PNAS
www.pnas.org

**Supplementary Information for**

# An aphid RNA transcript migrates systemically within plants and is a virulence factor

Yazhou Chen[1$], Archana Singh[2$], Gemy G. Kaithakottil[2], Thomas C. Mathers[1], Matteo Gravino[1], Sam T. Mugford[1], Cock van Oosterhout[3], David Swarbreck[2] and Saskia A. Hogenhout[1*]

[1]Crop Genetics, John Innes Centre, Norwich Research Park, Norwich, NR4 7UH, UK; [2]Earlham Institute, and [3]School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7UH, UK.

[$]Y.C and A.S. contributed equally to the work.

*To whom correspondence may be addressed:
Saskia A. Hogenhout, Department of Crop Genetics, John Innes Centre, Norwich Research Park, Norwich NR4 7UH, United Kingdom
Email: saskia.hogenhout@jic.ac.uk
ORCID: 0000-0003-1371-5606

**This PDF file includes:**

- Supplementary file S1
- Figures S1 to S19
- Tables S1 to S3
- SI References

**Other supplementary materials for this manuscript include the following:**

- Datasets S1-S5



www.pnas.org/cgi/doi/10.1073/pnas.1918410117

## Supplementary Methods

### Plants and growth conditions

Seeds of *B. rapa*, *A. thaliana*, *N. benthamiana* were provided by John Innes Centre horticultural department. *Solanum tuberosum* (Desiree) were purchased from Marks & Spencer (Norwich, UK). Seeds of *C. indicum* (polar star), *H. annuus* (Helios X Helios Flame), *P. vulgaris* (NYFB) were purchased from Thompson Morgan (Ipswich, UK). Seeds of *P. sativum* (JI3253) were provided by Claire Domoney (John Innes Centre, UK). Seeds of *Z. mays* (Early sunglow corn) were provided by Ian Bedford (John Innes Centre, UK). Seeds were directly sown into soil (peat-based compost). Plants were grown in a controlled environmental room at a constant temperature of 22°C and 70% humidity under a 10 h day /14 h night cycle.

### *M. persicae* transfer to 9 plant species

A colony of *M. persicae* clone O that started from a single female was established on *B. rapa* in a growth chamber (14 h light, 10 h dark at constant 20 °C, 75% humidity) in 2010. From this founder colony on *B. rapa*, approximately 500 asexual individuals were transferred to each of 9 plant species, including *B. rapa* (used as reference) and *A. thaliana*, as a representative of another plant species of the family Brassicaceae, and plant species belonging to 4 additional plant families, including *N. benthamiana* and *S. tuberosum* (Solanaceae), *C. indicum* and *H. annuus* (Asteraceae), *P. sativum* and *P. vulgaris* (*C. indicum* and *H. annuus*), and the monocot maize (*Z. mays*) (Poaceae). The aphids were maintained for about 4 generations on these plants, except for *C. indicum* and *P. vulgaris*, on which the aphids were reared for 10 generations. Aphid colonies on all 9 plant species were maintained in the same growth chamber (14 h light, 10 h dark at constant 20 °C, 75% humidity).

Stable aphid populations on the 9 hosts were analyzed for development time in days between births and emergence of adults, longevities in days between births to deaths, weights in mg per adult and reproduction rates measured by numbers

of nymphs divided by the number of adults on days 7, 9 and 11 for colonies on *B. rapa*, *A. thaliana*, *N. benthamiana* and *S. tuberosum* and on days 7, 9, 11, 13, 15, 17 and 18 for colonies on *C. indicum*, *H. annuus*, *P. sativum*, *P. vulgaris* and *Z. mays*.

**Transcriptome sequencing of *M. persicae* on 9 hosts**

Approximately 100 aphids were harvested from each plant species, at 5 independent biological replicates per plant species, and snap frozen in liquid nitrogen. Aphid samples were ground to powder using 5 mm stainless steel beads (Qiagen, Germany) and a TissueLyser II (Qiagen, Germany), and RNA was extracted using Trizol reagent (Sigma, UK) and the Qiagen RNeasy MinElute Cleanup Kit (Qiagen, Germany), which included and on-column DNase digestion.

**RNA sequencing and transcriptome assembly**

Strand-specific libraries were constructed from mRNAs isolated from 1 ug of total aphid RNA using the poly-A method of the Illumina TruSeq RNA Library Preparation kit (Illumina, US) following the manufacturer's procedures. cDNA was synthesized by 10 cycles of PCR to amplify the fragments. Libraries were then pooled and sequenced on a HiSeq 2000 generating 150-bp paired-end sequences (Earlham Institute, Norwich, UK). A genome-guided transcriptome assembly was generated with RNA-seq data of the 45 libraries of the nine host experiments, five replicates each (data generated herein as described above) and RNA-seq data generated from library LIB1777 (1), as follows. Reads were trimmed for low quality and adapter using Trim Galore! v0.4.0 with default settings (2). Trimmed reads were aligned to *M. persicae* reference genome G006 v1 (1) by HISat2 version-2.0.5 (3). The RNA-seq reads of all 46 libraries were merged together in one BAM file (Binary version of Sequence Alignment/Map) using Samtools (v0.1.18) (4) and assembled to create transcript models by StringTie version-1.3.3 (5) guided by the reference genome (6). A consensus assembly was produced using StringTie merge. Transcripts with Fragments Per Kilobase of transcript per Million (FPKM) ≤ 0.2 were removed from downstream analyses. Details of all transcriptomic libraries

generated for this study are listed in (***SI Appendix*, Table S1**). GFF files of the transcriptome assembly were submitted to GEO (GSE129667, Transcriptome_Assembly_G006_V2.gff).

**Functional annotation and *M. persicae* lncRNAs identification**

The computational workflow for the lncRNAs identification of the *M. persicae* is shown in ***SI Appendix*, Fig. S2A**. lncRNA identification was performed on transcript models obtained from the transcriptome assembly described above. Mikado compare (7) was used to identify and subsequently exclude transcripts overlap over 10% of the length of annotated miRNA, tRNA, rRNA and transposons features. To identify putative Open Reading Frame (ORFs), we used TransDecoder (https://transdecoder.github.io) with the default parameters. To further maximize the sensitivity, we scanned all ORFs for homology to curated protein sequence (Arthropods) in the Swiss-Prot database (8), downloaded November 20, 2018 from http://www.uniprot.org/downloads. This was done by BlastP (v2.7.1, evalue 1e-5). HAMMER (v-3.0) was run against the Pfam database (9) with default parameter to search for protein domains. The coding potentials of the remaining transcripts were assessed using CPC2 (10) and those with a coding potential score < 0.5 were selected. To consider transcript as a candidate lncRNA, transcript must be larger than 200nt, not have a hit in the SwissProt, Pfam database, considered non-coding by the CPC2, and not be already classified as another class of functional RNA (rRNA, miRNA, tRNA, transposons). The candidate lncRNA were submitted as a separate gff file (GSE129667, Candidate_lncRNA.gff).

Functional annotation for the protein coding genes were generated using annotF v1.02 (https://github.com/EI-CoreBioinformatics/AnnotF) and was submitted to GEO (GSE129667, Functional_annotation.txt.gz). Assessments of whether *M. persicae* genes are expressed in the salivary glands and guts was done by performing blastn against the EST datasets of Ramsey et al., 2007 (28).

**Differential gene expression analysis**

Differentially expressed transcripts of *M. persicae* colonies on the 9-plant species was determined by comparing transcript expression levels of *M. persicae* colonies on *B. rapa* (original host) with those of colonies on the 8 other plant species (new hosts) (see experimental design in Fig. 1A) using the DESeq2 package in R (v1.2.10) (12) and transcript count per million (TPM) generated by Kallisto v0.42.3 (https://pachterlab.github.io/kallisto/) (***SI* Dataset S1**). Additional filtering was employed in DESeq2 to remove lowly expressed transcripts (mean count < 10) on the basis of normalized counts. Transcripts were considered differentially expressed if they had a *p* value less than 0.05 after accounting for a 5% FDR according to the Benjamini-Hochberg procedure and if log2Fold change was greater than 1.

**Co-expression analysis**

Weighted gene co-expression network analysis (WGCNA) was used to generate unsigned co-expression networks on nine host swap data (13). Genes with normalized count (TPM) > 5 in at least one sample per plant host were used for the co-expression analysis and clustered into network modules using the topological overlap measure (TOM). Genes were grouped by hierarchical clustering on the basis of dissimilarity of gene connectivity (1-TOM). The co-expression clusters were produced by cutreeDynamic in which the minimum size of modules was kept at 20 genes. The modules were randomly colour-labelled. An adjacency matrix was built by applying a power function ($\beta$) on the Pearson correlation matrix. A $\beta$ value of 18 was found to be optimal for balancing the scale-free property of the co-expression network and the sparsity of connections between genes.

**Manual annotation of *Ya* genes**

A 148 bp nucleotide sequence (***SI Appendix*, Fig. S8**) that includes the entire exon 2 of the *Ya* genes and that was found to be conserved among 23 lncRNA genes in the darkslateblue module was used to search the *M. persicae* reference genome G006 with Blastn (v2.22) using default parameters. Blast hits,

with coverage more than 80 percent were converted to GFF and loaded to the Apollo browser (11) along with the annotated gene models (22) and all assembled transcripts herein before the merging step (see above). Gene models and corresponding transcripts that aligned to the 148-bp nucleotide sequence were selected and further curated by manually annotating the 3' ends of each of the transcripts based on the presence of a poly-A tail. The 5' ends were identified based on the most conserved sequence among all transcripts combined with existing RT-PCR data for *Ya1* (***SI Appendix*, Fig. S12**). The curated annotation of the *Ya* family enabled more accurate transcript quantifications among its members. We updated the gene models in our previous gff file for the *Ya* locus and submitted as a separate file (GSE129667, Manual_Anno.V3.gff). In addition to *M. persicae*, we also did manual annotation of *Ya* genes in five other aphid species including *A. pisum*, *M. cerasi*, *A. glycines*, *R. padi* and *D. noxia* using publicly available RNA-seq data and genome assemblies (***SI* Dataset S5**) and was submitted to GEO (GSE129667).

**Identification *M. persicae* RNAs in plants**

To assess if aphid translocate transcripts into plants, twenty adult *M. persicae* were caged on rosette leaves of 4-week-old *A. thaliana* plants for 24 hrs. The caged leaf area with aphids was assigned 'aphid feeding site'. Leaves on plants caged with empty cages were included as controls. Leaf areas covered by the cages were carefully washed three times with deionized water and three time with nuclease-free water. RNAs were isolated from four independent biological replicates of aphid-exposed leaves and non-exposed control leaves and processed for RNA-seq library synthesis and sequenced on the Ilumina HiSeq 25000 (Novogene, Beijing, China). Reads were trimmed to remove sequencing adapters and aligned to *A. thaliana* genome (TAIR10 database, http://arabidopsis.org) and the *M. persicae* G006 genome (22) with HISAT2 v2.0.5. Reads mapped to the *M. persicae* genome were retrieved and subjected to further filtering by mapping them back to the *A. thaliana* genome. Reads that did not align to the *A. thaliana* genome in the last step were considered as unique *M. persicae*

mapping reads. Transcripts with TPM ≥ 50 in at least one sample and that were present in at least three samples were selected for further analysis.

**RT-PCR analyses to detect systemic migration of aphid transcripts in plants**

Systemic migration of aphid transcripts was determined by caging a leaf section with aphids and detection of aphid transcripts in the caged area (feeding site), next to the caged area of the same leaf (near-feeding site) and a distal leaf (distal site). See experimental setups shown in **Fig. 4A** and ***SI Appendix*, Fig. S14**. Plants exposed to cages without aphids were used as controls. For *A. thaliana* plants, aphids were caged at the distal halves of the 8$^{th}$ rosette leaf (14) of 4-week-old *A. thaliana* plants for 24 hrs. The proximal leaf area of the 8$^{th}$ leaf next to the cage near the petiole was assigned near-feeding site and the 5$^{th}$ leaf that is likely phloem-connected with the 8$^{th}$ leaf (14) of *A. thaliana* plants the distal site. Similar setup as for *A. thaliana* were used for *B. rapa*, *P. sativum* and maize plants, except that for maize the near-feeding site was the middle of the leaf next to the caged area and distal site the part of this leaf near the stem (***SI Appendix*, Fig. S14**). Upon the 24 hr period of exposure to the cages with or without aphids, sections of leaves that were caged were immediately separated from the non-caged parts with scissors and the cages and aphids removed. Then the remaining parts of the leaves and distal leaves were detached from the plants. Leaf sections detached from plants were cleaned with the brush to remove aphids and visible debris. The leaf tissues were submerged in 5 mL of MilliQ water in a 15 mL tube. The tube was shaken for 30 sec after which the MilliQ water was removed. This was repeated two more times with MilliQ water and three additional times with nuclease-free water. Samples were snap-frozen in liquid nitrogen and storage at -80 °C.

Total RNAs were isolated from the leaf tissues by RNeasy Plant Mini Kit (Qiagen, Germany) followed by a DNase treatment with RNase-free DNase I (Thermo Fisher Scientific, US). cDNA was synthesized from 1 µg total RNA at 20 µL reaction volume with poly(A) primers using the RevertAid First Strand cDNA Synthesis Kit (Thermo Fisher Scientific, US). The qRT-PCRs reactions were

performed on a CFX96 Touch™ Real-Time PCR Detection System using transcript-specific primers (***SI Appendix*, Table S3**). Each reaction was performed in a 20 µL reaction volume containing 10 µL SYBR Green (Maxima SYBR Green/ROX qPCR Master Mix, Thermo Fisher Scientific, USA), 0.4 µL Rox Reference Dye II, 1 µL of each primer (10 µM), 1 µL of 20 µL sample cDNA, and 7.6 µL UltraPure Distilled water (Invitrogen, US). The PCR cycles were: 95 $^{o}$C for 10 s, 40 cycles at 95 $^{o}$C for 20 s, 63 $^{o}$C for 30 s. The identities of PCR products were verified by Sanger sequencing.

To compare *Ya1* transcript concentrations in the feeding and distal sites, a standard *Ya1* concentration curve was generated (***SI Appendix*, Fig. S13**). For this, the 273-nt Ya1 fragment cloned into the plasmid pBI121 under promoter 35S was used as the PCR template. The concentration curve was generated with a serial of dilution of the plasmid pBI121_35S::Ya1 from the highest concentration 1.44E$^{-10}$ g/ µL to the lowest concentration 1.44E$^{-18}$ g/ µL and primers (Ya1 primer6 and Ya1 primer9*,* ***SI Appendix*, Table S3**) using PCR conditions as described above.

**Primer design for amplification of *M. persicae Ya* transcripts**

Primers to specifically amplify *M. persicae Ya* transcripts were designed with the PrimerQuest Tool (Integrated DNA Technologies, IA, USA) that predicted five to ten primers for each of the *Ya* transcripts. The primer pairs were aligned to the sequences of all *Ya* transcripts and the ones that matched unique sequences of one *Ya* transcript selected (***SI Appendix*, Table S3**). *Ya* transcripts for which no unique primers were available were excluded from further analyses.

**Detection of aphid transcripts by northern blotting**

For northern blot analyses, 1 ug of total plant and aphid RNA were separated on a 6% denaturing polyacrylamide gel (PAGE) with 1 X TBE buffer (10 X TBE buffer stock, Thermo Fisher Scientific, USA) at 100 V for 60 min. The 273-nt *Ya1* fragment was cloned with a SP6 sequence at the 3 the end of *Ya1* and used to synthesize a 291-nt *Ya1-SP6* RNA of which 100 ng was ran alongside the total

RNAs from aphids and plants on polyacrylamide gels. RNAs were transferred to a nylon membrane (Hybond N, Amersham, UK) by electroblotting at 0.8 A for 2 hrs (BIO-RAD, USA) using 0.5 X TBE buffer at 60 V for 60 min. RNAs were cross-linked to the membranes by a UV cross linker (UVP Inc., CA, USA) using auto-crosslink function, twice on the side of the blot exposed to the gel and one time on the other slide of the blot.

*Ya1* transcript in aphids were detected via hybridization of the northern blots to a biotin-labelled anti-sense sequence of *Ya1*. Biotin-labelled anti-sense probe of *Ya1* was synthesized from anti-sense sequences of *Ya1* with the MAXIscript™ SP6/T7 Transcription Kit (Thermo Fisher Scientific, USA) and Biotin-16-dUTP (Roche, USA). The northern blots were performed according to the manufacturing manual of North2South Chemiluminescent Hybridization and Detection Kit (Thermo Fisher Scientific, US). Briefly, the blot was washed in North2South Hybridization Stringency Wash Buffer at room temperature for one time 20 min, in Wash Buffer containing 2X SSC/0.1% SDS at room temperature for three times 20 min, and Stringency Wash Buffer at 65 $^{o}$C for one time 20 min. The washing was done in 0.2 ml per $cm^2$ of blot. The blot was then incubated in the Substrate Working Solution containing equal volumes of the Luminol/Enhancer Solution and Stable Peroxide Solution for 5 min. To visualize the hybridization signal, the membrane was exposed to an X-ray film for an appropriate exposure time. Prestain RNA markers, DynaMarker Prestain Marker for RNA High (catalog number DM260, BioDynamics Laboratory Inc. Japan) and DynaMarker® Prestain Marker for Small RNA Plus (catalog number DM253, BioDynamics Laboratory Inc. Japan) were used to estimate the size of RNAs.

*Ya1* transcript in plants were detected via hybridization of the northern blots to a 273-nt *Ya1* fragment that was labeled using 3000 Ci/mmol of [a-32P] dATP (PerkinElmer Life Sciences. USA) with the Klenow DNA polymerase reaction as per manufacturer's instruction (Megaprime DNA Labeling System, GE Healthcare).The radioactively-labelled PCR probe was denatured at 95ºC for 5 min, transferred to ice and then incubated with the blot in 50 ml Hybridization buffer (Sigma-Aldrich) at 42°C for overnight. Washing was done three times at 42ºC with

washing buffer (2X SSC/0.1% SDS). The membranes were exposed to storage Phosphor Screens (GE Healthcare) and hybridization signals were visualized using Typhoon Trio (GE Healthcare) scanner.

**Sequencing of 3' ends of aphid *Ya1* transcript**

Total RNA was extracted from *M. persicae* aphids reared on *B. rapa* and 3 ug of aphid total RNAs was added to a 80 µL ligation mixture containing 8 µL T4 RNA ligase buffer and 4 µL T4 RNA ligase (Thermo Fisher Scientific, USA), 8 µL ATP (Invitrogen, USA), 8 µL BSA (Invitrogen, USA), and 10 pmol 3' RACE RNA adaptor (***SI Appendix,* Table S3**). RNA ligation was carried out at 16 ºC overnight. The ligated RNA was converted into cDNA using oligo sequences complementary to the 3' RACE adaptor with RevertAid First Strand cDNA Synthesis Kit (Thermo Fisher Scientific, USA) (***SI Appendix,* Table S3**). RACE PCRs were performed with the Ya1 forward primer GGACAAGTCCAATCTGC and the adapter primer CAAGCAGAAGACGGCATACGA (***SI Appendix,* Table S3**) in a 50 µL reaction volume containing 0.5 µL Phusion DNA polymerase (NEB), 10 µL 5X Phusion HF buffer, 1 µL 10 mM dNTPs, 1 µL of each primer (10 µM), 1 µL of cDNA sample. The cycle programs were: 98 ºC for 30 s, 35 cycles at 98 ºC for 10 s, 60 ºC for 30 s, 72 ºC for 15 s, final extension 72 ºC 10 mins. PCR produces were separated by 3% agarose gel. DNA bands were visualized under UV light, then cut to extract DNA using QIAquick Gel Extraction Kit (Qiagen, Germany). DNA were ligated to pGEM-T (Promega, USA) and Sanger sequenced with M13 primer.

**Plasmid construction**

To generate pJawohl8-RNAi constructs, a 273 bp of Ya1 was amplified from *M. persicae* cDNA by PCR with specific primers containing additional attB1 and attB2 linkers (***SI Appendix,* Table S3**) for cloning with the Gateway system (Invitrogen, USA). The 273-bp Ya1 fragment was introduced into pDONR207 (Invitrogen, USA) plasmid using Gateway BP reaction and transformed into DH5α competent cells (Invitrogen, USA). Subsequent clones were sequenced to verify correct size and sequence of inserts. Via the Gateway LB reaction, inserts were

transferred from pDONR207 into the plant transformation vector, pJawohl8-RNAi (kindly provided by I.E. Somssich, Max Planck Institute for Plant Breeding Research, Germany), which is a plasmid that enables the expression of the transgene as a double-stranded hairpin transcript, generating plasmid pJawohl8-RNAi_Ya1.

Plasmids pBI121_35S::Ya1 and pBI121_35S::Ya1_3TAG that produced *Ya1* and *Ya1_3UAG* transcripts were constructed as follows. The fragment corresponding to the 273 nt Ya1 transcript was amplified from *M. persicae* cDNA by PCR with specific primers containing *Bam*HI and *Sac*I restriction sites in the forward and reverse primers, specifically (***SI Appendix,* Table S3**). The *Bam*HI and *Sac*I digested PCR fragments were introduced into pBI121 to generate pBI121_35S::Ya1. Plasmid pBI121_35S::Ya1 was used in overlap PCR reactions (15) to generate the Ya1_3TAG construct in which the three ATGs at positions 41, 62, and 146 of Ya1 were converted into TAG stop codons. This involved the amplification of fragments 1, 2, 3 and 4 with primer pairs MutateF and ATG41_R, ATG41_F and ATG62_R, ATG62_F and ATG146_R and ATG146_F and MutateR, respectively (see ***SI Appendix*, Table S3 for primer sequences**), and subsequent amplification of a 81 bp fragment from fragments 1 and 2 with primers MutateF and ATG62_R, a 230 bp fragment from fragments 3 and 4 with primers ATG62_F and MutateR, and finally the amplification of the 273 nt Ya1 mutant in which three ATGs were mutated to three TAGs by combining the 81 bp and 230 bp fragments and primers MutateF and MutateR. The resulting Ya1_3TAGs PCR fragment was digested with *Bam*HI and *Sac*I and introduced into pBI121 to generate pBI121_35S::Ya1_3TAG.

The sequences of plasmid inserts were verified by Sanger sequencing using specific primers (***SI Appendix*, Table S3**).

**Generation of transgenic *A. thaliana* plants**

pJawohl8-RNAi_Ya1, pBI121_35S::Ya1, and pBI121_35S::Ya1_3TAG plasmids were introduced into *A. tumefaciens* strain GV3101 that carried the helper plasmid pMP90RK for subsequent transformation of *A. thaliana* Col-0 using

the floral dip method (16). Transgenic seeds were selected on Murashige and Skoog (MS) medium supplemented with 20 µg/mL phosphinothricin (BASTA) to select dsRNA_Ya1 transformants or on MS containing 50 µg/mL kanamycin to selection of 35S::Ya1 and 35S::Ya1_3TAG transformants. F2 seeds were germinated on MS medium supplemented with 20 µg/mL BASTA or 50 µg/mL kanamycin for selection. F2 seedlings with 3:1 alive/dead segregation (evidence of single insertion) were taken forward to the F3 stage. Seeds from F3 plants were sown on MS with BASTA (for dsRNA_Ya1 plants) or on MS with kanamycin (for 35S::Ya1 and 35S::Ya1_3TAG) and lines with 100% survival ratio (homozygous) were selected.

**Knock down of aphid transcripts by plant-mediated RNA interference (RNAi)**

Seeds of the dsRNA_Ya1 homozygous lines (expressing dsRNA corresponding to *Ya1*) were sown on MS medium and, after one week, seedlings were transferred to single pots (8 cm diameter) and transferred to a controlled environmental growth room at temperature 24 $^0$C day/20 $^0$C night under 10 hours of light. One *M. persicae* adult was confined to single 4-weeks-old Arabidopsis lines in sealed experimental cages (15.5 cm diameter and 30 cm height) containing the entire plant. One day later, the adult was removed, five nymphs remained on the plants and become adults in five days. Adults were harvested for qRT-PCR to confirm *Ya1* silencing compared to adults on dsGFP plants.

For qRT-PCRs, total RNA was isolated from aphids using Trizol reagent (Sigma) and subsequent DNase treatment using an RNase-free DNase I (Thermo Fisher Scientific, US). cDNA was synthesized from 1 µg total RNA with RevertAid First Strand cDNA Synthesis Kit (Thermo Fisher Scientific, US). The qRT-PCRs reactions were performed on CFX96 Touch™ Real-Time PCR Detection System using gene-specific primers. Each reaction was performed in a 20 µL reaction volume containing 10 µL SYBR Green (Thermo Fisher Scientific, US), 0.4 µL Rox Reference Dye II, 1 µL of each primer (10 µM), 1 µL of sample cDNA, and 7.6 µL UltraPure Distilled water (Invitrogen, US). The cycle programs were: 95 $^o$C for 10 s, 40 cycles at 95 $^o$C for 20 s, 60 $^o$C for 30 s. Relative quantification was calculated

using the comparative $2^{-\triangle}$Ct method (17). All data were normalized to the level of elongation factor gene (MYZPE13164_G006_v1.0_000087220) from the same sample.

### *M. persicae* fecundity assay

Seeds of Arabidopsis lines were sown on MS medium and, after one week, seedlings were transferred to single pots (8 cm diameter) and transferred to a controlled environmental growth room at temperature 24 $^0$C day/20 $^0$C night under 10 hours of light. One *M. persicae* adult was confined to single 4-weeks-old Arabidopsis lines in sealed experimental cages that contained the entire plant. One day later, the adult was removed, and one nymph remained on the plants. Offspring produced on the 7$^{th}$, 9$^{th}$, 11$^{th}$ day of the experiment were scored and removed. This experiment was repeated three times to create data from three independent biological replicates with four to six plants per line per replicate.

### Statistical analyses

All the data analyses were performed in R (v3.5.2). All statistical tests are described in the figure legends.

## Supplementary Figures and Tables

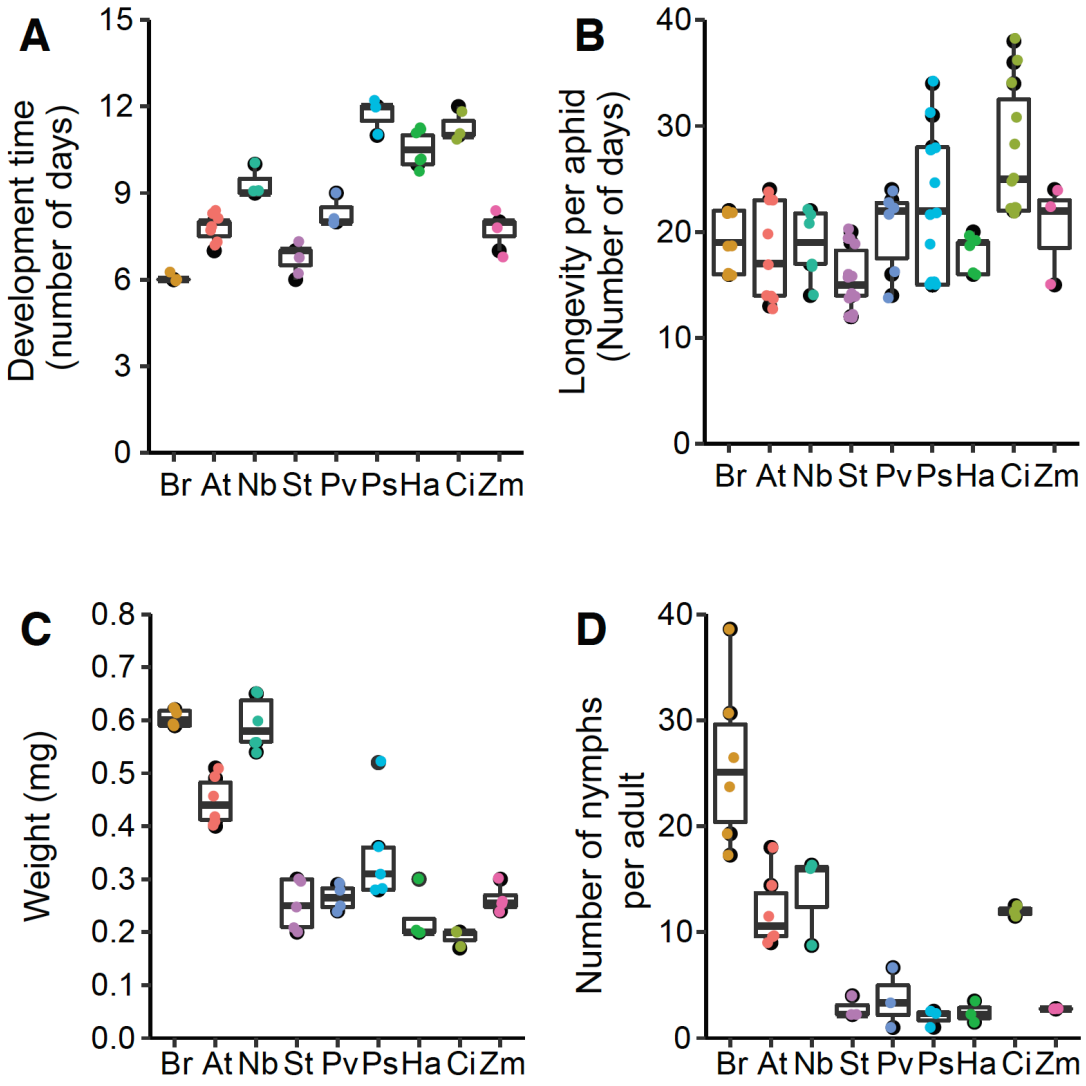


**Fig. S1.** Performance parameters of stable colonies of *M. persicae* clone O on 9 divergent plant species. Colonies on *B. rapa* (Br), *A. thaliana* (At), *N. benthamiana* (Nb), *S. tuberosum* (St), *C. indicum* (Ci), *H. annuus* (Ha), *P. sativum* (Ps), *P. vulgaris* (Pv) and *Zea mays* (Zm), established as shown in **Fig. 1A**, were analysed for development time (A) in days between birth and emergence of adults (n = 3-7 plants with one aphid each), longevity (B) in days between birth to death (n = 3-15 plants with one aphid each), weight (C) in mg per adult (n = 3-6 plants with 100 adult aphids each), and number of nymphs per

adult (D) calculated by dividing the number of nymphs by the number of adults on days 7, 9 and 11 (Br, At, Nb and St) or days 7, 9, 11, 13, 15, 17 and 18 (Ci, Ha, Ps, Pv and Zm) (n = 3-6 plants seeded with one aphid.

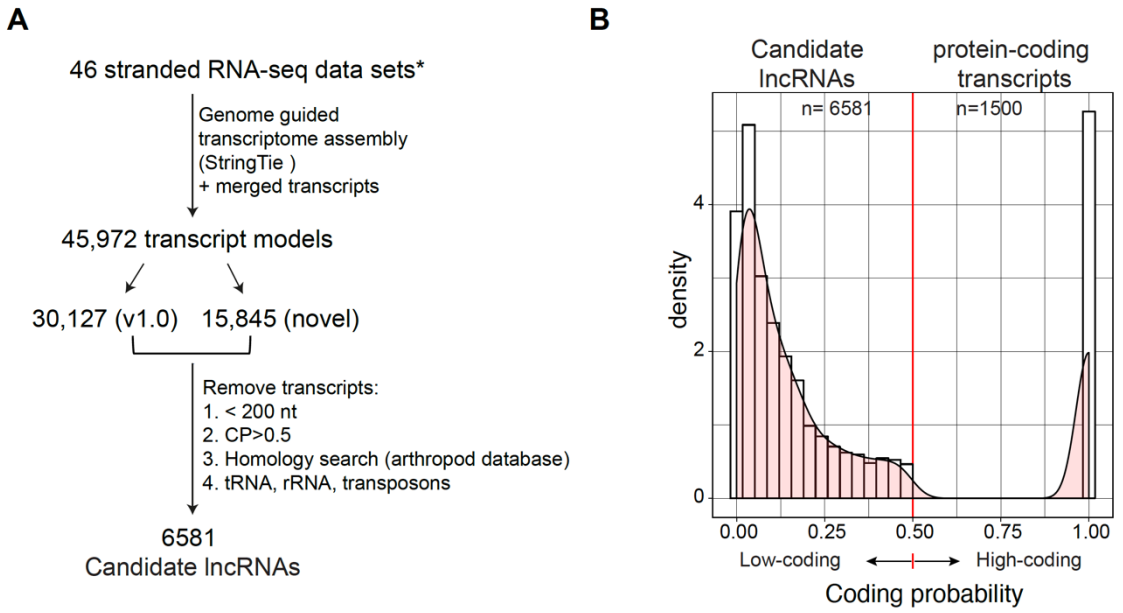


**Fig. S2.** Annotation of *M. persicae* transcript models and candidate lncRNAs. (A) Transcript annotation pipeline. RNA-seq reads from 45 libraries generated from *M. persicae* on nine divergent hosts (data generated herein, ***SI Appendix*, Fig. S1A**) and LIB1771 derived from a *M. persicae* colony on *B. rapa* (1) were assembled into transcripts using a genome-guided approach with StringTie (5). Of the 45,972 transcripts in total, 30,127 were annotated previously (1) and 15 845 are novel. The 6 581 transcripts that are candidate lncRNAs were identified upon removing transcripts of < 200 nt in size, transcripts with coding potential (CP) >0.5, as determined using the Coding Potential Calculator 2 (CPC2, http://cpc2.cbi.pku.edu.cn/) (10), and that have similarities of deduced protein sequences to known arthropods proteins, house-keeping RNAs (rRNA and tRNA) and transposons. (B) Distribution of CPC2 coding probability scores of 6,581 candidate lncRNAs identified in Fig. S2A and 1,500 randomly selected transcripts from *M. persicae* protein-coding sequences.

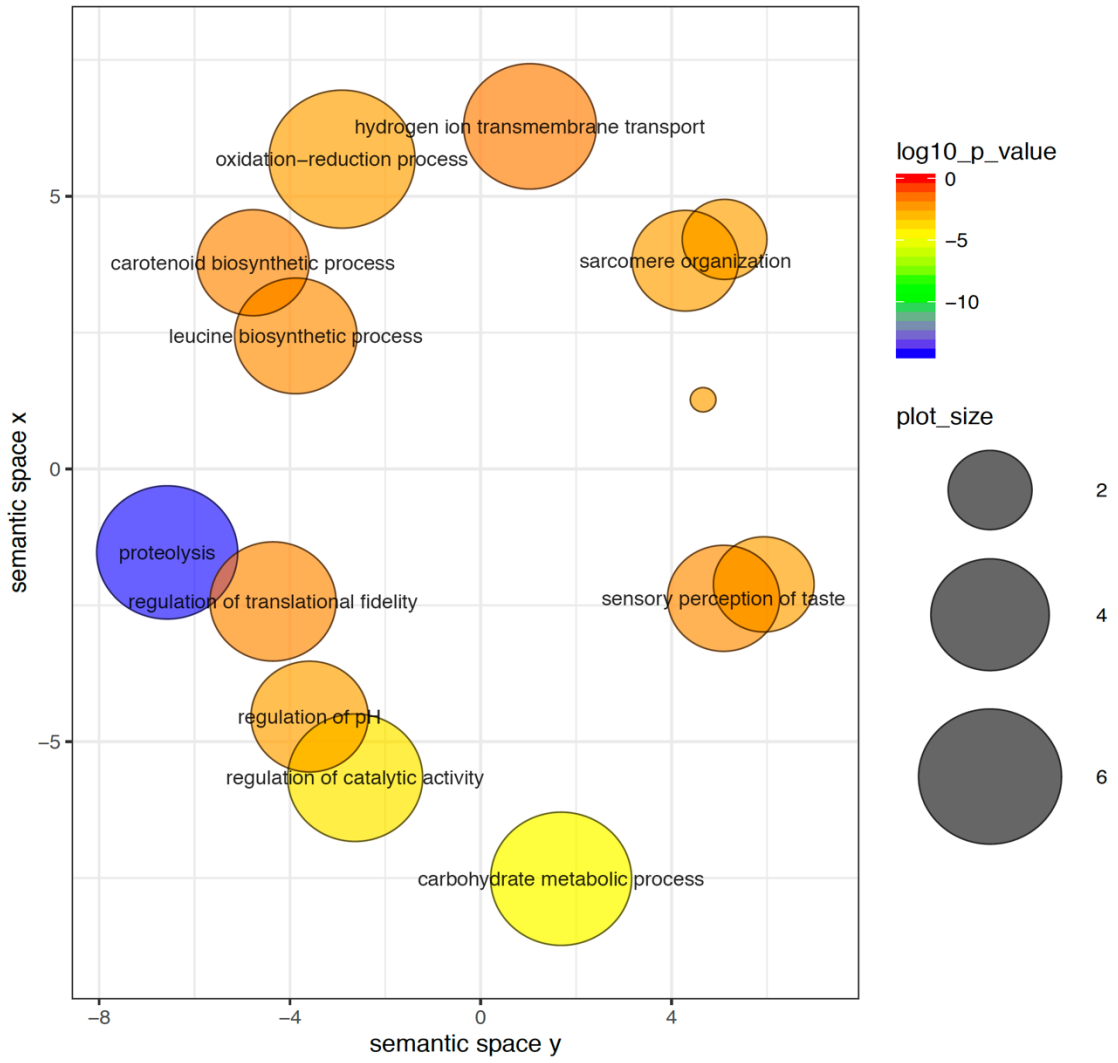


**Fig. S3**. Biological process GO terms that are enriched among DE genes of *M. persicae* on 9 plants species. Graphs was generated from data shown in **Fig. 1B**. Enriched GO terms are represented by circles and are clustered according to semantic similarities to other GO terms in the gene ontology. Distance between the circles indicates the semantic similarity between the corresponding GO terms. Circle size is proportional to the frequency of the GO term, whereas colour indicates the $\log_{10}(p$ value) for the enrichment calculated using GO-seq (red higher, blue lower). Only GO-terms having $p$ value and padj < 0.05 are shown. Data were summarized and visualized by REVIGO online tool (http://revigo.irb.hr/) (18).

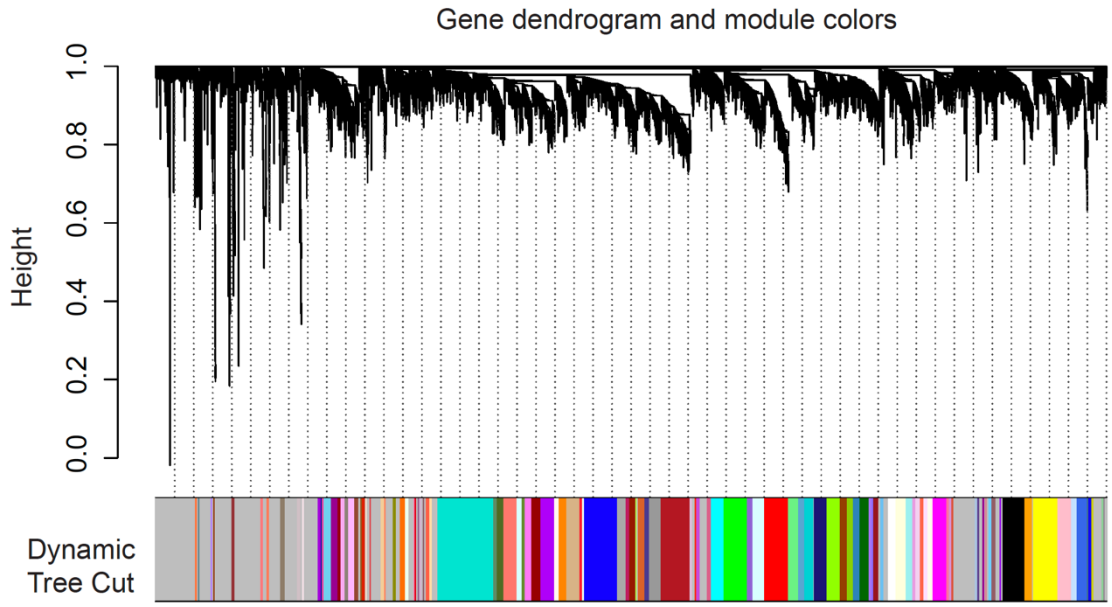


**Fig. S4**. WGCNA gene co-expression modules of *M. persicae* on 9 plant species. The dendogram groups 11,824 *M. persicae* genes into 77 distinct modules (color-coded) using hierarchical clustering. Dynamic cut tree was used to identify modules, dividing modules at significant branch points in the dendrogram. The x-axis shows the width of each module that is defined by the number of genes. The y-axis corresponds to distance determined by the extent of topological overlap (1-TOM). Genes not assigned to a module are labelled in grey.

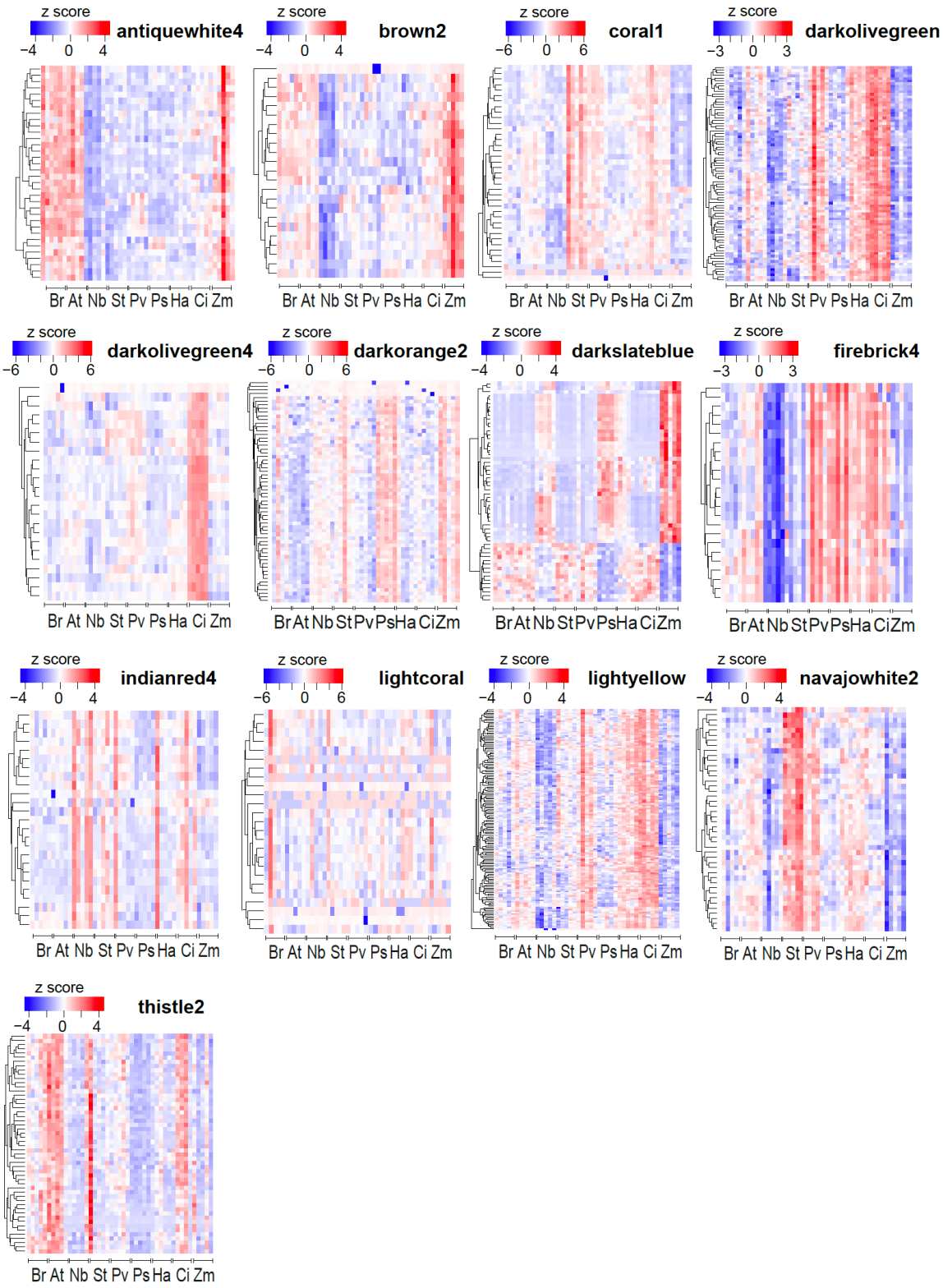


**Fig. S5.** Heatmaps of 13 modules enriched for DE genes of *M. persicae* colonies on 9 plant species. The colour codes of the modules are indicated above the

heatmaps. Hierarchical clustering was done based on plant species as follows: Br, *B. rapa*; At, *A. thaliana*; Nb, *N. benthamiana*; St, *S. tuberosum*; Ci, *C. indicum*; Ha, *H. annuus*; Ps, *P. sativum*; Pv, *P. vulgaris*; Zm, *Z. mays*. Rows are scaled based log transformed TPM values, which are shown using a z-score as indicated.

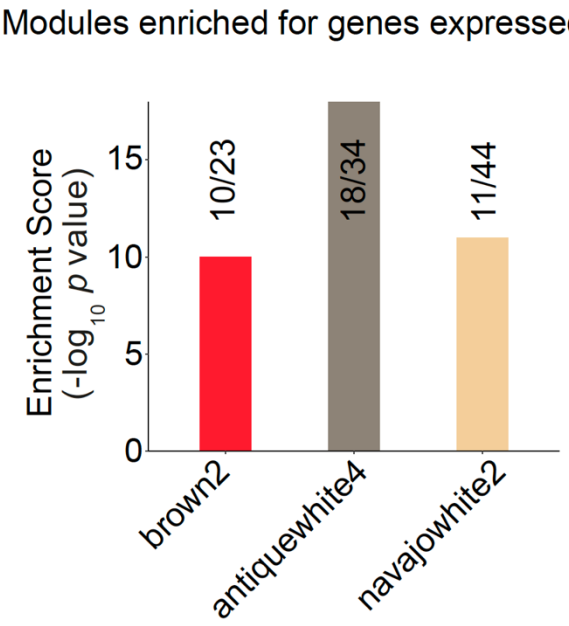


**Fig. S6**. Modules enriched for genes expressed in *M. persicae* guts. *M. persicae* DE genes were searched against a gut EST dataset (28). The x/y numbers above the bars indicate the number of genes in the enriched category (x) and the total number of genes in the module (y).

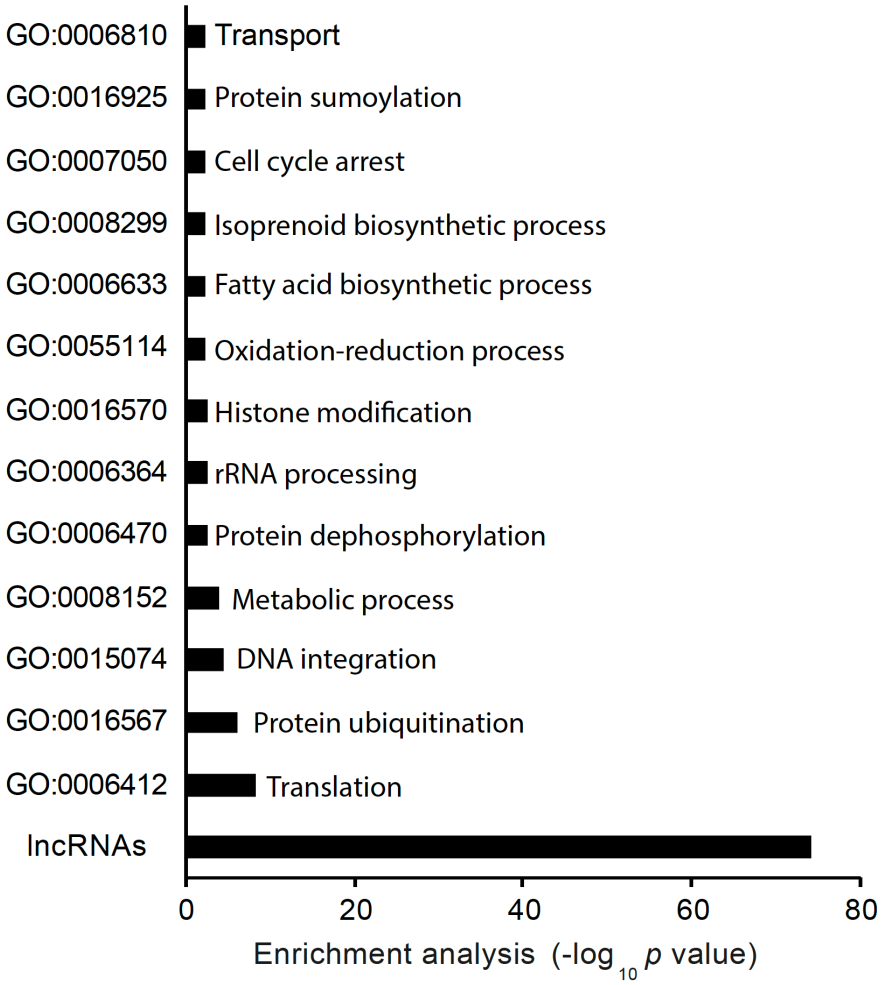


**Fig. S7.** Enrichment analysis of aphid transcripts found in the feeding sites. Gene functional annotations were used. Enrichment analysis was calculated with Fisher's Exact Test by comparing the ratios of genes that belong to a GO category versus the total number of genes between the feeding site and the whole aphid.

```
>Ya1_exon2_148nt
TTTCTCTTTTAAACCTAAAAAACCAACCAACAAATCAAAAATGGGCGCTGAAAAGGTATCCATGAACA
TCGTTGTCGTCGGACAAGTCCAATCTGCCAAGGCCATCAAGACCGTCGCCAAGGCCTCCCCAGCCACC
AACCAATTGATG
```

**Fig. S8.** Nucleotide sequence of *Ya1* exon 2 used to identify *Ya* family members in *M. persicae* and other aphid species**.**

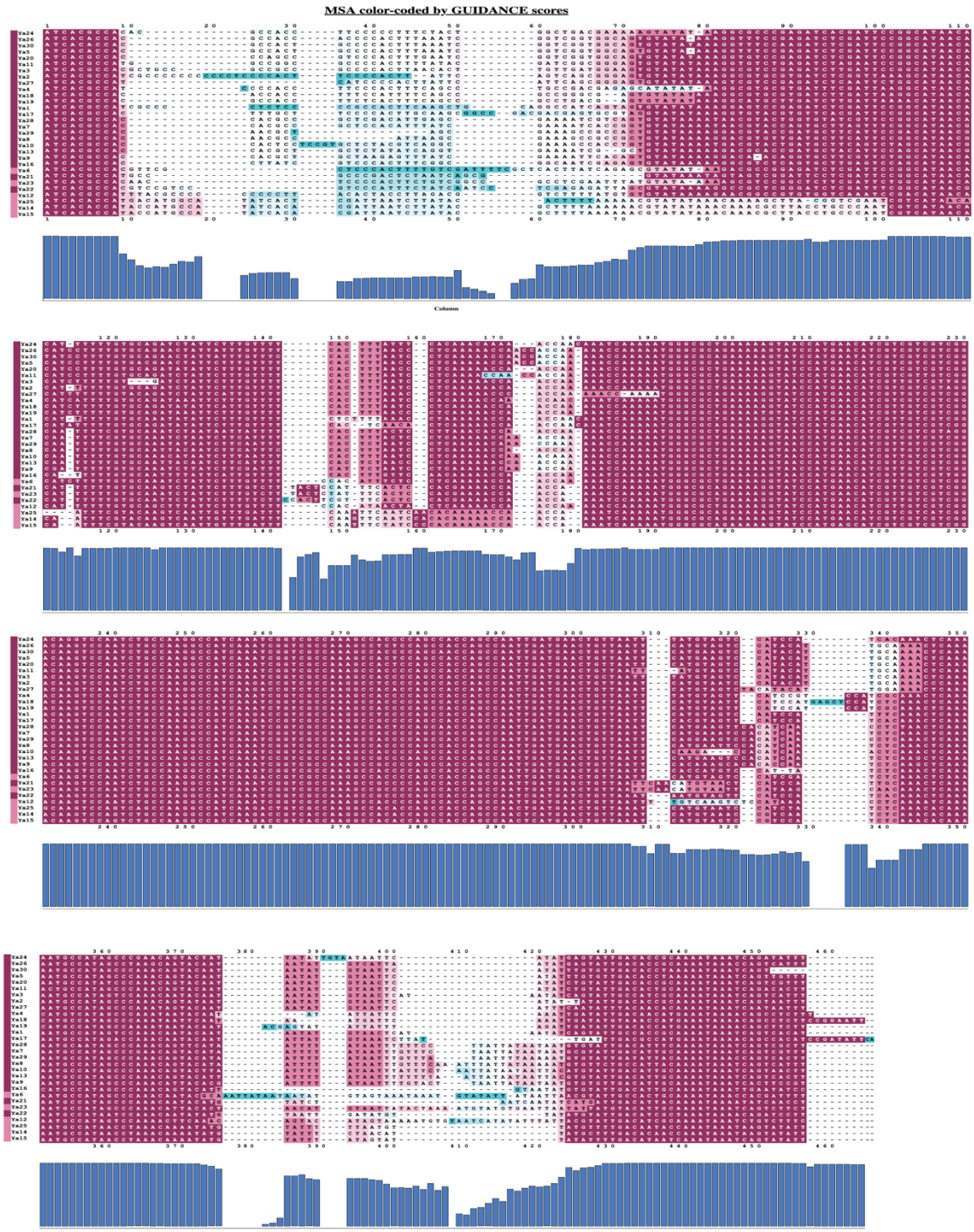


**Fig. S9.** Alignment of transcripts of 30 *Ya* gene family members. The alignment was used to generate Fig. 3A and the phylogenetic tree of Fig. 3B. Alignment was performed on online version of MAFFT (version 7, https://mafft.cbrc.jp/alignment/server/index.html) (19) with default settings. The

multiple alignment was subjected to GUIDANCE2 (http://guidance.tau.ac.il/ver2/) (20) to compute the residue-wise confidence scores and extract well-aligned residues. Nucleotides are highlighted (colour scale between red and turquoise) to indicate confidence of alignment (red is high confidence and turquoise is low). The bottom bar graph displays the scores of aligned nucleotides, as determined by GUIDANCE2.

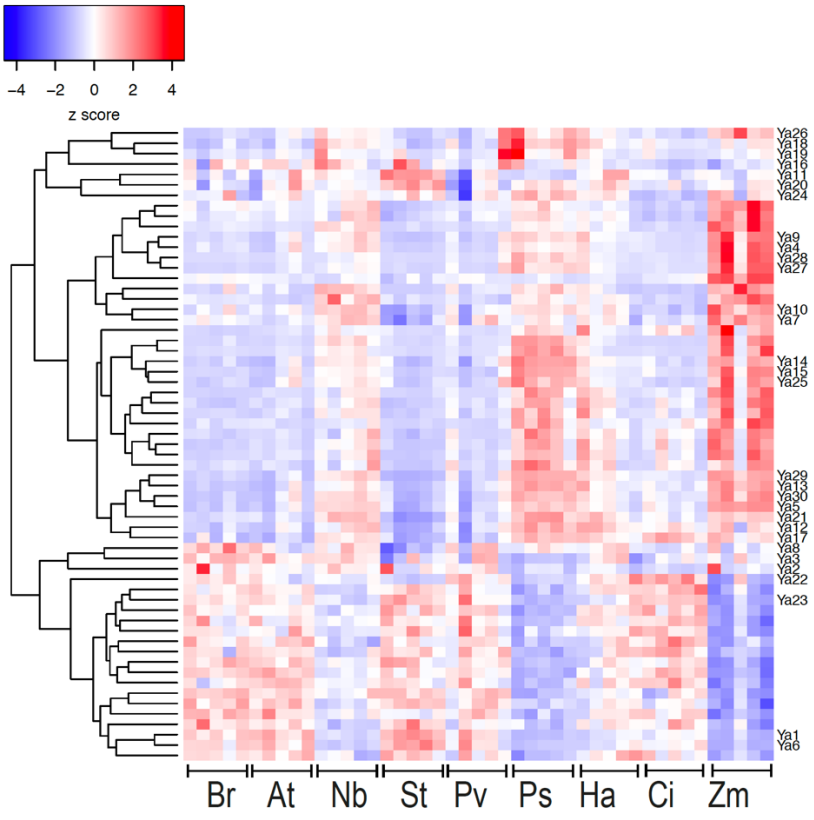


**Fig. S10.** Heatmap of darkslateblue module after manual annotation of the *Ya* family. WGCNA analysis was repeated with the corrected set of *Ya* genes. Heatmap was generated using log-transformed TPM values of all genes in the module. Br, *B. rapa*; At, *A. thaliana*; Nb, *N. benthamiana*; St, *S. tuberosum*; Ci, *C. indicum*; Ha, *H. annuus*; Ps, *P. sativum*; Pv, *P. vulgaris*; Zm, *Z. mays*. Rows are scaled and represented as z-score.

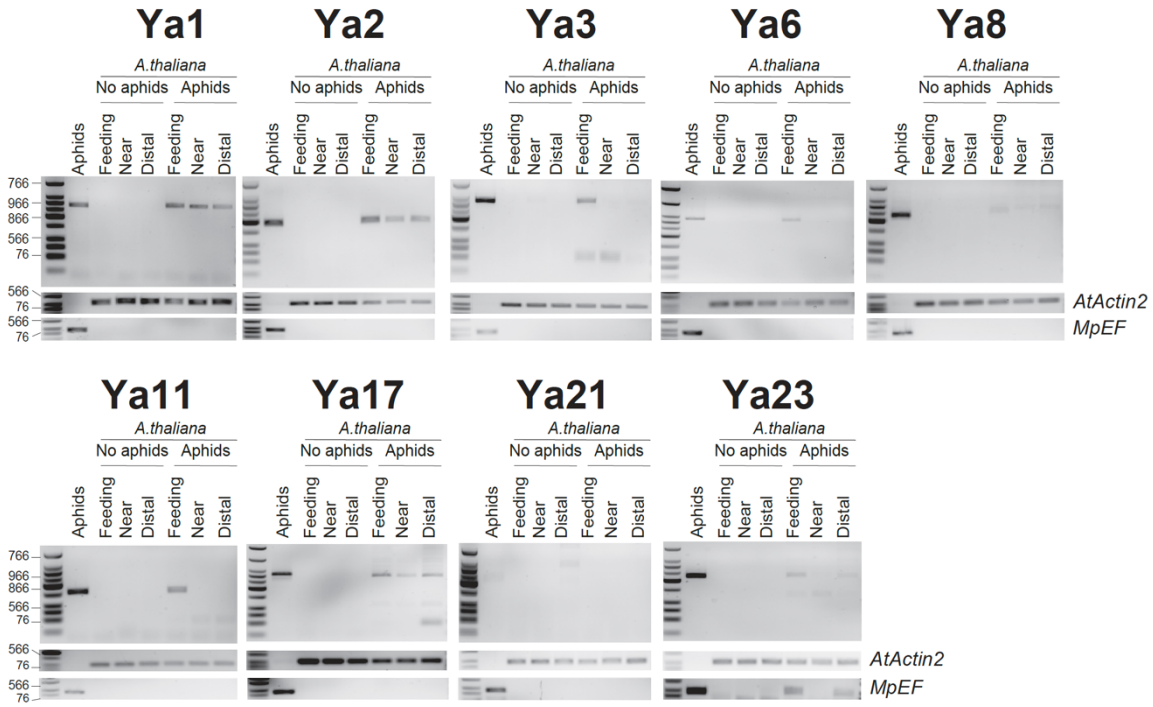


**Fig. S11.** RT-PCR experiments showing translocation of 9 *Ya* transcripts into plants and systemic migration of *Ya1*, *Ya2*, and *Ya17* transcripts. Experimental setup is shown in Fig. 4A. The *Ya* transcripts were amplified by RT-PCR using gene-specific primers (***SI Appendix*, Table S3**) and PCR products were separated by 3% agarose gel. The bands corresponding to the PCR products were extracted from the gel and sequenced directly by forward and/or reverse primers to verify the identity of the *Ya* sequence.

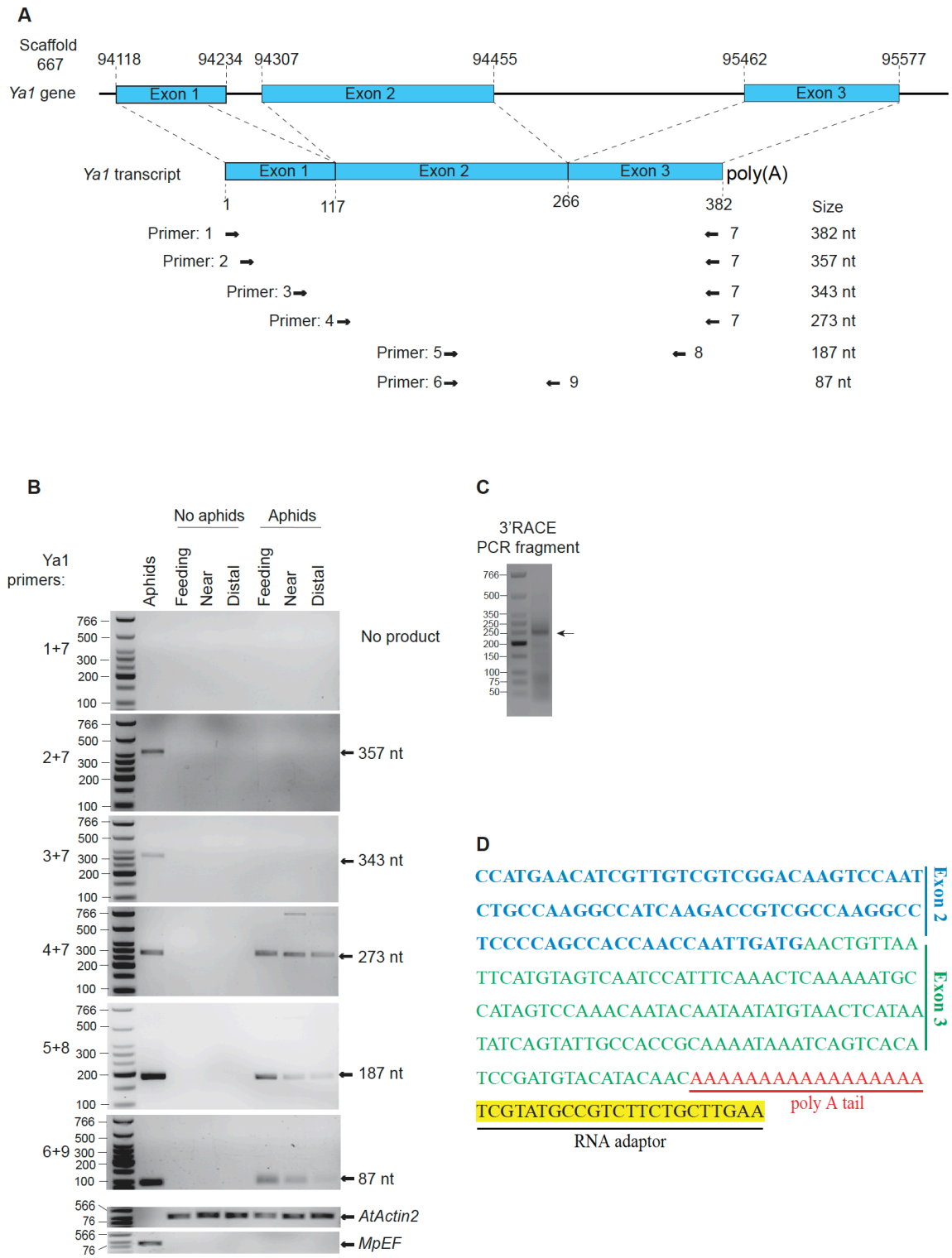


**Fig. S12.** Identification of *Ya1* transcript sizes in aphids and plants. (**A**) Transcript model of the *Ya1* gene. The three exons of the *Ya1* gene that form the transcript

are indicated as blue rectangles. The numbers above the Ya gene indicate the first and last nucleotide of each exon on scaffold 667 of the *M. persicae* genome assembly. The locations of the Ya1 primers 1-9 (***SI Appendix*, Table S2**) used to assess the presence of *Ya1* transcripts in aphids and in plants by RT-PCR are indicated as black arrows. (**B**) A 357-nt *Ya1* transcript was detected in aphids and not in plants, whereas the largest *Ya1* transcript detected in plants at aphid feeding sites and distal leaves is 273-nt (primers 4 and 7). Experimental setup is shown in Fig. 4A. RT-PCR products generated with primers shown in A were separated on 3% agarose gels. Fragment sizes are indicated at right of the gels and correspond with the expected sizes as shown in A. Primers 1 and 7 did not generate a product in aphids nor in plants. (**C**) Results of a 3' RACE experiment. The 3' RACE PCR products were separated a 3% agarose gel. The band isolated from the gel and sequenced is shown with an arrow at right of the gel. (**D**) Sequence of the 3' RACE PCR fragment of the band shown in C. The sequence is identical to the sequences of the end of exon 2 (letters in blue font) and the entire exon 3 (letters in green font) of the annotated *Ya1* gene in the *M. persicae* genome. The *Ya1* transcript has a poly(A) tail (letters in red font). The RNA adaptor sequence used in 3' RACE protocol is indicated in yellow.

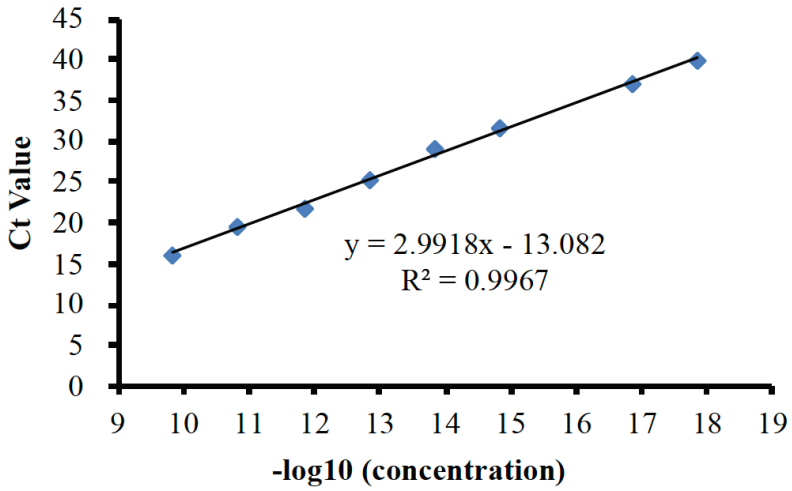


**Fig. S13.** Standard concentration curve of *Ya1* and cycles of PCR amplification. The 273-nt Ya1 fragment in the plasmid pBI121 was used as the PCR template. qPCR were performed on a serial of dilution of the plasmid pBI121_35S::Ya1 from the highest concentration $1.44E^{-10}$ g/ µL to the lowest concentration $1.44E^{-18}$ g/ µL. x axis are minus log 10 transferred concentrations.

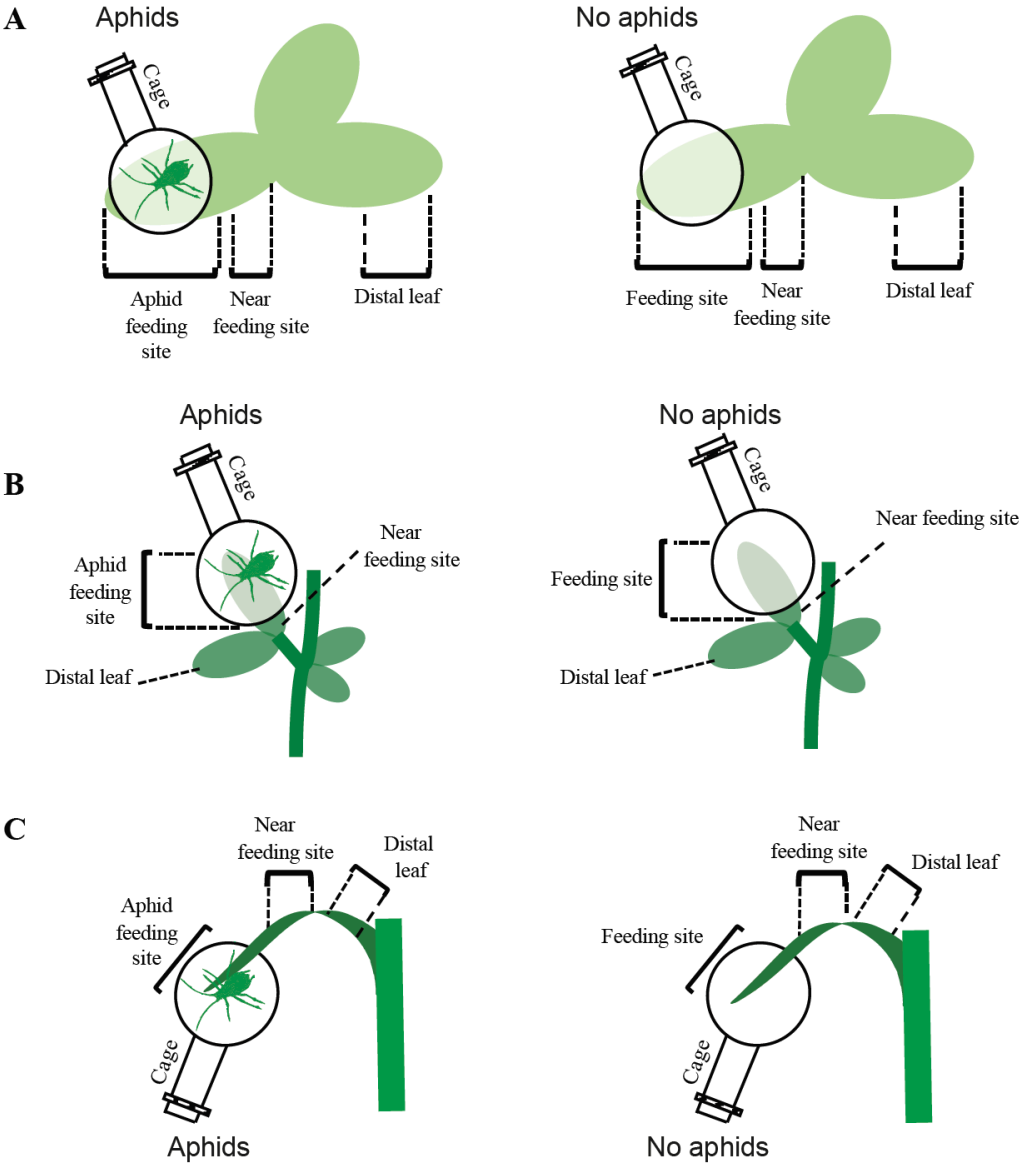


**Fig. S14**. Schematic overview of experimental setups to investigate systemic migration of aphid *Ya1* transcripts in *Brassica rapa*, *Pisum sativum* and *Zea mays*. (A) *Brassica rapa.* (B) *Pisum sativum.* (C) *Zea mays*. Aphid feeding site is the caged parts of the leaf. Distal sites include near-feeding sites that are located just externally of the caged leaf areas and more distal locations that are different leaves for *B. rapa* and *P. sativum* and further away of the caged area on the same leaf of *Z. mays*.

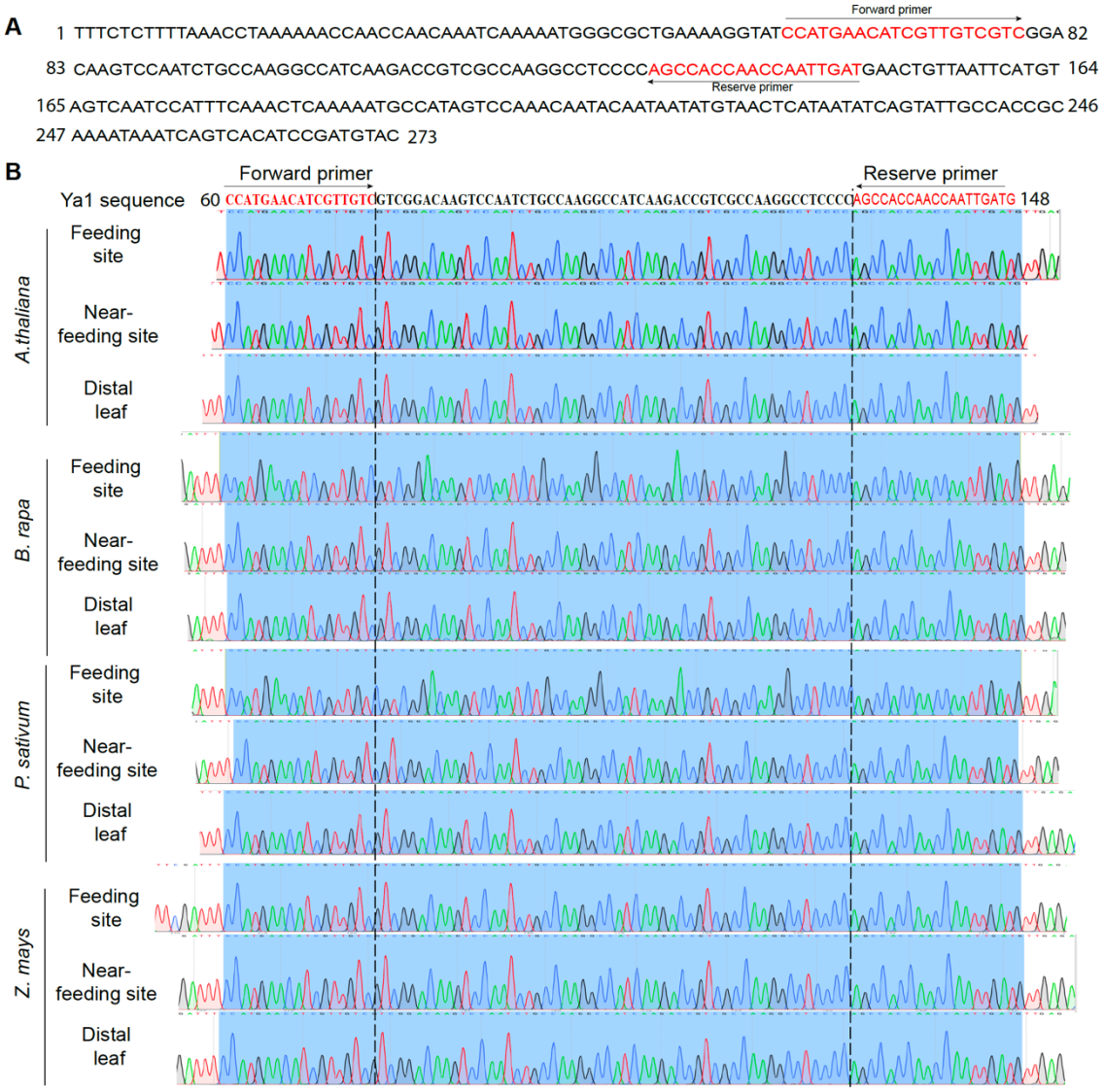


**Fig. S15.** Sequences of RT-PCR products of *M. persicae Ya1* transcripts amplified from plants. The *Ya1* transcripts were amplified with Ya1 forward and reverse primers from feeding sites, near-feeding sites, and distal sites of *A. thaliana*, *B. rapa*, *P. sativum* and *Z. mays* as per experimental setups shown in **Fig. 4A** and **Suppl. Fig. S13**. The sequences shown were identical to the corresponding region within the *Ya1* gene.

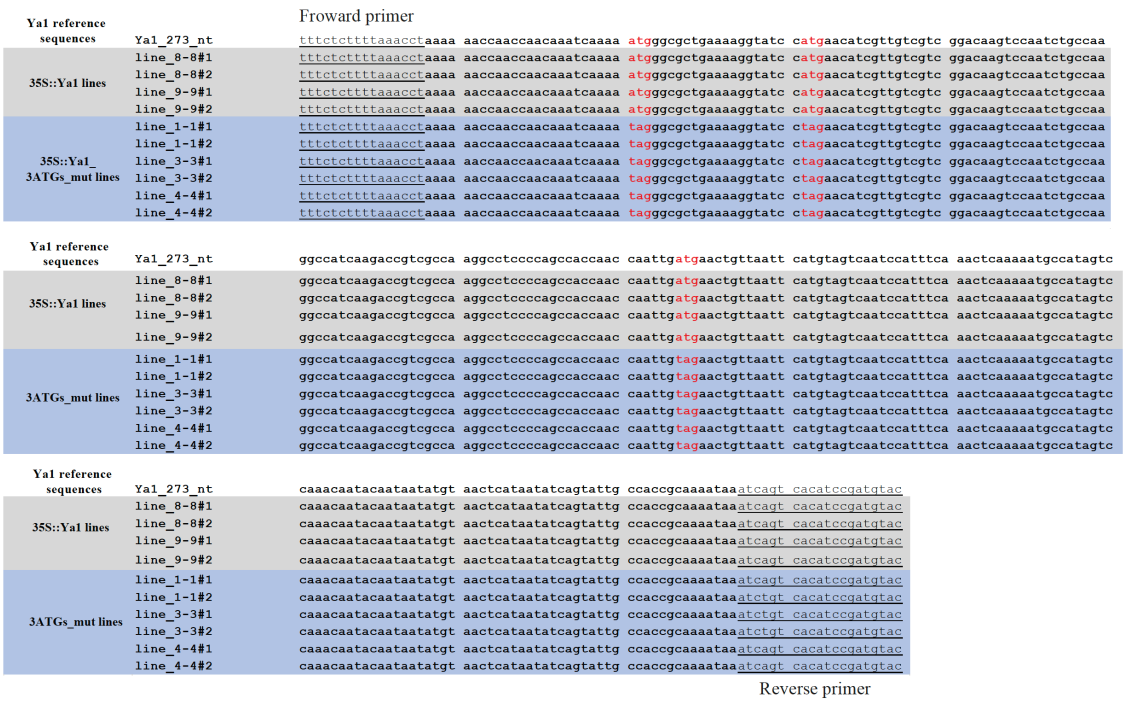


**Fig. S16.** Sequences of RT-PCR products of *Ya1* and *Ya1_3ATGs_mut* transcripts amplified from transgenic *A. thaliana* plants (Figure 5D). Nucleotides underlined lowercase are sequences of forward and reverse primer. ATG and TAG (mutated ATG) are in red.

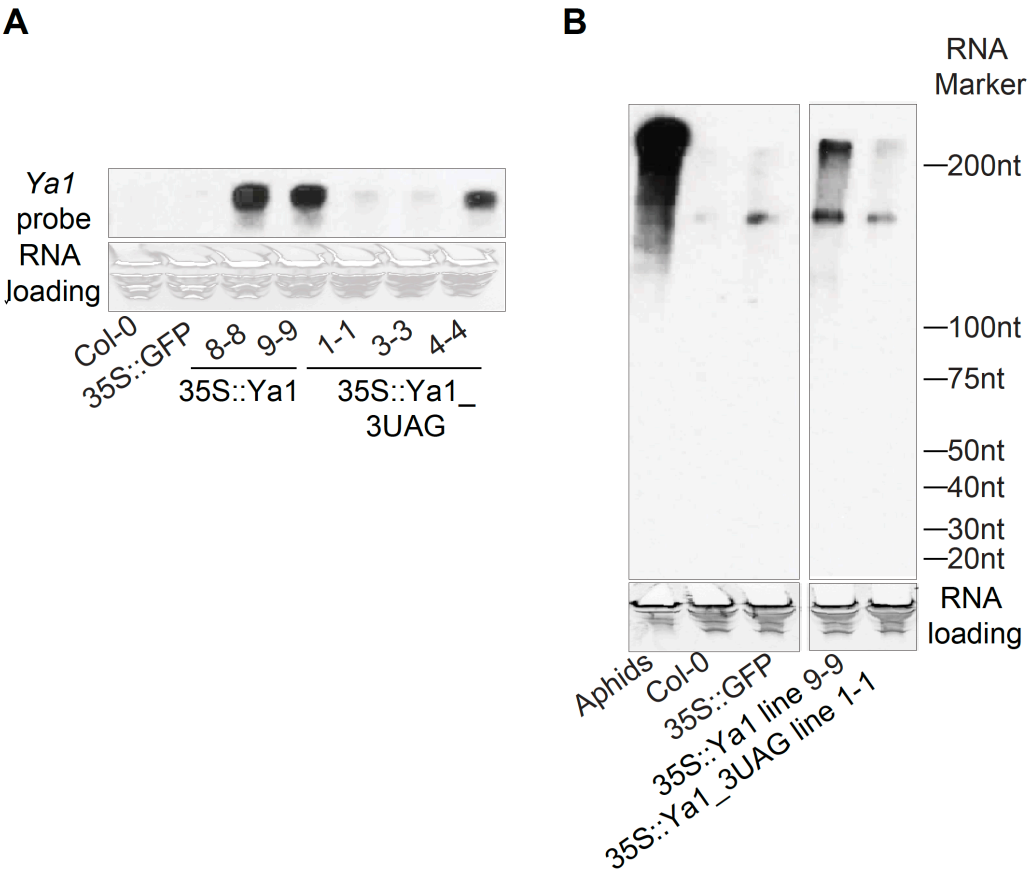


**Fig. S17.** Northern blot to confirm expression of *Ya1* and *Ya1_3UAG* expressed in transgenic *A. thaliana* plants. (A) Northern blot hybridizations with a *Ya1* probe to detect *Ya1* and *Ya1_3UAG* expressed in transgenic *A. thaliana* plants. (B) Full length of *Ya1* and *Ya1_3UAG* were detected in the transgenic plants. Prestain RNA markers (DynaMarker) were used to estimate the size of RNAs.

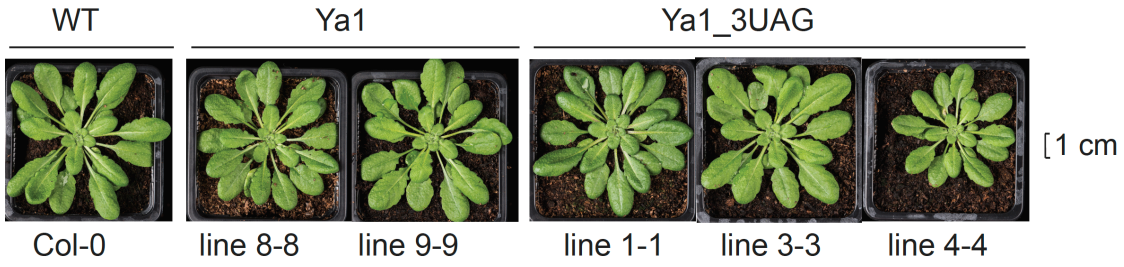


**Fig. S18.** Comparison of phenotypes of *Ya1* transgenic *A. thaliana* plants. Transgenic *A. thaliana* (Col-0) plants that express *Ya1* and *Ya1_3UAG* under control of the *35S* promoter did not show obvious morphological differences compared to the non-transformed wild type Col-0 plants.

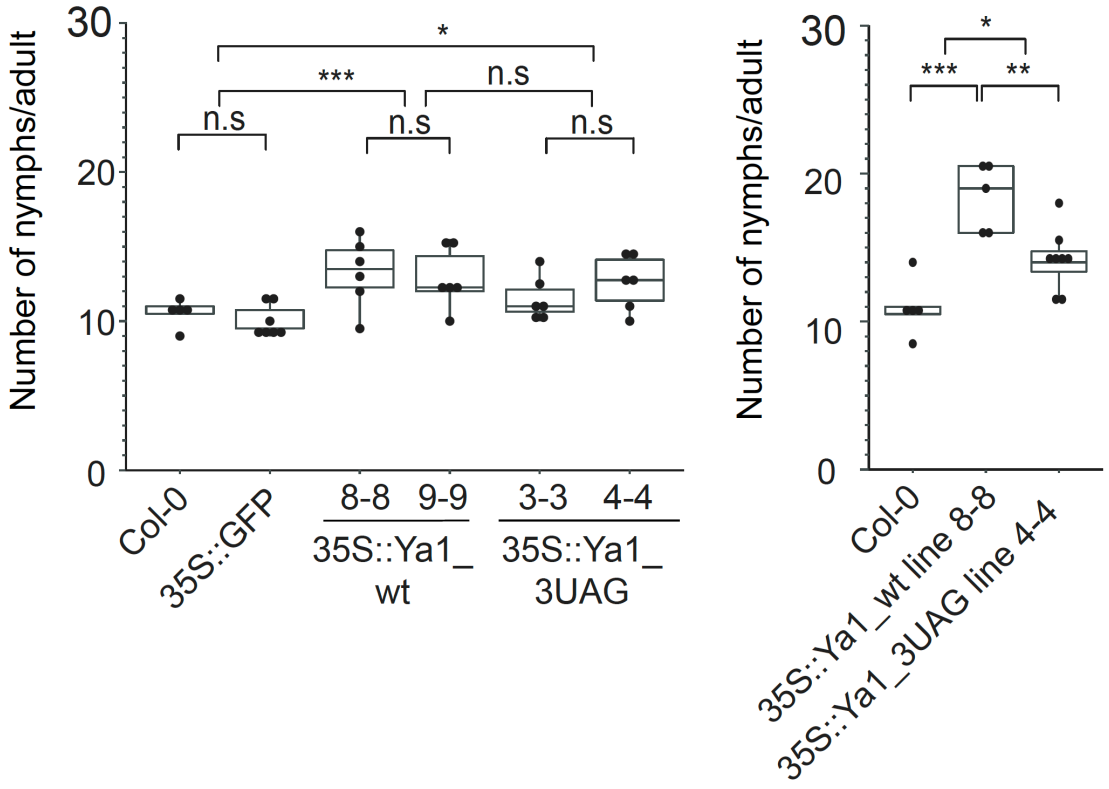


**Fig. S19.** Repeats showing that *M. persicae Ya* lncRNA promotes *M. persicae* colonization on *A. thaliana*. Stable expression of *Ya1_wt* and *Ya1_3UAG* promotes *M. persicae* reproduction on plants. Each data point (black dot) represents number of nymphs produced by an adult female aphid per plant. Box plots show distribution of data points collected from n = 5-8 female aphids per *A. thaliana* line. *$p < 0.05$, ** $p < 0.01$, ***$p < 0.001$, ANOVA followed by a Tukey–Kramer post-hoc test.

**Table S1.** Transcriptome assembly statistics. RNA-seq reads were derived from samples LIB1777 (1) and colonies on 5 independent plants (1-5) of *Arabidopsis thaliana* (At); Br, *Brassica rapa* (Br), *Nicotiana benthamiana* (Nb), *Solanum tuberosum* (St), *Chrysanthemum indicum* (Ci), *Helianthus annuus* (Ha), *Pisum sativum* (Ps), *Phaseolus vulgaris* (Pv) and *Zea mays* (Zm). The reads were mapped to the *M. persicae* G006 genome assembly (1).

[a] Total transcripts were assembled from mapped reads.

[b,c] Transcripts that correspond to genes identified previously in Mathers et al. 2017 (1).

[d] Novel transcript as identified in the transcriptome assembly reported herein.

| Samples | Total reads | Mapped reads | Percentage of mapped reads | Transcripts | | | |
|---|---|---|---|---|---|---|---|
| | | | | Total[a] | Known[b] | Genes[c] | Novel[d] |
| LIB1777 | 182928264 | 117579937 | 64.28 | 87444 | 18750 | 11332 | 68694 |
| At1 | 35027735 | 31609619 | 90.24 | | | | |
| At2 | 30317626 | 27641464 | 91.17 | | | | |
| At3 | 32270976 | 29367540 | 91 | 61983 | 17781 | 10486 | 44202 |
| At4 | 33889504 | 30765104 | 90.78 | | | | |
| At5 | 29871796 | 27369312 | 91.62 | | | | |
| Ps1 | 33997244 | 31321585 | 92.13 | | | | |
| Ps2 | 33572880 | 30992391 | 92.31 | | | | |
| Ps3 | 30706147 | 28148247 | 91.67 | 60443 | 17737 | 10528 | 42706 |
| Ps4 | 38851016 | 35655100 | 91.77 | | | | |
| Ps5 | 23736763 | 20424890 | 86.05 | | | | |
| Br1 | 36105440 | 33390772 | 92.48 | | | | |
| Br2 | 27368493 | 23985300 | 87.64 | | | | |
| Br3 | 28484604 | 26110873 | 91.67 | 58293 | 17750 | 10489 | 40543 |
| Br4 | 27486534 | 25035464 | 91.08 | | | | |
| Br5 | 30044024 | 27584912 | 91.81 | | | | |
| Ci1 | 31213071 | 28395530 | 90.97 | | | | |
| Ci2 | 26059223 | 23430480 | 89.91 | | | | |
| Ci3 | 27469301 | 24621304 | 89.63 | 65917 | 17856 | 10621 | 48061 |
| Ci4 | 32534564 | 29376935 | 90.29 | | | | |
| Ci5 | 36510146 | 33342272 | 91.32 | | | | |
| Zm1 | 23898955 | 21330499 | 89.25 | | | | |
| Zm2 | 32663254 | 30362632 | 92.96 | | | | |
| Zm3 | 38833395 | 32998965 | 84.98 | 61157 | 17886 | 10610 | 43271 |
| Zm4 | 36223785 | 33199732 | 91.65 | | | | |
| Zm5 | 37985100 | 34900276 | 91.88 | | | | |
| Nb1 | 26749263 | 23973979 | 89.62 | | | | |
| Nb2 | 28476921 | 25844725 | 90.76 | | | | |
| Nb3 | 30481687 | 28348224 | 93 | 57533 | 17567 | 10444 | 39966 |
| Nb4 | 31075574 | 28936577 | 92.12 | | | | |
| Nb5 | 25055052 | 22819867 | 91.08 | | | | |

| Samples | Total reads | Mapped reads | Percentage of mapped reads | Transcripts | | | |
|---|---|---|---|---|---|---|---|
| | | | | Total[a] | Known[b] | Genes[c] | Novel[d] |
| Ps1 | 22202831 | 19882967 | 89.55 | | | | |
| Ps2 | 26907680 | 24899984 | 92.54 | | | | |
| Ps3 | 30923478 | 28678037 | 92.74 | 56431 | 17590 | 10488 | 38841 |
| Ps4 | 37835896 | 34845885 | 92.10 | | | | |
| Ps5 | 30745946 | 28130700 | 91.49 | | | | |
| St1 | 22647808 | 19265819 | 85.07 | | | | |
| St2 | 26239515 | 23358283 | 89.02 | | | | |
| St3 | 29668113 | 27521716 | 92.77 | 60662 | 17731 | 10537 | 42931 |
| St4 | 32534760 | 30241012 | 92.95 | | | | |
| St5 | 32905595 | 30331494 | 92.18 | | | | |
| Ha1 | 34475345 | 31538679 | 91.48 | | | | |
| Ha2 | 39135872 | 36179328 | 92.45 | | | | |
| Ha3 | 32762153 | 29883530 | 91.21 | 58152 | 17744 | 10490 | 40408 |
| Ha4 | 20135006 | 18438242 | 91.57 | | | | |
| Ha5 | 24340322 | 22404095 | 92.05 | | | | |
| **Total** | | | | 45972 | 30127 | 18529 | 15845 |

1

**Table S2.** Statistics of analyses of RNA-seq data retrieved from aphid-exposed (feeding sites) and non-exposed (control) leaves. RNA-seq reads derived from four biological replicates per treatment were mapped to both *A. thaliana* Col-0 (TAIR10) and *M. persicae* (Mp) genomes (1). *Reads mapping to Mp were realigned to At to find uniquely mapping Mp reads. The unique reads were then assigned to the transcripts obtained from the transcriptome assembly v2. ^ represents transcripts at TPM ≥ 50 in at least one biological replicate and presence in at least three replicates.

| | Biological Replicate | Total reads (million) | Reads mapped to At (million) | Reads mapped to At (%) | Reads mapped to Mp | Reads mapped to Mp (%) | Reads mapped to At | Unique Mp reads* | Mp transcripts (TPM ≥ 50) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | Number of transcripts | Number of low-coding potential transcripts |
| Aphid feeding sites | 1 | 28 | 25 | 90.61 | 18007 | 0.0631 | 5 | 18002 | 2928 | 170 |
| | 2 | 21 | 19 | 90.02 | 242628 | 1.1091 | 0 | 242628 | 2599 | 280 |
| | 3 | 24 | 22 | 90.39 | 3847 | 0.0153 | 2 | 3845 | 1837 | 79 |
| | 4 | 24 | 21 | 88.11 | 657024 | 2.6969 | 3 | 657021 | 3154 | 336 |
| | | | | | | | | | **3186^** | **201^** |
| No aphid control sites | 1 | 31 | 28 | 90.53 | 35 | 0.0001 | 23 | 12 | 0 | 0 |
| | 2 | 22 | 20 | 91.08 | 9 | 0.0000 | 5 | 4 | 0 | 0 |
| | 3 | 28 | 26 | 90.79 | 5 | 0.0000 | 5 | 0 | 0 | 0 |
| | 4 | 21 | 19 | 90.55 | 17 | 0.0001 | 14 | 3 | 0 | 0 |

**Table S3**. **Primers used in this study.**

**Primers for qRT-PCR for RNAi**

| **Primer names** | **Sequence (5'> 3')** |
|---|---|
| Ya1 F | ATCAAGACCGTCGCCAAG |
| Ya1 R | TTGCGGTGGCAATACTGATA |

**Primers for qRT-PCR for detection in plants**

| **Primer names** | **Sequence (5'> 3')** |
|---|---|
| Ya1 primer1 | ATCACGCCATCGCCCCTCTC |
| Ya1 primer2 | AAGCTGCAGGCCATCAGT |
| Ya1 primer3 | AGTAGTGTATATAAAGGAGTCC |
| Ya1 primer4 | CGATTCCGGCATAACACAT |
| Ya1 primer5 | GTTTTCTCTTTTAAACCT |
| Ya1 primer6 | CCATGAACATCGTTGTCGTC |
| Ya1 primer7 | ATGTGACTGATTTATTTTGC |
| Ya1 primer8 | GCGGTGGCAATACTGATATTAT |
| Ya1 primer9 | CATCAATTGGTTGGTGGCT |
| Ya1 Forward | CCATGAACATCGTTGTCGTC |
| Ya1 Reverse | GCGGTGGCAATACTGATATTAT |
| Ya2 Forward | ACAAGTCCAATCTGCCAAGG |
| Ya2 Reverse | TGCGGTGTCAATATAATATG |
| Ya3F Forward | CTTGATTCCGGCATAACACATC |
| Ya3 Reverse | AAATGACTGATTTATTTTGC |
| Ya6 Forward | CATAACACATCTTTTTGTG |
| Ya6 Reverse | AAAATCGTGGTTTATTTTGC |
| Ya8 Forward | CCATGAACATCGTTGTCGTC |
| Ya8 Reverse | AAATGACTGATTTATTTTGT |
| Ya11 Forward | CCACCAAAAACCAAAAATGG |
| Ya11 Reverse | AAGTGTCTGATTTATTTTGC |
| Ya17 Forward | GGCATAACACATATTTTTGT |
| Ya17 Reverse | TGAATATCGGATGTGGCTGTT |
| Ya23 Forward | CTTCTGTCGGCCGCCTCGAATT |
| Ya23 Reverse | AGAAGACTGATTTATTTTGCAA |

**Primers for cloning**

**Underlined lower case nucleotides indicate restriction enzyme sequences**

| **Primer names** | **Sequence (5'> 3')** |
|---|---|
| Ya1_F (273 nt) | TTTCTCTTTTAAACCTAAAAAA |
| Ya1_R (273 nt) | GTACATCGGATGTGACTGATT |
| attB1_Ya1 | acaagtttgtacaaaaaagcaggctagtTTTCTCTTTTAAACCTAAAAAA |
| attB2_Ya1 | accactttgtacaagaaagctgggt GTACATCGGATGTGACTGATT |
| Ya1_BamHI | cgggggactctagaggatccagtTTTCTCTTTTAAACCTAA |
| Ya1_SacI | aaattcgagctcGTACATCGGATGTGACTGATTT |
| MutateF | cgggggactctagaggatccagtTTTCTCTTTTAAACCTAA |
| ATG41_F | TTTAAACCTAAAAAACCAACCAACAAATCAAAATAGGGCGCTGAAAAGGTAT |
| ATG41_R | ATACCTTTTCAGCGCCCTATTTTGATTTGTTGGTTGGTTTTTTAGGTTTAAA |
| ATG62_F | GGCGCTGAAAAGGTATCCTAGAACATCGTTGTCGTCGG |
| ATG62_R | CCGACGACAACGATGTTCTAGGATACCTTTTCAGCGCC |
| ATG146_F | CCCCAGCCACCAACCAATTGTAGAACTGTTAATTCATGTAGT |
| ATG146_R | ACTACATGAATTAACAGTTCTACAATTGGTTGGTGGCTGGGG |
| MutateR | aaattcgagctcGTACATCGGATGTGACTGATTT |

**Probes for northern blot hybridizations**

| **Probe** | **Sequence (5'> 3')** |
|---|---|
| Probe of AtU6 | CTCGATTTATGCGTGTCATCCTTGC |

**Adapter and primers for 3'RACE**

| **Names** | **Sequence (5'> 3')** |
|---|---|
| Ya1 Forward primer | GGACAAGTCCAATCTGC |
| 3' RNA Adapter (5' Phosphorylation and 3' Inverted dT) | P-UCGUAUGCCGUCUUCUGCUUGUidT |
| oligo sequence complementary to 3'RNA adapter | CAAGCAGAAGACGGCATACGA |

## Datasets

**Dataset S1**. List of differentially expressed transcripts of *M. persicae* colonies on 8 divergent plant species compared to *M. persicae* on *Arabidopsis thaliana* (At); Br, *Brassica rapa* (Br), *Nicotiana benthamiana* (Nb), *Solanum tuberosum* (St), *Chrysanthemum indicum* (Ci), *Helianthus annuus* (Ha), *Pisum sativum* (Ps), *Phaseolus vulgaris* (Pv) and *Zea mays* (Zm). 1, differentially expressed (DE) transcript; 0, not DE. Transcript ID refers to the transcriptome assembly *M. persicae* G006 v2 (GSE129669) generated herein.

**Dataset S2**. Results of WGCNA analyses with 13 modules enriched in differentially expressed genes (Tab A) and the 64 other co-expression modules (Tab B). Genes of tandem repeats in the scaffold are highlighted in yellow or green and numbers of tandemly repeated groups identified in each module are summarized in the table at right. ^ indicates whether genes are differentially expressed (DE), encode candidate lncRNAs, and are expressed in salivar glands and guts of *M. persicae*.

**Dataset S3**. List of *M. persicae* candidate lncRNAs (Tab A) and other transcripts (Tab B) found *in A. thaliana* leaves exposed to *M. persicae* (feeding sites).

**Dataset S4**. Characteristics of the 30 *Ya* gene family members upon manual annotation. Gene IDs correspond to version 3 GFF generated after manual correction.

**Dataset S5**. Annotation of *Ya* gene in five aphid species in addition to *M. persicae*. Accession numbers and versions of genome assemblies and RNASeq data used for annotations of the *Ya* genes (Tab A); Genomic locations, strand information and annotation of *Ya* genes (Tab B).

## SI References


1. T. C. Mathers, Y. Chen, G. Kaithakottil, F. Legeai, S. T. Mugford, P. Baa-Puyoulet, A. Bretaudeau, B. Clavijo, S. Colella, O. Collin, T. Dalmay, T. Derrien, H. Feng, T. Gabaldón, A. Jordan, I. Julca, G. J. Kettles, K. Kowitwanich, D. Lavenier, P. Lenzi, S. Lopez-Gomollon, D. Loska, D. Mapleson, F. Maumus, S. Moxon, D. R. Price, A. Sugio, M. van Munster, M. Uzest, D. Waite, G. Jander, D. Tagu, A. C. Wilson, C. van Oosterhout, D. Swarbreck, S. A. Hogenhout, Rapid transcriptional plasticity of duplicated gene clusters enables a clonally reproducing aphid to colonise diverse plant species. *Genome Biol.* **18**, 27 (2017).
2. F. Krueger, Trim Galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries. URL http://www. bioinformatics. babraham. ac. uk/projects/trim_galore/. (2012).
3. M. Pertea, D. Kim, G. M. Pertea, J. T. Leek, S. L. Salzberg, Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650-1667 (2016).
4. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R .Durbin, 1000 Genome Project Data Processing Subgrou, The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
5. M. Pertea, G. M. Pertea, C. M. Antonescu, T. C. Chang, J. T. Mendell, S. L. Salzberg, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290-295 (2015).
6. A. Roberts, H. Pimentel, C. Trapnell, L. Pachter, Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27**, 2325-2329 (2011).
7. L. Venturini, S. Caim, G. G. Kaithakottil, D. L. Mapleson, D. Swarbreck, Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *Gigascience* **7** (2018).

8. A. Bairoch, R. Apweiler, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic. Acids Res.* **28**, 45-48 (2000).
9. R. D. Finn, P .Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, A. Bateman, The Pfam protein families database: towards a more sustainable future. *Nucleic. Acids Res.* **44**, D279-285 (2016).
10. Y. J. Kang., D. C. Yang, L. Kong 1, M. Hou, Y. Q. Meng, L. Wei, G. Gao, CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic. Acids Res.* **45**, W12-W16 (2017).
11. S. E. Lewis, S. M. Searle, N. Harris, M. Gibson, V. Lyer, J. Richter, C. Wiel, L. Bayraktaroglu, E. Birney, M. A. Crosby, J. S. Kaminker, B. B. Matthews, S. E. Prochnik, C. D. Smithy, J. L. Tupy, G. M. Rubin, S. Misra, C. J. Mungall, M. E. Clamp, Apollo: a sequence annotation editor. *Genome Biol.* **3**, RESEARCH0082 (2002).
12. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
13. P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
14. J. S. Ramsey, A. C. Wilson, M. de Vos, Q. Sun, C. Tamborindeguy, A. Winfield, G. Malloch, D. M. Smith, B. Fenton, S. M. Gray, G. Jander, Genomic resources for *Myzus persicae*: EST sequencing, SNP identification, and microarray design. *BMC Genomics* **8**: 423 (2007).
15. M. D. Nelson, D. H. Fitch, Overlap extension PCR: an efficient method for transgene construction. *Methods Mol. Biol.* **772**, 459-470 (2011).
16. N .Bechtold, J. Ellis, G. Pelletier, In planta Agrobacterium-mediated gene transfer by infiltration of adult *Arabidopsis thaliana* plants. *C. R. Acad. Sci. Ser. III Sci. Vie. Life Sci*. **316**,1194–1199 (1993).
17. T. D. Schmittgen, K. J. Livak, Analyzing real-time PCR data by the comparative C(T) method. *Nat. Protoc.* **3**, 1101-1108 (2008).
18. F. Supek, M. Bosnjak, N. Skunca, T. Smuc, REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**, e21800 (2011).

19. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol*. **30**, 772-780 (2013).
20. I. Sela, H. Ashkenazy, K. Katoh, T. Pupko, GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic. Acids Res.* **43**, W7-14 (2015).