

# Expansion of the miRNA Pathway in the Hemipteran Insect *Acyrtosiphon pisum*

Stéphanie Jaubert-Possamai,<sup>\*</sup>1 Claude Rispe,<sup>1</sup> Sylvie Tanguy,<sup>1</sup> Karl Gordon,<sup>2</sup> Thomas Walsh,<sup>3</sup> Owain Edwards,<sup>3</sup> and Denis Tagu<sup>1</sup>

<sup>1</sup>Institut de la Recherche Agronomique-INRA, UMR1099 BiO3P, Domaine de la Motte, F-35653 Le Rheu, France

<sup>2</sup>Commonwealth Scientific and Industrial Research Organization Entomology, GPO Box 1700, Canberra ACT 2601, Australia

<sup>3</sup>Commonwealth Scientific and Industrial Research Organization Entomology, Center for Environment and Life Sciences, Floreat Park WA 6014, Australia

**\*Corresponding author:** E-mail: stephanie.jaubert@rennes.inra.fr.

**Associate editor:** Douglas Crawford

## Abstract

The pathways that allow short noncoding RNAs such as the microRNAs (miRNAs) to mediate gene regulation and control critical cellular and developmental processes involve a limited number of key protein components. These proteins are the Dicer-like RNases, double-stranded RNA (dsRNA)-binding proteins, and the Argonaute (AGO) proteins that process stem-loop hairpin transcripts of endogenous genes to generate miRNAs or long dsRNA precursors (either exogenous or endogenous). Comparative genomics studies of metazoans have shown the pathways to be highly conserved overall; the major difference observed is that the vertebrate pathways overlap in sharing a single Dicer (DCR) and AGO proteins, whereas those of insects appear to be parallel, with distinct Dicers and AGOs required for each pathway. The genome of the pea aphid is the first available for a hemipteran insect and discloses an unexpected expansion of the miRNA pathway. It has two copies of the miRNA-specific *dicer-1* and *ago1* genes and four copies of *pasha* a cofactor of *droscha* involved in miRNA biosynthesis. For three of these expansions, we showed that one copy of the genes diverged rapidly and in one case (*ago1b*) shows signs of positive selection. These expansions occurred concomitantly within a brief evolutionary period. The pea aphid, which reproduces by viviparous parthenogenesis, is able to produce several adapted phenotypes from one single genotype. We show by reverse transcriptase-polymerase chain reaction that all the duplicated copies of the miRNA machinery genes are expressed in the different morphs. Investigating the function of these novel genes offers an exciting new challenge in aphid biology.

**Key words:** pea aphid, phenotypic plasticity, positive selection, parthenogenesis.

## Introduction

Small noncoding RNAs (ncRNAs) are now known to play a central role in the regulation of gene expression, affecting cellular processes critical for development in eukaryotes (Amaral et al. 2008). The short (~22 nt) RNAs that mediate gene regulation are small interfering RNAs (siRNAs) and microRNAs (miRNAs). Work over the past decade has shown not only how they are produced and act to regulate gene expression in model systems but has also begun to show the diversity of pathways involved in different organisms. The miRNAs are processed from endogenous genes encoding stem-loop hairpin primary miRNA (pri-miRNA) transcripts, which are then processed in the nucleus by a multiprotein complex composed of Drosha (an RNase III) and Pasha, a double-stranded RNAs (dsRNAs)-binding protein (dsRBP) (Denli et al. 2004). The resulting 70 nt miRNA precursors (pre miRNAs) are exported from the nucleus by Exportin-5. In contrast, siRNAs are derived from either exogenous viral dsRNAs or endogenous dsRNA long precursors from various origins such as transposable elements, *cis*-natural antisense transcripts (*cis*-NATs), *trans*-NATs, and hairpin RNA transcripts (Okamura et al. 2008). In the cytoplasm, RNase III enzymes called

Dicer, in association with dsRBPs, process both the pre-miRNAs and the siRNA precursors to yield the mature short RNAs. These are then loaded into different multiprotein RNA-induced silencing complexes (RISC) containing members of the Argonaute (AGO) protein family, leading either to mRNA degradation or to the repression of mRNA translation.

Studies in vertebrate and model insect systems have allowed comparison of their miRNA and siRNA pathways. Vertebrates have a single Dicer responsible for the processing of both siRNA and miRNAs (Kim et al. 2009). However, *Drosophila melanogaster* possesses two Dicers *gtr* with distinct ncRNA specificities (Foerstemann et al. 2007; Jaubert et al. 2007): DCR-1 is associated with the dsRBP protein Loquacious and is specific to the miRNA pathway, whereas DCR-2 is specific to the siRNA pathway and can be associated either with the dsRBP protein Loquacious or R2D2 depending on the origin (endogenous or exogenous) of the siRNA (Lee et al. 2004; Czech et al. 2008) (supplementary fig. S1, Supplementary Material online). The role of the RISC is defined by their core protein—the AGO protein—that binds to either the miRNAs or the siRNAs (Peters and Meister 2007). The Ago family has multiple members,

which have distinct activities of RNA cleavage or translation repression. In vertebrates, miRNAs may be loaded onto any of the four AGO proteins, of which only AGO2 has RNase H activity (Meister and Tuschl 2004). Some organisms present an expansion of the AGO protein family. An impressive expansion of the Ago protein family (27 members) in the nematode *Caenorhabditis elegans* involves subfunctionalization and neofunctionalization of the various proteins, with different functions demonstrated in several pathways, including chromosome segregation and fertility (Yigit et al. 2006). Despite this expansion, only the Ago genes *alg-1* and *alg-2* are linked to miRNAs. An expansion of the AGO protein family has also been observed in insects such as the mosquitoes *Culex pipiens* and *Aedes aegypti* (Campbell et al. 2008), but its biological significance is still not understood. In *D. melanogaster*, five AGO proteins have been identified but only AGO1 and AGO2 are involved in the RISC. Perfectly complementary RNA duplexes (mainly siRNAs) are loaded in the AGO2–RISC complex that targets RNA degradation, whereas imperfectly complementary duplexes (mainly miRNAs) are loaded in the AGO1–RISC complex that directs translation repression (Ghildiyal and Zamore 2009; Kim et al. 2009). Therefore, in the model insect *Drosophila*, DrosHa, Pasha, DCR-1, Loquacious, and AGO1 act in the miRNA pathway, whereas DCR-2, R2D2/Loquacious, and AGO2 are involved in the siRNA pathway (supplementary fig. S1, Supplementary Material online).

Insect evolution is characterized by an ancient radiation about 250–350 million years that yielded the main orders. The small RNA machinery has been studied in the dipterans, including three mosquito species in addition to *D. melanogaster* (Campbell et al. 2008), as well as in the coleopteran *Tribolium castaneum* (Tomoyasu et al. 2008). These studies showed both conserved mechanisms and diverged functions in insect ncRNA pathways. Agos appear to have differentially evolved in these insects: Whereas only one gene copy of *ago1* and *ago2* is found in *D. melanogaster* and *Anopheles gambiae*, two gene copies of *ago2* are found in *T. castaneum* and in *C. pipiens* and two gene copies of *ago1* are found in *A. aegypti*. This divergence is even more pronounced for the AGO proteins unrelated to siRNAs and miRNAs such as Aubergine–Piwi and AGO3, for which an expansion has been identified in both *C. pipiens* and *A. aegypti*. However, it is important to note that the biosynthesis pathways of both miRNAs (involving unique *drosha*, *pasha*, and *dcr-1* genes) and siRNAs (a single *dcr-2* gene) have remained unchanged in these insect orders. The same appears to be true for the other insect orders for which genome sequences are now available (Obbard et al. 2009), including the hymenopteran honeybee (Honeybee Genome Sequencing Consortium 2006) and for the lepidopteran silk moth (International Silkworm Genome Consortium 2008). The recent sequencing of the complete genome of the pea aphid *Acyrtosiphon pisum* (International Aphid Genomics Consortium 2009) has provided the opportunity to analyze the miRNAs and siRNAs machinery in a new insect order, the Hemiptera. The goal of this study was to identify and annotate the full complement of genes involved in the siRNA and miRNA pathways in the pea aphid.

In this paper, our results show an unexpected gene expansion specific to the miRNA pathway. We identify duplications of the *dcr-1* and *ago1* genes and four copies of *pasha* all genes specific to the miRNA pathway. We also identify similar expansions in other aphid species. Although *ago* genes are known to be duplicated in other insect species (Campbell et al. 2008; Tomoyasu et al. 2008) and other basal metazoans (de Jong et al. 2009), this is the first example of such a broad expansion of the key miRNA biosynthetic pathway genes in a Coelomata. We study the patterns of substitutions for each of the pea aphid duplicated genes as an evaluation of altered selective pressures (a possible signature of changes in function). We observe that several of the duplications of miRNA machinery genes in *A. pisum* are characterized by marked shifts in selective pressures and extreme levels of divergence. Finally, the presence of multiple duplications in the whole pathway in *A. pisum* incited us to explore the possibility that different copies of the miRNA pathway genes are expressed at various stages of the aphid life cycle. Aphids show a high degree of phenotypic plasticity and are able to produce several distinct morphs in response to environmental cues such as parthenogenetic individuals in spring and summer and sexual males and females in autumn (Le Trionnaire et al. 2008). We show here that the different duplicated genes are expressed in both parthenogenetic and sexual morphs of the pea aphid.

## Materials and Methods

### Manual Annotation of Genes of the siRNA and miRNA Pathways

Orthologs of siRNA and miRNA machinery genes were identified by mining the genomic data for the *A. pisum* genome (Acyr 1.0 version of the assembly) at AphidBase ([www.aphidbase.com](http://www.aphidbase.com)). This was done using the corresponding *D. melanogaster* sequences as bait and the collection of predicted proteins (program BlastP) or the genomic scaffolds (program TBlastN) of *A. pisum* as targets. The first hits were all included in a preliminary phylogenetic analysis using Neighbor-Joining (NJ), which allowed us to unambiguously distinguish between homologs and genes that were more distantly related (e.g., *dcr-2* and *ago3*). Given the relatively high level of length and sequence conservation of the various genes studied, homologs always corresponded to hits with a very high *e* value (the lowest one being 1e-83 for *exportin-5*). Gene models from prediction programs were checked, resulting in only a few modifications. All annotated genes are listed in supplementary table S1 (Supplementary Material online). Amino acid sequences were then deduced for the curated pea aphid gene models for all the candidate genes. The domain distribution of the deduced *A. pisum* proteins was predicted by using the Pfam software (Finn et al. 2008) and Interproscan (Hunter et al. 2009).

### Phylogenetic Analysis

For each gene, we collected homolog sequences from other insects with sequenced genome (either using all insect

sequences available or choosing one species in each order) and from outgroups (using two chordates, *Homo sapiens* and *Ciona intestinalis*, and other arthropods if available). Amino acid sequences were aligned using T-Coffee (Notredame et al. 2000): when DNA alignments were analyzed, they were obtained by reporting the amino acid alignment on DNA sequences. Both NJ and maximum likelihood (ML) methods were used for all analyses; as both methods gave largely similar results (and identical groupings for the aphid duplicates), we chose to comment only the ML results obtained with PHYML because this method is more accurate (Guindon and Gascuel 2003). Protein analyses were done using the Jones–Taylor–Thornton model of substitution. For DNA-level analyses (used for *dcr-1* and *ago1*), parameters of the ML model were optimized using Modeltest (Posada and Crandall 1998) to test 56 different models of substitution. For *dcr-1*, the best model was that of Hasegawa, Kishino, and Yano with variable sites (gamma parameter = 0.46). For *ago1*, the best model was found to be Tamura–Nei with a gamma distribution (gamma parameter = 1.46) and a proportion of 0.322 of invariable sites. Bootstrap tests were performed using 500 pseudoreplicates of the ML phylogenies. In addition, relative rate tests were used to compare evolutionary rates of the different gene copies found in *A. pisum* using an outgroup species (tests performed using MEGA3; Kumar et al. 2004). Analyses of the synonymous (dS) and nonsynonymous rates of substitutions (dN) were performed using Codeml (Yang 1997) (ML estimates of pairwise rates). We also used Codeml to evaluate shifts in the ratio of nonsynonymous to synonymous substitutions, in relation to duplication events. The average dN/dS ratio was evaluated using the one-ratio model; a free-ratio model (with a specific ratio on each branch) was then evaluated and compared with the one-ratio model using likelihood ratio tests (LRT).

### Partial Sequencing of *dcr-1* and *ago1* in Different Aphid Species

Partial sequences of *dcr-1* were cloned from other aphid species to evaluate differences in mutation rates in relation to speciation and duplication events. We chose one species from the same genus as the pea aphid (*Acyrtosiphon kondoi*), one species belonging to the same tribe Macrosiphini (*Myzus persicae*), and two species (*Rhopalosiphum padi* and *Aphis gossypii*) belonging to the Aphidini tribe. Genomic DNA from adult parthenogenetic female aphids was extracted using a “salting out” protocol (Sunnucks and Hales 1996). Partial *Ap-dcr-1a* and *Ap-dcr-1b* genomic sequences were amplified simultaneously by polymerase chain reaction (PCR) by using primers that flank the first RNase III domain in the two copies of *Ap-dcr-1* Dic1abF1 (TGGGAGTAAATTCAAACACTGG)/Dic1abR1 (CGATTGGGGTAATAAGAAGCA). Partial *Ap-ago 1a* and *Ap-ago 1b* genomic sequences were amplified by PCR by using Ago1aF1 (TTGCAATTGGAAAATGGT)/Ago1aR1 (GTTGGACATTTAATCCTCCC) and Ago1bF1 (CTGCAAGAAAAACAATG)/Ago1bR1 (GTTGGGCATTTAATTTTT) primers, respectively. All the amplified fragments were cloned in *Escherichia coli* by using the pGemT-easy cloning sys-

tem (Promega). For each cloned fragment, two clones were sequenced in both strands.

### RNA Expression Profiles

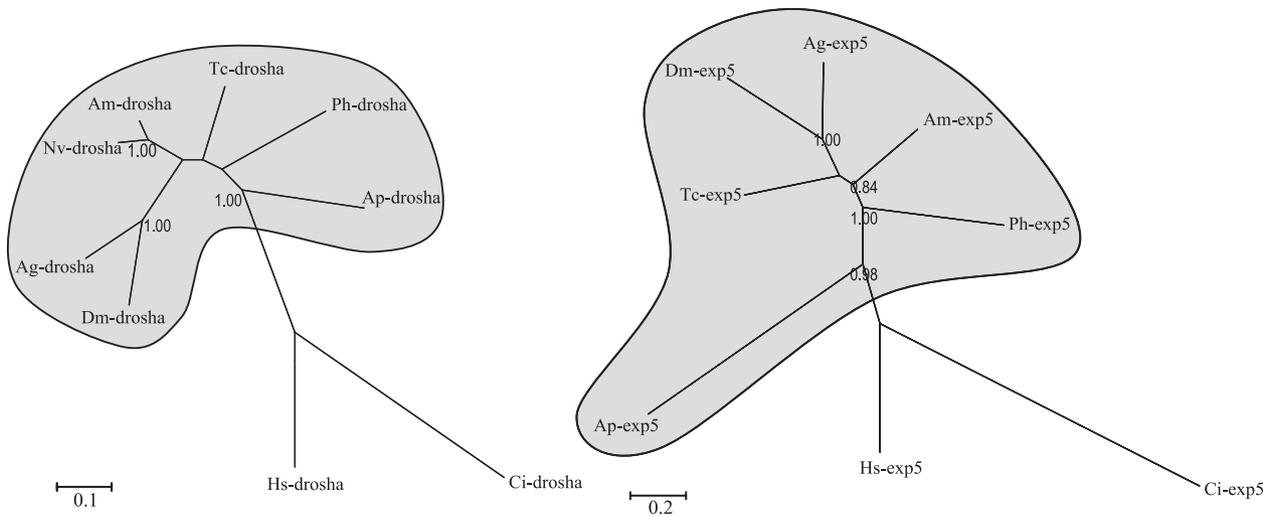
The LSR1-A1-G1 clone of the pea aphid *A. pisum* (International Aphid Genomics Consortium 2010) was reared on broad bean (*Vicia fabae*) at 18 °C. Parthenogenetic reproduction was maintained at 16 h of light, and aphids were reared at low density (one to five individuals per plant). Production of sexual morphs was obtained by rearing aphids at 12 h light for two generations (Le Trionnaire et al. 2007). Total RNA was extracted by using the SV Total RNA Isolation System (Promega) from each of the *A. pisum* morphs: adult parthenogenetic females reared under long-day photoperiod (called virginoparae) and producing parthenogenetic clones, adult parthenogenetic females reared under short-day photoperiod (called sexuparae) and producing males and sexual female clones, adult sexual females, and adult sexual males. The concentration and quality of the extracted RNA was estimated with a NanoDrop (Thermo Scientific). First-strand cDNAs were produced from 500 ng total RNA using the SuperscriptIII Reverse Transcriptase (Invitrogen) and Random Nonamers (Promega) following the supplier's instructions. DNA contamination was removed by treating RNA extraction products with RNase-free DNase (Promega).

The expression of *Ap-dcr-1a* and *Ap-dcr-1b* was investigated by reverse transcriptase (RT)-PCR by using Dic1abF2 (TGGGAGTAAATTCAAACACTGG)/DicR2 (CGATTGGGGTAATAAGAAGCA) and Dic1bF5 (CAGCAGCCAAATGTGCTTTA)/Dic1bR5 (CAATTCCTCTGATCAATCTATTCAAA) PCR primers, respectively. *pasha1–4* expression was analyzed by RT-PCR by using Pas1F2 (TTCTGGAGTATCTGATGATGATG)/Pas1R2 (GCAGTCTCCACTTTGGCATT), Pas2F2 (AAAACAAACCTCACAATGAACA)/Pas2R2 (TCCTTGATGTTTTTAGCAT), Pas3F1 (CTGAAACCGGCAGCTCTAGT)/Pas3R1 (TGATCTTCGGGATTGGATGT) and Pas4F1 (CGACAGCGATGATGAATAC)/Pas4R1 (CATGGCTCAATGTCAAAGGA), respectively. The expression of *Ap-ago 1a* and *Ap-ago-1b* was investigated by RT-PCR by using Ago1aF1 (TTGCAATTGGAAAATGGT)/Ago1aR2 (AGCCATTGCGCCTGGTGTCT) and Ago1bF1 (CTGCAAGAAAAACAATG)/Ago1bR2 (AAGAGTCAACGGTGTGCTAG) primers, respectively.

## Results

### Genes of the siRNA and miRNA Pathways in the Pea Aphid

Aphid orthologs of genes encoding components of the siRNA and miRNA pathways were identified in the *A. pisum* genome and compared with those previously reported in *D. melanogaster*, *A. gambiae*, *T. castaneum*, and *Apis mellifera*. As previously observed for most other insects with sequenced genome, *A. pisum* possesses one copy of each gene in the siRNA pathway: one *dcr-2*, one *r2d2*, and one *ago2*. For the miRNA pathway, one copy of *droscha* and one *exportin-5* copy were found in the pea aphid genome, which is also the situation in other insect species



**Fig. 1.** Phylogenetic trees (ML method on amino acid alignments) for Drosha (right) and Exportin-5 (left) for insect species (Ap: *Acyrtosiphon pisum*, Ph: *Pediculus humanus*, Am: *Apis mellifera*, Nv: *Nasonia vitripennis*, Tc: *Tribolium castaneum*, Ag: *Anopheles gambiae*, Dm: *Drosophila melanogaster*) and two outgroups, an urochordate (Ci: *Ciona intestinalis*) and a vertebrate (Hs: *Homo sapiens*). Insect taxa are shaded in gray. Bootstrap values at the nodes are shown only if greater than 0.80.

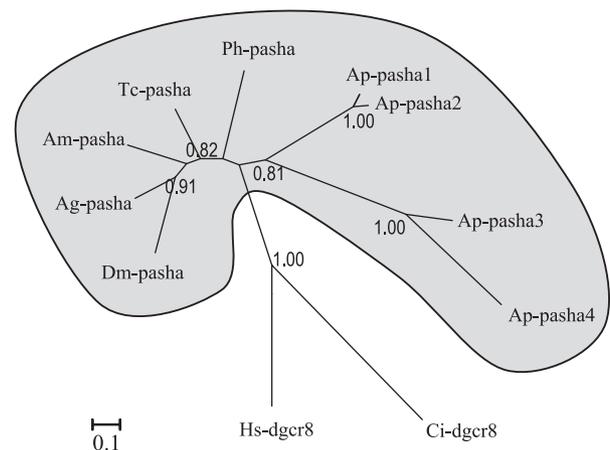
analyzed so far. However, we found duplications of *pasha* (four copies), *dcr-1* (two copies), *loquacious* (two copies), and *ago1* (two copies) in the pea aphid genome (supplementary fig. S1 and table S1, Supplementary Material online). All these copies encode for complete polypeptide sequences except for one copy of the Loquacious gene, which had multiple frameshifts and/or premature stops and may be a pseudogene or encode for a nonfunctional protein.

**Drosha and Exportin-5.** Both *drosha* and *exportin-5* are found in *A. pisum* and are represented by a single copy orthologous to those of other insects (fig. 1). The phylogeny (based on amino acid sequences) of Drosha and Exportin-5 largely parallels the expected species phylogeny. For example, the clustering of the two dipteran species (*A. gambiae* and *D. melanogaster*) and the two hymenopteran species (*Nasonia vitripennis* and *A. mellifera*) are strongly supported. However, whereas we would expect a grouping of *Pediculus humanus* and *A. pisum* (both are paraneopteran), *A. pisum* appears basal in each case. For Exportin-5, a particularly long branch suggests a very high divergence of this protein in aphids (see also below).

**Pasha.** Pasha, also named DGCR8 in vertebrates, is the cofactor of the RNase III Drosha in forming the “microprocessor” multiprotein complex involved in the processing of pri-miRNAs into pre-miRNA. The primary role of Pasha is to recognize the substrate pri-miRNA, whereas Drosha cleaves the pri-miRNA (Han et al. 2006). All the insect species studied so far possess only a single *pasha* copy, but four *pasha-like* genes were identified in *A. pisum*. Phylogenetic analysis shows a rather solidly supported (bootstrap value of 0.81) grouping of the four aphid copies, suggesting that all these copies arose through successive lineage-specific duplications (fig. 2). The particularly long branch for *Ap-pasha3* and *Ap-pasha4* suggests an accelerated divergence of these copies following duplication.

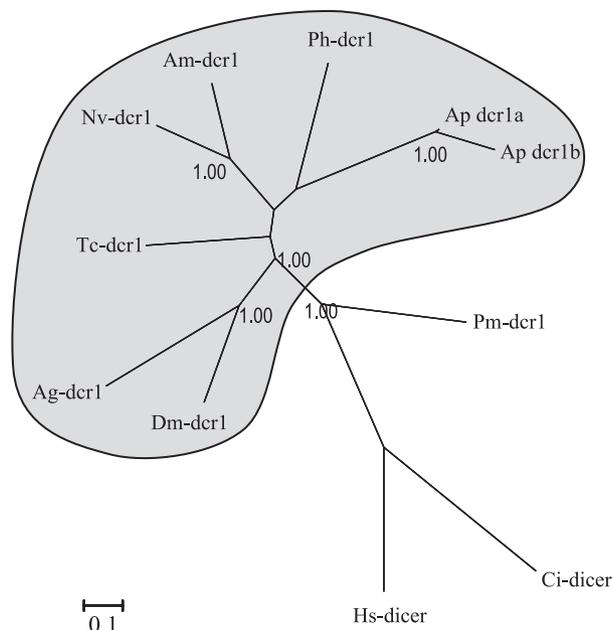
The DGCR8/Pasha proteins are composed of an N-terminal WW domain important for the nuclear localization of the protein and two dsRNA-binding domain crucial for interaction with Drosha and processing of pre-miRNAs (Landthaler et al. 2004). All four aphid Pasha proteins possess these three functional domains (supplementary fig. S2, Supplementary Material online) and are homologous in their C-terminal region. The four proteins differ mainly by the presence or absence of a block of 95 AA at their N-terminal end: this block is repeated three times in Pasha1, is present only once in Pasha2, and is absent from Pasha3 and Pasha4. This part of the sequence has no homology in the Interpro database, and its functional significance is unknown.

**Dicer.** Dicers belong to classIII of RNase III enzymes, involved in the synthesis of small RNA duplexes. All the insects studied so far possess two Dicers: DCR-1 specifically involved in miRNA biosynthesis and DCR-2 responsible



**Fig. 2.** Phylogenetic tree (methods and species suffix as for fig. 1) of Pasha in insects and outgroup species.

Downloaded from <http://mbe.oxfordjournals.org/> by guest on September 22, 2016

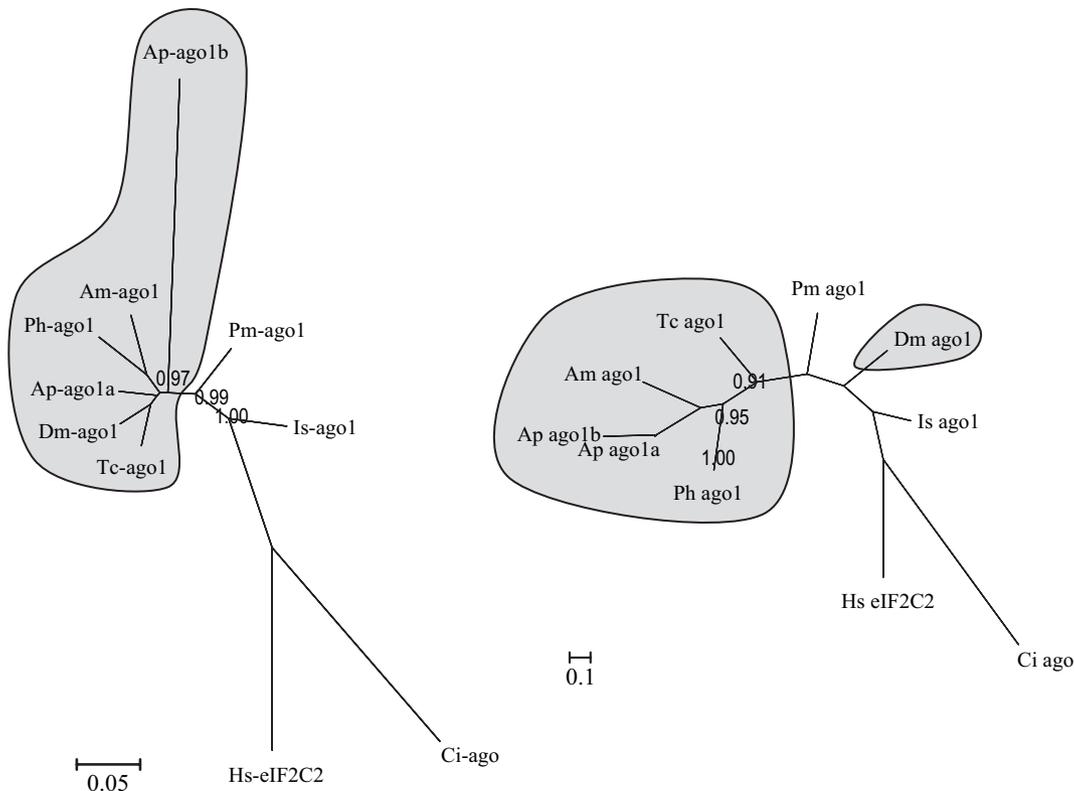


**Fig. 3.** Phylogenetic tree (methods and species suffix as for fig. 1 with in addition: Pm: *Pennaeus monodon*, a shrimp) of Dicer-1 in insects and outgroup species.

for siRNA processing. The phylogenetic reconstruction of Dicer (fig. 3) shows a lineage-specific duplication of *dcr-1* in *A. pisum*. Dicer proteins are composed of several conserved motifs (Du et al. 2008): a N-terminal DExD/H-box helicase domain, a small domain of unknown function (DUF283), a PAZ (Piwi/Ago/Zwille) domain, two tandem RNase III domains (RNase IIIa and IIIb), and a C-terminal dsRNA-binding domain (Du et al. 2008). Both aphid DCR-1 proteins possess these conserved domains (supplementary fig. S3, Supplementary Material online). However, the main difference between the two copies of DCR-1 is an inserted/deleted region of 141 nt that corresponds to 47 AA in the first RNase IIIa domain. RNase III domains are thought to play a major role in the RNase activity of Dicer proteins (Zhang et al. 2004). The occurrence of this deletion is well supported, as there is good trace coverage for each copy over the full region of the deletion (four genomic traces for *dcr-1a* and six traces for *dcr-1b*). Moreover, these two sequences were confirmed by RT-PCR by using primers that flank the area including the first RNase III domains in the two *Dcr-1* copies (supplementary fig. S3, Supplementary Material online). Sequencing of the amplified fragments confirmed the deletion within the first RNase III domain in the *dcr-1b* copy (data not shown). Finally, amplification and sequencing of an ortholog sequence of *dcr-1b* in *A. kondoi* (see below) indicating that the same deletion was present, demonstrating that this duplication of this truncated copy exists in other aphid species. Despite this deletion within the RNase IIIa domain, both *dcr-1* from *A. pisum* possess the four catalytic residues of both RNase III domains (Zhang et al. 2004) (supplementary fig. S3, Supplementary Material online), suggesting that both *dcr-1* are catalytically active. However, the difference within

the RNase IIIa domain might reflect a difference in cleavage activity or small RNA specificity between the two *dcr-1* copies.

**Argonaute.** AGOs belong to a multigene family with multiple paralogs that can be divided into two subgroups: AGO and Piwi. Only the AGO-type proteins contribute to the RISC belonging to the siRNA and miRNA pathways. The other AGOs (Piwi type) are involved in transcriptional silencing and are related to a different type of small ncRNA such as piRNAs. Among the five AGO proteins of *D. melanogaster*, only AGO1 and AGO2 are associated with RISC. AGO1–RISC loads predominantly miRNAs, whereas AGO2–RISC is linked mostly to the siRNA pathway (Kim et al. 2009). AGO proteins are composed of an N-terminal DUF1785 domain of unknown function, a PAZ domain thought to be involved in protein–protein interaction and a C-terminal PIWI domain. The precise function of the PAZ and PIWI domains are not well understood, but these domains are thought to allow the alignment and the stabilization of small RNAs to their target sequences (Murphy et al. 2008). Although most of the insects possess a single *ago1* and a single *ago2* gene, we identified two paralogs of *ago1* and one of *Ago2* in the pea aphid genome. A duplication of *Ago1* has been identified in the mosquito *A. aegypti* (Campbell et al. 2008), but the two sequences are identical at the protein level, suggesting a strong purifying selection on both copies and/or a recent origin for the duplication. By contrast, a phylogeny of the AGO1 protein (fig. 4) shows that the *A. pisum ago1b* copy is highly divergent (long branch) and appears basal to the rest of the insect sequences. A basal duplication of *ago1* would imply several losses in different orders of insects (at least in *P. humanus* and in an ancestor of holometabola), which is nonparsimonious but remains a possible scenario. However, the ML tree based on DNA sequences (fig. 4) shows a strong support for a group formed by the two *A. pisum* copies (supported by both the NJ and ML methods), suggesting instead that they arose by a lineage-specific duplication. It is likely that the accelerated evolution of the second copy may have obscured the phylogenetic reconstruction at the protein level, whereas the DNA-level analysis shows that the two aphid copies are indeed related. By contrast, the DNA-level analysis shows an unexpected positioning of the *D. melanogaster* sequence, which appears in a basal position and more distant from other insect sequences than the sequence of *Pennaeus monodon*, a shrimp. We believe that this is largely explained by the fact that DNA-level analyses are sensitive to compositional bias when comparing distant species (the *Drosophila* genome is more GC rich to that of insects from other orders) and are thus not suited to reconstruct deep events. Finally, the status of *ago1a* and *ago1b* as aphid-specific paralogs was also confirmed by the PhylomeDB database that has been recently built to compare all *A. pisum* genes with other complete genomes through different phylogenetic methods (International Aphid Genomics Consortium 2010; [www.aphidbase.com](http://www.aphidbase.com)).



**Fig. 4.** Phylogenetic tree (at left, methods and species suffix as for fig. 1) of *ago1* in insects and outgroup species (Is: *Ixodes scapularis*, an arthropod). At right: ML tree for a DNA alignment.

**Comparative Evolutionary Rates**

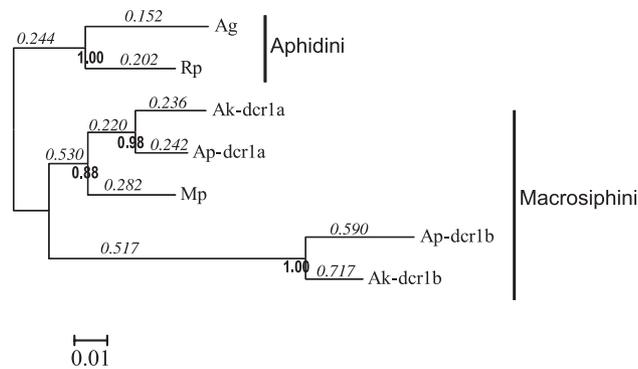
To get some insight into the specific evolutionary pressures acting on the miRNA machinery genes in *A. pisum*, we performed relative rate tests on both the single copy and multiple copy genes (Tajima 1993)—comparing the sequences of *A. pisum*, *D. melanogaster*, and an outgroup species (*H. sapiens*, or, when available, another arthropod, *P. monodon*, table 1). We first confirmed that *exportin-5* has evolved significantly faster in *A. pisum* compared with *D. melanogaster* (similar results were obtained when *A. pisum* was compared with any other insect). This was also the case for the two *pasha* copies, *pasha3* and *pasha4*, for one of the *ago1* copies (*Ap-ago1b*) and for one of the *dcr-1* copies (*Ap-dcr1b*).

To probe further these differences in selective pressures among gene copies, we estimated the ratios of nonsynonymous to synonymous substitutions. For *dcr-1*, two partial sequences were obtained for *A. kondoi* (copy –a and –b, respectively), whereas a single sequence was obtained for three species, *M. persicae*, *R. padi*, and *A. gossypii*. The high level of bootstrap support (0.88) for the grouping of the *M. persicae* sequence with *Ap-dcr-1a* and *Ak-dcr-1a* sequences shows that the duplication almost certainly occurred before the *Myzus/Acyrtosiphon* divergence, that is, it is at least a few tens of million years old (fig. 5). The existing data provide no additional insight as to whether the duplication of *dcr-1* was basal to Aphidini and Macrosiphini

**Table 1.** Results from the Tajima Relative Rate Test for Three Sequences.

	<i>drosha</i>	<i>exportin-5</i>	<i>pasha1</i>	<i>pasha2</i>	<i>pasha3</i>	<i>pasha4</i>	<i>ago1a</i>	<i>ago1b</i>	<i>dcr-1a</i>	<i>dcr-1b</i>
Identical sites in all three sequences ( $m_{iii}$ )	457	208	142	142	119	97	754	617	583	523
Divergent sites in all three sequences ( $m_{ijk}$ )	179	475	132	135	134	157	7	34	342	339
Unique differences in Dm ( $m_{ijj}$ )	67	121	31	30	24	21	18	14	96	89
Unique differences in Ap ( $m_{iji}$ )	68	192	43	43	57	78	8	144	120	161
Unique differences in outgroup ( $m_{ijj}$ )	155	141	91	89	51	45	50	27	143	119
Chi-square statistic	0.01	16.11	1.95	2.32	13.44	32.82	3.85	106.96	2.67	20.74
P	0.931	0.00006	0.163	0.128	0.0002	0.00000	0.05	0.00000	0.0102	0.00001
Significance	NS	***	NS	NS	***	***	NS	***	NS	***

NOTE.—The reference sequence was the ortholog from *Drosophila melanogaster* (Dm)—always a single copy. The column titles are the name of the compared sequence in *Acyrtosiphon pisum* (single copies for *drosha* and *exportin-5* and multiple copies for *pasha*, *ago1* and *dcr-1*). The outgroup species was *H. sapiens* for *Drosha*, *Exportin-5* and *pasha* and *Pennaes monodon* for *ago1* copies and *dcr-1* copies. NS, not significant.



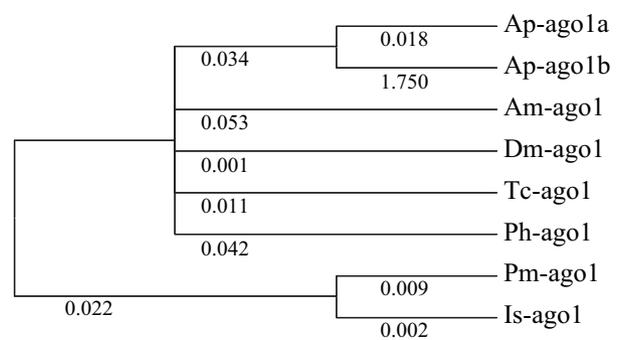
**Fig. 5.** Phylogenetic tree for a DNA alignment including partial sequences of *dcr-1* in different aphid species and showing estimates of the nonsynonymous to synonymous rates (dN/dS) above branches (free-ratio model of Codeml). Bootstrap values are given below the nodes if greater than 0.80. Suffixes indicate species name, Ap: *Acyrtosiphon pisum*, Ak: *Acyrtosiphon kondoi*, Mp: *Myzus persicae*, Rp: *Rhopalosiphum padi*, Ag: *Aphis gossypii*. Ap, Ak, and Mp belong to Macrosiphini, whereas Rp and Ag belong to Aphidini.

(the two tribes represented in the data set) or occurred in an ancestor of Macrosiphini. We evaluated the substitution ratios along a trifurcated tree and found that a free-ratio model was significantly better than a one-ratio model (LRT,  $P = 0.05$ ). The values of dN/dS for the  $-b$  copy in the two *Acyrtosiphon* species are two to three times higher than the  $-a$  copy, suggesting considerable relaxation of selection for that copy.

For *ago1* and *pasha*, only a few partial sequences (usually orthologs to just one of the *A. pisum* copies) could be obtained in other aphids (results not shown), so comparisons could only be made with nonaphid sequences. For *ago1* (fig. 6), a striking increase of the dN/dS ratio, well above 1, again suggested accelerated evolution of rates after duplication. A large majority of substitutions were nonsynonymous, suggesting positive selection on the *ago1b* copy.

### Estimates of the Age of Aphid-Specific Duplications

We estimated the ML pairwise synonymous distances among the duplicated genes of the miRNA machinery in *A. pisum* as a way to evaluate the age of each duplication. The duplication giving rise to *Ap-Pasha1/2* and *Ap-Pasha3/4* appears to be the most ancient (table 2), followed by the duplication separating *Ap-Pasha3* and *Ap-Pasha4*. The duplication separating *Ap-Pasha1* and *Ap-Pasha2* appears to be the most recent. Interestingly, the duplications of both *Dcr-1* and *Ago1* occurred in the same window of evolutionary time as the *Ap-Pasha1/2* duplication. An estimation of the mean synonymous distance among orthologs of different aphid species can be derived from Brisson and Nuzhdin (2008) ( $dS = 0.25$  for *A. pisum*–*M. persicae*,  $dS = 0.35$  for *A. pisum*–*A. gossypii*). Although  $dS$  is not a strict measure of time and is subject to variation among genes, this suggests that the three coincident duplications (*Pasha1/2*, *Dcr-1a/b*, and *Ago1a/b*) probably occurred close to the time of the divergence between *A. pisum* and



**Fig. 6.** Topological tree of *ago1* in different insects and other arthropods, showing estimates of the dN/dS ratio (free-ratio model of Codeml, PAML) on each branch.

*M. persicae*. Finally, we note that these estimates of  $dS$  are consistent with the tree topologies and branch lengths in the phylogenies described above for the different genes.

### Expression Profiles

RT-PCR was performed on RNA extracted from whole bodies of adult parthenogenetic virginoparae females (produced under long-day photoperiod), adult parthenogenetic sexualoparae females (produced under short-day photoperiod), sexual females, and males (supplementary fig. S4, Supplementary Material online). The four copies of *pasha*, the two *ago1* as well as the two *dcr-1* copies, were expressed in all four morphs. Some differences in expression patterns between sexuals and asexuals were observed for *dcr-1b* that need to be quantitatively analyzed further on a tissue-specific basis. Altogether, this demonstrates that the different gene copies are expressed at the transcript level in all morphs examined, a first necessary step for the functional characterization of this set of gene copies.

### Discussion

In this paper, we report a comprehensive analysis of the genes encoding components of the pea aphid small ncRNA pathways, which surprisingly revealed that the miRNA machinery is duplicated in aphids. In the basal metazoan Placozoans, duplications of Dicer genes (both *dcr-1* and *dcr-2*) were recently reported (de Jong et al. 2009). However, the genome appears to lack an important gene in the miRNA pathway (*Pasha*) and may not be able to produce miRNAs at all, suggesting that the lineage-specific duplication of Dicer in this taxon could be involved in defence against viruses. We compared also the evolution of the miRNA machinery genes in aphids and other insects with sequenced

**Table 2.** ML Pairwise Estimates of Synonymous Distances among Duplicated Genes of the miRNA Machinery in *Acyrtosiphon pisum*.

	dS
Pasha (1/2 vs. 3/4)	2.145
Pasha (3 vs. 4)	0.744
Pasha (1 vs. 2)	0.192
dicer1 (a vs. b)	0.198
ago1 (a vs. b)	0.159

NOTE.—For *Pasha1/2* versus 3/4, mean of all four  $dS$  estimates among copies one or two and copies three or four.

genome in order to determine whether the expansion of miRNA machinery genes occurred early during the evolution of the aphid family.

### Evolution of the miRNA Machinery

The common pattern to emerge from our analysis of the evolution of pea aphid miRNA genes is one of striking acceleration of the rate of substitutions following duplication, with the acceleration being concentrated in one copy of the gene. Phylogenetic and distance analysis indicates that the three major duplication events in the pea aphid miRNA machinery (*pasha1* vs. *pasha2*, *dcr-1a* vs. *dcr-1b*, and *ago1a* vs. *ago1b*) occurred in approximately the same time window.

Gene duplication has been increasingly seen as a source of evolutionary novelty. However, the evolutionary fate of specific duplicates is variable. The most frequent fate would be the elimination of one copy; alternatively, the two copies may be retained, whereas one or both would acquire differentiated profiles of expressions or even different functions. Strikingly, we have found that for each of these three genes, one of the copies diverged in an accelerated way, whereas the other remained more conserved. In one case (*ago1b*), the evolution of the more divergent copy seems even to be driven by positive selection. This suggests the acquisition of a new function for AGO1B, which remains to be confirmed. In the nematode *C. elegans*, the expansion of the AGO protein family has been linked to subfunctionalization (distinct types of AGO proteins act sequentially in the different step of the exo-siRNA or endo-siRNA pathways) (Yigit et al. 2006) or neofunctionalization (like the AGO protein nuclear RNAi defective-3 (NRDE-3) that transports specific classes of small regulatory RNAs to distinct cellular compartments to regulate gene expression) (Guang et al. 2008). The expansion of a part of the miRNA machinery in the pea aphid, and in particular of *ago1*, could be linked to a similar subfunctionalization and/or neofunctionalization processes. In a second case (*dcr-1*), a striking difference in the structure of the protein has been found between DCR-1A and DCR-1B, affecting a functional domain that could reflect a difference in activity. Recent work in *Drosophila* has shown that novel miRNAs arise by accumulation of nucleotide mutations in non-miRNA transcripts (Lu, Fu, et al. 2008; Lu, Shen, et al. 2008), which also appears to be the case in vertebrates (Liu et al. 2008). Very few new miRNAs are paralogous to existing miRNAs, and none are derived by inversion of duplications (Lu, Shen, et al. 2008). It is possible that by having an additional, modified Dicer associated with a duplicated miRNA pathway, the pea aphid may have an enhanced ability among insects to experiment with generating miRNAs from new transcripts, that is, to explore the miRNA evolutionary space as effectively as it can explore the gene evolutionary space through duplication/amplification. Evidence supporting this hypothesis could be obtained by searching the aphid genome for aphid-specific miRNAs that appear to be derived from recently inverted hairpins or from complementary strand transcripts. Additional insights might be obtained by predicting the effect of the 47

AA insertion in DCR-1B on its function, in particular whether it would affect its ability to cleave perfect (bulge free) hairpins that might release new miRNAs. Investigating the function of this novel Dicer (DCR-1B) offers an exciting new challenge in aphid biology.

The analysis of the gene repertoire of *A. pisum* has shown that its genome shows a striking excess of duplicated genes when compared with other insects with sequenced genome (bee and *Drosophila*), especially with respect to recent to moderately recent duplications (International Aphid Genomics Consortium 2010). This is raising the possibility that the duplication of the miRNA machinery could be attributed to a genome-wide increased level of duplication. Yet, the distribution of paralogs in *A. pisum* shows that most of them are eliminated over time, a pattern observed in every genome. Therefore, the fact that duplicates have been conserved for three genes involved in the miRNA machinery over a long evolutionary period of time is intriguing and does suggest that the new copies have retained functionality and probably acquired a specificity of function that remains to be exactly deciphered. Given their phylogenetic position, and based on distance analyses among copies, these duplications are clearly posterior to the acquisition of polyphenisms common to all aphids (reproductive mode polyphenism and winged/wingless polyphenism), which are basal to the group.

### Conclusions

In this article, we described an expansion of a part of the miRNA machinery in the pea aphid, *A. pisum*. We showed a rapid divergence for three of these genes and for *Ago1b* a positive selection. This expansion in aphids of important proteins of the miRNA pathway may reflect an expansion of the functions of the miRNA pathway in this organism. The next challenge will be to understand the biological significance of such an expansion in aphid biology, which was shown clearly to be posterior to the acquisition of polyphenism by aphids. A deep functional analysis is now required to propose hypotheses for the role of such an expansion in aphid biology. This should include analysis of gene expression by real-time PCR and in situ hybridization, inhibition of gene expression by RNAi, or identification of the small ncRNAs associated with proteins encoded by expanded miRNA machinery genes.

### Supplementary Material

Supplementary materials are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

### Acknowledgments

Mme Agnès Méreau (Center National de la Recherche Scientifique, Rennes) is acknowledged for discussion. This work has been funded by INRA-Santé des Planks et environnement and INRA “Projets Communs INRA/INRIA.”

### References

Amaral PP, Dinger ME, Mercer TR, Mattick JS. 2008. The eukaryotic genome as an RNA machine. *Science* 319:1787–2178.

- Brisson JA, Nuzhdin SV. 2008. Rarity of males in pea aphids results in mutational decay. *Science* 319:58.
- Campbell CL, Black WC, Hess AM, Foy BD. 2008. Comparative genomics of small RNA regulatory pathway components in vector mosquitoes. *BMC Genomics*. 9:425.
- Czech B, Malone CD, Zhou R, et al. (12 co-authors). 2008. An endogenous small interfering RNA pathway in *Drosophila*. *Nature* 453:798–802.
- de Jong D, Eitel M, Jakob W, Osigus HJ, Hadrys H, DeSalle R, Schierwater B. 2009. Multiple *Dicer* genes in the early-diverging metazoa. *Mol Biol Evol*. 26:1333–1340.
- Denli AM, Tops BBJ, Plasterk RHA, Ketting RF, Hannon GJAM, Tops BBJ, Plasterk RHA, Ketting RF, Hannon GJ. 2004. Processing of primary microRNAs by the microprocessor complex. *Nature* 432:231–235.
- Du Z, Lee JK, Tjhen R, Strould RM, James TL. 2008. Structural and biochemical insights into the dicing mechanism of mouse *Dicer*: a conserved lysine is critical for dsRNA cleavage. *Proc Natl Acad Sci USA*. 105:2391–2396.
- Finn RD, Tate J, Mistry J, et al. (11 co-authors). 2008. The Pfam protein families database. *Nucleic Acids Res*. 36:D281–D288.
- Foerstemann K, Horwich MD, Wee LM, Tomari Y, Zamore PD. 2007. *Drosophila* microRNAs are sorted into functionally distinct argonaute complexes after production by *Dicer*-1. *Cell* 130:287–297.
- Ghildiyal M, Zamore PD. 2009. Small silencing RNAs: an expanding universe. *Nat Rev Genet*. 10:94–108.
- Guang S, Bochner AF, Pavelec DM, Burkhardt KB, Harding S, Lachowicz J, Kennedy S. 2008. An Argonaute transports siRNAs from the cytoplasm to the nucleus. *Science* 321:537–541.
- Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 52:696–704.
- Han JJ, Lee Y, Yeom KH, Nam JW, Heo I, Rhee JK, Sohn SY, Cho YJ, Zhang BT, Kim VN. 2006. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* 125:887–901.
- Honeybee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* 443:931–949.
- Hunter S, Apweiler R, Attwood TK, et al. (38 co-authors). 2009. InterPro: the integrative protein signature database. *Nucleic Acids Res*. 37:D211–D215.
- International Aphid Genomics Consortium. Forthcoming. 2010. The genome of the pea aphid *Acyrtosiphon pisum*.
- International Silkworm Genome Consortium. 2008. The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. *Insect Biochem Mol Biol*. 38:1036–1045.
- Jaubert S, Méreau A, Antoniewski C, Tagu D. 2007. MicroRNAs in *Drosophila*: the magic wand to enter the chamber of secrets? *Biochimie*. 89:1211–1220.
- Kim VN, Han J, Siomi MC. 2009. Biogenesis of small RNAs in animals. *Nat Rev Mol Cell Biol*. 10:126–139.
- Kumar S, Tamura K, Nei M. 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform*. 5:150–163.
- Landthaler M, Yalcin A, Tuschl T. 2004. The human DiGeorge syndrome critical region gene 8 and its *D. melanogaster* homolog are required for miRNA biogenesis. *Curr Biol*. 14:2162–2167.
- Le Trionnaire G, Hardie J, Jaubert-Possamai S, Simon J-C, Tagu D. 2008. Shifting from asexual to sexual reproduction in aphids: physiological and developmental aspects. *Biol Cell*. 100:441–451.
- Le Trionnaire G, Jaubert S, Sabater-Munoz B, Benedetto A, Bonhomme J, Prunier-Leterme N, Martinez-Torres D, Simon JC, Tagu D. 2007. Seasonal photoperiodism regulates the expression of cuticular and signalling protein genes in the pea aphid. *Insect Biochem Mol Biol*. 37:1094–1102.
- Lee YS, Nakahara K, Pham JW, Kim K, He Z, Sontheimer EJ, Carthew RW. 2004. Distinct roles for *Drosophila* *Dicer*-1 and *Dicer*-2 in the siRNA/miRNA silencing pathways. *Cell* 117:69–81.
- Liu N, Okamura K, Tyler DM, Phillips MD, Chung WJ, Lai EC. 2008. The evolution and functional diversification of animal microRNA genes. *Cell Res*. 18:985–996.
- Lu J, Fu Y, Kumar S, Shen Y, Zeng K, Xu A, Carthew R, Wu CI. 2008. Adaptive evolution of newly emerged micro-RNA genes in *Drosophila*. *Mol Biol Evol*. 25:929–938.
- Lu J, Shen Y, Wu Q, Kumar S, He B, Shi S, Carthew RW, Wang SM, Wu CI. 2008. The birth and death of microRNA genes in *Drosophila*. *Nat Genet*. 40:351–355.
- Meister G, Tuschl T. 2004. Mechanisms of gene silencing by double-stranded RNA. *Nature* 431:343.
- Murphy D, Dancis B, Brown JR. 2008. The evolution of core proteins involved in microRNA biogenesis. *BMC Evol Biol*. 8:92.
- Notredame C, Higgins D, Heringa J. 2000. T-Coffee: a novel method for multiple sequence alignments. *J Mol Evol*. 302:205–217.
- Obbard DJ, Gordon KH, Buck AH, Jiggins FM. 2009. The evolution of RNAi as a defence against viruses and transposable elements. *Philos Trans R Soc Lond B Biol Sci*. 364:99–115.
- Okamura K, Chung WJ, Ruby JG, Guo H, Bartel DP, Lai EC. 2008. The *Drosophila* hairpin RNA pathway generates endogenous short interfering RNAs. *Nature* 453:803–806.
- Peters L, Meister G. 2007. Argonaute proteins: mediators of RNA silencing. *Mol Cell*. 26:611–623.
- Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14:817–818.
- Sunnucks P, Hales DF. 1996. Numerous transposed sequences of mitochondrial cytochrome oxidase I-II in aphids of the genus *Sitobion* (Hemiptera: Aphididae). *Mol Biol Evol*. 13:510–524.
- Tajima F. 1993. Statistical analysis of DNA polymorphism. *Jpn J Genet*. 68:567–595.
- Tomoyasu Y, Miller SC, Tomita S, Schoppmeier M, Grossmann D, Bucher G. 2008. Exploring systemic RNA interference in insects: a genome-wide survey for RNAi genes in *Tribolium*. *Genome Biol*. 9:R10.
- Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*. 13:555–556.
- Yigit E, Batista PJ, Bei Y, Pang KM, Chen CC, Tolia NH, Joshua-Tor L, Mitani S, Simard MJ, Mello CC. 2006. Analysis of the *C. elegans* Argonaute family reveals that distinct Argonautes act sequentially during RNAi. *Cell* 127:747–757.
- Zhang HD, Kolb FA, Jaskiewicz L, Westhof E, Filipowicz W. 2004. Single processing center models for human *Dicer* and bacterial RNase III. *Cell* 118:57–68.