Insect Biochemistry and Molecular Biology 38 (2008) 661–676

Contents lists available at ScienceDirect

Insect Biochemistry and Molecular Biology

journal homepage: www.elsevier.com/locate/ibmb

Extensive gene amplification and concerted evolution within the CPR family of cuticular proteins in mosquitoes

R. Scott Cornman *, Judith H. Willis

Department of Cellular Biology, University of Georgia, Athens, GA 30602, USA

ARTICLE INFO

Article history:
Received 26 February 2008
Received in revised form
27 March 2008
Accepted 3 April 2008

Keywords:
Anopheles gambiae
Aedes aegypti
Drosophila melanogaster
Cuticular protein
Concerted evolution
CPR family
Rebers and Riddiford Consensus

ABSTRACT

Annotation of the *Anopheles gambiae* genome has revealed a large increase in the number of genes encoding cuticular proteins with the Rebers and Riddiford Consensus (the CPR gene family) relative to *Drosophila melanogaster*. This increase reflects an expansion of the RR-2 group of CPR genes, particularly the amplification of sets of highly similar paralogs. Patterns of nucleotide variation indicate that extensive concerted evolution is occurring within these clusters. The pattern of concerted evolution is complex, however, as sequence similarity within clusters is uncorrelated with gene order and orientation, and no comparable clusters occur within similarly compact arrays of the RR-1 group in mosquitoes or in either group in *D. melanogaster*. The dearth of pseudogenes suggests that sequence clusters are maintained by selection for high gene-copy number, perhaps due to selection for high expression rates. This hypothesis is consistent with the apparently parallel evolution of compact gene architectures within sequence clusters relative to single-copy genes. We show that RR-2 proteins from sequence-cluster genes have complex repeats and extreme amino-acid compositions relative to single-copy CPR proteins in *An. gambiae*, and that the amino-acid composition of the N-terminal and C-terminal sequence flanking the chitin-binding consensus region evolves in a correlated fashion.

1. Introduction

Proteins with the Rebers and Riddiford Consensus ('R&R Consensus' hereafter; Rebers and Riddiford, 1998) are a major component of insect cuticle (Andersen et al., 1995; Willis et al., 2005). These proteins, called CPR proteins, include two major evolutionary groups defined by variants of the R&R Consensus (RR-1 and RR-2). The RR-2 consensus sequence is approximately 64 amino acids long and is well conserved, whereas the RR-1 consensus is more variable in length and sequence. The two variants are readily distinguished by Hidden Markov Models (Karouzou et al., 2007). Although examples of both variants have been shown to bind chitin *in vitro* (Rebers and Willis, 2001; Togawa et al., 2004), Rebers and Willis (2001) also showed that the affinity of the R&R Consensus to chitin may be disrupted *in vitro* by non-conservative substitutions at particular positions. Such substitutions are actually present in known CPR proteins and it is reasonable to postulate that CPR proteins differ in their affinity to chitin under physiological conditions. Nonetheless, the vast majority of *Anopheles gambiae* CPR proteins have been detected in cuticle by proteomic analyses, including those with potentially disrupted chitin affinity (He et al., 2007; He, personal communication). Thus, the presence of CPR proteins within cuticular structures appears to be a general feature of this gene family.

It is therefore remarkable that recent characterizations of the CPR gene family within the genomes of *Drosophila melanogaster* (Karouzou et al., 2007) and *An. gambiae* (Cornman et al., 2008; Togawa et al., 2008) have confirmed a large diversity of expressed CPR genes and identified very few pseudogenes. The complement of RR-2 genes is particularly large in mosquitoes. The PEST strain of *An. gambiae* has 101 annotated RR-2 CPR genes whereas only 35 were identified in *D. melanogaster*. Comparing the phylogeny and genomic organization of RR-2 genes in *An. gambiae* (Cornman et al., 2008) and *D. melanogaster* (Karouzou et al., 2007), it is clear that much of the difference in RR-2 gene number is due to the presence in *An. gambiae* of sets of highly similar genes that are often but not exclusively arranged in complex clusters within larger tandem arrays (Cornman et al., 2008) and which are co-expressed (Togawa et al., 2008).

Here we identify orthologous genes and syntenic regions between *An. gambiae* and both *Ae. aegypti* and *D. melanogaster* in order to investigate the origins and evolution of these RR-2 'sequence clusters.' We further investigate whether sequence clusters are evolving in a non-independent manner through some combination of unequal crossing-over and intergenic gene conversion. This process is commonly referred to as concerted

* Corresponding author.
E-mail addresses: scornman@uga.edu (R.S. Cornman), jhwillis@cb.uga.edu (J.H. Willis).

0965-1748/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ibmb.2008.04.001

evolution and is a potentially important component of multigene family evolution (Ohta, 1983; Walsh, 1987; Innan, 2003). Nei and Rooney (2005) have argued that the contribution of concerted evolution to the maintenance of highly similar paralogs is generally weak; they posit that strong purifying selection and a background level of gene turnover are sufficient to explain most such cases (the 'birth-and-death' model). Nonetheless, gene conversion events among paralogs have been identified in a diverse array of taxa (Drouin, 2002a; Teshima and Innan, 2004) and appear to be relatively frequent at the genome scale in mice and rats (Ezawa et al., 2006), rice (Wang et al., 2007), and yeast (Drouin, 2002b) but less so in nematodes (Semple and Wolfe, 1999). Drouin et al. (1999) and Drouin (2002a) evaluated different approaches for detecting gene conversion in the absence of population-genetic data, i.e., when few alleles for each gene-family member are known from a given species (as is often the case). They found that several methods are effective and appear to give low rates of false-positives. We utilize several of these approaches to show that concerted evolution is more important than birth-and-death evolution in maintaining RR-2 sequence clusters in *An. gambiae*. Finally, to gain insight into the possible functional significance of these 'sequence clusters,' we compared properties of these genes and their encoded proteins to 'single copy' RR-2 genes and to RR-1 genes.

2. Methods

2.1. Phylogeny, organization, and properties of RR-2 genes

In *An. gambiae*, some CPR genes occur in isolation, but most occur in tandem arrays in which genes are usually spaced a few kb apart and not more than 20 kb based on our operational cutoff. Many of these tandem arrays include sets of genes that are alignable across the length of the protein, which may be due to recent duplication, purifying selection, and/or gene conversion (see Sections 3 and 4). In this paper, we use the term 'sequence cluster' in contradistinction to 'tandem array' to denote a set of highly similar paralogs that are often but not exclusively contiguous. Following Koonin (2005), we use 'co-orthologous region' to describe a chromosomal region in two species that is syntenic and contains putatively orthologous genes although the orthology of individual genes is uncertain due to the differential gain or loss of genes or due to gene-conversion events. In this study, *An. gambiae* sequence clusters are named by their order on chromosomes, e.g., 2LA, 2LB, and 2LC are the three sequence clusters on chromosome 2L ordered by Ensembl coordinates (see Cornman et al. (2008) for details).

Most *An. gambiae* CPR genes are available from Ensembl (http://www.ensembl.org/index.html) and all their conceptual translations are available from the cuticleDB database (http://bioinformatics.biol.uoa.gr/cuticleDB/; Magkrioti et al., 2004). Cornman et al. (2008) provide contig coordinates for all *An. gambiae* CPR genes as well as supporting evidence for their annotation. *Ae. aegypti* genes were obtained by examining all gene predictions of Ensembl v. 40 that contained the Pfam00379 domain (http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF00379), which defines the extended R&R Consensus sequence, and identifying putative RR-2 genes (Supplementary File 1). We corrected 26 gene predictions for annotation errors that were evident based on alignment with other mosquito RR-2 genes (Supplementary File 2). We removed two predicted genes of *Ae. aegypti* (AAEL014925 and AAEL011037) from the analysis because of major annotation concerns that could not be resolved by inspection. Six *Ae. aegypti* CPR genes not annotated by Ensembl v. 40 were identified by BLAST searches or dot-plots (listed in Supplementary File 3 and named *AeCPR A-F*). Our confidence in the annotated *An. gambiae* RR-2 genes allows us to confirm the overall accuracy of the *Ae. aegypti* RR-2 genes used in this study with respect to the consensus region, despite uncertainties regarding the 5′ boundaries of some genes and whether nearly identical genes on different contigs are distinct genes. We included all such duplicates in this study because the intergenic regions were not similar when examined by dot-plot, which shows that these putative paralogs are unlikely to be alleles of a single locus. Exclusion of these sequences did not qualitatively change the results of any analysis. The annotation of the CPR family in *D. melanogaster* has been recently updated (Karouzou et al., 2007) and the sequences are available from Flybase (http://www.flybase.org) and the cuticleDB website given above.

We developed a phylogeny of all RR-2 genes of the three species based on the Pfam00379 domain alignment, but excluding the first five positions of that alignment because they do not align well across all three species. We used amino-acid sequence rather than nucleotide sequence because there is clear evidence of mutational saturation (pairwise synonymous substitution rates $\gg 1$, see Cornman et al. (2008)). A neighbor-joining tree with 1000 bootstrap replicates was generated in MEGA3 (Kumar et al., 2004) using the JTT cost matrix (Jones et al., 1992) and performing pairwise deletion of indels. We assumed a gamma distribution of rates with $\beta = 1.0$, as an evolutionary model with rate variation among sites gives a significantly better fit to the data than does a single ratio of non-synonymous to synonymous substitutions (Ka/Ks) for *An. gambiae* RR-2 genes (Cornman et al., 2008).

We identified all probable orthologs of *An. gambiae* RR-2 genes in *Ae. aegypti* and *D. melanogaster*. While the R&R Consensus is strongly conserved, flanking regions are often of low complexity and typically are not alignable among paralogs. However, these flanking regions are conserved among orthologs in *An. gambiae* and *Ae. aegypti* and often share sequence motifs within tandem arrays that aid in the identification of syntenic chromosomal regions in *D. melanogaster* (Cornman et al., 2008). We therefore used both consensus-region phylogenetic analysis and reciprocal BLAST scores of the full protein sequence, including the signal peptide, to identify putatively orthologous genes or co-orthologous regions. Importantly, conserved orthologous single-copy genes (whether CPR genes or otherwise) serve as genomic landmarks that provide strong support for the putative co-orthology of sequence clusters in the two species (see Section 3).

For two single-copy *An. gambiae* genes, *CPR59* and *CPR68*, we identified two potential orthologs each in *Ae. aegypti* on different contigs; Ka and Ks estimates between *An. gambiae* and *Ae. aegypti* were obtained by averaging the values for the two potential orthologs. Three other *An. gambiae* genes, *CPR47*, *CPR63*, and *CPR156* were ambiguous as to whether they belonged in sequence clusters and their placement in the phylogenetic tree was sensitive to changes in phylogenetic method. *CPR47* and *CPR156* are physically adjacent to the sequence clusters in question and *CPR47* showed evidence of non-orthologous recombination with other sequences in 2LB (assessed with the RDP2 program (Martin et al., 2005)), but this was not true of *CPR156* nor *CPR63*. We conservatively included *CPR47* in the 2LB cluster and *CPR156* in the 3RB cluster but excluded *CPR63* from 2LC, although these placements do not materially affect our results.

2.2. Molecular evolution of sequence-cluster genes

To infer intergenic conversion as a mode of molecular evolution within sequence clusters, we used three approaches. We first examined patterns of synonymous polymorphisms and intron polymorphism to determine whether purifying selection or