Insect Molecular Biology

Insect Molecular Biology (2010), **19** (Suppl. 2), 13–21 doi: 10.1111/j.1365-2583.2009.00947.x

The pea aphid phylome: a complete catalogue of evolutionary histories and arthropod orthology and paralogy relationships for *Acyrthosiphon pisum* genes

J. Huerta-Cepas*, M. Marcet-Houben*, M. Pignatelli†, A. Moya† and T. Gabaldón*

**Bioinformatics and Genomics Programme, Centre de Regulació Genòmica, Doctor Aiguader, Barcelona, Spain; and †Institut Cavanilles de Biodiversitat i Biología Evolutiva, Universitat de València, Avenida Blasco Ibáñez, València (Spain), CIBER en Epidemiología y Salud Pública (CIBEResp) and Centro Superior de Salud Pública (CSISP), Consellería de Sanitat (Generalitat Valenciana), Avenida de Cataluña, València, Spain*

Abstract

Phylogenetic analyses serve many purposes, including the establishment of orthology relationships, the prediction of protein function and the detection of important evolutionary events. Within the context of the sequencing of the genome of the pea aphid, *Acyrthosiphon pisum*, we undertook a phylogenetic analysis for every protein of this species. The resulting phylome includes the evolutionary relationships of all predicted aphid proteins and their homologues among 13 other fully-sequenced arthropods and three out-group species. Subsequent analyses have revealed multiple gene expansions that are specific to aphids and have served to transfer functional annotations to 4058 pea aphid genes that display one-to-one orthology relationships with *Drosophila melanogaster* annotated genes. All phylogenies and alignments are accessible through the PhylomeDB database. Here we provide a description of this dataset and provide some examples on how can it be exploited.

Keywords: Phylome, phylogeny, aphid, gene duplication, orthology.

Correspondence: Toni Gabaldón, Bioinformatics and Genomics Programme, Centre de Regulació Genòmica, Doctor Aiguader, 88. 08003 Barcelona, Spain. Tel.: +34 933160281; fax: +34 93 3969983; e-mail: tgabaldon@crg.es

Introduction

The phylogenetic analysis of molecular sequences has numerous applications. Among many other purposes, the availability of phylogenetic trees is instrumental for establishing reliable orthology and paralogy predictions, for elucidating the function of uncharacterized proteins or for the detection of several evolutionary events. In recent years, the development of faster algorithms and automated pipelines for phylogenetic inference has paved the way for the computation of large sets of multiple sequence alignments and phylogenetic trees, including the reconstruction of the evolutionary history of all genes encoded in a given genome, i.e. the *phylome* (Huerta-Cepas *et al*., 2007). The availability of such large datasets provides us with a genome-wide view of the evolution of a given organism from the perspective of all the individual components of its proteome. Questions that can be addressed through the use of a phylome range from evaluating the level of genome-wide support for alternative evolutionary scenarios in a species phylogeny to the study of how gene duplication events have shaped a particular genome. Moreover, the analysis of all gene phylogenies can be used to produce a set of highly reliable predictions of orthology and paralogy relationships among the genomes considered (Huerta-Cepas *et al*., 2007, Gabaldón, 2008). This application is of particular relevance in the context of newly sequenced genomes, since it allows for reliable automated transfers of functional annotations based on clear orthology, rather than just homology, relationships.

Within the context of the sequencing of the complete genome of the pea aphid, *Acyrthosiphon pisum*, and in order to improve the automated and manual functional annotation of the predicted gene set for this insect species, we undertook the reconstruction of the pea aphid phylome. This large-scale phylogenetic collection includes the evolutionary relationships of all *A. pisum* proteins and their homologues among thirteen other arthropods with fully-sequenced genomes and three out-group species. All the resulting phylogenies, multiple sequence alignments and orthology and paralogy predictions are made

accessible through phylomeDB (Huerta-Cepas *et al*., 2008), providing a powerful resource for biologists who want to explore the evolutionary history of particular proteins of interest. Here, we describe the details of the reconstruction of the pea aphid phylome and provide an overview of how it can be exploited to gain insight into aphid biology. Among the analyses performed, we highlight the use of the *A. pisum* phylome to detect aphid-specific gene family expansions and to transfer high-quality functional annotations to 4058 pea aphid genes that display one-to-one orthology relationships with *Drosophila melanogaster* annotated genes. Finally, an arthropod species phylogeny is reconstructed based on sequence data from the seventeen species included in the phylome.

Results and discussion

Reconstruction of the pea aphid phylome

The combination of manual and automated gene annotation of the first draft assembly of the *A. pisum* genome, Acyr_1.0, produced a consensus gene set of 34 600 genes (International Aphid Genomics Consortium, 2010). We compared the proteins encoded in this set with those encoded in seventeen fully-sequenced genomes (see Table 1). This species set includes twelve other insects, representing all major insect groups with sequenced genomes, including: representatives from paraneoptera, hymenoptera, coleoptera, amphiesmenoptera, nematocera and brachycera; the crustacean *Daphnia pulex*; and three non-arthropod out-groups, including the nematode *Caenorhabditis elegans* and the chordates *Ciona intestinalis* and *Homo sapiens*. Our sequence searches revealed that 12 885 genes in the Acyr_1.0 gene set (37%) do not present significant similarity (e-value < 10^{-3}) with genes in other species included in the analysis. This large number of putative species-specific genes might be in part due to high false-positive rates in gene prediction programs. Genes encoded in the *A. pisum* genome have been predicted using a combination of the NCBI evidence-based RefSeq annotation pipeline, which uses evidence from expressed sequence tag data and protein homology to support a given gene structure, and a combination of *ab initio* gene prediction programs combined into a single prediction with GLEAN (International Aphid Genomics Consortium, 2010). Of the 34 604 genes in the Acyr_1.0 gene set, 12 251 are based on RefSeq annotation, thus the level of predicted genes based on *ab initio* approaches (more error prone) is quite high and would be compatible with a high rate of false positives. An abundance of transposable elements in *A. pisum* might be an additional reason for the high specific gene count. Although several insect genome projects include pipelines to detect transposable elements, these are rarely eliminated from the initial, automatically generated consensus gene sets at least for the gene models that are predicted *ab initio*. These gene-finding programs usually mask repetitive regions but do predict the protein-coding parts of the transposable elements. These can be eliminated in subsequent annotation phases. The use of our phylome pipeline in the first annotation phase of the genome prevented us from discarding putative transposable elements from the analysis. This could be accounted for in future genome projects, at least for the easily detectable transposable elements families, thus saving valuable time in the phylogenetic computations. Alternatively, as we will discuss below, the phylogenomic pipeline used here, could also serve to help in the identification of transposable elements, since they tend to involve many lineage-specific duplications. Taking into account these considerations, the analyses of aphid-specific genes might identify true genetic specificities of aphids as compared to other insects. Shared gene sets, in contrast, may provide information on the genetic similarities of different organisms. Our, sequence comparison analyses showed that *A. pisum* shares a range of 30–53% of its gene repertoire with the other insects (Fig. 1). The two species sharing the highest percentage of aphid genes were the wasp *Nasonia vitripennis* and the beetle *Tribolium castaneum* (53% in both cases). Interestingly, the closest relative among insects with sequenced genomes, the body louse *Pediculus humanus*, shares only 38% of the pea aphid genes. This low percentage is probably related to an extreme reduction in the size of the genome of this human parasite (Johnston *et al*., 2007), since genome size, and not just evolutionary distance, is one of the strongest

Table 1. Table showing the genomes used in the phylome reconstruction

Species	Source DB	Version	# proteins
Acyrthosiphon pisum	Aphidbase	1.0	34 600
Aedes aegypti	ENSEMBL	V49	16 789
Anopheles gambiae	ENSEMBL	v49	12 646
Apis mellifera	NCBI	4.0	9 257
Bombyx mori	SilkDB	N/A	14 622
Culex pipiens	VectorBase	1.1	20 307
Drosophila melanogaster	Flybase	5.9	21 064
Drosophila mojavensis	Flybase	1.2	14 595
Droso. pseudoobscura	Flybase	2.0	16 071
Drosophila yakuba	Flybase	1.3	16 082
Nasonia vitripennis	NCBI	1.0	9 254
Pediculus humanus	VectorBase	1.1	11 198
Tribolium castaneum	NCBI	3.0	9 833
Daphnia pulex	IUBIO	jgi060905	30 940
Homo sapiens	ENSEMBL	v49	21 926
Ciona intestinalis	ENSEMBL	v49	21 548
Caenorhabditis elegans	ENSEMBL	v49	20 140

The species, database source, release version and the gene count for that version is presented. The last three genomes (*Homo sapiens, Ciona intestinalis* and *Caenorhabditis elegans*) were used as out-groups in the phylogenetic analysis.

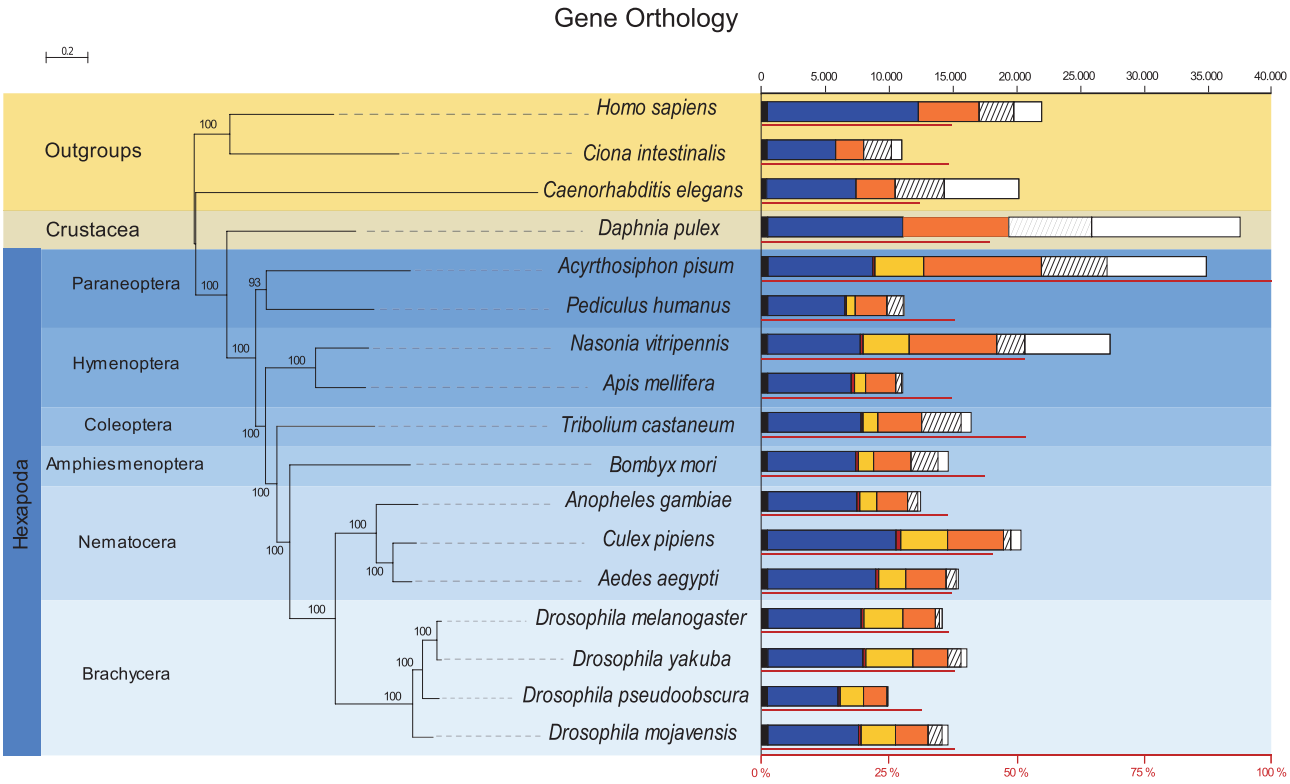

Figure 1. Comparative genomics of insect species. The species phylogeny is based on maximum likelihood analysis of a concatenated alignment of 197 widespread, single-copy proteins. The tree has been rooted using chordates as the most basal out-group. Different background colours represent taxonomic groupings within the species used to make the tree. Bars represent the total number of genes for each species (scale on the top). These have been divided to indicate different types of homology relationships. Black: widespread genes that are found with a one-to-one orthology in at least 16 of the 17 species; Blue: widespread genes that can be found in at least 16 of the 17 species and are sometimes present in more than one copy; Red: widespread but insect-specific genes present in at least 12 of the 13 insect species; Yellow: non-widespread insect-specific genes (present in less than 12 insect species); Orange: genes present in insects and other groups but with a patchy distribution; White: species-specific genes with no (detectable) homologues in other species (striped section corresponds to species-specific genes present in more than one copy). The thin red line under each bar represents the percentage of *A. pisum* genes that have homologues in a given species.

determinants of shared gene content between related species (Snel *et al.*, 1999).

We subsequently applied a similar pipeline to the one used for the human phylome (Huerta-Cepas *et al.*, 2007) to reconstruct the phylogenies of every single aphid gene, obtaining a total of 23 523 phylogenetic trees and multiple sequence alignments (see Experimental procedures). First, significant hits (*e*-value $< 10^{-3}$) that overlapped with more than 50% of the query aphid sequence were selected to reconstruct the phylogeny. Multiple sequence alignments of homologous proteins were obtained with MUSCLE v.3.6 (Edgar, 2004) and then trimmed with trimAl (Capella-Gutierrez *et al.*, 2009) to filter out gap-rich columns. Phylogenetic analyses were performed using Neighbor Joining (NJ) and Maximum Likelihood (ML) approaches as implemented in PhyML (Guindon & Gascuel, 2003) (see Experimental procedures section for more details). The resulting alignments, phylogenies and orthology predictions can be accessed through phylomeDB (http://phylomedb.org) and AphidBase (Legeai *et al.*, 2009) (http://www.aphidbase.com) databases. PhylomeDB is a public database for complete collections of gene phylogenies (phylomes) that allows users to explore the evolutionary history of genes through the visualization of phylogenetic trees and alignments, and to obtain their phylogeny-based orthology and paralogy relationships across a number of species. Since these trees and alignments are generated automatically, it is recommended to inspect the protein alignments to judge the quality of the data. As explained in the methods section, these alignments can be refined and expanded for further analyses.

Detection of orthology and paralogy relationships across insect genomes

As discussed above, one of the main applications of complete phylomes is the possibility of obtaining high-quality homology relationships based on phylogenetic analyses. In order to generate a complete catalogue of orthology and paralogy predictions among aphid genes and their homologues in the other genomes considered, we scanned the pea aphid phylome with a previously

described orthology prediction algorithm (Huerta-Cepas *et al*., 2007). In brief, this algorithm scans a phylogenetic tree, and uses the level of species overlap between sister tree branches to detect and mark duplication and speciation events. Then, using the prediction of speciation and duplication nodes in the tree, the algorithm establishes orthology and paralogy relationships according to the original evolutionary definition of these terms (Fitch, 1970). This algorithm has been shown to produce highly reliable orthology predictions and to be superior to the alternative phylogeny-based method based on tree reconciliation with the species phylogeny (Marcet-Houben & Gabaldón, 2009). A table of orthology and paralogy predictions of *A. pisum* genes and their homologues in the 16 other species included in the analysis can be retrieved from phylomeDB download section. Additionally, the visualization of phylogenetic trees through the phylomeDB interface allows the inspection of speciation and duplication events, indicated with different colours.

High-quality transfer of functional annotations through phylogeny-based orthology prediction

Functional annotation in newly sequenced genomes is usually performed using blast searches against related organisms or public repositories and then transferring the annotations from the top hits. The use of homology, rather than orthology, to infer the function of a protein presents known caveats, which may lead to wrong annotations that are propagated in the databases (Jones *et al*., 2007). Paralogous genes, as opposed to orthologues, are less likely to share a particular function because duplications may promote functional divergence among duplicated genes through processes of neo- and sub-functionalization (Conant & Wolfe, 2008). Therefore the use of phylogeny-based orthology prediction is more likely to produce reliable transfers of functional annotations, especially when related species are used. Transferring functional information is most reliable when achieved using one-to-one orthology relationships, meaning that there is, respectively, only one orthologue of a given gene in the other species. When one or more genes are co-orthologous to a set of genes in another genome (one-to-many or many-to-many orthology relationships), duplications that occurred within one of the lineages considered might have been associated with functional shifts, thereby affecting the reliability of the functional transfer.

In order to produce a high confidence set of functional predictions for *A. pisum* genes, we used our previously described catalogue of evolutionary relationships to obtain the subset of one-to-one orthologies between the pea aphid and *Drosophila melanogaster*, the most intensively studied model insect. Using this phylogeny-based approach we could transfer functional Gene Ontology (GO) annotations to 4059 pea aphid genes (see Experimental procedures and Fig. 2). These annotations have been included in the corresponding AphidBase entries. An additional set of 2282 *A. pisum* genes showed orthologies of the type one-to-many, many-to-one or many-to-many with annotated *D. melanogaster* genes. Although less reliable than those based on one-to-one orthology relationships, annotation transfers based on other type of orthology relationships can provide important hints to predict the actual function of aphid genes. For instance, the functions of a group of co-orthologous genes can be transferred to a single gene in a many-to-one relationship. To our knowledge, this is the first newly sequenced genome for which phylogeny-based orthology predictions have been used in the annotation pipeline.

Detection of lineage-specific duplications and losses

Another advantage of the availability of the phylome, is that we can readily obtain a picture of the gene duplications that have occurred within the *A. pisum* lineage. To do so, we used the above mentioned phylogenetic algorithm to detect all paralogy relationships that were specific for aphid (in-paralogies). 2459 gene families presented lineage-specific duplications. Most of these gene family expansions are small-to-moderate in size, resulting in a total of two to 10 in-paralogs (2239 families). The remaining 220 families seem to have experienced massive expansions resulting in in-paralogues groups with 10–50 members (196 families) and 50 to 209 members (19 families). Sequence analyses of members of the latter groups have identified reverse-transcriptase and transposase domains, suggesting that these may represent expansions of transposable elements. Indeed, 1921 genes coding for these activities were found among families with more than five in-paralogues, representing 49% of the total genes in families of that size. The remaining families, have functions that cannot be associated with transposable elements (see supplementary material Table S1). These families, which underwent extremely large aphid-specific expansions, as well as those families that appear to be lost in this species might be correlated with aphid morphological or physiological specificities. A functional enrichment analysis of pea-aphid gene expansions and losses suggests that this is the case (Table 2). Interestingly, the expansion of families involved in amino acid, oligopeptide and carbohydrate transport might be related to the highly specific food source for aphids, which feed from plant phloem sap. The particular composition of this diet, which is sugar-rich and amino-acid poor, might require specific adaptations in the number and specificities of the corresponding transporters (Douglas, 2006). Other expanded families involve those participating in

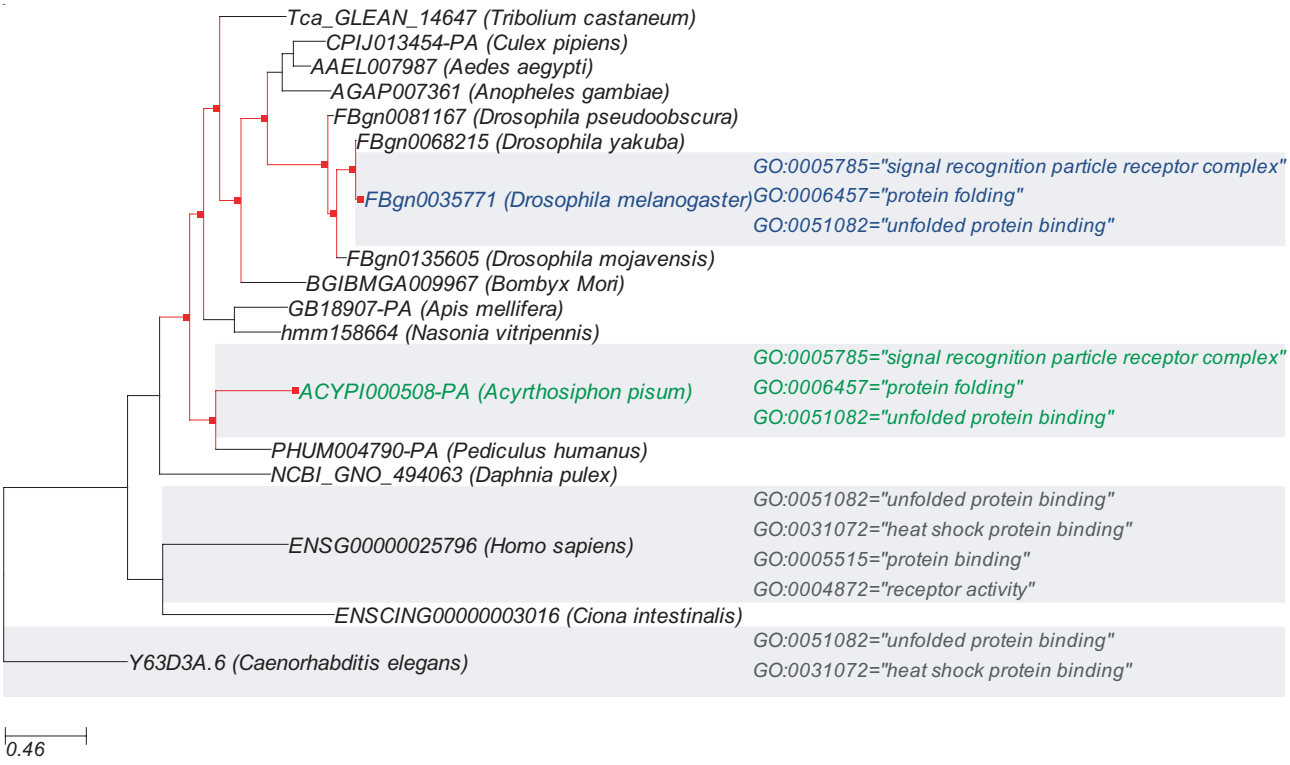

Figure 2. Example of an automated transfer of functional annotation from *Drosophila melanogaster* to an *Acythosiphon pisum* gene. The figure shows the reconstructed phylogeny of the pea aphid gene ACYPI000508-PA, including its evolutionary relationships with homologues in other species. Speciation nodes and associated branches that connect *A. pisum* and *D. melanogaster* genes, and in which the inference of one-to-one orthology relationship is based, are marked in red. The right side of the tree indicates Gene Ontology (GO) terms associated to the *Drosophila* gene Fbgn0035771 (in blue), and those inferred for *A. pisum* (in green). Additionally, functional terms associated with orthologues in other species are shown in grey and provide further support for the annotation transfer.

Table 2. Table showing over-represented Gene Ontology terms (biological process) in families that have expanded in the pea aphid lineage (more than 10 in-paralogues) and in insect genes that have been lost, specifically, in the pea aphid lineage

Over-represented terms in aphid-specific family expansions		
GO:0006508	Proteolysis	1.745540e-76
GO:0006857	Oligopeptide transport	5.530970e-06
GO:0006865	Amino acid transport	2.431340e-16
GO:0006916	Anti-apoptosis	6.657330e-19
GO:0006979	Response to oxidative stress	1.112340e-04
GO:0007608	Sensory perception of smell	2.338690e-11
GO:0008643	Carbohydrate transport	1.550560e-27
GO:0042048	Olfactory behavior	1.728790e-05
GO:0050790	Regulation of catalytic activity	7.487550e-04
GO:0050896	Response to stimulus	1.825330e-15
Over-represented terms in aphid-specific gene losses		
GO:0006952	Defense response	1.036770e-04
GO:0006955	Immune response	5.877600e-13
GO:0009253	Peptidoglycan catabolic process	8.563480e-19
GO:0016045	Detection of bacterium	3.201630e-05
GO:0016998	Cell wall macromolecule catabolic process	1.778300e-09
GO:0019730	Antimicrobial humoral response	3.440020e-06
GO:0048096	Chromatin-mediated maintenance of transcription	1.208550e-05
GO:0050909	Sensory perception of taste	3.686630e-04

For both groups the columns indicate the Gene Ontology (GO) term code, description statistical significance (*P*-value) of the over-representation (see methods).

processes such us perception of smell, olfactory behaviour and response to stimulus. This might be interpreted in the context of the necessity of aphids to recognize a specific type of plant host. Interestingly, this expansion of smell-related pathways is coupled to a significant loss of genes related to the perception of taste. This might again be related to a very simple and stable diet in the aphids, which would not rely on their taste to select the food source, in contrast to other organisms such as flies that use many different sources. Additional processes that have been significantly reduced include those related to immune and antimicrobial response and detection of bacterium. As it is discussed in an additional companion paper (Gerardo *et al.*, 2009), this softened immune response might be related to the association of aphids with bacterial endosymbionts such as *Buchnera aphidicola*.

The detailed analysis of specific duplications can help to derive testable hypotheses. For instance one of the largest expansions found by our analysis corresponds to a group of co-orthologues to the *D. melanogaster* gene *kelch*, a protein involved in oogenesis and ovarian

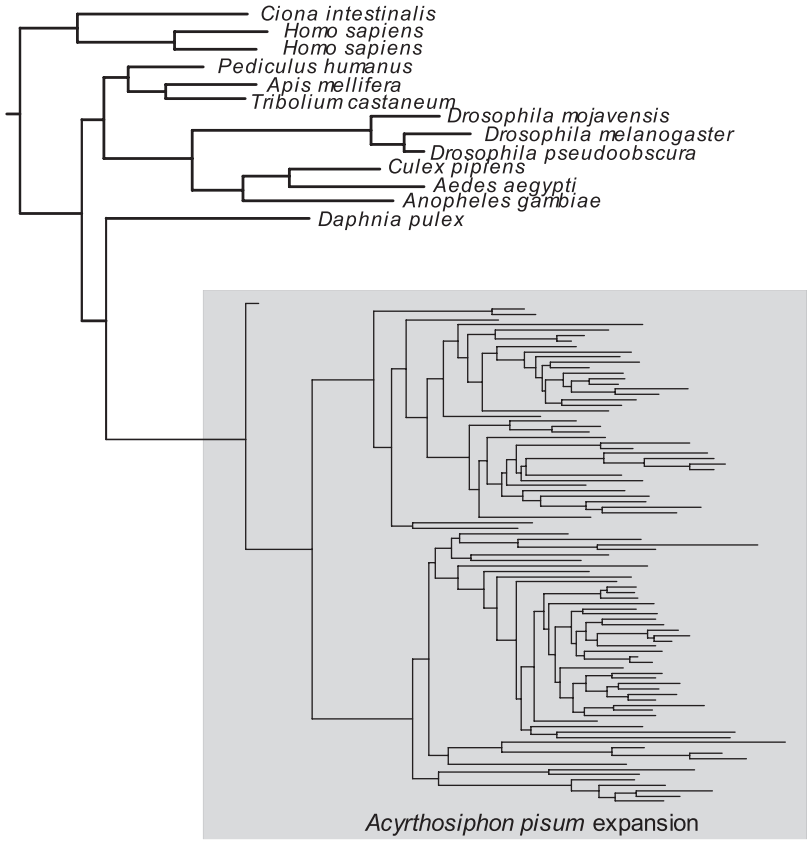

Figure 3. Phylogenetic tree depicting the intraspecific expansion of the kelch protein family in *Acyrthosiphon pisum*. The protein ACYPI007299-PA was used as a seed to build this phylogenetic tree using the phylome pipeline described in the text. An ancient duplication preceding the root of the shown tree has been omitted in the figure but the full tree can be seen at phylomeDB. All branch tips within the grey square represent in-paralogous members of the kelch family in the pea aphid. For simplicity, their names and IDs have been omitted.

organization (Fig. 3). This gene presents one-to-one orthology relationships with other insect species, whereas in the pea aphid lineage, several rounds of gene duplications created a large paralogous group with about 200 members. Although additional research should be carried out to understand the role of this protein family in aphids, it is tempting to speculate that it might be involved in aphid-specific processes such as reproductive poliphenism and the large morphological differences observed between oviparous and viviparous ovaries in aphid females. In aphids, sexual and asexual females display large differences in the morphology of their ovaries and the oocyte differentiation cycles. Ovaries from different morphotypes differ in many characteristics such as the presence or absence of embryos, accessory glands or spermathecae, and in the relative size of their oocytes and, finally, in how and when the eggs are arrested and extruded (Pyka-Fosciak & Szklarzewicz, 2008). In *Drosophila*, the protein encoded by *kelch* has been shown to be necessary to maintain actin organization in ovarian ring canals and for oocyte migration (Xue & Cooley, 1993). A high degree of diversification of this protein family in aphids might have thus facilitated the emergence of complex and diversified ovarian structures and oocyte cycles.

An insect phylogeny based on 197 widespread genes

Insects belong to one of the most successful major lineages of metazoans. With an estimated number of extant species ranging from 2 to 5 million (Mayhew, 2007), this group of arthropods possesses one of the highest level of diversity among animals. Attempts to elucidate the phylogenetic relationships among major taxa of arthropods, including insects, have presented frequent challenges, including the possible paraphyli of some of the groups (Whitfield & Kjer, 2008). *A. pisum* is the first hemipteran insect for which we have a complete genome sequence. Hemiptera, which includes aphids, leafhoppers, whiteflies, psylloids and other insects, is the largest of the non-endopterygote orders, comprising more than 50 000 species grouped in approximately 100 families. The availability of a complete genome sequence of an hemipteran allows us to better reconstruct a reliable phylogeny representing the evolutionary relationships of this group with other insect species. To do so, we performed a ML analysis of 197 concatenated alignments of genes with a single-copy orthologue in all species considered (see Experimental procedures and Fig. 1). The resulting phylogeny clusters major insect groups according to previously established taxonomy, including the Brachicera and

Nematocera suborders (Richards *et al*., 2008). This phylogeny correctly places the pea aphid as a sister group of *P. humanus*, also a member of the the para-neopthera clade, and places them at the base of the insect phylogeny. The long branch leading to *A. pisum* is indicative of a very long evolutionary distance and therefore significant genomic differences with its closer relatives. This is in line with the large differences found in terms of gene content and the significant high rate of aphid-specific gene expansions (see above).

Concluding remarks

With the sequencing of the *A. pisum* genome, aphid research enters the genome era. The reconstruction of the complete collection of evolutionary histories of aphid genes and their homologues in other sequenced arthropods has not only helped in the annotation of the genome sequence but has also already provided important insights into the evolution of this specialized hemipteran. So far, one of the major applications of the pea aphid phylome has been the use of phylogeny-based orthology predictions to transfer functional annotations to *A. pisum* genes. To our knowledge this is the first time that such a reliable methodology has been applied in the annotation pipeline of a newly sequenced genome. The availability of the pea aphid phylome to the research community is likely to produce further insights in the future.

Experimental procedures

Phylome reconstruction

We reconstructed the complete collection of phylogenetic trees for all *A. pisum* protein-coding genes. For this we used a similar automated pipeline to that described earlier for the first reconstruction of the human phylome (Huerta-Cepas *et al*., 2007). A database was created containing *A. pisum* proteome and that of 16 other species (see Table 1). For each protein encoded in *A. pisum* genome, a Smith-Waterman (Smith and Waterman, 1981) search (*e*-val <= 10^{-3}) was performed against the above mentioned proteomes. Sequences that aligned with a continuous region longer than 50% of the query sequence were selected and aligned using MUSCLE 3.6 (Edgar, 2004) with default parameters. Positions in the alignment with a high number of gaps were removed using trimAl v1.0 (Capella-Gutierrez *et al*., 2009) (http://trimal.cgenomics.org), using a gap threshold of 25% and a conservation thresohld of 50%. Phylogenetic trees were derived using NJ trees using scoredist distances as implemented in BioNJ (Gascuel, 1997) and ML as implemented in PhyML v2.4.4 (Guindon & Gascuel, 2003). In all cases, JTT was used as an evolutionary model, assuming a discrete gamma-distribution model with four rate categories and invariant sites, where the gamma shape parameter and the proportion of invariant sites were estimated form the data. Support for the different partitions was computed by approximate likelihood ratio test as implemented in PhyML aLRT version (Guindon & Gascuel, 2003; Anisimova & Gascuel, 2006), using the option 'Minimum of SH-like and Chi2-based'. All trees and alignments have been deposited in phylomeDB.

Orthology determination

Orthology and paralogy relationships among *A. pisum* genes and those encoded in the other genomes included in the analysis were inferred by a phylogenetic approach that uses a previously described species-overlap algorithm (Huerta-Cepas *et al*., 2007). Basically, this algorithm uses the level of species overlap between the two daughter partitions of a given node to define it as a duplication (if there is species overlap) or speciation (if there is no overlap). After mapping all duplication and speciation events on the phylogenetic tree of a given gene family, all orthology and paralogy relationships are inferred accordingly. Resulting orthology and paralogy predictions for *A. pisum* genes can also be accessed through phylomedDB.

Detection of aphid-specific gene expansions

The duplication events defined by the above mentioned species overlap algorithm that only comprised paralogues from *A. pisum* were considered lineage-specific duplications. Whenever more than one round of duplication followed the *A. pisum* speciation event (family expansion), all resulting paralogues were grouped into a single group of 'in-paralogues'. Results from all the trees in the phylome were merged into a non-redundant list of in-paralogues groups, by merging sets sharing a significant fraction of their members (50%).

Orthology-based functional annotation

A list of orthology-based transfer of functional annotations to *A. pisum* genes was built based on orthology relationships with annotated *D. melanogaster* genes. These were grouped according to the type of orthology relationship. 4058 aphid genes could be annotated based on a clear one-to-one orthology relationship with a *Drosophila* gene. Additional 2315 genes presented a many-to-one orthology relationship with annotated *Drosophila* genes and thus could tentatively be annotated with the GO terms associated with the fly genes but with the cautionary remark that processes of neo- and sub-functionalization may have occurred. To provide additional information on the reliability of the transfer we provide information on whether an orthologue in a species out-group was also annotated with that function (as in Fig. 2). Supplementary Table S2 includes all functional transfers performed from *D. melanogaster* and other insect orthologues.

Functional enrichment analyses

For all pea aphid families with more than 10 in-paralogues, we extracted the list of homologous genes in four other well annotated genomes: *D. melanogaster, Caenorhabditis elegans, Anopheles gambiae* and *Anopheles aegypti*. Functional terms associated with the annotated genes in the pea aphid expanded families were compared with those in the non-expanded families. Enrichment analyses of overrepresented GO terms in pea aphid expanded families were performed by using the FatiGO program

(Al-Shahrour *et al*., 2004) using the two-sided Fisher test and *e*-value cut-off of 10^{-3}.

Species tree reconstruction

197 genes having a single-copy orthologue in all the species included in the analyses were selected to infer a species phylogeny. Alignments performed with MUSCLE (Edgar, 2004) were concatenated into a super-alignment containing 144 922 positions. The removal with trimAl of columns with gaps in more than 50% of the sequences resulted in a final alignment of 90 512 positions. This alignment was used for ML tree reconstruction as implemented in PhyML v2.4.4 (Guindon & Gascuel, 2003), using JTT as an evolutionary model and assuming a discrete gamma-distribution model with four rate categories and invariant sites, where the gamma shape parameter and the fraction of invariant sites were estimated from the data. Bootstrap analysis was performed on the basis of 100 replicates.

Accessing pea aphid phylome data through phylomeDB

The pea aphid phylome comprises a total of 23 350 ML gene phylogenies and multiple sequence alignments. These data can easily be browsed through the main search panel at the phylomeDB web interface (http://phylomedb.org) or downloaded at convenience. In order to find gene-specific resources, both RefSeq and AphidBase gene identifiers are supported for querying phylomeDB entries. Additionally, a BLAST-based sequence search may help users finding their proteins of interest. Even if the protein of interest is encoded in an insect genome not present in the phylome, the users can use the blast-based search to localize homologous *A. pisum* proteins and subsequently explore or download the corresponding trees and alignments. This allows users not only to search for a specific gene phylogeny but also to expand the analysis at convenience. A possible application, for instance, could consist of downloading the multiple sequence alignment from phylomeDB, re-align it with the new protein of interest to subsequently reconstruct a new phylogenetic tree. This can be done locally or through dedicated servers such as Phylemon (Tarraga *et al*., 2007), which implements the same phylogenetic methods as those used in the phylome pipeline.

All data hosted in phylomeDB can be linked from external sources through a simple URL based system (see Table S2 for the accepted syntax). Aphid related databases such as AphidBase (http://www.aphidbase.com) and AcypiCyc (http://acypicyc.cycadsys.org) already link their data to the evolutionary information stored in phylomeDB. Finally, the complete set of pea aphid gene trees, alignments and orthology predictions are also available for large-scale analyses. The 'Downloads' section at the phylomeDB website includes links to the most commonly used datasets but any other type of data is available upon request.

Acknowledgements

The authors are grateful to the International *Pediculus humanus* and *Nasonia vitripenis* genome sequencing consortia for providing the data before publication.

References

Al-Shahrour, F., Diaz-Uriarte, R. and Dopazo, J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of Genes. *Bioinformatics* **20**: 578–580.

Anisimova, M. and Gascuel, O. (2006) Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst Biol* **55**: 539–552.

Capella-Gutierrez, S., Silla-Martinez, J.M. and Gabaldon, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.

Conant, G.C. and Wolfe, K.H. (2008) Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet* **9**: 938–950.

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

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113.

Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Syst Zool* **19**: 99–113.

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

Gascuel, O. (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* **14 (7)**: 685–695.

Gerardo, N.M., Altincicek, B., Anselme, C., Atamian, H., Barribeau, S.M., De Vos, M. *et al.* (2009) Immunity and defense in pea aphids, *Acyrtosiphon pisum*. Accepted.

Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: 696–704.

Huerta-Cepas, J., Dopazo, H., Dopazo, J. and Gabaldón, T. (2007) The human phylome. *Genome Biol* **8**: R109.

Huerta-Cepas, J., Bueno, A., Dopazo, J. and Gabaldón, T. (2008) PhylomeDB: a database for genome-wide collections of gene phylogenies. *Nucleic Acids Res* **36**: D491–D496.

International Aphid Genomics Consortium (2010) Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biology*, DOI 10.1371/journal.pbio.1000313.

Johnston, J.S., Yoon, K.S., Strycharz, J.P., Pittendrigh, B.R. and Clark, J.M. (2007) Body lice and head lice (Anoplura: Pediculidae) have the smallest genomes of any hemimetabolous insect reported to date. *J Med Entomol* **44**: 1009–1012.

Jones, C.E., Brown, A.L. and Baumann, U. (2007) Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinformatics* **8**: 170.

Legeai, F., Shigenobu, S., Gauthier, J., Colbourne, J., Rispe, C., Collin, O. *et al.* (2009) AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome. (submitted).

Marcet-Houben, M. and Gabaldón, T. (2009) The tree versus the forest: the fungal tree of life and the topological diversity within the yeast phylome. *PLoS ONE* **4**: e4357.

Mayhew, P.J. (2007) Why are there so many insect species? Perspectives from fossils and phylogenies. *Biol Rev Camb Philos Soc* **82**: 425–454.

Pyka-Fosciak, G. and Szklarzewicz, T. (2008) Germ cell cluster formation and ovariole structure in viviparous and oviparous generations of the aphid *Stomaphis quercus*. *Int J Dev Biol* **52**: 259–265.

Richards, S., Gibbs, R.A., Weinstock, G.M., Brown, S.J., Denell, R., Beeman, R.W. *et al.* (2008) The genome of the model beetle and pest *Tribolium castaneum*. *Nature* **452**: 949–955.

Smith, T.F. and Waterman, M.S. (1981) Identification of Common Molecular Subsequences. *Journal of Molecular Biology* **147**: 195–197.

Snel, B., Bork, P. and Huynen, M.A. (1999) Genome phylogeny based on gene content. *Nat Genet* **21**: 108–110.

Tarraga, J., Medina, I., Arbiza, L., Huerta-Cepas, J., Gabaldon, T., Dopazo, J. *et al.* (2007) Phylemon: a suite of web tools for molecular evolution, phylogenetics and phylogenomics. *Nucleic Acids Res* **35**: W38–W42.

Whitfield, J.B. and Kjer, K.M. (2008) Ancient rapid radiations of insects: challenges for phylogenetic analysis. *Annu Rev Entomol* **53**: 449–472.

Xue, F. and Cooley, L. (1993) kelch encodes a component of intercellular bridges in Drosophila egg chambers. *Cell* **72**: 681–693.

Supporting Information

Additional Supporting Information may be found in the online version of this article under the DOI reference: DOI 10.1111/j.1365-2583.2009.00947.x

Table S1. Functional analysis of expanded protein families in the aphid genome. Annotations are based on terms associated to best hits in blast searches against non-redundant NCBI database.

Table S2. Examples of URL accepted syntaxes to link to Phylome DB.