Shared Transcriptional Control and Disparate Gain and Loss of Aphid Parasitism Genes

Peter Thorpe[1,2], Carmen M. Escudero-Martinez[1,2], Peter J.A. Cock[2,3], Sebastian Eves-van den Akker[4,*], and Jorunn I.B. Bos[1,2,5,*]

[1]Cell and Molecular Sciences, The James Hutton Institute, Dundee, United Kingdom

[2]Dundee Effector Consortium, The James Hutton Institute, Dundee, United Kingdom

[3]Information and Computational Sciences, The James Hutton Institute, Dundee, United Kingdom

[4]Department of Plant Sciences, University of Cambridge, Cambridge, United Kingdom

[5]Division of Plant Sciences, School of Life Sciences, University of Dundee, Dundee, United Kingdom

*Corresponding authors: E-mails: j.bos@dundee.ac.uk; se389@cam.ac.uk.

Accepted: August 23, 2018

Data deposition: This project has been deposited at NCBI under the BioProject accessions PRJEB24287, PRJEB24204, PRJEB24317 and PRJEB24338. Genome assemblies and gene calls are available at http://bipaa.genouest.org/is/aphidbase/ and DOI:10.5281/zenodo.125293410.5281/zenodo.1252934.

Abstract

Aphids are a diverse group of taxa that contain agronomically important species, which vary in their host range and ability to infest crop plants. The genome evolution underlying agriculturally important aphid traits is not well understood. We generated draft genome assemblies for two aphid species: *Myzus cerasi* (black cherry aphid) and the cereal specialist *Rhopalosiphum padi*. Using a de novo gene prediction pipeline on both these, and three additional aphid genome assemblies (*Acyrthosiphon pisum, Diuraphis noxia*, and *Myzus persicae*), we show that aphid genomes consistently encode similar gene numbers. We compare gene content, gene duplication, synteny, and putative effector repertoires between these five species to understand the genome evolution of globally important plant parasites. Aphid genomes show signs of relatively distant gene duplication, and substantial, relatively recent, gene birth. Putative effector repertoires, originating from duplicated and other loci, have an unusual genomic organization and evolutionary history. We identify a highly conserved effector pair that is tightly physically linked in the genomes of all aphid species tested. In *R. padi*, this effector pair is tightly transcriptionally linked and shares an unknown transcriptional control mechanism with a subset of ~50 other putative effectors and secretory proteins. This study extends our current knowledge on the evolution of aphid genomes and reveals evidence for an as-of-yet unknown shared control mechanism, which underlies effector expression, and ultimately plant parasitism.

Key words: aphids, effectors, genome evolution, shared transcriptional control, horizontal gene transfer.

Introduction

Among the over 5,000 aphid species described to date, about 250 are important agricultural pests (Blackman and Eastop 2000). These aphid species are highly diverse with regards to many phenotypic and ecological traits. Interestingly, while host specialization on a single or few plant species is common, some aphid species have evolved to infest a wide range of plant species, including from different families. How interactions with biotic factors have shaped aphid diversity is a complex and unanswered question. With increasing numbers of aphid genomes becoming available, it is possible to interrogate the evolution of genes that are predicted to play a role in aphid–environment interactions, such as host parasitism.

Genome sequences have become available for four different aphid species, *Acyrthosiphum pisum* (pea aphid)

(IAGC 2010), *Myzus persicae* (green-peach aphid) (Mathers et al. 2017), *Diuraphis noxia* (Russian wheat aphid) (Nicholson et al. 2015), and most recently, *Aphis glycines* (soybean aphid) (Wenger et al. 2017). Already, this has led to important discoveries, such as the association of duplicated gene cluster transcriptional plasticity in the broad host range *M. persicae* with colonization of diverse host species (Mathers et al. 2017), and the discovery that genes involved in carotenoid biosynthesis in the pea aphid were acquired by horizontal gene transfer (HGT) from fungi (Moran and Jarvik 2010). Screening of the pea aphid genome for genes putatively acquired by HGT from bacteria identified only 12 candidates, of which at least eight appeared to be functional based on expression data (Nikoh et al. 2010). HGT could therefore have played a role in the acquisition of novel important aphid traits, but the extent of its impact on aphid genome evolution, and host–parasite interactions, remains unclear.

Recent progress in the field revealed that a molecular dialog takes place between plants and aphids leading to activation of plant defenses in resistant plants (reviewed by Jaouannet et al. [2014]), or the suppression of host defenses and/release of nutrients in susceptible plants (Girousse et al. 2005; Will et al. 2007; Wilson et al. 2011). Aphid effectors, which are molecules delivered inside host plant cells and the apoplast during probing and feeding, play an important role in the infestation process in that they contribute to host susceptibility by targeting host cell processes (reviewed by Rodriguez and Bos 2013; Elzinga and Jander 2013; Rodriguez et al. 2017). Recent progress in aphid transcriptomics and proteomics facilitated the identification of effectors in several important species (Harmel et al. 2008; Bos et al. 2010; Carolan et al. 2011; Atamian et al. 2013; Rao et al. 2013; Boulain et al. 2018), and revealed overlap and diversity between species (Thorpe et al. 2016). Expanding comparative analyses of aphid effectors to the genome level promises to provide new insight into their evolution. For example, in the case of plant parasitic nematodes and filamentous plant pathogens, effectors tend to be located in gene-sparse regions, which are repeat-rich to allow for adaptive evolution (Dong et al. 2015; Eves-van den Akker et al. 2016).

In this study, we sequenced the genomes of *Myzus cerasi* (black-cherry aphid), which is closely related to *M. persicae* but in contrast has a limited host range, and *Rhopalosiphum padi* (bird-cherry oat aphid), which is a cereal specialist. Together with three previously published aphid genomes (*A. pisum, D. noxia* and *M. persicae*), we compare gene content, duplication, putative HGT events, and effector repertoires. Importantly, our gene model (re-)prediction approach revealed that the different aphid genomes have more consistent gene number than previously reported (IAGC 2010; Mathers et al. 2017), between 25,726 and 28,688 genes predicted across the different genomes. A combination of gene duplication, gene birth, as well as putative HGT events has shaped aphid genomes, and contributed to the acquisition of predicted aphid effector genes. Strikingly, we found that expression of a subset of these aphid effector genes is tightly coregulated, reflecting the presence of an unknown transcriptional control mechanism that likely underpins plant parasitism.

Materials and Methods

All data are available under accession numbers PRJEB24287, PRJEB24204, PRJEB24338, and PRJEB24317. Assembled genomes and gene calls are available at http://bipaa.genouest.org/is/aphidbase/ and doi:10.5281/zenodo.1252934. All custom python scripts used to analyze the data using Biopython (Cock et al. 2009) are available on Github and are cited in the text where appropriate.

Aphids Stocks and Material

Aphids were maintained in growth rooms at 18 °C with a 16 h light and 8 h dark period. *Myzus persicae* (JHI_genotype O) was maintained on oil seed rape, a clonal line of *M. cerasi* (JHI1) was maintained on American Land Cress (*Barbarea verna*), and a clonal line of *R. padi* (JHI_JB) was maintained on barley (*Hordeum vulgare* cv. *Optic*).

DNA Extraction and Sequencing

Aphids were collected and subjected to one ethanol wash, with agitation, to help remove fungal and bacterial contamination, followed by three sterile distilled water washes. DNA was extracted using Qiagen Blood Tissue extraction kit following manufacturer's protocol, followed by a DNA ethanol precipitation step to improve DNA purity. DNA quality was assessed using a Nanodrop (Thermo Scientific) prior to sending the Earlham Institute, Norwich, for PCR-free library preparation and sequencing (insert size ~395 bp). Illumina-HiSeq 2X250bp (and 2X150bp for *M. cerasi*) paired-end sequencing was performed.

Filtering, Quality Control, and Genome Assembly

The raw reads were assessed for quality before and after trimming using FastQC (Andrews 2010). For quality control, the raw reads were quality trimmed using Trimmomatic (minimum phred Q15) (Bolger et al. 2014). An iterative process of assembly and contaminant removal was performed. For early iterations of the assembly, CLC (version 4.1.0) was used due to rapid assembly and coverage mapping. To remove contaminant reads, the assembly was compared with the nonredundant database (nt) using BlastN (megablast), and the assembly was also searched against the genome sequence of *A. pisum* to facilitate the identification of Arthropoda contigs, SWISS-Prot database, and GenBank NR using DIAMOND (v0.7.9.58) in sensitive mode

(Buchfink et al. 2015). The DIAMOND-BLAST versus NR data were taxonomically annotated using https://github.com/peterthorpe5/public_scripts/tree/master/Diamond_BLAST_add_taxonomic_info. The resulting taxonomically annotated BLAST results and genomic read coverage generated by CLC (mapper) were used as input to BlobTools (Kumar et al. 2013). Reads that contributed to the assembly of contigs similar to bacteria, fungal, or virus sequences were removed in an iterative approach using Mirabait $K = 99$ (Chevreux 2005). This was repeated eight times for *M. cerasi* and five times for *R. padi*.

The final "cleaned" data sets were converted from .fastq to .bam files using custom python scripts and were assembled using DISCOVAR (Weisenfeld et al. 2014). All assemblies were assessed for "completeness" using Core Eukaryotic Genes Mapping Approach (CEGMA) (Parra et al. 2007) and Benchmarking Universal Single-Copy Orthologs (BUSCO) using Arthropoda hidden Markov models (Simão et al. 2015). Statistics on genome assemblies were generated using https://github.com/sujaikumar/assemblage/blob/master/scaffold_stats.pl. All scripts and commands used for genome assembly are available at https://github.com/peterthorpe5/Methods_M.cerasi_R.padi_genome_assembly

Gene Prediction and Annotation

Due to a lack of publically available known genes from both *M. cerasi* and *R. padi*, the approach of using known sequences to train MAKER (Cantarel et al. 2007) was not used. A preliminary approach was taken. Augustus (Stanke and Waack 2003) gene prediction, using RNAseq hints for each species was performed using the "Pea_aphid species config files" bundled with Augustus (IAGC 2010) (Gene models: v0.9-JHI). RNAseq was mapped to the genomes using splice aware aligner STAR (Dobin et al. 2013) allowing a maximum of seven mismatches, and RNAseq intron hints were generated with bam2hints (a script bundled with Augustus). Additional RNAseq data for each species were obtained from: *R. padi, M. persicae* genotype O, *M. cerasi* (PRJEB24317) and PRJEB9912 (Thorpe et al. 2016), *A. pisum* (PRJNA209321) (IAGC 2010), and *D. noxia* SRR1999270 (Nicholson et al. 2015) (supplementary table S8, Supplementary Material online). Once a set of gene models were predicted, the RNAseq for each species was mapped back to the nucleotide coding sequence gene prediction (exome) of that species using SNAP (Zaharia et al. 2011), to determine the percentage of RNAseq that maps. This did not allow reads that spanned the start and stop codons, and SNAP is a DNA aligner, thus spliced reads resulted in a lower Q mapping score. All mapping was performed in the same way for comparison purposes only. Predicted proteins were DIAMOND-BlastP (Buchfink et al. 2015) searched against NR, and taxonomically annotated as described above. Due to poor RNAseq mapping results, the "Pea_aphid config files" guided gene models were deemed unsatisfactory (see Results). Therefore an alternative, de novo, approach was taken.

The final gene models for all species (*R. padi, M. cerasi, D. noxia, A. pisum*, and *M. persicae* genotype O) were predicted using BRAKER (version 1.8) (Hoff et al. 2015) and intron RNAseq-guided hints (see above) (Gene models: v1.0-JHI). BRAKER uses Genemark-ET (Lomsadze et al. 2014), with the RNAseq hints and Eukaryote hidden Markov models to predict genes and retrain Augustus. Trained Augustus was used in conjunction with RNAseq intron hints to predict gene models v1.0-JHI. Gene models were annotated using Blast2GO version 2.8, database September 2015 (Conesa et al. 2005), Interproscan (Quevillon et al. 2005), PFAM (Finn et al. 2013), DIAMOND-BlastP versus NR (Buchfink et al. 2015). The DIAMOND BLAST output was taxonomically annotated as mentioned above. BLAST output was taxonomically filtered to remove Pea aphid "hit" using https://github.com/peterthorpe5/public_scripts/blob/master/blast_output/top_BLAST_hit_filter_out_tax_id.py

Endosymbiont Genome Assembly

To assemble the *Buchnera* spp. genome from the genomic data, raw reads were trimmed of adapter sequences and low-quality bases (Phred <30), and assembled using SPAdes (version 3.5) using $k = 77,99,127$ (Bankevich et al. 2012). From this assembly, one of the contigs corresponded to the expected genome size of the endosymbiont and shared considerable sequence similarity to other *Buchnera* genomes. The *Buchnera* spp. genomes were annotated with the web-server instance of RAST (Aziz et al. 2008). Assemblies and annotation are available at doi:10.5281/zenodo.1252934.

Transposon-Like Sequence Prediction and Repeat Masking

To predict transposons and repetitive regions, an aphid-specific database was generated using RepeatModeller (version 1.0.8) (Smit and Hubley 2014). The database was classified using Censor (Bao et al. 2015). Repeatmasker (version 4.0.6) (Smit et al. 2014) using this classified database and Repbase was used to identify repetitive regions and transposons. LTRharvest (genometools-1.5.8) (Ellinghaus et al. 2008) and TransposonPSI (version 08222010) (Haas 2007) were also used to identify transposons. A consensus prediction was generated and .gff formatted (https://github.com/HullUni-bioinformatics/TE-search-tools). Transposon and gene distances were calculated using https://github.com/peterthorpe5/public_scripts/tree/master/transposon_analysis.

Alien Index—Detection of HGT Events and Putative Contamination

To detect candidate HGT events, an Alien Index (AI) was calculated as described by Gladyshev et al. (2008) and Flot et al. (2013). All predicted proteins were compared with NR using

DIAMOND-BlastP, with kingdom and tax_id assignment, and an *e*-value threshold of $1e^{-5}$. An AI could only be calculated for a protein returning at least one hit in either a metazoan or a nonmetazoan species, as stated in the following formula: AI = log((best *E*-value for metazoa) + e-200) − log((best *E*-value for nonmetazoa) + e-200).

When neither metazoan nor nonmetazoan BLAST results were identified, the query sequence was removed from downstream analysis. BLAST results in the phylum Arthropoda (which the aphids of interest belong) were ignored for the calculation of AI to allow the detection of putative HGT events that may be shared with other related species. An AI > 30 corresponds to a difference of magnitude e^{10} between the best nonmetazoan and best metazoan *e*-values and is estimated to be indicative of a potential HGT event (Flot et al. 2013). Sequences with an AI> 30 and >70% identity to a nonmetazoan sequence were considered putative contaminants and removed from further analyses (supplementary table S7, Supplementary Material online). HGT prediction tool set is available at Github: https://github.com/peterthorpe5/public_scripts/tree/master/Lateral_gene_transfer_prediction_tool. Intron splice sites were extracted using https://github.com/DRL/GenomeBiology2016_globodera_rostochiensis/tree/master/scripts, and log plots were generated using MEMEsuite (Bailey et al. 2009). Metabolic pathways were predicted using the entire predicted proteome of each species using the KEGG Automatic Annotation Server (Moriya et al. 2007).

Transcriptomic Analyses upon Aphid Exposure to Host and Nonhost Plants and Artificial Diets

To determine the extent transcriptional plasticity contributes to aphid interactions with host and nonhost plants, we sequenced the transcriptome of *R. padi* and *M. persicae* after feeding on an artificial diet for 3 or 24 h, a host plant for 3 or 24 h, and a nonhost plant for 3 or 24 h. For *R. padi*, barley is considered a host and *Arabidopsis* is considered a nonhost (Jaouannet et al. 2015). For *M. persicae*, *Arabidopsis* is considered a host and barley is considered a nonhost (Escudero-Martinez et al. 2017). For both species, the artificial diet consisted of 15% sucrose, 100 mM L-serine, 100 mM L-methionine, and 100 mM L-aspartic acid with a pH of 7.2 (KOH) (Will et al. 2012).

Barley plants (cv Optic) were pregerminated in Petri dishes with wet filter paper for 3 days in the dark. Plants were moved to a growth room and grown for 7 days prior to aphid infestation. *Arabidopsis* plants were sown directly in soil and grown for 5 weeks prior to aphid infestation. Artificial diets were prepared and placed between Parafilm sheets according to Thorpe et al. (2016). Plant growth as well as aphid exposure to plant and diet were carried out under 8 h of light (125 µmol photons/m^2.s), at 22 °C and 70% humidity.

For transfer of *R. padi* and *M. persicae* aphids from stock plants to barley and *Arabidopsis*, 15 mixed-aged apterous aphids were enclosed in a single clip cage, with one clip cage per plant, and six plants per plant–aphid combination per time point (3 and 24 h). The clip cage was placed in the middle of the first leaf for barley, and it was covered 1–2 fully expanded leaves for *Arabidopsis*. For the artificial diet treatment, 100 mixed-aged apterous aphids were used per time point in a single artificial diet container. One single batch of artificial diet was prepared and stored in aliquots at −20 °C, and thawed aliquots were used for the different biological replicates. All aphids were collected 3 and 24 h after exposure to plants or diet and flash frozen in liquid nitrogen, and aphids from the six individual plants per plant–aphid combination per time point were pooled into one single tube. In total, five independent biological replicates were performed of the whole experiment. Individual replicates were set up at the same time of day to avoid variability due to the aphid or plant circadian cycle. Replicates of host and nonhost plant treatments were set in different weeks over a 2-month period at ~9 AM, with the 3 h time point collected at 12 noon the same day, and the 24 h time point at ~9 AM the next day. Artificial diet treatments were not set up in parallel to the plant treatments, but on consecutive days, between 10 AM and 12 PM, with collection of the 3 h time point occurring between 1 and 1.30 PM the same day, and collection of the 24 h time point between 11 AM and 12 PM the next day.

RNA was extracted from 70 to 90 aphids with the Qiagen RNeasy Plant Mini Kit following the manufacturer's protocol. RNA quality was assessed using agarose gel electrophoresis and the Agilent 2100 Bioanalyzer. Approximately, 2.5 µg of total RNA per sample (60 samples total) was submitted to TGAC (The Genome Analysis Centre, Norwich Research Park) for Illumina TrueSeq library preparation and sequencing (100-bp paired end).

Temporal RNAseq data described above were analyzed with spatial RNAseq data of a previous study (PRJEB9912; Thorpe et al. 2016). All raw RNAseq reads were assessed using FastQC (Andrews 2010), and low-quality bases were removed using Trimmomatic (Minimum Phred score 22) (Bolger et al. 2014). Reads were mapped to the corresponding genome using STAR version2.5.1b (Dobin et al. 2013). The resulting bam file was assembled using Trinity (version 2.1.1) (Haas et al. 2013). The assembly was subjected to quality control using Transrate (Smith-Unna et al. 2016). Transcript abundance was quantified using Kallisto (Bray et al. 2016). Differential expression analysis was conducted using EdgeR (Robinson et al. 2010), using minimum threshold of Log2-fold change and a false discovery rate (FDR) $P < 0.001$. Coding sequence from transcripts was predicted using TransDecoder (Haas et al. 2013).

Effector Identification and Comparisons

To compare aphid effector repertoires across the five different species, we (re)-predicted effector loci contained within the

v1.0-JHI annotations based on three modes of evidence, and as described previously (Thorpe et al. 2016). In brief, we predicted effectors based on 1) upregulation in aphid head tissues (containing salivary glands) compared with aphid bodies without heads in combination with the presence of signal peptide coding sequences (data set described by Thorpe et al. 2016), 2) presence in aphid saliva as determined by proteomics (data set described by Thorpe et al. [2016]), and 3) similarity to previously described putative effectors (Bos et al. 2010; Carolan et al. 2011; Elzinga et al. 2014). Aphid genes with at least one mode of evidence were considered putative effector loci (supplementary table S4, Supplementary Material online). These approaches were not applied to *D. noxia* due to the lack of tissue-specific gene expression and saliva proteomics data. The effector repertoire network of all species was generated by calculating the BlastP bit score of pairwise comparisons between effectors of all species. An array of pairwise bit scores was parsed to gefx format using a custom python script (https://github.com/sebastianevda/SEvdA_Gephi_array_to_gefx) and visualized using Gephi (Bastian et al. 2009).

Promoter Analyses

The genomic 5′ region to genes of interest was obtained using custom python script (https://github.com/peterthorpe5/public_scripts/tree/master/genomic_upstream_regions). Motif enrichment was performed using the differential motif discovery algorithm HOMER (Heinz et al. 2010).

Comparative Genomics

An MCL all-versus-all network was generated using the predicted proteomes of *R. padi, D. noxia, A. pisum, M. persicae, M. cerasi*, and the outgroup model insect *Drosophila melanogaster.* Similarity was assessed using DIAMOND-BlastP (1e-31) and clustered using MCL (inflation value of 6). All MCL analyses were performed using Biolinux 7 (Field et al. 2006). Individual sequence alignments were carried out using Muscle v3.8.31 (Edgar 2004) and visualized using the BoxShade web server (https://www.ch.embnet.org/software/BOX_form.html).

Gene Duplication and Synteny Analyses

Gene duplication and synteny analysis was performed using the similarity searches from DIAMOND-BlastP (*e*-value 1e-5) with MCSanX toolkit (Wang et al. 2012). Synteny between scaffolds was visualized using Circos 0.67-7 (Krzywinski et al. 2009).

Phylogenetic Inference

Single-copy orthologous genes were identified using in all five aphid species studies, and the outgroup *D. melanogaster*. Only those sequences identified in all genome assemblies, and classified as single copy loci in all assemblies, were studied ($n = 386$). For a given BUSCO gene in a given species, if the gene length deviated by more than 5% from the average for that BUSCO gene in all other species, that BUSCO gene was not analyzed further for any species (remaining $n = 123$). The amino acid sequences of the remaining 123 highly conserved BUSCO genes were aligned and refined using MUSCLE (supplementary file 1, Supplementary Material online). Individual BUSCO alignments were concatenated and a partition file generated using a custom python script (https://github.com/sebastianevda/SEvdA_Gephi_array_to_gefx/blob/master/cat_alignments_rename_names_write_partition_file.py). Model selection for each partition, and phylogenetic inference, was carried out using the IQ-TREE webserver to generate a consensus tree of 1,000 bootstraps (Trifinopoulos et al. 2016).

DN/DS Analysis

A 1:1 Reciprocal Best BLAST Hit network was generated from the predicted amino acid sequences, using a minimum threshold of 70% identity and 50% query coverage (Cock et al. 2015) and clustered using MCL (version 12-135) (Enright et al. 2002) with an inflation value of 6. The number of species contained in a cluster was obtained using mcl_to_cafe.py (De Bie et al. 2006). DN/DS values for each cluster that contained a predicted effector were calculated. Within each cluster, deduced protein sequences were aligned using MUSCLE (version 3.8.31) (Edgar 2004), and the nucleotide sequences were back-translated onto the alignments (https://github.com/peterjc/pico_galaxy/tree/master/tools/align_back_trans) (Cock et al. 2009). Alignments were manually curated using Jalview (Waterhouse et al. 2009) by removing nonconsensus, possibly miss-predicted 5′ and 3′ regions. Modified alignments were subjected to DN/NS analysis using CodonPhyml (version 1.0) (Gil et al. 2013).

Results and Discussion

We sequenced the genomes of a clonal line of *M. cerasi* established on secondary host species *Barbarea verna* (Land Cress) and of a clonal line of *R. padi* established on *Hordeum vulgare* (Barley) using Illumina 2X250bp pair-end libraries (and 2X150 bp for *M. cerasi*) to a depth of 233× and 129×, respectively. Using these data, the genome of *M. cerasi* was assembled to 406 Mb contained in 49,349 contigs and the *R. padi* genome assembled to 319 Mb contained in 15,616 contigs. These assemblies are of a similar size to those reported for *Diuraphis noxia* (393 Mb; Nicholson et al. 2015), *Myzus persicae* genotype O (347/356 Mb; Mathers et al. 2017), and *Acyrthosiphum pisum* (533 Mb; IAGC 2010) (table 1). When compared with previously published assemblies of aphid genomes, the *M. cerasi* and *R. padi* genomes have good continuity [contig N50 of 19,701 and 98,943 bp, respectively] (table 1). The endosymbiont

Table 1
Genome Statistics

	Acyrthosiphon pisum	*Diuraphis noxia*	*Myzus cerasi*	*Myzus persicae*	*Rhopalosiphum padi*
Assembly size (Mb)	533	395	406	356	319
Scaffolds (*n*)	12,969	5,614	49,349	13,509	15,616
Scaffold N50 (bp)	530,744	397,774	23,265	164,460	116,185
Longest scaffold (bp)	3,073,041	2,142,037	265,361	1,018,155	616,405
Contig N50 (bp)	29,034	13,141	19,701	59,031	98,943
Longest contig (bp)	424,120	147,337	209,856	421,714	570,536
N (bp)	41,784,240	98,534,451	194,118	11,542,805	54,488
GC (%)	29.8	29.1	29.9	30.2	27.8
CEGMA: $N = 248$ (complete/partial)	92%/98%	86%/94%	86%/96%	94%/100%	93%/97%
BUSCO: $N = 2,675$ (complete, duplicated, fragmented, missing)	83%, 10%, 7.8%, 8.3%	76%, 6.3%, 11%, 11%	80%, 8.1%, 10%, 9.3%	84%, 9.3%, 7.4%, 8.1%	82%, 8.1%, 7.8%, 9.4%
Transposable elements: % of genome/number/avg. len.	31%/313,339/510 bp	11%/123,792/350 bp	7%/61,812/470 bp	14%/137,377/376 bp	12%/113,457/342 bp
Genes per Mb	51, 9	65, 8	70, 7	72, 3	82, 4
Genes (*n*)	27,676	25,987	28,688	25,726	26,286
BlastP hit in NR (1e-5)	25,313 (91%)	21,818 (84%)	21,576 (75%)	19,816 (77%)	20,368 (77%)

genomes (*Buchnera aphidicola*) of *M. cerasi* and *R. padi* were assembled as single contigs of 641,811 and 643,950 bp, respectively. BUSCO (Simão et al. 2015) and CEGMA (Parra et al. 2007) were used to estimate an assembly completeness of 80% and 86%, respectively, for *M. cerasi*, and 82% and 93%, respectively, for *R. padi* nuclear genomes (table 1). The GC content of the *M. cerasi* and *R. padi* genomes (29.9% and 27.8%, respectively) is consistent with other aphid genomes (IAGC 2010; Mathers et al. 2017; Wenger et al. 2017), and they contain a high proportion of repeat rich and/or transposon-like sequence (table 1). Altogether, these data indicate that, especially in the case of *R. padi*, high-quality draft genome assemblies were generated. With a number of aphid genomes available, we are able to perform detailed comparative analyses to understand the evolution of aphid parasitism genes.

Gene Model Prediction and Reprediction Indicate that Aphid Genomes Encode Similar Gene Numbers and, in the Case of *A. pisum*, Fewer and Larger Genes than Previously Reported

To annotate the genome assemblies generated, we initially used the configuration files for the related aphid *A. pisum* (bundled with AUGUSTUS; IAGC 2010) guided by two sets of evidence: 1) The 36,939 *A. pisum* gene models (IAGC 2010) and 2) extensive species-specific RNAseq data (described later). Using this approach, we predicted 35,316 genes for *M. cerasi* (Mc_v0.9-JHI), comparable to the 36,939 genes predicted for *A. pisum* (IAGC 2010; supplementary table S1, Supplementary Material online). However, these gene models describe a minority of the expressed genes, with only 29% of the RNAseq read pairs mapped to the predicted Mc_v0.9-JHI exome (exome here is defined as the predicted coding genes in the genome). To address this, a subsequent RNAseq-guided de novo approach was adopted, generating 28,688 loci for *M. cerasi* (Mc_v1.0-JHI). When compared with v0.9-JHI gene models, the de novo v1.0-JHI gene models are longer (means of 772 vs. 952, respectively), encode almost exactly the same total exome size (27,332,397 vs. 27,278,139 nt), contain approximately 35% fewer genes without RNAseq support (8,301 vs. 5,494), and describe more than twice as much of the total RNAseq reads (29% vs. 60%, fig. 1*A*). Comparing the two sets of gene models with one another revealed a markedly different size distribution (fig. 1*B*). Version 0.9-JHI encodes ~8,000 more very short gene models in the size range 0–300 bp than v1.0-JHI and contains 10,411 "unique" loci with no overlap in genomic coordinates with any locus in v1.0-JHI (~29%). The loci unique to v0.9-JHI contribute a larger proportion of the small 0–300 bp gene models than any other gene size category. In contrast to this, the loci unique to v1.0-JHI are evenly distributed across individual size categories and each category is similar to the total proportion in v1.0-JHI (fig. 1*B*).

Taken together, our results suggest that using gene models of other aphid species to facilitate the annotation of new genomes produces a similar number of loci. However, the majority of these loci are not supported by RNAseq data (even though the RNAseq data were used to facilitate prediction). To avoid propagating errors, we annotated the genome of *R. padi*, and reannotated all other available aphid genomes, using the RNAseq-guided de novo approach described above (supplementary table S1, Supplementary Material online). Despite being entirely independent, de novo annotation

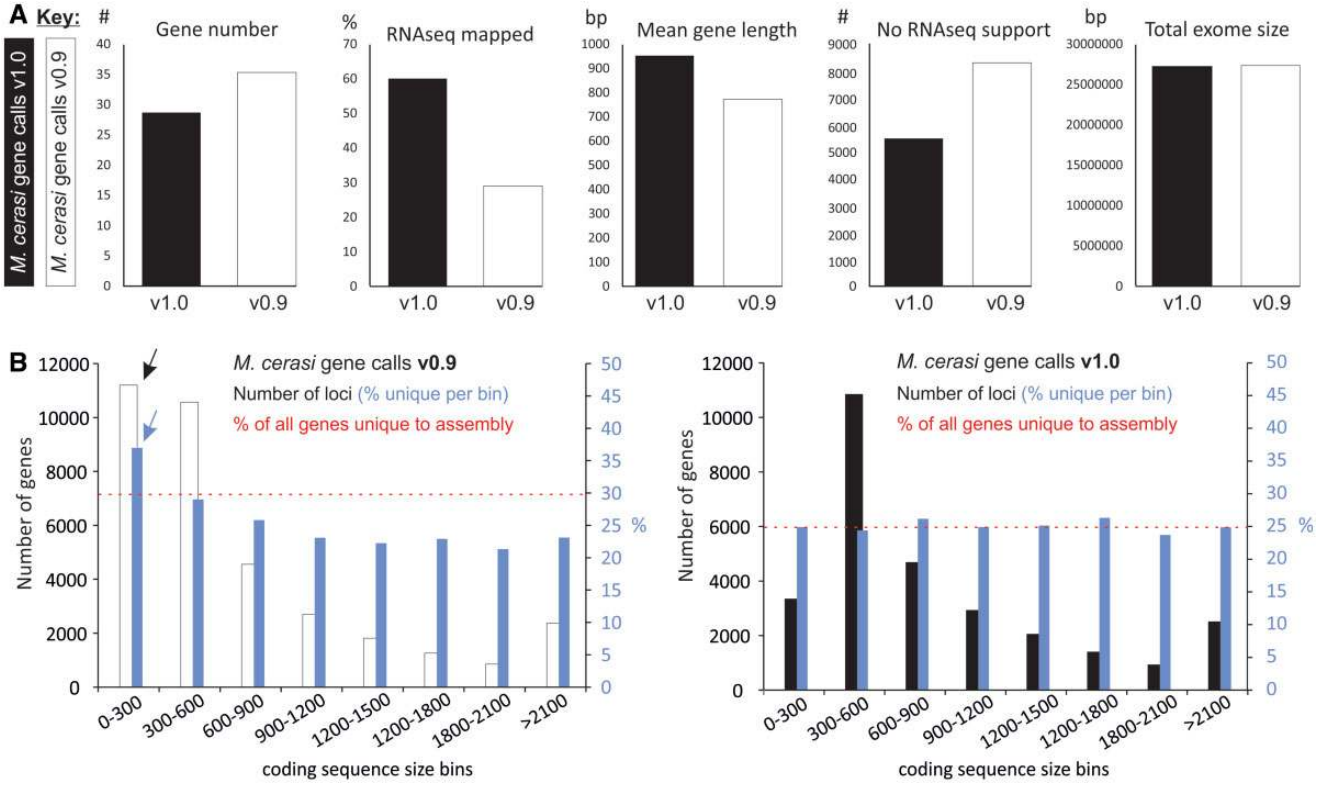

Fig. 1.—Comparison of *M. cerasi* gene models v0.9-JHI and v1.0-JHI. An initial homology- and RNAseq-guided gene model prediction (v0.9-JHI, white bars) compared with a subsequent RNAseq-guided de novo approach using BRAKER (v1.0-JHI, black bars). (*A*) v1.0-JHI predictions contained fewer loci, improved mapping of RNAseq reads, were longer on average (mean), had fewer loci with no RNAseq support, and yet had an almost identical total exome size to v0.9-JHI. (*B*) Markedly different frequency distribution of coding sequence length of v1.0-JHI predictions (black) compared with v0.9-JHI predictions (white): v0.9-JHI contains ~8,000 very short gene models in the size range 0–300 bp (black arrow). Genes predicted in v1.0-JHI with no corresponding prediction in v0.9-JHI (blue) are evenly distributed across coding sequence size bins. Genes predicted in v0.9-JHI with no corresponding prediction in v1.0-JHI are preferentially contained within the 0–300 bp coding sequence size bin (blue arrow).

produced remarkably consistent gene counts for all aphid species (between 25,726 and 28,688), ~25% less loci overall, and a more complete representation of their individual transcriptomes (supplementary table S1, Supplementary Material online). Importantly, our approach reduced the possibility that direct comparison of gene content between species is confounded by an inherent bias in different gene prediction methods. Gene models for all species have been made publically available via AphidBase (http://bipaa.genouest.org/is/aphidbase/) and doi:10.5281/zenodo.1252934. For the remainder of the article, all comparisons are between the genomes and re/predicted gene content of *M. cerasi, M. persicae* (genotype O)*, A. pisum, D. noxia*, and *R. padi*.

Aphid Genomes Show Signs of Extensive Gene Duplication and Recent Gene Birth

Gene content of aphid genomes is extensively duplicated and ranges from around 55% multicopy loci in *M. persicae* to nearly 70% in *A. pisum* (fig. 2*A*, supplementary table S2, Supplementary Material online). Although most duplicated loci were classed as "dispersed" rather than "tandem" or "segmental," we appreciate actual values may deviate from those reported here due to the limit of current assembly contiguity. Gene duplication was previously described in aphids, mostly in *A. pisum*, including for genes encoding amino acid transporters (Price et al. 2011; Duncan et al. 2016), Cytochrome P450 (Puinean et al. 2010), chemosensory receptors (Smadja et al. 2009), and carotenoid biosynthesis genes (Nováková and Moran 2012).

To explore the origins of this gene duplication, a robust phylogenetic framework was generated using a multigene phylogeny of 123 highly conserved BUSCO genes present as single-copy loci in all aphid genomes tested, and the distant outgroup *D. melanogaster* (fig. 2*A*). The entire predicted proteomes of all species were clustered based on sequence similarity using MCL, and cross-referenced with the phylogenetic, and gene duplication analyses. This revealed that for genes present in clusters with at least one representative from all aphid species but excluding *D. melanogaster* (hereafter referred to as aphid-specific) 54% are duplicated (52% dispersed, 1% tandem, and 1% proximal), whereas 46% are

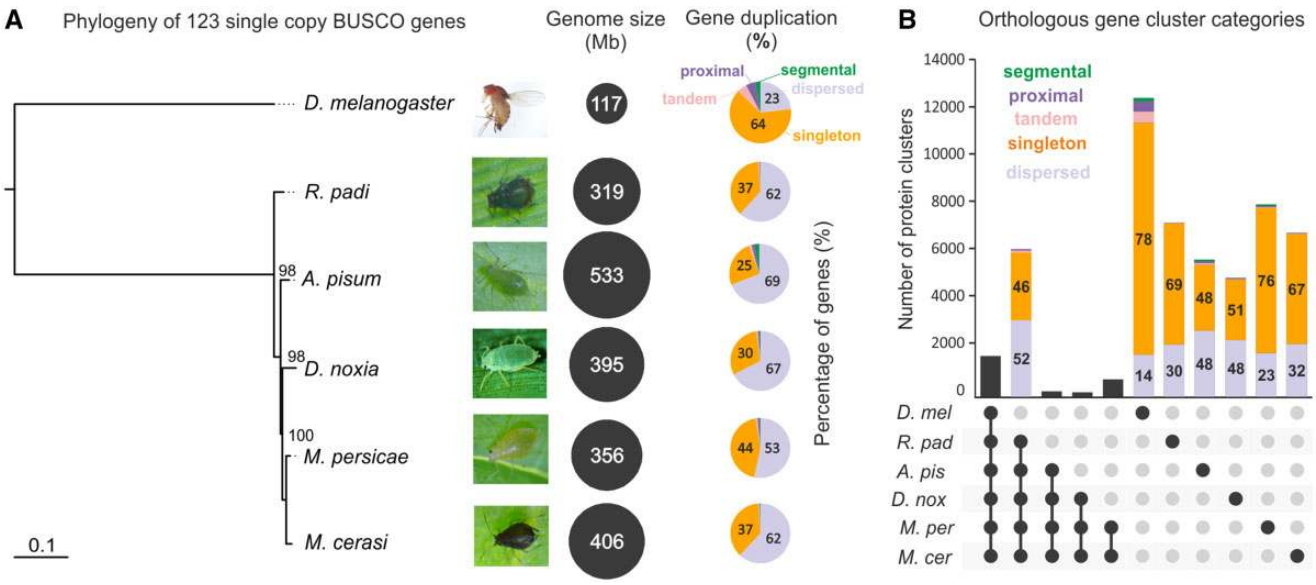

Fig. 2.—An overview of aphid genomes and gene content. (*A*) A multigene phylogeny derived from a concatenated alignment of 123 highly conserved BUSCO nuclear genes classified as single copy in five aphid species (*R. padi, D. noxia, A. pisum, M. persicae*, and *M. cerasi*) and the outgroup model insect *D. melanogaster*. Node values indicate boot strap support of 1,000 iterations. For each species, black circles are scaled by genome assembly size, and pie charts show the proportion of the genes that belong to various duplication categories (singleton, dispersed, segmental, proximal, and tandem, see supplementary table S2, Supplementary Material online, for all values). (*B*) The predicted protein sets of the five aphids were compared with that of the model insect *D. melanogaster*. The histogram shows the number of orthologous gene clusters shared uniquely between the species highlighted below. Dark dots indicate the species contained within each set (e.g., the first column contains orthologous gene clusters with representatives from all species, the second all aphid species but not containing any representative from *D. melanogaster*, etc.). Selected histograms are divided by the proportion of gene duplication categories, where internal numbers refer to the percentage of that category in that cluster. A total of 6,121 clusters contain at least one sequence from each aphid but do not contain any sequence from *D. melanogaster*. Of the genes within these clusters, 54% are duplicated (52% dispersed, 1% tandem, and 1% proximal) while 46% are singletons.

singletons (fig. 2*B*). These most likely reflect duplication events that occurred before speciation, followed by retention of multiple copies to present date.

In stark contrast, genes present in clusters that exclude all other species (hereafter referred to as species-specific) are often dominated by single-copy loci, with 48% and 51% of genes present as singletons in *A. pisum* and *D. noxia*, respectively, and 67–76% in the remaining three aphid species (fig. 2*B*). Given the relatedness of these aphid species (indicated by short branch lengths in fig. 2*A*), this observation most likely reflects large scale and relatively recent gene birth in most aphid species, after speciation. Such lineage-specific gene birth and death has also been reported in other insect species and likely is driven by unrelated traits (Hahn et al. 2007; Heger and Ponting 2007). Taken together, it is likely that a combination of extensive gene multiplication and relatively recent gene birth has shaped the evolution of aphid genomes.

With aphids having rather complex lifecycles, many traits could be driving these features of aphid genomes. However, juxtaposed to the recent discovery that duplicated genes play a role in parasitism of the broad host range *M. persicae* (Mathers et al. 2017), the implication of large-scale species-specific gene birth in the context of broad and narrow host-range aphids is intriguing.

Disparate Gain and Loss of Loci Putatively Acquired via HGT

To determine whether HGT events have contributed to the unusual distribution of gene cluster categories, a systematic genome-wide putative HGT-identification approach was employed. Putative HGT events were predicted by their ratio of sequence similarity to metazoan and nonmetazoan sequences (termed the AI; Gladyshev et al. 2008; Flot et al. 2013; Rancurel et al. 2017). Using a conservative approach, predicted proteins with an AI >30 and <70% identity to nonmetazoan sequences were classed as putative HGT, while those with more than 70% identity to nonmetazoan sequences were classed as putative contaminants, and not further interrogated. As a consequence of the efforts to avoid classifying contaminants as putative HGT, recent HGT events would be excluded from the analysis.

Using these criteria, we provide an estimate that ~1–2% of aphid loci may be of nonmetazoan origin (between 212 [*M. persicae*] and 338 [*D. noxia*], fig. 3*A* and supplementary

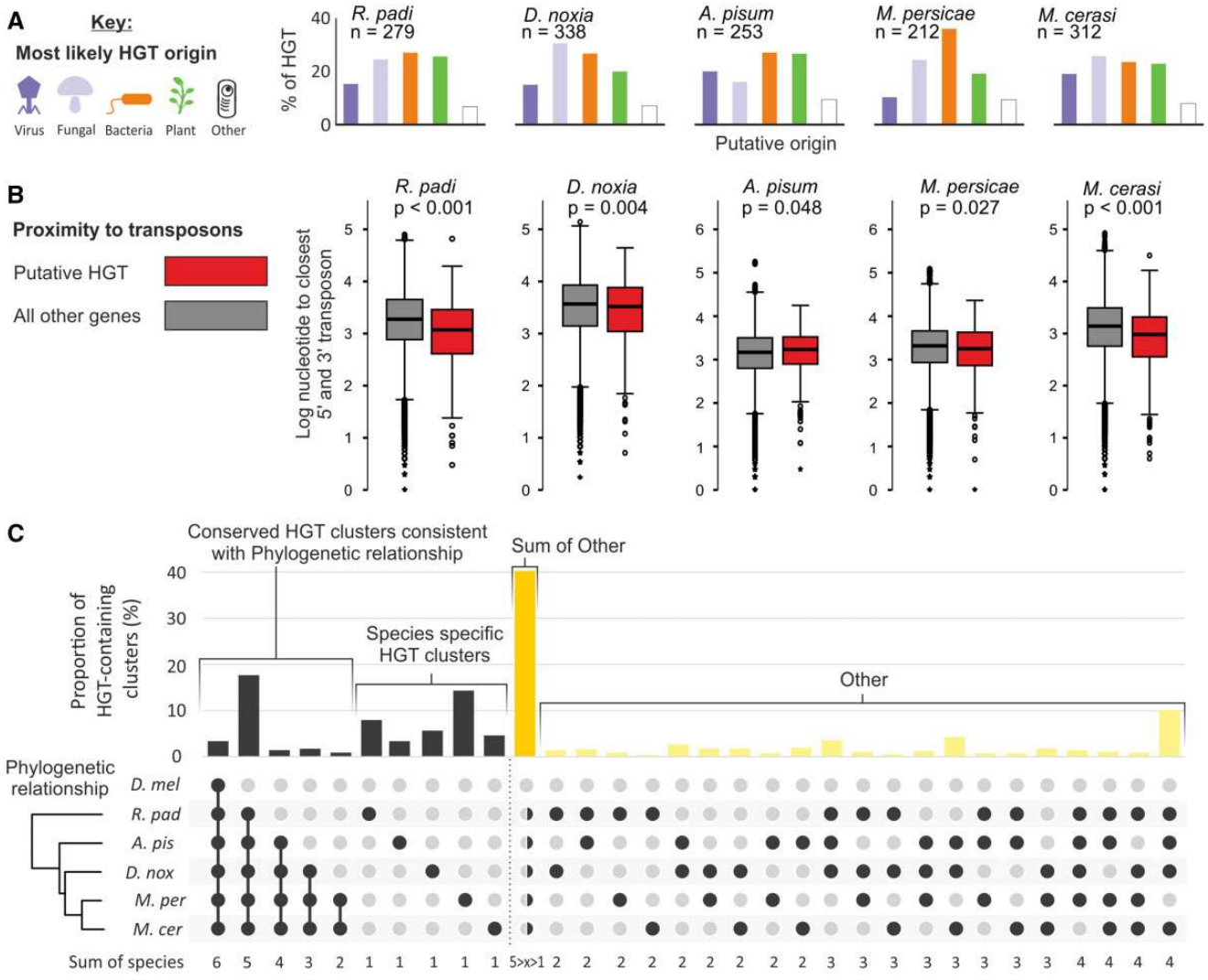

Fig. 3.—Putative HGT: origins and acquisition. We deployed a systematic genome-wide approach to identify putative HGT events from nonmetazoans, based on AI calculations. (*A*) The number of putative HGT events varies from 212 (*M. persicae*) to 338 (*D. noxia*). Histograms show the number of putative HGT events of viral (dark purple), fungal (light purple), bacterial (orange), plant (green), or nonmetazoan Eukaryotes (white) for each aphid species. (*B*) With the exception of *A. pisum*, putative HGT events (red) in aphid genomes are typically closer to their neighboring 5′ and 3′ transposons than all other genes in the genome (gray). Mann–Whitney *U* test *P*-values range from 0.027 to <0.001. (*C*) The histogram shows the proportion of putative HGT-containing clusters shared uniquely between the aphid species highlighted below. Dark dots indicate the species contained within each set (e.g., the first contains all species, the second all species except *D. melanogaster*, etc.). Approximately 40% of putative HGT-containing clusters are not consistent with the phylogenetic relationships of the different aphid species (dark yellow), but neither do they predominantly support any one other alternative (light yellow).

table S3, Supplementary Material online). While it is a relatively modest contribution, this estimate expands upon previous reports in both absolute number and donor taxa (Moran and Jarvik 2010; Nikoh et al. 2010), and so further experimentation would be required to confirm these predictions.

Putative HGT events were detected from diverse donor taxa (including plantae, fungi, bacteria, viruses, and other nonmetazoan eukaryotes), but were primarily similar to sequences from the fungal and bacterial kingdoms (fig. 3*A*). This approach reidentified previously characterized cases of HGT, the carotenoid biosynthesis genes (Moran and Jarvik 2010), although the complete pathway is apparently missing in some of the aphid genomes studied (supplementary fig. S1, Supplementary Material online). Intriguingly, predicted HGT events are generally closer to their 5′ and 3′ transposable elements when compared with the remainder of the genes in aphid genomes (Mann–Whitney *U* test, *P*-values range between 0.001 and 0.048; fig 3*B*). This observation was similarly described for putative HGT events predicted in a plant-parasitic nematode using the same methods (reviewed by

Kikuchi et al. [2017]), and perhaps is indicative of a general characteristic of HGT acquisition or prediction by this method. Less than 20% of predicted HGT loci are present in aphid-specific gene clusters (fig. 3C). This minority of putative HGT events likely originates before speciation and has been conserved to present day. Similarly, <15% of predicted HGT events are specific to a single aphid species, and likely originate after speciation events (fig. 3C). Remarkably, 40% of putative HGT-containing clusters are not consistent with the phylogenetic positions of the different aphid species, but neither do they predominantly support any one other alternative (fig. 3C). Based on these observations we propose that most of these putative HGT events may have complex evolutionary histories characterized by disparate gain and perhaps unsurprisingly frequent loss, and that HGT does not explain the large scale and recent gene birth observed in the aphid genomes.

Following transfer, predicted HGT events are apparently "normalized" to the host genome (Lawrence and Ochman 1997). The AT content of putative HGT events predicted herein is largely consistent with the remainder of the genome (fig. 4B). In contrast, the majority of putative HGT events have on average ~1 less intron per gene compared with the remainder of the genome (Mann–Whitney *U* test, $P < 0.000$ for all aphids tested). Although the corresponding 5′ donor and 3′ acceptor splice sites are indistinguishable from the remainder of the genes in the aphid genomes, and are largely consistent with canonical CAG:GTAAGT (exon:intron) splicing (fig. 4C). The predicted *D. melanogaster* splice sites are in line with previous splice site predictions (Korf 2004; Lomsadze et al. 2005). Finally, the vast majority of putative HGT events have evidence of transcription; however, the proportion of putative HGT events that have no measurable RNAseq expression is slightly higher than the remainder of the genome, with the notable exception of *A. pisum* (supplementary fig. S2. Supplementary Material online). Taken together, this set of otherwise sequence-unrelated putative HGT events has some characteristics consistent with the remainder of the genome, and others that are inconsistent. Nevertheless, by comparing HGT predictions with a proteomics data set that identified proteins present in saliva secretions (Thorpe et al. 2016), we are able to detect evidence of translation for a number of genes in these sets in *M. cerasi* ($n = 11$) and *M. persicae* ($n = 3$). Further experimental evidence would be required to confirm whether any of these candidates represented bonafide HGT genes, as it is likely our list of putative HGT events contains false positives.

The Unusual Genomic Organization and Evolutionary History of Predicted Aphid Effector Repertoires

To compare aphid effector repertoires across four different aphid species, we (re)-predicted effector loci contained within the v1.0-JHI annotations based on three modes of evidence, and as described previously (Thorpe et al. 2016). In brief, we predicted effectors based on 1) upregulation in aphid head tissues (containing salivary glands) compared with aphid bodies without nymphs/heads in combination with the presence of signal peptide coding sequences (data set described by Thorpe et al. 2016), 2) presence in aphid saliva as determined by proteomics (data set described by Thorpe et al. 2016), and 3) similarity to previously described putative effectors (Bos et al. 2010; Carolan et al. 2011; Elzinga et al. 2014). Aphid genes with at least one mode of evidence were considered putative effector loci (supplementary table S4, Supplementary Material online). It is likely that selection based on these criteria does not cover the full effector complements and also will lead to some false positives due to (technical) limitations previously described (Thorpe et al. 2016). Our approach identified 484 putative effectors for *R. padi*, 225 for *M. cerasi*, 240 for *M. persicae*, and 226 for *A. pisum* (supplementary table S4, Supplementary Material online). The differences in numbers of predicted effectors are most likely due to differences in quality of the genome assemblies, as well as the transcriptomics and proteomics data sets. These data provide a platform for follow-up functional validation to validate effector activity. Effector predictions were not applied to *D. noxia* due to the lack of tissue-specific gene expression and saliva proteomics data.

Clustering only the putative effector gene content between aphid species (fig. 5A) revealed a different pattern to clustering the entire proteomes (fig. 2). Specifically, the pan-genome of putative aphid effector repertoires is dominated by singletons, with few highly connected clusters (fig. 5A, singletons represented by single not-connected dots). However, when looking at gene duplication of predicted effectors we noted that most were classified as multicopy loci, with 53–73% of genes within the dispersed category across the aphid species (supplementary fig. S3, Supplementary Material online). Therefore, while many effectors are part of paralogous/homologous gene families within a species, as evidenced by the gene duplication data, often only one member of this family is predicted to be an effector. This can be a hallmark of either neofunctionalization following gene duplication (Lilley et al. 2018) or loss of an effector gene following recognition by the plant immune system, and will need to be further explored. A role of gene duplication in shaping the *A. pisum* effector repertoire was also recently reported by Boulain et al. (2018) who found more duplicated genes among predicted effector sets than expected, as well as evidence of positive selection on several duplicated predicted effector genes. Interestingly, in *R. padi*, *M. cerasi* and *M. persicae*, approximately 2.4%, 2.7% and 3.6% of putative effectors have a predicted AI >30: Approximately two and a half times the relative contribution to the remainder of the predicted proteome (supplementary fig. S4, Supplementary Material online).

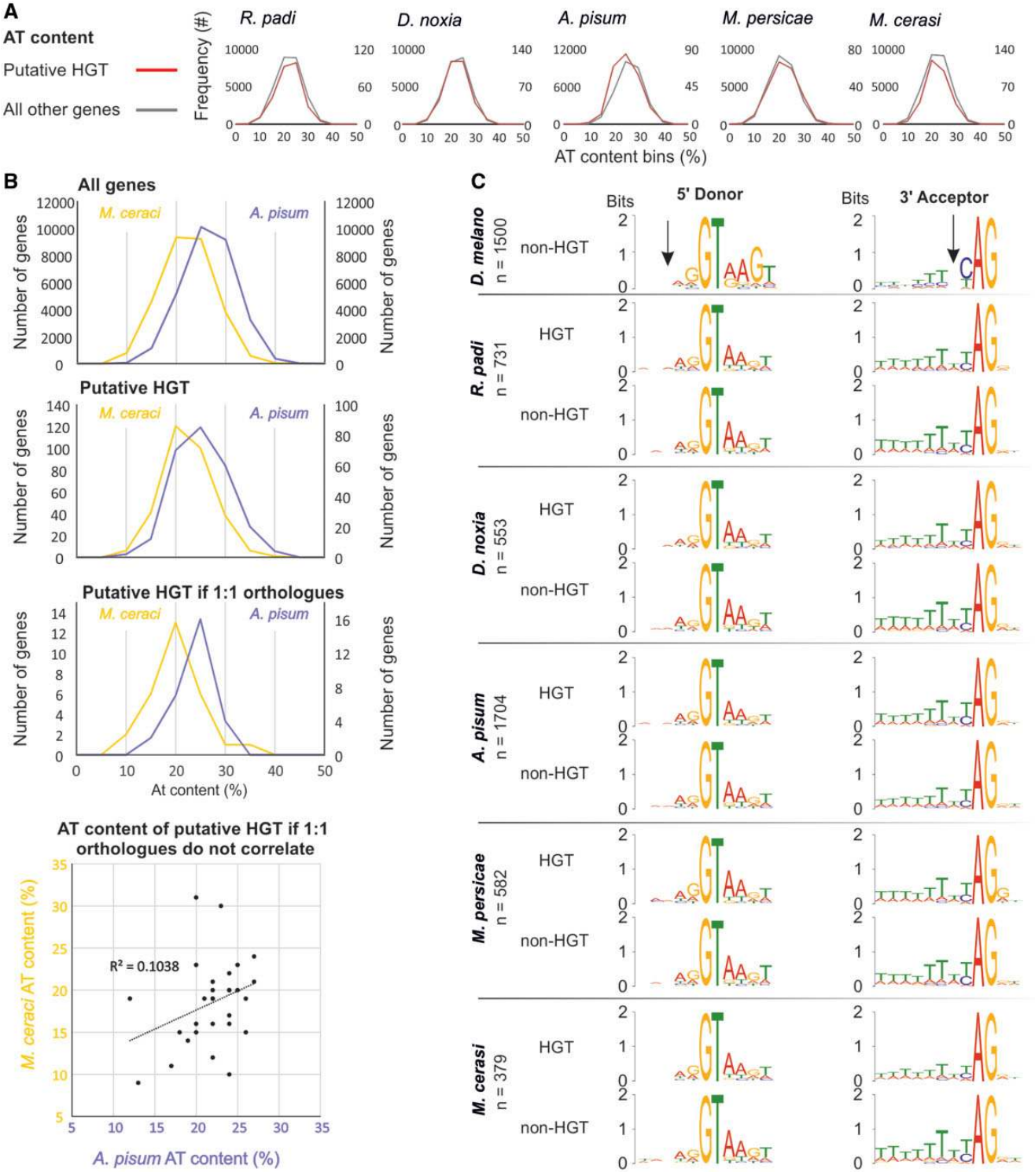

Fig. 4.—"Normalization" following HGT. (*A*) The frequency distribution plots of AT content for putative HGT events (red) compared with T content of all other genes (gray) for *R. padi, D. noxia, A. pisum, M. persicae*, and *M. cerasi*. (*B*) Comparison of AT content frequency distributions between *M. cerasi* (yellow) and *A. pisum* (purple) for all genes, putative HGT events, and putative HGT events that are putative 1:1 orthologs. (*C*) Base composition of 5′ donor and 3′ acceptor splice sites for a random selection of 1,500 *D. melanogaster* genes is compared with that of all putative HGT events, and a randomly selected equal number of non-HGT events, from *R. padi, D. noxia, A. pisum, M. persicae*, and *M. cerasi*. Black arrows indicate a consistent deviation from canonical splice sites.

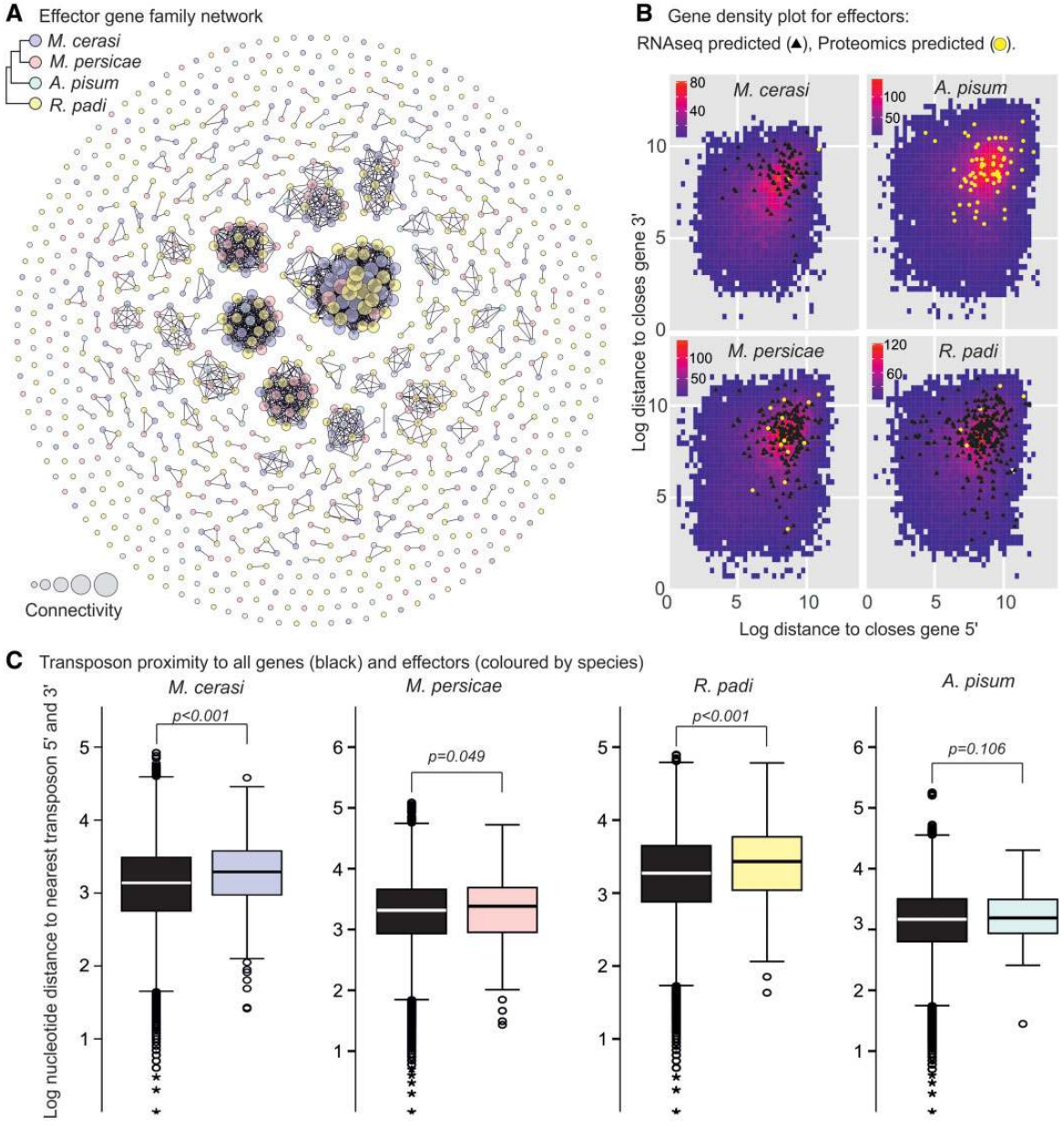

Fig. 5.—Effector repertoire and genomic organization. (*A*) Putative effector loci from all species were clustered with one another using BLAST. A network of sequence similarity was produced, where each node represents an individual effector locus. A schematic representation of the relationship between species is inset (top left), colored by species. Connections between nodes are made if the similarity between sequences has a minimum bit score of 91. Node size is scaled by connectivity as shown below the network. (*B*) The log nucleotide distance of each gene to its neighbor, 5′ (*x* axis) and 3′ (*y* axis) direction, colored by density from blue to red. Putative effectors are highlighted, colored by prediction method (RNAseq predicted—black triangle, proteomics predicted—yellow circle). Salivary proteomics was conducted for *M. cerasi*, all but two of the corresponding genes (represented by two yellow circles) were located at the ends of contigs, and therefore were excluded *M. cerasi* here. Head and body tissue RNAseq was not conducted for *A. pisum* and so it does not have black triangles, only yellow circles. *D. noxia* was not included in this analysis due to a lack of available data. (*C*) Box and whisker plots show the distance to nearest 5′ and 3′ transposable elements for putative effectors (colored by species), and all other noneffectors (black). Distributions were compared using the Mann–Whitney *U* test.

Putative effectors are not randomly distributed across the aphid genomes, but are apparently partitioned into less gene-dense subdomains (fig. 5*B*). Compared with all other genes in the aphid genomes, putative effectors are significantly further from their neighboring genes in both the 3′ and 5′ directions (Mann–Whitney *U* test, $P < 0.000$, fig. 5*B*). Similarly, effectors from distinct eukaryotic plant pathogens are often located in less-gene dense regions within the genomes (Haas et al.

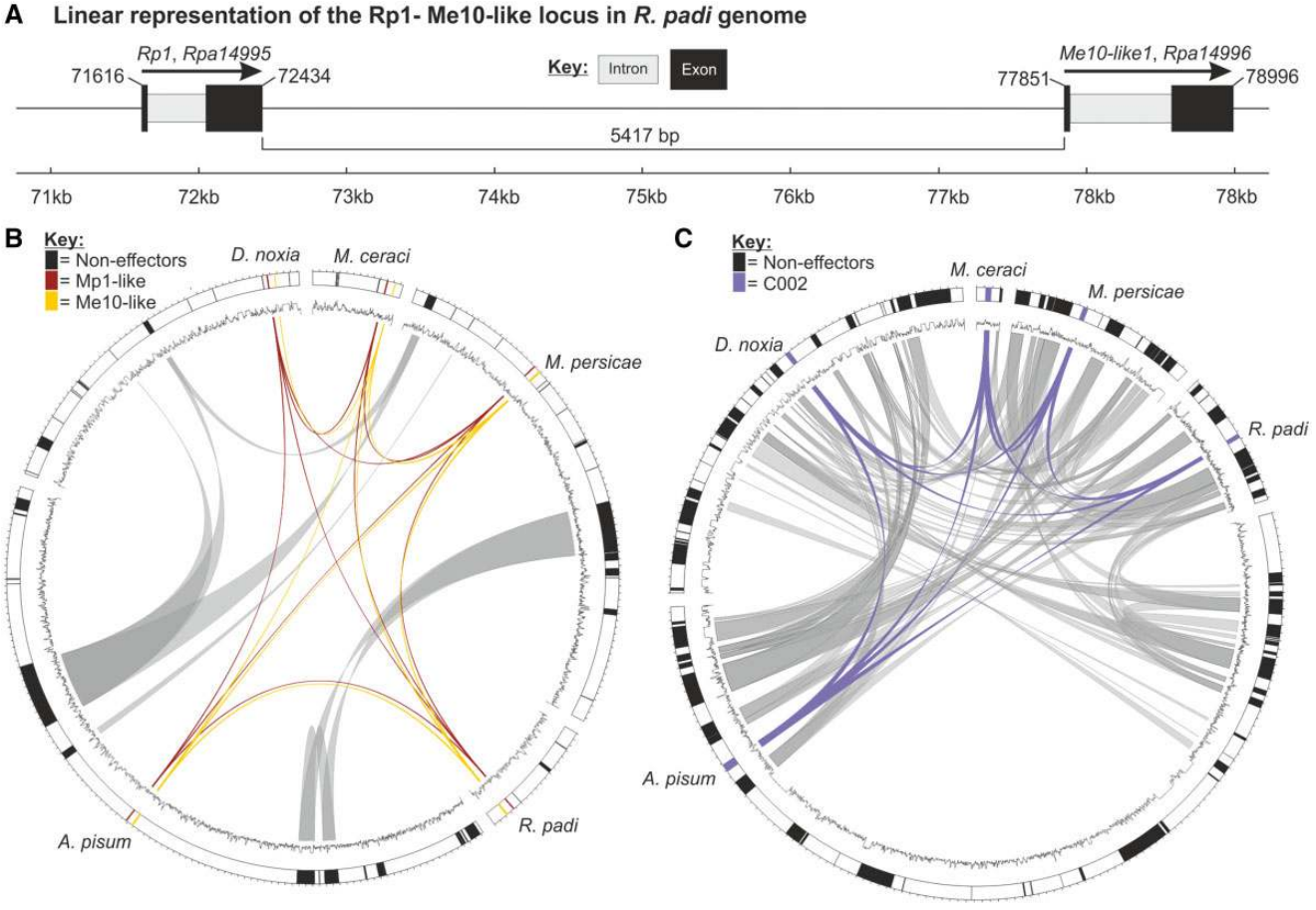

Fig. 6.—Tight genomic association of *Mp1*- and *Me10-like* effectors. (*A*) A linear representation of the Rp10:Me10-like pair in the genome assembly of *R. padi*. (*B* and *C*) Circos plots showing selected scaffolds/contigs of the *R. padi, D. noxia, A. pisum, M. persicae*, and *M. cerasi* genome assemblies. (*B*) The characterized effectors Mp1-like and Me10-like are conserved as an adjacent pair in all species. Every neighboring gene and transposon is different, in all genome assemblies. (*C*) The characterized effector C002-like is present in a syntenic block that is largely conserved in all species as evident by the gray lines in the plot.

2009; Rouxel et al. 2011; Eves-van den Akker et al. 2016). However, the classical signature of a close genetic association of putative effectors and transposable elements, as reported for oomycetes, nematodes and fungi, does not manifest in the aphid genomes (Haas et al. 2009; Rouxel et al. 2011; Eves-van den Akker et al. 2016). With the exception of *A. pisum*, putative effectors are actually further from their nearest transposable element in both the 3′ and the 5′ direction when compared with the remainder of the genes in the genome ($P < 0.01$ and $P < 0.05$, respectively, Mann–Whitney *U* test; fig. 5*C*).

The presence of effectors in less-gene dense regions of the genome is hypothesized to coincide with regions of high mutability, thus providing a means for rapid evolution of genes under high selection pressure from the plant host immune system (Dong et al. 2015). Consistent with this, we identified 30 orthologous gene clusters, containing 170 putative effectors, as being under diversifying selection (DN/DS >1.0; supplementary table S5, Supplementary Material online), notably including well-characterized effectors C002 (DN/DS = 2.35) and Me10-like (DN/DS = 1.60) (Thorpe et al. 2016).

Physical Linkage of an Effector Pair across Five Different Aphid Genomes

We initially noted that putative orthologs of the previously characterized effectors Me10 (Atamian et al. 2013) and Mp1 (Bos et al. 2010; Pitino and Hogenhout 2013; Rodriguez et al. 2017) (here referred to as *Me10-like* and *Rp1*) are tightly physically linked in the *R. padi* genome. These two genes are present in a head-to-tail orientation, 5,417 bp from the end of the first (*Rp1*, *Rpa14995*) to the start of the last (*Me10-like*, *Rpa14996*, fig. 6*A*). This same physical linkage, and an identical genomic organization, is conserved in all five aphid species (fig. 6*B*). Remarkably, genes and transposons adjacent to this effector pair are different in

every aphid species, indicating a lack of synteny in the corresponding genomic regions. Effector gene colocation does appear to be a feature of effectors, albeit not universal: Effector *COO2* is present in a large syntenic block of noneffector loci conserved in all aphids (fig. 6*C*), while 25.8% of *R. padi* putative effectors have another putative effector as an adjacent genomic neighbor (cf. ~3% expected by chance). We observed that the promoter region of the 5′ gene of each of the *Mp1–Me10-like* pair is highly similar and may be indicative of shared transcriptional control (supplementary fig. S5, Supplementary Material online).

Shared Transcriptional Control of Predicted Aphid Effectors

To assess the role of aphid transcriptional plasticity to aphid interactions with host versus nonhost plant species, and determine whether effectors (in particular, the *Mp1–Me10-like* pair) are potentially coregulated, we sequenced the transcriptomes of *R. padi* and *M. persicae* after feeding on an artificial diet for 3 or 24 h, a host plant for 3 or 24 h, and a nonhost plant for 3 or 24 h (each with five replicates prepared in environment-controlled growth cabinets, conducted at the same time of day on sequential days). Transcript abundance was quantified, and differential expression analyses were performed to identify aphid genes differentially expressed across the different aphid treatments (host, nonhost plant or diet). Cluster analyses of the aphid transcriptional responses from this and previous work reporting on differential aphid gene expression in head versus body tissues (Thorpe et al. 2016) revealed only limited variation across different treatments. Specifically, while we observed a clear difference in gene expression patterns between samples collected from aphid head versus body tissues, overall expression profiles corresponding to aphids collected from host, nonhost plants or diets were largely indistinguishable (supplementary figs. S6 [Rp], S7 [Mp], S8 [new data sets only] and supplementary table S6, Supplementary Material online). Moreover, differential expression analyses (FDR < 0.01, 2-fold change), which were performed only using the new data sets generated in this study, revealed low numbers of aphid differentially expressed genes across treatments (supplementary table S6, Supplementary Material online). These data suggest that aphids show only a limited transcriptional response in the first 24 h upon transfer to a host, nonhost plant or artificial diet. The limited changes in gene expression are consistent with data presented by Mathers et al. (2017), where only a small set (171) of differentially expressed genes were found in *M. persicae* adapted to different host plants, based on a >1.5-fold change, 10% FDR,

We made use of the RNAseq generated here as well as in our previous work to determine whether the physically linked effector pair, Me10-like and Mp1-like, is under shared transcriptional control. For this, we focused on the aphid species *R. padi* and *M. persicae* based on the availability of genome, proteome, and transcriptome data sets (Thorpe et al. 2016; this work). In the case of *R. padi*, expression *Rp1/Me10-like* gene pair was almost perfectly correlated (fig. 7*A*): Measuring variation in the expression of *Me10-like* describes 99% of the variation in expression of *Rp1* (R^2 = 0.99, fig. 7*A*). No such correlation was observed when comparing the expression of *Rp1* with its adjacent noneffector gene in the opposite direction (Rpa14994, R^2 = 0.06, fig. 7*A*). We identified five other pairs of effector genes that are adjacent in all aphid species, but their expression did not correlate to the same extend as the *Rp1/Me10-like* pair (supplementary fig. S9, Supplementary Material online), and these did not necessarily share the same orientation. Similarly, in *M. persicae* and *M. cerasi* expression of the Mp1(-like)/Me10-like gene pair was correlated (supplementary fig. S10, Supplementary Material online). The fact that the genetic linkage of the effector pair has persisted throughout evolution in spite of considerable local rearrangements, coupled with shared transcriptional control, is strongly indicative of functional linkage.

Using the physically linked Rp1/Me10-like and Mp1/Me10-like effector pair, we sought to identify other genes that are similarly transcriptionally, but not physically linked, in the *R. padi* and *M. persicae* genomes. Although our transcriptome data of aphids exposed to different plants or diet did not reveal patterns of distinct transcriptional responses, there were nevertheless 213 loci, in the case of *R. padi*, that mirrored the co-regulation of the Rp1/Me10-like effector pair with a Pearson's correlation of >90% (fig. 7*B*). Similarly, for *M. persicae* we identified 114 loci showing tight coregulation with the Mp1/Me-10-like pair (supplementary fig. S10, Supplementary Material online). Of the 213 *R. padi* loci, 32% were predicted to encode secretory proteins (a 4.5-fold enrichment over the remainder of the genome). Of these 69, 71% were already predicted to be effectors ($n = 49$, fig. 7*C*). Of these 49 effectors, 36% were present in gene clusters specific to *R. padi* ($n = 18$). Taken together, this suggests that with just two criteria, 1) concerted expression with a highly conserved effector pair and 2) the presence of a signal peptide for secretion, a 71% accuracy of effector identification can be achieved. These predictions work similarly, albeit to a lesser extent in *M. persicae* (57% were predicted secreted, 16% of which are already predicted to be effectors). Remarkably, of the 213 that correlate >90% with the *Rp1:Me10-like* pair that are not predicted to encode a secretion signal, 46 have been detected in the saliva of *R. padi* using a proteomics approach (Thorpe et al. 2016). This is a substantial proportion of all proteins detected in the saliva of *R. padi* (30%), is numerically more than those with a classical signal peptide for secretion, and may question the suitability of canonical secretory protein prediction pipelines. Concerted expression of effectors has been reported in a plant-pathogenic fungus, and likely relies on an epigenetic control

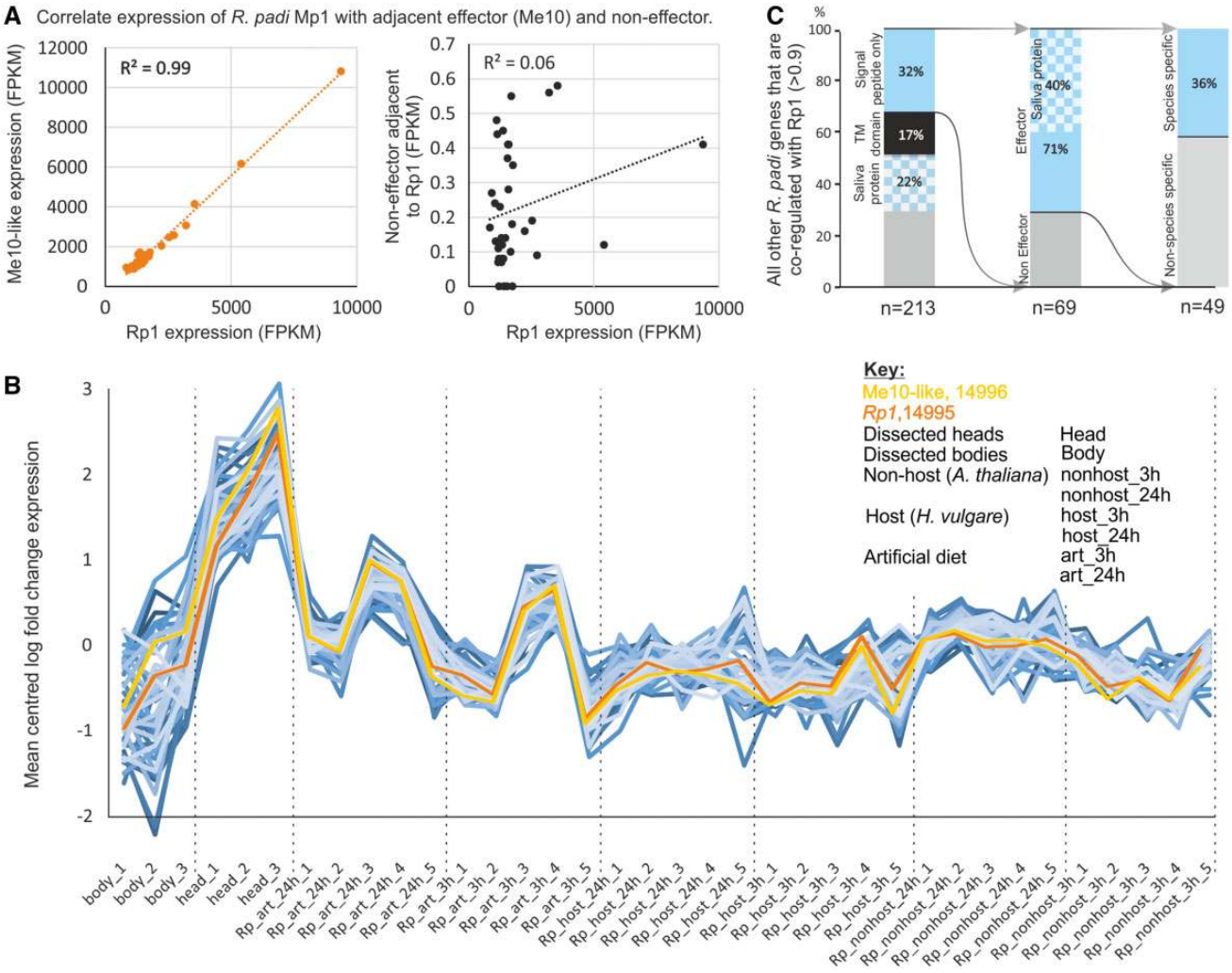

Fig. 7.—Shared transcriptional control of a subset of the effector repertoire. (*A*) Correlating the normalized RNAseq expression of the *Rp1* effector with its adjacent effector *Me10-like* (orange) reveals almost perfect concerted expression across a range of diverse stimuli (supplementary fig. S7, Supplementary Material online). No such correlation is observed with the adjacent noneffector (black). (*B*) Identification of all other genes in the *R. padi* genome that are coregulated with the Rp1:Me10-like pair based on a >90% Pearson's correlation (blue, $n = 213$). (*C*) Of the 213 genes in the *R. padi* genome that are coregulated with the Rp1:Me10-like pair, 32% encode a signal peptide ($n = 69$, blue first bar), 17% encode a transmembrane domain (black first bar), and 22% were detected in the salivary proteomics (blue and white checker box first bar). Of the 32% that encode a signal peptide, 71% are predicted to be effectors ($n = 49$, blue second bar). Of these 71%, 36% are present in MCL clusters that exclude all other aphid species ($n = 18$, blue third bar).

mechanism (Soyer et al. 2014). Whether epigenetic control is also responsible for the tight coregulation of a significant subset of aphid effectors remains to be elucidated.

Conclusions

In this study, we reveal a complex history of ancient gene duplication and relatively recent gene birth in aphids. We identified several aphid effector pairs that are physically linked across aphid species genomes, one of which also showed tight coregulation of transcription with a substantial subset of putative effectors. Exploiting transcriptional linkage for utility, we develop a series of criteria to expand the putative effector repertoire of aphids, and potentially implicate nonclassical secretion in aphid parasitism.

Data Access

Assembled genomes and gene calls are available at http://bipaa.genouest.org/is/aphidbase/ and doi:10.5281/zenodo.1252934. The raw genomic reads for *M. cerasi* and *R. padi* are available at study accession numbers PRJEB24287 and PRJEB24204, respectively. The raw RNAseq reads for *R. padi* and *M. persicae* reared on host, nonhost and artificial diet at time points 3 h and 24 h are available at study accession number PRJEB24317. RNAseq data for *M. cerasi* used to

assist gene call is available at study accession number PRJEB24338. Most custom python scripts used to analyze the data use Biopython, all scripts and methods used throughout this study are available at https://github.com/peterthorpe5/Methods_M.cerasi_R.padi_genome_assembly and https://github.com/sebastianevda/SEvdA_Gephi_array_to_gefx

Supplementary Material

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

Acknowledgments

We thank Brian Fenton and Gaynor Malloch for kindly providing the *Rhopalosiphum padi* and *Myzus persicae* aphids and for support in generating a *Myzus cerasi* stock culture. We also thank Brian Fenton for useful discussion on results, Dr Matt Clark (Earlham Institute) for conceptual design of the sequencing approach, and Dr Dominik Laetsch for advice on using BlobTools. This work was supported by the Biotechnology and Biological Sciences Research Council (BB/M014207/1 to S.E.V.D.A.), European Research Council (310190-APHIDHOST to J.I.B.B.), and Royal Society of Edinburgh (fellowship to J.I.B.B.).

Author Contributions

P.T. carried out aphid culture maintenance, collected and prepared material for DNAseq analysis, genome assemblies, annotation, transposons prediction, effector prediction, RNAseq analysis, and comparative genomic analysis. C.M.E.M. carried out aphid culture maintenance, collected and prepared material for RNAseq analysis. P.T. and S.E.V.D.A. analyzed the data. P.T., S.E.V.D.A., and J.I.B.B. wrote the manuscript. J.I.B.B. conceived and coordinated the project, secured the funding, and contributed to the analyses of data. P.J.A.C. assisted and advised on python scripts and conceptual design. All authors read and approved the final manuscript.

Literature Cited

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Atamian HS, et al. 2013. In planta expression or delivery of potato aphid *Macrosiphum euphorbiae* effectors Me10 and Me23 enhances aphid fecundity. Mol Plant Microbe Interact. 26(1):67–74.

Aziz RK, et al. 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9:75.

Bailey TL, et al. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37(Web Server issue):W202–W208.

Bankevich A, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 19(5):455–477.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6(1):1.

Bastian M, Heymann S, Jacomy M. 2009. Gephi: an open source software for exploring and manipulating networks. ICWSM 8:361–362.

Blackman RL, Eastop VF. 2000. Aphids on the world's trees—an identification and information guide. Wallingford, UK: CABI.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.

Bos JIB, et al. 2010. A functional genomics approach identifies candidate effectors from the aphid species *Myzus persicae* (green peach aphid). PLoS Genet. 6(11):e1001216.

Boulain H, et al. 2018. Fast evolution and lineage-specific gene family expansions of aphid salivary effectors driven by interactions with host-plants. Genome Biol Evol. 10(6):1554–1572.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 34(5):525–527.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 12(1):59–60.

Cantarel BL, et al. 2007. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18(1):188–196.

Carolan JC, et al. 2011. Predicted effector molecules in the salivary secretome of the pea aphid (*Acyrthosiphon pisum*): a dual transcriptomic/proteomic approach. J Proteome Res. 10(4):1505–1518.

Chevreux B. 2005. MIRA: an automated genome and EST assembler. Heidelberg (Germany): Ruprecht-Karls University.

Cock PJ, Chilton JM, Grüning B, Johnson JE, Soranzo N. 2015. NCBI BLAST+ integrated into Galaxy. GigaScience 4(1):1–7.

Cock PJ, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25(11):1422–1423.

Conesa A, et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18):3674–3676.

De Bie T, Cristianini N, Demuth JP, Hahn MW. 2006. CAFE: a computational tool for the study of gene family evolution. Bioinformatics 22(10):1269–1271.

Dobin A, et al. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15–21.

Dong S, Raffaele S, Kamoun S. 2015. The two-speed genomes of filamentous pathogens: waltz with plants. Curr Opin Genet Dev. 35:57–65.

Duncan RP, Feng H, Nguyen DM, Wilson AC. 2016. Gene family expansions in aphids maintained by endosymbiotic and nonsymbiotic traits. Genome Biol Evol. 8(3):753–764.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792–1797.

Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9(1):18.

Elzinga DA, De Vos M, Jander G. 2014. Suppression of plant defenses by a *Myzus persicae* (green peach aphid) salivary effector protein. Mol Plant Microbe Interact. 27(7):747–756.

Elzinga DA, Jander G. 2013. The role of protein effectors in plant-aphid interactions. Curr Opin Plant Biol. 16(4):451–456.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30(7):1575–1584.

Escudero-Martinez CM, Morris JA, Hedley PE, Bos JIB. 2017. Barley transcriptome analyses upon interaction with different aphid species identify thionins contributing to resistance. Plant Cell Environ. 40(11):2628–2643.

Eves-van den Akker S, et al. 2016. The genome of the yellow potato cyst nematode, *Globodera rostochiensis*, reveals insights into the basis of parasitism and virulence. Genome Biol. 17(1):124.

Field D, et al. 2006. Open software for biologists: from famine to feast. Nat Biotechnol. 24(7):801–804.
Finn RD, et al. 2013. Pfam: the protein families database. Nucleic Acids Res. 42: 222–230.
Flot J-F, et al. 2013. Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. Nature 500(7463):453–457.
Gil M, Zanetti MS, Zoller S, Anisimova M. 2013. CodonPhyML: fast maximum likelihood phylogeny estimation under codon substitution models. Mol Biol Evol. 30(6):1270–1280.
Girousse C, Moulia B, Silk W, Bonnemain JL. 2005. Aphid infestation causes different changes in carbon and nitrogen allocation in alfalfa stems as well as different inhibitions of longitudinal and radial expansion. Plant Physiol. 137(4):1474–1484.
Gladyshev EA, Meselson M, Arkhipova IR. 2008. Massive horizontal gene transfer in bdelloid rotifers. Science 320(5880):1210–1213.
Haas B. 2007. TransposonPSI: an application of PSI-Blast to mine (retro-)transposon ORF homologies. http://transposonpsi.sourceforge.net.
Haas BJ, et al. 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. Nature 461(7262):393–398.
Haas BJ, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 8(8):1494–1512.
Hahn MW, Han MV, Han SG. 2007. Gene family evolution across 12 *Drosophila* genomes. PLoS Genet. 3(11):e197.
Harmel N, et al. 2008. Identification of aphid salivary proteins: a proteomic investigation of *Myzus persicae*. Insect Mol Biol. 17(2):165–174.
Heger A, Ponting CP. 2007. Evolutionary rate analyses of orthologs and paralogs from 12 *Drosophila* genomes. Genome Res. 17(12):1837–1849.
Heinz S, et al. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 38(4):576–589.
Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2015. BRAKER1: unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32:767–769.
IAGC. 2010. Genome sequence of the pea aphid *Acyrthosiphon pisum*. PLoS Biol. 8(2):e1000313.
Jaouannet M, et al. 2014. Plant immunity in plant-aphid interactions. Front Plant Sci. 5:663.
Jaouannet M, Morris JA, Hedley PE, Bos JI. 2015. Characterization of *Arabidopsis* transcriptional responses to different aphid species reveals genes that contribute to host susceptibility and non-host resistance. PLoS Pathog. 11(5):e1004918.
Kikuchi T, Eves-van den Akker S, Jones JT. 2017. Genome evolution of plant-parasitic nematodes. Annu Rev Phytopathol. 55:333–354.
Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5:59.
Krzywinski M., et al. 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19(9):1639–1645.
Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M. 2013. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. Front Genet. 4:237.
Lawrence JG, Ochman H. 1997. Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol. 44(4):383–397.
Lilley CJ, et al. 2018. Effector gene birth in plant parasitic nematodes: neofunctionalization of a housekeeping glutathione synthetase gene. PLoS Genet. 14(4):e1007310.
Lomsadze A, Burns PD, Borodovsky M. 2014. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. 42(15):e119.
Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33(20):6494–6506.
Mathers TC, et al. 2017. Rapid transcriptional plasticity of duplicated gene clusters enables a clonally reproducing aphid to colonise diverse plant species. Genome Biol. 18(1):27.
Moran NA, Jarvik T. 2010. Lateral transfer of genes from fungi underlies carotenoid production in aphids. Science 328(5978):624–627.
Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35(Web Server issue):W182–W185.
Nicholson SJ, et al. 2015. The genome of *Diuraphis noxia*, a global aphid pest of small grains. BMC Genomics 16(1):1.
Nikoh N, et al. 2010. Bacterial genes in the aphid genome: absence of functional gene transfer from *Buchnera* to its host. PLoS Genet. 6(2):e1000827.
Nováková E, Moran NA. 2012. Diversification of genes for carotenoid biosynthesis in aphids following an ancient transfer from a fungus. Mol Biol Evol. 29(1):313–323.
Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23(9): 1061–1067.
Pitino M, Hogenhout SA. 2013. Aphid protein effectors promote aphid colonization in a plant species-specific manner. Mol Plant Microbe Interact. 26(1):130–139.
Price DR, Duncan RP, Shigenobu S, Wilson AC. 2011. Genome expansion and differential expression of amino acid transporters at the aphid/*Buchnera* symbiotic interface. Mol Biol Evol. 28(11):3113–3126.
Puinean AM, et al. 2010. Amplification of a cytochrome P450 gene is associated with resistance to neonicotinoid insecticides in the aphid *Myzus persicae*. PLoS Genet. 6(6):e1000999.
Quevillon, E., et al. 2005. InterProScan: protein domains identifier. Nucleic Acids Res. 33:W116–120.
Rancurel C, Legrand L, Danchin EGJ. 2017. Alienness: rapid detection of candidate horizontal gene transfers across the tree of life. Genes (Basel) 8(10):248.
Rao SA, Carolan JC, Wilkinson TL. 2013. Proteomic profiling of cereal aphid saliva reveals both ubiquitous and adaptive secreted proteins. PLoS One 8(2):e57413.
Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139–140.
Rodriguez PA, Bos JI. 2013. Toward understanding the role of aphid effectors in plant infestation. Mol Plant Microbe Interact. 26(1):25–30.
Rodriguez PA, Escudero-Martinez C, Bos JI. 2017. An aphid effector targets trafficking protein VPS52 in a host-specific manner to promote virulence. Plant Physiol. 173(3):1892–1903.
Rouxel T, et al. 2011. Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. Nat Commun. 2:202.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210–3212.
Smadja C, Shi P, Butlin RK, Robertson HM. 2009. Large gene family expansions and adaptive evolution for odorant and gustatory receptors in the pea aphid, *Acyrthosiphon pisum*. Mol Biol Evol. 26(9):2073–2086.
Smit A, Hubley R. 2014. RepeatModeler Open-1.0. 2008–2010. Available online at http://www.repeatmasker.org
Smit A, Hubley R, Green P. 2014. RepeatMasker Open-4.0. 2013–2015. Available online at http://www.repeatmasker.org
Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. 2016. TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res. 26(8):1134–1144.
Soyer JL, et al. 2014. Epigenetic control of effector gene expression in the plant pathogenic fungus *Leptosphaeria maculans*. PLoS Genet. 10(3):e1004227.

Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19(Suppl 2): ii215–ii225.
Thorpe P, Cock PJ, Bos J. 2016. Comparative transcriptomics and proteomics of three different aphid species identifies core and diverse effector sets. BMC Genomics 17(1):1.
Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ. 2016. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res. 44(W1):W232–W235.
Wang Y, et al. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40(7):e49.
Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. 2009. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics 25(9):1189–1191.
Weisenfeld NI, et al. 2014. Comprehensive variation discovery in single human genomes. Nat Genet. 46(12):1350–1355.
Wenger JA, et al. 2017. Whole genome sequence of the soybean aphid, *Aphis glycines*. Insect Biochem Mol Biol.
Will T, Steckbauer K, Hardt M, van Bel AJ. 2012. Aphid gel saliva: sheath structure, protein composition and secretory dependence on stylet-tip milieu. PLoS One 7(10):e46903.
Will T, Tjallingii WF, Thönnessen A, van Bel AJ. 2007. Molecular sabotage of plant defense by aphid saliva. Proc Natl Acad Sci U S A. 104(25):10536–10541.
Wilson AC, Sternberg LDS, Hurley KB. 2011. Aphids alter host-plant nitrogen isotope fractionation. Proc Natl Acad Sci U S A. 108(25): 10220–10224.
Zaharia M, et al. 2011. Faster and more accurate sequence alignment with SNAP. arXiv preprint arXiv:1111.5572.

Associate editor: Nancy Moran