OPEN ACCESS Freely available online

PLOS ONE

Comparison of Ultra-Conserved Elements in Drosophilids and Vertebrates

Igor V. Makunin[1,2]*, Viktor V. Shloma[2], Stuart J. Stephen[3], Michael Pheasant[1], Stepan N. Belyakin[2]

1 Research Computing Centre, The University of Queensland, Brisbane, Queensland, Australia, **2** Institute of Molecular and Cellular Biology SD RAS, Novosibirsk, Russia, **3** Computational Biology Group, CSIRO Plant Industry, Canberra, Australian Capital Territory, Australia

Abstract

Metazoan genomes contain many ultra-conserved elements (UCEs), long sequences identical between distant species. In this study we identified UCEs in drosophilid and vertebrate species with a similar level of phylogenetic divergence measured at protein-coding regions, and demonstrated that both the length and number of UCEs are larger in vertebrates. The proportion of non-exonic UCEs declines in distant drosophilids whilst an opposite trend was observed in vertebrates. We generated a set of 2,126 Sophophora UCEs by merging elements identified in several drosophila species and compared these to the eutherian UCEs identified in placental mammals. In contrast to vertebrates, the Sophophora UCEs are depleted around transcription start sites. Analysis of 52,954 *P-element*, *piggyBac* and *Minos* insertions in the *D. melanogaster* genome revealed depletion of the *P-element* and *piggyBac* insertions in and around the Sophophora UCEs. We examined eleven fly strains with transposon insertions into the intergenic UCEs and identified associated phenotypes in five strains. Four insertions behave as recessive lethals, and in one case we observed a suppression of the marker gene within the transgene, presumably by silenced chromatin around the integration site. To confirm the lethality is caused by integration of transposons we performed a phenotype rescue experiment for two stocks and demonstrated that the excision of the transposons from the intergenic UCEs restores viability. Sequencing of DNA after the transposon excision in one fly strain with the restored viability revealed a 47 bp insertion at the original transposon integration site suggesting that the nature of the mutation is important for the appearance of the phenotype. Our results suggest that the UCEs in flies and vertebrates have both common and distinct features, and demonstrate that a significant proportion of intergenic drosophila UCEs are sensitive to disruption.

Citation: Makunin IV, Shloma VV, Stephen SJ, Pheasant M, Belyakin SN (2013) Comparison of Ultra-Conserved Elements in Drosophilids and Vertebrates. PLoS ONE 8(12): e82362. doi:10.1371/journal.pone.0082362

Editor: Denis Dupuy, Inserm U869, France

Received July 8, 2013; **Accepted** October 24, 2013; **Published** December 13, 2013

Funding: IMV and MP were supported by the Genomics Virtual Lab (GVL) project and NeCTAR,the Australian Government projects conducted as part of the Super Science initiative and financed by the Education Investment Fund. The work in Novosibirsk was supported by grants from Russian Foundation for Basic Research #11-04-01344, 12-04-01007, 12-04-00160, 12-04-01030, 12-04-00874-a. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: makunin@hotmail.com

Introduction

Comparative analysis of mammalian and insect genomes have demonstrated that the majority of the evolutionarily constrained sequences in these lineages are located outside of protein coding regions [1,2]. Comparison of the human and mouse sequences on chromosome 21 showed that many non-coding sequences are even more conserved than protein-coding regions [3]. Subsequent studies then identified numerous highly conserved non-coding elements (CNEs) in species as evolutionarily distant as human and fish [4,5], which are clustered around genes involved in regulation of transcription and development [6].

The Ultra-Conserved Elements (UCEs) are arguably the most constrained sequences in the human genome. The UCEs were first identified as sequences at least 200 bp long identical between the human, mouse and rat genomes [7]. Another study described 13,736 human UCEs identical over at least 100 bp in at least 3 of 5 placental mammals [8], and shorter UCEs were identified between human and phylogenetically distant species such as sponge, sea anemone, fruit fly and sea urchin [9]. An alignment-independent method was used for the identification of both syntenic and non-syntenic Long Identical Multispecies Elements in vertebrates and plants [10] including some elements omitted in earlier studies due to gap alignment deficiencies.

While UCEs were identified on the basis of sequence identity, they likely fall into several functional categories. The majority of UCEs do not overlap with protein-coding regions. Some exonic UCEs are apparently involved in regulation of splicing [11,12]. Non-exonic UCEs can act as enhancers [13,14]. Many UCEs are transcribed [15–17]. It is still unknown why UCEs maintain 100% sequence conservation over such long stretches of DNA, considering that the majority of known protein-binding sites are relatively very short. One possible explanation would be an overlap of several constraints, for example enhancer/protein binding and non-coding RNA [18] or enhancer and protein-coding regions [19]. It seems that distances between the UCEs are also conserved [20] raising the possibility that such sequences could be involved in the maintenance of higher-order genome structure. To some extent the conservation of distances between the UCEs can be explained by the absence of annotated transposons in the vicinity of many UCEs [21]. Such transposon-free regions are maintained in bony vertebrates [22] and often

coincide with so-called chromatin bivalent domains [23] hinting at a possible link between UCEs and chromatin. CNEs are apparently linked to maintenance of long syntenic regions in both vertebrates and insects [6,24].

The observed extreme conservation within UCEs suggests a strong negative selection pressure, implying that any disruption of the UCEs may have dramatic negative consequences for an organism. Indeed, some mutations outside of protein-coding regions are very harmful [25]. However, deletion of four individual UCEs had no visible effect on mice [26]. Similarly, deletion of long "gene desert" regions containing many CNEs also had no detectible effect on mice viability or phenotype [27]. On the other hand, over evolutionary history, deletions of ultra-conserved-like elements are observed to be over 300-fold less likely than deletions of neutral DNA [28]. Insertion of 16 bp sequences into the highly conserved Dc2 enhancer of *DACH1* gene did not cause any detectable changes in expression of a marker gene [29]. One study described an association between SNPs in UCEs and breast cancer [30], but the results were not reproduced in a different population [31]. While some UCEs are altered in cancer [15] it is not clear whether such changes represent driver or passenger mutations. SNPs in UCEs associate with complex traits such as BMI and height but show no transmission bias from parents to children, ruling out any strongly deleterious effect of these rare alleles [32].

It is plausible that the disruption or deletion of UCEs may have a strong deleterious impact on survival in complex natural environments but would have very little effect on fitness under controlled lab conditions as observed for non-coding RNA BC1 [33]. Indeed, an SNP in an ultraconserved regulatory sequence is linked with *Dlx5/Dlx6* expression in the forebrain [34]. Another paper reported an enrichment of UCEs in chromosomal rearrangements, especially pathogenic deletions, identified in 200 people with idiopathic neurodevelopmental disorders [35].

Despite the extreme conservation in mammals, none of the non-coding human UCEs were traced outside the vertebrate lineage [7] even though thousands of human protein-coding genes have detectable orthologs in insects [36]. None of the 481 UCEs identified by Bejerano and co-authors [7] overlap with the UCEs identified between phylogenetically distant species such as human and demosponge, hydra or sea anemone [9]. However, a recent study has described 183 CNEs conserved between mammals, fishes and tunicates, of which 145 overlap with an extended set of 5,404 vertebrate UCEs [37]. This contradiction between the extremely high conservation of UCEs in mammals and amniotes and the nearly complete lack of homologs even in distant Chordata species could be explained by a non-uniform substitution rate in such sequences over evolutionary time. Analysis of the extended set of 13,736 UCEs revealed an extremely low substitution rate within amniotes whilst in amphibian and bony fish lineages these regions evolved with higher substitution rates [8]. Another study found that many UCEs identified in Eutherian species can be found in a cartilaginous fish, the elephant shark, but it also confirmed that a significant number of non-exonic UCEs had first appeared in tetrapods and amniotes [38].

It seems the changes in substitution rates within CNEs has been very common over evolutionary history [39]. Non-uniform evolutionary substitution rates can reflect changes of function by a sequence, either gaining of a new biological functionality or loss of the existing. Such change in function may transform selection pressure on the sequence. The term "exaptation" was coined to describe acquisition, or "cooption" of a new function with a positive effect on fitness [40]. Consistent with this idea, some UCEs originated from ancient mobile elements [41,42] and their current sequence constraints are presumably no longer due to their original mobile element functionality. It is interesting to note that while there is no similarity between CNEs in distant species, some orthologous genes acquired highly conserved CNEs in different lineages: out of 156 human CNE-associated genes with invertebrate orthologs, 40 are also associated with CNEs in worms and flies [43].

UCEs have also been identified in insects [44,45]. Accepting that a very limited number of insect species genome assemblies have been analyzed, the UCEs in these insects show somewhat different properties compared to vertebrates. They are less frequently associated with genes encoding transcriptional factors, the UCE elements are shorter and the longest UCEs tend to overlap splice sites or reside in exons [44]. The longest identified drosophila UCE, at an exon-intron junction of the *homothorax* (*hth*) gene, has a complementary sequence located downstream in an intron of the gene and potentially may form an alternative RNA secondary structure and regulate its alternative splicing [46].

In this work we compare UCEs in insects and vertebrates. Using available genome sequences from several Drosophila and vertebrate species we have identified and compared UCEs in insect and vertebrate species with similar phylogenetic distances estimated from protein-coding sequences. We analyzed transposon insertions within eleven intergenic UCEs and identified visible phenotypes in five fly strains including four lethals. To prove the link between transposon insertion into intergenic UCEs and lethality we performed "phenotype rescue" experiments for two fly stocks with *P-elements* insertions. The removal of the transposons from the intergenic UCEs restored viability, confirming the association between the insertions and lethality.

Materials and Methods

Sequences

The 15-way drosophila dm3 centric and the 28-way human hg18 centric alignment were downloaded from the UCSC Genome Browser website [47] and used as the reference datasets when identifying putative UCEs. A sliding window of 100 bp over the multiZ alignments was used to identify minimal length seeds which were 100% identical between the relevant subset of aligned species (*e.g., D. melanogaster vs D. yakuba vs D. erecta*). Identified minimal length seeds were then maximally extended until either a base mismatch or the extent of the multiZ alignment region containing the species subset was reached. The following genome assemblies were used: dm3, droEre2, dp4, droAna4, droGri2, droMoj3, droVir3, droWil1, droYak2, hg18, canFam2, anoCar1, bosTau3, danRer4, fr2, galGal3, gasAcu1, mm8, monDom4, ornAna1, oryLat1, tetNig1, xenTro2. The vertebrate UCEs coordinates were converted to the hg19 by liftOver [48]. An UCE was considered conserved if at least 20 nucleotides were aligned in other species. We used FlyBase 5.12 and the refSeq genes annotations (dm3, July 30, 2012; hg19, August 1, 2012). A UCE with any overlap with an annotated exon was considered exonic.

The Phylogenetic Distances for Insects and Vertebrates

The dm3 and hg18 genome annotations on the UCSC Genome browser [47] were queried with "rpl" and "rps" to identify ribosomal genes and the refSeq IDs were extracted. For the non-conserved gene set we arbitrary selected 306 Drosophila genes on chromosome 2L without annotated alignments to the *Anopheles gambiae* genome. Only one protein-coding isoform per refSeq gene was used (Dataset S1). The bed files for selected genes were extracted from the UCSC Table Browser and were uploaded to

the Galaxy web site [49]. The alignments of coding regions were extracted using Stitch Gene blocks on 28-way multiZ alignment of hg18 or 15-way dm3 alignment and concatenated using the "Concatenate FASTA alignment by species" function. The concatenated alignments were imported into MEGA4 phylogenetic software [50]. Four-fold degenerate sites were exported as separate alignments and used for reconstruction of the phylogenetic trees in the PhyML software [51] with default parameters except that the ratio of transversions and transitions were autocalculated. The pairwise distances for non-synonymous substitutions within ribosomal genes were calculated in MEGA4 with the Pamilo-Bianchi-Li model and the complete deletion option for gaps and missing data [50].

Syntenic Blocks

We used the 22 largest syntenic blocks (HCBs) identified with conserved gene order (GO) criterion [52] and containing at least 21 independent gene anchors. The coordinates were converted to dm3 using the liftOver function. For the Chi-square test we assumed a uniform distribution of UCEs.

Statistics

The expected proportion of insertions or TSS features in any region of interest was calculated as the proportion of that region's total length to the length of the genome (*i.e.*, assuming a random distribution of features across the genome). For UCEs outside of intercalary heterochromatin regions we used TSS and insertions located outside of those regions. Chi-square tests were calculated in Excel, and for 2×2 contingency tables we used an online calculator: http://faculty.vassar.edu/lowry/tab2×2.html. Gene Ontology analysis was performed using FuncAssociate 2.0 [53] at http://llama.med.harvard.edu/funcassociate/.

Mutations/insertions in UCEs

Coordinates of insertion sites were downloaded from FlyBase [54] as following: the Insertions section of FlyBase was queried with "P{*", "pBac{*" and "Mi{*" for *P-element*, *piggyBac* and *Minos* inserts, respectively. The data on the integration sites were downloaded using HitList Conversion tools. Data for the *P-elements* were downloaded on February 11, 2010, and *piggyBac*s and *Minos* on January 5, 2012. In total we used 33,481 *P-elements*, 15,355 *piggyBac* and 4,118 *Minos* inserts mapped with precision of less than 10 bp within euchromatin. Analysis of insertion and UCEs distribution was done using the UCSC Table Browser [47].

Genetics and Molecular Biology

Flies were raised on a standard drosophila cornmeal–yeast–agar medium at 25°C. The fly stocks carrying insertions within intergenic Sophophora UCEs were ordered from Exelixis: *PBac{WH}f06142* (stock ID: f06142), *PBac{PB}c00059* (c00059), *PBac{WH}f02223* (f02223), *PBac{WH}f05912* (f05912), *P{XP}d07857* (d07857), *PBac{WH}f07151* (f07151), *PBac{WH}f02632* (f02632) and *PBac{PB}c06670* (c06670) [55]. Three other fly stocks were ordered from the Bloomington stock center: *P{SUPor-P}KG10325* (stock ID 15254), *PBac{WH}Rdl*^{f02994} (18606), *P{SUPor-P}KG02042* (14258) [55,56]. The *y w; Ki, P{*π*}*$^{+}$*, Δ2–3}99B* (y[1] w[1]; Ki[1] P{ry[+t7.2] = Delta2–3}99B) flies used as source of the transposase (Bloomington stock center, http://flybase.org/reports/FBst0004368.html). The *y, w; If/CyO; MKRS/TM6b* stock was used for balancing chromosomes after *P-element* excisions. DNA around insertion site after excision of the *P-element* in *d07857* flies was amplified and sequenced with the following primers: d07857_1_d CACCCCTCCACCTAACCAA and d07857_1_r CGATCTGTGATCTTGTGATTGATC. The sequences were aligned to the *D. melanogaster* genome by BLAST at the FlyBase site [54].

Results

Comparison of Phylogenetic Distances in Vertebrates and Flies

The rate of substitution is higher in insect than in vertebrate genomes. For example, the *Sophophora* and *Drosophila* subgenera diverged ~36 million years ago [57] and the Amniota diverged ~310 million years ago, yet the phylogenetic distance (estimated from fourfold degenerate sites) between *D. melanogaster* (*Sophophora* subgenus) and *D. virilis* (*Drosophila* subgenus) exceeds the phylogenetic distance between the amniotes human and chicken [58]. In our work we focus on the comparison of UCEs in sets of species with similar phylogenetic distances.

Phylogenetic distance can be measured by the neutral substitution rate, commonly estimated using divergence in assumed non-functional regions of the genome such as four-fold degenerate (synonymous) sites or syntenic ancient transposons, but this can be problematic comparing highly divergent genomes where sites become saturated with substitutions (*e.g.*, when the number of substitutions per site exceeds 1). Alternatively, relative phylogenetic distance can be estimated using non-synonymous substitutions in conserved protein-coding genes. Both approaches have problems. Drosophilids have very few transposons in the euchromatic part of their genomes; hence ancient transposons essentially cannot be used for estimation of the neutral substitution rate in flies. Divergence at four-fold degenerate sites has been used for comparative analysis of vertebrates and flies [58]. However, a very strong codon bias linked to gene expression level [59] results in a significant variation of the substitution rate at these sites in Drosophila genomes. To demonstrate the significance of this phenomena, we calculated the phylogenetic distances in four-fold degenerate sites of highly expressed genes encoding ribosomal proteins and 306 genes from *D. melanogaster* chromosome 2L without annotated orthologs in the *Anopheles gambiae* genome (Figure 1) by the PhyML software [51] used in comparative analysis of drosophilids and vertebrates [58]. While the divergence of vertebrates measured at four-fold degenerate sites of genes encoding ribosomal proteins is very similar to the whole genome estimation [58], the phylogenetic distances for drosophilids are significantly smaller. The phylogenetic distance measured at four-fold degenerate sites of 306 non-conserved genes exceeds the genome average data (Figure 1) and is 1.7–2 times greater than the distances determined for highly expressed ribosomal genes. A similar tendency was observed with pairwise distances estimated using the MEGA4 software [50] but the difference was less pronounced (data not shown). The data suggest that highly expressed genes can introduce a significant bias in estimation of the neutral substitution rate in drosophilids and that the phylogenetic distances reported for drosophilids [58] might be underestimated.

As a comparison we used the divergence of protein sequences for estimation of the relative phylogenetic divergence, calculating pairwise phylogenetic distances of vertebrate and drosophilid species using non-synonymous substitutions in ribosomal genes (Figure 1B). This showing that the relative distances between *D. melanogaster* and other species from the *Drosophila* subgenus (such as *D. virilis*) are similar to or just slightly greater than the distance between human and reptiles, in agreement with results obtained by whole genome analysis [58]. However, it is possible that the highly expressed genes have a lower substitution rate in

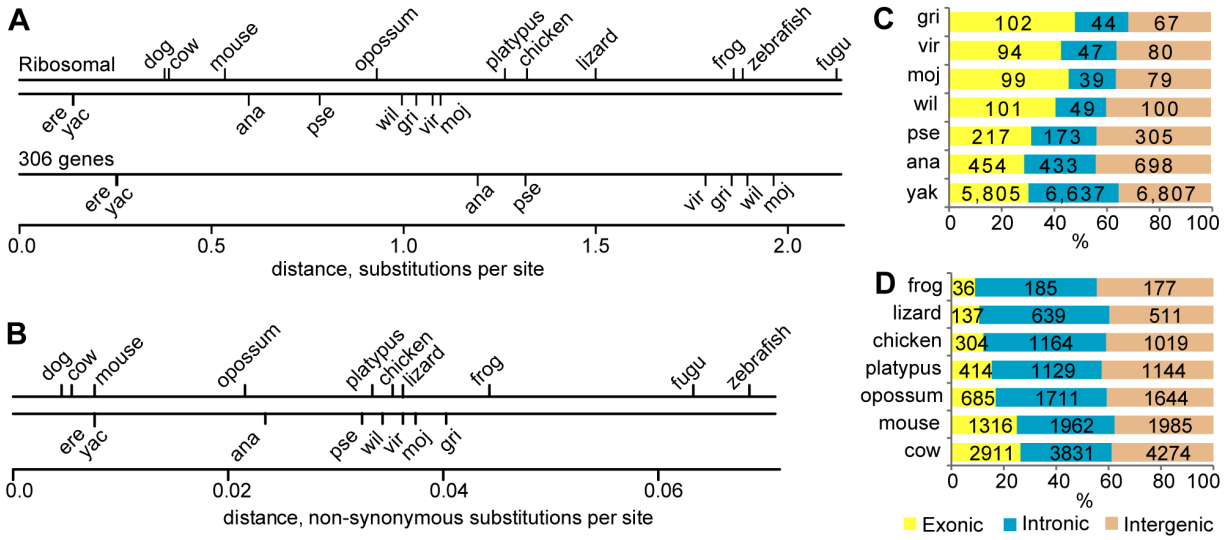

Figure 1. Comparison of phylogenetic distances and UCEs in drosophilids and vertebrates. (**A**) The pairwise distances at four-fold degenerate sites were estimated from the phylogenetic trees generated by PhyML package. The human and *D. melanogaster* genomes were used as a master sequence. Note the difference in distance estimations between the highly expressed genes encoding ribosomal proteins and 306 non-conserved genes in flies. Abbreviation for the Drosophila species: ere – *D. erecta*, yac – *D. yakuba*, ana – *D. ananassae*, pse – *D. pseudoobscura*, wil - *D. willistoni*, vir – *D. virilis*, moj – *D. mojavensis*, gri – *D. grimshawi*. (**B**) The pairwise distances for non-synonymous sites within the ribosomal genes. Only one transcript isoform per gene was used. (**C**) UCEs identified in different drosophila sets split into exonic, intronic and intergenic according to the refSeq genes (September 16, 2013). (**D**) UCEs identified in vertebrates. The fish sets were excluded due to a small number of the elements.
doi:10.1371/journal.pone.0082362.g001

drosophilids similar to the bias at four-fold degenerate sites (Figure 1A). Our results demonstrate that estimations of phylogenetic divergences in drosophilids and vertebrates vary between different approaches.

Comparison of UCEs sets between Vertebrate and Fly Species with Similar Phylogenetic Distance

To investigate how the phylogenetic divergence of species affects the number and features of UCEs we identified sequences 100 nt or longer identical in sets of three species of drosophilids or vertebrates (Table 1, Datasets S2 and S3). The use of three species instead of pairwise comparisons reduces the chance of identification of cross-contamination and other errors present in genome assemblies. For flies we used *D. melanogaster*-centric alignments, for vertebrates we used human-centric alignments. *D. erecta* and dog were used as the second species for insect and vertebrate sets, and the phylogenetic difference was associated with divergence of a third species used (Table 1). For convenience the name of the third species was used as the name of the set. As an approximate measure of divergence we used the total length of branches for three species in each set calculated at four-fold degenerate sites in ribosomal genes for vertebrates and in 306 non-conserved Drosophila genes by the PhyML software [51]. As discussed in the previous section these numbers should be regarded as a very approximate estimation.

The divergence of the drosophilid sets is either smaller or larger than the divergence of the placental mammals sets (Table 1). However, the divergence in the ananassae and pseudoobscura sets is comparable to the divergence of species in the opossum, platypus and chicken sets, and the remaining drosophila sets are comparable to the lizard and frog sets. Comparison of the sets with similar divergence shows that flies have fewer UCEs than the matching vertebrate species (Table 1), and the elements in flies are shorter (Figure 2). For example, the pseudoobscura set has 695 UCEs compared to 2,687 and 2,487 UCEs in the platypus and chicken sets, respectively. The pseudoobscura set has just 6 UCEs longer or equal to 200 bp, while both platypus and chicken sets have 50 times more elements of the same length (Table 1). In sets with a larger divergence, such as mojavensis and frog, the difference is less pronounced. The number of UCEs declines dramatically in fish sets, apparently due to different selection pressure in this lineage [8]. While the estimations of divergence at four-fold degenerate sites of the mojavensis and zebrafish sets are very close to each other (Table 1), it is worth pointing out that the estimation of relative phylogenetic distances (non-synonymous substitutions) indicates that the fish is more divergent (Figure 1B).

Some sets with similar phylogenetic divergence show a dramatic difference in the number of identified UCEs. For example, the expected number of substitutions per four-fold degenerate site in the ananassae and pseudoobscura sets is 1.3 and 1.4, respectively, but the number of UCEs in these sets differs by more than a factor of two (Table 1). A similar situation is observed in the chicken and lizard sets. We concluded that the number of UCEs does not always reflect the phylogenenetic distance measured by neutral substitutions. We tested whether discordance between a phylogenetic distance and number of observed UCEs could be attributed to a potential difference in the genome assembly quality. We selected 1,136 UCEs from the ananassae set that do not overlap with the UCEs from the pseudoobscura set and mapped these UCEs to the pseudoobscura genome using the liftOver function. Out of 1,136 sequences 1,070 (92%) mapped to pseudoobscura indicating that the orthologous sequences are present in *D. pseudoobscura* for the majority of ananassae UCEs but many sequences do not fit the strict criteria used for UCEs identification due to the presence of substitutions or indels.

The maximal length of UCEs identified in the insect and amniote sets remains more or less the same while the number of UCEs decreases significantly with increasing phylogenetic distance (Table 1). For example, the longest elements identified in the ananassae and mojavensis sets differ in length by just ~3% despite

Table 1. UCEs identified in different species.

Species	Distance sbt/ site#	UCEs, count	Average length, bp	Median length, bp	UCEs 200+ bp, count	Max length, bp
melanogaster-erecta-						
yakuba	0.348	19,249	124	116	412	520
ananassae	1.322	1,585	117	110	17	301
pseudoobscura	1.445	695	116	110	6	246
virilis	1.915	221	117	110	2	210
grimshawi	1.983	213	119	110	3	210
willistoni	2.024	250	116	110	3	293
mojavensis	2.091	217	119	110	3	293
human-dog-						
cow	0.561	11,016	149	129	1,466	770
mouse	0.745	5,263	142	126	520	770
opossum	1.139	4,040	147	129	518	653
platypus	1.47	2,687	146	129	315	586
chicken	1.504	2,487	148	131	322	610
lizard	1.71	1,287	144	130	138	615
frog	2.068	398	132	122	15	391
zebrafish	2.093	22	119	112.5	0	168
fugu	2.337	20	116	114	1	215

Distance, in substitutions per four-fold degenerate site, was calculated for ribosomal genes (vertebrates) and 306 non-conserved genes on chromosome 2L (drosophilids).
doi:10.1371/journal.pone.0082362.t001

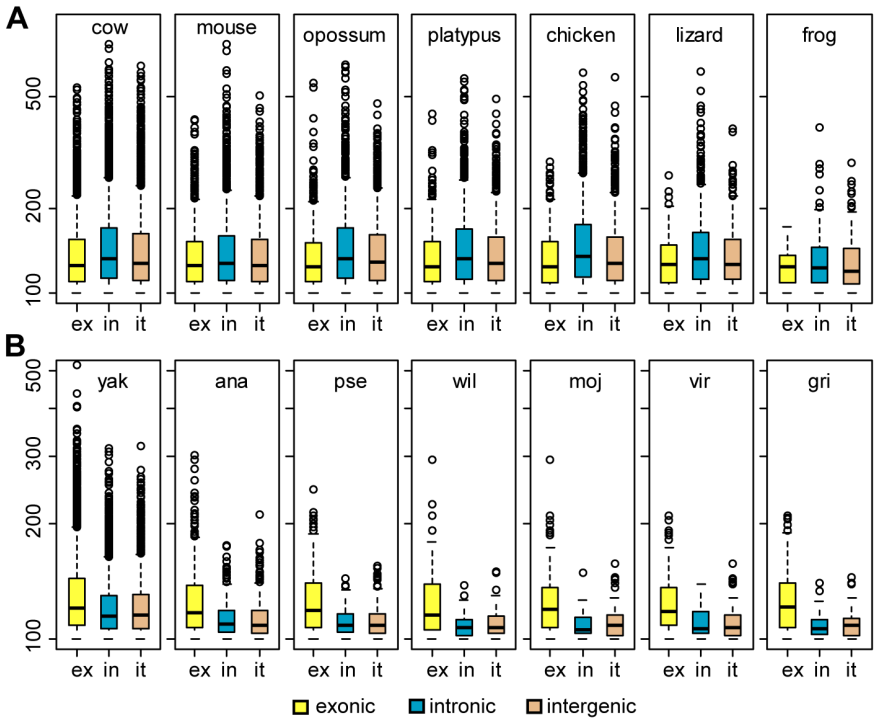

Figure 2. Comparison of UCEs length in vertebrates and flies. (**A**) Vertebrate sets. Data for the UCEs identified in fish are not shown due to a small number of the elements. (**B**) Drosophilids. The abbreviations are the same as on Figure 1. The length of UCEs in bp is shown on the y-axes. The UCEs are split into exonic (ex), intronic (in) and intergenic (it) based on the refSeq genes model. For drosophila the annotation was downloaded on September 16, 2013. Boxes represent upper and lower quartiles, with median shown as a black line. The whiskers were drawn using the R boxplot default 1.5 of interquartile range.
doi:10.1371/journal.pone.0082362.g002

the fact that the ananassae set contains only species from the *Sophophora* subgenus while the mojavensis set contains species from both the *Sophophora* and *Drosophila* subgenera. The maximal length of UCEs is very similar in all analyzed amniotes sets but declines significantly in the amphibian and fish sets (Table 1). This is in agreement with our previous observation of the elevated substitution rate within eutherian UCEs in the amphibian and fish lineages [8].

We compared UCEs in each set with refSeq gene models in the human and *D. melanogaster* genomes. The majority of UCEs identified in vertebrates and insects are non-exonic, but significant differences exist between the two groups. In drosophilids the proportion of exonic UCEs increases with phylogenetic distance while in tetrapods the fraction of exonic UCEs decreases in distant species (Figure 1). This indicates that in tetrapods the constraints leading to maintenance of the UCEs are stronger in non-coding regions. In all drosophilid sets the density of UCEs in intergenic regions is greater than in introns while the opposite is observed in vertebrates (Table S1). In flies the exonic UCEs are longer than the non-exonic and in mammals the opposite is observed (Figure 2). In fact, the 10 longest UCEs in every fly set analyzed overlap exons of FlyBase genes except for the virilis set in which 9 of the 10 longest UCEs are exonic. In flies the median length of intronic and intergenic UCEs is very similar while in mammals and amniotes intronic UCEs are longer than intergenic. Despite changes in the relative frequencies of exonic, intronic and intergenic UCEs and the general decline in UCE numbers in sets with distant species, the average length of UCEs remains essentially identical in each class of UCEs (Table 1).

Combined Set of Sophophora UCEs 100+

The criteria used for the identification of UCEs require 100% identity between three distant species. However, UCEs do evolve – albeit at a very slow rate – and such simple criteria would miss some elements. To overcome this problem we identified UCEs in four sets of drosophila species with the phylogenetic distances comparable to those in distant mammals (melanogaster-erecta-ananassae, melanogaster-erecta-pseudoobscura, melanogaster-yakuba-ananassae and melanogaster-yakuba-pseudoobscura) from *Sophophora* subgenus and merged these UCE sets into one superset similar to the approach used for the identification of eutherian UCEs in five placental mammals [8]. The resulting Sophophora UCEs set (Dataset S4) contains 2,126 ultra-conserved elements covering 249.1 kb, with the longest element being just 301 nt long. Almost half of Sophophora UCEs are intergenic (1,018 or 47.9%), 575 (27.0%) are intronic and 533 (25.1%) are exonic (using the refSeq gene models). The combined Sophophora UCEs set has slightly smaller fraction of exonic UCEs than any of the individual UCEs sets suggesting that in flies either exonic UCEs are more conserved or many UCEs avoid detection due to either alignment or assembly problems.

The distribution of Sophophora UCEs varies significantly between chromosomes (Table 2). Only two UCEs were identified on chromosome 4, and the density of UCEs on chromosome X is half of that on chromosome 3R. Regions on long chromosomes adjacent to pericentric heterochromatin are depleted of UCEs (Figure S1). For example, there are no UCEs in the first 2.4 Mb of chromosome 2R assembly or the last 1.6 Mb of chromosome 3L. However, Sophophora UCEs are enriched in regions of intercalary heterochromatin scattered along euchromatin [60]. These regions occupy 14.1% of the *D. melanogaster* genome but contain 450 (21.2%) Sophophora UCEs (1.5 fold enrichment, Chi-square test, *P*-value 1.6E-20). The Sophophora UCEs are enriched in long intergenic regions, at significant distance from annotated genes, *e.g.*, 548 (25.4%) Sophophora UCEs are located at least 5 kb away from any annotated refSeq gene, and such regions occupy 16.2 Mb (13.5%) of the Drosophila genome (1.9 fold enrichment, *P*-value 2.7E-58, Chi-square test). Only 37 Sophophora UCEs overlap known copy number variants, or CNVs [61], 1.5 times less than expected (Chi-square test, *P*-value 0.01).

The vast majority of Sophophora UCEs are conserved in other drosophila species. Even in distant species from the *Drosophila* subgenus (*D. mojavensis*, *D. virilis*, *D. grimshawi*) orthologous sequences were found for about 97% of Sophophora UCEs, with average identity slightly above 91% to the *D. melanogaster* sequence, very similar to the conservation of eutherian UCEs in amniotes [8]. The conservation of Sophophora UCEs declines dramatically outside of drosophilids: only 499, 385 and 429 elements are conserved in mosquito, red flour beetle and honey bee, with an average identity 74%, 69% and 69%, respectively. Among 299 Sophophora UCEs conserved in all three non-drosophilid species, 289 overlap exons of FlyBase protein coding genes indicating that very few non-exonic Sophophora UCEs are conserved in distant insects.

Sophophora UCEs are Enriched in Conserved Syntenic Blocks but Depleted around Promoters

Many non-exonic UCEs in vertebrates associate with genes involved in regulation of development and transcription [7] and are often embedded within large conserved regions [14,27]. We analyzed the distribution of Sophophora UCEs in the 22 largest conserved syntenic regions, or homologous collinear blocks, HCBs [52]. Out of 2,126 Sophophora UCEs, 204 (9.6%) overlap with 22 HCBs covering approximately 7.4 Mb (6.1%) of the *D. melanogaster* genome, 1.6 fold more than expected.

The transcription start sites (TSSs) of human refSeq genes are enriched in eutherian UCEs and their flanking regions (Figure 3). The enrichment of the eutherian UCEs with the TSSs remains strong even after exclusion of 32 TSSs corresponding to annotated miRNAs. In contrast, in the *D. melanogaster* genome, TSSs of refSeq genes are under-represented within the Sophophora UCEs and flanking regions up to 5 kb (Figure 3). Out of 25 TSSs mapped within the Sophophora UCEs, 12 correspond to 5′ ends of miRNA, snoRNA and snRNAs, and one TSS corresponds to a short non-coding RNA *tre-3* from the *bxd* locus [62]. The TSSs of protein coding FlyBase genes show a threefold decrease in UCEs (data not shown). Within a distance up to 3 kb from the UCEs the number of annotated TSSs is observed to be only half that expected. The Sophophora UCEs are over-represented in gene-poor regions of the intercalary heterochromatin. However, under-representation of TSS near the Sophophora UCEs outside of the intercalary heterochromatin is nearly identical to that observed for the whole genome but with slightly higher *P*-values presumably due to a slightly smaller number of TSSs (Table S2). It seems that in mammals UCEs have somewhat different properties compared to those in flies.

In the human genome 767 refSeq gene TSSs map within 1 kb of the eutherian UCEs. These TSSs belong to 613 genes, 69 miRNAs and 5 snoRNAs. The Gene Ontology annotations [63] were available for 542 genes, and these genes are strongly enriched in the categories linked to development and regulation of transcription (Table S3). Among these genes 164 (30.3%) are assigned to the *transcription factor activity* category (GO:0003700), so 17.2% of 956 genes encoding transcription factors have at least one TSS within 1 kb of eutherian UCEs. Interestingly, among genes not assigned to the GO terms, we noticed four non-coding anti-sense transcripts, DLX6AS, HOXA11AS, OTX2OS1 and

Table 2. Distribution of the Sophophora UCEs in the *D. melanogaster* genome.

Chromosome	Size, Mb	refSeq genes (isoforms)	UCEs, count	UCEs per Mb	UCEs per 100 genes
chrX	22.4	4,208	272	12.1	6.5
chr2L	23.0	4,728	366	15.9	7.7
chr2R	21.1	5195	333	15.7	6.4
chr3L	24.5	4,786	476	19.4	9.9
chr3R	27.9	6,153	677	24.3	11.0
chr4	1.4	291	2	1.5	0.7
Total	120.4	25,361	2,126	17.7	8.4

doi:10.1371/journal.pone.0082362.t002

SOX2OT, which highlights a putative link between ultra-conservation and regulatory non-coding RNAs [64].

Drosophila genes assigned to categories related to regulation of transcription are also overrepresented among these in proximity to UCEs (Table S3), even though the TSSs of the *D. melanogaster* genes generally do not overlap Sophophora UCEs and flanking regions. Out of 160 protein coding FlyBase genes with TSSs located within or less than 1 kb from the Sophophora UCEs, 122 are assigned to GO categories and 27 (22.1%) are assigned to the category *transcription factor activity* (GO:0003700), corresponding to a 5.9 fold enrichment. However, these 27 genes represent only 7.2% of 374 genes assigned to that category, a significantly smaller proportion than in human.

Distribution of *P-element*, *piggyBac* and *Minos* Insertions in and around the Sophophora UCEs

We analyzed insertions of transposon-based gene vectors into Sophophora UCEs and neighboring regions in *D. melanogaster*. Out of 52,954 *P-element*, *piggyBac* and *Minos* insertions, 59 are mapped to 52 Sophophora UCEs (Table 3). Both *P-elements* and *piggyBac*s are depleted in regions adjacent to the Sophophora UCEs, and *P-elements* are also depleted from UCEs (Table 4) while the distribution of *Minos* inserts is close to the expected (data not shown). It is well known that transposons are distributed non-uniformly in the genome (Figure S1). Both *P-elements* and *piggyBac*s are under-represented in the intercalary heterochromatin [65], and *P-elements* are strongly biased to TSSs [56] while distribution of the Sophophora UCEs shows the opposite bias. However, the UCEs located outside of the intercalary heterochromatin regions show a low density of insertions in the adjacent regions (Table S4).

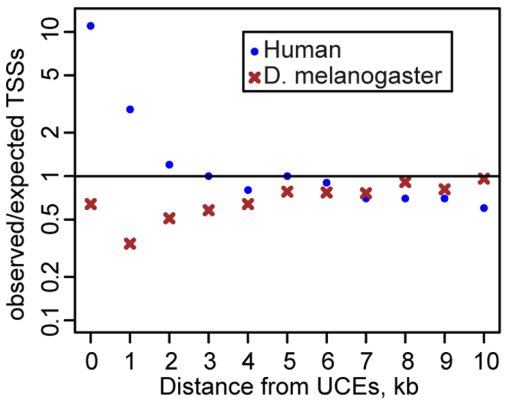

Figure 3. Ratio of the observed and expected refSeq genes transcription start sites in the UCEs and adjacent regions. The eutherian and Sophophora UCEs were used for the human and *D. melanogaster* genomes, respectively. Distance between the UCEs and nearest TSS is on the x-axis. The observed ratio is show on log scale. Expected TSSs were calculated assuming a random distribution.
doi:10.1371/journal.pone.0082362.g003

It is possible that the observed low insertion density around the Sophophora UCEs is caused by an inactivation of the marker gene in transposons that may prevent detection of insertions as hypothesized for the intercalary heterochromatin [65]. To test this hypothesis we analyzed the distribution of 2,852 *P-elements* with known suppression status. This set contains 383 insertions with partial suppression of the mini-*white* marker gene [65]. In total, 196 *P-elements* from this dataset are located within 5 kb from the Sophophora UCEs. Among these, 36 are suppressed, which is 1.4 fold more than expected (Chi-square test, *P*-value 0.04). While it does suggest a statistically significant enrichment of suppressed transgenes in proximity to the Sophophora UCEs, the numbers are too small to be definitive.

Phenotypes Associated with Transposon Insertions into UCEs

Very little is known about the effects of the disruption of non-exonic UCEs in animals. We studied the consequences of transposon integration into Sophophora UCEs. Because it is clear that insertions into genes (exons or introns) could impair gene function due to interference of the insertion with transcription or disruption of the protein-coding region, we focused only on insertions into intergenic UCEs as defined by the FlyBase genes annotation 5.12 (Table 5). Of the 18 intergenic UCEs with insertions, four overlap snRNAs. An insertion in one fly stock, *f02994*, is annotated as an allele of the *Rdl* gene, and it behaves as a recessive lethal [55,66]. None of the remaining insertions were characterized as lethal or causing sterility in FlyBase. We examined 11 fly stocks carrying insertions in 11 intergenic Sophophora UCEs and identified visible phenotypes in five stocks. In one strain the files have a variegating eye color due to a partial suppression of the mini-*white* marker gene within the transposon integrated into UCE chr2L.116 (Figure S2), three stocks including *f02994* behave as recessive lethals, and a reduced viability was observed in *KG02042* flies (Table 5). However, after rebalancing of the *KG02042* insertion with two different balancers, TM6B-Tb and MKRS, the insertion also behaved as a recessive lethal.

To prove the lethality is caused by the insertions, we performed "phenotype rescue" experiments and removed the transposons from the intergenic UCEs in two fly stocks on chromosome 3L, *d07857* and *KG02042*. Both stocks carry *P-elements* with mini-*white* marker gene responsible for the red eye color in w^- genetic background. After excision of the transposon flies would have

Table 3. Transposon insertions into the Sophophora UCEs.

Transposon type	Analyzed insertions	in exonic UCEs#	in intronic UCEs#	in intergenic UCEs#	Total	%
P-element	33,481	5	7	12	24	0.07
piggyBac	15,355	9	9	11	29	0.19
Minos	4,118	5	1	0	6	0.15
All inserts	52,954	19	17	23	59	0.11

As defined by FlyBase Genes 5.12.

doi:10.1371/journal.pone.0082362.t003

white eyes. If the lethality is caused by the insertion, we expect to see disappearance of the phenotype in the case of perfect excision of the transposon ("phenotype rescue"), while in the case of partial excision or deletion of DNA adjacent to the integration site we may see a preservation of the phenotype.

The flies with the inserts were crossed to the flies carrying an active transposase to activate the transgenes (Figure 4), and the flies with white eyes were recovered in the F2 progeny for both *d07857* and *KG02042* stocks indicating that the *P-elements* had jumped out of the original integration sites. Among F4 flies with white eyes we found both restoration of viability (reversion of the phenotype) and maintenance of lethality for both original stocks. We established the *Re1-d07857* fly stock with a complete reversion of the phenotype as a homozygote for the third chromosome and *Re3-d07857* stock with the recessive lethal with *CyO* balancer chromosome and analyzed DNA around the insertion site in these flies.

DNA around the integration site in *Re1-d07857* flies with a complete reversion of the phenotype was amplified by PCR and sequenced. Comparison with the reference sequence revealed an additional 47 bp sequence at the integration site (Figure 4B) suggesting that the presence of the very short insertion in the UCE does not affect the viability. In *Re3-d07857/CyO* flies with recessive lethal the PCR produced a single DNA fragment identical to wild type, presumably from the balancer chromosome.

Discussion

We compared the UCEs in vertebrates and drosophilids and found that both the number and length of UCEs are smaller in insects, in agreement with previous reports [44,45]. The abundance of UCEs in vertebrates can be partially explained by two rounds of whole genome duplications and the subsequent retention of duplicated copies of the key developmental genes, such as HOX clusters or DLX genes alongside their conserved regulatory elements including UCEs [7]. UCEs can be derived from transposable elements [41] through the process of exaptation, and transposons are abundant in vertebrate genomes. Similar to a recent report [9], the number of UCEs in drosophilid and in vertebrate species negatively correlates with the evolutionary distance between species, however, some sets of species with similar levels of sequence divergence differ significantly in their numbers of UCEs (*e.g.* the ananassae and pseudoobscura sets, Table 1).

A comparison of phylogenetic distances in insects and vertebrates is complicated by differences in their genome structures, such as a shortage of ancient transposons in drosophilids and a strong codon bias in highly expressed Drosophila genes. In addition, methylated CpG sites show a high substitution rate in vertebrates but Drosophila, having lost the CpG methylation system, has a somewhat different mutation pattern. While comparison of divergences between drosophilids and vertebrates is challenging, it is less problematic within each lineage. We also would like to point out that the fly species used in the analysis come from one genus, *Drosophila* (order Diptera, class Insecta), while the vertebrate species include representatives of different classes, *e.g.*, Mammalia, Aves, *etc*. We identified numerous UCEs in species belonging to different classes of vertebrates but none were found between drosophilids and another Diptera species, *Anopheles gambiae* (data not shown). On the evolutionary time scale, the divergence between the most distant drosophila species used in the analysis occurred about 36 million years ago

Table 4. Distribution of the *P-element* and *piggyBac* inserts in the UCEs and flanking regions.

Regions	Size,kb	P-element inserts	Obs/Exp	P-value, Chi test	PBac inserts	Obs/Exp	P-value, Chi test
UCEs	249	24	0.35	5.2E-8	29	0.91	0.6
1 kb	3,880	483	0.45	5.7E-76	260	0.53	7.1E-27
1–2 kb	3,352	536	0.57	1.5E-39	303	0.71	1.0E-9
2–3 kb	3,007	468	0.56	4.4E-38	245	0.64	7.7E-13
3–4 kb	2,748	495	0.65	6.4E-23	237	0.68	8.5E-10
4–5 kb	2,566	450	0.63	1.9E-23	232	0.71	1.0E-7
5–6 kb	2,395	538	0.81	5.5E-7	249	0.82	1.1E-3
6–7 kb	2,251	394	0.63	7.8E-21	216	0.75	2.3E-5
7–8 kb	2,135	432	0.73	2.1E-11	224	0.82	3.1E-3
8–9 kb	2,018	509	0.91	2.6E-2	198	0.77	1.9E-4
9–10 kb	1,927	439	0.82	2.4E-5	216	0.88	5.5E-2

doi:10.1371/journal.pone.0082362.t004

Table 5. Transposon insertions into intergenic Sophophora UCEs.

Insertion	UCE position (dm3)	Phenotype
PBac{WH}f06142	chrX:20868423–20868526	None
PBac{PB}c00059	chr2L:7357956–7358059	Variegating eye colour
P{SUPor-P}KG10325	chr2L:12796474–12796574	None
P{EPgy2}snRNA:U2:34ABa^{EY07636}	chr2L:13212004–13212116	Not examined
P{EP}snRNA:U2:34ABb^{G2309}	chr2L:13215838–13215951	Not examined
P{GSV6}GS15147	chr2L:13244450–13244567	Not examined
P{SUPor-P}snRNA:U2:34ABc^{KG07625}	chr2L:13244450–13244567	Not examined
PBac{WH}f02223	chr2L:15466371–15466474	None
P{GSV6}GS14066	chr2R:15626716–15626818	Not examined
PBac{WH}f05912	chr3L:3890591–3890691	None
PBac{WH}Rdl^{f02994}	chr3L:9176791–9176898	Lethal
P{SUPor-P}KG02042	chr3L:9294864–9294963	Semi-lethal/lethal
P{EP}EP3253	chr3L:10369759–10369863	Not examined
P{EP}EP2008b	chr3L:10369759–10369863	Not available
PBac{WH}f00315	chr3L:10730347–10730465	Not examined
PBac{RB}e03261	chr3L:10730347–10730465	Not available
P{EP}EP1091	chr3L:10730347–10730465	Not examined
P{XP}d07857	chr3L:12686673–12686786	Lethal
PBac{WH}f07151	chr3L:20943084–20943183	Lethal
PBac{WH}f02632	chr3R:1894272–1894388	None
PBac{PB}c06670	chr3R:9261319–9261447	None
PBac{PB}c06574	chr3R:9261319–9261447	Not examined
P{wHy}snRNA:U1:95Ca^{DG12112}	chr3R:19685186–19685357	Not examined

doi:10.1371/journal.pone.0082362.t005

[57] while Eutherian mammals, such as human and cow, diverged over 90 million years ago, according to estimates on the TimeTree [67].

Numerous conserved sequences identical over at least 100 nucleotides are present in sets containing very divergent species (*e.g.*, virilis or frog sets, about two substitutions per neutral site),

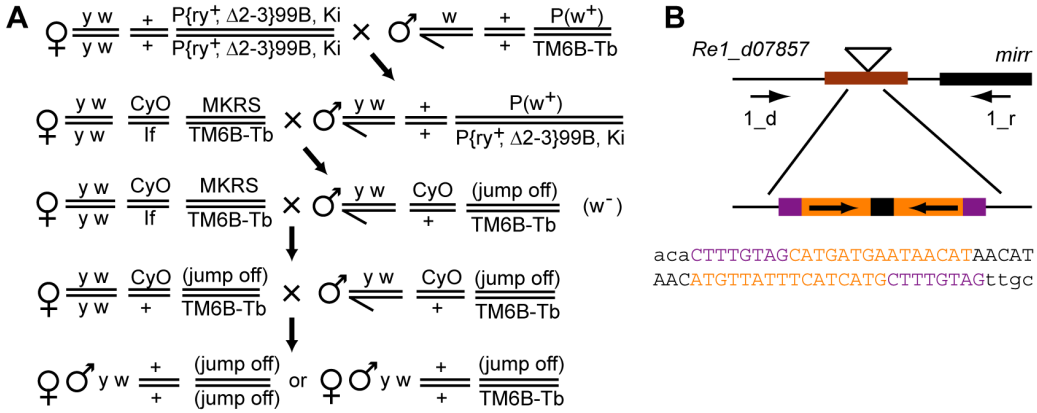

Figure 4. Phenotype rescue experiment. (**A**) Genetic scheme used to remove the transposons from the intergenic UCEs in *d07857* and *KG02042* flies. Males bearing *P-element* insertions were crossed with females carrying the source of transposase on the chromosome 3. In the progeny transposase initiates «jump out» of the *P-element*-based transposons resulting in progeny with white eyes due to absense of mini-*white* marker gene. In the next generation *CyO Tb* males with the white eyes resulted from a complete or partial excision of the transposons were selected and individually crossed with ''four balancers'' females. The resulted F3 *Tb* progeny were intercrossed and F4 flies were examined on presence of homozigotes with the normal third chromosome. (**B**). Molecular map of the remaining insertion in *Re1-d07857* flies (not to scale). Black bar represent 5′ UTR of *mirr* gene, triangle represent the remaining insertion after the excision of the transposon from the UCE chr3L.296 (brown color). Arrows indicate position of primers used to amplify and sequence the DNA after excision of the transposon. The structure and the sequence of the remaining insertion are shown below. The purple blocks represent 8 bp target site duplication after insertion of the transposon, the orange blocks correspond to remnants of the terminal inverted repeats, and the remaining nucleotides apparently appeared as a results of incomplete excision.
doi:10.1371/journal.pone.0082362.g004

however, only a few non-exonic UCEs can be traced beyond drosophilids. This suggests that the constraints acting on the non-exonic UCEs were either relaxed outside of drosophilids, or these sequences only came under strong selection pressure in flies after the acquisition of new functions as suggested for amniotes [8].

As in vertebrates, the majority of the UCEs in drosophilds are in non-exonic regions (Figure 1) but significant differences exist between these two groups. In flies the longest UCEs are exonic, and the proportion of exonic UCEs increases in sets with higher phylogenetic distances between member species, whilst in tetrapods the longest UCEs are non-exonic, and the proportion of non-exonic UCEs increases with higher phylogenetic distances between member species. The difference in UCE distribution is also observed between non-coding regions. In contrast to vertebrates, in flies the density of UCEs is lower in introns relative to intergenic regions. The prevalence of exonic UCEs in distant drosophilids might be partially attributed to a strong codon bias for highly expressed genes in drosophila [59]. The abundance of non-exonic UCEs identified in distant vertebrates may indicate a sophisticated regulatory mechanism existing in this lineage.

Sophophora UCEs are depleted near TSSs in the *D. melanogaster* genome whilst the opposite is observed in the human genome. Additionally, in the human genome UCEs are depleted within segmental duplications and copy number variants (CNV), with the strongest depletion observed for non-exonic UCEs [68]. The Sophophora UCEs are under-represented among CNVs but at low statistical significance. In contrast to the human genome, in *D. melanogaster* both exonic and non-exonic UCEs show similar levels of depletion around CNVs: among 37 Sophophora UCEs overlapping CNVs, 10 are exonic – consistent with the proportion of exonic Sophophora UCEs in the whole genome.

Human genes annotated as involved in splicing and RNA binding are enriched with exonic UCEs [7] whilst there is no evidence for a similar enrichment in *D. melanogaster*. The GO analysis of the *D. melanogaster* genes overlapping Sophophora UCEs demonstrates no enrichment for either splicing factors or RNA binding proteins (Table S3) suggesting that this group of genes has no significant contribution to the pool of UCEs in drosophilids. Whilst the GO category 'Transcription factor activity' is enriched in genes with exonic Sophophora UCEs, the other enriched categories reference membrane and channel complex proteins which mirrors those of earlier observations [44] and the enrichment is remarkably different from the eutherian UCEs [8].

The Sophophora UCEs are non-uniformly distributed along the *D. melanogaster* genome (Figure S1). The UCEs are depleted in regions adjacent to pericentric heterochromatin but enriched in regions of intercalary heterochromatin with low gene density [60,69]. These regions are enriched with the Polycomb-group proteins [70], components of the silenced chromatin. In a similar manner, large arrays of highly conserved noncoding elements coincide with Polycomb binding sites and the conserved syntenic blocks in insects [6,52]. In agreement with this we found an enrichment of the suppressed transgenes integrated in the vicinity of the Sophophora UCEs, as well as a suppression of the marker gene in the *c00059* fly stock carrying the *piggyBac* transposon inserted into an intergenic UCE. While it is tempting to speculate about the involvement of some UCEs in silenced chromatin, the analysis of chromatin data is beyond the scope of this work.

Four out of eleven studied transposon insertions into the intergenic Sophophora UCEs behave as recessive lethals (Table 5) implying important associated functionality in flies. The UCE chr3L.214 associated with lethality in the *KG02042* strain is located within a long intergenic region, more than 13 kb away from the 3' end of the nearest gene, *glutamate receptor IB* (*Glu-RIB*), and more than 30 kb away from the closest promoter of the *PGRP-LA* gene. Removal of a transposon from its integration site restored viability suggesting that the observed lethality is apparently associated with the integration of the transposon. The UCE chr3L.205 associated with lethality in the *f02994* strain is located ~1.5 kb upstream of the *Rdl* gene. The *piggyBac* insert in this strain is annotated as being an allele of *Rdl* [55,66]. The two remaining UCEs associated with lethality are located within 1 kb of the transcription start sites of FlyBase annotated genes. The UCE chr3L.296 associated with lethality in the *d07857* fly strain overlaps the promoter region of the current refSeq annotation of the *mirr* gene (checked on March 2013). The UCE chr2L.116 associated with the suppression of the marker gene is located more than 5 kb away from the nearest gene. Our results show that the disruption of UCEs, irrespective of distance from a nearby protein-coding gene, can produce a visible phenotype. Four out of six studied intergenic Sophophora UCEs without any visible phenotype resultant through transposon integration are located more than 5 kb from any nearby gene. For comparison, the mean distance between protein coding FlyBase 5.12 genes in the euchromatic part of the *D. melanogaster* genome is 4.1 kb while the median distance is less than 700 nucleotides.

The "phenotype rescue" experiments performed on two independent fly strains have confirmed that lethality is linked to the disruption of the UCEs by the insertions. Viability was restored by a partial excision of the transposon in the *Re1-d07857* flies, indicating that a small insertion in the same position does not cause lethality. We did not observe a reduction of viability, or any other visible phenotype, in six out of eleven fly strains with intergenic UCEs disrupted by transposons. It is plausible that intact UCEs are only required for survival under natural environmental conditions and hence no visible phenotypes were observed under controlled laboratory conditions. Nonetheless, the high conservation of these sequences still remains a mystery.

Supporting Information

Figure S1 Distribution of Sophophora UCEs in the *D. melanogaster* genome. The data is shown only for the euchromatic part of the genome. The instances were counted in 100 kb bins. (**A**) Distribution of the UCEs annotated as exonic, intronic or intergenic using protein coding FlyBase gene 5.12 models. Color-coding is shown on each panel. (**B**) Unique transcription start sites of FlyBase genes 5.12. (**C**) *P-element* and (**D**) *piggyBac* insertions with integration sites shorter than 10 nucleotides.
(PDF)

Figure S2 Partial suppression of mini-*white* marker gene of the *piggyBac* transposon inserted into intergenic UCE. (**A**) Modified screenshot of the UCSC Genome Browser showing DNA around *PBac{PB}c00059* transposon integrated into UCE chr2L.116 (14 kb region, chr2L:7,351,501–7,365,500). Orange marks show integration sites of known transgenes from FlyBase. Black boxes correspond to the Sophophora UCEs, and the refSeq genes are shown in blue. The conservation plot is shown at the bottom. (**B**) Eye color in fly carrying a non-supressed transgene with the mini-*white* gene in w^- background. (**C**) Mosaic pigmentation in *c00059* fly. (**D**) Eye of w^- fly lacking any pigmentation.
(TIF)

Table S1 Density of UCEs in introns.
(DOC)

Table S2 Distribution of the Sophophora UCEs and TSSs outside of the intercalary heterochromatin regions.
(DOC)

Table S3 Gene Ontology analysis.
(XLS)

Table S4 Distribution of transposons and Sophophora UCEs outside of the intercalary heterochromatin regions.
(DOC)

Dataset S1 List of the refSeq gene IDs used for phylogenetic trees.
(XLS)

Dataset S2 BED file with coordinates of UCEs identified in drosophilids. UCEs identified in different drosophila sets (Table 1). Each BED file contain positions of sequences at least 100 nt long identical in *D. melanogaster*, *D. erecta*, and the third species: yak set, *D. yakuba*; ana set, *D. ananassae*; pse set, *D. pseudoobscura*; wil set, *D. willistoni*; moj set, *D. mojavensis*; vir set, *D. virilis*; gri set, *D. grimshawi*. The coordinates are for the dm3 (BDGP R5) assembly of the *D. melanogaster* genome.
(BED)

Dataset S3 BED file with coordinates of vertebrates UCEs. UCEs identified in different vertebrates (Table 1). The coordinates are for the hg19 assembly of the human genome. The vertebrate UCEs were identified in the hg18-centric alignment and the coordinates were liftOvered to the hg19.
(BED)

Dataset S4 BED file containing Sophophora UCEs for the dm3 genome assembly.
(BED)

Acknowledgments

We thank Exelixis and the Blumington stock center for providing the fly stocks, Tatyana Kolesnikova for productive discussion, and two anonymous reviewers for useful comments and suggestions.

Author Contributions

Conceived and designed the experiments: IVM. Performed the experiments: VVS. Analyzed the data: IVM SJS. Contributed reagents/materials/analysis tools: MP SNB. Wrote the paper: IVM.

References

1. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034–1050.
2. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520–562.
3. Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, et al. (2003) Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). Science 302: 1033–1035.
4. Sandelin A, Bailey P, Bruce S, Engstrom PG, Klos JM, et al. (2004) Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics 5: 99.
5. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, et al. (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 3: e7.
6. Engstrom PG, Ho Sui SJ, Drivenes O, Becker TS, Lenhard B (2007) Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res 17: 1898–1908.
7. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, et al. (2004) Ultraconserved elements in the human genome. Science 304: 1321–1325.
8. Stephen S, Pheasant M, Makunin IV, Mattick JS (2008) Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. Mol Biol Evol 25: 402–408.
9. Ryu T, Seridi L, Ravasi T (2012) The evolution of ultraconserved elements with different phylogenetic origins. BMC Evol Biol 12: 236.
10. Reneker J, Lyons E, Conant GC, Pires JC, Freeling M, et al. (2012) Long identical multispecies elements in plant and animal genomes. Proc Natl Acad Sci U S A 109: E1183–1191.
11. Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE (2007) Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. Nature 446: 926–929.
12. Ni JZ, Grate L, Donohue JP, Preston C, Nobida N, et al. (2007) Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev 21: 708–718.
13. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, et al. (2006) In vivo enhancer analysis of human conserved non-coding sequences. Nature 444: 499–502.
14. Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, et al. (2008) Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat Genet 40: 158–160.
15. Calin GA, Liu CG, Ferracin M, Hyslop T, Spizzo R, et al. (2007) Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. Cancer Cell 12: 215–229.
16. Licastro D, Gennarino VA, Petrera F, Sanges R, Banfi S, et al. (2010) Promiscuity of enhancer, coding and non-coding transcription functions in ultraconserved elements. BMC Genomics 11: 151.
17. Scaruffi P, Stigliani S, Coco S, Valdora F, De Vecchi C, et al. (2010) Transcribed-ultra conserved region expression profiling from low-input total RNA. BMC Genomics 11: 149.
18. Feng J, Bi C, Clark BS, Mady R, Shah P, et al. (2006) The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. Genes Dev 20: 1470–1484.
19. Dong X, Navratilova P, Fredman D, Drivenes O, Becker TS, et al. (2010) Exonic remnants of whole-genome duplication reveal cis-regulatory function of coding exons. Nucleic Acids Res 38: 1071–1085.
20. Sun H, Skogerbo G, Chen R (2006) Conserved distances between vertebrate highly conserved elements. Hum Mol Genet 15: 2911–2922.
21. Simons C, Pheasant M, Makunin IV, Mattick JS (2006) Transposon-free regions in mammalian genomes. Genome Res 16: 164–172.
22. Simons C, Makunin IV, Pheasant M, Mattick JS (2007) Maintenance of transposon-free regions throughout vertebrate evolution. BMC Genomics 8: 470.
23. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125: 315–326.
24. Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engstrom PG, et al. (2007) Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Res 17: 545–555.
25. Lettice LA, Hill AE, Devenney PS, Hill RE (2008) Point mutations in a distant sonic hedgehog cis-regulator generate a variable regulatory output responsible for preaxial polydactyly. Hum Mol Genet 17: 978–985.
26. Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, et al. (2007) Deletion of ultraconserved elements yields viable mice. PLoS Biol 5: e234.
27. Nobrega MA, Zhu Y, Plajzer-Frick I, Afzal V, Rubin EM (2004) Megabase deletions of gene deserts result in viable mice. Nature 431: 988–993.
28. McLean C, Bejerano G (2008) Dispensability of mammalian DNA. Genome Res 18: 1743–1751.
29. Poulin F, Nobrega MA, Plajzer-Frick I, Holt A, Afzal V, et al. (2005) In vivo characterization of a vertebrate ultraconserved enhancer. Genomics 85: 774–781.
30. Yang R, Frank B, Hemminki K, Bartram CR, Wappenschmidt B, et al. (2008) SNPs in ultraconserved elements and familial breast cancer risk. Carcinogenesis 29: 351–355.
31. Catucci I, Verderio P, Pizzamiglio S, Manoukian S, Peissel B, et al. (2009) SNPs in ultraconserved elements and familial breast cancer risk. Carcinogenesis 30: 544–545; author reply 546.
32. Chiang CW, Liu CT, Lettre G, Lange LA, Jorgensen NW, et al. (2012) Ultraconserved elements in the human genome: association and transmission analyses of highly constrained single-nucleotide polymorphisms. Genetics 192: 253–266.
33. Lewejohann L, Skryabin BV, Sachser N, Prehn C, Heiduschka P, et al. (2004) Role of a neuronal small non-messenger RNA: behavioural alterations in BC1 RNA-deleted mice. Behav Brain Res 154: 273–289.
34. Poitras L, Yu M, Lesage-Pelletier C, Macdonald RB, Gagne JP, et al. (2010) An SNP in an ultraconserved regulatory element affects Dlx5/Dlx6 regulation in the forebrain. Development 137: 3089–3097.
35. Martinez F, Monfort S, Rosello M, Oltra S, Blesa D, et al. (2010) Enrichment of ultraconserved elements among genomic imbalances causing mental delay and congenital anomalies. BMC Med Genomics 3: 54.

36. Richards S, Gibbs RA, Weinstock GM, Brown SJ, Denell R, et al. (2008) The genome of the model beetle and pest Tribolium castaneum. Nature 452: 949–955.
37. Sanges R, Hadzhiev Y, Gueroult-Bellone M, Roure A, Ferg M, et al. (2013) Highly conserved elements discovered in vertebrates are present in non-syntenic loci of tunicates, act as enhancers and can be transcribed during development. Nucleic Acids Res.
38. Wang J, Lee AP, Kodzius R, Brenner S, Venkatesh B (2009) Large number of ultraconserved elements were already present in the jawed vertebrate ancestor. Mol Biol Evol 26: 487–490.
39. Kim SY, Pritchard JK (2007) Adaptive evolution of conserved noncoding elements in mammals. PLoS Genet 3: 1572–1586.
40. Gould SJ, Vrba ES (1982) Exaptation; a missing term in the science of form. Paleobiology 8: 4–15.
41. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, et al. (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 441: 87–90.
42. Santangelo AM, de Souza FS, Franchini LF, Bumaschny VF, Low MJ, et al. (2007) Ancient exaptation of a CORE-SINE retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene. PLoS Genet 3: 1813–1826.
43. Vavouri T, Walter K, Gilks WR, Lehner B, Elgar G (2007) Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. Genome Biol 8: R15.
44. Glazov EA, Pheasant M, McGraw EA, Bejerano G, Mattick JS (2005) Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. Genome Res 15: 800–808.
45. Papatsenko D, Kislyuk A, Levine M, Dubchak I (2006) Conservation patterns in different functional sequence categories of divergent Drosophila species. Genomics 88: 431–442.
46. Glazov EA, Pheasant M, Nahkuri S, Mattick JS (2006) Evidence for control of splicing by alternative RNA secondary structures in Dipteran homothorax pre-mRNA. RNA Biol 3: 36–39.
47. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, et al. (2009) The UCSC Genome Browser Database: update 2009. Nucleic Acids Res 37: D755–761.
48. Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, et al. (2013) The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res 41: D64–69.
49. Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11: R86.
50. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596–1599.
51. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696–704.
52. von Grotthuss M, Ashburner M, Ranz JM (2010) Fragile regions and not functional constraints predominate in shaping gene organization in the genus Drosophila. Genome Res 20: 1084–1096.
53. Berriz GF, Beaver JE, Cenik C, Tasan M, Roth FP (2009) Next generation software for functional trend analysis. Bioinformatics 25: 3043–3044.
54. McQuilton P, St Pierre SE, Thurmond J (2012) FlyBase 101–the basics of navigating FlyBase. Nucleic Acids Res 40: D706–714.
55. Thibault ST, Singer MA, Miyazaki WY, Milash B, Dompe NA, et al. (2004) A complementary transposon tool kit for Drosophila melanogaster using P and piggyBac. Nat Genet 36: 283–287.
56. Bellen HJ, Levis RW, Liao G, He Y, Carlson JW, et al. (2004) The BDGP gene disruption project: single transposon insertions associated with 40% of Drosophila genes. Genetics 167: 761–781.
57. Morales-Hojas R, Vieira J (2012) Phylogenetic patterns of geographical and ecological diversification in the subgenus Drosophila. PLoS One 7: e49552.
58. Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, et al. (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 450: 219–232.
59. Shields DC, Sharp PM, Higgins DG, Wright F (1988) "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 5: 704–716.
60. Belyakin SN, Christophides GK, Alekseyenko AA, Kriventseva EV, Belyaeva ES, et al. (2005) Genomic analysis of Drosophila chromosome underreplication reveals a link between replication control and transcriptional territories. Proc Natl Acad Sci U S A 102: 8269–8274.
61. Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M (2008) Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. Science 320: 1629–1631.
62. Sanchez-Elsner T, Gou DW, Kremmer E, Sauer F (2006) Noncoding RNAs of trithorax response elements recruit Drosophila Ash1 to ultrabithorax. Science 311: 1118–1123.
63. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, et al. (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32: D258–261.
64. Amaral PP, Neyt C, Wilkins SJ, Askarian-Amiri ME, Sunkin SM, et al. (2009) Complex architecture and regulated expression of the Sox2ot locus during vertebrate development. RNA 15: 2013–2027.
65. Babenko VN, Makunin IV, Brusentsova IV, Belyaeva ES, Maksimov DA, et al. (2010) Paucity and preferential suppression of transgenes in late replication domains of the D. melanogaster genome. BMC Genomics 11: 318.
66. Liu X, Krause WC, Davis RL (2007) GABAA receptor RDL inhibits Drosophila olfactory associative learning. Neuron 56: 1090–1102.
67. Kumar S, Hedges SB (2011) TimeTree2: species divergence times on the iPhone. Bioinformatics 27: 2023–2024.
68. Derti A, Roth FP, Church GM, Wu CT (2006) Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. Nat Genet 38: 1216–1220.
69. Belyakin SN, Babenko VN, Maksimov DA, Shloma VV, Kvon EZ, et al. (2010) Gene density profile reveals the marking of late replicated domains in the Drosophila melanogaster genome. Chromosoma 119: 589–600.
70. Zhimulev IF, Belyaeva ES, Makunin IV, Pirrotta V, Semeshin VF, et al. (2003) Intercalary heterochromatin in Drosophila melanogaster polytene chromosomes and the problem of genetic silencing. Genetica 117: 259–270.