Received: 7 February 2019 | Revised: 24 September 2019 | Accepted: 30 September 2019

DOI: 10.1111/mec.15268

ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

Molecular signatures of selection associated with host plant differences in *Pieris* butterflies

Yu Okamura[1,2] | Ai Sato[2] | Natsumi Tsuzuki[2] | Masashi Murakami[2] | Hanna Heidel-Fischer[1,3] | Heiko Vogel[1]

[1]Department of Entomology, Max Planck Institute for Chemical Ecology, Jena, Germany

[2]Community Ecology Lab, Faculty of Science, Chiba University, Chiba, Japan

[3]Leibniz Institute for Natural Product Research and Infection Biology, Hans Knöll Institute (HKI), Jena, Germany

Correspondence
Yu Okamura, Department of Entomology, Max Planck Institute for Chemical Ecology, Hans-Knöll-Str. 8, 07745 Jena, Germany.
Email: 0707yu@gmail.com

Funding information
Japan Society for the Promotion of Science, Grant/Award Number: 15J00320

Abstract

Adaptive traits that enable organisms to conquer novel niches and experience subsequent diversification are ecologically and evolutionarily important. The larvae of *Pieris* butterflies express nitrile-specifier proteins (*NSP*s), a key innovation for overcoming the glucosinolate (GLS)-myrosinase-based defence system of their Brassicales host plants. Nitrile-specifier proteins are a member of the *NSP*-like gene family, which includes the major allergen (*MA*) protein, a paralog of *NSP* with a GLS-disarming function, and a single domain major allergen (*SDMA*) protein, whose function is unknown. The arms-race between GLS-based defences and the *NSP*-like gene family is suggested to mediate diversification in both Pierid butterflies and Brassicales plants. Here, we tested whether the expected strong selection on *NSP*-like gene family correlates with shifts in host plant spectra among Pierid butterflies. We combined feeding experiments using 25 Brassicaceae plants and five *Pieris* species with larval transcriptome data to investigate the patterns of selection acting on *NSP*-like gene family members. Although we observed significantly elevated nonsynonymous to synonymous substitution rate ratios in *NSP*s on branches associated with changes in patterns of host plant usage, no such pattern was observed in *MA*s or *SDMA*s. Furthermore, we found evidence for positive selection of *NSP* at a phylogenetic branch which reflects different host plant spectra. Our data indicate that the *NSP*-related gene members have evolved differently: *NSP*s have accumulated more amino acid changes in response to shifting preferences for host plants, whereas *MA*s and *SDMA*s appear to be more conserved. Further detailed functional assays of these genes would provide important insights to understand their role in the chemical arms-race between *Pieris* butterflies and their Brassicales host plants.

KEYWORDS

arms-race, host plant adaptation, insects, selection

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2019 The Authors. Molecular Ecology published by John Wiley & Sons Ltd

1 | INTRODUCTION

Key innovations that enable organisms to acquire novel niches and experience subsequent radiation are ecologically and evolutionarily important (Bond & Opell, 1988; Hunter, 1998). In plant–herbivore interactions, a number of key innovations were identified that enabled herbivores to overcome specific plant defence mechanisms and colonize novel host plants. For instance, gene duplications in the cytochrome P450 family are known to enable adaptation of *Papilio* butterflies to furanocoumarin-based defences, and *Plutella xylostella* larvae utilize glucosinolate (GLS) sulfatase enzymes to disarm the major defence system of their Brassicales host plants (Berenbaum, Favret, & Schuler, 1996; Heidel-Fischer et al., 2019; Janz, 2011; Ratzka, Vogel, Kliebenstein, Mitchell-Olds, & Kroymann, 2002; Wheat et al., 2007).

Pieris butterfly larvae also feed on plants containing GLSs, redirecting toxic breakdown products to less toxic metabolites using gut-expressed nitrile-specifier proteins (*NSP*s) (Wittstock et al., 2004). Nitrile-specifier proteins are known to be a key innovation of *Pieris* butterflies: the acquisition of *NSP*s enabled *Pieris* to colonize GLS-containing Brassicales, followed by higher speciation rates compared to those of sister butterfly clades.

Nitrile-specifier proteins are members of the small *NSP*-like gene family, which includes major allergen (*MA*) proteins and single domain major allergen (*SDMA*) proteins. Although the function of *MA* and *SDMA* is mostly unclear, the structures of *MA* and *NSP* are known to be similar: both proteins contain three replicated domains which originated from *SDMA* (Fischer, Wheat, Heckel, & Vogel, 2008). In addition, although *SDMA* is generally expressed in the guts of Lepidopteran larvae, *NSP* and *MA* are only found in Pierid butterflies feeding on Brassicales (Fischer et al., 2008; Randall, Perera, London, & Mueller, 2013). These findings suggest that in *Pieris*, *MA*s, like *NSP*s, have a function related to disarming GLSs. The ability of *MA* to redirect GLS hydrolysis was recently documented in one Brassicales-feeding Pierid, *Anthocharis cardamines*, which seems to have *MA* genes only, that is, it lacks *NSP* (Edger et al., 2015; In this study, "*NSP*" was used to name *NSP*-like gene family members, including *MA*). Furthermore, *MA* (but not *NSP*) expression is known to be upregulated in *Pieris* larvae responding to the presence of certain types of GLSs (Okamura, Sato, et al., 2019a). Thus, although the function of *MA* in Pieridae is largely unknown, especially in those species which have *NSP*s and *MA*s, *MA*s also appear to be ecologically important for overcoming the host plant's GLS-based defence system.

Previous studies indicated that the co-evolutionary diversification of Brassicales plants and Pierid butterflies was mediated by the chemical arms-race between the glucosinolate-myrosinase defence system and members of the *NSP*-like gene family (Edger et al., 2015). Past increases of GLS complexity in Brassicales were followed by frequent gene birth–death events of *NSP*-like gene family members in Pierid butterflies. This suggests that members of the *NSP*-like gene family would potentially be under strong selection pressure, were Pieridae butterflies to expand or shift their host plants. Such a scenario is supported by recent findings of signatures of positive selection in partial *NSP* sequences of a pair of *Pieris* butterflies in comparison with the signatures of 70 randomly selected genes (Heidel-Fischer, Vogel, Heckel, & Wheat, 2010). However, the evolutionary forces acting on all *NSP*-like gene family members, especially when considering the associated host plant spectrum, remain unknown.

Besides *NSP*-like gene family members, a number of detoxification-related genes are either hypothesized or were shown to be directly involved in overcoming chemical challenges of host plants, such as glutathione S-transferases, UDP-glycosyltransferases or cytochrome P450 enzymes (Feyereisen, 2012; Krempl et al., 2016; Simon et al., 2015). The expression patterns of numerous putative detoxification-related genes in larvae feeding on different host plants have been broadly tested in both specialist and generalist herbivores (Celorio-Mancera et al., 2016; Heidel-Fischer et al., 2009; Mao et al., 2007; Nallu et al., 2018). However, in most of these cases there is a lack of data pertaining to field-observed host plant associations (or associations of the larvae with specific groups of secondary metabolites). We thus not only need reliable host plant data but also more in-depth analyses of enzymatic activities and patterns of selection of these host plant chemistry-induced genes.

Here, we focus on five Japanese butterfly species (*Pieris napi*, *P. melete*, *P. rapae*, *P. brassicae* and *P. canidia*) in the genus *Pieris*, which has both *NSP* and *MA* genes and feed on Brassicaceae plants

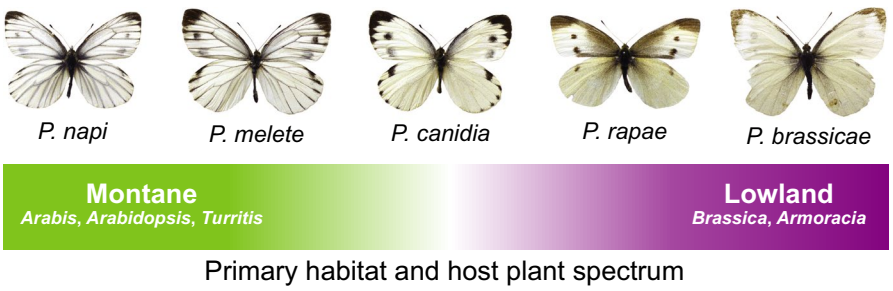

FIGURE 1 Field observations of primary habitat and larval host plant spectra of five *Pieris* butterflies in Japan. *Pieris napi* and *Pieris melete* tend to be found in montane habitat and rely mostly on Brassicaceae plants in forests; these include *Arabis*, *Arabidopsis* or *Turritis*. *Pieris rapae* and *Pieris brassicae* are known as *Brassica* crop pests. In Japan, *Pieris canidia* can only be found in a restricted area and uses *Cardamine* or *Lepidium* as host plants [Colour figure can be viewed at wileyonlinelibrary.com]

with the highest GLS diversity among the Brassicales. The five *Pieris* species have different host plant spectra according to field observations (Figure 1), with *P. napi* and *P. melete* frequently using wild Brassicaceae plants (such as *Arabis* or *Arabidopsis*), whereas *P. rapae* and *P. brassicae* tend to feed on Brassicaceae crops and are known as major pests (Benson, Pasquale, Van Driesche, & Elkinton, 2003; Kitahara, 2016; Ohsaki & Sato, 1994; Ueno, 1997). In contrast, in Japan, *P. canidia* can be found only in the southern islands (Yonaguni Island, Okinawa), relying on the limited number of host plants, such as *Cardamine* or *Lepidium*, in their habitat range. We aim to identify patterns of selection of *NSP*-like gene family members correlating with different host plant spectra among the five *Pieris* species used in this study (Figure 1).

To this end, we conducted feeding experiments with 25 Brassicaceae plants to acquire patterns of host plant utilization in *Pieris* species. With larval transcriptome (RNA-seq) data from the five *Pieris* species, we analysed the divergence in amino acid sequences of orthologs based on nonsynonymous (dN) and synonymous substitution (dS) rates. We investigated signatures of selection on members of the *NSP*-like gene family compared with other larval-expressed orthologs. We also performed tests to detect evidence of positive selection in *NSP*-like gene family members. Additionally, we searched for potential genes more generally related to host plant detoxification with signatures of selection which correlate with the observed larval performance based on gene ontology (GO) and dN/dS analyses. By combining these approaches, we were able to investigate whether there are correlations between host plant spectra and signatures of selection on ecologically important *NSP*-like gene family members or other detoxification-related genes in *Pieris* (Figure 2). The obtained results provide important insights into the evolution of adaptive key innovations in *Pieris* butterflies.

2 | MATERIALS AND METHODS

2.1 | Feeding experiments

We used four *Pieris* butterfly species for the feeding assay, leaving out *P. canidia*, which is endemic and rather rare in Japan. We collected 7–10 female butterflies of three *Pieris* butterfly species (*P. napi*, *P. melete*, *P. rapae*) from wild populations in Chiba and Hokkaido, Japan. Most wild-caught female butterflies were already fertilized. We released the female butterflies into cages containing cabbage (*Brassica oleracea* var. *capitata*) or *Cardamine leucantha* under high-intensity light conditions and waited for eggs to be laid. For *P. brassicae*, final-instar larvae were caught in the wild (Hokkaido, Japan), fed on cabbage and reared to the adult stage. After eclosion, 10 female butterflies were hand-paired with males and eggs were collected as they were from the other species. Eggs of the four *Pieris* butterfly species were incubated at 25°C until they hatched.

For experimental plants, we collected seeds of 25 Brassicaceae plant species, covering a phylogenetically broad range (Table S1) (Beilstein, Al-Shehbaz, Mathews, & Kellogg, 2008; Couvreur et al., 2010; Franzke, Lysak, Al-Shehbaz, Koch, & Mummenhoff, 2011). The seeds of 19 Brassicales plant species were collected from the

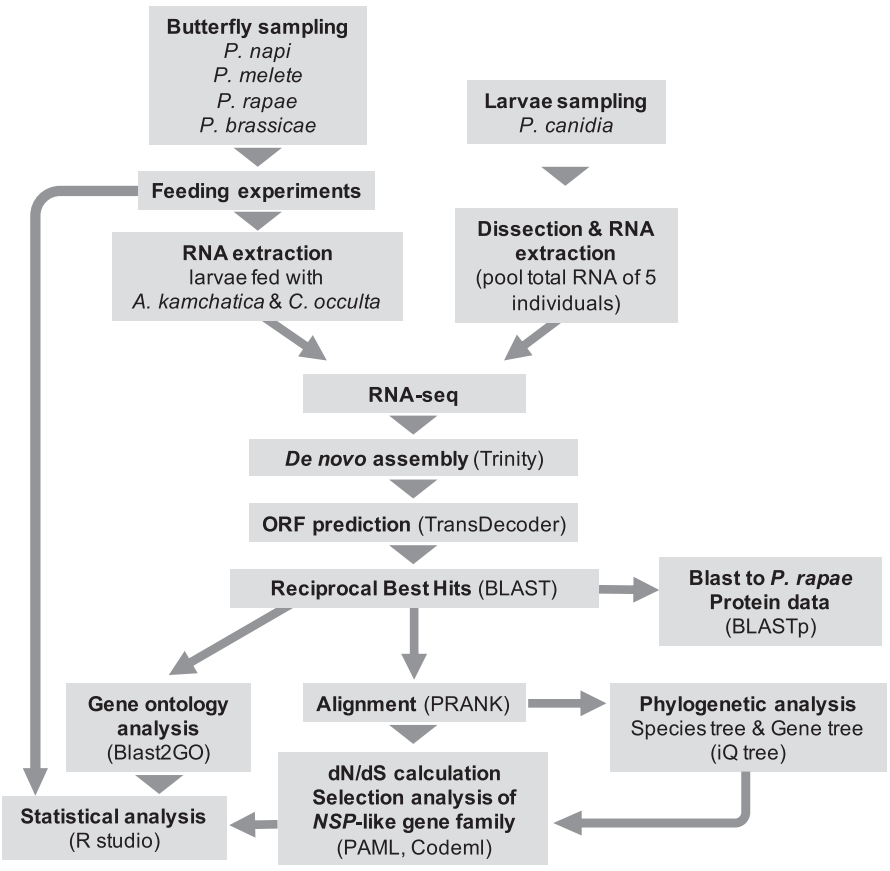

FIGURE 2 Analysis pipeline used to compare dN/dS ratios of *NSP*-like gene family members with all observed ortholog sets from the reciprocal best hit using BLAST across five *Pieris* butterflies. Signatures of selection on *NSP*-like gene family members were investigated in each phylogenetic branch and compared with the results of the feeding assay

wild and the others were acquired from commercial suppliers (Table S1). We grew the plants in the greenhouse at 25°C, with 60% relative humidity and L16:D8. Plants were watered and fertilized every week with a 2,000× diluted solution of Hyponex (N:P:K = 6:10:5; Hyponex, Osaka, Japan). After 2 months of cultivation, plants were used for the feeding experiments.

Neonate larvae were collected within 12 hr after they hatched for the feeding experiment. We transferred three neonate larvae to each of two plants per plant species using a soft-haired brush ($n = 6$). To minimize changes in the condition of the experimental plants, experimental trials were carried out within 5 days for all four *Pieris* species. We conducted feeding experiments under the same temperature and light conditions used for plant growth. We measured the weight of each larva individually (within 0.1 mg) after 120 hr of feeding. Since there was no significant difference of larval performance between the two plants replicates (ANOVA; $p \geq .05$), we used the average weight of larval individuals from each plant species as an index of the performance of each *Pieris* butterfly species.

Larval weights were standardized as z-scores to enable comparison between species. We calculated the mean scores of each plant treatment and used these for the comparative analysis. We conducted Pearson's correlation test and hierarchical clustering analysis to assess differences in larval performances among the four *Pieris* species. The possible clustering was evaluated with the gap statistics (Tibshirani, Walther, & Hastie, 2001). All of these analyses were performed on R studio ver. 1.1.453 (RStudioTeam, 2016).

2.2 | RNA sequencing

From four *Pieris* butterfly species (*P. napi*, *P. melete*, *P. rapae* and *P. brassicae*), excluding *P. canidia*, we collected larvae that we used for the feeding experiments for transcriptome analysis (Figures 1 and 2). We used larvae that fed on *Arabidopsis kamchatica* and *Cardamine occulta* as representatives. The larvae were flash-frozen in liquid nitrogen and stored at −80°C until RNA extraction. We selected a single representative larva for each of the four *Pieris* and plant species combinations, and RNA was extracted using the RNeasy Mini Kit (QIAGEN). RNA sample quantity and quality were checked by Agilent 2100 Bioanalyzer. Illumina libraries of individual larva were prepared by Sure Select Strand-Specific RNA Library Preparation Kit for Illumina Multiplexed Sequencing, and RNA sequencing was performed on an Illumina HiSeq 1500 Genome Analyzer platform using a 2 × 100 bp paired-end approach. For *P. canidia*, we collected larvae directly from wild *Lepidium virginicum* on Yonaguni Island, Okinawa, Japan. The collected larvae were dissected, and gut tissues were stored at −80°C in solution until RNA extraction. Five larvae were randomly selected, and RNA was extracted with the RNeasy Mini Kit (QIAGEN). *Pieris canidia* RNA concentrations were quantified on a Qubit 2 Fluorometer (Invitrogen), and a fraction of the RNA from each of the five larvae was pooled as a single sample for RNA-seq. Paired-end (2 × 150 bp) sequencing was performed by the Max Planck Genome Center Cologne on an Illumina HiSeq 2500 Genome Analyzer platform.

2.3 | De novo assembly, searching for reciprocal best hits (RBHs) using BLAST

Acquired reads of RNA-seq data were pooled for each species after filtering out bad quality reads by trimmomatic with the following options (LEADING:10 TRAILING:10 SLIDINGWINDOW:4:20 MINLEN:40) (Bolger, Lohse, & Usadel, 2014). The quality of reads was checked by FastQC (Andrews, 2010). Pooled reads were de novo assembled by Trinity ver. 2.0.6 (Grabherr et al., 2011). We used TransDecoder (http://transdecoder.github.io/) to predict open reading frames (ORFs) from the assembled contigs and subsequently looked for reciprocal best hits (RBHs) using BLAST alignment methods to analyse amino acid sequences (longer than 100 amino acids) predicted by TransDecoder (Camacho et al., 2009; Cock, Chilton, Grüning, Johnson, & Soranzo, 2015). We used RBH BLAST software with default settings (minimum percentage identity for BLAST matches = 0.7, minimum percentage query coverage for BLAST matches = 0.5) on all possible species pairs (10 pairs) and subsequently extracted *P. rapae* orthologs from this RBH result and ran blastp on the amino acid sequences against a *P. rapae* protein database to confirm the ORF prediction from TransDecoder. Orthologs in the RBH result without any BLAST hits to the *P. rapae* protein database (Shen et al., 2016) were removed since these amino acid sequences may have resulted from wrong ORF predictions by TransDecoder. We used PRANK to conduct codon-based alignment of each ortholog set acquired from the RBH result (Loytynoja & Goldman, 2005). Since the *P. canidia* sample was obtained from dissected gut tissue only, the entire RBH result was likely biased to gut-expressed proteins, that is the interface between larvae and their plant diet.

2.4 | Phylogenetic tree construction

We reconstructed an unrooted phylogeny of the five *Pieris* species using the transcriptome data by concatenating all aligned orthologous nucleotide sequences into one sequence for each species, generating an maximum-likelihood (ML) phylogenetic tree by IQ tree (Nguyen, Schmidt, Von Haeseler, & Minh, 2015) after removing gaps with TrimAl (2,063,074 bp remaining) (Capella-Gutiérrez, Silla-Martínez, & Gabaldón, 2009). We used the GTR + gamma substitution model and set ultrafast bootstrap approximation iterations as 1,000, using -bnni options to construct a phylogeny of the five *Pieris* species (Hoang, Chernomor, Von Haeseler, Minh, & Vinh, 2018).

2.5 | Comparing patterns of divergence of *NSP*-like gene family members with other *Pieris* species orthologs

We used the acquired unrooted tree for estimating dN/dS ratios of all the orthologs at each branch using PAML 4.8 (Yang, 2007). We used runmode = 0, model = 1 and NSsites = 0 option in codeml implemented in PAML and estimated dN/dS ratios using the ML method. The estimated dN/dS values of *NSP*-like gene family members were

compared with the entire dN/dS distributions of all ortholog sets in each phylogenetic branch. We discarded orthologs which had estimated dS below 0.01 from this analysis, since too low dS values can cause unreliable dN/dS value estimation (Villanueva-Cañas, Laurie, & Alba, 2013).

2.6 | Tests for positive selection on *NSP*-like gene family members

We used the branch-site model test (Zhang, Nielsen, & Yang, 2005) to identify cases of positive selection on specific sites of *NSP*-like gene family members at a specific branch. We prepared ML molecular phylogeny of a combined ML gene tree of *NSP* and *MA* from our RNA-seq data with additional sequences of *MA*s from *Anthocharis* spp. and *Pontia* spp. (Accession nos: EU137117.1, EU137133.1, EU137132.1) using IQtree. Regarding *SDMA*, we added *SDMA* sequences from *A. cardamines*, *Eucheira socialis*, *Dixeia pigea*, *Colias eurytheme*, *P. xylostella* (Accession nos: EU137118.1, EU137122.1, EU137121.1, EU137119.1, EU137131.1) for generating an *SDMA* gene tree. We tested all the branches in *Pieris* using codeml model 2 with NSsites = 2 option and ran an alternative model: varied dN/dS ratios across sites as well as lineages were allowed (fixed_omega = 0), and null model: fixed dN/dS (fixed_omega = 1). We conducted a likelihood-ratio test (LRT) with the chi-square distribution to evaluate significant differences between the alternative and null models. Acquired *p* values were corrected with false discovery rates (FDRs) in each analysis. Signs of positive selection on each site were identified by the Bayes empirical Bayes (BEB) analysis (.90 cut-offs).

Since the branch-site model in codeml can cause false positives in case of multinucleotide mutations (MNMs, Venkat, Hahn, & Thornton, 2018), we also performed more conservative branch-site model tests covering MNM situations (BS + MNM). In BS + MNM, the additional parameter δ represents the relative instantaneous rate of double mutations compared to that of single mutations. We ran null models and alternative models in BS + MNM and conducted LRTs to evaluate significance.

2.7 | GO categories with elevated dN/dS values at the branch highlighting host plant differences in *Pieris*

We used *P. rapae* contigs from the RBH result for GO annotation and ran these genes against the NCBI nonredundant protein sequence database in Galaxy (Blastx, e-value = 10e−4). We subsequently used the Blast2GO platform to load the resulting Blast-xml file and to conduct mapping and annotation steps based on the BLAST result for acquiring GO annotations for each contig (Götz et al., 2008). To test significantly elevated dN/dS ratios among genes associated

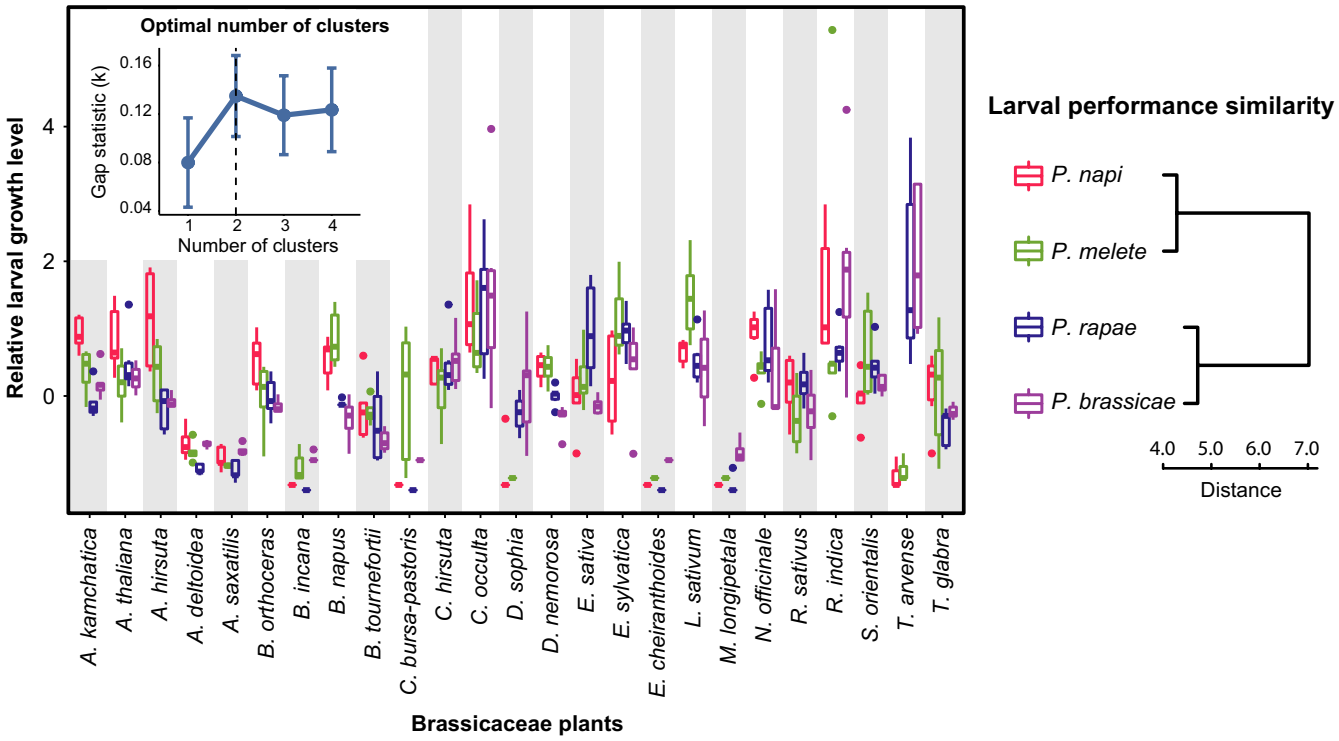

FIGURE 3 Feeding assays of four *Pieris* butterfly larvae on 25 different Brassicaceae plants (*n* = 6). The four *Pieris* butterfly species generally grew better on *Cardamine occulta* but could not use *B. incana* or *E. cheiranthoides* as optimal hosts. Gap statistic (inbox) was highest at clustering number 2, suggesting overall larval performance patterns of the four *Pieris* species could be best clustered in two groups. The hierarchical clustering analysis suggested that the two groups are as follows: *Pieris napi*– *Pieris melete* and *Pieris rapae*– *Pieris brassicae* shown here as larval performance similarity cladogram. The largest performance differences were observed on *Thlaspi arvense* on which *Pieris rapae* and *Pieris brassicae* larvae grew better than the other two species [Colour figure can be viewed at wileyonlinelibrary.com]

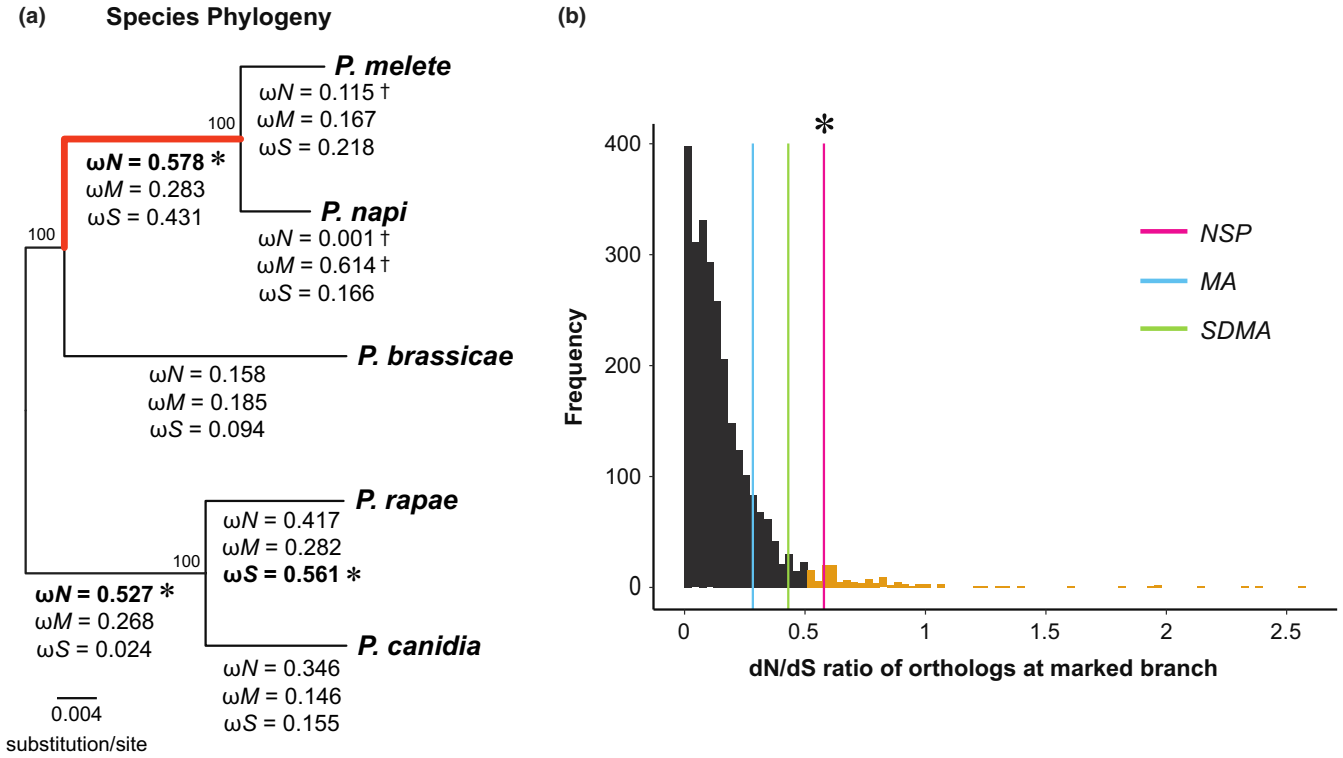

FIGURE 4 (a) Unrooted species phylogeny of five *Pieris* species used in this study. The tree was constructed based on all the aligned orthologs from reciprocal best blast hit analyses (2,063,074 bp total). Statistical supports from bootstrapping tests are shown at each node. The branch that reflects host plant difference is marked with colour. Each branch has estimated dN/dS (=ω) values of *NSP*-like gene family members (ω*N*: ω*NSP*, ω*M*: ω*MA*, ω*S*: ω*SDMA*). Values with "*" mean that the value is located in the top 5% of the entire dN/dS distribution. Values with "†" showed too low estimated dS (dS < 0.01) and were removed from the analysis. (b) The distribution of dN/dS values among orthologs at the marked branch of the species phylogeny (a). The top 5% values in the histogram are coloured orange. The vertical lines show dN/dS values of *NSP*-like gene family members; *NSP* (pink), *MA* (blue) and *SDMA* (green). "*"shows the line is in the top 5% [Colour figure can be viewed at wileyonlinelibrary.com]

with specific GO terms, we selected those that contained at least 20 orthologs and tested their dN/dS distributions with those of all the observed orthologs (background) using a Wilcoxon test. We performed this analysis based on the estimated dN/dS at the two internal branches: (*P. melete*, *P. napi*) (*P. brassicae*, *P. rapae*, *P. canidia*) branch, which highlighted the larval performance differences, and (*P. melete*, *P. napi*, *P. brassicae*) (*P. rapae*, *P. canidia*), which does not explain differential host plant use. We compared the GO categories with elevated dN/dS between these two branches to potentially identify genes with signatures of selection which correlate with the observed larval performance. All statistical analyses were performed in R studio ver. 1.1.453, and *p* values acquired were adjusted by FDR (RStudioTeam, 2016).

3 | RESULTS

3.1 | Performance of four *Pieris* butterflies on 25 Brassicaceae plants

We obtained larval weights for four *Pieris* butterfly species (*P. napi*, *P. melete*, *P. rapae* and *P. brassicae*) feeding on 25 different Brassicaceae plant species (Figure 3). The gap statistics for the given number of clusters were as follows: Gap_1 = 0.080, Gap_2 = 0.135, Gap_3 = 0.119, Gap_4 = 0.123 (Figure 3). Our analysis showed that larval performance of the four *Pieris* species could be best clustered into two groups: the *P. napi*-*P. melete* group and the *P. rapae*–*P. brassicae* group, which was also expected from field observations. The largest performance differences were observed on *Thlaspi arvense*, on which *P. rapae* and *P. brassicae* performed better than *P. napi* and *P. melete* (Figure 3). However, since each of the four species also has their species-specific host spectra but also has shared host (or non-host) plants, Gap statistics of other cluster numbers were also higher (e.g., Gap_4).

3.2 | RNA-seq, reciprocal best hit (RBH) BLAST analysis of *Pieris* butterflies

We obtained 32–40 million Illumina 100 bp pair-end reads for the four species (*P. napi*, *P. melete*, *P. rapae* and *P. brassicae*) and 64 million Illumina 150 bp pair-end reads for *P. canidia*. De novo transcriptome assemblies using Trinity resulted in 64,279; 62,054; 59,327; 53,004; and 149,481 contigs and in N50 values of 2,048 bp; 2,132 bp; 2,060 bp; 2,594; and 2,075 bp for *P. napi*, *P. melete*, *P. rapae*, *P. brassicae*, and *P. canidia*, respectively. Using RBH BLAST on the five *Pieris* species, we obtained transcriptome data resulted in 2,723 ortholog sets.

3.3 | Phylogeny of *Pieris* and dN/dS ratios of *NSP*-like gene family members across *Pieris* branches

The unrooted tree from all the aligned orthologs of *Pieris* species displays solid statistical support for all major nodes (Figure 4a), with *P. napi* and *P. melete* forming a distinct clade while a different clade consists of *P. rapae* and *P. canidia*. We estimated dN/dS ratios for all ortholog sets at all phylogenetic branches with PAML 4.8 (Yang, 2007). We found *NSP* had a significantly elevated dN/dS value at the (*P. melete*, *P. napi*) (*P. brassicae*, *P. rapae*, *P. canidia*) branch (dN/dS = 0.578), the branch that is consistent with major host plant differences (Figures 3 and 4a). The complete distribution of estimated dN/dS values at this branch is shown in Figure 4b (mean dN/dS = 0.105). We also observed that (*P. melete*, *P. napi*, *P. brassicae*) (*P. canidia*, *P. rapae*) branch had higher dN/dS values for *NSP* (dN/dS = 0.527) (Figure 4a). Regarding *MA*, we could not find elevated dN/dS values at any of the branches, except for the *P. napi* branch. However, this higher dN/dS value at the *P. napi* branch was caused by quite low dS values and is likely not reliable (dS ≤ 0.01). *SDMA* showed elevated dN/dS only at the *P. rapae* branch (dN/dS = 0.561).

3.4 | Signatures of clade-specific positive selection on *NSP*-like gene family members correlating with larval performance differences

The ML gene trees of *NSP*-like gene family members are shown in Figure 5. A branch-site model approach identified positively selected sites on *NSP* at branch1 and branch3, both of which could explain the observed host plant differences between (*P. melete*, *P. napi*) and (*P. brassicae*, *P. rapae*) (FDR adjusted p = .0236 and .0010, LRT; Figure 5a, Table 1). The BEB analysis suggested that one codon site had signs of positive selection in *NSP*s at branch1 (Table 1, posterior probability >.9). This site was located in the second domain of *NSP*s (position 304 of the amino acid sequence), where also the positively selected site identified in previous work is located (position 379) (Heidel-Fischer et al., 2010). BS + MNM analysis also confirmed the significance of positive selection at both of the branches (FDR adjusted p = .0001 and .0071, LRT; Table 1). Regarding *MA* and *SDMA*, we found no sign of positive selection even at the branches at which we found higher dN/dS values of these genes compared to other orthologs (Figure 5a,b).

3.5 | GO terms with elevated dN/dS ratios associated with differential host plant use

After GO annotations of all *P. rapae* RBH contigs, we obtained 1,457 GO terms in our data sets. These included 680 terms related to biological process, 540 to molecular function and 237 to cellular component GOs. We conducted the Wilcoxon test for the GO terms that had more than 20 assigned orthologs. Based on the estimated dN/dS values at (*P. melete*, *P. napi*) (*P. brassicae*, *P. rapae*, *P. canidia*) branch, which highlights the host plant differences (Figure 4a), we found that one biological process—"proteolysis"—two processes associated with molecular function—"hydrolase activity" and "serine-type endopeptidase activity"—and two cellular component terms—"extracellular region" and "membrane"—had significantly elevated dN/dS values when compared to the entire dN/dS distribution of all contigs (Figure 6, Table 2). This test also showed that 9 GO terms had significantly lower dN/dS values in the three categories at this branch (Table 2). For the other internal branch (*P. melete*, *P. napi*, *P. brassicae*) (*P. rapae*, *P. canidia*), which does not reflect differential host plant use (Figure 4a), we could not find any GO terms with elevated dN/dS values, although 10 GO terms showed significantly lower dN/dS values (Table 2).

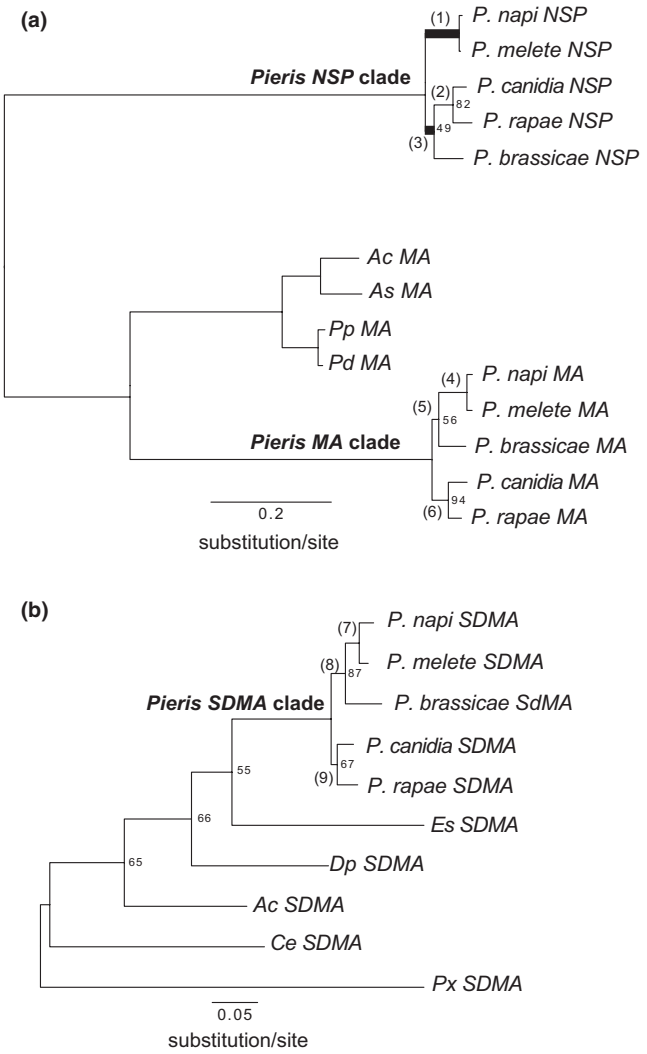

FIGURE 5 Maximum-likelihood tree of (a) *NSP* and *MA*, and (b) *SDMA*. *Ac*: *Anthocharis cardamines*, *As*: *Anthocharis scolymus*, *Pp*: *Pontia protodice*, *Pd*: *Pontia daplidice*, *Es*: *Eucheira socialis*, *Dp*: *Dixeia pigea*, *Ce*: *Colias eurytheme*, *Px*: *Plutella xylostella*. Numbers next to nodes show bootstrap support if below 100% (1,000 replicates). Numbers in brackets show the phylogenetic branches that were tested for positive selection. Branches with bold indicate evidence of positive selection

TABLE 1 Branch-site model tests on *NSP*-like gene family members by codeml

Gene	Branch	lnL *Model null*	lnL *Model alt*	delta L	*p* value (FDR adjust.)	BS + MNM *p* (FDR adjust.)	BEB site *p* > .9
NSP	branch1	−10,913.56	−10,908.77	9.58	.0236	.0001	304 I 0.984
	branch2	−10,914.81	−10,914.81	0	n.s.	n.s.	
	branch3	−10,915.57	−10,907.21	16.72	.0010	.0071	
	Pieris napi	−10,916.45	−10,916.45	0	n.s.	n.s.	
	Pieris melete	−10,916.45	−10,916.45	0	n.s.	n.s.	
	Pieris brassicae	−10,916.45	−10,916.45	0	n.s.	n.s.	
	Pieris rapae	−10,916.45	−10,916.45	0	n.s.	n.s.	
	Pieris canidia	−10,916.45	−10,916.45	0	n.s.	n.s.	
MA	branch4	−10,916.45	−10,916.45	0	n.s.	n.s.	
	branch5	−10,916.37	−10,916.37	0	n.s.	n.s.	
	branch6	−10,916.38	−10,916.38	0	n.s.	n.s.	
	Pieris napi	−10,916.45	−10,916.45	0	n.s.	n.s.	
	Pieris melete	−10,916.45	−10,916.45	0	n.s.	n.s.	
	Pieris brassicae	−10,916.45	−10,916.45	0	n.s.	n.s.	
	Pieris rapae	−10,916.08	−10,914.79	2.58	n.s.	n.s.	
	Pieris canidia	−10,916.45	−10,916.45	0	n.s.	n.s.	
SDMA	branch7	−3,289.82	−3,289.82	0	n.s.	n.s.	
	branch8	−3,290.22	−3,290.22	0	n.s.	n.s.	
	branch9	−3,290.22	−3,290.22	0	n.s.	n.s.	
	Pieris napi	−3,289.65	−3,289.65	0	n.s.	n.s.	
	Pieris melete	−3,289.27	−3,287.60	3.34	n.s.	n.s.	108 M 0.917
	Pieris brassicae	−3,290.22	−3,290.22	0	n.s.	n.s.	
	Pieris rapae	−3,287.38	−3,287.38	0	n.s.	n.s.	
	Pieris canidia	−3,289.79	−3,289.79	0	n.s.	n.s.	

Note: lnL *Model null*: log likelihood for null model with fixed dN/dS ratios. lnL *Model alt*: log likelihood for alternative model which allows having unfixed dN/dS values at the branch. Delta L: 2(lnL *Model alt* − lnL *Model null*) for the likelihood-ratio test (LRT). *p* values are from LRT and adjusted for multiple testing. BS + MNM *p*: Adjusted *p* values acquired from branch-site test covering situations of multiple nucleotide mutations. BEB analysis shows the specific sites which have significant signatures of positive selection with posterior probability. Positions are based on *Pieris rapae* protein sequences.

4 | DISCUSSION

Focusing on five Japanese *Pieris* butterflies, we tested host plant spectra and investigated signatures of selection on *NSP*-like genes, which are a key innovation of these butterflies to overcome the GLS defence system of their Brassicales host plants (Edger et al., 2015; Wheat et al., 2007). We acquired RBH ortholog sets expressed in larvae of the five *Pieris* species based on transcriptome data and compared the calculated dN/dS ratios of each ortholog or performed evolutionary tests in order to investigate the effect of evolutionary forces on *NSP*-like gene family members. We also combined ecological approaches for acquiring performance data on larvae of *Pieris* species by conducting a feeding experiment using 25 Brassicaceae plant species. These approaches yielded four major findings. First, we observed that *Pieris* species showed phylogenetically conserved differences in larval host performance. Second, we observed that *NSP* had significantly elevated dN/dS ratios compared to other genes at some phylogenetic branches; however, its sister gene *MA* did not show this trend. Third, evidence of positive selection on *NSP*s was observed at a phylogenetic branch which showed differences in larval performance according to our feeding assays, but no evidence of positive selection was found in *MA* or *SDMA*. Last, we observed significantly elevated dN/dS ratios in GO terms which are associated with potential detoxification-related genes and could correlate with larval performance differences at the branch.

According to our feeding experiments with four Japanese *Pieris* species (*P. napi*, *melete*, *rapae* and *brassicae*) and 25 Brassicaceae plant species, *P. napi* and *P. melete* larvae performed similarly, as did *P. rapae* and *P. brassicae* larvae (Figure 3). Observations in the field suggest that these four *Pieris* species have slightly different host preferences: *P. napi* and *P. melete* feed on wild and montane Brassicaceae plants, such as *Arabis* or *Turritis*, and *P. rapae* and *P. brassicae* use Brassicaceae crops more often than the other two species (Figure 1) (Harvey, Poelman, & Gols, 2010; Ohsaki & Sato, 1994). In addition, feeding assays also showed that *P. napi* and *P. melete* have similar larval performance trends on a set of Brassicaceae plants compared to *P. rapae*, supporting our results (Okamura, Tsuzuki, et al., 2019b). Although the clustering analysis did

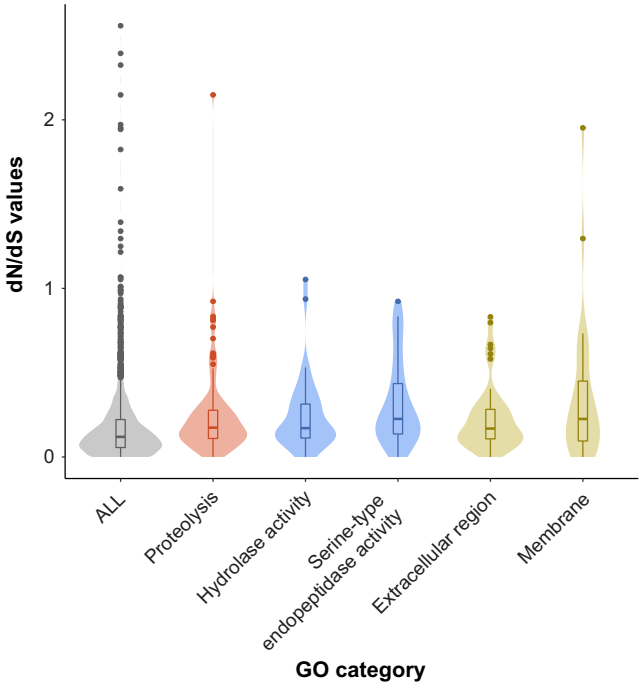

FIGURE 6 GO terms which have significantly elevated dN/dS values compared to those of entire ortholog sets at (*Pieris melete*, *Pieris napi*) (*Pieris brassicae*, *Pieris rapae*, *Pieris canidia*) branch, which highlights host plant differences. Significantly elevated dN/dS values were observed in "proteolysis" from biological process (red), and "hydrolase activity" and "serine-type endopeptidase activity" from molecular function (blue), and "extracellular region" and "membrane" from cellular component (yellow), as compared to the entire distribution of all the observed contigs. Comparisons with other enriched GO terms are shown in Table 2 [Colour figure can be viewed at wileyonlinelibrary.com]

not fully support this grouping (*Gap4* was also higher), this was expected since each *Pieris* species can also have its own, discrete host plant spectrum but at the same time are also known to share some host or nonhost plant species. The main objective of this larval performance analysis was to see which species tend to perform similar in more controlled feeding assays and how these results correlate with field observation. Thus, our feeding assay results confirm that *P. napi* and *P. melete* have more similar host plant spectra as do *P. rapae* and *P. brassicae* (Figure 3).

Overall, phylogenetic analysis showed that the species phylogeny seemed to correspond with larval performance (Figures 3 and 4), suggesting that the larval host preferences of the four *Pieris* butterflies are phylogenetically conserved. In this study, we did not perform any physical or chemical defence analyses on the different Brassicaceae plants species we used; however, a number of previous studies revealed that the GLS profiles of Brassicaceae plants can differ dramatically among Brassicaceae species (Agerbirk & Olsen, 2012; Fahey, Zalcmann, & Talalay, 2001; Olsen et al., 2016). Our results suggest that *Pieris* species might not always be capable of fully adapting to the defences of the complete range of their potential (i.e., available in their geographic distribution) host plants and so likely evolved to feed on a subset of Brassicaceae plants. This insight is consistent with the prediction of the arms-race hypothesis of their diversification.

Comparing dN/dS ratios of each ortholog at all branches, we found that *NSP*s had higher dN/dS values at two branches as ranked in the top 5% among all tested orthologs (Figure 4a). Although we filtered out a number of genes by RBH processes and therefore compared only a subset of the entire orthologs, our findings suggest that *NSP*s show evidence for positive selection—or, strongly relaxed purifying selection—among the five *Pieris* butterfly species. Interestingly, we also found that *MA*s had lower dN/dS values compared to *NSP*s (Figures 4 and 5), and their dN/dS values did not reach the top 5% among all tested orthologs, suggesting that in this genus *MA*s are under stronger purifying selection than are *NSP*s. *NSP*s and *MA*s are known as paralogs, and only *NSP* was confirmed to have GLS-disarming activity in *Pieris*. However, *MA*s also disarm GLSs in another Brassicaceae-feeding Pierid genus, *Anthocharis*, which has only *MA*s (Edger et al., 2015); this overlap strongly suggests that in *Pieris MA*s act like *NSP*s. Our results show that selection on these two paralogous genes, both of which have a similar repeat domain structure and can potentially disarm GLSs, can differ strikingly. This could imply that these paralogs have been differentially (sub-)functionalized in *Pieris*, where *NSP*s have more derived functions, whereas *MA*s have more conserved functions. Finally, for *SDMA*, we also observed elevated dN/dS ratios at one branch (*P. rapae* branch). Expressed in the gut, *SDMA*s are known to be found in all Lepidoptera, supporting the hypothesis that their function is related to digestion and not to disarming GLS (Fischer et al., 2008; Randall et al., 2013). However, we still lack reliable information about their role in the Lepidopteran gut environment. Focusing on the *SDMA* branch with elevated dN/dS ratios might provide additional information to understand its function.

Using the branch-site test implemented in both PAML and BS + MNM, we detected evidence of positive selection only in *NSP* at branch1 and branch3 in the ML gene tree (Figure 5a). These branches highlight and support the results of our feeding experiment, in which we found that the *P. napi* and *P. melete* clade had different host preferences from *P. rapae* and *P. brassicae* (Figures 2 and 4). Interestingly, branch2, which did not reflect the result of our feeding assays, also had elevated dN/dS ratios of *NSP* compared to other orthologs (Figure 4) but did not show any evidence of positive selection in PAML and BS + MNM. For both *MA* and *SDMA*, no evidence for positive selection was detected (Figure 5a,b, Table 1). Thus, the patterns of nucleotide substitution rate ratios and results of the branch-site tests suggest that host plant preferences in *Pieris* are correlated with evidence for positive selection of *NSP*s, but not *MA*s or *SDMA*s.

In this study, we did not test the functional differences of *NSP*s among the five *Pieris* species. Furthermore, we could not determine whether the differences in larval performance that we observed among the four *Pieris* species were caused by the dissimilarity among the GLS profiles of the host plants. However, our findings imply a strong relationship between the molecular evolution of *NSP*s and host-utilization patterns among *Pieris* butterflies. Moreover, it is also important to note that only *NSP*s showed evidence for positive selection correlating with different host plant use, suggesting that *NSP*s have been functionalized to detoxify GLSs specific to certain plant species; in contrast, *MA*s may have evolved to disarm the much

TABLE 2 GO terms with elevated or decreased dN/dS values corresponding to the complete set of orthologs at branch highlighting different host plant use (*Pieris melete*, *Pieris napi*) (*Pieris brassicae*, *Pieris rapae*, *Pieris canidia*) and the background internal branch (*Pieris melete*, *Pieris napi*, *Pieris brassicae*) (*Pieris rapae*, *Pieris canidia*) which does not reflect host plant differences

	Branch highlighting different host plant use					Background internal branch				
GO term	*N*	dN/dS	*p* (FDR adjst.)		Up/down	*N*	dN/dS	*p* (FDR adjst.)		Up/down
ALL	2,652	0.172				2,713	0.129			
Biological process										
Oxidation–reduction process	127	0.169	.487			130	0.128	.409		
Proteolysis	116	0.244	≤.001	***	**Up**	116	0.140	.091		
Regulation of transcription, DNA-templated	99	0.118	≤.001	***	Down	106	0.071	≤.001	***	Down
Transmembrane transport	77	0.139	.487			78	0.103	.417		
Ribosome biogenesis	42	0.119	.094			44	0.050	≤.001	***	Down
Carbohydrate metabolic process	47	0.179	.386			47	0.126	.409		
Translation	35	0.129	.240			37	0.046	≤.001	***	Down
Signal transduction	31	0.120	.206			33	0.088	.409		
Protein phosphorylation	39	0.118	.206			39	0.077	.091		
Phosphorylation	37	0.169	.411			37	0.177	.459		
Methylation	26	0.237	.411			26	0.154	.584		
Purine nucleobase metabolic process	25	0.172	.786			25	0.114	.935		
Molecular function										
ATP binding	200	0.120	≤.001	***	Down	203	0.103	.006	**	Down
Nucleic acid binding	136	0.194	.603			139	0.151	.121		
Zinc ion binding	124	0.157	.984			125	0.134	.442		
Metal ion binding	115	0.171	.524			117	0.139	.181		
DNA binding	108	0.160	.119			111	0.108	.115		
RNA binding	84	0.142	.901			86	0.101	.207		
Structural constituent of ribosome	38	0.130	.227			40	0.044	≤.001	***	Down
Oxidoreductase activity	34	0.239	.063			35	0.127	.582		
Hydrolase activity	41	0.246	.010	**	**Up**	42	0.175	.080		
GTP binding	38	0.108	.030	*	Down	39	0.086	.090		
Serine-type endopeptidase activity	41	0.316	≤.001	***	**Up**	41	0.164	.306		
Transmembrane transporter activity	36	0.119	.545			36	0.122	.878		
Calcium ion binding	39	0.090	.017	*	Down	40	0.070	.085		
DNA-binding transcription factor activity	29	0.095	.057			31	0.062	.006	**	Down
Sequence-specific DNA binding	28	0.050	≤.001	***	Down	30	0.045	≤.001	***	Down
Transferase activity	30	0.178	.545			31	0.133	.853		
Ligase activity	30	0.149	.345			31	0.136	.853		
GTPase activity	26	0.115	.143			27	0.090	.138		
Kinase activity	31	0.173	.545			31	0.190	.585		
Helicase activity	24	0.167	.751			26	0.115	.585		
Methyltransferase activity	24	0.240	.545			24	0.150	.940		
Iron ion binding	23	0.162	.557			23	0.143	.442		
Cellular component										
Integral component of membrane	607	0.176	.408			624	0.115	.311		
Nucleus	197	0.146	.006	**	Down	203	0.098	≤.001	***	Down

(Continues)

TABLE 2 (Continued)

GO term	Branch highlighting different host plant use					Background internal branch				
	N	dN/dS	*p* (FDR adjst.)		Up/down	*N*	dN/dS	*p* (FDR adjst.)		Up/down
Cytoplasm	77	0.125	.047	*	Down	76	0.099	.071		
Ribosome	65	0.129	.408			68	0.079	.005	**	Down
Extracellular region	51	0.229	.009	**	**Up**	51	0.152	.089		
Intracellular	41	0.119	.140			42	0.116	.517		
Transcription factor complex	33	0.087	.009	**	Down	35	0.060	.002	**	Down
Membrane	28	0.363	.009	**	**Up**	29	0.176	.217		
Mitochondrion	22	0.161	.868			23	0.175	.108		

Note: ALL: all the orthologs with assigned GO term. *N*: number of orthologs in the GO term. *p* values are adjusted with false discovery rates. GO terms with elevated dN/dS values are in bold. FDR adjusted *p* value: "*" < .05, "**" < .01, "***" < .001.

more widespread types of GLSs such as are found universally across Pieridae host plants. In addition, in this study we found positively selected sites in the second domain of *NSP*s (exon 5), as well as in earlier population genetic work using *P. rapae* (Heidel-Fischer et al., 2010). Although the molecular mechanisms of the GLS-disarming function of *NSP*s and *MA*s are still unclear, our results suggest that the second domains of *NSP*s are important for substrate specificity.

Besides individual *NSP*-like gene family members, elevated dN/dS values were also more broadly observed at the two internal branches of the unrooted *Pieris* phylogeny (Figure 4a): the (*P. melete*, *P. napi*) (*P. brassicae*, *P. rapae*, *P. canidia*) branch highlighting differential host plant use, and the (*P. melete*, *P. napi*, *P. brassicae*) (*P. rapae*, *P. canidia*) branch which does not reflect host plant differences. Surprisingly, we could only find GO terms with elevated dN/dS at the branch highlighting differential host plant use, including "proteolysis" (biological process); "serine-type endopeptidase activity" and "hydrolase activity" (molecular function); "extracellular region" and "membrane" (cellular component). These GO categories with elevated dN/dS values were broadly consistent with potential candidates of positive selection or relaxed purifying selection along with differential host plant use in herbivorous insects in general. In Lepidopteran larvae, most of the digestive enzymes are involved in proteolysis (Simon et al., 2015) and several classes of digestive enzymes are necessary for insect herbivores to acquire essential nutrients in appropriate amounts (Broadway, 1989). In *Pieris*, these proteolytic activities are dominated by serine endopeptidases (Broadway, 1996). Since plants also have varied species-specific protease inhibitors to inhibit protease activity in herbivores, herbivores need to have evolved inhibitor-resistant proteinases as a counter adaptation (Bolter & Jongsma, 1997). Our findings showed elevated dN/dS values and evidence for positive or relaxed purifying selection in protease-related genes at the branches which are correlated with different host plant spectra. This suggests that these genes have accumulated more functional changes or have been released from stronger functional constraints as a consequence of interactions with plants in their specific host plant ranges. A number of genes with hydrolase activity are included in genes related to detoxification in herbivores (Simon et al., 2015). In addition, several detoxification-related proteins, including *NSP* and *MA*, are secreted and thus display extracellular localization. Therefore, although the GO category "extracellular region" appears to be a very general term, the observed elevated dN/dS values for these genes would also be consistent with their potential role in interactions with host plant-derived compounds, including complex polysaccharides and proteins, but also toxic metabolites. Previous research has uncovered differential regulation of genes associated with this GO term in several herbivore species responding to different host plants (Schweizer, Heidel-Fischer, Vogel, & Reymond, 2017). Utilizing different host plants can cause specific functional changes of detoxification-related genes or releasing them from strict functional constraints. Although both scenarios can cause elevated dN/dS values, the observed GO terms with elevated dN/dS may relate to the challenges *Pieris* butterflies encounter on the respective host plants in their natural environment. Broader analyses of patterns of nucleotide substitutions (e.g., of detoxification-related genes) in the context of host plant associations would help to understand more general patterns of selection in specialist and generalist herbivores.

To uncover the co-evolutionary diversification of plants and herbivores, it is important to understand the evolutionary interactions between all involved partners. We found evidence for positive selection on *NSP*s in *Pieris*, suggesting that the evolution of host plant adaptive genes is correlated with patterns of host plant usage in this butterfly genus. Moreover, we also observed that *MA*s, which are paralogs of *NSP*s, are subject to more strict purifying selection than *NSP*s. Our findings combine results from genetic and ecological assays to focus on how the evolution of these two paralogous genes may affect the arms-race between Brassicales and *Pieris* butterflies and their consequent diversification. Functional assays focusing on selected sites will increase our understanding of the evolution and functional differentiation of *NSP*s and *MA*s and how *Pieris* butterflies adapted evolutionarily to diverse glucosinolates in their host plants.

ACKNOWLEDGEMENTS

We are grateful to Takashi Tsuchimatsu (Chiba University) for useful discussions and comments on this study. We thank Emily Wheeler, Boston, for editorial assistance. We also thank Itsuzai Aoki for his help in *P. canidia* field collection. This work was supported by

a Grant-in-Aid for Scientific Research from the Japanese Society for the Promotion of Science (15J00320 to Y.O.) and partially by Max–Planck–Gesellschaft.

AUTHOR CONTRIBUTIONS

Y.O., A.S. and N.T. carried out the laboratory work. Y.O., M.M., H.H.F. and H.V. conceived, designed and coordinated the study. Y.O., M.M., H.H.F. and H.V. wrote the manuscript. All authors, drafted parts of the manuscript, gave approval for publication and agree to be accountable for the content.

DATA AVAILABILITY STATEMENT

The RNA-seq short read data have been deposited in the EBI short read archive (SRA) with the following sample Accession nos: ERX2829492-ERX2829499, ERX3552761. The complete study can also be accessed directly using the following URL: http://www.ebi.ac.uk/ena/data/view/PRJEB29048 & http://www.ebi.ac.uk/ena/data/view/PRJEB34531.

ORCID

Yu Okamura https://orcid.org/0000-0001-6765-4998

REFERENCES

Agerbirk, N., & Olsen, C. E. (2012). Glucosinolate structures in evolution. *Phytochemistry*, *77*, 16–45. https://doi.org/10.1016/j.phytochem.2012.02.005

Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data*. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Beilstein, M. A., Al-Shehbaz, I. A., Mathews, S., & Kellogg, E. A. (2008). Brassicaceae phylogeny inferred from phytochrome A and *ndhF* sequence data: Tribes and trichomes revisited. *American Journal of Botany*, *95*(10), 1307–1327. https://doi.org/10.3732/ajb.0800065

Benson, J., Pasquale, A., Van Driesche, R., & Elkinton, J. (2003). Assessment of risk posed by introduced braconid wasps to *Pieris virginiensis*, a native woodland butterfly in New England. *Biological Control*, *26*(1), 83–93. https://doi.org/10.1016/S1049-9644(02)00119-6

Berenbaum, M. R., Favret, C., & Schuler, M. A. (1996). On defining "key innovations" in an adaptive radiation: Cytochrome P450s and Papilionidae. *The American Naturalist*, *148*, 139–155. https://doi.org/10.1086/285907

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bolter, C., & Jongsma, M. A. (1997). The adaptation of insects to plant protease inhibitors. *Journal of Insect Physiology*, *43*(10), 885–895. https://doi.org/10.1016/S0022-1910(97)00040-1

Bond, J. E., & Opell, B. D. (1988). Testing adaptive radiation and key innovation hypotheses in spiders. *Evolution*, *52*(2), 403–414. https://doi.org/10.1111/j.1558-5646.1998.tb01641.x

Broadway, R. M. (1989). Characterization and ecological implications of midgut proteolytic activity in larval *Pieris rapae* and *Trichoplusia ni*. *Journal of Chemical Ecology*, *15*(7), 2102–2113. https://doi.org/10.1007/BF01207441

Broadway, R. M. (1996). Dietary proteinase inhibitors alter complement of midgut proteases. *Archives of Insect Biochemistry and Physiology*, *32*(1), 39–53. https://doi.org/10.1002/(SICI)1520-6327(1996)32:1<39:AID-ARCH3>3.0.CO;2-S

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 1–9. https://doi.org/10.1186/1471-2105-10-421

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)*, *25*(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Celorio-Mancera, M., de la, P., Wheat, C. W., Huss, M., Vezzi, F., Neethiraj, R., Reimegård, J., … Janz, N. (2016). Evolutionary history of host use, rather than plant phylogeny, determines gene expression in a generalist butterfly. *BMC Evolutionary Biology*, *16*, 59. https://doi.org/10.1186/s12862-016-0627-y

Cock, P. J. A., Chilton, J. M., Grüning, B., Johnson, J. E., & Soranzo, N. (2015). NCBI BLAST+ integrated into Galaxy. *GigaScience*, *4*, 39. https://doi.org/10.1186/s13742-015-0080-7

Couvreur, T. L. P., Franzke, A., Al-Shehbaz, I. A., Bakker, F. T., Koch, M. A., & Mummenhoff, K. (2010). Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). *Molecular Biology and Evolution*, *27*(1), 55–71. https://doi.org/10.1093/molbev/msp202

Edger, P. P., Heidel-Fischer, H. M., Bekaert, M., Rota, J., Glöckner, G., Platts, A. E., … Wheat, C. W. (2015). The butterfly plant arms-race escalated by gene and genome duplications. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, 8362–8366. https://doi.org/10.1073/pnas.1503926112

Fahey, J. W., Zalcmann, A. T., & Talalay, P. (2001). The chemical diversity and distribution of glucosinolates and isothiocyanates among plants. *Phytochemistry*, *56*(1), 5–51. https://doi.org/10.1016/S0031-9422(00)00316-2

Feyereisen, R. (2012). Insect CYP genes and P450 enzymes. In L. I. Gilbert (Ed.), *Insect molecular biology and biochemistry* (pp. 236–316). London, UK: Elsevier.

Fischer, H. M., Wheat, C. W., Heckel, D. G., & Vogel, H. (2008). Evolutionary origins of a novel host plant detoxification gene in butterflies. *Molecular Biology and Evolution*, *25*(5), 809–820. https://doi.org/10.1093/molbev/msn014

Franzke, A., Lysak, M. A., Al-Shehbaz, I. A., Koch, M. A., & Mummenhoff, K. (2011). Cabbage family affairs: The evolutionary history of Brassicaceae. *Trends in Plant Science*, *16*(2), 108–116. https://doi.org/10.1016/j.tplants.2010.11.005

Gotz, S., Garcia-Gomez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., … Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, *36*(10), 3420–3435. https://doi.org/10.1093/nar/gkn176

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Harvey, J. A., Poelman, E. H., & Gols, R. (2010). Development and host utilization in *Hyposoter ebeninus* (Hymenoptera: Ichneumonidae), a solitary endoparasitoid of *Pieris rapae* and *P. brassicae* caterpillars (Lepidoptera: Pieridae). *Biological Control*, *53*(3), 312–318. https://doi.org/10.1016/j.biocontrol.2010.02.004

Heidel-Fischer, H. M., Freitak, D., Janz, N., Söderlind, L., Vogel, H., & Nylin, S. (2009). Phylogenetic relatedness and host plant growth form influence gene expression of the polyphagous comma butterfly (*Polygonia c-album*). *BMC Genomics*, *10*, 506. https://doi.org/10.1186/1471-2164-10-506

Heidel-Fischer, H. M., Kirsch, R., Reichelt, M., Ahn, S.-J., Wielsch, N., Baxter, S. W., … Kroymann, J. (2019). An insect counteradaptation against host plant defenses evolved through concerted

neofunctionalization. *Molecular Biology and Evolution*, *36*(5), 930–941. https://doi.org/10.1093/molbev/msz019

Heidel-Fischer, H. M., Vogel, H., Heckel, D. G., & Wheat, C. W. (2010). Microevolutionary dynamics of a macroevolutionary key innovation in a Lepidopteran herbivore. *BMC Evolutionary Biology*, *10*, 60. https://doi.org/10.1186/1471-2148-10-60

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, *35*(2), 518–522. https://doi.org/10.1093/molbev/msx281

Hunter, J. P. (1998). Key innovation and the ecology of macroevolution. *Trends in Ecology and Evolution*, *13*(1), 31–36.

Janz, N. (2011). Ehrlich and Raven revisited: Mechanisms underlying codiversification of plants and enemies. *Annual Review of Ecology, Evolution, and Systematics*, *42*(1), 71–89. https://doi.org/10.1146/annurev-ecolsys-102710-145024

Kitahara, H. (2016). Oviposition plants and seasonal migratory movements of sympatric *Pieris melete* and *P. napi japonica* (Lepidoptera, Pieridae). *Lepidoptera*. *Science*, *67*(1), 32–40.

Krempl, C., Sporer, T., Reichelt, M., Ahn, S.-J., Heidel-Fischer, H., Vogel, H., … Joußen, N. (2016). Potential detoxification of gossypol by UDP-glycosyltransferases in the two Heliothine moth species *Helicoverpa armigera* and *Heliothis virescens*. *Insect Biochemistry and Molecular Biology*, *71*, 49–57. https://doi.org/10.1016/j.ibmb.2016.02.005

Loytynoja, A., & Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences*, *102*(30), 10557–10562. https://doi.org/10.1073/pnas.0409137102

Mao, Y., Cai, W., Wang, J., Hong, G., Tao, X., Wang, L., Chen, X. (2007). Silencing a cotton bollworm P450 monooxygenase gene by plant-mediated RNAi impairs larval tolerance of gossypol. *Nature Biotechnology*, *25*(11), 1307–1313. https://doi.org/10.1038/nbt1352

Nallu, S., Hill, J. A., Don, K., Sahagun, C., Zhang, W., Meslin, C., … Kronforst, M. R. (2018). The molecular genetic basis of herbivory between butterflies and their host plants. *Nature Ecology and Evolution*, *2*(9), 1418–1427. https://doi.org/10.1038/s41559-018-0629-9

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Ohsaki, N., & Sato, Y. (1994). Food plant choice of *Pieris* butterflies as a trade-off between parasitoid avoidance and quality of plants. *Ecology*, *75*(1), 59–68. https://doi.org/10.2307/1939382

Okamura, Y. U., Sato, A. I., Tsuzuki, N., Sawada, Y., Hirai, M. Y., Heidel-Fischer, H., … Vogel, H. (2019a). Differential regulation of host plant adaptive genes in *Pieris* butterflies exposed to a range of glucosinolate profiles in their host plants. *Scientific Reports*, *9*, 7256. https://doi.org/10.1038/s41598-019-43703-8

Okamura, Y., Tsuzuki, N., Kuroda, S., Sato, A., Sawada, Y., Hirai, M. Y., & Murakami, M. (2019b). Interspecific differences in the larval performance of *Pieris* butterflies (Lepidoptera : Pieridae) are associated with differences in the glucosinolate profiles of host plants. *Journal of Insect Science*, *19*(3), 1–9. https://doi.org/10.1093/jisesa/iez035

Olsen, C. E., Huang, X. C., Hansen, C. I. C., Cipollini, D., Ørgaard, M., Matthes, A., … Agerbirk, N. (2016). Glucosinolate diversity within a phylogenetic framework of the tribe Cardamineae (Brassicaceae) unraveled with HPLC-MS/MS and NMR-based analytical distinction of 70 desulfoglucosinolates. *Phytochemistry*, *132*, 33–56. https://doi.org/10.1016/j.phytochem.2016.09.013

Randall, T. A., Perera, L., London, R. E., & Mueller, G. A. (2013). Genomic, RNAseq, and molecular modeling evidence suggests that the major allergen domain in insects evolved from a homodimeric origin. *Genome Biology and Evolution*, *5*(12), 2344–2358. https://doi.org/10.1093/gbe/evt182

Ratzka, A., Vogel, H., Kliebenstein, D. J., Mitchell-Olds, T., & Kroymann, J. (2002). Disarming the mustard oil bomb. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(17), 11223–11228. https://doi.org/10.1073/pnas.172112899

RStudioTeam (2016). *RStudio: Integrated development for R*. Retrieved from http://www.rstudio.com

Schweizer, F., Heidel-Fischer, H., Vogel, H., & Reymond, P. (2017). *Arabidopsis* glucosinolates trigger a contrasting transcriptomic response in a generalist and a specialist herbivore. *Insect Biochemistry and Molecular Biology*, *85*, 21–31. https://doi.org/10.1016/j.ibmb.2017.04.004

Shen, J., Cong, Q., Kinch, L. N., Borek, D., Otwinowski, Z., & Grishin, N. V. (2016). Complete genome of *Pieris rapae*, a resilient alien, a cabbage pest, and a source of anti-cancer proteins. *F1000Research*, *5*, 2631. https://doi.org/10.12688/f1000research.9765.1

Simon, J.-C., d'Alencon, E., Guy, E., Jacquin-Joly, E., Jaquiery, J., Nouhaud, P., … Streiff, R. (2015). Genomics of adaptation to host-plants in herbivorous insects. *Briefings in Functional Genomics*, *14*(6), 413–423. https://doi.org/10.1093/bfgp/elv015

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *63*(2), 411–423. https://doi.org/10.1016/j.scico.2012.08.004

Ueno, M. (1997). A note on the large white, *Pieris brassicae*. (I). *Yadoriga*, *169*, 25–41.

Venkat, A., Hahn, M. W., & Thornton, J. W. (2018). Multinucleotide mutations cause false inferences of lineage-specific positive selection. *Nature Ecology and Evolution*, *2*, 1280–1288. https://doi.org/10.1038/s41559-018-0584-5

Villanueva-Cañas, L., Laurie, S., & Alba, M. M. (2013). Improving genome-wide scans of positive selection by using protein isoforms of similar length. *Genome Biology and Evolution*, *5*(2), 457–467. https://doi.org/10.1093/gbe/evt017

Wheat, C. W., Vogel, H., Wittstock, U., Braby, M. F., Underwood, D., & Mitchell-Olds, T. (2007). The genetic basis of a plant-insect coevolutionary key innovation. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(51), 20427–20431. https://doi.org/10.1073/pnas.0706229104

Wittstock, U., Agerbirk, N., Stauber, E. J., Olsen, C. E., Hippler, M., Mitchell-Olds, T., … Vogel, H. (2004). Successful herbivore attack due to metabolic diversion of a plant chemical defense. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(14), 4859–4864. https://doi.org/10.1073/pnas.0308007101

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Zhang, J., Nielsen, R., & Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution*, *22*(12), 2472–2479. https://doi.org/10.1093/molbev/msi237

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Okamura Y, Sato A, Tsuzuki N, Murakami M, Heidel-Fischer H, Vogel H. Molecular signatures of selection associated with host plant differences in *Pieris* butterflies. *Mol Ecol*. 2019;28:4958–4970. https://doi.org/10.1111/mec.15268

MOLECULAR ECOLOGY

Supplemental Information for:

Molecular signatures of selection associated with host-plant differences in *Pieris* butterflies

Yu Okamura[1,2], Ai Sato[2], Natsumi Tsuzuki[2], Masashi Murakami[2], Hanna Heidel-Fischer[1,3] Heiko Vogel[1]

1 Department of Entomology, Max Planck Institute for Chemical Ecology, Hans-Knöll-Str. 8, 07745, Jena, Germany

2 Community Ecology Lab, Faculty of Science, Chiba University, 263-8522, Chiba, Japan

3 Leibniz Institute for Natural Product Research and Infection Biology

Hans Knöll Institute (HKI), Beutenbergstraße 11a, 07745 Jena, Germany

MOLECULAR ECOLOGY

Table S1

25 Brassicaceae plants used for the feeding assay

Plant species	Seed collection site
Arabidopsis kamchatica	Nagano, Japan (36° 31' N, 138° 20' E)
Arabidopsis thaliana	Hokkaido, Japan (42° 50' N, 141° 41' E)
Arabis hirsuta	Nagano, Japan (36° 31' N, 138° 20' E)
Aubrieta deltoidea	-
Aurinia saxatilis	-
Barbarea orthoceras	Toyama, Japan (36° 34' N, 137° 26' E)
Berteroa incana	Burlington, Canada (43° 30' N, 79° 79' W)
Brassica napus	Chiba, Japan (35° 30' N, 140° 50' E)
Brassica tournefortii	Chiba, Japan (35° 54' N, 139° 56' E)
Capsella bursa-pastoris	Chiba, Japan (35° 30' N, 140° 50' E)
Cardamine hirsuta	Chiba, Japan (35° 30' N, 140° 50' E)
Cardamine occulta	Chiba, Japan (35° 30' N, 140° 50' E)
Descurainia sophia	Inner Mongolia, China (43° 37' N, 116° 42' E)
Draba nemorosa	Nagano, Japan (36° 20' N, 137° 50' E)
Eruca sativa	-
Eruca sylvatica	-
Erysimum cheiranthoides	Hokkaido, Japan (42° 50' N, 141° 41' E)
Lepidium sativum	-
Matthiola longipetala	-
Nasturtium officinale	Chiba, Japan (35° 30' N, 140° 50' E)
Raphanus sativus var. raphanistroides	Chiba, Japan (35° 07' N, 140° 11' E)
Rorippa indica	Chiba, Japan (35° 30' N, 140° 50' E)
Sisymbrium orientale	Chiba, Japan (35° 54' N, 139° 56' E)
Thlaspi arvense	Hokkaido, Japan (42° 56' N, 142° 01' E)
Turritis glabra	Nagano, Japan (36° 31' N, 138° 20' E)