Selection following gene duplication shapes recent genome evolution in the pea aphid *Acyrthosiphon pisum*

Rosa Fernández[*1,2], Marina Marcet-Houben[*1,2], Fabrice Legeai[3,4], Gautier Richard[3,5], Stéphanie Robin[3,6], Valentin Wucher[7], Cinta Pegueroles[–1,2], Toni Gabaldón[–1,2,8,9], Denis Tagu[–,3]

[1] Bioinformatics and Genomics Unit, Center for Genomic Regulation, Carrer del Dr. Aiguader 88, 08003 Barcelona, Spain.
[2] Current address: Department of Life Sciences, Barcelona Supercomputing Center, Carrer de Jordi Girona 29, 08034 Barcelona, Spain.
[3] IGEPP, INRA, Agrocampus Ouest, Université de Rennes 1, 35653 Le Rheu, France
[4] INRIA, IRISA, Genscale, Campus Beaulieu, Rennes, France.
[5] Max Planck Institute of Immunobiology and Epigenetics, Stübeweg 51, 79108 Freiburg im Breisgau, Germany
[6] INRIA, IRISA, GenOuest Core Facility, Campus Beaulieu, Rennes, France
[7] Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Dr. Aiguader 88, 08003 Barcelona, Spain.
[8] Universitat Pompeu Fabra. 08003 Barcelona, Spain. [9] Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010 Barcelona, Spain.

[*]Contributed equally
[–]Contributed equally and share correspondence
Corresponding authors: CP, cintapq@gmail.com, TG, toni.gabaldon.bcn@gmail.com, and DT, denis.tagu@inra.fr

Running title: Gene duplications and adaptation in the pea aphid

Abstract

Ecology of insects is as wide as their diversity, which reflects their high capacity of adaptation in most of the environments of our planet. Aphids, with over 4,000 species, have developed a series of adaptations including a high phenotypic plasticity, and the ability to feed on the phloem-sap of plants, which is enriched in sugars derived from photosynthesis. Recent analyses of aphid genomes have indicated a high level of shared ancestral gene duplications that might represent a basis for genetic innovation and broad adaptations. In addition, there is a large number of recent, species-specific gene duplications whose role in adaptation remains poorly understood. Here, we tested whether duplicates specific to the pea aphid *Acyrthosiphon pisum* are related to genomic innovation by combining comparative genomics, transcriptomics, and chromatin accessibility analyses. Consistent with large levels of neofunctionalization, we found that most of the duplicated genes evolved asymmetrically, showing different positive selection and gene expression profiles. Genes under selection involved a plethora of biological functions, suggesting that neofunctionalization, tissue specificity and other evolutionary mechanisms have orchestrated the evolution of recent paralogs in the pea aphid.

Keywords (4 to 6)

Chromatin; FAIRE-Seq; Insect; Neofunctionalization; Phylogenomics; Positive selection

Introduction

Aphids are insect pests belonging to the order Hemiptera, which diverged some 280-250 million years ago. They feed exclusively on plant phloem sap, a trait that involved specific adaptations such as an obligatory symbiosis with bacteria of the genus *Buchnera,* which supplies aphids with essential amino acids that are missing in the phloem sap. In addition, to adapt to stressful environments such as cold, predation and parasitism (Vellichirammal *et al.* 2016), aphids have developed several plastic phenotypic traits, involving winged and apterous morphs, or sexual oviparous and parthenogenetic viviparous morphs. Although several studies have tried to unravel the genetic mechanisms of these adaptations at the molecular level, the evolutionary forces underlying these genomic changes are still poorly understood. Today, several aphid genomes are publically available and all show a high level of gene duplication and expansions (The International Aphid Genomics Consortium 2010; Mathers *et al.* 2017). Some of these

duplications are shared between aphid species, but most of them are lineage-specific (IAGC 2010). The mechanisms of gene duplication occurring in these species are not yet fully determined. Whether or not duplicated or expanded gene families are in relation with the above-mentioned or other functional innovations enabling adaptive evolution in aphids is still largely unknown (Huerta-Cepas *et al.* 2010b; Simon *et al.* 2011).

There are at least four different outcomes for gene duplicates (reviewed in Innan & Kondrashov 2010). First, while one duplicate keeps the original function, the other acquires a new function (neofunctionalization). Second, each of the two duplicated genes keeps part of the functions of the ancestral gene, so that they jointly cover the original functions (subfunctionalization). Third, when the increase in gene dosage is beneficial, the two copies are maintained in the absence of functional divergence. And fourth, the most common output of gene duplication is the inactivation by accumulation of mutations of one of the duplicated genes (pseudogenization). Several evolutionary forces can drive these different outcomes, for instance relaxed selection for subfunctionalization, purifying selection for neofunctionalization or deleterious mutations for pseudogenization (Lynch & Conery 2000; Han *et al.* 2009; Innan & Kondrashov 2010). These different scenarios can be addressed by scrutinizing patterns of variation of gene families including lineage-specific duplications (Han *et al.* 2009; Innan & Kondrashov 2010; Pegueroles *et al.* 2013; Pich i Roselló & Kondrashov 2014).

More recently, sub- or neofunctionalization have started to be assessed by epigenetic regulation (Robin and Riggs 2003). Acquiring and losing functions can occur, among other means, by modification of chromatin states that drives transcriptional activities of genes. The so-called 'open chromatin', in which accessible DNA allows for active transcription, can be opposed to the so-called ''closed' chromatin, which is compact and transcriptionally repressed. Little is known about the role of chromatin in determining the fate of duplicated genes, but it is intuitive to think that two duplicated gene copies could have spatially or temporally different chromatin states, thus resulting in different transcription patterns. For instance, (Keller & Yi 2014) showed that the DNA methylation of gene promoters of both copies of young duplicates in humans is higher than that of old duplicates. This observation stands for different tested tissues, indicating that this trait is not related to tissue-specificity regulation, as DNA methylation is known to regulate transcription. Thus, it could be hypothesized that chromatin state influences the expression of duplicated copies - and thus consequently their evolution - possibly as a protection against possible misregulations by dosage compensation (Chang & Liao 2012), before mutations occur and the genetic selection operates.

Here, we test the hypothesis that gene duplication - particularly recent duplicates - in the pea aphid *Acyrthosiphon pisum* is a source of innovation fueled by selection. For this, we anchor our study on a phylogenomic approach exploring for the first time ten hemipteran genomes, including six aphid species. We show that (i) a large proportion of gene duplications are under positive selection in *A. pisum* and affect a large number of biological functions (most notably immunological response, reproduction, morphogenesis, dosage compensation and resistance to insecticides), (ii) asymmetrical rates of young paralogs coupled to positive selection suggest neofunctionalization is a main force reshaping the pea aphid genome, (iii) a third of young duplicates show divergent tissue expression patterns, consistent in some cases with subfunctionalization by tissue specialization, and (iv) chromatin accessibility of the transcription start site (TSS) can change between genes in duplicated gene pairs, although it cannot directly explain their transcriptional state in *A. pisum*.

Material and Methods

1. *Identification and selection of duplications in the pea aphid genome*

The phylome (*i.e.*, the complete collection of phylogenetic trees for each gene in its genome) of *A. pisum* Mordvilko, 1914 was reconstructed in the context of Hemiptera evolution. In addition to this species, belonging to the suborder Sternorrhyncha and to the family Aphididae and tribe Macrosiphini, we selected representatives of several hemipterans based on phylogenetic position and availability of a fully-sequenced genome: *Diaphorina citri* Kuwayama, 1908 (Sternorrhyncha, Psylloidea), *Bemisia tabaci* (Gennadius, 1889) (Sternorrhyncha, Aleyrodoidea), *Daktulosphaira vitifoliae* (Fitch, 1855) (Sternorrhyncha, Phylloxeridae), *Cinara cedri* (Curtis, 1835) (Sternorrhyncha, Aphidoidea), *Diuraphis noxia* (Kurdjumov, 1913) (Sternorrhyncha, Aphidoidea), *Aphis glycines* Matsumara, 1917 (Sternorrhyncha, Aphidoidea), *Myzus persica*e (Sulzer, 1776) (Sternorrhyncha, Aphidoidea) and *Rhopalosiphum padi* (Stal, Linnaeus, 1758) (Aphidinae, Aphidini) (Fig. 1). Genome versions and number of predicted proteins are indicated in **Table S1**.

Phylomes were reconstructed using the PhylomeDB pipeline (Huerta-Cepas *et al.* 2011a). For each protein encoded in the *A. pisum* genome, a BLAST search was performed against the custom proteome database built from the genomes listed above. Results were filtered using an e-value of 1e-05 and a minimum overlapping region of 0.5. Multiple sequence alignments were reconstructed in both directions using three different programs (MUSCLE v3.8

(Edgar 2004), MAFFT v6.712b (Katoh 2005), and Kalign (Lassmann & Sonnhammer 2005)) and combined using M-COFFEE (Wallace *et al.* 2006). A trimming step was performed using trimAl v1.3 (Capella-Gutierrez *et al.* 2009), consistency-score cutoff = 0.1667 and gap-score cutoff = 0.9). Following model selection, the best model in terms of likelihood as selected by the Akaike Information Criterion (AIC) was chosen for tree reconstruction. Phylogenetic trees were inferred using PhyML v3.0 (Guindon *et al.* 2010). Four rate categories were used and invariant positions were inferred from the data. Branch support was computed using an aLRT (approximate likelihood ratio test) based on a chi-square distribution. Resulting trees and alignments are stored in PhylomeDB 4.0 (Huerta-Cepas *et al.* 2014), http://phylomedb.org).

A species-overlap algorithm, as implemented in ETE v3 (Huerta-Cepas *et al.* 2010a) was used to infer orthology and paralogy relationships from the phylogenetic trees reconstructed in the phylome. The algorithm scans the tree and calls speciation or duplication events at internal nodes based on the presence of common species at both daughter partitions defined by the node. Gene gains and losses were calculated on this basis. Duplication ratios per node were calculated by dividing the number of duplications observed in each node by the total number of gene trees containing that node: theoretically, a value of 0 would indicate no duplication, a value of 1 an average of one duplication per gene in the genome, and >1 multiple duplications per gene and node.

To build the species tree, one-to-one orthologs present in all species were selected, resulting in a final alignment with 1,047 genes and 635,610 amino acid positions after concatenation. To ensure a congruent phylogenetic hypothesis under different models, a series of approaches were followed to infer the species tree. First, an ML tree was reconstructed with PhyML under the best selected model of amino acid evolution (LG; Le *et al.* 2008). Second, a supertree was reconstructed using DupTree (Wehe *et al.* 2008) based on all the trees reconstructed in the phylome. Both phylogenies were congruent (**Fig. 1**).

2. Detection and selection of gene duplications

For each gene tree, we first selected with ETE v3 (Huerta-Cepas *et al.* 2010a) the nodes that exclusively contained multiple *A. pisum* sequences. These were considered species specific duplications in *A. pisum*. Overlapping species specific duplications were fused when they shared more than 50% of their members. Only families with at least 2 single-copy orthologs were kept so as to ensure that duplications were indeed lineage-specific. Trees were then further scanned for the presence of, at least, two single-copy orthologs. Species specific duplicated genes and

selected orthologs were grouped and used to build a second ML tree. The purpose of this tree was to ensure that the resulting topology still contained the species specific duplication. Pairs of duplicates with incongruent CDS annotation, unsatisfactory topology and overlapping with larger families were discarded. This resulted in a final number of scrutinized duplications of 572. For each duplication, we obtained multiple protein sequence alignments with PASTA v1.8.3 (Mirarab *et al.* 2015), estimated median similarity for each protein sequence in the alignment using trimAl v1.3 (-sident option; Capella-Gutierrez *et al.* 2009) and back-translated into nucleotidic multiple sequence alignments with trimAl.

3. Age of the selected duplications and classification into fast and slow copies

The relative age of the selected duplications was calculated using the number of synonymous substitutions per synonymous site (dS) as a proxy. dS, the number of non-synonymous substitutions per non-synonymous site (dN) and dN/dS ratio were estimated using the "free ratio branch model" implemented in codeML from PAML v. 4.9 (Yang 2007), using model = 1, CodonFreq = 3, Nsites = 0 as options. This software allows to estimate dS, dN and dN/dS for each internal and terminal branch of a given tree. Analyses were computed for the 572 selected duplications and subsequently duplications with dS >2 (which may indicate problems in the orthology identification, 64 duplications) and dS < 0.01 (which may lead to high dN/dS ratios with no biological sense, 390 duplications) were filtered out. A total of 166 duplications remained after the filtering for age inference analysis. We also used the dS estimates to classify the two copies of each selected duplication into fast and slow, by comparing their dS values, the copy with the lowest dS value being classified as slow and the other as fast.

4. Positive selection tests

We tested for positive selection using the "branch-site" test 2 implemented in codeML from PAML v.4.9 (Yang 2007). We compared the null hypothesis where dN/dS is fixed in all branches (model = 2, NSsites = 2, fix_omega = 1, omega = 1) and the alternative hypothesis where the branch that is being tested for positive selection may include codons evolving at dN/dS>1 (model = 2, NSsites = 2, fix_omega = 0, omega = 1.5). The two models were compared using a likelihood ratio test (LRT) and p-values were adjusted for multiple comparisons using the Bonferroni method.

5*. Functional annotation and GO term enrichment analysis and visualization*

To assign Gene Ontology (GO) terms to the genes in the pea aphid genome, GO terms based on orthology relationship were propagated with eggNOG-mapper (Huerta-Cepas *et al.* 2017). For that, we selected the eukaryotic eggNOG database (euNOG, (Huerta-Cepas *et al.* 2019)) and prioritised coverage (*i.e.*, GO terms were propagated if any type of orthologs to a gene in a genome were detected). Functional enrichment of the selected duplications as well as genes under positive selection was explored with FatiGO (Al-Shahrour *et al.* 2004). We tested enrichment against two different backgrounds: all the genome and the remaining genes in the genome (*i.e.*, non-expanded genes and non-positively selected ones, respectively). In addition, enrichment of GO terms expressed in the different tissues (see below) was explored as well, following the same steps. Sets of GO terms were summarized and visualized in REVIGO (Supek *et al.* 2011).

6. Tissue expression diverge between duplicates

Messenger RNA (mRNA) expression data was obtained from 106 different samples from the *A. pisum* LSR1 lineage (The International Aphid Genomics Consortium 2010). We obtained RNA-Seq libraries from 18 different conditions. Some of them were retrieved from the public databases and others newly generated for this study (**Table S2**). These were sequenced using Illumina technology as paired-end of 100 bp size, containing more than 25 million raw reads per library. Reads from all the RNA libraries were mapped on the version 2 of the pea aphid genome assembly (Acyr_2.0, ID NCBI : 246238) using STAR version 2.5.2a (Dobin *et al.* 2013) with the default parameters except the following parameters: outFilterMultimapNmax = 5, outFilterMismatchNmax = 3, alignIntronMin = 10, alignIntronMax = 50000 and alignMatesGapMax = 50000. The number of reads covering each gene prediction (NCBI Annotation release ID: 102) was then counted using FeatureCounts version 1.5.0-p3 (Liao *et al.* 2014) with the default parameters except the following parameters : -g gene -C -p -M --fraction. For each counting, RPKM calculation was performed using edgeR (Robinson *et al.* 2010; McCarthy *et al.* 2012) with gene.length = sum of exons size for each gene.

RNA-Seq values for each individual gene were divided into four quartiles. Each RNA-Seq experiment was processed independently. Replicates were then joined by collapsing the different values obtained in the different experiments of the same tissue. If more than 50% of the experiments placed the RNA-Seq data into the same, this bin was assigned to the overall tissue. On the other hand, if none of the bins had enough representation across experiments, no value

was assigned (NA). Once each tissue was assigned a value, a profile was created for each individual gene. The profiles consisted of 0 and 1 in which 0 represented not-expressed and were values located in the lowest of the four bins. 1 represented expressed genes and consisted of values located in the other three bins. These expression profiles were used to calculate the tissue expression divergence between pairs of duplicates using three different methods: i) Normalized Hamming distance, which counts the number of differences between two profiles and divides it by the total number of considered tissues. A tissue is not considered when it's value is NA for either gene. ii) Tissue expression complementarity (TEC) distance (Huerta-Cepas *et al.* 2011b), which compares the relative number of tissues in which only one set but not the other was expressed over the total number of tissues in which each gene is expressed. iii) Tissue expression divergence (dT) (Pegueroles *et al.* 2013), which subtracts tissues were one or the two copies are expressed from tissues were the two copies are expressed divided by tissues were one or the two copies are expressed. Values for the three distances range from 0 to 1, where 0 means no differences in gene expressions between duplicates (in other words, the two copies tend to be expressed in the same tissues) and 1 means that the two copies have totally different expression patterns.

7. *FAIRE-Seq data analysis*

FAIRE-Seq data for samples for males and females adults was taken from (Richard *et al.* 2017). FAIRE-Seq samples for embryos were newly generated for another, unpublished, study (Richard 2017). Subsequently to sequencing, FAIRE and Control reads were mapped using bowtie2 with default parameters (Langmead *et al.* 2009; Langmead & Salzberg 2012) on the pea aphid genome assembly v2.1 (AphidBase: http://bipaa.genouest.org/is/aphidbase/acyrthosiphon_pisum/). Only uniquely mapped reads with a mapping quality over or equal to 30 in the phred scale were kept using SAMtools (Li *et al.* 2009), following the IDR recommendations (Li *et al.* 2011), https://sites.google.com/site/anshulkundaje/projects/idr/deprecated). MACS2 (Zhang *et al.* 2008) was used to perform the peak calling with the following parameters using control samples: --gsize 541675471 --nomodel --extsize 500 -p 0.05 --keep-dup all -f BEDPE, followed by IDR analyses using a threshold of 0.01 for original replicates, of 0.02 for self-consistency replicates and of 0.0025 for pooled pseudoreplicates. Replicates consistency was then assessed using the Irreproducible Discovery Rate (IDR) algorithm (Li *et al.* 2011) and the two most correlated FAIRE replicates out of the three in each condition were pooled in order to reduce the noise, as widely recommended for ChIP-Seq or ATAC-Seq data. Input-normalized FAIRE-Seq signals were

calculated using deepTools2 `*bamCompare*` (Ramírez *et al.* 2016) across the whole genome for each condition by calculating the average log2 (Pooled FAIRE/Input) in windows of 10 bp. Both Pooled FAIRE and Input read counts were normalized by sequencing depth using `*--normalizeTo1x*`. Using deepTools2 `*multiBigwigsummary*`, the average FAIRE signal was extracted 900 bp around the beginning of genes (450 bp in 5' and 450 bp in 3') in all samples. We then used a threshold of 1 for the average log2 (FAIRE/Input) to define genes whose TSS is open (above the threshold) or closed (below the threshold).

Embryos and adults RNA-Seq data were related to the FAIRE-Seq data for each condition and individual gene. According to the data, genes were classified in four categories: (i) open and expressed, (ii) open and not expressed, (iii) closed and expressed and (iv) closed and not expressed. Each gene expressed in at least one embryo or adult sample was assigned to one of the four categories by averaging the number of tissues in each category. 13,858 genes were assigned to one of the categories.

Results and Discussion

1. Young paralogs in *A. pisum* are under neofunctionalization and involve diverse biological functions

We built a phylome (*i.e.* the complete collection of phylogenetic trees for each gene encoded in a genome) for *A. pisum* in the context of hemipteran evolution, including five additional aphid species and three basal Sternorrhyncha species (**Fig. 1A**). The phylome was then scanned for the presence of species-specific duplications. A total of 5,300 species/specific duplication events were detected in the *A. pisum* phylome that were clustered in 1,834 paralogous families. Due to the complexity of analysing and interpreting highly expanded and old duplications, we focused on simple duplications by identifying the families fulfilling the following three conditions: i) families only contained one duplication event (*i.e.* family of paralogs of size 2), ii) there were at least two orthologs among the other species included in the phylome, iii) those orthologs were not duplicated themselves. A total of 2,352 proteins contained one single duplication and as such fulfilled our first requirement (9.8% of the *A. pisum* proteome). These proteins were grouped in 1,176 families. Of those, 604 pairs of proteins did not fulfill the orthology requirements and therefore were discarded. Thus, the final set of duplicated proteins we

selected for further analysis contained 1,144 genes clustered into 572 families (see **Table S3** for the the complete list of selected genes and families).

We calculated the relative age of the selected duplications using the number of synonymous substitutions per synonymous site (dS) as a proxy. We estimated dS for each internal and terminal branch of each gene tree using the "free ratio branch model" from codeML and we filtered out duplications with dS > 2 and dS < 0.01 (see Material and Methods for details). By comparing the distribution dS in each copy of the selected duplications (*A. pisum* Post-Dup) with the pre-duplication branches (*A. pisum* Pre-Dup) and single-copy orthologs (**Fig. 1B**), we showed that lineage-specific genes in *A. pisum* are enriched in recent duplications represented by their low dS values compared to other species (**Fig. 2**).

In an initial characterization of our set of recent gene duplications, we estimated the median similarity for each protein sequence of each gene family alignment using trimAl v1.3. We observed that *A. pisum* duplicates were consistently (and significantly) less similar at the sequence level between them than when compared to single-copy orthologs, suggesting that their sequences are diverging faster (**Fig. S1A**).

To test the hypothesis of faster evolution of recently duplicated genes and to evaluate the pace of evolution in our set of recent gene duplications, we calculated the rate of evolution (dN/dS) for each gene family using codeML software from PAML package v4.9 (see Material and Methods for details). This software computes individual estimations for each branch of a given tree, allowing us to distinguish the evolutionary rate before and after the duplication (hereafter called as pre-duplication (Pre-Dup) and post-duplication (Post-Dup) branches, see **Fig. 1B**).

Paralogs have significantly faster rates as compared to their pre-duplicated ancestors and also to single copy orthologs (**Fig. 3A**). We then classified paralogous copies of each duplicated gene pair into fast and slow evolving copies according to the dS values of the branch subtending each copy (see Material and Methods), which allows to distinguish between subfunctionalization and neofunctionalization scenarios (Sandve *et al.* 2018). Evolutionary rates are not homogeneous in the two copies, since we observed that the fast post-duplication copy is evolving significantly more rapidly than both the slow post-duplication copy and the pre-duplication ancestor (**Fig. 3B**). Asymmetrical evolution of gene duplicates has been observed in several organisms, such as fungi, *Drosophila melanogaster*, *Caenorhabditis elegans* and human, which was attributed to relaxed selective constraints and, in some cases, to the action of

adaptive selection (Conant 2003; Zhang 2005; Scannell & Wolfe 2008; Pegueroles *et al.* 2013; Pich i Roselló & Kondrashov 2014). To further evaluate whether this pattern may be related to positive selection, we tested for positive selection using codeML (see Material and Methods for details). We detected positive selection in 189 genes distributed in 176 duplications (**Table S3, S4**), which supports that positive selection contributed to the asymmetrical acceleration of a substantial fraction of duplicates (at least ~31%). It is worth noting that our estimate of positive selective cases is conservative due to the strict filtering applied and the inherent difficulty of detecting positive selection since this often acts during short periods of evolutionary time (Zhang 2005; Pegueroles *et al.* 2013; Pich I Roselló & Kondrashov 2014) . In addition, in most duplications, only one duplicate was under positive selection, with a few exceptions (n=10, **Table S4**) where both duplicates showed signs of selection (p-value <0.01). Interestingly, post-duplication branches under positive selection have significantly different (and faster) rates than both the post-duplication branch without positive selection and the pre-duplication branch (**Fig. 3C**), which is in agreement with the lower levels of similarity detected for branches under positive selection (**Fig. S1B**, yellow boxplot).

The fraction of duplicates under positive selection is higher for the fast paralogs as compared to their slow counterparts (26% *vs* 9%, **Fig. 4A**) which supports that the asymmetrical increase in rates may be due to adaptive selection, at least in a fraction of the simple duplications analysed. We also observed that fast post-duplication copies tend to have shorter sequence lengths (**Fig. 4B**) and the evolutionary rate of the fast evolving copies is significantly higher for duplicates that are not consecutively positioned in the genome (p-value = 0.003, wilcoxon test) (**Fig. 4C**). In other words, in tandem repeats, the less common scenario affecting 56 out of 572 duplicates, both copies seem to evolve at the same evolutionary rate (p-value = 0.58, wilcoxon test). Altogether, these results point to neofunctionalization as the most likely scenario to explain the pace of evolution of *A. pisum* recent duplicates that are not duplicated in tandem.

To understand the putative functions of the duplicated and positively selected genes, we tested whether the resulting paralogs were enriched in any particular functions. While only a few functions were enriched when compared to all the genome, a vast number of them were enriched when compared to the non-expanded portion of the genome, including metabolism, development, immunological response and reproduction, among others. Functions enriched in our initial set of 1,144 duplicated genes yielded very similar results in terms of the nature of the enriched functions, indicating that these ‘young’ duplicates affected many different aspects of the

biology of *A. pisum* as well (**Fig. S2**). Interestingly, duplications under positive selection did not result in any functions enriched when compared to the whole genome, the non-expanded genome, all the expansions or the non-positively selected ones. All in all, our results indicate that positively selected genes encompass a wide range of functions in the pea aphid that are not statistically enriched when compared to the rest of the genome: neofunctionalization is therefore globally affecting biology of the pea aphid, at least for the selected species-specific duplicates (**Fig. S3**).

Some interesting biological functions affected by neofunctionalization in the pea aphid include resistance to insecticides, reproduction, morphogenesis and development, immunology, dosage compensation and ecdysis and metamorphosis, among others (**Fig. S4**). A first example of a gene positively selected that may have undergone neofunctionalization is the gene encoding the protein maelstrom 2 (UniProtKB - B3MZY6 MAEL2_DROAN), that in *Drosophila ananassae* has been predicted to play a central role during oogenesis by repressing transposable elements and preventing their mobilization, essential for maintaining the germline integrity (Sato *et al.* 2011). It is also the case of *shade*, a gene encoding an ecdysone 20-monooxygenase (UniProtKB - Q9VUF8 CP314_DROME) involved in the ecdysteroid hormone pathway and which has been proposed to be involved in the breakdown of synthetic insecticides in *D. melanogaster (Giesen et al. 2003; Petryk et al. 2003)* . In addition, ecdysone pathways are involved in the development of wings in aphids, therefore being directly implicated in dispersal polyphenism where *shade* could play a role (Vellichirammal *et al.* 2017). Likewise, evolution of dosage compensation may also be linked to neofunctionalization, since the male-specific lethal 3 (*msl3*) homolog gene (UniProtKB - P50536 MSL3_DROME) is positively selected in our dataset. This gene is known to be part of the MSL complex in *Drosophila*, which is responsible for dosage compensation in male by doubling the transcription of their single X chromosome (Tanaka *et al.* 1976; Belote & Lucchesi 1980; Samata & Akhtar 2018). Dosage compensation is highly suspected to happen on the males unique X chromosome in the pea aphid with a mechanism resembling the one from Drosophila (X-upregulation by increased chromatin accessibility) (Richard et al. 2017). Since MSL3 is highly conserved in *A. pisum* compared to *D. melanogaster* (Richard et al. 2017), and as shown here is positively selected, we propose that msl3 might be involved dosage compensation. This could thus hint towards the involvement of a MSL-like complex in the pea aphid dosage compensation despite the millions of years of evolution separating Hemiptera from Diptera. Overall, these three examples illustrate how key biological functions (oo- and morphogenesis, resistance to insecticides and dosage

compensation) might have been reshaped through duplication followed by neofunctionalization in the pea aphid.

2. Tissue divergence patterns in duplicated genes range from low to high

We have shown that recent *A. pisum* duplicates have different evolutionary rates and that a substantial fraction of them are evolving under positive selective pressure. The different behaviour of the two copies may result in differences in gene expression levels. To evaluate this hypothesis, we compiled RNA-Seq data from a total of 106 libraries grouped into 18 different conditions, and for each selected gene we calculated its expression profile, being 0 for not-expressed and 1 for expressed (**Table S5**, see Material and Methods for details). Interestingly, we observed that 198 duplicates (*i.e.* ~32% of the 572 selected duplicates) showed differences in their tissue expression pattern. In order to measure the expression divergence between duplicates in the 18 conditions, we computed three different statistics using a binary profiling binning approach: hamming distance, tissue expression complementarity (TEC) distance and tissue divergence (dT, **Table S4**, see Material and Methods for details). The three methods show similar results, which is in agreement with the high correlation between them (**Fig. S5**). Overall, tissue divergence between duplicates is low, with mean values ranging from 0.08 to 0.16 and the median being 0 in the three methods, which was expected since ~68% of the duplicates have the same expression profile. As expected, when considering merely pairs with differences in the expression profile we obtained higher values (median values ranged from 0.40 to 0.20). Interestingly, the maximum value detected is 1 for the three methods, meaning that some pair of duplicates have totally opposite expression patterns. For each tissue, GO term enrichment was also investigated. Enrichment was tested comparing the genes expressed in each tissue against genes in the entire genome, in the non-expanded genome, in all expansions, in positively-selected duplications, and in the 1,144 genes in the selected duplications. No enrichment was detected in any case.

3. Positive selection may modulate differences in gene expression

Positive selection might be correlated with sub- or neofunctionalization by acquiring a new expression profile. To test this hypothesis, we compared the tissue expression patterns between gene duplicates. We found different expression patterns in 10% (1 out of 10) of pairs with two

copies under selection, 25.4% (43 out of 169) of pairs with only one copy under selection and 36.1% (142 out of 393) of pairs with no copy under selection. This suggests that positive selection plays a role in gene transcription regulation but other factors are also involved, since in the absence of positive selection, differences in gene expression were also detected. When focusing on duplications that have different expression patterns in at least one of the studied conditions, we observed that tissue expression divergence levels were similar for duplications having or not copies under selection (**Fig. S6**). For the 169 duplications with positive selection in one copy we quantified the cases in which a gene expression was gained or lost in any of the tissues considering the expression profile of the copy with absence of positive selection as background (**Table S5**). The number of losses was higher than that of gains (35 and 8 respectively), meaning that in most cases the gene expression profile of the copy under selection is reduced as compared to the non selected copy. In other words, the selected copy is expressed in a subset of tissues. These cases are consistent with a specialization scenario, in which one copy is expressed in all (or most) tissues but at least one copy is not, since the median number of tissues in which the non-selected and the selected copies are expressed is 18 and 7 respectively. This scenario, which can be considered a particular case of subfunctionalization, has been proposed to be the main fate after whole genome duplication (Marlétaz *et al.* 2018) and may influence the evolution of young duplicates (Huerta-Cepas *et al.* 2011b). In addition, we detected 8 cases in which the selected copy is expressed in at least a tissue in which the non-selected copy have no expression (gain cases). These cases are candidates that may have undercome neofunctionalization after gene duplication. Notably, most genes positively selected with a gain in function in some tissues concentrate in the embryos and the heads (**Table 1**). From these eight duplications, four did not yield any annotations neither with BLAST or InterProScan (duplications 231, 258, 617, 647). Duplications 679 and 481 were annotated as both general transcription factor II-I related protein superfamily and ribonuclease H-like superfamily, which encompass a series of enzymes related to transposons (*i.e.*, transposases, reverse transcriptases, integrases, etc. (Rice *et al.* 1996; Reddy *et al.* 2012). In humans, the general transcription factor II-I is required for the formation of functional DNA-binding complexes and for activation of immunoglobulin heavy-chain transcription upon B-lymphocyte activation (Rajaiya *et al.* 2006), suggesting that this positively-selected young paralog may have acquired putatively an immunological function. Duplication 594 was annotated as major facilitator and MFS transporter families (**Table 1**), which are single-polypeptide secondary carriers transporting small solutes in response to chemiosmotic ion gradients (Reddy *et al.* 2012). Lastly, duplication 498 was annotated as a member of the 26S proteasome non-

ATPase regulatory subunit 12 superfamily. Proteasomes participate in numerous cellular processes, such as cell cycle progression, apoptosis, or DNA damage repair. The 26S complex plays a key role in the maintenance of protein homeostasis by removing misfolded or damaged proteins and by removing proteins whose functions are no longer required (Kanayama *et al.* 1992). Duplication followed by neofunctionalization therefore may have potentially helped in reshaping vital functions such as DNA repair, immunological response and protein homeostasis in the pea aphid.

4. Chromatin accessibility changes in young duplicated genes but is not correlated to gene expression

RNA-Seq and FAIRE-Seq data were analysed together for each of the predicted genes in the *A. pisum* genome (**Table S6**). In order to explore the correlation between transcriptional status and chromatin accessibility, genes were classified in four categories depending on their expression and TSS log2 (FAIRE/Input) values (see Methods): (i) open and expressed, (ii) open and not expressed, (iii) closed and expressed, and (iv) closed and not expressed. Concerning all genes, we found that most (8,870 genes; 64%) belonged to the category “closed and expressed”, which could potentially be due to the threshold applied to define “open genes” and to the higher sensitivity of RNA-Seq compared to FAIRE-Seq. From the remaining, 18% of the genes were “open and expressed”, and the other 18% “closed and not expressed” (2,497 and 2,491 genes, respectively). The “open and not expressed” category was barely represented in the dataset (0.3% of the assigned genes). This result reflects the quality of the FAIRE-Seq data processing since it is expected that “open and not expressed” genes are virtually absent as they will violate the common rules of gene transcription, therefore corresponding to false positives. GO term enrichment analysis for each of the four categories revealed that genes “open and expressed” were enriched in transcription factor activity (GO:0003700, molecular function) and sequence-specific DNA binding (GO:0043565, molecular function). Genes “closed and not expressed” were enriched in nucleic acid binding (GO:0003676, molecular function). The two remaining categories were not significantly enriched in any functions, despite the high number of genes in the “closed and expressed” category: this highlights the biological relevance of gene classes displaying coordinated expression and TSS accessibility in this dataset.

Regarding the set of young gene duplicates, we found that 56% of them had different chromatin states (275 out of 490 duplicates, **Table S7**) and therefore, they potentially have

different expression patterns. To test this hypothesis, we searched for duplicates where each paralog belonged to a different category (*i.e.*, categories (i) to (iv) combining chromatin state and expression pattern, as described above) for the embryos and adult morphs conditions for which FAIRE-Seq data were available (**Table S7**). Sixty-nine pairs of duplicated genes did belong to different categories. From those, in 50 pairs both duplicates were expressed, with one being open and all the other closed. In 18 pairs, both duplicates were closed but one gene was expressed and the other one not. In one single case (duplication 490) one of the genes in the pair was closed and not expressed whereas the other one was open and expressed. Both genes in the pair were annotated as uncharacterized proteins though they contain a conserved domain (PF08598), involved in repression of gene transcription mediated by histone deacetylases containing repressor-co-repressor complexes, which are recruited to promoters of target genes via interactions with sequence-specific transcription factors. The two proteins are located in differents contigs and none of the copies are under positive selection.

Our results indicate that the number of genes with closed chromatin (n=11,322) was higher than for open chromatin (n=2,536). This reduces overall the possible correlation of genes expressed and accessible at the same time, consequently hampering any gene-by-gene comparisons such as in the case of duplicated genes. Our results are in line with those of the integration of RNA-Seq and ATAC-Seq (similar to FAIRE-Seq) performed in a recent study (Ackermann *et al.* 2016), who discussed that the poor correlation between RNA-Seq and ATAC-Seq data in human tissues may be due to gene activation depending on multiple regulatory regions, possibly being located far from the gene locus itself. Indeed, FAIRE-Seq in whole body individuals or embryos is far less precise than at the level of cells or tissues, thus making the correlation between expression and accessibility even trickier. Also, since FAIRE-Seq only allows to test for *cis*-regulatory interactions, it may be hypothesized that most of the genes may be trans-regulated, which was impossible to determine with the data at hand considering the non-completeness of the pea aphid genome assembly. Moreover, our analysis was centered on putative TSS which have not been validated experimentally, for instance by CAGE-Seq (Takahashi *et al.* 2012) or NET-Seq (Churchman & Weissman 2012). Nevertheless we identified that the chromatin accessibility of the TSS of duplicated genes was different between the pairs in more than half of the cases. This shows that the chromatin states of promoters of simple duplicated genes can evolve independently in each pair. This could correspond to the differential chromatin states recently identified in new genes in nematodes (Werner *et al.* 2018), since in each pair, one gene is more recent than the other.

Conclusions

Our study shows that recent gene duplicates in *A. pisum* evolved asymmetrically, with one more conserved and one more divergent paralogous copy. The conserved copy was likely maintained mainly through purifying selection pressures and hardly ever under the effect of positive selection. The divergent copy was usually positively-selected and showed a faster evolutionary rate. This trend was stressed in copies that are located in different genomic positions, in contrast to tandem repeats, where both duplicates tend to evolve in a similar way. Altogether, these results suggest that neofunctionalization may be one of the driving forces affecting young gene duplicates in *A. pisum.* In addition, genes under positive selection were putatively related to a large, and diverse number of functions, indicating that neofunctionalization has a broad impact in many functions of the pea aphid biology. Concerning the expression patterns of the duplicated genes, we observed that one third of the duplicates showed different expression patterns, with some of them under adaptive selection. This suggests that positive selection might not be the main or the only factor driving these differences in gene expression. For those duplicates with signals of positive selection, we found that a loss of function in a specific tissue is the most likely outcome, consistent with a scenario of tissue specialization and/or subfunctionalization. In contrast, we also found examples of genes under positive selection that gained their function in some tissues, compatible with a scenario of neofunctionalization.

Lastly, we did not find a relationship between chromatin accessibility and gene expression, which may potentially be explained by technical issues such as a limited prediction of TSS in the pea aphid genome coupled to the inherent low signal over background ratio of FAIRE-Seq data (Ackermann *et al.* 2016). Moreover, although this discrepancy may be due to the different sensitivity of both RNA-Seq and FAIRE-Seq, it may also reflect a pervasive level of *trans* regulation in the pea aphid genome (as seen as well in humans (Ackermann *et al.* 2016)). Nevertheless, we showed that more than half of the young duplicated genes selected had different chromatin states. This indicates that FAIRE-Seq technique is sensitive to differences in chromatin dynamics even in recent gene duplicates.

To conclude, our results indicate that gene duplication provided an arena of genetic novelty to reshape the genome of the pea aphid through positive selection, neofunctionalization and tissue-specific expression in young duplicated species-specific gene families. The

relationships between these evolutionary scenarios are complex and difficult to disentangle. We emphasize that phylogenomic-centered studies are therefore most needed to further understand genome evolution in nonmodel organisms.

References

Ackermann AM, Wang Z, Schug J, Naji A, Kaestner KH (2016) Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. *Molecular metabolism*, **5**, 233–244.

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

Belote JM, Lucchesi JC (1980) Male-specific lethal mutations of *Drosophila melanogaster*. *Genetics*, **96**, 165–186.

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

Chang AY-F, Liao B-Y (2012) DNA methylation rebalances gene dosage after mammalian gene duplications. *Molecular biology and evolution*, **29**, 133–144.

Churchman LS, Weissman JS (2012) Native elongating transcript sequencing (NET-seq). *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, **Chapter 4**, Unit 4.14.1–17.

Conant GC (2003) Asymmetric Sequence Divergence of Duplicate Genes. *Genome research*, **13**, 2052–2058.

Dobin A, Davis CA, Schlesinger F *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* , **29**, 15–21.

Edgar RC (2004) MUSCLE: multiple sequence alignment with improved accuracy and speed. *Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004.*

Giesen K, Lammel U, Langehans D *et al.* (2003) Regulation of glial cell number and differentiation by ecdysone and Fos signaling. *Mechanisms of development*, **120**, 401–413.

Guindon S, Dufayard J-F, Lefort V *et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, **59**, 307–321.

Han MV, Demuth JP, McGrath CL, Casola C, Hahn MW (2009) Adaptive evolution of young gene duplicates in mammals. *Genome research*, **19**, 859–867.

Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP *et al.* (2011a) PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic acids research*, **39**, D556–60.

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 research*, **42**, D897–902.

Huerta-Cepas J, Dopazo J, Gabaldón T (2010a) ETE: a python Environment for Tree Exploration. *BMC Bioinformatics*, **11**.

Huerta-Cepas J, Dopazo J, Huynen MA, Gabaldón T (2011b) Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication. *Briefings in bioinformatics*, **12**, 442–448.

Huerta-Cepas J, Forslund K, Coelho LP *et al.* (2017) Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution*, **34**, 2115–2122.

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

Huerta-Cepas J, Szklarczyk D, Heller D *et al.* (2019) eggNOG 5.0: a hierarchical, functionally

and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, **47**, D309–D314.

Innan H, Kondrashov F (2010) The evolution of gene duplications: classifying and distinguishing between models. *Nature reviews. Genetics*, **11**, 97–108.

Kanayama HO, Tamura T, Ugai S *et al.* (1992) Demonstration that a human 26S proteolytic complex consists of a proteasome and multiple associated protein components and hydrolyzes ATP and ubiquitin-ligated proteins by closely linked mechanisms. *European journal of biochemistry / FEBS*, **206**, 567–578.

Katoh K (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, **33**, 511–518.

Keller TE, Yi SV (2014) DNA methylation and evolution of duplicate genes. *Proceedings of the National Academy of Sciences*, **111**, 5932–5937.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods*, **9**, 357–359.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol10: R25.

Lassmann T, Sonnhammer ELL (2005) Kalign--an accurate and fast multiple sequence alignment algorithm. *BMC bioinformatics*, **6**, 298.

Le SQ, Lartillot N, Gascuel O (2008) Phylogenetic mixture models for proteins. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **363**, 3965–3976.

Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* , **30**, 923–930.

Li Q, Brown JB, Huang H, Bickel PJ (2011) Measuring reproducibility of high-throughput experiments. *The annals of applied statistics*, **5**, 1752–1779.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* , **25**, 2078–2079.

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.

Marlétaz F, Firbas PN, Maeso I *et al.* (2018) Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature*, **564**, 64–70.

Mathers TC, Chen Y, Kaithakottil G *et al.* (2017) Rapid transcriptional plasticity of duplicated gene clusters enables a clonally reproducing aphid to colonise diverse plant species. *Genome biology*, **18**, 27.

McCarthy DJ, Chen Y, Smyth GK (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic acids research*, **40**, 4288–4297.

Mirarab S, Nguyen N, Guo S *et al.* (2015) PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences. *Journal of computational biology: a journal of computational molecular cell biology*, **22**, 377–386.

Pegueroles C, Laurie S, Albà MM (2013) Accelerated evolution after gene duplication: a time-dependent process affecting just one copy. *Molecular biology and evolution*, **30**, 1830–1842.

Petryk A, Warren JT, Marqués G *et al.* (2003) Shade is the Drosophila P450 enzyme that mediates the hydroxylation of ecdysone to the steroid insect molting hormone 20-hydroxyecdysone. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 13773–13778.

Pich i Roselló O, Kondrashov FA (2014) Long-term asymmetrical acceleration of protein evolution after gene duplication. *Genome biology and evolution*, **6**, 1949–1955.

Rajaiya J, Nixon JC, Ayers N *et al.* (2006) Induction of immunoglobulin heavy-chain transcription through the transcription factor Bright requires TFII-I. *Molecular and cellular biology*, **26**, 4758–4768.

Ramírez F, Ryan DP, Grüning B *et al.* (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic acids research*, **44**, W160–5.

Reddy VS, Shlykov MA, Castillo R, Sun EI, Saier MH (2012) The major facilitator superfamily (MFS) revisited. *FEBS Journal*, **279**, 2022–2035.

Rice P, Craigie R, Davies DR (1996) Retroviral integrases and their cousins. *Current opinion in structural biology*, **6**, 76–83.

Richard G (2017) Régulations chromatiniennes et transcriptionnelles impliquées dans le cycle de vie du puceron du pois. Rennes, Agrocampus Ouest.

Richard G, Legeai F, Prunier-Leterme N *et al.* (2017) Dosage compensation and sex-specific epigenetic landscape of the X chromosome in the pea aphid. *Epigenetics & chromatin*, **10**, 30.

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* , **26**, 139–140.

Samata M, Akhtar A (2018) Dosage Compensation of the X Chromosome: A Complex Epigenetic Assignment Involving Chromatin Regulators and Long Noncoding RNAs. *Annual review of biochemistry*, **87**, 323–350.

Sandve SR, Rohlfs RV, Hvidsten TR (2018) Subfunctionalization versus neofunctionalization after whole-genome duplication. *Nature genetics*, **50**, 908–909.

Sato K, Nishida KM, Shibuya A, Siomi MC, Siomi H (2011) Maelstrom coordinates microtubule organization during Drosophila oogenesis through interaction with components of the MTOC. *Genes & development*, **25**, 2361–2373.

Scannell DR, Wolfe KH (2008) A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome research*, **18**, 137–147.

Simon J-C, Pfrender ME, Tollrian R, Tagu D, Colbourne JK (2011) Genomics of environmentally induced phenotypes in 2 extremely plastic arthropods. *The Journal of heredity*, **102**, 512–525.

Supek F, Bošnjak M, Škunca N, Šmuc T (2011) REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PloS one*, **6**, e21800.

Takahashi H, Kato S, Murata M, Carninci P (2012) CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks. *Methods in molecular biology* , **786**, 181–200.

Tanaka A, Fukunaga A, Oishi K (1976) Studies on the sex-specific lethals of Drosophila melanogaster. II. Further studies on a male-specific lethal gene, maleless. *Genetics*, **84**, 257–266.

The International Aphid Genomics Consortium (2010) Genome Sequence of the Pea Aphid Acyrthosiphon pisum. *PLoS biology*, **8**, e1000313.

Vellichirammal NN, Gupta P, Hall TA, Brisson JA (2017) Ecdysone signaling underlies the pea aphid transgenerational wing polyphenism. *Proceedings of the National Academy of Sciences of the United States of America*, **114**, 1419–1423.

Vellichirammal NN, Madayiputhiya N, Brisson JA (2016) The genomewide transcriptional response underlying the pea aphid wing polyphenism. *Molecular ecology*, **25**, 4146–4160.

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

Wehe A, Bansal MS, Burleigh JG, Eulenstein O (2008) DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics* , **24**, 1540–1541.

Werner MS, Sieriebriennikov B, Prabh N *et al.* (2018) Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation. *Genome research*, **28**, 1675–1687.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, **24**, 1586–1591.

Zhang J (2005) Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. *Molecular Biology and Evolution*, **22**, 2472–2479.

Zhang Y, Liu T, Meyer CA *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome biology*, **9**, R137.

Acknowledgements

Drs. Akiko Sugio, Julie Jaquiéry, Gaël Le Trionnaire and Jean-Christophe Simon (INRA, UMR 1349 Igepp, Rennes, France) are acknowledged for the access to unpublished data of RNA-Seq. *Daktulosphaira vitifoliae* data were provided by the Phylloxera Genome Project (https://bipaa.genouest.org/is/aphidbase/): funding for *D. vitifoliae* clone Pcf genomic sequencing was provided by INRA (AIP Bioressources) and BGI Biotech in the frame of i5k initiative. Parts of the transcriptomic resources were obtained within the 1KITE projects (Bernhard Misof, Bonn, Germany). RF was funded by a Juan de la Cierva-Incorporación Fellowship (Government of Spain, IJCI-2015-26627) and a Marie Skłodowska-Curie Fellowship (747607). TG acknowledges support from the Spanish Ministry of Economy, Industry, and Competitiveness (MEIC) for the EMBL partnership, and grants 'Centro de Excelencia Severo Ochoa 2013-2017' SEV-2012-0208, and BFU2015-67107 co-funded by European Regional Development Fund (ERDF); from the CERCA Programme / Generalitat de Catalunya; from the Catalan Research Agency (AGAUR) SGR857, and grant from the European Union's Horizon 2020 research and innovation programme under the grant agreement ERC-2016-724173 the Marie Sklodowska-Curie grant agreement No H2020-MSCA-ITN-2014-642095. DT was funded by "Severo Ochoa visiting scientific programme" for a 6 months stay at the Center for Genomic Regulation - Barcelona - to start the project, supported as well by INRA SPE.

Data Accessibility

The *Acyrthosiphon pisum* phylome can be accessed at PhylomeDB 4.0 under phylome number 441.

RNA-Seq data are accessible at NCBI, see Table S2

Author contributions

RF designed research, analysed data, wrote the manuscript and prepared figures and tables. MMH designed research, analysed data, wrote the manuscript and prepared figures and tables. FL and SR performed bioinformatic analyses on RNA-Seq methods. GR produced FAIRE-Seq data and participated in the analysis and discussion of that section. VW was involved in the early steps of RNA-Seq analyses. CP designed research, analysed data, wrote the manuscript and prepared figures and tables. TG, DT designed and supervised research, coordinated the production of data and supervised the writing of the manuscript.

Figures and Tables

Figure 1A. Phylogenetic hypothesis of Sternorrhyncha interrelationships. Systematic classifications (superfamily, family, subfamily and tribe) are shown in each node/branch. Images selected from PhyloPic. **1B.** Example of individual gene tree showing a duplication in *A. pisum*, as the genes selected from the present study (see Material and Methods). Pre- and post-duplication branches as defined for the positive selection analysis are highlighted in red.

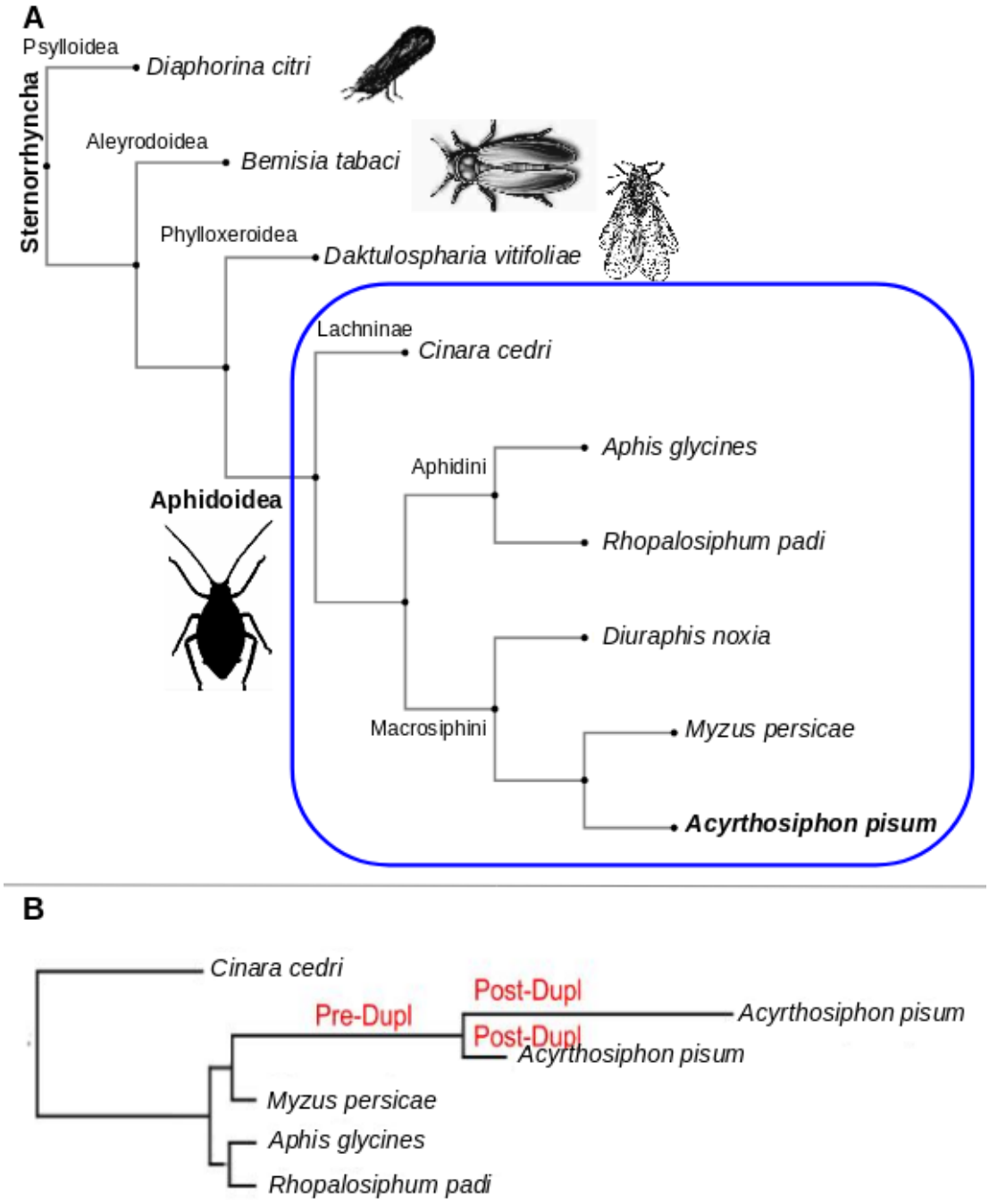

Figure 2. dS values for the selected duplications after filtering duplications with dS >2 and dS <0.01 (2A), and zoom by limiting x-axis to 0.5 (2B). dS for *A. pisum* was calculated before (*A. pisum* Pre-Dup) and after (*A. pisum* Post-Dup) the duplication took place. See Figure 1B and below for "Pre-Dup" and "Post-Dup" explanation.

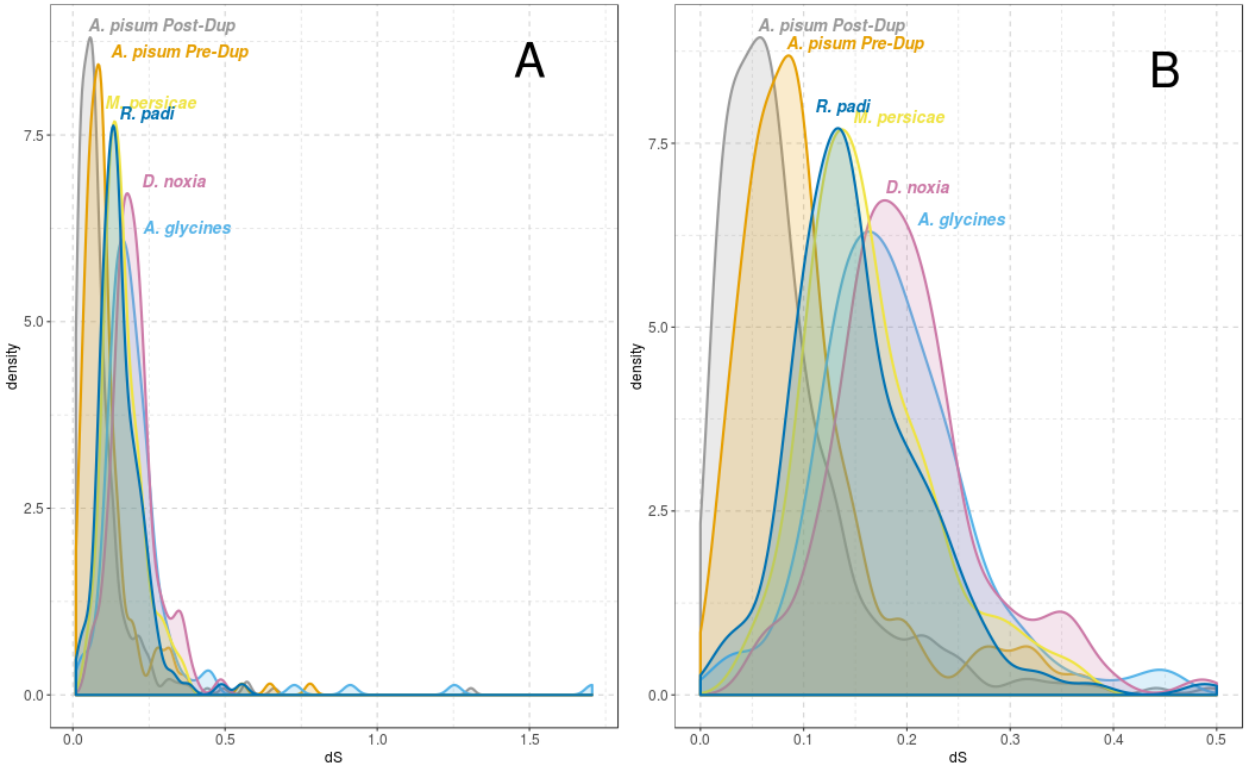

Figure 3. Evolutionary rate (dN/dS) for the selected duplications after filtering duplicates with dS >2 and dS <0.01. A: *A. pisum* pre-duplication (Pre-Dup) and post-duplication (Post-Dup) branches are shown. B: *A. pisum* post-duplication branches were classified as Fast (F) or Slow (S) according to dS. P-values were estimated using wilcox.test function from R. C: *A. pisum* post-duplication branches were classified as having (PS=1) or not (PS=0) signals of positive selection.

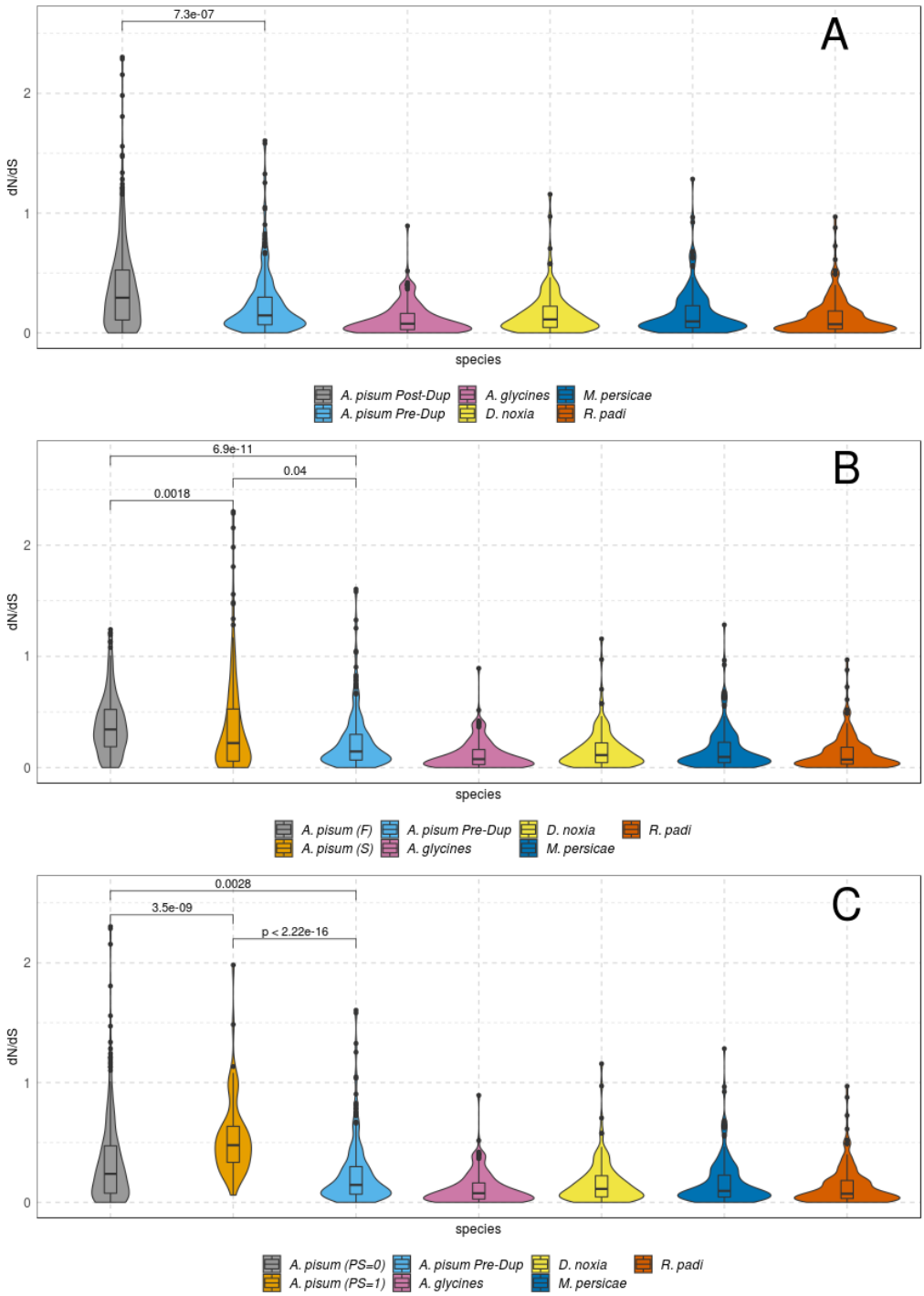

Figure 4. *A. pisum* genes from the selected duplications classified as Fast (F) or Slow (S) according to dS, after filtering duplications with dS >2 and dS <0.01. A: percentage of genes under positive selection (PS=1 in ochre) or with no signal of positive selection (PS=0 in grey); B: cDNA length (in aa); C: dN/dS after classifying duplicates according to their relative location (*i.e.* tandem and non tandem duplicates). P-values were estimated using wilcox.test function from R.

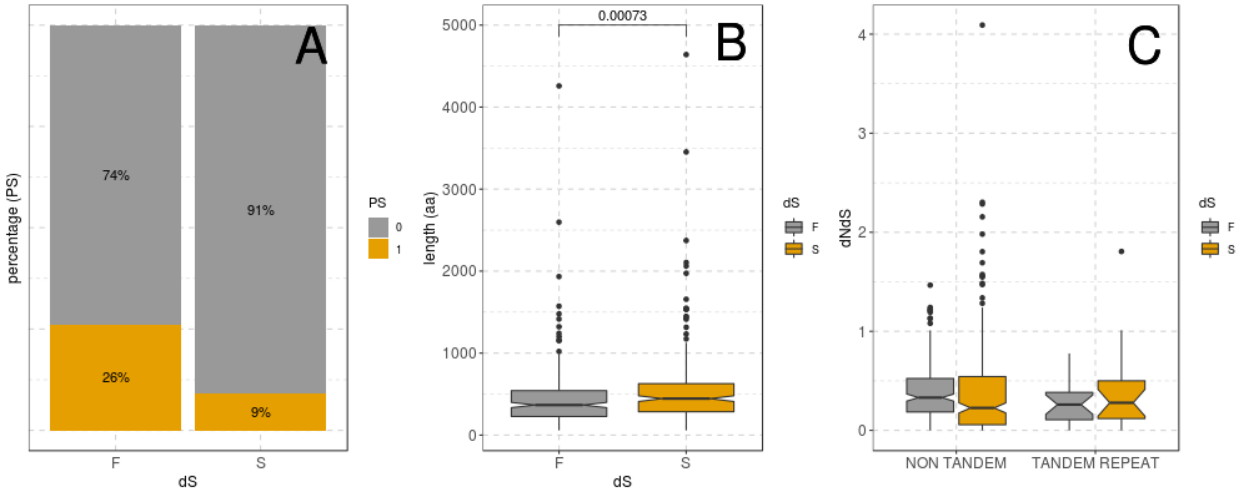

Table 1. Duplications in which the copy under positive selection (PS=1) is expressed in at least a tissue in which the non-selected copy (PS=0) have no expression (highlighted in grey). NA, no information on gene expression available. BP: biological process. MM: molecular function. CC: cellular component (see Material and Methods for further details about each condition type).

Duplication	Gene code	PS	No. tissues	Embryos_T0	Embryos_T1A	Embryos_T1K	Embryos_T2A	Embryos_T2K	Embryos_T3A	Embryos_T3K	Bacteriocyte	Head	Head_PT	Head_R2	Head_R4	Legs_PP	Gut	Salivary_glands	Adult_Male	Adult_Ovipare	Adult_Partheno	Superfamily	InterPro Code	GO term	GO description
duplication_617	LOC100571978	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	NA	1	0	0	-	-	-	
duplication_617	LOC107884413	1	2	0	NA	NA	0	NA	NA	NA	0	NA	NA	1	1	0	NA	NA	0	0	0	-	-	-	
duplication_679	LOC100161705	0	4	NA	0	0	NA	1	1	NA	0	1	0	0	NA	0	1	1	0	0	1	General transcription factor II-I related protein	IPR026625	GO:0046983 (MF)	protein dimerization activity
duplication_679	LOC103309036	1	8	1	NA	1	1	1	1	1	NA	NA	NA	0	0	0	NA	1	1	0	0	Ribonuclease H like	IPR012337		
duplication_258	LOC103310866	0	0	NA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	-	-	-
duplication_258	LOC100571229	1	12	1	1	1	1	1	1	1	1	0	NA	NA	0	1	NA	NA	1	1	1	-	-	-	-
duplication_647	LOC100569858	0	1	NA	NA	0	0	0	0	0	0	NA	0	0	0	0	NA	NA	1	0	0	-	-	-	-
duplication_647	LOC100574180	1	4	NA	NA	0	NA	NA	0	NA	0	0	NA	1	NA	0	1	1	0	0	0	-	-	-	-
duplication_594	LOC107882727	0	4	0	0	0	0	0	0	0	NA	NA	1	1	1	1	NA	NA	0	0	NA	Major facilitator superfamily	IPR011701	GO:0055085 (BP)	transmembrane transport
duplication_594	LOC100164217	1	17	NA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	MFS transporter family	IPR036259	GO:0016021 (CC)	integral component of membrane
duplication_481	LOC103309550	0	0	0	0	0	0	0	0	0	0	NA	0	0	0	0	NA	NA	0	0	0	General transcription factor II-I related protein	IPR026625	-	-
duplication_481	LOC103307965	1	1	0	NA	0	0	NA	0	0	1	NA	NA	NA	0	0	NA	NA	0	NA	NA	Ribonuclease H like	IPR012337	-	-
duplication_231	LOC103307906	0	1	0	0	NA	0	0	0	0	0	NA	0	0	0	0	1	NA	0	0	0	-	-	-	-
duplication_231	LOC103309408	1	4	NA	0	NA	NA	NA	0	0	0	1	1	1	NA	0	1	NA	0	0	0	-	-	-	-
duplication_498	LOC100568916	0	9	1	1	1	1	1	1	1	NA	NA	NA	0	0	NA	NA	NA	1	1	NA	26S Proteasome non-ATP ase regulatory subunit 12	IPR035297	-	-
duplication_498	LOC100160128	1	15	1	1	1	1	1	1	1	1	NA	1	1	1	1	NA	NA	1	1	1	COP9 signalosome complex subunit 4	IPR040134	-	-