Published in final edited form as:

Insect Biochem Mol Biol. 2008 May ; 38(5): 508–519.

Developmental expression patterns of cuticular protein genes with the R&R Consensus from *Anopheles gambiae*

Toru Togawa, W. Augustine Dunn, Aaron C. Emmons, John Nagao, and Judith H. Willis^{*}
Department of Cellular Biology, University of Georgia, Athens GA 30602, USA

Abstract

CPR proteins are the largest cuticular protein family in arthropods. The whole genome sequence of *Anopheles gambiae* revealed 156 genes that code for proteins with the R&R Consensus and named CPRs. This protein family can be divided into RR-1 and RR-2 subgroups, postulated to contribute to different regions of the cuticle. We determined the temporal expression patterns of these genes throughout post-embryonic development by means of real-time qRT-PCR. Based on expression profiles, these genes were grouped into 21 clusters. Most of the genes were expressed with sharp peaks at single or multiple periods associated with molting. Genes coding for RR-1 and RR-2 proteins were found together in several co-expression clusters. Twenty-five genes were expressed exclusively at one metamorphic stage. Five out of six X-linked genes showed equal expression in males and females, supporting the presence of a gene dosage compensation system in *An. gambiae*. Many RR-2 genes are organized into sequence clusters whose members are extremely similar to each other and generally closely associated on a chromosome. Most genes in each sequence cluster are expressed with the same temporal expression pattern and at the same level, suggesting a shared mechanism to regulate their expression.

Keywords

Cuticle; Cuticular protein; CPR; R&R Consensus; Real-time qRT-PCR; Gene cluster; Gene dosage compensation

1. Introduction

Arthropod cuticle functions as the exoskeleton, which maintains the body structure, inhibits the evaporation of water and serves as a barrier to the environment. Insect cuticle is mainly composed of the polysaccharide chitin and a few groups of cuticular proteins (for review see Andersen et al., 1995; Willis et al., 2005). Hundreds of cuticular proteins have now been identified from numerous insect species and several other arthropods. Their sequences are available at the cuticleDB website (<http://bioinformatics2.biol.uoa.gr/cuticleDB/index.jsp>) (Magkrioti et al., 2004). The largest cuticular protein family is defined by the presence of a conserved domain called the “R&R Consensus” first recognized by Rebers and Riddiford (1988). An extended version of the consensus was subsequently described and we, like most others, now refer to it as the R&R Consensus; it is also recognized as pfam00379. Proteins with the R&R Consensus have been classified as belonging to the CPR protein family. This family has two distinct groups, RR-1 and RR-2 (Andersen, 1998; Karouzou et al., 2007). CPRs

^{*}Corresponding author: Judith H. Willis, Tel: +1-706-542-0802; Fax: +1-706-542-4271, E-mail: jhwillis@cb.uga.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

with the RR-1 type domain have been attributed to soft (flexible) cuticles, while RR-2 proteins have been associated with rigid (hard) cuticle, although this classification is tentative (Andersen, 2000; Willis et al., 2005). A third group, RR-3, was also proposed (Andersen, 2000), but a precise definition has not been established (Karouzou et al., 2007). Predictions that the R&R Consensus serves to bind to chitin have been supported in various ways. The secondary structure of the R&R Consensus was proposed and experimentally analyzed; homology models for the tertiary structure exist (Iconomidou et al., 1999; Iconomidou et al., 2001; Hamodrakas et al., 2002; Iconomidou et al., 2005). Biochemical analyses using recombinant fusion proteins with the R&R Consensus confirmed that it can function as a chitin-binding domain (Rebers and Willis, 2001; Togawa et al., 2004).

We have annotated 156 CPR genes from the whole genome sequence of *Anopheles gambiae* (Cornman et al., 2007). The CPR genes are found as singletons and in tandem arrays of linked genes. Genes coding for RR-1 and RR-2 proteins do not co-exist in the same tandem array. Within tandem arrays there are also sequence clusters of genes that code for very similar RR-2 proteins (Cornman et al., 2007). Members of a sequence cluster are generally sequentially arranged on a chromosome, but a few were found at a distance from the main cluster with other less closely related RR-2 genes intervening.

Proteomics analyses by MS/MS of various cuticle preparations identified peptides, either unique or shared among similar proteins, for 93% of the 156 CPR proteins, validating their authenticity as cuticular proteins (Cornman et al., 2007; He et al., 2007). The present study was designed to answer the following questions that might help us to understand why *An. gambiae* devotes over 1% of its genes to coding for CPR proteins: i) Is each of the 156 CPR genes actually expressed? ii) When is each gene expressed and are there similar patterns of expression of genes linked in tandem arrays and/or sequence clusters? iii) Do any of the CPR genes show stage-specific expression? iv) Do any of the CPR genes show sex-specific expression? v) Do levels of expression provide clues about the evolution/maintenance of sequence clusters?

We have analyzed the developmental expression patterns of all *An. gambiae* CPR genes that were detected in the G3 strain. While some would suggest using microarrays for this purpose, we were concerned that this technique would not be suitable for groups of genes that are so similar to each other, because it would be difficult/impossible to design gene-specific oligonucleotides to use under chip hybridization conditions. Therefore, we adopted quantitative real-time RT-PCR (qRT-PCR) that allows more stringent criteria to assure that each primer pair was indeed gene-specific.

2. Materials and methods

2.1. Description of gene and protein sequences

All of the genes discussed in this paper, their location on contigs, and corresponding proteins are available at the website (http://may2005.archive.ensembl.org/Anopheles_gambiae/submission). Protein sequences are also available at cuticleDB (<http://bioinformatics2.biol.uoa.gr/cuticleDB/index.jsp>). A detailed discussion of annotation and similarity among sequences can be found in Cornman et al. (2007). A summary of the genes arranged in chromosomal order and showing tandem arrays, sequence clusters, and VectorBase names is given in Supplementary Table 1.

2.2. Mosquito sampling, RNA isolation, and reverse transcription

The G3 strain of *Anopheles gambiae* was used even though the genomic sequence data come from the PEST strain; the PEST strain no longer exists. Mosquitoes were reared at 27°C under

a 14 h light/10 h dark cycle. Total RNA was isolated from developmentally synchronized animal as described previously (Togawa et al., 2007). In brief, newly ecdysed animals were collected at each molt, reared in small groups, sampled at 12 h intervals, and the RNA was isolated with Trizol (Invitrogen). RNA isolation was followed by removal of contaminating DNA with TURBO DNA-free (Ambion). The concentration of RNA was measured with RiboGreen RNA Quantitation Reagent (Molecular Probes) on the Bio-Rad iCycler. First strand cDNA was synthesized from 3 µg of total RNA with SuperScript III (Invitrogen) using an oligo (dT)₁₂₋₁₈ primer in a 20 µl reaction. In order to confirm that there was no contamination of genomic DNA in the cDNA preparations, we carried out RT-PCR for genes for ribosomal protein S7 (*RpS7*, AGAP010592; Salazar et al., 1993) and the chitin synthase that codes for the protein involved in epidermal chitin synthesis (AGAP001748, called *CHS2* in Arakane et al., 2004). For each gene we used primers that span a small intron. No cDNA sample showed amplification of a fragment with the intron.

2.3. Primer design

It was essential to design gene specific primer pairs for qRT-PCR in order to reveal expression patterns for each individual gene. Although genes in a sequence cluster often have almost identical coding sequences, the 5' and 3' UTRs are usually unique. Therefore most primers were designed in the 3' UTR or around the stop codon. Primers were designed using Primer3 software (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi) (Rozen and Skaletsky, 2000) to yield products between 50 and 150 bp. We checked the validity of each primer pair in terms of gene specificity and amplification efficiency. In order to confirm the specificity of primer pairs, we used three tests: i) a sharp melting curve of the PCR product, ii) a single band of the product on an agarose gel, and iii) amplification kinetics against genomic DNA that confirmed that only a single gene was being amplified.

For this final test of gene copy number, real-time PCR was performed for target genes and known single copy genes using genomic DNA as template. Gene copy number relative to the single copy genes was calculated with following equation (Liu and Saint, 2002):

$$R_{0,T}/R_{0,R} = (1 + E)^{\Delta C_T}$$

where $R_{0,T}$ and $R_{0,R}$ are initial fluorescence representing the amounts of target and reference gene sequences in the template DNA samples, respectively; ΔC_T is equal to $C_{T,R} - C_{T,T}$ ($C_{T,R}$ and $C_{T,T}$ are threshold cycles of reference and target genes); E is amplification efficiency of PCR (set at a default value of 1.0). As single copy gene standards, *RpS7* and chitin synthase were used. We accepted primer pairs whose gene copy number relative to these reference genes was > 0.5 and < 2.0 . The amplification efficiency of primer pairs was determined from the slope of the curve generated by amplification from serially diluted genomic DNA with the following equation: Efficiency = $[10^{(-1/\text{slope})}] - 1$ (MyiQ Single-Color Real-Time PCR Detection System Instruction Manual, Bio-Rad; Lekanne Deprez et al., 2002). Efficiency had to be at least 0.9 for a primer pair to be accepted. Gene specific primer pairs which passed the criteria above, with a few slight exceptions (Supplementary Table 2), were made for 147 out of the 156 CPR genes. For the other nine genes, primer pairs which amplified two or three genes were used. In total, 153 primer pairs were used. The information for primer pairs used in each reaction and the actual primer sequences are listed in Supplementary Tables 2 and 3.

2.4. Real-time qRT-PCR

qRT-PCR was performed with Bio-Rad's iCycler and MyiQ Single-Color Real-Time PCR Detection System. All reactions were carried out in triplicate in a 20 µl reaction containing 5 µl of 1/100 diluted cDNAs (equivalent to starting with 7.5 ng of total RNA), 500 nM of each primer, and 1X iQ SYBR Green Supermix (Bio-Rad). PCR conditions were 95°C for 3 min

followed by 40 cycles of 95°C for 15 sec and 57°C for 1 min. After the PCR reactions were complete, melt curve analyses were done. Each set of triplicate measurements was carried out on two different cDNA samples, made from RNA isolated from animals collected on different occasions. In cases where there appeared to be sex-specific differences in expression, experiments were done with three different starting RNA samples. We made multiple preparations of cDNA from the same starting RNA. The transcript level of *RpS7* was measured on every run, and the results show that the multiplicity of cDNA preparations did not affect our conclusion (Supplementary Fig. 1).

2.5. Data analyses

Although *RpS7* is often used as a normalizer in gene expression studies of *An. gambiae* (Richman et al., 1997; Vizioli et al., 2001; Nikou et al., 2003), we found that expression of this gene fluctuated across the array of developmental stages we tested (Supplementary Fig. 2). We also tested genes for ribosomal protein L32 (AGAP002122), elongation factor 2 (AGAP009441), and the ubiquitin-ribosomal protein L40 fusion protein (AGAP007927), however, we were unable to identify a suitable “housekeeping gene” to normalize our data across the developmental stages we examined. Therefore, we adopted very careful quantification of RNA to standardize our samples. The initial amount of cDNA was determined as R_0 calculated by following equation (Livak and Schmittgen, 2001; Liu and Saint, 2002):

$$R_0 = \frac{R_{C_T}}{(1 + E)^{C_T}}$$

where R_0 is initial fluorescence representing target cDNA quantity; C_T is threshold cycle; R_{C_T} is fluorescence at threshold cycle; E is amplification efficiency (set at a default value of 1.0). The threshold was determined automatically with the Bio-Rad software. The mean R_0 from two or three cDNA samples (biological replicates) for each time point was used.

Expression profiles were grouped by self-organizing maps (SOM) (Tamayo et al., 1999) with GenePattern software (Reich et al., 2006). This algorithm is often used to summarize microarray data. We obtained expression profiles in terms of relative expression rather than absolute expression levels by dividing the R_0 of each developmental time point for each gene by the sum of R_0 throughout development for that gene.

3. Results

3.1. How many CPR genes are expressed in the G3 strain?

The primers we designed allowed us to obtain data for 147 CPR genes with gene specific primer pairs. The other nine genes were analyzed with six primer pairs that recognize multiple genes (Supplementary Table 2). Careful analyses of gene copy number (qPCR on genomic DNA) and expression data indicated that our primers designed based on PEST strain sequences did not detect four genes in the G3 strain we used. These are *CPR65*, *CPR91*, *CPR99* or *142*, *CPR133* or *153*; see Supplementary Table 2 – comments). For the last two pairs of genes, we had evidence for only one gene from each pair. These four genes (*CPR65*, *CPR91*, and one of each pair) might be missing in the G3 strain, or they were not detected because of strain differences such as SNPs and indels. We arbitrarily chose to use the names *CPR99* and *CPR133* to represent their respective pair. For *CPR94*, we obtained an unambiguous expression pattern because we could subtract the data with specific primers from the data with multi-gene primers. For *CPR99*, the same procedure produced data that would fit this gene. The expression data of two other genes (*CPR12* and *CPR13*) analyzed with common primer pairs did not reveal the precise expression of each single gene. Gene copy number analysis revealed that the common primer set was amplifying two genes. Thus the data presented in this paper provide

developmental expression information for 152/156 CPR genes that had been annotated in the PEST strain of *An. gambiae*.

3.2. Developmental expression pattern and expression magnitude of CPR genes

The expression profiles of each of the 152 CPR genes are given in Supplementary Fig. 3. These expression profiles were grouped into 21 co-expression clusters by SOM (Fig. 1). Asking the program to select more clusters did not reveal any qualitatively different clusters, and designating fewer clusters lost unique clusters. Because only six genes had different expression between sexes (described below), the average of male and female data was used for this analysis. Even though CPR genes are expressed with various patterns as shown in Fig. 1A, these patterns generally share some characteristics. First, half of these clusters showed very sharp peak(s) at one or multiple stages linked to ecdysis. These are clusters 0, 5, 6, 7, 12, 14, 17, 18, 19, and 20. This indicates that expression of these genes is strictly regulated and the transcripts are rapidly turned over. The other clusters had broader expression peaks, showing that these genes are transcribed over a longer span or their transcripts are more stable. Second, based on the assumption that translation/secretion occurs without a lag after transcription, the co-expression clusters can be divided into three functional groups. Some genes are expressed exclusively in pharate stages (clusters 0, 1, 5, 6, 7, 8, 12, 13, 14, 17, 18, 19, and 20), indicating that these products contribute solely to the pre-ecdysial exocuticle. For others, expression begins in the pharate stage and mRNA is present after ecdysis (clusters 2, 3, 10, 11, and 15). Genes in a few co-expression clusters are only expressed after ecdysis (clusters 4, 9, and 16), suggesting the exclusive contribution of their products to the post-ecdysial endocuticle.

The CPR transcript levels ranged from $R_0 \times 10^7 = 5$ (*CPR155*) to 46,600 (*CPR5*). The lower level appears to be authentic because we verified the very low expression levels of two CPR genes, *CPR112* and *CPR155* whose maximum levels were $R_0 \times 10^7 = 6$ and 5, respectively (level 1 in Fig. 1B). In order to confirm that this was not from genomic DNA contamination or another artifact, we performed real-time PCR with RT minus control samples, which were prepared exactly the same as cDNA preparation but with water instead of reverse transcriptase. When *CPR112* and *CPR155* were analyzed on these RT minus controls, no consistent amplification was observed (data not shown). Hence among CPR genes we have a dynamic range of expression of approximately 4 orders of magnitude. The range for a single gene can be even greater since several genes were highly expressed at one stage ($R_0 \times 10^7$ was greater than 10,000) and had no expression at other stages. Another way to appreciate the massive expression of some CPR genes is to relate it to the expression of *RpS7* ($R_0 \times 10^7$ is 626 to 4,950; Supplementary Fig. 2). Genes for ribosomal proteins are generally considered to be highly expressed. Over a third of the CPR genes had 2-fold higher expression than *RpS7* in at least one developmental stage, and 19 were 10-fold higher.

3.3. Stage specificity of CPR genes

An accurate assessment of metamorphic stage requires that we include the pharate stage (period when the cuticle of the next stage is being deposited) along with the post-ecdysial form of that stage. Hence in our designation of a metamorphic stage we added the last two data points of the fourth larval instar, 36 and 48 h after ecdysis, to the pupal stage, and animals aged 12 and 24 h after pupation were included with adults. When this is done, it appears that many genes are predominantly expressed in only one metamorphic stage (co-expression clusters 0, 1, 2, 4, 5, 6, 9, 11, 13, 14, 15, 16, and 20). In order to learn whether there are any CPR genes expressed exclusively at a specific stage, we examined the data to find genes whose expression levels at a particular stage were at least two orders of magnitude higher than at the other two metamorphic stages. We also required that the R_0 of stages with low expression was not higher than 10^{-6} , because that was the maximum R_0 of *CPR112* and *CPR155*, whose maximum

expression was the least of all CPR genes (Fig. 1B). As shown in Table 1, by these criteria, expression of 25 genes was specific to a single metamorphic stage.

3.4. Sexually different expression of CPR genes

Given that there are many sex-specific cuticular structures and that post-hematophagy cuticle expansion occurs in females but not in the non-hematophagous males, we investigated sex-specific expression of CPR genes in pupae and pharate and young adults. Only the six genes shown in Fig. 2 showed significantly different expression in the two sexes. Two of these (*CPR106* and *CPR148*), were predominantly expressed in larvae, suggesting that their sex specific contributions to adult female cuticle are unlikely to be important. *CPR152* had clear male-specific expression, albeit at low levels, suggesting it may contribute to male-specific cuticle structures such as male genitalia and antennae.

In *An. gambiae*, the male is the heterogametic sex, possessing X and Y chromosomes. Six CPR genes are coded on the X chromosome of *An. gambiae* (*CPR125-CPR130*). Only *CPR125* is expressed at different levels between the two sexes and just after pupation. At other periods, sex differences were not seen. The other five genes showed comparable expression in both sexes. It is not clear whether *An. gambiae* has a dosage compensation system, even though it is implied by the presence of orthologous genes involved in the system (Zdobnov et al., 2002). Our data on these six X-linked genes supports the presence of dosage compensation in *An. gambiae*.

3.5. Types of genes in co-expression clusters

There are two groups of CPR genes, RR-1 and RR-2, and we examined whether these two are ever expressed with the same expression pattern. Ten of the co-expression clusters, with 52% of the genes, were a mixture of RR-1 and RR-2 (Fig. 1B). Four clusters were exclusively RR-1; all of these had prominent expression within an instar, indicating that their products contributed exclusively to post-ecdysial endocuticle. All seven clusters that had exclusively RR-2 genes had sharp peaks of expression in pharate stages, suggesting that they contribute to pre-ecdysial exocuticle. We will need to study spatial expression patterns to learn whether genes within a cluster contribute to the same body regions.

3.6. Expression patterns of genes in tandem arrays and sequence clusters

An. gambiae CPR genes can also be classified based on genome organization and sequence similarity. Only 32 of the 156 genes exist as singletons; all others are found in tandem arrays (genes linked on a chromosome no more than 20 kb from their nearest CPR neighbors). RR-1 and RR-2 genes do not co-exist in the same tandem array. Tandem arrays were not named when they were first described (Cornman et al., 2007), but we have now numbered them in their order on each chromosomal arm (Fig. 1B). Within some tandem arrays with RR-2 genes are sequence clusters of genes closely related in sequence; sequence clusters can also have members outside an array. Our data revealed that genes within a single tandem array are found in several co-expression clusters (Fig. 1B, gray highlights), indicating that they are regulated independently. On the other hand, the majority of genes of a sequence cluster showed the same expression pattern. These shared expression patterns are shown in Fig. 3. One exception is sequence cluster 3RA that consists of only five genes, and they showed two different patterns in terms of expression in pharate adults. One (3RAa) representing the first two genes in the cluster, the other (3RAb) the last three. There were five genes left out of Fig. 3 because they did not conform to the pattern of the other members of their cluster. Three of these non-conformers were outside (*CPR72*) or at the ends of their respective sequence clusters (*CPR47* and *CPR156*). The two remaining (*CPR119* and *CPR145*) were embedded in their respective clusters and had the same orientation on the chromosome as some other members.

Exceptions may also be due to differences between gene position in the sequenced PEST strain and the G3 strain that we analyzed.

Genes in each sequence cluster showed comparable expression not only for patterns but also for their expression levels. Fig. 4 presents box plots of the sum of R_0 through all developmental stages for genes in each sequence cluster with the composite of all genes. The consistent expression patterns and levels reveal that transcription of almost all genes in each sequence cluster is regulated in same manner.

4. Discussion

4.1. Comparison with published microarray studies

Data are available from two microarray analyses that used RNA from whole animals collected at different developmental ages. Details are provided below, but the comparison with our data revealed the importance of collecting animals at precise periods after molts and problems that arise because of the large number of very similar CPR genes.

Koutsos et al. (2007) carried out a microarray analysis using 20,000 ESTs on a chip and probing with cDNA from embryos, 5 larval stages taken at 48 h intervals after oviposition, pupae, and newly emerged adult males and females. All data were normalized to RNA produced from their EST set. We were able to retrieve 18 CPRs by searching with IPR000618 (that retrieves sequences with the R&R Consensus). We examined data in the Supplementary Figures of the paper and in the link it provided to VectorBase (<http://agambiae.vectorbase.org/ExpressionData/>). VectorBase now uses the AGAP designations and reported far more “multiple reporters” than the original data. Of the 18 CPR genes, data from 8 were consistent with what we found. In most cases the discrepancies were due to the microarray analyses revealing expression at stages where our analysis did not. Furthermore, the sharp peaks of expression we found for most genes were not observed with the fewer developmental periods they examined and their failure to use more than egg laying as a point of synchronization. Finally, the variance we found for replicate time points was far below that seen with the microarray data.

In an earlier study, Marinotti et al. (2006) examined gene expression with particular emphasis on changes after blood feeding. They used the Affymetrix Gene Chip *Plasmodium/Anopheles* Genome Array and RNA taken from the Pink-eye strain of *Anopheles*. Their analysis, available at <http://www.angagepuci.bio.uci.edu/>, has had the annotation revised to reflect Ensembl's release 45 (with AGAP identifiers). The surprising finding of this analysis was the large number of CPR genes with maximal, and in many cases only, expression at BF3h, three hours after blood feeding. Of the 156 CPR genes, 145 have corresponding AGAP numbers (Supplementary Table 1). Many of the genes were so similar that the Affymetrix analysis combined them under one AGAP number; there were 15 such combinations ranging from 2-9 genes each. All six combinations with signals higher than 1,000 had their maximum values at BF3h. Of the 72 CPR genes represented by a unique AGAP number only 29 had maximal expression levels higher than 1,000; and of these almost 60% had maximal expression at BF3h. These belonged to 10 different co-expression clusters. Low levels of the epidermal chitin synthase transcript (AGAP001748) were also observed at this time. Obviously further work is needed to learn if this surprising adult expression of CPR genes can be confirmed and whether the resulting proteins are secreted into the cuticle.

4.2. Expression patterns

Expression profiles of 152 CPR genes were grouped into 21 co-expression clusters (Fig. 1). The majority (13) of the co-expression clusters had maximum expression during pharate stages

when synthesis of the cuticle for the next stage is underway. (See Section 3.3.) Our willingness to assign even 12 h pupae as pharate adults was based on the short period between pupal and adult eclosion (less than 36 h). Also, it was reported that the pupal peak of ecdysteroid, presumed to trigger pupal-adult apolysis, was observed at 6 h of the pupal stage that lasts 48 h in another mosquito *Aedes aegypti* (Margam et al., 2006). All but two co-expression clusters (clusters 9 and 16) have their expression peaks at molting periods, i.e. pharate stages and/or just after each molt, consistent with cuticle synthesis being very active at these periods. Clusters that displayed significant cDNA levels in larvae had similar patterns at all larval stages.

Andersen (2000) was the first to recognize that pre-ecdysial cuticle had RR-2 proteins, whereas post-ecdysial cuticle had RR-1 proteins. He acknowledged that this generalization was tentative because of limited evidence. Our data provide additional support for this idea. Genes in the seven co-expression clusters that have exclusively RR-2 genes (clusters 0, 1, 6, 7, 8, 13, and 18) are all expressed only at pharate stages. The large co-expression cluster 15, which had only RR-1 genes, might be an exception because its members were also expressed in pharate stages. In contrast to the RR-2 clusters, however, cluster 15 mRNAs were also present after the molt. The other three co-expression clusters with only RR-1 genes (clusters 4, 9, and 16) had cDNA only in inter-molt or non-molt stages.

4.3. Stage-specific genes

Wigglesworth (1961) proposed that metamorphic stages were underwritten by stage-specific genes. Cuticular proteins would be good candidates for stage-specificity for they are essential components of the different structures that characterize each metamorphic stage. In accord with this idea, cuticular proteins were often named based on the stage when they were first found, such as LCP standing for **L**arval **C**uticular **P**rotein. Later on, however, evidence marshaled against stage-specific genes was reported based on lepidopteran and coleopteran cuticular proteins, first with observations of electrophoretic banding patterns and subsequently by analyses of when genes coding for cuticular proteins were active. It appeared that the physical properties of cuticle (rigid or flexible) were more important than metamorphic stage in determining which cuticular proteins would be present. The data were scanty, however, based on just a few genes in any species (for review see Willis, 1996). Now with the current study, we had the opportunity to ask again whether there are genes expressed only in one metamorphic stage. We used stringent criteria for stage-specificity and identified 25 stage-specific genes (Table 1). Does this restore the gene set hypothesis for metamorphosis? Not yet. Our analysis was based on mRNA extracted from the whole animal at 12 h intervals. It remains possible that an “adult-specific” gene might be expressed in epidermal cells secreting a particular larval and/or pupal structure that made a tiny contribution to the total RNA pool, or that the periods we sampled did not include a time when the gene was active. Even if some genes are absolutely stage-specific, it must be remembered that the study that first used data from cuticular proteins to challenge the gene set hypothesis acknowledged that there would be some “stage-specific” genes that were used to build stage-specific structures (Willis, 1986). The original hypothesis anticipated that the polymorphic metamorphic stages would be underwritten by stage-specific genes. No one expected genes for housekeeping functions to be stage-specific, but the supporters of the hypothesis anticipated that the structures characteristic of each stage would be built with products from stage-specific genes. Our data showed that the vast majority (84%) of the CPR genes are used at more than one stage. Therefore, metamorphosis is underwritten by a complex pattern of usage of cuticular protein genes. Obviously, additional information about this issue will come from an analysis of spatial distribution of cuticular proteins.

4.4. Coordination of regulation

We looked for putative transcription factor binding sites that might underlie the variety of expression patterns of the CPR genes. We used TESS, Transcription Element Search System (<http://www.cbil.upenn.edu/cgi-bin/tess/tess>) (Schug and Overton, 1997) to search for binding sites for insect transcription factors from the TRANSFAC database in the 1 kb upstream region of CPR genes. Some CPR genes are so close to each other that the 1 kb region overlaps the coding region of an adjacent gene. For these genes shorter upstream sequences were used. Numerous sites were found in most genes. All the CPR genes had sites for E74A, Eve, Hb, and Zen. E74A is known as one of the ecdysone signal transducers (Fletcher and Thummel, 1995; Buszczak and Segraves, 2000). Binding sites for some other ecdysone signal-related transcription factors, EcR, Ftz-F1, and one or more of the multiple isoforms of the Broad-Complex, were found in 96%, 89%, and 92% of CPR genes, respectively. Their presence indicates that all of the CPR genes are regulated by ecdysteroids directly or indirectly across post-embryonic development, something consistent with the expression patterns we found. Surprisingly, no correlation was observed between the distribution of these binding sites and co-expression clusters. Several other factors also had binding sites in the majority of genes; these are: Abd-B, AP-1, Bcd, B-factor, C/EBP, DEP2, dl, Elf1/NTF-1, Ems, GAGA factor, Gt, HSTF, Oct-2, Prd, SGF-1, SGF-2, Sn, TBP, Tll, Twi, Ubx, and Zeste. These factors may be conveying temporal and/or spatial information. Given how many putative binding sites are present in association with each gene, it is impossible with the present analysis to draw any conclusions.

It has been reported that genes in gene clusters are often expressed in a concerted pattern. Homeotic (Hox) gene complexes are arranged in the same order in the genome as the domain where they function along the antero-posterior axis (Lewis, 1978; Harding et al., 1985; Duboule, 1998). In contrast, genes in the human β -globin cluster are lined in the order of temporal timing of expression (Jane and Cunningham, 1996; Li et al., 2006). Several models to explain the regulatory mechanism of this concerted expression have been proposed (Kmita and Duboule, 2003; Li et al., 2006; Negre and Ruiz, 2007). With the CPR genes, there was no coordinated regulation of genes in tandem arrays except for the subsets of some RR-2 genes that are found in sequence clusters. Most RR-2 genes in each sequence cluster had the same temporal expression pattern; exceptions were generally in the genes at the edges of the cluster or in genes that were not tightly linked. It is still necessary to learn the spatial expression pattern of these RR-2 genes. We anticipate that CPR genes of each sequence cluster that have the same temporal expression pattern will also be expressed in the same regions because the proteins they code for are so similar to one another. These genes are good candidates for investigating whether there might be a mechanism to regulate gene clusters at the chromatin level.

Why does *An. gambiae* possess so many paralogs that are expressed in same pattern? It may be relevant that most or all of the genes in many of the sequence clusters are being expressed at the two highest levels of expression we measured (4 or 5, Fig. 1B). This suggests that once maximal expression levels were reached, the only way to get more transcripts would be to increase gene copy number. The rapid development of mosquitoes requires massive synthesis of cuticular proteins over a brief period of time. It is thus prudent to postulate that sequence clusters are maintained to increase transcript levels.

To obtain the complete story regarding what types of cuticle are composed of which cuticular proteins, spatial expression analyses are required. Furthermore, we may be wrong to anticipate that transcription is followed quickly by translation and secretion. Nevertheless, the precise data on the developmental expression patterns of the 152 CPR genes in *An. gambiae* presented here provide important indications of cuticular protein distribution and provocative information to guide future studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We are grateful to Dr. Mark Brown (University of Georgia) for providing eggs from his *Anopheles* colony. We thank Frank Bizouarn from Bio-Rad for early help with handling qRT-PCR data, and Dr. R. Scott Cornman for his valuable advice on data analyses and insightful comments on the MS. This work was supported by a grant from the National Institutes of Health (AI55624) to JHW.

References

- Andersen SO. Amino acid sequence studies on endocuticular proteins from the desert locust, *Schistocerca gregaria*. *Insect Biochem Mol Biol* 1998;28:421–434. [PubMed: 9692242]
- Andersen SO. Studies on proteins in post-ecdysial nymphal cuticle of locust, *Locusta migratoria*, and cockroach, *Blaberus craniifer*. *Insect Biochem Mol Biol* 2000;30:569–577. [PubMed: 10844249]
- Andersen SO, Hojrup P, Roepstorff P. Insect cuticular proteins. *Insect Biochem Mol Biol* 1995;25:153–176. [PubMed: 7711748]
- Arakane Y, Hogenkamp DG, Zhu YC, Kramer KJ, Specht CA, Beeman RW, Kanost MR, Muthukrishnan S. Characterization of two chitin synthase genes of the red flour beetle, *Tribolium castaneum*, and alternate exon usage in one of the genes during development. *Insect Biochem Mol Biol* 2004;34:291–304. [PubMed: 14871625]
- Buszczak M, Segraves WA. Insect metamorphosis: out with the old, in with the new. *Curr Biol* 2000;10:R830–R833. [PubMed: 11102824]
- 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*. 2007IN PRESS
- Duboule D. Vertebrate hox gene regulation: clustering and/or colinearity? *Curr Opin Genet Dev* 1998;8:514–518. [PubMed: 9794816]
- Fletcher JC, Thummel CS. The *Drosophila E74* gene is required for the proper stage- and tissue-specific transcription of ecdysone-regulated genes at the onset of metamorphosis. *Development* 1995;121:1411–1421. [PubMed: 7789271]
- Hamodrakas SJ, Willis JH, Iconomidou VA. A structural model of the chitin-binding domain of cuticle proteins. *Insect Biochem Mol Biol* 2002;32:1577–1583. [PubMed: 12530225]
- Harding K, Wedeen C, McGinnis W, Levine M. Spatially regulated expression of homeotic genes in *Drosophila*. *Science* 1985;229:1236–1242. [PubMed: 3898362]
- He N, Botelho JM, McNall RJ, Belozarov V, Augustine Dunn W, Mize T, Orlando R, Willis JH. Proteomic analysis of cast cuticles from *Anopheles gambiae* by tandem mass spectrometry. *Insect Biochem Mol Biol* 2007;37:135–146. [PubMed: 17244542]
- Iconomidou VA, Willis JH, Hamodrakas SJ. Is β -pleated sheet the molecular conformation which dictates formation of helicoidal cuticle? *Insect Biochem Mol Biol* 1999;29:285–292. [PubMed: 10319442]
- Iconomidou VA, Chryssikos GD, Gionis V, Willis JH, Hamodrakas SJ. “Soft”-cuticle protein secondary structure as revealed by FT-Raman, ATR FT-IR and CD spectroscopy. *Insect Biochem Mol Biol* 2001;31:877–885. [PubMed: 11439247]
- Iconomidou VA, Willis JH, Hamodrakas SJ. Unique features of the structural model of ‘hard’ cuticle proteins: implications for chitin-protein interactions and cross-linking in cuticle. *Insect Biochem Mol Biol* 2005;35:553–560. [PubMed: 15857761]
- Jane SM, Cunningham JM. Molecular mechanisms of hemoglobin switching. *Int J Biochem Cell Biol* 1996;28:1197–1209. [PubMed: 9022279]
- 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]

- Kmita M, Duboule D. Organizing axes in time and space; 25 years of colinear tinkering. *Science* 2003;301:331–333. [PubMed: 12869751]
- Koutsos AC, Blass C, Meister S, Schmidt S, MacCallum RM, Soares MB, Collins FH, Benes V, Zdobnov E, Kafatos FC, Christophides GK. Life cycle transcriptome of the malaria mosquito *Anopheles gambiae* and comparison with the fruitfly *Drosophila melanogaster*. *Proc Natl Acad Sci USA* 2007;104:11304–11309. [PubMed: 17563388]
- Lekanne Deprez RH, Fijnvandraat AC, Ruijter JM, Moorman AF. Sensitivity and accuracy of quantitative real-time polymerase chain reaction using SYBR green I depends on cDNA synthesis conditions. *Anal Biochem* 2002;307:63–69. [PubMed: 12137780]
- Lewis EB. A gene complex controlling segmentation in *Drosophila*. *Nature* 1978;276:565–570. [PubMed: 103000]
- Li Q, Barkess G, Qian H. Chromatin looping and the probability of transcription. *Trends Genet* 2006;22:197–202. [PubMed: 16494964]
- Liu W, Saint DA. A new quantitative method of real time reverse transcription polymerase chain reaction assay based on simulation of polymerase chain reaction kinetics. *Anal Biochem* 2002;302:52–59. [PubMed: 11846375]
- Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ Method. *Methods* 2001;25:402–408. [PubMed: 11846609]
- Magkrioti CK, Spyropoulos IC, Iconomidou VA, Willis JH, Hamodrakas SJ. cuticleDB: a relational database of Arthropod cuticular proteins. *BMC Bioinformatics* 2004;5:138. [PubMed: 15453918]
- Margam VM, Gelman DB, Palli SR. Ecdysteroid titers and developmental expression of ecdysteroid-regulated genes during metamorphosis of the yellow fever mosquito, *Aedes aegypti* (Diptera: Culicidae). *J Insect Physiol* 2006;52:558–568. [PubMed: 16580015]
- Marinotti O, Calvo E, Nguyen QK, Dissanayake S, Ribeiro JM, James AA. Genome-wide analysis of gene expression in adult *Anopheles gambiae*. *Insect Mol Biol* 2006;15:1–12. [PubMed: 16469063]
- Negre B, Ruiz A. HOM-C evolution in *Drosophila*: is there a need for Hox gene clustering? *Trends Genet* 2007;23:55–59. [PubMed: 17188778]
- Nikou D, Ranson H, Hemingway J. An adult-specific CYP6 P450 gene is overexpressed in a pyrethroid-resistant strain of the malaria vector, *Anopheles gambiae*. *Gene* 2003;318:91–102. [PubMed: 14585502]
- 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. *Nature Genet* 2006;38:500–501. [PubMed: 16642009]
- Richman AM, Dimopoulos G, Seeley D, Kafatos FC. *Plasmodium* activates the innate immune response of *Anopheles gambiae* mosquitoes. *EMBO J* 1997;16:6114–6119. [PubMed: 9321391]
- Rozen, S.; Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz, S.; Misener, S., editors. *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. 132. Humana Press; Totowa, NJ: 2000. p. 365–386.
- Salazar CE, Mills-Hamm D, Kumar V, Collins FH. Sequence of a cDNA from the mosquito *Anopheles gambiae* encoding a homologue of human ribosomal protein S7. *Nucleic Acids Res* 1993;21:4147. [PubMed: 8371989]
- Schug, J.; Overton, GC. Technical Report CBIL-TR-1997-1001-v0.0. Computational Biology and Informatics Laboratory, School of Medicine, University of Pennsylvania; 1997. TESS: Transcription Element Search Software on the WWW. URL: <http://www.cbil.upenn.edu/tess>
- Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. 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]
- 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, Dunn WA, 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]
- Vizioli J, Bulet P, Hoffmann JA, Kafatos FC, Müller HM, Dimopoulos G. Gambicin: a novel immune responsive antimicrobial peptide from the malaria vector *Anopheles gambiae*. *Proc Natl Acad Sci USA* 2001;98:12630–12635. [PubMed: 11606751]
- Wigglesworth, VB. Insect polymorphism - a tentative synthesis. In: Kennedy, JS., editor. *Insect Polymorphism*. Royal Entomological Society; London: 1961. p. 103-113.
- Willis JH. The paradigm of stage specific gene sets in insect metamorphosis: Time for revision! *Arch Insect Biochem Physiol* 1986:47–57.
- Willis, JH. Metamorphosis of the cuticle, its proteins and their genes. In: Gilbert, LI.; Tata, JR.; Atkinson, BG., editors. *Metamorphosis: Post-Embryonic Reprogramming of Gene Expression in Amphibian and Insect Cells*. Academic Press; New York: 1996. p. 253-282.
- Willis, JH.; Ionomidou, VA.; Smith, RF.; Hamodrakas, SJ. Cuticular Proteins. In: Gilbert, LI.; Iatrou, K.; Gill, SS., editors. *Comprehensive Molecular Insect Science*. 4. Elsevier; Oxford: 2005. p. 79-110.
- Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, Copley RR, Christophides GK, Thomasova D, Holt RA, Subramanian GM, Mueller HM, Dimopoulos G, Law JH, Wells MA, Birney E, Charlab R, Halpern AL, Kokoza E, Kraft CL, Lai Z, Lewis S, Louis C, Barillas-Mury C, Nusskern D, Rubin GM, Salzberg SL, Sutton GG, Topalis P, Wides R, Wincker P, Yandell M, Collins FH, Ribeiro J, Gelbart WM, Kafatos FC, Bork P. Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* 2002;298:149–159. [PubMed: 12364792]

A

Fig. 1. SOM co-expression clusters of CPR genes

Expression pattern of 152 CPR genes (153 primer pairs) were grouped into 21 co-expression clusters by SOM with 3×7 node geometry and iterations of 500,000. A) Expression pattern of co-expression clusters. The number of genes included in each cluster is indicated in parentheses. The Y-axis shows expression level represented by the centroid of genes in the cluster, which is almost identical to the mean of data after 500,000 iterations. Error bars are S.D. of centroid. The X-axis shows time points along development. Each post-ecdysial developmental stage is indicated with either a white or gray area representing 1st through 4th instar larvae (L1-4), pupae (P), and adults (A) from left to right, labeled only in the first panel for cluster 0. Each stage was sampled at 12 h intervals beginning immediately after ecdysis,

so the last point or two in each stage represents the pharate condition of the following stage (see Section 3.3). B) The names of CPR genes in co-expression clusters. Stage indicates synopsis of stages when genes are expressed: A, adult; PA, pharate adult; P, pupa; PP, pharate pupa; L, post-ecdysial larva; PL, pharate larval instar; IM, inter molt; L4, 4th instar larva. Multiple genes in one cell in name column indicate genes analyzed with common primers. A “+” indicates both genes were expressed, an “or” means that only one of the two genes is present in the G3 strain. (See Supplementary Table 2 – comments for details). Designation of class, sequence cluster, and tandem array are from Cornman et al. (2007). Each sequence cluster has a different color. Tandem arrays are named by chromosomal arms and their order of appearance. Three tandem arrays are shown with gray of different intensities. Singletons are indicated with just chromosomal arms in the column of tandem array. Expression levels show rounded up values of $\text{Log}_{10}(\text{Max}R_0 \times 10^7)$, where $\text{Max}R_0$ is maximum R_0 value across the developmental stages examined.

Fig. 2. Sexually different expression of six CPR genes

Expression during pupal and adult stages of six genes expressed differently in two sexes are shown. The X-axis is developmental stage with numbers showing hours after pupation and adult emergence: P, pupa; A, adult. The Y-axis shows $R_0 \times 10^7$. Open bars and gray bars show males and females, respectively. Error bars indicate S.E. of the mean from three independent cDNA preparations, and stars indicate statistically significant differences ($p < 0.05$) by t -test. The two genes with a # showed predominant expression during larval life.

Fig. 3. Consistent expression pattern of genes within a sequence cluster

Expression data of genes were combined for each sequence cluster. The X-axis is developmental stage with numbers showing hours after each molt: L1-4, 1st-4th instar larva; P, pupa; A, adult. The Y-axis shows the mean of relative expression level, which is the ratio of R_0 to summation of R_0 through all stages examined. Solid bars are larval samples that were not sexed, open bars and gray bars are males and females, respectively. Error bars show S.D. of the mean. Numbers in parentheses indicate numbers of genes included in figure followed by the total number of genes in that sequence cluster. Because they did not conform to the standard expression pattern, the following genes were excluded from the figure: 2RB – *CPR119*; 2LB – *CPR47*; 2LC – *CPR72*, *CPR145*; 3RB – *CPR156*.

Fig. 4. Distribution of expression level of genes in sequence clusters

Expression level of each gene (summation of R_0 across development) in each sequence cluster is plotted in the box plot. Box plot for all CPR genes is also shown (ALL). The data from the 3 primer pairs that recognize multiple genes were divided by gene numbers to obtain expression level per single gene. The Y-axis shows R_0 logarithmically. Each box represents the 25%-75% range. The horizontal lines in the boxes show the median. "Whiskers" show the lowest and highest data without outliers. Open circles and stars indicate mild and extreme outliers, respectively. These lie at a distance of 1.5 (mild outliers) or 3.0 (extreme outliers) times the interquartile range shown by the box. The same 5 genes that were excluded in Fig. 3 were excluded in this analysis.

Table 1

Stage specific genes.

Larval-specific	Pupal-specific	Adult-specific
<i>CPR11</i> <i>CPR31</i> <i>CPR49</i> <i>CPR50</i> <i>CPR51</i> <i>CPR52</i> <i>CPR53</i> <i>CPR54</i> <i>CPR61</i> <i>CPR129</i> <i>CPR133</i> or <i>153</i> <i>CPR136</i>	<i>CPR96</i>	<i>CPR10</i> <i>CPR17</i> <i>CPR18</i> <i>CPR19</i> <i>CPR20</i> <i>CPR26</i> <i>CPR55</i> <i>CPR56</i> <i>CPR76</i> <i>CPR119</i> <i>CPR132</i> <i>CPR156</i>

Expression levels are at least two orders of magnitude higher than at the other two metamorphic stages and any minor expression at other stages is not higher than $R_0 = 10^{-6}$.