The Dca Gene Involved in Cold Adaptation in Drosophila melanogaster Arose by Duplication of the Ancestral regucalcin Gene

Carlos E. Arboleda-Bustos¹,² and Carmen Segarra*,¹,²

¹Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain
²Institut de Recerca de la Biodiversitat (IrBio), Universitat de Barcelona, Barcelona, Spain

*Corresponding author: E-mail: csegarra@ub.edu.

Abstract

The Drosophila cold acclimation gene (Dca) is involved in the adaptive response to low temperatures. This gene is upregulated at the transcriptional level when D. melanogaster flies are exposed 1 day to 15 °C. Dca (or smp-30) is a member of the SMP-30/Glucolactonase/LRE-like family. In the current study, we characterized the members of this gene family in the 12 Drosophila species with available complete genomes sequences. Two paralogous genes, Dca and regucalcin, were identified in all the Sophophora subgenus species (9 of the 12 species), and their presence was further confirmed in three other species of the subgenus (D. subobscura, D. madeirensis, and D. guanche). However, only regucalcin was present in the species of the Drosophila subgenus (D. grimshawi, D. virilis, and D. mojavensis). The phylogenetic analysis and the molecular organization of Dca that is a nested intronic gene support that Dca arose by a duplication event from the ancestral regucalcin gene after the split of the Sophophora and Drosophila subgenera but before the Sophophora radiation. After the duplication event, the nonsynonymous fixation rate increased in the branch leading to Dca (but not to regucalcin), suggesting the neofunctionalization of the former duplicate. Thus, regucalcin would have maintained the ancestral gene function, and Dca would have acquired a new function likely related to Ca²⁺ homeostasis and cold acclimation. Molecular evolution of Dca has been affected by its implication in the adaptive response to cold temperatures. Indeed, the gene has evolved under stronger purifying selection in the temperate than in the tropical Sophophora species, as reflected by the ratio of nonsynonymous to synonymous substitutions. This result is consistent with functional constraints acting on the DCA protein to keep species adaptation to temperate climates. Dca and regucalcin also differ in their expression patterns. The expression profile of regucalcin is similar to that of the anterior fat body protein gene (AFP) of Sarcophaga peregrina and Calliphora vicina, which is also a member of the SMP-30/Glucolactonase/LRE-like gene family. Sequence similarity and expression profile suggest that AFP and regucalcin are indeed orthologous genes.

Key words: cold acclimation, gene duplication, Drosophila, molecular evolution, DNA divergence, Dca.

Introduction

Temperature is a major environmental factor that can influence the behavior and fitness of animals and plants. In nature, living organisms are exposed to temperature changes, which generally impose stress and may result in the evolution of adaptive genetic mechanisms to cope with temperature extremes. Drosophila species can be found in a wide range of habitats and have thus been a valuable model for understanding the physiological and genetic basis of thermal tolerance variation in ectotherms (reviewed in Hoffmann et al. 2003; Dillon et al. 2009). Thermotolerance has been extensively studied after experimentally inducing heat stress in the laboratory. The molecular mechanisms involved in thermo tolerance after heat shock stresses are rather well characterized. They are based on the synthesis of several families of heat shock proteins (HSP) that prevent, as chaperon proteins, the formation of misfolded proteins after the exposure to high temperatures or other stresses (Hartl 1996; Feder and Hoffmann 1999; Sorensen et al. 2003). Although recent research provides a better understanding of cold stress resistance, a generalized cellular cold shock response similar to that associated to HSPs has not been yet characterized (Sinclair and Roberts 2005).

Drosophila adaptations to survive at mildly cold temperatures or after cold stress are complex interdependent processes (Ayrinhac et al. 2004; Macdonald et al. 2004; Rako and Hoffmann 2006). Recovery after a cold shock treatment is affected by the rearing temperature (rearing acclimation) and by previous exposures to suboptimal temperatures during long (cold acclimation) or short periods (cold hardening). These plastic responses might affect the distribution area of Drosophila species (Gibert et al. 2001; Gibert and Huey 2001; Kimura 2004; Kellermann et al. 2009). They might also contribute to the latitudinal clines detected for cold tolerance in D. melanogaster species.
(Hoffmann et al. 2002; Ayrinhac et al. 2004) and D. serrata (Hallas et al. 2002). The presence in both species of clinal variation over a similar geographic gradient would favor the adaptive character of the detected clines. The genetic basis of cold adaptation has been analyzed in experiments of selection for cold resistance measured either by mortality after a cold shock or by the recovery time from a chill coma (Chen and Walker 1994; Watson and Hoffmann 1996; Anderson et al. 2005; Mori and Kimura 2008; MacMillan et al. 2009).

The number of well-known genes involved in survival at low temperatures is still scarce in insects (Clark and Worland 2008). Subtractive hybridization allowed the identification of a candidate gene for cold tolerance in Drosophila (Goto 2000). This gene was named Dca (Drosophila cold acclimation gene), and it is upregulated at the transcription level in D. melanogaster flies acclimated to 15 °C for 1 day. Differential gene expression of Dca in D. subobscura flies reared at either 13 °C or 22 °C has also been detected in a microarray experiment (Laayouni et al. 2007). By contrast, microarrays and real time-polymerase chain reaction (RT-PCR) techniques showed that Dca is downregulated after a cold shock at 0 °C for 1–3 h (Qin et al. 2005; Sinclair et al. 2007). Dca is most likely to be involved in cold adaptation but not in the response after a cold shock. In addition, different quantitative trait loci (QTLs) for thermotolerance have been identified in Drosophila (Morgan and Mackay 2006; Norry et al. 2008). Both studies identified a QTL for chill coma recovery that included genes known to be involved in cold tolerance, such as Fst that is upregulated after a cold shock at 0 °C (Goto 2001). This QTL in the Norry et al. (2008) study also includes Dca. More recently, nucleotide variation at Dca has been related to phenotypic variation in cold tolerance and to the clinal pattern for wing size detected in D. melanogaster (Clowers et al. 2010; McKechnie et al. 2010).

In D. melanogaster, Dca (smp-30 in flybase; www.flybase.org) is located in the right arm of chromosome 3 (3R) at cytological position 88D2. The coding region of the gene is 909 nt long, and it is interrupted by a small intron. The gene presents an additional intron in the 5′ untranslated region (5′UTR) and has two annotated transcripts that differ from the length of the 3′ UTR. Dca is a nested intronic gene that together with CG7362 is located in the long intron of CG3563. Both Dca and CG7362 are found on the opposite strand and in an antisense orientation relative to the host gene.

The role of the protein encoded by Dca (DCA) is unknown, but it shows similarity to several proteins of the SMP-30/Gluconolactonase/LRE-like family (SGL, pf08450) present in a variety of eukaryotic and prokaryotic species. According to the Pfam database (Finn et al. 2008), this family includes different proteins, such as the mammalian senescence marker protein 30 (SMP-30) involved in Ca^{2+} homeostasis and present in amounts that decrease with aging (Fujita 1999; Yamaguchi 2005), the prokaryotic gluconolactonase (GLN) required to interconvert lactone intermediates (Chen et al. 2008), and the firefly luciferin-regenerating enzyme (LRE) that regenerates D-luciferin after its modification during the bioluminescence process (Day and Bailey 2003). In the flesh fly (Sarcophaga peregrina), a protein preferentially present in the anterior part of the larval fat body (anterior fat body protein [AFP]) was also identified as a member of the SMP-30 family. This family shares with other families the β-propeller structure consisting of six 4-stranded β-sheet motifs (Murzin 1992).

Here, we characterize the members of the SMP-30/ Gluconolactonase/LRE-like family in D. melanogaster and extend the study to the 11 Drosophila species with available complete genome sequences as well as to three additional species of the obscura group. These species differ in their climatic origin, and some of them are typically temperate species, such as D. pseudoobscura and D. persimilis, and others are typically tropical species, such as D. sechellia and D. grimshawi. The different climatic distributions of the species are important to analyze the interspecific divergence in a gene involved in cold acclimation as Dca.

The results obtained indicate that 1) the SMP-30/ Gluconolactonase/LRE-like family in D. melanogaster includes two paralogous genes: Dca and regucalcin, 2) Dca originated by a duplication event after the divergence of the Sophophora and the Drosophila subgenera, 3) Dca has evolved differentially in the lineages of the tropical and temperate Drosophila species, and 4) after the duplication event, there was an acceleration of the nonsynonymous substitution rate at Dca.

Materials and Methods

Fly Samples

Drosophila mauritiana, D. yakuba, and D. erecta lines were kindly provided by F. Lemeunier. Highly inbred lines of these species were established after ten generations of sib mating. Inbred lines of D. simulans, D. madeirensis, and D. guanche were already available in the laboratory, as were frozen single males of D. subobscura (Munte et al. 2000) and the DNA from an isochromosomal line for the O chromosome of this species (Rozas et al. 1995).

DNA Extraction, PCR Amplification, and DNA Sequencing

DNA from the highly inbred lines or from single males was purified using a modification of protocol 48 in Ashburner (1989). The available sequences of the Dca and regucalcin genes from D. melanogaster and D. pseudoobscura (www.flybase.org) were used to design the PCR primers with the OLIGO v4.1 program (Rychlik 2007). In order to PCR amplify the complete coding region of each gene, specific primer sets for the melanogaster and obscura species groups were designed in noncoding flanking regions relatively conserved between D. melanogaster and D. pseudoobscura. The Dca and regucalcin genes were PCR amplified from the purified DNA of the highly inbred lines. In D. subobscura, DNA purified from either single males or the O chromosome isochromosomal line were used to PCR amplify the X-linked regucalcin gene and the Dca gene, respectively. The PCR amplified fragments were purified with Qiaquick columns.
(QIAGEN, Chatsworth, CA). Both strands of the purified fragments were completely sequenced by primer walking using internal primers. Sequencing reactions were performed using the ABI Prism BigDye Terminators 3.0 Cycle sequencing kit (Applied Biosystems) and run on an ABI PRISM 3,700 automated DNA sequencer. The sequences of the amplification and sequencing primers and the PCR conditions are available upon request. Partial sequences were assembled using the SeqMan program of the DNASTAR Lasergene package (Burland 2000). The newly reported sequences are deposited in the European Molecular Biology Laboratory data library under accession numbers FR774034–FR774043.

**Sequence Analysis**

Blast searches against the available genome sequence from *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. willistoni*, *D. pseudoobscura*, *D. persimilis*, *D. virilis*, *D. mojavensis*, and *D. grimshawi* ([www.flybase.org](http://www.flybase.org)) were performed to identify the *Dca* and *regucalcin* genes in these species. Sequences were multiply aligned using the PROBCONS program (Do et al. 2005) and subsequently edited with the MacClade program (Maddison and Maddison 1992).

The maximum likelihood method (Nielsen and Yang 1998; Yang et al. 2000) implemented with the CODEML program of PAML version 4 (Yang 2007) was used to estimate interspecific nucleotide sequence divergence at *Dca* and *regucalcin*. Phylogenetic analysis was performed with the maximum likelihood approach implemented in the PHYML package (Guindon and Gascuel 2003). Sequences were assumed to evolve according to the Hasegawa–Kishino–Yano (HKY) nucleotide substitution model (Hasegawa et al. 1985), and the bootstrap analysis was based on 1,000 replicates. The DnaSP v5 program (Librado and Rozas 2009) was used to estimate the degree of departure from a random use of synonymous codons (codon bias) according to the scaled $\chi^2$ (Shields et al. 1988), the codon bias index (CBI; Morton 1993) and the effective number of codons (ENC; Wright 1990). High values of the scaled $\chi^2$ and CBI indicate a strong deviation from a random use of synonymous codons, whereas the highest ENC value (ENC = 61) means that all synonymous codons are used equally, and the lowest value (ENC = 20) that only one codon is used in each synonymous class. The relative rate test proposed by Tajima (1993) was used to compare the nucleotide substitution rate in different lineages. This test contrasts whether the number of nucleotide substitutions in each of two lineages (inferred from an outgroup species) differs significantly by a $\chi^2$ test with 1 degrees of freedom. The analysis was applied independently to synonymous and nonsynonymous substitutions.

Different evolutionary models with alternative assumptions on the ratio of the nonsynonymous to synonymous divergence ($\omega = d_{ns}/d_s$) were compared with the PAML program (Yang 2007) to analyze the selective pressures acting on the *Dca* and *regucalcin* genes. If synonymous and nonsynonymous mutations are assumed to be neutral, they will be fixed at the same rate with $\omega = 1$. In contrast, $\omega < 1$ if nonsynonymous mutations are deleterious and purifying selection reduces its fixation rate. Only the fixation of advantageous nonsynonymous mutations can result in $d_{ns} > d_s$ and, consequently, in $\omega > 1$. *D. willistoni* was excluded from this analysis, as it presents a codon bias that is different from that of the other *Drosophila* species (Clark et al. 2007), which might affect synonymous divergence, and thus, $\omega$ estimates in this lineage. In addition, the AFP gene sequence from *Calliphora vicina* (blue blowfly) and *Sarcophaga peregrina* (flesh fly) with a high similarity to *regucalcin* were included in this analysis. Three different branch tests were used to analyze whether the distribution of selective constraints acting on the *Dca* and *regucalcin* genes varies across lineages. First, the null M0 model with one $\omega$ value for all branches was compared with the free ratio (FR) model that estimates the $\omega$ parameter independently for each branch. Second, the null M0 model was contrasted with the branch Br-Trop/Temp model to analyze whether *Dca* evolves differentially in temperate than in tropical lineages. Third, the null M0 model was compared with the branch Br-dup-Dca and Br-dup-RC models to detect putative changes in the functional constraints after the duplication that originated the *Dca* gene. On the other hand, the modified branch-site model A (Zhang et al. 2005) that allows variation in the $\omega$ value across sites in those lineages specified as foreground branches was applied to detect the presence of positively selected sites. This model includes a site class under positive selection with $\omega_2 \geq 1$ in foreground branches, and it is compared with the null model with a fixed $\omega_2 = 1$ in these branches (branch-site test of positive selection or test 2). This test was applied to *Dca* choosing as foreground branches those leading to the tropical (Test 2 Trop) or, alternatively, to the temperate (Test 2 Temp) species. For the *Dca* and *regucalcin* joint data set, the branch leading to *Dca* after the duplication event was specified as the foreground branch (Test 2 dup-Dca). The Bayes Empirical Bayes (BEB) method was used to identify sites under positive selection in foreground branches. In all cases, the likelihood ratio test (LRT) was used to contrast the fit to the data of two nested models, assuming that twice the log likelihood difference between two models follows a $\chi^2$ distribution with a number of degrees of freedom equal to the difference in the number of free parameters (Whelan and Goldman 1999). To prevent incorrect parameter estimates caused by local optima, the CODEML program was run multiple times for the same model, specifying different initial values. The PAML package was also used to infer the ancestral sequences at particular nodes. The EvoRadical program (Wong et al. 2006) was applied to contrast whether *Dca* and *regucalcin* evolved under positive selection acting on codon substitutions that change the physicochemical properties of amino acids. Model A1 includes site classes with $\gamma > 1$, where $\gamma$ is equivalent to $\omega$ ($\omega = d_{ns}/d_s$) but considering only those nonsynonymous mutations that alter amino acid properties. Indeed, only conservative nonsynonymous mutations are considered when defining $\omega$ in this model. The null model A2
only includes site classes with $γ \leq 1$. The A1 and A2 models are contrasted by a LRT for amino acid partitions according to particular properties, such as hydrophobicity, volume, or charge. In addition, the TestBranchDNDS.bf method implemented in HyPhy program (Pond et al. 2005) was applied to contrast whether the branch leading to either Dca or regucalcin after the duplication event evolved under different selection pressures than the rest of the branches regarding the amino acids physicochemical properties. The analysis was performed choosing the Dca (or alternatively the regucalcin) branch as foreground branch and using the MG94xHKY85 nucleotide substitution model, the general discrete distributions of synonymous and nonsynonymous substitutions with three categories each, and the polarity/charge/hydrophobicity amino acid class model.

The secondary and tertiary structures of the DCA and REGUCALCIN proteins were inferred by PredictProtein (Rost et al. 2004) and I-TASSER (Zhang 2008), respectively. A search with the predicted 3D structure was performed using the DALI server (Holm and Rosenström 2010) to identify other proteins with similar structures to DCA and REGUCALCIN.

**Results**

**Dca and regucalcin in Drosophila Species**

Blast searches indicated that the SMP-30/Gluconolactonase/LRE-like family of *D. melanogaster* includes only two genes: *Dca* (3R chromosomal arm) and *regucalcin* (X linked). The coding region of the *Dca* gene was subsequently sequenced in four species of the *melanogaster* subgroup (*D. simulans*, *D. mauritiana*, *D. yakuba*, and *D. erecta*) and in three species included in the *obscura* group (*D. subobscura*, *D. madeirensis*, and *D. guanche*). The *regucalcin* gene coding region was also sequenced in these latter species. A homology Blast search using the *Dca* and *regucalcin* genes of *melanogaster* against *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. willistoni*, *D. pseudoobscura*, and *D. persimilis* genomes allowed us to identify both genes in these species. However, a single region with a higher level of similarity to *regucalcin* than to *Dca* was found in the *D. virilis*, *D. mojavensis*, and *D. grimshawi* genes, suggesting the presence of only one of these genes in the three species included in the *Drosophila* subgenus. This single region was located in scaffolds of the X chromosome in the three species, which confirmed that it corresponded to the *regucalcin* orthologous gene. The absence of *Dca* in *D. virilis*, *D. mojavensis*, and *D. grimshawi* was further corroborated by the analysis of the 3R syntenic regions in these species (i.e., the long intron of CG3563). No sequence similarity to *Dca* was detected in any of these regions. On the other hand, several linked regions with high nucleotide similarity to the *Dca* gene were identified in *D. ananassae*, *D. pseudoobscura*, and *D. willistoni*. In *D. ananassae*, three copies of the *Dca* gene clustered in approximately 3.5 kb. Two of these copies corresponded to functional genes (*ana1* and *ana3*), and the other copy (*ana2*) had a premature stop codon and was thus considered a pseudogene. The two functional *Dca* copies detected in *D. pseudoobscura* (pse1 and pse2) and *D. willistoni* (wil1 and wil2) were separated by approximately 10 and 3 kb, respectively. The *Dca* and *regucalcin* genes have a coding region of 909 nt (303 amino acids) in all species analyzed, except in the two *Dca* functional copies of *D. ananassae* that present one codon insertion (fig. 1). Furthermore, the structure of *Dca* and *regucalcin* is maintained across all species and consists of two exons separated by a short intron.

The obtained *Dca* coding regions of *D. simulans*, *D. yakuba*, and *D. erecta* were aligned and compared with those available in the complete genome of these species. The two aligned sequences differ at 12 nt sites both in *D. simulans* and *D. yakuba*, giving three nonsynonymous and one nonsynonymous differences in each species, respectively. Nucleotide polymorphism within species could explain the detected differences. In addition, a single base pair deletion at the end of the coding region in the available *D. simulans* sequence generates a protein that lacks three last residues. In this case, a sequencing error in the available genome sequence cannot be discarded. No nucleotide differences were detected in the sequences of *D. erecta*, but they present two 3 bp indels that do not modify the length of the encoded protein. On the other hand, *Dca* and the 5′ adjacent gene CG7362 are annotated as a single quimeric gene in *D. yakuba* (www.flybase.org). However, at *Dca*, the length of the coding region and the boundaries of the intron are well conserved between *D. yakuba* and *D. melanogaster*. In addition, the last exon of CG7362 can be identified in the intron that fuses both genes, although the CG7362 coding region would generate a shorter protein (20 residues less). Therefore, CG7362 is likely missannotated in *D. yakuba*, and consequently, *Dca* has been considered in this study as a single functional gene in this species.

**Sequence Divergence at Dca**

Synonymous and nonsynonymous divergence ($d_{s}$ and $d_{N}$, respectively) between the 13 *Drosophila* species of the *Sophophora* subgenus was estimated using the PAML software (supplementary table S1, Supplementary Material online). Higher $d_{s}$ than $d_{N}$ estimates were detected in all comparisons indicating that *Dca* is under purifying selection ($ω = 0.110$). Synonymous divergence differs considerably in comparisons between species of the *melanogaster* and *obscura* groups. Thus, $d_{s}$ values between the *melanogaster* subgroup and the *obscura* group species range from 0.753 to 0.986. Despite the similar divergence time, these estimates are higher than those detected between the functional copies of *D. ananassae* and the *obscura* group species (ranging from 0.559 to 0.700 for *ana1*, and from 0.595 to 0.672 for *ana3*). This result could be explained by differences in the departure from a random use of synonymous codons between the *melanogaster* group species (supplementary table S2, Supplementary Material online), as synonymous divergence is negatively correlated with codon bias. Thus, the *obscura* group species present a rather high level of codon bias at *Dca* regardless of the estimate (e.g., CBI ranges from 0.591 in *D. madeirensis* to 0.662 in *D. pseudoobscura*). These estimates are rather similar to those observed in the
D. ananassae functional genes (i.e., CBI = 0.691 in ana3 and CBI = 0.720 in ana1) but much higher than those detected in the species of the melanogaster group (i.e., CBI ranges from 0.425 in D. melanogaster to 0.489 in D. erecta). D. willistoni is the species with the lowest codon bias at Dca, as expected given the genome wide change in codon usage reported in this species (Clark et al. 2007). Maintenance of a higher codon bias in the obscura group species and D. ananassae relative to the other melanogaster group species would thus explain the detected differences in synonymous divergence at Dca. However, the Tajima’s relative rate test (Tajima 1993) did not detect significant differences in the number of synonymous mutations fixed in the D. ananassae lineage, and the lineages leading to the other melanogaster group species (supplementary table S3, Supplementary Material online). This excess is even stronger for the ana2 nonfunctional copy. The maximum likelihood tree based on nonsynonymous divergence also reflects this result, as the branches leading to the Dca paralogs of D. ananassae are rather long, specially the branch leading to the ana2 copy (result not shown).

Sequence Divergence Between Dca Paralogs

The Dca gene has suffered duplication events in three different Drosophila lineages. In the D. ananassae lineage, two successive duplications render the three current Dca paralogous (ana1, ana2, and ana3), and a posterior pseudogenization process inactivated ana2. The ana2 pseudogene has a premature stop codon that generates a protein with only 89 residues, and an insertion and a deletion (both 15 nt long) that do not affect the frameshift. The maximum likelihood phylogenetic tree suggests that ana1 is the result of the first duplication event and that the second duplication gave rise ana2 and ana3 (result not shown). The average of the nonsynonymous to synonymous divergence
Table 1. LRT and Bayesian Prediction of Amino Acid Sites Under Positive Selection.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Model</th>
<th>LRT $P$ value</th>
<th>Positively Selected Sites BEB$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dca</td>
<td>FR versus M0</td>
<td>$8.20 \times 10^{-9}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Br-Trop/Temp versus M0</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Test 2 Temp</td>
<td>0.008</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Test 2 Trop</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Dca-regucalcin</td>
<td>FR versus M0</td>
<td>$2.57 \times 10^{-12}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Br-dup-Dca versus M0</td>
<td>$1.07 \times 10^{-5}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Br-dup-RC versus M0</td>
<td>0.921</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Test 2 dup-Dca</td>
<td>0.026</td>
<td>655, 141L, 143T, 146K, 191L, 283N</td>
</tr>
</tbody>
</table>

$^a$Positively selected sites were inferred using the cutoff posterior probability $P = 0.95$.

Climatic Adaptation Analysis at Dca

The two paralogous Dca genes of D. pseudoobscura (ps1 and ps2) differ only at eight synonymous and two non-synonymous sites. The low divergence between copies indicates that the duplication event in this lineage is rather recent, and it can be dated after the split of D. persimilis that occurred some 0.55 Mya (Wang et al. 1997), as this species presents a single Dca gene. In contrast, the two D. willistoni Dca paralogs (wil1 and wil2) present 105 synonymous and 79 non-synonymous differences, which indicates that the duplication in this lineage is quite ancient. In addition, it can be inferred that purifying selection acting on wil2 is relaxed because divergence estimates (both synonymous and non-synonymous) between D. willistoni and the other species is always higher in comparison with wil1 than with wil1 (supplementary table S1, Supplementary Material online).

Sequence Divergence at regucalcin

The regucalcin gene is present in all species analyzed. However, the available regucalcin gene sequence of D. persimilis presents a gap of about 120 nt in the coding region, and it was not included in the analysis. Divergence estimates of the regucalcin coding region between all species are presented in supplementary table S4, Supplementary Material online. Synonymous divergence ($d_s$ estimates) is always higher than non-synonymous divergence ($d_N$ estimates), species. Gibert et al. (2001) showed that Drosophila species can be unambiguously classified as tropical or temperate depending on the recovery time after a chill coma. In that study, flies grown at 21 °C were submitted to cold treatment during several hours at 0 °C, and then let to recover at ambient temperature. Temperate species recovered rapidly (average 1.8 min, range 0.15–4.9 min), whereas tropical species needed a much longer recovery time (average 56 min, range 24–120 min). According to their results, species of the melanogaster subgroup were considered as tropical species and those of the obscura group as temperate species. Although D. mauritiana, D. sechellia, and D. erecta were not included in the previous study, they were treated as tropical species given their origin and distribution area. The same criterion was used to consider D. madeirensis and D. guanche as temperate species. This classification allowed the design of a branch model (Br-Trop/Temp) with three $\omega$ values: one for tropical lineages ($\omega_{Trop}$), one for temperate lineages ($\omega_{Temp}$), and one for the outgroup lineages ($\omega_{Rest}$). The Br-Trop/Temp branch model fits better the data than the M0 model ($P = 0.005$) (table 1). The $\omega$ estimate in the tropical lineages ($\omega_{Trop} = 0.128$) is higher than in the temperate lineages ($\omega_{Temp} = 0.097$), which would suggest that purifying selection is relaxed, at least to some extent, in the tropical but not in temperate species. In order to detect the putative action of positive selection, the modified branch-site A model was applied using tropical (or alternatively temperate) lineages as foreground branches. Only when temperate lineages were considered as foreground branches (Test 2 Temp), the modified branch-site model A with a site class under positive selection was found to be significantly better than the null model without this class ($P = 0.008$) (table 1). However, the BEB analysis did not detect any codon with posterior probabilities high enough to be considered under positive selection in the temperate lineages.
The Paralogous Genes Dca and regucalcin

doi:10.1093/molbev/msr040

The maximum likelihood tree inferred from the multiple alignment of the Dca and regucalcin genes in the Drosophila species, except D. willistoni, is shown in figure 2. The alignment also includes the sequences of the AFP gene from C. vicina and S. peregrina that were used as outgroup. The reconstructed tree is consistent with the origin of Dca by a duplication of the ancestral regucalcin gene after the split of the Sophophora and Drosophila subgenera but before the Sophophora radiation. This duplication event had to be associated with a translocation from the X chromosome to Muller’s element E.

The phylogenetic tree in figure 2 was used in the analysis with the PAML software, but the branching pattern of the species in the Dca and regucalcin clusters was modified to reflect the accepted phylogenetic relationships among the species studied. The FR model fits the data better than the M0 model (P = 2.57 × 10⁻¹⁵) (table 1). In order to detect the putative action of positive selection after the duplication event, a branch model (Br-dup-Dca or Br-dup-RC) with two ω values was contrasted with the M0 model with an equal ω for all branches (table 1). In the Br-dup-Dca model, the branch just after the duplication event that leads to Dca (dDca in fig. 2) was assumed to evolve with a ω different from the rest of the branches. Alternatively, a different ω for the branch leading to regucalcin (dRC in fig. 2) was assumed in the Br-dup-RC model. The null hypothesis (M0 model) was rejected by the two ω values model only for the branch leading to

indicating the action of purifying selection against replacement mutations (ω = 0.071). The maximum likelihood tree based on synonymous substitutions is consistent with the accepted phylogenetic relationships of the studied species, except for D. willistoni (Sophophora subgenus) that groups with D. grimshawi (Drosophila subgenus) (results not shown). The genome wide change in codon bias in D. grimshawi (Clark et al. 2007) could explain this result. Indeed, codon bias at regucalcin is much lower in D. willistoni (CBI = 0.312) than in the other species (average CBI = 0.563).

Nonsynonymous divergence estimates between the melanogaster species group, and the three Palaearctic species included in the obscura group (from 0.092 to 0.108) were much higher than those obtained between the melanogaster species group and D. pseudoobscura (from 0.053 to 0.077). Tajima’s relative rate test (Tajima 1993) indicated a highly significant excess of nonsynonymous (but not of synonymous) substitutions in the D. subobscura, D. madeirensis, and D. guanche lineages relative to the D. pseudoobscura lineage, independently of the melanogaster group species used as the outgroup (supplementary table SS, Supplementary Material online).

Divergence of the DCA and REGUCALCIN Proteins

The maximum likelihood phylogenetic tree inferred from the multiple alignment of the Dca and regucalcin genes in the Drosophila melanogaster species group, and the three Palaearctic species included in the subgenus) (results not shown). The

Fig. 2. Maximum likelihood tree of the Dca and regucalcin genes reconstructed using the PHYML program (Guindon and Gascuel 2003) using Calliphora vicina and Sarcophaga peregrina as the outgroups. Dmel, Drosophila melanogaster; Dsim, D. simulans; Dmau, D. mauritiana; Dsec, D. sechellia; Dyak, D. yakuba; Dere, D. erecta; Dana, D. ananassae; Dpse, D. pseudoobscura; Dper, D. persimilis; Dsub, D. subobscura; Dmad, D. madeirensis; Dgu, D. guanche; Dvir, D. virilis; Dmoj, D. mojavensis; Dgrim, and D. grimshawi. Branches classified as tropical and temperate in the Dca gene are depicted in thick gray and thin black, respectively. The thick black branches Dca and dRC are those used as foreground branches in the PAML analysis. a, b, and c indicate those nodes from which the ancestral sequences were reconstructed using PAML. Numbers on each node indicate bootstrap percentages after 1,000 replicates. Only bootstrap values higher than 80 are shown. The scale bar corresponds to the in the PAML analysis. a, b, and c indicate those nodes from which the ancestral sequences were reconstructed using PAML. Numbers on each node indicate bootstrap percentages after 1,000 replicates. Only bootstrap values higher than 80 are shown. The scale bar corresponds to the
Dca ($P = 1.07 \times 10^{-5}$, $\omega_{Dca} = 0.3836$ and $\omega_{Rest} = 0.0838$). In addition, the modified branch-site model A (Test 2 dup-Dca) fitted the data better than the null model ($P = 0.026$), suggesting the presence of sites with $\omega > 1$ in this branch (table 1). Finally, the BEB method identified six codons with high posterior probabilities of having evolved under positive selection after the duplication event that gave rise to Dca (sites 65S, 141I, 143T, 146K, 191L, and 283N in fig. 1). The PAML program was also used to reconstruct the ancestral protein sequences in nodes a, b, and c (fig. 2). According to these sequences, 59 amino acid replacements were fixed in the branch leading to Dca after the duplication but only seven in the branch leading to regucalcin. This result further supports an accumulation of adaptive changes in the Dca copy after the duplication event.

The Dca and regucalcin data set was also analyzed to detect selective pressures acting on particular amino acid physicochemical properties using the codon-based likelihood models implemented in Evoradical program (Wong et al. 2006). Model A1 fits the data better than model A2 for the examined amino acid partitions (hydrophobicity, $P = 1.726 \times 10^{-13}$; volume, $P = 1.414 \times 10^{-8}$; and polarity, $P = 0.0002$). A fair proportion of sites ($p$) were identified in the $\omega < 1$ and $\gamma > 1$ class for the three partitions (hydrophobicity, $p = 0.150$; volume, $p = 0.131$; and polarity, $p = 0.481$) as well as in the $\omega > 1$ and $\gamma > 1$ class for the polarity partition ($p = 0.124$). These results suggest the action of positive selection on these particular physicochemical properties during the evolution of DCA and REGUCALCIN. Moreover, the TestBranchDNDS.bf method indicated that the dDca branch, but not the dRC branch (fig. 2), had a significantly different evolution than the rest of the branches after the duplication event when the amino acid physicochemical properties are considered ($P = 0.005$ and $P = 0.940$, respectively).

The secondary structure of the ancestral DCA and REGUCALCIN proteins do not present strong differences and are consistent with a $\beta$-propeller fold. Indeed, the best template selected by I-TASSER for both proteins was the human SMP30 with a six-bladed $\beta$-propeller structure (Chakraborti and Bahnsin 2010). Four of the six amino acid residues detected to be under positive selection at DCA by the BEB approach were located between $\beta$-structures. In addition to SMP-30, the DALI server identified other $\beta$-propeller proteins with structural similarity to DCA and REGUCALCIN, such as the GLN from Xanthomonas campestris (Chen et al. 2008), the Staphylococcus aureus drug resistance protein (Drp35) (Tanaka et al. 2007), the diisopropylfluorophosphatase (DFPase) from Loligovulgaris (Scharff et al. 2001), and the human serum paraoxonase 1 (PON1) (Harel et al. 2004).

**Discussion**

The Dca and regucalcin genes of D. melanogaster are two paralogous genes of the SMP30/Gluconolactonasa/LRE-like gene family. They are present in all the species of the Sophophora subgenus analyzed. However, only the regucalcin gene is present in the species included in the Drosophila subgenus (D. grimshawi, D. virilis, and D. mojavensis). The phylogenetic analysis (fig. 2) suggests that Dca arose by a duplication event after the divergence of the Drosophila–Sophophora subgenera, but before the radiation of the Sophophora species. The genomic organization of Dca would also support that this gene originated by duplication from the ancestral regucalcin gene. Dca is a nested gene, and its coding region lies within the large intron of CG3563. Genome wide analysis of nested intronic genes has revealed that most of them have originated by the insertion of a duplicated DNA sequence into the intron of a preexisting gene (Assis et al. 2008; Kumar 2009). In addition, the duplication event that gave rise to Dca had to be associated with a translocation from the X chromosome to Muller’s element E, as Dca is located on chromosome arm 3R and regucalcin is X linked. Gene movement out of the X chromosome is a pervasive pattern in Drosophila that was first observed in several genes duplicated by retroposition in D. melanogaster (Betrán et al. 2002). More recently, it has been also detected in genes duplicated via DNA when analyzing the complete genomes of 12 Drosophila species (Meisel et al. 2009; Vibranovski et al. 2009). The conservation of introns in Dca indicates that its origin was not due to retroposition but to duplication at the DNA level. Therefore, it can be concluded that regucalcin is the ancestral gene from which the derived Dca gene arose and later relocated into a different chromosome by a DNA-based movement.

Gene duplication has generally been thought to be a primary source of material for the origin of evolutionary novelties in all kinds of life forms (Sidow 1996; Kondrashov et al. 2002) providing a mechanism for evolution of divergent protein function (Ohno 1970; Piatigorsky and Wistow 1991; Ohta 1993; Hughes 1994), novel gene expression patterns (Force et al. 1999; Lynch and Force 2000), or both (Gibert 2002; Hughes 2002). The acquisition of a new function by one of the duplicates while the other retains the ancestral function is a process known as neofunctionalization. Present data would support the neofunctionalization of Dca after its origin. First, maximum likelihood estimates of the ratio of nonsynonymous to synonymous divergence are much higher for the branch leading to Dca than for the regucalcin branch when these branches are specified as foreground branches in the Br-dup-Dca and Br-dup-RC models ($\omega = 0.384$ and $\omega = 0.098$, respectively). This result is consistent with strong purifying selection acting on regucalcin and an accelerated rate of nonsynonymous substitutions in Dca. Second, the footprint of positive selection was detected by the BEB method in six sites of the DCA protein soon after its origin. Four of these sites are located between $\beta$-structures, which is consistent with the trend toward an excess of adaptive changes in disordered regions of the proteins detected in Drosophila (Ridout et al. 2010). Third, the number of fixed amino acid replacements was much higher in the branch leading to Dca than in the regucalcin branch according to either the FR model of the PAML package (71.6 vs. 7.8 nonsynonymous substitutions, respectively) or the inferred sequences in the ancestral
nodes (59 vs. 7 amino acid replacements, respectively). Finally, putative selection on changes that alter the physicochemical properties of amino acids was detected during the early evolution of the Dca branch (dDca in fig. 2) but not in the regucalcin branch (dRC) after the duplication. Therefore, these results are consistent with a neofunctionalization process and would suggest that Dca has acquired a new function after its origin through the fixation of advantageous mutations driven by positive selection. This new function would be related to cold acclimation.

The implication of Dca in cold acclimation seems to have also affected its molecular evolution in Drosophila. Indeed, the \( \omega \) ratio differs in the lineages of the tropical and temperate species. The lower \( \omega \) value in the temperate than in the tropical lineages would suggest stronger purifying selection against amino acid replacements in the former lineages. Negative selection might thus have contributed to keep the adaptation of the temperate species to cope with extreme low temperatures. In addition, Dca has suffered recurrent duplication events suggesting a dynamic evolution of this gene in the Sophophora species. Similarity of the paralogous genes in the three lineages with multiple Dca copies indicates differences in the timing of the duplication events. Indeed, the duplication is quite ancient in the D. ananassae and D. willistoni lineages but quite recent in the D. pseudoobscura lineage. Although it can be argued that gene conversion might have contributed to the homogenization of the copies in the latter lineage, this possibility seems unlikely as Dca is a single copy gene in the closely related species D. persimilis.

The Dca and regucalcin genes also diverged in their expression pattern as the expression profile of both genes differs in time and space. Hence, available data at Flybase (http://flybase.org) show that during metamorphosis regucalcin is highly expressed, but Dca has only a low expression level. Moreover, according to FlyAtlas (Chintapalli et al. 2007; http://www.flyatlas.org), Dca is mainly expressed in adult structures, such as the tubule, midgut, spermatheca, and fat body. In contrast, the highest expression level of regucalcin is found in the larval fat body, where the Dca transcript is almost absent. The expression pattern of regucalcin is similar to that of the AFP gene of S. peregrina and C. vicina that, indeed, was isolated given the abundance of the encoded protein in the anterior fat body of larvae in the former species (Nakajima and Natori 2000, Hansen et al. 2002). AFP is also a member of the Smp-30/Gluconolactonase/LRE-like family. A Blast search using the AFP gene of S. peregrina and C. vicina against the Drosophila species genomes only detected sequence similarity to the Dca and regucalcin genes. In addition, similarity of the AFP protein is substantially higher to REGUCALCIN than to DCA (e.g., 75.2% and 59.2% in the C. vicina–D. melanogaster comparison, respectively). The similar expression pattern of the AFP and regucalcin genes and the similarity of the encoded proteins would suggest that AFP and regucalcin are indeed orthologous genes, which would further support that regucalcin, and not Dca, has retained the ancestral function in Diptera.

Although the molecular function of the Dca and RC genes in Drosophila is not well characterized, the roles of Dca in cold acclimation (Goto 2000) and likely in wing size (McKechnie et al. 2010) have not been attributed to regucalcin. Indeed, this gene was named regucalcin in D. melanogaster given its similarity to the smp-30 gene of mammals in which the SMP-30 protein was originally identified as a \( \text{Ca}^{2+} \)-binding protein with a key regulatory role in calcium homeostasis whose expression decreased with age (Fujita 1999; Yamaguchi 2005). More recently, it has also been characterized as a GLN involved in the L-ascorbic acid biosynthesis pathway in mice liver (Kondo et al. 2006). Therefore, SMP-30 accomplishes both enzymatic and nonenzymatic functions in mammals. However, present data indicate that the regucalcin orthologous gene in Diptera is the gene coding the AFP protein that has been shown to have no calcium binding activity in S. peregrina (Nakajima and Natori 2000). Therefore, the relation of regucalcin with \( \text{Ca}^{2+} \) homeostasis can be questioned in Drosophila. Instead, the putative role of AFP in S. peregrina and C. vicina is the incorporation of storage proteins, mainly hexamerins, from the haemolymph to the fat body cells by interacting with the hexamerin receptor before the formation of the puparium (Nakajima and Natori 2000; Hansen et al. 2002). In this latter study, substantial amounts of AFP were also detected in larval hemocytes, and interestingly, REGUCALCIN has also been identified as a larval hemolymph protein in a proteomics survey in D. melanogaster (Vierstraete et al. 2003). Therefore, it can be suggested that after the duplication event of the ancestral regucalcin gene, this gene maintained the AFP function in Drosophila, whereas Dca acquired the new function involved in cold acclimation, likely related to the regulation of intracellular \( \text{Ca}^{2+} \) levels that may be important during different stresses in insects as the rapid cold hardening response (Teets et al. 2008). However, further functionally and biochemical analyses in Drosophila are essential for a better understanding of the molecular function of Dca and regucalcin.

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

Acknowledgments
We thank M. Aguadé for critical comments on the manuscript. We also thank Serveis Científico-Tècnics, Universitat de Barcelona, for automated DNA sequencing facilities. This work was supported by a predoctoral fellowship from Ministerio de Ciencia y Tecnología, Spain, to C.E.A-B.; and grants BFU2004-2253, BFU2007-63228 from Comisión Interdepartamental de Ciencia y Tecnología, Spain and 2005SGR-166, 2009SGR-1287 from Comissió Interdepartamental de Recerca i Innovació Tecnològica, Catalonia, Spain, to M. Aguadé.
References


