

# Selection following Gene Duplication Shapes Recent Genome Evolution in the Pea Aphid *Acyrtosiphon pisum*

Rosa Fernández,<sup>†,§,1</sup> Marina Marcet-Houben,<sup>†,§,1,2</sup> Fabrice Legeai,<sup>3,4</sup> Gautier Richard,<sup>3,5</sup> Stéphanie Robin,<sup>3,6</sup> Valentin Wucher,<sup>1</sup> Cinta Pegueroles,<sup>\*,†,§,1,2</sup> Toni Gabaldón,<sup>\*,†,§,1,2,7,8</sup> and Denis Tagu<sup>\*,†,3</sup>

<sup>1</sup>Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Barcelona, Spain

<sup>2</sup>Institute for Research in Biomedicine (IRB Barcelona), Barcelona, Spain

<sup>3</sup>IGEPP, INRAE, Agrocampus Ouest, Université de Rennes 1, Le Rheu, France

<sup>4</sup>INRIA, IRISA, Genscale, Campus Beaulieu, Rennes, France

<sup>5</sup>Max Planck Institute of Immunobiology and Epigenetics, Freiburg im Breisgau, Germany

<sup>6</sup>INRIA, IRISA, GenOuest Core Facility, Campus Beaulieu, Rennes, France

<sup>7</sup>Universitat Pompeu Fabra, Barcelona, Spain

<sup>8</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

<sup>§</sup>Present address: Animal Biodiversity and Evolution, Institute of Evolutionary Biology (CSIC-UPF), Barcelona, Spain

<sup>¶</sup>Present address: Department of Life Sciences, Barcelona Supercomputing Center, Barcelona, Spain

<sup>†</sup>These authors contributed equally to this work.

<sup>‡</sup>These authors contributed equally to this work as the corresponding authors.

\*Corresponding authors: E-mails: cintapq@gmail.com; toni.gabaldon.bcn@gmail.com; denis.tagu@inra.fr, denis.tagu@rennes.inra.fr.

Associate editor: Rebekah Rogers

## 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 are 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 *Acyrtosiphon 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 recent pairs of gene duplicates evolved asymmetrically, showing divergent patterns of positive selection and gene expression. Genes under selection involved a plethora of biological functions, suggesting that neofunctionalization and tissue specificity, among other evolutionary mechanisms, have orchestrated the evolution of recent paralogs in the pea aphid and may have facilitated host–symbiont cooperation. Our comprehensive phylogenomics analysis allowed us to tackle the history of duplicated genes to pave the road toward understanding the role of gene duplication in ecological adaptation.

**Key words:** gene duplicates, FAIRE-Seq, insect, neofunctionalization, phylogenomics, positive selection.

## Introduction

Aphids are insect pests belonging to the order Hemiptera, which diverged some 280–250 Ma. They feed exclusively on plant phloem sap, a trait that involves 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 female morphs. Although several studies have addressed 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 (International Aphid Genomics Consortium 2010; Mathers et al. 2017; Li et al. 2019). Some of these duplications are shared between aphid species, but most of them are lineage specific (International Aphid Genomics Consortium 2010). The modes of evolution of gene duplicates 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, Dopazo, et al.

© The Author(s) 2020. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

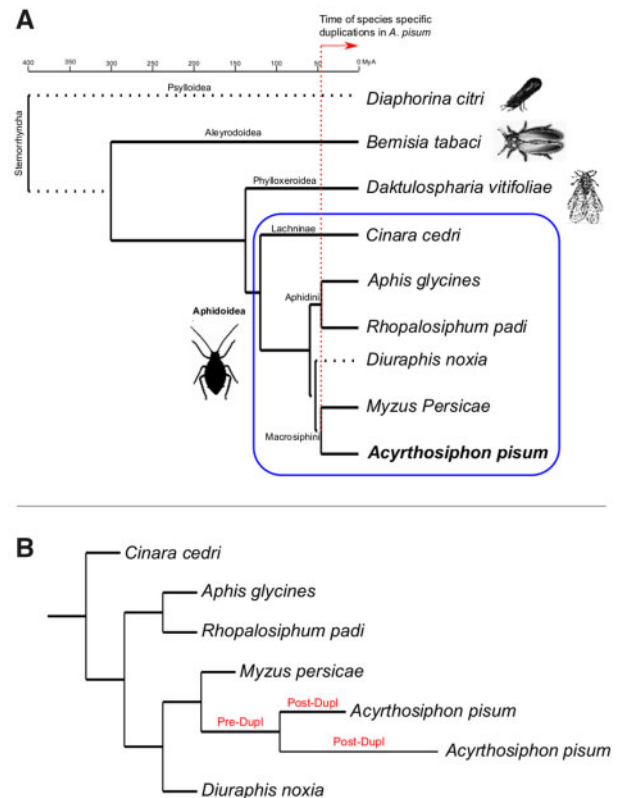
Open Access

2011; Huerta-Cepas, Marcet-Houben, et al. 2010; Simon et al. 2011).

There are at least four different outcomes for gene duplicates (reviewed by Capella-Gutierrez et al. [2009] and Innan and Kondrashov [2010]). First, although 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 and Conery 2000; Han et al. 2009; Innan and 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 and Kondrashov 2010; Pegueroles et al. 2013; Pich I Roselló and Kondrashov 2014).

More recently, sub- or neo-functionalization have started to be assessed by epigenetic regulation. Acquiring and losing functions can occur, among other means, by modification of chromatin states, which drive the 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 and 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 and Liao 2012), before mutations occur and genetic selection operates. It is worth noting that divergent epigenetic environments may result in subfunctionalization (e.g., through changed expression patterns), but the epigenetic differences themselves do not result from functional differences between the copies.

Here, we test the hypothesis that gene duplication—particularly recent duplicates—in the pea aphid *Acyrtosiphon 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 1) a large proportion of gene duplications are under positive selection in *A. pisum* and affect a large number of biological functions (most notably oo- and



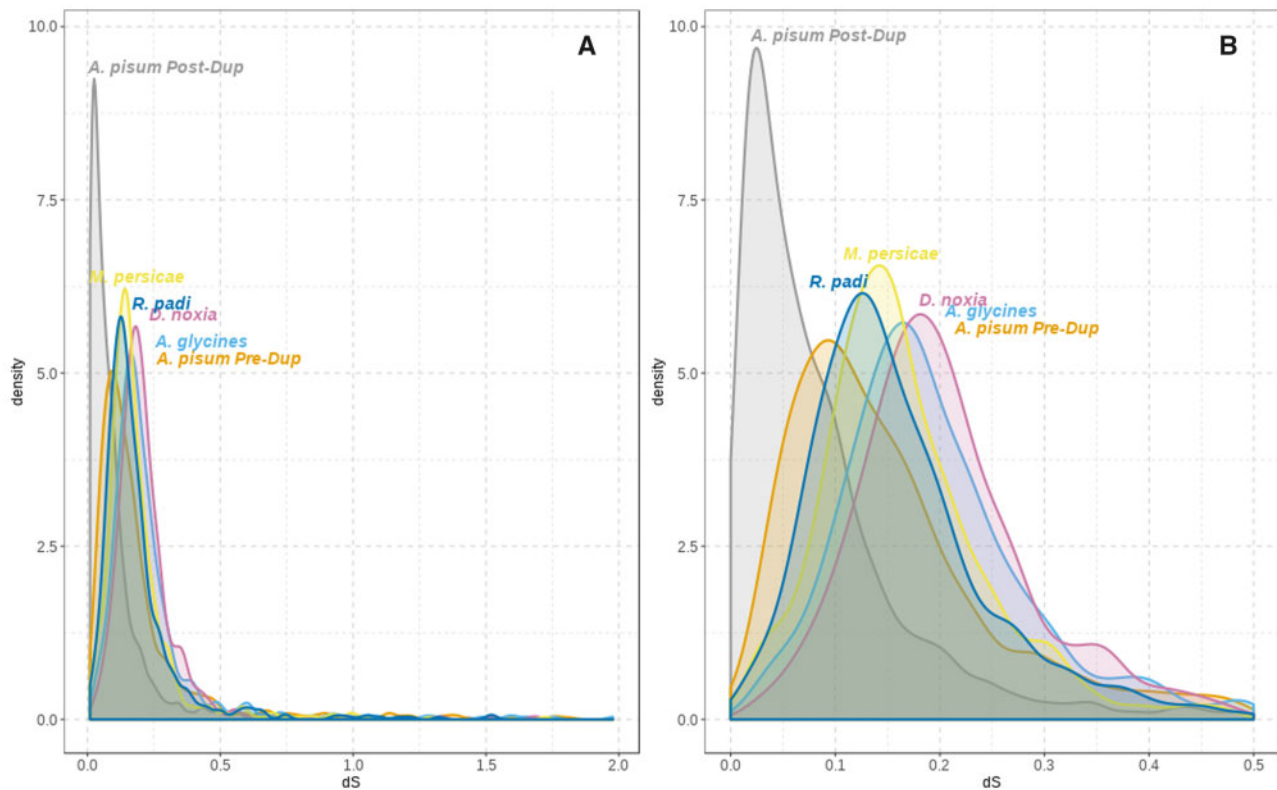
**FIG. 1.** (A) Chronogram of Sternorrhyncha interrelationships. Systematic classifications (superfamily, family, subfamily, and tribe) are shown in each node/branch. Images selected from PhyloPic. Divergence times taken from TimeTree (Kumar et al. 2017). Dotted branches represent lineages for which divergence times were not available. (B) Example of individual gene tree showing a duplication in *Acyrtosiphon pisum*, as the genes selected from the present study (see Materials and Methods). Pre- and post-duplication branches as defined for the positive selection analysis are highlighted in red.

morpho-genesis and host–symbiont cooperation), 2) asymmetrical rates of young paralogs coupled to positive selection suggest that neofunctionalization is a main force reshaping the pea aphid genome, 3) a third of young duplicates show divergent tissue expression patterns, consistent in some cases with subfunctionalization by tissue specialization and others with neofunctionalization through gain of gene expression, and 4) 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*.

## Results and Discussion

### 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

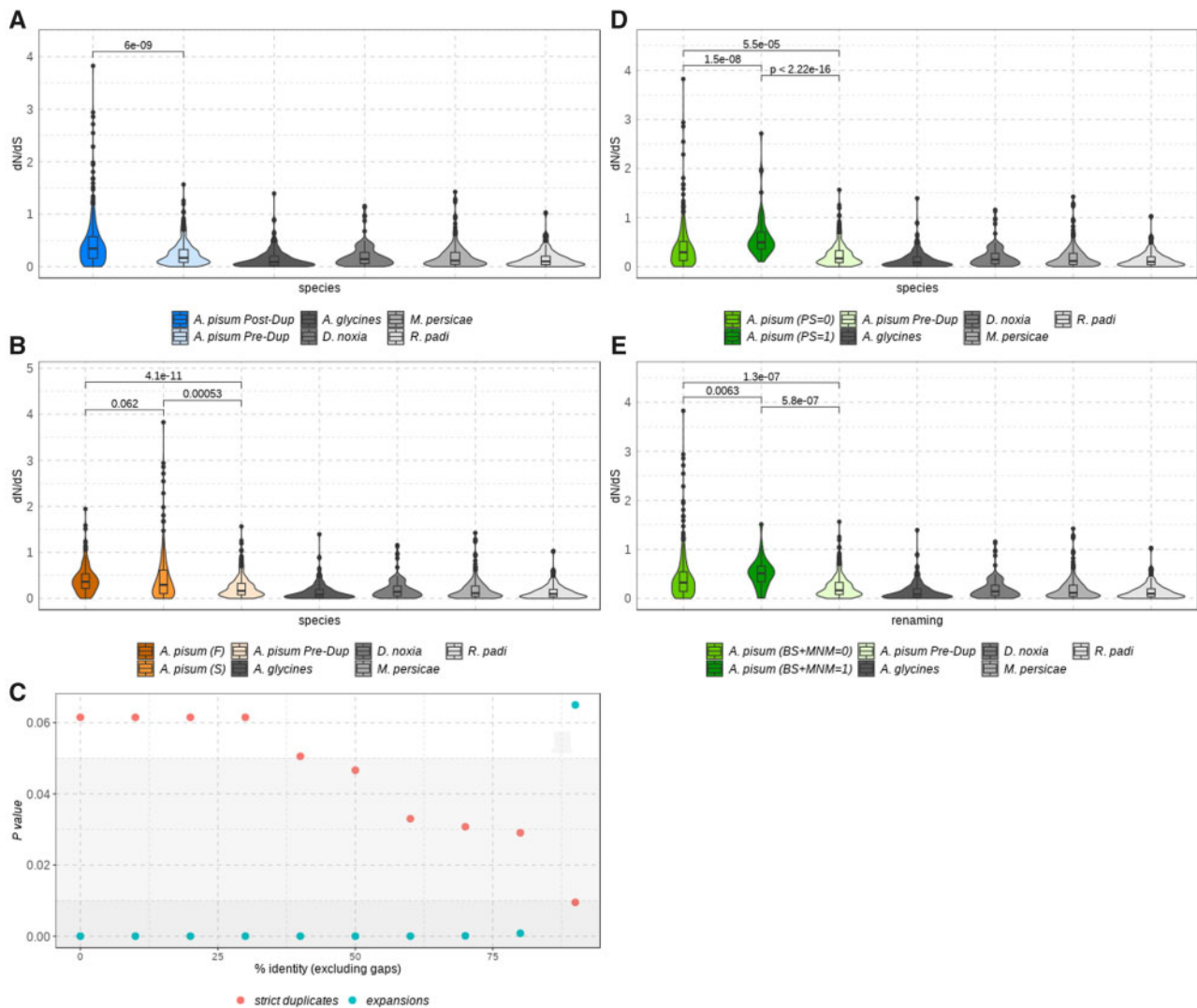


**Fig. 2.** dS values for the selected duplications after filtering genes with  $dS > 2$  and  $dS < 0.01$  (A) and zoom by limiting x-axis to 0.5 (B). *Cinara cederi* was removed from this plot for visualization. dS for *Acyrtosiphon pisum* was calculated before (*A. pisum* Pre-Dup) and after (*A. pisum* Post-Dup) the duplication took place. See figure 1B for “Pre-Dup” and “Post-Dup” explanation.

duplication events were detected in the *A. pisum* phylome that were clustered in 1,834 paralogous families. Due to the complexity of analyzing and interpreting highly expanded and old duplications, we divided the duplications into two different sets. The first set consisted of pairs of genes specifically duplicated in *A. pisum* which fulfilled the following criteria: 1) they presented a single-copy ortholog in at least two of the other species in the phylome, 2) the node where the two sequences duplicated in the tree was well supported, and 3) the tree reconstructed only using the selected paralogs and orthologs had to retain the duplication event. A second data set focused on genes that were duplicated more than once specifically in *A. pisum*. Since our goal was to understand the evolution of recent duplicates, in the case of these larger gene family expansions only pairs of genes found at the tips were considered. These pairs also had to fulfill the previous requirements. A total of 843 duplication events containing 1,686 genes were further analyzed. Among those, 606 came from single duplication events (the first set), whereas the remaining 237 were extracted from larger duplication events (from the second set; see [supplementary table S3, Supplementary Material](#) online, for the complete list of selected genes and families). Note also that in some cases (8.5% of cases), an ortholog to the closest relative to *A. pisum* (*Myzus persicae*) was missing, and hence, there is the possibility that these particular duplications are older.

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 specific genes with  $dS > 2$  and  $dS < 0.01$  (see Materials and Methods for details). By comparing the distribution dS in each copy of the selected duplications (*A. pisum* Post-Dup) with the preduplication 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 with other species (fig. 2). Since it is known that gene conversion (GC) may decrease the divergence between paralogs, we scanned the respective coding sequences for the presence of GC tracts using GENECONV software (Sawyer 1989). We detected that 187 duplications (22.2%) showed evidence of a GC event between the two *A. pisum* sequences that remained significant after multiple-comparison correction. From those, 27 occurred in tandem duplicates (28.7% of total duplications in tandem), 31 in duplications in the same contig (31%) and 117 in duplications in different contigs (19.3%). Thus, tandem duplicates do not seem to be particularly enriched in GC; however, we cannot rule out that some of the selected duplicates may be older than inferred due to GC. In addition, due to the fragmentation of the genome assembly used, it is possible that some tandem duplicates are not detected in our analyses. In an initial characterization of our set of recent gene duplications, we estimated the median identity for each protein sequence of each gene family alignment using trimAl



**Fig. 3.** Evolutionary rate ( $dN/dS$ ) for the selected duplications after filtering duplications with  $dS > 2$  and  $dS < 0.01$  in any of the genes (201 strict duplicates remained). *Acyrtosiphon pisum* genes are colored, whereas single-copy orthologs are shown in gray scale. (A) *A. pisum* preduplication (Pre-Dup) and postduplication (Post-Dup) branches are shown. (B) *A. pisum* postduplication branches were classified as Fast (F) or Slow (S) according to  $dS$ . (C)  $P$  values comparing Fast (F) or Slow (S) copies from duplicates binned according to their percentage of identity (increasing by 10 the percentage of identity between bins). Background was colored according to the  $P$  value: dark gray for  $P$  value  $< 0.01$ , light gray for  $P$  value  $> 0.01$  and  $< 0.05$ , and white for  $P$  value  $> 0.05$ . (D) *A. pisum* postduplication branches were classified as having ( $PS = 1$ ) or not ( $PS = 0$ ) signals of positive selection. (E) *A. pisum* postduplication branches were classified as having ( $BS + MNM = 1$ ) or not ( $BS + MNM = 0$ ) signals of positive selection after  $BS + MNM$  model.  $P$  values for all plots were estimated using `wilcox.test` function from R.

v1.3. We observed that *A. pisum* duplicates were significantly less similar at the sequence level between them than when compared with single-copy orthologs, suggesting that their sequences are diverging faster (supplementary fig. S1A, Supplementary Material online). Thus, despite the presence of GC tracts, this process was not enough to homogenize the sequence of the gene duplicates. As a consequence, we did not discard GC tracts from the sequences because, according to the literature, DNA sequences are useful to detect the presence of GC but not to correctly infer their length (Mansai et al. 2011).

To test the hypothesis that recently duplicated genes evolved faster and to evaluate the pace of evolution in our set of recent gene duplications, we calculated evolutionary rates as the ratio of nonsynonymous to synonymous

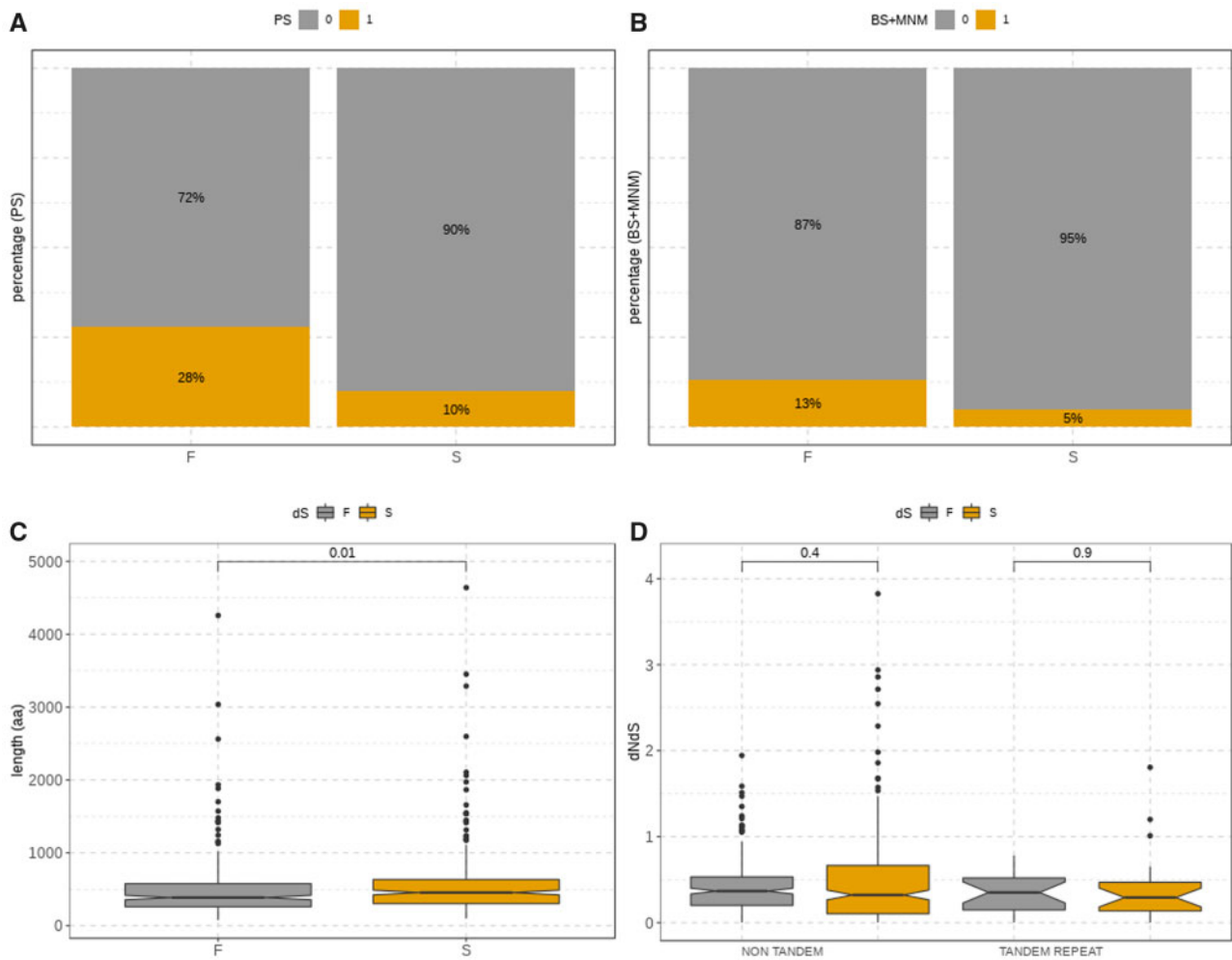
substitution rates ( $dN/dS$ ) for each gene family using codeML software from PAML package v4.9 (see Materials and Methods for details). This software computes individual estimations for each branch of a given tree, allowing us to calculate  $dN/dS$  before and after the duplication (hereafter called as preduplication [Pre-Dup] and postduplication [Post-Dup] branches, see fig. 1B). To facilitate the interpretation of the results, gene duplicates were divided into two groups: “strict duplicates” (i.e., genes with only two copies, ~72% of the selected duplicates) and “expansions” (genes with more than two copies, see Materials and Methods for details). Paralogs of both “strict duplicates” and “expansions” had significantly faster rates as compared with their preduplicated ancestors as well as to single-copy orthologs (fig. 3A and supplementary fig. S2, Supplementary Material online). 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 Materials and Methods), which allows to distinguish between subfunctionalization and neofunctionalization scenarios (Sandve et al. 2018).  $dN/dS$  was not homogeneous in the two copies, since the fast postduplication copy is evolving more rapidly than both the slow postduplication copy and the preduplication ancestor in the “strict duplicates” subset (fig. 3B). The statistical significance of these differences in  $dN/dS$  values depends on the percentage of identity between duplicate copies (fig. 3D). The “expansions” subset showed the opposite pattern, in which the slow copy had a higher  $dN/dS$  ratio than the fast one (supplementary fig. S2B, Supplementary Material online). It is important to highlight that expansions are complex families more prone to include missannotated genes or pseudogenes, which may influence our results. Thus, conclusions from this subset should be taken with caution. 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 and Wolfe 2007; Pegueroles et al. 2013; Pich I Roselló and Kondrashov 2014). To further evaluate whether asymmetrical evolution may be related to positive selection, we tested for positive selection using codeML (see Materials and Methods for details). We detected positive selection in 388 genes distributed in 316 duplications (supplementary tables S3 and S4, Supplementary Material online), which supports that positive selection contributed to the acceleration of a substantial fraction of duplicates (at least  $\sim 37\%$ ). In addition, in most duplications, only one duplicate was under positive selection, with some exceptions where both duplicates showed signs of selection (28 and 44 for “strict duplicates” and “expansions,” respectively, supplementary table S4, Supplementary Material online). Interestingly, postduplication branches under positive selection have significantly different (and faster) rates (i.e.,  $dN/dS$  values) than both the postduplication branch without positive selection and the preduplication branch in both subsets (fig. 3C), which is in agreement with the lower levels of identity detected for branches under positive selection (supplementary fig. S1B, Supplementary Material online, yellow boxplot). It is worth noting that our estimate of positive selection cases may be conservative due to the strict applied filtering 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 and Kondrashov 2014). However, a recent paper showed that the branch-site test cannot distinguish which sequence patterns have been caused by positive selection or by the neutral fixation of nonsynonymous multinucleotide mutations (MNM) (Venkat et al. 2018). To evaluate this, we identified MNMs in our gene duplicates. First, we reconstructed the ancestral sequence using codeML (see Materials and Methods for details), and then we compared the derived and ancestral sequences codon by codon and counted the number of changes. Those codons having

more than one change were considered as MNMs. Supplementary figure S3, Supplementary Material online, shows the percentage of MNMs per sequence for the duplicated genes, classified as having or not signals of positive selection. We observed that genes under positive selection were enriched in MNMs in both subsets. In order to test whether MNMs were biasing our estimates of positively selected genes, we evaluated the detection of genes under positive selection using a model that accounts for the presence of MNMs (BS + MNM model; see Materials and Methods). The new analysis recovered 155 genes under selection under the BS + MNM (supplementary fig. S3, Supplementary Material online). Therefore,  $\sim 40\%$  of the genes originally inferred as positively selected still maintained the signal of positive selection after accounting for MNMs. These genes thus represent a conservative set of genes under positive selection in the pea aphid. However, it should be noted that MNM-bearing genes can indeed be truly under positive selection, and as such we did not exclude them from our discussion.

To evaluate the impact of positive selection in pairs of gene duplicates, we discarded the duplicates with  $dS > 2$  or  $dS < 0.01$  in any of the duplicated genes (fig. 4A and supplementary fig. S4, Supplementary Material online). The fraction of duplicates under positive selection is higher for the fast paralogs as compared with their slow counterparts (fig. 4A and supplementary fig. S4, Supplementary Material online) which supports that the asymmetrical increase in rates may be due to adaptive selection, at least in a fraction of the duplications analyzed, especially in the “strict duplicates” subset. We also observed that the fast postduplication copies tend to have shorter sequence lengths in the “strict duplicates” subset (fig. 4B, not in the “expansions” subset as shown in supplementary fig. S4B, Supplementary Material online).  $dN/dS$  values of the fast and slow evolving copies were similar independently of being consecutively positioned in the genome or not, particularly in the “strict duplicates” set (fig. 4C). It is worth noting that there are more cases of positive selection in the “expansions” subset, and consequently positive selection is more prone to be wrongly assigned to this subset due to its complexity.

We further evaluated the presence of selective pressures on *A. pisum* genome by estimating Tajima's  $D$  for all annotated genes. We obtained variation data from three pea aphid host races specialized on different crops: alfalfa, pea, and genista (Grigorescu et al. 2018; Nouhaud et al. 2018). We observed differences in Tajima's  $D$  values between races, but overall values are negative in the three of them (supplementary figs. S5–S8, Supplementary Material online). Negative values indicate an excess of rare alleles, which may be due to purifying selection or a recent population expansion after a bottleneck. We splitted the genome into “strict duplicates,” “expansions,” and the rest of the genes (which are mostly not duplicated) and estimated Tajima's  $D$  in these subsets of genes. We observed the same trend in the three host races. When considering all genes as a whole, “expansions” have significantly higher values than both “strict duplicates” and the rest of the genes (with the only exception of genista population), which may indicate that the selective



**Fig. 4.** *Acyrthosiphon pisum* genes from the selected duplications classified as Fast (F) or Slow (S) according to dS, after filtering duplications with  $dS > 2$  and  $dS < 0.01$  (201 strict duplicates remained, see Materials and Methods for details). (A) Percentage of genes under positive selection (PS = 1 in ochre) or with no signal of positive selection (PS = 0 in gray); (B) percentage of genes positively selected after BS + MNM model (BS + MNM = 1 in ochre) or not (BS + MNM = 0 in gray); (C) cDNA length (in aa); and (D) dN/dS after classifying duplications according to their relative location (i.e., tandem and nontandem duplicates). *P* values were estimated using wilcox.test function from R.

pressure to constrain “expansions” is lower ([supplementary fig. S5, Supplementary Material](#) online). However, no significant differences were found when comparing Tajima’s *D* for the fast and slow evolving copies in none of the subsets, indicating that both copies may have undergone similar levels of purifying selection. Remarkably, we observed that exons have significantly lower values compared with introns. Thus, negative values are more likely to be driven by purifying selection rather than a recent population expansion, even if changes in population size are expected in a species that reproduces by cyclical parthenogenesis, as is the case in *A. pisum*. We also investigated codon usage index in gene duplicates, since departures from optimal codons in one of the copies would suggest relaxed purifying selection. Most of the duplicates (62%) have an optimal codon usage, whereas in 32% of the cases both copies have a nonoptimal codon usage. For pairs of duplicates in which one of the copies was under positive selection, we further explored whether there were differences in optimal codon usage. Out of the 244

duplicates in which only one gene is under positive selection in the pair of duplicates, only nine pairs showed a low codon adaptation index in the positively selected gene. A similar number (eight pairs) was recovered for the opposite scenario (i.e., in pairs of genes where the gene under selection was the one showing a better codon usage). Thus, our set of recent duplicates does not seem to have a differential relaxation of purifying selection among copies. Altogether, these observations suggest that, at least for the “strict duplicates” subset, neofunctionalization fueled by positive selection is the most likely scenario underlying the observed evolution patterns of recent duplicates in *A. pisum*.

To identify the putative functions of the duplicated and positively selected genes, we tested whether the resulting paralogs were enriched in any particular functions through Gene Ontology (GO) enrichment analyses. We explored enrichment only in strict duplicates as described above, which accounts for 72% of all duplicates. Similar results including enrichment in neurotransmitter metabolism, neural retina

development, biosynthesis, and metabolism of glutamate, quinone, and ammonia (supplementary fig. S9, Supplementary Material online) were found when the list of strict duplicates under selection was compared with all the genome or with the complementary portion of the genome (i.e., all genes but the pairs of strict duplicates where at least one gene is under positive selection). This suggests that neofunctionalization may be affecting all these functions. Strict duplicates under positive selection did not result in any functions enriched when compared with all strict duplicates.

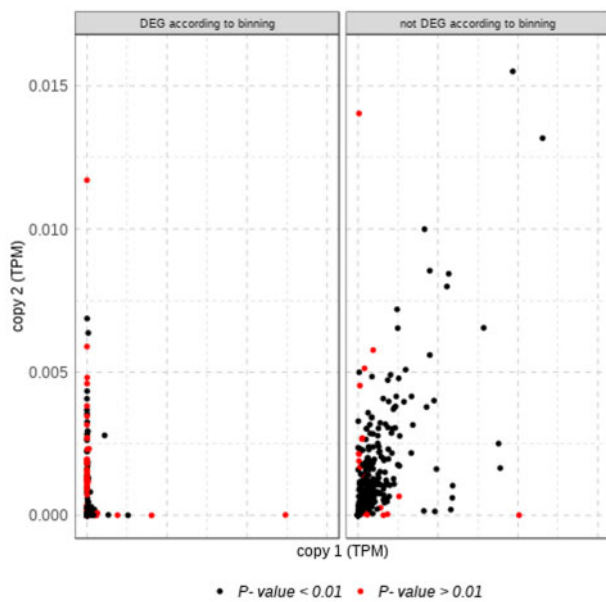
A first example of a gene positively selected that may have undergone neofunctionalization is the gene encoding the protein *maelstrom 2* (UniProtKB B3MZY6 MAEL2\_DROAN), that in *Drosophila ananassae* has been predicted to play a central role during oogenesis by repressing transposable elements and preventing their mobilization, essential for maintaining the germline integrity (Sato et al. 2011). It is also the case of the genes encoding for glutamine synthetase 2 (UniProtKB J9JML2\_ACYPI) and other genes involved in glutamate metabolism such as the genes encoding aspartate aminotransferase 2 (UniProtKB J9JIS1\_ACYPI), glutamate dehydrogenase (UniProtKB J9KB74\_ACYPI), and glutamate decarboxylase (UniProtKB DCE\_DROME). These genes belong to the same pathway of incorporation of ammonium nitrogen into glutamate cycle to assimilate ammonia into glutamate: Those genes have been shown to be upregulated in bacteriocytes in *A. pisum*, which function as specialized symbiont-bearing organs of amino acid production (Hansen and Moran 2011). Therefore, positive selection and neofunctionalization may have facilitated host–symbiont cooperation in the production of amino acids between the pea aphid and *Buchnera*, such as for the amino acid transporters as shown by Duncan et al. (2016). Overall, these two examples illustrate how key biological functions (oo- and morpho-genesis and host–symbiont cooperation) might have been reshaped through duplication followed by neofunctionalization in the pea aphid.

### Tissue Divergence Patterns in Duplicated Genes Range from Low to High

We have shown that recent *A. pisum* duplicates have different evolutionary rates (using dN/dS values as a proxy) and that a substantial fraction of them are evolving under positive selective pressure. The different behavior 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 (see Materials and Methods). A principal component analysis showed that samples are clustered by tissues when considering all annotated genes in the *A. pisum* genome as well as when subtracting the genes of the “strict duplicates” and “expansions” subsets (supplementary fig. S10, Supplementary Material online). To compare gene expression profiles across tissues, we profiled the expression of each selected gene, being 0 for not expressed and 1 for expressed (supplementary table S5, Supplementary Material online, see Materials and Methods for details). Interestingly, we observed

that 329 duplicates (i.e., ~39% of the 843 selected duplicates, 212 of them “strict duplicates” and 117 “expansions”) showed differences in their tissue expression pattern in at least one tissue. When considering their location, we observed that corresponds to 40.4% of the 705 DNA-based pairs (i.e., duplicated by copying and pasting a DNA sequence from one genomic region to another), 36.4% of the 44 retrocopies (i.e., mRNA-mediated gene duplicates), and 29.8% of the 94 tandem duplicates (i.e., duplicated one after the other in the genome). Thus, most tandem duplicates (~70%) tend to have the same expression patterns, as expected (Lercher et al. 2003). 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 distance and tissue divergence (dT, supplementary table S4, Supplementary Material online, see Materials and Methods for details). The three methods show similar results in both subsets (“strict duplicates” and “expansions”), which is in agreement with the high correlation between them (supplementary fig. S11, Supplementary Material online). Overall, tissue divergence between duplicates is low, with mean values ranging from 0.09 to 0.20 and the median being 0 in the three methods, which was expected since ~71% 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 0.33 to 0.20). Interestingly, the maximum value detected is 1 for the three methods, meaning that some pairs of duplicates have totally opposite expression patterns.

We also compared gene expression between copies estimated as transcripts per million (TPMs). It is worth noting that overall gene expression values for “expansions” is significantly lower than for “strict duplicates” (median values were 1.07e-04 and 2.43e-05, respectively,  $P$  value = 2.64e-07). This may reflect real biological differences such as higher tissue specificity or lower expression of genes that are part of large family expansions, although it may also be due to the difficulty to assign gene expression to a given copy in these complex expansions and/or to the likely higher amount of missannotations or pseudogenes in this subset, as discussed above. We compared gene expression across tissues, by computing Pearson correlations and building linear models within gene duplicates (see Materials and Methods for details). If the two copies have similar expression patterns across tissues we should expect high Pearson correlations and  $r^2$  values. Overall, our findings are in line with the binning approach, since both Pearson correlations and  $r^2$  values are high (0.91 and 0.82, respectively, for “strict duplications” and 0.77 and 0.60 for “expansions”; supplementary table S6, Supplementary Material online). In addition, the subset of differentially expressed genes obtained in the binning analysis is enriched in differentially expressed genes according to our models (fig. 5 and supplementary fig. S12, Supplementary Material online). Thus, we can conclude that the two subsets tend to have similar expression patterns in the two copies, despite the fact that there are some interesting differences as we will discuss below.



**Fig. 5.** Scatterplot showing median gene expression between pairs of duplicates for the “strict duplications” subset. Colors correspond to the  $P$  values of a Pearson correlation (see Materials and Methods for details). The significance of the test should be interpreted with caution for those genes that are lowly expressed (i.e., those located in the lower left corner). For visualization purposes, we discarded five gene pairs that were outliers in the subset “not DEG according to binning.”

### 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 50% of pairs with two copies under selection (35.7% of 14 “strict duplicates” and 59% of “expansions”), 36.9% of pairs with only one copy under selection (30% of “strict duplicates” and 50.7% of “expansions”), and 39.6% of pairs with no copy under selection (37% of “strict duplicates” and 47.2% of “expansions”). 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 in both subsets (“strict duplicates” and “expansions”) as well as overall duplicates (supplementary figs. S13 and S14, Supplementary Material online). For the 244 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 (supplementary table S5, Supplementary Material online). The number of losses was higher than that of gains in the “strict duplicates” subset (40 and 15, respectively), meaning that in most cases the gene expression profile of the copy under selection is reduced as

compared with the nonselected copy. In other words, the selected copy is expressed in a subset of tissues at least in the “strict duplicates” subset, since the median number of tissues in which the nonselected and the selected copies are expressed is 11 and 8.5, respectively, in this set of duplicates. In the “expansions” subset the amount of gains and losses is quite similar (21 and 19, respectively), as well as the median number of tissues in which the nonselected and the selected copies are expressed (9 and 10, respectively). The pattern observed in the “strict duplicates” subset is consistent with a specialization scenario, in which one copy is expressed in all (or most) tissues but at least one copy is not. 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, Dopazo, et al. 2011). In addition, the 15 cases in which the selected copy is expressed in at least a tissue in which the nonselected copy has no expression (gain cases) are candidates that may have undergone neofunctionalization after gene duplication (table 1). From these 15 duplications, 9 showed similar predicted annotation between the pairs (duplications 288, 322, 397, 460, 489, 576, 633, 795, and 840). The six other duplications include pairs with different predicted annotations. For the 15 duplications (i.e., 30 genes), 16 genes are uncharacterized with no predicted functions, 5 are annotated as zinc finger putative proteins, and 3 are dynein-like proteins.

### Chromatin Accessibility Is Altered in Young Duplicated Genes but Does Not Correlate to Gene Expression

RNA-Seq and FAIRE-Seq data were analyzed together for each of the predicted genes in the *A. pisum* genome (supplementary table S7, Supplementary Material online). FAIRE-Seq is a molecular technique that allows to detect nucleosome-depleted regions of the genome, which are usually found in open chromatin. We validated the overall correlation of transcription and chromatin accessibility (supplementary fig. S15, Supplementary Material online, for embryos and Richard et al. 2017, fig. 5 for whole body of males and females), ensuring the quality of the data sets. In order to explore the correlation between transcriptional status and chromatin accessibility, genes were classified in four categories depending on their expression and TSS  $\log_2$  (FAIRE/Input) values (see Materials and Methods): 1) open and expressed, 2) open and not expressed, 3) closed and expressed, and 4) 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 with 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 data set (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



**Table 1.** Duplications in Which the Copy under Positive Selection (PS = 1) Is Expressed in at Least One Tissue in Which the Nonselected Copy (PS = 0) Has No Expression (highlighted in dark gray).

Duplication Name	Gene Code	Putative Function	PS	BS-M	AM	AP	AO	E0	E1A	E1K	E2A	E2K	E3A	E3K	H	HP	HR2	HR4	LP	G	SG	B	
duplication_159	LOC103311559	Uncharacterized	0	0	0	0	0	0	0	0	0	0	0	0	NA	0	0	0	0	0	0	NA	0
duplication_159	LOC100166252	PiggyBac transposable element-derived protein 4-like	1	0	NA	NA	NA	NA	NA	NA	NA	NA	1	1	NA	NA	0	0	0	0	1	NA	NA
duplication_213	LOC100570324	Uncharacterized	0	0	1	0	NA	NA	1	1	NA	1	NA	NA	0	1	NA	0	0	0	NA	NA	NA
duplication_213	LOC100574933	PAX-interacting protein 1-like	1	0	1	1	1	1	1	1	1	1	1	1	NA	NA	1	1	1	1	NA	1	1
duplication_288	LOC100572588	Uncharacterized	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1
duplication_288	LOC103308356	Uncharacterized	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	NA	1	1	1	1
duplication_322	LOC107883251	Uncharacterized	0	0	0	NA	NA	1	NA	1	1	1	1	1	NA	1	1	1	NA	1	1	1	1
duplication_322	LOC107883068	Uncharacterized	1	1	1	1	NA	1	1	1	1	1	1	NA	1	1	1	1	1	1	1	1	1
duplication_340	LOC103310866	Uncharacterized RING finger protein C32D5.10-like	0	0	0	0	0	NA	0	0	0	0	0	0	0	0	0	0	0	0	NA	0	0
duplication_340	LOC100571229	E3 ubiquitin-protein ligase Topors-like	1	0	1	1	1	1	1	1	1	1	1	1	0	NA	NA	0	1	1	NA	NA	1
duplication_397	LOC100162340	Dynein heavy chain 1, axonemal	0	0	1	0	0	0	0	0	0	NA	0	0	0	0	0	NA	0	0	NA	0	0
duplication_397	LOC107882216	Dynein heavy chain 1, axonemal-like	1	0	0	NA	1	0	0	0	NA	0	NA	0	NA	1	0	NA	0	0	NA	NA	NA
duplication_4	LOC100568916	Uncharacterized	0	0	1	NA	1	1	1	1	1	1	1	1	NA	NA	0	0	0	0	NA	NA	NA
duplication_4	LOC100160128	26S proteasome non-ATPase regulatory subunit 12-like	1	0	1	1	1	1	1	1	1	1	1	1	NA	1	1	1	1	1	NA	NA	1
duplication_460	LOC107882168	Zinc finger protein 134-like	0	0	0	0	NA	0	NA	NA	0	0	NA	NA	0	0	0	0	0	0	NA	NA	0
duplication_460	LOC103310004	Zinc finger protein 134-like	1	0	1	0	1	NA	NA	0	0	NA	0	0	0	0	0	0	0	0	NA	NA	0
duplication_489	LOC107882711	Uncharacterized	0	0	NA	1	1	1	1	1	1	1	1	1	NA	1	1	1	0	1	NA	1	1
duplication_489	LOC100162929	Uncharacterized	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
duplication_576	LOC100569858	Uncharacterized	0	0	1	0	0	NA	NA	0	0	0	0	0	NA	0	0	0	0	0	NA	NA	0
duplication_576	LOC100574180	Uncharacterized	1	0	1	0	0	NA	NA	0	NA	NA	0	NA	0	NA	1	NA	0	0	1	1	0
duplication_633	LOC103309550	Zinc finger MYM-type protein 1-like	0	0	0	0	0	0	0	0	0	0	0	0	NA	0	0	0	0	0	NA	NA	0
duplication_633	LOC103307955	Zinc finger MYM-type protein 1-like	1	0	0	NA	NA	0	NA	0	0	0	0	0	NA	NA	0	0	0	0	NA	NA	1
duplication_663	LOC107884962	Uncharacterized	0	0	0	0	0	1	NA	1	NA	NA	1	1	NA	0	0	0	0	0	1	1	NA
duplication_663	LOC100570263	Uncharacterized	1	0	NA	1	1	1	1	1	1	1	1	1	NA	1	NA	0	1	1	NA	1	1
duplication_693	LOC100165085	Uncharacterized	0	0	1	0	0	NA	0	NA	0	NA	0	0	NA	0	NA	NA	NA	NA	NA	NA	1
duplication_693	LOC100165046	Dynein heavy chain 7, axonemal-like	1	1	NA	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
duplication_795	LOC107882727	Sialin-like	0	0	0	NA	0	0	0	0	0	0	0	0	NA	1	1	1	1	1	1	1	1
duplication_795	LOC100164217	Putative inorganic phosphate cotransporter	1	0	1	1	1	NA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
duplication_840	LOC100571978	Uncharacterized	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
duplication_840	LOC107884413	Uncharacterized	1	0	0	0	0	0	NA	NA	0	NA	NA	0	NA	NA	1	1	0	0	NA	NA	0

NOTE.—NA, no information on gene expression available (see Materials and Methods for further details about each condition type); PS, positive selection; BS-M, branch-site test model accounting for MNM; AM, adult males; AP, adult females parthenogenetic; AO, adult female oviparae; E0, embryos stage 17; E1A, embryos stage 18 sex; E1K, embryos stage 18 sex; E2A, embryos stage 19 sex; E2K, embryos stage 19 sex; E3A, embryos stage 20 sex; E3K, embryos stage 20 sex; H, Head\_Mix; HP, head adult female parthenogenetic; HR2, head larvae 2; HR4, head larvae 4; LP, legs adult female parthenogenetic; G, gut; SG, salivary glands; B, bacteriocyte.

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 data set.

Regarding the set of young gene duplicates, we found that 26% of them had different chromatin states in each paralog (i.e., one paralog was detected as open chromatin and the other as closed) (173 out of 666 duplicates, [supplementary table S8, Supplementary Material](#) online; see some examples in [supplementary fig. S16, Supplementary Material](#) online). If we divide young duplicates into “strict duplicates” and “expansions” we observed similar patterns (26% and 27%, respectively). To test whether different expression patterns correlated to different chromatin states in each paralog, we searched for duplicates where each paralog belonged to a different category (i.e., categories [1]–[4] combining chromatin state and expression pattern, as described above) for the embryos and adult morphs, conditions for which FAIRE-Seq data were available ([supplementary table S8, Supplementary Material](#) online). Seventy-three pairs of duplicated genes did belong to different categories. From those, in 54 pairs both duplicates were expressed, with one copy being open and the other closed. In 19 pairs, both duplicates were closed but one gene was expressed and the other one not.

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 study in human tissues ([Ackermann et al. 2016](#)). They discussed that the poor correlation between RNA-Seq and ATAC-Seq data 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 noncompleteness of the pea aphid genome assembly. Moreover, our analyses were centered on putative TSS which have not been validated experimentally. 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 state of promoters of simple pairs of

duplicated genes can evolve independently in each member of the pair. A similar pattern has been observed in nematodes, where differential chromatin states were observed in new pairs of duplicated genes ([Werner et al. 2018](#)).

## Conclusions

Our study shows that recent simple gene duplicates—those formed by two duplicated genes—evolved asymmetrically in *A. pisum*, with one paralogous copy being more conserved and the other more divergent. The conserved copy was under the effect of purifying selection and hardly ever under the effect of positive selection. The divergent copy was usually positively selected and showed a faster evolutionary rate. 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 multiple functions of the pea aphid biology. Remarkably, neofunctionalization may also be involved in symbiosis functioning, facilitating host–symbiont cooperation between *A. pisum* and *Buchnera*. More complex scenarios may have driven the evolution of large expansions, which remain difficult to analyze due to confounding factors, such as possible missannotations and pseudogenes. Concerning the expression patterns of the duplicated genes, we observed that more than one third of the duplicates showed different expression patterns, with some of them being under adaptive selection. This suggests that positive selection might not be the main or the only factor driving such 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 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.

Altogether, 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 genes. 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.

## Materials and Methods

### 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* (Matsumura, 1917) (Sternorrhyncha, Aphidoidea), *M. persicae* (Sulzer, 1776) (Sternorrhyncha, Aphidoidea), and *Rhopalosiphum padi* (Stal, Linnaeus, 1758) (Aphidinae, Aphidini) (fig. 1). Genome versions and number of predicted proteins are indicated in [supplementary table S1, Supplementary Material](#) online.

Phylomes were reconstructed using the PhylomeDB pipeline (Huerta-Cepas, Capella-Gutierrez, et al. 2011). 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 v2.04 [Lassmann and 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 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 approximate likelihood ratio test (aLRT) based on a chi-square distribution. Resulting trees and alignments are stored in PhylomeDB 4.0 (Huerta-Cepas et al. 2014, <http://phylo-medb.org>). The phylomeID is 441 ([http://phylo-medb.org/phylome\\_441](http://phylo-medb.org/phylome_441)).

A species-overlap algorithm, as implemented in ETE v3 (Huerta-Cepas, Dopazo, et al. 2010; Huerta-Cepas, Marcet-Houben, et al. 2010), 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, a maximum likelihood (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).

### Detection and Selection of Gene Duplications

For each gene tree, we first selected with ETE v3 (Huerta-Cepas, Dopazo, et al. 2010; Huerta-Cepas, Marcet-Houben, et al. 2010) the nodes that exclusively contained multiple *A. pisum* sequences. These were considered as species-specific duplications in *A. pisum*. Overlapping species-specific duplications were fused when they shared more than 50% of their members. Trees were then scanned for the presence of pairs of duplicates in *A. pisum* whose duplication node was highly supported (aLRT > 0.95) and which had at least two single-copy orthologs. Note that the selected pairs of duplicates are not limited to genes that duplicated just once, some of them belong to a (larger) expansion in which case the chosen pairs were always at the tips of the tree. 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 coding sequences annotation or unsatisfactory topology were discarded. This resulted in a final number of scrutinized duplications of 843. For each duplication, we obtained multiple protein sequence alignments with PASTA v1.8.3 (Mirab et al. 2015). For each alignment, we computed a gene tree (using the tree-estimator RAXML option) that was used for the codeml analyses (see below) and we estimated the median identity score for each protein sequence in the alignment compared with all the others using trimAl v1.3 (-sident option; Capella-Gutierrez et al. 2009). We finally back-translated protein multiple sequence alignments into nucleotide with trimAl (using -phylip\_paml -nogaps -backtran options). GC was estimated from back-translated sequences using GENECONV software (Sawyer 1989), by considering fragments with evidence of a GC event between the ancestors of two *A. pisum* sequences that remained significant after multiple-comparison correction.

### dN/dS and Additional Filtering

We estimated the number of synonymous substitutions per synonymous site (dS), the number of nonsynonymous substitutions per nonsynonymous site (dN) and dN/dS ratio

using the “free ratio branch model” implemented in codeML from PAML v. 4.9 (Yang 2007), using model = 1, CodonFreq = 3, and Nsites = 0 as options. This software allows to estimate dS, dN, and dN/dS for each internal and terminal branch of a given tree and also to reconstruct the ancestral sequence before the duplication occurred. Analyses were computed for the 1,020 selected duplications, which contained a specific duplication in *A. pisum* and at least two single-copy orthologs for any of the other eight species included in the phylome. We noticed that dS for the three more distant species was much higher (the percentage of sequences with dS > 2 was 82.1%, 86.2%, and 63.9% in *Diaphorina citri*, *Bemisia tabaci*, and *D. vitifoliae*, respectively) than in the closely related species (16.5%, 6.4%, 5.8%, 3.3%, 4.4%, and 0.7% in *C. cedri*, *Aphis glycines*, *Rhopalosiphum padi*, *Diuraphis noxia*, *M. persicae* and *A. pisum*, respectively). Since such large dS values may indicate problems in the orthology identification, we discarded duplications with only single-copy orthologs in the three most distant species. A total of 843 duplications remained after this filtering.

### 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. From the total of 5,589 genes from six species in 843 duplications, we filtered out genes with dS > 2 (which may indicate problems in the orthology identification, 242 genes) and dS < 0.01 (which may lead to high dN/dS ratios with no biological sense, 908 genes). 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.

### Classification of Gene Duplications

Gene duplicates can be divided into two groups: strict duplicates (606), including *A. pisum* genes that derive from a recent common ancestor and duplicated specifically in this species only once (i.e., genes with only two copies) and expansions (237), including *A. pisum* genes that also derive from a recent common ancestor but duplicated multiple times (i.e., genes with more than two copies). In addition, duplicates were further classified as tandem (defined as duplicates with no genes in between; 94 in total), located in the same or different contig (dispersed duplicates, 100 and 605, respectively) and retrotransposed (defined as dispersed duplicates in which one copy lacks introns; 44 in total).

### 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, and 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, and omega = 1.5). The two models were compared using a

likelihood ratio test, and *P* values were adjusted for multiple comparisons using the Holm, Hochberg, SidakSS, SidakSD, BH, and BY methods using multtest package for R. We considered that a given gene is under selection if any of the adjusted *P* values computed using the different methods was <0.01.

Selection tests taking into account MNMs were explored in HyPhy (Pond et al. 2020) following the pipeline described by Venkat et al. (2018). Null and alternative hypotheses were compared as described before.

We obtained population data from three different *A. pisum* host races specialized on different crops: alfalfa (population size = 60), genista (population size = 34), and pea (population size = 60) (Grigorescu et al. 2018; Nouhaud et al. 2018). To compute Tajima’s *D* for each gene, we mapped the reads with bwa and generated a pileup with samtools mpileup (Etherington et al. 2015). Polymerase chain reaction duplicates were removed and the primary alignments with a mapping quality higher than 20 were kept. From the pileup file, we run the script subsample-pileup.pl from Popoolation (Kofler et al. 2011), with the option –target-coverage 15 –max-coverage 150 –method withoutreplace –fastq-type sanger. Then, Tajima’s *D* was calculated using the script “Variance-at-position.pl –measure *D*” from Popoolation, with the parameter pool size set to 120, 120, and 68 for the alfalfa, pea, and genista sagittalis, respectively, on a GTF file with all the protein coding genes in the *A. pisum* genome. We computed Tajima’s *D* for whole genes, exons, and introns.

The codon adaptation index for each gene of interest was estimated using CaiCAL (Puigbò et al. 2008). The codon usage table for *A. pisum* was obtained from CoCoPUTs (Athey et al. 2017). CaiCAL was also used to calculate the expected Codon Adaptation Index (CAI) value based on a 1,000 randomly created sequences. Genes with a CAI value above the eCAI value were considered as optimized.

### Functional Annotation and GO Term Enrichment Analysis and Visualization

To assign 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 prioritized coverage (i.e., GO terms were propagated if any type of orthologs to a gene in a genome was detected). See supplementary table S9, Supplementary Material online, for the full annotation of the selected genes. Functional enrichment of the selected duplications 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., nonexpanded genes and nonpositively selected ones, respectively). Sets of GO terms were summarized and visualized in REVIGO (Supek et al. 2011).

### Tissue Expression Diverge between Duplicates

Messenger RNA (mRNA) expression data were obtained from 106 different samples from the *A. pisum* LSR1 lineage. They correspond to RNA-Seq libraries from 18 different conditions. Some of them were retrieved from the public databases and

others newly generated for this study (supplementary table S2, Supplementary Material online). All samples 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 = 50,000`, and `alignMatesGapMax = 50,000`. 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`. TPMs were calculated from RPKM using the equation  $TPM(i) = [FPKM(i)/\sum (FPKM \text{ all transcripts})] \times 10^6$ . Principal component analysis was performed using `prcomp` function from R.

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 consisted of values located in the lowest of the four bins. The value “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: 1) 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 its value is NA for either gene. 2) Tissue expression complementarity distance (Huerta-Cepas, Dopazo, et al. 2011), 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. 3) Tissue expression divergence (dT) (Pegueroles et al. 2013), which subtracts tissues where one or two copies are expressed from tissues where the two copies are expressed divided by tissues where one or two copies are expressed. Values for the three distances range from 0 to 1, where 0 means no differences in gene expression 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. In addition, for each replicate of a given tissue, we computed the median expression value in TPM. We then rescaled the expression across the tissues using the `rescale` function from `plyr` package from R and subsequently we calculated the median expression value, the Pearson correlation and its *P* value, and the  $r^2$

and the slope of a linear model using gene expression across the 18 tissues.

### FAIRE-Seq Data Analysis

FAIRE-Seq data for samples for males and females adults were 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 and Salzberg 2012) on the pea aphid genome assembly Acyr\_2.0 (ID NCBI: 246238, AphidBase: [http://bipaa.genouest.org/is/aphidbase/acyrthosiphon\\_pisum/](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 irreproducible discovery rate (IDR) recommendations (Li et al. 2011, <https://sites.google.com/site/anshulkundaje/projects/idr#TOC-Latest-pipeline>). 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, 0.02 for self-consistency replicates, and 0.0025 for pooled pseudoreplicates. Replicates consistency was then assessed using the 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  $\log_2$  (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  $\log_2$  (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: 1) open and expressed, 2) open and not expressed, 3) closed and expressed, and 4) closed and not expressed. For each gene, the percentage of tissues representing each category was calculated. If this average reached at least 75%, a single category was assigned to the gene. In total, 13,858 genes were assigned to one of the categories (supplementary table S6, Supplementary Material online).

### Supplementary Material

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

### Acknowledgments

R.F. was funded by a Juan de la Cierva-Incorporación Fellowship (Government of Spain, IJCI-2015-26627) and a

Marie Skłodowska-Curie Fellowship (747607). T.G. 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 cofunded 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 program under the grant agreement ERC-2016-724173 and the Marie Skłodowska-Curie (Grant Agreement No. H2020-MSCA-ITN-2014-642095). D.T. was funded by “Severo Ochoa visiting scientific program” for a 6 months stay at the Center for Genomic Regulation—Barcelona—to start the project, supported as well by INRA SPE. Dr Roderic Guigo (CRG, Barcelona) is warmly acknowledged for his welcome in his lab. Dr Akiko Sugio, Dr Julie Jaquiéry, Dr Gaël Le Trionnaire, and Dr Jean-Christophe Simon (INRAE, UMR 1349 Igepp, Rennes, France) are acknowledged for 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 Bioresources) and BGI Biotech in the frame of i5k initiative. Parts of the transcriptomic resources were obtained within the 1KITE projects (Bernhard Misof, Bonn, Germany). The *Acyrtosiphon pisum* phylome can be accessed at PhylomeDB 4.0 under phylome number 441. RNA-Seq data are accessible at NCBI, see [supplementary table S2, Supplementary Material](#) online.

### Author Contributions

R.F., M.M.-H., and C.P. designed research, analyzed data, wrote the manuscript, and prepared figures and tables. F.L. and S.R. performed bioinformatic analyses on RNA-Seq methods. G.R. produced FAIRE-Seq data and participated in the analysis and discussion of that section. V.W. was involved in the early steps of RNA-Seq analyses. T.G. and D.T. designed and supervised research, coordinated the production of data, and supervised the writing of the manuscript.

### 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. *Mol Metab.* 5(3):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(4):578–580.
- Athey J, Alexaki A, Osipova E, Rostovtsev A, Santana-Quintero LV, Katneni U, Simonyan V, Kimchi-Sarfaty C. 2017. A new and updated resource for codon usage tables. *BMC Bioinformatics* 18(1):391.
- Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972–1973.
- Chang A-F, Liao B-Y. 2012. DNA methylation rebalances gene dosage after mammalian gene duplications. *Mol Biol Evol.* 29(1):133–144.
- Conant GC. 2003. Asymmetric sequence divergence of duplicate genes. *Genome Res.* 13(9):2052–2058.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1):15–21.
- Duncan RP, Feng H, Nguyen DM, Wilson A. 2016. Gene family expansions in aphids maintained by endosymbiotic and nonsymbiotic traits. *Genome Biol Evol.* 8(3):753–764.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with improved accuracy and speed. In: CSB 2004. Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference; 2004. Available from: <http://dx.doi.org/10.1109/csb.2004.1332560>.
- Etherington GJ, Ramirez-Gonzalez RH, MacLean D. 2015. bio-samtools 2: a package for analysis and visualization of sequence and alignment data with SAMtools in Ruby: fig. 1. *Bioinformatics* 31(15):2565–2567.
- Grigorescu AS, Renoz F, Sabri A, Foray V, Hance T, Thonart P. 2018. Accessing the hidden microbial diversity of aphids: an illustration of how culture-dependent methods can be used to decipher the insect microbiota. *Microb Ecol.* 75(4):1035–1048.
- Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 59(3):307–321.
- Han MV, Demuth JP, McGrath CL, Casola C, Hahn MW. 2009. Adaptive evolution of young gene duplicates in mammals. *Genome Res.* 19(5):859–867.
- Hansen AK, Moran NA. 2011. Aphid genome expression reveals host-symbiont cooperation in the production of amino acids. *Proc Natl Acad Sci U S A.* 108(7):2849–2854.
- Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, Denisov I, Kormes D, Marcet-Houben M, Gabaldón T. 2011. PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res.* 39(Database):D556–D560.
- Huerta-Cepas J, Capella-Gutiérrez S, Pryszcz LP, Marcet-Houben M, Gabaldón T. 2014. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.* 42(D1):D897–D902.
- Huerta-Cepas J, Dopazo J, Gabaldón T. 2010. ETE: a python Environment for Tree Exploration. *BMC Bioinformatics* 11(1):11.
- Huerta-Cepas J, Dopazo J, Huynen MA, Gabaldón T. 2011. Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication. *Briefings Bioinf.* 12(5):442–448.
- Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol Biol Evol.* 34(8):2115–2122.
- Huerta-Cepas J, Marcet-Houben M, Pignatelli M, Moya A, Gabaldón T. 2010. The pea aphid phylome: a complete catalogue of evolutionary histories and arthropod orthology and paralogy relationships for *Acyrtosiphon pisum* genes. *Insect Mol. Biol.* 19:13–21.
- Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, et al. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47(D1):D309–D314.
- Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet.* 11(2):97–108.
- International Aphid Genomics Consortium. 2010. Genome sequence of the pea aphid *Acyrtosiphon pisum*. *PLoS Biol.* 8:e1000313.
- Katoh K. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 33(2):511–518.
- Keller TE, Yi SV. 2014. DNA methylation and evolution of duplicate genes. *Proc Natl Acad Sci U S A.* 111(16):5932–5937.
- Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, Kosiol C, Schlötterer C. 2011. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One* 6(1):e15925.
- Kumar S, Stecher G, Suleski M, Heddes SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol.* 34(7):1812–1819.

- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 9(4):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 Biol*. 10(3):R25.
- Lassmann T, Sonnhammer E. 2005. Kalign—an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics* 6(1):298.
- Le SQ, Lartillot N, Gascuel O. 2008. Phylogenetic mixture models for proteins. *Philos Trans R Soc B* 363(1512):3965–3976.
- Lercher MJ, Blumenthal T, Hurst LD. 2003. Coexpression of neighboring genes in *Caenorhabditis elegans* is mostly due to operons and duplicate genes. *Genome Res*. 13(2):238–243.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.
- Li Q, Brown JB, Huang H, Bickel PJ. 2011. Measuring reproducibility of high-throughput experiments. *Ann Appl Stat*. 5(3):1752–1779.
- Li Y, Park H, Smith TE, Moran NA. 2019. Gene family evolution in the pea aphid based on chromosome-level genome assembly. *Mol Biol Evol*. 36(10):2143–2156.
- Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30(7):923–930.
- Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290(5494):1151–1155.
- Mansai SP, Kado T, Innan H. 2011. The rate and tract length of gene conversion between duplicated genes. *Genes* 2(2):313–331.
- Marlétaz F, Firbas PN, Maeso I, Tena JJ, Bogdanovic O, Perry M, Wyatt CDR, de la Calle-Mustienes E, Bertrand S, Burguera D, et al. 2018. Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature* 564(7734):64–70.
- Mathers TC, Chen Y, Kaithakottil G, Legeai F, Mugford ST, Baa-Puyoulet P, Bretaudeau A, Clavijo B, Colella S, Collin O, et al. 2017. Rapid transcriptional plasticity of duplicated gene clusters enables a clonally reproducing aphid to colonise diverse plant species. *Genome Biol*. 18(1):27.
- McCarthy DJ, Chen Y, Smyth GK. 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res*. 40(10):4288–4297.
- Mirarab S, Nguyen N, Guo S, Wang L-S, Kim J, Warnow T. 2015. PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *J Comput Biol*. 22(5):377–386.
- Nouhaud P, Gautier M, Gouin A, Jaquiéry J, Peccoud J, Legeai F, Mieuzyet L, Smadja CM, Lemaitre C, Vitalis R, et al. Forthcoming 2018. Identifying genomic hotspots of differentiation and candidate genes involved in the adaptive divergence of pea aphid host races. *Mol Ecol*. 27(16):3287–3300.
- Pegueroles C, Laurie S, Albà MM. 2013. Accelerated evolution after gene duplication: a time-dependent process affecting just one copy. *Mol Biol Evol*. 30(8):1830–1842.
- Pich I, Roselló O, Kondrashov FA. 2014. Long-term asymmetrical acceleration of protein evolution after gene duplication. *Genome Biol Evol*. 6(8):1949–1955.
- Pond SLK, Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Lance Hepler N, Murrell B, Shank SD, Magalis BR, Bouvier D, et al. 2020. HyPhy 2.5—a customizable platform for evolutionary hypothesis testing using phylogenies. *Mol Biol Evol*. 37(1):295–299.
- Puigbò P, Bravo IG, Garcia-Vallve S. 2008. CAlcal: a combined set of tools to assess codon usage adaptation. *Biol Direct* 3(1):38.
- Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res*. 44(W1):W160–W165.
- Richard G. 2017. Régulations chromatiniques et transcriptionnelles impliquées dans le cycle de vie du puceron du pois. Available from: <http://www.theses.fr/2017NSARB130>.
- Richard G, Legeai F, Prunier-Leterme N, Bretaudeau A, Tagu D, Jaquiéry J, Le Trionnaire G. 2017. Dosage compensation and sex-specific epigenetic landscape of the X chromosome in the pea aphid. *Epigenet Chromatin* [Thesis]. France: University Rennes 1.
- Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1):139–140.
- Sandve SR, Rohlfis RV, Hvidsten TR. 2018. Subfunctionalization versus neofunctionalization after whole-genome duplication. *Nat Genet*. 50(7):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 Dev*. 25(22):2361–2373.
- Sawyer S. 1989. Statistical tests for detecting gene conversion. *Mol Biol Evol*. 6(5):526–538.
- Scannell DR, Wolfe KH. 2007. A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Res*. 18(1):137–147.
- Simon J-C, Pfreder ME, Tollrian R, Tagu D, Colbourne JK. 2011. Genomics of environmentally induced phenotypes in 2 extremely plastic arthropods. *J Hered*. 102(5):512–525.
- Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6(7):e21800.
- Vellichirammal NN, Madayiputhiya N, Brisson JA. 2016. The genome-wide transcriptional response underlying the pea aphid wing polyphenism. *Mol Ecol*. 25(17):4146–4160.
- Venkat A, Hahn MW, Thornton JW. 2018. Multinucleotide mutations cause false inferences of lineage-specific positive selection. *Nat Ecol Evol*. 2(8):1280–1288.
- Wallace IM, O'Sullivan O, Higgins DG, Notredame C. 2006. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res*. 34(6):1692–1699.
- Wehe A, Bansal MS, Burleigh JG, Eulenstein O. 2008. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics* 24(13):1540–1541.
- Werner MS, Sieriebriennikov B, Prabh N, Loschko T, Lanz C, Sommer RJ. 2018. Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation. *Genome Res*. 28(11):1675–1687.
- Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24(8):1586–1591.
- Zhang J. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* 22(12):2472–2479.
- Zhang Y, Liu T, Meyer CA, Eickhout J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 9(9):R137.