for differentially enriched RNAs in samples of interest. Similarly, each non-RBP sample was tested for differentially enrichment as compared to both the validated control samples as well as non-RBP samples. As each non-RBP experiment was conducted in biological replicate these replicates were tested together. None of the non-RBP samples produced more than a few (10) significantly bound (adjusted p-value < 0.05) RNAs and thus all (5 samples each in biological replicate) non-RBP samples were used for testing in the samples of interest.

Sequence-based Transfection Validation

In addition to western gel validation of RBP transfection the following sequence based method was applied to all samples. The raw sequence output of each experiment was queried for the exact FLAG-HA tag sequence immediately adjacent to the RBP of interest. A consensus sequence was created for each experiment by anchoring each read at the FLAG-HA tag sequence and recording the most frequent base at each position starting adjacent to the tag sequence. BLASTN[99] was run on the consensus sequence against the drosophila “nr” database. All RBP samples of interest presented in this paper were confirmed via this method. Five non-RBP samples were used as negative controls as described in the previous methods section.

Identification of Differentially Bound RNAs

In order to confidently identify the differentially captured targets of each RBP of interest the following pipeline was implemented. Mapped reads were binned into gene counts with the htseq python package script htseq-count with setting intersection-strict against most current functionally characterized FlyBase annotation r5.57[23]. The DESeq[82] tool (R version 3.0.2 and DESeq version 1.14.0) was used as the basis of the pipeline to identify differentially enriched RNAs. We note that differentially enriched RNAs have the same sequence signature as differential transcription in many extant studies, except that differential binding only results in more abundant transcripts. As such loci in which both replicates did not show a normalized fold change greater than one were filtered from downstream analysis. Also loci that did not show a sequencing depth greater than 1 read per kilobase per million mapped reads (RPKM) in the controls or any sample were removed from downstream analysis. We note that although some RNAs show low expression in the control samples we find significant evidence for strong binding at a select few loci (e.g. *Ccn*) and this does not use a strict RPKM cutoff from the controls alone. DESeq is thus an applicable tool in this setting after the appropriate filters are applied.

We note that all biochemical steps for the experiments in this dataset are identical aside from the RBP of interest’s RNA is transfected into the cells. Thus dispersion estimates at a particular locus are comparable across samples, up to sequence depth as required by the DESeq model. To take advantage of this replication schema gene level dispersion estimates were computed across all samples not being tested as well as controls at each locus for each RBP sample of interest (corresponding to the DESeq “per-condition” setting). Differential enrichment statistical significance values were then calculated on each biological replicate separately, thus producing two p-values for each RBP-RNA combination. Only RNAs that showed adjusted p-values less than one in both biological replicates as compared to the control samples were considered in further analysis.

In order to identify RNAs that are both strongly and reproducibly bound the IDR (R package; version 1.1.1) model[83], a copula mixture model, was fit on the significance values (across all combinations of valid RBPs and RNAs). The IDR method has been extensively applied to assess reproducibility in biological experiments[84]. Note that IDR cutoff values are quite reproducible when fit on each sample individually. Differentially enriched RNAs are defined as those that produce a local IDR value of less than 10% (corresponding to <10% chance of having resulted from the irreproducible component) as well as a minimal 50% increased fold change in both biological replicates.

GO Term Enrichment

In order to identify gene ontology (GO) terms that are identified within particular groups of genes we compute enrichment p-values for each GO term. The enrichment p-value is a hypergeometric p-value for the number of genes annotated with a particular GO term within a set of interest compared to all expressed genes (defined as in the previous section; at least one sample with greater than 1 RPKM). GO terms annotated to less than 5 genes are removed from analysis. All reported enrichment p-values are adjusted via the Benjamini, Hochberg[100] correction to control false discovery rate (FDR).

Global RBP Binding Profile Comparison

In order to visualize the binding partners, related characteristics of all RBPs as well as sequence binding preferences we applied dimension reduction techniques. The multi-dimensional scaling (MDS) algorithm[101], which minimizes the difference between the input distance and the plotted two-dimensional representation, was applied to several different definitions of distance based on the bound RNAs of each RBP (using the R “cmdscale” function for metric MDS and “isoMDS” function for non-metric MDS that is part of the MASS package; version 7.3-33). Note that MDS output coordinates are only unique up to centering and dilation, thus coordinate values are omitted in plots.

The first distance is the Jaccard distance, one minus the size of the intersection divided by the size of the union, between the set of bound RNAs for any two RBPs. Note that non-metric MDS (which optimizes the rank of distances as opposed to the true distances) was applied to this distance definition since a reasonable fit could not be achieved with metric MDS. The second distance is defined on the vector of negative log (base 10) biological GO term enrichment values annotated to all bound RNAs. Note that values are capped at 10 to avoid outlier enrichment from driving distances. The “functional” distance between any two RBPs is defined as the cosine distance between the vectors of GO term enrichments for each RBP. The third distance is defined on the motifs discovered for each RBP as described in the next section as well as motifs discovered by the *in vivo* method RNAcompete method. In order to assess the distance between two motifs we used the Kullback Leibler divergence as has been done in previous studies with good success[95]. We define the distance by the smallest divergence between two motifs across all possible offsets requiring that at least four overlapping positions.

In order to compare relative clustering of RBP classifications we used the ratio of mean distances within a class to the mean distances between a class and all other RBPs. Thus lower ratios indicate tighter clusters (smaller distances within a class of RBPs than between that group and other RBPs). Note that these ratios are computed from the raw distance measures not the Euclidian distances from the MDS plots.

Specifically for the functional characterization we were interested in the GO terms that "drive" the clustering observed for a particular class. In order to identify those terms, we leave a single GO term out of the distance calculation and calculate the within class versus between class ratio. A positive differential with respect to the ratio including all terms indicates that a term causes a group to cluster more tightly. Those terms that are more prevalent outside a particular class of RBPs than within are removed, as these terms are not indicative of an attribute for that class of RBPs. These terms are driving the tighter clustering of a class by virtue of existing in all other RBPs. More prevalent is defined as a higher mean enrichment value within a class that in all other RBPs.

RBP Motif Discovery

Given that our differential enrichment analysis pipeline identifies loci, as opposed to linear sequences as in similar studies[102,103], we have implemented the following pipeline in order to identify enriched motifs amongst the bound set of transcripts for each RBP in this study. We note that discovery of bound motifs in previous studies has proven particularly difficult and cannot be accurately obtained using existing software such as HOMER[104] that takes linear, genomically disjoint sequences or MEME[105] that in addition does not take into account enrichment, but only presence of a motif within a set of sequences.

The first step in the algorithm is to represent each locus as a vector of 7-mer counts. As the *Drosophila melanogaster* transcriptome contains a large number of alternative events (alternative transcript starts, stop as well as alternative splicing events) the set of 7-mers for each locus are defined as all unique genomic locations of 7-mers across all transcripts within each gene model. We note that we are not performing differential transcript isoform analysis that has proven to be a difficult problem to solve[9] and thus must take into account all annotated transcripts in our motif analysis. Note also that we are removing all global low complexity sequences from this analysis.

For each RBP the total count of each 7-mer across all bound loci is computed. Then the hypergeometric p-value for the enrichment of each 7-mer is calculated as compared to the background 7-mer totals amongst all expressed loci. We note that this distribution does not strictly follow the hypergeometric distribution as genes may contain multiple copies of a particular 7-mer, but note that randomly selected sets of genes produce reasonably uniform distribution of p-values under the hypergeometric model. The set of 7-mers that are significantly more enriched in the differentially enriched set of loci than one would expect at random are used to create the desired motif. Under the null model, of no sequence enrichment, p-values follow a uniform distribution. Thus we have that the minimal p-value follows a beta distribution with parameters alpha=1 and beta=1/# of 7-mers. We define a cutoff for a significant 7-mer as the one-sided 1% enrichment p-value of the beta distribution ($6.13*10^{-7}$). We require that a bound set of interest produce greater than ten 7-mers which show significant enrichment (which all of our 20 RBPs do) in order to produce a motif. We only take the 50 most enriched 7-mers.

In order to produce a motif of variable width from the enriched set of 7-mers we generate a set of sequences with five copies of each 7-mer surrounded by ten bases of random sequence, thus approximately following the null distribution of the MEME model. This set of sequences is passed to the MEME algorithm[105] along with weights corresponding to the transformed negative log enrichment p-values, thus forcing more enriched 7-mers to drive the motif signal. The transformation raises the minimum of the negative log p-value and 200 to the 0.75th power and scales these values between 1/50th and 1. The MEME algorithm produced a single motif per sample between 3 and 11 bases in width and is required to include each sequence in the produced motif (corresponding to the "OOPS" setting). The produced motif is taken as the experimentally discovered motif (Figure 2.8). A post-hoc filter was applied to trim the motif if less than 5% of the total information content lay in the outer-most positions of the motif.

Gene Region Motif Enrichment

Because the RIP-seq protocol does not include a cross-linking step we are able to observe only those transcripts that are differentially bound. In order to identify the gene region (i.e. un-translated region; UTR or coding sequence; CDS) preferences for each RBP we used the learned motif. First we identified the top 0.1% of hits (defined probabilistically by the position specific weight matrix) to the *Drosophila* transcriptome. We then intersected these locations with those regions that are only ever observed as a particular gene region (5' UTR, CDS or 3' UTR). Only genes in which each region type composed 2% of the total gene length were included. Also genes were required to have at least 20 hits for a particular motif to be included in order to remove genes with a small sample of motif locations. For these genes we compute a z-score by first computing the binomial test p-value, where the test statistic is the fraction of motif hits in a region type and the expected fraction is the sequence length for that region type over the total gene length. This p-value is transformed to a z-score by taking the inverse of the survival function for the Gaussian distribution.

Software Implementation

All statistical procedures were completed using the R software program (version 3.0.2). Rank sum p-values are computed using the one-sided "wilcox.test" function that is part of the stats package. Poisson-binomial p-values are computed using the "poibin" package, version 1.2 where the parameters are the fraction of expressed genes observed as differentially bound for each RBP.

Chapter 3: The Early Response to Ecdysone in 41 Diverse *Drosophila* Cell Lines

Preface

The contents of this chapter are derived from the in submission paper, “The Early Response to Ecdysone in 41 Diverse *Drosophila* Cell Lines” with the consent from primary contributing co-authors. The contents presented here represent analyses conducted by myself. Specifically, the Cherbas and Cherbas labs conducted data producing experiments. The data production methods have been omitted from this manuscript and can be found in the above referenced paper (available here in pre-print http://brownlab.lbl.gov/marcus.stoiber/preprint_manuscripts/). Additionally, supplementary tables have been omitted from this manuscript and can be found in the above referenced manuscript.

Abstract

Endocrine signals transduced by nuclear receptors elicit major cell state changes that include cytodifferentiation, profound modulation of immune responses, neoplastic growth, and insect metamorphosis. Responses alter gene expression and cell types respond differently to a single, common endocrine signal. In *Drosophila*, the molting hormone 20E, ecdysone, directs major developmental transitions. Here we survey the early ecdysone responses of 41 *Drosophila* cell lines, representing diverse cell states. We observe genes that are *widespread* in their responsiveness, those responding in most lines, and many more whose responsiveness is *restricted* to one or a few lines. Genes in the widespread class include those previously identified in ecdysone responses studies in few tissues and genetic analyses. Many restricted genes are induced in some cell lines, repressed in others and fail to respond in still others. Expression of the ecdysone receptor (*EcR*) expression level predicts both the extent and the velocity of the global magnitude of cellular responses, and hence EcR titer appears to be rate limiting for ecdysone transduction. Promoter motif compositions combined with transcription factor titer provide significant predictive power for the identification of restricted responses. We characterize the conditional responsiveness for genes with shared promoter architecture and find that transcripts initiating from a bidirectional promoter can be independently controlled in ecdysone response. These findings provide the basis for decoding the specificity of ecdysone responses, and for understanding the pathways of type-II nuclear receptors.

Introduction

In animals, steroid hormones induce development and differentiation, regulate immune function and inflammation, modulate cell cycle and osmoregulation, and are broadly critical to organismal health. Errors in steroid hormone signaling pathways are linked to disease states ranging from oncogenesis to mood disorders. Steroid hormones bind nuclear receptors, which are deeply conserved across metazoans[106] and share a common structure[107]: an unconserved N-terminal A/B region including a transcriptional activation domain (AF-1), followed by a highly conserved 66-68 residue DNA binding domain (DBD), a short hinge region, a conserved ligand binding domain (LBD), and unconserved sequence of variable extent called the F domain. Receptors function by binding to highly conserved response elements (HREs), where they act as powerful transcriptional effectors. Early targets of transcriptional regulation are known in a variety of cell types in several organisms. Although, the complete cell-type-specific genomic response has not been elucidated.

Binding of the hormone to its nuclear hormone receptor (NHR), activates transcription of a limited set of direct target genes including transcription factors (TFs). Within hours this generally leads to a secondary response of activation and repression of hundreds of downstream genes. These changes activate expression of cell-fate-specific structural genes and lead to cellular differentiation. In *Drosophila*, responses to the steroid hormone 20-hydroxyecdysone (20E, hereafter referred to as ecdysone) provide an excellent model for study of regulatory principles[108]. Ecdysone acts through a type 2 heterodimeric NHR composed of the products of *Ecdysone receptor* and *ultraspiracle* (*usp*) bound to a typical response element (EcRE) at distributed sites in the genome[109-111]. At least one-third of *Drosophila* genes respond to ecdysone signals in some cell at one stage or another (L. Cherbas and P. Cherbas, unpublished observations). The number of responders in any one cell at any particular stage is much smaller. Because the effects of the hormone are global, hormones are distributed throughout the organism by the endocrine system, the nature of an individual cell’s stage-specific response varies greatly[112,113] ecdysone may drive cyto-differentiation or it may drive apoptosis, as in most larval tissues at metamorphosis. The wide array of specific cellular effects include the modulation of cell cycle[114], the induction of apoptosis[115,116], neurite elongation[117]. Ecdysone is known to activate and/or repress both protein coding and non-coding genes, including microRNAs[118,119].

In the EcR/USP system, only the heterodimer binds ligand[111] thus USP is an allosteric regulator with respect to ligand binding by EcR. Perhaps related, DNA binding modifies ligand binding by the heterodimer[120]. The “canonical” EcRE is an inverted repeat 5’-AGGTCA/TGACCT-3’[121], but EcR/USP also binds direct repeats and inverted repeats of different spacings[122,123].

Numerous EcR/USP coregulators have been identified. Davis *et al.*[124] carried out a bioinformatic search looking for potential coregulators based on the LXXLL motif common to many hormone receptors. *Trithorax-related* (TRR) is known to interact with EcR/USP and to methylate H3K4[125]. Cryptocephal (*Drosophila* ATF4) is known to interact directly with isoform B2[126], Taiman (TAI) a p160 homolog and also Alien co-localize with the receptor[122]. There is evidence implicating the products of *Rig*, *Ash2*, *βFtz-F1*, and the histone chaperone DEK as coregulators or critical components of coregulator complexes[127-129]. *Drosophila* SMRTER (a relative of SMRT and NCoR) is known to be crucial to ligand-independent repression. There is ample evidence that remodeling factors including SWI/SNF and the NURF complex interact with EcR/USP and play key roles in ecdysone response[130-134]. There is also evidence that ecdysone induced expression is associated with acetylation of H3K23[135].

However, the central question that remains is that of specificity: How are responding genes selected from the broad array of potential targets? Few genome-wide studies have been conducted of the ecdysone response. Following initial work using subsets of genes and microarrays[136,137]. Gauhar et al.[138] employed low-resolution methods (enzymatic tagging) to provide initial data of the receptor binding sites in Kc167 cells and identified ecdysone responsive genes. Kellner *et al*.[139] showed that JIL-1 kinase is present at both enhancers and promoters of ecdysone induced genes in (Kc167 cells) and argue that it phosphorylates nearby histone H3. They find that JIL-1's presence is required for CREB-induced acetylation of H3K27 and is also required for recruitment of the 14-3-3 scaffold protein that is involved in multi-protein regulation. Shlyueva et al[140] performed the STARR-seq assay that identifies regions with enhancer activity in S2 and OCS cell lines before and 24 hours after ecdysone exposure. RNA-seq was performed in S2 cells before and after 24 hours of ecdysone exposure. These studies together provide a set of 3,415 ecdysone responsive genes from genome-wide ecdysone exposure studies from a small set of two cell lines (S2 and Kc) and two organ cultures (salivary gland and third instar larvae organs).

We performed a survey of the early transcriptional responses in 41 *Drosophila* cell lines to ecdysone transduction to identify responsive genes in a diverse set of cell types representing embryonic, larval, and adult tissues, and including the female germ line. We find that the extent of the transcriptional response – the number of genes induced at five hours, is highly correlated with the steady-state expression level of *EcR*, and not *usp* or any other TF, suggesting that the *EcR* expression level is rate-limiting in the early cellular response. The set of responsive genes differs substantially between cell lines: most genes are induced in only a few cell lines. Responsive genes cluster in neighborhoods, and tandem arrays of genes on the same strand are more likely to be jointly induced than pairs of genes transcribed from bidirectional promoters. Lastly, we identify a network of TFs that explains much of the restricted ecdysone responses, and sets the stage for subsequent, hypothesis based 'omics interrogation of this model steroid hormone.

Results

Overview of Study Design

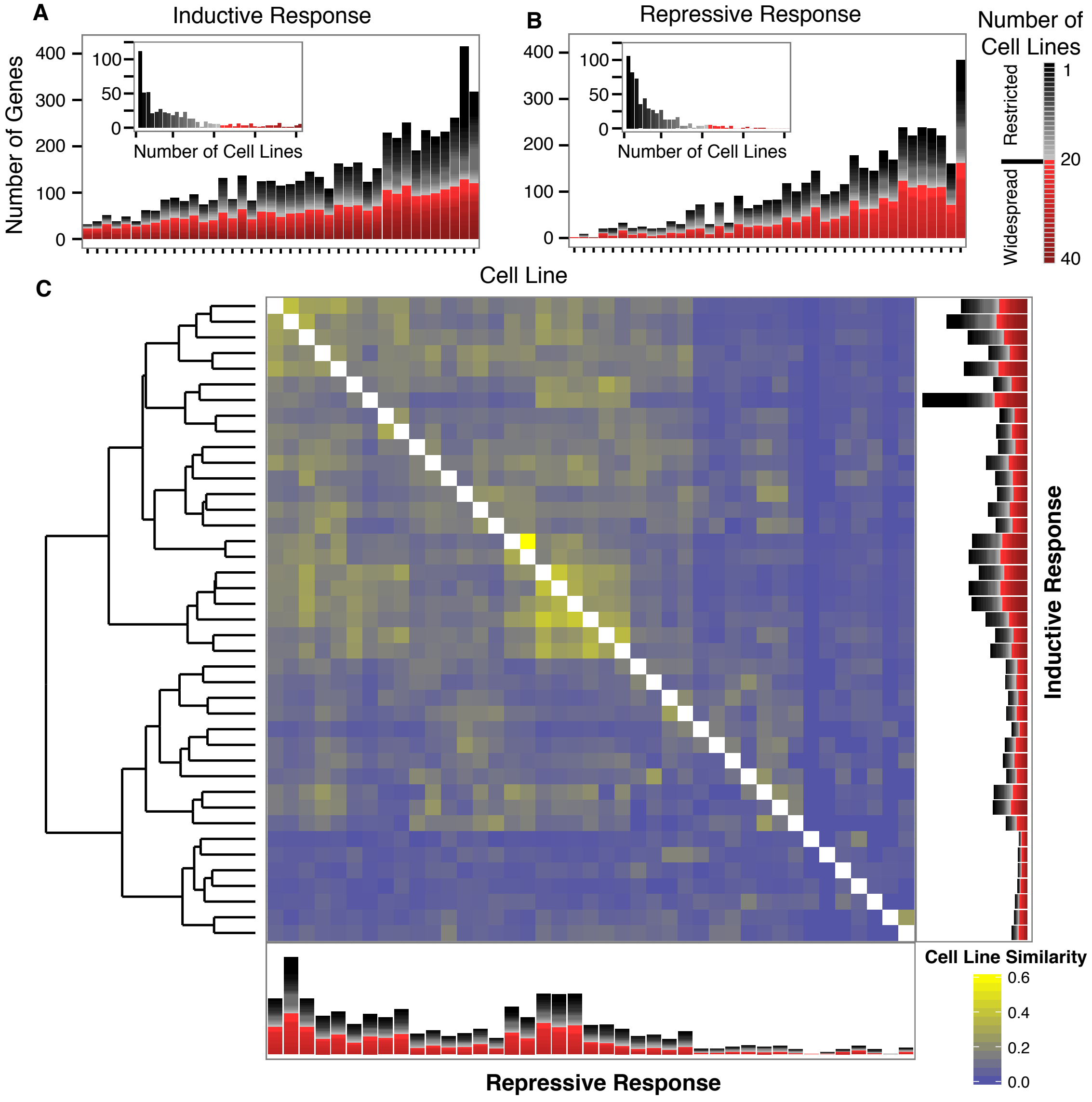
RNA samples were collected from 41 cell lines (Table 3.1) immediately before and at a five hour exposure to ecdysone (20E) at a biologically relevant concentration of 10^{-6} M. Transcription levels were measured by single-end poly(A)+ RNA-sequencing with 100 base pair (bp) reads. For four of the cell lines, CCa, Kc, MCW12 and BG3-c2, duplicate samples were collected in order

Table 3.1 41 Cell Lines
Table includes the official cell line names, short names used throughout this paper, the tissue/stage of origin, whether the cell line was included in the 25 cell lines analysis paper[1] and if the cell line is included in the time course studied here.

Cell Line	Short Name	Origin	In 25 Cell Line Paper[1]	Extended Time Course
1182-4H	1182-4H	embryonic – hapl.	🟦	
CCa	Cca	embryonic		
CME L1	L1	v. prothoracic d.	🟦	
CME W1 CL8+	Cl.8	d. mesothoracic d.	🟦	
CME W2	W2	d. mesothoracic d.	🟦	
D1	D1	embryonic		
DX	DX	embryonic		
E-CS	E-CS	embryonic		
E-OR	E-OR	embryonic		
G1	G1	embryonic		
G2	G2	embryonic		
GM2	GM2	embryonic	🟦	
GM3	GM3	embryonic		
Jupiter	Jupiter	embryonic		
Kc167	Kc	embryonic	🟦	🟦
mbn2	mbn2	larval circ. system	🟦	
MCW12	MCW12	d. mesothoracic d.		
ML83-26	83-26	embryonic		
ML-DmBG1-c1	BG1-c1	CNS	🟦	
ML-DmBG2c2	BG2-c2	CNS	🟦	
ML-DmBG3-c2	BG3-c2	CNS	🟦	🟦
ML-DmD1-c4	D1-c4	d. mesothoracic d.		
ML-DmD11	D11	eye-ant. d.	🟦	
ML-DmD17-c3	D17-c3	d. metathoracic d.	🟦	
ML-DmD20-c5	D20-c5	ant. d.	🟦	
ML-DmD21	D21	d. mesothoraic d.	🟦	
ML-DmD23-c4	D23-c4	d. mesothoracic d.		
ML-DmD4-c1	D4-c1	imaginal d.	🟦	
ML-DmD8	D8	d. mesothoracic d.	🟦	
ML-DmD9	D9	d. mesothoracic d.	🟦	
OSS	OSS	ovary -- somatic		
PR-8	PR-8	embryonic		
Pten X	Pten X	embryonic		
Ras[v12];wts[RNAi]	Ras-wts:RNAi	embryonic		
Ras[v12]-H3	Ras-H3	embryonic		
Ras[v12]-H7	Ras-H7	embryonic		
Rumi[26]Ras[v12]-4	Rumi-Ras	embryonic		
S1	S1	embryonic	🟦	
S2-DRSC	S2-DRSC	embryonic	🟦	🟦
S3	S3	embryonic	🟦	
Sg4	Sg4	embryonic	🟦	

to estimate the biological variation present in this system. Samples were taken from a time course of exposure at one, three, five and seven hours of exposure to ecdysone for three cell lines, Kc, BG3-c2 and S2-DRSC. Data for nine cell lines was biologically replicated in triplicate on microarrays.

Figure 3.1 Cell Line Responses: Breadth and Similarity
Using the thresholds defined in the text the inductive (**a**) and repressive (**b**) response within each cell line is represented. The red shaded bars represent widespread genes, responsive in greater than half of cell lines, and black shaded bars indicate restricted response genes, responsive in less than half of cell lines, as noted in the legend. Main histograms show the response of each cell line, ordered by total number of responsive genes E-OR, Ras-H3, D11, BG3-c2, D23-c4, D21, D8, MCW12, Sg4, S3, ML83-26, PR-8, D4-c1, D20-c5, G1, Rumi-Ras, Pten X, Jupiter, Ras-H7, GM3, D1-c4, 1182-4H, S2-DRSC, S1, CCa, W2, L1, D9, E-CS, DX, Cl.8, Kc, mbn2, BG1-c1, GM2, OSS, D1, D17-c3, BG2-c2, G2, Ras-wts:RNAi. Inset histograms show the total number of responsive genes by the number of cell lines responding. **C**. Cell line similarity, as measured by the Jaccard similarity (size of the intersection of responsive genes divided by the size of the union) within the restricted response, is used to cluster the cell lines as shown in the dendrogram on the left. Repressive (lower left) and inductive (upper right) response similarity corresponds to the scale indicated in the lower right. Stacked barplots are the same as those in Figure 3.11a and 3.1b, reordered by the dendrogram



Gene and exon level transcriptional quantifications were assessed using the FlyBase r5.57 annotation. Differential expression calculations were carried out using the DESeq[82] R package (Methods). Significantly induced or repressed genes were identified by applying a biological relevance fold change threshold of two and a statistical significance threshold of 0.01 differential expression adjusted p-value. A relaxed statistical significance threshold of 0.01 unadjusted p-value is applied for genes that show strongly significant tendency for induction or repression across many cell lines (Fisher's method p-value < 10^{-8}, Methods).

As reported previously[1], cell lines are remarkably transcriptionally diverse, though not necessarily representative of *in vivo* transcriptional diversity. This study includes 21 cell lines with previously characterized transcriptomes[1]. Greater than 70% of all *Drosophila* genes (11,884) are detectibly expressed at RPKM (reads per kilobase per million mapped reads) greater than one prior to ecdysone exposure. Of those, 5,846 are constitutively expressed in all cell lines, while 1,459 are expressed in only a single cell line. The number of expressed transcription factors (TFs) per cell line ranges from 406 to 450, and 595 (85% of all TFs in fly[141]) are expressed in at least one cell line. After ecdysone exposure, an additional 305 genes are expressed that were not basally detectable (at one BPKM) in any cell line, and 360 genes are constitutively inactivated (expressed at greater than one BPKM in at least one cell line before and exposure and none after exposure). We define genes with significant responses in more than half of the cell lines as "widespread", and others as "restricted". The vast majority of restricted genes are expressed in only a few cell lines; only 100 genes fall into the widespread class (Figure 3.1a-b). Indeed, few pairs of cell lines overlap in their restricted responses by more than 20% (Figure 3.1c). Therefore this *in vitro* system provides the opportunity to study diverse and distinct ecdysone response dynamics as a function of initial cell states.

The Ecdysone Responsive Transcriptome

A total of 1,645 genes are significantly transcriptionally responsive in at least one cell line. Fifty-nine TFs are induced in response to ecdysone, and 35 of these are responsive in five or fewer cell lines. Several of these are known ecdysone responsive TFs, while many are newly identified and point to new hormone-responsive pathways. Among responsive genes, the most strongly

		Fold Change Threshold						
		log2(1.2)	**log2(1.5)**	**log2(1.66)**	**1**	**log2(2.5)**	**log2(3)**	**log2(4)**
P-Value Threshold	**0.001**	0.9862	0.9862	0.9862	0.9889	0.9938	0.9900	0.9714
	0.0025	0.9916	0.9916	0.9916	0.9954	0.9954	0.9898	0.9684
	0.005	0.9938	0.9938	0.9938	0.9985	0.9955	0.9885	0.9665
	0.01	0.9950	0.9950	0.9960	1.0000	0.9949	0.9875	0.9659
	0.025	0.9929	0.9929	0.9966	0.9978	0.9922	0.9862	0.9633
	0.05	0.9915	0.9934	0.9960	0.9949	0.9890	0.9846	0.9633
	0.1	0.9899	0.9923	0.9913	0.9841	0.9834	0.9809	0.9629

Table 3.2 Responsive Gene Count Robustness to Thresholds
Table shows Pearson correlations between count of responsive genes with chosen thresholds and a wide range of biologically applicable fold change values and statistically meaningful significance values.

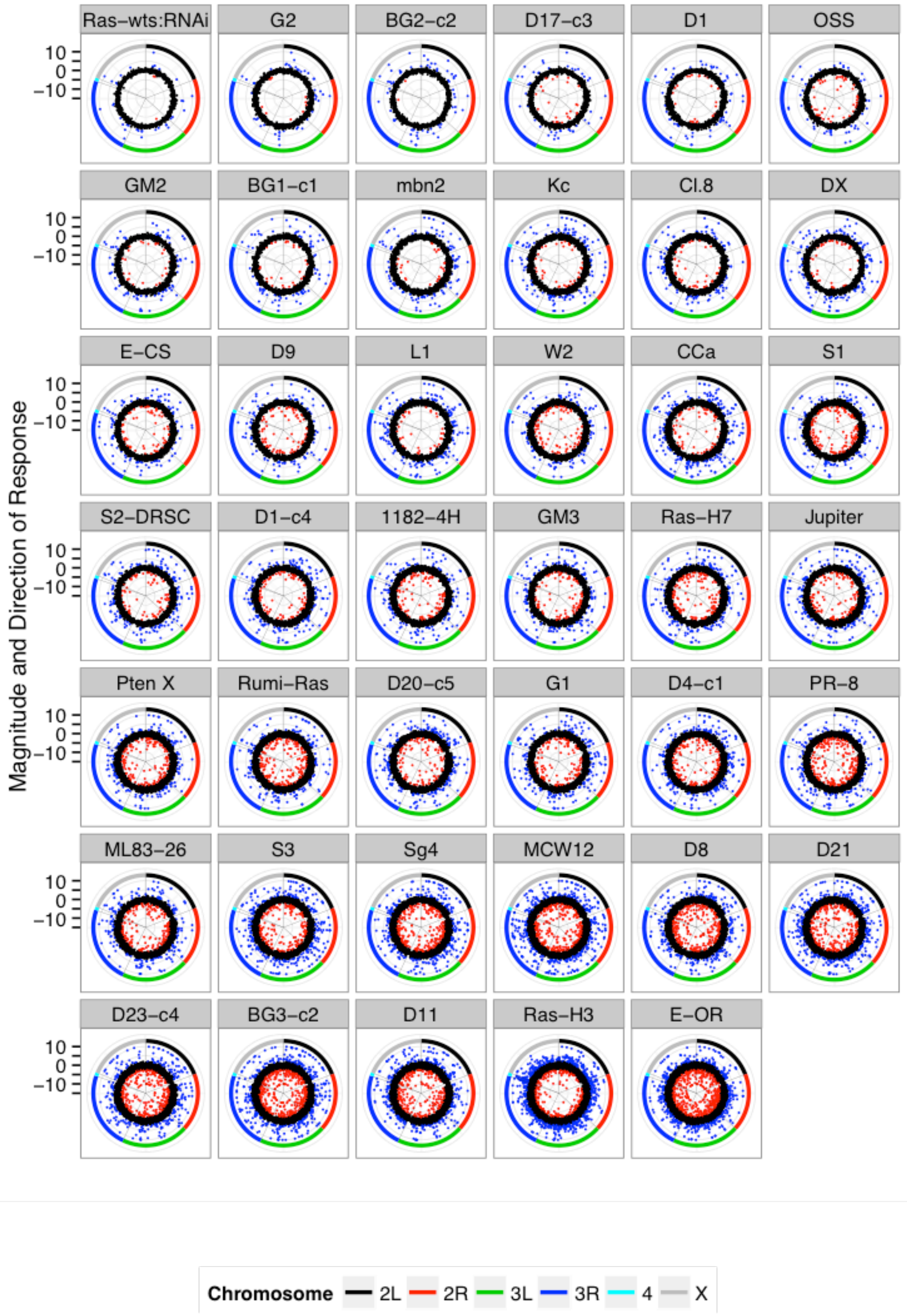


Figure 3.2 Genomic Locations of Differentially Expressed Genes
Each panel represents the ecdysone responsive behavior for a cell line (ordered by the total number of responsive genes). The genomic position is represented on the radial axis. The magnitude and direction of response, as measured by the negative log10 of the differential expression p-value times the direction of response, is represented on the polar axis. Red and blue points are significantly repressed and induced, respectively, in response to ecdysone

enriched GO term is "imaginal disc-derived wing morphogenesis" (adj. p-value < 0.001). When induced and repressed genes are analyzed separately, the GO terms “protein catabolic process”, “salivary gland autophagic cell death” and “axon guidance” are enriched among induced genes, and “mesoderm development” among the repressed.

It is well known that some tissues and cell types are more responsive to ecdysone than others. We measure the responsiveness of a cell line as the count of genes significantly induced or repressed five hours after induction, and refer to this as the Responsive Gene Count (RGC). While RGC is threshold-dependent, the rank-order of cellular responsiveness is well preserved across a broad range of biologically and statistically meaningful parameterizations (Methods; Table 3.2). The RGC varies by two orders of magnitude across cell lines and is driven by responsive genes with highly restricted expression patterns (Figure 3.1 and 3.2).

The RGC does not correlate with the count of basally expressed TFs (r ~ -0.04, p = 0.85) and correlates only weakly with the total number of genes expressed per cell line (r ~ -0.35, p = 0.03). We assessed association of the RGC with the basal expression level of each gene in the genome in a multiple testing setting (Methods). Among all genes, the expression level of *EcR* is by far the most strongly correlated with RGC (r ~ 0.71, FDR < 0.001, minimum FDR for other genes > 0.03). We further assessed this observation within both induced and repressed genes (Figure 3.3) and find that *EcR* expression level is strongly correlated with both. In fact, *EcR* is the only gene highly statistically significantly (p-value < 0.01) correlated with the number of both induced and repressed genes (*EcR* FDR < 0.001, minimum FDR for other genes > 0.02). The *EcR* heterodimer partner, *usp*, exhibits weak correlation with RGC (r > 0.35, FDR > 0.9; Figure 3.3).

The residuals left after correcting for the effect of *EcR* expression level on RGC, as well as on the counts of induced and repressed genes (separately) show little correlation with any gene (maximal correlation < 0.61; minimum FDR >

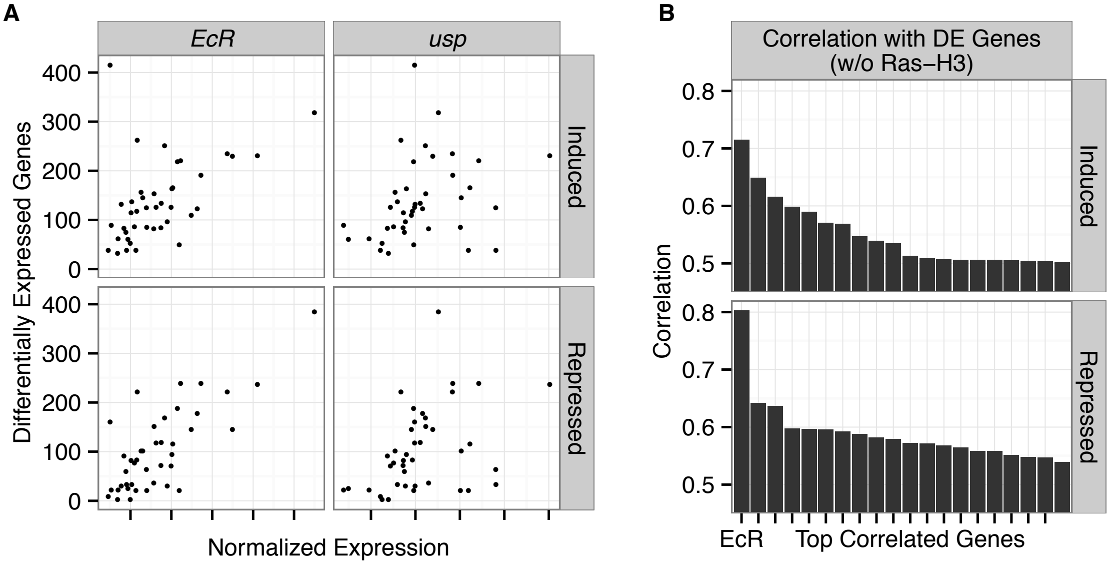


Figure 3.3 Ecdysone Receptor Expression Correlations
A. Scatter plots comparing the global number of induced (upper panels) and repressed (lower panel) genes with the normalized expression of the canonical ecdysone heterodimer receptor (EcR and usp panels). **B.** Barplots showing the correlation of normalized expression with number of induced or repressed genes for the genes with 20 highest correlations. EcR shows the highest correlation for both induced and repressed genes

0.6). Taken together, these results indicate that *EcR* titer is rate limiting for the transcriptional response to ecdysone.

Correspondence with Previous Studies

We reviewed existing literature on genome-wide transcriptional responses to ecdysone and found that 3,415 genes (1,755 induced and 1,800 repressed) have been identified as ecdysone-responsive. Of these, we observe only 730 among our 1,645 (overlap significance, hypergeometric p-value < 10^{-100}). Our set corresponds most similarly with the study from Shlyueva *et al.*[140], which also collected RNA-seq data from S2 cells (although a different isolate) collected before and 24 hours after ecdysone exposure. We find that 70% of genes induced in more than half of the cell lines surveyed are also reported induced in Shlyueva *et al.* We see minimal correspondence with samples taken from organ cultures, which conflate the responses of many distinct cell types.

Widespread Response to Ecdysone

The 100 genes with widespread responses include 68 induced and 32 repressed in more than half of our cell lines. Five genes, *Hr4*, *Hormone receptor-like in 46 (Hr46)*, *Ecdysone-induced protein 75B (Eip75B)*, *CG44004* and *bip1*, are induced in all 41 cell lines. All five have been previously identified in other genome-wide surveys of ecdysone exposure response[136-138,140]. There are no genes repressed in all cell lines, with *fruitless (fru)* being the most constitutively repressed in 33 of the 41 cell lines. More broadly, widespread induced genes are significantly enriched for biological GO terms “metamorphosis,” “salivary gland cell autophagic cell death” and “steroid hormone mediated signaling” (p-value < 0.001), as expected, since the majority of these genes have been previously reported in other studies of ecdysone response. I note that these genes may not respond in the different transcriptional and chromosomal landscapes of all tissues and developmental stages, but these observations suggest that their responses are global in *Drosophila* cells.

The promoter regions of widespread induced genes, defined as 500 bp upstream of each TSS, are 36% enriched over background for the EcRE motif (p-value < 0.02) representing the most significantly enriched of all known motifs in the HOMER library[104] (Methods).

Notably, 11 widespread induced genes lack GO annotations, and these include four long non-coding RNA (lncRNA) genes, *CR43432*, *CR43626*, *CR45391* and *CR45424*. These lncRNAs are each expressed in the salivary gland and fat body and at low levels in most other tissues. Only *CR43432* is expressed at high levels during development, with maximal expression (>100 RPKM) in the 4-14 hour embryos. This is in contrast to the majority of lncRNAs in *Drosophila* (and indeed mammals), which are expressed predominantly in tissues of the nervous system and the gonads[7,21]. Notably, this lncRNA is induced at levels comparable the well-known response *polished rice*, which encodes short (11 aa) peptides critical for ecdysone transduction in the epidermis[142]. *CR43432* encodes three short, ultra-conserved ORFs, and hence constitutes a candidate protein-coding gene.

The set of widespread repressed genes is much smaller than the set of induced genes, as is the overall repressive response in most cell lines. There are no statistically enriched GO terms among this set of genes.

Diversity of the Ecdysone Response: The Restricted Set

We find that 93% (a total of 863) of induced genes and 96% (a total of 765) of repressed genes are significantly ecdysone responsive in fewer than half our cell lines. These restricted responses form the molecular basis of the diverse transcriptional and physiological effects that the ecdysone hormone induces throughout developmental and within distinct cell types. A total of 400 and 241 genes are induced and repressed respectively in exactly one cell line, with a large fraction (31%) responding only in the Ras-H3 cell line, an outlier in this study (Methods).

Eighty-three genes are significantly induced in some cell lines and significantly repressed in others. Sixteen are induced and repressed in at least two cell lines. One striking example is the TF, *CG9932*, which is significantly induced in five cell lines and significantly repressed in six. *CG9932* is differentially expressed across development with peaks at 20 hours and late L3 stage and shows strong expression in embryonic fat body and salivary glands. It is likely that some promoters respond in distinct directions based on prior epigenetic state. This phenomenon has also been noted in mammalian response to glucocorticoids[143].

The restricted set includes genes that respond weakly in several cell lines. Some of these genes do not pass our statistical criteria in any single sample, but by aggregating information across cell lines we obtain sufficient power to confidently annotate weak, reproducible induction or repression (Methods). A total of 635 genes, 335 induced and 300 repressed, of this type are present genome-wide. Weakly induced genes are strongly enriched for GO terms including “protein binding”, "vesicle-mediated transport" and "macroautophagy" and weakly repressed genes are enriched for "rRNA processing" and "calmodulin binding". Weakly induced genes are found within 2.5 kb of a significantly induced gene more often than expected by chance (84% enrichment, binomial p-value $<10^{-5}$). This pattern does not hold for weakly repressed genes.

Responsive-Proximal Genes Tend to Respond Similarly

The proximity of weakly induced to strongly induced genes indicates that local genome architecture may play a role in ecdysone response. To explore the effects of proximity, we modeled the average fraction of neighboring responsive genes (as well as the direction of response) as a function of genomic distance (Figure 3.4a, raw data shown in Figure 3.5). Additionally, we stratified these proximity relationships based on the positional relationships of gene-pairs, i.e. upstream versus downstream and same versus opposite strand (Methods). We then smooth these values to produce an interpretable plot of the average effect of local genome architecture on neighborhoods including genes that respond to ecdysone.

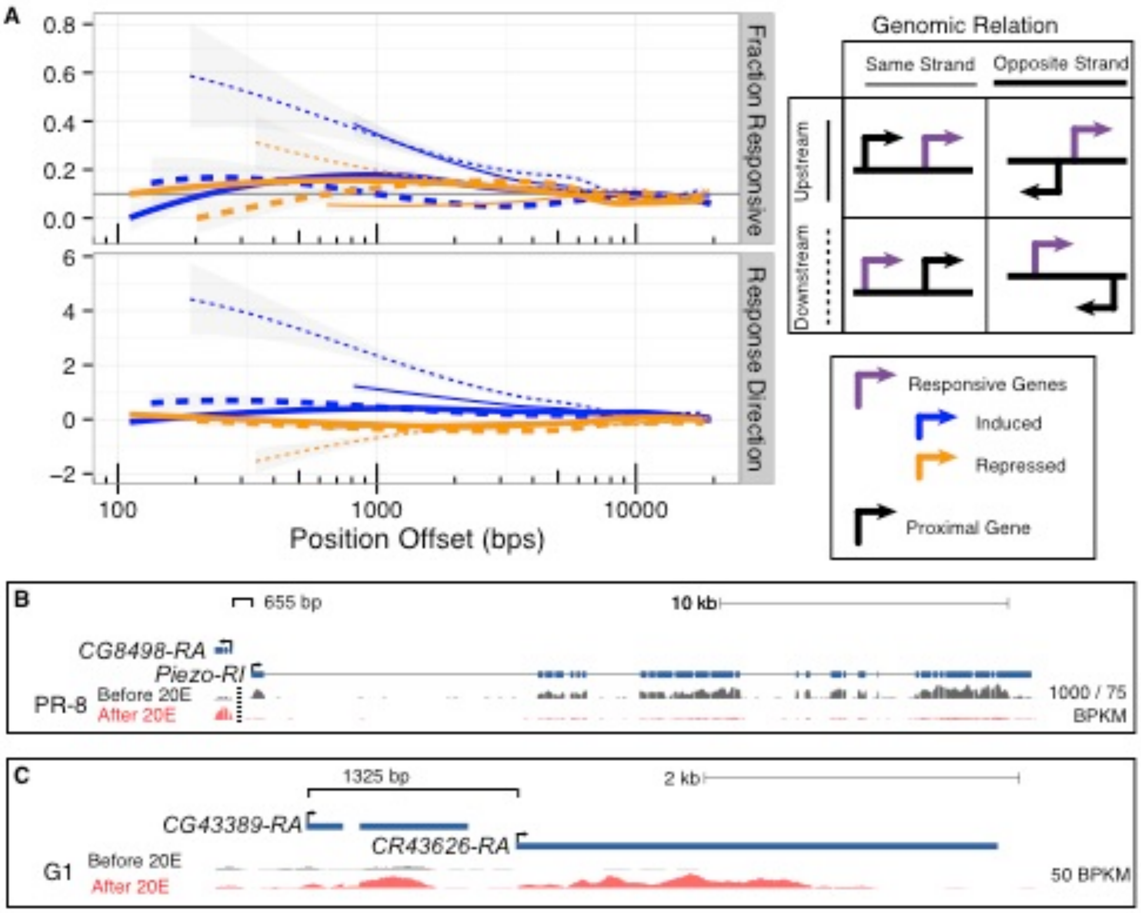


Figure 3.4 Genomic Positional Dependence of Ecdysone Response
A. Each line represents the smoothed either fraction of responsive genes (upper panel) or response direction (lower panel). Response direction is measured by the mean of the magnitude and direction of response for genes within a moving window across genomic position. Genes within 20 kbp of a significantly responsive gene contribute to smoothed lines grouped by their genomic positional relation, see key. Red and black lines summarize genes proximal to a repressed or induced gene respectively. In the first panel the blue line represents the genome-wide average as measured by the 10kb to 20kb proximal region. **B.** The CG8498 and Piezo locus is an example of divergent promoters responding in opposing directions. This exemplifies the trend shown in Figure 3.4a where divergent promoters do not tend as strongly to respond consistently. Note that since CG8498 and Piezo are expressed at different levels, the maximal height to the left of the vertical dashed line is 1000 BPKM and the maximal height to the left is 75 BPKM. **C.** The CG43389 and CR43626 locus is an example of “operon-type” promoter structured genes responding in the same direction. This exemplifies the trend shown in Figure 3.4a where “operon-type” promoters tend to respond consistently, particularly for induced genes

We find that genes proximal to responsive genes tend to be responsive (p-value < 10^{-15}; sample t-test against median for genes between 10 kb and 20 kb) and additionally they tend to respond in the same direction (p-value < 10^{-100} induced and p-value < 10^{-19} repressed; one-sample t-test). The response of opposite-strand gene pairs (bidirectional and convergent) is less coordinated than genes with operon-type architecture (same strand pairs, p-value < 10^{-10}). One striking example of a divergently responsive bidirectional pair is *Piezo* and *CG8498*. The TSSs of these genes are separated by only 655 bps (Figure 3.4b) and the responses in most cell lines are strong and in opposite directions. *CG8498* is a widespread induced gene with 38 cell lines responding significantly

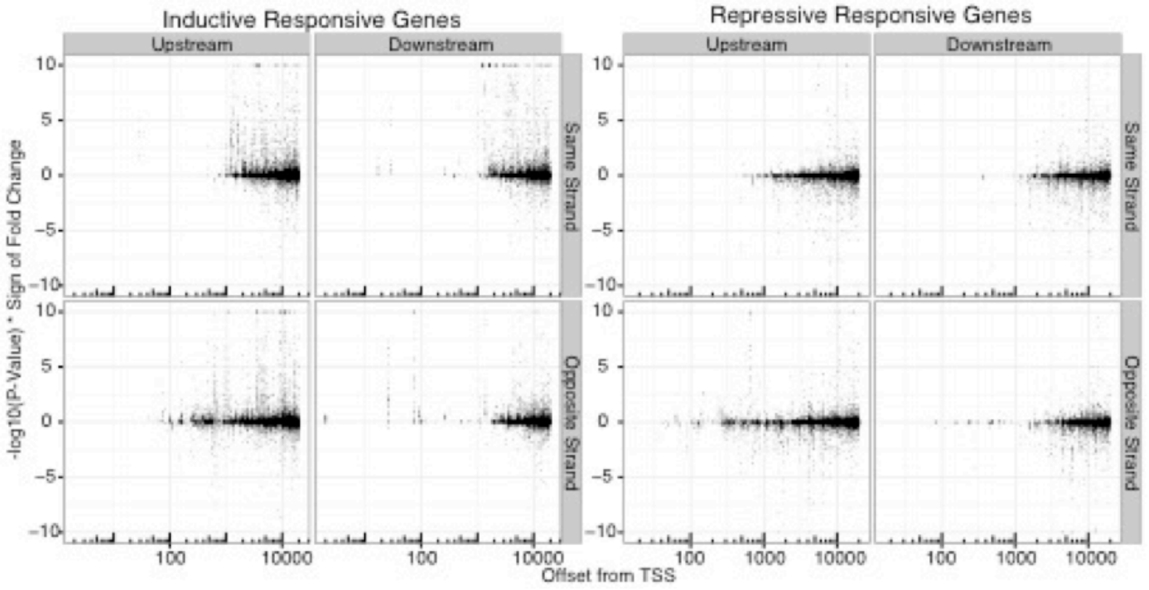


Figure 3.5 Raw Genomic Positional Dependence Data
Each point represents the response, as measured by the transformed p-value; that is the negative log 10 of the DESeq p-value times the sign of the log 2 fold change, of a gene proximal to a significantly responsive gene. Genes proximal to significantly induced gene are shown on the left and genes near repressed genes are shown on the right. The facets show the responses grouped by promoter architecture (upstream vs. downstream and same versus opposite strand). The data from each panel is used to produce each line in Figure 3.4a

and all but two cell lines showing repression of *Piezo*. One example of co-responsive genes in an operon-type configuration is the pair *CG43389* and the non-coding gene *CR43626*, which are significantly induced in 32 and 35 cell lines respectively (Figure 3.4c).

Taken together, these results indicate that response to ecdysone may involve or depend upon local chromatin organization or modification. Indeed, the bromodomain protein *toutatis* (*tou*), a gene involved in chromatin remodeling[144], is strongly induced in 15 cell lines. Both an acetyllysine binding domain (bromodomain) and a methyl-CpG binding domain exist in *tou*, a BAZ (bromodomain adjacent to zinc finger) protein. As a class, these genes appear to be involved in the integration of information encoded in DNA methylation and post-translational histone modifications[145]. The involvement of acetyllysine binding factors is consistent with previous reports demonstrating that ecdysone transduction impacts H3K23 acytelation[135].

Exon Level Analysis

Along with gene level events the induction and repression of specific exons, predominantly promoter switching events, has been previously reported[140]. The widespread responsive gene *Eip75B* shows the strongest exon level event in the genome across all cell lines (p-value < 10^{-100}), consistent with previous reports[140,146]. In total 35 genes show significant exon level induction events and 31 genes show significant repression events (thresholds: adj. p-value < 0.01 and 50% fold change). Six genes, including *Eip75B* show significant induced and repressed exons representing promoter-switching events.

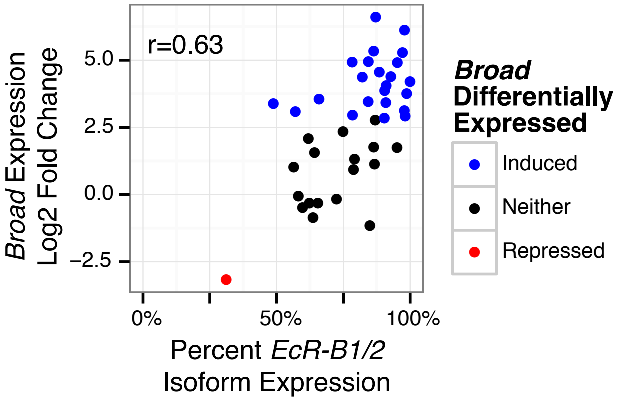


Figure 3.6 Ecdysone Receptor Isoform Correlation
Scatter plot of showing the correlation between *EcR* isoform expression ratio and the log2 fold change of expression *broad* upon ecdysone exposure. *Broad* response shows the largest dynamic range amongst the genes with the highest correlation to *EcR* isoform ratio

Ecdysone Receptor Isoforms

Cherbas *et al.*[147] demonstrated that the alternative *EcR* isoforms (*EcR-A*, consisting of transcripts A, D and E and *EcR-B1/2*, consisting of transcripts B, C and G) play important roles in development. Gonsalves *et al.*[137] showed that *EcR* isoforms are differentially expressed between Kc cells and salivary gland cells, indicating that alternate *EcR* isoforms elicit different transcriptional responses. Cell lines in this study show a wide range of *EcR* isoform expression. The length normalized fraction of the *EcR-B1/2* isoform expression (Methods) ranges from 0.31 (BG3-c2 cell line) to 1 (S2 cell line) with most cell lines expressing predominantly the *EcR-B1/2* isoform. We do not see strong correlation to the total number of induced or repressed genes, or the residuals after correcting for the main *EcR* expression effect, but we do observe strong correlation between the *EcR-B1/2* isoform fraction and the expression of many individual genes. The most significantly correlated genes are *gliolectin (glec)*, *squeeze (sqz)*, *CG5059*, *Eip55E* and *broad* (*br)*. Only *glec* and *Xbp1* show increased expression with increased *EcR-A* isoform levels amongst the top 10 most correlated, consistent with a predominantly repressive role for *EcR-A* (Wilcox rank-sum p-value $< 10^{-10}$). Of the highly correlated genes, the TF *br* shows the largest dynamic range (two orders of magnitude); twenty-three cell lines show significant induction and one, BG3-c2, shows significant repression (Figure 3.6).

Transcription Factor Expression Predicts Restricted Responses

While the global level of responsiveness (RGC) is well characterized by *EcR* titer, other effects are clearly at work producing diverse restricted responses. TFs are likely candidate *EcR* cooperative factors. We developed a statistical machine-learning model aimed at identifying these factors: we used basal expression levels and TF binding motifs to predict restricted responses. Specifically, we used normalized TF expression levels as our covariates, and a TF's expression level to affect the prediction of a gene's response if (and only if) we observe an instance of the cognate binding motif in the gene's promoter (Methods). We also supplied the model with information about the basal expression level of each gene since genes with low expression are intuitively less likely to be repressed, and genes expressed at high levels are less likely to be induced by orders of magnitude.

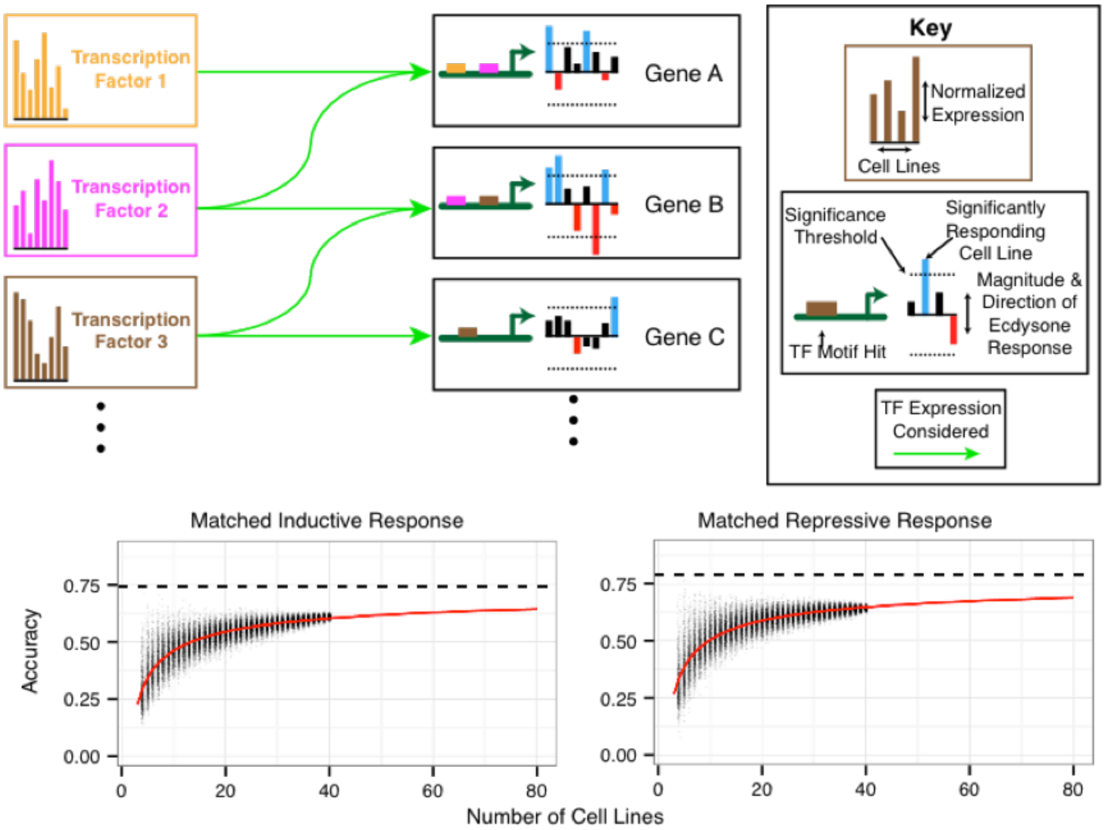


Figure 3.7 Graphical Abstract of Restricted Set Response Prediction
A. Transcription factor expression levels (left bar plots) combined with known binding motif preference (green arrows indicate promoter motif match) are used to predict whether a cell line is significantly induced (blue bars) or most repressed/lowest fold change (red bars) at a locus for genes in the induced restricted set and similarly for the repressed restricted set. Random forests are trained and used to predict response direction on either left out sets of cell lines or left out sets of genes. **B.** Scatterplots showing predictive power when cell lines are left out as a test set for the matched inductive (left panel) and the matched repressive response (right panel). The x-axis indicates the number of cell lines included in the sub-sampled data (note that points are randomly shifted in the x-direction for visual clarity) and the y-axis indicates the average predictive accuracy when each cell line in the sample is left out as test set. Fitted lines in red show average predictive accuracy. Fitted horizontal asymptotes (dashed lines), representing the average predictive accuracy with infinite cell lines available to predict the response of an unseen cell line, are 0.74 for the matched inductive response and 0.79 for the matched repressive response

We fit models for induction and repression. The inductive model will be explained here, as the construction is analogous for the repressive model. We construct mutually exclusive training and test sets by selecting gene-cell line pairs, as follows. First, we select a gene with an induced, restricted expression pattern. Then we identify the subset of cell lines in which it is responsive. These data points are gene-cell line pairs, and will enter either the training or the test set. For the same gene, we select an equal number of cell lines with the most opposed responses. This means that if a given gene is induced in five cell lines and repressed in twelve, data points corresponding induction levels in the five induced and also the five most repressed cell lines will enter either our training or test sets. We fit the model to predict which gene-cell line pairs correspond to inductive versus repressive responses, and test on held-out data. A notable feature of our modeling strategy is that we hold out entire cell lines, so when we assess our models predictive performance, it is assessed on cell lines it has not previously seen. This ensures that the rules we learn about ecdysone response are generalizable. We find a predictive accuracy on held-out cell lines of 61% and 64% for induced and repressed genes respectively, indicating that we have weak, but significant (binomial p-value < 10^{-15}) power to predict the direction of a gene's response.

We used feature selection to compute the relative importance of each covariate in our model (Methods). Our measure of importance for a given covariate is the average percent loss of predictive accuracy when the values of the covariate are randomly permuted (as described,[148]). We find that, for models of both induction and repression, the rank of basal gene expression level is the most important covariate. This variable importance may be due to the biological inability to suppress a gene that is already not expressed or to increase the expression of a gene already highly expressed. I note that this observation may be a statistical detection artifact, but may also represent true biological insight (i.e. the response to ecdysone must achieve a certain level of expression as opposed to a particular fold change). For the repression model, we see that gene rank is 2.5 times more important than the most important TF. In the induction model, the gene rank is only 1.3 times more important. However, in both models a number of TFs are also statistically significantly important for the prediction problem. Several of the most important TFs have known roles in ecdysone response, including *br* and *Eip74EF*. TFs not previously implicated in ecdysone are also important in both prediction problems, including *longitudinals lacking (lola)* and *Chorion factor 2* (*Cf2)*. We note nearly all covariates have positive importance values in both prediction problems, indicating that the cooperation of many factors may be involved in the restricted responses.

This approach also enables us to assess the power of transcriptional data along with promoter and TF binding site information for elucidating the basis of hormone responses. We fit and assessed this model successively using different numbers of cell lines, between four and forty. For each count of cell lines, we randomly selected (with replacement) 1000 training/test set combinations (Figure 3.7). For small numbers of cell lines, the model does worse than random guessing in the test set, and this is not surprising: the responsiveness defined as the RGC of cell lines varies by two orders of magnitude – hence small subsets of cell lines provide poor generalizability. In particular, the marginal distribution of induction versus repression is, on average, substantially different between the training and test set for small sample sizes. For larger samples, as above, we find that the average accuracy reaches 61% for the induced genes and 64% for the repressed genes. We extrapolate from this a theoretical maximal accuracy (asymptote) for identifying the response of an unseen cell line of 74% and 79% (respectively) given an infinite number of cell lines in the training set (Methods). I note that an alternative model comprised of only transcriptional data may provide increased predictive power.

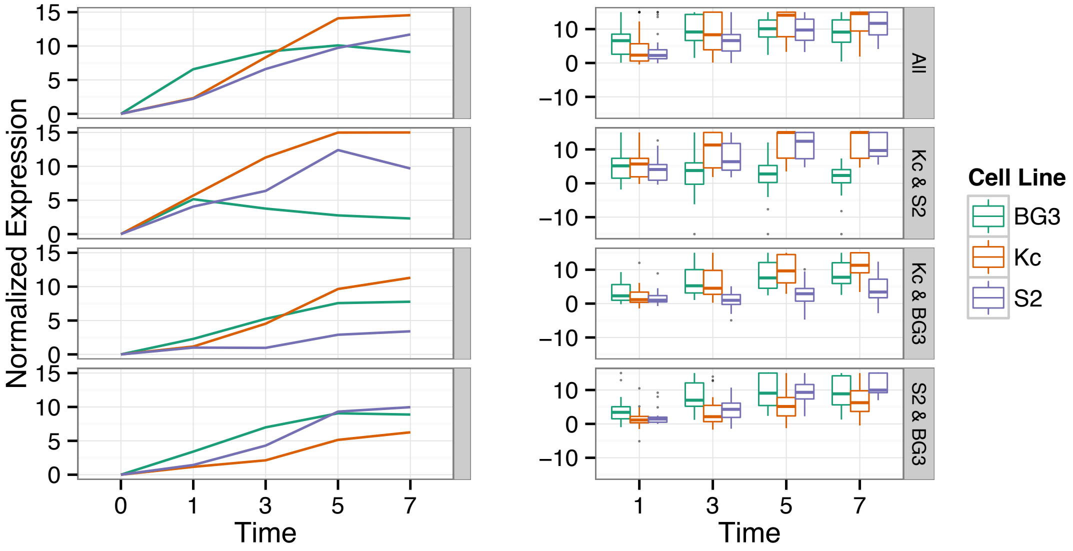


Figure 3.8 Extended Time Course. Four panels show normalized expression (Methods) over an extended time course of zero, one, three, five and seven hours for subsets of genes identified as significantly ecdysone responsive in the zero to five hour analysis for these three cell lines. The left panels show the median normalized expression at each time point and the right panels show the full set of normalized expression values at each time point

Dynamics of Extended Temporal Response

In addition to the response detected at the five hour time point, we explored the response to ecdysone for an extended temporal range, including one, three, five and seven hours after exposure for three cell lines, BG3-c2, Kc and S2. We normalized responses, setting the basal expression level of each gene to zero, and then quantified changes at subsequent time points in multiples of fitted standard deviations (Methods). This intuitive representation captures much of the same information as z-scores, and has the advantage that each gene is set to the same (zero) value in basal conditions. We used this representation to identify structures in both the scales and directionalities of temporal responses (Figure 3.8).

Among genes that show significant induction in all three cell lines, genes in the BG3-c2 cell line respond systematically more rapidly. Expression levels in this cell line also quench (level off) more quickly than in S2 and Kc, which show steadily rising expression through all seven hours.

There are 15 genes that are responsive at five hours after induction only in Kc and S2 cells, and these show a strong and consistent expression pattern in the BG3-c2 time course: they are responsive at one hour, and then reduce expression level at each time-point thereafter. Six of these 15 genes are widespread responsive (p-value < $1e^{-10}$), including *br* and *Eip71CD.* The rapid (at 1 hour) induction of these genes indicates that BG3-c2 is in a more ecdysone sensitive basal state. Notably, the total responsiveness of these three cell lines differs substantially: the RGC value for BG3-c2 cells is more than twice that of S2 or Kc cells. Taken together, these data indicate that the observed increased responsiveness of BG3-c2 cells at five hours may be due to an accelerated ecdysone response relative to other cells.

Discussion

The overall responsiveness of a cell line to ecdysone induction correlates strongly with EcR mRNA titer. This suggests that the availability of ECR, not USP or other factors, is the primary rate-limiting step in the ecdysone response. Furthermore, the EcRE is enriched in the promoters of genes in the widespread, but not the restricted class. Widespread genes are predominantly immediate and direct targets of the EcR complex and we provide strong evidence that the response of restricted genes is conditionally dependent on additional factors. Additionally, EcR isoform titer is a powerful predictor of response direction and magnitude for several genes. Among these, the TF *br* exhibits the largest dynamic ecdysone response range (three orders of magnitude), suggesting strong dependence on ECR isoform as well as net titer.

The expression of restricted genes correlates significantly with predictors based on TFs together with their discovered sequence specific motifs. In particular, *br* and *lola*, are strong predictors of restricted responses across all 41 cell lines. We estimate that more than a hundred TFs are needed to achieve maximal predictive accuracy indicating that the interaction of many TFs and co-factors convey divergent gene responses.

The organization of responsive genes in the genome supports the idea that epigenetic state, including chromatin context may be important: weakly responsive genes tend to be near more significantly responsive genes in "responsive neighborhoods". These neighborhoods may be sensitized to induction or repression via chromatin modifications, e.g. the generation of docking sites for chromatin-binding transcriptional co-factors. Responsive neighborhoods tend to be organized in an "operon" configuration, meaning that neighboring genes transcribed on the same strand tend to be induced or repressed together, with the upstream gene tending to respond more strongly than the downstream neighbor. This is remarkable, since genes transcribed from bidirectional promoters are often under independent control, indicating that the spatial resolution of ecdysone transduction along the genome is on the order of hundreds of base pairs. Ecdysone induction spreads directionally along the genome, in the direction of transcription, suggesting a role for Pol2-associated chromatin modifications. Furthermore, the widespread and early induction of chromatin remodeling factors like *tou* supports the idea that the secondary targets may be determined in part by chromatin remodeling, and that co-factors, in addition to TFs, play essential roles in specifying cellular responses.

Extended time courses with additional early and late time points revealed that the less responsive cell lines may simply be responding more slowly, where EcR titer determines not only the extent of the response but also the timing. Our data are consistent with a kinetic model of transduction, which would predict that the widespread response constitutes the primary and broadly conserved targets of EcR, and that the timing of the onset of secondary inductions alters the trajectory of cell fate and specifies much of the restricted responses. Assessing this possibility will require perturbation experiments where the EcR titer is manipulated and time courses are taken at finer resolution.

We have produced the most expansive transcriptional ecdysone-response atlas of distinct cell states to date. These data point towards the molecular basis of the diverse responses of *Drosophila* cells to ecdysone. It is clear that transcriptional data as analyzed by our model will not fully elucidate the transduction of this model steroid: even given sequencing of every cell in the fly, it is unlikely that we could exceed 79% accuracy in our *in silico* model of response based on our analysis of the asymptotics of our predictive power. This is not surprising; we are missing information about enhancers, chromatin state, and the 3D structure of the nucleus. Recent studies of ECR binding patterns reveal that only 30% of binding sites are promoter-proximal[140]. Though an alternative model including only transcriptional information may provide greater predictive accuracy, we postulate that a minimally sufficient dataset that could fully elucidate the molecular machinery of the ecdysone response will include *in vivo* binding site maps for relevant transcription factors, extensive chromatin state profiling, enhancer activity data and structural information. Furthermore, since chromatin-binding co-factors are as yet largely unknown, a second round of primary data production will likely be needed to identify specific actuators once an initial integrative model has been formed. Early responders also include RNA binding proteins and many small RNAs. Understanding secondary effects will require models of post-transcriptional regulation. Additionally, multi-cellular effects, including auxiliary cell-cell signaling and mechanotransduction across fields of cells in tissues may play essential roles, and these will likely be missed *ex vivo*.

It is important to recall that the large number of ecdysone-responsive genes in each cell line is in apparent contrast to the results of genetic interaction experiments ([147] and unpublished observations), which suggest that the primary ecdysone response in individual tissues is controlled by a relatively small number of critical genes. The latter hypothesis is based on genetic experiments, which show that each ecdysone receptor isoform, as well as many co-regulators, are required in only a few tissues. While it is possible that cell lines are aberrant in displaying rapid ecdysone regulation of a large number of genes, studies of primary organ cultures[137] and whole animals[136] support the notion that very large numbers of genes are ecdysone-responsive. It may be that most ecdysone-responsive genes are not “critical” to the cell’s response, in the sense that their response is not required for the cell to achieve its developmental fate. Such a phenomenon would be expected if homeostatic mechanisms buffer the effects of most fluctuations in transcription, as has been described for the very well-studied folate cycle of vertebrates, where biochemistry and mathematical modeling indicates that most components of the pathway are remarkably insensitive to variations in the level of individual enzymes and substrates[149]. Thus a change in the level of a particular transcript may be critical for the developmental response of a cell to ecdysone or may have no detectable consequences for the physiology of the cell; functional studies are required to distinguish these possibilities for any given gene.

It may also be that, if the response of a particular gene is critical in one tissue, and merely not harmful in another, that gene may show a strong ecdysone response in cells for which the response is of little or no physiological significance. We find only a few genes with promoters that are both inducible and repressible via ecdysone, indicating that the evolution of this plasticity may be difficult. We must also consider the possibility that some ecdysone responses are accidental in the sense that the receptor binds (directly or indirectly) to a promoter or enhancer sequence that evolved for other purposes. Or, alternatively, as is indicated by the co-responsiveness of proximal genes transcribed on the same strand, that local chromatin modifications needed to active a target gene have off-target effects in the genomic neighborhood. A response that is physiologically unnecessary but not harmful will be maintained if the components of the response are needed for other purposes.

This survey provides a foundation for understanding the context dependence of steroid hormone signaling. We produced a draft map of critical transcription factors and co-factors important for both the early and secondary responses. Hypothesis-based experimentation along with additional ‘omic studies will be needed if ecdysone is to be the first fully mapped metazoan hormone.

Methods

Differential Expression Analysis

Gene and exon level counts were computed using the python package HTSeq[150] (version 0.6.1p1) using the FlyBase annotation version r5.57[44]. Exon level counts were analyzed using the DEXSeq R package[151] (version 1.12.1). Exon level ecdysone exposure effects are reported only for a model fit across all cell lines. Thus exon p-values and fold changes are not reported for each cell line individually. Gene level analysis is completed using the DESeq R package[82] (version 1.18.0). As only a portion of the samples were completed in biological duplicate, gene level dispersion estimates were made using the replicated samples and applied to all cell lines. The statistical assumption underlying this analysis is that gene-level biological dispersion is consistent across cell lines.

EcR Isoform Analysis

In order to investigate the effects of the differential isoforms at the *EcR* locus, exons were assigned to either the long *EcR-A* or short *EcR-B1/2* isoforms. Constitutive exons were ignored in this analysis. The ratio of *EcR-B1/2* to *EcR-A* is calculated for each cell line as the ratio of the length normalized (total collapsed isoform-specific exonic regions) number of reads assigned to the *EcR-B1/2* exons to the total length normalized read count to isoform specific exons. These ratios are compared to RGCs in the results.

Identification of Widespread and Restricted Ecdysone-responsive Genes

In order to leverage the breadth of cell lines examined in this study while identifying ecdysone-responsive gene we employ a threshold based on the responsiveness across all cell lines. This threshold allows genes that show a trend toward significance, while possibly not achieving standard statistical significance within any particular cell line, to be confidently identified as ecdysone-responsive. Formally this is measured by the Fisher's Method test for trend in significance across all cell lines.

Each gene is associated with two Fisher's Method values corresponding to a trend towards induction and repression across all cell lines. We note that it is possible to achieve significance in both responsive behaviors under this structure. The procedure to produce these tests is analogous for induction and repression, so we will describe the procedure for induction here. In order to compute a Fisher's Method significance value the p-values produced by the differential expression analysis are used to construct a ranked list of induced genes within each cell line. Genes that are repressed are assigned a p-value of one and thus are tied at the bottom of that cell line's rank list. These rank values thus represent a uniform marginal distribution for each cell line, as required in order to apply Fisher's Method. For each gene the rank within each cell line is combined using Fisher's Method. Genes that tend towards the top of the rank list in many cell lines will produce significant values, while gene randomly distributed amongst each list will produce less significant values.

Two types of thresholds, biological relevance and statistical significance, are applied to each gene within each cell line. The biological relevance threshold is defined by a fold change upon ecdysone exposure greater than two fold (inductive or repressive). Statistically significant responsive genes are those that achieved either an adjusted p-value less than 0.01 regardless of Fisher's Method p-value or an unadjusted p-value of 0.01 and a Fisher's Method p-value less than 10^{-8}.

Enrichment of Motifs

Enrichment of motifs within promoter region DNA sequence was carried out using the homer2 program [104] with the "known" command against the supplied all.motifs database wihch contains the EcRE motif of interest. Scripts and database are available online http://homer.salk.edu/homer/motif/.

Summary Analysis

All statistical analyses are computed using R (version 3.1.2) using custom scripts. Gene lengths for length-normalized expression are taken as the mean of the lengths of the transcripts for that gene. GO term enrichment was produced using the fb_2014_03 version of the flybase gene ontology[152]. Only genes with at least one annotated ontology term were used for enrichment calculates. All GO term enrichment p-values are calculated using the hypergeometric distribution.

Responsive Proximal Genes

For all significantly responsive genes the fraction of responsive genes and the average response direction of nearby genes are analyzed. In order to determine the distance between two genes the ecdysone relevant transcription start sites (TSS) are first determined. For genes with multiple transcription start sites the TSS is determined to be the TSS associated with a significant exon level response to ecdysone if one exists, or the exon with the highest length normalized expression in the relevant time point (i.e. five hour time point for induced genes, zero hour time point for repressed genes or the average for non-responsive genes).

Each gene within 20,000 base pairs of a responsive gene is associated with two values, first if the gene is responsive and second the direction of response, taken as the negative log10 of the p-value times the sign of the log fold change after ecdysone exposure divided by before (these values are trimmed to plus and minus 10 to avoid outlier effects). A moving window of 100 gene, cell line combinations is used to calculate the fraction of responsive genes and the average response direction within each bin. The binned points are grouped according to the response direction of the responsive gene as well as the shared promoter architecture (upstream/downstream and same/opposite strand). These points are then smoothed using loess with a local linear fit over binned points. The plot produced is found in Figure 3.4.

Restricted Ecdysone-response Prediction

In order to predict the restricted response the following two models, referred to as matched induction and matched repression, were fit. The models are symmetric, so only matched induced model will be described in full detail.

At each gene in the restricted induced set we aim to predict which cell lines respond significantly and which respond most repressively, defined as the cell lines showing the smallest log2 fold change of ecdysone exposure expression / pre-exposure expression. A matching number of the most repressive responsive cell lines are chosen such that each gene contains a balanced number of induced and matched repressed cell lines.

In order to predict which gene-cell lines combinations belong to the above described sets the model is provided with normalized TF expression masked by known TF motif from the TOMTOM database[153] presence in the promoter of the gene to be predicted. TF expression is included only if at least one cell line expresses the TF above the 20th percentile of genes with at least one read. A TF motif is considered significant by setting the threshold on the allowed mismatches to the known PWM such that just less than 5% of non-responsive genes' promoters contain a hit to the motif. There are 270 TFs with known motifs and valid expression. Additionally, the model is given the rank of the normalized expression of the gene to be predicted as lowly expressed genes are intuitively less likely to be repressed and more likely to be induced and conversely for highly expressed genes.

This model is then fit using random forests[148,154]. Important variables are determined from a model fit on all 41 cell lines. Accuracy measures are obtained by constructing the above outcome and predictor variables and then removing each cell line as a test set. The data from the remaining cell lines are used to train the random forest and the data from the left out cell line is used to test the accuracy of the model's predictions. Accuracy measures from all cell lines are averaged and reported.

In order to determine the effect of increased numbers of cell lines on prediction accuracy samples of cell lines are taken randomly and restricted responding genes, outcome and predictor data are constructed. The left out cell line method is again used to determine the accuracy of the model. Since there are many subsets of cell lines which may be chosen the subsetting procedure is repeated 1000 times for each number of cell lines and all average accuracy values are reported.

In order to estimate the asymptotic predictive accuracy given infinite cell lines the predictive accuracy values described above were fit to the function $Accuracy\sim -a\times\#CellLines^{-b}+c$ using an alternating least squares linear fit. The fitted c parameter indicates the estimated asymptotic predictive accuracy.

Extended Time Course Analysis

For a subset of three cell lines, Kc, BG3 and S2, that have an extended time course, including the zero, one, three, five and seven hour time points the following analysis pipeline was conducted. In order to compare the transcriptional responses across the time course for each cell line a normalization that allows comparison of genes with dissimilar steady state expression levels, but may share ecdysone response “shape”. This normalization begins by applying the robust median library size normalization from the DESeq R package[82]. Then a mean centering is applied at each gene and cell line across all five time points. The gene cell line expression is then divided by the fitted standard deviation across all gene-cell lines combinations in order to adjust for the known increase in variance at higher expression loci. These normalized expression measures can thus be interpreted as a response shape across time for each cell line and response shape is comparable for genes at distance mean expression levels, but the trend and relative scale of response over time is maintained. Units represent standard deviations from the mean across time. These normalized expression values are analyzed in the context of the response at the five hour time point.

References

1 Cherbas, L. *et al.* The transcriptional diversity of 25 Drosophila cell lines. *Genome research* **21**, 301-314, doi:10.1101/gr.112961.110 (2011).

2 Kotake, Y. *et al.* Splicing factor SF3b as a target of the antitumor natural product pladienolide. *Nature chemical biology* **3**, 570-575, doi:10.1038/nchembio.2007.16 (2007).

3 Mendel, G. & Vlaamse Academie voor Wetenschappen Letteren en Schone Kunsten van Belgie. [from old catalog]. *Gregor Mendel herdacht naar aanleiding van de honderdste verjaring van zijn Versuche über Pflanzenhybriden <1865> Openbare vergadering van 9 oktober 1965*. (Koninklijke Vlaamse Academie voor Wetenshappen).

4 Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5463-5467 (1977).

5 Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 560-564 (1977).

6 Morgan, T. H. & Bridges, C. B. *Sex-linked inheritance in Drosophila*. (Carnegie Institution of Washington, 1916).

7 Brown, J. B. *et al.* Diversity and dynamics of the Drosophila transcriptome. *Nature* **512**, 393-399, doi:10.1038/nature12962 (2014).

8 Graveley, B. R. *et al.* The developmental transcriptome of Drosophila melanogaster. *Nature* **471**, 473-479, doi:10.1038/nature09715 (2011).

9 Boley, N. *et al.* Genome-guided transcript assembly by integrative analysis of RNA sequence data. *Nature biotechnology* **32**, 341-346, doi:10.1038/nbt.2850 (2014).

10 Hoskins, R. A. *et al.* Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome research* **21**, 182-192, doi:10.1101/gr.112466.110 (2011).

11 Celniker, S. E. & Rubin, G. M. The Drosophila melanogaster genome. *Annual review of genomics and human genetics* **4**, 89-117, doi:10.1146/annurev.genom.4.070802.110323 (2003).

12 Stapleton, M. *et al.* A Drosophila full-length cDNA resource. *Genome biology* **3**, RESEARCH0080 (2002).

13 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628, doi:nmeth.1226 [pii]
10.1038/nmeth.1226 (2008).

14 Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344-1349, doi:1158441 [pii]
10.1126/science.1158441 (2008).

15 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* **786**, 181-200, doi:10.1007/978-1-61779-292-2_11 (2012).

16 Mangone, M. *et al.* The landscape of C. elegans 3'UTRs. *Science* **329**, 432-435, doi:science.1191244 [pii]
10.1126/science.1191244 (2010).

17 Jan, C. H., Friedman, R. C., Ruby, J. G. & Bartel, D. P. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. *Nature* **469**, 97-101, doi:nature09616 [pii]
10.1038/nature09616 (2011).

18 Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-578, doi:nprot.2012.016 [pii]
10.1038/nprot.2012.016 (2012).

19 Collins, J. E., White, S., Searle, S. M. & Stemple, D. L. Incorporating RNA-seq data into the zebrafish Ensembl genebuild. *Genome Res* **22**, 2067-2078, doi:gr.137901.112 [pii]
10.1101/gr.137901.112 (2012).

20 Carninci, P. *et al.* Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res* **13**, 1273-1289, doi:10.1101/gr.1119703
13/6b/1273 [pii] (2003).

21 Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research* **22**, 1775-1789, doi:10.1101/gr.132159.111 (2012).

22 Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101-108, doi:nature11233 [pii]
10.1038/nature11233 (2012).

23 St Pierre, S. E., Ponting, L., Stefancsik, R., McQuilton, P. & FlyBase, C. FlyBase 102--advanced approaches to interrogating FlyBase. *Nucleic acids research* **42**, D780-788, doi:10.1093/nar/gkt1092 (2014).

24 Spradling, A. C. *et al.* The Berkeley Drosophila Genome Project gene disruption project: Single P-element insertions mutating 25% of vital Drosophila genes. *Genetics* **153**, 135-177 (1999).

25 Hansen, K. D. *et al.* Genome-wide identification of alternative splice forms down-regulated by nonsense-mediated mRNA decay in Drosophila. *PLoS genetics* **5**, e1000525, doi:10.1371/journal.pgen.1000525 (2009).

26 McBrayer, Z. *et al.* Prothoracicotropic hormone regulates developmental timing and body size in Drosophila. *Developmental cell* **13**, 857-871, doi:10.1016/j.devcel.2007.11.003 (2007).

27 Horton, P. *et al.* WoLF PSORT: protein localization predictor. *Nucleic acids research* **35**, W585-587, doi:10.1093/nar/gkm259 (2007).

28 Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-476, doi:10.1038/nature07509 (2008).

29 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 genetics* **40**, 1413-1415, doi:10.1038/ng.259 (2008).

30 Schmucker, D. *et al.* Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101**, 671-684 (2000).

31 Hodge, T. & Cope, M. J. A myosin family tree. *Journal of cell science* **113 Pt 19**, 3353-3354 (2000).

32 Chen, Z. X. *et al.* Comparative validation of the D. melanogaster modENCODE transcriptome annotation. *Genome research* **24**, 1209-1223, doi:10.1101/gr.159384.113 (2014).

33 Lipshitz, H. D., Peattie, D. A. & Hogness, D. S. Novel transcripts from the Ultrabithorax domain of the bithorax complex. *Genes & development* **1**, 307-322 (1987).

34 Tupy, J. L. *et al.* Identification of putative noncoding polyadenylated transcripts in Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 5495-5500, doi:10.1073/pnas.0501422102 (2005).

35 Young, R. S. *et al.* Identification and properties of 1,119 candidate lincRNA loci in the Drosophila melanogaster genome. *Genome biology and evolution* **4**, 427-442, doi:10.1093/gbe/evs020 (2012).

36 Kondo, T. *et al.* Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nature cell biology* **9**, 660-665, doi:10.1038/ncb1595 (2007).

37 Katayama, S. *et al.* Antisense transcription in the mammalian transcriptome. *Science* **309**, 1564-1566, doi:10.1126/science.1112009 (2005).

38 Duncan, D. M., Burgess, E. A. & Duncan, I. Control of distal antennal identity and tarsal development in Drosophila by spineless-aristapedia, a homolog of the mammalian dioxin receptor. *Genes & development* **12**, 1290-1303 (1998).

39 Schwartz, C., Locke, J., Nishida, C. & Kornberg, T. B. Analysis of cubitus interruptus regulation in Drosophila embryos and imaginal disks. *Development* **121**, 1625-1635 (1995).

40 Misra, S. *et al.* Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. *Genome Biology* **3**, research0083 (2002).

41 Lipovich, L. *et al.* Activity-dependent human brain coding/noncoding gene regulatory networks. *Genetics* **192**, 1133-1148, doi:10.1534/genetics.112.145128 (2012).

42 Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218-223, doi:10.1126/science.1168978 (2009).

43 Banfai, B. *et al.* Long noncoding RNAs are rarely translated in two human cell lines. *Genome research* **22**, 1646-1657, doi:10.1101/gr.134767.111 (2012).

44 St Pierre, S. E., Ponting, L., Stefancsik, R., McQuilton, P. & Consortium, F. FlyBase 102-advanced approaches to interrogating FlyBase. *Nucleic acids research* **42**, D780-D788, doi:Doi 10.1093/Nar/Gkt1092 (2014).

45 Marchler-Bauer, A. *et al.* CDD: NCBI's conserved domain database. *Nucleic acids research* **43**, D222-226, doi:10.1093/nar/gku1221 (2015).

46 Marchler-Bauer, A. & Bryant, S. H. CD-Search: protein domain annotations on the fly. *Nucleic acids research* **32**, W327-331, doi:10.1093/nar/gkh454 (2004).

47 Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275-282, doi:10.1093/bioinformatics/btr209 (2011).

48 Blankenberg, D., Taylor, J., Nekrutenko, A. & Galaxy, T. Making whole genome multiple alignments usable for biologists. *Bioinformatics* **27**, 2426-2428, doi:10.1093/bioinformatics/btr398 (2011).

49 Team, R. C. *R: A Language and Environment for Statistical Computing*, 2014).

50 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).

51 Draper, D. E. Protein-RNA recognition. *Annual review of biochemistry* **64**, 593-620, doi:10.1146/annurev.bi.64.070195.003113 (1995).

52 Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS letters* **582**, 1977-1986, doi:10.1016/j.febslet.2008.03.004 (2008).

53 Lewis, H. A. *et al.* Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the fragile X syndrome. *Cell* **100**, 323-332 (2000).

54 Hall, M. P. *et al.* Quaking and PTB control overlapping splicing regulatory networks during muscle cell differentiation. *Rna* **19**, 627-638, doi:10.1261/rna.038422.113 (2013).

55 Kielkopf, C. L., Lucke, S. & Green, M. R. U2AF homology motifs: protein recognition in the RRM world. *Genes & development* **18**, 1513-1526, doi:10.1101/gad.1206204 (2004).

56 Jin, Y. *et al.* A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. *The EMBO journal* **22**, 905-912, doi:10.1093/emboj/cdg089 (2003).

57 Mili, S. & Steitz, J. A. Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. *Rna* **10**, 1692-1694, doi:10.1261/rna.7151404 (2004).

58 McHugh, C. A., Russell, P. & Guttman, M. Methods for comprehensive experimental identification of RNA-protein interactions. *Genome biology* **15**, 203, doi:10.1186/gb4152 (2014).

59 Strein, C., Alleaume, A. M., Rothbauer, U., Hentze, M. W. & Castello, A. A versatile assay for RNA-binding proteins in living cells. *Rna* **20**, 721-731, doi:10.1261/rna.043562.113 (2014).

60 Mitchell, S. F., Jain, S., She, M. & Parker, R. Global analysis of yeast mRNPs. *Nature structural & molecular biology* **20**, 127-133, doi:10.1038/nsmb.2468 (2013).

61 Ishizuka, A., Siomi, M. C. & Siomi, H. A Drosophila fragile X protein interacts with components of RNAi and ribosomal proteins. *Genes & development* **16**, 2497-2508, doi:10.1101/gad.1022002 (2002).

62 Chen, E., Sharma, M. R., Shi, X., Agrawal, R. K. & Joseph, S. Fragile X mental retardation protein regulates translation by binding directly to the ribosome. *Molecular cell* **54**, 407-417, doi:10.1016/j.molcel.2014.03.023 (2014).

63 Bechara, E. G. *et al.* A novel function for fragile X mental retardation protein in translational activation. *PLoS biology* **7**, e16, doi:10.1371/journal.pbio.1000016 (2009).

64 Licatalosi, D. D. *et al.* HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464-469, doi:10.1038/nature07488 (2008).

65 Han, H. *et al.* MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* **498**, 241-245, doi:10.1038/nature12270 (2013).

66 Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature methods* **7**, 1009-1015, doi:10.1038/nmeth.1528 (2010).

67 Buckanovich, R. J. & Darnell, R. B. The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo. *Molecular and cellular biology* **17**, 3194-3201 (1997).

68 Brooks, A. N. *et al.* Conservation of an RNA regulatory map between Drosophila and mammals. *Genome research* **21**, 193-202, doi:10.1101/gr.108662.110 (2011).

69 Shepard, P. J. & Hertel, K. J. The SR protein family. *Genome biology* **10**, 242, doi:10.1186/gb-2009-10-10-242 (2009).

70 Zhao, J. *et al.* Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Molecular cell* **40**, 939-953, doi:10.1016/j.molcel.2010.12.011 (2010).

71 Brazao, T. F. *et al.* A new function of ROD1 in nonsense-mediated mRNA decay. *FEBS letters* **586**, 1101-1110, doi:10.1016/j.febslet.2012.03.015 (2012).

72 Chen, L. *et al.* Global regulation of mRNA translation and stability in the early Drosophila embryo by the Smaug RNA-binding protein. *Genome biology* **15**, R4, doi:10.1186/gb-2014-15-1-r4 (2014).

73 Lu, Z., Guan, X., Schmidt, C. A. & Matera, A. G. RIP-seq analysis of eukaryotic Sm proteins identifies three major categories of Sm-containing ribonucleoproteins. *Genome biology* **15**, R7, doi:10.1186/gb-2014-15-1-r7 (2014).

74 Hogan, D. J., Riordan, D. P., Gerber, A. P., Herschlag, D. & Brown, P. O. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS biology* **6**, e255, doi:10.1371/journal.pbio.0060255 (2008).

75 Kurosaki, T. & Maquat, L. E. Rules that govern UPF1 binding to mRNA 3' UTRs. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 3357-3362, doi:10.1073/pnas.1219908110 (2013).

76 Le Hir, H., Gatfield, D., Izaurralde, E. & Moore, M. J. The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *The EMBO journal* **20**, 4987-4997, doi:10.1093/emboj/20.17.4987 (2001).

77 Sabin, L. R. *et al.* Ars2 regulates both miRNA- and siRNA- dependent silencing and suppresses RNA virus infection in Drosophila. *Cell* **138**, 340-351, doi:10.1016/j.cell.2009.04.045 (2009).

78 Will, C. L. & Luhrmann, R. Spliceosome structure and function. *Cold Spring Harbor perspectives in biology* **3**, doi:10.1101/cshperspect.a003707 (2011).

79 Han, S. P., Tang, Y. H. & Smith, R. Functional diversity of the hnRNPs: past, present and perspectives. *The Biochemical journal* **430**, 379-392, doi:10.1042/BJ20100396 (2010).

80 Kosti, I., Radivojac, P. & Mandel-Gutfreund, Y. An integrated regulatory network reveals pervasive cross-regulation among transcription and splicing factors. *PLoS computational biology* **8**, e1002603, doi:10.1371/journal.pcbi.1002603 (2012).

81 Guruharsha, K. G. *et al.* A protein complex network of Drosophila melanogaster. *Cell* **147**, 690-703, doi:10.1016/j.cell.2011.08.047 (2011).

82 Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome biology* **11**, R106, doi:10.1186/gb-2010-11-10-r106 (2010).

83 Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. 1752-1779, doi:10.1214/11-AOAS466 (2011).

84 Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research* **22**, 1813-1831, doi:10.1101/gr.136184.111 (2012).

85 Lewis, J. D., Izaurralde, E., Jarmolowski, A., McGuigan, C. & Mattaj, I. W. A nuclear cap-binding complex facilitates association of U1 snRNP with the cap-proximal 5' splice site. *Genes & development* **10**, 1683-1698 (1996).

86 Kumar, S. & Lopez, A. J. Negative feedback regulation among SR splicing factors encoded by Rbp1 and Rbp1-like in Drosophila. *The EMBO journal* **24**, 2646-2655, doi:10.1038/sj.emboj.7600723 (2005).

87 Sidman, R. L., Dickie, M. M. & Appel, S. H. Mutant Mice (Quaking and Jimpy) with Deficient Myelination in the Central Nervous System. *Science* **144**, 309-311 (1964).

88 Hilgers, V., Lemke, S. B. & Levine, M. ELAV mediates 3' UTR extension in the Drosophila nervous system. *Genes & development* **26**, 2259-2264, doi:10.1101/gad.199653.112 (2012).

89 Smibert, P. *et al.* Global patterns of tissue-specific alternative polyadenylation in Drosophila. *Cell reports* **1**, 277-289, doi:10.1016/j.celrep.2012.01.001 (2012).

90 Ray, D. *et al.* Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature biotechnology* **27**, 667-670, doi:10.1038/nbt.1550 (2009).

91 Mount, S. M. & Steitz, J. A. Sequence of U1 RNA from Drosophila melanogaster: implications for U1 secondary structure and possible involvement in splicing. *Nucleic acids research* **9**, 6351-6368 (1981).

92 Borg, I. & Groenen, P. J. F. *Modern multidimensional scaling : theory and applications*. 2nd edn, (Springer, 2005).

93 Harrison, M. M., Li, X. Y., Kaplan, T., Botchan, M. R. & Eisen, M. B. Zelda binding in the early Drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. *PLoS genetics* **7**, e1002266, doi:10.1371/journal.pgen.1002266 (2011).

94 Kaplan, T. *et al.* Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. *PLoS genetics* **7**, e1001290, doi:10.1371/journal.pgen.1001290 (2011).

95 Aerts, S., Van Loo, P., Thijs, G., Moreau, Y. & De Moor, B. Computational detection of cis - regulatory modules. *Bioinformatics* **19 Suppl 2**, ii5-14 (2003).

96 Nogueira, T. & Springer, M. Post-transcriptional control by global regulators of gene expression in bacteria. *Current opinion in microbiology* **3**, 154-158 (2000).

97 Keene, J. D. RNA regulons: coordination of post-transcriptional events. *Nature reviews. Genetics* **8**, 533-543, doi:10.1038/nrg2111 (2007).

98 Jaeger, J. The gap gene network. *Cellular and molecular life sciences : CMLS* **68**, 243-274, doi:10.1007/s00018-010-0536-y (2011).

99 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).

100 Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. & Golani, I. Controlling the false discovery rate in behavior genetics research. *Behavioural brain research* **125**, 279-284 (2001).

101 Cox, T. F. & Cox, M. A. A. *Multidimensional scaling*. 2nd edn, (Chapman & Hall/CRC, 2001).

102 Li, Y., Zhao, D. Y., Greenblatt, J. F. & Zhang, Z. RIPSeeker: a statistical package for identifying protein-associated transcripts from RIP-seq experiments. *Nucleic acids research* **41**, e94, doi:10.1093/nar/gkt142 (2013).

103 Riordan, D. P., Herschlag, D. & Brown, P. O. Identification of RNA recognition elements in the Saccharomyces cerevisiae transcriptome. *Nucleic acids research* **39**, 1501-1509, doi:10.1093/nar/gkq920 (2011).

104 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).

105 Bailey, T. L. & Elkan, C. The value of prior knowledge in discovering motifs with MEME. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* **3**, 21-29 (1995).

106 Laudet, V., Hanni, C., Coll, J., Catzeflis, F. & Stehelin, D. Evolution of the nuclear receptor gene superfamily. *The EMBO journal* **11**, 1003-1013 (1992).

107 Kumar, R. & Thompson, E. B. The structure of the nuclear hormone receptors. *Steroids* **64**, 310-319 (1999).

108 Ashburner, M. Sequential gene activation by ecdysone in polytene chromosomes of Drosophila melanogaster. I. Dependence upon ecdysone concentration. *Developmental biology* **35**, 47-61 (1973).

109 Koelle, M. R. *et al.* The Drosophila EcR gene encodes an ecdysone receptor, a new member of the steroid receptor superfamily. *Cell* **67**, 59-77 (1991).

110 Cherbas, P., Cherbas, L., Lee, S. S. & Nakanishi, K. 26-[125I]iodoponasterone A is a potent ecdysone and a sensitive radioligand for ecdysone receptors. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 2096-2100 (1988).

111 Yao, T. P. *et al.* Functional ecdysone receptor is the product of EcR and Ultraspiracle genes. *Nature* **366**, 476-479, doi:10.1038/366476a0 (1993).

112 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).

113 Andres, A. J. & Cherbas, P. Tissue-specific regulation by ecdysone: distinct patterns of Eip28/29 expression are controlled by different ecdysone response elements. *Developmental genetics* **15**, 320-331, doi:10.1002/dvg.1020150403 (1994).

114 Fallon, A. M. & Gerenday, A. Ecdysone and the cell cycle: investigations in a mosquito cell line. *Journal of insect physiology* **56**, 1396-1401, doi:10.1016/j.jinsphys.2010.03.016 (2010).

115 Cakouros, D., Daish, T. J. & Kumar, S. Ecdysone receptor directly binds the promoter of the Drosophila caspase dronc, regulating its expression in specific tissues. *The Journal of cell biology* **165**, 631-640, doi:10.1083/jcb.200311057 (2004).

116 Kilpatrick, Z. E., Cakouros, D. & Kumar, S. Ecdysone-mediated up-regulation of the effector caspase DRICE is required for hormone-dependent apoptosis in Drosophila cells. *The Journal of biological chemistry* **280**, 11981-11986, doi:10.1074/jbc.M413971200 (2005).

117 Tominaga, M. *et al.* Neurite elongation from Drosophila neural BG2-c6 cells stimulated by 20-hydroxyecdysone. *Neuroscience letters* **482**, 250-254, doi:10.1016/j.neulet.2010.07.049 (2010).

118 Sempere, L. F., Dubrovsky, E. B., Dubrovskaya, V. A., Berger, E. M. & Ambros, V. The expression of the let-7 small regulatory RNA is controlled by ecdysone during metamorphosis in Drosophila melanogaster. *Developmental biology* **244**, 170-179, doi:10.1006/dbio.2002.0594 (2002).

119 Garbuzov, A. & Tatar, M. Hormonal regulation of Drosophila microRNA let-7 and miR-125 that target innate immunity. *Fly* **4**, 306-311 (2010).

120 Azoitei, A. & Spindler-Barth, M. DNA affects ligand binding of the ecdysone receptor of Drosophila melanogaster. *Molecular and cellular endocrinology* **303**, 91-99, doi:10.1016/j.mce.2009.01.022 (2009).

121 Cherbas, L., Lee, K. & Cherbas, P. Identification of ecdysone response elements by analysis of the Drosophila Eip28/29 gene. *Genes & development* **5**, 120-131 (1991).

122 Nakagawa, Y. & Henrich, V. C. Arthropod nuclear receptors and their role in molting. *The FEBS journal* **276**, 6128-6157, doi:10.1111/j.1742-4658.2009.07347.x (2009).

123 Braun, S., Azoitei, A. & Spindler-Barth, M. DNA-binding properties of Drosophila ecdysone receptor isoforms and their modification by the heterodimerization partner ultraspiracle. *Archives of insect biochemistry and physiology* **72**, 172-191, doi:10.1002/arch.20328 (2009).

124 Davis, M. B., SanGil, I., Berry, G., Olayokun, R. & Neves, L. H. Identification of common and cell type specific LXXLL motif EcR cofactors using a bioinformatics refined candidate RNAi screen in Drosophila melanogaster cell lines. *BMC developmental biology* **11**, 66, doi:10.1186/1471-213X-11-66 (2011).

125 Sedkov, Y. *et al.* Methylation at lysine 4 of histone H3 in ecdysone-dependent development of Drosophila. *Nature* **426**, 78-83, doi:10.1038/nature02080 (2003).

126 Gauthier, S. A., VanHaaften, E., Cherbas, L., Cherbas, P. & Hewes, R. S. Cryptocephal, the Drosophila melanogaster ATF4, is a specific coactivator for ecdysone receptor isoform B2. *PLoS genetics* **8**, e1002883, doi:10.1371/journal.pgen.1002883 (2012).

127 Carbonell, A., Mazo, A., Serras, F. & Corominas, M. Ash2 acts as an ecdysone receptor coactivator by stabilizing the histone methyltransferase Trr. *Molecular biology of the cell* **24**, 361-372, doi:10.1091/mbc.E12-04-0267 (2013).

128 Zhu, J., Chen, L., Sun, G. & Raikhel, A. S. The competence factor beta Ftz-F1 potentiates ecdysone receptor activity via recruiting a p160/SRC coactivator. *Molecular and cellular biology* **26**, 9402-9412, doi:10.1128/MCB.01318-06 (2006).

129 Sawatsubashi, S. *et al.* A histone chaperone, DEK, transcriptionally coactivates a nuclear receptor. *Genes & development* **24**, 159-170, doi:10.1101/gad.1857410 (2010).

130 Badenhorst, P. *et al.* The Drosophila nucleosome remodeling factor NURF is required for Ecdysteroid signaling and metamorphosis. *Genes & development* **19**, 2540-2545, doi:10.1101/gad.1342605 (2005).

131 Kugler, S. J., Gehring, E. M., Wallkamm, V., Kruger, V. & Nagel, A. C. The Putzig-NURF nucleosome remodeling complex is required for ecdysone receptor signaling and innate immunity in Drosophila melanogaster. *Genetics* **188**, 127-139, doi:10.1534/genetics.111.127795 (2011).

132 Zraly, C. B. & Dingwall, A. K. The chromatin remodeling and mRNA splicing functions of the Brahma (SWI/SNF) complex are mediated by the SNR1/SNF5 regulatory subunit. *Nucleic acids research* **40**, 5975-5987, doi:10.1093/nar/gks288 (2012).

133 Zraly, C. B., Middleton, F. A. & Dingwall, A. K. Hormone-response genes are direct in vivo regulatory targets of Brahma (SWI/SNF) complex function. *The Journal of biological chemistry* **281**, 35305-35315, doi:10.1074/jbc.M607806200 (2006).

134 Ables, E. T. & Drummond-Barbosa, D. The steroid hormone ecdysone functions with intrinsic chromatin remodeling factors to control female germline stem cells in Drosophila. *Cell stem cell* **7**, 581-592, doi:10.1016/j.stem.2010.10.001 (2010).

135 Bodai, L. *et al.* Ecdysone induced gene expression is associated with acetylation of histone H3 lysine 23 in Drosophila melanogaster. *PloS one* **7**, e40565, doi:10.1371/journal.pone.0040565 (2012).

136 Beckstead, R. B., Lam, G. & Thummel, C. S. The genomic response to 20-hydroxyecdysone at the onset of Drosophila metamorphosis. *Genome biology* **6**, R99, doi:10.1186/gb-2005-6-12-r99 (2005).

137 Gonsalves, S. E., Neal, S. J., Kehoe, A. S. & Westwood, J. T. Genome-wide examination of the transcriptional response to ecdysteroids 20-hydroxyecdysone and ponasterone A in Drosophila melanogaster. *BMC genomics* **12**, 475, doi:10.1186/1471-2164-12-475 (2011).

138 Gauhar, Z. *et al.* Genomic mapping of binding regions for the Ecdysone receptor protein complex. *Genome research* **19**, 1006-1013, doi:10.1101/gr.081349.108 (2009).

139 Kellner, W. A., Ramos, E., Van Bortle, K., Takenaka, N. & Corces, V. G. Genome-wide phosphoacetylation of histone H3 at Drosophila enhancers and promoters. *Genome research* **22**, 1081-1088, doi:10.1101/gr.136929.111 (2012).

140 Shlyueva, D. *et al.* Hormone-responsive enhancer-activity maps reveal predictive motifs, indirect repression, and targeting of closed chromatin. *Molecular cell* **54**, 180-192, doi:10.1016/j.molcel.2014.02.026 (2014).

141 Hammonds, A. S. *et al.* Spatial expression of transcription factors in Drosophila embryonic organ development. *Genome biology* **14**, R140, doi:10.1186/gb-2013-14-12-r140 (2013).

142 Chanut-Delalande, H. *et al.* Pri peptides are mediators of ecdysone for the temporal control of development. *Nature cell biology* **16**, 1035-1044, doi:10.1038/ncb3052 (2014).

143 Chodankar, R., Wu, D. Y., Schiller, B. J., Yamamoto, K. R. & Stallcup, M. R. Hic-5 is a transcription coregulator that acts before and/or after glucocorticoid receptor genome occupancy in a gene-selective manner. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 4007-4012, doi:10.1073/pnas.1400522111 (2014).

144 Vanolst, L., Fromental-Ramain, C. & Ramain, P. Toutatis, a TIP5-related protein, positively regulates Pannier function during Drosophila neural development. *Development* **132**, 4327-4338, doi:10.1242/dev.02014 (2005).

145 Filippakopoulos, P. & Knapp, S. Targeting bromodomains: epigenetic readers of lysine acetylation. *Nature reviews. Drug discovery* **13**, 337-356, doi:10.1038/nrd4286 (2014).

146 Bernardo, T. J., Dubrovskaya, V. A., Jannat, H., Maughan, B. & Dubrovsky, E. B. Hormonal regulation of the E75 gene in Drosophila: identifying functional regulatory elements through computational and biological analysis. *Journal of molecular biology* **387**, 794-808 (2009).

147 Cherbas, L., Hu, X., Zhimulev, I., Belyaeva, E. & Cherbas, P. EcR isoforms in Drosophila: testing tissue-specific requirements by targeted blockade and rescue. *Development* **130**, 271-284 (2003).

148 Breiman, L. Random forests. *Mach Learn* **45**, 5-32, doi:Doi 10.1023/A:1010933404324 (2001).

149 Nijhout, H. F., Reed, M. C., Budu, P. & Ulrich, C. M. A mathematical model of the folate cycle: new insights into folate homeostasis. *The Journal of biological chemistry* **279**, 55008-55016, doi:10.1074/jbc.M410818200 (2004).

150 Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169, doi:10.1093/bioinformatics/btu638 (2015).

151 Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome research* **22**, 2008-2017, doi:10.1101/gr.133744.111 (2012).

152 Gene Ontology, C. *et al.* Gene Ontology annotations and resources. *Nucleic acids research* **41**, D530-535, doi:10.1093/nar/gks1050 (2013).

153 Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome biology* **8**, R24, doi:10.1186/gb-2007-8-2-r24 (2007).

154 Abraham, A. *et al.* Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics* **8**, 14, doi:10.3389/fninf.2014.00014 (2014).