Published in final edited form as:
Insect Mol Biol. 2009 October ; 18(5): 607–622. doi:10.1111/j.1365-2583.2009.00902.x.

Annotation and analysis of low-complexity protein families of *Anopheles gambiae* that are associated with cuticle

R. S. Cornman and **J. H. Willis**
Department of Cellular Biology, University of Georgia, Athens, GA, USA

Abstract

We have characterized four new families of homologous genes of the mosquito, *Anopheles gambiae*, all of which include members shown by previous work to be cuticular in nature. The CPLCG, CPLCW, CPLCP, and CPLCA families (where CPLC is 'cuticular protein of low complexity') encode proteins with a high proportion of low-complexity sequence. We have also annotated the *An. gambiae Tweedle* genes, a family of cuticular protein genes first described in *Drosophila*, and additional ungrouped *An. gambiae* cuticular proteins identified by proteomics. Our annotations reveal multiple gene-family expansions that are specific to Diptera or Culicidae. The CPLCG and CPLCW families occur within a large and dynamic tandem array on chromosome 3R that includes sets of concertedly evolving genes. Most gene families exhibit two or more different expression profiles during development.

Keywords

cuticular protein; *Anopheles gambiae*; gene expression; gene family; concerted evolution

Introduction

The insect cuticle is a sophisticated exoskeleton and environmental interface with remarkable biomechanical properties. It is composed of chitin fibres embedded in an ordered matrix of proteins. Recent protein-level and gene-level analyses have demonstrated a surprising diversity of cuticular proteins within and amongst species. Many of these proteins belong to the CPR family, which is defined by a well-conserved domain termed the Rebers and Riddiford Consensus (Rebers & Riddiford, 1988) that has chitin-binding properties (Rebers & Willis, 2001; Togawa *et al.*, 2004). This gene family has been annotated for several insect genomes and is found in other arthropods, implying an ancient origin.

Correspondence: Scott Cornman, USDA-ARS Bee Research Lab, BARC-West Bldg 476, Beltsville, MD 20705, USA. Tel.: +1 301 364 7011; fax (c/o Judith Willis): +1 706 542 4271; scott.cornman@gmail.com.

Supporting Information
Additional Supporting Information may be found in the online version of this article:
Appendix S1. Sequences of low-complexity cuticular proteins used in phylogenies.
Appendix S2. Sequences of genes in Table S3 that lack Ensembl or Vectorbase identifiers.
Figure S1. Four replicates of chitin synthase expression showing results normalized to Table S4 (top) and total RNA (bottom).
Table S1. Annotation summaries of *Anopheles gambiae* genes described in this paper.
Table S2. Summary of quantitative gene expression.
Table S3. List of genes in CPLCG/CPLCW tandem arrays (where CPLC is 'cuticular protein of low complexity') of *Culex pipiens* and *Aedes aegypti* that are shown in Fig. S1.
Table S4. Genes in each of the four larval expression clusters described in the text.
Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Numerous cuticular proteins have been identified that lack the Rebers and Riddiford Consensus. These include the *Tweedle* gene family (Guan *et al.*, 2006), the *Dacp* (Qiu & Hardin, 1995) and *Edg91* (Apple & Fristrom, 1991) genes first identified in *Drosophila*, and the CPF gene family first identified in *Tenebrio molitor* (Andersen *et al.*, 1997). More recently, 'apidermins' have been described in *Apis mellifera* (Kucharski *et al.*, 2007) and a 'CPF-like' family that is present in a number of insect groups (Togawa *et al.*, 2007). How these gene families are functionally related to each other and to the more abundant CPR gene family is not known. An important initial step in addressing this question is to characterize fully all of the genes of a suitable model organism that contribute structurally or enzymatically to cuticle. Annotation and studies of the spatial and temporal expression of these genes will shed light on the range of potential protein interactions during cuticle assembly. These data will also aid our understanding of the evolutionary origins and developmental control of cuticular components. Here we extend previous annotations of the 'cuticleome' of the mosquito *Anopheles gambiae* (Togawa *et al.*, 2007; Cornman *et al.*, 2008) to include novel gene families identified by proteomic analyses of *An. gambiae* cuticle (He *et al.*, 2007), as well as families identified in other species but not yet annotated in *An. gambiae*. We have focused on this malaria vector in part because cuticular proteins could potentially underlie adaptive responses to human-associated selective pressures, such as aridity (White *et al.*, 2007) or insecticides (Vontas *et al.*, 2007). Continued rapid advancement in comparative and functional genomics has begun to allow powerful studies of cuticular protein function and regulation in these important disease vectors.

Results and discussion

Naming convention

We followed a consistent naming convention for all *An. gambiae* cuticular protein genes characterized in this paper. Most gene family names use the prefix 'CPLC' (for 'cuticular protein of low complexity') plus a single-letter identifier, with each gene numbered in chromosomal order according to Ensembl coordinates (http://www.ensembl.org/Anopheles_gambiae/index.html). The one exception is the use of the prefix 'TWDL' to identify *An. gambiae* homologues of the *Tweedle* family. Note that the phrase 'low complexity' refers in a general sense to the length and composition of the conserved regions of these families relative to the CPR family of cuticular proteins. We do not mean to imply that all proteins with a CPLC prefix are less complex, in a probabilistic sense (Wootton, 1994), than all CPR proteins, as many CPR proteins also have regions of low sequence complexity.

A recurrent feature of cuticular protein genes in *An. gambiae* is the presence of highly similar genes, often as subsets of larger tandem arrays (members of a gene family in close proximity on a chromosome). We use the term 'sequence cluster' to distinguish such similar genes from a tandem array of genes *per se* (see below and Cornman & Willis, 2008).

Phylogenetic subgroups of gene families that are discussed in the text are referred to as 'Group A', 'Group B', etc. according to their order in the text, regardless of the gene family to which they belong. These groups are marked as such in the relevant figures and tables, and may refer to close homologues in other species as well as to *An. gambiae* genes.

Annotation support

Our annotations of the PEST genome sequence are summarized in Supporting Information Table S1, including Ensembl coordinates and gene features. All gene families characterized here are based on sequence homology to genes experimentally verified as cuticular in nature. Although sequence homology does not prove that a related protein is also cuticular, homology and not function is the criterion for grouping genes into evolutionary families.

Indeed, contrasts between function and evolutionary history are of great interest for understanding the molecular basis of adaptation. We are careful to denote in the text and in Supporting Information Table S1 those predicted gene products that are not independently supported by proteomics or other data as cuticular in nature, and excluded these genes from comparative analyses specific to cuticular components. Of course, current methods of mass spectrometry for proteomics are not free of false positives or false negatives arising from computational or biological limitations (see Colinge & Bennett, 2007 for a review).

Sequences of homologues in other species used in phylogenies are provided in Supporting Information Appendix S1.

Quantitative expression data

Using real-time quantitative reverse transcriptase PCR (qRT-PCR), we obtained quantitative estimates of expression for almost all genes annotated in this paper. However, we could not design unique primers for a number of sets of very similar genes. In such cases, we designed shared primers and calculated the average expression of all genes in that set (Supporting Information Table S2).

For most analyses, we report transcript levels normalized to an external reference, total RNA, following the methods of Togawa *et al.* (2007). This normalization was chosen because we have not identified an internal reference gene with constant expression across all developmental stages of interest. Because externally normalized values are contingent on the particular reagents and equipment used, transcript levels are presented in arbitrary fluorescence units. The difference between any two measurements is proportional to the difference in transcript abundances in the original samples. To group genes into expression clusters, however, we normalized expression to the ribosomal protein gene S7 because it has higher reproducibility (see Supporting Information Fig. S1). The approximately fivefold change in S7 expression estimated by Togawa *et al.* (2007, 2008) during the life history does not affect the outcome of gene clustering with S7-normalized data because genes are compared within time points and not between time points.

In our hands, the dynamic range of variation in transcript level appears to be up to five orders of magnitude. Of course, the lower bound is dictated not only by transcript abundance but also by the rate of nonspecific amplification, the complete absence of which is both unlikely and very difficult to confirm. However, in our data, genes with high maxima have consistently higher transcript levels than do those with low maxima across the entire cycle of expression associated with larval moults (see results below). If the gene-expression minima were substantially influenced by nonspecific amplification, we would expect the rank order of genes at lowest abundance to be unrelated to the rank order of genes at highest abundance. Instead, these genes appear to maintain their relative abundance (measured at the whole-animal level) in both up-regulated and down-regulated states.

A few genes did not show the characteristically cyclical expression of cuticular protein genes expressed in larvae. It is possible that nontarget amplicons of similar size and melting temperature affected the quantitative measurement, or that the gene products are not actually cuticular in nature. However, levels of these transcripts are within the range of other CPLC genes, and one gene product (TWDL1, formerly CPLC1) has been detected in cast cuticle by proteomic analyses (He *et al.*, 2007).

The CPLCG gene family

Members of the CPLCG family were first identified in *An. gambiae* by proteomic analysis of cuticle preparations (He *et al.*, 2007; N. He, unpubl. data). Blast searches revealed the gene family to be homologous to cuticular protein genes identified in *Drosophila*

melanogaster by Qiu & Hardin (1995). Those authors named two *Drosophila* genes, *Dacp-1* and *Dacp-2*, for 'Drosophila adult cuticle protein.' We have not continued that nomenclature here because it is now clear that several cuticular protein families have members that are expressed in adults, and homologous genes in *An. gambiae* are expressed in larvae as well. We instead chose 'CPLCG' in reference to two invariant glycine residues in the conserved domain separated by eight amino acids (Fig. 1).

We identified a total of 27 CPLCG genes in *An. gambiae*, all linked in an expansive array on chromosome 3R. Within this array, the CPLCG gene family is interspersed with a second cuticular protein gene family described below (the CPLCW family). Phylogenetic analysis of *An. gambiae* CPLCG genes identified two distinct subgroups (Fig. 2). One group consists of 12 genes that are on average 86.4% identical at the nucleotide level (labelled 'Group A' in Fig. 2). Genes in this sequence cluster contain substantially more glycine than other *An. gambiae* CPLCG genes (16.0 vs. 4.9%). To the extent that unique primers could be designed, all 'Group A' genes have very similar expression patterns that peak in pharate and newly moulted larvae (Fig. 3). The remaining CPLCG genes are more variable in expression, although there is a phylogenetically distinct subgroup of five genes that is highly expressed in adults and/or pharate adults ('Group B' in Fig. 2). This subgroup is interesting because the genes are physically adjacent and lie within a haplotype region of the PEST genome assembly; ie a presumed structural polymorphism segregating in *An. gambiae* (Sharakhova *et al.*, 2007). Blast searches confirmed that these genes are also present in the genome sequences of the Mali-NIH (M-form) and Pimperena (S-form) strains available from Vectorbase (http://www.vectorbase.org). The subgroup includes *CPLCG3*, formerly called *CPLC8*, the putative orthologue of a gene identified by Vontas *et al.* (2007) as the most highly up-regulated gene in pyrethroid-resistant *Anopheles stephensi*. Group B is sister to the *Drosophila* CPLCGs in the gene tree and thus is presumably more ancestral in character. Qiu & Hardin (1995) performed *in situ* hybridization for one of these *Drosophila* genes and found expression throughout the epidermis of the adult head and thorax.

In terms of CPLCG gene number and organization, the co-orthologous regions in *Aedes aegypti* and *Culex pipiens* differ substantially from *An. gambiae* and from each other. Fig. 4). However, all three species contain sets of very similar CPLCG genes that cluster as paralogues within the gene tree (Fig. 2). No CPLCG genes were found outside of the Diptera. The approximate coordinates of the *Ae. aegypti* and *C. pipiens* genes shown in Fig. 4 are provided in Supporting Information Table S3, and the predicted sequences of those genes that lack Ensembl or Vectorbase names are provided in Supporting Information Appendix S2.

The CPLCW gene family

Members of this gene family were also originally identified from *An. gambiae* cuticle by He *et al.* (2007). We have annotated nine genes in the PEST genome, which are interspersed with the CPLCG family in the large tandem array described above (Fig. 4). One gene, *CPLCW5*, has a frameshift in the genome assembly but no other evident defect. We did not detect the frameshift in the M and S strains of *An. gambiae*.

The coding sequences of *An. gambiae* CPLCW genes average a remarkable 98.9% pairwise nucleotide identity. Thus, comparisons with other species were necessary to determine what part of the sequence defines the gene family across taxa. The region of sequence that is well aligned and well conserved amongst CPLCW genes of all three mosquito species is relatively short (Fig. 5, bottom), even though co-orthology is unambiguously supported by the synteny of the larger tandem array and short, shared amino-acid motifs. The name 'CPLCW' refers to an invariant tryptophan within the conserved domain. No other examples of this gene family were found outside of mosquitoes, suggesting an origin after the

divergence of these taxa from the *Drosophila* lineage approximately 250 million years ago (Gaunt & Miles, 2002), in contrast to the implicitly older CPLCG family with which it is physically linked. Alternatively, the CPLCW family may have been lost in *Drosophila*.

Remarkably, the CPLCW gene tree (Fig. 6) shows complete clustering by species, a pattern also seen for a large subset of the CPLCG phylogeny, as noted above. This pattern implies concerted evolution by unequal crossing-over or intergenic gene conversion (discussed below), despite the complex arrangement of the CPLCG and CPLCW gene families within the array (Fig. 4).

Primers for real-time qRT-PCR were designed for a single CPLCW gene (*CPLCW7*) and for the entire gene family as a group. The expression profiles measured with these two primer pairs were identical, and the relative transcript levels were close to the expected ratio (1:9) assuming equal expression of all genes (Fig. 3). The CPLCW family has the same expression profile as the 'Group A' CPLCG genes described above. We have initiated *in situ* hybridization studies to determine whether these two groups of genes in fact contribute to the same cuticular structure.

The CPLCP gene family

He *et al.* (2007) identified peptides from cuticle that match four predicted genes encoding proteins rich in proline, valine, lysine, and tyrosine. We grouped these proteins together as a single 'proline-rich' gene family named CPLCP. Blast searches with these sequences identified numerous potential homologues, although the low complexity of these regions inflates significance scores of sequence alignments, such that homology can be more difficult to judge. Fortunately, other conserved sequence features outside of the highly repetitive, proline-rich region provide further evidence of homology (Fig. 7). The N-terminus of the mature peptide contains three distinct sequence features in a conserved order: (1) a region rich in polar and acidic residues; (2) a GLW[D/E] motif; and (3) a region rich in glycine, tyrosine, and histidine. Also characteristic is a histidine-rich region near the C-terminus (not shown). Both *Drosophila* CPLCP homologues that have embryonic *in situ* hybridization images in the FlyExpress database (CG30101 and CG16885, http://www.flyexpress.net) are expressed in cuticular structures such as tracheae, spiracles, and pharynx.

Most *An. gambiae* CPLCP gene annotations are based on sequence homology alone, as uniquely identifying peptides were not detected from cuticular structures in the proteomics analyses of He *et al.* (2007) despite measurable expression of these genes in larvae, pupae, and adults (see below). This is an intriguing finding, as all other cuticular protein families identified in *An. gambiae* have proteomics support for a high proportion of homologues (He *et al.* 2007; Cornman *et al.* 2008; Supporting Information Table S1). One of the proteins not detected by proteomics (CPLCP3) is the putative orthologue of the *Drosophila* gene CG30101 mentioned above.

Overall, 28 *An. gambiae* CPLCP genes were found concentrated in several tandem arrays, most notably a 16-gene tandem array that includes a 14-gene sequence cluster ('Group C' in Fig. 8). Smaller co-orthologous arrays were found in the other two mosquito genomes as well, and constitute the bulk of a mosquito-specific expansion of this gene family. The gene family was found in all other insect genomes that we searched, but in smaller numbers.

Several CPLCP genes contain frameshifts in the PEST genome sequence. However, like *CPLCW5* mentioned above, these frameshifts were not detected in the M- and S-strain sequence reads and may be assembly artefacts or mutations arising in culture. Our annotations of all these genes terminate with those bases that would constitute the stop

codon in the absence of these frameshifts, and frameshifts were necessarily ignored in the construction of all gene trees.

Expression profiles of *An. gambiae* CPLCP genes are shown in three panels of Fig. 3. The three panels distinguish all genes for which there is proteomics support for their localization in cuticle, all Group C genes (which have highest expression in larvae), and all other CPLCP genes, which have expression profiles similar to those of confirmed cuticular protein genes, particularly with regards to the strong peak in 12-h pupae.

The proline-rich region of CPLCP proteins is strikingly similar (Fig. 9) to a group of protozoan proteins called articulins, which, although intracellular, form fibrils and are believed to provide support, pattern formation, and some elasticity to the cytoskeleton (Huttenlauch *et al.*, 1998). These properties also characterize insect cuticle and thus the sequence similarity suggests a case of functional convergence.

CPLCA gene family

Two genes identified by He *et al.* (2007) and an additional Ensembl-annotated gene not detected by proteomics, AGAP006148, constitute a small family of alanine-rich cuticular proteins that we have accordingly named CPLCA (Fig. 10). The three genes are adjacent to each other within a larger array that includes another cuticular protein gene (CPLCX1, see below) detected by He *et al.* (2007) that has a similar amino-acid composition. However, this latter gene does not align well with the others and we do not consider it homologous.

We identified CPLCA homologues in other mosquitoes and in *Drosophila* but none were found outside the Diptera (Fig. 11). The *D. melanogaster* homologues are embedded within a larger array of genes that predominantly code for alanine-rich proteins of unknown function and which contain signal peptides. These homologues include the gene *retinin*, which is known to be expressed in the cornea (Kim *et al.*, 2008). In fact, the conserved sequence that defines the CPLCA family matches a previously recognized domain termed the retinin domain (Pfam04527/IPR007614). In the gene-family phylogeny, however, *retinin* lies on a long branch outside the main body of genes and thus is not a very representative member of the gene family.

We obtained quantitative expression data for two of the three CPLCA genes in *An. gambiae*. The two genes differed several fold in transcript level but had identical expression profiles (Fig. 3). Pharate pupal and pupal expression peaks were 10-fold higher than larval and adult peaks, yet in general these proteins showed the smallest differences between up-and down-regulated states of the gene families studied here.

The TWDL gene family

The *Tweedle* gene family was originally identified by a mutant screen for body shape in *Drosophila*. The first mutant identified (*TwdlD*) was shown to cause a cuticular defect (Guan *et al.*, 2006). Twenty-seven members of this family were identified in *Drosophila* and temporal and spatial expression patterns indicated that all seven genes examined could contribute to cuticular structures (Guan *et al.*, 2006). Those authors also reported that *Tweedle* homologues were present in other insects.

We identified 12 *Tweedle* homologues in the PEST genome sequence. Six TWDL genes occur as a tandem array on the X chromosome and are part of a mosquito-specific subfamily (‘Group D’ in Fig. 12). The two genes at the upstream end of the array are nearly identical at the nucleotide level (98.8%) and there are nearly identical copies of an *AG-RTE1* retrotransposon inserted approximately 1.5 kb from the 3′ end of the genes. Thus, it is unclear whether this region is a recent tandem duplication or an assembly artefact. We were

able to design primers for each gene that were unique at the 3′ end of each primer, however, and obtained single-copy kinetics for both primer sets, implying that both sequences are present in the G3 strain. Comparisons with the sequence reads of the M and S strains indicate that both genes are present in S but may not be both present in M (not shown).

In total, nine TWDL genes were identified in the mosquito *Ae. aegypti* and 10 in *C. pipiens*, compared with the 27 known in *D. melanogaster* (Guan *et al.*, 2006). In almost all other insect species that we searched, we found only two *Tweedle* genes. *Bombyx mori* is a minor exception in that these two genes have been duplicated for a total of four *Tweedle* genes. A phylogenetic analysis (Fig. 12) identifies three well-supported groups: a *Drosophila*-specific expansion, a mosquito-specific expansion, and the ancestral clade that includes all *Tweedle* genes from non-Dipteran species. A comparative study of these ancestral genes would provide important insights into the evolutionary origins of the *Tweedle* family and the functional significance of the gene-family expansion in the Diptera. Guan *et al.* (2006) found that *Drosophila* Tweedles varied spatially and temporally in their expression, but did not investigate whether they also differed in their biochemical properties.

The Group D genes do not show a level of sequence similarity comparable to the ‘sequence clusters’ found in other mosquito cuticular protein families (see above and Cornman & Willis 2008), with pairwise nucleotide similarities averaging 68.2% and numerous indels. Nonetheless, *An. gambiae* paralogues in this clade cluster together and not with presumably orthologous genes in the other mosquitoes. This contrasts strikingly with the phylogenetic pattern of the ‘ancestral’ genes, which, notably, are not tandemly arrayed in the genome. Group D proteins have more ordered amino-acid repeats than do other Tweedles and are alanine-rich (23.0 vs. 8.5% in *An. gambiae*).

Expression profiles of the *Tweedle* gene family are summarized in Fig. 3. The Group D genes are more variable in expression than other *An. gambiae* sequence clusters. The two genes upstream of *AG-RTE1* retrotransposons are shifted 12 to 24 h in peak expression relative to other genes in Group D. Because the coding sequences of these retroelements appear intact (not shown), the insertion event is likely to be a recent one.

Ungrouped cuticular protein genes

He *et al.* (2007) identified three additional genes that code for novel proteins of low complexity. These could not be placed with other *An. gambiae* genes, nor were any homologues detected in other species. These genes were assigned the prefix ‘CPLCX’ and indexed by chromosome coordinates. *CPLCX1* and *CPLCX3* encode proteins of similar length and amino-acid composition (high in alanine, proline, tyrosine, and valine) and have very similar expression patterns (Fig. 13), yet are found on opposite arms of chromosome 2. The two genes lack strong evidence of homology by sequence alignment and differ in exon structure. Thus, while *CPLCX1* and *CPLCX3* are potentially homologous this remains to be clarified.

Patterns of larval expression

Based on the analysis of Togawa *et al.* (2008) and the data presented here, it is apparent that most larval-expressed genes for cuticle proteins have characteristic patterns relative to ecdysis that can be meaningfully described as ‘early’, ‘middle’, or ‘late’. We therefore used the self-organizing map (SOM) method of Tamayo *et al.* (1999) to cluster the expression profiles of larval-expressed genes into four groups, anticipating that this clustering approach would discriminate amongst early, middle, and late genes, as well as identify residual genes that do not fall into this trichotomy. For this analysis, we excluded the final two L4 stages to minimize the influence of gene expression associated with pupal structures during this

interval. We also renormalized the expression of each gene to its maximum value in the reduced data set, because we were considering only the timing of expression peaks and not their magnitude.

An important goal of this clustering approach is to relate the expression of 'low-complexity' cuticular proteins to the expression of the larger and older CPR family, which to date is the only group for which chitin-binding potential has been shown (Rebers & Willis, 2001; Togawa *et al.*, 2004, 2007). We therefore combined the present data with previous results for the CPR (Togawa *et al.*, 2008) and CPF/CPFL (Togawa *et al.*, 2007) families, which were derived from the same cDNA series.

Figure 14A shows the expression profiles of each larval SOM cluster, whereas Fig. 14B shows the number of genes in each expression cluster by family. The cycles of expression for three of the SOM clusters are initially coincident, rising to high levels at the final time point of each larval stage from minima that occur in the preceding time point. These clusters diverge only at the 0-h time points, at which expression either drops to a new minimum, declines but remains substantial, or continues to increase. These three patterns are labelled 'Early', 'Middle', and 'Late' in Fig. 14A and B, indicating the inferred timing of peak expression. Thus, at this level of resolution, genes differ less with respect to their initial activation and more with respect to rate of increase and persistence of message. The fourth expression cluster is out of phase, having maxima at time points that are minima for the other clusters. Genes in this 'irregular' expression cluster tend to have their highest expression in the final larval stage, but the variance around the mean is large at each time point (not shown) and no general pattern is evident. This cluster has the fewest number of genes, whereas the 'Early' cluster has the largest number of genes, largely because of the number of CPR genes. The names of genes in each expression cluster are listed in Supporting Information Table S4.

Concerted evolution

The recurring pattern of concerted evolution within *An. gambiae* cuticular protein families is, to our knowledge, the most extensive example of this phenomenon yet identified amongst protein-coding genes. Most families contain one or more sets of genes that have low synonymous variation and which cluster phylogenetically with within-species paralogues, rather than with genes from co-orthologous regions in other species (identifiable by shared sequence features and synteny). These 'sequence clusters' are evident in the alignments and phylogenies of the CPLCG, CPLCW, and CPLCP families described here, although in the latter family, sequence-cluster genes have not been experimentally demonstrated to be cuticular in nature. Sequence clusters are also found in the *An. gambiae* CPFL family (AgamCPFL2 toAgamCPFL7, see Togawa *et al.*, 2007) and are particularly common in the CPR family (Cornman & Willis, 2008). The fact that sequence clusters need not be on the same strand or adjacent within a tandem array implies that simple tandem duplication and loss is not the major mechanism homogenizing genes, although dot-plot analysis does demonstrate that tandem duplication occurs (not shown). There is also no evidence that any genes characterized in this paper derive from reverse-transcribed mRNA, because introns and promoter sequences are intact (see Supporting Information Table S1).

It is remarkable that we have yet to identify an obviously degenerate pseudogene in any *An. gambiae* cuticular protein family (but see Cornman *et al.*, 2008). Although the ascertainment bias against highly degenerate pseudogenes is not trivial, our data clearly contrast with the evolutionary dynamics of some other large gene families such as mammalian odorant receptors (Glusman *et al.*, 2001) that show extensive pseudogenization. Given the many sets of cuticular protein genes that temporally co-express very similar protein products, the lack

of degenerate pseudogenes is surprising. Comparative studies with other *Anopheles* species would further clarify the molecular evolution of these regions.

Conclusions

This study has shown that the diversity of cuticular proteins within *An. gambiae* is even greater than previously realized. The present study increases the count of experimentally supported cuticular proteins by 54 over the previously characterized 156 CPR proteins (Cornman *et al.*, 2008), 11 CPF/CPFL proteins (Togawa *et al.*, 2007), four 'CPTC' genes (He *et al.*, 2007), and four 'CPR+C' genes (He *et al.*, 2007). An additional 28 homologues of these genes (24 CPLCPs, one CPLCG, one CPLCA, and two Tweedle genes) were identified by sequence analysis but not detected in existing proteomics data. The count of cuticular protein genes may be even higher in *Ae. aegypti* and *C. pipiens* (this study and Cornman *et al.*, 2008). This figure does not include genes associated with cuticle but not believed to be structural in nature, such as those contributing to chitin metabolism, or the pigmentation, sclerotization, or digestion of cuticle.

The functional significance of this structural protein diversity remains to be explored. It is known that the amino-acid composition of cuticular proteins is an important determinant of the biomechanical properties of cuticle. For example, histidine and lysine are important substrates of cross-linking reactions that affect cuticle stiffness (Andersen *et al.*, 1995) and are conspicuous contributors to the overall variation in amino-acid composition amongst *An. gambiae* cuticular proteins (this study and Cornman *et al.*, 2008). The high elasticity of the resilin protein, a CPR protein originally identified by Weis-Fogh (1960), has been shown to derive from a short repeated motif flanking the conserved domain (Elvin *et al.*, 2005). Furthermore, some cuticular proteins are expressed only in specific tissues and presumably contribute to their specialized function, such as the *Drosophila* protein crystallin that is specific to eyes (Janssens & Gehring, 1999). However, for the large majority of cuticular proteins, *in situ* hybridization data and functional analyses are lacking. Of course, gene expression and proteomic studies have highlighted broad distinctions amongst proteins in terms of where and when they are expressed. Yet at any given period of cuticle synthesis, message for many different cuticular proteins can be detected in whole-animal extracts.

It remains to be seen whether mosquitoes are exceptional in terms of the proportion of their genome devoted to cuticle. Whereas *Ap. mellifera* (Honeybee Genome Sequencing Consortium, 2006), *D. melanogaster* (Karouzou *et al.*, 2007), and *Nasonia vitripennis* (J. H. Willis, unpubl. data) all have fewer CPR genes, *B. mori* (Futahashai *et al.*, 2008) has virtually the same number. Furthermore, based on this study and work in other species (eg Apple & Fristrom, 1991, Kucharski *et al.*, 2007), a significant portion of the cuticleome is narrowly distributed, taxonomically speaking. This finding implies a relatively rapid turnover not only in gene number but in the gene families themselves, and complicates the use of a single species as a general model for functional analyses.

Experimental procedures

Existing Ensembl annotations were inspected and manually revised if necessary to correct errors evident by comparison with other homologues, or to identify missing sequence features such as a signal peptide. Novel candidate genes were identified by Blast searches with previously characterized cuticular proteins, or from peptides identified by mass spectrometric analyses of *An. gambiae* cuticle preparations (He *et al.*, 2007). Intron–exon junctions were inferred from protein-level homology and appropriate intron sequence features, a task simplified by the fact that few genes in this study had more than one intron. Other supporting features that we identified include the TATA box sequence, initiator/cap

site motifs (Cherbas & Cherbas, 1993), the Kozak motif (Kozak, 1999), start and stop codons, signal peptides, and canonical polyadenylation sites. Ambiguous annotations were confirmed by sequencing transcripts of the G3 strain of *An. gambiae*, using standard PCR cloning or rapid amplification of cDNA ends (RACE). RACE was performed with the Invitrogen GeneRacer kit (Invitrogen, Carlsbad, CA, USA) with a modified 3′ RACE primer as described in Cornman *et al.* (2008).

Homologues in other insect species were identified by Blast and, if not previously annotated, partial or complete coding sequences were obtained or modified from Genscan (Burge & Karlin, 1997) gene predictions. Sequence logos were generated by WebLogo (Crooks *et al.*, 2004). Gene-family alignments were constructed from amino-acid sequence using ClustalW (Thompson *et al.*, 1994) with default parameters. Similarity shading was based on the BLOSUM62 matrix. Neighbor-joining phylogenies of amino-acid sequences were created with Mega4 (Tamura *et al.*, 2007) using the substitution model of Jones *et al.* (1992), pairwise deletion of indels, and ignoring rate heterogeneity amongst sites. Bootstrap support was based on 1000 resamples.

Real-time qRT-PCR was performed using the BioRad SYBR-GREEN fluorescent marker, iCycler thermocycler, and MyIQ software (BioRad, Hercules, CA, USA) following the methods of Togawa *et al.* (2007). Briefly, RNA was extracted from animals collected at 19 different time points, quantified using RiboGreen RNA quantitation reagent (Molecular Probes, Eugene, OR, USA), and reverse-transcribed. For larval and pupal stages, animals were collected at discrete intervals approximately 12 h apart, whereas the adult stage was represented by pooling individuals that were between 0 and 12 h old. We identify the four larval stages by the abbreviations L1 to L4, with P and A representing pupal and adult, respectively. These are combined with animal age using a hyphen, such that ‘L1–12’, for example, designates L1 larvae that are 12 h old. All primers were evaluated for efficiency, correct amplicon size, melt-curve dynamics, and single-copy dynamics on genomic DNA. Transcript concentration was normalized to total RNA and expressed in the arbitrary units of raw fluorescence (‘R0’). Separate measurements made for male and female pupae and adults were pooled for analysis. Because two different combinations of iCycler and MyIQ software were used for these analyses, we ran replicate plates under both conditions to normalize the raw fluorescence values.

We performed all SOM clustering with the GenePattern (Reich *et al.*, 2006) SOM module under default parameters. Genes were excluded if the maximum expression during the larval period (L1-0 to L4–24) was less than 1% of maximum for the whole time series. This was carried out to filter out quantitative measurements at the lower end of the dynamic range, as this variation is potentially less informative for clustering purposes. The data for each gene were then re-normalized to a maximum value of 1.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Toru Togawa and Aaron Emmons provided technical assistance. Toru Togawa also provided access to previously published data in raw format. Jay Evans and Monica Poelchau provided helpful comments, and the manuscript was further improved by the critique of three anonymous reviewers. This work was supported by a grant from the National Institutes of Health (A155624) to J. H. W.

References

Andersen SO, Hojrup P, Roepstorff P. Insect cuticular proteins. Insect Biochem Mol Biol. 1995; 25:153–176. [PubMed: 7711748]

Andersen SO, Rafn K, Roepstorff P. Sequence studies of proteins from larval and pupal cuticle of the yellow meal worm, *Tenebrio molitor*. Insect Biochem Mol Biol. 1997; 27:121–131. [PubMed: 9066122]

Apple RT, Fristrom JW. 20-Hydroxyecdysone is required for, and negatively regulates, transcription of *Drosophila* pupal cuticle protein genes. Dev Biol. 1991; 146:569–582. [PubMed: 1713868]

Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997; 268:78–94. [PubMed: 9149143]

Cherbas L, Cherbas P. The arthropod initiator: the capsite consensus plays an important role in transcription. Insect Biochem Mol Biol. 1993; 23:81–90. [PubMed: 8485519]

Colinge J, Bennett KL. Introduction to computational proteomics. PLoS Comput Biol. 2007; 3:e114. [PubMed: 17676979]

Cornman RS, Willis JH. Extensive gene amplification and concerted evolution within the CPR family of cuticular proteins in mosquitoes. Insect Biochem Mol Biol. 2008; 38:661–676. [PubMed: 18510978]

Cornman RS, Togawa T, Dunn WA, He N, Emmons AC, Willis JH. Annotation and analysis of a large cuticular protein family with the R&R Consensus in *Anopheles gambiae*. BMC Genomics. 2008; 9:22. [PubMed: 18205929]

Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004; 14:1188–1190. [PubMed: 15173120]

Elvin CM, Carr AG, Huson MG, Maxwell JM, Pearson RD, Vuocolo T, et al. Synthesis and properties of crosslinked recombinant pro-resilin. Nature. 2005; 437:999–1002. [PubMed: 16222249]

Futahashai R, Okamoto S, Kawasaki H, Zhong YS, Iwanaga M, Mita K. Genome-wide identification of cuticular protein genes in the silkworm, *Bombyx mori*. Insect Biochem Mol Biol. 2008; 38:1138–1142. [PubMed: 19280704]

Gaunt MW, Miles MA. An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. Mol Biol Evol. 2002; 19:748–761. [PubMed: 11961108]

Glusman G, Yanai I, Rubin I, Lancet D. The complete human olfactory subgenome. Genome Res. 2001; 11:685–702. [PubMed: 11337468]

Guan X, Middlebrooks BW, Alexander S, Wasserman SA. Mutation of TweedleD, a member of an unconventional cuticle protein family, alters body shape in *Drosophila*. Proc Natl Acad Sci USA. 2006; 103:16794–16799. [PubMed: 17075064]

He N, Botelho JM, McNall RJ, Belozerov V, Dunn WA, Mize T, et al. Proteomic analysis of cast cuticles from *Anopheles gambiae* by tandem mass spectrometry. Insect Biochem Mol Biol. 2007; 37:135–146. [PubMed: 17244542]

Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature. 2006; 443:931–949. [PubMed: 17073008]

Huttenlauch I, Peck RK, Stick R. Articulins and epiplasmins: two distinct classes of cytoskeletal proteins of the membrane skeleton in protists. J Cell Sci. 1998; 111:3367–3378. [PubMed: 9788878]

Janssens H, Gehring WJ. Isolation and characterization of drosocrystallin, a lens crystallin gene of *Drosophila melanogaster*. Dev Biol. 1999; 207:204–214. [PubMed: 10049575]

Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 1992; 8:275–282. [PubMed: 1633570]

Karouzou MV, Spyropoulos Y, Iconomidou VA, Cornman RS, Hamodrakas SJ, Willis JH. *Drosophila* cuticular proteins with the R&R Consensus: annotation and classification with a new tool for discriminating RR-1 and RR-2 sequences. Insect Biochem Mol Biol. 2007; 37:754–760. [PubMed: 17628275]

Kim E, Choi Y, Lee S, Seo Y, Yoon J, Baek K. Characterization of the *Drosophila melanogaster* retinin gene encoding a cornea-specific protein. Insect Mol Biol. 2008; 17:537–543. [PubMed: 18828839]

Kozak M. Initiation of translation in prokaryotes and eukaryotes. Gene. 1999; 234:187–208. [PubMed: 10395892]

Kucharski R, Maleszka J, Maleszka R. Novel cuticular proteins revealed by the honey bee genome. Insect Biochem Mol Biol. 2007; 37:128–134. [PubMed: 17244541]

Qiu J, Hardin PE. Temporal and spatial expression of an adult cuticle protein gene from *Drosophila* suggests that its protein product may impart some specialized cuticle function. Dev Biol. 1995; 167:416–425. [PubMed: 7875368]

Rebers JE, Riddiford LM. Structure and expression of a *Manduca sexta* larval cuticle gene homologous to *Drosophila* cuticle genes. J Mol Biol. 1988; 203:411–423. [PubMed: 2462055]

Rebers JE, Willis JH. A conserved domain in arthropod cuticular proteins binds chitin. Insect Biochem Mol Biol. 2001; 31:1083–1093. [PubMed: 11520687]

Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nat Genet. 2006; 38:500–501. [PubMed: 16642009]

Sharakhova MV, Hammond MP, Lobo NF, Krzywinski J, Unger MF, Hillenmeyer ME, et al. Update of the *Anopheles gambiae* PEST genome assembly. Genome Biol. 2007; 8:R5. [PubMed: 17210077]

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA. 1999; 96:2907–2912. [PubMed: 10077610]

Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol. 2007; 24:1596–1599. [PubMed: 17488738]

Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994; 22:4673–4680. [PubMed: 7984417]

Togawa T, Nakato H, Izumi S. Analysis of the chitin recognition mechanism of cuticle proteins from the soft cuticle of the silkworm, *Bombyx mori*. Insect Biochem Mol Biol. 2004; 34:1059–1067. [PubMed: 15475300]

Togawa T, Augustine Dunn W, Emmons AC, Willis JH. CPF and CPFL, two related gene families encoding cuticular proteins of *Anopheles gambiae* and other insects. Insect Biochem Mol Biol. 2007; 37:675–688. [PubMed: 17550824]

Togawa T, Dunn WA, Emmons AC, Nagao J, Willis JH. Developmental expression patterns of cuticular protein genes with the R&R Consensus from *Anopheles gambiae*. Insect Biochem Mol Biol. 2008; 38:508–519. [PubMed: 18405829]

Vontas J, David JP, Nikou D, Hemingway J, Christophides GK, Louis C, et al. Transcriptional analysis of insecticide resistance in *Anopheles stephensi* using cross-species microarray hybridization. Insect Mol Biol. 2007; 16:315–324. [PubMed: 17433071]

Weis-Fogh T. A rubber-like protein in insect cuticle. J Exp Biol. 1960; 37:889–907.

White BJ, Hahn MW, Pombi M, Cassone BJ, Lobo NF, Simard F, et al. Localization of candidate regions maintaining a common polymorphic inversion (2La) in *Anopheles gambiae*. PLoS Genet. 2007; 3:e217. [PubMed: 18069896]

Wootton JC. Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem. 1994; 18:269–285. [PubMed: 7952898]

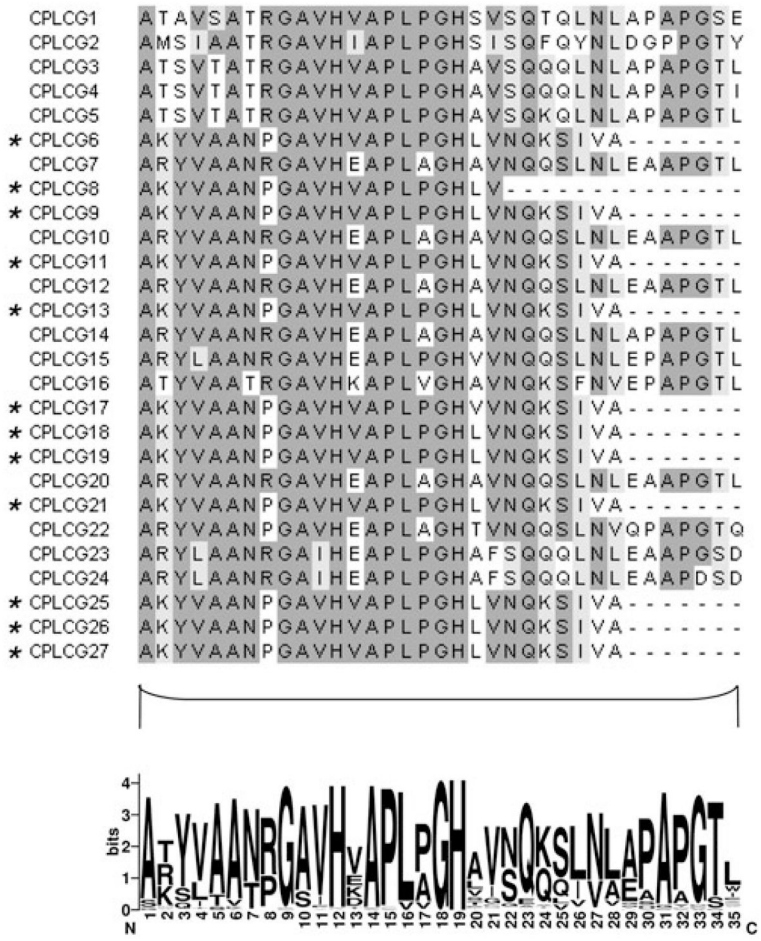

Figure 1.

Similarity-shaded alignment of the C-terminus of *Anopheles gambiae* CPLCG predicted proteins (where CPLC is ‘cuticular protein of low complexity’) showing the region of conserved sequence. The sequence logo represents the consensus of all three mosquito and *Drosophila* predicted proteins. Group A genes (see text) are denoted by an asterisk.

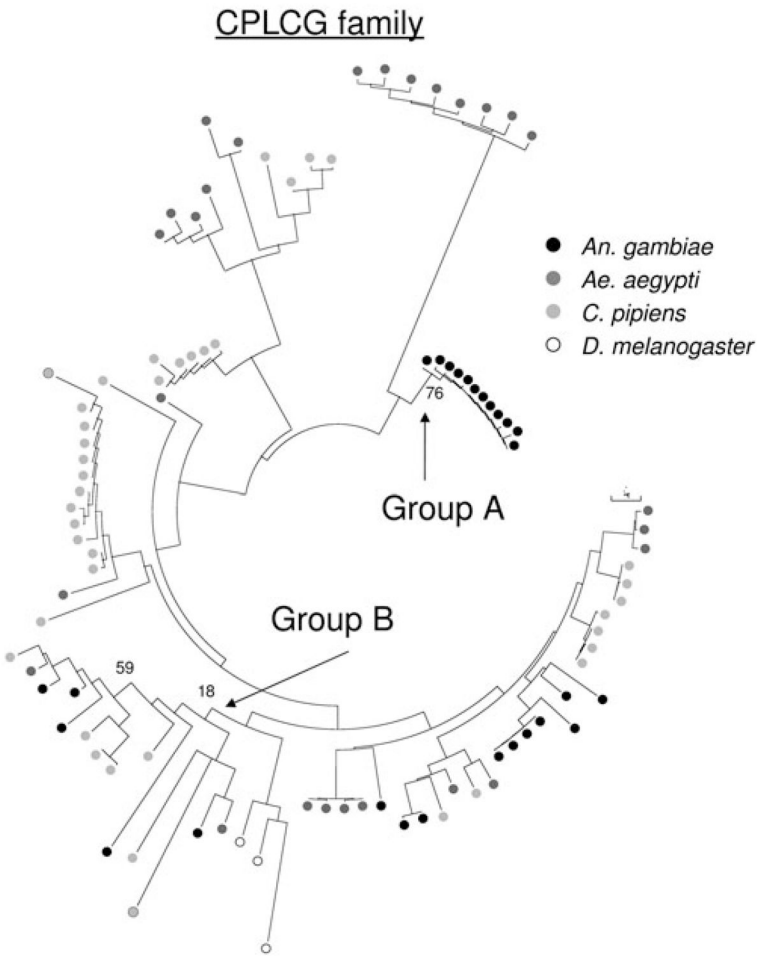

Figure 2.

Neighbor-joining tree of CPLCG amino-acid sequences (where CPLC is ‘cuticular protein of low complexity’), aligned with ClustalW using default parameters. Tree created in Mega4 (Tamura *et al.*, 2007) using the substitution matrix of Jones *et al.* (1992) and pairwise deletion of indels, assuming rate homogeneity. The clades marked ‘Group A’ and ‘Group B’ are discussed in the text. Bootstrap support is shown for select nodes based on 1000 replicates.

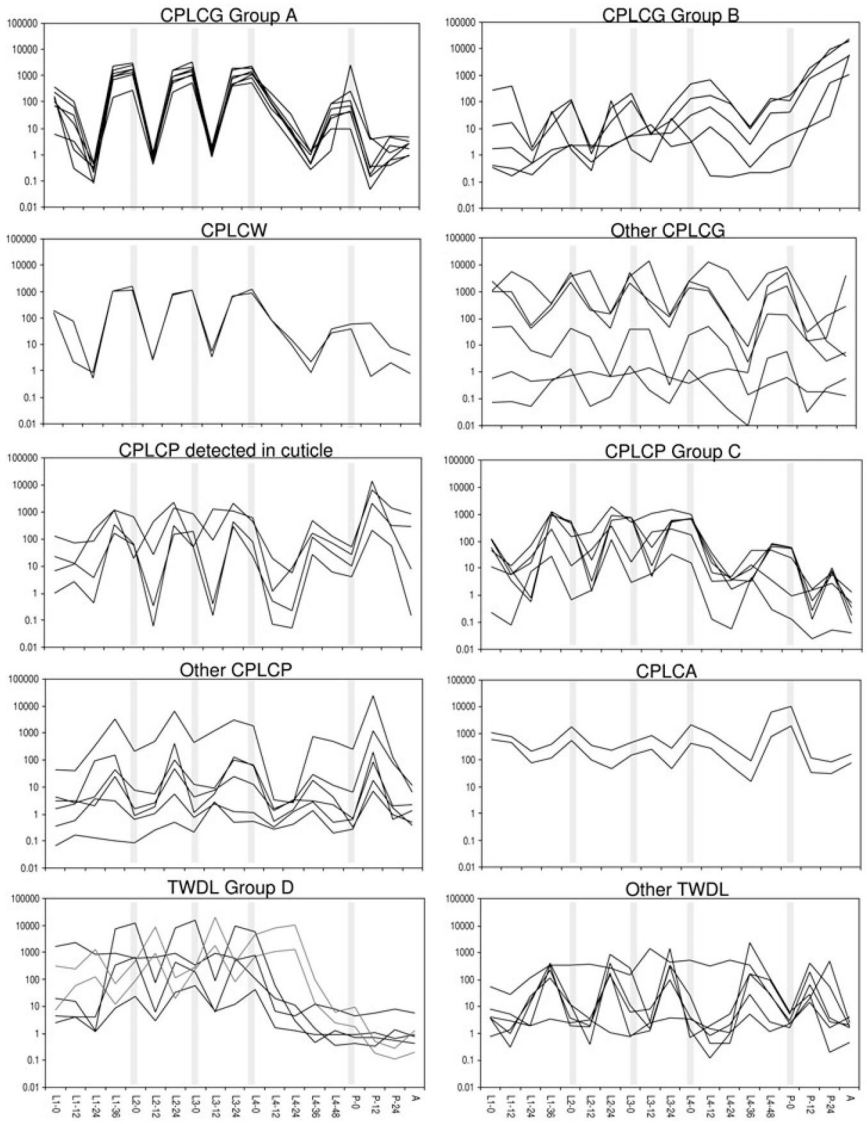

Figure 3.

Expression profiles of *Anopheles gambiae* genes. Each box corresponds to a set of genes described in the text. Transcript abundance measured by quantitative reverse-transcriptase PCR and normalized to total RNA, in arbitrary fluorescence units. The horizontal axis represents the developmental stage and approximate age in hours of animals at each collection point; see Experimental procedures for details. The vertical axis represents raw fluorescence ($R_0 \times 10^7$) and is proportional to initial transcript level. The two genes in the TWDL Group D box that are represented by grey lines each have an intact retrotransposon 3′ of the gene (see text). The grey vertical bars mark the 0-h time point of each life history stage from L2 to P and are intended to aid comparisons amongst graphs.

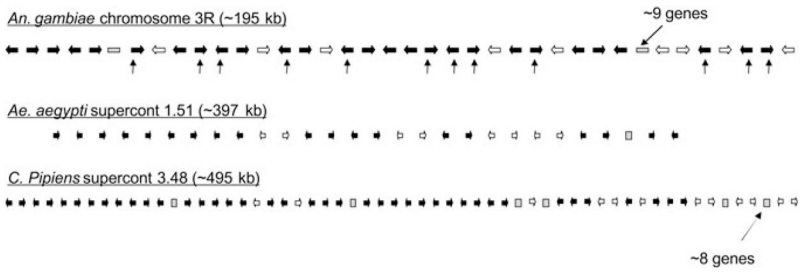

Figure 4.

Schematic of the tandem array of CPLCG and CPLCW genes (where CPLC is 'cuticular protein of low complexity'; black and white arrows, respectively) from three mosquito species, *Anopheles gambiae*, *Culex pipiens*, and *Aedes aegypti*. Directions of arrows indicate coding strand. Grey boxes represent one or more genes that belong to neither gene family (number estimated from Blast search results and existing Ensembl annotations). Vertical arrows identify members of 'Group A' of Fig. 2, in order to highlight their dispersed organization relative to other CPLCG and CPLCW genes. The chromosome or scaffold containing the array and the approximate length of the array are also given.

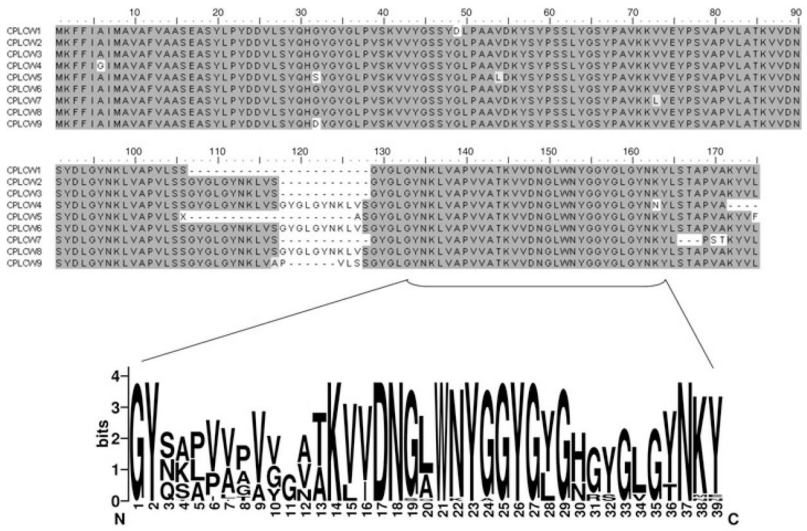

Figure 5.

Similarity-shaded alignment of predicted CPLCW proteins (where CPLC is ‘cuticular protein of low complexity’), performed with ClustalW using default parameters. The sequence logo represents the consensus of the conserved region from all three mosquito species. The CPLCW5 sequence includes an X where a truncating frameshift mutation is present in the PEST strain genome assembly.

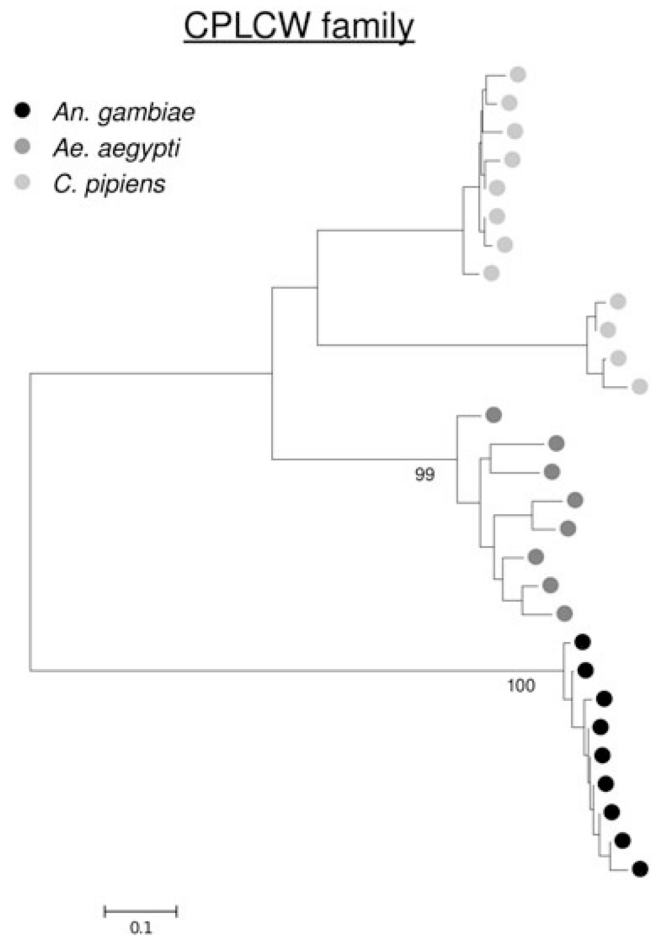

Figure 6.
Neighbor-joining tree of predicted CPLCW proteins (where CPLC is 'cuticular protein of low complexity'), aligned with ClustalW using default parameters. Tree created with Mega4 (Tamura *et al.*, 2007) using the substitution matrix of Jones *et al.* (1992) and pairwise deletion of indels, assuming rate homogeneity. Bootstrap support is shown for select nodes based on 1000 replicates.

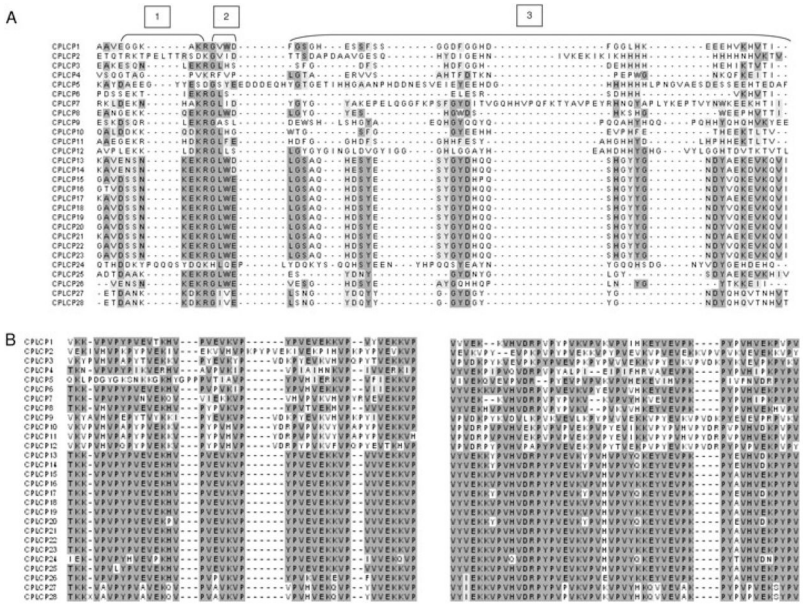

Figure 7.

Similarity-shaded alignments of *Anopheles gambiae* predicted proteins of the CPLCP family (where CPLC is ‘cuticular protein of low complexity’), performed with ClustalW using default parameters. (A) N-terminal region of aligned proteins with signal peptide removed. The numbered regions indicate sequence motifs characteristic of the proposed CPLCP family: (1) polar/acidic region; (2) GLW[D/E] motif; and (3) region rich in glycine, tyrosine, and histidine. (B) Representative blocks of the proline-rich repetitive region.

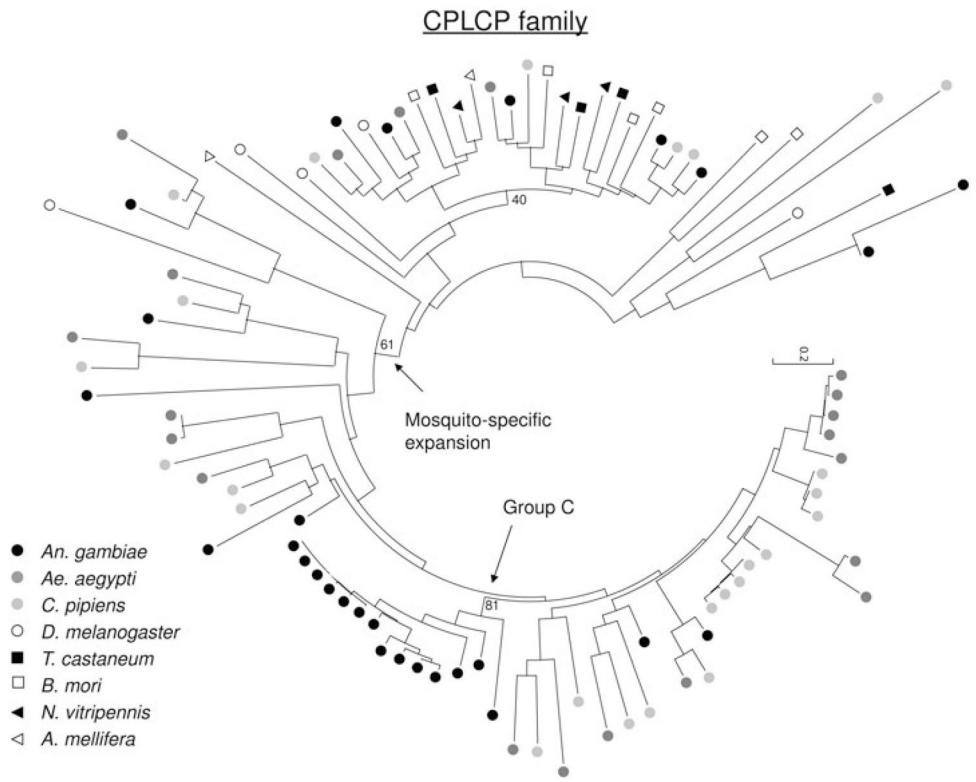

Figure 8.

Neighbor-joining phylogeny of predicted CPLCP proteins (where CPLC is ‘cuticular protein of low complexity’), aligned with ClustalW using default parameters. Tree were created with Mega4 (Tamura *et al.*, 2007) using the substitution matrix of Jones *et al.* (1992) and pairwise deletion of indels, assuming rate homogeneity. Node marked ‘Group C’ is discussed in the text. A mosquito-specific clade is also indicated. Bootstrap support is shown for select nodes based on 1000 replicates.

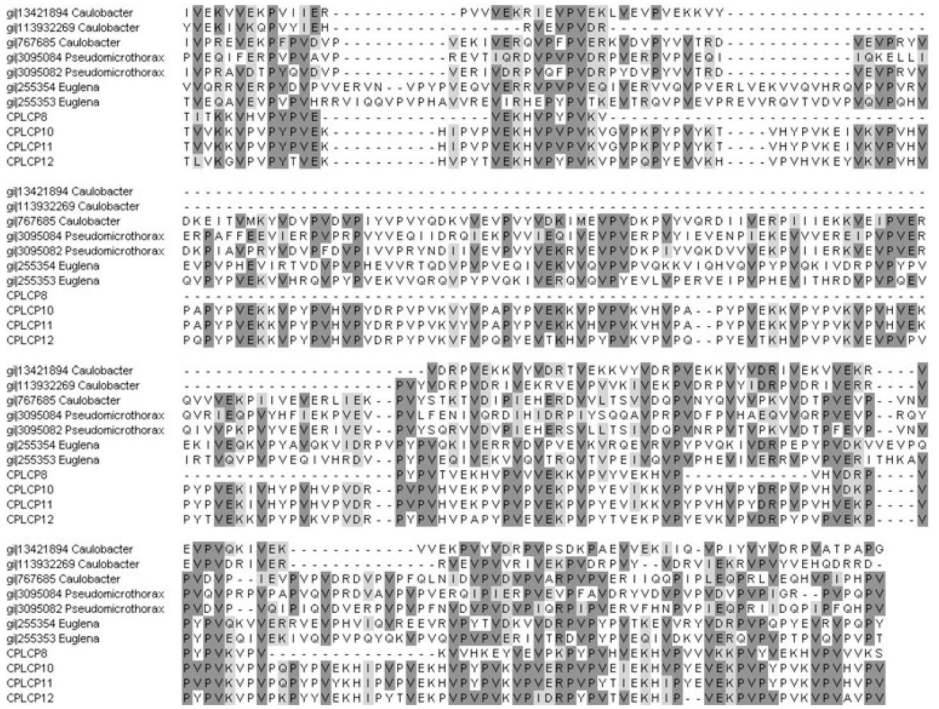

Figure 9.

Amino-acid alignment showing a region of sequence similarity amongst the four *Anopheles gambiae* CPLCP proteins (where CPLC is 'cuticular protein of low complexity') detected in cuticle by He *et al.* (2007) and select articulin proteins from three protist genera. Articulins are intracellular proteins of protists that form ordered, filamentous, cytoskeletal structures, whereas cuticular proteins are extracellular but organized into laminae of ordered filaments. Both families have extensive proline-rich repeats with very similar spacing of proline, valine, and acidic/polar residues. The overall amino-acid compositions of the two groups are highly correlated: the mean pairwise correlation coefficient is 0.98 amongst the CPLCP proteins, 0.93 amongst the articulin proteins, and 0.85 between the two families.

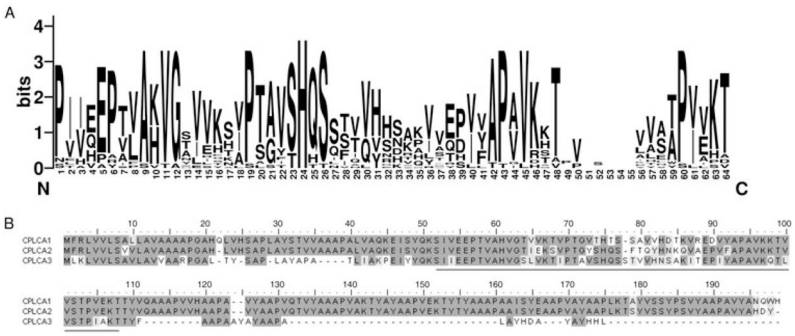

Figure 10.

Sequence features of the CPLCA family (where CPLC is ‘cuticular protein of low complexity’). (A) A sequence logo represents sequence conservation amongst all homologues in mosquitoes and *Drosophila*. The retinin domain alignment, Pfam04527/IPR007614, is approximately equal to positions 2 to 64 of the sequence logo. (B) Similarity-shaded alignment of the predicted proteins of the *Anopheles gambiae* CPLCA family. The region corresponding to the sequence logo is underlined.

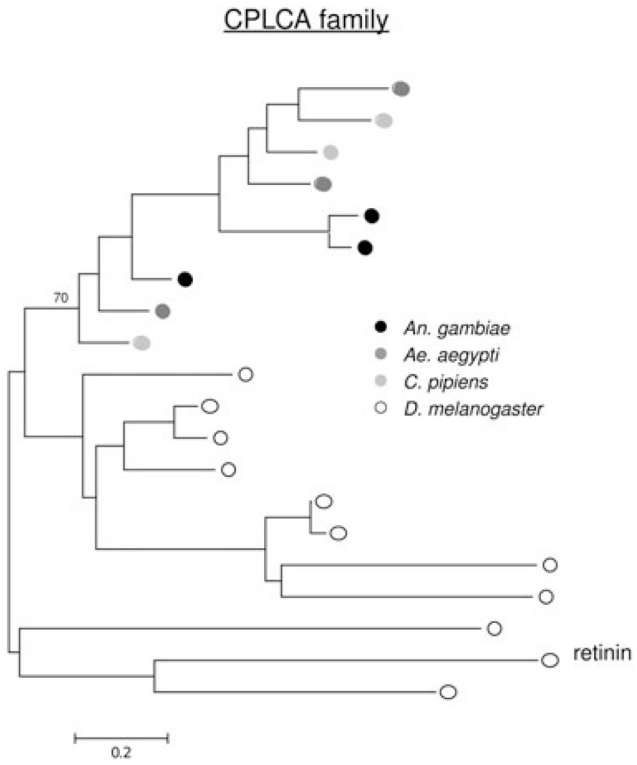

Figure 11.
Neighbor-joining phylogeny based on predicted protein sequence of the CPLCA gene family (where CPLC is 'cuticular protein of low complexity'). Sequences were aligned with ClustalW using default parameters. Trees were created with Mega4 (Tamura *et al.*, 2007) using the substitution matrix of Jones *et al.* (1992) and pairwise deletion of indels, assuming rate homogeneity. Bootstrap support is shown for select nodes based on 1000 replicates.

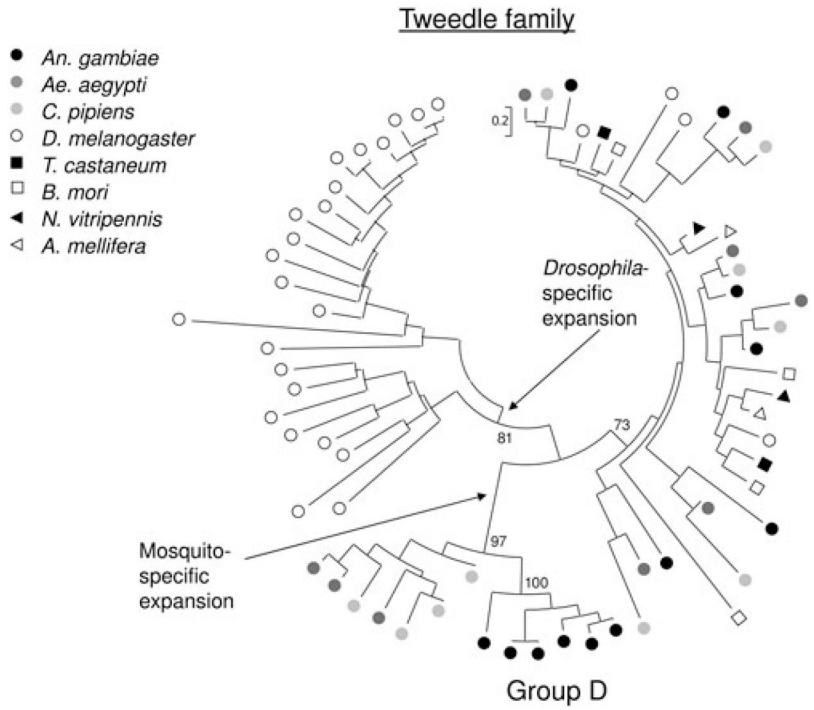

Figure 12.

Neighbor-joining phylogeny based on predicted protein sequence of the Tweedle gene family. *Drosophila* annotations are from Ensembl, and *Anopheles gambiae* sequences are as annotated in this paper. Partial or complete coding sequences from other insect genomes were inferred from homology searches. Sequences were aligned with ClustalW using default parameters. Tree were created with Mega4 (Tamura *et al.*, 2007) using the substitution matrix of Jones *et al.* (1992) and pairwise deletion of indels, assuming rate homogeneity. *Drosophila*- and mosquito-specific expansions are indicated, and 'Group D' is discussed in the text. Bootstrap support is shown for select nodes based on 1000 replicates.

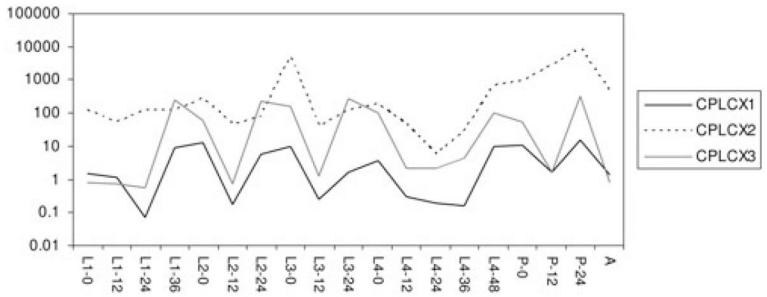

Figure 13.

Expression profiles of three genes encoding *Anopheles gambiae* cuticular protein genes that have not been grouped into families of homologous genes and are designated *CPLCX1–CPLCX3* (where CPLC is 'cuticular protein of low complexity'). Transcript abundance measured by quantitative reverse-transcriptase PCR and normalized to total RNA, in arbitrary fluorescence units. The horizontal axis represents the developmental stage and approximate age of animals at each collection point; see Experimental procedures for details. Note logarithmic scale of vertical axis.

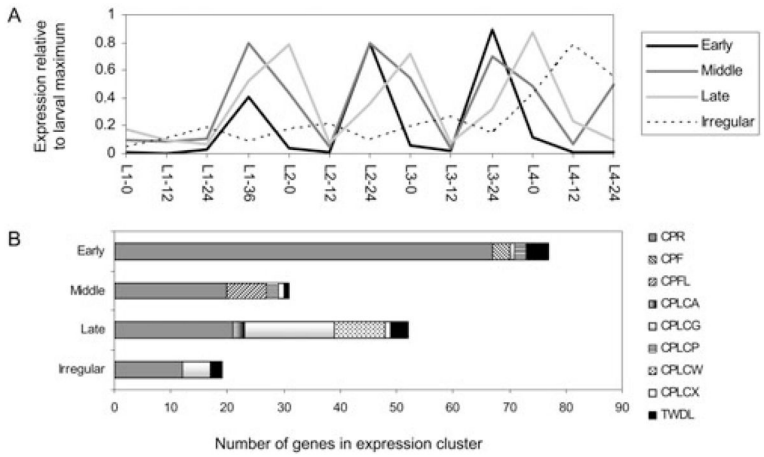

Figure 14.

Clusters of genes with similar expression patterns during larval development as determined by self-organizing map. (A) Mean expression within each cluster expressed as the proportion of maximum expression during larval development. (B) Number of genes in each cluster by gene family; see Supporting Information Table S4 for gene names.