Abstract

Symbiosis is well known to influence bacterial symbiont genome evolution and has recently been shown to shape eukaryotic host genomes. Intriguing patterns of host genome evolution, including remarkable numbers of gene duplications, have been observed in the pea aphid, a sap-feeding insect that relies on a bacterial endosymbiont for amino acid provisioning. Previously, we proposed that gene duplication has been important for the evolution of symbiosis based on aphid-specific gene duplication in amino acid transporters (AATs), with some paralogs highly expressed in the cells housing symbionts (bacteriocytes). Here, we use a comparative approach to test the role of gene duplication in enabling recruitment of AATs to bacteriocytes. Using genomic and transcriptomic data, we annotate AATs from sap-feeding and non sap-feeding insects and find that, like aphids, AAT gene families have undergone independent large-scale gene duplications in three of four additional sap-feeding insects. RNA-seq differential expression data indicate that, like aphids, the sap-feeding citrus mealybug possesses several lineage-specific bacteriocyte-enriched paralogs. Further, differential expression data combined with quantitative PCR support independent evolution of bacteriocyte enrichment in sap-feeding insect AATs. Although these data indicate that gene duplication is not necessary to initiate host/symbiont amino acid exchange, they support a role for gene duplication in enabling AATs to mediate novel host/symbiont interactions broadly in the sap-feeding suborder Sternorrhyncha. In combination with recent studies on other symbiotic systems, gene duplication is emerging as a general pattern in host genome evolution.

Keywords: aphid, bacteriocyte, functional evolution, gene duplication, mealybug, sap-feeding insect

Received 10 July 2013; revision received 3 December 2013; accepted 8 December 2013
example, bacterial nutritional endosymbionts have undergone drastic genome reduction and gene loss in response to evolving an obligate endosymbiotic lifestyle (Shigenobu et al. 2000; Nakabachi et al. 2006; McCutcheon & Moran 2007, 2012; McCutcheon et al. 2009; Sabree et al. 2009, 2012a; McCutcheon & von Dohlen 2011; Nikoh et al. 2011; Sloan & Moran 2012; Bennett & Moran 2013). Historically, symbiont genomes have received more attention than the genomes of their hosts, but as deep sequencing becomes cheaper and assembly technology advances, host genomes are providing insight into how symbiosis shapes genomes in eukaryotic hosts (International Aphid Genomics Consortium 2010; Kirkness et al. 2010; Nygaard et al. 2011; Young et al. 2011; Husnik et al. 2013). Four interesting and novel (given current sampling) features of host genomes include (i) metabolic complementarity with symbionts in essential nutrient biosynthesis (Shigenobu et al. 2000; Wilson et al. 2010; Hansen & Moran 2011; McCutcheon & von Dohlen 2011; Nygaard et al. 2011; Husnik et al. 2013), (ii) loss or modulation of immune pathways (Gerardo et al. 2010; Kim et al. 2011b; Ratzka et al. 2013), (iii) maintenance and expression of functional genes acquired horizontally from bacteria other than the obligate symbiont (Nikoh & Nakabachi 2009; Nikoh et al. 2010; Husnik et al. 2013) and (iv) duplication of genes with functions that may facilitate symbiosis (Ganot et al. 2011; Price et al. 2011; Young et al. 2011; Shigenobu & Stern 2013). Although these features suggest a role for symbiosis in shaping host genomes, some genomic attributes of eukaryotic hosts come from isolated examples and a role for symbiosis in their evolutionary origin remains untested. One way to test the role of symbiosis in shaping host genome evolution is by evaluating specific genomic traits within an evolutionary framework.

An evolutionary framework is especially powerful in evaluating the extensive gene duplication and differential expression of amino acid transporters (AATs) observed in the genome of the pea aphid, *Acyrthosiphon pisum*. This evolutionary pattern may be influenced by the relationship between *A. pisum* and its obligate bacterial endosymbiont, *Buchnera aphidicola*. *A. pisum*, a member of the insect order Hemiptera, feeds on plant phloem sap, a diet deficient in key nutrients such as essential amino acids (Douglas 1993, 2006; Sandstrom & Pettersson 1994; Wilkinson & Douglas 2003). Essential amino acids – that is, amino acids that animals are unable to synthesize *de novo* – are provided to aphids by *Buchnera* in exchange for nonessential amino acids (Shigenobu et al. 2000). Supply of nonessential amino acids to *Buchnera* and distribution of essential amino acids from *Buchnera* to host tissues is mediated by amino acid transport across three key membrane barriers that we collectively refer to as the symbiotic interface: (i) the plasma membrane of the specialized aphid cells that house *Buchnera* (bacteriocytes), (ii) the host-derived symbiosomal membrane surrounding individual *Buchnera* cells and (iii) the bacterial inner and outer membranes of individual *Buchnera* (Shigenobu & Wilson 2011). Analyses of transcripts (Hansen & Moran 2011; Price et al. 2011; Macdonald et al. 2012) and proteins (Poliakov et al. 2011) enriched in aphid bacteriocytes suggest that amino acid flux at the aphid/*Buchnera* symbiotic interface is mediated by several aphid AATs from two gene families: the amino acid polyamine organocation (APC) family (transporter classification (TC) #2.A.3) and the amino acid-auxin permease (AAAP) family (TC #2.A.18) (Castagna et al. 1997; Saier 2000; Saier et al. 2006, 2009). These two AAT families play important nutritional roles in insects (Martin et al. 2000; Dubrovsky et al. 2002; Colombani et al. 2003; Jin et al. 2003; Goberdhan et al. 2005; Attardo et al. 2006; Evans et al. 2009). Some aphid AATs enriched in bacteriocytes are paralogs derived from within an aphid-specific gene expansion. The membership of bacteriocyte-enriched AATs to an aphid-specific expansion intrigues us because gene duplication can be a critical source of raw genetic material for evolutionary innovation. While gene duplication is random, duplicates can be maintained in a genome for many reasons, including the evolution of novel functions and/or the spatiotemporal partitioning of ancestral function across paralogs (reviewed in Kondrashov 2012). Finding AAT gene duplicates with enriched expression in bacteriocytes is consistent with the hypothesis that gene duplication plays an important, possibly adaptive, role in recruiting AATs to the symbiotic interface of aphids and other sap-feeders (Price et al. 2011). This hypothesis predicts that other sap-feeders with obligate bacterial endosymbionts also maintain duplicated AATs with similar patterns of bacteriocyte enrichment.

Most sap-feeding insects are hemipterans, and thus, testing the role of gene duplication in recruiting AATs to the symbiotic interface can be facilitated with genomic data from sap-feeding and non sap-feeding hemipteran taxa. Despite difficulties resolving higher-level hemipteran relationships (Campbell et al. 1995; von Dohlen & Moran 1995; Grimaldi & Engel 2005; Cryan & Urban 2011; Song et al. 2012), current understanding of hemipteran suborders can facilitate the selection of appropriate taxa to evaluate whether symbiosis influences AAT evolution. Ideal taxon sampling will span the three major hemipteran suborders of Sternorrhyncha, Auchenorrhyncha and Heteroptera (see Fig. 1). Sternorrhyncha, the suborder that includes aphids, will enable the determination of whether the AAT duplications we discovered in the pea aphid (Price et al. 2011)
Bactericera cockerelli, the kissing bug unrelated to diet. A transcriptome is available for a test of whether gene duplication in AATs is sap-feeders, and inclusion of this suborder will provide evolutionary patterns. Here, we generate a transcriptome for an auchenorrhyan, the cicada sister to kissing bug (consistent with Hennig 1981; Song et al. 2012). Lastly, Heteroptera comprises mostly non-sap-feeders, and inclusion of this suborder will provide tests of independence (Weber & Agrawal 2012) in AAT evolution. Together, these data indicate that gene duplication has broadly played a role in recruiting amino acid transporters to operate at the symbiotic interface of sternorrhynchan taxa.

Materials and methods

Insect collection and cultivation

Adult female cicadas (Diceroprocta semicincta) were collected in Tucson, AZ, and preserved in RNAlater (Ambion). Citrus mealybugs (Planococcus citri) were collected from coleus plants in the Utah State University greenhouse in Logan, Utah (von Dohlen et al. 2001; McCutcheon & von Dohlen 2011), and raised on coleus plants in the laboratory at 25 °C. Pea aphids (Acyrthosiphon pisum) from the genome line LSR1 (Caillaud et al. 2002) were raised on fava plants at 20 °C. Both insect colonies were maintained under a photoperiodicity of 16:8 (L:D).

Transcriptome sequencing and assembly

For cicada transcriptomes, total RNA was purified from either (i) bacteriocytes or (ii) a combination of head, legs and wing muscles (hereafter referred to as ‘insect’) dissected from RNAlater-preserved adult female cicadas according to the manufacturer’s protocols (MoBio PowerBiofilm RNA Isolation Kit). RNA was sent to Hudson Alpha Institute for Biotechnology for barcoded library preparation and Illumina HiSeq sequencing. Paired-end 100-nt reads were filtered to a minimum quality of 20 over 95% of the read, and 5 nt were trimmed from the 5’ end. Insect (42 688 895 read pairs) and bacteriocyte (53 510 432 read pairs) reads were assembled into separate insect and bacteriocyte transcriptomes in Trinity (25 January 2012 release) (Haas et al. 2013) using kmer_length = 25 and min_contig_length = 48.

Mealybug whole body, paired-end, 100-nt reads (Husnik et al. 2013) from a mixed population of adult and penultimate instar females were filtered to a minimum quality of 30 over 95% of the read. The resulting 58 812 530 read pairs were assembled with two different assembly packages. First, reads were assembled in...
**Table 1** Hemipteran taxa and associated symbionts

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Diet</th>
<th>Obligate symbiont(s)</th>
<th>Symbiont classification</th>
<th>Symbiont localization</th>
<th>Symbiosomal membrane</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sternorrhyncha</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pea aphid</td>
<td>Phloem sap</td>
<td>Buchnera aphidicola^a</td>
<td>γ-Proteobacteria</td>
<td>Bacteriocytes</td>
<td>Yes</td>
</tr>
<tr>
<td>Citrus mealybug</td>
<td>Phloem sap</td>
<td>Tremblaya princeps^b</td>
<td>β-Proteobacteria</td>
<td>Bacteriocytes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Moranella endobia^c</td>
<td>γ-Proteobacteria</td>
<td>Nested within</td>
<td>Not applicable</td>
</tr>
<tr>
<td>Whitefly</td>
<td>Phloem sap</td>
<td>Portiera aleyroidarum^a</td>
<td>γ-Proteobacteria</td>
<td>Bacteriocytes</td>
<td>Yes</td>
</tr>
<tr>
<td>Potato psyllid</td>
<td>Phloem sap</td>
<td>Carsonella ruddii^d</td>
<td>γ-Proteobacteria</td>
<td>Bacteriocytes</td>
<td>Yes</td>
</tr>
<tr>
<td>Auchenorrhyncha</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cicada</td>
<td>Xylem sap</td>
<td>Sulcia muelleri^e</td>
<td>Bacteroidetes</td>
<td>Bacteriocytes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hodgkinia cicadica^f</td>
<td>α-Proteobacteria</td>
<td>Bacteriocytes</td>
<td>Yes</td>
</tr>
<tr>
<td>Heteroptera</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kissing bug</td>
<td>Vertebrate blood</td>
<td>Rhodococcus rhodnii^g</td>
<td>Actinobacteria</td>
<td>Gut lumen</td>
<td>No</td>
</tr>
</tbody>
</table>

References: (Munson et al. 1991)^a; (Thao et al. 2002)^b; (McCutcheon & von Dohlen 2011)^c; (von Dohlen et al. 2001)^d; (Thao & Baumann 2004)^e; (Thao et al. 2000)^f; (Moran et al. 2005)^g; (McCutcheon et al. 2009)^h; (Goodfellow & Alderson 1977)^i.

**VELVET** (v.1.2) (Zerbino & Birney 2008) and **OASES** (v.0.2) (Schulz et al. 2012) using variable k-mer lengths (between 33 and 63 nt), and resulting assemblies were merged into one master assembly. Second, reads were assembled in **TRINITY** using default parameters (**kmer_length** = 25).

The whitefly (Bemisia tabaci) transcriptome was re-assembled using 170 884 234 RNA-seq read pairs from a mixed population of adult males and females (NCBI BioProject PRJNA89143). Reads were assembled with RNA assemblers **VELVET**/**OASES** v.1.2.03/v.0.2.06 (2012.02), **SOAPDENOVO-TRANS** v.2011.12.22 and **TRINITY** (17 March 2012 release), using multiple options. **EVIDENTIALGENE TR2AACDS** pipeline software was used to process the many resulting assemblies by coding sequences, translate to proteins, score gene evidence and classify/reduce to a biologically informative transcriptome of primary and alternate transcripts. The gene set is publicly available at http://arthropods.eugenes.org/EvidentialGene/arthropods/whitefly/whitefly1eg6/.

**De novo identification of hemipteran amino acid transporters**

Amino acid transporters (AATs) were identified using **HMMER** (v.3.0) (Eddy 2009) from transcriptomes of cicada, mealybug, whitefly, the potato psyllid Bactericera cockerelli (Nachappa et al. 2012; mixed population of adult males and females) and the kissing bug Rhodnius prolixus (NCBI BioProject PRJNA191820; mixed developmental stages and sexes). Briefly, using a stand-alone PERL script underlying the open reading frame (ORF) prediction webserver hosted by the Proteomics/Genomics Research Group at Youngstown State University (http://proteomics.ysu.edu/tools/OrfPredictor.html), transcripts were translated into all six reading frames. As described previously (Price et al. 2011), translated transcripts were searched for conserved functional domains associated with the APC (TC # 2.A.3) and AAAP (TC # 2.A.18) families of amino acid transporters (Castagna et al. 1997; Saier 2000; Saier et al. 2006, 2009) in **HMMER** v.3.0 (Eddy 2009; Finn et al. 2011). Transcripts significantly matching APC or AAAP domains (e ≤ 0.001) were verified by BLASTX searches against the NCBI refseq database and retained for further analyses if they showed a significant (e ≤ 0.001) similarity to the APC or AAAP sequences from the fruit fly Drosophila melanogaster and/or A. pisum.

Alleles and splice variants were collapsed into a conservative set of representative transcripts for each insect by one of the following two methods depending on availability of genome sequence data: (i) draft genome assemblies are available for mealybugs (Husnik et al. 2013) and kissing bugs (unpublished; hosted at vectorbase.org and NCBI), so we validated loci by mapping transcripts to genomic scaffolds by BLASTN searches. Of the transcripts mapping to the same region of a particular scaffold(s), the transcript encoding the longest protein was kept to represent the gene locus. In a few cases, 2-3 partial transcripts were merged into a single locus for phylogenetic analyses (Tables S1–S4 in Appendix S1, Supporting Information). In all cases, partial transcripts mapped side by side to genomic scaffolds on the same strand. Additionally in all cases, the partial transcript mapping upstream in the genome aligned to the 5’ end of other, full-length AAT loci and the downstream partial transcript aligned to the 3’ end. (ii) In contrast, whiteflies, psyllids and cicadas lack draft genome assemblies. In these insects, transcripts that have been diverging for a short period of time were...
were estimated in RAXML v.7.2.8 (Stamatakis 2006; Ott 2009) and Bayesian methods. ML phylogenies were reconstructed in MrBayes v.3.1.2 (Huelsenbeck & Ronquist 2001; Ronquist & Huelsenbeck 2003) using two runs with 4 chains per run. The LG protein substitution matrix is not available in MrBayes, so phylogenies were inferred using WAG+G. Analyses were allowed to run until the standard deviation of split frequencies between runs dropped below 0.05. Convergence of estimated parameters was confirmed in TRACER v.1.5 (Rambaut & Drummond 2007) and of topology in AWTY (Nylander et al. 2008), assuming a burn-in of 10% of generations. The criteria supported convergence, so the first 10% of generations were discarded and phylogenies sampled in the remaining generations were used to estimate a 50% majority-rule consensus tree.

The AAAP family contained a large amount of sequence divergence, preventing convergence of the Markov chain Monte Carlo (MCMC) in Bayesian phylogenetic analyses. Therefore, we estimated the phylogeny of a reduced set of AAAP genes corresponding to a monophyletic clade supported in a preliminary maximum-likelihood analysis (Fig. S1 in Appendix S1, Supporting Information).

**Gene conversion analyses**

Lineage-specific AAT expansions were assessed for the possibility of gene conversion using the program GENECONV (Sawyer 1989). Codon alignments were produced by the CLUSTALW plugin of SEAVIEW (Gouy et al. 1999) and run in GENECONV using three different mismatch penalties, g0, g1 and g2. Applying different mismatch penalties to the analysis facilitates the identification of recent gene conversion and ancient gene conversion that may be partially masked by the accumulation of different substitutions between paralogs.

**Expression analysis by quantitative reverse transcriptase PCR**

Expression profiles of select AATs were measured by quantitative reverse transcriptase PCR (qRT-PCR) in whole bodies and bacteriocytes of adult female LSR1 pea aphids, a mixture of adult and penultimate female citrus mealybugs and adult female potato psyllids from the same colonies used for the potato psyllid transcriptome (Nachappa et al. 2012). Bacteriocytes were dissected from 100 female aphids, mealybugs or psyllids in 0.9% RNase-free NaCl and immediately stabilized by placing in TRI Reagent (Ambion). Total RNA was extracted from dissected bacteriocytes and whole female

Phylogenetic analyses

Gene phylogenies for the APC and AAAP amino acid transporter families were estimated using sequences from citrus mealybug, potato psyllid, whitefly, cicada and kissing bug as well as previously annotated AATs (Price et al. 2011) from the pea aphid, the human body louse (Pediculus humanus), the fruit fly (D. melanogaster), a tick (Ixodes scapularis) and humans (Homo sapiens). Outgroup sequences were aphid and/or fruit fly genes closely related to APC and AAAP gene families and members of the same transporter superfamly (Price et al. 2011). Full-length protein sequences were aligned in MAFFT (Katoh et al. 2002) using default parameters, and resulting alignments were trimmed in TRIMAL v.1.2 (Capella-Gutierrez et al. 2009) using a gap threshold of 25%.

Phylogenies were estimated using maximum-likelihood (ML) and Bayesian methods. ML phylogenies were estimated in RAXML v.7.2.8 (Stamatakis 2006; Ott et al. 2007) using the protein evolution model LG+G (the best-fit model as determined by PROTEST v.2.4 using the Akaike Information Criterion (Abascal et al. 2005)) and the fast bootstrap option. The number of bootstrap replicates for each analysis was chosen by the bootstrap convergence criterion ‘autofc’ implemented in RAXML. Bayesian phylogenies were reconstructed in MrBayes v.3.1.2 (Huelsenbeck & Ronquist 2001; Ronquist & Huelsenbeck 2003) using two runs with 4 chains per run. The LG protein substitution matrix is not available in MrBayes, so phylogenies were inferred using WAG+G. Analyses were allowed to run until the standard deviation of split frequencies between runs dropped below 0.05. Convergence of estimated parameters was confirmed in TRACER v.1.5 (Rambaut & Drummond 2007) and of topology in AWTY (Nylander et al. 2008), assuming a burn-in of 10% of generations. The criteria supported convergence, so the first 10% of generations were discarded and phylogenies sampled in the remaining generations were used to estimate a 50% majority-rule consensus tree.

The AAAP family contained a large amount of sequence divergence, preventing convergence of the Markov chain Monte Carlo (MCMC) in Bayesian phylogenetic analyses. Therefore, we estimated the phylogeny of a reduced set of AAAP genes corresponding to a monophyletic clade supported in a preliminary maximum-likelihood analysis (Fig. S1 in Appendix S1, Supporting Information).

Gene conversion analyses

Lineage-specific AAT expansions were assessed for the possibility of gene conversion using the program GENECONV (Sawyer 1989). Codon alignments were produced by the CLUSTALW plugin of SEAVIEW (Gouy et al. 2011) and run in GENECONV using three different mismatch penalties, g0, g1 and g2. Applying different mismatch penalties to the analysis facilitates the identification of recent gene conversion and ancient gene conversion that may be partially masked by the accumulation of different substitutions between paralogs.

Expression analysis by quantitative reverse transcriptase PCR

Expression profiles of select AATs were measured by quantitative reverse transcriptase PCR (qRT-PCR) in whole bodies and bacteriocytes of adult female LSR1 pea aphids, a mixture of adult and penultimate female citrus mealybugs and adult female potato psyllids from the same colonies used for the potato psyllid transcriptome (Nachappa et al. 2012). Bacteriocytes were dissected from 100 female aphids, mealybugs or psyllids in 0.9% RNase-free NaCl and immediately stabilized by placing in TRI Reagent (Ambion). Total RNA was extracted from dissected bacteriocytes and whole female

© 2014 John Wiley & Sons Ltd
bodies of each insect using the TRI Reagent procedure (Ambion), treated with DNase I in solution and cleaned up using the RNaseasy Mini Kit (Qiagen). First-strand cDNA was synthesized from 500 ng of RNA from each tissue, using qScript cDNA Supermix (Quanta Biosciences) and following the manufacturer’s protocol. qRT–PCR assays were performed as previously described (Price et al. 2011) using one biological replicate and three technical replicates for each gene/tissue. Primers were subject to BLASTN searches against genomic and/or transcriptomic data sets using a word length of 7, an expect threshold of 1000 and without the low-complexity filter. In all cases, only the target sequence was returned as a hit for each pair of forward and reverse primers. To confirm that primers amplified only one locus, we analysed melt curves from our qRT–PCR results. With the exception of one gene, which was discarded from analysis, all melt curves showed one clear peak, indicating a single product. No template controls and no reverse transcriptase controls (controlling for RNA contaminated with gDNA) were run in parallel with unknown samples. Identifiers, sequences, amplification efficiency and optimization details for primers used in qRT–PCR assays are listed in Table S6 (Appendix S1, Supporting Information). Expression for target genes within a particular insect was compared between whole insect and bacteriocytes using $2^{- \Delta \text{ACT}}$ methodology (Livak & Schmittgen 2001) with expression normalized to either glyceraldehyde-3-phosphate dehydrogenase (GAPDH) in aphids or the 60S ribosomal protein L7 (RPL7) in mealybugs and psyllids. Expression data within each insect were collectively normalized by converting $\Delta \text{CT}$ to $z$-scores as follows:

$$z = -10 \times \left( \frac{\Delta \text{CT} - \Delta \text{CT}_{\text{ref}}}{\sigma_{\Delta \text{CT}}} \right)$$

Normalized expression values were compiled into a heat map where $z > 0$ (high expression) is represented as yellow and $z < 0$ (low expression) is represented as blue.

**Differential expression quantification**

Global differential expression between mealybug insect and bacteriocyte tissues was quantified for mealybug AATs using the whole body transcriptome data from this study and previously published bacteriocyte transcriptome data (Husnik et al. 2013). Differential expression analyses were conducted with the PERL script pipeline implemented in TRINITY. Briefly, raw RNA-seq reads were mapped to transcripts using BOWTIE v.0.12.7 (Langmead et al. 2009), and mapped reads were counted by RSEM v.1.1.18 (Li & Dewey 2011). Data were normalized by TMM (trimmed mean of M values), and transcripts significantly differentially expressed between whole body and bacteriocytes were identified using the Bioconductor package edger v.2.10 (Robinson et al. 2010). Digital expression values of differentially expressed transcripts are presented in Appendix S3 (Supporting Information) as ‘fragments per kilobase of exon per million fragments mapped’ (FPKM). Differential expression was not quantified for cicada bacteriome vs. insect tissues because cicadas lacked gene duplications.

**Results and Discussion**

*Nutrient amino acid transporter families are expanded in the Sternorrhyncha*

Consistent with our pea aphid work (Price et al. 2011), all sternorrhynchan hemipterans we sampled (Table 1, Fig. 1) possessed expanded amino acid transporter (AAT) families relative to non sap-feeding insects (kissing bug, human body louse and the fruit fly) (Table 2). In particular, citrus mealybugs, potato psyllids and whiteflies possessed 36-38 AAT loci across both gene families; relatively large AAT numbers compared with the 20 AAT loci in the non sap-feeding hemipteran annotated here (kissing bug; Table 2) and 22-28 AAT loci in other insects annotated by Price et al. (2011) (the fruit fly *D. melanogaster*, the body louse *P. humanus*, the honey bee *Apis mellifera*, the flour beetle *Tribolium castaneum*, the silkworm moth *Bombyx mori*, the wasp *Nasonia vitripennis* and the mosquito *Anopheles gambiae*). In contrast, we identified only 26 AAT loci in cicada, a sap-feeder belonging to the hemipteran suborder Auchenorrhyncha (Table 1, Fig. 1).

<table>
<thead>
<tr>
<th>Amino acid transporters in sampled insects</th>
<th>APC Loci</th>
<th>AAAP Loci</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pea aphid</td>
<td>18</td>
<td>22</td>
<td>40</td>
</tr>
<tr>
<td>Citrus mealybug</td>
<td>10</td>
<td>28</td>
<td>38</td>
</tr>
<tr>
<td>Whitefly</td>
<td>12</td>
<td>24</td>
<td>36</td>
</tr>
<tr>
<td>Potato psyllid</td>
<td>13</td>
<td>25</td>
<td>38</td>
</tr>
<tr>
<td>Cicada</td>
<td>10</td>
<td>16</td>
<td>26</td>
</tr>
<tr>
<td>Kissing bug</td>
<td>7</td>
<td>13</td>
<td>20</td>
</tr>
<tr>
<td>Human body louse</td>
<td>8</td>
<td>13</td>
<td>21</td>
</tr>
<tr>
<td>Fruit fly</td>
<td>10</td>
<td>17</td>
<td>27</td>
</tr>
</tbody>
</table>

$^a$Distinct loci confirmed by mapping transcripts to genomic scaffolds (Price et al. 2011).

$^b$Distinct loci confirmed by mapping transcripts to genomic scaffolds (this study).

$^c$Estimated number of loci based on the rate of synonymous substitutions (dS) between paralogs being greater than 0.25.
Amino acid transporter expansions in sap-feeding insects result from both ancient and recent gene duplication events

To clarify the evolutionary mechanism and timing of events that led to AATs expanding in sap-feeding insects, we estimated phylogenies for the APC and AAAP amino acid transporter families (Figs 2a and 3a). The phylogenies revealed that gene duplications occurred on two timescales. First, two ancient gene duplication events (one in each gene family) pre-date hemipteran diversification (marked by pale green boxes in Figs 2a and 3a). Second, consistent with our previous observation in aphids (Price et al. 2011), multiple, more recent, gene duplications occurred independently in sternorrhynchan taxa following their divergence from a common ancestor (marked by grey boxes in Figs 2a and 3a). In contrast, our analyses failed to support any Auchenorrhyncha-specific gene duplications in either AAT family. That said, in four instances (marked in Figs 2a and 3a with asterisks), we found phylogenetic support for close relationships between 2-3 cicada (Auchenorrhyncha) loci and one kissing bug (Heteroptera) locus. Two scenarios could explain these close relationships. First, AAT duplication could have taken place independently in the lineage leading to cicadas and no gene duplication took place in kissing bugs, but sequence similarity among orthologs prevents resolution of cicada-specific clades. Second, assuming species tree B (Fig. 1b), gene duplications took place in the common ancestor of cicadas and kissing bugs, but paralogs were only retained in cicada. Of the two scenarios, the second is the least parsimonious, requiring that cicadas retain their paralogs and that kissing bugs lose all but one paralog in three independent instances.

In our pea aphid work (Price et al. 2011), we found that aphid AAT paralogs were tandemly arrayed in the genome. Although new AATs in this study were annotated from transcriptome data, a draft genome assembly for the citrus mealybug (Husnik et al. 2013) enabled us to preliminarily assess paralog arrangement in that genome. In the mealybug AAAP expansion (Fig. 3a), three pairs of paralogs map to different regions of the same scaffold within ~4 kbp or less of each other (Fig. 4, Table S2 in Appendix S1, Supporting Information), indicating that these paralogs are tandemly arrayed in the mealybug genome. These tandemly arrayed paralogs thus resulted from localized gene duplication (as opposed to whole genome duplication). No other mealybug AAT loci shared a scaffold (Tables S1-S2 in Appendix S1, Supporting Information), which could at least in part be due to the poor quality of the assembly (Husnik et al. 2013). In the kissing bug genome, several transcripts mapped to the same scaffold (Tables S3-S4 in Appendix S1, Supporting Information), but were usually separated by large genomic regions between 19 kbp and 1.2 Mbp. The only exception was that two loci were separated by 5.7 kbp (Tables S3-S4 in Appendix S1, Supporting Information). Despite the short distance between those two loci, our phylogeny (Fig. 3a) indicates that they did not result from a recent gene duplication event in the lineage leading to kissing bugs.

Amino acid transporter evolution within the Sternorrhyncha

One unexpected result of this work is finding that AATs have undergone gene family expansions independently in each of the sternorrhynchans we sampled (aphids, mealybugs, psyllids and whiteflies). Consistent molecular and morphological phylogenetic support for the monophyly of Sternorrhyncha (Hennig 1981; Campbell et al. 1995; von Dohlen & Moran 1995; Grimald & Engel 2005; Cryan & Urban 2011; Song et al. 2012) indicates that aphids, mealybugs, whiteflies and psyllids inherited sap-feeding from their common ancestor. We thus assume that the common ancestor also had an amino acid-provisioning symbiont that was later replaced in three, or perhaps all four, lineages we sampled (explaining why each lineage has a different symbiont, Table 1). Importantly, this common ancestor required that AATs mediate host/symbiont amino acid exchange. Retention of independently duplicated paralogs in sternorrhynchans could be explained in four ways. First, our understanding that symbiosis pre-dates sternorrhynchan diversification could be wrong, and each lineage independently evolved symbiosis and comparable symbiotic interfaces. Second, the importance of different transporters could depend on the symbiont lineage. Third, some AAT gene duplications in these taxa could appear to be more recent than they truly are if tandem arrays of paralogs have undergone concerted evolution through gene conversion or nonhomologous crossing-over after the major sternorrhynchan lineages (aphids, mealybugs and other scale insects, whiteflies and psyllids) began diversifying (e.g. Colbourne et al. 2011). We found evidence of gene conversion only among a few paralogs in aphids and whiteflies (Table S5 in Appendix S1, Supporting Information), indicating that AAT paralogs are largely evolving independently of one another. However, we cannot rule out the possibility that gene conversion played a more important role in paralog diversification at some time in the past. Fourth, consistent with our previous discovery of male-biased AAT paralogs in aphids (Duncan et al. 2011), many paralogs are probably retained in sternorrhynchan genomes for lineage-specific roles not related to symbiosis. Most aphid and mealybug AAT paralogs are
Psyllid-specific expansion
Mealybug-specific expansion
Aphid-specific expansion
Hemipteran-specific duplication (Fig. 2B)

Posterior probability:
- ≥ 0.95
- 0.80–0.95
- < 0.80

Bootstrap support:
- ≥ 95%
- 80–95%
- < 80%
Data not available
Ambiguous expression

Differential expression (fill):
- > 2-fold upregulated in bacteriocyte
- > 2-fold downregulated in bacteriocyte
- No difference
- Data not available

Differential expression source (outline):
- This study only
- This study and previous study
- Previous study only

Outgroups:
- Human
- Tick
- Body louse
- Fruit fly
- Pea aphid
- Citrus mealybug
- Potato psyllid
- Whitefly
- Cicada
- Kissing bug

Hemipterans:
- ≥ 0.95
- 0.80–0.95
- < 0.80

© 2014 John Wiley & Sons Ltd
not enriched in bacteriocytes, supporting a role for non-symbiotic factors in driving the maintenance of AAT paralogs in these insects. Given the important role that AATs play broadly in animals, it is not surprising that AAT paralog maintenance in sternorrhynchan genomes is not only driven by symbiosis. For example, nutrient AATs also mediate amino acid uptake from the gut into hemolymph (insect blood) (Colombani et al. 2003; Morris et al. 2009; Price et al. 2011). Further, some nutrient AATs play a role in nutrient sensing (Colombani et al. 2003; Attardo et al. 2006). Lastly, some AATs transport neurotransmitters, likely explaining their expression in aphid heads (Price et al. 2011). Accepting their many roles, it is not surprising that some insects without intracellular, amino acid-provisioning symbionts maintain lineage-specific AAT duplications (e.g. Fig 3 and Price et al. 2011). However, that sternorrhynchan sap-feeding insects maintain more AAT duplications in their genomes than other, non sap-feeding insects is compelling and suggests that gene duplication has facilitated AAT recruitment to bacteriocytes – at least in the Sternorrhyncha. Nevertheless, the absence of duplicates in cicada indicates that gene duplication is not a prerequisite for initiation of host/symbiont amino acid exchange in sap-feeding insects, an interpretation consistent with the fact that some single-copy AATs also operate at the symbiotic interface in aphids (Price et al. 2011) and mealybugs (Fig. 3c).

Amino acid transporter recruitment to the symbiotic interface is dynamic

We measured the expression of paralogs resulting from both ancient and recent gene duplication events because both could play a role in recruiting AATs to the symbiotic interface. Although most examples of gene duplication giving rise to novelty involve gene duplication evolving concurrently with or after the origin of new traits, there are some examples of gene duplication pre-dating the evolution of novelty (Ben Trevaskis et al. 1997; Arnegard et al. 2010). Expression patterns in both the anciently and recently duplicated AATs (Figs 2 and 3) indicate that AATs were recruited independently to the bacteriocytes of different sap-feeding insect lineages. In the ancient duplications pre-dating hemipteran diversification, qRT–PCR results for aphids, mealybugs and psyllids indicate that bacteriocyte enrichment in one psyllid AAT (Bcoc-APC12; Fig. 2b and Fig. S2, Supporting Information) and one aphid AAT (ACYPI000536; Fig. 3b and Fig. S2 in Appendix S1, Supporting Information) is derived. This finding is consistent with sap being a derived diet within Hemiptera (Cobben 1979; Zrzavy 1990, 1992), requiring that AATs be independently recruited to the symbiotic interface after hemipteran suborders (and these orthologs) diverged from their common ancestor. Biological replication within each sternorrhynchan lineage (psyllids, mealybugs and aphids) would provide finer resolution of the extent to which expression is or is not conserved in these taxa. However, lack of within-species biological replication does not compromise our finding that expression of orthologous AATs is not conserved.

Similarly with respect to the recent taxon-specific gene duplications, qRT–PCR from this study (Fig. 2 and 3; Appendix S1, Supporting Information) and Price et al. (2011; adult female aphids) together with RNA-seq differential expression data from this study (Fig. 2, 3; Appendix S3, Supporting Information), Hansen & Moran (2011; fourth-instar female aphids) and Macdonald et al. (2012; 7-day-old female aphids) support independent AAT recruitment to the symbiotic interface in pea aphids and citrus mealybugs (Figs 2c and 3c). Notably, bacteriocyte expression in aphid AATs is remarkably consistent across qRT–PCR and RNA-seq studies that together include data from four different pea aphid lineages at different developmental stages (this present study; Price et al. 2011; Hansen & Moran 2011; Macdonald et al. 2012). Similar to what was previously reported for pea aphids (Hansen & Moran 2011; Price et al. 2011), six mealybug-specific paralogs have enriched bacteriocyte expression (Fig. 3c; Appendix S3, Supporting Information). As we reported previously, expression profiles among aphid APC paralogs (Fig. 2c) are most parsimoniously explained by bacteriocyte enrichment evolving after (and potentially being enabled by) gene duplication, an argument based on
Fig. 3 Partial AAAP (TC # 2.A.18) phylogeny and bacteriocyte expression. (a) Bayesian gene phylogeny for amino acid transporters (AATs) in the AAAP family. Hemiptera-specific gene duplications and taxon-specific expansions are highlighted with green or grey boxes, respectively. Asterisks denote possible cicada-specific paralogs. Branches are colour-coded based on taxon and clade support ≥ 50% (posterior probability and ML bootstrap support) is indicated on branches/nodes as described in the key. (b) qRT–PCR expression data generated in this study for Hemiptera-specific gene duplication are presented both as a heat map for whole insect (‘WI’) and bacteriocyte (‘B’) and as differential expression (‘DE’) between bacteriocyte and whole insect. Heat map expression data are normalized across all tissues and genes within each insect, but not across insects. (c) Differential expression between whole insect and bacteriocyte is indicated for aphid and mealybug genes in boxes to the right of gene IDs, as indicated in the key. RNA-seq differential expression data for aphids are from Hansen and Moran (2011) and Macdonald et al. (2012). qRT–PCR data not generated here are from Price et al. (2011). Expression is marked as ambiguous (‘A’) if different transcripts or data sets show inconsistent relative bacteriocyte expression.

![Schematic illustrating the arrangement of mealybug AAAP paralogs along genomic scaffolds. Grey arrows depict the position and 5′–3′ direction of representative transcripts (including introns) along three mealybug genomic scaffolds. Each row represents a different scaffold. The top two scaffolds are depicted at the same scale (upper scale bar), and the bottom scaffold is depicted at a different scale (bottom scale bar).](image)

Fig. 4 Paralogs in mealybug-specific AAAP expansion are tandemly arrayed in the genome. Schematic illustrating the arrangement of mealybug AAAP paralogs along genomic scaffolds. Grey arrows depict the position and 5′–3′ direction of representative transcripts (including introns) along three mealybug genomic scaffolds. Each row represents a different scaffold. The top two scaffolds are depicted at the same scale (upper scale bar), and the bottom scaffold is depicted at a different scale (bottom scale bar).

The fact that bacteriocytes are a novel, derived tissue and most aphid APC paralogs, like their insect orthologs, are highly expressed in gut (Price et al. 2011). In contrast, the distribution of bacteriocyte enrichment among mealybug AAAP paralogs (Fig. 3c) lacks a clear most parsimonious explanation. Bacteriocyte enrichment/expression could be derived or ancestral, consistent with either neofunctionalization or subfunctionalization of duplicated paralogs. Furthermore, some paralogs may be functionally redundant and are maintained for dosage reasons or are differentially expressed across time and space. Indeed, some aphid AAT paralogs are enriched in head and gut tissues (Price et al. 2011), and others have male-biased expression (Duncan et al. 2011). Distinguishing between these explanations will be facilitated with functional data for paralogs of this expansion and their orthologs in other insects. However, that multiple paralogs show bacteriocyte enrichment together with substantial sequence divergence among paralogs (indicated by long branches) strongly suggests that at least some mealybug paralogs have evolved novel functional roles. Our results are thus consistent with the hypothesis that gene duplication played a role in recruiting mealybug AATs to the symbiotic interface, enabling them to carry out novel, symbiotic functions.

Interestingly, expression patterns indicate that aphids and mealybugs use different AATs at their symbiotic interface. For example, bacteriocyte-enriched AATs are not orthologous between aphids and mealybugs (Figs 2 and 3). Recruitment of different AATs in aphids and mealybugs could reflect differences in nutritional demand between these insects or could simply result from chance. Alternatively, AATs could be functionally dynamic, with similar environmental pressures experienced by aphids and mealybugs resulting in distinct AAT loci converging upon common functional roles.

Differential AAT expansion among sap-feeding hemipterans is consistent with co-evolutionary patterns of host/symbiont metabolic collaboration

Despite evidence that gene duplication has facilitated the recruitment of AATs to the symbiotic interface in the Sternorrhyncha, cicadas demonstrate that gene duplication is not necessary to initiate novel sap-feeding/symbiont amino acid exchange. Cicadas did not experience expansions in their AATs, a pattern that may relate to a dietary difference between cicadas and sternorrhynchan sap-feeders. While sternorrhynchans feed on plant phloem sap (Gullan et al. 2003), the source of sap for cicadas is the plant xylem (White & Strehl 1978), a more dilute source of nitrogen than phloem (Redak et al. 2004). It is unclear how amino acid concentration per se could influence host insect AAT evolution and recruitment to the symbiotic interface. However, differences in individual amino acid content could potentially influence the nutritional demands of different sap-feeding insects and thus the evolutionary trajectory of amino acid transporters operating at the symbiotic interface. However, recent sequencing of the symbiont genomes of a phloem-feeding auchenorrhynchan suggests that differences in AAT copy number...
between sternorrhynchans and auchenorrhynchans are not driven by diet.

Bennett and Moran (2013) recently sequenced _Sulcia muelleri_ and _Nasua deltocephalinicola_, the obligate symbionts of the phloem-feeding auchenorrhynchan _Macrostelus quadrilineatus_. Their work highlights an important genomic difference between the obligate symbioses of sternorrhynchans and auchenorrhynchans. Obligate symbionts of both sternorrhynchans and auchenorrhynchans play a major role in providing their hosts with essential amino acids. However, while symbionts of both phloem-feeding and xylem-feeding auchenorrhynchans retain relatively autonomous metabolic pathways (Wu et al. 2006; McCutcheon & Moran 2007; McCutcheon et al. 2009; Bennett & Moran 2013), sternorrhynchans symbionts lack some genes for crucial metabolic steps – metabolic steps that the host has been demonstrated to complement (Russell et al. 2013). For example, sternorrhynchans symbionts typically lack genes necessary to complete the terminal steps in branch-chain amino acid and phenylalanine biosynthesis as well as the step required to synthesize homocysteine for methionine biosynthesis (Shigenobu et al. 2000; Nakabachi et al. 2006; McCutcheon & von Dohlen 2011; Sabree et al. 2012b; Sloan & Moran 2012; Husnik et al. 2013). These missing steps are carried out by host insect enzymes (Wilson et al. 2010; Hansen & Moran 2011; McCutcheon & von Dohlen 2011; Poliakov et al. 2011; Shigenobu & Wilson 2011; Macdonald et al. 2012; Husnik et al. 2013; Russell et al. 2013). This within-metallic pathway host/symbiont collaboration likely necessitates host/symbiont exchange of intermediate metabolites, a step that is not required in auchenorrhynchans that possess metabolically autonomous symbionts (Wu et al. 2006; McCutcheon & Moran 2007; McCutcheon et al. 2009; Bennett & Moran 2013). Therefore, gene duplication could have enabled, through neofunctionalization of paralogs, the evolution of novel transporters capable of transporting intermediate metabolites in amino acid biosynthesis pathways, facilitating pathway partitioning between sternorrhynchans hosts and their symbionts. Once functional data are available for these transporters, this hypothesis can be tested. Thus, current evidence suggests that differences in AAT copy number between sternorrhynchans and auchenorrhynchans are driven by differences in the extent of host/symbiont metabolic independence.

**Gene duplication and the evolution of novel, symbiotic interactions**

The generation of genomic resources for nonmodel organisms, including the partners of symbiotic systems, makes it possible to understand how intimate symbiotic relationships have influenced genome evolution in both symbionts (Shigenobu et al. 2000; Nakabachi et al. 2006; McCutcheon & Moran 2007; McCutcheon et al. 2009; Sabree et al. 2009; McCutcheon & von Dohlen 2011; Nikoh et al. 2011; McCutcheon & Moran 2012; Sabree et al. 2012a; a) and hosts (International Aphid Genomics Consortium 2010; Kirkness et al. 2010; Nygaard et al. 2011; Young et al. 2011; Husnik et al. 2013). The pea aphid/Buchnera symbiosis was the first symbiotic system to have both host and symbiont genomes sequenced, providing the first insights into how host genomes are shaped by symbiosis. Here, we provide evidence that one of those insights applies more broadly to sternorrhynchans sap-feeding insects: gene duplication plays a role in recruiting amino acid transporters to operate at the host/symbiont interface. Further, recent studies in other, very divergent, symbiotic systems also invoke gene duplication in the evolution of genes with symbiotic functions. For example, in legumes, an ancient whole-genome duplication event in the ancestor of the major papilionoid subfamily was followed by some paralogs evolving enriched expression in symbiotic root nodules. This pattern correlates with the evolution of many important Nod factor signalling components that are critical for legume/Rhizobium recognition and the initiation of nodulation in this subfamily (Young et al. 2011). Additionally, gene duplication may have facilitated the origin of leghaemoglobin, a special haemoglobin protein that legumes use to remove oxygen from symbiotic root nodules, facilitating symbiotic nitrogen fixation (Anderson et al. 1996; Ben Trevaskis et al. 1997). Similarly, in an anemone/dinoflagellate symbiosis, cnidarian-specific paralogs gave rise to three genes proposed to function in symbiosis. All three of these cnidarian-specific paralogs are both enriched in individuals hosting symbionts (as opposed to individuals lacking symbionts) and preferentially expressed in the gastroderm, where symbionts are housed (Canot et al. 2011). Together with the results presented here, these plant and cnidarian studies suggest that gene duplication facilitates the recruitment of nonsymbiotic genes to play a role in symbiosis broadly across symbiotic systems. The independent evolution in diverse symbiotic systems of gene duplication followed by expression in tissues that host symbionts, however intriguing, does not in itself provide insight into the potential adaptive significance of gene duplication in the evolution of symbiosis-related genes. The crucial next step to deciphering the role of gene duplication in the evolution of symbiotic interactions will be functional characterization within a phylogenetic framework, which will reveal whether paralogs preferentially expressed at the host/symbiont interface have also evolved novel symbiotic functions.
Acknowledgements

We thank Carol von Dohlen for helpful discussion and for supplying mealybugs. Cecilia Tambourindey and Joseph Hancock helped with potato psyllid dissections. The assembled kissing bug transcriptome was provided by Pedro Lagerblad de Oliveira and Gloria Braz, and the assembled potato psyllid transcriptome was provided by Cecilia Tambourindey. Jack Min kindly provided the stand-alone ORF prediction PERL script. We are also grateful to Dan Price, Angela Douglas, Jacob Russell and three anonymous reviewers for helpful discussion and feedback on the manuscript. This research was supported by a University of Miami Department of Biology Evoey award (RPD), a Sigma Xi Grant-in-Aid of Research (RPD), NSF Graduate Research Fellowship DGE-0951792 (RPD), the Czech Science Foundation 13-01878S (FH), NSF DBI-0640462 (DDG), NSF IOS-1256680 (JPM), NSF IOS-1121847 (ACCW) and start-up funds from the University of Miami (ACCW).

References


R.P.D. and A.C.C.W. conceived of and designed the project. R.P.D. assembled the citrus mealybug whole insect transcriptome, mapped transcripts to genome scaffolds and performed dS analyses, gene conversion analyses and qRT–PCR experiments. F.H. assembled the mealybug bacteriocyte transcriptome and conducted the mealybug differential expression analysis. J.P.M. collected cicadas and generated the cicada RNA-seq data. J.T.V.L. assembled cicada whole insect and bacteriocyte transcriptomes. R.P.D., A.C.C.W., F.H. and L.M.D. designed the phylogenetic analyses, and R.P.D. conducted the phylogenetic analyses. D.G.G. conducted the whitefly transcriptome re-assembly. R.P.D. and A.C.C.W. drafted the manuscript, and all authors edited the manuscript. All authors approved the final version of the manuscript.

Data accessibility

Raw sequence reads and assemblies: Mealybug – raw sequence reads for transcriptomes and genome are available under NCBI BioProject PRJNA196641. Psyllid – the full psyllid transcriptome assembly is publicly available at http://psyllid.org/download. Whitefly – raw sequence reads are available under NCBI BioProject PRJNA89143. The full transcriptome assembly is publicly available at http://arthropods.eugenes.org/EvidentialGene/arthropods/whitefly/whitefly1eg6/. Cicada – raw sequence reads are available in the NCBI Sequence Read Archive (SRR952383). Insect and bacteriocyte transcripts are pooled and can be separated by the index sequences CAGATC (insect) and ACTTGA (bacteriocyte). Kissing bug – raw sequence reads are available under NCBI BioProject PRJNA191820. Assembled contigs: transcripts for amino acid transporters are available in Appendix S6 (Supporting Information). Mealybug genome scaffolds associated with amino acid transporters are available in Appendix S7 (Supporting Information). Kissing bug genome scaffolds are available on vectorbase.org under the scaffold IDs reported in Appendix S1 (Supporting Information). Data for phylogenetic analyses: protein sequences used for phylogenetic analyses: Appendix S4 (APC) and S5 (AAAP) (Supporting Information).

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Maximum-likelihood phylogeny of full AAAP family.

Fig. S2 qRT–PCR expression results for hemipteran-specific gene duplications.

Table S1 Mealybug loci for ACP family and representative transcripts.

Table S2 Mealybug loci for AAAP family and representative transcripts.

Table S3 Kissing bug loci for ACP family and representative transcripts.

Table S4 Kissing bug loci for AAAP family and representative transcripts.

Table S5 Gene conversion results.

Table S6 qRT–PCR primers.

Appendix S1 Tables S1–S6, Figures S1–S2.

Appendix S2 dS analysis results.

Appendix S3 Mealybug bacteriocyte-whole insect differential expression results.

Appendix S4 APC protein sequences used for phylogenetic analyses.

Appendix S5 AAAP protein sequences used for phylogenetic analyses.

Appendix S6 Assembled contigs for transcripts referenced in this study.

Appendix S7 Assembled mealybug genome scaffolds referenced in this study.