Combinatorial gene regulatory functions underlie ultraconserved elements in Drosophila

Warnefors, Maria, Hartmann, Britta, Thomsen, Stefan and Alonso, Claudio R (2016) Combinatorial gene regulatory functions underlie ultraconserved elements in Drosophila. Molecular Biology and Evolution, 33 (7). msw101. ISSN 0737-4038

This version is available from Sussex Research Online: http://sro.sussex.ac.uk/id/eprint/61891/

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the URL above for details on accessing the published version.

Copyright and reuse:
Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

http://sro.sussex.ac.uk
Ultraconserved elements (UCEs) are discrete genomic elements conserved across large evolutionary distances. Although UCEs have been linked to multiple facets of mammalian gene regulation their extreme evolutionary conservation remains largely unexplained. Here, we apply a computational approach to investigate this question in Drosophila, exploring the molecular functions of more than 1,500 UCEs shared across the genomes of 12 Drosophila species. Our data indicate that Drosophila UCEs are hubs for gene regulatory functions and suggest that UCE sequence invariance originates from their combinatorial roles in gene control. We also note that the gene regulatory roles of intronic and intergenic UCEs (iUCEs) are distinct from those found in exonic UCEs (eUCEs). In iUCEs, transcription factor (TF) and epigenetic factor binding data strongly support iUCE roles in transcriptional and epigenetic regulation. In contrast, analyses of eUCEs indicate that they are two orders of magnitude more likely than the expected to simultaneously include protein-coding sequence, TF-binding sites, splice sites, and RNA editing sites but have reduced roles in transcriptional or epigenetic regulation. Furthermore, we use a Drosophila cell culture system and transgenic Drosophila embryos to validate the notion of UCE combinatorial regulatory roles using an eUCE within the Hox gene Ultrabithorax and show that its protein-coding region also contains alternative splicing regulatory information. Taken together our experiments indicate that UCEs emerge as a result of combinatorial gene regulatory roles and highlight common features in mammalian and insect UCEs implying that similar processes might underlie ultraconservation in diverse animal taxa.

Key words: ultraconserved elements, UCEs, alternative splicing, epigenetic regulation, transcriptional regulation, Hox genes, organismal development.

Introduction

Evolutionary conservation of genomic sequences is highly heterogeneous: poorly conserved regions are commonly intermingled with sequences that show perfect conservation across large evolutionary distances. This latter category includes the remarkable class of ultraconserved elements (UCEs), originally defined as sequences of at least 200 nt that are identical across the human, mouse, and rat genomes (Bejerano et al. 2004). In spite of over a decade of research on UCEs and their detection in vertebrates, insects, and other animals, as well as yeasts and plants (Kellis et al. 2003; Glazov et al. 2005; Siepel et al. 2005; Kritsas et al. 2012; Reneker et al. 2012; Ryu et al. 2012), there is no unifying molecular mechanism that satisfactorily explains their extreme evolutionary conservation (Harmston et al. 2013).

Mammalian UCEs have been linked to a diverse set of regulatory functions including transcriptional enhancers (Woolfe et al. 2005; Pennacchio et al. 2006; Lampe et al. 2008; Visel et al. 2008; Viturawong et al. 2013) and noncoding RNAs (ncRNAs) (Feng et al. 2006; Calin et al. 2007; Mestdagh et al. 2010; Berghoff et al. 2013; Liz et al. 2014; Nielsen et al. 2014). UCEs that overlap with protein-coding genes have also been implicated in RNA regulatory processes, such as alternative splicing, nonsense-mediated RNA decay, and RNA editing (Bejerano et al. 2004; Siepel et al. 2005; Larea et al. 2007; Ni et al. 2007). Nonetheless, given that none of these mechanisms seems sufficient to explain the phenomenon of ultraconservation on its own, it has been proposed that superimposed functional constraints might contribute to the generation of UCEs (Siepel et al. 2005; Lampe et al. 2008; Viturawong et al. 2013).
The genomes of Drosophila melanogaster and related species harbor an independent set of UCEs to those found in mammals (Glazov et al. 2005). These sequences are shorter than UCEs shared across equally divergent vertebrate genomes (Makunin et al. 2013), but were shown to experience a higher degree of selective constraint than mammalian UCEs (Kern et al. 2015).

Despite the independent evolutionary origin of mammalian and Drosophila UCEs, their comparison reveals a number of common and distinct features. In regards to the latter, while in mammals it has been difficult to link mutations in UCEs to phenotypic outcomes (Drake et al. 2006; Ahituv et al. 2007; Chen et al. 2007; Yang et al. 2008; Catucci et al. 2009; Poitras et al. 2010; Chiang et al. 2012), a recent survey of 11 insertions into Drosophila UCEs identified four as recessive lethal (Makunin et al. 2013), illustrating a more straightforward association between UCEs and phenotypes in the fly, making Drosophila a promising model system to study the mechanisms that lead to ultraconservation. As for similarities, fly and mammalian UCEs tend to cluster around genes involved in developmental processes (Bejerano et al. 2004; Boffelli et al. 2004; Sandelin et al. 2004; Glazov et al. 2005) and seem associated to related regulatory mechanisms including alternative splicing and RNA editing (Glazov et al. 2005, 2006; Kern et al. 2015); however, to date, Drosophila UCEs were not found to overlap with transcription factor (TF) binding sites (Glazov et al. 2005). A deeper understanding of Drosophila UCEs combined with the use of the powerful genetic tools available in the fruitfly might therefore reveal general and fundamental properties of ultraconservation and its links to the genetic programs encrypted in the genome.

Here, we investigate the functional roles of 1,516 UCEs shared across 12 Drosophila genomes. In contrast to previous findings we find that intronic and intergenic Drosophila UCEs (iUCEs) are indeed overrepresented within annotated regulatory elements and show significant enrichments and depletions for binding of specific TFs. Our analysis of individual DNA-binding factors also shows that UCEs are enriched for Polycomb-group (PcG) protein binding, suggesting a role for UCEs in chromatin regulation during development. We further show that exonic UCEs (eUCEs) are strongly enriched for multifunctional sequences and are about 100-fold more likely to combine protein-coding capacity with the presence of RNA editing sites, splicing regulators and TF binding sites compared with randomly chosen exonic elements of identical size. To explore whether such predicted multi-functionality was relevant to gene expression we studied one of the Drosophila genes bearing the highest number of UCEs—the Hox gene Ultrabithorax (Ubx)—and observed that mutation of one ultraconserved exon of Ubx affects alternative splicing in Drosophila cells in culture and during embryonic development.

This study therefore contributes to the understanding of the mechanisms that lead to the existence of UCEs and suggests that constraints derived from their multiple enrolment in diverse gene regulatory processes can explain the high level of evolutionary conservation observed in UCEs.

Results

A Novel Approach for UCE Identification Finds More than 1,500 UCEs Shared across 12 Drosophila Genomes

In this study we define UCEs as DNA elements of at least 50 nt that are identical across the genomes of 12 Drosophila species. Our design is such that the evolutionary distance between the most divergent fruitfly species in our dataset exceeds that of humans and reptiles (Stark et al. 2007; Makunin et al. 2013). Although the original criteria for UCE annotation relied on fewer species (Bejerano et al. 2004; Glazov et al. 2005), we reasoned that the inclusion of more species would be especially valuable for the identification and analysis of eUCEs which might overlap with coding sequences under strong purifying selection. To identify the full set of Drosophila UCEs, we extracted all unique 50-mers from the D. melanogaster genome and checked for their presence in the other 11 genomes using the short read mapper Bowtie (Langmead et al. 2009). All universal 50-mers were then extracted and, where appropriate, reassembled into longer UCEs (see Materials and Methods). In this manner, we identified a total of 1,516 Drosophila UCEs (fig. 1; supplementary table 1, Supplementary Material online).

We assessed the accuracy of our method by manually inspecting each of the 466 UCEs detected on the D. melanogaster chromosome arm 3R (31% of our total dataset) in the 15-way multiple alignment available from the dm3 release of the UCSC Genome Browser (https://genome.ucsc.edu/) (Methods and Materials; supplementary table 1, Supplementary Material online). We found that 345 out of the 466 UCEs (74.0%) were intact in the alignment: these elements aligned across the 12 Drosophila species without mismatches or gaps (fig. 1D). A further 115 UCEs (24.7%) could be recovered after taking into account different types of alignment ambiguities and errors, which might have prevented their detection using an alignment-based approach: 86 were disrupted due to the insertion of a gap at an ambiguous position, 19 had been split across alignment blocks although the sequence was contiguous in all 12 species and 10 appeared nonconserved due to more extensive alignment or assembly errors (supplementary table 1, Supplementary Material online). A total of six UCEs (1.3%) could not be verified to be present in a syntenic location in all genomes, typically because they occurred in multiple copies, due to duplications and/or assembly errors. Thus, our alignment-free method greatly increases the sensitivity of UCE detection in these 12 Drosophila genomes, at the cost of a small false positive rate in the order of 1%. As a case in point, a recent study of the same 12 species identified 98 UCEs of at least 80 nt based on the multiple alignment used above (Kern et al. 2015), whereas our method identified 131 such UCEs, an increase of 34%.

Notably, the UCEs discovered were not randomly distributed: they occurred in clusters within the D. melanogaster genome (observed median distance between UCEs: 18 kb; expected: 56 kb; \( P < 10^{-15} \), Mann–Whitney test). The largest UCE clusters overlapped with key developmental factors,
including the pair-rule gene odd-skipped (odd), the Hox genes Antennapedia (Antp) and Ultrabithorax (Ubx) (see below) and the Hox co-factor homothorax (hth) (table 1; Materials and Methods). The association between UCEs and developmental genes was further confirmed by gene ontology (GO) analysis (Ashburner et al. 2000; Eden et al. 2009) which revealed significant enrichments of GO categories such as “organ development”, “pattern specification process”, and “regulation of transcription from RNA polymerase II promoter” among the most UCE-rich genes (supplementary table 2, Supplementary Material online; Materials and Methods). Thus, our novel approach applied to distantly related Drosophila species points toward a general and robust association between ultraconservation and the genetic control of animal development, in line with earlier findings regarding Drosophila and mammalian UCEs (Sandelin et al. 2004; Glazov et al. 2005; Kern et al. 2015).
the fact that in mammals many UCEs serve as developmental enhancers (de la Calle-Mustienes et al. 2005; Woolfe et al. 2005; Pennacchio et al. 2006; Visel et al. 2008) prompted us to consider a plausible role of Drosophila UCEs in transcriptional regulation.

In contrast to the findings of an earlier study (Glazov et al. 2005) we found an enrichment of Drosophila UCEs in annotated regulatory regions: we detected 21 UCEs that overlapped with known regulators in the ORegAnno database (Griffith et al. 2008), which represented a 1.7-fold enrichment compared with randomly chosen genomic elements (P = 0.015, $\chi^2$ test). Motivated by this finding, which suggests a functional parallel between mammalian and Drosophila UCEs, we decided to undertake a more detailed investigation of the contributions of UCEs to transcriptional regulation in Drosophila.

For this we intersected our UCE annotations with data from the modENCODE consortium (Celniker et al. 2009) providing the binding sites of 34 TFs in early development (see Materials and Methods). UCEs that overlapped with annotated ncRNAs were excluded from this analysis, since many ncRNAs are highly expressed and therefore prone to give rise to false positives in chromatin immunoprecipitation (ChIP) experiments (Teytelman et al. 2013). We assessed each TF for enrichment or depletion within UCEs compared with reference elements and detected several cases where patterns of TF binding differed significantly between UCEs and reference elements (fig. 2A) suggesting that ultraconservation is associated with specific regulatory networks. The list of significantly enriched TFs included developmental factors such as Hairy, which plays a critical role in the segmentation of the early embryo (Nusslein-Volhard and Wieschaus 1980), Fruitless, which promotes male-specific neural development and behavior (Manoli et al. 2005), as well as Homothorax, a Hox protein co-factor (Ryoo et al. 1999).

We further reasoned that, if the enrichment of TFs within UCEs is biologically relevant in the context of Drosophila development, we might observe differential TF enrichment at various developmental time points. For one of the enriched TFs, Caudal (Cad) (Mlodzik et al. 1985), data were available for different developmental stages enabling us to use this protein to explore the ways in which TF binding relates to UCE function at different developmental time points. This analysis showed that in young embryos Cad binding was significantly enriched in iUCEs (intronic), while in adult flies there was a depletion of Cad binding in this type of UCE (fig. 2C). This observation is consistent with a role for UCEs in the dynamic transcriptional processes that control development. Taken together, our results indicate that many Drosophila UCEs, especially iUCEs, act as enhancers in particular during early Drosophila development.

UCEs Are Bound by Polycomb-Group Proteins

The protein Polycomblike (Pcl) was one of the most consistently enriched TFs in our analysis (fig. 2A). Because Pcl binds to Polycomb response elements (PRES) (Papp and Muller,

| Cluster ID  | Total UCEs | Intergenic | Intronic | Junction | Exonic | ncRNA | Flybase genes |
|------------|------------|------------|----------|----------|--------|-------|---------------|
| cluster_5  | 10         | 6          | 4        | –        | –      | –     | Cyp12e1       |
| cluster_10 | 10         | 10         | –        | –        | –      | –     | Hth           |
| cluster_94 | 10         | 10         | –        | –        | –      | –     | Fkh           |
| cluster_157| 10         | 10         | –        | –        | –      | –     | CG33648       |
| cluster_187| 10         | 10         | –        | –        | –      | –     | CG4218        |
| cluster_195| 10         | 10         | –        | –        | –      | –     | bru-3         |
| cluster_208| 10         | –          | 10       | –        | –      | –     | Annp2        |
| cluster_253| 10         | 11         | –        | –        | –      | –     | CG17025       |
| cluster_5  | 10         | 6          | 4        | –        | –      | –     | Slo           |
| cluster_10 | 10         | 10         | –        | –        | –      | –     | CG2267        |
| cluster_94 | 10         | 10         | –        | –        | –      | –     | CG31013       |
| cluster_157| 10         | 10         | –        | –        | –      | –     | PH4alphaPV    |
| cluster_187| 10         | 10         | –        | –        | –      | –     | CG34432/Spn100A|
| cluster_195| 10         | 10         | –        | –        | –      | –     | CG34433       |
| cluster_208| 10         | –          | 10       | –        | –      | –     | CG1342        |
| cluster_253| 10         | 11         | –        | –        | –      | –     | CG12069       |
| cluster_5  | 10         | 6          | 4        | –        | –      | –     | CG12066/Pka-C2|
| cluster_10 | 10         | 10         | –        | –        | –      | –     | CG131010      |
| cluster_94 | 10         | 10         | –        | –        | –      | –     | CG5397        |
| cluster_157| 10         | 10         | –        | –        | –      | –     | CG30647       |
| cluster_187| 10         | 10         | –        | –        | –      | –     | CG18822       |
| cluster_195| 10         | 10         | –        | –        | –      | –     | CG33259       |
| cluster_208| 10         | –          | 10       | –        | –      | –     | Hth           |
| cluster_253| 10         | 11         | –        | –        | –      | –     | CG13137       |
| cluster_253| 10         | 11         | –        | –        | –      | –     | CG14370       |
we speculated that UCEs might be associated with PcG proteins, which mediate epigenetic silencing of gene expression and target many genes with critical roles in development including the Hox genes (Steffen and Ringrose, 2014). To examine this possibility we looked at the binding patterns of Polycomb (Pc), Posterior sex combs (Psc), and Sex combs extra (Sce), which form part of the Polycomb repressive complex 1 (PRC1), Enhancer of zeste (E(z)), which forms part of Polycomb repressive complex 2 (PRC2), as well as the Trithorax-group protein Absent, small, or homeotic discs 1 (Ash1), which is associated with nonrepressed PcG targets (Steffen and Ringrose 2014). All four PcG proteins were significantly enriched at iUCEs (fig. 2B). Pc and E(z) were also enriched at eUCEs, while Psc and Sce were depleted from

![Figure 2](http://mbe.oxfordjournals.org/)

**Fig. 2.** Involvement of Drosophila UCEs in transcriptional regulation. (A) Significant (P < 0.05) enrichment or depletion of 34 TFs in UCEs relative to reference elements in early Drosophila development. A $\chi^2$-test was applied to each factor and UCE type, followed by Benjamini–Hochberg correction for multiple tests. The datasets used to generate this figure are listed in supplementary table 5, Supplementary Material online. Bab1, Bric a brac 1; Cad, Caudal; Chinmo, Chronologically inappropriate morphogenesis; Cnc, Cap-n-collar; CTCF, CTCF; D, Dichaete; Disco, Disconnected; Dll, Distal-less; En, Engrailed; Fru, Fruitless; Ftz-f1, Ftz transcription factor 1; GATAe, GATAe; H, Hairy; Hkb, Huckebein; Hr46, Hormone receptor-like in 46; Hth, Homothorax; Inv, Invected; Jumu, Jumeau; Kn, Knot; Kr, Kruppel; Lola, Longitudinals lacking; Pan, Pangolin; Pcl, Polycomblike; Prd, Paired; Run, Runt; Sc, Scute; Sens, Senseless; Sin3A, Sin3A; Stat92E, Signal-transducer and activator of transcription protein a t92E; Su(H), Suppressor of Hairless; Zfh1, Ttk, Tramtrack; Ubx, Ultrabithorax; Usp, Ultraspiracle; Zfh1, Zn finger homeodomain 1. (B) Enrichment and depletion of four PcG and one Trithorax-group proteins. Analyses were performed as in (A). Ash1, absent, small, or homeotic discs 1; E(z), Enhancer of zeste; Pc, Polycomb; Psc, Posterior sex combs; Sce, Sex combs extra. (C) Enrichment and depletion of Cad binding at five points of Drosophila development. Analyses were performed as in (A). Double asterisks indicate 0.01 < P < 0.001 and triple asterisks P < 0.001.

2006), we speculated that UCEs might be associated with PcG proteins, which mediate epigenetic silencing of gene expression and target many genes with critical roles in development including the Hox genes (Steffen and Ringrose, 2014). To examine this possibility we looked at the binding patterns of Polycomb (Pc), Posterior sex combs (Psc), and Sex combs extra (Sce), which form part of the Polycomb repressive complex 1 (PRC1), Enhancer of zeste (E(z)), which forms part of Polycomb repressive complex 2 (PRC2), as well as the Trithorax-group protein Absent, small, or homeotic discs 1 (Ash1), which is associated with nonrepressed PcG targets (Steffen and Ringrose 2014). All four PcG proteins were significantly enriched at iUCEs (fig. 2B). Pc and E(z) were also enriched at eUCEs, while Psc and Sce were depleted from
Multifunctional Sequences Are Strongly Enriched among eUCEs

Due to their involvement in protein encoding eUCEs are likely to be subjected to distinct functional requirements from those found in iUCEs. Apart from constraints on coding sequences, previous work highlighted RNA editing and splicing as two individual processes associated with eUCEs in flies and mammals (Bejerano et al. 2004; Glazov et al. 2005; Lareau et al. 2007; Ni et al. 2007), but their cumulative contribution had not been considered. We hypothesized that combinations of several molecular functions within a single sequence might collectively constrain sequence evolution and explain the presence of eUCEs in Drosophila.

To test whether this was the case we calculated the degree of multifunctionality observed for UCEs and reference elements by assessing each UCE in terms of protein-coding potential, presence of RNA editing sites (Ramaswami and Li 2014), overlap with intron–exon boundaries and presence of TF binding sites (Celniker et al. 2009). We observed a clear shift in the UCE distribution towards a larger number of molecular functions per element (fig. 3A; \( P = 1.9 \times 10^{-8} \), Mann–Whitney test) suggesting that multifunctionality is a core property of UCEs. As an example, our analysis showed that eUCEs are ~100-fold more likely than reference elements to be tetrafunctional, i.e., they are located in a protein-coding region, overlap with a splice site, contain an RNA editing site and are bound by at least one TF (fig. 3B, supplementary table 3, Supplementary Material online).

Notably, our analysis of eUCEs also revealed that almost all (96%) eUCEs overlap with alternatively spliced exons (expectation based on reference elements: 47%; \( P = 6.8 \times 10^{-12} \), \( \chi^2 \)-test) suggesting an important role for ultraconservation in the generation of alternative transcripts. This observation prompted us to consider specific eUCEs to experimentally test the hypothesis that ultraconservation might be related to multifunctional roles played by UCEs (see below).

Multifunctionality Underlies Ultraconservation of an Alternatively Spliced Exon in the Hox Gene Ubx

One of the eUCEs in our dataset overlaps with a small exon (51 nt) in the Drosophila Hox gene Ubx (fig. 4A). Given that the Hox genes represent a gene class characterized for its high number of UCEs in mammals (Lampe et al. 2008; Lin et al. 2008) we decided to investigate the Ubx eUCE in higher detail. Alternative splicing of this Ubx exon, known as microexon I (ml), as well as of an additional small exon, microexon II (mII), generates functionally distinct Ubx isoforms (Reed et al. 2010; de Navas et al. 2011). The demonstrated molecular and developmental relevance of ml alternative splicing together with the evolutionary conservation of Ubx splicing patterns and motifs across distantly related Drosophila species (Bomze and Lopez 1994; Hatton et al. 1998) brought us to hypothesize that ml ultraconservation might be a consequence of overlapping constraints derived from the necessity of maintaining both a particular protein-coding sequence and specific splicing regulatory elements within ml.

To evaluate this hypothesis, we first investigated to what extent the ml protein-coding sequence showed signs of purifying selection. Using BLAST (Altschul et al. 1990), we were able to locate the ml sequence in the genomes of three outgroup species to the Drosophila clade: the common housefly Musca domestica, the tsetse fly Glossina morsitans, and the Mediterranean fruit fly Ceratitis capitata (Materials and
**Fig. 4.** Analysis of an ultraconserved exon within the *Ubx* gene. (A) The *Ubx* gene (purple) contains 12 UCEs (red triangles) within the transcribed region. Coding sequences are shown as thick boxes and untranslated regions (UTRs) as narrow boxes. The gene is transcribed from left to right. One of the UCEs overlaps with the alternatively spliced mI exon. An alignment of the mI-UCE sequence from 12 Drosophila species and three more distantly related fly species is shown. Substitutions relative to the Drosophila sequence are shown in orange and insertions as a vertical line, with the number of inserted bases added above the sequence. The light blue box highlights the mI exon, while the surrounding sequences are intronic. The amino acid sequence (corresponding to the Drosophila nucleotide sequences) is shown above the alignment. For positions with observed substitutions, it is noted below the alignment whether these are synonymous (s) or nonsynonymous (n). (B) To explore the roles of microexon mI ultraconservation in *Ubx* splicing control we engineered a series of *Ubx* minigene constructs so that they included wild type (wt) or mutated versions of microexon mI (mutA) in which the protein coding potential of the gene was maintained while the ultraconserved nucleotide sequence of mI was disrupted by means of synonymous mutations (red). Approximate positions of splicing primers *Ubx* E1F (forward) and *Ubx* 3′ U (reverse) and expression primers expF/R (forward/reverse) are indicated. (C) Experiments in *Drosophila* Schneider 2 (S2) cells. Semi-quantitative RT-PCR analysis of wild type and mutA *Ubx* minigenes expressed in S2 cells reveals distinct patterns of *Ubx* mRNA splicing where the mutA minigene construct shows a marked reduction of *Ubx* Ia isoform production. *Ubx*.AS refers to the signal detected with primers *Ubx* E1F and *Ubx* 3′R (see B) which detects all alternative splicing variants of the gene; *Ubx*.exp denotes signal amplified with primers expF/R (see B) which are
We then estimated the rate of nonsynonymous substitutions relative to the rate of synonymous substitutions \((dN/dS)\) with codeml \((\text{Yang } 1997)\). The \(dN/dS\) value was 0.045, consistent with strong purifying selection acting on the ml coding sequence. The conservation of ml at nonsynonymous sites can therefore be fully or partly explained by coding constraints. Next, we tested whether the conservation at synonymous sites was due to selective constraints or a consequence of insufficient divergence times between species, which might not have allowed mutations at synonymous sites to occur. To this end, we compared the ml exon with the Ubx homeodomain, a sequence encoding 60 amino acids (aa) identical across all 12 investigated Drosophila species. Focusing on the third position of each codon, we found that 29 out of 57 sites within the homeodomain had synonymous substitutions. Using this as our reference we were thus able to exclude that the lack of synonymous mutations at the 15 synonymous sites within the ml exon was due to chance \((P = 0.0002, \text{Fisher’s exact test})\) supporting our hypothesis that coding constraints contribute to, but cannot fully explain, ultraconservation within the ml exon.

Building on these observations we decided to use the UCE in Ubx ml (ml-UCE) to test the possibility that it functions as a protein-coding element as well as a docking region for splicing factors. We reasoned that if the latter were true, an experiment where coding sequence capacity is maintained while the nucleotide sequence is modified should potentially expose such splicing-related functions. To test this model, we decided to carry out a series of experiments where the original sequence of the ml-UCE was replaced by one in which the coding potential was unaffected but the nucleotide composition distorted (fig. 4B). Both versions of the ml-UCE (wild type and mutated (mutA)) were subcloned within a Ubx splicing minigene previously shown to produce a relatively complex pattern of alternatively spliced products (Hatton et al. 1998). We then proceeded to test the alternative splicing patterns that resulted from these constructs in Drosophila S2 cells in culture. Here, we observed that the expression level of ml-bearing mRNAs was significantly reduced in the presence of the ml mutation (fig. 4C). These results together with the fact that overall expression levels of Ubx mRNAs do not differ across the genotypes suggest that ml mutation indeed affects Ubx alternative splicing in Drosophila cells.

To explore the extent to which our observations in cultured cells were also valid in the physiological context of the developing fruitfly embryo, we created a series of transgenic Drosophila lines carrying wild type and mutA versions of ml-UCE in Ubx minigenes whose expression could be controlled via the Drosophila UAS/Gal4 system \((\text{Brand and Perrimon } 1993)\). For our analysis we chose to activate the expression of the Ubx minigenes within the embryonic central nervous system (CNS) given that the splicing patterns of Ubx are complex within this tissue \((\text{Lopez and Hogness } 1991; \text{Artero et al. } 1992; \text{Reed et al. } 2010; \text{Thomsen et al. } 2010)\). Remarkably, the results of these experiments in developing embryos (fig. 4D and E) nicely match those observed in cultured cells: lower levels of ml-bearing mature mRNAs were detected in the presence of the ml mutation.

Furthermore, the biological significance of the observed changes in the pattern of Ubx alternative splicing generated via UCE mutation is evident given that Ubx splicing isoforms have been shown to: \((1)\) have different abilities to bind to DNA targets \(\text{in vitro (Reed et al. } 2010)\), \((2)\) display isoform-specific gene activation patterns \(\text{in vivo}\) in two developmental contexts: within the developing embryo, \((\text{regulation of both: the decapentaplegic (dpp) promoter and endogenous gene})\), and during the formation of adult appendages—\(\text{is that is regulation of wingless, araucan, and spalt genes in wing and haltere imaginal discs (Reed et al. } 2010; \text{de Navas et al. } 2011)\), \((3)\) perform different roles during the development of the peripheral nervous system (PNS) in the embryo \((\text{Reed et al. } 2010)\) and during haltere development in the adult \((\text{de Navas et al. } 2011)\), and \((4)\) induce different patterns of neural differentiation during the formation of the embryonic nervous system \((\text{Rogulja-Ortmann et al. } 2014)\).

All in all, the data of our experimental analysis in cultured cells and \(\text{in vivo}\) support a model by which the ml-UCE performs functions that affect the process of alternative splicing; these data support our hypothesis that UCEs may have retained their sequences over long evolutionary periods due to their intrinsic multifunctionality in regards to gene expression control.

**Discussion**

Despite a decade of research since the discovery of UCEs the molecular basis of ultraconservation remains unknown. So far, most UCE studies have been carried out in mammals, yet Drosophila UCEs are currently receiving renewed attention \((\text{Makunin et al. } 2013, 2014; \text{Kern et al. } 2015)\), not least because of the clear phenotypic effects that can be observed.

Fig. 4. (Continued) positioned in the 3’ exon, a constitutive segment of Ubx mRNAs. (D) Expression of Ubx wild type and Ubx mutA minigenes in the Drosophila embryo. We produced HA-tagged UAS versions of wt and mutA Ubx minigenes \((\text{see A})\) and generated independent transgenic UAS-lines with insertions in identical chromosomal loci by means of site-specific recombination. The resulting UAS-Ubx lines \((\text{wt and mutA})\) were crossed with the elav-gal4 \((\text{elav})\) driver to express Ubx transgenes selectively within the developing embryonic nervous system. Note that expression patterns obtained with anti-HA antibodies in the embryonic CNS and PNS were identical across genotypes confirming comparable gene expression conditions. (E) Semi-quantitative RT-PCR analysis of wt and mutA Ubx minigene expression in the embryonic Drosophila nervous system reveals effects of ml on Ubx splicing control. In line with the results obtained in S2 cells \((\text{see C})\) we observed that the mutA minigene produced a reduced amount of Ubx isoform when compared with its wild-type counterpart. (\text{See C for definition of labels Ubx_AS and Ubx.exp and text for further details}) Statistical analyses: \(**P < 0.01\) and \(*P < 0.05\) obtained in one-tailed t-test \((P\text{-value } S2\text{ cells } = 0.0035 (**); \text{P-value embryos } = 0.0318 (*))\). Error bars indicate standard error of the mean. HA, haemagglutinin tag; B, Ubx B-element; ml, microexon I; mlI, microexon II.
in UCE mutants in fruitflies (Makunin et al. 2013). Nevertheless, the molecular functions of Drosophila UCEs have not been extensively investigated, leaving open the question of whether similar molecular mechanisms underlie ultraconservation in insects and mammals. Here, we combine a computational approach with molecular experiments in Drosophila cells and embryos and gene regulatory data from the modENCODE consortium (Celniker et al. 2009) to provide a modern and comprehensive functional overview of more than 1,500 UCEs shared across 12 Drosophila genomes.

Previous work suggested that Drosophila UCEs are not involved in transcriptional regulation (Glazov et al. 2005; Kern et al. 2015). In contrast, our analysis of the binding sites of 34 TFs revealed that many iUCEs likely serve as enhancers in early fly development, similarly to what has been observed in mammals (Pennacchio et al. 2006; Visel et al. 2008). Furthermore, at least in some cases, UCE enhancer activity appears to be temporally restricted as suggested by our observations of Cad binding at different developmental time points. We also observed significant enrichment of PcG proteins at UCEs, thus extending previous findings of high conservation at individual PREs (Dellino et al. 2002) and proteins at UCEs, thus extending previous findings of high conservation at different developmental time points. We further observed distinctive enrichment of transduction of PcG proteins at UCEs in mouse embryonic stem cells (Viturawong et al. 2013); alternatively, these seemingly distinct results might reflect the dynamic nature of PcG-mediated silencing, especially given that another study reported an association between PcG proteins and highly conserved mammalian sequences (Lee et al. 2006). Taken together, our analysis of the binding sequences of TFs and Polycomb proteins support an integral role for iUCEs in developmental gene regulation.

In contrast, eUCEs show less pronounced patterns of TF binding enrichment and depletion, suggesting that they do not primarily function as transcriptional regulators. Indeed, eUCEs have been associated with other regulatory processes, such as alternative splicing and RNA editing (Glazov et al. 2005; Kern et al. 2015). In this regard, our data show a statistically significant association between eUCEs and alternatively spliced exons. Although we sought to determine whether UCEs had a distinctive link with splicing regulatory elements the size of our eUCE dataset was too small to probe this possibility using computational methods (Materials and Methods).

Furthermore, our analysis shows that multiple functional layers are frequently superimposed on a single eUCE sequence. For example, we show that eUCEs are nearly 100-fold more likely than expected to simultaneously contain protein-coding sequence, TF binding sites, splice sites, and RNA editing sites. This finding adds to the growing literature on regulatory sequences embedded within coding regions (Lin et al. 2011; Stergachis et al. 2013; Birnbaum et al. 2014) and suggests that many eUCEs represent extreme cases of “genomic multitasking”.

We experimentally evaluated one such multitasking element, an eUCE overlapping the short mI exon within the Hox gene Ubx. Although the protein-coding sequence is under strong purifying selection, our comparison between the mI sequence and that of the well-conserved homeodomains showed that protein-coding constraints alone were insufficient to explain the ultraconservation of this exon. As the mI exon is alternatively spliced, we used a previously developed splicing minigene system (Hatton et al. 1998) to analyze splicing patterns of the wild-type Ubx gene and a version of Ubx where the mI exon contained synonymous substitutions. The results of these experiments showed clear differences in Ubx splicing, both in cell culture and fly embryos, demonstrating that the mI exon carries out two functions in parallel: it encodes an evolutionarily conserved protein sequence and regulates its own splicing. Interestingly, many alternatively spliced short exons overlap with UCEs also in mammalian genomes (Bejerano et al. 2004), suggesting that our findings might be relevant to the study of UCEs in other animal groups.

The work presented above shows several common features shared between UCEs in mammals and Drosophila. An intriguing possibility that emerges from this study is that these similarities reflect common mechanisms underlying ultraconservation in distantly related species. In addition, our findings support the hypothesis that UCEs are sculpted by overlapping functional constraints, in particular for eUCEs, and suggest that further functional dissection of Drosophila UCEs will lead to general insights into the selective forces that shape gene regulatory elements in animal genomes.

In summary, we have performed a functional survey of 1,516 UCEs that are shared across 12 Drosophila genomes. Our findings support a role for iUCEs in the transcriptional and epigenetic regulation of genes involved in early fly development. In addition, we found that eUCEs are shaped by cumulative functional constraints and are two orders of magnitude more likely than expected to contain protein-coding sequence, TF binding sites, RNA editing sites, and splice sites within a single element. We experimentally characterized an eUCE found in the Hox gene Ultrabithorax and showed that the extreme conservation of this element is due to both protein-coding constraints and the presence of splicing regulations that modulate the balance of biologically distinct Ubx isoforms in cell culture and developing fly embryos. Our results highlight similarities between UCEs in Drosophila and mammals, pointing to a shared molecular mechanism underlying these independently evolved elements.

**Materials and Methods**

Identification of Ultraconserved Elements in 12 Genomes

Genome assemblies for *D. ananassae* (droAna3), *D. erecta* (droEre2), *D. grimshawi* (droGri2), *D. melanogaster* (dm3), *D. mojavensis* (droMoj3), *D. persimilis* (droPer1), *D. pseudoobscura* (dp4), *D. sechellia* (droSec1), *D. simulans* (droSim1), *D. virilis* (droVir3), *D. willistoni* (droWil1), and *D. yakuba* (droYak2) were downloaded from the UCSC
Genome Browser (Adams et al. 2000; Kent et al. 2002; Richards et al. 2005; Drosophila 12 Genomes Consortium 2007). We extracted all 50-mers that occurred a single time in the D. melanogaster genome and mapped them to the other 11 genomes using Bowtie release 0.12.7 (Langmead et al. 2009). Only 50-mers with perfect matches (bowtie -v 0) in all genomes were kept for further analysis. To retrieve full-size UCEs, we fused overlapping 50-mers into longer sequences and verified that the reconstituted elements were found in all genomes.

Earlier studies of Drosophila UCEs have defined ultraconservation using different criteria (Glazov et al. 2005; Makunin et al. 2013; Kern et al. 2015). We chose a cutoff of 50 nt to be consistent with the original study by Glazov et al (2005) that focused on UCEs in the D. melanogaster and D. pseudoobscura genomes. Based on the overall similarity of these two genomes, Glazov et al (2005) estimated the false positive rate in their study to be only 0.4%. Given that our analysis included ten additional species, of which three are more distantly related, the proportion of UCEs in our dataset that are explained by overall sequence similarity should be substantially lower than 0.4% and can therefore be considered negligible.

Unlike previous approaches, our method does not rely on whole-genome alignments and does not incorporate information on synteny. To assess whether our dataset included UCEs that did not occur at syntenic positions across the 12 genomes, we carefully inspected the 466 UCEs located on the D. melanogaster chromosome arm 3R in the Multiz 15-way whole-genome alignment available from the dm3 release of the UCSC Genome Browser (http://genome.ucsc.edu/), which was previously used by Kern et al (2015). We first divided UCEs into four groups: (1) perfect correspondence between our annotation and the alignment, (2) perfect correspondence after the position of an ambiguously placed gap had been adjusted, (3) perfect correspondence, but due to an outgroup species the alignment spanned two or more alignment blocks, and (4) incomplete correspondence. For the UCEs in the fourth group, we assessed whether our annotated UCE occurred in a syntetic position, which we defined as the region between the two UCEs that neighbored the UCE in D. melanogaster. Instances where we did not find our annotated UCE in a syntetic position, and where this could not be explained by alignment or assembly errors, were due to multiple occurrences of the UCE and highly similar sequences in one or more of the nonmelanogaster genomes. In these cases, the annotated UCE in D. melanogaster had been aligned to a UCE-like sequence, which typically contained only a single mismatch (supplementary table 1, Supplementary Material online).

Annotation of UCE Clusters
We grouped UCEs if they occurred within less than the median distance (18 kb) of each other. This resulted in 288 UCE clusters, comprising 1,043 UCEs (supplementary table 4, Supplementary Material online). To explore the association between UCEs and protein-coding genes, we chose to focus on the 15 largest clusters (each comprising at least 10 UCEs), as the limited quality of some of the analyzed genomes prevented us from performing a global clustering and synteny analysis for all species. For these clusters, we checked the overlap with protein-coding genes in the D. melanogaster, D. pseudoobscura, and D. virilis genomes (table 1).

Comparisons of UCEs and Genomic Reference Elements
We created a reference dataset by dividing the D. melanogaster genome into fragments with the same length distribution as the UCE set. All reference elements were required to map uniquely to the genome (bowtie -v 0 -m 1). The reference elements were grouped into functional classes (intergenic, intronic, exonic, or ncRNA) in the same manner as the UCEs. We used BEDTools (Quinlan and Hall 2010) to define the overlap between UCEs or reference elements with various genomic annotations, and carried out the statistical analyses with R version 2.12.2 (R Development Core Team 2011).

GO Analysis
We searched for GO categories that were associated with the most UCE-rich genes using the GOrrila tool (Ashburner et al. 2000; Eden et al. 2009), which identifies enriched GO terms for ranked gene lists. We ranked genes based on the number of UCEs within each gene, including flanking regions of 10 kb upstream and downstream, divided by the number of reference elements in the same interval.

Analysis of modENCODE Data
We analyzed ChIP-chip and ChIP-seq data from the modENCODE consortium (Celniker et al. 2009). TFs were chosen based on the association with the GO term GO:0003700 (sequence-specific DNA binding TF activity) (Ashburner et al. 2000) and the availability of data from embryos younger than 12 h. PcG and TrxG proteins were chosen based on annotations provided by Steffen and Ringrose (2014). For the PcG/TrxG analysis, we did not limit our analysis to a specific developmental stage, but merged all available datasets for each protein. A list of the precise datasets we used is included in supplementary table 5, Supplementary Material online. We tested for enrichment or depletion of each factor within intergenic, iUCEs or eUCEs relative to reference elements (see above) using a \( \chi^2 \)-test. Correction for 117 multiple tests was performed using the Benjamini–Hochberg method.

Global Analysis of Alternative Splicing
To evaluate the overlap between eUCEs and alternatively spliced exons, we downloaded D. melanogaster coordinates for constitutive and nonconstitutive exons from Ensembl release 79 (Cunningham et al. 2015) and intersected these with our set of eUCEs and exonic reference elements using BEDTools (Quinlan and Hall 2010).

We obtained sequences of 99 putative exonic splicing enhancers from Brooks et al. (2011). For each motif, we counted the number of matching eUCEs and exonic reference elements (see above) and evaluated the statistical significance a \( \chi^2 \)-test. Correction for 99 multiple tests was performed using the Benjamini–Hochberg method. In addition, we used
Phylogenetic Analysis of the m1 Exon

We used the m1 aa sequence in a tblastn BLAST search (Altschul et al. 1990) against the NCBI nucleotide collection (http://blast.ncbi.nlm.nih.gov) and recovered m1 in the common housefly M. domestica (Scott et al. 2014) and the Mediterranean fruit fly C. capitata (http://www.hgscbcm.edu), but not in more distantly related species, such as the mosquitoes Anopheles gambiae and Aedes aegypti, the flour beetle Tribolium castaneum or the honey beeApis mellifera.

We used the same method to identify the m1 exon in the genome of the tsetse flyG. morsitans (International Glossina Genome Initiative 2014), made available through VectorBase (Mey et al. 2012). To generate an alignment of the surrounding introns, we extended the sequences by including 50 nt on each side of the m1 exon and aligned the sequences with MUSCLE (Edgar 2004). For the coding part of the alignment, we estimated dN/dS for the whole tree using codeml and standard settings (Yang 1997).

Fly Stocks and Embryo Collections

Virgins of transgenic UAS-lines were crossed to males of 69B-Gal4 (Bloomington 1774) or elav-Gal4 (a gift of Matthias Soller, Birmingham, United Kingdom) fly lines. Embryos were collected at 25°C in the dark on apple juice plates supplemented with yeast paste following standard procedures.

Generation of Ubx-mutA Constructs

The Ubx.4 plasmid, which carries a Ubx wild-type minigene, was originally developed in the laboratory of Javier Lopez (Hatton et al. 1998). We introduced synonymous mutations into the ml exon (fig. 4B) by PCR-driven overlap extension (Heckman and Pease 2007). The PCR fragment was cloned into the pGEM-T Easy vector (Promega), which was sequentially digested with AffI and PmlI to release a 255 nt fragment. The fragment was then cloned into the Ubx.4 plasmid to generate the derivative construct Ubx.mutA.

Subcloning and Transgenesis

To create UAS-Ubx minigene expression constructs bearing wild-type (wt) or mutated versions of the ml microexon (mutA) the following procedures were employed. For the creation of pUAS.Ubx.mini.HA.short attB (HA.short) the Ubx wild type minigene was released from Ubx.4 via a partial NruI digest followed by a SacII digestion; the resulting 6241 nt fragment was cloned into pBluescript to form pBSII.SK(+)Ubx.mini.HA.short. Subsequently, the minigene was released by sequential SpeI/blunting through T4 Polynuclrase/KpnI digests and transferred to the transformation vector pUASP.K10.attB (a gift from Beat Suter, Bern, Switzerland) (Koch et al. 2009) that had been sequentially treated with NdeI/blunting through T4 Polynuclrase/KpnI. To create the expression construct pUAS.Ubx.mini.HA.short.mlmutAattB (HA.short.mlmutA), a mutant version of ml was transferred from Ubx.4.mlmutA to pBSII.SK(+).

Ubx.mini.HA.short using a Ndel/NsiI digest; the mutant minigene was then transferred to pUASP.K10.attB as described above. All enzymes were from NEB. Injection of UASP.attB constructs as well as the screening for and balancing of transformants was performed by BestGENE (http://www.thebestgene.com/) using the ZH-attB-51C landing site (Bischof et al. 2007).

Antibody Labeling

At the protein level Ubx-mini transgene expression was detected by antibody-stains using enzymatic, alkaline phosphatase detection using standard protocols. In brief, we used anti-hemagglutinin (HA) (Covance; 1:400) followed by anti-mouse-biotin (Sigma-Aldrich, 1:200) and streptavidin-alkaline phosphatase conjugates (Roche, 1:5,000); enzymatic detection was performed using NBT/BCIP (Roche) substrate.

S2 Cell Experiments

Drosophila S2 cells were transfected using Effectene transfection reagent (Qiagen, Valencia, CA) according to the manufacturer’s protocol. Typically, 1.5 million cells were transfected with 200 ng DNA (10 and 40 ng of Plasmid-minigene, Ubx-wt, or mutant, were transfected) and incubated for ~65 h before collection of RNA.

RT-PCR Analysis

Total RNA was isolated (Sigma RNA Minikit) followed by DNase treatment. cDNA was generated with oligo-dT primer using either 1 or 3 μg of total RNA with M-MuLV reverse transcriptase from NEB (20 μl reaction volume; 1 h at 42°C). 1 μl of cDNA was used as a template for PCR to amplify isoforms. PCR primers Ubx_E1F: TGGAATGCCAATTGAGTGGAAGGAGCGCAGATTA3′ and Ubx_E1R: 5′CGCGTCTTTGCAGACCATTT3′ were used to detect the different isoforms [PCR cycles 31× (94°C 45 s, 56°C 1 min, 72°C 30 s)]. More PCR cycles amplified other isoforms but gave inconsistent results indicating that the reaction was not in linear range and were not considered further. Minigene expression level was monitored by primers expF: 5′AGTGGAAGGAGCGCAGATTG3′/Ubx3′R 5′CGCGTCTTTGCAGACCATTT3′ [PCR cycles 25× (94°C 30 s, 56°C 20 s, 72°C 20 s) amplifying a product of 103 nt. Products were separated on an 8% nondenaturing polyacrylamide gel and quantified with MultiGauge (Fujifilm). Background corrected intensity values were normalized to the general gene level and the average and standard error from the three replicas was calculated presented in log2 scale.

Supplementary Material

Supplementary tables S1–S5 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Adam Eyre-Walker, Margarida Cardoso Moreira, and members of the Alonso Lab for helpful discussions. This study was funded by an EMBO long-term fellowship to M.W. (ALTF 1589-2011) and a Wellcome Trust Investigator Award to C.R.A. (098410/Z/12/Z).
