Evolution of a genomic regulatory domain: The role of gene co-option and gene duplication in the Enhancer of split complex

Elizabeth J. Duncan and Peter K. Dearden

*Genome Res.*, published online May 10, 2010
Access the most recent version at doi:10.1101/gr.104794.109

**Supplemental Material**
http://genome.cshlp.org/content/suppl/2010/05/06/gr.104794.109.DC1.html

**Published online** May 10, 2010 in advance of the print journal.

**Email alerting service**
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](http://genome.cshlp.org/content/suppl/2010/05/06/gr.104794.109.DC1.html)

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to: [http://genome.cshlp.org/subscriptions](http://genome.cshlp.org/subscriptions)

Copyright © 2010 by Cold Spring Harbor Laboratory Press
Evolution of a genomic regulatory domain: The role of gene co-option and gene duplication in the Enhancer of split complex

Elizabeth J. Duncan and Peter K. Dearden

Laboratory for Evolution and Development, Genetics Otago and the National Research Centre for Growth and Development, Biochemistry Department, University of Otago, Dunedin 9054, New Zealand

The Drosophila Enhancer of split complex [E(spl)-C] is a remarkable complex of genes many of which are effectors or modulators of Notch signaling. The complex contains different classes of genes including four bearded genes and seven basic helix-loop-helix (bHLH) genes. We examined the evolution of this unusual complex by identifying bearded and bHLH genes in the genome sequences of Arthropods. We find that a four-gene E(spl)-C, containing three bHLH genes and one bearded gene, is an ancient component of the genomes of Crustacea and Insects. The complex is well conserved in insects but is highly modified in Drosophila, where two of the ancestral genes of the complex are missing, and the remaining two have been duplicated multiple times. Through examining the expression of E(spl)-C genes in honeybees, aphids, and Drosophila, we determined that the complex ancestrally had a role in Notch signaling. The expression patterns of genes found inserted into the complex in some insects, or that of ancestral E(spl)-C genes that have moved out of the complex, imply that the E(spl)-C is a genomic domain regulated as a whole by Notch signaling. We hypothesize that the E(spl)-C is a Notch-regulated genomic domain conserved in Arthropod genomes for around 420 million years. We discuss the consequence of this conserved domain for the recruitment of novel genes into the Notch signaling cascade.

[Supplemental material is available online at http://www.genome.org.]

The Enhancer of split complex [E(spl)-C] of Drosophila melanogaster is a well characterized genetic locus containing 12 genes on chromosome 3R, most of which are to be effectors or modulators of Notch signaling. The E(spl)-C contains a number of different Notch responsive genes, some of which are related in sequence (Wurmbach et al. 1999). The largest class of genes in the region encodes basic helix-loop-helix (bHLH) transcription factors. There are seven of these bHLH transcription factors in the D. melanogaster E(spl)-C [HLHm2, HLHm4, HLHm5, HLHm6, HLHm7, and E(spl)] (Delidakis and Artavanis-Tsakonas 1992; Knust et al. 1992). These bHLH proteins have a distinctive “orange” domain and a C-terminal WRPW motif. This motif is a protein interaction domain (Fisher et al. 1996; Allifrégis et al. 1997), allowing these proteins to interact with Groucho, a transcriptional corepressor that interacts with histone deacetylases to repress gene expression (Chen et al. 1999). E(spl)-C bHLH proteins act as hetero- or homodimeric transcription factors by binding either to specific enhancer sequences, or to other DNA bound transcription factors (Allifaréjis et al. 1997), recruiting groucho to those regions (Giagtzoglou et al. 2003). The recruitment of groucho leads to changes in chromatin conformation and transcriptional repression (Palaparti et al. 1997). E(spl)-C bHLH proteins act as transcriptional repressors and do not activate gene expression in response to Notch signaling (de Celis et al. 1996). In D. melanogaster the groucho gene lies at the telomeric end of the E(spl)-C (Hartley et al. 1998). In D. melanogaster four bearded class genes also lie in the E(spl)-C (m4, m2, and m6) (Lai et al. 2000b). Bearded proteins have an N-terminal amphipathic α-helix, but overall sequence similarity is low (Lai et al. 2000b) implying rapid evolution of these genes. Three additional bearded class genes are present in the D. melanogaster genome (Tom, Brd, and Ocho) and are present in a separate complex on chromosome 3L (Lai et al. 2000b). The D. melanogaster E(spl)-C contains one other gene, m1. The function of this gene is unknown although it is similar in sequence to Kazal class protease inhibitors (Wurmbach et al. 1999).

All the genes of the E(spl)-C, except m1, are Notch responsive (Wurmbach et al. 1999). During embryonic neurogenesis the E(spl)-C bHLH genes are expressed in the neuroectoderm in response to activated Notch signaling (Jennings et al. 1994) and repress key regulators of neural cell fate including proneural genes and the Notch ligand Delta (Heitzler et al. 1996). Cells expressing the E(spl)-C bHLH proteins suppress neural cell fate, allowing cells to take up a secondary epidermal fate, a process known as lateral inhibition (Tata and Hartley 1995; Nakao and Campos-Ortego 1996).

The bearded class genes of the E(spl)-C and bearded complex (Lai et al. 2000b) act in adult sensory precursor formation as antagonists of Notch signaling (Apidianakis et al. 1999; Lai et al. 2000a). Bearded proteins interact with the E3 ubiquitin ligase neurelalized to promote degradation of Delta (Lai et al. 2000a; Debléndre et al. 2001; Pavlopoulo et al. 2001). Little is known about bearded class proteins from other species. There are low levels of sequence conservation between family members (Lai et al. 2000b); thus the evolution of this gene family is unclear, and no bearded class genes have been identified in vertebrates. Recently a Daphnia pulex bearded protein has been shown to interact with neuronalized implying conservation of function in the absence of sequence similarity (Fontana and Posakony 2009).

In Drosophila the expression of both the E(spl)-C bHLH proteins and bearded class genes are regulated by Suppressor of Hairless [SU(H)] (Eastman et al. 1997; Nellesen et al. 1999; Lai et al. 2000b; Maeder et al. 2007), and by miRNA binding to conserved sites (GY-box, Brd-Box, and K-box) located in the 3′ untranslated
The E(spl)-C bHLH genes are related in sequence to other bHLH genes in the D. melanogaster genome. The gene, Her (hairy-E(spl)-related), is closely related to the E(spl)-C bHLH genes (Moore et al. 2000) but not linked. Other, also unlinked, more distantly related bHLH genes are hairy (h), deadpan (dpm), similar to deadpan (Side), hairy/E(spl)-related with YRPW motif (Hey), and clockwork orange (cwo). Related bHLH genes in vertebrate genomes, named the HES (hairy-enhancer of split) genes, have functions in somitogenesis, neurogenesis, and stem cell maintenance (for review, see Kageyama et al. 2007). HES genes are regulated by Notch signaling and require groucho for their activities.

Examination of the genomes of the mosquito Anopheles gambiae and the honeybee Apis mellifera has led to the suggestion that E(spl)-C in Drosophila evolved from an ancestral “Urcomplex” consisting of a single E(spl)-C bHLH gene and a single bearded class gene, as seen in the Anopheles gambiae genome (Schlatter and Maier 2005). The authors point out, however, that the Apis mellifera genome contains three E(spl)-C bHLH genes closely linked to a bearded gene and ascribe these to Apis-specific duplications (Schlatter and Maier 2005).

The E(spl)-C is unlike most other eukaryote gene complexes as it contains a number of unrelated genes involved in a single developmental process, regulating Notch signaling. How did such a complex evolve? Here we utilize the sequenced genomes of arthropods to discover the organization and origins of the E(spl)-C genes. We use the expression patterns of E(spl)-C genes in human, bee and aphid to determine the ancestral functions of these genes. We argue that the E(spl)-C is an ancient complex of genes that evolved from an ancestral Urcomplex consisting of a single E(spl)-C bHLH and a single bearded class gene. We use the expression patterns of E(spl)-C genes in honeybee and aphid to determine the ancestral functions of these genes. We argue that the E(spl)-C is an ancient complex of genes present in the genomes of the common ancestors of insects and crustaceans, and that it is a remarkable evolutionarily conserved genomic domain that has been regulated by Notch signaling for at least 420 million yr (Myr).

**Results**

**Phylogenetics of E(spl)-C-related genes**

**Phylogenetics of arthropod bHLH-orange genes**

In most Drosophilid species there are 13 bHLH genes encoded in the genome. Bayesian phylogenetic analysis groups these proteins into robust phylogenetic groups, all of which are represented in D. melanogaster. This implies that this gene family has been stable since the divergence of the Drosophila lineage 60 million yr ago (Mya) (Supplemental Fig. 1). To examine more distant phylogenetic relationships, the bHLH proteins related to E(spl)-C bHLHs, h, dpm, Side, and cwo were extracted from sequenced arthropod genomes (Ixodes scapularis [chelicerate]; Daphnia pulex [crustacean]; Acrythosiphon pisum [hemipteran]; Apis mellifera [insect]) and from the genome of the cephalochordate Branchiostoma floridae (Holland et al. 2008) (as an example of a non-arthropod set of HES genes). The full-length predicted proteins were aligned and analyzed by Bayesian phylogenetic techniques (Fig. 1; Ronquist and Huelsenbeck 2003). This tree is rooted with the cephalochordate bHLH proteins, which cluster together to the exclusion of the arthropod bHLHs, indicating that these genes have arisen by duplication of an ancestral hairy/E(spl)-C protein and are independent from the duplications in arthropods; thus, chordates do not have direct orthologs of the E(spl)-C bHLH genes.

The arthropod sequences are grouped into five clades; cwo is the most deeply branching and is found in all arthropod genomes including the chelicerate Ixodes, thought to be the most distant group of arthropods to the insects (Cook et al. 2001; Hwang et al. 2001). dpm and Side are also found in all arthropod genomes examined. However, h, while present in all other genomes, appears to be missing from Ixodes. This loss is accompanied by

**Figure 1.** Phylogram of arthropod and Branchiostoma E(spl)-C-related bHLH proteins. Bayesian phylogeny of bHLH proteins from sequenced Arthropod genomes rooted with Branchiostoma HES genes. Posterior probabilities are shown at nodes. Names of proteins and their respective species are shown in Figure 4 and in Supplemental Table 1. The phylogenetic analysis recovers six well-supported clades, a clade of Branchiostoma proteins, a clade of proteins that cluster with Drosophila CWO, a clade that contains both Hairy and Deadpan proteins, a clade clustering with Drosophila SIDE, and a final, poorly resolved clade containing proteins clustering with Drosophila E(spl)-C bHLH and HER. Drosophila HER is circled in red. Ixodes ISCW016537 and ISCW016540, apparent tandem duplications of a Deadpan-like protein, are circled in blue.
C-type bHLHs fall into three major clades with high posterior phylogenetics (Fig. 2; Ronquist and Huelsenbeck 2003). E(spl)-these proteins and examined their relationships using Bayesian proteins and HER, we extracted the bHLH and orange domains of 1) and E(spl)-C proteins are closely related. It is clear that HER (circled in red in Fig. 2000, 2005). The final clade includes the E(spl)-C bHLH and HER proteins, which group together. This clade has poorly supported internal structure and a number of long branches making it difficult to assign relationships. It is clear that HER (circled in red in Fig. 2) and E(spl)-C proteins are closely related.

To better resolve relationships between the E(spl)-C bHLH proteins and HER, we extracted the bHLH and orange domains of these proteins and examined their relationships using Bayesian phylogenetics (Fig. 2; Ronquist and Huelsenbeck 2003). E(spl)-C-type bHLHs fall into three major clades with high posterior probabilities. The first of these clades we name E(spl)-C bHLH2. This clade contains E(spl)-C-related genes from honeybees, Bombyx, aphids, and Daphnia. A second, much larger clade we name E(spl)-C bHLH1; this clade contains all the Drosophila E(spl)-C bHLHs and a single bHLH protein from each of Bombyx, mosquitoes, Tribolium, honeybees, Pediculus, Acrystosiphon, and Daphnia. The final clade contains proteins from honeybees, Tribolium, Bombyx, Pediculus, and Acystosiphon. This clade also contains HER, a non-E(spl)-C bHLH from Drosophila. The only protein that does not clearly fall into these clades is NCBL_GNO_130833, encoded by a Daphnia E(spl)-C gene that appears on the branch leading to E(spl)-C bHLH2. The placement of this gene close to Her, and its position in the Daphnia genome (see below) makes it most likely an ortholog of Her. While the three major clades of bHLH proteins group together robustly, the relationships between genes within the clades are either unresolved, or have low posterior probability, reflecting a lack of sequence divergence. The exception to this is the genes of the Drosophila E(spl)-C and HER, which have long branch lengths and are unusually derived members of the bHLH family.

**Phylogenetics of non-bHLH proteins of the Drosophila E(spl)-C**

The Drosophila m1 gene encodes a protein with similarity to Kazal protease inhibitors; phylogenetic analysis of all insect Kazal protease inhibitors revealed that m1 is found only in Drosophilid species (Supplemental Fig. 2).

The relationships between bearded class genes are more difficult to assess due to sequence divergence. Members of this class are defined only by a basic amphipathic α-helix at the N terminus (Bailey and Posakony 1995) and a bearded motif (N-motif [NXAN(E/K/R)(H/L/M)]i) (Lai et al. 2000b) and cannot be identified by sequence similarity. We identified potential bearded class genes due to their proximity to E(spl)-C bHLH1 and bHLH2 genes. ORFs and annotated genes surrounding bHLH genes were examined for bearded class characteristics (Supplemental Fig. 2). We identified bearded class genes linked with the E(spl)-C in Daphnia, Acystosiphon, Apis, Tribolium, Bombyx, Aedes, and Anopheles. We aligned predicted proteins from these genes, and the known bearded class proteins from Drosophila and subjected them to Bayesian phylogenetic analysis. Trees derived from this analysis separate the Drosophila bearded class proteins into seven clades, (m1, m4, m2, m6, Tom, Brd, and Ocho) (Fig. 3). This analysis produces a star-shaped phylogeny with most of the nondrosophilid sequences forming a clade, implying that the Drosophila bearded class genes all derive from duplications of the single bearded gene seen in nondrosophilid insects. Two exceptions to this, the bearded class genes
from aphid and Daphnia, branch with m6 and Brd, respectively; but it is likely that these associations are spurious due to long branch attraction.

Identification of Arthropod E(spl) complexes

Drosophilid E(spl) complexes

Comparisons of the E(spl)-C across Drosophila species indicates that it is very stable (data not shown). In D. melanogaster the complex consists of 12 genes. Her—the most closely related gene to the E(spl)-C bHLHs (Fig. 1)—lies on chromosome X, and the bearded cluster, Tom, Brd, and Ocho, is on chromosome 3L (Fig. 4). In other Drosophila species, the E(spl)-C is identical to that of D. melanogaster (data not shown). This conservation does not extend to Her, which is missing from the genomes of D. grimshawi, majewsiensis, and yakuba.

Dipteran E(spl) complexes

We re-examined the E(spl)-C in the Anopheles gambiae genome. In this species, as reported previously (Schlatter and Maier 2005) only a single E(spl)-C-like bHLH gene can be identified [AGAP012342, an E(spl)-C bHLH-1 gene]; this gene lies next to a single bearded class gene AGAP012341. This single bHLH class 1 gene next to a bearded class gene is repeated in the related Culex pipens genome and that of Aedes aegypti, although the direction of transcription of the bearded gene relative to the bHLH is reversed in Culex and Aedes (Fig. 4).

Holometabolous E(spl) clusters

The simple E(spl)-C seen in mosquito genomes is not observed in other holometabolous insects. We examined the genomes of A. mellifera, T. castaneum, Nasonia vitripennis, and Bombyx mori for E(spl)-C genes (Fig. 4). In B. mori, a cluster of four E(spl)-C-related genes lies on linkage group 3 within 240 kb. Two bHLH genes (BGIBMGA98916 [E(spl)-C bHLH2] and BGIBMGA98915 [Her]) lie in a head to tail arrangement separated by three genes [unrelated to Drosophila E(spl)-C genes] from a bearded class gene (BGIBMGA98964); next to this gene is a third bHLH gene (BGIBMGA98915 [E(spl)-C bHLH1]).

An identical arrangement of genes is present in the honeybee on linkage group 14 contained within 202 kb: two upstream bHLH genes (GB17028 [E(spl)-C bHLH2] and GB10585 [Her]) in this case separated by a gene encoding a tubulin tyrosine ligase-like protein with no similarity to any Drosophila E(spl)-C gene), a bearded class gene (GB15253), and a final bHLH gene (GB19475 [E(spl)-C bHLH1]) (Fig. 4).

The current assembly of the Nasonia genome encodes no E(spl)-C bHLH genes nor can they be found in unassembled reads. bHLHs with orange domains are present in the genome, but these are orthologs of ippn, h, and cwo. It is not clear if the E(spl)-C has been lost in the lineage leading to Nasonia, or if the absence is due to gaps in the genome sequence.

In Tribolium, the complex is 70 kb and encodes only two bHLH genes (XP_972493 [Her] and XP_972685 [E(spl)-C bHLH1]) flanking a bearded class gene (glean6579). The bHLH genes are both separated from the bearded class gene by two intervening genes with no homology with Drosophila E(spl)-C genes.

E(spl)–C in hemimetabolous insects

The conclusion that E(spl) complexes are larger and more evolutionarily stable than previously thought led us to examine the newly sequenced genomes of the pea aphid (Acyrthosiphon pisum) and louse (Pediculus humanus) (Fig. 4). In Pediculus, only two linked bHLH genes (PHUM000172 [E(spl)-C bHLH1] and PHUM9020 [Her]) could be found (on contig 484) with no evidence for a linked bearded class gene (Fig. 4).

In the aphid genome, a larger cluster of E(spl)-C genes can be found but these are distributed over two contigs. The first contig, EQ122792, contains two E(spl)-C bHLH genes (ACYPI005974 [E(spl)-C bHLH2] and ACYPI009737 [Her]); the second EQ124787 contains a bearded class gene (ACYPI003071) (Supplemental Fig. 3) and a final bHLH gene (ACYPI003697 [E(spl)-C bHLH1]) (Fig. 4).

The Daphnia cluster

As hemimetabolous insects have large E(spl)-C we examined the genome of D. pulex, a crustacean, for any evidence of an E(spl)-C.
The evolution of the Enhancer of split complex

In *Daphnia* a four-gene E(spl)-C exists on scaffold 170. Two bHLH genes lie upstream (NCBI_GNO_234834 [E(spl)-C bHLH2] and NCBI_GNO_299514 [Her]) of a bearded class gene (NCBI_GNO_230834) and a final bHLH gene (NCBI_GNO_228834 [E(spl)-C bHLH1]) (Fig. 4).

The *Ixodes* E(spl)-C genes

Phylogenetic evidence implies that chelicerates are the most distal group of arthropods from the crustacean/insect clade (Cook et al. 2001; Hwang et al. 2001). Examination of bHLH genes in the *Ixodes scapularis* genome did not uncover any sign of a complex of E(spl)-C genes. Two bHLH genes lie next to each other in the genome, but our phylogeny (Fig. 1) indicates that these are dpn related rather than E(spl)-C bHLHs.

This analysis reveals that the E(spl)-C is an ancient feature of insect and crustacean genomes, deriving from the common ancestor of these subphyla. The analysis also indicates that *Her* is ancestrally contained within the complex and has been lost from the E(spl)-C lineage leading to Diptera.

Noncoding sequences of the E(spl) complex

Our studies have identified a four-gene E(spl)-C as a conserved feature of both crustacean and insect genomes. In *Drosophila* the expression of the E(spl)-C genes is regulated by Notch signaling via Su(H) (Bailey and Posakony 1995), by proneural genes (Heitzler et al. 1996), and by miRNA binding to conserved sites (known as Brd-box, GY-box and K-box) in the 3' UTR of these genes (Lai et al. 2005). To determine whether these regulatory mechanisms are evolutionarily conserved we identified putative regulatory motifs to the 5' and 3' of the *Drosophila*, honeybee, aphid, and *Daphnia* E(spl)-C genes (Supplemental Fig. 4). Su(H) sites are found upstream of the start codon in all Arthropod E(spl)-C genes with the exception of the honeybee GB17028. These Su(H) sites are usually coupled with a proneural A site, raising the possibility that these genes may be activated by Notch signaling similar to *Drosophila* E(spl)-C genes. In both aphid and honeybee the ortholog of *Her* has a paired Su(H) site and an A-box proneural site in the 5' region of the gene, and a GY-box and K-box in the sequence to the 3' of the stop codon. However, in the lineage leading to *Drosophila* these sites have been lost with only a conserved GY box. Motifs corresponding to the GY- and K-boxes were identified in the genome sequence downstream from the coding regions of a number of E(spl)-C genes in honeybee, aphid, and *Daphnia*, indicating that regulation of these transcripts by miRNA may be conserved. No Bd-box sequences were found associated with *Daphnia* genes, indicating that regulation by this miRNA may have evolved in the lineage leading to insects or that both the seed sequence and miRNA have diverged such that it cannot be detected.

Expression of E(spl)-C-related genes in honeybees

To determine whether genes of the E(spl)-C have a conserved role in Notch signaling and lateral inhibition we cloned partial coding sequences of the genes of the honeybee complex, (including GB14765, a bee-specific tubulin tyrosine ligase-like protein) and examined their expression via in situ hybridization. The four classical E(spl)-C genes of the honeybee E(spl)-C (Fig. 4) have similar expression patterns (Fig. 5). All four genes, including *Her*, are expressed initially between stages 5 and 6 of development and are expressed in the neuroectoderm as neurogenesis begins. Expression of these genes is limited to those neuroectodermal cells that do not take up neural cell fate, consistent with the role of E(spl)-C genes in *Drosophila*. All of the honeybee E(spl)-C genes are also expressed in a complex series of stripes across the ectoderm, in tracheal pits, and all except *Her* are expressed in the gnathal limb buds. GB14765, the intervening gene in the honeybee complex, is first expressed later in development than the other genes in the E(spl)-C, but it is expressed in a complex pattern that includes regions of the embryo that express other members of the honeybee E(spl)-C, including the neuroectoderm, and tracheal pits (Fig. 5F–I). Expression of the *Drosophila* ortholog of this gene, CG16833, was also determined with in situ hybridization. CG16833 RNA is expressed ubiquitously in *Drosophila* with a clear maternal

Figure 4. Arthropod E(spl) complexes. Genomic architecture of the arthropod E(spl)-C. E(spl)-C complexes from Arthropod genome sequences. Names of gene and their respective species are shown in Supplemental Table 1. To the left is a phylogenetic tree of the species represented. (Squares) bHLH genes; (hexagons) bearded class genes; (light gray octagon) groucho; (ovals) intervening genes in arthropod E(spl) complexes with no similarity to Drosophila E(spl)-C genes. Two bHLH genes lie next to each other in the genome, but our phylogeny (Fig. 1) indicates that these are dpn related rather than E(spl)-C bHLHs.

In *Daphnia* a four-gene E(spl)-C exists on scaffold 170. Two bHLH genes lie upstream (NCBI_GNO_234834 [E(spl)-C bHLH2] and NCBI_GNO_299514 [Her]) of a bearded class gene (NCBI_GNO_230834) and a final bHLH gene (NCBI_GNO_228834 [E(spl)-C bHLH1]) (Fig. 4).
contribution but is not specifically expressed in the neuroectoderm (Fig. 6J–L).

This analysis shows that genes within the honeybee E(spl)-C (including Her and the tubulin tyrosine ligase) are expressed in a pattern consistent with a role in lateral inhibition in the neuroectoderm.

Expression of Aphid and *Drosophila Her*

Honeybee Her, unlike *Drosophila Her*, lies within the E(spl)-C (Fig. 4) and is regulated like the other bHLH genes in the complex (Fig. 5). To determine if this expression pattern is unusual in insects or a honeybee-specific pattern we re-examined the expression of *Drosophila Her*, previously reported as ubiquitously expressed (Moore et al. 2000) and compared this with the expression of Her in the aphid. *Drosophila Her* RNA is expressed in all cells of the embryo (Fig. 6A–C), from very early stages where RNA appears to be maternally provided, though weak staining can be seen in negative controls at this stage (Fig. 6 D,E). This expression pattern is markedly different from that seen with honeybee Her (Fig. 5J–M).

In contrast, aphid Her (ACYPI009737) RNA is not ubiquitously distributed (Fig. 6F–G). Coincident with the specification of the CNS aphid Her RNA is detected in cells of the ventral midline and in cells surrounding the midline, with diffuse staining throughout the head of the developing aphid (Fig. 6G). This expression pattern is similar to that seen for aphid orthodenticle (otd), which we have used to help describe the embryo (Fig. 6I; Huang et al. 2010), although there are differences; expression of Ap-Her is more diffuse and there is clear evidence for staining in cells either side of the midline. No expression of Her RNA can be detected in the mature CNS.

---

**Figure 5.** Expression of honeybee E(spl)-C genes. Honeybee embryos stained for E(spl)-C RNA (blue) using in situ hybridization. All embryos are oriented with anterior to the left, and, unless otherwise stated, viewed ventrally. Scale bars, 100 μm. Staging as per DuPraw (1967). (ne) Neuroectoderm. (A–E) Expression of GB17028, an E(spl)-C bHLH-2 ortholog. (A) Stage 6 embryo (lateral view). GB17028 RNA is expressed in ectoderm on either side of the gastrulation furrow, in the anterior and in a posterior domain. By late stage 6, after the gastrulation furrow has closed, RNA is present in a complex pattern of cells in the anterior neuroectoderm. This expression spreads from anterior to posterior by stage 7 (B), and by stage 8 (C) RNA is present in a segmentally patterned array of cells throughout the lateral neuroectoderm, and strongly in the developing brain. (D) A stage 9 embryo showing expression in paired domains on either side of the ventral midline, including the gnathal limb buds. (E) A dorsal view of the embryo shown in D. GB17028 RNA is present in the developing brain. (F–I) Expression of GB14765 RNA, a gene that encodes a tubulin tyrosine ligase-like protein. (F) Expression in a stage 7 embryo. GB14765 RNA is present in segmental paired domains at the lateral edge of the germ band, probably the tracheal primordia. In stage 8 embryos (G, H [G damaged at the anterior left hand side]) RNA is present at low levels generally but also in segmental paired domains on either side of the ventral midline. This expression fades in late stage 8 embryos (H). (I) Stage 9 embryo (lateral view) showing expression in segmental stripes, with highest expression at the dorsal edge of the germ band in each segment. (J–M) Expression of GB10585 RNA, a honeybee ortholog of Her. In stage 6 embryos (J,K) GB10585 RNA is expressed in the neuroectoderm in a complex pattern of cells with the expression spreading from anterior to posterior, with the entire the pattern extending all the way to the posterior by late stage 6 (K). By stage 8 (L) this expression is lost, and GB10585 RNA is expressed in segmental paired domains on either side of the ventral midline and in the developing brain. (M) Stage 9 embryo, lateral view, showing GB10585 RNA in paired domains on either side of the midline and in patches of cells in each segment at the dorsal edge of the germ band, the tracheal primordia. (N–R) Expression of GB15253 RNA, a bearded ortholog. (N) Expression in a stage 5 embryo (lateral view), just before the onset of gastrulation. GB15253 RNA is present in a series of broad stripes in the anterior lateral ectoderm and in a single narrow stripe in the posterior. (O) Stage 6 embryo (lateral view) as gastrulation begins, GB15253 RNA is present broadly in the lateral ectoderm, but expression is modulated in a series of narrow stripes. (P) Late stage 6 embryo (ventral view; embryo damaged in the posterior right side). As the gastrulation furrow closes in the posterior, GB15253 RNA expression is present in broad domains at the edge of the gastrulation furrow, the neuroectoderm, as gastrulation proceeds this broad domain fades, and in anterior to posterior sequence, leaving a complex pattern of cells expressing GB15253 RNA in the neuroectoderm at stage 7 (Q). (R) Stage 8 embryo (anterior regions damaged) showing GB15253 RNA in paired domains on either side of the midline, including in the gnathal limb buds; expression can also be seen in a small number of cells in the ventral midline. (S–W) Expression of GB19475 RNA, an E(spl)-C bHLH-1 orthologous gene. (S) Early stage 6 embryo. GB19475 RNA is present in the lateral neuroectoderm in broad domains modulated by possible segmental stripes. As the gastrulation furrow closes, in the anterior in this specimen, this domain becomes narrower and focuses down to just the edges of the gastrulation furrow. (T) In mid stage 6 embryos, broad lateral expression has faded, but RNA is becoming expressed in a complex set of cells on either side of the gastrulation furrow, in the neuroectoderm, and in a posterior cap. (U) In late stage 6 embryos this complex neuroectodermal staining stretches the length of the embryo to the posterior cap. (V) Stage 8 embryo. GB19475 RNA is present in a complex pattern of cells in the neuroectoderm and ventral midline. Segmental domains of expression can also be seen in cells at the dorsal edge of the germ band. (W) Stage 9 embryo (lateral view) GB19475 RNA is expressed in paired groups of cells on either side of the ventral midline, including the gnathal limb buds, and faintly in the tracheal primordia at the dorsal edge of the germ band.
Together the expression of Her in aphids and honeybees implies that the ancestral expression of Her is neuroectodermal, with a probable role in lateral inhibition. The ubiquitous expression seen in *Drosophila* is clearly unusual and may be related to its location outside of the E(spl)-C.

**Discussion**

**Evolution of the arthropod E(spl) complex**

Our analysis of the architecture of the E(spl)-C across insect and non-insect arthropods clearly contradicts the previous assertion that the E(spl)-C of *Drosophila* has evolved from a simple two-gene Urcomplex (Schlatter and Maier 2005). Indeed this four-gene complex has a long evolutionary history predating the divergence of insects from crustaceans ~420 Mya (Gaunt and Miles 2002).

This “ancestral” complex is not present in the chelicerate *Ixodes*, and so likely arose between 550 and 420 Mya (Gaunt and Miles 2002) after the split of myriapod and chelicerate lineages. This makes the E(spl)-C both more ancient, and more complex, than previously described. Parsimony ancestral state reconstruction was used to visualize the key steps in the evolution of the E(spl)-C (Fig. 7). Ancestrally the complex was likely to consist of two E(spl)-C–type bHLH genes [*E(spl) bHLH1* and *E(spl) bHLH2*], a single Her-type bHLH gene, and a bearded class gene (Fig. 7). Although, as bHLH2 has been lost in the Diptera, *Tribolium*, and *Pediculus* the parsimony reconstruction is ambivalent as to whether the ancestral complex included bHLH2. The phylogenetic evidence (Fig. 2), however, precludes the reevolution of bHLH2 from a duplication of bHLH1; thus, repeated losses of this gene have occurred in these lineages. The conservation of gene organization and direction of transcription of genes in the E(spl)-C in Crustacea, hemimetabolous...
Figure 7. Parsimony ancestral state reconstructions for the evolution of arthropod E(spl) complexes. Ancestral character states were inferred using parsimony analysis under an Mk1 model in Mesquite. Character states were mapped onto a cladogram representing the currently accepted phylogenetic relationships between insect species (Krauss et al. 2005; Robertson 2005; Savard et al. 2006). Reconstructions of the ancestral E(spl)-C are shown at the nodes and genes are color coded as for Figure 4. bHLH2 is absent from the Diptera, Tribolium, and Pediculus and the parsimony reconstruction is ambiguous as to whether the ancestral complex included bHLH2. However, we argue that the phylogenetic evidence (Fig. 2) precludes the re-evolution of bHLH2 from a duplication of bHLH1, and it is likely that bHLH2 has been lost independently. However, this uncertainty is represented by the dotted line surrounding bHLH2 in the diagrams.
insects (with the caveat that the Acyrthosiphon contigs are not currently joined), and holometabolous insects is strong evidence that this four-gene structure is the ancestral organization of this complex and was present in the last common ancestor of insects and crustaceans.

**Drosophila E(spl) complexes**

The E(spl) complex of Drosophilids is unusual. Most surprisingly it is missing two of the ancestral genes: Her, which has moved out of the complex and is now expressed in a ubiquitous manner in *D. melanogaster* but is missing from the genomes of *D. grimshawi, D. mojavensis,* and *D. yakuba*; and E(spl)-C bHLH2, which is not present in any Drosophilid genome. Based on our analyses these genes have remained in the E(spl)-C throughout insect evolution before being lost in Diptera. This loss of the Her and E(spl)-C bHLH2 genes is associated with the expansion of the E(spl)-C bHLH1 class from one to eight genes in *Drosophila.* Bearded genes have also duplicated with four partially redundant copies in the complex, and three others in another genomic location (the bearded complex) (Lai et al. 2000b). These duplication events appear to be associated with significant sequence divergence with both bHLH and bearded class genes having long branch lengths in the phylogenetic analysis. It is unclear why the E(spl)-C has undergone such duplication and divergence in *Drosophila.* This expansion has not been driven solely by the loss of Her and E(spl)-C bHLH2 in *Drosophila* as both genes are also missing from mosquito genomes, where there is no evidence for complex expansion. None of the *Drosophila* E(spl)-C genes (except m1) contain introns, while those in other species do; perhaps the loss of introns has facilitated repeated gene duplication. It is possible that the increase in E(spl)-C genes in *Drosophila,* mirrored as it is by an increase in the number of proneural genes (Schlatter and Maier 2005), may provide more specific control over neurogenesis in *Drosophila* and may reflect a need to place neural elements such as bristles in more stereotyped locations in the Drosophilidae than in mosquitoes and other holometabolous insects (Simpson et al. 1999). Consistent with this idea, polymorphisms in the *Drosophila* E(spl)-C are known to be associated with bristle number variation (Macdonald et al. 2005).

The *Drosophila* E(spl)-C complex is tightly linked to two other genes, m1 and groucho. m1 is not Notch responsive (Wurmbach et al. 1999), while Groucho is a key component of transcriptional repression mediated by Notch (Parouzh et al. 1994). These genes do not appear in the complex in any other species and are likely to be serendipitously inserted into, or near, the complex. It is interesting to speculate, however, that given the key role of groucho in Notch signaling and the function of the E(spl) bHLH genes, it is extraordinary that this gene now sits next to the E(spl) complex. Perhaps this close linkage to the E(spl)-C represents the incipient capture of another Notch signaling component by the E(spl)-C.

**GR14765** is a honeybee-specific insertion into the E(spl)-C and is expressed in a pattern consistent with it being regulated, in part, by Notch signaling, while its ortholog in *D. melanogaster* is not. Conversely bHLH genes, like Her in *Drosophila,* that move out of the complex, lose their *E(spl)*-type expression pattern and presumably their ability to be Notch responsive. These data lead us to hypothesize that the E(spl)-C, at least in early embryos, may act as a genomic regulatory domain that can be regulated as a whole by Notch signaling. Support for this hypothesis also comes from the recent finding that the E(spl)-C may be regulated in an unusual way (Schaf et al. 2009). Chromatin immunoprecipitation experiments demonstrate that the E(spl)-C is bound by both cohesin and polycomb group proteins, and knock-down of the VTD (also known as RAD21) subunit of cohesin in these cells causes many of the E(spl)-C genes’ expression to be up-regulated (Schaf et al. 2009). This is consistent with the idea that the complex is a genomic domain, regulated in concert by cohesin and polycomb in some developmental contexts. Consistent with this Her, the E(spl)-C like bHLH that has moved out of the complex in *Drosophila* is not sensitive to cohesin knockdown (Schaf et al. 2009).

**Function of the E(spl) complex**

The function of the E(spl)-C genes in the honeybee embryo, based on their expression pattern, is similar to their role in *Drosophila.* Expression during neurogenesis, in cells of the neuroectoderm that do not delaminate, implies that honeybee E(spl)-C genes are acting to repress neural cell fate. This expression pattern is the same for bHLH genes and bearded genes.

Earlier, during gastrulation (stage 6), expression of RNA from these genes is in a complex series of stripes across the ectoderm, which are not of a periodicity, or at a stage, that suggests a role in segmentation. Yet Notch signaling triggered by Delta does not regulate segmentation, despite both Delta and fringe being expressed in clear segmental stripes (Mj Wilson, BH McKeelvey, S van der Heide, and PK Dearden, unpubl.). The expression patterns of the E(spl)-C genes may reflect a general tendency for genes to be expressed in stripes during segmentation with no functional consequence, as seen for many *Drosophila* genes (Liang and Biggin 1998). Later expression of honeybee E(spl)-C genes indicates a role in brain, mouthpart, and tracheal development.

The conservation of expression patterns of these genes across the 300 Myr divergence between honeybees and *Drosophila* implies that the function of the E(spl)-C is stable over evolutionary time. Phylogenetic and phylogenomic evidence supports the idea that the Hymenoptera are the most basally branching group of the holometabolous insects (Krauss et al. 2005; Savard et al. 2006). This phylogenetic placement implies that the function of the E(spl)-C is conserved in holometabolous insects. Aphid Her is also expressed in a pattern implying a role in lateral inhibition, suggesting that the E(spl)-C in aphids, and possibly all holometabolous insects, also acts in neurogenesis.

**Gene complexes and gene co-option**

The recent sequencing of a number of arthropod genomes has revealed a significant difference between arthropod and vertebrate genome evolution. In vertebrates, gene organization and synteny is often conserved over long evolutionary distances (Barbazuk et al. 2000; Kohn et al. 2006; Kikuta et al. 2007). In arthropods this is not the case. The relationships between genes changes more rapidly in arthropod lineages, perhaps reflecting shorter average generation times. Gene complexes in insects are thus less likely to remain together by serendipity than in vertebrates. Therefore, it is likely that any highly conserved complex in arthropods must be functionally constrained. In arthropods few such complexes exist, the best described being the *Hox* complex, in which gene organization and function are linked (Hughes and Kaufman 2002), the Runt complex, where it is not clear why the genes remain together over long evolutionary periods (Duncan et al. 2008), and now the E(spl)-C.

The E(spl)-C is a remarkable example of a conserved gene complex as it contains, at its most simple, two functionally different
classes of genes. This structure suggests that the complex is maintained for functional reasons.

Data presented here show that genes that are inserted in the complex appear to become controlled by Notch signaling: GB14765 has moved into the honeybee E(spl)-C and is now expressed in an E(spl)-C-like way, unlike its ubiquitous expression in Drosophila. Those that move out seem to lose that Notch responsiveness: Her has moved out of the E(spl)-C in Drosophila and is now expressed ubiquitously, a pattern of expression dissimilar to sponsiveness: E(spl)-C–like bHLH gene regulated by Notch became close to a bearded sequence and captured it, causing it to become regulated by Notch signaling, and eventually to act in it. Such capture of a sequence into a signaling pathway due to a local chromatin domain may be a general mechanism whereby novel genes are recruited into ancient cell signaling pathways.

Methods

Gene identification

E(spl)-C genes were identified using BLASTP or TBLASTN searches (Altschul et al. 1990) on whole insect genome sequence databases (BeeBase, http://www.bioinformatics.ksu.edu/BeeBase; L. scaptula VectorBase, http://www.vectorbase.org; Nasonia Genome Project, http://www.hgsb.bcm.tmc.edu/projects/nasonia/; P. humanus VectorBase, http://www.vectorbase.org; vFeaBase: Daphnia Genome project, http://vfeabase.org/; Colbourne et al. 2005; Wang et al. 2005, 2007; Nene et al. 2007; Drysdale 2008; Tribolium Genome Sequencing Consortium 2008; Lawson et al. 2009; The International Aphid Genomics Consortium 2008; Legeai et al. 2010; The Nasonia Genome Working Group 2010). When E(spl)-C bHLH genes were identified we searched the regions around those genes using BLAST searches against the Drosophila genome to identify other possible components of a complex. For identification of bearded class genes, the secondary structure of the predicted protein sequence was assessed using the multivariate linear regression combination (MLRC) algorithm (Guermeur et al. 1999) at the Network Protein Sequence Analysis website (NPS) (Combet et al. 2003). Phylogenetic relationships were reconstructed using the Jones (bHLH proteins; Jones et al. 1992) or WAG model (Whelan and Goldman 2001), which were found to be the most appropriate after preliminary investigations using mixed models. The first 25% of trees were discarded as burn-in and the remaining trees summarized and visualized using Dendroscope (Huson et al. 2007).

Ancestral state reconstruction of the E(spl)-C was performed with Mesquite (v2.72) (http://mesquiteproject.org). Reconstruction was performed using parsimony methods with the Mk1 model (Markov K-state 1 parameter model), which assumes equal probability for changes between states. Character states were mapped onto the currently accepted phylogeny of arthropods (Krauss et al. 2005; Robertson 2005; Savard et al. 2006) assuming equal branch lengths.

In situ hybridization

Clones of E(spl)-C genes were amplified using PCR (see Supplemental Table 2 for primer sequences) and cloned into PCRII Vector (Promega). Fragments of E(spl)-C genes were amplified using PCR (see Supplemental Table 2 for primer sequences) and cloned into PCRII Vector (Promega). The nature of predicted helical regions was determined using the multivariate linear regression combination (MLRC) algorithm (Guermeur et al. 1999) at the Network Protein Sequence Analysis website (NPS) (Combet et al. 2003). Phylogenetic relationships were reconstructed using the Jones (bHLH proteins; Jones et al. 1992) or WAG model (Whelan and Goldman 2001), which were found to be the most appropriate after preliminary investigations using mixed models. The first 25% of trees were discarded as burn-in and the remaining trees summarized and visualized using Dendroscope (Huson et al. 2007).

Ancestral state reconstruction of the E(spl)-C was performed with Mesquite (v2.72) (http://mesquiteproject.org). Reconstruction was performed using parsimony methods with the Mk1 model (Markov K-state 1 parameter model), which assumes equal probability for changes between states. Character states were mapped onto the currently accepted phylogeny of arthropods (Krauss et al. 2005; Robertson 2005; Savard et al. 2006) assuming equal branch lengths.

Acknowledgments

We thank Megan Wilson and Nathan Kenny for critical comments on this manuscript. P.K.D. thanks David Hartley for first introducing him to the E(spl)-C and for patiently waiting for a paper. We also thank Gabrielle Drayton and David Teulon from Plant and Food Research for providing us with the pea aphid cultures. This work was supported by a grant to P.K.D. from the National Research Centre for Growth and Development (MP04) and by a UORG grant from the University of Otago.

References


The evolution of the Enhancer of split complex


Palapart A, Baratz A, Stifani C, Hediger M, The grouch/transducin-like Enhancer of split transcriptional repressors interact with the genetically defined


Received January 1, 2010; accepted in revised form April 29, 2010.