Phylogenomics Identifies an Ancestral Burst of Gene Duplications Predating the Diversification of Aphidomorpha

Irene Julca,[†,1] Marina Marcet-Houben,[†,1] Fernando Cruz, [2] Carlos Vargas-Chavez,[3] John Spencer Johnston,[4] Jèssica Gómez-Garrido,[2] Leonor Frias,[2] André Corvelo,[2,5] Damian Loska,[1] Francisco Cámara,[1] Marta Gut,[2,6] Tyler Alioto,[2,6] Amparo Latorre,[3,7] and Toni Gabaldón*[†,1,6,8]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain

[2]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain

[3]Institute for Integrative Systems Biology (I[2]SysBio), University of Valencia and CSIC, Valencia, Spain

[4]Department of Entomology, Texas A&M University, College Station, TX

[5]New York Genome Center, New York, NY

[6]Universitat Pompeu Fabra (UPF), Department of Experimental and Health Sciences, Barcelona, Spain

[7]Joint Unit in Genomics and Health, Foundation for the Promotion of Sanitary and Biomedical Research (FISABIO) and University of Valencia, Valencia, Spain

[8]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

[†]Present address: Barcelona Supercomputing Centre (BSC-CNS), and the Institute for Research in Biomedicine, Barcelona, Spain

***Corresponding author:** E-mail: tgabaldon@crg.es.

Associate editor: Fabia Ursula Battistuzzi

The genome, annotation and sequencing reads have been deposited at the European Nucleotide Archive (ENA) under the project accession PRJEB33415.

Abstract

Aphids (Aphidoidea) are a diverse group of hemipteran insects that feed on plant phloem sap. A common finding in studies of aphid genomes is the presence of a large number of duplicated genes. However, when these duplications occurred remains unclear, partly due to the high relatedness of sequenced species. To better understand the origin of aphid duplications we sequenced and assembled the genome of *Cinara cedri*, an early branching lineage (Lachninae) of the Aphididae family. We performed a phylogenomic comparison of this genome with 20 other sequenced genomes, including the available genomes of five other aphids, along with the transcriptomes of two species belonging to Adelgidae (a closely related clade to the aphids) and Coccoidea. We found that gene duplication has been pervasive throughout the evolution of aphids, including many parallel waves of recent, species-specific duplications. Most notably, we identified a consistent set of very ancestral duplications, originating from a large-scale gene duplication predating the diversification of Aphidomorpha (comprising aphids, phylloxerids, and adelgids). Genes duplicated in this ancestral wave are enriched in functions related to traits shared by Aphidomorpha, such as association with endosymbionts, and adaptation to plant defenses and phloem-sap-based diet. The ancestral nature of this duplication wave (106–227 Ma) and the lack of sufficiently conserved synteny make it difficult to conclude whether it originated from a whole-genome duplication event or, alternatively, from a burst of large-scale segmental duplications. Genome sequencing of other aphid species belonging to different Aphidomorpha and related lineages may clarify these findings.

Key words: **gene duplication, aphids, Aphidomorpha.**

Article

Introduction

Large-scale gene duplication, including whole-genome duplication (WGD), is a very common phenomenon in eukaryotic genomes. Bursts of gene duplications are considered a major source of evolutionary innovation and have been associated with the increase in biological complexity and adaptive radiations of species (Zhang 2003). In particular, large-scale gene duplications, generally associated with WGDs, have been reported for many eukaryotic lineages including plants (Van de Peer et al. 2017), fungi (Marcet-Houben and Gabaldón 2015), and animals (Taylor et al. 2001). Although large-scale duplication seems less pervasive in animals than in plants, a growing number of studies report such events in animals. Among other lineages, putative WGDs have been described at the base of vertebrates (Ohno 1970; Dehal and Boore 2005; Putnam et al. 2008), and in several lineages of fish (Christoffels et al. 2004; Glasauer and Neuhauss 2014), amphibians (Mable et al. 2011; Session et al. 2016), and arthropods (Jacobson et al. 2013; Kenny et al. 2016; Schwager et al. 2017; Li et al. 2018).

Aphids belong to the infraorder Aphidomorpha that includes three families: Aphididae, Adelgidae, and

© The Author(s) 2019. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

Phylloxeridae (Favret 2013; Nováková et al. 2013; Blackman and Eastop 2000). Aphids and related (Aphidomorpha) species (Becker-Migdisova and Aizenberg 1962) are hemipteran insects that feed on plant sap (Tjallingii 1995). This specialized diet, rich in carbohydrates but poor in nitrogen compounds, has resulted in several adaptations including the establishment of tight relationships with bacterial endosymbionts (Scarborough et al. 2005; Moya et al. 2008; von Dohlen et al. 2017). There are more than 5,000 described aphid species, of which, about 450 have been collected from crop plants, and 100 are considered of significant economic importance (Van Emden and Harrington 2017). Genomes of several aphid species of agricultural interest have been sequenced, including *Acyrthosiphon pisum*, *Myzus persicae*, *Diuraphis noxia*, *Aphis glycines*, and *Sipha flava* (International Aphid Genomics Consortium 2010; Nicholson et al. 2015; Mathers et al. 2017; Wenger et al. 2017). However, except for *S. flava* (subfamily Chaitophorinae), the sequenced aphids belong to a single subfamily, Aphidinae, limiting our understanding of the genomic diversity in this group of insects. Remarkably, most genome analyses in these species have revealed an important number of paralogous sequences and expanded gene families, including amino acid transporters, odorant and gustatory receptor genes, miRNA-specific dicer-1, ago1 genes, and pasha, among others (Smadja et al. 2009; Huerta-Cepas et al. 2010; Jaubert-Possamai et al. 2010; Duncan et al. 2016; Mathers et al. 2017). However, the close relatedness of the sequenced species provides little resolution to the phylogenetic placement of the duplication events, particularly the ancestral ones.

Recent studies have focused on assessing patterns of sequence and expression divergence among recently duplicated genes in *A. pisum* (Fernández et al. 2019) or *M. persicae* (Mathers et al. 2017). They have also inspected the distribution of old and young *A. pisum* paralogs along chromosomes, by categorizing the age of genes that are best-reciprocal hits of each other based on the amount of synonymous substitutions (Li et al. 2019). However, we still lack a proper understanding of when the ancestral duplications occurred, and whether they can be linked to phenotypic innovations shared by aphids or related species. To better assess the origin of the paralogous genes of aphids we sequenced the genome of *Cinara cedri* (Lachninae subfamily, tribe *Eulachnini*), the first representative genome from an early-branching lineage of the Aphididae family. *Cinara* species (and most Lachninae) are particular among aphids as they feed on conifers (gymnosperms), whereas all the other genome-sequenced aphids feed on angiosperms. Another clear difference between the Lachninae and the rest of aphids is that two co-obligate endosymbionts (*Buchnera aphidicola*, *Serratia symbiotica*) are present in this group, whereas only *B. aphidicola* is obligate for the rest of aphids (Latorre and Manzano-Marín 2017). We used a phylogeny-based approach (Huerta-Cepas and Gabaldón 2011) to provide the relative timing of aphid duplications in a phylogenetic framework that includes 21 other fully sequenced genomes and two transcriptomes. Our results provide compelling evidence for an ancestral wave of gene duplications, whose origin predates the diversification of all sequenced aphids, adelgids, and phyloxerids, but are subsequent to their divergence from the Coccoidea lineage, ~106–227 Ma.

Results and Discussion

Genome Sequence of *C. cedri*

The haploid genome sizes for *C. cedri* and two other Lachninae species (*C. tujafiina* and *Tuberolagnus salignus*, tribes *Eulachnini* and *Tuberolachnini*, respectively) were measured using flow cytometry (Johnston et al. 2019) which resulted in estimates of ~592, 713, and 494 Mb, respectively. For reference, the genome size of *A. pisum* is 520.8 Mb (International Aphid Genomics Consortium 2010). We used an Illumina pair-end sequencing approach to produce a draft assembly of the *C. cedri* genome (see Materials and Methods). A rough estimate of the genome size obtained by dividing the total number of 17-mers by the peak 17-mer coverage results in an estimate of 508.6 Mb (supplementary fig. S1, Supplementary Material online), slightly smaller than the flow cytometry estimate. However, the K-mer profiles indicated an appreciable amount of repeated sequences, which makes the assembly from short reads challenging. To obtain a more precise estimate, we used GenomeScope v1.0 (Vurture et al. 2017) and fit the previous K-mer profile to a mixture model. This provided a haploid genome size estimate of 399.76 Mb, which was used to guide our assembly strategy. This analysis also inferred the amount of unique (223.34 Mb) and repetitive (175.4 Mb) content. Separate assemblies, exploring different K-mer sizes, were done with ABySS v1.5.2 (Simpson et al. 2009), and later merged with ASM (Cruz et al. 2016). The continuity of the merged assembly was improved through several rounds of scaffolding, first with ABySS and later with SSPACEv3.0 (Boetzer et al. 2011). Gaps were closed with GapFiller (Boetzer and Pirovano 2012). The length of the final assembly (see Materials and Methods) is 396.03 Mb, and its contig and scaffold N50 are 104,784 bp and 1.23 Mb, respectively.

The gene completeness of our assembly is high, as evaluated by BUSCO v3.0.2 (93.9% of 1,658 single-copy, conserved genes in *insecta_odb9* data set were present) and CEGMA (100% of 248 eukaryotic core genomes) (Parra et al. 2007; Simão et al. 2015). Notably, 2.5% of the BUSCO genes were duplicated in our assembly. The postassembly K-mer analysis (Mapleson et al. 2017) suggests that these are real paralogs and not the result of assembly artifacts (supplementary fig. S2, Supplementary Material online).

The final protein-coding annotation (see Materials and Methods) resulted in 16,996 genes, whose 24,835 transcripts (1.46 transcripts/gene) encode 22,503 unique protein products. Attempts to detect selenoprotein genes with selenoprofiles (Santesmasses et al. 2018) failed, which indicates that the previously described loss of selenoproteins in some aphids (International Aphid Genomics Consortium 2010; Mariotti et al. 2015) is ancient, and had already occurred at the base of the Aphididae lineage. Similarly, the immune repertoire in *C. cedri* resembles that of other sequenced aphids, which

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

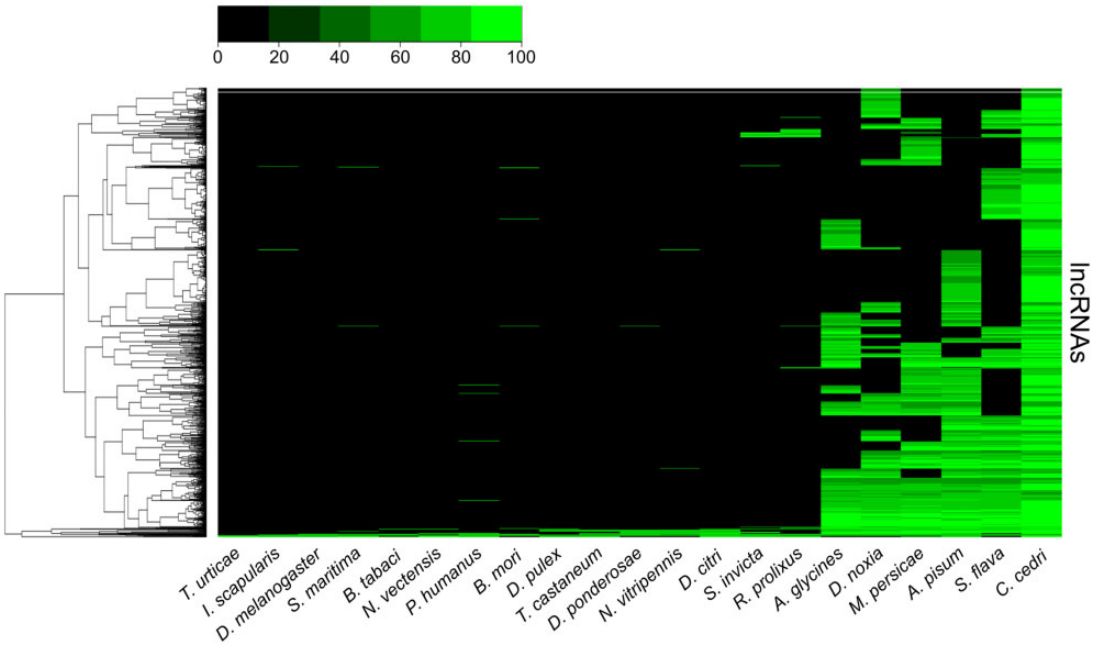

Fig. 1. Heatmap of *Cinara cedri* lncRNA conservation among 20 metazoans. The input lncRNA sequences come from *C. cedri*. The rows represent *C. cedri* lncRNAs and the columns represent species. Each cell is colorized based on the level of conservation of the lncRNA (green—100% conservation, black—0% conservation [absence of the lncRNA]).

indicates that the reported streamlining of the immune system in aphids (Gerardo et al. 2010) appears at the base of Aphididae lineage (supplementary table S1, Supplementary Material online). Previous aphid genome annotations do not report long-noncoding RNAs (lncRNAs) (International Aphid Genomics Consortium 2010; Nicholson et al. 2015; Mathers et al. 2017). To gain insight on the potential lncRNA content in aphids, we used RNAseq to predict lncRNAs (see Materials and Methods). A total of 13,478 lncRNAs were predicted in the genome of *C. cedri*. Importantly, 706 lncRNAs are shared between *C. cedri* and other aphids. Of these 191 appear to form a conserved core within aphids, and some are conserved across insects (fig. 1). Altogether, given its key phylogenetic position, the *C. cedri* genome provides an important resource to study genome evolution in aphids.

Aphid Phylomes and Species-Specific Gene Duplications

As our main focus was to assess gene duplication dynamics in aphids, we reconstructed the complete collection of evolutionary gene histories (i.e., the phylome) of *C. cedri*, *A. pisum*, *M. persicae*, *D. noxia*, *Ap. glycines*, and *S. flava* in the context of other sequenced species (supplementary tables S2 and S3, Supplementary Material online, see Materials and Methods). These genes were scanned to infer duplication and speciation events and derive orthology and paralogy relationships among homologous genes per each phylome (Gabaldón 2008). All of the resulting gene trees, alignments, and orthology and paralogy predictions are available for download or browsing at PhylomeDB (PhylomeIDs: *C. cedri*—701, *S. flava*—702, *Ap. glycines*—703, *D. noxia*—704, *M. persicae*—705, *A. pisum*—706) (Huerta-Cepas et al. 2014). To reconstruct the evolutionary relationships among all considered species, we concatenated the protein alignments of 57 gene trees that are present across all considered species (see Materials and Methods). The resulting highly supported topology (fig. 2*a*) was congruent with current views on aphids phylogeny (Nováková et al. 2013; Chen et al. 2016; Rebijith et al. 2017) and places *C. cedri* as the earliest branching lineage from our set of aphids.

We next focused on gene duplications, including large expansions, that occurred specifically in the lineage leading to each aphid. Interestingly, *C. cedri*, *A. pisum*, *M. persicae*, *Ap. glycines*, and *S. flava* have similar proportions of proteins that have an in-paralog (resulting from a species-specific duplication): *C. cedri*—4,670 (28% of the proteome), *S. flava*—2,832 (21%), *Ap. glycines*—3,232 (17%), *M. persicae*—4,097 (22%), *A. pisum*—5,431 (29%). These events can be assigned to a similar number of inferred specific gene duplication events: *C. cedri*—1,420, *S. flava*—899, *Ap. glycines*—1,153, *M. persicae*—1,543, *A. pisum*—1,889. On the contrary, *D. noxia* only presented a total of 685 proteins (6% of the proteome) with an in-paralog, corresponding to 315 gene duplication events. In all six aphids, the majority of the gene duplication events result in a moderate number of paralogs (2–5 in-paralogs; supplementary fig. S3, Supplementary Material online), and only few represent large gene family expansions ($\geq$10 in-paralogs). The large expansions could be due to the presence of expanded transposable element families (Huerta-Cepas et al. 2010). In the six aphids, an average of 9% of the total number of annotated protein-coding genes are associated with transposons, with *A. pisum* and *M. persicae* (*Macrosiphini*) containing the highest percentages (fig. 2*b*). Moreover, larger expansions in *C. cedri* and *A. pisum* ($>$50 proteins) often include proteins associated

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

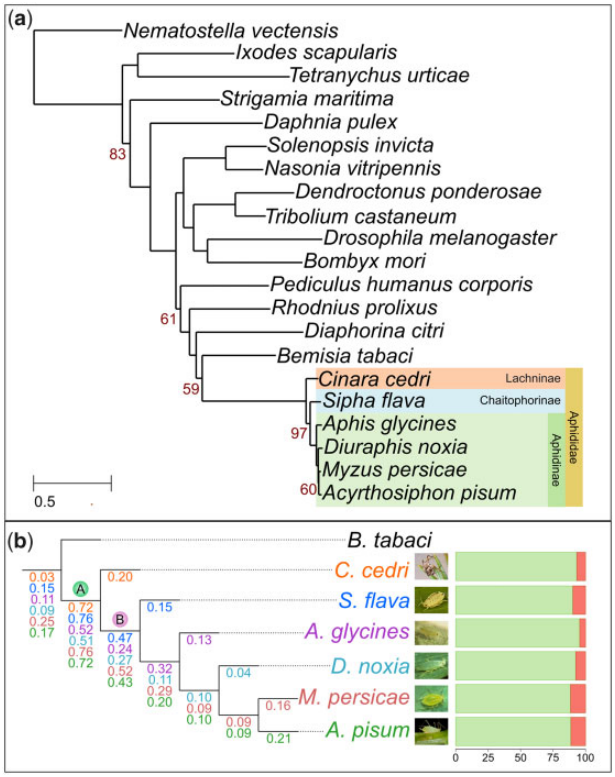

FIG. 2. Species tree and duplication ratios of the six phylomes. (*a*) Phylogenetic tree obtained from the concatenation of 57 widespread gene families. In yellow, all the individuals included in this study that belong to the family Aphididae; in green, light blue, and orange, the aphids that belong to the subfamily Aphidinae, Chaitophorinae, and Lachninae, respectively. All omitted bootstrap values are maximal (bootstrap 100%). (*b*) Zoom out showing the duplication ratios per each phylome: *Cinara cedri*—orange, *Sipha flava*—blue, *Aphis glycines*—purple, *Diuraphis noxia*—light blue, *Myzus persicae*—red, *Acyrthosiphon pisum*—green. The two branches with the higher duplication ratio are marked as A (ancestral to all six aphids) and B (after the divergence of *C. cedri* and ancestral to the other five aphids). Bars on the right show the percentage of proteins (orange) associated with transposons in each aphid species. *Bemisia tabaci* is the outgroup.

with transposons. However, after removing expansions containing at least one paralog annotated with a PFAM domain or a gene ontology (GO) term associated with transposable elements or viruses, the number of duplications remained high (supplementary fig. S3, Supplementary Material online).

We performed a functional GO term enrichment analysis of these transposon-free, species-specific paralogs (table 1) for each proteome. DNA and RNA processing terms were enriched among sets of in-paralogs of all species except *D. noxia*. Moreover, *C. cedri* in-paralogs were enriched in GO terms associated with olfactory receptor activity, odorant binding, acetyl-CoA transporter activity, and CCR4-NOT. For *S. flava*, peroxidase activity, methyltransferase activity, beta-glucosidase activity, lipid droplet, CCR4-NOT complex, and response to oxidative stress were enriched. For *Ap. glycines*, fatty acid synthase activity, SUMO transferase activity, and regulation of JAK-STAT cascade were enriched. For *D. noxia*, fucose metabolic process and protein glycosylation were enriched. For *M. persicae*, peroxidase activity was enriched, and for *A. pisum*, enoyl-reductase, oleoyl-hydrolase, myristoyl-hydrolase, palmitoyl-hydrolase, odorant binding, and response to stress were enriched. These results are consistent with previous results restricted to *A. pisum* and *M. persicae* (Huerta-Cepas et al. 2010; International Aphid Genomics Consortium 2010; Mathers et al. 2017).

In order to detect parallel duplications, we searched for orthologs between *C. cedri* and the other aphids with species-specific duplications. A total of 909 *C. cedri* genes (26% of the total proteins with in-paralogs) with species-specific duplications have parallel species-specific duplications in at least one of the other aphids (*S. flava*—364, *Ap. glycines*—181, *D. noxia*—57, *M. persicae*—235, *A. pisum*—375). Specifically, 694 *C. cedri* genes share unique parallel duplications with one of the other aphids: *S. flava*—252, *Ap. glycines*—91, *D. noxia*—14, *M. persicae*—120, *A. pisum*—217. Interestingly, *C. cedri* parallel paralogs show enrichments only in four aphids. The parallel duplications shared with *S. flava* show enrichment for aconitate hydratase activity, L-amino acid transmembrane transporter activity, tricarboxylic acid cycle, aromatase activity, and CCR4-NOT complex. *Acyrthosiphon pisum* species-specific duplications shared with *C. cedri* show enrichment for oxidoreductase activity and L-ascorbic acid binding. *Cinara cedri* duplications shared with *Ap. glycines* and *M. persicae* show only five and three enriched terms, respectively (table 2). Interestingly, two proteins show parallel duplications in all the considered aphid species, from which only one has a functional annotation. This protein is associated with UDP-*N*-acetylglucosamine–peptide *N*-acetylglucosaminyltransferase 110 kDa subunit-like, which catalyzes the transfer of a single *N*-acetylglucosamine from UDP-GlcNAc to a serine or threonine residue (*O*-GlcNAc glycosylation) (Lazarus et al. 2012; Ding et al. 2015). In insects, this type of glycosylation has been shown to be central to a variety of physiological processes, including regulation of the cell cycle, expression of developmental genes, nutrient sensing, response to starvation, insulin signaling, or specification of body size (Vandenborre et al. 2011; Walski et al. 2017). Altogether, these results indicate a high dynamism of aphid gene repertoire and suggest that gene duplication may play a major role in the adaptation of aphid species to their respective environments.

High Number of Ancient Gene Duplications Suggests One Ancestral Burst of Large-Scale Genome Duplication

In order to detect waves of ancestral duplications in the evolutionary history of aphids, we used a phylogeny-based phylostratigraphic approach based on a species-overlap algorithm (Huerta-Cepas and Gabaldón 2011) to detect gene duplications and map them onto the species tree (see Materials and Methods). After excluding large expansions (duplications resulting in >5 paralogs), we computed ratios of gene duplications (average number of duplications per gene detected in a given branch of the species tree) for each phylome (fig. 2*b*). Interestingly, in the aphid lineage two branches have high duplication ratios: one present in the ancestral branch of all six aphids (Aphididae family,

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

Table 1. List of the GO Terms Enriched in the Expanded Protein Families Specific to *Cinara cedri*, *Sipha flava*, *Aphis glycines*, *Diuraphis noxia*, *Myzus persicae*, and *Acyrthosiphon pisum*.

Term Category	Term	Term Level	Adj. *P*-Value	Term Name
Cinara cedri				
molecular_function	GO:0001227	1	3.56E-18	Transcriptional repressor activity, RNA polymerase II transcription regulatory region sequence-specific binding
molecular_function	GO:0003676	1	4.52E-14	Nucleic acid binding
molecular_function	GO:0003677	1	8.29E-05	DNA binding
molecular_function	GO:0003690	1	4.57E-13	Double-stranded DNA binding
molecular_function	GO:0003715	1	5.02E-28	Obsolete transcription termination factor activity
molecular_function	GO:0003723	1	3.25E-07	RNA binding
molecular_function	GO:0003725	1	2.56E-07	Double-stranded RNA binding
molecular_function	GO:0003964	1	9.54E-11	RNA-directed DNA polymerase activity
molecular_function	GO:0003994	1	5.41E-04	Aconitate hydratase activity
molecular_function	GO:0004190	1	2.57E-05	Aspartic-type endopeptidase activity
molecular_function	GO:0004356	1	1.36E-04	Glutamate-ammonia ligase activity
molecular_function	GO:0004497	1	3.44E-06	Monooxygenase activity
molecular_function	GO:0004525	1	6.85E-06	Ribonuclease III activity
molecular_function	GO:0004618	1	3.52E-04	Phosphoglycerate kinase activity
molecular_function	GO:0004984	1	6.99E-06	Olfactory receptor activity
molecular_function	GO:0005506	1	6.42E-13	Iron ion binding
molecular_function	GO:0005549	1	1.04E-04	Odorant binding
molecular_function	GO:0008521	1	3.14E-05	Acetyl-CoA transporter activity
molecular_function	GO:0016705	1	1.08E-18	Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
molecular_function	GO:0018024	1	2.57E-05	Histone-lysine *N*-methyltransferase activity
molecular_function	GO:0020037	1	3.00E-05	Heme binding
molecular_function	GO:0031177	1	3.00E-04	Phosphopantetheine binding
molecular_function	GO:0031490	1	8.29E-05	Chromatin DNA binding
molecular_function	GO:0042302	1	2.70E-05	Structural constituent of cuticle
cellular_component	GO:0000786	1	2.76E-10	Nucleosome
cellular_component	GO:0030015	1	5.41E-04	CCR4-NOT core complex
cellular_component	GO:0070877	1	8.29E-05	Microprocessor complex
biological_process	GO:0006278	1	1.60E-09	RNA-dependent DNA biosynthetic process
biological_process	GO:0006353	1	1.79E-26	DNA-templated transcription, termination
biological_process	GO:0006807	1	7.05E-05	Nitrogen compound metabolic process
biological_process	GO:0009452	1	8.38E-05	7-Methylguanosine RNA capping
biological_process	GO:0015074	1	4.11E-08	DNA integration
biological_process	GO:0016075	1	1.86E-06	rRNA catabolic process
Sipha flava				
molecular_function	GO:0000166	1	2.39E-08	Nucleotide binding
molecular_function	GO:0001227	1	6.76E-23	Transcriptional repressor activity, RNA polymerase II transcription regulatory region sequence-specific binding
molecular_function	GO:0003676	1	2.44E-37	Nucleic acid binding
molecular_function	GO:0003678	1	1.23E-44	DNA helicase activity
molecular_function	GO:0003690	1	9.51E-18	Double-stranded DNA binding
molecular_function	GO:0003696	1	2.71E-22	Satellite DNA binding
molecular_function	GO:0003697	1	1.53E-11	Single-stranded DNA binding
molecular_function	GO:0003715	1	1.72E-12	Obsolete transcription termination factor activity
molecular_function	GO:0003723	1	6.15E-37	RNA binding
molecular_function	GO:0003730	1	1.15E-19	mRNA 3′-UTR binding
molecular_function	GO:0003887	1	1.15E-05	DNA-directed DNA polymerase activity
molecular_function	GO:0004535	1	6.78E-04	Poly(A)-specific ribonuclease activity
molecular_function	GO:0004601	1	1.99E-11	Peroxidase activity
molecular_function	GO:0004666	1	8.59E-06	Prostaglandin-endoperoxide synthase activity
molecular_function	GO:0004801	1	3.80E-04	Sedoheptulose-7-phosphate: D-glyceraldehyde-3-phosphate glyceronetransferase activity
molecular_function	GO:0005200	1	1.10E-10	Structural constituent of cytoskeleton
molecular_function	GO:0008168	1	1.26E-25	Methyltransferase activity
molecular_function	GO:0008408	1	7.74E-04	3′–5′ Exonuclease activity
molecular_function	GO:0010521	1	8.42E-07	Telomerase inhibitor activity
molecular_function	GO:0020037	1	1.05E-07	Heme binding
molecular_function	GO:0032947	1	8.59E-06	Protein complex scaffold activity
molecular_function	GO:0043141	1	2.72E-06	ATP-dependent 5′–3′ DNA helicase activity
molecular_function	GO:0043169	1	5.84E-06	Cation binding

(continued)

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

Table 1. Continued

Term Category	Term	Term Level	Adj. *P*-Value	Term Name
cellular_component	GO:0000792	1	1.22E-20	Heterochromatin
cellular_component	GO:0005657	1	3.02E-04	Replication fork
cellular_component	GO:0005701	1	3.26E-23	Polytene chromosome chromocenter
cellular_component	GO:0005811	1	8.41E-04	Lipid droplet
cellular_component	GO:0005858	1	3.51E-04	Axonemal dynein complex
cellular_component	GO:0005874	1	7.40E-07	Microtubule
cellular_component	GO:0030014	1	7.64E-05	CCR4-NOT complex
cellular_component	GO:0030015	1	1.47E-07	CCR4-NOT core complex
cellular_component	GO:0030529	1	2.48E-14	Intracellular ribonucleoprotein complex
biological_process	GO:0000002	1	3.02E-04	Mitochondrial genome maintenance
biological_process	GO:0000288	1	8.59E-06	Nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay
biological_process	GO:0000289	1	1.41E-04	Nuclear-transcribed mRNA poly(A) tail shortening
biological_process	GO:0000723	1	1.67E-50	Telomere maintenance
biological_process	GO:0001510	1	4.54E-36	RNA methylation
biological_process	GO:0005975	1	1.29E-05	Carbohydrate metabolic process
biological_process	GO:0006260	1	7.46E-04	DNA replication
biological_process	GO:0006281	1	2.55E-24	DNA repair
biological_process	GO:0006353	1	4.95E-12	DNA-templated transcription, termination
biological_process	GO:0006370	1	1.83E-05	7-Methylguanosine mRNA capping
biological_process	GO:0006954	1	3.78E-08	Inflammatory response
biological_process	GO:0006979	1	2.18E-10	Response to oxidative stress
biological_process	GO:0007017	1	1.85E-08	Microtubule-based process
biological_process	GO:0007059	1	4.42E-13	Chromosome segregation
biological_process	GO:0008217	1	8.59E-06	Regulation of blood pressure
biological_process	GO:0009452	1	4.36E-45	7-Methylguanosine RNA capping
biological_process	GO:0016070	1	3.65E-13	RNA metabolic process
biological_process	GO:0019371	1	8.59E-06	Cyclooxygenase pathway
biological_process	GO:0030261	1	7.60E-17	Chromosome condensation
biological_process	GO:0031507	1	3.26E-23	Heterochromatin assembly
biological_process	GO:0032211	1	8.42E-07	Negative regulation of telomere maintenance via telomerase
biological_process	GO:0032259	1	1.86E-04	Methylation
biological_process	GO:0044806	1	8.42E-07	G-quadruplex DNA unwinding
biological_process	GO:0045727	1	4.44E-20	Positive regulation of translation
biological_process	GO:0051258	1	3.88E-05	Protein polymerization
biological_process	GO:0051974	1	8.42E-07	Negative regulation of telomerase activity
biological_process	GO:1901657	1	5.82E-05	Glycosyl compound metabolic process
Aphis glycines				
molecular_function	GO:0001227	1	2.36E-04	Transcriptional repressor activity, RNA polymerase II transcription regulatory region sequence-specific binding
molecular_function	GO:0003676	1	3.08E-37	Nucleic acid binding
molecular_function	GO:0003677	1	9.92E-05	DNA binding
molecular_function	GO:0003715	1	4.51E-19	Obsolete transcription termination factor activity
molecular_function	GO:0003887	1	1.12E-04	DNA-directed DNA polymerase activity
molecular_function	GO:0004312	1	5.19E-08	Fatty acid synthase activity
molecular_function	GO:0004553	1	8.62E-05	Hydrolase activity, hydrolyzing *O*-glycosyl compounds
molecular_function	GO:0004866	1	5.12E-04	Endopeptidase inhibitor activity
molecular_function	GO:0004869	1	1.04E-05	Cysteine-type endopeptidase inhibitor activity
molecular_function	GO:0008408	1	5.21E-05	3′–5′ Exonuclease activity
molecular_function	GO:0008422	1	1.12E-04	Beta-glucosidase activity
molecular_function	GO:0008521	1	5.75E-08	Acetyl-CoA transporter activity
molecular_function	GO:0019789	1	1.11E-08	SUMO transferase activity
molecular_function	GO:0031177	1	4.71E-05	Phosphopantetheine binding
molecular_function	GO:0043027	1	1.55E-05	Cysteine-type endopeptidase inhibitor activity involved in apoptotic process
molecular_function	GO:0043169	1	1.21E-07	Cation binding
molecular_function	GO:0044390	1	5.12E-04	Ubiquitin-like protein conjugating enzyme binding
molecular_function	GO:0061663	1	1.50E-04	NEDD8 ligase activity
molecular_function	GO:0089720	1	1.55E-05	Caspase binding
cellular_component	GO:0005652	1	2.75E-04	Nuclear lamina
cellular_component	GO:0005705	1	4.07E-04	Polytene chromosome interband
cellular_component	GO:0005876	1	1.91E-04	Spindle microtubule
cellular_component	GO:0008537	1	1.55E-05	Proteasome activator complex
cellular_component	GO:0035012	1	2.18E-05	Polytene chromosome, telomeric region
cellular_component	GO:0070776	1	5.82E-08	MOZ/MORF histone acetyltransferase complex

(continued)

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

Table 1. Continued

Term Category	Term	Term Level	Adj. *P*-Value	Term Name
biological_process	GO:0001510	1	7.60E-13	RNA methylation
biological_process	GO:0005975	1	1.18E-08	Carbohydrate metabolic process
biological_process	GO:0006353	1	1.61E-18	DNA-templated transcription, termination
biological_process	GO:0007289	1	5.12E-04	Spermatid nucleus differentiation
biological_process	GO:0007446	1	3.21E-05	Imaginal disc growth
biological_process	GO:0009452	1	6.31E-18	7-Methylguanosine RNA capping
biological_process	GO:0030261	1	9.74E-06	Chromosome condensation
biological_process	GO:0043154	1	1.13E-04	Negative regulation of cysteine-type endopeptidase activity involved in apoptotic process
biological_process	GO:0046425	1	6.74E-04	Regulation of JAK-STAT cascade
biological_process	GO:0046426	1	4.76E-04	Negative regulation of JAK-STAT cascade
biological_process	GO:0070936	1	4.71E-05	Protein K48-linked ubiquitination
biological_process	GO:0090307	1	6.55E-04	Mitotic spindle assembly
biological_process	GO:0097340	1	1.50E-04	Inhibition of cysteine-type endopeptidase activity
biological_process	GO:1901657	1	2.18E-05	Glycosyl compound metabolic process
biological_process	GO:1990001	1	2.82E-07	Inhibition of cysteine-type endopeptidase activity involved in apoptotic process
biological_process	GO:2001271	1	1.50E-04	Negative regulation of cysteine-type endopeptidase activity involved in execution phase of apoptosis
Diuraphis noxia				
molecular_function	GO:0003678	1	6.51E-05	DNA helicase activity
molecular_function	GO:0004827	1	2.22E-05	Proline-tRNA ligase activity
molecular_function	GO:0008424	1	1.34E-04	Glycoprotein 6-alpha-L-fucosyltransferase activity
molecular_function	GO:0046921	1	6.24E-05	Alpha-(1->6)-fucosyltransferase activity
biological_process	GO:0000723	1	5.82E-06	Telomere maintenance
biological_process	GO:0006433	1	2.22E-05	Prolyl-tRNA aminoacylation
biological_process	GO:0033578	1	1.34E-04	Protein glycosylation in Golgi
biological_process	GO:0036071	1	1.34E-04	*N*-glycan fucosylation
biological_process	GO:0046368	1	6.24E-05	GDP-L-fucose metabolic process
Myzus persicae				
molecular_function	GO:0001227	1	1.86E-13	Transcriptional repressor activity, RNA polymerase II transcription regulatory region sequence-specific binding
molecular_function	GO:0003676	1	1.79E-25	Nucleic acid binding
molecular_function	GO:0003677	1	2.30E-11	DNA binding
molecular_function	GO:0003690	1	1.17E-09	Double-stranded DNA binding
molecular_function	GO:0003777	1	5.98E-04	Microtubule motor activity
molecular_function	GO:0004149	1	2.89E-04	Dihydrolipoyllysine-residue succinyltransferase activity
molecular_function	GO:0004601	1	1.38E-04	Peroxidase activity
molecular_function	GO:0004818	1	6.76E-05	Glutamate-tRNA ligase activity
molecular_function	GO:0004827	1	1.36E-04	Proline-tRNA ligase activity
molecular_function	GO:0020037	1	5.22E-06	Heme binding
biological_process	GO:0006424	1	2.02E-05	Glutamyl-tRNA aminoacylation
biological_process	GO:0006433	1	1.36E-04	Prolyl-tRNA aminoacylation
biological_process	GO:0006596	1	3.16E-06	Polyamine biosynthetic process
biological_process	GO:0016925	1	2.30E-07	Protein sumoylation
Acyrthosiphon pisum				
molecular_function	GO:0003676	1	4.01E-14	Nucleic acid binding
molecular_function	GO:0003678	1	2.18E-11	DNA helicase activity
molecular_function	GO:0003715	1	4.88E-05	Obsolete transcription termination factor activity
molecular_function	GO:0004177	1	8.98E-13	Aminopeptidase activity
molecular_function	GO:0004252	1	3.43E-19	Serine-type endopeptidase activity
molecular_function	GO:0004313	1	7.14E-05	[acyl-carrier-protein] *S*-acetyltransferase activity
molecular_function	GO:0004317	1	7.14E-05	3-Hydroxypalmitoyl-[acyl-carrier-protein] dehydratase activity
molecular_function	GO:0004319	1	7.14E-05	Enoyl-[acyl-carrier-protein] reductase (NADPH, B-specific) activity
molecular_function	GO:0004320	1	7.14E-05	Oleoyl-[acyl-carrier-protein] hydrolase activity
molecular_function	GO:0004601	1	5.92E-23	Peroxidase activity
molecular_function	GO:0004748	1	7.45E-05	Ribonucleoside-diphosphate reductase activity, thioredoxin disulfide as acceptor
molecular_function	GO:0004888	1	2.47E-05	Transmembrane signaling receptor activity
molecular_function	GO:0005506	1	6.66E-07	Iron ion binding
molecular_function	GO:0005549	1	4.97E-04	Odorant binding
molecular_function	GO:0008234	1	9.62E-10	Cysteine-type peptidase activity
molecular_function	GO:0008237	1	1.20E-10	Metallopeptidase activity
molecular_function	GO:0016295	1	7.14E-05	Myristoyl-[acyl-carrier-protein] hydrolase activity
molecular_function	GO:0016296	1	7.14E-05	Palmitoyl-[acyl-carrier-protein] hydrolase activity

(continued)

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

Table 1. Continued

Term Category	Term	Term Level	Adj. *P*-Value	Term Name
molecular_function	GO:0016705	1	1.03E-06	Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
molecular_function	GO:0020037	1	1.23E-23	Heme binding
molecular_function	GO:0030170	1	8.55E-04	Pyridoxal phosphate binding
molecular_function	GO:0043169	1	5.10E-04	Cation binding
cellular_component	GO:0005581	1	1.03E-06	Collagen trimer
cellular_component	GO:0035012	1	7.45E-05	Polytene chromosome, telomeric region
cellular_component	GO:0042600	1	9.51E-11	Chorion
biological_process	GO:0000723	1	2.18E-11	Telomere maintenance
biological_process	GO:0005975	1	1.46E-06	Carbohydrate metabolic process
biological_process	GO:0006260	1	4.02E-04	DNA replication
biological_process	GO:0006353	1	2.29E-05	DNA-templated transcription, termination
biological_process	GO:0006508	1	1.12E-33	Proteolysis
biological_process	GO:0006857	1	1.00E-04	Oligopeptide transport
biological_process	GO:0006979	1	1.27E-14	Response to oxidative stress
biological_process	GO:0007166	1	1.04E-04	Cell surface receptor signaling pathway
biological_process	GO:0009263	1	6.11E-04	Deoxyribonucleotide biosynthetic process
biological_process	GO:0035194	1	6.11E-04	Posttranscriptional gene silencing by RNA

branch A), probably related to their adaptation to specific diet and life-style, and the other after the divergence of *C. cedri* and ancestral to the other five aphids (branch B), which could be related to a nonconifer-specific diet. To validate these findings, we analyzed the relative age of the duplications by plotting the ratio of transversions at 4-fold degenerate sites (4DTv) of paralogs mapped at the two branches with high duplication ratios (supplementary fig. S4, Supplementary Material online). We also mapped two speciation events per each phylome by plotting the 4DTv of orthologous gene pairs (see Materials and Methods). Unexpectedly, the distribution of the 4DTv of both waves of duplications detected by phylostratigraphy was fully overlapping within the period of time corresponding to the most ancestral duplication. This suggests that the most recent peak of duplications detected through topological analyses of gene trees may result from more ancestral duplications followed by loss of both paralogs in *C. cedri*, therefore rendering a topology, that indicates a more recent duplication event. Consistent with this interpretation, 70% of the genes duplicated in branch B do not have an ortholog in *C. cedri*. From these observations, we conclude that the second apparent duplication peak is actually the result of differential retention of duplicates and our limited sampling of early-branching lineages. Such large levels of differential retention of duplicates have also been observed in other organisms, such as *Paramecium* (McGrath et al. 2014) and *Brassica* (Mun et al. 2009).

To test for the robustness of the detected ancestral wave of duplications, we applied stronger filters, by considering gene trees containing ancestral aphid duplications with a maximum of five genes per aphid. In all five phylomes, an average of 76% of gene trees passed this filter: *C. cedri*—11,304 (78%), *S. flava*—9,781 (75%), *Ap. glycines*—13,170 (80%), *D. noxia*—10,379 (85%), *M. persicae*—11,759 (70%), and *A. pisum*—12,485 (71%). When the duplication ratios were calculated using this more restricted set of gene trees, only the duplication ratio at the ancestral branch of all six aphids was still apparent (see Materials and Methods; supplementary fig. S5, Supplementary Material online). Taken together, these results suggest that there was one large-scale genome duplication in the evolutionary history of aphids predating the divergence of the Aphididae family, which could be related to adaptive innovations. A functional enrichment analysis of the proteins duplicated in the ancestral branch of the six aphids showed enrichment for annotations related to carbohydrate metabolic process, response to stimulus, olfactory receptor activity, odorant binding, glucuronidation, transmembrane transporter activity, and DNA and RNA processing, among others (table 3).

The Ancestral Wave of Duplications Predates the Divergence of Aphids and Adelgids

Given the long branch subtending Aphididae, and to provide a narrower placement of the ancestral duplication wave, we expanded our taxonomic sampling by including the transcriptomes of two additional hemipteran insects from the suborder Sternorrhincha of taxonomic importance for our group of study: the adelgid (Adelgidae) *Adelges tsugae* (accession number: PRJNA242203) and the scale insect (Coccoidea) *Paratachardina pseudolobata* (Christodoulides et al. 2017). Most phylogenies show that the Adelgidae family is a sister group of the Phylloxeridae family (Heie and Wegierek 2009; Vilcinskas 2016). Thus, with the inclusion of *Ad. tsugae* we can obtain a general image of the Aphidomorpha lineage. With this increased species set, we reconstructed an expanded *C. cedri* phylome and species tree (fig. 3*a*). The duplication analysis on the expanded data set initially resulted in two

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

Table 2. List of GO Terms Enriched in the Parallel *Cinara cedri* Species-Specific Duplications Uniquely Shared with One of the Other Aphids (*Sipha flava*, *Aphis glycines*, *Myzus persicae*, and *Acyrthosiphon pisum*).

Term Category	Term	Term Level	Adj. *P*-Value	Term Name
Sipha flava				
molecular_function	GO:0003994	1	1.93E-13	Aconitate hydratase activity
molecular_function	GO:0004497	1	6.19E-11	Monooxygenase activity
molecular_function	GO:0004535	1	1.84E-04	Poly(A)-specific ribonuclease activity
molecular_function	GO:0005200	1	1.34E-08	Structural constituent of cytoskeleton
molecular_function	GO:0005506	1	1.23E-23	Iron ion binding
molecular_function	GO:0015179	1	3.88E-04	L-amino acid transmembrane transporter activity
molecular_function	GO:0016705	1	1.36E-27	Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
molecular_function	GO:0020037	1	1.46E-22	Heme binding
molecular_function	GO:0032947	1	3.88E-04	Protein complex scaffold activity
molecular_function	GO:0051539	1	1.23E-05	4 Iron, 4 sulfur cluster binding
molecular_function	GO:0070330	1	1.14E-04	Aromatase activity
cellular_component	GO:0005874	1	3.19E-05	Microtubule
cellular_component	GO:0030014	1	4.67E-05	CCR4-NOT complex
cellular_component	GO:0030015	1	4.91E-05	CCR4-NOT core complex
biological_process	GO:0000288	1	3.88E-04	Nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay
biological_process	GO:0000289	1	3.41E-05	Nuclear-transcribed mRNA poly(A) tail shortening
biological_process	GO:0006099	1	6.27E-06	Tricarboxylic acid cycle
biological_process	GO:0006402	1	4.17E-04	mRNA catabolic process
biological_process	GO:0007017	1	8.35E-08	Microtubule-based process
biological_process	GO:0017148	1	1.46E-04	Negative regulation of translation
biological_process	GO:0055085	1	9.30E-04	Transmembrane transport
Acyrthosiphon pisum				
molecular_function	GO:0004020	1	7.10E-05	Adenylylsulfate kinase activity
molecular_function	GO:0004190	1	6.34E-04	Aspartic-type endopeptidase activity
molecular_function	GO:0004656	1	1.64E-09	Procollagen-proline 4-dioxygenase activity
molecular_function	GO:0004719	1	1.57E-04	Protein-L-isoaspartate (D-aspartate) *O*-methyltransferase activity
molecular_function	GO:0016702	1	3.02E-06	Oxidoreductase activity, acting on single donors with incorporation of molecular oxygen, incorporation of two atoms of oxygen
molecular_function	GO:0031418	1	1.99E-07	L-ascorbic acid binding
biological_process	GO:0000103	1	9.52E-05	Sulfate assimilation
Aphis glycines				
molecular_function	GO:0003676	1	1.43E-07	Nucleic acid binding
molecular_function	GO:0003924	1	8.18E-04	GTPase activity
cellular_component	GO:0005741	1	3.64E-05	Mitochondrial outer membrane
biological_process	GO:0008053	1	6.78E-13	Mitochondrial fusion
biological_process	GO:0048662	1	1.99E-04	Negative regulation of smooth muscle cell proliferation
Myzus persicae				
molecular_function	GO:0003777	1	3.06E-06	Microtubule motor activity
molecular_function	GO:0003964	1	3.91E-04	RNA-directed DNA polymerase activity
biological_process	GO:0007018	1	2.38E-04	Microtubule-based movement

waves of duplications (fig. 3*b*): one (0.18 duplications/gene) still specific to the Aphididae lineage, and another (0.36 duplications/gene) at the base of the Aphidomorpha lineage (Aphididae and Adelgidae families). However, it has been previously observed that, due to their incompleteness, transcriptomic data sets make difficult the correct placement of duplications (Jiménez-Guri et al. 2013). To account for this, we repeated the analysis considering only gene trees that included the two species with transcriptomic data sets. In this stringent set, the duplication ratio in the ancestral branch of the Aphidomorpha was still high (0.29), whereas the one in the branch subtending Aphididae disappeared (0.07). To confirm the presence of a single ancestral peak we again analyzed ratios of 4DTv for the pairs of paralogs mapped at the ancestral branch of the Aphidomorpha lineage, and for the orthologous pairs found between *C. cedri* and *Ad. tsugae*, and *P. pseudolobata* (see Materials and Methods). The distribution of the 4DTv values of paralogs and orthologs is used to estimate the relative order of duplication and speciation events, respectively. Consistent with our phylogenomic analyses described above, the ancestral peak of duplications is placed before the divergence of *C. cedri* and *Ad. tsugae*, and after the divergence of *C. cedri* and *P. pseudolobata*. From these results, we conclude that the large-scale gene duplication observed at the long branch subtending Aphididae in the full-genome data set occurred before the divergence of the Aphidomorpha group, and after the separation of this lineage from *P. pseudolobata* (Coccoidea). A dating analysis (see Materials and Methods) situates this duplication wave over a putative long temporal period, 106–227 Ma. As these times are molecular estimates, additional analysis should be necessary to place a more accurate time scale for this duplication event.

A functional analysis of the *C. cedri* proteins duplicated at the base of the Aphidomorpha lineage was largely consistent

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

Table 3. List of the GO Terms Enriched in the Duplicated Protein Families at the Base of All Six Aphids per Each Phylome.

Term Category	Term	Term Level	Adj. *P*-Value	Term Name
Cinara cedri				
molecular_function	GO:0001227	1	1.35E-05	Transcriptional repressor activity, RNA polymerase II transcription regulatory region sequence-specific binding
molecular_function	GO:0003676	1	1.59E-30	Nucleic acid binding
molecular_function	GO:0003964	1	5.96E-28	RNA-directed DNA polymerase activity
molecular_function	GO:0004032	1	8.29E-05	Alditol:NADP+ 1-oxidoreductase activity
molecular_function	GO:0004185	1	3.69E-04	Serine-type carboxypeptidase activity
molecular_function	GO:0004197	1	1.23E-09	Cysteine-type endopeptidase activity
molecular_function	GO:0004316	1	4.99E-05	3-Oxoacyl-[acyl-carrier-protein] reductase (NADPH) activity
molecular_function	GO:0004497	1	7.28E-05	Monooxygenase activity
molecular_function	GO:0004523	1	3.51E-06	RNA–DNA hybrid ribonuclease activity
molecular_function	GO:0004553	1	7.45E-11	Hydrolase activity, hydrolyzing *O*-glycosyl compounds
molecular_function	GO:0004555	1	1.14E-05	Alpha, alpha-trehalase activity
molecular_function	GO:0004984	1	1.03E-12	Olfactory receptor activity
molecular_function	GO:0005215	1	5.81E-08	Transporter activity
molecular_function	GO:0005254	1	6.47E-05	Chloride channel activity
molecular_function	GO:0005355	1	6.09E-04	Glucose transmembrane transporter activity
molecular_function	GO:0005506	1	2.92E-09	Iron ion binding
molecular_function	GO:0005542	1	1.12E-04	Folic acid binding
molecular_function	GO:0005549	1	3.15E-07	Odorant binding
molecular_function	GO:0008194	1	3.76E-04	UDP-glycosyltransferase activity
molecular_function	GO:0008234	1	1.18E-07	Cysteine-type peptidase activity
molecular_function	GO:0008417	1	1.12E-04	Fucosyltransferase activity
molecular_function	GO:0008422	1	9.39E-04	Beta-glucosidase activity
molecular_function	GO:0008518	1	1.12E-04	Reduced folate carrier activity
molecular_function	GO:0008521	1	3.90E-17	Acetyl-CoA transporter activity
molecular_function	GO:0015020	1	1.96E-13	Glucuronosyltransferase activity
molecular_function	GO:0015171	1	1.55E-04	Amino acid transmembrane transporter activity
molecular_function	GO:0015295	1	1.03E-06	Solute:proton symporter activity
molecular_function	GO:0015297	1	9.39E-04	Antiporter activity
molecular_function	GO:0015299	1	9.39E-04	Solute:proton antiporter activity
molecular_function	GO:0015528	1	3.51E-06	Lactose:proton symporter activity
molecular_function	GO:0016705	1	1.25E-13	Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
molecular_function	GO:0016758	1	1.23E-20	Transferase activity, transferring hexosyl groups
molecular_function	GO:0016788	1	9.89E-04	Hydrolase activity, acting on ester bonds
molecular_function	GO:0017110	1	3.60E-04	Nucleoside-diphosphatase activity
molecular_function	GO:0018024	1	4.89E-06	Histone-lysine *N*-methyltransferase activity
molecular_function	GO:0019799	1	3.70E-05	Tubulin *N*-acetyltransferase activity
molecular_function	GO:0020037	1	2.79E-11	Heme binding
molecular_function	GO:0022857	1	7.61E-25	Transmembrane transporter activity
molecular_function	GO:0022891	1	9.33E-06	Substrate-specific transmembrane transporter activity
molecular_function	GO:0031490	1	3.60E-04	Chromatin DNA binding
molecular_function	GO:0035197	1	2.40E-04	siRNA binding
molecular_function	GO:0043169	1	3.94E-06	Cation binding
molecular_function	GO:0090482	1	1.12E-04	Vitamin transmembrane transporter activity
molecular_function	GO:0102336	1	9.83E-04	3-Oxo-arachidoyl-CoA synthase activity
molecular_function	GO:0102337	1	9.83E-04	3-Oxo-cerotoyl-CoA synthase activity
molecular_function	GO:0102338	1	9.83E-04	3-Oxo-lignoceronyl-CoA synthase activity
cellular_component	GO:0008537	1	1.12E-04	Proteasome activator complex
biological_process	GO:0005975	1	6.65E-05	Carbohydrate metabolic process
biological_process	GO:0005991	1	1.14E-05	Trehalose metabolic process
biological_process	GO:0006278	1	6.80E-24	RNA-dependent DNA biosynthetic process
biological_process	GO:0006310	1	1.25E-05	DNA recombination
biological_process	GO:0006508	1	2.74E-10	Proteolysis
biological_process	GO:0006629	1	7.28E-05	Lipid metabolic process
biological_process	GO:0006820	1	8.83E-05	Anion transport
biological_process	GO:0007283	1	2.35E-05	Spermatogenesis
biological_process	GO:0007608	1	8.86E-07	Sensory perception of smell
biological_process	GO:0015074	1	1.08E-16	DNA integration
biological_process	GO:0016973	1	9.39E-04	Poly(A)+ mRNA export from nucleus
biological_process	GO:0030162	1	9.39E-04	Regulation of proteolysis
biological_process	GO:0035428	1	3.03E-04	Hexose transmembrane transport

(continued)

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

Table 3. Continued

Term Category	Term	Term Level	Adj. *P*-Value	Term Name
biological_process	GO:0046323	1	3.03E-04	Glucose import
biological_process	GO:0050790	1	2.49E-07	Regulation of catalytic activity
biological_process	GO:0051180	1	1.12E-04	Vitamin transport
biological_process	GO:0051603	1	4.08E-04	Proteolysis involved in cellular protein catabolic process
biological_process	GO:0052696	1	1.12E-04	Flavonoid glucuronidation
biological_process	GO:0055085	1	9.23E-34	Transmembrane transport
biological_process	GO:0055114	1	2.21E-04	Oxidation–reduction process
biological_process	GO:0071929	1	3.70E-05	Alpha-tubulin acetylation
biological_process	GO:1901657	1	3.70E-05	Glycosyl compound metabolic process
Sipha flava				
molecular_function	GO:0001227	1	1.78E-45	Transcriptional repressor activity, RNA polymerase II transcription regulatory region sequence-specific binding
molecular_function	GO:0003676	1	2.51E-19	Nucleic acid binding
molecular_function	GO:0003677	1	2.55E-06	DNA binding
molecular_function	GO:0003678	1	5.62E-15	DNA helicase activity
molecular_function	GO:0003690	1	1.36E-30	Double-stranded DNA binding
molecular_function	GO:0003715	1	4.74E-16	Obsolete transcription termination factor activity
molecular_function	GO:0003964	1	7.50E-07	RNA-directed DNA polymerase activity
molecular_function	GO:0004316	1	7.50E-07	3-Oxoacyl-[acyl-carrier-protein] reductase (NADPH) activity
molecular_function	GO:0004497	1	3.17E-04	Monooxygenase activity
molecular_function	GO:0004553	1	1.08E-08	Hydrolase activity, hydrolyzing *O*-glycosyl compounds
molecular_function	GO:0004555	1	2.62E-06	Alpha, alpha-trehalase activity
molecular_function	GO:0004565	1	1.64E-04	Beta-galactosidase activity
molecular_function	GO:0004984	1	1.45E-14	Olfactory receptor activity
molecular_function	GO:0005215	1	8.75E-09	Transporter activity
molecular_function	GO:0005351	1	2.79E-06	Sugar:proton symporter activity
molecular_function	GO:0005355	1	2.59E-07	Glucose transmembrane transporter activity
molecular_function	GO:0005506	1	2.01E-05	Iron ion binding
molecular_function	GO:0005549	1	1.14E-11	Odorant binding
molecular_function	GO:0008194	1	4.71E-07	UDP-glycosyltransferase activity
molecular_function	GO:0008521	1	1.51E-10	Acetyl-CoA transporter activity
molecular_function	GO:0010521	1	7.20E-04	Telomerase inhibitor activity
molecular_function	GO:0015020	1	4.52E-10	Glucuronosyltransferase activity
molecular_function	GO:0016614	1	9.74E-05	Oxidoreductase activity, acting on CH–OH group of donors
molecular_function	GO:0016705	1	9.69E-08	Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
molecular_function	GO:0016758	1	8.93E-18	Transferase activity, transferring hexosyl groups
molecular_function	GO:0020037	1	4.91E-12	Heme binding
molecular_function	GO:0022857	1	3.25E-21	Transmembrane transporter activity
molecular_function	GO:0022891	1	1.02E-06	Substrate-specific transmembrane transporter activity
molecular_function	GO:0042302	1	3.45E-07	Structural constituent of cuticle
molecular_function	GO:0043169	1	4.58E-06	Cation binding
cellular_component	GO:0043231	1	3.60E-04	Intracellular membrane-bounded organelle
biological_process	GO:0000723	1	1.59E-18	Telomere maintenance
biological_process	GO:0005991	1	2.62E-06	Trehalose metabolic process
biological_process	GO:0006281	1	1.66E-05	DNA repair
biological_process	GO:0006353	1	1.20E-16	DNA-templated transcription, termination
biological_process	GO:0006508	1	6.34E-06	Proteolysis
biological_process	GO:0006857	1	8.42E-05	Oligopeptide transport
biological_process	GO:0007608	1	4.80E-06	Sensory perception of smell
biological_process	GO:0032211	1	7.20E-04	Negative regulation of telomere maintenance via telomerase
biological_process	GO:0035428	1	2.59E-07	Hexose transmembrane transport
biological_process	GO:0044806	1	7.20E-04	G-quadruplex DNA unwinding
biological_process	GO:0046323	1	2.59E-07	Glucose import
biological_process	GO:0050909	1	4.30E-05	Sensory perception of taste
biological_process	GO:0051974	1	7.20E-04	Negative regulation of telomerase activity
biological_process	GO:0052696	1	4.87E-08	Flavonoid glucuronidation
biological_process	GO:0055085	1	6.34E-26	Transmembrane transport
Aphis glycines				
molecular_function	GO:0001227	1	1.41E-09	Transcriptional repressor activity, RNA polymerase II transcription regulatory region sequence-specific binding
molecular_function	GO:0003676	1	5.84E-43	Nucleic acid binding
molecular_function	GO:0003690	1	3.51E-05	Double-stranded DNA binding

(continued)

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

Table 3. Continued

Term Category	Term	Term Level	Adj. *P*-Value	Term Name
molecular_function	GO:0003715	1	3.33E-20	Obsolete transcription termination factor activity
molecular_function	GO:0003964	1	1.45E-05	RNA-directed DNA polymerase activity
molecular_function	GO:0004197	1	1.97E-06	Cysteine-type endopeptidase activity
molecular_function	GO:0004497	1	9.76E-06	Monooxygenase activity
molecular_function	GO:0004553	1	7.60E-08	Hydrolase activity, hydrolyzing *O*-glycosyl compounds
molecular_function	GO:0004984	1	2.42E-20	Olfactory receptor activity
molecular_function	GO:0005351	1	1.26E-06	Sugar:proton symporter activity
molecular_function	GO:0005355	1	7.12E-08	Glucose transmembrane transporter activity
molecular_function	GO:0005506	1	2.39E-07	Iron ion binding
molecular_function	GO:0005549	1	5.98E-17	Odorant binding
molecular_function	GO:0008194	1	1.17E-06	UDP-glycosyltransferase activity
molecular_function	GO:0008234	1	6.93E-04	Cysteine-type peptidase activity
molecular_function	GO:0008521	1	2.76E-21	Acetyl-CoA transporter activity
molecular_function	GO:0015020	1	7.06E-13	Glucuronosyltransferase activity
molecular_function	GO:0015295	1	3.29E-07	Solute:proton symporter activity
molecular_function	GO:0016705	1	7.31E-10	Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
molecular_function	GO:0016758	1	4.95E-25	Transferase activity, transferring hexosyl groups
molecular_function	GO:0019789	1	2.14E-08	SUMO transferase activity
molecular_function	GO:0019799	1	9.15E-05	Tubulin *N*-acetyltransferase activity
molecular_function	GO:0020037	1	2.22E-15	Heme binding
molecular_function	GO:0022857	1	7.22E-14	Transmembrane transporter activity
molecular_function	GO:0022891	1	9.76E-06	Substrate-specific transmembrane transporter activity
molecular_function	GO:0042626	1	1.89E-05	ATPase activity, coupled to transmembrane movement of substances
molecular_function	GO:0043169	1	1.95E-04	Cation binding
molecular_function	GO:0050660	1	3.55E-04	Flavin adenine dinucleotide binding
molecular_function	GO:0090482	1	6.11E-04	Vitamin transmembrane transporter activity
cellular_component	GO:0008537	1	5.67E-06	Proteasome activator complex
cellular_component	GO:0035012	1	1.47E-05	Polytene chromosome, telomeric region
biological_process	GO:0005975	1	5.43E-06	Carbohydrate metabolic process
biological_process	GO:0006353	1	3.25E-18	DNA-templated transcription, termination
biological_process	GO:0006508	1	1.08E-05	Proteolysis
biological_process	GO:0006629	1	8.02E-04	Lipid metabolic process
biological_process	GO:0007095	1	1.30E-04	Mitotic G2 DNA damage checkpoint
biological_process	GO:0007446	1	2.35E-06	Imaginal disc growth
biological_process	GO:0007608	1	1.34E-10	Sensory perception of smell
biological_process	GO:0030097	1	9.79E-04	Hemopoiesis
biological_process	GO:0032968	1	6.93E-04	Positive regulation of transcription elongation from RNA polymerase II promoter
biological_process	GO:0035428	1	7.12E-08	Hexose transmembrane transport
biological_process	GO:0042176	1	6.93E-04	Regulation of protein catabolic process
biological_process	GO:0046323	1	7.12E-08	Glucose import
biological_process	GO:0046425	1	1.27E-06	Regulation of JAK-STAT cascade
biological_process	GO:0050790	1	1.27E-06	Regulation of catalytic activity
biological_process	GO:0050909	1	6.62E-04	Sensory perception of taste
biological_process	GO:0051180	1	6.11E-04	Vitamin transport
biological_process	GO:0052696	1	2.03E-07	Flavonoid glucuronidation
biological_process	GO:0055085	1	2.74E-19	Transmembrane transport
biological_process	GO:0071929	1	9.15E-05	Alpha-tubulin acetylation
biological_process	GO:1901657	1	6.11E-04	Glycosyl compound metabolic process
Diuraphis noxia				
molecular_function	GO:0000064	1	5.19E-04	*L*-ornithine transmembrane transporter activity
molecular_function	GO:0003676	1	8.65E-17	Nucleic acid binding
molecular_function	GO:0003678	1	3.25E-04	DNA helicase activity
molecular_function	GO:0003715	1	5.79E-09	Obsolete transcription termination factor activity
molecular_function	GO:0003964	1	8.84E-05	RNA-directed DNA polymerase activity
molecular_function	GO:0004197	1	4.18E-06	Cysteine-type endopeptidase activity
molecular_function	GO:0004316	1	3.19E-04	3-Oxoacyl-[acyl-carrier-protein] reductase (NADPH) activity
molecular_function	GO:0004396	1	3.28E-04	Hexokinase activity
molecular_function	GO:0004497	1	1.51E-10	Monooxygenase activity
molecular_function	GO:0004553	1	1.67E-07	Hydrolase activity, hydrolyzing *O*-glycosyl compounds
molecular_function	GO:0004555	1	3.76E-09	Alpha, alpha-trehalase activity
molecular_function	GO:0004601	1	2.82E-05	Peroxidase activity

(continued)

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

Table 3. Continued

Term Category	Term	Term Level	Adj. *P*-Value	Term Name
molecular_function	GO:0004984	1	1.80E-04	Olfactory receptor activity
molecular_function	GO:0005215	1	3.17E-07	Transporter activity
molecular_function	GO:0005351	1	2.13E-10	Sugar:proton symporter activity
molecular_function	GO:0005355	1	1.01E-11	Glucose transmembrane transporter activity
molecular_function	GO:0005506	1	1.93E-09	Iron ion binding
molecular_function	GO:0005536	1	3.28E-04	Glucose binding
molecular_function	GO:0005549	1	5.44E-04	Odorant binding
molecular_function	GO:0008194	1	1.78E-07	UDP-glycosyltransferase activity
molecular_function	GO:0008234	1	5.08E-05	Cysteine-type peptidase activity
molecular_function	GO:0008424	1	3.28E-04	Glycoprotein 6-alpha-*L*-fucosyltransferase activity
molecular_function	GO:0008521	1	3.96E-22	Acetyl-CoA transporter activity
molecular_function	GO:0015020	1	5.59E-13	Glucuronosyltransferase activity
molecular_function	GO:0015174	1	8.84E-05	Basic amino acid transmembrane transporter activity
molecular_function	GO:0015181	1	5.19E-04	Arginine transmembrane transporter activity
molecular_function	GO:0015189	1	5.19E-04	L-lysine transmembrane transporter activity
molecular_function	GO:0015295	1	2.27E-05	Solute:proton symporter activity
molecular_function	GO:0015326	1	8.84E-05	Basic amino acid transmembrane transporter activity
molecular_function	GO:0016614	1	3.76E-09	Oxidoreductase activity, acting on CH–OH group of donors
molecular_function	GO:0016705	1	2.96E-16	Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
molecular_function	GO:0016758	1	1.58E-24	Transferase activity, transferring hexosyl groups
molecular_function	GO:0016788	1	3.00E-04	Hydrolase activity, acting on ester bonds
molecular_function	GO:0016872	1	8.84E-05	Intramolecular lyase activity
molecular_function	GO:0019799	1	2.27E-05	Tubulin *N*-acetyltransferase activity
molecular_function	GO:0020037	1	1.52E-22	Heme binding
molecular_function	GO:0022857	1	3.73E-21	Transmembrane transporter activity
molecular_function	GO:0022891	1	3.09E-08	Substrate-specific transmembrane transporter activity
molecular_function	GO:0042626	1	1.91E-07	ATPase activity, coupled to transmembrane movement of substances
molecular_function	GO:0043169	1	9.84E-06	Cation binding
molecular_function	GO:0050660	1	2.64E-06	Flavin adenine dinucleotide binding
cellular_component	GO:0005874	1	7.95E-04	Microtubule
cellular_component	GO:0016021	1	1.56E-06	Integral component of membrane
cellular_component	GO:0031461	1	3.19E-04	Cullin-RING ubiquitin ligase complex
cellular_component	GO:0043231	1	1.28E-05	Intracellular membrane-bounded organelle
biological_process	GO:0000723	1	2.83E-04	Telomere maintenance
biological_process	GO:0001678	1	3.28E-04	Cellular glucose homeostasis
biological_process	GO:0005975	1	9.76E-06	Carbohydrate metabolic process
biological_process	GO:0005991	1	3.76E-09	Trehalose metabolic process
biological_process	GO:0005993	1	3.28E-04	Trehalose catabolic process
biological_process	GO:0006096	1	8.93E-05	Glycolytic process
biological_process	GO:0006352	1	6.52E-05	DNA-templated transcription, initiation
biological_process	GO:0006353	1	5.79E-09	DNA-templated transcription, termination
biological_process	GO:0006508	1	3.17E-08	Proteolysis
biological_process	GO:0006629	1	1.16E-06	Lipid metabolic process
biological_process	GO:0006865	1	8.84E-05	Amino acid transport
biological_process	GO:0007352	1	3.28E-04	Zygotic specification of dorsal/ventral axis
biological_process	GO:0009452	1	3.28E-04	7-Methylguanosine RNA capping
biological_process	GO:0033578	1	3.28E-04	Protein glycosylation in Golgi
biological_process	GO:0035428	1	1.01E-11	Hexose transmembrane transport
biological_process	GO:0046323	1	1.01E-11	Glucose import
biological_process	GO:0050790	1	5.25E-04	Regulation of catalytic activity
biological_process	GO:0050909	1	4.68E-04	Sensory perception of taste
biological_process	GO:0052696	1	1.60E-08	Flavonoid glucuronidation
biological_process	GO:0055085	1	6.40E-25	Transmembrane transport
biological_process	GO:0055114	1	1.71E-04	Oxidation-reduction process
biological_process	GO:0071929	1	2.27E-05	Alpha-tubulin acetylation
biological_process	GO:1903352	1	5.19E-04	*L*-ornithine transmembrane transport
Myzus persicae				
molecular_function	GO:0001227	1	3.22E-16	Transcriptional repressor activity, RNA polymerase II transcription regulatory region sequence-specific binding
molecular_function	GO:0003676	1	3.81E-65	Nucleic acid binding
molecular_function	GO:0003677	1	6.51E-16	DNA binding
molecular_function	GO:0003690	1	1.01E-08	Double-stranded DNA binding

(continued)

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

Table 3. Continued

Term Category	Term	Term Level	Adj. *P*-Value	Term Name
molecular_function	GO:0003950	1	5.43E-07	NAD+ ADP-ribosyltransferase activity
molecular_function	GO:0004185	1	3.40E-04	Serine-type carboxypeptidase activity
molecular_function	GO:0004197	1	3.56E-09	Cysteine-type endopeptidase activity
molecular_function	GO:0004252	1	1.33E-05	Serine-type endopeptidase activity
molecular_function	GO:0004316	1	1.38E-04	3-Oxoacyl-[acyl-carrier-protein] reductase (NADPH) activity
molecular_function	GO:0004396	1	1.87E-05	Hexokinase activity
molecular_function	GO:0004497	1	2.72E-07	Monooxygenase activity
molecular_function	GO:0004553	1	1.59E-07	Hydrolase activity, hydrolyzing *O*-glycosyl compounds
molecular_function	GO:0004555	1	4.59E-09	Alpha, alpha-trehalase activity
molecular_function	GO:0004565	1	1.86E-04	Beta-galactosidase activity
molecular_function	GO:0004601	1	7.64E-06	Peroxidase activity
molecular_function	GO:0004984	1	1.49E-21	Olfactory receptor activity
molecular_function	GO:0005215	1	2.40E-04	Transporter activity
molecular_function	GO:0005351	1	8.69E-06	Sugar:proton symporter activity
molecular_function	GO:0005355	1	1.92E-07	Glucose transmembrane transporter activity
molecular_function	GO:0005506	1	7.01E-10	Iron ion binding
molecular_function	GO:0005536	1	1.87E-05	Glucose binding
molecular_function	GO:0005549	1	9.69E-19	Odorant binding
molecular_function	GO:0008194	1	1.02E-05	UDP-glycosyltransferase activity
molecular_function	GO:0008234	1	3.73E-08	Cysteine-type peptidase activity
molecular_function	GO:0008236	1	3.71E-04	Serine-type peptidase activity
molecular_function	GO:0008521	1	2.99E-20	Acetyl-CoA transporter activity
molecular_function	GO:0015020	1	8.81E-14	Glucuronosyltransferase activity
molecular_function	GO:0015174	1	1.86E-04	Basic amino acid transmembrane transporter activity
molecular_function	GO:0015295	1	6.02E-05	Solute:proton symporter activity
molecular_function	GO:0015326	1	1.86E-04	Basic amino acid transmembrane transporter activity
molecular_function	GO:0016705	1	2.44E-17	Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
molecular_function	GO:0016758	1	3.11E-27	Transferase activity, transferring hexosyl groups
molecular_function	GO:0016765	1	5.72E-05	Transferase activity, transferring alkyl or aryl (other than methyl) groups
molecular_function	GO:0016872	1	1.59E-07	Intramolecular lyase activity
molecular_function	GO:0019799	1	1.87E-05	Tubulin *N*-acetyltransferase activity
molecular_function	GO:0020037	1	4.37E-20	Heme binding
molecular_function	GO:0022857	1	1.36E-16	Transmembrane transporter activity
molecular_function	GO:0022891	1	1.30E-04	Substrate-specific transmembrane transporter activity
molecular_function	GO:0031177	1	3.23E-04	Phosphopantetheine binding
molecular_function	GO:0033777	1	5.89E-04	Lithocholate 6beta-hydroxylase activity
molecular_function	GO:0042302	1	5.89E-04	Structural constituent of cuticle
molecular_function	GO:0043169	1	1.28E-04	Cation binding
cellular_component	GO:0032590	1	2.21E-04	Dendrite membrane
biological_process	GO:0001678	1	1.87E-05	Cellular glucose homeostasis
biological_process	GO:0005975	1	2.35E-05	Carbohydrate metabolic process
biological_process	GO:0005991	1	4.59E-09	Trehalose metabolic process
biological_process	GO:0005993	1	1.86E-04	Trehalose catabolic process
biological_process	GO:0006508	1	5.25E-17	Proteolysis
biological_process	GO:0006857	1	5.03E-07	Oligopeptide transport
biological_process	GO:0006865	1	1.86E-04	Amino acid transport
biological_process	GO:0007608	1	6.70E-12	Sensory perception of smell
biological_process	GO:0035428	1	1.92E-07	Hexose transmembrane transport
biological_process	GO:0046323	1	1.92E-07	Glucose import
biological_process	GO:0050790	1	2.38E-08	Regulation of catalytic activity
biological_process	GO:0050896	1	7.80E-04	Response to stimulus
biological_process	GO:0050909	1	6.70E-12	Sensory perception of taste
biological_process	GO:0051603	1	5.47E-05	Proteolysis involved in cellular protein catabolic process
biological_process	GO:0052696	1	1.59E-07	Flavonoid glucuronidation
biological_process	GO:0055085	1	1.33E-13	Transmembrane transport
biological_process	GO:0071929	1	1.87E-05	Alpha-tubulin acetylation
Acyrthosiphon pisum				
molecular_function	GO:0000064	1	5.99E-05	L-ornithine transmembrane transporter activity
molecular_function	GO:0001227	1	7.63E-08	Transcriptional repressor activity, RNA polymerase II transcription regulatory region sequence-specific binding
molecular_function	GO:0003676	1	7.77E-87	Nucleic acid binding
molecular_function	GO:0003678	1	1.82E-10	DNA helicase activity

(continued)

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

Table 3. Continued

Term Category	Term	Term Level	Adj. *P*-Value	Term Name
molecular_function	GO:0003715	1	3.16E-08	Obsolete transcription termination factor activity
molecular_function	GO:0003743	1	9.37E-05	Translation initiation factor activity
molecular_function	GO:0003950	1	1.71E-07	NAD+ ADP-ribosyltransferase activity
molecular_function	GO:0003964	1	3.65E-04	RNA-directed DNA polymerase activity
molecular_function	GO:0004197	1	4.77E-08	Cysteine-type endopeptidase activity
molecular_function	GO:0004252	1	5.21E-21	Serine-type endopeptidase activity
molecular_function	GO:0004316	1	3.66E-04	3-Oxoacyl-[acyl-carrier-protein] reductase (NADPH) activity
molecular_function	GO:0004497	1	8.16E-07	Monooxygenase activity
molecular_function	GO:0004553	1	2.92E-06	Hydrolase activity, hydrolyzing *O*-glycosyl compounds
molecular_function	GO:0004555	1	3.35E-05	Alpha, alpha-trehalase activity
molecular_function	GO:0004565	1	2.51E-05	Beta-galactosidase activity
molecular_function	GO:0004748	1	2.47E-04	Ribonucleoside-diphosphate reductase activity, thioredoxin disulfide as acceptor
molecular_function	GO:0004984	1	1.05E-24	Olfactory receptor activity
molecular_function	GO:0005215	1	3.82E-04	Transporter activity
molecular_function	GO:0005351	1	3.61E-04	Sugar:proton symporter activity
molecular_function	GO:0005355	1	9.37E-05	Glucose transmembrane transporter activity
molecular_function	GO:0005506	1	4.78E-10	Iron ion binding
molecular_function	GO:0005549	1	1.29E-27	Odorant binding
molecular_function	GO:0008234	1	6.87E-11	Cysteine-type peptidase activity
molecular_function	GO:0008408	1	3.61E-04	3′–5′ exonuclease activity
molecular_function	GO:0008417	1	4.30E-04	Fucosyltransferase activity
molecular_function	GO:0008424	1	6.80E-05	Glycoprotein 6-alpha-L-fucosyltransferase activity
molecular_function	GO:0008521	1	4.46E-25	Acetyl-CoA transporter activity
molecular_function	GO:0015020	1	1.56E-15	Glucuronosyltransferase activity
molecular_function	GO:0015174	1	8.07E-06	Basic amino acid transmembrane transporter activity
molecular_function	GO:0015181	1	5.99E-05	Arginine transmembrane transporter activity
molecular_function	GO:0015189	1	5.99E-05	L-lysine transmembrane transporter activity
molecular_function	GO:0015295	1	8.16E-07	Solute:proton symporter activity
molecular_function	GO:0015326	1	8.07E-06	Basic amino acid transmembrane transporter activity
molecular_function	GO:0016705	1	2.52E-16	Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
molecular_function	GO:0016758	1	4.79E-27	Transferase activity, transferring hexosyl groups
molecular_function	GO:0016872	1	5.92E-04	Intramolecular lyase activity
molecular_function	GO:0019789	1	1.64E-06	SUMO transferase activity
molecular_function	GO:0019799	1	2.47E-06	Tubulin *N*-acetyltransferase activity
molecular_function	GO:0020037	1	1.33E-16	Heme binding
molecular_function	GO:0022857	1	1.26E-16	Transmembrane transporter activity
molecular_function	GO:0022891	1	5.49E-05	Substrate-specific transmembrane transporter activity
molecular_function	GO:0031177	1	3.66E-04	Phosphopantetheine binding
molecular_function	GO:0033777	1	6.80E-05	Lithocholate 6beta-hydroxylase activity
molecular_function	GO:0042302	1	4.06E-17	Structural constituent of cuticle
molecular_function	GO:0043169	1	5.68E-05	Cation binding
molecular_function	GO:0046921	1	2.03E-04	Alpha-(1->6)-fucosyltransferase activity
molecular_function	GO:0080019	1	9.50E-07	Fatty-acyl-CoA reductase (alcohol-forming) activity
cellular_component	GO:0005868	1	3.65E-04	Cytoplasmic dynein complex
cellular_component	GO:0016281	1	1.57E-04	Eukaryotic translation initiation factor 4F complex
cellular_component	GO:0032580	1	9.48E-04	Golgi cisterna membrane
cellular_component	GO:0034388	1	3.65E-04	Pwp2p-containing subcomplex of 90S preribosome
cellular_component	GO:0043186	1	3.58E-04	P granule
biological_process	GO:0000082	1	9.14E-04	G1/S transition of mitotic cell cycle
biological_process	GO:0000723	1	2.00E-11	Telomere maintenance
biological_process	GO:0005975	1	3.92E-04	Carbohydrate metabolic process
biological_process	GO:0005991	1	3.35E-05	Trehalose metabolic process
biological_process	GO:0006352	1	2.41E-04	DNA-templated transcription, initiation
biological_process	GO:0006353	1	8.24E-09	DNA-templated transcription, termination
biological_process	GO:0006508	1	2.72E-21	Proteolysis
biological_process	GO:0006629	1	9.50E-07	Lipid metabolic process
biological_process	GO:0006857	1	2.68E-11	Oligopeptide transport
biological_process	GO:0006865	1	2.47E-04	Amino acid transport
biological_process	GO:0007095	1	3.77E-05	Mitotic G2 DNA damage checkpoint
biological_process	GO:0007179	1	6.47E-05	Transforming growth factor beta-receptor signaling pathway
biological_process	GO:0007608	1	2.29E-11	Sensory perception of smell

(continued)

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

Table 3. Continued

Term Category	Term	Term Level	Adj. *P*-Value	Term Name
biological_process	GO:0009166	1	3.82E-04	Nucleotide catabolic process
biological_process	GO:0009452	1	2.42E-07	7-Methylguanosine RNA capping
biological_process	GO:0009953	1	3.48E-05	Dorsal/ventral pattern formation
biological_process	GO:0010025	1	9.50E-07	Wax biosynthetic process
biological_process	GO:0010629	1	2.41E-04	Negative regulation of gene expression
biological_process	GO:0033578	1	6.80E-05	Protein glycosylation in Golgi
biological_process	GO:0035336	1	9.50E-07	Long-chain fatty-acyl-CoA metabolic process
biological_process	GO:0035428	1	9.37E-05	Hexose transmembrane transport
biological_process	GO:0045705	1	9.37E-05	Negative regulation of salivary gland boundary specification
biological_process	GO:0046323	1	9.37E-05	Glucose import
biological_process	GO:0046368	1	2.03E-04	GDP-L-fucose metabolic process
biological_process	GO:0050790	1	1.45E-08	Regulation of catalytic activity
biological_process	GO:0050909	1	1.01E-05	Sensory perception of taste
biological_process	GO:0051603	1	1.36E-05	Proteolysis involved in cellular protein catabolic process
biological_process	GO:0052696	1	2.47E-04	Flavonoid glucuronidation
biological_process	GO:0055085	1	1.29E-14	Transmembrane transport
biological_process	GO:0071929	1	2.47E-06	Alpha-tubulin acetylation
biological_process	GO:1903352	1	5.99E-05	L-ornithine transmembrane transport

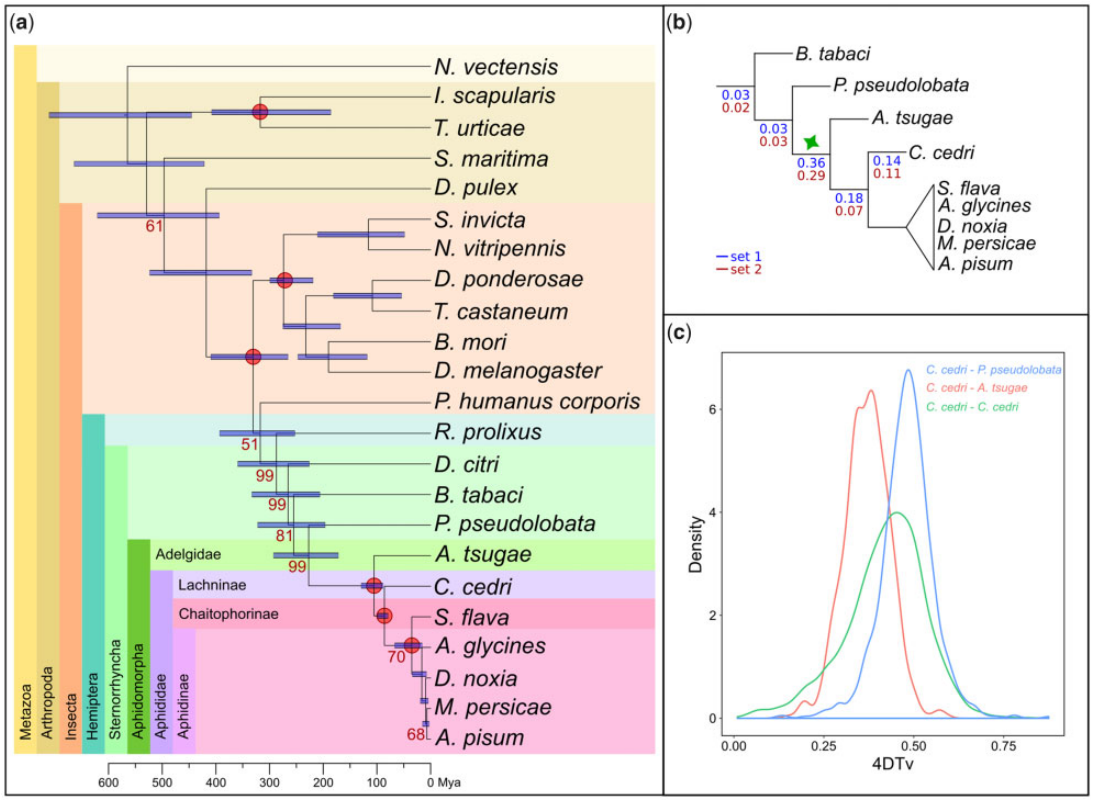

Fig. 3. Species tree, duplication ratio, and 4DTv of *Cinara cedri*. (*a*) Phylogenetic tree which included the two species with transcriptomes (*Adelges tsugae* and *Paratachardina pseudolobata*). Taxonomic groups are indicated by different colors. The bottom line represents the divergence time in Ma. Bars in the nodes indicate the uncertainty around mean age estimates based on 95% credibility intervals. All omitted bootstrap values are maximal (bootstrap 100%). Red dots mark the calibration points used to estimate the divergence times. (*b*) Zoom out of the Sternorrhyncha group. Duplication ratios are indicated in each branch for each set of gene trees: set 1 and set 2. The green star marks the position of the large-scale duplication event. (*c*) 4DTv of paralogous genes of *C. cedri* in the branch where the large-scale duplication event is marked in figure 3*b*. 4DTv of orthologous pairs between *C. cedri* and *Adelges tsugae*, and *P. pseudolobata* are shown with different colors.

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

with the analysis of the data set that included only complete genomes (see above, table 3). A general analysis of both results (tables 3 and 4) shows many GO terms enriched with key functions for aphid, phylloxerid, and adelgid biology. These insects base their diets strictly on phloem sap, which requires very specific adaptations (Douglas 2006). In this regard, genes duplicated ancestrally are enriched in carbohydrate metabolism and metabolite transporters, which may be related to the need for efficient exploitation of phloem sap, which is rich in sugar but poor in other essential nutrients (Douglas 2006). Essential amino acids in aphids are provided by microbial symbionts (Baumann 2005). In this context, ancestral duplications are also enriched in amino acid transporters, which may allow reallocation of these essential nutrients to enhance the amino acid supply (Hansen and Moran 2011). Other important adaptations of phloem-sap feeding are the adaptation to plant secondary metabolites. Glutathione *S*-transferases play an important role in the detoxification of many substances including allelochemicals from plants (Francis et al. 2005). Genes associated with glutathione *S*-transferases are also duplicated at the ancestral branch of Aphidomorpha (e.g., see supplementary fig. S6*a*, Supplementary Material online). Similarly, genes duplicated ancestrally are enriched in functions related to UDP-glycosyltransferases, which are a major class of drug-metabolizing enzymes and play an important role in the detoxification of a large number of xenobiotics (Bock 2016). In aphids, UDP-glycosyltransferase may confer tolerance to thiamethoxam (Pan et al. 2015). Other ancestral duplications of genes are involved in wax biosynthesis, which may be related to maintaining water balance, and preventing desiccation (Chung and Carroll 2015). Enrichment in fatty acyl reductases (e.g., see supplementary fig. S6*b*, Supplementary Material online) may be related to not only wax biosynthesis but also components of insect cuticular hydrocarbons and pheromones (Tupec et al. 2019). Moreover, ancestral duplications of genes involved in smell and sugar taste perception may have facilitated detection of suitable plant host or development of alarm pheromones (Zhang et al. 2017). This ancestral wave of duplication is also enriched in functions related to growth and molting. One such example is the genes associated with ecdysis triggering hormone receptor (supplementary fig. S6*c*, Supplementary Material online), which are crucial for the activation of the ecdysis sequence (Roller et al. 2010). Other functional classes enriched among ancestral paralogs are associated with DNA and RNA processing, and may be fundamental to the maintenance of the genomic and phenotypic plasticity observed in aphids (Mathers et al. 2017).

Use of a Chromosome-Level Assembly of *A. pisum*

Recently, a chromosome-level assembly for the *A. pisum* genome became available (Li et al. 2019). The presence of high numbers of paralogs complicates the process of genome assembly. Depending on the nature of the data and the assembly algorithms and parameters used, recently duplicated paralogs can be (partially) collapsed into a single sequence or, conversely, divergent alleles of the same loci can be separated into distinct sequences (Gabaldón and Alioto 2016). These issues had always been a concern when assessing the high levels of duplications in *A. pisum* and other aphid species (International Aphid Genomics Consortium 2010). To confirm our previous results with a more contiguous version of the assembly, we repeated the *A. pisum* phylome (*A. pisum*2, PhylomeID 707), this time using the annotation of the newly released, chromosome-level assembly (Li et al. 2019).

The results of the analysis of this *A. pisum*2 phylome are in agreement with the results of the *A. pisum*1 phylome (PhylomeID 706). The duplication ratio at the ancestral branch of all aphids is still high (0.69). Also, the number of specific gene duplication events (1,825) is similar to that in the phylome 706 (1,889), and the percentage of proteins that have an in-paralog is the same (29%). However, the availability of the AL4 assembly allowed us to analyze the paralogs in the context of chromosomes. Interestingly, a high percentage of the duplicated proteins are present in the assembled chromosomes: For the aphid-ancestral wave of duplication, 82% of duplicated proteins are present in the chromosomes, and for the species-specific *A. pisum* wave, 76%. If we analyze the number of pairs of paralogs (expansions will form more than one pair) present in the same chromosome, a higher percentage was observed in the species-specific duplications (chromosome X—47%, A1—33%, A2—42%, A3—34%) with respect to the aphid ancestral duplications (chromosome X—34%, A1—28%, A2—19%, A3—12%). In both cases chromosome X has the highest percentage, whereas A3 has the lowest. Moreover, from the total number of proteins duplicated in both waves, chromosome X has the highest percentage, followed by chromosome A1 (fig. 4*a*). These results are in agreement with recent studies (Li et al. 2019).

In order to analyze the distribution of paralogs along the chromosomes, we plotted all the pairs per chromosome and per wave of duplication (fig. 4*b* and *c*). As previously noticed (Li et al. 2019), the paralogs are distributed throughout all the chromosomes. However, some blocks of paralogous pairs can also be observed. An evident block is shared between chromosomes X and A2 in both waves of duplications (fig. 4*b* and *c*).

Finally, we searched for footprints of these large-scale duplication events in the relative gene order of paralogs (i.e., synteny conservation), which rendered no significant result. A comparison of *C. cedri* against itself using Symap did not reveal any conserved region (see supplementary fig. S7, Supplementary Material online). The analysis was repeated by comparing *C. cedri* with *A. pisum* and *Bemisia tabaci*. We observed that some conserved blocks between *C. cedri* and *A. pisum*. However, when we compare *C. cedri* with *Be. tabaci*, a more distant relative, the gene order conservation disappears (see supplementary fig. S8, Supplementary Material online). Similarly a comparison of the chromosome-level assembly of *A. pisum* to *Be. tabaci* revealed no apparent conserved synteny block.

We repeated the analysis using i-ADHore (Proost et al. 2012) and found few collinear segments (9) (see Materials and Methods). We repeated both analyses for the chromosome-level assembly of *A. pisum* with similar results. We also compared gene order between *C. cedri* and

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

Table 4. List of GO Terms Enriched in the Duplicated Protein Families at the Base of the Aphidomorpha Group.

Term Category	Term	Term Level	Adj. *P*-Value	Term Name
molecular_function	GO:0000064	1	2.38E-04	L-ornithine transmembrane transporter activity
molecular_function	GO:0003730	1	2.38E-04	mRNA 3′-UTR binding
molecular_function	GO:0004032	1	9.45E-06	Alditol:NADP+ 1-oxidoreductase activity
molecular_function	GO:0004035	1	9.08E-04	Alkaline phosphatase activity
molecular_function	GO:0004185	1	3.29E-04	Serine-type carboxypeptidase activity
molecular_function	GO:0004197	1	5.74E-16	Cysteine-type endopeptidase activity
molecular_function	GO:0004497	1	8.22E-08	Monooxygenase activity
molecular_function	GO:0004553	1	1.02E-15	Hydrolase activity, hydrolyzing *O*-glycosyl compounds
molecular_function	GO:0004555	1	5.24E-08	Alpha, alpha-trehalase activity
molecular_function	GO:0005215	1	2.31E-16	Transporter activity
molecular_function	GO:0005254	1	3.45E-06	Chloride channel activity
molecular_function	GO:0005351	1	2.46E-06	Sugar:proton symporter activity
molecular_function	GO:0005355	1	5.52E-07	Glucose transmembrane transporter activity
molecular_function	GO:0005506	1	8.41E-12	Iron ion binding
molecular_function	GO:0005542	1	4.33E-07	Folic acid binding
molecular_function	GO:0008194	1	2.43E-05	UDP-glycosyltransferase activity
molecular_function	GO:0008234	1	9.57E-12	Cysteine-type peptidase activity
molecular_function	GO:0008237	1	1.46E-06	Metallopeptidase activity
molecular_function	GO:0008260	1	6.12E-04	3-Oxoacid CoA-transferase activity
molecular_function	GO:0008422	1	1.50E-04	Beta-glucosidase activity
molecular_function	GO:0008518	1	4.33E-07	Reduced folate carrier activity
molecular_function	GO:0008521	1	9.33E-25	Acetyl-CoA transporter activity
molecular_function	GO:0015020	1	2.42E-15	Glucuronosyltransferase activity
molecular_function	GO:0015116	1	1.50E-04	Sulfate transmembrane transporter activity
molecular_function	GO:0015171	1	3.58E-05	Amino acid transmembrane transporter activity
molecular_function	GO:0015174	1	3.81E-05	Basic amino acid transmembrane transporter activity
molecular_function	GO:0015181	1	2.38E-04	Arginine transmembrane transporter activity
molecular_function	GO:0015189	1	2.38E-04	L-lysine transmembrane transporter activity
molecular_function	GO:0015295	1	1.61E-09	Solute:proton symporter activity
molecular_function	GO:0015297	1	2.33E-06	Antiporter activity
molecular_function	GO:0015326	1	3.81E-05	Basic amino acid transmembrane transporter activity
molecular_function	GO:0015528	1	9.87E-09	Lactose:proton symporter activity
molecular_function	GO:0016298	1	6.12E-04	Lipase activity
molecular_function	GO:0016491	1	6.42E-04	Oxidoreductase activity
molecular_function	GO:0016620	1	2.71E-05	Oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor
molecular_function	GO:0016705	1	1.95E-14	Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
molecular_function	GO:0016717	1	1.21E-06	Oxidoreductase activity, acting on paired donors, with oxidation of a pair of donors resulting in the reduction of molecular oxygen to two molecules of water
molecular_function	GO:0016747	1	6.42E-04	Transferase activity, transferring acyl groups other than amino-acyl groups
molecular_function	GO:0016758	1	5.44E-27	Transferase activity, transferring hexosyl groups
molecular_function	GO:0016787	1	5.09E-05	Hydrolase activity
molecular_function	GO:0017110	1	7.12E-06	Nucleoside-diphosphatase activity
molecular_function	GO:0019799	1	2.63E-07	Tubulin *N*-acetyltransferase activity
molecular_function	GO:0020037	1	8.50E-19	Heme binding
molecular_function	GO:0022857	1	1.30E-38	Transmembrane transporter activity
molecular_function	GO:0022891	1	3.70E-10	Substrate-specific transmembrane transporter activity
molecular_function	GO:0042626	1	5.04E-07	ATPase activity, coupled to transmembrane movement of substances
molecular_function	GO:0043169	1	4.04E-10	Cation binding
molecular_function	GO:0050660	1	5.48E-04	Flavin adenine dinucleotide binding
molecular_function	GO:0052689	1	8.97E-05	Carboxylic ester hydrolase activity
molecular_function	GO:0080019	1	8.02E-07	Fatty-acyl-CoA reductase (alcohol-forming) activity
molecular_function	GO:0090482	1	4.33E-07	Vitamin transmembrane transporter activity
molecular_function	GO:0102336	1	2.97E-04	3-Oxo-arachidoyl-CoA synthase activity
molecular_function	GO:0102337	1	2.97E-04	3-Oxo-cerotoyl-CoA synthase activity
molecular_function	GO:0102338	1	2.97E-04	3-Oxo-lignoceronyl-CoA synthase activity
cellular_component	GO:0005887	1	8.49E-11	Integral component of plasma membrane
cellular_component	GO:0008537	1	1.27E-06	Proteasome activator complex
cellular_component	GO:0016021	1	2.85E-43	Integral component of membrane
cellular_component	GO:0043186	1	1.97E-04	P granule
cellular_component	GO:0043231	1	3.81E-05	Intracellular membrane-bounded organelle

(continued)

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

Table 4. Continued

Term Category	Term	Term Level	Adj. *P*-Value	Term Name
biological_process	GO:0001510	1	8.02E-07	RNA methylation
biological_process	GO:0005975	1	4.78E-13	Carbohydrate metabolic process
biological_process	GO:0005991	1	5.24E-08	Trehalose metabolic process
biological_process	GO:0006508	1	1.84E-22	Proteolysis
biological_process	GO:0006629	1	6.45E-11	Lipid metabolic process
biological_process	GO:0006633	1	8.18E-04	Fatty acid biosynthetic process
biological_process	GO:0006857	1	9.96E-05	Oligopeptide transport
biological_process	GO:0006865	1	3.81E-05	Amino acid transport
biological_process	GO:0007283	1	4.73E-05	Spermatogenesis
biological_process	GO:0008152	1	2.80E-08	Metabolic process
biological_process	GO:0009166	1	3.12E-07	Nucleotide catabolic process
biological_process	GO:0009452	1	2.51E-07	7-Methylguanosine RNA capping
biological_process	GO:0010025	1	1.73E-07	Wax biosynthetic process
biological_process	GO:0016973	1	9.45E-06	Poly(A)+ mRNA export from nucleus
biological_process	GO:0035336	1	1.73E-07	Long-chain fatty-acyl-CoA metabolic process
biological_process	GO:0035428	1	2.63E-07	Hexose transmembrane transport
biological_process	GO:0042759	1	6.18E-05	Long-chain fatty acid biosynthetic process
biological_process	GO:0046323	1	2.63E-07	Glucose import
biological_process	GO:0050790	1	3.19E-11	Regulation of catalytic activity
biological_process	GO:0051180	1	4.33E-07	Vitamin transport
biological_process	GO:0051603	1	1.40E-08	Proteolysis involved in cellular protein catabolic process
biological_process	GO:0052696	1	7.61E-06	Flavonoid glucuronidation
biological_process	GO:0055085	1	4.51E-56	Transmembrane transport
biological_process	GO:0055114	1	1.35E-15	Oxidation-reduction process
biological_process	GO:0070507	1	2.09E-04	Regulation of microtubule cytoskeleton organization
biological_process	GO:0071929	1	2.63E-07	Alpha-tubulin acetylation
biological_process	GO:1901657	1	9.80E-06	Glycosyl compound metabolic process
biological_process	GO:1903352	1	2.38E-04	L-ornithine transmembrane transport

M. persicae and between *C. cedri* and the more distantly related *Be. tabaci*. The comparison between the first two showed a moderate gene order conservation, though the fragmentation of both genomes makes it difficult to assess whether the conservation is real or just an artifact. i-ADHore detected only 43 conserved segments between the two species, much less than the 509 segments found between more closely related *A. pisum* and *M. persicae*, indicating a quick degradation of synteny. Comparison between *C. cedri* and the less related *Be. tabaci* shows that all signs of gene order conservation have been lost between these two genomes, which is confirmed by i-ADHore. This result was also observed between the chromosome-level *A. pisum* and *Be. tabaci*, which indicates that the loss of gene order conservation is unlikely to be due to the fragmentation of the two genomes. The patterns observed in the 4DTv analysis (see above) indicate that the duplication event likely occurred soon after the divergence between aphids and *P. pseudolobata*. Therefore, if in fact there was a WGD, it is likely that further rearrangements and additional duplications have blurred the syntenic conservation between duplicated genes. These results may be influenced by the fragmentation of the genomes, as only the genome of *A. pisum* is at chromosome level. Yet, from our observations we do not see many differences when comparing *C. cedri* to *A. pisum* and *M. persicae*, leading us to believe that although we may be missing syntenic blocks and we could not provide an exact number of such blocks, the observed trend would remain similar if we had chromosome-level assemblies.

Altogether, our results point to the presence of one major wave of ancestral duplications in the aphid lineage, predating the diversification of Aphidomorpha. This ancestral wave of duplications occurred in addition to other lineage-specific duplications, and to many recent species-specific duplications, highlighting a high genomic plasticity in aphids.

Conclusions

Recently, many WGDs have been described in insects (Li et al. 2018). Moreover, several independent studies have shown a burst of gene duplication in different species of aphids (Huerta-Cepas et al. 2010; Mathers et al. 2017; Li et al. 2018), but the origin of these duplications has been thus far unclear. Here we present the genome sequence of an early-diverging aphid (*C. cedri*) and its comparison with 20 complete animal genomes, including five sequenced aphid genomes, and the transcriptomes of two other phylogenetically important species. Taken together, our phylogenomic results provide compelling evidence for the existence of a large-scale gene duplication event predating the divergence of aphids, adelgids, and presumably phylloxerids (i.e., of Aphidomorpha). Genes duplicated at this large-scale event are enriched in functions that are relevant to aphids, adelgids, and phylloxerids, which share traits such as a phloem-sap-based diet or tight association with endosymbionts.

The availability of the genome sequence of *C. cedri*, belonging to the subfamily Lachninae, an early-branching lineage within Aphididae, which feeds on gymnosperm, helps situate

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

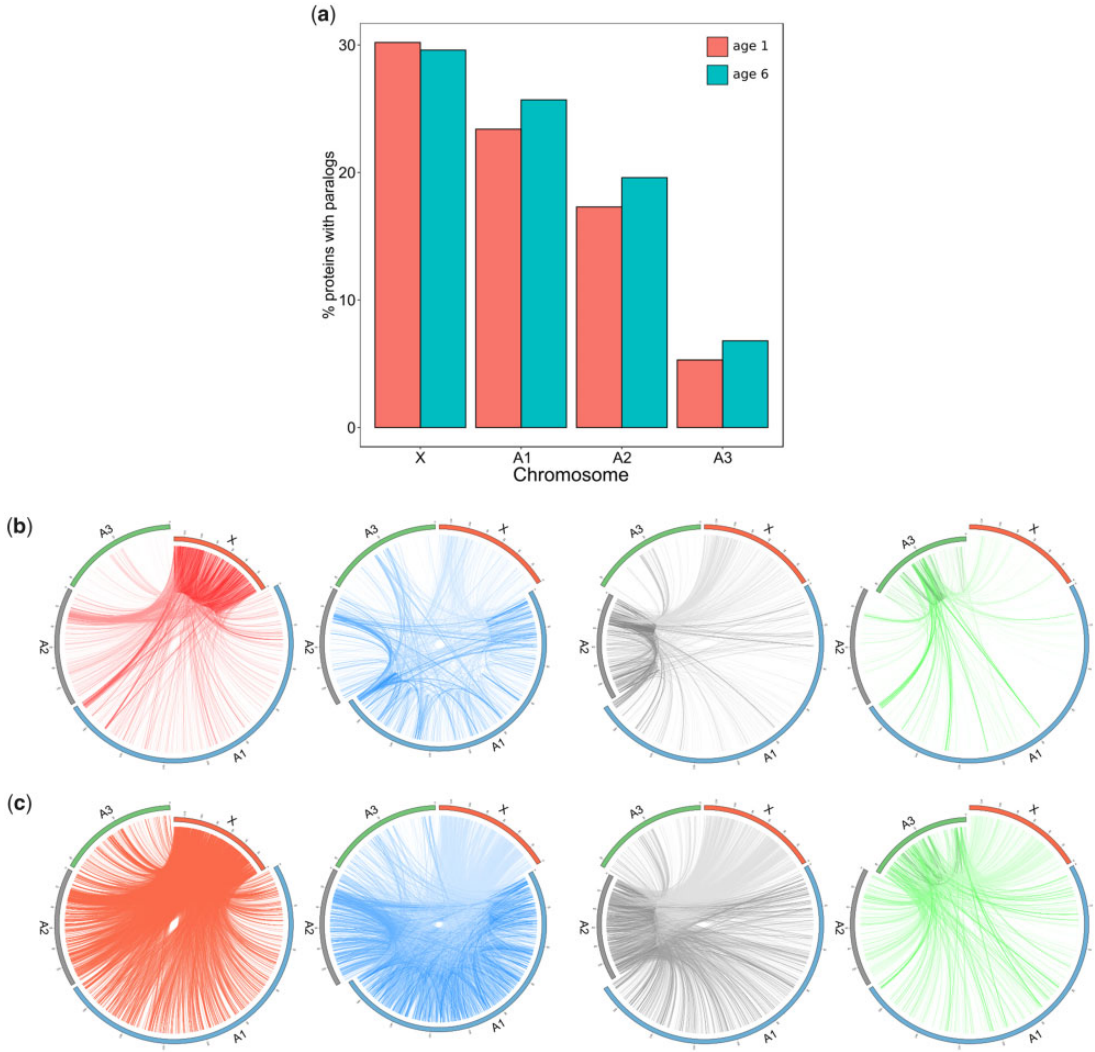

FIG. 4. *Acyrthosiphon pisum* paralogs mapped onto the chromosomes. (*a*) Percentage of proteins with paralogs at the two waves of duplications (age1: *A. pisum* species-specific, age6: ancestral to all aphids) per each chromosome. (*b*) Links between pairs of paralogs at age1. (*c*) Links between pairs of paralogs at age6. The chromosomes are shown in different colors: X—red, A1—blue, A2—gray, A3—green. Each chromosome was plotted independently for each age.

other major genomic events, such as the loss of selenoproteins and the streamlining of the immune repertoire closer to the base of the aphid lineage. Additional genome sequences, particularly of species of the Phylloxeridae family would help to confirm these findings and provide additional evolutionary insights. Our results underscore the use of phylogenomic approaches to study ancient duplication events (Marcet-Houben and Gabaldón 2015; Julca et al. 2018), and we do not discard the possibility of at least one ancestral WGD in the evolutionary history of aphids, even in the absence of syntenic conservation. Although synteny has been the traditional approach to uncover ancient WGDs (Ohno 1970; Wolfe 2015), it relies on a signal that is blurred by subsequent genomic rearrangements and may not apply equally well to different lineages. It has been suggested that the holocentric nature of aphid chromosomes may promote rapid reshuffling of gene order (Blackman 1980; Mandrioli et al. 2016; Li et al. 2019), and this is consistent with our findings. In our case, the lack of syntenic conservation between *Be. tabaci* and the aphids indicates that—at the level of resolution of our methods and the current level of completion of the compared genomes—a large number of rearrangements have occurred. This, coupled with large amounts of lineage-specific gene expansions, has likely degraded the initial syntenic relationships originated at the putative ancestral wave of gene duplications.

Materials and Methods

Heterozygosity Analysis and DNA Extraction

A *C. cedri* population collected from a single cedar tree (*Cedrus libani*) in the spring of 2011 in Lliria (Valencia, Spain, 39°38′47,0826″N, 0°37′ 51,3006″O) was introduced and maintained on a cedar tree at the facilities of the Institut Cavanilles de Biodiversitat I Biologia Evolutiva at the University of Valencia (Spain). For the DNA extraction, all the individuals were collected during March 2012 from a colony from the tree, and because they are parthenogenetic, it is

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

likely that all come from the same female. To analyze heterozygosity in the population, the cytochrome oxidase I (COI) mitochondrial gene of 45 individual insects, collected between July 2011 and March 2012, was amplified by polymerase chain reaction (PCR) using the primers LCO1490 and HCO21980 (Folmer et al. 1994) and sequenced. All analyzed sequences were identical. Approximately 50–60 apteral adult females were gently homogenized and used for DNA extraction, using the "Purification of DNA from insects using the DNeasy Blood & Tissue Kit" (Quiagen) according to the manufacturer's instructions.

Flow Cytometric Genome Size Estimates

Flow cytometric genome size was estimated for nuclei isolated from whole insect tissue stained with propidium iodide and scored by flow cytometry for relative red fluorescence of the 2C peaks of the sample and a co-prepared standard (see Johnston et al. 2019).

In brief, the anterior one-third of each sample was placed into 1 ml of Galbraith buffer in a 2-ml Kontes Dounce tissue homogenizer along with the head of a lab-reared *Drosophila virlis* standard (1C = 328 Mb). Nuclei were released from the sample and standard by 15 strokes of the "A" Dounce pestle. The resultant solution was filtered through a 40-µm nylon mesh, stained with 25 mg/ml propidium iodide and held for 2 h in the dark and cold to allow comparable levels of dye saturation in the sample and standard. The average channel number of the 2C peaks of the sample and standard were scored and the 1C (gametic) DNA calculated as the ratio of the 2C peaks of the sample and standard times 328 Mb. At least 500 nuclei were scored under each peak with a maximum coefficient of variation (CV) of 3.0 for each. The mean and standard error were based on estimates for individuals and standards in separate co-preparations.

Genome Sequencing and Assembly

The short-insert paired-end libraries were prepared with the NO-PCR protocol. TruSeqDNA Sample Preparation Kit v2 (Illumina Inc.) and the KAPA Library Preparation Kit (Kapa Biosystems) were used. In short, 2.0 µm of sheared genomic DNA was end-repaired, adenylated, and ligated to Illumina-specific indexed paired-end adaptors. The DNA was size selected with AMPure XP beads (Agencourt, Beckman Coulter) in order to reach the fragment size of 220–550 bp. The final libraries were quantified by Library Quantification Kit (Kapa Biosystems).

The library was sequenced using the TruSeq SBS Kit v3-HS (Illumina Inc.) in paired-end mode, 2 × 101 bp, in one sequencing lane of a HiSeq2000 (Illumina Inc.) according to standard Illumina operation procedures with a yield of >30 Gb of raw data. Primary data analysis, image analysis, base calling, and quality scoring of the run were processed using the manufacturer's software Real Time Analysis (1.13.48) and followed by the generation of FASTQ sequence files by CASAVA.

In addition, two Mate Pair (MP) libraries with insert sizes of 3,000 (MP3000) and 5,000 bp (MP5000) were constructed according to a modified Illumina protocol. In brief, after genomic DNA fragmentation, circularization of the DNA was performed in the presence of a biotinylated 454 double-stranded linker. Thereafter, the standard Illumina mate-pair preparation method was followed. The libraries were sequenced on the Illumina HiSeq2000 platform in paired-end mode, which outputs 101-bp reads (2 × 101 bp). The yield was at least 11 Gb for both MP libraries. Postprocessing of sequence reads involved trimming of the linker sequence (TCGTATAACTTCGTATAATGTATGCTATACGAAGTTAT TACG and reverse complement) using cutadapt v2.5 (Martin 2011) with -e 0.05 and -O 10 options. Only pairs for which at least one mate was trimmed (i.e., contained the linker and was thus a true mate pair (MP) and not paired-end (PE) contamination) were kept for scaffolding. Then, we used gem-mapper (Marco-Sola et al. 2012) to detect reads matching contaminants. Contaminated reads were filtered before the assembly. The genome size and complexity were estimated using Jellyfish v1.1 (Marçais and Kingsford 2011) and GenomeScope v1.0 (Vurture et al. 2017), resulting in estimates of 508.6 and 399.76 Mb, respectively. The latter estimate was used to guide our assembly strategy.

The genome was assembled as follows. SGA preqc analysis (Simpson 2014) was used to determine optimal K-mer length ($k = 65$) for de Bruijn graph construction. Bracketing around this optimum, multiple assemblies with shorter and longer K-mers were made using ABySS v1.5.2 (Simpson et al. 2009) merged with ASM, an OLC-like assembly-merging software to obtain contigs (Frias and Ribeca, 2016; Cruz et al. 2016). The merging parameters were: –anchor 125 –anchor-spacing 10 –min-anchor 25 –coverage 2 –divergence 0.03 –anchorsxchunk 50000000 –repeat-resolution-depth 0 –path-expansion-depth 0 –consensus-type majority. The sequencing libraries were mapped with gem-mapper (Marco-Sola et al. 2012), and these mappings were used for scaffolding the merged contigs with ABySS using parameters -n 4 -s 200 -N 8 -S 130-2000 -k 67 -l 36 -q 10. This intermediate assembly was refined by decontamination, consistency check (Cruz et al. 2016), and then discarding scaffolds shorter than 4 kb already contained in longer scaffolds (i.e., unique mappings with 0% mismatches detected with gem-mapper). This refined assembly was re-scaffolded with SSPACEv3.0 (Boetzer et al. 2011; Cruz et al. 2016), and gaps closed with GapFiller (Boetzer and Pirovano 2012). Afterwards, the PE library (PE400) was mapped against the target assembly (gap-filled and decontaminated) with BWA mem v0.7.7 (Li and Durbin 2009) and then performed variant calling with samtools v0.1.19 mpileup (Li et al. 2009). Only single-nucleotide substitutions or indels having at least ten reads supporting the alternative allele were used to produce an alternative reference. The alternative FASTA sequence was obtained using GATK v3.5 *FastaAlternateReferenceMaker* (McKenna et al. 2010). Finally, the assembly was named internally cinced3, and the gene completeness was evaluated with CEGMA v2.4 (Parra et al. 2007), which searches for 248 core-eukaryotic genes, and BUSCO v3.0.2 (Simão et al. 2015) using the insecta_odb9 that includes 42 species and 1,658 genes.

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

RNA Extraction and Transcriptome Sequencing

RNA samples were prepared from 400 adults of *C. cedri* females from the clonal population mentioned above. Samples were obtained from aphid heads (absence of endosymbionts) and dissected bacteriocytes (presence of the two endosymbionts, *B. aphidicola* and *Se. symbiotica*). Total RNA extraction was performed using the "TRI Reagent Solution" Kit (Ambion), following the manufacturer's instructions. In addition, two SOLID libraries were prepared from the same tissues.

Genome Annotation

A combination of the Program to Assemble Spliced Alignments (PASA v2.0.2) and Evidence Modeler (EVM v1.1.1) (Haas et al. 2008) was used to obtain consensus coding sequence models using three main sources of evidence: aligned transcripts, aligned proteins, and gene predictions. Finally, noncoding RNAs were annotated employing CMsearch (Cui et al. 2016), tRNAscan-SE (Lowe and Eddy 1997), and lncRNAs were obtained from the PASA-assemblies without protein-coding gene annotations that were longer than 200 bp. The Piper-R NF pipeline (https://github.com/cbcrg/piper-nf) was used to detect lncRNA conservation between *C. cedri* and 20 other metazoans (supplementary table S3, Supplementary Material online). Then, a heatmap was plotted including all the lncRNAs that are present in at least one of the other species using the R package gplots (Warnes et al. 2016).

Identification of Genes of the Immune System of *C. cedri*

A database of the genes involved in the immune system of well-studied insects was generated by updating and expanding the database from Insect Innate Immunity Database (Brucker et al. 2012). The source references for the jewel wasp (*Nasonia vitripennis* [Werren et al. 2010]), the honeybee (*Apis mellifera* [Evans et al. 2006]), the fruit fly (*Drosophila melanogaster* [De Gregorio et al. 2001]), the African malaria mosquito (*Anopheles gambiae* [Christophides et al. 2002]), and the pea aphid (*A. pisum* [Gerardo et al. 2010]) were revised, and the genes they described were retrieved. To increase completeness, the immune system genes from the red flour beetle (*Tribolium castaneum* [Zou et al. 2007]), the diamondback moth (*Plutella xylostella* [Xia et al. 2015]), the tobacco hornworm (*Manduca sexta* [Cao et al. 2015]), the head louse (*Pediculus humanus* [Kang et al. 2015]), the Florida carpenter ant (*Camponotus floridanus* [Gupta et al. 2015]), the Asian citrus psyllid (*Diaphorina citri* [Arp et al. 2016]), and the silkworm (*Bombyx mori* [Tanaka et al. 2008]) were also added. EggNOG-mapper (Huerta-Cepas et al. 2017) was used to identify orthologs among the selected species using the Arthropoda (artNOG) data set followed by manual curation where discrepancies were observed.

Aphid Phylomes Reconstruction

The phylomes of *C. cedri*, *A. pisum*, *M. persicae*, *D. noxia*, *Ap. glycines*, and *S. flava* were reconstructed using the PhylomeDB pipeline (Huerta-Cepas et al. 2011). For *A. pisum*, two phylomes were reconstructed using the two proteomes available (Acyr 2.0 [International Aphid Genomics Consortium 2010] and AL4 [Li et al. 2019]). In brief, for each protein-coding gene in each aphid genome we searched for homologs (Smith–Waterman BLAST search, *e*-value cutoff $<$1e-05, minimum contiguous overlap over the query sequence cutoff 50%) in a protein database containing the proteomes of 21 species for *C. cedri* and a subset of 14 species for the other aphids (supplementary table S2, Supplementary Material online). The most similar 150 homologs were aligned using three different programs (MUSCLE [Edgar 2004], MAFFT [Katoh et al. 2005], and KALIGN [Lassmann and Sonnhammer 2005]) in both forward and reverse orientations. These six alignments were combined using M-COFFEE (Wallace et al. 2006) and trimmed with trimAl v.1.3 (Capella-Gutiérrez et al. 2009) using a consistency cutoff of 0.16667 and a gap threshold of 0.1. Phylogenetic trees were built using maximum likelihood approach as implemented in PhyML v3.0 (Guindon et al. 2010) using the best fitting model among seven different ones (JTT, LG, WAG, Blosum62, MtREV, VT, and Dayhoff). The two models best fitting the data were determined based on likelihoods of an initial neighbor joining tree topology and using the AIC criterion. We used four rate categories and inferred fraction of invariant positions and rate parameters from the data. All alignments and trees are available for browsing or download at PhylomeDB with the PhylomeID: *C. cedri*—701, *S. flava*—702, *Ap. glycines*—703, *D. noxia*—704, *M. persicae*—705, *A. pisum* 1—706, and *A. pisum*2—707 (Huerta-Cepas et al. 2014).

Prediction of Gene Duplications, and Orthology and Paralogy Relationships

Orthology and paralogy relationships were predicted based on phylogenetic evidence from each aphid phylome. We used ETE v3.0 (Huerta-Cepas et al. 2016) to infer duplication and speciation relationships using the species overlap method (Gabaldón 2008) and a topology-based phylostratigraphic method to date duplication events (Huerta-Cepas and Gabaldón 2011). In brief the species-overlap algorithm identifies internal nodes as duplications if the two daughter clades show any overlap in species, and the phylostratigraphic method assigns a relative age to that duplication as the last common ancestor of all the taxa contained in the two daughter clades. Species-specific duplications (expansions) were computed as duplication that map only to the seed species of each phylome (*C. cedri*, *A. pisum*, *M. persicae*, *D. noxia*, and *Ap. glycines*). In order to reduce redundancy, expansions that overlap in more than 50% of their sequences were fused together using a UPGMA clustering. Duplication ratios were calculated by dividing the number of duplications mapped to a given node in the species tree by all the gene trees that contain that node. In all the cases, duplication frequencies that include expansions larger than five in each phylome were excluded. Due to ancestral expansions in other species that can affect the duplication ratio, an additional filter was applied by removing all the gene trees that contained aphid duplications with more than five sequences in any of the aphid species included (*C. cedri*, *A. pisum*, *M. persicae*,

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

D. noxia, *Ap. glycines*, and *S. flava*). Duplication ratios were calculated again using this new subset of gene trees. All orthology and paralogy relationships are available through PhylomeDB (Huerta-Cepas et al. 2014).

Incorporation of Transcriptomic Data in the Phylome

In order to increase the taxonomic sampling in the Sternorrhyncha group we decided to include two species, which have their transcriptome available: *Ad. tsugae* (accession number: PRJNA242203) and *P. pseudolobata* (Christodoulides et al. 2017). *Adelges tsugae* belongs to the family Adelgidae, a clade inside the Aphidomorpha. *Paratachardina pseudolobata* is a scale insect that belongs to the superfamily Coccoidea. The transcriptome of *P. pseudolobata* was obtained from the whole body of female samples, and the assembly is mostly complete according to BUSCO (89% of highly conserved arthropod sequences were present as single-copy or duplicated transcripts in the assembly) (Christodoulides et al. 2017). The transcriptome of *Ad. tsugae* was also obtained from the whole body of female samples. Although these data are not published, the gene completeness was evaluated with BUSCO 3.1.0 (Simão et al. 2015) and we found that 79% of the conserved genes of the insecta_odb9 data set were present as single-copy or duplicated transcript in this assembly.

The Transcriptome Shotgun Assembly (TSA) file of both species was downloaded, and the prediction of proteins was obtained by selecting the longest open reading frame for each transcript (>100 amino acids). The incorporation of the transcriptomic data was done using the following pipeline. First, a BLASTP was performed from the seed protein against a database that contained the two transcriptomes. Then, the results were filtered based on three thresholds: *e*-value < 1e-05. Overlap between query and hit had to be at least 0.3, and sequence identity > 40.0%. Proteins that passed these filters were incorporated into the raw alignment of the phylome using MAFFT v7.222 (–add and –reorder options) (Katoh et al. 2005). Then, gene trees were reconstructed using this new alignment following the same procedure as described above. Finally, these gene trees were filtered in order to remove unreliably placed transcriptome sequences (set1) and filtered again to keep only trees that contained both species, *Ad. tsugae* and *P. pseudolobata* (set2).

GO Term Enrichment

GO terms were assigned to the five aphid proteomes using Interproscan v5.34-73.0 (Jones et al. 2014) and the annotation of orthologs from the PhylomeDB database (Huerta-Cepas et al. 2014). GO term enrichment analysis was performed in each phylome using an in-house adaptation of FatiGO (Al-Shahrour et al. 2007) by comparing annotations of the duplicated proteins specific to the aphid (and at the branches subtending the wave of duplications) against all the other proteins encoded in the aphid genome.

Species Tree Reconstruction

All the gene trees that contained one-to-one orthologs per each of the 21 species included in the *C. cedri* phylome were kept (57 gene trees). The alignments used for the reconstruction of these gene trees were concatenated, and a species tree was reconstructed using the amino acid substitution model LG implemented in RAxML v8.1.17 (Stamatakis 2014) and 1,000 bootstrap replicates. In addition, a species tree including the two transcriptome data (*Ad. tsugae* and *P. pseudolobata*) was also obtained by concatenating, from the previous set, the gene trees that included at least one of the new species (52 gene trees).

Transversion Rate at 4-Fold Degenerate Sites

The distribution of the 4DTv was used to estimate the relative age of speciation events and duplications. Per each phylome, the 4DTv of pairs of paralogs at the branches subtending the high duplication rates (branches marked as A and B, fig. 2*b*) were calculated. Also, for the *C. cedri* phylome we included the 4DTv of orthologs of *C. cedri* with *S. flava*, and *Be. tabaci*. For the other five phylomes (*S. flava*, *Ap. glycines*, *D. noxia*, *M. persicae*, *A. pisum*1) the 4DTv of orthologs for each aphid with *Be. tabaci*, and *C. cedri* were also included. After the two transcriptome data sets were included, the 4DTv was calculated for pairs of paralogous genes of *C. cedri* at the base of the Aphidomorpha group, and pairs of orthologous genes between *C. cedri* and *P. pseudolobata*, and *Ad. tsugae*.

Divergence Times

To place a time scale on the maximum likelihood phylogeny, we used the Bayesian-relaxed molecular clock approach as implemented in PhyloBayes v4.1c (Lartillot et al. 2013). An uncorrelated relaxed clock model was applied, and six constraints specified to the most recent common ancestor, including fossil specimens and secondary calibrations: Aphidomorpha (135 Ma, fossil evidence [Heie 1987; Havill et al. 2007]), Aphididae (80–100 Ma, previous molecular date estimates [Von Dohlen 2000] and fossil remains [Heie 1987, 1999; Heie and Wegierek 2011]), Aphidinae (70 Ma, fossil record [Heie 1987; Hong 2002]), Hexapoda (425 Ma, fossil evidence [Grimaldi and Engel 2005]), Holometabola (300 Ma, fossil evidence [Labandeira and Phillips 1996]), and Acari (410 Ma, fossil record [Hirst 1923; Dubinin 1962; Dunlop and Selden 2009]). These calibration constraints were used with soft bounds (Yang and Rannala 2006) under a birth–death prior, and a prior on the root of the tree (a mean of 560 Ma and a standard deviation of 100 Ma) (Liu et al. 2014). Two independent MCMC chains were run for 100,000 cycles, sampling posterior rates and dates every ten cycles. The initial 25% were discarded as burn-in. Posterior estimates of divergence dates and associated 95% credibility intervals were then computed from the remaining samples of each chain.

Physical Mapping of *A. pisum* Paralogs into the Chromosomes

Paralogs obtained from the *A. pisum*2 phylome (PhylomeID 707) were mapped into the four chromosomes (X, A1, A2, and A3). Links between the pairs of paralogs were visualized using Circos v0.69-6 (Krzywinski et al. 2009).

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

Synteny

SyMap v4.2 (Soderlund et al. 2011) was used to search for gene order conservation within the *C. cedri* genome, between *C. cedri* and its ancestor *Be. tabaci*, between *C. cedri* and *A. pisum* and between *A. pisum* and *Be. tabaci*. Only scaffolds of 0.5 Mb or longer were used for the comparison. Default parameters were used.

i-ADHore (Proost et al. 2012) was also used to search for collinear segments within genomes and between different genomes. It was run for the individual genomes of *C. cedri* and the chromosome-level assembly of *A. pisum*. Then it was run between *C. cedri* and *M. persicae* and *C. cedri* and *Be. tabaci*, and finally, repeated between *A. pisum* and the same three species. i-ADHore was run using gg2 as the alignment method, with a tandem_gap of 75, gap_size of 30, cluster_gap of 35, q_value = 0.75, prob_cutoff = 0.01, and anchor_points = 3.

Data Availability

The assembly and annotation are also hosted at http://denovo.cnag.cat/ccedri where a jbrowse genome browser with most of the data can also be accessed.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

This research was funded by European Regional Development Fund (ERDF) and Ministerio de Economía y Competitividad (Spain) (Grant Nos. PGC2018-099344-B-100 and BFU2015-67107). T.G. group also acknowledges support from the Catalan Research Agency (AGAUR) SGR857, and grants from the European Union's Horizon 2020 research and innovation program under the grant agreements ERC-2016-724173 and MSC-747607. T.G. also receives support from an INB (Grant No. PT17/0009/0023—ISCIII-SGEFI/ERDF). The authors want to thank Sophia Derdak for her help in the genome polishing step.

References

Al-Shahrour F, Minguez P, Tárraga J, Medina I, Alloza E, Montaner D, Dopazo J. 2007. FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic Acids Res.* 35(Web Server issue):W91–W96.

Arp AP, Hunter WB, Pelz-Stelinski KS. 2016. Annotation of the Asian citrus psyllid genome reveals a reduced innate immune system. *Front Physiol.* 7:570.

Baumann P. 2005. Biology bacteriocyte-associated endosymbionts of plant sap-sucking insects. *Annu Rev Microbiol.* 59(1):155–189.

Becker-Migdisova EE, Aizenberg EE. 1962. Infraotryad Aphidomorpha. [Infraorder Aphidomorpha]. 194–199. In: Rohdendorf BB (ed) Osnovy palontologii. Chlenistonogie. Trakheïnye i Khelicerovye. [Fundamentals of Palaeontology. Arthropoda. Tracheata and Chelicerata.] 9. Izdatel'stvo Akademii Nauk SSSR, Moscow. [Published in English as: Becker-Migdisova EE, Aizenberg EE (1991). Infraorder Aphidomorpha. In: Rohdendorf BB (eds) Fundamentals of Paleontology. Arthropoda. Tracheata and Chelicerata 9. 218–289. General Editor English Translation. Davis DR Smithsonian Institution Libraries and The National Science Foundation, Washington D.C., pp 1–894].

Blackman RL. 1980. Chromosome numbers in the Aphididae and their taxonomic significance. *Syst Entomol.* 5(1):7–25.

Blackman RL, Eastop VF. 2000. Aphids of the World's Crops - an identification and information guide. England: Wiley & Sons. 466 p.

Bock KW. 2016. The UDP-glycosyltransferase (UGT) superfamily expressed in humans, insects and plants: animal-plant arms-race and co-evolution. *Biochem Pharmacol.* 99:11–17.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27(4):578–579.

Boetzer M, Pirovano W. 2012. Toward almost closed genomes with GapFiller. *Genome Biol.* 13(6):R56.

Brucker RM, Funkhouser LJ, Setia S, Pauly R, Bordenstein SR. 2012. Insect Innate Immunity Database (IIID): an annotation tool for identifying immune genes in insect genomes. *PLoS One* 7(9):e45125.

Cao X, He Y, Hu Y, Wang Y, Chen Y-R, Bryant B, Clem RJ, Schwartz LM, Blissard G, Jiang H. 2015. The immune signaling pathways of *Manduca sexta*. *Insect Biochem Mol Biol.* 62:64–74.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972–1973.

Chen W, Hasegawa DK, Kaur N, Kliot A, Pinheiro PV, Luan J, Stensmyr MC, Zheng Y, Liu W, Sun H. 2016. The draft genome of whitefly *Bemisia tabaci* MEAM1, a global crop pest, provides novel insights into virus transmission, host adaptation, and insecticide resistance. *BMC Biol.* 14(1):110.

Christodoulides N, Van Dam AR, Peterson DA, Frandsen RJN, Mortensen UH, Petersen B, Rasmussen S, Normark BB, Hardy NB. 2017. Gene expression plasticity across hosts of an invasive scale insect species. *PLoS One* 12(5):e0176956.

Christoffels A, Koh EGL, Chia J-M, Brenner S, Aparicio S, Venkatesh B. 2004. Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol Biol Evol.* 21(6):1146–1151.

Christophides GK, Zdobnov E, Barillas-Mury C, Birney E, Blandin S, Blass C, Brey PT, Collins FH, Danielli A, Dimopoulos G, et al. 2002. Immunity-related genes and gene families in *Anopheles gambiae*. *Science* 298(5591):159–165.

Chung H, Carroll SB. 2015. Wax, sex and the origin of species: dual roles of insect cuticular hydrocarbons in adaptation and mating. *Bioessays* 37(7):822–830.

Cruz F, Julca I, Gómez-Garrido J, Loska D, Marcet-Houben M, Cano E, Galán B, Frias L, Ribeca P, Derdak S, et al. 2016. Genome sequence of the olive tree, *Olea europaea*. *Gigascience* 5:29.

Cui X, Lu Z, Wang S, Jing-Yan Wang J, Gao X. 2016. CMsearch: simultaneous exploration of protein sequence space and structure space improves not only protein homology detection but also protein structure prediction. *Bioinformatics* 32(12):i332–i340.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* 3(10):e314.

De Gregorio E, Spellman PT, Rubin GM, Lemaitre B. 2001. Genome-wide analysis of the *Drosophila* immune response by using oligonucleotide microarrays. *Proc Natl Acad Sci U S A.* 98(22):12590–12595.

Ding X, Jiang W, Zhou P, Liu L, Wan X, Yuan X, Wang X, Chen M, Chen J, Yang J, et al. 2015. Mixed lineage leukemia 5 (MLL5) protein stability is cooperatively regulated by O-GlcNac transferase (OGT) and ubiquitin specific protease 7 (USP7). *PLoS One* 10(12):e0145023.

Douglas AE. 2006. Phloem-sap feeding by animals: problems and solutions. *J Exp Bot.* 57(4):747–754.

Dubinin VB. 1962. Class Acaromorpha: mites or gnathosomic chelicerate arthropods. In: Rodendorf BB, editor. Fundamentals of palaeontology. Moscow (Russia): Academy of Sciences of the USSR. p. 447–473(In Russian).

Duncan RP, Feng H, Nguyen DM, Wilson A. 2016. Gene family expansions in aphids maintained by endosymbiotic and nonsymbiotic traits. *Genome Biol Evol.* 8(3):753–764.

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

Dunlop JA, Selden PA. 2009. Calibrating the chelicerate clock: a paleontological reply to Jeyaprakash and Hoy. *Exp Appl Acarol*. 48(3):183–197.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32(5):1792–1797.

Evans JD, Aronstein K, Chen YP, Hetru C, Imler JL, Jiang H, Kanost M, Thompson GJ, Zou Z, Hultmark D. 2006. Immune pathways and defence mechanisms in honey bees *Apis mellifera*. *Insect Mol Biol*. 15(5):645–656.

Favret C. 2013. Aphid species file (version 5.0/5.0). Available from: http://aphid.speciesfile.org; last accessed January 20, 2019.

Fernández R, Marcet-Houben M, Legeai F, Richard G, Robin S, Wucher V, Pegueroles C, Gabaldón T, Tagu D. 2019. Selection following gene duplication shapes recent genome evolution in the pea aphid *Acyrthosiphon pisum*. BioRxiv. doi: https://doi.org/10.1101/643544.

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit 1 from diverse metazoan invertebrates. *Mol Mar Biol Biotechnol*. 3(5):294–299.

Francis F, Vanhaelen N, Haubruge E. 2005. Glutathione S-transferases in the adaptation to plant secondary metabolites in the *Myzus persicae* aphid. *Arch Insect Biochem Physiol*. 58(3):166–174.

Frias L, Ribeca P. 2016. ASM scripts. Available from: https://github.com/lfrias81/anchor-asm/tree/master/wrapper; last accessed July 17, 2018.

Gabaldón T. 2008. Large-scale assignment of orthology: back to phylogenetics? *Genome Biol*. 9(10):235.

Gabaldón, T, Alioto TS. 2016. Whole-Genome Sequencing Recommendations. In Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing. In: Aransay, AM, Lavín Trueba JL. editors. Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing. Cham: Springer International Publishing, pp. 13–41.

Gerardo NM, Altincicek B, Anselme C, Atamian H, Barribeau SM, de Vos M, Duncan EJ, Evans JD, Gabaldón T, Ghanim M, et al. 2010. Immunity and other defenses in pea aphids, *Acyrthosiphon pisum*. *Genome Biol*. 11(2):R21.

Glasauer SMK, Neuhauss S. 2014. Whole-genome duplication in teleost fishes and its evolutionary consequences. *Mol Genet Genomics*. 289(6):1045–1060.

Grimaldi, DA, Engel MS. 2005. Evolution of the Insects. New York: Cambridge University Press.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 59(3):307–321.

Gupta SK, Kupper M, Ratzka C, Feldhaar H, Vilcinskas A, Gross R, Dandekar T, Förster F. 2015. Scrutinizing the immune defence inventory of *Camponotus floridanus* applying total transcriptome sequencing. *BMC Genomics* 16(1):540.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol*. 9(1):R7.

Hansen AK, Moran NA. 2011. Aphid genome expression reveals host-symbiont cooperation in the production of amino acids. *Proc Natl Acad Sci U S A*. 108(7):2849–2854.

Havill NP, Foottit RG, von Dohlen CD. 2007. Evolution of host specialization in the Adelgidae (Insecta: Hemiptera) inferred from molecular phylogenetics. *Mol Phylogenet Evol*. 44(1):357–370.

Heie OE. 1987. The evolutionary history of aphids and a hypothesis on the coevolution of aphids and plants. *Boll Zool Agr Bachic*. 28:149–155.

Heie OE. 1999. Aphids of the past (Hemiptera: Sternorrhyncha). In: Proceedings of the First Palaeoentomological Conference . Moscow: AMBA, Bratislava, pp. 49–55.

Heie OE, Wegierek P, 2009. A classification of the Aphidomorpha (Hemiptera Sternorrhyncha) under consideration of the fossil taxa. *Redia* 92:69–77.

Heie OE, Wegierek P. 2011. A list of fossil aphids (Hemiptera, Sternorrhyncha, Aphidomorpha). Department of Natural History Upper Silesian Museum.

Hirst S. 1923. *On some Arachnid remains from the Old Red Sandstone (Rhynie Chert Bed, Aberdeenshire)*. *Ann Mag Nat Hist*. 12(70):455–474.

Hong YC. 2002. Atlas of amber insects of China. Beijing (China): Beijing Science and Technology Press.

Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, Denisov I, Kormes D, Marcet-Houben M, Gabaldón T. 2011. PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res*. 39(Database issue):D556–D560.

Huerta-Cepas J, Capella-Gutiérrez S, Pryszcz LP, Marcet-Houben M, Gabaldón T. 2014. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res*. 42(D1):D897–D902.

Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol Biol Evol*. 34(8):2115–2122.

Huerta-Cepas J, Gabaldón T. 2011. Assigning duplication events to relative temporal scales in genome-wide studies. *Bioinformatics* 27(1):38–45.

Huerta-Cepas J, Marcet-Houben M, Pignatelli M, Moya A, Gabaldón T. 2010. The pea aphid phylome: a complete catalogue of evolutionary histories and arthropod orthology and paralogy relationships for *Acyrthosiphon pisum* genes. *Insect Mol Biol*. 19(Suppl 2):13–21.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol*. 33(6):1635–1638.

International Aphid Genomics Consortium. 2010. Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biol*. 8:e1000313.

Jacobson AL, Johnston JS, Rotenberg D, Whitfield AE, Booth W, Vargo EL, Kennedy GG. 2013. Genome size and ploidy of *Thysanoptera*. *Insect Mol Biol*. 22(1):12–17.

Jaubert-Possamai S, Rispe C, Tanguy S, Gordon K, Walsh T, Edwards O, Tagu D. 2010. Expansion of the miRNA pathway in the hemipteran insect *Acyrthosiphon pisum*. *Mol Biol Evol*. 27(5):979–987.

Jiménez-Guri E, Huerta-Cepas J, Cozzuto L, Wotton KR, Kang H, Himmelbauer H, Roma G, Gabaldón T, Jaeger J. 2013. Comparative transcriptomics of early dipteran development. *BMC Genomics* 14(1):123.

Johnston JS, Bernardini A, Hjelmen CE. 2019. Genome size estimation and quantitative cytogenetics in insects. *Methods Mol Biol*. 1858:15–26.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30(9):1236–1240.

Julca I, Marcet-Houben M, Vargas P, Gabaldón T. 2018. Phylogenomics of the olive tree (*Olea europaea*) reveals the relative contribution of ancient allo- and autopolyploidization events. *BMC Biol*. 16(1):15.

Kang JS, Cho Y-J, Kim JH, Kim SH, Yoo S, Noh S-J, Park J, Yoon KS, Marshall Clark J, Pittendrigh BR, et al. 2015. Comparison of the genome profiles between head and body lice. *J Asia Pac Entomol*. 18(3):377–382.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*. 33(2):511–518.

Kenny NJ, Chan KW, Nong W, Qu Z, Maeso I, Yip HY, Chan TF, Kwan HS, Holland PWH, Chu KH, et al. 2016. Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs. *Heredity* 116(2):190–199.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res*. 19(9):1639–1645.

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

Labandeira CC, Phillips TL. 1996. A Carboniferous insect gall: insight into early ecologic history of the Holometabola. *Proc Natl Acad Sci U S A*. 93(16):8470–8474.
Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol*. 62(4):611–615.
Lassmann T, Sonnhammer E. 2005. Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics* 6(1):298.
Latorre A, Manzano-Marín A. 2017. Dissecting genome reduction and trait loss in insect endosymbionts. *Ann NY Acad Sci*. 1389(1):52–75.
Lazarus MB, Jiang J, Gloster TM, Zandberg WF, Whitworth GE, Vocadlo DJ, Walker S. 2012. Structural snapshots of the reaction coordinate for O-GlcNAc transferase. *Nat Chem Biol*. 8(12):966–968.
Liu AG, Matthews JJ, Menon LR, McIlroy D, Brasier MD. 2014. *Haootia quadriformis* n. gen., n. sp., interpreted as a muscular cnidarian impression from the Late Ediacaran period (approx. 560 Ma). *Proc Biol Sci*. 281:pii: 20141202.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.
Li Y, Park H, Smith TE, Moran NA. 2019. Gene family evolution in the pea aphid based on chromosome-level genome assembly. *Mol Biol Evol*.
Li Z, Tiley GP, Galuska SR, Reardon CR, Kidder TI, Rundell RJ, Barker MS. 2018. Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proc Natl Acad Sci U S A*. 115(18):4713–4718.
Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*. 25(5):955–964.
Mable BK, Alexandrou MA, Taylor MI. 2011. Genome duplication in amphibians and fish: an extended synthesis. *J Zool*. 284(3):151–182.
Mandrioli M, Rivi V, Nardelli A, Manicardi GC. 2016. Genomic and cytogenetic localization of the carotenoid genes in the aphid genome. *Cytogenet Genome Res*. 149(3):207–217.
Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. 2017. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* 33(4):574–576.
Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27(6):764–770.
Marcet-Houben M, Gabaldón T. 2015. Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biol*. 13(8):e1002220.
Marco-Sola S, Sammeth M, Guigó R, Ribeca P. 2012. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods*. 9(12):1185–1188.
Mariotti M, Santesmasses D, Capella-Gutierrez S, Mateo A, Arnan C, Johnson R, D'Aniello S, Yim SH, Gladyshev VN, Serras F, et al. 2015. Evolution of selenophosphate synthetases: emergence and relocation of function through independent duplications and recurrent subfunctionalization. *Genome Res*. 25(9):1256–1267.
Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J*. 17(1):10.
Mathers TC, Chen Y, Kaithakottil G, Legeai F, Mugford ST, Baa-Puyoulet P, Bretaudeau A, Clavijo B, Colella S, Collin O, et al. 2017. Rapid transcriptional plasticity of duplicated gene clusters enables a clonally reproducing aphid to colonise diverse plant species. *Genome Biol*. 18(1):27.
McGrath CL, Gout J-F, Johri P, Doak TG, Lynch M. 2014. Differential retention and divergent resolution of duplicate genes following whole-genome duplication. *Genome Res*. 24(10):1665–1675.
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 20(9):1297–1303.
Moya A, Peretó J, Gil R, Latorre A. 2008. Learning how to live together: genomic insights into prokaryote-animal symbioses. *Nat Rev Genet*. 9(3):218–229.
Mun J-H, Kwon S-J, Yang T-J, Seol Y-J, Jin M, Kim J-A, Lim M-H, Kim JS, Baek S, Choi B-S, et al. 2009. Genome-wide comparative analysis of the *Brassica rapa* gene space reveals genome shrinkage and differential loss of duplicated genes after whole genome triplication. *Genome Biol*. 10(10):R111.
Nicholson SJ, Nickerson ML, Dean M, Song Y, Hoyt PR, Rhee H, Kim C, Puterka GJ. 2015. The genome of *Diuraphis noxia*, a global aphid pest of small grains. *BMC Genomics* 16:429.
Nováková E, Hypša V, Klein J, Foottit RG, von Dohlen CD, Moran NA. 2013. Reconstructing the phylogeny of aphids (Hemiptera: Aphididae) using DNA of the obligate symbiont *Buchnera aphidicola*. *Mol Phylogenet Evol*. 68(1):42–54.
Ohno S. 1970. Evolution by gene duplication. Berlin/Heidelberg: Springer Berlin Heidelberg.
Pan Y, Peng T, Gao X, Zhang L, Yang C, Xi J, Xin X, Bi R, Shang Q. 2015. Transcriptomic comparison of thiamethoxam-resistance adaptation in resistant and susceptible strains of *Aphis gossypii* Glover. *Comp Biochem Physiol Part D Genomics Proteomics*. 13:10–15.
Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23(9):1061–1067.
Proost S, Fostier J, De Witte D, Dhoedt B, Demeester P, Van de Peer Y, Vandepoele K. 2012. i-ADHoRe 3.0–fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res*. 40(2):e11.
Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu J-K, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature* 453(7198):1064–1071.
Rebijith KB, Asokan R, Hande HR, Joshi S, Surveswaran S, Ramamurthy VV, Krishna Kumar NK. 2017. Reconstructing the macroevolutionary patterns of aphids (Hemiptera: Aphididae) using nuclear and mitochondrial DNA sequences. *Biol J Linn Soc*. 121(4):796–814.
Roller L, Žitňanová I, Dai L, Šimo L, Park Y, Satake H, Tanaka Y, Adams ME, Žitňan D. 2010. Ecdysis triggering hormone signaling in arthropods. *Peptides* 31(3):429–441.
Santesmasses D, Mariotti M, Guigó R. 2018. Selenoprofiles: a computational pipeline for annotation of selenoproteins. *Methods Mol Biol*. 1661:17–28.
Scarborough CL, Ferrari J, Godfray H. 2005. Aphid protected from pathogen by endosymbiont. *Science* 310(5755):1781.
Schwager EE, Sharma PP, Clarke T, Leite DJ, Wierschin T, Pechmann M, Akiyama-Oda Y, Esposito L, Bechsgaard J, Bilde T, et al. 2017. The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. *BMC Biol*. 15(1):62.
Session AM, Uno Y, Kwon T, Chapman JA, Toyoda A, Takahashi S, Fukui A, Hikosaka A, Suzuki A, Kondo M, et al. 2016. Genome evolution in the allotetraploid frog *Xenopus laevis*. *Nature* 538(7625):336–343.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19):3210–3212.
Simpson JT. 2014. Exploring genome characteristics and sequence quality without a reference. *Bioinformatics* 30(9):1228–1235.
Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Res*. 19(6):1117–1123.
Smadja C, Shi P, Butlin RK, Robertson HM. 2009. Large gene family expansions and adaptive evolution for odorant and gustatory receptors in the pea aphid, *Acyrthosiphon pisum*. *Mol Biol Evol*. 26(9):2073–2086.
Soderlund C, Bomhoff M, Nelson WM. 2011. SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res*. 39(10):e68.

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.

Tanaka H, Ishibashi J, Fujita K, Nakajima Y, Sagisaka A, Tomimoto K, Suzuki N, Yoshiyama M, Kaneko Y, Iwasaki T, et al. 2008. A genome-wide analysis of genes and gene families involved in innate immunity of *Bombyx mori*. *Insect Biochem Mol Biol*. 38(12):1087–1110.

Taylor JS, Van de Peer Y, Braasch I, Meyer A. 2001. Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philos Trans R Soc Lond B*. 356(1414):1661–1679.

Tjallingii WF. 1995. Regulation of phloem sap feeding by aphids. In: Chapman RF, de Boer G, editors. Regulatory mechanisms in insect feeding. Boston: Springer. p. 190–209.

Tupec M, Buček A, Janoušek V, Vogel H, Prchalová D, Kindl J, Pavlíčková T, Wenzelová P, Jahn U, Valterová I, et al. 2019. Expansion of the fatty acyl reductase gene family shaped pheromone communication in *Hymenoptera*. *Elife* 8:pii: e39231.

Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. *Nat Rev Genet*. 18(7):411–424.

Van Emden HF, Harrington R. 2017. Aphids as crop pests. 2nd ed. Wallingford/Oxford: CABI.

Vandenborre G, Smagghe G, Ghesquière B, Menschaert G, Nagender Rao R, Gevaert K, Van Damme E. 2011. Diversity in protein glycosylation among insect species. *PLoS One* 6(2):e16682.

Vilcinskas A. 2016. Biology and Ecology of Aphids. UK: CRC Press.

Von Dohlen C. 2000. Molecular data support a rapid radiation of aphids in the Cretaceous and multiple origins of host alternation. *Biol J Linn Soc*. 71(4):689–717.

von Dohlen CD, Spaulding U, Patch KB, Weglarz KM, Foottit RG, Havill NP, Burke GR. 2017. Dynamic acquisition and loss of dual-obligate symbionts in the plant-sap-feeding *Adelgidae* (Hemiptera: Sternorrhyncha: Aphidoidea). *Front Microbiol*. 8:1037.

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. 2017. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33(14):2202–2204.

Wallace IM, O'Sullivan O, Higgins DG, Notredame C. 2006. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res*. 34(6):1692–1699.

Walski T, De Schutter K, Van Damme EJM, Smagghe G. 2017. Diversity and functions of protein glycosylation in insects. *Insect Biochem Mol Biol*. 83:21–34.

Warnes GR, Bolker B, Bonebakker L, et al. 2016. gplots: various R programming tools for plotting data. R package version 3.0. 1. The Comprehensive R Archive Network.

Werren JH, Richards S, Desjardins CA, et al. 2010. Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. *Science* 327(5963):343–348.

Wenger JA, Cassone BJ, Legeai F, Johnston JS, Bansal R, Yates AD, Coates BS, Pavinato VAC, Michel A. 2017. Whole genome sequence of the soybean aphid, *Aphis glycines*. *Insect Biochem Mol Biol*. S0965-1748(17)30005-X.

Wolfe KH. 2015. Origin of the yeast whole-genome duplication. *PLoS Biol*. 13(8):e1002221.

Xia X, Yu L, Xue M, Yu X, Vasseur L, Gurr GM, Baxter SW, Lin H, Lin J, You M. 2015. Genome-wide characterization and expression profiling of immune genes in the diamondback moth, *Plutella xylostella* (L.). *Sci Rep*. 5:9877.

Yang Z, Rannala B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol Biol Evol*. 23(1):212–226.

Zhang J. 2003. Evolution by gene duplication: an update. *Trends Ecol Evol (Amst)*. 18(6):292–298.

Zhang R, Wang B, Grossi G, Falabella P, Liu Y, Yan S, Lu J, Xi J, Wang G. 2017. Molecular basis of alarm pheromone detection in aphids. *Curr Biol*. 27(1):55–61.

Zou Z, Evans JD, Lu Z, Zhao P, Williams M, Sumathipala N, Hetru C, Hultmark D, Jiang H. 2007. Comparative genomic analysis of the *Tribolium* immune system. *Genome Biol*. 8(8):R177.

Downloaded from https://academic.oup.com/mbe/article/37/3/730/5614851 by CSIC user on 07 October 2020