Molecular cloning and comparative analysis of transcripts encoding chemosensory proteins from two plant bugs, *Lygus lineolaris* and *Lygus hesperus*

J. Joe Hull1,* 1, Omaththage P. Perera2,* 2 and Mei-Xian Wang1, 3

1 USDA-ARS Arid Land Agricultural Research Center, Maricopa, Arizona, USA; 2 USDA-ARS, Southern Insect Management Research Unit, Stoneville, Mississippi, USA and 3 College of Animal Sciences, Zhejiang University, Hangzhou, China

**Abstract**  Chemosensory proteins (CSPs) are soluble carrier proteins typically characterized by a six-helix bundle structure joined by two disulfide bridges and a conserved Cys spacing pattern (C1-X6-8-C2-X16-21-C3-X2-C4). CSPs are functionally diverse with reported roles in chemosensation, immunity, development, and resistance. To expand our molecular understanding of CSP function in plant bugs, we used recently developed transcriptomic resources for *Lygus lineolaris* and *Lygus hesperus* to identify 17 and 14 CSP-like sequences, respectively. The *Lygus* CSPs are orthologous and share significant sequence identity with previously annotated CSPs. Three of the CSPs are predicted to deviate from the typical CSP structure with either five or seven helical segments rather than six. The seven helix CSP is further differentiated by an atypical C3-X3-C4 Cys spacing motif. Reverse transcriptase PCR-based profiling of CSP transcript abundance in adult *L. lineolaris* tissues revealed broad expression for most of the CSPs with antenna specific expression limited to a subset of the CSPs. Comparative sequence analyses and homology modeling suggest that variations in the amino acids that comprise the *Lygus* CSP binding pockets affect the size and nature of the ligands accommodated.

**Key words**  chemosensation; chemosensory protein; homology modeling; *Lygus* plant bug; mired; transcriptome

**Introduction**

The chemosensory protein (CSP) family in insects comprises a diverse group of relatively small (100–135 amino acids) globular polypeptides characterized by a hydrophobic binding pocket (Pelosi et al., 2006, 2014) bounded by two disulfides with a highly conserved cysteine spacing motif (C1-X6-8-C2-X16-21-C3-X2-C4) (Zhou et al., 2006). CSPs have been identified from multiple insect orders (Pelosi et al., 2014) with the number expressed varying widely across species with as few as four in *Drosophila melanogaster* (Vieira & Rozas, 2011) to as many as 70 in *Locusta migratoria* (Zhou et al., 2013).

Although RNAi-mediated knockdown of CSPs has been reported to result in odor-specific attenuation of antennal responses (Yi et al., 2014; Song et al., 2018), unequivocal *in vivo* validation of CSP function in chemosensation remains to be broadly demonstrated. However, chemosensory tissue expression (Pelosi et al., 2006, 2014) and affinity for plant volatiles (Gu et al., 2012; Liu et al., 2014; Sun et al., 2014; Yi et al., 2014, 2015), cuticular hydrocarbons (Ozaki et al., 2005; González et al., 2009), and pheromonal components (Briand et al., 2002; Ban et al., 2003; Li et al., 2016) are suggestive of a *bona fide* role for CSPs in chemical communication. Despite this, expression in other tissues (Zhou et al., 2006, 2013; Gong et al., 2007) suggests that functionality likely also extends beyond chemosensation with roles proposed in pheromone transport and release (Jacquin-Joly et al., 2001; Vogel

© 2018 Institute of Zoology, Chinese Academy of Sciences
et al., 2010; Dani et al., 2011; Iovinella et al., 2011), immunity and xenobiotic degradation (Oduol et al., 2000; Sabatier et al., 2003; Hou et al., 2013; Xuan et al., 2015; Liu et al., 2016b), tissue regeneration (Kitabayashi et al., 1998), development (Picimbon et al., 2001; Wanner et al., 2005; Maleszka et al., 2007), locust phase transition (Guo et al., 2011), and reduction in proboscis cavity surface tension (Liu et al., 2014). In a majority of these reports, CSP functionality involves binding and transport/protection of various hydrophobic compounds. This broad substrate range can be attributed to the unique conformation afforded by the helical bundle and the location of the disulfide bridges, which together provide greater structural flexibility than odorant binding proteins (OBPs) without compromising the stability of the polypeptide (Lartigue et al., 2002; Campanacci et al., 2003; Mosbah et al., 2003; Tegoni et al., 2004; Tomaselli et al., 2006).

Although CSPs have been most extensively characterized in dipterans and lepidopterans, advances in transcriptomic resources have facilitated their identification in a number of hemipterans (Jacobs et al., 2005; Zhou et al., 2006, 2010, 2014, 2015; Xu et al., 2009; Gu et al., 2012, 2013; Hua et al., 2012, 2013; Futahashi et al., 2013; Ribeiro et al., 2014; Sun et al., 2015; Cui et al., 2017; Wang et al., 2016; Wu et al., 2016; Xue et al., 2016; Liu et al., 2016a). Similar to CSPs in other insect orders, hemipteran CSPs are broadly expressed (Zhou et al., 2006, 2014, 2015; Wang et al., 2016). Functional characterization studies, however, have largely focused on putative chemosensory roles. Among mirid plant bugs, CSPs in Adelphocoris lineolatus have been reported to bind host-related compounds and pheromonal components (Gu et al., 2012; Sun et al., 2015), whereas Apolygus luctorum CSPs bound both plant volatiles (Hua et al., 2013) and secondary metabolites of cotton (Hua et al., 2012, 2013).

Lygus species represent a complex of morphologically similar polyphagous hemipteran plant bugs (Miridae) (Schwartz & Footitt, 1998; Wheeler, 2001) that cause significant economic losses in diverse food, fiber, and seed crops (Scott, 1977; Wheeler, 2001; Ritter et al., 2010; Naranjo et al., 2011). Typically employing a “lacerate and flush” (also referred to as “macerate and flush”) strategy, Lygus feeding damage can manifest in organ abscission, deformation of developing fruits, feeding site necrosis, and reduced vegetative growth (Strong, 1970). Although ~40 species occur worldwide, two species dominate (with some degree of geographical overlap) different regions of the continental United States—L. lineolaris (tarnished plant bug) in the mid-southern states and L. hesperus (western tarnished plant bug) in the western states (Ellsworth & Barkley, 2001; Musser et al., 2007). Both species are characterized by multiple generations per season with each generation consisting of five nymphal instars. In recent years, L. lineolaris has become the dominant pest species of cotton in the mid-south and has transitioned from historically being an early-season pest of cotton to a mid- to late-season pest that requires increasing numbers of insecticide applications (Fleming et al., 2016). L. hesperus is likewise a key pest of cotton in addition to strawberries and forage alfalfa (Schwartz & Footitt, 1998; Strand, 2008). While management strategies for both species have traditionally relied on broad-spectrum insecticides, reports of resistance in field populations (Snodgrass, 1996; Snodgrass & Scott, 2002; Snodgrass et al., 2009) underscore the need for alternative control tactics. One promising area for potential development involves targeted disruption of the chemosensory system (Soffan et al., 2016; Andersson & Newcomb, 2017). Both L. lineolaris and L. hesperus are strongly influenced by environmental chemical cues (Blackmer et al., 2004; Innocenzi et al., 2005; Byers et al., 2013; Fountain et al., 2014) that trigger antennal responses (Chinta et al., 1994; Dickens et al., 1995; Ho & Millar, 2002; Williams et al., 2010). Our knowledge of the molecular mechanisms underlying Lygus chemosensation, however, is limited to L. lineolaris OBPs (Dickens et al., 1995; Vogt et al., 1999; Hull et al., 2014b) and the olfactory receptor coreceptor (Orco) (Hull et al., 2012). To address this limitation, we mined recent L. lineolaris and L. hesperus transcriptome assemblies (Hull et al., 2013, 2014a, 2014b; Tassone et al., 2016) for CSP-like sequences, examined sequence conservation between the species, and profiled CSP expression in L. lineolaris. In addition, phylogenetic relationships of the respective transcripts were examined across multiple insect orders and with other hemipteran sequences. This is the first report of CSPs in Lygus and as such fills a knowledge gap for this economically important pest species.

Materials and methods

Insect rearing

L. lineolaris and L. hesperus were obtained from in-house laboratory stock colonies maintained at the USDA-ARS Southern Insect Management Research Unit (Stoneville, MS, USA) and the USDA-ARS Arid Land Agricultural Research Center (Maricopa, AZ, USA), respectively. Colonies were maintained at 27.5–29.0 °C, ~40% humidity (L. lineolaris) or <20% humidity (L. hesperus) under a L14 : D10 photoperiod on green beans (Phaseolus vulgaris) and disposable artificial diet packs (Debolt, 1982; Patana, 1982).
Annotation and bioinformatic analysis of transcripts encoding putative CSPs

Putative CSP encoding transcripts were initially annotated with Blast2GO (Conesa et al., 2005; Götz et al., 2008) using transcriptomic data generated from multiple *L. lineolaris* developmental stages (Hull et al., 2014b). *L. lineolaris* unigene sequences annotated as CSPs were then used to search *L. hesperus* transcriptome assemblies (Hull et al., 2013, 2014a; Tassone et al., 2016). The resulting hits were then re-submitted as queries in a subsequent BLAST-based search of the respective assemblies. The unigene sequences were curated to remove duplicates and the longest isoforms were evaluated via BLASTx against the NCBI nonredundant (nr) database. To confirm the veracity of the CSP annotations, sequences were screened for the presence of the characteristic C1-X5,-6-C2-X18-19-C3-X2-C4 Cys motif (Xu et al., 2009).

Domain analyses were performed using the HMMscan module on the HMMER webserver (Finn et al., 2011) with Pfam and Superfamily databases. Signal peptide predictions were made with SignalP4.0 (Petersen et al., 2011). For comparative purposes, multiple sequence alignments consisting of the respective *L. lineolaris* and *L. hesperus* sequences either alone or in conjunction with CSPs from multiple representative insect orders (Hymenoptera—*Apis mellifera*, Lepidoptera—*Bombyx mori*, Diptera—*D. melanogaster*, and Coleoptera—*Tribolium castaneum*) or five hemipteran species (*Nilaparvata lugens*, *A. lineolatus*, *Sogatella furcifera*, *Laodelphax striatella*, and *A. lucorum*) were generated using default settings for MUSCLE (Edgar, 2004) in Geneious R9.0.02 (Kearse et al., 2012). Accession numbers for the non-*Lygus* sequences used are listed in Table S1. Phylogenetic relationships were inferred from maximum likelihood, minimum evolution, neighbor joining, and UPGMA analyses with support values based on 1000 bootstrap iterations in MEGA6 v6.06 r6140220 (Tamura et al., 2013). Secondary structure predictions were performed using online servers hosting JPred4 (Drozdetskiy et al., 2015) and YASPIN (Lin et al., 2005).

Cloning full-length CSP ORFs from *L. lineolaris* and *L. hesperus*

*L. lineolaris* CSP transcripts were initially amplified using primers (Table S2) designed as described (Hull et al., 2014b) using a full-length cDNA library derived from an RNA pool of all life stages. Full-length sequences for *L. lineolaris* and *L. hesperus* CSPs were subsequently amplified using primer sets capable of annealing to both species (Table S2). For *L. hesperus*, total RNAs from 7–9-d-old mixed gender adult bodies were isolated using TRI Reagent (Life Technologies, Carlsbad, CA, USA). The resulting RNAs were quantified and quality assessed spectrophotometrically using the Take3 module on a Synergy H4 Hybrid Multi-Mode Microplate Reader (Biotek Instruments, Winoski, VT, USA). Residual genomic DNA was removed with DNase I (New England Biolabs, Ipswich, MA, USA) and first-strand cDNAs generated from 500 ng DNase I-treated total RNAs using Superscript III reverse transcriptase (Life Technologies/ThermoFisher, Carlsbad, CA, USA) with custom-made random pentadecamers (IDT, San Diego, CA, USA). Multiple independent reactions were performed using Sapphire Amp Fast PCR Master Mix (Clontech Laboratories/Takara Bio USA Inc., Mountain View, CA, USA) in a 20 μL reaction volume with 0.5 μL cDNA template and 0.2 μmol/L of each primer (Table S2). Thermocycler conditions were 95 °C for 2 min followed by 40 cycles at 95 °C for 20 s, 53 °C for 20 s, 72 °C for 30 s, and a final extension at 72 °C for 5 min. Amplimers were separated on 1.5% agarose gels using a Tris/acetate/EDTA buffer system and visualized with SYBR Safe (Life Technologies). Products from each reaction were subcloned into a pcR2.1-TOPO TA cloning vector (Life Technologies/ThermoFisher). Clones were sequenced at either the Arizona State University DNA Core Laboratory (Tempe, AZ, USA) or the USDA-ARS Genomics and Bioinformatics Research Unit sequencing facility (Stoneville, MS, USA). GenBank accession numbers for LlinCSP1-17 are KX950019-KX950035, LhesCSP1-12 are KU194348-KU194359, LhesCSP13 is KU524880, and LhesCSP14 is KX950018.

RT-PCR-based expression profile of *L. lineolaris* CSPs

To assess the expression of *L. lineolaris* CSP transcripts in early adult life (i.e., the host seeking period), total RNAs were isolated using TriZol (Life Technologies/ThermoFisher) in duplicate from immature 2-d-old adult *L. lineolaris* male and female antenna, proboscis, leg, heads, bodies, midgut/hindgut, and fat body. To provide insights into the role sexual maturity may have on antennal CSP expression, total RNAs were also isolated from antenna of reproductively mature (Brent, 2010) 8-d-old mixed gender adult bodies. Total RNAs were isolated using TRIzol (Life Technologies, Carlsbad, CA, USA) in Geneious 10.1.3 to amplify 100–150 bp fragments of the *L. lineolaris* CSP transcripts.
along with a ubiquitous expression control gene, rpL29 (GDAW01003327), and the olfactory receptor coreceptor, Orco (JQ639214). Amplification of single discrete products from the cDNAs described above was confirmed by end-point PCR with Sapphire Amp Fast PCR Master Mix in a 15 μL reaction with thermocycler conditions consisting of 95 °C for 2 min followed by 35 cycles at 95 °C for 20 s, 62 °C for 20 s, and 72 °C for 20 s and terminated with a final 5 min extension. Products were electrophoresed for 30 min at 100 V using 2% agarose gels stained with SYBR safe and then imaged using an Alphalmager gel documentation system (ProteinSimple, San Jose, CA, USA). Images were processed (auto contrast and despeckle) with Photoshop CS6 v13.0 (Adobe Systems Inc., San Jose, CA, USA). Representative products for each primer set were subcloned into the pCR2.1-TOPO TA cloning vector and sequenced as described before.

Molecular modeling

To assess the structural features of select Lygus CSPs, three-dimensional models of LlinCSP1/LhesCSP1, LlinCSP3/LhesCSP3, and LlinCSP6/LhesCSP6 were generated using the Phyre2 web portal (Mezulis et al., 2015). The respective Lygus CSP structures were constructed using spatial coordinates for M. brassicae CSP6 (PDB id 1KX9) (Campanacci et al., 2003) with 100% confidence and high sequence coverage despite varying degrees of conservation with the template: LlinCSP1/LhesCSP1 80% coverage and 22% sequence identity; LlinCSP3/LhesCSP3 79% coverage and 54% sequence identity; LlinCSP6/LhesCSP6 50% coverage and 29% sequence identity. The quality of the resulting structures was assessed using PROSESS (Berjanskii et al., 2010), ProSA (Sippl, 1993; Wiederstein & Sippl, 2007), and RAMPAGE (Lovell et al., 2003), the latter of which performs a Ramachandran analysis of the peptide backbone angles. The solvent accessible surface area was calculated using ProtSA (Bernadó et al., 2006; Estrada et al., 2009). Models were displayed with Swiss-PDB viewer (Guex & Peitsch, 1997) (http://www.expasy.org/spdbv/). Images were processed for publication using pov-ray (http://www.povray.org/).

Results

Identification of putative CSP transcripts

Using available Lygus transcriptomic resources, we identified 17 L. lineolaris transcripts encoding proteins that either exhibited significant similarity with annotated CSP sequences or had the characteristic C1-X5,-C2-X18-19-C3-X2-C4 Cys motif (Xu et al., 2009). We also identified 14 L. hesperus transcripts with 97.3%–100% sequence identity (Table S3) to the L. lineolaris sequences. While the number of putative CSPs identified in the transcriptomes of the two Lygus species is comparable to that reported for a number of other hemipterans (Xu et al., 2009; Vieira & Rozas, 2011; Zhou et al., 2014, 2015; Sun et al., 2015; Cui et al., 2017), temporally, spatially, and/or conditionally restricted transcripts that were not represented in the assemblies may have been missed. Consequently, deeper RNA sequencing and/or generation of genome assemblies for the respective species may further expand the CSP repertoire.

BLASTx analyses of the Lygus sequences using the NCBI nonredundant database revealed highest similarities with hemipteran CSP sequences (Tables S4–S5). Similar to that reported for L. lineolaris OBPs (Hull et al., 2014b), A. luctorum and A. lineolatus transcripts were among the most highly represented BLAST alignments. This degree of conservation is not unexpected as the four species belong to the Miridae family and have broadly overlapping host ranges. Furthermore, unlike OBPs, which typically exhibit limited cross-species sequence identities, the top Lygus CSP BLAST alignment identities ranged from 35% to 100% with a median of ~59% (Tables S4–S5). Sequence identities for the Lygus CSPs varied from 14% to 89% in L. lineolaris (Table S3, Fig. 1) and 14% to 79% in L. hesperus (Table S3, Fig. 2). The sequences for most of the CSPs in the two species were sufficiently conserved that primer sets designed to one species could amplify the orthologous transcript in the other species. The consensus sequences have been deposited with GenBank under accession numbers KX950019–KX950032 for LlinCSP1-14, KU194348–KU194359 for LhesCSP1-12, KU524880 for LhesCSP13, and KX950018 for LhesCSP14.

Comparison of the consensus CSP sequences revealed synonymous mutations in seven CSPs, whereas nonsynonymous mutations that introduced conserved amino acid changes were found for five CSPs; only LlinCSP4/LhesCSP4 and LlinCSP11/LhesCSP11 were 100% identical (Table 1). LlinCSP6 and LhesCSP6 exhibited the greatest degree of sequence divergence with three nonsynonymous mutations. Although LlinCSP15 was specifically amplified from L. lineolaris whole body cDNAs using primers for LlinCSP4/LhesCSP4, the respective CSPs are only 85% identical at the amino acid level with LlinCSP15 containing 11 nonconserved changes relative to the LlinCSP4/LhesCSP4 sequences (Fig. 1). Despite repeated attempts, we were unable to amplify the LlinCSP15 sequence from L. hesperus cDNAs suggesting that it may be unique to L. lineolaris or
Lygus chemosensory proteins

Fig. 1 Amino acid sequence alignment of *Lygus lineolaris* chemosensory proteins (CSPs). The conserved Cys residues (C1–C4) in the “classic” CSP motif are indicated. Circles indicate highly conserved residues potentially critical to ligand binding, whereas asterisks denote conserved aromatic residues thought to function as gates to the ligand-binding pocket. Shading represents conservation of sequence identity. In the sequence logo stacks (Crooks et al., 2004), the height of each stack corresponds to the degree of sequence conservation at that position.

that transcriptional regulation of the gene differs between the two species. LlinCSP16 and LlinCSP17 are internal fragments lacking both start and stop codons that were specific to the *L. lineolaris* transcriptomic dataset, which, unlike the *L. hesperus* assemblies, was generated from all developmental stages. LlinCSP16 is 72% identical with LlinCSP2 at the nucleotide level, whereas LlinCSP17 is 98% identical to LlinCSP7, but has four nucleotide insertions and one deletion. Attempts to amplify the two sequences from diverse *L. lineolaris* tissues failed to

© 2018 Institute of Zoology, Chinese Academy of Sciences, 00, 1–21
yield the desired products; we were unable to amplify LlinCSP16 and putative LlinCSP17 products were indistinguishable at the nucleotide level from LlinCSP7. These findings suggest that expression of the two CSP transcripts may be developmentally and/or conditionally regulated.

Most of the Lygus CSP sequences encode 12–15 kDa proteins composed of 111–135 amino acids with amino terminal signal peptide sequences and the characteristic C1-X₆-C2-X₁₈-C3-X₂-C4 Cys spacing motif (Table 2). Exceptions to the typical CSP structure are LlinCSPs 6, 16, 17, and LhesCSP6. The partial LlinCSP16 and
Table 1 Comparative analysis of *L. lineolaris* and *L. hesperus* CSP clones.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Number of clones</th>
<th>Syn/Nonsyn$^1$</th>
<th>Nature of Nonsyn substitution$^1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>L. lineolaris</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSP1</td>
<td>7</td>
<td>1/0</td>
<td>–</td>
</tr>
<tr>
<td>CSP2</td>
<td>5</td>
<td>4/2</td>
<td>L/P35I; R97/K</td>
</tr>
<tr>
<td>CSP3</td>
<td>9</td>
<td>2/0</td>
<td>–</td>
</tr>
<tr>
<td>CSP4</td>
<td>7</td>
<td>0/0</td>
<td>–</td>
</tr>
<tr>
<td>CSP5</td>
<td>5</td>
<td>7/2</td>
<td>T7L; I10T</td>
</tr>
<tr>
<td>CSP6</td>
<td>10</td>
<td>5/3</td>
<td>V/L10A; I143V; S145T</td>
</tr>
<tr>
<td>CSP7</td>
<td>6</td>
<td>1/0</td>
<td>–</td>
</tr>
<tr>
<td>CSP8</td>
<td>8</td>
<td>1/0</td>
<td>–</td>
</tr>
<tr>
<td>CSP9</td>
<td>4</td>
<td>2/1</td>
<td>R114K</td>
</tr>
<tr>
<td>CSP10</td>
<td>6</td>
<td>2/0</td>
<td>–</td>
</tr>
<tr>
<td>CSP11</td>
<td>7</td>
<td>0/0</td>
<td>–</td>
</tr>
<tr>
<td>CSP12</td>
<td>5</td>
<td>4/0</td>
<td>–</td>
</tr>
<tr>
<td>CSP13</td>
<td>6</td>
<td>4/0</td>
<td>–</td>
</tr>
<tr>
<td>CSP14</td>
<td>8</td>
<td>2/0</td>
<td>–</td>
</tr>
</tbody>
</table>

$^1$Number of synonymous (Syn) and nonsynonymous (Nonsyn) substitutions present in the *L. lineolaris* and *L. hesperus* CSP sequences.

The *L. lineolaris* amino acid affected by the nonsynonymous change is indicated as well as its position within the protein sequence and the corresponding amino acid in the *L. hesperus* CSP.

LlinCSP17 sequences encode internal CSP fragments that have the typical Cys spacing motif and high sequence identity with CSPs identified in *A. lucorum* and *A. lineolatus* respectively (Table S4). The orthologous LlinCSP6 and LhesCSP6 sequences encode atypical CSPs consisting of 196 amino acids (~22.5 kDa) that deviate from the C3-X2-C4 Cys motif by the inclusion of a third residue between the two Cys (Table 2; Figs. 1 and 2). While CSPs extending beyond 180 amino acids have been reported (Forêt et al., 2007; Zhan et al., 2011; Zhang et al., 2014, 2015; Derks et al., 2015; Gu et al., 2015; Li et al., 2015), we found no formal reports describing deviation from the C3–C4 pattern. The insertion is not a sequencing error or an artifact introduced during transcriptome assembly, as multiple full-length clones with the C3-X3-C4 motif were amplified from both species and sequence verified.

Comparative sequence analyses revealed the presence of three additional CSP motifs defined by Wanner et al. as A-C (2004). Among the *Lygus* CSPs, the carboxyl terminal C motif (KYDP) was the most conserved (mean conservation 69%). Absolute conservation of individual motifs was only observed in four CSPs (motif A, 1 CSP; motif B, 1 CSP; motif C, 2 CSPs) with none of the *Lygus* CSPs exhibiting conservation across all three motifs (Table 2; Figs. 1 and 2). LlinCSP9/LhesCSP9 and LlinCSP10/LhesCSP10 have the greatest degree of motif conservation (mean 92%), whereas LlinCSP1/LhesCSP1 and LlinCSP13/LhesCSP13 have the least (mean 25%). LlinCSP12/LhesCSP12 are differentiated from the other CSPs by a seven-amino acid insertion in motif A and LlinCSP13/LhesCSP13 by a three-amino acid deletion. In addition to sequence motifs, CSPs are also typically characterized by the presence of six helical segments that comprise a portion of the ligand binding pocket (Tegoni et al., 2004; Pelosi et al., 2014). Secondary structure determination algorithms predicted six helices for 11 of the full-length *Lygus* CSPs (Fig. S1). LlinCSP1/LhesCSP1 and LlinCSP13/LhesCSP13 deviated from the expected profile with five helices, while LlinCSP6/LhesCSP6 are predicted to contain seven helical segments.

Phylogenetic analyses

To assess the relationship of the *Lygus* CSPs with other insects, we constructed a maximum likelihood tree incorporating the complete repertoire of CSPs from species representing four additional insect orders: Diptera (*D. melanogaster*), Coleoptera (*Tribolium castaneum*), Hymenoptera (*Apis mellifera*), and Lepidoptera (*B. mori*). Similar to other reports (Vieira & Rozas, 2011; Kulmuni & Havukainen, 2013; Pelosi et al., 2014; Zhou et al., 2015), we found poor bootstrap support for deeper branches. Most CSPs grouped according to order and/or species with clear indications of gene expansion in *T. castaneum* and *B. mori* (Fig. 3). LlinCSP10/LhesCSP10
Table 2 Bioinformatic analyses of *L. lineolaris* and *L. hesperus* CSPs.

<table>
<thead>
<tr>
<th>Protein</th>
<th># Residues</th>
<th>kDa</th>
<th>SP cleavage (^1)</th>
<th>Cys Spacing</th>
<th>Conserved CSP motifs (^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LlinCSP1</td>
<td>111</td>
<td>12.2</td>
<td>21-22</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>RSVSDEALEAAL, VGRRLKTYAIVLQ VYAG</td>
</tr>
<tr>
<td>LlinCSP2</td>
<td>135</td>
<td>15.4</td>
<td>22-23</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>YTDKYDNDIIDEIL, DGAELKESPDAL VYDP</td>
</tr>
<tr>
<td>LlinCSP3</td>
<td>128</td>
<td>14.5</td>
<td>19-20</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>YTTKYDNDLDDIL, DGKEKELPDAL IYDP</td>
</tr>
<tr>
<td>LlinCSP4</td>
<td>130</td>
<td>15.1</td>
<td>22-23</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>YTDQYDKMDVDVIL, YGVLKETPDAL IYDP</td>
</tr>
<tr>
<td>LlinCSP5</td>
<td>133</td>
<td>15.5</td>
<td>22-23</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>YTHHDYLVQVL, EAKERDDKPEAL KYPD</td>
</tr>
<tr>
<td>LlinCSP6</td>
<td>196</td>
<td>22.5</td>
<td>18-19</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>FFRQLEGINVDISL, TMKDLRLALPLIL LYPD</td>
</tr>
<tr>
<td>LlinCSP7</td>
<td>127</td>
<td>14.6</td>
<td>19-20</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>YIKVFEEVDPEFIL, EVMKKSIPDVL IYDP</td>
</tr>
<tr>
<td>LlinCSP8</td>
<td>123</td>
<td>14.0</td>
<td>16-17</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>YIKVFEEIDPDIL, HAEALKESPDVL IYDP</td>
</tr>
<tr>
<td>LlinCSP9</td>
<td>130</td>
<td>14.8</td>
<td>18-18</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>YTTKYDNDLDEIL, DGKEKELPDAL IYDP</td>
</tr>
<tr>
<td>LlinCSP10</td>
<td>123</td>
<td>13.9</td>
<td>16-17</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>YTSKDYNDVDDKIL, EGKELKKTPLDALK YYPD</td>
</tr>
<tr>
<td>LlinCSP11</td>
<td>132</td>
<td>15.2</td>
<td>19-20</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>YPDKDYNDLDEIL, DGAELDKPEAL KDNP</td>
</tr>
<tr>
<td>LlinCSP12</td>
<td>131</td>
<td>14.8</td>
<td>18-17</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>WVTGYDQKGLQFNIDVDAI, PSEAEKKYLGEIF KYNP</td>
</tr>
<tr>
<td>LlinCSP13</td>
<td>113</td>
<td>12.5</td>
<td>16-17</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>FLFTSE—VKLI, LGNQKLPAEVL RYRI</td>
</tr>
<tr>
<td>LlinCSP14</td>
<td>126</td>
<td>14.4</td>
<td>16-17</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>YTRVFEEVDPEFIL, HAEVKKSPDVL IYDP</td>
</tr>
<tr>
<td>LlinCSP16</td>
<td>130</td>
<td>15.2</td>
<td>18-23</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>YTDKYDNDLVIDVIL, YGVLKETPDAL IYDP</td>
</tr>
<tr>
<td>LlinCSP16</td>
<td>90</td>
<td>10.1</td>
<td>n/a</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>D GAELKESPDAL SQHL</td>
</tr>
<tr>
<td>LlinCSP17</td>
<td>98</td>
<td>11.4</td>
<td>n/a</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>—FEEVDPEFIL, HAEVKKSPDVL IYDP</td>
</tr>
<tr>
<td>LhesCSP1</td>
<td>111</td>
<td>12.2</td>
<td>21-22</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>RSVSDEALEAAL, VGRRLKTYAIVLQ VYAG</td>
</tr>
<tr>
<td>LhesCSP2</td>
<td>135</td>
<td>15.4</td>
<td>22-23</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>YTDKYDNDLDEIL, DGAELKESPDAL VYDP</td>
</tr>
<tr>
<td>LhesCSP3</td>
<td>138</td>
<td>15.5</td>
<td>19-20</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>YTTKYDNDLDDIL, DGKEKELPDAL IYDP</td>
</tr>
<tr>
<td>LhesCSP4</td>
<td>130</td>
<td>15.2</td>
<td>22-23</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>YTDQYDKMDVDVIL, YGVLKETPDAL IYDP</td>
</tr>
<tr>
<td>LhesCSP5</td>
<td>133</td>
<td>15.5</td>
<td>20-21</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>YTHHDYLVQVL, EAKERDDKPEAL KYPD</td>
</tr>
<tr>
<td>LhesCSP6</td>
<td>196</td>
<td>22.6</td>
<td>18-19</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>FFRQLEGINVDISL, TMKDLRLALPLIL LYPD</td>
</tr>
<tr>
<td>LhesCSP7</td>
<td>127</td>
<td>14.6</td>
<td>19-20</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>YIKVFEEVDPEFIL, HAEVKKSPDVL IYDP</td>
</tr>
<tr>
<td>LhesCSP8</td>
<td>123</td>
<td>14.0</td>
<td>16-17</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>YIKVFEEIDPDIL, HAEALKESPDVL IYDP</td>
</tr>
<tr>
<td>LhesCSP9</td>
<td>130</td>
<td>14.8</td>
<td>17-18</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>YTTKYDNDLDEIL, DGKEKELPDAL IYDP</td>
</tr>
<tr>
<td>LhesCSP10</td>
<td>123</td>
<td>13.9</td>
<td>16-17</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>YTSKDYNDVDDKIL, EGKELKKTPLDALK YYPD</td>
</tr>
<tr>
<td>LhesCSP11</td>
<td>132</td>
<td>15.2</td>
<td>19-20</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>YPDKDYNDLDEIL, DGAELDKPEAL KDNP</td>
</tr>
<tr>
<td>LhesCSP12</td>
<td>131</td>
<td>14.8</td>
<td>24-25</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>WVTGYDQKGLQFNIDVDAI, PSEAEKKYLGEIF KYNP</td>
</tr>
<tr>
<td>LhesCSP13</td>
<td>113</td>
<td>12.5</td>
<td>16-17</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>FLFTSE—VKLI, LGNQKLPAEVL RYRI</td>
</tr>
<tr>
<td>LhesCSP14</td>
<td>126</td>
<td>14.4</td>
<td>16-17</td>
<td>C1-X(_6)-C2-X(_8)-C3-X(_2)-C4</td>
<td>YTRVFEEVDPEFIL, HAEVKKSPDVL IYDP</td>
</tr>
</tbody>
</table>

\(^1\) Signal peptide cleavage determined with SignalP 4.1 Server (Petersen et al., 2011).

\(^2\) Wanner et al. (2004); Motif A: YTTKYDN[X/V/I][N/D][L/V]DEIL, Motif B: DGKELKXX[I/L]PDAL, Motif C: KYDP.

\(^3\) Fragment identified in transcriptome.

Bold: conserved residue; underline: sequence insertion; box: divergence in conserved Cys spacing.
Fig. 3 Maximum likelihood tree of CSPs from two Lygus species and representative species from four other insect orders. CSP sequences were aligned and the evolutionary history inferred by the maximum likelihood method. The tree with the highest log likelihood is shown. Numbers at the branch point of each node represent support values. Species abbreviations and color coding are: Amel, *Apis mellifera* (purple: Hymenoptera); Bmor, *Bombyx mori* (teal: Lepidoptera); Dmel, *Drosophila melanogaster* (pink: Diptera); Lhes, *Lygus hesperus* (dark green: Hemiptera); Llin, *Lygus lineolaris* (light green: Hemiptera); and Tcas, *Tribolium castaneum* (orange: Coleoptera). Accession numbers for the CSP sequences used are listed in Table S1.

sorted to a smaller branch with two *T. castaneum* CSPs and LlinCSP12/LhesCSP12 aligned, albeit with poor bootstrap support, to a *B. mori* dominant branch. The lone exception to the order/specific groups was a well-supported clade of putatively orthologous sequences encompassing the pentahelix *Lygus* CSPs, LlinCSP1/LhesCSP1 and LlinCSP13/LhesCSP13. The other sequences in this clade are likewise predicted to form five-helix bundles. The clustering of LlinCSP7/LhesCSP7, LlinCSP8/LhesCSP8, LlinCSP14/LhesCSP14, and LlinCSP17 could be indicative of alternative splicing, or alternatively, an indication that the CSPs arose from gene duplication prior to divergence of the two *Lygus* species.

We further examined the phylogenetic relationships of the *Lygus* CSPs within the context of hemipteran sequences by including three planthopper species (*Nilaparvata lugens*, *Sogatella furcifera*, and *Laodelphax striatella*) (Family Delphacidae) and two additional mirids.
Fig. 4 Maximum likelihood tree of CSPs from two *Lygus* species and five additional hemipteran species. CSP sequences were aligned and the evolutionary history inferred by the maximum likelihood method. The tree with the highest log likelihood is shown. Numbers at the branch point of each node represent support values. Species abbreviations and color coding are: Aluc, *Apolygus lucorum* (closed pink), Alin, *Adelphocoris lineolatus* (closed blue), Sfur, *Sogatella furcifera* (open teal), Lstr, *Laodelphax striatella* (open orange); Nlug, *Nilaparvata lugens* (open purple); Lhes, *Lygus hesperus* (closed dark green); and Llin, *Lygus lineolaris* (closed light green). Accession numbers for the CSP sequences used are listed in Table S1.

(*A. lineolatus* and *A. lucorum*). In contrast to that seen with the higher order phylogenetic analysis, multiple clades were identified with fair to moderate bootstrap support (>40) that we have designated CSP-A through CSP-H (Fig. 4). As before, the five-helix CSPs clustered in a single clade (i.e., CSP-A). Consistent with previous reports, branches within the clades were largely lineage specific, with mirid and planthopper CSPs clustering based on
their phylogenetic Family (mirid specific = CSP-G and CSP-H; planthopper specific = CSP-F). Clade CSP-C was characterized by unusually large CSP sequences (>175 amino acids) and included LlinCSP6/LhesCSP6 as well as three planthopper sequences, none of which exhibited the atypical C3-X3-C4 spacing of the Lygus sequences. Potential orthologous relationships observed among the mirid CSPs include: LlinCSP7/LhesCSP7 with AlinCSP4, LlinCSP10/LhesCSP10 with AlucCSP3/AlinCSP3, LlinCSP5/LhesCSP5 with AlinCSP5, LlinCSP9/LhesCSP9 with AlinCSP1/AlucCSP1, LlinCSP3/LhesCSP3 with AlucCSP4, and LlinCSP11/LhesCSP11 with AlinCSP2/AlucCSP7.

**RT-PCR-based profiling of LlinCSP tissue expression**

To gain insights into potential functionality, we used RT-PCR to sex-specifically assess the abundance of the LlinCSP transcripts in various tissues from reproductively immature adults (i.e., day 2 adults) and antennae from reproductively mature adults (i.e., day 8 adults). Transcripts for a majority of the LlinCSPs were amplified from both chemosensory (antennae, leg, proboscis) and nonchemosensory (fat body, gut) associated tissues (Fig. 5), which suggest biological roles that extend beyond chemosensation. Among these “broad expression CSPs,” LlinCSP3, LlinCSP9, and LlinCSP10 appear to be ubiquitously expressed, whereas LlinCSP7 and LlinCSP14 are antennal dominant. In contrast, six CSPs exhibited a narrower expression profile with four CSPs (LlinCSP5, LlinCSP10, LlinCSP13, and LlinCSP15) chemosensory tissue specific/dominant, and two CSPs (LlinCSP4 and LlinCSP8) antennae specific (Fig. 5). Sexual dimorphism in CSP expression was observed for four LlinCSPs: LlinCSP4 was specific to female antennae; LlinCSP5 and LlinCSP6 were amplified from female proboscis but not male; and LlinCSP13 exhibited sex specific expression in female legs and male antennae. Age-related differences in
antennal expression were largely limited to LlinCSP4 in female antennae and LlinCSP14 in male antennae with each transcript more abundant in immature adults (Fig. 5). No LlinCSP16 amplimers were generated from any of the tissues assayed (Fig. 5), and primers designed to amplify LlinCSP17 yielded LlinCSP7 sequences (data not shown).

Structural characterization

While CSP solution structures are limited (e.g., Campanacci et al., 2003; Mosbah et al., 2003; Tomaselli et al., 2006), comparative analyses using those conformations as templates can provide insights into the ligand binding characteristics and functional diversity of CSPs. Using molecular coordinates for ligand-bound Mamestra brassicae CSP6 (MbraCSP6, 1KX9) and a distance criteria of <5 Å, Kulmuni and Havukainen (2013) identified 34 potential ligand interaction sites. We aligned the respective Lygus CSPs to that dataset and examined the distribution of amino acids in the Lygus CSPs based on side chain bulk: small—Gly, Ala, Val, Pro, Ser, Thr, and Cys; intermediate—His, Asp, Glu, Asn, Gln, Ile, Leu, and Met; or large—Arg, Lys, Phe, Tyr, and Trp (Table 3). Compared with MbraCSP6, binding pockets among the Lygus CSPs shifted away from small residues in favor of amino acids with larger sidechains, potentially shrinking the binding pocket relative to MbraCSP6. A similar shift in size distribution was also reported for ant CSPs (Kulmuni & Havukainen, 2013). On a protein-by-protein basis, however, the binding pocket of LlinCSP1/LhesCSP1 and LlinCSP13/LlinCSP13 is larger than MbraCSP6 as both sets of CSPs showed a bias towards smaller amino acids (Table 3). The binding pocket of LlinCSP6/LhesCSP6 is similarly unique with a marked increase in the number of intermediate sized residues, suggesting that the nature of their ligand interactions differs from that of the other Lygus CSPs.

While CSPs are structurally flexible with evidence for cooperative binding (Tegoni et al., 2004; Pelosi et al., 2006; 2014), comparative protein modeling can provide insights into potential structure–function relationships, interspecies CSP diversity, and the effects of sequence variation on conformation. We thus used the spatial coordinates for MbraCSP6 to examine the potential conformational space of the atypical Lygus CSPs (LlinCSP1/LhesCSP1 and LlinCSP6/LhesCSP6) relative to a “typical” Lygus CSP (LlinCSP3/LhesCSP3) and MbraCSP. All of the Lygus sequences were modeled with a high degree of confidence and coverage despite minimal sequence identity (22%–54%), supporting conservation of the CSP folding pattern. Ramachandran plots of the modeled CSPs revealed the backbone dihedral angles of LlinCSP3/LhesCSP3 and LlinCSP6/LhesCSP6 were within favored regions; Cys75 and Ala109 in LlinCSP1/LhesCSP1 mapped to allowed regions (data not shown). Further evaluation of our Lygus CSP models revealed highest scores for the

<table>
<thead>
<tr>
<th>CSP</th>
<th>Small</th>
<th>Intermediate</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Ser, Thr, Cys, Gly, Pro, Ala, Val)</td>
<td>(His, Asp, Glu, Asn, Gln, Ile, Leu, Met)</td>
<td>(Arg, Lys, Phe, Tyr, Trp)</td>
<td></td>
</tr>
<tr>
<td>MbraCSP6</td>
<td>11</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>LlinCSP1</td>
<td>14</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>LlinCSP2</td>
<td>8</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>LlinCSP3</td>
<td>8</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>LlinCSP4</td>
<td>8</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>LlinCSP5</td>
<td>8</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>LlinCSP6</td>
<td>7</td>
<td>22</td>
<td>5</td>
</tr>
<tr>
<td>LlinCSP7</td>
<td>10</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>LlinCSP8</td>
<td>9</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>LlinCSP9</td>
<td>9</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>LlinCSP10</td>
<td>10</td>
<td>17</td>
<td>7</td>
</tr>
<tr>
<td>LlinCSP11</td>
<td>8</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>LlinCSP12</td>
<td>10</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>LlinCSP13</td>
<td>13</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>LlinCSP14</td>
<td>11</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>LlinCSP15</td>
<td>7</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>Avg</td>
<td>9.3</td>
<td>16.1</td>
<td>7.7</td>
</tr>
<tr>
<td>LhesCSP1</td>
<td>13</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>LhesCSP2</td>
<td>8</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>LhesCSP3</td>
<td>8</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>LhesCSP4</td>
<td>8</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>LhesCSP5</td>
<td>8</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>LhesCSP6</td>
<td>7</td>
<td>22</td>
<td>5</td>
</tr>
<tr>
<td>LhesCSP7</td>
<td>10</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>LhesCSP8</td>
<td>9</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>LhesCSP9</td>
<td>9</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>LhesCSP10</td>
<td>10</td>
<td>17</td>
<td>7</td>
</tr>
<tr>
<td>LhesCSP11</td>
<td>8</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>LhesCSP12</td>
<td>10</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>LhesCSP13</td>
<td>12</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>LhesCSP14</td>
<td>11</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>Avg</td>
<td>9.4</td>
<td>15.7</td>
<td>7.6</td>
</tr>
</tbody>
</table>

Note: Amino acids assessed are those predicted to be ≤5 Å from bound ligand in the MbraCSP6 binding pocket.
“typical” LlinCSP3/LhesCSP3, which is consistent with the higher predicted structural similarity. Coverage for the atypical LlinCSP6/LhesCSP6 was lowest of the three CSPs at ~ 50%. Although the resulting model lacked the seventh predicted helix, it did encompass the predicted binding pocket, consequently we concluded that the models, as well as those for LlinCSP1/LhesCSP1 and LlinCSP3/LhesCSP3, were of sufficient quality for the purposes of our study.

Comparison of the three structures revealed LlinCSP1/LhesCSP1 had the smallest solvent accessible surface area (5821 Å²), followed by LlinCSP6/LhesCSP6 (~6208 Å²), and then LlinCSP3/LhesCSP3 (6511 Å²), which is largely consistent with the predicted sizes of the proteins. The smaller surface area for the large LlinCSP6/LhesCSP6 is likely due to incomplete modeling as the structural template lacked the seventh helix. Indeed, no template with the necessary sequence homology was available in the databases to completely accommodate the extended LlinCSP6/LhesCSP6 conformation. All of the Lygus CSP structures exhibited the typical CSP fold with helices 1–2 and 4–5 forming a V-shaped structure capped at one end by a perpendicular helix 3 (Fig. 6). In LlinCSP3/LhesCSP3 and LlinCSP6/LhesCSP6, helix 6

---

**Fig. 6** Ribbon model representation of *Lygus* CSP structures. Homology-based conformations for a typical (LlinCSP3/LhesCSP3) and two atypical (LlinCSP1/LhesCSP1 and LlinCSP6/LhesCSP6) CSPs are shown along with the MbraCSP6 solution structure used to generate the respective models. Helical segments are color-coded: helix 1, light blue; helix 2, light purple; helix 3, turquoise; helix 4, orange; helix 5, light green, and helix 6, light pink. The upper panel shows different perspectives of the respective models with the front and bottom views indicated. The Tyr residue in helix 2 that is predicted to form the bottom of the channel is colored bright pink and depicted in stick mode. The Cys residues and accompanying disulfide bonds are colored yellow and are likewise depicted in stick mode. The lower panel shows differences in size of the putative binding pocket. The residues predicted to comprise the binding pocket, which is based on distance criteria (<5Å) generated from a ligand bound MbraCSP6 structure, are shown in purple and depicted in stick mode. The lack of a helix 6 and substitution of amino acids with smaller sidechains likely contributes to a larger binding pocket in LlinCSP1/LhesCSP1. Based on the spatial coordinates used to generate the respective models, the amino acid insertion between C3 and C4 in LlinCSP6/LhesCSP6 prevented formation of the characteristic disulfide bridge. LlinCSP1 and LhesCSP1 are 100% identical as is LlinCSP3 and LhesCSP3, thus only one model is shown for each set. Although LlinCSP6 and LhesCSP6 are 98% identical (four amino acid differences), the residues affected reside in portions of the sequence that were not modeled. The figure was created using Swiss-pbd viewer (Guex & Peitsch, 1997) and pov-ray (http://www.povray.org/).
is parallel to helix 3 and partially occludes the carboxyl terminal opening between helices 4–5. Consequently, the absence of the sixth helix in LinCSP1/LhesCSP1 creates a larger opening, which may allow accommodation of bulkier substrates. The Tyr in helix 2 (defined as Tyr26 in the MbraCSP6 model) that caps one the end of the ligand pocket (Campanacci et al., 2003) is rotated away from the pocket in LinCSP1/LhesCSP1 and LinCSP6/LhesCSP6, but is facing the pocket in LinCSP3/LhesCSP3 (Fig. 6). This difference could indicate that LinCSP1/LhesCSP1 and LinCSP6/LhesCSP6 accommodate longer/larger ligands. Unexpectedly, the second disulfide bridge linking C3 and C4 in LinCSP6/LhesCSP6 is not present in our model (Fig. 6). We speculate that the structural constraints used to generate the models limited the conformational space available to the respective Cys residues, which, because of the unique third amino acid insertion, oriented the sidechains > 5Å apart and thus prevented the linkage. While this almost certainly is not the case with the actual solution structure, it further highlights the divergence of the LinCSP6/LhesCSP6 sequences.

To further investigate potential differences in ligand binding among the three Lygus CSPs, we used the respective models to assess the effects that varied side chain sizes have on the putative binding pockets. As discussed above, the expansion of small residues in the LinCSP1/LhesCSP1 pocket (Table 3) resulted in a more compact structure with potentially greater ligand access than MbraCSP6 or LinCSP3/LhesCSP3 (Fig. 6). In contrast, the increased number of intermediate sized amino acids lining the LinCSP6/LhesCSP6 pocket yielded a more occluded pocket (Fig. 6). Consequently, it appears the variation in the ligand binding pocket residues affects not only the physicochemical properties of the respective CSPs, but also the size of potential ligands.

Discussion

To expand our knowledge of the Lygus chemosensory system, we used available transcriptome resources to identify 17 CSP-like sequences in L. lineolaris and 14 sequences in L. hesperus. Although the extent of the CSP repertoire varies depending on species, the number identified in the two Lygus species is comparable to that reported for other hemipterans: Acrythonephon pismun 12; Aphis gossypii 9; A. suturalis 8; A. lineolatus 8; N. lugens 11; S. furcifera 9; L. striatella 12; and A. lucorum 8. The increased number of CSP in the Lygus species compared to other mirids likely reflects methodological differences rather than gene expansion. The L. lineolaris dataset was generated using whole bodies across a range of developmental stages, and the L. hesperus assemblies were generated from intact adults. In contrast, studies in other mirids focused on either antenna (A. lineolatus and A. suturalis) or proboscis/leg (A. lucorum) specific transcriptomes/cDNA libraries. Consequently, expanding those datasets to include other tissues and/or developmental stages would likely reveal additional CSPs. Indeed, the absence of the evolutionarily conserved pentahelical CSPs (Fig. 4) suggests that those mirid datasets are incomplete.

Based on comparative analyses, the 14 L. hesperus sequences are likely orthologs of the L. lineolaris CSPs. Although LinCSP15 was amplified from L. lineolaris cDNAs using primers for LinCSP4/LhesCSP4, we were unable to amplify the transcript from L. hesperus. This could indicate that LinCSP15 transcription is spatially or temporally regulated. Alternatively, the absence of the transcript could reflect physiological adaptation and/or genetic differentiation of L. lineolaris. The LinCSP16 and LinCSP17 sequence fragments were likewise specific to the L. lineolaris datasets. However, the sequence similarity between LinCSP16/LinCSP2 and LinCSP17/LinCSP7, 8, and 14 could be indicative of alternative splicing, which has been reported for L. lineolaris OBPs (Hull et al., 2014b) and CSPs in Holotrichia parallelata (Ju et al., 2014), or alternatively an example of tissue specific CSP RNA editing as reported in B. mori (Xuan et al., 2014).

LinCSP1/LhesCSP1 and LinCSP13/LhesCSP13 are structurally and phylogenetically differentiated from the other Lygus CSPs. Unlike most CSPs, which are typically 100–135 amino acid polypeptides characterized by six helices, LinCSP1/LhesCSP1 and LinCSP13/LhesCSP13 are smaller pentahelical proteins. Furthermore, the conserved Tyr identified in MbraCSP6 (Tyr26 in the model, Tyr42 in the full-length sequence) that is critical for ligand binding (Campanacci et al., 2003) has been replaced with Gln. The disparate properties of these two residues could indicate different ligand interaction kinetics. Additionally, motif A in the pentahelical CSPs, which is predicted to contribute to potential protein-protein interactions, is poorly conserved (14% sequence identity) relative to the other Lygus CSPs (43%–100% identity). As a result, the residues predicted to comprise and/or surround the ligand-binding pocket are smaller (Table 3) and may thus facilitate binding of larger ligands (Fig. 6; Kulmuni & Havukainen, 2013). Indeed, A. mellifera CSP2, which sorts to the same monophyletic clade as LinCSP1/LhesCSP1 and LinCSP13/LhesCSP13, showed a preference for larger aromatic compounds (Dani et al., 2010). The presence of this clade in virtually all taxa in which CSPs have
been annotated (Forêt et al., 2007; Vieira & Rozas, 2011; Kulmuni & Havukainen, 2013; Pelosi et al., 2014) is consistent with the nonchemosensory role proposed by Kulmuni et al. (2013). In support of this, AmelCSP5 has been shown to function in embryonic development (Maleszka et al., 2007) and AmelCSP2 and BmorCSP16 (two other members of the pentahelical clade) are, like LlinCSP1, broadly expressed (Forêt et al., 2007; Qiao et al., 2013). Similarly, FlyAtlas data indicate broad expression for the pentahelical DmelCSP1 (Chintapalli et al., 2007). The discovery that all of the annotated CSPs in Daphnia pulex, an aquatic crustacean that is evolutionarily distant to insects, are pentahelical suggests that the pentahelical CSP clade pre-dates terrestrial colonization and has thus likely gained new functionalities in development and cellular homeostasis. The presence of two different genes encoding these CSPs in L. lineolaris and L. hesperus along with T. castaneum (Vieira & Rozas, 2011), A. mellifera (Wanner et al., 2004; Forêt et al., 2007; Vieira & Rozas, 2011), A. pismum (Vieira & Rozas, 2011), Anopheles gambiae (Vieira & Rozas, 2011), and Diaphorina citri (Wu et al., 2016) suggests that the lone genes reported in D. melanogaster (Vieira & Rozas, 2011) and Pediculus humanus (Vieira & Rozas, 2011) may be the result of gene loss.

In addition to the pentahelical CSPs, we identified a third atypical CSP (LlinCSP6/LhesCSP6) in both Lygus species that is larger than normal (22.5 kDa, 196 aa), has seven helices (Fig. S1), and deviates from the highly conserved C3-X2-C4 spacing motif. Although large CSPs (170–250 aa) have been identified in other species, they comprise only a fraction of the annotated sequences in the NCBI database. Curiously, the atypical CSPs in the database largely consist of lepidopterans (four species) and hemipterans (five species). Whether or not this distribution is indeed order specific or is an artifact arising from limited datasets remains to be fully explored. However, it is interesting that among the hemipteran species that comprised our phylogenetic dataset we found six CSPs of 170 aa or greater in length with three (NlugCSPP 189 aa, LstrCSPP 178 aa, and SfurCSP4 177 aa) forming a moderately supported clade (CSP-C) with LlinCSP6/LhesCSP6 (Fig. 4). While all three are predicted to be heptahelical, none exhibit the C3-X3-C4 deviation present in LlinCSP6/LhesCSP6. These CSPs (i.e., LlinCSP6, S. furcifera CSP4 and CSP7; L. striatella CSP6 and CSP8) tend to be leg dominant (Fig. 6 and Zhou et al., 2015), suggesting a potential role in contact chemoreception. This tissue distribution, however, is not shared by all large CSPs as the 21.6 kDa BmrCSP9 (ABH88202, also referred to as BmCSP10) is broadly expressed (Qiao et al., 2013) and undergoes insecticide-dependent upregulation in antennae and leg (Xuan et al., 2015). Similarly, the large CSPs in Aphis gossypii (CSP1 and CSP9) are expressed in multiple tissues and developmental stages (Gu et al., 2013).

Deviations from the CSP Cys spacing (C1-X2-C2-X18-C3-X2-C4) proposed by Wanner et al. (2004) were initially thought to be uncommon, but class/order-specific variations have since been reported. Orthopteran CSPs frequently have an insertion between C1 and C2 (C1-X2-C2-X18-C3-X2-C4), hymenopterans have an insertion between C2 and C3 (C1-X18-C2-X19-C3-X2-C4), and a CSP from Manduca sexta has a deletion between C2 and C3 (C1-X3-C2-X17-C3-X2-C4). Consequently, Xu et al. (2009) proposed order-specific motifs that follow the general C1-X5-3-C2-X15-19-C3-X2-C4 pattern with the C3–C4 spacing the most highly conserved. Surprisingly, LlinCSP6/LhesCSP6 have an extra amino acid between these two Cys, which yields a C3-X2-C4 motif (also see Figs. 1 and 2). Multiple independent clones amplified from each species confirmed that the insertion was not an artifact of transcriptome assembly. A search of available databases identified only two additional sequences, both derived from mirid species, with the C3-X2-C4 motif—a partial sequence from A. lineolatus (AlinCSP9, accession no. AMD02858) and an unannotated transcript shotgun assembly sequence (GASV02024394) from Notostira elongate. This deviation was not reported for CSPs from the mirids A. suturalis (Cui et al., 2017) and A. lucorum (Hua et al., 2012, 2013). The absence of LlinCSP6/LhesCSP6 orthologs in those datasets may reflect spatial and/or temporal specific expression. While the function of LlinCSP6/LhesCSP6 remains to be addressed, the predicted binding pocket is likely larger than typical CSPs (e.g., LlinCSP3/LhesCSP3), but smaller than the pentahelical CSPs, and is thus expected to accommodate a different range of substrate sizes (Table 3).

The phylogenetic positioning of Lygus CSPs with other mirid CSPs (Fig. 4) may provide insights into the nature of possible ligands. LlinCSP3/LhesCSP3, LlinCSP5/LhesCSP5, LlinCSP9/LhesCSP9, LlinCSP10/LhesCSP10, and LlinCSP11/LhesCSP11 aligned with A. lineolatus and A. lucorum CSPs. LlinCSP9/LhesCSP9 clustered with AlinCSP1 and AlucCSP1 (Fig. 4), which are predominantly expressed in the antenna (Gu et al., 2012; Hua et al., 2012) but exhibit different ligand binding profiles. AlinCSP1 binds a range of host plant volatiles released in response to herbivore damage (Gu et al., 2012), some of which have been shown to be attractive to L. hesperus females (Blackmer et al., 2004) and/or trigger positive electroantennograph responses (Williams et al., 2010). In contrast, AlucCSP1, which was cloned from a proboscis cDNA library, binds secondary cotton
metabolites (Hua et al., 2012), suggesting a gustatory role. LlinCSP10/LhesCSP10 aligned with AlinCSP3 and AlucCSP3. AlinCSP3 binds alfalfa and cotton volatiles (Gu et al., 2012), whereas AlucCSP3 binds secondary cotton metabolites (Hua et al., 2013). The two transcripts are also differentially expressed with AlinCSP3 most abundant in antennae (Gu et al., 2012) and AlucCSP3 in female wings (Hua et al., 2012). LlinCSP11/LhesCSP11 sorted with AlinCSP2 and AlucCSP7. AlinCSP2 is predominantly expressed in the antennae and preferentially binds multiple green plant volatiles including linalool (Gu et al., 2012), an aliphatic terpenoid reported to have repellency effects on Lygus males (Chinta et al., 1994; Williams et al., 2010). Tissue expression and functional characterization of the AlucCSP7 transcript have yet to be determined. LlinCSP5/LhesCSP5 and LlinCSP3/LhesCSP3 are potentially orthologs of AlinCSP5 and AlucCSP4, respectively. AlinCSP5 is an antenna dominant CSP with a relatively narrow ligand spectrum of green plant and cotton volatiles (Sun et al., 2015) that trigger L. hesperus female antennal responses (Williams et al., 2010). However, localization of AlinCSP5 to the outer sensillum lymph of short sensilla basiconica, which lack neuron dendrites, suggests that it may function as an odorant sink (Sun et al., 2015). AlucCSP4 is highly expressed in female wings and antennae, and, like the other characterized AlucCSPs, preferentially binds secondary cotton metabolites (Hua et al., 2013). LlinCSP7/LhesCSP7 sorted to clade CSP-B with AlinCSP4, an antenna dominant CSP that binds multiple compounds including various cotton and green plant volatiles as well as components of the A. lineolatus sex pheromone blend (Sun et al., 2015), some of which (i.e., trans-2-hexenyl-butyrate and hexyl butyrate) are also active in L. lineolaris and L. hesperus (Byers et al., 2013). While the potential ligand specificities of the orthologous sequences are intriguing, conclusions will require functional analyses of the Lygus CSPs. Furthermore, the variations in ligand binding reported in the A. lineolatus and A. lucorum studies (Gu et al., 2012; Hua et al., 2012, 2013; Sun et al., 2015) may be a result of methodological differences (the use of different fluorescent binding pocket probes) as opposed to actual biological/physiological differentiation of the phylogenetically related CSPs.

Acknowledgments

We thank Calvin A. Pierce III (USDA-ARS SIMRU) for assistance with insect rearing, RNA extractions, and cDNA synthesis. We also thank Lynn Jech (USDA-ARS ALARC) for assistance with plasmid isolation and sequencing. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U. S. Department of Agriculture. USDA is an equal opportunity provider and employer.

Disclosures

The authors declare no conflicts of interest.

References


© 2018 Institute of Zoology, Chinese Academy of Sciences, 00, 1–21


© 2018 Institute of Zoology, Chinese Academy of Sciences, 00, 1–21


Liu, G.X., Ma, H.M., Xie, H.Y., Xuan, N., Guo, X., Fan, Z.X. *et al.* (2016b) Biotype characterization, developmental profiling, insecticide response and binding property of *Bemisia*


Manuscript received August 13, 2018
Final version received November 21, 2018
Accepted November 26, 2018

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**Fig. S1.** Secondary structure prediction of *L. lineolaris* and *L. hesperus* CSPs. Predictions performed using JPRED4 (Drozdetskiy et al., 2015) (upper blue) and YASPIN (Lin et al., 2005) (lower red and orange). Putative signal peptides are shown boxed, H indicates predicted helical regions, E indicates predicted strand regions.

**Table S1.** Accession numbers of sequences used in phylogenetic analyses.

**Table S2.** Oligonucleotide primers used in this study.

**Table S3.** Heat map of *L. lineolaris* and *L. hesperus* CSP percent amino identities.

**Table S4.** TOP BLASTx hits for putative *L. lineolaris* CSPs.

**Table S5.** TOP BLASTx hits for putative *L. hesperus* CSPs.